Detection and identification of papillomavirus sequences in NGS data of human DNA samples : a bioinformatic approach

Alexis Robitaille

N°d'ordre NNT : 2019LYSE1358

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale ED 340
BIOLOGIE MOLÉCULAIRE INTÉGRATIVE ET CELLULAIRE (BMIC)

Spécialité de doctorat : Bioinformatique

Soutenue publiquement le 18/12/2019, par :
Alexis Robitaille

Detection and identification of papillomavirus sequences in NGS data of human DNA samples: a bioinformatic approach

---

Devant le jury composé de :

Legras-Lachuer Catherine, Professeure des Universités, Université Lyon	Présidente
Parish Joanna, Directrice de Recherche, Université de Birmingham	Rapporteure
Montanini Barbara, Professeure Associé, Université de Parme	Rapporteure
Legras-Lachuer Catherine, Professeure des Universités, Université Lyon	Examinatrice
Grundhoff Adam, Directeur de Recherche, Henrich Pette Institute	Examinateur
Tommasino Massimo, Directeur de Recherche/Chef de Section, IARC	Directeur de thèse
Olivier Magali, Chercheur, IARC	Co-directrice de thèse

Laboratory

Infection and Cancer Biology Group (ICB)
Section of Infections (INF)
International Agency for Research on Cancer (IARC)
World Health Organization (WHO)
150 cours Albert Thomas
69372 Lyon CEDEX 08
FRANCE

Résumé

Les papillomavirus humains (HPV) constituent une famille de petits virus à double brin d'ADN qui ont un tropisme pour les cellules épithéliales de la peau et des muqueuses. Plus de 200 types d'HPV ont été découverts, et classifiés en plusieurs genres taxonomiques en fonction de la constitution de leur séquence ADN. De part le rôle de certains HPV dans les maladies affectant les humains, allant de l'apparition de verrues anogénitales bénignes jusqu'au développement d'un cancer, il est nécessaire de développer des méthodes de détection et de caractérisation de la population d'HPV dans un échantillon d'ADN. Elles sont nécessaires à la clarification du rôle de l'HPV dans les différentes étapes de la progression de la maladie. Cette détection d'HPV lors d'approches ciblées en laboratoire a principalement reposé sur des méthodes de PCR couplées avec du séquençage Sanger. Avec l'introduction des nouvelles technologies de séquençage haut débit (NGS), ces approches peuvent être revisitées afin d'intégrer la puissance de séquençage de ces technologies. Alors que des outils d'analyse *in-silico* ont été développés pour la recherche de virus, connus ou nouveaux, à partir de données de NGS, aucun outil approprié n'est disponible pour la classification et l'identification de nouvelles séquences virales à partir de données produites par des méthodes de séquençage d'amplicons. Dans cette thèse, la première partie présente cinq nouveaux génomes d'HPV isolés via l'utilisation d'amorces d'amplification dégénérées ciblant le gène L1 à partir d'échantillons de peau humaine. Puis, dans une seconde partie, nous présentons **PVAmpliconFinder**, un outil d'analyse de données conçu pour identifier et classifier rapidement des séquences connues et potentiellement nouvelles de la famille *Papillomaviridae,* à partir de données de NGS d'amplicons générées par PCR via l'utilisation d'oligonucleotides dégénérés ciblants les HPV. Enfin, les caractéristiques de **PVAmpliconFinder** sont présentées**,** ainsi que plusieurs applications sur des données biologiques obtenues lors du séquençage d'amplicons de spécimens humains. Ces applications ont permis la découverte de nouveaux types d'HPV.

Abstract

Human Papillomaviruses (HPV) are a family of small double-stranded DNA viruses that have a tropism for the mucosal and cutaneous epithelia. More than 200 types of HPV have been discovered so far and are classified into several genera based on their DNA sequence. Due to the role of some HPV types in human disease, ranging from benign anogenital warts to cancer, methods to detect and characterize HPV population in DNA sample have been developed. These detection methods are needed to clarify the implications of HPV at the various stages of the disease. The detection of HPV from targeted wet-lab approaches has traditionally used PCR-based methods coupled with cloning and Sanger sequencing. With the introduction of next generation sequencing (NGS) these approaches can be improved by integrating the sequencing power of NGS. While computational tools have been developed for metagenomic approaches to search for known or novel viruses in NGS data, no appropriate bioinformatic tool has been available for the classification and identification of novel viral sequences from data produced by amplicon-based methods. In this thesis, we initially describe five fully reconstructed novel HPV genomes detected from skin samples after amplification using degenerate L1 primers. Then, in the second part, we present **PVAmpliconFinder**, a data analysis workflow designed to rapidly identify and classify known and potentially new *Papillomaviridae* sequences from NGS amplicon sequencing with degenerate PV primers. This thesis describes the features of **PVAmpliconFinder** and presents several applications using biological data obtained from amplicon sequencing of human specimens, leading to the identification of new HPV types.

Acknowledgments

At first place, I wish to thank Dr. Massimo Tommasino and Dr. Magali Olivier, my supervisors during this three years of PhD at the IARC. I have special thanks for Magali, who have initially been my Master 2 internship supervisor, providing me with her trust, and proposing my candidacy to be a PhD student in Massimo's group in 2016. I'm grateful to Massimo who have in turn trust me, allowing me to join his team and realize my PhD in his lab. Both of you brought me lot during these years, and have actively participated not only in my professional development as a scientist, but also on my personal growing, and I will always be grateful to you.

I have a special thank for Massimo, who have been a continuously supportive supervisor, and greatly improved my knowledge thanks to plenty of insightful discussions. You gave me responsibilities that went over my duties, and I'm really thankful because even if scary at first place, it has permit me to grow faster and better, and to understand many facets of a scientific career.

I would like to express my sincere gratitude to Magali, who not only recommended me to Massimo, but also trust me at first place as a Master 2 student. During this 6 months at MMB, it has sometimes been hard time, but you have always been patient and took time to repeat many times your advices until I applied them. This period bring me the rigor and the motivation that I since put every day into my work. Moreover, you have constantly been available to answer my many questions, and I thank you a lot for your patience.

I would like to thank ICB members that have particularly been always encouraging and sources of precious guidance. I'm thinking here of Dr. Tarik Gheit who have been a really important figure during this three years of PhD, always available to answer any single questions that I had. Many thanks to you, Tarik. I'm also thinking of the other ones that were available to answer my numerous questions, especially Shankadeep, Maria Grazia, Assunta, Rosario and Maria. Thanks a lot, it has always been really great to work with you.

I'm also grateful to all the other MMB and ICB members I met during my time at IARC. You have created a great scientific emulsion, continuously pushing me forward.

I have special thought for the others bioinformaticians in the Agency, and I wish to specifically express my gratitude to two of them: Vincent Cahais for being always willing to share his experience with me, and Dr. Matthieu Foll for his help as a bioinformatic leader in the agency, organizing the unseminar, proposing courses and implying the different bioinformatician together in many scientific projects. You both have make me feel less isolated during this years as the only ICB bioinformatician and I'm grateful for this.

I wish to say thank you to many colleagues in the Agency, and I cannot be exhaustive, but Nicole Suty and Sylvie Nouveau deserve many of my acknowledgement for their huge support, on administration-related issues, and on many other facets.

I have also a special thank for Stephen Martin, who took time to proofread this manuscript. Thanks a lot, you have bring a lot of value to my work, and greatly improved my English writing and speaking skills this last years.

I thank the thesis committee members for their guidance during this years, and especially James McKay, who has been available at short notice to provide me with his advice, twice.

I'm also thinking of Nicole Fisher and Adam Grundhoff, that have welcome me in Hamburg for few months and make me discover their lab and their team. It was a great time to be there with you, and I have many thanks for this opportunity you offered me, and for the many scientific discussions we had.

I would like to recognize the member of my PhD jury for accepting reviewing my work. I hope that this project will bring its contribution to the Science.

As importantly as the scientific support I receive during my PhD, I'm really grateful to my parents. Being a PhD students is also a lot of personal development, and some periods are easier than others: knowing that you will always be unconditionally supportive and keep expressing your confidence on my success has been the best motivation I ever had. This has always been the case since my first day of school, and you're the basis of all my accomplishments. I'm also really grateful to my sisters and to Jean-Baptiste, always cheering me up and encouraging me throughout this experience.

Finally, I wish to thanks her for her constant love, dedication, patience and support.

Résumé en français

Les papillomavirus humains (HPV) constituent une famille de petits virus à double brin d'ADN qui ont un tropisme pour les cellules épithéliales de la peau et des muqueuses. Plus de 200 types d'HPV ont été découverts, et classifiés en plusieurs genres taxonomiques en fonction de la constitution de leur séquence ADN. Certains HPV possèdent un génotype qui les rend responsables de l'apparition et du développement de pathologies affectant les humains, allant de l'apparition de verrues anogénitales bénignes jusqu'au développement d'un cancer. Dans la majorité des cas, l'infection est transitoire car le système immunitaire permet d'éliminer spontanément les virus. Néanmoins, un groupe de 12 HPV infectant les muqueuses a été défini comme «à haut risque» de par leur association avec le développement du cancer du col de l'utérus. HPV16 et HPV18 sont responsables de 50 % et 20 %, respectivement, du nombre total de cas de cancer du col de l'utérus dans le monde, et cette pathologie est la cause de plus de 236 000 morts dans le monde chaque année. Les HPV étant particulièrement contagieux, on estime que près de 80 % des hommes et des femmes sexuellement actifs ont été au moins une fois en contact avec un HPV infectant les muqueuses. De par leur pathogénicité, il est nécessaire de développer des méthodes de détection et de caractérisation de la population d'HPV dans un échantillon d'ADN. Elles sont nécessaires à la clarification du rôle de l'HPV dans les différentes étapes de la progression de la maladie. Cette détection d'HPV lors d'approches ciblées en laboratoire a principalement reposé sur des méthodes de PCR. Ces PCR sont effectuées à partir d'amorces construites dans des régions génomiques conservées entre les HPVs, et contiennent souvent des oligonucléotides dégénérés afin d'augmenter la sensibilité de la détection. Initialement, ces méthodes de détection ont été développées afin d'amplifier les HPV mucosaux, et dans le même temps la connaissance sur la diversité des HPV s'est élargie, menant à l'établissement de nouveaux couples d'amorces ciblant les HPV cutanés. L'élaboration de ces nouvelles amorces d'amplifications a été motivée par la découverte d'associations entre les HPV cutanés et l'apparition de maladies. Ainsi, les HPV cutanés du genre taxonomique β-3 ont été associés avec le développement de carcinomes épidermoïdes, en particulier lors d'une co-exposition avec un autre agent carcinogène. Par exemple l'exposition aux UV de l'épiderme de rongeurs infectés par les oncogènes E6 et E7 de HPV38 favorise l'apparition de cancer de la

peau. Par ailleurs, il semblerait que l'infection ne soit nécessaire qu'aux stades précoces du développement du cancer, et aide l'accumulation de mutations somatiques causées par les UV. Ce phénomène, dénommé «frapper et courir», met en exergue l'importance de la sensibilité de la détection des papillomavirus chez les individus immunocompétents, pour lesquels la charge virale peut être très faible. Suivant leur détection, la reconstruction de nouveaux génomes d'HPV a originairement été possible via le clonage du nouveau génome dans un vecteur de type plasmide, grâce à l'utilisation d'amorces spécifiques de la nouvelle cible, construites dans la région initialement amplifiée par les amorces non-spécifiques. Ce clonage est communément suivi d'un séquençage Sanger afin d'obtenir la composition nucléotidique du génome au complet. L'ensemble de ce protocole est long et nécessite beaucoup de labeur. Avec l'introduction des nouvelles technologies de séquençage haut débit (NGS), ces approches peuvent être revisitées afin d'intégrer la puissance de séquençage de ces technologies. Pour cela, des méthodes d'analyses spécifiques de ces données doivent être mises en place, dans l'objectif précis de la découverte et la caractérisation de nouveaux HPV. Alors que des outils d'analyse *in-silico* ont été développés pour la recherche de virus, connus ou nouveaux, à partir de données de NGS, aucun outil approprié n'est disponible pour la classification et l'identification de nouvelles séquences virales à partir de données produites par des méthodes de séquençage d'amplicons. Ce projet de recherche est axé autour de deux parties : la première partie présente cinq nouveaux génomes d'HPV isolés soit via l'utilisation d'amorces d'amplification dégénérées ciblant le gène L1, soit à partir d'une méthode d'amplification spécifique de l'ADN circulaire. Lors de ces travaux, les amorces d'amplification utilisées sont celles déjà publiées : leur utilisation permettant de confirmer leur usage dans les échantillons de peau humaine. La reconstruction de ces génomes a été effectuée *in-silico* à partir de données de séquençage Sanger, via l'utilisation d'une stratégie nommée «Arpentage chromosomique». Puis, dans une seconde partie, nous présentons **PVAmpliconFinder**, un outil d'analyse de données conçu pour identifier et classifier rapidement des séquences connues et potentiellement nouvelles de la famille *Papillomaviridae,* à partir de données de NGS d'amplicons générées par PCR via l'utilisation d'oligonucléotides dégénérés ciblant les HPV. Dans cette partie, la description de nouveaux oligonucléotides partiellement dégénérés est présentée, puis leur utilisation en combinaison du NGS est décrite, ayant mené au

développement d'un outil d'analyse dédié : **PVAmpliconFinder.** Les caractéristiques de **PVAmpliconFinder** sont présentées, reposant sur des méthodes de similarité et de phylogénie, afin de définir les séquences potentiellement nouvelles d'HPV, mais aussi d'obtenir une inférence de leur classification taxonomique via l'utilisation de l'état des lieux de la connaissance sur la diversité des HPV. Finalement, plusieurs applications sur des données biologiques obtenues lors du séquençage d'amplicons de spécimens humains sont commentées. Ces applications ont permis la découverte de nouveaux types d'HPV.

List of acronyms

AK	Actinic Keratoses
BCC	Basal Cell Carcinomas
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
BWA	Burrows-Wheeler Aligner
cds	Coding DNA Sequence
CODEHOP	Consensus-degenerate hybrid oligonucleotide primers
CpG	5'—C—phosphate—G—3'
cSCC	Cutaneous Squamous Cell Carcinoma
DNA	Deoxyribonucleic Acid
DOCK8	Dedicator of Cytokinesis 8
dsDNA	Double Strand Deoxyribonucleic Acid
EM	Expectation-Maximization algorithm
EPA	Evolutionary Placement Algorithm
EV	Epidermodysplasia verruciformis
FM indexes	Ferragina-Manzini indexes
HMM	Hidden Markov Model
HPV	Human Papillomaviruses
HR	High Risk
HTML	Hypertext Markup Language
HTS	High-Throughput Sequencing
IARC	International Agency for Research on Cancer
ICB	Infection and Cancer Biology Group
ICTV	International Committee for the Taxonomy of Viruses
indels	Insertion and Deletion
INF	Section of Infections
KA	Keratoacanthomas
kb	Kilobase
kDa	Kilo Dalton
LANL	Los Alamos National Laboratory
LCA	Last Common Ancestor

LCR	Long Control Region
LR	Low Risk
LZW	Lempel-Ziv-Welch
MIDs	Multiplex Identifiers
NCBI	National Center for Biotechnology Information
NCR	Non-Coding Region
NGS	Next Generation Sequencing
NMSC	Non-Melanoma Skin Cancer
ORFs	Open Reading Frames
pAE	Early Polyadenylation signal
PaVE	Papillomavirus Episteme
PCR	Polymerase Chain Reaction
PV	Papillomaviruses
QC	Quality Control
qPCR	Quantitative Polymerase Chain Reaction
RaxML	Randomized Axelerated Maximum Likelihood
RCA	Rolling circle amplification
RNA	Ribonucleic Acid
RPM	Reads Per Million
rRNA	Ribosomal Ribonucleic Acid
S phase	Synthesis Phase
SCC	Squamous Cell Carcinoma
SNPs	Single Nucleotide Polymorphisms
sRNA	Small interfering RNA
SV	Structural Variants
SVM	Support Vector Machine
TA	Transit Amplifying cells
tRNA	Transfer ribonucleic acid
URR	Upstream Regulatory Region
UV	Ultraviolet
WGA	Whole Genome Amplification
WGS	Whole Genome Sequencing
WHIM	Warts, Hypogammaglobulinemia, Infections, Myelokathexis syndrome
WHO	World Health Organization

Table of Contents

Introduction

Human Papillomaviruses (HPV) belong to a large family of viruses including hundreds of members, which are classified into several genera based on their deoxyribonucleic acid (DNA) sequence. HPV infections can, depending on the HPV genotype, lead to different clinical outcomes ranging from genital warts to cancer lesions. Based on available biological data, a group of 12 mucosal HPV types have been defined as high-risk (HR) types, having clear evidence of their association with the development of cervical cancer. HPV16 and HPV18, the most carcinogenic types, are responsible for approximately 50% and 20% of all cervical cancers worldwide, respectively. Despite the availability of a vaccine and the implementation of screening programs, cervical cancer remains a public health problem on a global level, causing 236,000 deaths worldwide. Papillomaviruses are particularly contagious and nearly 80% of sexually active men and women are in contact with a mucosal HPV at least once during their lifetime. Contact with infected hands is also a route of cutaneous HPV transmission between individuals. Due to the pathogenicity of some HPV types, it is crucial to develop rapid, specific and sensitive methods to identify and characterize the HPV population in human clinical samples.

I. The *papillomaviridae* family

The *papillomaviridae* family belongs to the phylum *Incertae sedis,* and the taxonomic class and taxonomic order of the same name. *Incertae sedis* refers to a complex taxonomic group in which the relationship between representatives of this taxonomic group is still unknown or poorly defined. It includes organisms as disparate as some fossils (1), or the HeLa cells (2). In 2016, a PV in fish (SaPV1) was characterized, rendering this family of viruses much older than expected, with an emergence 450 million years ago (3). Papillomaviruses (PVs) are small non-enveloped icosahedral viruses, presenting circular double-stranded DNA ranging from 5.7 kb for the smallest genome (3) to 8.6 kb for the longest genome (4). The capsid is constituted of 72 pentameric capsomers that are composed of two structural proteins - L1 (55 kDa in size) and L2 (70 kDa) (**Figure 1**). PVs are widely distributed across vertebrates, and have a tropism for mucosal and cutaneous epithelia of human and other vertebrates

(e.g. mammalians, reptiles, birds, etc.) (5). They are traditionally described as ‘‘types’’ based on their genome sequences and identified by a number provided by the International HPV Reference Center, the Karolinska Institute (https://ki.se/en/labmed/international-hpv-reference-center https://ki.se/en/labmed/international-hpv-reference-center). So far, more than 300 PVs types have been characterized, completely sequenced and referenced, including more than 200 HPVs. In addition, more than 100 PV types have recently been identified using *in-silico* methods from metagenomics data (https://pave.niaid.nih.gov/#explore/reference_genomes). HPV types have been shown to be ubiquitous and widely distributed in the human population, where they can infect various anatomical sites depending on their tropism, and cause lesions with distinctive clinical pathologies. There are five major known HPV genera: ∝-papillomavirus, β-papillomavirus, γ-papillomavirus, mu-papillomavirus and nu-papillomavirus (6). Alpha-PVs preferentially infect the oral or anogenital mucosa, but have also been found in lesions of cutaneous sites, and some members are considered oncogenic in view of their regular presence in malignant tissue. Beta-PVs typically cause latent infections, but can also cause warts or progress to cutaneous squamous cell carcinomas (SCC) in immunocompromised individuals. They are the most common HPV types found in the human oral cavity and some have been prospectively associated with oropharyngeal cancer development (7). Gamma-PVs are typically found on the skin and oral mucosa and their infections are usually asymptomatic. However, some infections cause cutaneous lesions in their host. Mu and Nu-PVs are also associated with cutaneous lesions, which are sometimes malignant but most often only benign.

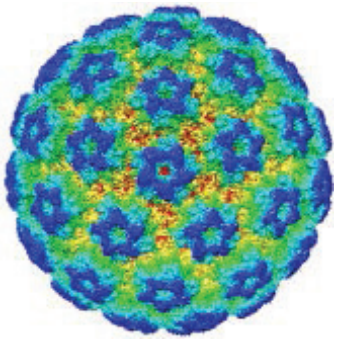


Figure 1: 3D representation of HPV16 particle

Adapted from (8).

I.1 - The genomic structure of PVs

The circular dsDNA genome is approximately 8 kb in size. All PVs share a common genetic structure that generally contains eight ORFs, all transcribed from a single DNA strand. The viral genome can be divided into three functional parts (**Figure 2**):

- An upstream regulatory region (URR), also called the long control region (LCR), containing early promoter and transcription factor binding sites and controlling gene expression, located between the L1 and E6 open reading frames.
- An early region, that encodes for six genes (E1, E2, E4, E5, E6 and E7) involved in multiple functions such as viral gene expression, replication and cell transformation.
- A late region, encoding for two capsid proteins (L1 and L2) which yields to the virion structure.

A smaller non-coding region (denoted by NCR) is also located between the E5 and L2 ORFs and harbors an early polyadenylation signal (pAE) required for gene expression from the early promoter, including alternatively spliced early transcripts and their gene products (9). Both the NCR and the URR viral regions display variability and are useful in assessing genetic heterogeneity (10).

The LCR, and the 4 following ORFs - namely E1, E2, L1 and L2 - are required to ensure viral replication and regulation (11). These 4 proteins are sufficient for the release of progeny virions (12). In addition, PVs have the potential to express an E8^E2C transcript for which the protein acts as a transcriptional repressor of E1/E2-dependent replication of the viral origin (13-16). Although HPV genera have basically the same genome organization, the majority of beta and gamma HPV types lack of the E5 protein (17).

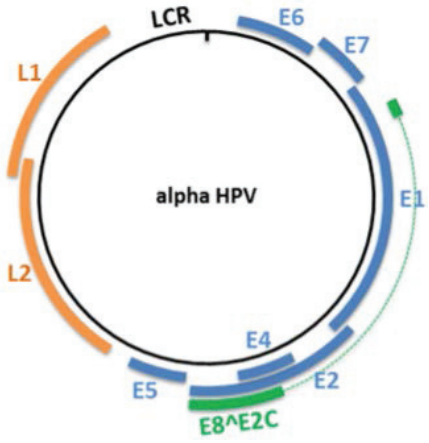


Figure 2: Genomic organization of PVs dsDNA genome

(in black), schematized on HPV16, presenting the LCR, the early genes (in blue), the late genes (in orange) and the E8^E2C transcript (in green). Adapted from (18).

I.2 – PV taxonomic classification

The International Committee for the Taxonomy of Viruses (ICTV; http://ictvonline.org/index.asp)http://ictvonline.org/index.asp defines the rules for the nomenclature and taxonomic classification of PVs. The taxonomy of PVs is based on the nucleotide sequence of the L1 ORF: the nucleotide sequence of L1 must differ by more than 10% of the closest known PV type to define a new type, more than 30% to define a new species, and more than 40% to define a new genus (6, 19). Closely related PVs types based on this classification can present different phenotypes, prevalence and tropisms (20, 21). In particular, HPVs are classified into five genera, namely Alpha, Beta, Gamma, Mu and Nu (22) that are subdivided into species and then into types. The majority of the HPV types of genus alpha have a mucosal tropism, while beta and gamma HPV types appear to preferentially infect the skin (23).

The IARC classifies PV types based on the evidence of their carcinogenicity (**Table 1**). PV carcinogenicity is evaluated based on information from case reports and epidemiological studies, as well as biological data on humans and animal models.

Group 1 Carcinogenic to Humans	**Group 2A** Probably Carcinogenic to Humans	**Group 2B** Possibly Carcinogenic to Humans	**Group 3** Not classifiable	**Group 4** Probably not Carcinogenic to Humans
Sufficient evidence of carcinogenicity in humans and in experimental animals	Limited evidence of carcinogenicity in humans and sufficient evidence of carcinogenicity in experimental animals	Limited evidence of carcinogenicity in humans and insufficient evidence of carcinogenicity in experimental animals	Inadequate evidence of carcinogenicity in humans and in experimental animals	Evidence suggesting lack of carcinogenicity in humans and in experimental animals
111 agents, including 8 biological agents: - Epstein-Barr virus - Helicobacter pylori (infection with) - Hepatitis B virus (chronic infection with) - Hepatitis C virus (chronic infection with) - Human immunodeficiency virus type 1 (infection with) - Human papillomavirus types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 - Human T-cell lymphotropic virus type I -Kaposi sarcoma herpesvirus	**65 agents, including 3 biological agents:** -Human papillomavirus type 68 -Malaria (caused by infection with Plasmodium falciparum in holoendemic areas) -Merkel cell polyomavirus	**274 agents, including 6 biological agents:** -BK polyomavirus -Human immunodeficiency virus type 2 (infection with) -Human papillomavirus types 5 and 8 (in patients with epidermodysplasia verruciformis) -Human papillomavirus types 26, 53, 66, 67, 70, 73, 82 -Human papillomavirus types 30, 34, 69, 85, 97 (Classified by phylogenetic analogy to the HPV genus alpha types classified in Group 1) -JC polyomavirus	**504 agents, including 5 biological agents:** -Human papillomavirus genus beta (except types 5 and 8) and genus gamma -Human papillomavirus types 6 and 11 -Human T-cell lymphotropic virus type II -SV40 polyomavirus -Hepatitis D virus	**1 agent, no biological agent**

Table 1: Classification of PV type carcinogenicity
as defined by the International Agency for Research on Cancer (IARC Monographs vol. 100B and 104) - Adapted from (24).

To be recognized as a novel PV type, in addition to presenting at least 10% dissimilarity to any other PV type on its L1 ORF, a viral genome needs to be referenced and must meet a strict set of requirements defined as followed: the entire viral genome must be cloned, as a whole or as an overlapping fragment, and this cloned genome must be submitted and reviewed by the International Human Papillomavirus Reference Center or the Animal Papillomavirus Reference Center (https://pave.niaid.nih.gov/#explore/taxonomy/taxonomy_concept; https://ki.se/en/labmed/international-hpv-reference-center; http://vandoorslaer.info/Ref).

I.3 – PV life cycle

PVs sustain their life cycle through the keratinocyte proliferation and differentiation process (**Figure 3**). The HPV life cycle can be divided into two stages: non-productive and productive. The non-productive stage occurs in proliferating basal layers of the epithelium where the virus replicates its genome at a low copy number. The productive stage occurs in differentiated layers of the epithelium (25, 26).

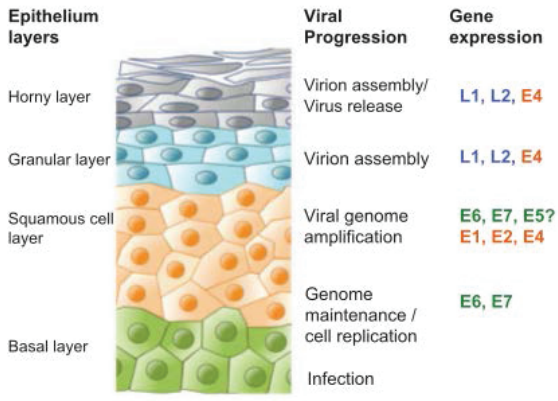
The basal layer of the keratinocyte is initially targeted by PV virions, through microwounds or hair follicles (27-29). Cell surface receptors allow the binding of the viral capsid and facilitate virion entry into the targeted cells, but this is not sufficient to trigger production of virions or to allow the virus to replicate (30-32). The viral particles interact with the cell surface via interaction of the major capsid protein, L1, with heparin sulfate proteoglycans. Evidence also suggests the involvement of a secondary receptor and a possible role for the minor capsid protein, L2, in cell surface interaction (33).

The replication of the viral genome is performed by high-fidelity polymerase and allows viral replication in the nucleus of infected cells after incorporation of the viral genome, maintaining approximately 50-100 copies per cell (34).

Viral gene expression is regulated differently depending on the degree of differentiation of the infected cells. In the skin, the only actively dividing cells are located in the basal layers, and are composed of two main cell types: the transit amplifying cells (TA), which are proliferating cells that can undergo terminal differentiation, and the stem cells, that rarely divide in order to refresh the TA pool. Upon cell division, TA cells produce daughter cells which migrate away from the

basal layer and start to differentiate (35). As infected cells divide, viral DNA is equally distributed between both daughter cells. One of the daughter cells migrates and initiates a program of differentiation, while the other continues to divide in the basal layer and provides a reservoir of viral DNA for further cell division (36). Keratinocytes migrate upward as they enter through the differentiation process.

In the early phase of the viral life cycle, limited viral DNA amplification is supported by the viral E1 and E2 replication proteins (37), while E5, E6, and E7 are key to stimulating and enhancing viral proliferation. When HPV-infected cells divide and leave the basal layer undergoing differentiation, activation of the late viral promoter occurs and the level of viral proteins increases dramatically. As a result, the viral copy number soars from 50-200 copies to several thousand copies per cell (38). When reaching the upper layer of the epithelium, the capsid proteins L1 and L2 are involved in the encapsidation of the newly replicated genome, resulting in virion release during desquamation (39). Virion release is facilitated by E4 that interacts with the keratin network. The entire HPV life cycle is completed without causing cell death, viremia or apparent inflammation, in order to prevent alerting the immune responses (37, 40).



Locus	Function
URR	-Harbours transcription factor-binding sites -Controls gene expression
E6	-Promotes cellular immortalization by degradation of p53 -Modifies cell adhesion and differentiation by degradation of TAp63 and p73
E7	-Promotes pRb degradation, permitting cell progression to S-phase of cell cycle -Induces chromosomal instability
E5	-Induces cell proliferation -Contributes to evasion of apoptosis -Downregulates MHC expression
E1	-ATP-dependent DNA helicase activity -Role in viral DNA replication
E2	-Coactivator of viral DNA replication -Transcrition repressor of HPV E6 and E7 -Regulates cell cycle and apoptosis -Interacts with chromatin for segregation of viral genome
E4	Binds cytoskeletal proteins and disrupts cytoskeletal structure of the G2 arrested cell
L2	-Minor capsid protein -Recruits L1 protein for virus assembly
L1	-Major capsid protein -Can autoassemble into VLPs

Figure 3: Summary of PV gene function and schematic view of PV life cycle over the squamous epithelium with corresponding PV gene expression at each stage of the keratinocyte differentiation program. Adapted from (24).

I.4 – Clinical implications

Most PV infections are asymptomatic, indicating a commensalism between the PV and its host (41)(41). Some Gamma-PV infections can lead to benign lesions such as warts affecting children in boundary epithelia in the fingers, lips or eyelids. Sexual transmission of certain Alpha-PVs is also recognized as a cause of anogenital warts, possibly the most common sexually transmitted disease (42). The majority of viral infections are cleared by the host's immune system within 1 or 2 years (43) as a result of a cell-mediated immune response. However, a minority of infections become persistent, which increases the risk of cancer (44). A number of mucosal HPV types belonging to the alpha genus are classified as high-risk (HR) and low-risk (LR) HPVs based on their ability to induce malignant lesions. The International Agency for Research on Cancer (IARC) has classified 12 different HR HPV types as carcinogenic to humans, i.e. types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58 and 59 (45). HPV types 16 and 18 are the most frequently found in cervical cancers worldwide (in approximately 50% and 20% of squamous cell carcinoma, respectively, and in 35% of cervical adenocarcinoma for HPV18) (46-48). High-risk HPV types are also involved in a subset of other genital cancers, such as vulvar (around 40%), vaginal (70% to 90%), anal (around 80%) and penile (around 50%) cancers, as well as head and neck cancers where HPV16 is responsible for the majority (86-95%) of HPV-positive oropharyngeal carcinomas (49).

Several factors promote viral persistence of the oncogenic types, e.g. viral genomic variation, the host's genetics and the lifestyle behavior of the infected organism. In immunocompromised hosts, such as people infected with HIV, persistent high-risk infections are more frequent and severe (50). Some other sexually transmitted Alpha-PVs are also responsible for anogenital warts (42), but these PVs are also found in children before sexual debut, suggesting they are transmitted in other ways, for example through mother and child contact during labor (51-53).

The progression from HPV-associated lesion to invasive cervical cancer generally requires more than one decade and all of these steps are facilitated by deregulation and over-expression of the high-risk E6 and E7 viral oncoproteins together with the E5 oncoprotein. These viral oncoproteins target tumor suppressor pathways as well as paths involved in evasion from the host immune system in order to ensure viral genome replication and create a cellular environment at risk for oncogenic

transformation. Checkpoint mechanisms ensuring cell viability and integrity are hijacked by PVs' E6 and E7 proteins, causing uncontrolled cell proliferation (**Figure 3**) (54). E2F transcription factor family proteins are released and activated due to the E7 protein that causes degradation of Retinoblatoma family proteins through binding, leading to an unexpected re-entry into the S-phase. The cellular response to this unscheduled S-phase entry is cell apoptosis, but the E6 protein prevents programmed cell death through P53 degradation (55). Lastly, infected cells hyperproliferate due to E5 protein expression, facilitating malignant progression (56). Immune surveillance is also reduced by the E5 protein activity of oncogenic PVs (57). The outcome is uncontrolled proliferation activity and facilitation of mutations accumulation over time, leading to cancers. Thus, E1, E2, E4 and E5 expression is triggered in combination with E6 and E7, resulting in viral amplification of up to thousands of copies per cell (38, 58).

Approximately 75% of human HPV types are cutaneotropic, represented mainly by the beta and gamma genera, which are widely present in the skin of normal individuals. Beta HPV types were originally isolated in patients suffering from a rare autosomal recessive disorder, *Epidermodysplasia verruciformis* (EV), and several studies have proved the role of HPV infection in the pathogenesis of skin cancer associated with EV (59).

Cutaneous PV infections occur early in the lifetime (60, 61) and can cause benign lesions, but are mostly eradicated by the immune system in the following years (62). Several means of transmission have been described, including contact with infected hand or linens (63).

Many findings support the role of cutaneous beta HPV, together with ultraviolet (UV) irradiation, in the development of non-melanoma skin cancer (NMSC) (64, 65). One meta-analysis showed that five beta HPV types - HPV5, 8, 17, 20, and 38 - are significantly associated with the risk of cutaneous squamous cell carcinoma (cSCC) in immunocompetent subjects (66). However, experiments in animal models have provided evidence for the cooperation of the viral proteins with UV radiation in promoting cSCC (64, 65, 67).

In contrast to mucosal HPV infections in cervical cancer, where E6 and E7 expression is required for the initiation and maintenance of cellular transformation, beta HPV appears to play a role only at an early stage of cancer development.

Indeed, a higher copy number of beta HPV genomes is found in pre-malignant actinic keratosis lesions compared to skin squamous cell carcinoma (SSC) (68, 69). However, because of the plurality of HPV types found in the normal tissue and in the lesion of a single skin biopsy, the role of cutaneous HPV types in skin cancer remains unclear and requires further investigation.

I.5 – PV Evolution

A virus-host coevolution is suggested as being the main driving force of PV evolution, leading to an evolution directed towards adaption to their hosts, but only accounting for about one-third of the evolutionary events explaining PV-host evolutionary histories (70, 71). Even though PVs do not encode for DNA polymerase and use high-fidelity cellular polymerase, their mutation rates and mutational biases, as well as their codon usage preferences, do not match those of their host genome. Indeed, extreme codon usage preferences have been observed in PV genomes, specifically in E2, E6, L1 and L2 ORFs, in which a bias correlates positively with A+T content at the codon 3rd position (72). This codon usage preference seems to be conserved across the different PV types, regardless of the infected host, and it is thought to provide the replicational fitness of PVs in mammalian epithelial cells, even though this process still needs to be better understood. This codon usage preference can also be related to some selective processes, and several adaptive explanations have been proposed. One of these hypotheses is that PV codon usage lowers viral protein synthesis, enabling decreased immune exposure (73). Another adaptive process that could explain PV codon usage preference is the fitting of this codon usage with the variation in the keratinocyte tRNA profile during the differentiation program (74). A third explanation could be accommodation to the complex differential methylation occurring in the convoluted PV genome, including overlapping genes and regulatory regions, in order to avoid CpG island accumulation and thus escape immune response (75).

As PVs are dsDNA viruses, they present the slowest substitution rate among viruses, estimated between $2x10^{-8}$ and $5x10^{-9}$ substitutions per site per year for coding regions, although non-coding regions accumulate mutations faster (76-78). However, it seems that PV genomes are subjected to endogenous DNA mutation processes common to those of their host and leading to a specific type of DNA mutation. For

example, C>T mutation biases due to APOBEC3 cytidine deaminase similar to those retrieved in cancer genomes have been observed in PV genomes (79). Moreover, the viral genome is also exposed to the same exogenous processes as the host cells, especially to the pyrimidine cyclobutane dimers due to UV radiation, as viral replication occurs in the upper layer of the skin (80).

Another molecular event – recombination - plays a role in the dynamics of PVs and is suspected of having important implications for the emergence of oncogenic phenotypes, as well as for the colonization of new niches. Homologous recombination occurs during PV replication, during the productive stages of the infection when the rolling circle-like replication produces concatenated viral genomes that require excising and re-circularizing into individual plasmid genomes (81, 82). In rare cases, non-homologous recombination during natural infection has also been observed in HPV16 (83). One particularly important recombination event occurring between the early and late regions has oriented the evolution of PVs, is related to the integration of the E5 ORF, and has been associated with the infection phenotype (57, 84). Thus, viruses associated with similar clinical phenotypes cluster together when phylogeny is inferred using the early genes, but this pattern is not observed when phylogeny is inferred using the late genes. Acquisition of the E5 gene helps to sustain growth and escape immune response elicitation. This adaptive radiation has generated most of the PVs responsible for genital or cutaneous warts, as well as those responsible for mucosal lesions, some of them having carcinogenic potential (57). A few cases of distantly relative PVs presenting recombination events have also been observed, but these remain unusual (85, 86).

II. Detection of PV sequences

For several decades, PCR has been shown to be the most sensitive method for identifying HPV infection in clinical samples (87-91). Different sets of consensus or/and degenerate primers have been developed and used to detect HPV (92). However, primers amplifying DNA fragments in the conserved L1 region have become the most widely used in clinical and epidemiological studies.

II.1 PCR methods

The MY09/MY11 primer set-mediated PCR (MY-PCR) (93) and the GP5/GP6 primer set-mediated PCR (GP-PCR) (94) have been the most frequently used amplification systems for the detection of HPV DNA in clinical samples. Both sets of primers were developed in the early 1990s in order to detect mucosal HPV types in the genital tract (95, 96).

II.1.1 MY09/MY11 primer set

In 1989, the MY09-MY11 primer set was developed. This PCR protocol uses a set of 25 degenerate primers. The sensitivity of the MY09/MY11 primers set was improved few years later (92).

II.1.2 Nested-PCR protocols for genital HPV

In 1995, Ylitalo *et al.* developed a nested-PCR protocol (97). Conserved regions were retrieved on the L1 and the E1 ORF after the alignment of 19 different mucosal HPV types, using a multiple alignment program like Megalign (98) (**Figure 4A**). The panel of HPV types considered showing great diversity, the primer set has been synthesized with several degenerate nucleotides (**Figure 4B**). When compared to MY09/MY11 primers, these new primers enable identification of all genital HPV types known at this time with better sensitivity (HPV6, 11, 16, 18, 30, 31, 33, 34, 35, 39, 40, 42, 45, 51, 52, 53, 56, 57, and HPV58).

II.1.3 MY09/MY11 primer set improvement

In the beginning of 21st century, MY09/11 primers were redesigned to increase the sensitivity of amplification across the different PV types, still using the same kind of primers targeting the L1 ORF (99). For this purpose, multiple primer sequences were designed and combined in order to increase sequence heterogeneity capture ability, thereby avoiding the use of degenerate bases that could yield to irreproducible primer syntheses. These new PGMY09/11 primers appeared to be significantly more sensitive for HPV26, 35, 42, 52, 54, 55, 59, 66 and HPV73 detection than the MY09/11 system when tested on a set of cervicovaginal specimens.

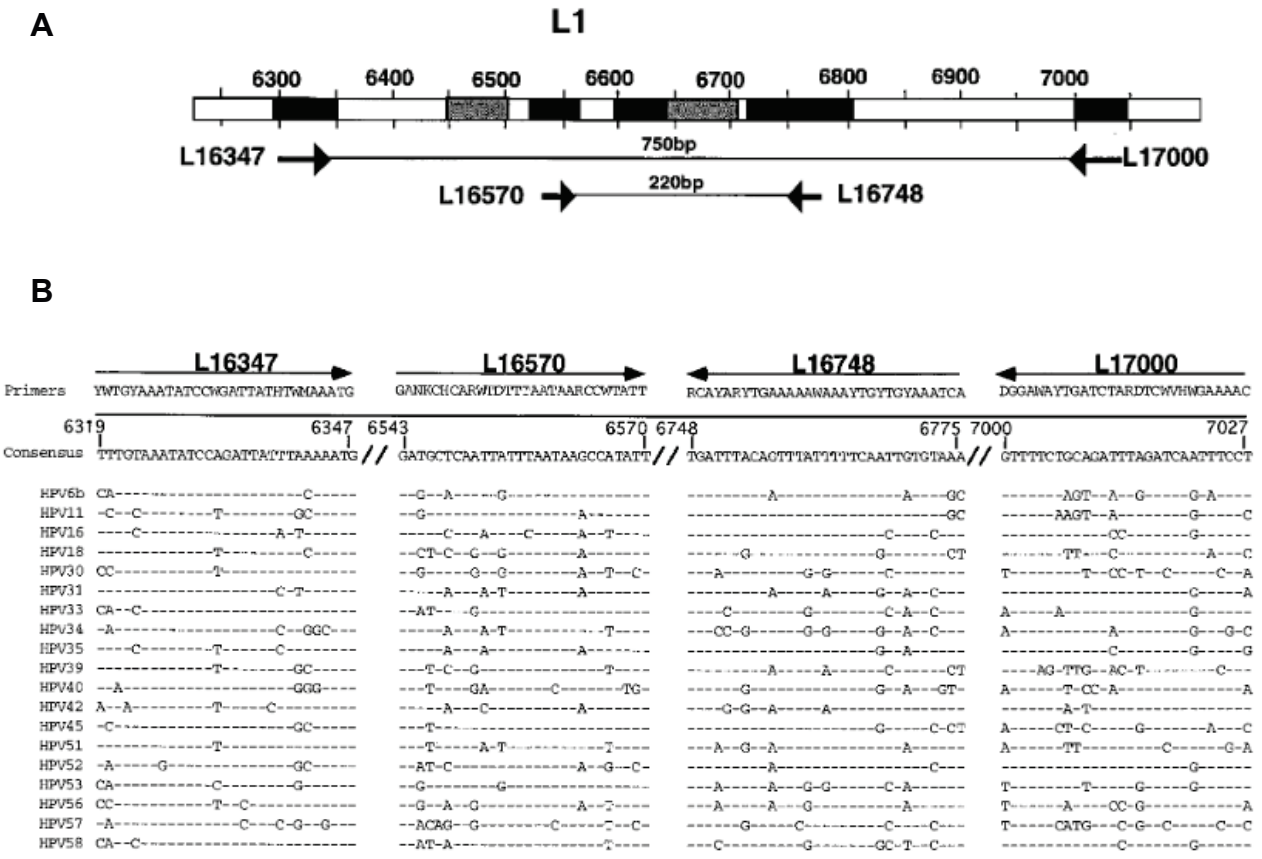


Figure 4: (A) Schematic representation of the location of the MY09/MY11 primers along the L1 ORF - (B) Alignment of the conserved L1 open reading frames regions among 19 PV

The arrows mark the locations of the primer pairs. The position number is relative to the HPV16 genome. **(B)** Alignment of the conserved L1 open reading frames regions among 19 PV, showing the positions of the primers, and the top line shows the resulting degenerate primer sequences (S = G or C; R = A or G; N = A, C, G, or T; Y = T or C; W = T or A; H = A, T, or C; M = A or C; K = G or T; D = T, G, or A). Deletions are indicated by asterisks, and nucleotides identical to those in the consensus sequence are indicated by dashes. Adapted from (97).

II.1.4 GP5/GP6 primer set

In 1990, the L1 conserved primers GP5 and GP6 were developed by Snijders *et* al., allowing the detection of 11 HPV (HPV1a, HPV6, HPV8, HPV11, HPV13, HPV16, HPV18, HPV30, HPV31, HPV32 and HPV33) (94). These primers consist of a fixed nucleotide sequence and detect a wide range of HPV types. Using a lower annealing temperature during PCR, there can be up to three mismatches between primers and

target DNA without affecting the efficiency of the assay. Five years later, the same research group published an improved version of these GP primers, GP5+ and GP6+, elongated at their 3' ends with adjacent highly conserved L1 sequences (100) (**Figure 5**). Multiple alignment of putative amino acid sequences from the L1 region flanked by both GP5 and GP6 of 24 mucosotropic HPVs revealed the consensus sequences Thr-Arg-Ser-Thr-Asn (TRSTN) immediately downstream of the GP5 (forward primer) region and Arg-His-X-Glu-Glu (RHXEE) upstream of the GP6 (backward primer) region (101). Moreover, increased primer length contributes to more efficient amplification, and has helped to improve the initial GP primers. When compared to their previous versions, these new primers present 10- to 100-fold higher sensitivity.

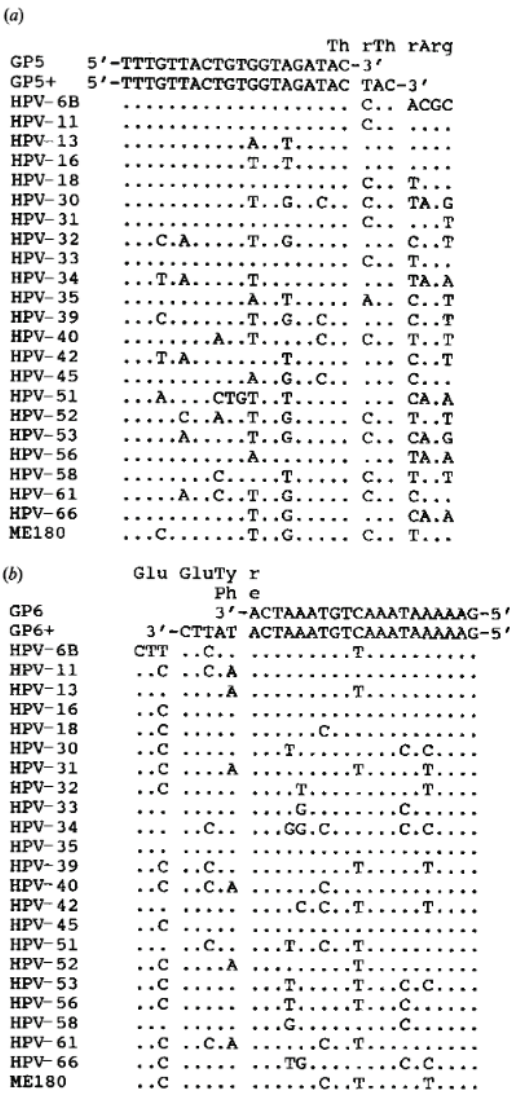


Figure 5: Alignment of GP5 and GP5+ (A) and GP6 and GP6+ primers (B)
with corresponding regions of L1 ORF of 23 mucosotropic HPV genotypes. Characters indicate mismatched pairs and nucleotides identical to those in the consensus sequence are indicated by dots. Adapted from (100).

II.1.5 BSGP5+/BSGP6+ primer set

More recently, the GP-PCR protocol has been improved to include L1 sequence information of up to 48 HPV types (102). This new BSGP5+/BSGP6+ protocol increased plasmid amplification of genital HPV types 10- to 1000-fold compared to GP5+/GP6+. Suitable for epidemiological and diagnostic applications, integration of internal Beta-globulin PCR allows simultaneous DNA quality control without affecting

the sensitivity of HPV detection. This protocol has been combined with Luminex technology to perform large-scale epidemiological studies (103, 104).

II.1.6 Primer set targeting mucosal and cutaneous PVs

As the diversity of identified HPVs increased over the years, and their biology was deciphered, a few types were suspected of being associated with skin cancer (105). Thus, the interest in detecting not only genital but also cutaneous PVs, and even all potential pathogenic PV types in general, increased over time. Thus, in 1994, Shamanin et al. published new primers aiming at detecting not only the known PV types, but also very distantly related types (106). Looking for the most conserved regions between L1 ORF of 45 HPV and 9 animal PV, it appeared that this region happened to coincide with the previously described GP-PCR region (94). These degenerate primers were successfully used to detect known HPVs and animal PVs, but also generated some false positive signals, raising concerns about the specificity of the assay. Subsequent DNA sequencing of the obtained PCR product appears to be the only way to confirm the presence of PV sequences. Two years later, De Villiers et al. used several combinations of these primers, enabling detection of 18 cutaneous HPVs associated with E*pidermodysplasia Verruciformis* (EV) (107).

II.1.7 Primer set targeting EV-associated HPV

Due to the increased interest in EV-associated HPV, Berkhout et al. developed a nested PCR protocol that enables detection of all known EV-associated HPV types at relatively low-copy-number levels (108). Combined with radioactive sequencing, these degenerate primers reveal many multiple infections never characterized before. A slightly modified version of these primers was subsequently developed by Boxman et al. (109).

All of these PCR methods allowed the detection of a wide range of cutaneous PV types, but their combined usage over time has resulted in discrepant reports (110).

II.1.8 FAP primer set

Hence, a general PCR method using single pair of degenerate primers and enabling detection of a broad range of HPV was developed in 1999 by Forslund et al. (111).

Using information from the L1 ORF of 80 HPV types from the HPV Sequence Database Compendia (112), two regions were defined with a relatively high degree of nucleotide sequence identity, leading to the construction of two degenerate primer sequences, FAP59 and FAP64 (**Figure 6**).

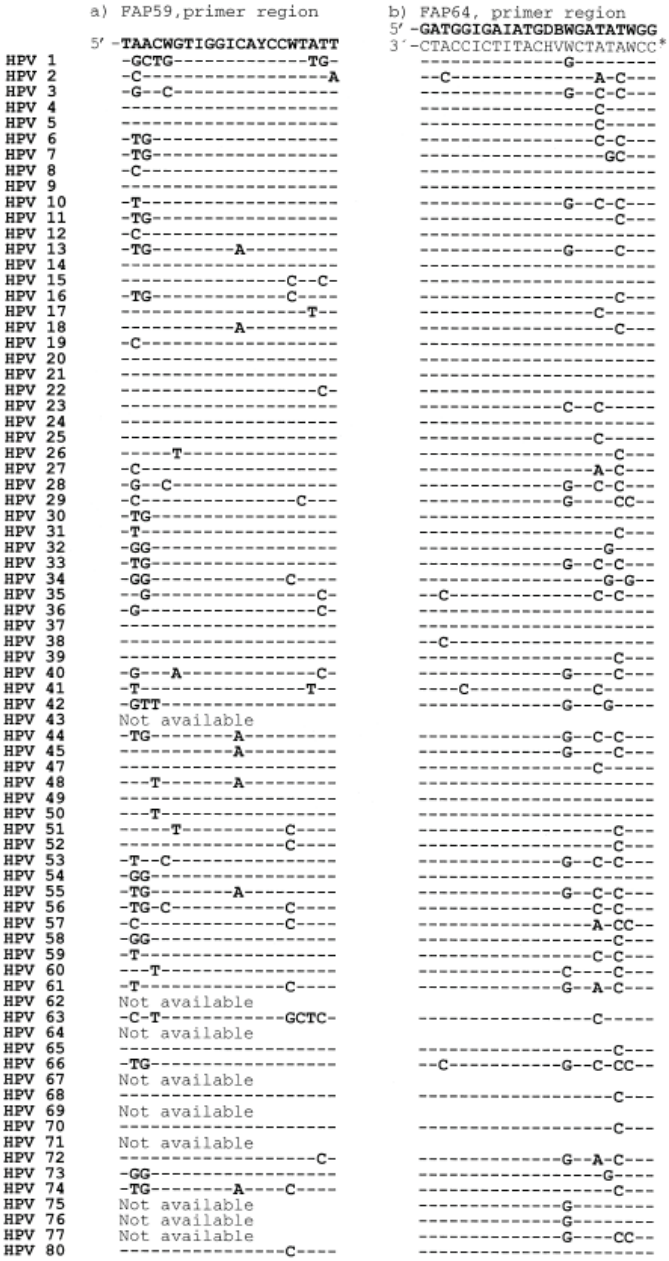


Figure 6: Alignment of the FAP59 (a) and FAP64 (b) sequences.

Lines and characters represent identical and mismatched nucleotides, respectively.

Degenerate nucleotides :W = T, A ; I = inosine ; Y = C, T ; D = A, G, T ; B = G, C, T ; H = A, C, T ; V = A, C, G. Adapted from (111).

Generating an amplicon of 478 bp, these primers allowed the characterization of 12 new HPV candidates. The method presents high sensitivity and allows detection of fewer than 10 copies of certain cloned HPV genomes. The technique was also found to detect significantly higher numbers of HPV from skin samples when compared to the nested PCR test described by Berkhout et al. (108). The main limitation is that direct sequencing of the amplicons from clinical material was unsuccessful due to the

presence of more than one type of HPV template in each sample. Moreover, the same authors observed an underestimated HPV prevalence in biopsies from immunocompetent patients, raising the need to increase the method's sensitivity (113). In any case, this method was successfully used in the following years after its development, and has led to characterization of 30 putative new human PVs (41) and 53 putative new animal PVs (114).

Four years later, the same author published an improved version of these primers using single-tube nested 'hanging droplet' PCR (115). The 15 EV associated HPVs known at this time all belonged to the β1 species but other skin-related HPV types (HPV4, HPV48, HPV50, HPV60 and HPV65) were found to belong to the β2 species (116). The goal of this primer improvement was to be able to capture more skin-related HPVs, in particular from the β genus. The objective was also to increase the primers' sensitivity through the use of nested primer pairs without increasing the risk of cross contamination from the primary amplification product due to the two-tube system needed to perform a nested PCR (117). Thus, a single-tube nested 'hanging droplet' PCR was developed. A region of 235bp between the original FAP 59/64 primers was identified when looking for a conserved region between the L1 genes of the 1996 HPV Sequence Database compendia (112) and two candidate HPVs, HPV92 and HPV93. A couple of new primers, FAP6085F and FAP6319R, were designed in this region (**Figure 7**). The sensitivity of these newly designed primers was increased 10-fold compared to the initial FAP primer protocol. The FAP primers have since been actively used to study HPV diversity in the skin (118, 119) as well as to develop new HPV detection methods (120, 121) that can be combined with Luminex (122) or with NGS (123, 124). Moreover, many animal PVs have also been characterized using these primers (125-129). In 2015, at least twenty-eight sub-genomic fragments identified using the FAP primer (FA-fragments) were recognized as novel HPV types (130) (**Table 2**).

```
              (a) FAP6085F, primer region   (b) FAP 6319R, primer region
                                            5'-TTGAYCCHAAACAANTNCAAATGT

              5'-CCWGATCCHAATMRRTTTGC       3'-AACTRGGDTTTGTTIAIGTTTACA*
HPV 1            G-------------------          -----G-------G-C--------
HPV 4            -----C-----C--------          -A-------------C---G----
HPV 5            -----------C--------          -------------G----------
HPV 8            --------------------          ----------G-------------
HPV 9            --------------------          -------------G----------
HPV 12           -C---------C--------          ------------------------
HPV 14           -----C--------------          ----------G-------------
HPV 15           --------------------          ----------G--G----------
HPV 17           --------------------          -------------------G----
HPV 19           -----C--------C-----          -------------G----------
HPV 20           -----------C--------          ------------------------
HPV 21           --------------------          ------------------------
HPV 22           -----------C--------          -------------G----------
HPV 23           -----C--------------          ----------G--------G----
HPV 24           -----C--------------          ------------------------
HPV 25           --------------T-----          -------------------G----
HPV 36           --------------------          ------------------------
HPV 37           --------------------          -------------------G----
HPV 38           --------------------          -------------------G----
HPV 41           ---A---------CC-----          ----------G----C-GCTGCTG
HPV 47           --------------------          -------------------G----
HPV 48           -----------C--------          -G--------G--G-G---G--AC
HPV 49           -----------C--------          ------------------------
HPV 50           -----------C--------          -A-------------C---GT-A-
HPV 60           -----------C--------          -G------------------T---
HPV 63           -----------C--------          -----G---------C--------
HPV 65           --------------------          -A-----------G-C--------
HPV 75           --------------------          ----------G--G----------
HPV 76           -----------------C--          -------------G----------
HPV 80           --------------------          -------------G----------
candHPV92        --------------------          ----------G--------GT-A-
candHPV93        -----C--------------          ------------------------
```

Figure 7: Alignment of the FAP6085F (a) and FAP6319R (b) sequences

Lines and characters represent identical and mismatched nucleotides, respectively. β1 species have an HPV name in bold font. Degenerate nucleotides: W = T, A ; I = inosine ; Y = C, T ; D = A, G, T ; B = G, C, T ; H = A, C, T ; V = A, C, G. Adapted from (115).

FA-fragment	Novel HPV type	PV genus	PV species
FA1.2	HPV134	*Gamma*	7
FA5	HPV110	*Beta*	2
FA8	HPV136	*Gamma*	11
FA13	HPV147	*Gamma*	8
FA16.1	HPV120	*Beta*	2
FA23.1	HPV124	*Beta*	1
FA28	HPV135	*Gamma*	15
FA35	HPV146	*Gamma*	15
FA44	HPV133	*Gamma*	15
FA47	HPV96	*Beta*	5
FA51	HPV111	*Beta*	2
FA53	HPV100	*Beta*	2
FA67	HPV142	*Gamma*	10
FA69	HPV180	*Gamma*	10
FA73	HPV141	*Gamma*	11
FA75	HPV104	*Beta*	2
FA83	HPV144	*Gamma*	17
FA85	HPV107	*Beta*	2
FA119	HPV98	*Beta*	1
FA136	HPV123	*Gamma*	7
FA137	HPV109	*Gamma*	7
FA147	HPV130	*Gamma*	10
FA155	HPV119	*Gamma*	7
FA164	HPV112	*Gamma*	7
FAIMVS2	HPV92	*Beta*	4
FAIMVS6.1	HPV93	*Beta*	4
FA.X	HPV153	*Gamma*	13
FA.X	HPV154	*Gamma*	11

Table 2: Twenty-eight sub-genomic FA-fragments recognized as novel HPV types in 2015

Adapted from (130).

II.1.9 CUT primer set

In 2010, Chouhy et. al. developed a new generic primer system targeting both mucosal/genital HPV and cutaneous HPV, leading to the characterization of HPV115, a novel β3 species type (131). The L1 ORF of 88 cutaneotropic and mucosotropic HPVs were aligned to find conserved regions, and led to the design of 4 forward primers, namely CUT1Fw, CUT1AFw, CUT1BFw and CUT1CFw, and one reverse primer, CUT1BRv (**Figure 8**).

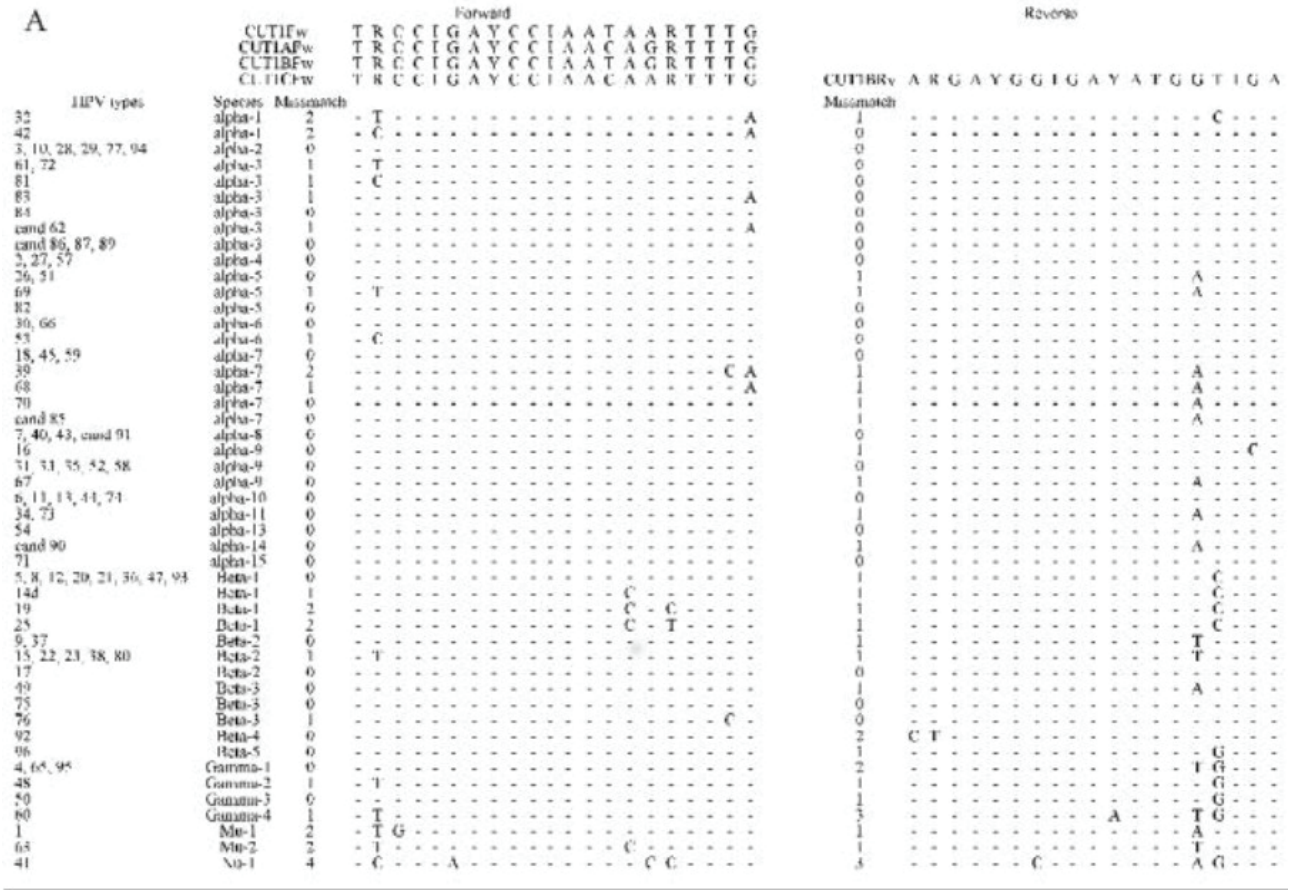


Figure 8: Alignments of the forward and reverse CUT primer sequences with the corresponding region in the L1 ORF of 88 selected HPV types

Lines and characters represent identical and mismatched nucleotides, respectively. Degenerate nucleotides of primers: I = inosine, Y= C or T, R= A or G. Adapted from (131).

The 5 primers can be prepared as a mix (CUT1ABCFw) and yield an amplicon of 370 bp. When comparing the efficiency of CUT primers with that of FAP primers using

skin samples, it appears that they have a differential specificity, with only a subset of samples tested being positive using both primers. Moreover, the sample that came back positive for both primer systems only presents the same viruses in 22% of cases (12/55), for a total of 69 different HPV types or putative new types identified, including 10 detected by both primer systems (**Figure 9**).

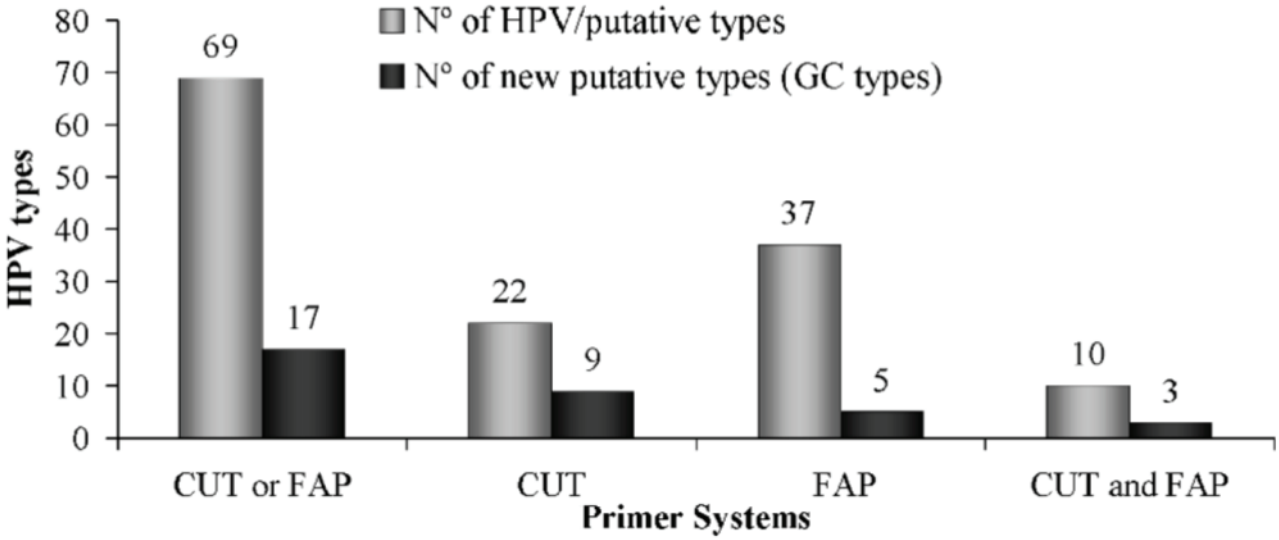


Figure 9: Differential capacity of CUT and FAP primer systems for HPV type/novel putative type detection
Adapted from (131).

Presenting a better capacity to identify novel HPVs, the CUT primer system also showed a broader capacity to detect HPVs from different genera and species with respect to the FAP primer system. This primer system was thus used in the following years for prevalence studies (132), as well as in combination with NGS (133).

II.1.10 Consensus-degenerate hybrid oligonucleotide primers (CODEHOP)

In 2004 Baines et. al. also developed new consensus-degenerate hybrid oligonucleotide primers named CODEHOP to detect novel PVs (134). Clues indicating the possibility that some HPVs may be more closely phylogenically related to animal PVs than to human genital or mucosal PVs motivated the synthesis of these primers. PVs sequences from the Human Papillomaviruses Database hosted by the Los Alamos National Laboratory (LANL) between 1994 and 1997 were used to design the primers. Conserved amino acid sequences from the PV L1 protein of

representatives from many different PV types (human and animal) were retrieved, and blocks were defined using the Blocks program (135). Two blocks (E and H) yielded primers that could amplify plasmids containing representative PVs from three of the six supergroups defined as follows: (A) for mucosal/genital HPV types, (B) for cutaneous/EV-associated HPV types, (C) for bovine and deer PVs, (D) including BPV4, (E) for rabbit PVs and variants of HPV1 and, lastly, the group (S) to represent *Mastomys natalensis* PV (MnPV). Degenerate primers Mayo E (ME) and Mayo H (MH) were derived from these blocks, and subsequent consensus partially degenerate primers were created for each group, including a clamp region to match the specific group of papillomaviruses (**Figure 10**).

Supergroup	Primer	Sequence
Downstream		Clamp Core
A	ME/A1	5′ G A A T T A A T A A C T T G T G T T A T A **S A R G A Y G G N G A N A T G** 3′
	ME/A2	5′ G A G T T T A T T A C T A C A M C T A T T **S A R G A Y G G N G A N A T G** 3′
	ME/A3	5′ G A A T T A A T A A A C A C A G T T A T T **S A R G A Y G G N G A N A T G** 3′
B	ME/B1	5′ G A A T T A A W A A A T A C A G T T A T T **S A R G A Y G G N G A N A T G** 3′
	ME/B2	5′ G A A C T A G T A A A T A C T G T T A T T **S A R G A Y G G N G A N A T G** 3′
C/D	ME/C	5′ G A A T T A A A A A A C A R A Y A T A T A **S A R G A Y G G N G A N A T G** 3′
	ME/D	5′ G A G C T C A A A A A C A C A A A A A T A **S A R G A Y G G N G A N A T G** 3′
E	ME/E	5′ G A A C T R A T A A A C A C A G T C A T A **S A R G A Y G G N G A N A T G** 3′
	ME/41	5′ G A G C T T A A G T C C T C A T A C A T T **S A R G A Y G G N G A N A T G** 3′
S	ME/S	5′ C A G C G G A T G T C T G G G A T G A T T **S A R G A Y G G N G A N A T G** 3′
Upstream		Clamp Core
A	MH/A	5′ A A Y T G A T T A C C C C A R C A **N A Y N C C R T T R T T** 3′
B	MH/B	5′ A T Y T G A T T R C C C C A A M A **N A Y N C C R T T R T T** 3′
C/D	MH/C1	5′ A A T A A A T T A T T C C A T G C **N A Y N C C R T T R T T** 3′
	MH/C2/D	5′ A G T G G A T T A T T C C A G C A **N A Y N C C R T T R T T** 3′
E	MH/E	5′ A A C T G A T T G Y T C C A G C C **N A Y N C C R T T R T T** 3′
	MH/41	5′ G C C T C G T T G T G C C A C A G **N A Y N C C R T T R T T** 3′
S	MH/S	5′ A A G T C T T T G T T C C A C A G **N A Y N C C R T T R T T** 3′

Figure 10: CODEHOP PCR sequences
for each group. Adapted from (134).

Even though they present equivalent sensitivity for HPV detection in the genital/mucosal group (group A) when compared to previous MY09/11 primers,

CODEHOP primers show greater ability to detect PVs from the other 5 groups when tested on plasmid and clinical samples from esophageal and tonsillar cancer.

In 2013, approximately 200 to 250 putative HPV types were identified using one of the described primer systems. However, less than 25% of these putative HPV types were fully cloned and characterized (136). To obtain longer sequences, primers located in the L1 and E1 primers were used together to reach around 4 kb amplicon and enabled the characterization of HPV156 (136). Using this primer in order to isolate and reconstruct a full genome was successful when combined with cloning and Sanger-sequencing-based strategies. However, this approach is quite laborious and time-consuming, and enables identification of only the most represented amplicons. In particular, this strategy is ineffective in the context of multiple infections, or in the event of a very low amount of viral DNA in the initial sample.

II.2 Rolling circle amplification (RCA)

Identification of novel PV types has been greatly facilitated by the use of a special isothermal DNA amplification technique, RCA, enabling amplification of any circular single- or double-stranded DNA molecule using bacteriophage phi29 DNA polymerase and random hexamer primers (**Figure 11**) (137). This method does not require any prior knowledge of the nucleotide sequence content and allows the detection and high-fidelity amplification of circular viral genomes, the polymerase being highly efficient and less error-prone. Used in combination with genus-specific primers, RCA can be directed towards specific PV types (138).

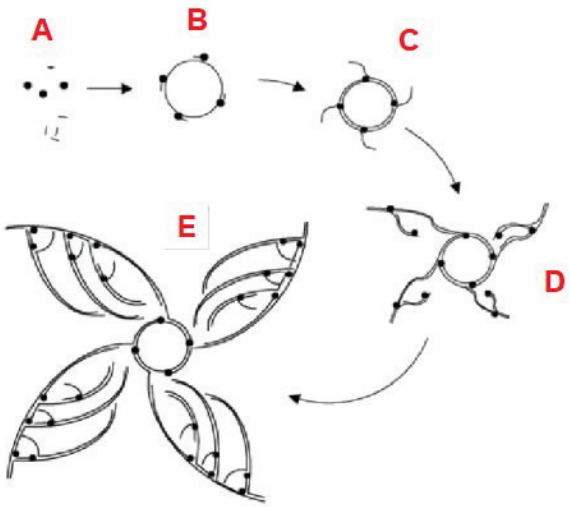


Figure 11: Amplification of circular DNA matrix by RCA

(A) Primer hybridization on circular denatured dsDNA. (B) DNA synthesis by phi29 DNA polymerase following the matrix. (C) Non-stop amplification of concatemeric repetitions of the circular genome (D) DNA strand movement permitting exponential amplification (E).

Restriction analysis, cloning and Sanger sequencing of RCA led to the identification of novel HPVs (139). One limitation of the RCA technique is its capacity to also amplify the host genome due to RCA's ability to also amplify linear DNA to a lesser extent. Thus, pretreatment with exonuclease or separation of the targeted 8kb fragment by gel electrophoresis must be applied (140) to select for the amplified viral DNA material. Optimized RCA protocols followed by restriction enzyme analysis and qPCR have been applied to improve HPV16 detection in human keratinocytes (137).

II.3 Combination with NGS

The first robust DNA sequencing method was developed during the 1970s at the Medical Research Council Center in Cambridge by Frederick Sanger (141), who was awarded the Nobel Prize in 1980 for this innovative discovery. The method is based on incorporating complementary nucleotides during in vitro DNA replication, which can be determined at the end of the experiment. The Sanger sequencing method was used to produce the first human genome in 2001 (142). Nevertheless, the main disadvantage of this sequencing technique is its limited throughput.

This need has driven the development of new high-throughput sequencing methods, known as High-Throughput Sequencing (HTS) or Next Generation Sequencing (NGS) (143). NGS, or "deep sequencing", refers to high-throughput sequencing technologies that allow massive parallel sequencing of different DNA molecules in a short period of time. Several NGS platforms are available, and differ in the technology used, the costs of sequencing and the amount of sequence data generated (144).

Whatever the sequencing technology, the NGS DNA sequencing method follows three steps: library preparation, amplification of DNA fragments and sequencing of these fragments (143). The resulting entity is called a sequencing read, which identifies the sequence of a particular DNA segment, and an NGS experiment can create a massive amount of reads, up to several hundred million. Most NGS technologies can generate "paired-end" reads (i.e. read from the two ends of a particular DNA fragment) but some are limited to "single-end" reads.

The capacity of NGS technologies to sequence millions of DNA fragments is a major improvement when comparing NGS and traditional Sanger sequencing: NGS offers the unique ability to detect many diverse DNA sequences simultaneously.
The raw sequencing only outputs small segments corresponding to initial DNA, but these small fragments need to be mapped to the reference DNA to then entirely reconstruct the DNA sequence or mapped to an annotated DNA sequence database to characterize the sequencing reads. There are multiple methods of NGS alignment based on different algorithms. The main challenges justifying these developments are computing time and accuracy of the method (145).

II.3.1 Metagenomics

In recent years, the genome sequences of several novel PV types have been identified using NGS in various specimens such as the skin (123, 140, 146, 147), head and neck mucosal sites (148-150) and genital tissues (151). NGS has similarly been commonly used to analyze amplicons containing PV sequences generated using the original RCA method (152), in combination with PV type-specific primers (138) or using junction probes (153).
In the fall of 2018, Pastrana et. al. published an important metagenomic study leading to the discovery of 83 novel HPVs in immunodeficient patients (154). Using virion enrichment, RCA and NGS, they identified all circular DNA viruses in 48 patients presenting immunodeficiencies (warts, hypogammaglobulinemia, infections, myelokathexis (WHIM) syndrome or EV) and compared them with healthy adults and non-immunodeficient individuals. Such a study has required crucial bioinformatics support and the building of a data processing strategy (https://github.com/BUCK-LCO-NCI/VirConTaxa). Here, contigs were *de-novo* assembled using SPAdes (155) after quality trimming of the reads. Contigs longer than 300 nt were subsequently identified using the MegaBlast algorithm against the nr/nt NCBI database and all sequences presenting at least 95% identity with their matches were considered as known sequences. To identify distantly related previously unknown viral sequences, a viral protein database was created utilizing a custom library of ORFs that are known to be well conserved, such as E1 and L1 of HPV16, and many other virus proteins. tBlastn was used to query this database, Blastn against Genbank enabled false positive elimination and criteria such as contig length, coverage and percentage

of identify were defined to select for more promising putative new HPV sequences. Circularity of the promising sequence identified was confirmed by re-mapping the raw reads to their respective contigs in order to find reads overlapping both ends of the candidate genome. Incomplete sequences (non-circular and with a length different than the expected 8kb) were extended at their ends by re-mapping of raw reads if coverage, extension length, overlapping region length and percentage of identity on the overlapping region criteria were met. An iterative process was used until the aforementioned criteria were no longer met, and Sanger sequencing was used to confirm the low covered novel HPVs. Lastly, phylogenetic analysis was performed to better characterize putative new HPV genomes (**Figure 12**). Sixty-nine out of the 83 new HPVs are Gamma types, 8 are Beta types, 1 is Mu, 1 is Alpha and 4 are potential new HPV types (<70% identity). In addition, the authors characterized 35 incomplete genomes representing potential new types.

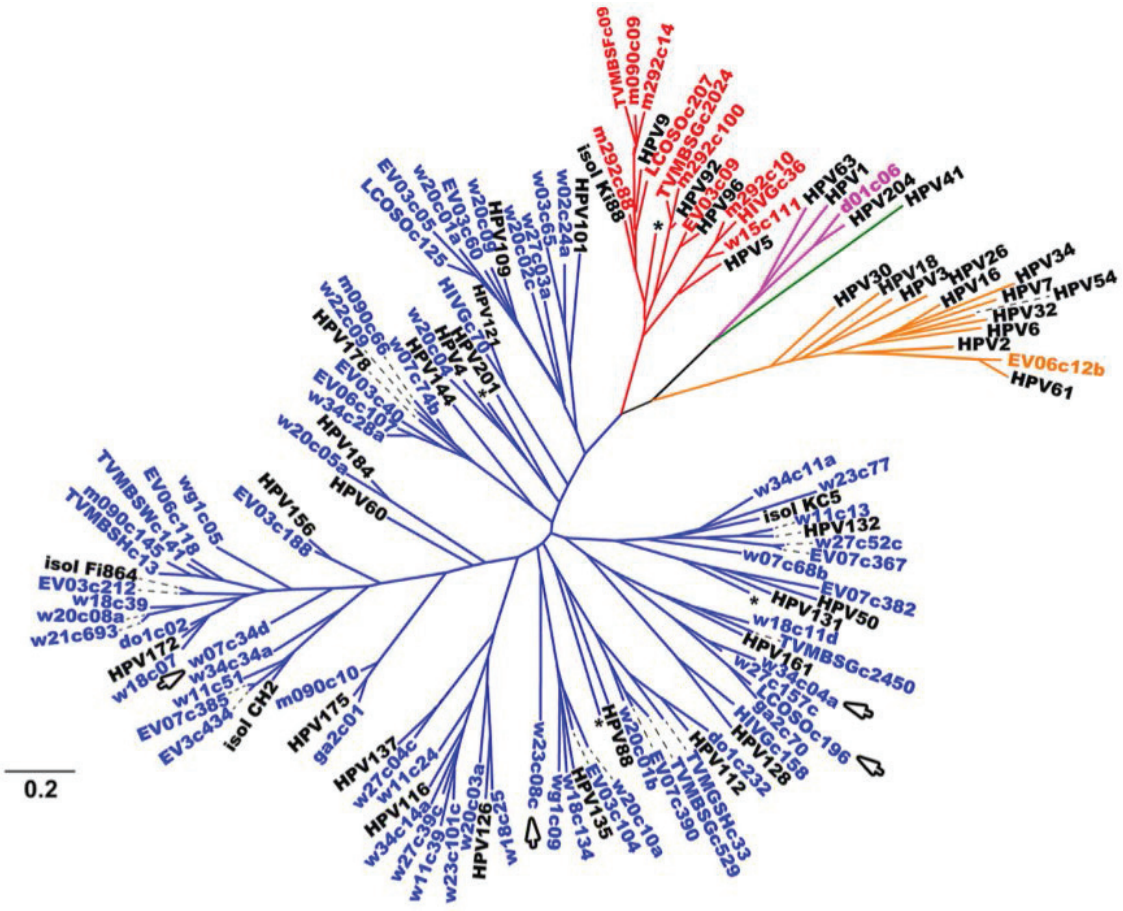


Figure 12: Phylogenetic analysis of 83 reported novel HPV types based on the L1 protein sequence

Asterisks show known representative species that did not fit due to the figure's space restrictions (counterclockwise, Beta 3-HPV49, Gamma 20-HPV163, Gamma 21-

HPV167, and Gamma 2-HPV48). Orange, Alpha; red, Beta; blue, Gamma; pink, Mu; green, Nu. Dotted lines are not significant and arrows indicate potential new species. Adapted from (154).

Around the same time that the Pastrana et. al. results were published, Tirosh et.al. published deep metagenomic sequencing data from dedicator of cytokinesis 8 (DOCK8) -deficient skin samples, describing 250 putative new HPV genomes (156). DOCK8 deficiency is a rare primary human immunodeficiency characterized by recurrent cutaneous and systemic infections, as well as atopy and cancer susceptibility (157). Total DNA was isolated and library preparation was performed to generate shotgun metagenomic sequence data from skin sites. Reads were also *de-novo* assembled using SPAdes (155) to form contigs. Contigs less than 750 nt in length and contigs presenting aberrant GC content were filtered out, and remaining sequences were taxonomically classified to a taxon based on the BLASTN best hit against the nt database. In all, 8,631 of the contigs were identified as belonging to the *Papillomaviridae* family, including 3,725 contigs with a size comparable to an HPV genome (~7.9kb). Using the Papillomavirus Episteme (PaVE) Database (158) (www.pave.niaid.nih.gov), the contigs were masked if presenting more than 90% identity on a 500nt window, and subsequently excluded if more than 4000nt of the contigs was masked, resulting in 2,189 contigs. The mash algorithm (159) was used to build a graph of 95% identical contigs which were then clustered using SPICi (160) to remove redundant genomes. This clustering resulted in 208 clusters and 42 singleton genomes, for a total of 250 non-redundant HPV genomes. Using the L1 taxonomy tool available on the PaVe website (https://pave.niaid.nih.gov/#analyze/l1_taxonomy_tool), 205 out of the 250 novel genomes were depicted as novel types and 45 genomes were depicted as members of a single novel species. The novel species were subsequently verified by targeted PCR to the L1 region from the original DNA extracted from the skin swabs. In all, 229 of the 250 novel HPV genomes belonged to gamma genus of HPV, 19 belonged to the beta genus and 2 belonged to mu genus (**Figure 13**).

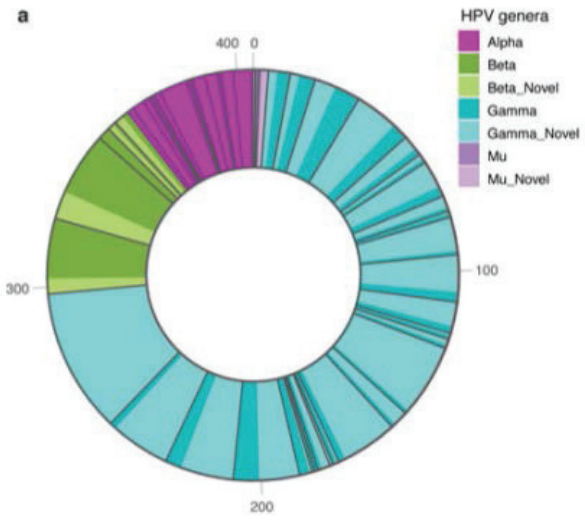


Figure 13: HPV diversity on DOCK8-deficient patients' skin

Each genus is depicted by a different color; pie slices represent different species within a genus and lighter shades represent new HPV types within the species. Numbers around the pie refer to the HPV type count. Adapted from (156).

The main pitfall of massive parallel sequencing is that contig assembly can result in artifactual assembly of chimeric genomes (161, 162), particularly when closely related viral strains are present in the same sample, and methods are still being developed to limit the occurrence of chimeric sequence reconstruction (163, 164). Pastrana et. al. found a way to address this issue, for the case of long chimeric reconstruction, by creating phylogenetic trees based on E1 protein sequences and L1 protein sequences, to enable checking of artifactual chimerization between early and late genome segments (156). Yet, small artifactual re-arrangement among closely related genomes remains undetectable. As well, as already stated in the RCA section, the capacity of RCA to additionally amplify the human genome may impair the amplification of HPV genomes if present in low amounts. The preference of WGA to amplify circular molecules is an advantage when detection of PVs is desired, but the lower efficiency in amplifying linear molecules impairs the ability to infer what amount of circular and linear viruses may have been present in the original sample (165-167). For example, in the study of Bzhalava D. et. al., NGS of FAP PCR amplicons of skin samples detected 352 different HPV types/putative novel types compared with only 26 different HPVs detected by NGS of RCA-amplified samples (140).

II.3.2 Amplicon-based experiment

Paired-end NGS has also been used to analyze amplicons of various PV-specific broad-range PCRs (123, 124, 146). In 2011, Ekström et. al. studied the diversity of HPV types in skin lesions such as squamous cell carcinoma (SCC), actinic keratoses (AK) and keratoacanthomas (KA) from fresh frozen biopsies (123). The FAP primers were used in combination with NGS, leading to the discovery of 44 novel putative HPV types. After the removal of primer sequences and all sequences at least 80% similar to the human genome, the MIRA assembler (168) was used to assemble contiguous sequences and, subsequently, a Blastn against GenBank was launched to identify putative novel HPVs presenting more than 10% dissimilarity to other HPV genomes. In this study, a large number of putative novel types were found in spite of the fact that the degenerate primer pair used was not able to detect all HPV types. Moreover, sequences from the genus Alpha - HPV 3, 16, and 77 - were also found. However, correct classification of the novel putative viruses can only be obtained by cloning and sequencing the full viral genomes.

In 2013 and 2014, the same research team published two other studies describing the use of PV-specific broad-range primer systems in combination with NGS (124, 146). The samples processed came from frozen biopsies and skin swabs from SCC, AK and KA, as well as from basal cell carcinomas (BCC) in the most recent study.

For their study in 2013, the FAP56/64 primer was used to amplify the L1 region in each individual sample, and pools of samples were constructed and bidirectionally sequenced at least twice, using multiplex identifiers (MIDs) allowing separation of individual samples from each other in a single sequencing run and without the use of MIDs (124). In order to increase the coverage of HPV types and investigate whether their broad detection by the FAP primers could be further improved, FAP primer sequences were aligned against all HPV sequences from GenBank. As well, five forward and four reverse partially degenerate new primers were designed. Used in combination with the FAP6085 and FAP64 primers, the reads generated were trimmed for low quality bases and primer sequences, and SSAHA2 software was used to screen for human and bacterial DNA (169). The remaining sequences were assembled using MIRA (168) and identified by Blastn against GenBank, and a protocols was set up to identify potential chimeric sequences. In short, the sequence that aligned to its most closely related sequence in GenBank was divided into three

equal segments. If at least one of the segments differed in similarity to the corresponding nearby parts by more than 5% (e.g., if segment 1 was 88% similar and segment 2 was 94% similar) the sequence was considered as a “possible chimera”. Lastly, HPV-related sequences, which had less than 90% identity over 90% of their length to known HPV genomes, were classified as putative new HPV types. Analysis of their sequencing runs led to the identification of 273 different HPV types or putative types, including 87 known HPV types, 139 sequences from previously known putative HPV types and 47 sequences of putative novel types (44 putative new Gamma PVs and 5 putative new Beta PVs). Only 17 of the 44 putative new sequences were detected in Luminex-based testing of the same samples, suggesting low copy numbers and/or mismatches to the PCR primer sequence. Their results proved that by using updated sequencing methods and PCR protocols, even HPV types present in low copy numbers can be found. Furthermore, twenty-five different known mucosal types from the Alpha genus were also detected, highlighting the value of updating the primer systems. Finally, the re-sequencing provided a clear distinction between HPV types detected by all MIDs, probably present at high copy numbers, and HPV types only picked up by a subset of the MIDs, and therefore probably present at a much lower abundance.

II.3.3 Evaluation of the methods

One year later, this research team published a third study following the same protocols, leading to the characterization of HPV197 (146). At the same time, they launched two other NGS (using MiSeq and HiSeq Illumina machine) without any prior PCR on the same samples in order to compare the efficiency of both methodologies. The main bioinformatic steps were as follows: quality control (QC) and trimming of raw reads, exclusion of reads presenting high similarity to the human genome over their length when aligned with BWA-MEM (170), normalization (redundant read exclusion and reduction of sampling variation and sequencing errors), and lastly *de-novo* assembly using the SOAPdenovo (171), SOAPdenovo-Trans (171), Trinity (172) and IDBA-UD (173) assemblers. Finally, reads were mapped back to the reconstructed contigs in order to increase their coverage. The same chimeric detection protocol described above was applied (124). The MiSeq produced a total of 0.03% viral reads, 75% of which were HPV-related reads, identified in 28 out of 91

specimens (four known HPV types, one putative known and one unknown). The HiSeq produced a total of 0.04% viral reads, 96.4% of which were HPV-related reads, identified in 31 out of 91 specimens (three known HPV types, two putative known and four unknown). Putative known types are types for which the International HPV Reference Center (https://ki.se/en/labmed/international-hpv-reference-center) did not receive a biological clone, did not re-sequence the genome, and thus are not part of the reference taxonomic classification of the *Papillomaviridae* sequences, and do not have any official HPV number. The author gave no information on the fraction of viral reads when using PCR-guided amplification before sequencing. Nonetheless, a total of 24 known HPV types, 13 putative known types and 3 novel types were detected using this methodology. Only two HPV types were detected both when sequencing without prior general primer PCR and when sequencing PCR amplicons, and HPV197 was detected only without prior HPV-specific amplification. The FAP PCR primer sequence had several mismatches with HPV197, indicating that this virus could not be amplified by this primer set.

To conclude, viruses with low similarity to primer sequences will not be amplified during the PCR reaction and thus, will remain undetected if a PCR reaction is used before the detection step. The unbiased method is not dependent on any particular sequence but is less sensitive in detecting low amounts of virus. Moreover, without prior HPV-specific amplification, most of the sequencing reads will be discarded due to the very low specificity of the protocol. The most sensitive method is to perform sequencing of PCR amplicons. The review of the different techniques to detect and characterize PVs from PCR methods highlights the power of NGS, and raises the need to update the HPV-PCR primers in order to accurately represent today's PV diversity, and improve their ability to detect a broad range of diverse PVs.

III. Bioinformatics workflows for virus detection

Rapid advances in NGS technologies have helped the development of new bioinformatics methods for identifying and characterizing pathogens (174), and more broadly to study the genetic material recovered directly from environmental samples. Amplification and sequencing of the 16S rRNA gene has allowed for high-throughput detection of prokaryotic communities, while shotgun metagenomic sequencing

approaches have enabled the capture of the composition and functional potential of multi-species populations.

The first approach required a good biomarker, constructed based on conserved features shared among the representatives of the group studied, but also presenting sufficient diversity to enable a clear distinction between them (175). Such a biomarker is not always available for all groups of organisms studied and is even less common when representatives of the group studied present a significant molecular divergence during the time. Moreover, the pertinence of using such biomarkers is also linked to the biomarker's capacity to reconstruct the phylogenetic history of the group studied (176).

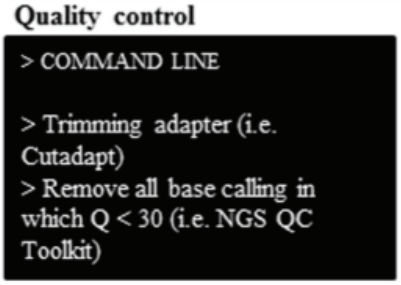
The second approach aims to identify any genetic material present in a given sample by employing non-specific amplification and sequencing of nucleic acid. Metagenomics has been used extensively to identify already-known and novel viruses in seawater, nearshore sediments, feces, serum, plasma and respiratory secretions (177). This approach has helped scientists to study extraordinarily diverse and previously unexplored microorganisms, which has led to the realization that there are many organisms yet to be discovered (178).

The fact that these two techniques are independent of lab culturing makes it easier to sequence large groups of microorganisms in many environments without the need for a wet lab approach. Because there is no gene that is common to every virus, viral metagenomics based on sequencing of non-specifically amplified genomic material has become more commonly used. Nonetheless, strategies have been developed to study subgroups of viruses sharing some genetic elements used as a biomarker (179, 180). In both methodologies, cost reductions paired with the increased amount of data thanks to the advent of next-generation sequencing has led to a rapidly growing demand for bioinformatics software. Sophisticated bioinformatics tools able to process millions of reads in a reasonable time are thus required and therefore must be available to the scientific community. Difficulties have emerged due to NGS instruments producing inherent platform-specific error profiles. Possible biases and artifacts introduced during the sequencing process need to be sufficiently understood. However, it could be shown that they can, if left untreated, have a falsifying impact on study results (181-187).

III.1 Quality control

Quality control (QC) must be performed prior to any other data manipulation steps to remove artifacts, such as low-quality and contaminant reads. Low-quality reads may compromise downstream analysis, and their removal therefore increases accuracy in detecting the microbial or viral diversity of the sample (188). Identification of low-quality reads is possible during the QC stage thanks to the PHRED score. The PHRED score is part of the files containing the DNA sequencing data, coming from the analyses of the base calls, and encompassing a quality value for each call based on the formula Qphred = − 10xlogPe, where Pe stands for the probability of error for that base call (189). Thus, the Qphred score corresponds to the probability that a base has been erroneously incorporated (**Figure 14**). A Qphred value ranging from 25 to 30 is commonly employed to guarantee sequence confidence.

The NGS QC Toolkit is one existing tool for quality checking, filtering, trimming, generating statistics and converting NGS data files to different formats (190). One of today's most commonly used tools to perform QC on NGS data is FastQC (191), which enables processing of the same step described above, and generates a user-friendly Hypertext Markup Language (HTML) report.



Qphred values and base calling confidence

Q	Pe	Confidence (%)
30	0,001	99,999
20	0,01	99,99
10	0,1	99,9

Figure 14: The main steps during QC
are trimming of adapters inserted during library preparation and the removal of reads possessing low-quality base calls identified based on the Qphred score value for each sequenced base. Adapted from (192).

Sequences from adapters, which are chemically synthesized linkers ligated to the end of the DNA fragment to be sequenced during library preparation and enabling binding of the DNA fragment to the sequencing support (flowcell), must also be removed. This step can be performed using Cutadapt, which aligns the reads with all

adapter sequences, depending on the sequencing platform (193). The algorithm penalizes alignments in which adapter sequences are aligned with the 3′ region of reads, and thus all sequences from the adapters are removed. On the other hand, if the adapter sequence is overlapped at the beginning of the read, the sequences prior to the overlap are removed. Trim Galore (194) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) is a wrapper tool incorporating the Cutadapt algorithm and the QC reports from FastQC, making it a convenient tool for the QC step.

III.2 Assembly algorithms for metagenomic data

In general, there are two approaches for WGS metagenomics: read-based and assembly- based metagenomics (**Figure 15**). The former aims to classify single reads with regard to taxonomy and function. It is well suited to answer questions concerning the taxonomic composition of a sample or related to the presence or absence of organisms, genes or metabolic pathways. A fragment of metagenome sequences can be aligned to known reference genomes or genes to examine their coverage and variation, but this method is not suitable for discovery of novel organisms.

In assembly-based metagenomics, reads are first *de-novo* assembled into contigs and then clustered into so-called “genome bins” during a binning process, in order to reconstruct scaffolds. Contigs are continuous stretches of sequence containing only A, C, G, or T bases without gaps, and scaffolds are created by chaining contigs together (sometimes separated by gaps) using additional information about the relative position and orientation of the contigs in the genome. Thereby, it is possible to reconstruct genomes of abundant taxa from a metagenomic sample. For this purpose, the corresponding workflow includes an assembler that is well suited for the reconstruction of long contigs and a genome binner to cluster such sequences from the same organism in order to reconstruct scaffolds.

The difficulty of metagenome assemblies lies in the biological complexity of the sample, being a mixture of genomic elements from different genomes at varying coverages. Such a mixture contains some low-complexity sequences (i.e. dimer repetitions), as well as orthologous and paralogous sequences of different organisms. Homologous sequences refer to historical continuity, in which biological features in

related taxa are similar in pattern or form because they evolved from a corresponding structure in a common ancestor. Paralogous sequences are related to biological features retrieved in the same organism by descent from a single ancestral feature that was duplicated, and that may have a different DNA sequence and function. The difficulties in differentiating such sequences can lead to intragenomic or intergenomic chimeric assemblies (195, 196).

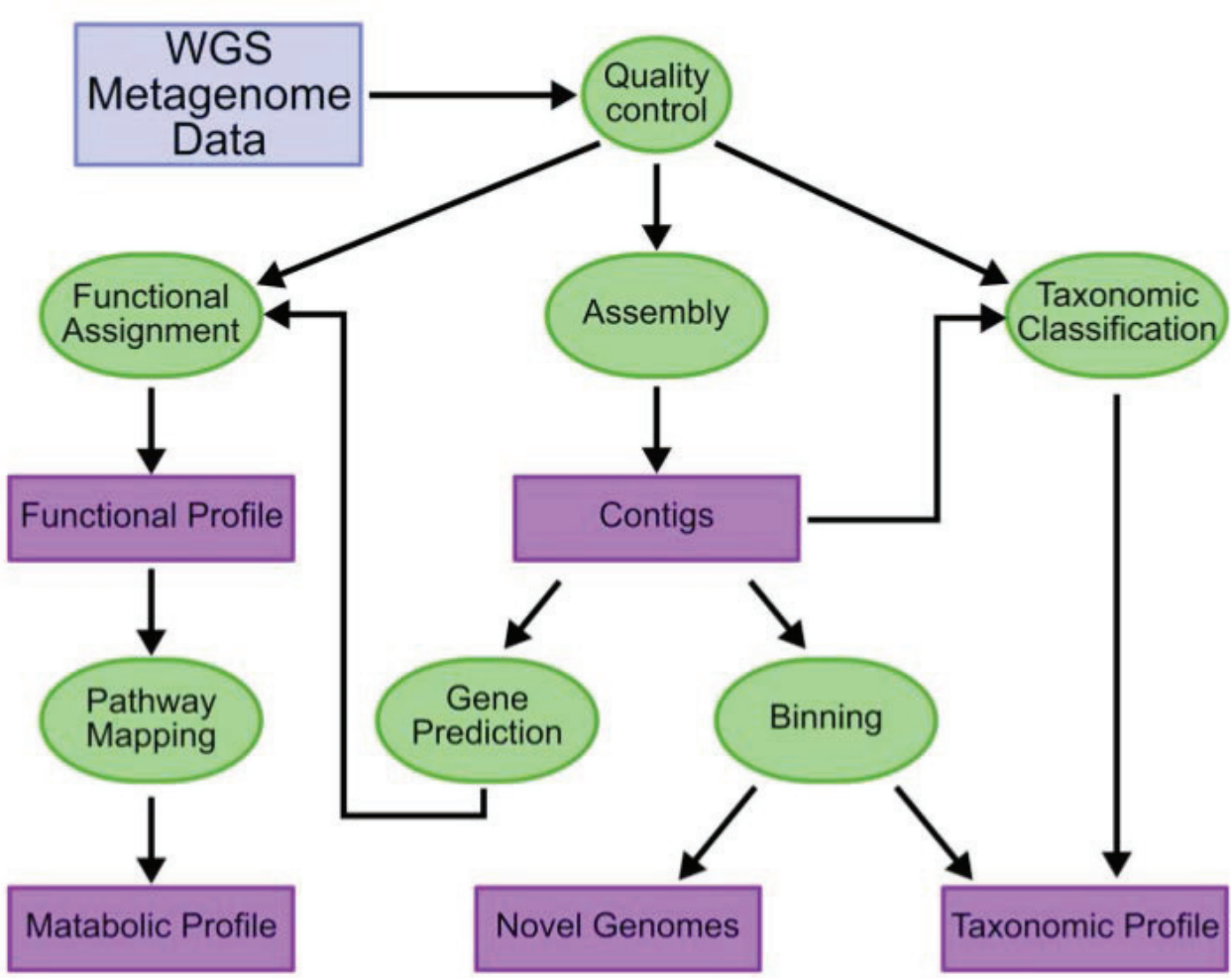


Figure 15: Schematic overview of the possible major steps in a metagenomic workflow

Square boxes represent data and results, and oval boxes represent processing steps. Adapted from (197).

Many tools are available to agglutinate reads into large contiguous segments, and most of them are based on *de-novo* assembly, defined as the reconstruction of long segments or genomes without a reference to guide the assembly, and thus, without the help of databases (198, 199). Contigs allow for multiple sequence alignments of

reads relative to a consensus sequence, and the performance of the genome assembly is evaluated based on the size of the smallest contigs in a set of contigs that makes up at least 50% of the assembly, named N50 (199).

Two main types of algorithms have been implemented in de-novo assembler tools: the greedy algorithm, which aims for a local optimum, and graph-based algorithms, which aim for a global optimum. In the former, identical or nearly identical reads are retrieved by constructing a matrix of pairwise distances between reads, and reads presenting the greatest overlap are assembled into contigs. The newly formed contigs then have their distance with each other evaluated, and aggregation is repeated until a certain distance threshold is met. An example of the greedy algorithm is the overlap-layout-consensus algorithm that was incorporated into assembly tools at the end of the 20th century to find overlaps between reads, in order to determine a layout of the reads to produce consensus sequences (200, 201). The limitations of this algorithm implementation are the time of execution when dealing with large read sets, and poor performance on repeated regions (202). In graph-based algorithms, a global optimum is reached instead of a local optimum, and these algorithms are mainly based on *de Bruijn* graph methods (**Figure 16**). *De-novo* assembly methods based on *de Bruijn* graphs rely on k-mers: the sequences are divided into pre-defined segments of size k which are overlapped to form a network of overlapping paths that interactively form the contigs, allowing the discovery and reconstruction of new genomes (198). In graph theory, k-mers are the nodes, and the overlapping parts are the edges, which can be weighted based on the overlapping length. Graph-based assemblers are nowadays the most commonly used type of *de-novo* assembly algorithm.

Graph-based *de-novo* assembly programs can be geared towards single genome assembly or towards metagenomic assembly. In the first case, the algorithm is optimized to reconstruct a single genome out of the sequencing reads but does not allow the capture of the polymorphisms among subspecies or genomic regions shared among several species. This can prevent the generation of long contigs, essential for the elucidation of full genomes, due to distinct ramifications in the *de Bruijn* graph.

A range of specialized short-read metagenome assemblers are now available, all utilizing different techniques to deal with data complexity.

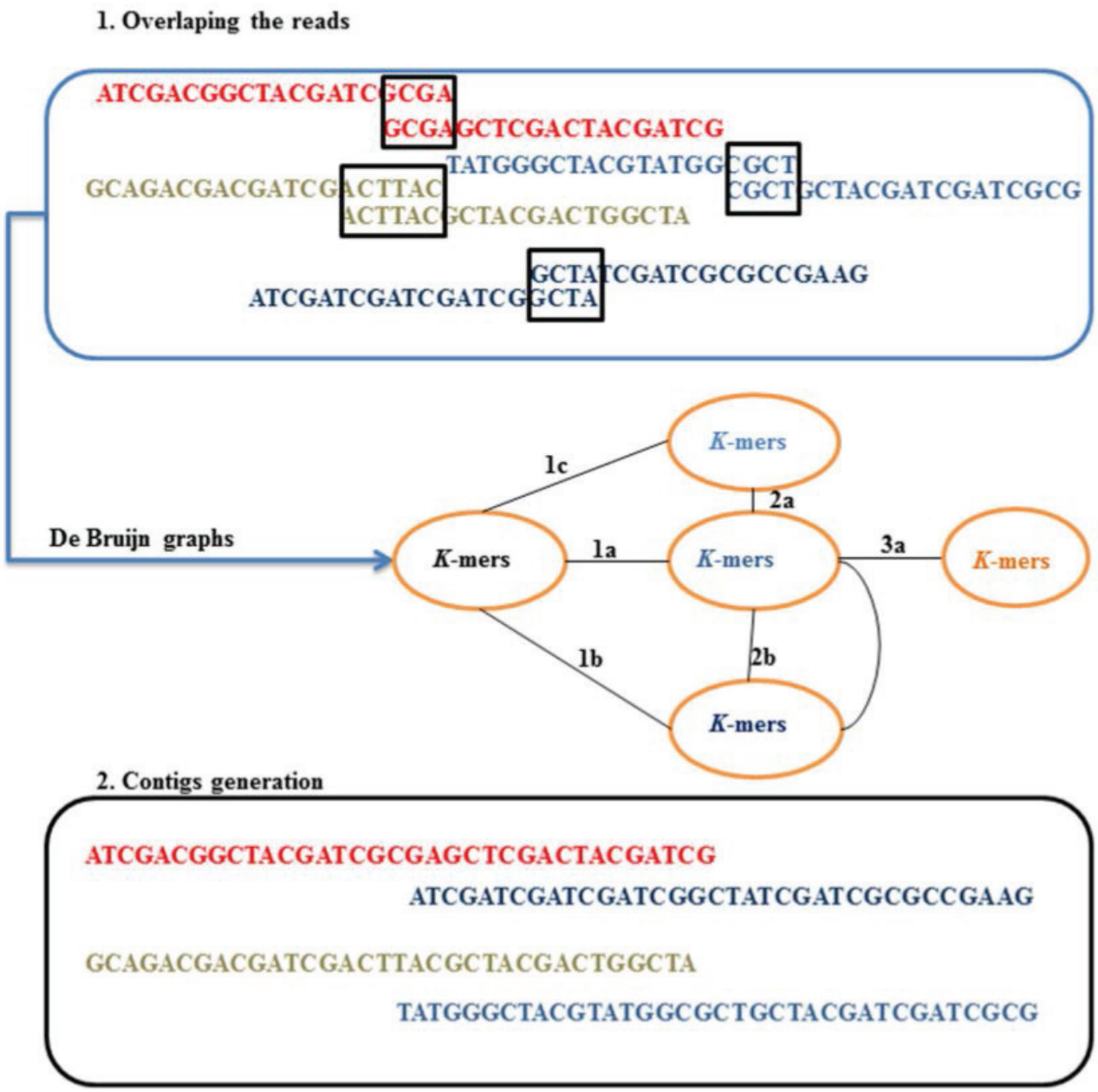


Figure 16: Genome assembly based on a de Bruijn graph

where overlaps among complementary reads are represented by *de Bruijn* graph pathways. As several genomes are present, many branches are expected in a *de Bruijn* graph, and following the pathway, the assembler identifies a large sequence. The results of the overlapped reads can be aligned against a dedicated database for taxonomic and functional annotation. Adapted from (192).

In 2011, Peng et al. developed Meta-IDBA, a metagenomic assembler for paired-end reads (203). Oriented towards the consideration that multi-species presenting polymorphic regions are present in the analyzed sample, the software identifies and

removes the branches originating from such regions with the aim of discriminating species and thus forming a *de Bruijn* graph with a set of connected components, each corresponding to a set of subspecies. Each component is transformed into a multiple alignment with a consensus sequence, representing the contigs of different polymorphisms of the same species to differentiate the diversity of microbial or viral groups in the sample. The assembler Meta-IDBA is still limited in its ability to separate low-complexity sequences from different species into independent components.

A year later, the same research group published IDBA-UD, a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth (173). Improvements in this new assembler include better handling of sequencing errors present in k-mer sequences that decrease *de Bruijn* graph accuracy as well as a better tackling of insertion/deletion (indels). A local assembly technique with paired-end read information was also used to solve the branch problem of low-depth short repeat regions.

The same year, Namiki et al. published MetaVelvet, an assembler for *de novo* metagenome assembly that prefers to decompose the multi-species *de Bruijn* graphs into individual subgraphs for each species instead of considering a unique *de Bruijn* graph incorporating information from many species (204) **(Figure 17**). Once the initial *de Bruijn* graph has been built, MetaVelvet uses the k-mer distribution frequency (node coverage) information as well as the information on the number of incoming and outgoing edges for each node to decompose the graphs into subgraphs and thus proceed with contig assembly.

Once again in 2012, a massively distributed metagenome assembler (Meta) was coupled with Ray Communities (205), which profiles microbiomes based on uniquely colored k-mers in a taxonomic tree, to related taxonomic information from the assembly into functional annotation (206). Each k-mer is directed to a higher taxon as the number of taxa with the same k-mer increases, classifying the taxa based on the closest common ancestor. This required a reference database (i.e. NCBI or GreenGenes) to color the k-mers, and thus enable taxonomic classification of the organisms as they are reconstructed.

Another software, MetAMOS, available starting in 2013, also allows assembly and taxonomic profiling, and is a concatenation of public tools covering those tasks (207). In this tool, contigs are constructed using software specific to the sequencing

platform and the library sizes are re-estimated based on the mapping of reads back to the contigs. Then, repeated regions are identified and scaffolds are assembled based on groups of contigs, followed by correction of the extended contig assemblies and detection of genomic variants. Finally, the scaffolds used to determine the taxonomic profile and functional annotation are visualized in an HTML report.

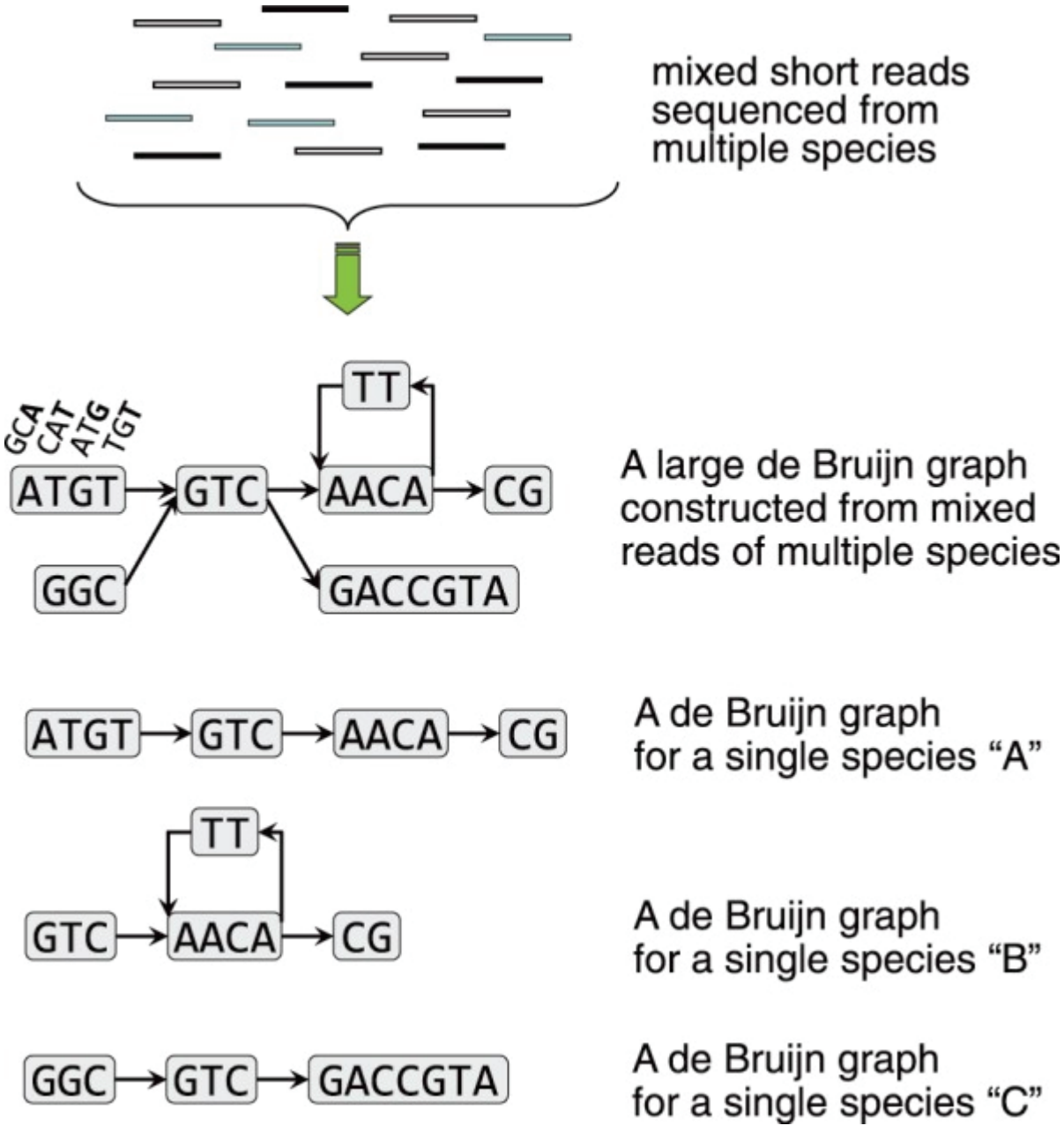


Figure 17: Decomposition of a mixed de Bruijn graph by MetaVelvet
Adapted from (204).

In 2015, computational time and memory performance were improved in MEGAHIT, through the use of succinct *de Bruijn* graphs, a compressed version of *de Bruijn* graphs (208). Moreover, MEGAHIT takes into consideration singletons, i.e. reads that do not participate in the reconstruction of any contigs, enabling better recovery of taxa with a low abundance.

One of the most recent assembly tools for metagenomics is MetaSPADES (209), which is built upon the commonly used SPAdes genome assembler (155) and combines graphs of different k values. Assemblers using a range of k-mers feature overall better performance compared to single k-mer assemblers because larger k-mers tend to facilitate the reconstruction of highly abundant genomes, whereas smaller k-mers are better suited for low abundant genomes (210). Of note, Spades includes an option to look for circular genomes, a highly important feature in the context of PV detection.

III.3 Taxonomic classification

Taxonomic classification of metagenomic reads consists in assigning sequences to a taxonomic group, and thus identifying the profile of the microbial/viral community in the analyzed sample. Two types of taxonomic classification can be applied: reference-based and reference-free.

Reference-based classification relies mainly on local alignment tools against a more or less complete database. MG-RAST was one of the first reference-based taxonomic classifiers available (211), predicting genes using FragGeneScan prediction of protein-coding regions in short read tools that combines sequencing error models and codon usages in a Hidden Markov Model (HMM) (212). Gene sequence predictions are then translated into amino acids using UCLUST (213) and aligned against the M5nr database (214) using BLAT, a BLAST-based alignment tool designed for taxonomic annotation (215). The last version of MG-RAST is available through a web server as a public resource and includes post-annotation analysis and visualization using a MG-RAST API web interface (216).

In 2016, MEGAN, a program that can work with contigs or directly with raw short reads, was developed. It also relies on BLAST (217) to compare the input sequence against known sequences (218). Using taxonomic information from the NCBI database, the tool assigns each read its Last Common Ancestor (LCA) from each hit

of the read with a reference taxon. This tool also allows integration of functional analysis by incorporating InterPro2GO annotation (219).

CARMA3 (220), an improvement of the initial CARMA (221), uses a reciprocal BLAST search to improve classification accuracy, and includes HMMER3-based variants to query the Pfam database (222). This tool is also available as a web server. All of these similarity-based methods usually offer a high classification resolution and accuracy but also lack processing speed.

Reference-based taxonomic classification tools can also rely on k-mers, making the process faster: in KRAKEN (223), fixed-length k-mers are extracted from the query sequences in a hash-based index structure built from the references, and taxonomy is inferred based on the query's individual k-mer matches to the prebuilt index. Ambiguous k-mers are assigned based on the LCA strategy, and reads are classified based on a path finder algorithm in a tree containing all matched k-mer taxa.

The main limitation of k-mer-based classifiers is the memory requirements. Thus, Centrifuge was implemented (224), indexing k-mer structures using compressed Burrows-Wheeler-transformed Ferragina-Manzini indexes. Short exact matches between a read and the index are identified and extended, and a score for each species hit is assigned (with longer segments having a higher score). Abundance of each taxa at any taxonomic rank is also computed using an Expectation-Maximization (EM) algorithm.

Reference-free classifiers are less common, and rely mainly on sequence composition, such as k-mer frequency. Frequency values can be used in supervised machine learning approaches, as developed in PhylopythiaS+, a Support Vector Machine (SVM) trained on a set of reference sequences (225). K-mer frequencies of the reads are then used to predict the taxon.

There are some other classifiers based on machine learning, but these are designed to work with full-length sequences to present accurate results (226).

III.4 Virome composition and virus discovery analysis tools

Several bioinformatic tools have been developed to analyze NGS data for the detection of viruses, but most of them are designed to analyze the virome composition of known viruses in clinical settings, or to discover new viruses from DNA or RNA shotgun sequencing.

In 2014, three bioinformatics tools for known pathogen identification in clinical samples were published, namely SURPI (Sequence-based Ultra Rapid Pathogen Identification (227), MePIC (Metagenomic Pathogen Identification for Clinical specimens) (228) and PathoScope 2.0 (229), an improvement on the initial PathoScope (230). SURPI, deployable on both cloud-based and standalone servers, focuses on process speed by proposing a “fast mode”, in contrast to its “comprehensive mode”, and relies on two existing alignment tools, SNAP (231) and RAPSearch (232). Once the common QC preprocessing step on the raw data is completed (removal of adapter, low-quality and low-complexity sequences), SURPI applies a methodology shared by nearly all bioinformatics tools for pathogen detection: the subtraction of host reads. This method consists in aligning the metagenomic reads against a database containing sequences from the host and excludes all of the reads that align against the host sequences. This method has several nomenclatures: digital subtraction, host subtraction and host read removal. In short, in its fast mode, remaining reads are aligned against a bacterial database, and reads that do not match are eventually aligned to a viral database (**Figure 18**). In its comprehensive mode, non-human reads are aligned against the full NCBI nt database, clustered based on the type of organism they matched (bacteria, fungi, viruses, etc.), and *de novo* aligned using AbySS (233) and Minimo (234). Finally, unmatched reads and contigs generated from *de novo* assembly are aligned to a viral protein database or all protein sequences in the NCBI nr collection using RAPSearch.

MePIC is a cloud-computing pipeline: once human reads have been removed, sequences are aligned against a comprehensive nucleotide database, and the strength of the tool relies on the use of the cloud for fast analysis.

Finally, Pathoscope 2.0 is based on several modules that perform all of the various computational analysis steps, including reference genome library extraction and indexing, read quality control and alignment, strain identification, and summarization and annotation of results. The main strength of the tool is the PathoID module, which uses a penalized statistical mixture model to reassign ambiguous reads to the most likely source genome, based on an EM algorithm.

These tools are designed for clinical settings, and thus favor the speed of the process over the accuracy or sensitivity of the method.

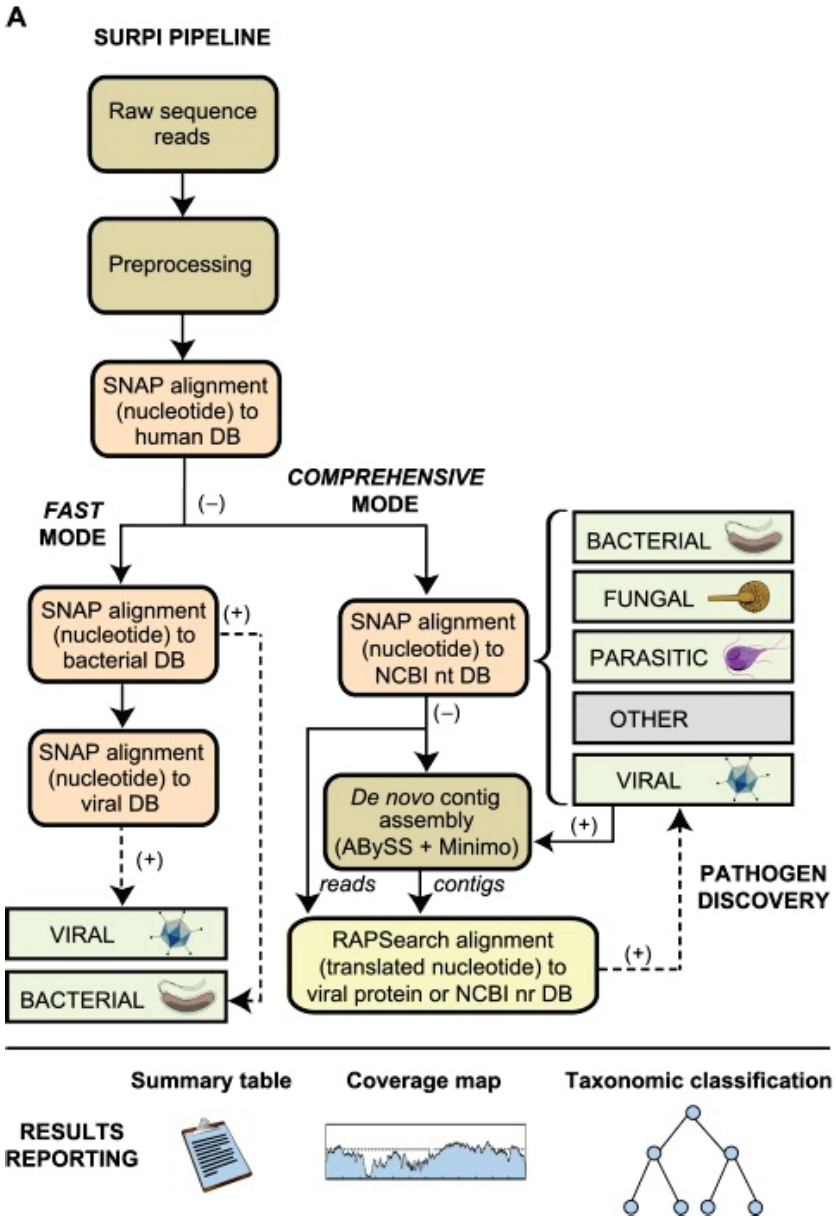


Figure 18: SURPI analysis workflow
Adapted from (227).

Metavir2 (235) and Virome (236) are two web-based tools for virome analysis. They focus heavily on data visualization of environmental samples and do not focus on virus discovery. In MetaVir2, reads are compared to the complete viral genomes of the RefSeq Virus database using BLAST, and taxonomy is inferred from the best hits. K-mer frequency bias is used to compute a hierarchical clustering and a non-metric multidimensional scaling, followed by phylogenetic tree construction. Many interactive plots are drawn at each step. Contigs can also be given as inputs and ORFs are inferred based on comparison to public database like PFAM (237). Virome

(the Viral Informatics Resource for Metagenome Exploration) classifies viral metagenome reads (predicted open-reading frames) based on homology search results. Functional and taxonomic information is derived from databases linked to the UniRef 100 database and environmental classifications are obtained from hits against a custom database named MetaGenomes On-Line. Web-server tools require an internet connection to operate and can pose concerns regarding data privacy.

CaPSID (238) is a platform made up of three components: a computational pipeline written in Python for executing digital subtraction, a core MongoDB database for storing reference sequences and alignment results, and a web application in Grails for visualizing and querying the data. The main feature is the use of the MongoDB that stores host genomes and target genomes, and it has been designed to be used as a platform for large collaborative projects, but not for pathogen discovery.

Bioinformatic solutions for virus detection have also been developed to specifically target some types of viruses, as in VirFind (239), focusing on plant virus discovery. Based on specific dsRNA amplification methods using barcoded PCR primers facilitating the multiplexing of NGS library generation, the first step of VirFind analysis is host read removal using Bowtie2 (240). *De novo* sequence assembly is then performed on unmapped reads using Velvet (241), and the resulting contigs are subjected to a Blastn search against the GenBank nt database. Sequences without any matches are then subjected to a Blastx search against all GenBank virus protein sequences. Taxonomy is inferred based on official ICTV taxonomic information, and remaining reads without matches are translated into the six possible frames before looking for conserved domains. VirFind is available through a web interface for the user to submit their sequences.

VirusDetect (242) was also initially applied to plant and animal viruses and is specifically designed for virus-derived small interfering RNA detection. The program first maps the sRNA reads to known virus reference sequences using BWA (170) and then assembles the reads into virus contigs using a reference-guided approach (**Figure 19**). *De novo* assembly of sRNAs is also performed in parallel using Velvet (241) with optimized k-mer lengths. The *de novo* assembled contigs are pooled together with those generated from reference-guided assemblies and redundancy is removed. BlastN is launched against a reference virus database to identify virus contigs, and remaining contigs are further compared against the reference virus protein sequences using the BLASTX program. The depth of each virus contig

covered by sRNA reads is calculated and normalized to reads per million (RPM), enabling abundance calculation and inference of the detected sequence's reliability.

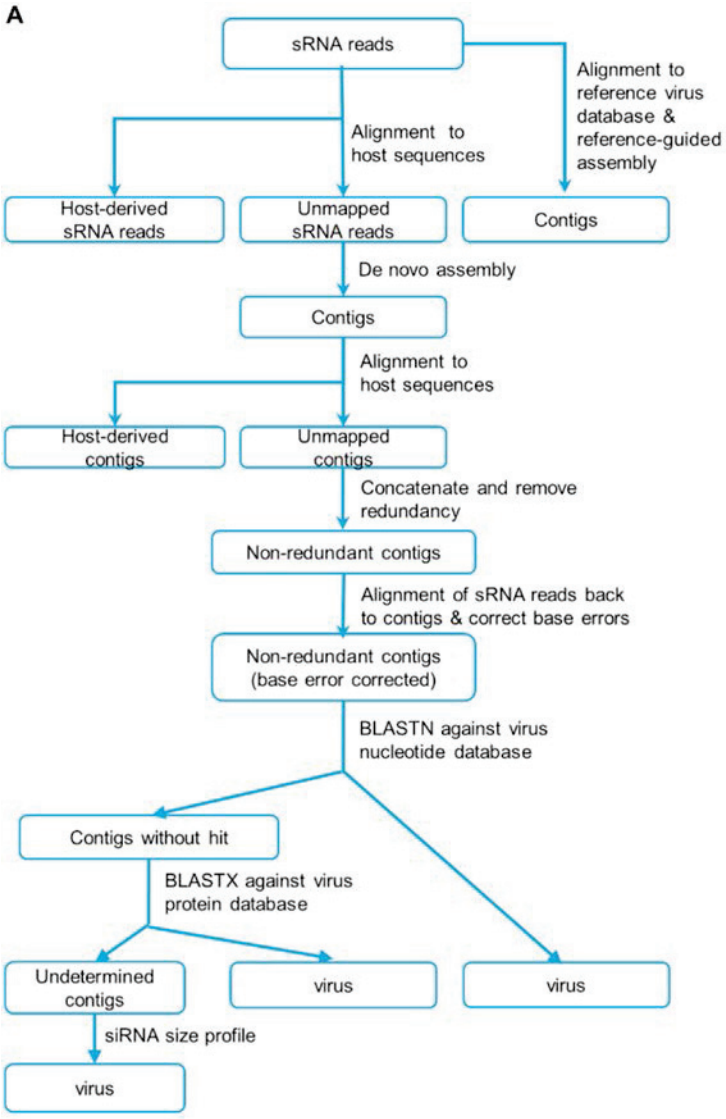


Figure 19: VirusDetect flowchart

Adapted from (242).

VirusSeq (243) and VERSE (244) are original tools using a custom reference database made up of the host genome combined with pathogens of interest genomes, enabling inference of the integration site. In both tools, computational subtraction is applied, and non-host sequences are subsequently aligned against a comprehensive database containing viral sequences from Genome Information Broker for Viruses (http://gib-v.genes.nig.ac.jp/) in VirusSeq, and against a concatenated virus-host reference genome in VERSE (containing a separate pseudo-chromosome name "chrVirus"). In the latter, inter-chromosomal structural

variants (SVs) are detected from aligned reads allowing inference of virus integration-harboring regions in the host genome, considering that high-quality consensus SNPs and indels detected from aligned reads at the different steps are used to modify the virus reference genome, thus increasing alignment accuracy. In VirusSeq, initial alignment enables quantification of identified viruses and removal of host reads. Then, remaining reads are aligned against the custom reference, and discordant read pairs (one mate aligning to the host genome and the second mate to the virus genome) that support the same integration (fusion) event are clustered and fusion candidates are reported.

VirusHunter (245) is a bioinformatic solution for novel virus identification of data coming from Roche/454 and other long read NGS platforms. CD-HIT (246) is used to remove redundant sequences using a clustering strategy, and quality control is applied, consisting of repetitive region masking with RepeatMasker (http://www.repeatmasker.org) and low-quality sequence removal. Then, remaining sequences are subjected to BLASTn alignment against the host genome and aligning sequences are removed from the analysis. Sequences retained from the previous step are queried against the NCBI nt database using BLASTn and subsequently taxonomically classified. If a sequence aligns to both a virus and a sequence derived from another organism type with the same e value, it is classified as “ambiguous”. Finally, sequences are queried against the NCBI nr database using BLASTx and findings are reported.

READSCAN (247) uses SMALT (H. Postingl 2012, personal communication; https://www.sanger.ac.uk/science/tools/smalt-0) to align chunks of sequences to chunks of k-mer indexed host and pathogen database sequences. The result of the mapping procedure is filtered and classified into several bins, namely, host, pathogen, ambiguous and unmapped. The use of indexation and many parallel chunks of sequences analyzed simultaneously allows the speed of the process to be significantly increased.

Rapid identification of non-human sequences (RINS) (248) is an intersection-based pathogen detection workflow. BLAT (215), a blast-like aligner, is used to align non-overlapping k-mers for each read against user-provided pathogen reference genomes. Reads with >80% identity are then aligned to the human genome using Bowtie (249) and reads with >97% identity are removed from the read set. Then, remaining reads are complexity filtered with LZW compression. The read set is finally

assembled into pathogen sequence contigs, and subsequently extended by mapping the original read set back to the contigs. A second alignment of the contigs against a non-human origin database is performed using BLAST. The author claims that RINS is faster than other similar tools, even though many time-consuming alignment steps are required.

PathSeq (250) is designed using the Apache Hadoop implementation of the MapReduce programming framework (http://hadoop.apache.org/mapreduce) and can be run on the Amazon Compute Cloud (EC2) (http://aws.amazon.com/ec2/). The workflow is comprised of three modules: pre-subtraction (corresponding to QC), subtraction (with MAQ (251) against a set of six human sequence databases, remaining reads being masked for repetition, and re-subtracted two times using MegaBlast and BlastN, respectively), and post-subtraction (BlastN and BlastX against NCBI viruses sequences). Finally, *de novo* assembly of the full set of remaining reads is performed before another BLAST alignment.

Finally, to our knowledge, the most recent tool for virus detection, published last year, is VirusSeeker (252), a BLAST-based NGS data analysis pipeline designed for virome composition description and novel virus discovery. Two modules have been specifically deployed for each purpose: VirusSeeker-Virome and VirusSeeker-Discovery, respectively (**Figure 20**).

.

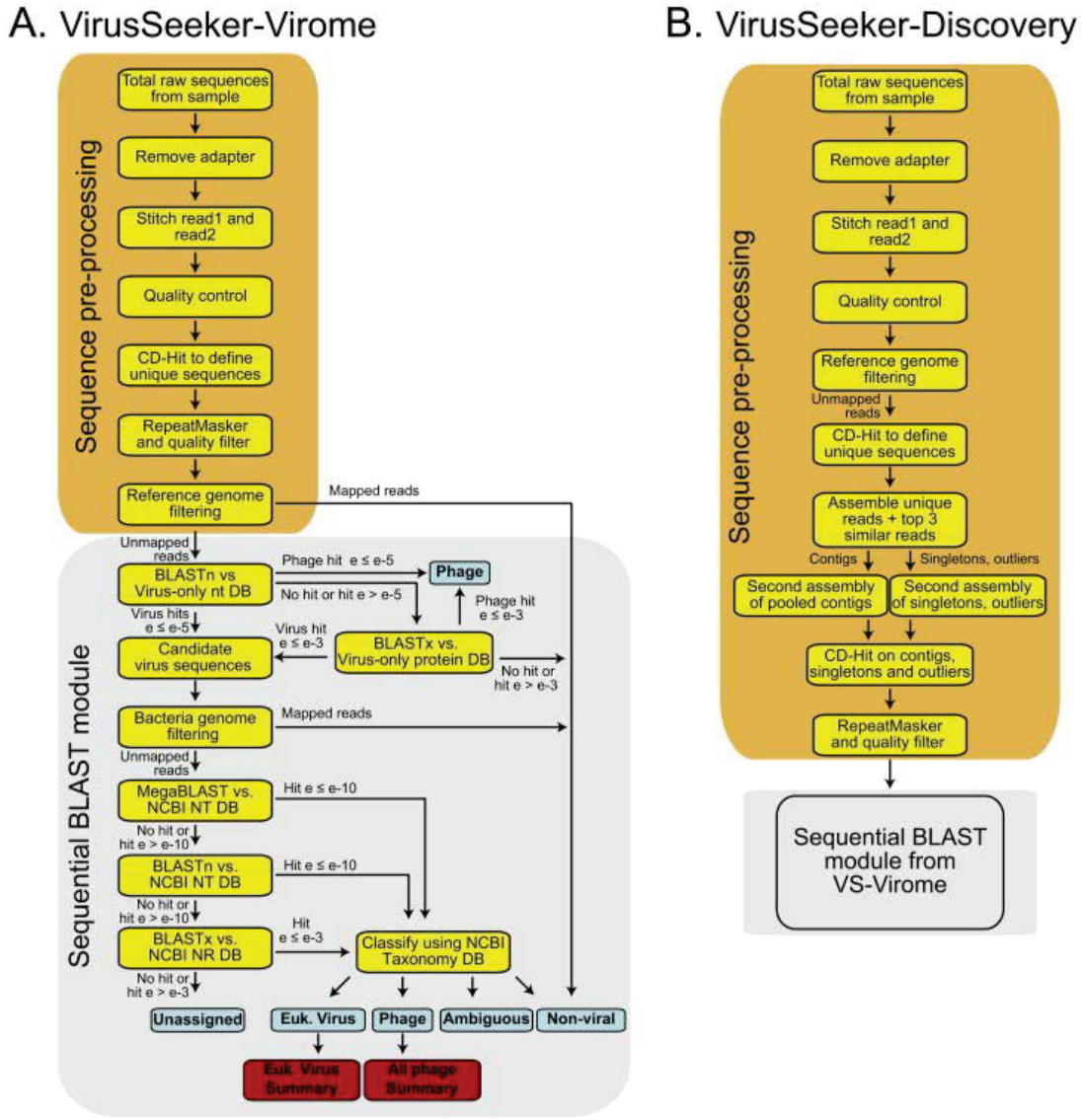


Figure 20: VS-Virome and VS-Discovery workflows

Adapted from (252)

VS-Virome consists in QC (trimming of adapter and low-quality sequences), merging of reads pairs, redundant sequence removal and masking of interspersed repeats and low-complexity DNA sequences. Next, digital subtraction is applied using BWA-MEM (170) and MegaBlast, and remaining reads are subjected to BLASTn alignment against the virus-only nucleotide database to detect those that share nucleotide sequence similarity to known viruses. Remaining sequences are then aligned using BLASTx against the virus-only protein database to detect viruses sharing protein

sequence similarity to known viruses. Candidate viral sequences are queried against the NCBI Bacteria reference genomes using BWA-MEM, against the nt database using MegaBLAST, BLASTn, and against the NCBI nr database using BLASTx. Finally, the sequences are taxonomically classified based on Blast best hit.

The VS-Discovery pipeline consists in the same pre-processing step with minor changes, and reads are assembled using Newbler, followed by two re-assemblies of contigs and singletons, respectively. CD-HIT is used to further remove redundant sequences and repeatmasker is applied. Finally, the sequential BLAST module from VS-Virome is applied.

None of these tools are specifically designed for PV detection even though, recently, a novel approach for characterization of HPV genomic variability and chromosomal integration has been published, but no associated bioinformatics tools are available yet (253). Moreover, none of these tools are designed to process amplicon data, and there is thus a need for bioinformatics tools able to detect new PVs from amplicon-based NGS data.

Aims and objectives

The *papillomaviridae* family includes a highly diverse roster of viruses having genomic features varying among the representatives, providing them with biological capabilities potentially pathogenic for their host, and classified into several genera based on their L1 gene nucleotide sequence composition. Evaluation of the likely scientific research interest for certain PV types is hard to decipher, due to taxonomy into genera and species based on the L1 gene not being in accordance with the oncogenic capabilities of the PVs, nor in agreement with the type of host organism infected. Detection of PVs, and particularly of HPVs, has always been an important public health concern, and PCR-based methods have proven to be the most sensitive technique over the years, in both clinical and fundamental research settings. This has led to the development of many amplification primer pairs, constructed by multiple alignments of a region of the L1 gene from papillomaviruses sharing the same tropisms, and sometimes bearing degenerate nucleotide bases in order to increase the range of targeted types. Over the past decades, these PCR-based methods used in combination with Sanger sequencing have enabled the detection and identification of many new HPVs, including some showing manifest oncogenic proprieties. However, the detection limits of these previously published L1 primers have been reached, and new primers are needed in order to consider the nucleotide diversity of the most recently discovered HPV types. Furthermore, thanks to the development of NGS, the sequencing power of these technologies can be used to increase experiment sensitivity and speed. Moreover, recently published metagenomic studies suggest that many more HPV types remain to be discovered and that additional research is required to characterize the biology and epidemiology of a vast number of HPV types that have been poorly investigated so far, with the final aim of clarifying their potential roles in human diseases. This requires the development of a novel laboratory protocol and specific *in-sillico* analysis tool to process the vast quantity of data generated from NGS. Though many bioinformatic tools have been developed in recent years, all have been designed to process shotgun metagenomic data, all lack specificity for the *Papillomaviridae* family, and all are oriented towards application in clinical settings. Moreover, most of them allow description of the known PV population in a biological sample and are not geared

towards discovering novel viruses. Thus, there is a need to deploy new bioinformatic methods able to handle *papillomaviridae* amplicon-based NGS data.

The aim of the thesis was to develop a new strategy for the detection of HPV in human samples, including the known PV population and potentially novel PV types. The objectives were to evaluate previously published PCR-based methodologies for HPV detection, to improve these methods in order to take into account the recently described diversity of HPVs, to deploy a strategy combining the use of amplicon sequencing with NGS, to implement a bioinformatic workflow to quickly process amplicon-based NGS data and, finally, to test the novel protocols on experimental data.

Results and application

I - HPV Genome reconstruction and identification

The following studies describe a total of five novel HPV types, comprising four Gammapapillomaviruses (**papers nr 1, 3, 4**) and one Betapapillomavirus (**paper nr 2**). In the three studies published in “Genome Announcements”, novel HPV types were identified in the skin of immunocompetent individuals, while for the study published in “Virus Research” the two novel HPV types were isolated from immunodeficient individuals. The use of RCA combined with restriction enzyme allowed the isolation of three different Gammapapillomaviruses (publications 1 and 4) as well as one Betapapillomavirus (publication 2), while FAP primer was used to amplify another novel Betapapillomavirus (publication 3). The five novel HPV types described in the publications below have since been cloned into a synthetic vector and have already obtained, or are in the process of obtaining, an official HPV number from the Karolinska HPV Reference Center, and thus are part of the official taxonomy on papillomaviruses. These studies have confirmed the use of previously published FAP primers and RCA methods, and have highlighted the need to develop a protocol using the power of NGS, the sequencing and reconstruction of the genome one by one being a slow and laborious process.

1) Dutta S., **Robitaille A**, Olivier M, Rollison DE, Tommasino M, Gheit T. (2017). Genome Sequence of a Novel Human Gammapapillomavirus Isolated from Skin. *Genome Announc.*, *5*(23), e00439-17.
2) Dutta S, **Robitaille A**, Rollison DE, Tommasino M, Gheit T. (2017). Complete Genome Sequence of a Novel Human Betapapillomavirus Isolated from a Skin Sample. *Genome Announc.*, *5*(13), e01642-16
3) Brancaccio RN, **Robitaille A**, Dutta S, Rollison DE, Fischer N, Grundhoff A, ..., Gheit T. (2017). Complete Genome Sequence of a Novel Human Gammapapillomavirus Isolated from Skin. *Genome Announc.*, *5*(34), e00833-17.
4) Dutta S, **Robitaille A**, Aubin F, Fouéré S, Galicier L, Boutboul D, ..., Gheit T. (2018). Identification and characterization of two novel Gammapapillomavirus

genomes in skin of an immunosuppressed *Epidermodysplasia Verruciformis* patient. *Virus research*, *249*, 66-68.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

VIRUSES

Genome Sequence of a Novel Human Gammapapillomavirus Isolated from Skin

Sankhadeep Dutta,[a] Alexis Robitaille,[a] Magali Olivier,[b] Dana E. Rollison,[c] Massimo Tommasino,[a] Tarik Gheit[a]

Infections and Cancer Biology Group, International Agency for Research on Cancer, Lyon, France[a]; Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, Lyon, France[b]; Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, Florida, USA[c]

ABSTRACT A new human gammapapillomavirus (HPV_MTS2) genome was isolated and fully cloned from a skin swab. The L1 open reading frame of HPV_MTS2 was 79% and 80% identical to those of its closest relatives, HPV type 149 (species Gamma-7 of the genus *Gammapapillomavirus*) and HPV isolate Dysk2 (GenBank accession no. KX781281), respectively, thus qualifying it as a new HPV type.

Received 7 April 2017 **Accepted** 18 April 2017 **Published** 8 June 2017

Citation Dutta S, Robitaille A, Olivier M, Rollison DE, Tommasino M, Gheit T. 2017. Genome sequence of a novel human gammapapillomavirus isolated from skin. Genome Announc 5:e00439-17. https://doi.org/10.1128/genomeA.00439-17.

Address correspondence to Tarik Gheit, gheitt@iarc.fr.

Human papillomaviruses (HPVs) belonging to the genus *Gammapapillomavirus* (gamma-HPVs) have traditionally been classified as cutaneotrophic (1). However, a growing body of evidence suggests a much broader tissue tropism with gamma-HPVs detected in mucocutaneous areas of the anogenital region, oral and nasal mucosa, and various cutaneous and genital lesions (2–9). With the identification of new gamma-HPV genomes, the genus *Gammapapillomavirus* has been growing rapidly in recent years and is currently divided into 27 species. Here, we report the complete genome sequence of a novel HPV type obtained from a skin swab of a healthy individual.

The complete genome of a new HPV type (HPV_MTS2; 7,319 bp) was obtained by amplifying DNA from a human forearm skin swab using multiply-primed rolling circle amplification (RCA) according to the manufacturer's instructions (illustra TempliPhi 100 amplification kit, GE Healthcare, USA). The amplified product was digested with *EcoRI* and cloned into the pUC19 vector for sequencing using a primer-walking strategy (GATC-Biotech, Germany), which covered nucleotides 647 to 6489 (5,843 bp) of the viral genome. To obtain the missing part of the viral genome, long-range PCR was performed on the RCA product template using TaKaRa LA Taq DNA polymerase and HPV_MTS2-specific primers (forward: 5′ TCCGCTTCTGTTACAATATACCA 3′; reverse: 5′ GTTTAGAAGCAGATATTCTTGC 3′). The amplicon was cloned in the pCR-XL-TOPO vector using the TOPO-XL PCR cloning kit (Invitrogen, USA) and sequenced. The sequence of the whole viral genome was confirmed by a strategy that implies the use of a proofreading *Pfu* DNA polymerase (Agilent Technologies, USA).

An HPV genome is considered to be a novel type if it shares less than 90% sequence similarity to the closest papillomavirus type in the L1 open reading frame (ORF) (1). The L1 ORF of HPV_MTS2 demonstrated 79% nucleotide homology to its closest relative, HPV type 149, belonging to species Gamma-7, and the newly identified HPV isolate Dysk2 (GenBank accession no. KX781281). However, the overall nucleotide homology between HPV_MTS2 and HPV isolate Dysk2 was 80%, and the homology between HPV_MTS2 and HPV type 149 was 79%. Overall, the G+C content of HPV_MTS2 was 37.8%. The genome contains five early (E1, E2, E4, E6, and E7) and two late (L1 and L2) ORFs, but no E5 ORF, a genomic organization typical of other gamma-HPVs. The long control region between L1 and E6 is 512 bp long and contains the TATA box (TATAAA), one polyadenylation site (AATAAA) for L1 and L2 transcripts, and four consensus palindromic E2-binding sites (ACC-N_6-GGT). Two conserved zinc-binding domains [CxxC$(x)_{29}$CxxC] separated by 36 amino acids were identified in E6 and one in E7 (10).

The consensus motif for binding to the pRB and its related proteins was observed in E7; however, serine was substituted for cysteine, thus forming the LxSxE motif (10). Such a modified LxSxE motif is common among members of the genus *Gammapapillomavirus*. In the carboxy terminus of E1 we identified a GPPNTGKS motif to be the putative ATP-binding site. Moreover, the E1 protein contained a cyclin interaction RXL motif required for viral replication (11).

To conclude, the genetic characterization of HPV_MTS2 expands the species composition of gamma-HPVs.

Accession number(s). The complete genome sequence of HPV_MTS2 is available in GenBank under the accession number KY780961.

ACKNOWLEDGMENTS

This work was partially supported by the European Commission FP7 Marie Curie Actions–People: Co-funding of Regional, National and International Programs (COFUND); by the National Cancer Institute R01 (grant no. 1-R01-CA177586-01A1 to D.E.R.); and by Foundation ARC (grant no. PJA 20151203192 to M.T.).

REFERENCES

1. de Villiers EM. 2013. Cross-roads in the classification of papillomaviruses. Virology 445:2–10. https://doi.org/10.1016/j.virol.2013.04.023.
2. Moscicki AB, Ma Y, Gheit T, McKay-Chopin S, Farhat S, Widdice LE, Tommasino M. 2016. Prevalence and transmission of beta and gamma human papillomavrius in heterosexual couples. Open Forum Infect Dis 4:ofw216. https://doi.org/10.1093/ofid/ofw216.
3. Bottalico D, Chen Z, Dunne A, Ostoloza J, McKinney S, Sun C, Schlecht NF, Fatahzadeh M, Herrero R, Schiffman M, Burk RD. 2011. The oral cavity contains abundant known and novel human papillomaviruses from the *Betapapillomavirus* and *Gammapapillomavirus* genera. J Infect Dis 204: 787–792. https://doi.org/10.1093/infdis/jir383.
4. Ekström J, Bzhalava D, Svenback D, Forslund O, Dillner J. 2011. High throughput sequencing reveals diversity of human papillomaviruses in cutaneous lesions. Int J Cancer 129:2643–2650. https://doi.org/10.1002/ijc.26204.
5. Foulongne V, Sauvage V, Hebert C, Dereure O, Cheval J, Gouilh MA, Pariente K, Segondy M, Burguière A, Manuguerra JC, Caro V, Eloit M. 2012. Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. PLoS One 7:e38499. https://doi.org/10.1371/journal.pone.0038499.
6. Ma Y, Madupu R, Karaoz U, Nossa CW, Yang L, Yooseph S, Yachimski PS, Brodie EL, Nelson KE, Pei Z. 2014. Human papillomavirus community in healthy persons, defined by metagenomics analysis of human microbiome project shotgun sequencing data sets. J Virol 88:4786–4797. https://doi.org/10.1128/JVI.00093-14.
7. Antonsson A, Erfurt C, Hazard K, Holmgren V, Simon M, Kataoka A, Hossain S, Håkangård C, Hansson BG. 2003. Prevalence and type spectrum of human papillomaviruses in healthy skin samples collected in three continents. J Gen Virol 84:1881–1886. https://doi.org/10.1099/vir.0.18836-0.
8. Sichero L, Pierce Campbell CM, Ferreira S, Sobrinho JS, Luiza Baggio M, Galan L, Silva RC, Lazcano-Ponce E, Giuliano AR, Villa LL. 2013. Broad HPV distribution in the genital region of men from the HPV infection in men (HIM) study. Virology 443:214–217. https://doi.org/10.1016/j.virol.2013.04.024.
9. Forslund O, Johansson H, Madsen KG, Kofoed K. 2013. The nasal mucosa contains a large spectrum of human papillomavirus types from the *Betapapillomavirus* and *Gammapapillomavirus* genera. J Infect Dis 208: 1335–1341. https://doi.org/10.1093/infdis/jit326.
10. Tommasino M. 2017. The biology of beta human papillomaviruses. Virus Res 231:128–138. https://doi.org/10.1016/j.virusres.2016.11.013.
11. Ding Q, Li L, Whyte P. 2013. Human papillomavirus 18 E1^E4 protein interacts with cyclin A/CDK 2 through an RXL motif. Mol Cell Biochem 373:29–40. https://doi.org/10.1007/s11010-012-1472-y.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

VIRUSES

Complete Genome Sequence of a Novel Human Betapapillomavirus Isolated from a Skin Sample

Sankhadeep Dutta,[a] Alexis Robitaille,[a] Dana E. Rollison,[b] Massimo Tommasino,[a] Tarik Gheit[a]

Infections and Cancer Biology Group, International Agency for Research on Cancer, Lyon, France[a]; Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, Florida, USA[b]

ABSTRACT We report the genetic characterization of a new papillomavirus (HPV_MTS1) isolated and fully cloned from a skin swab. The L1 open reading frame of HPV_MTS1 was 85% identical to its closest human papillomavirus (HPV) type 80, which belongs to the species beta-2 of the genus *Betapapillomavirus*, hence qualifying it as a new HPV type.

Received 22 December 2016 **Accepted** 24 January 2017 **Published** 30 March 2017

Citation Dutta S, Robitaille A, Rollison DE, Tommasino M, Gheit T. 2017. Complete genome sequence of a novel human betapapillomavirus isolated from a skin sample. Genome Announc 5:e01642-16. https://doi.org/10.1128/genomeA.01642-16.

Address correspondence to Tarik Gheit, gheitt@iarc.fr.

Human papillomaviruses (HPVs) that belong to the *Betapapillomavirus* genus (beta-HPVs) are broadly classified as cutaneotrophic viruses (1). More than 50 beta-HPV types have been identified to date and are widely prevalent in the skin of normal individuals (2, 3). However, this list continues to expand as many new beta-HPVs have recently been isolated from specimens of skin, oral cavities, and other anatomical sites (4, 5). Here, we report the complete genomic sequence of a novel HPV type obtained from the skin swab of a healthy individual.

The complete genome of a new HPV type (HPV_MTS1; 7,405 bp) was obtained by amplifying skin swab DNA using multiply primed rolling circle amplification (RCA) according to the manufacturer's instructions (Illustra TempliPhi 100 amplification kit, GE Healthcare, Piscataway, NJ). The amplified product was digested with *EcoRI* and cloned into the pUC19 vector for sequencing using the primer-walking strategy (GATC Biotech, Germany) which covered nucleotides (nt) 842 to 6,935 (6,094 bp) of the viral genome. Furthermore, a long-range PCR was performed on the RCA product as a template using TaKaRa LA Taq DNA polymerase (TaKaRa Bio Inc.) and HPV_MTS1 specific primers (forward: 5′-CATATTTTGTACCTTTGTCGTC-3′ and reverse: 5′-GTAAAGTACCTTTAAAAGC GGA-3′) to obtain the remaining part of the genome. Amplification was performed for 35 cycles at 94°C for 30 s, 56°C for 30 s, and 72°C for 5 min and cloned in pCR-XL-TOPO vector using the TOPO XL PCR cloning kit (Invitrogen, Carlsbad, CA) and sequenced. The sequence was checked using the proof reading Pfu DNA polymerase (Agilent Technologies, Santa Clara, CA, USA).

A novel HPV type shares less than 90% sequence similarity to the closest papillomavirus type in the L1 open reading frame (ORF) (1). Pairwise comparison of the L1 ORF of HPV_MTS1 demonstrated 85% nucleotide homology to its closest HPV type 80, which belongs to the genus *Betapapillomavirus*, species beta-2. The overall nucleotide homology between HPV_MTS1 and HPV80 was 89% with a G+C content of 39.8%. The genomic organization of this virus is typical of cutaneotrophic HPVs, containing five early (E1, E2, E4, E6, and E7) and two late (L1 and L2) genes but no E5 ORF. The long control region (LCR) is 384 bp and located between the L1 and E6 genes. LCR contains one polyadenylation site (AATAAA) for L1 and L2 transcripts, four consensus palindromic E2-binding sites (ACC-N_6-GGT = 2 and slightly modified ACC-N_5-GGT = 2) and two TATA Boxes (TATAAA) for the downstream early promoter. The two conserved zinc-binding domains of the viral E6 protein [CxxC$(x)_{29}$CxxC and CxxC$(x)_{30}$CxxC] are

separated by 36 amino acids (6). The zinc-binding domain [CxxC(x)$_{29}$CxxC] and the LxCxE motif for interaction with pRB and its related proteins are present in the E7 protein (6). The presence of ATP-binding site (GPPDTGKS) in the carboxy-terminal of the E1 protein confirms its ATP-dependent helicase activity. The E4 protein is encoded from an internal start codon of the E2 ORF and completely overlaps it.

To conclude, the genetic characterization of HPV_MTS1 expands the species composition of beta-2 papillomaviruses.

Accession number(s). The complete genome sequence of HPV_MTS1 is available in GenBank under the accession no. KY349817.

ACKNOWLEDGMENTS

This work was partially supported by the European Commission FP7 Marie Curie Actions–People-Co-funding of regional, national, and international programs (COFUND); National Cancer Institute R01 grant (R01-CA177586-01A1) to D.E.R., and Foundation ARC (PJA 20151203192) to M.T.

REFERENCES

1. de Villiers EM. 2013. Cross-roads in the classification of papillomaviruses. Virology 445:2–10. https://doi.org/10.1016/j.virol.2013.04.023.
2. Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, Gopalan V, Mohamoud Y, Huyen Y, McBride AA. 2013. The papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. Nucleic Acids Res 41:D571–D578. https://doi.org/10.1093/nar/gks984.
3. Bottalico D, Chen Z, Dunne A, Ostoloza J, McKinney S, Sun C, Schlecht NF, Fatahzadeh M, Herrero R, Schiffman M, Burk RD. 2011. The oral cavity contains abundant known and novel human papillomaviruses from the *Betapapillomavirus* and *Gammapapillomavirus* genera. J Infect Dis 204:787–792. https://doi.org/10.1093/infdis/jir383.
4. Kocjan BJ, Hosnjak L, Seme K, Poljak M. 2013. Complete genome sequence of a novel human *Betapapillomavirus*, HPV-159. Genome Announc 1(3):e00298-13. https://doi.org/10.1128/genomeA.00298-13.
5. Kocjan BJ, Steyer A, Sagadin M, Hosnjak L, Poljak M. 2013. Novel human papillomavirus type 174 from a cutaneous squamous cell carcinoma. Genome Announc 1(4):e00445-13. https://doi.org/10.1128/genomeA.00445-13.
6. Tommasino M. 14 November 2016. The biology of beta human papillomaviruses [Epub ahead of print.]. Virus Res https://doi.org/10.1016/j.virusres.2016.11.013.

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

VIRUSES

Complete Genome Sequence of a Novel Human Gammapapillomavirus Isolated from Skin

Rosario N. Brancaccio,[a] Alexis Robitaille,[a] Sankhadeep Dutta,[a] Dana E. Rollison,[b] Nicole Fischer,[c,d] Adam Grundhoff,[c,e] Massimo Tommasino,[a] Tarik Gheit[a]

Infections and Cancer Biology Group, International Agency for Research on Cancer, Lyon, France[a]; Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, Florida, USA[b]; German Center for Infection Research, Hamburg, Borstel, Lübeck, Riems, Germany[c]; Institute for Medical Microbiology, Virology and Hygiene, University Medical Center Hamburg-Eppendorf, Hamburg, Germany[d]; Heinrich Pette Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany[e]

ABSTRACT A novel human papillomavirus (HPV ICB1) was fully characterized from a skin swab by using a sensitive degenerate PCR protocol combined with next-generation sequencing. The L1 open reading frame of HPV ICB1 shares 70.54% nucleotide homology with its closest relative, HPV164, and thus constitutes a novel human gammapapillomavirus.

Received 4 July 2017 **Accepted** 18 July 2017 **Published** 24 August 2017

Citation Brancaccio RN, Robitaille A, Dutta S, Rollison DE, Fischer N, Grundhoff A, Tommasino M, Gheit T. 2017. Complete genome sequence of a novel human gammapapillomavirus isolated from skin. Genome Announc 5:e00833-17. https://doi.org/10.1128/genomeA.00833-17.

Address correspondence to Tarik Gheit, gheitt@iarc.fr.

Human papillomaviruses (HPVs) are nonenveloped double-stranded DNA viruses approximately 8 kb in size with an epithelial tropism. HPVs colonize normal skin and mucosa and can induce cutaneous and mucosal lesions (1–4). The L1 gene is well conserved among the papillomaviruses, and thus, it is used for taxonomic classification (5, 6). Here, we report the complete genome sequence of a novel HPV type isolated from a skin swab from a healthy individual.

Degenerate PCR primers (7) were used to screen a cohort of skin samples. The amplicons were purified, pooled, and sequenced by next-generation sequencing (NGS) using the NEBNext Ultra DNA library prep kit and MiSeq reagent kit version 2 (Illumina). NGS analysis revealed the presence of a sequence of approximately 205 bp from a putative new HPV.

The complete viral genome of a new HPV type (HPV ICB1, 7,233 bp), with a G+C content of 38.09%, was obtained by DNA amplification using multiply primed rolling circle amplification (RCA) according to the manufacturer's instructions (illustra TempliPhi 100 amplification kit; GE Healthcare, USA). RCA was combined with long-range PCR (LA Taq polymerase; TaKaRa Bio, Japan) performed with outward-directed primers specific for the putative new HPV (forward primer, 5′-CATTTTGCTCATCATCAC ATGGCC-3′; reverse primer, 5′-CTGGTGACTGTCCTCCTATCC-3′). An amplicon of approximately 8 kb in size was cloned in the pCR-XL-TOPO vector using the TOPO-XL PCR cloning kit (Invitrogen, USA) and sequenced by a primer walking strategy (GATC Biotech, Germany). The sequence was validated using a proofreading polymerase, followed by Sanger sequencing.

HPV L1 sequences that share less than 90% sequence similarity to the closest papillomavirus type are traditionally considered to be distinct HPV types (5, 8). The L1 open reading frame (ORF) of HPV ICB1 showed 70.54% nucleotide homology (9) with its closest HPV type, HPV164, belonging to species gamma-8 (GenBank accession no. JX413106). In addition, according to a BLASTn search, the overall nucleotide homology between HPV ICB1 and HPV119 (gamma-8; GenBank accession no. GQ845441) was 69%. Analysis of the HPV ICB1 genome showed the presence of five early (E1, E2, E4, E6, and E7) and two late (L1 and L2) ORFs. The E5 ORF was absent. The long control region between L1 and E6 has a length of 514 bp and contains the TATA box (TATAAA), one

polyadenylation site (AATAAA) for L1 and L2 transcripts, and four consensus palindromic E2-binding sites (ACC-N_6-GGT). Like all HPV types, E6 and E7 have zinc-binding domains [CxxC(x)29CxxC] containing two and one zinc-binding domains, respectively. In addition, E7 contains an LxSxE retinoblastoma (RB)-binding motif (10). Analysis of the E1 ORF revealed the presence of a putative ATP-binding site of the ATP-dependent helicase, a GPPDTGKS motif (11). Moreover, two cyclin interaction RXL motifs (10, 11) have been localized in the E1 protein. In conclusion, analysis of the complete nucleotide sequence showed that HPV ICB1 shares the features of other known gammapapillomaviruses.

Accession number(s). The complete genome sequence of HPV ICB1 is available in GenBank under the accession number MF356498.

ACKNOWLEDGMENTS

We are grateful to Karen Müller for editing the manuscript. We thank Malik Alawi for bioinformatics support.

This work was partially supported by the National Cancer Institute (grant no. R01-CA177586-01A1) to D.E.R. and by Foundation ARC (grant no. PJA 20151203192) to M.T.

REFERENCES

1. Antonsson A, Forslund O, Ekberg H, Sterner G, Hansson BG. 2000. The ubiquity and impressive genomic diversity of human skin papillomaviruses suggest a commensalic nature of these viruses. J Virol 74:11636–11641. https://doi.org/10.1128/JVI.74.24.11636-11641.2000.
2. Antonsson A, Erfurt C, Hazard K, Holmgren V, Simon M, Kataoka A, Hossain S, Håkangård C, Hansson BG. 2003. Prevalence and type spectrum of human papillomaviruses in healthy skin samples collected in three continents. J Gen Virol 84:1881–1886. https://doi.org/10.1099/vir.0.18836-0.
3. Doorbar J. 2005. The papillomavirus life cycle. J Clin Virol 32:7–15. https://doi.org/10.1016/j.jcv.2004.12.006.
4. Li L, Barry P, Yeh E, Glaser C, Schnurr D, Delwart E. 2009. Identification of a novel human *Gammapapillomavirus* species. J Gen Virol 90:2413–2417. https://doi.org/10.1099/vir.0.012344-0.
5. de Villiers EM, Fauquet C, Broker TR, Bernard HU, zur Hausen H. 2004. Classification of papillomaviruses. Virology 324:17–27. https://doi.org/10.1016/j.virol.2004.03.033.
6. de Villiers EM, Gunst K. 2009. Characterization of seven novel human papillomavirus types isolated from cutaneous tissue, but also present in mucosal lesions. J Gen Virol 90:1999–2004. https://doi.org/10.1099/vir.0.011478-0.
7. Forslund O, Antonsson A, Nordin P, Stenquist B, Hansson BG. 1999. A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. J Gen Virol 80:2437–2443. https://doi.org/10.1099/0022-1317-80-9-2437.
8. de Villiers EM. 2013. Cross-roads in the classification of papillomaviruses. Virology 445:2–10. https://doi.org/10.1016/j.virol.2013.04.023.
9. Van Doorslaer K, Li Z, Xirasagar S, Maes P, Kaminsky D, Liou D, Sun Q, Kaur R, Huyen Y, McBride AA. 2017. The papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Res 45:D499–D506. https://doi.org/10.1093/nar/gkw879.
10. Tommasino M. 2017. The biology of beta human papillomaviruses. Virus Res 231:128–138. https://doi.org/10.1016/j.virusres.2016.11.013.
11. Wohlschlegel JA, Dwyer BT, Takeda DY, Dutta A. 2001. Mutational analysis of the Cy motif from p21 reveals sequence degeneracy and specificity for different cyclin-dependent kinases. Mol Cell Biol 21:4868–4874. https://doi.org/10.1128/MCB.21.15.4868-4874.2001.

Virus Research 249 (2018) 66–68

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Identification and characterization of two novel *Gammapapillomavirus* genomes in skin of an immunosuppressed Epidermodysplasia Verruciformis patient

Sankhadeep Dutta[a], Alexis Robitaille[a], François Aubin[b], Sébastien Fouéré[c], Lionel Galicier[d], David Boutboul[d], Fabiola Luzi[e], Paola Di Bonito[f], Massimo Tommasino[a], Tarik Gheit[a,*]

[a] Infections and Cancer Biology Group, International Agency for Research on Cancer, Lyon, France
[b] Dermatology Department and EA3181, Centre National de Référence HPV, Centre Hospitalier Universitaire, Université de Franche-Comté, Besançon, France
[c] STD Unit (Centre des MST) Dermatology Department (Service de Dermatologie), Saint Louis University Hospital, Paris, France
[d] Department of Clinical Immunology, Hôpital Saint-Louis, Assistance Publique Hôpitaux de Paris (APHP), Paris, France
[e] Plastic and Reconstructive Surgery, San Gallicano Dermatologic Institute, IRCCS, Rome, Italy
[f] Department of Infectious Diseases, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy

ARTICLE INFO

Keywords:
New gamma-HPV
Human
Epidermodysplasia Verruciformis
Immunosuppression

ABSTRACT

Two novel human gamma-papillomavirus genomes (HPV_MTS3, and HPV_MTS4) were isolated from the skin of an immunosuppressed, late-onset Epidermodysplasia Verruciformis patient and fully cloned. The L1 open reading frames of HPV_MTS3 and HPV_MTS4 were 77% and 91% identical to their closest HPV full genome isolates w18c39 and EV03c60, which belong to the species gamma-22and gamma-7 of the genus *Gammapapillomavirus*, respectively.

Human papillomaviruses (HPVs) belonging to the beta and gamma genus have traditionally been classified as cutaneotropic types (de Villiers, 2013). However, a growing body of evidences suggest a much broader tissue tropism with these cutaneotropic HPVs detected in muco-cutaneous areas of anogenital region, oral and nasal mucosa, and various cutaneous and genital lesions (Moscicki et al., 2017; Bottalico et al., 2011; Ekstrom et al., 2011; Foulongne et al., 2012; Ma et al., 2014; Antonsson et al., 2003; Sichero et al., 2013; Forslund et al., 2013; Tommasino, 2017). Infection and prevalence of cutaneotropic-HPV is also reported to synergize with environmental factors, such as UV-radiations and individual immune status (Accardi and Gheit, 2014). Epidermodysplasia Verruciformis (EV) is a rare primarily autosomal recessive disorder in which the patients develop Pityriasis versicolor-like lesions along with atypical, flat-topped coalescing warts. Such warts frequently progress to cutaneous squamous cell carcinoma (SCC) in the sun-exposed areas of the body (Tommasino, 2017; McDermott et al., 2009; Vohra et al., 2010; Androphy et al., 1985; Orth, 2006). Furthermore, several cutaneotropic HPV types have been found in 90% of SCC lesions in EV patients which believed to synergize as co-carcinogen with UVR (Patel et al., 2010).

Here, we report the complete genomic sequence of two novel HPVs isolated from cutaneous lesions in a T-Cell lymphoma patient. The patient was a 29-years-old Kurdish woman with late onset EV-like phenotype and primary combined immune deficiency, characterized by naïve T-Cell and memory B-Cell lymphopenia and low immunoglobulin levels.

Previously, in an independent study, systemic infection of oncogenic alpha-HPV type 39, beta-HPV types 5, and 38, several gamma-HPV types, the Merkel Cell Polyomavirus, and herpesviruses EBV, HHV6 and HHV7 was detected in this patient (Fouéré et al., 2017 and unpublished data). This observation instigates us for further investigation on the possible presence of novel viral sequences in this patient. We extracted the DNA from cytobrush samples, collected from the flat warts on the back of hand, eyebrows, oral mucosa and vulva (Schowalter et al., 2010). Next, the DNA samples were subjected to multiple primed Rolling Circle Amplification (RCA) using the Illustra TempliPhi 100 Amplification Kit according to the manufacturer's recommendations (GE Healthcare, Piscataway, NJ), with supplementation of 450 µM dNTPs (Johne et al., 2009). The amplified products were digested with *EcoRI*; the ~7-8Kb bands obtained from the eyebrow hair DNA were cloned into pUC19 vector and sequenced using primer-walking strategy (GATC Biotech, Germany). Furthermore, a long-range PCR was performed on the RCA product for amplification of remaining part of one of the viral genomes using the Takara LA Taq HS polymerase (Takara

* Corresponding author at: Infections and Cancer Biology Group, International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France.
E-mail address: icb@iarc.fr (T. Gheit).

https://doi.org/10.1016/j.virusres.2018.03.003
Received 5 December 2017; Received in revised form 3 March 2018; Accepted 4 March 2018
Available online 08 March 2018
0168-1702/

Table 1
Characteristic features of HPV-MTS3 and HPV-MTS4 full genomes.

		HPV Type	
		MTS3	MTS4
	Genome size (bp)	7223	7320
	G + C content (%)	37.07	38.22
ORF			
E6	Nucleotide positions	1–423	1–414
	Zinc-binding domain [CxxC(x)29CxxC]	Two	Two
E7	Nucleotide positions	420–719	414–698
	Zinc-binding domain [CxxC(x)29CxxC]	One	One
	pRb binding motif	LxCxE	LxSxE
E1	Nucleotide positions	703–2499	688–2520
	C-terminus ATP-binding motif	GPSDTGKS	GPPNTGKS
	Cyclin interacting RXL motif	Present	Present
E2	Nucleotide positions	2435–3634	2462–3661
	E2 ORF is encoded from an internal start codon of E4	Yes	Yes
E4	Nucleotide positions	2967–3383	3039–3419
L2	Nucleotide positions	3637–5187	3664–5235
L1	Nucleotide positions	5198–6760	5171–6808
LCR	Nucleotide positions	6761–7223	6809–7320
	Polyadenylation site (AATAAA)	One	One
	Consensus palindromic E2-binding sites (ACC-N_6-GGT)	Four	Four
	TATA Box	Two	One
GenBank accession number		MG063749	MG520499

Bio Inc.) and specific sets of primers. The PCR product was cloned in pCR-XL-TOPO® vector using TOPO® XL PCR Cloning Kit (Invitrogen, Carlsbad, CA) and sequenced. The sequence has been confirmed using the proof reading Pfu DNA Polymerase (Agilent Technologies, Santa Clara, CA, USA).

The complete genomes of two novel HPVs (provisionally named, HPV_MTS3; 7223 bp and HPV_MTS4; 7320 bp) were identified from the eyebrow swab DNA samples using multiply primed RCA. A novel HPV type shares less than 90% sequence similarity to the closest papillomavirus type in the L1 open reading frame (ORF) (de Villiers, 2013). A BLASTn search against the Nucleotide collection nr/nt database version December 2017 revealed that the L1 ORF of HPV_MTS3 shares 77% nucleotide similarity to its closest HPV isolate w18c39, species Gamma-22 (accession number: MF588741). The L1 ORF of HPV_MTS4 showed 91% similarity with the HPV isolate EV03c60, species Gamma-7 (accession number: MF588699). The genomic sequences from these two novel HPVs were described in Table 1. The complete genome sequences of HPV_MTS3 and HPV_MTS4 are available in GenBank under the accession number MG063749 and MG520499, respectively.

To investigate the evolutionary history of HPV_MTS3 and HPV_MTS4, we constructed Maximum-likelihood (ML) phylogenetic tree using MEGA7 based on the MUSCLE alignment of the full length L1 ORF nucleotide sequences of 458 papillomaviruses (PVs) (288 full L1 ORF from HPV and 170 full L1 ORF from animal PVs) retrieved from PaVE database in January 2018 (Supplementary Fig. S1) (Van Doorslaer et al., 2017; Kumar et al., 2016; Nei and Kumar, 2000; Edgar, 2004). The alignments of the nucleotide and amino acid sequences are available as supplementary data S1a and S1b, respectively. HPV_MTS3 and HPV_MTS4 clustered significantly with HPV types belonging to the genus *Gammapapillomavirus,* and do not clustered with animal PVs. Furthermore, a comprehensive ML phylogenetic analysis, consisting of 159 full genomes of gamma-HPV, confirmed the HPV isolates w18c39 and EV03c60 as among the closest relatives to HPV_MTS3 and HPV_MTS4, respectively (Supplementary Fig. S2).

Furthermore, to study the molecular evolution of HPV_MTS3 and HPV_MTS4, a partition scheme was also explored consisting of six partitions corresponding to the E6, E7, E1, E2, L2 and L1 genes. The multigene genome partition scheme encompassed 159 multigene genome nucleotide sequences of *Gammapapillomavirus*. The MUSCLE multigene genome alignment of the 159 *Gammapapillomavirus* reference genomes is available at nucleotide level (Supplementary Data S2). The tree obtained by a ML phylogenetic inference confirmed the HPV isolates w18c39 and EV03c60 as among the closest relatives to HPV_MTS3 and HPV_MTS4, respectively (Supplementary Fig. S3).

Table 2
Primers used for HPV_MTS3 and HPV_MTS4 DNA screening.

Name	Oligonucleotide sequence (5′-3′)
MTS3 E6/E7 sense	CATCAGTGCTGTCGTCAAACAA
MTS3 E6/E7 anti-sense	AGGACCAGAATGTAATAGTGGT
MTS3 L1 sense	GACAGAAGTAAAGATCAA
MTS3 L1 anti-sense	TAGTTCTTAAAATCCTTAGCTTTGTATG
MTS4 E6/E7 sense	TTCAGCTTCAGTCACAACATAC
MTS4 E6/E7 anti-sense	CTTATCAGGATTTGTGTGCATTC
MTS4 L1 sense	TTGTTCCTCCTCCTGCTGGTGGAA
MTS4 L1 anti-sense	CCCTGAACTGCAGACCGTTTGCG

All the details regarding the length of the alignments, number of distinct patterns, the models of substitution, the rate of heterogeneity parameters and the numbers of bootstrap replicates used for each of the phylogenetic trees are available as Supplementary Materials and Methods.

Next, to ascertain if the two HPV sequences were also present in the other three anatomical sites (viz. flat warts on the back of hand, oral mucosa and vulva); the DNA samples were screened by PCR using E6/E7 specific primers to HPV_MTS3 and HPV_MTS4 (Table 2). The analyses confirmed the presence of HPV_MTS3 and HPV_MTS4 in the samples from all three anatomical sites. The specificity of the amplicons was confirmed by direct sequencing (GATC Biotech, Germany). Negative controls have been included during the DNA extraction and PCR analysis steps, and all tested negative. As the tissue architecture of oral mucosa and vulva is different from skin, HPV_MTS3 and HPV_MTS4 might have a broader tissue tropism. Recent studies suggested that some ‘cutaneotropic’ HPVs, might have a higher affinity to mucosal than to keratinized tissues (Bottalico et al., 2011; Hampras et al., 2017; Weissenborn et al., 2009; Pierce Campbell et al., 2013; Torres et al., 2015).

The analysis of healthy human skin scraping specimens collected in Rome using a sterile spatula showed that HPV_MTS3 and HPV_MTS4 were present in human skin, although, in varied frequencies. The study population comprised of 103 Italian subjects, consists of 51 females and 52 males, all were white and at the mean age of 71.3 years (age range 50–94 years). All participants gave signed informed consent and the study was approved by the Ethical Commissions of both NIHMP and San Gallicano Dermatologic Institute. The prevalence was evaluated by PCR using L1 specific primers (Table 2). In the analysis only one skin sample was tested positive for HPV_MTS3 (1/103, 0.97%), whereas, 10 out of 103 (10/103; 9.7%) were positive for HPV_MTS4. The sequence of the amplicons was validated by Sanger sequencing.

EV is manifested by cutaneous immunodeficiency and is extremely susceptible to repeated and persistent HPV infection (Patel et al., 2010). In this study, the prevalence of two novel gamma-HPV genomic sequences is discussed in an immunosuppressed late onset EV patient, co-infected with other oncogenic alpha and beta-HPV types. Despite the prevalence in oral, anogenital mucosa and skin the natural history of gamma-HPV have been poorly investigated. However, several studies contributed to a better knowledge of its biodiversity (Moscicki et al., 2017; Dona et al., 2015; Grace and Munger, 2017; Köhler et al., 2011; Li et al., 2013; Bolatti et al., 2016). On contrary, the association of gamma-HPV with non-melanoma skin cancer is well established (McLaughlin-Drubin, 2015; Hampras et al., 2014; Rahman et al., 2016; Deng et al., 2016; Nindl et al., 2007). Though, these two HPVs possess all important features to interact with the cellular partners like p53 and pRb; this might also be possible that the immunosuppressed subject act as a reservoir of subclinical HPV infection. Nevertheless, this study

increases the knowledge concerning the diversity and evolution of gamma-PV types.

Conflict of interest

None.

Acknowledgement

This work is supported by the Institut National de la Santé Et de la Recherche Médicale (no. ENV201610) (https://www.eva2.inserm.fr/EVA/jsp/AppelsOffres/CANCER/), and by a grant from Fondation ARC pour la recherche sur le cancer (no. PJA 20151203192) (https://www.fondation-arc.org/espace-chercheur) to MT.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.virusres.2018.03.003.

References

Accardi, R., Gheit, T., 2014. Cutaneous HPV and skin cancer. Presse Med. 43, e435–e443. http://dx.doi.org/10.1016/j.lpm.2014.08.008.

Androphy, E.J., Dvoretzky, I., Lowy, D.R., 1985. X-linked inheritance of epidermodysplasia verruciformis: genetic and virologic studies of a kindred. Arch. Dermatol. 121, 864–868.

Antonsson, A., Erfurt, C., Hazard, K., Holmgren, V., Simon, M., Kataoka, A., Hossain, S., Hakangard, C., Hansson, B.G., 2003. Prevalence and type spectrum of human papillomaviruses in healthy skin samples collected in three continents. J. Gen. Virol. 84, 1881–1886. http://dx.doi.org/10.1099/vir.0.18836-0.

Bolatti, E.M., Chouhy, D., Casal, P.E., Pérez, G.R., Stella, E.J., Sanchez, A., Gorosito, M., Bussy, R.F., Giri, A.A., 2016. Characterization of novel human papillomavirus types 157, 158 and 205 from healthy skin and recombination analysis in genus γ-Papillomavirus. Infect. Genet. Evol. 42, 20–29. http://dx.doi.org/10.1016/j.meegid.2016.04.018.

Bottalico, D., Chen, Z., Dunne, A., Ostoloza, J., McKinney, S., Sun, C., Schlecht, N.F., Fatahzadeh, M., Herrero, R., Schiffman, M., Burk, R.D., 2011. The oral cavity contains abundant known and novel human papillomaviruses from the Betapapillomavirus and Gammapapillomavirus genera. J. Infect. Dis. 204, 787–792. http://dx.doi.org/10.1093/infdis/jir383.

de Villiers, E.M., 2013. Cross-roads in the classification of papillomaviruses. Virology 445, 2–10. http://dx.doi.org/10.1016/j.virol.2013.04.023.

Deng, Q., Li, J., Pan, Y., Liu, F., He, Z., Liu, M., Liu, Y., Zhang, C., Abliz, A., Shen, N., Hang, D., Xu, Z., Wang, Q., Ning, T., Guo, C., Liang, Y., Xu, R., Zhang, L., Cai, H., Ke, Y., 2016. Prevalence and associated risk factors of human papillomavirus in healthy skin specimens collected from Rural Anyang, China, 2006–2008. J. Invest. Dermatol. 136, 1191–1198. http://dx.doi.org/10.1016/j.jid.2016.02.014.

Dona, M.G., Gheit, T., Latini, A., Benevolo, M., Torres, M., Smelov, V., McKay-Chopin, S., Giglio, A., Cristaudo, A., Zaccarelli, M., Tommasino, M., Giuliani, M., 2015. Alpha, beta and gamma Human Papillomaviruses in the anal canal of HIV-infected and uninfected men who have sex with men. J. Infect. 71, 74–84. http://dx.doi.org/10.1016/j.jinf.2015.02.001.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. http://dx.doi.org/10.1093/nar/gkh340.

Ekstrom, J., Bzhalava, D., Svenback, D., Forslund, O., Dillner, J., 2011. High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions. Int. J. Cancer 129, 2643–2650. http://dx.doi.org/10.1002/ijc.26204.

Forslund, O., Johansson, H., Madsen, K.G., Kofoed, K., 2013. The nasal mucosa contains a large spectrum of human papillomavirus types from the Betapapillomavirus and Gammapapillomavirus genera. J. Infect. Dis. 208, 1335–1341. http://dx.doi.org/10.1093/infdis/jit326.

Fouéré, S., Aubin, F., Péré, H., Galicier, L., Gheit, T., Tommasino, M., Ram Wolff, C., Boutboul, D., Bagot, M., 2017. Epidermodysplasia verruciformis in an adult patient with a germline Interleukin -2 inducible T-Cell Kinase mutation and lymphoma: the case of inherited versus acquired. J. Eur. Acad. Dermatol. Venereol. JEADV. http://dx.doi.org/10.1111/jdv.14756. [Epub ahead of print].

Foulongne, V., Sauvage, V., Hebert, C., Dereure, O., Cheval, J., Gouilh, M.A., Pariente, K., Segondy, M., Burguiere, A., Manuguerra, J.C., Caro, V., Eloit, M., 2012. Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. PLoS One 7, e38499. http://dx.doi.org/10.1371/journal.pone.0038499.

Grace, M., Munger, K., 2017. Proteomic analysis of the gamma human papillomavirus type 197 E6 and E7 associated cellular proteins. Virology 500, 71–81. http://dx.doi.org/10.1016/j.virol.2016.10.010.

Hampras, S.S., Giuliano, A.R., Lin, H.-Y., Fisher, K.J., Abrahamsen, M.E., Sirak, B.A., Iannacone, M.R., Gheit, T., Tommasino, M., Rollison, D.E., 2014. Natural history of cutaneous human papillomavirus (HPV) infection in men: the HIM study. PLoS One 9, e104843. http://dx.doi.org/10.1371/journal.pone.0104843.

Hampras, S.S., Rollison, D.E., Giuliano, A.R., McKay-Chopin, S., Minoni, L., Sereday, K., Gheit, T., Tommasino, M., 2017. Prevalence and Concordance of cutaneous beta human papillomavirus infection at mucosal and cutaneous sites. J. Infect. Dis. 216, 92–96. http://dx.doi.org/10.1093/infdis/jix245.

Johne, R., Muller, H., Rector, A., van Ranst, M., Stevens, H., 2009. Rolling-circle amplification of viral DNA genomes using phi29 polymerase. Trends Microbiol. 17, 205–211. http://dx.doi.org/10.1016/j.tim.2009.02.004.

Köhler, A., Gottschling, M., Manning, K., Lehmann, M.D., Schulz, E., Krüger-Corcoran, D., Stockfleth, E., Nindl, I., 2011. Genomic characterization of ten novel cutaneous human papillomaviruses from keratotic lesions of immunosuppressed patients. J. Gen. Virol. 1585–1594. http://dx.doi.org/10.1099/vir.0. 030593-0.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874. http://dx.doi.org/10.1093/molbev/msw054.

Li, J., Pan, Y., Deng, Q., Cai, H., Ke, Y., 2013. Identification and characterization of eleven novel human gamma-papillomavirus isolates from healthy skin, found at low frequency in a normal population. PLoS One 8, e77116. http://dx.doi.org/10.1371/journal.pone.0077116.

Ma, Y., Madupu, R., Karaoz, U., Nossa, C.W., Yang, L., Yooseph, S., Yachimski, P.S., Brodie, E.L., Nelson, K.E., Pei, Z., 2014. Human papillomavirus community in healthy persons, defined by metagenomics analysis of human microbiome project shotgun sequencing data sets. J. Virol. 88, 4786–4797. http://dx.doi.org/10.1128/JVI.00093-14.

McDermott, D.F., Gammon, B., Snijders, P.J., Mbata, I., Phifer, B., Howland Hartley, A., Lee, C.C.R., Murphy, P.M., Hwang, S.T., 2009. Autosomal dominant epidermodysplasia verruciformis lacking a known EVER1 or EVER2 mutation. Pediatr. Dermatol. 26, 306–310. http://dx.doi.org/10.1111/j.1525-1470.2008.00853.x.

McLaughlin-Drubin, M.E., 2015. Human papillomaviruses and non-melanoma skin cancer. Semin. Oncol. 42, 284–290. http://dx.doi.org/10.1053/j.seminoncol.2014.12.032.

Moscicki, A.B., Ma, Y., Gheit, T., McKay-Chopin, S., Farhat, S., Widdice, L.E., Tommasino, M., 2017. Prevalence and transmission of beta and gamma human papillomavirus in heterosexual couples. Open Forum Infect. Dis. 4. http://dx.doi.org/10.1093/ofid/ofw216. ofw216.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Oxford. http://dx.doi.org/10.1093/molbev/msw054.

Nindl, I., Gottschling, M., Stockfleth, E., 2007. Human papillomaviruses and non-melanoma skin cancer: basic virology and clinical manifestations. Dis. Mark. 23, 247–259. http://dx.doi.org/10.1155/2007/942650.

Orth, G., 2006. Genetics of epidermodysplasia verruciformis: insights into host defense against papillomaviruses. Semin. Immunol. 18, 362–374. http://dx.doi.org/10.1016/j.smim.2006.07.008.

Patel, T., Morrison, L.K., Rady, P., Tyring, S., 2010. Epidermodysplasia verruciformis and susceptibility to HPV. Dis. Mark. 29, 199–206. http://dx.doi.org/10.3233/DMA-2010-0733.

Pierce Campbell, C.M., Messina, J.L., Stoler, M.H., Jukic, D.M., Tommasino, M., Gheit, T., Rollison, D.E., Sichero, L., Sirak, B.A., Ingles, D.J., Abrahamsen, M., Lu, B., Villa, L.L., Lazcano-Ponce, E., Giuliano, A.R., 2013. Cutaneous human papillomavirus types detected on the surface of male external genital lesions: a case series within the HPV Infection in Men Study. J. Clin. Virol. 58, 652–659. http://dx.doi.org/10.1016/j.jcv.2013.10.011.

Rahman, S., Rollison, D.E., Pierce Campbell, C.M., Waterboer, T., Michel, A., Pawlita, M., Villa, L.L., Lazcano Ponce, E., Wang, W., Borenstein, A.R., Giuliano, A.R., 2016. Seroprevalence of cutaneous human papillomaviruses and the risk of external genital lesions in men: a nested case-control study. PLoS One 11, e0167174. http://dx.doi.org/10.1371/journal.pone.0167174.

Schowalter, R.M., Pastrana, D.V., Pumphrey, K.A., Moyer, A.L., Buck, C.B., 2010. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. Cell Host Microbe 7, 509–515. http://dx.doi.org/10.1016/j.chom.2010.05.006.

Sichero, L., Pierce Campbell, C.M., Ferreira, S., Sobrinho, J.S., Luiza Baggio, M., Galan, L., Silva, R.C., Lazcano-Ponce, E., Giuliano, A.R., Villa, L.L., 2013. Broad HPV distribution in the genital region of men from the HPV infection in men (HIM) study. Virology 443, 214–217. http://dx.doi.org/10.1016/j.virol.2013.04.024.

Tommasino, M., 2017. The biology of beta human papillomaviruses. Virus Res. 231, 128–138. http://dx.doi.org/10.1016/j.virusres.2016.11.013.

Torres, M., Gheit, T., McKay-Chopin, S., Rodríguez, C., Romero, J.D., Filotico, R., Doná, M.G., Ortiz, M., Tommasino, M., 2015. Prevalence of beta and gamma human papillomaviruses in the anal canal of men who have sex with men is influenced by HIV status. J. Clin. Virol. 67, 47–51. http://dx.doi.org/10.1016/j.jcv.2015.04.005.

Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., Sun, Q., Kaur, R., Huyen, Y., McBride, A.A., 2017. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Res. 45, D499–D506. http://dx.doi.org/10.1093/nar/gkw879.

Vohra, S., Sharma, N.L., Shanker, V., Mahajan, V.K., Jindal, N., 2010. Autosomal dominant epidermodysplasia verruciformis: a clinicotherapeutic experience in two cases. Indian J. Dermatol. Venereol. Leprol. 76, 557–561. http://dx.doi.org/10.4103/0378-6323.69092.

Weissenborn, S.J., De Koning, M.N.C., Wieland, U., Quint, W.G.V., Pfister, H.J., 2009. Intrafamilial transmission and family-specific spectra of cutaneous betapapillomaviruses. J. Virol. 83, 811–816. http://dx.doi.org/10.1128/JVI.01338-08.

II - Combining NGS & Amplicon sequencing: the need for a bioinformatics workflow

In the following papers, we present the new wet-lab and bioinformatics approach we have developed for the identification of HPV sequences using degenerated primers and NGS amplicon-sequencing.

In **paper nr 5**, we describe a novel protocol that combines the use of a novel L1 region amplification primer, an improvement on the original FAP primers, with the power of NGS. This novel method has allowed the discovery of 105 putative novel PV types from skin swabs and oral gargles of immunocompetent individuals. One of the 105 putative novel HPV types was fully reconstructed, and its clone was sent to the Karolinska HPV Reference Center during spring 2019.

In **paper nr 6**, we describe the novel Beta-2 HPV type isolated from skin samples based on the results of paper n°5. This novel genome is awaiting its official HPV number classification.

In **paper nr 7**, under review in "Nucleic Acid Research Genomic and Bioinformatics", we present the bioinformatics workflow we developed for the analysis of HPV amplicon data. The code for this worklflow is publicly available for the scientific community on the IARC GitHub platform (https://github.com/IARCbioinfo/PVAmpliconFinderhttps://github.com/IARCbioinfo/PVAmpliconFinder).

In **paper nr 8,** submitted to "The Journal of Infectious Diseases", we present an application of the method on healthy skin and actinic keratosis of the same individuals, allowing not only to identify novel HPV types, but also to observe PV composition differences based on clinical status.

5) Brancaccio RN, **Robitaille A**, Dutta S, Cuenin C, Santare D, Skenders G, ..., Grundhoff A. (2018). Generation of a novel next-generation sequencing-based method for the isolation of new human papillomavirus types. *Virology*, *520*, 1-10.

6) Brancaccio RN, **Robitaille A**, Dutta S, Rollison DE., Tommasino M, Gheit T. (2019). Isolation of a Novel Beta-2 Human Papillomavirus from Skin. *Microbiol Resour Announc*, *8*(9), e01628-18.

7) **Robitaille A**, Brancaccio RN, Dutta S, Rollison DE, Leja M, Fischer N, Grundhoff A, Gheit T, Tommasino M, Olivier M. (2019). PVAmpliconFinder: a package for the identification of human papillomaviruses from high throughput amplicon sequencing *NAR* (Submitted)

8) Galati L, Brancaccio R, **Robitaille A**, Cuenin C, Luzi F, Fiorucci G, Chiantore MV, Marascio N, Matera G, Liberto MC, Donà MG, Di Bonito P, Gheit T, Tommasino M. **Detection of human papillomaviruses in paired healthy skin and actinic keratosis by next generation sequencing.** *The Journal of Infectious Diseases* (Submitted)

Virology 520 (2018) 1–10

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/virology

Generation of a novel next-generation sequencing-based method for the isolation of new human papillomavirus types

Rosario N. Brancaccio[a,1], Alexis Robitaille[a,1], Sankhadeep Dutta[a], Cyrille Cuenin[a], Daiga Santare[b], Girts Skenders[b], Marcis Leja[b], Nicole Fischer[c,d], Anna R. Giuliano[e], Dana E. Rollison[e,f], Adam Grundhoff[d,g], Massimo Tommasino[a,*], Tarik Gheit[a,*]

[a] International Agency for Research on Cancer, Lyon, France
[b] Institute of Clinical and Preventive Medicine, University of Latvia, Riga, Latvia
[c] Institute for Medical Microbiology, Virology and Hygiene, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[d] German Center for Infection Research, partner site Hamburg, Borstel, Lübeck, Riems, Germany
[e] Center for Infection Research in Cancer, Moffitt Cancer Center, Tampa, FL, USA
[f] Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL, USA
[g] Heinrich-Pette Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany

ARTICLE INFO

Keywords:
Broad-spectrum HPV PCR primers
Next-generation sequencing
New papillomaviruses

ABSTRACT

With the advent of new molecular tools, the discovery of new papillomaviruses (PVs) has accelerated during the past decade, enabling the expansion of knowledge about the viral populations that inhabit the human body. Human PVs (HPVs) are etiologically linked to benign or malignant lesions of the skin and mucosa. The detection of HPV types can vary widely, depending mainly on the methodology and the quality of the biological sample. Next-generation sequencing is one of the most powerful tools, enabling the discovery of novel viruses in a wide range of biological material. Here, we report a novel protocol for the detection of known and unknown HPV types in human skin and oral gargle samples using improved PCR protocols combined with next-generation sequencing. We identified 105 putative new PV types in addition to 296 known types, thus providing important information about the viral distribution in the oral cavity and skin.

1. Introduction

Human papillomaviruses (HPVs) are non-enveloped viruses with double-stranded circular DNA of about 8 kb that can colonize the mucosal and cutaneous epithelia (Bernard et al., 2010; Bzhalava et al., 2013). To date, more than 200 PVs have been isolated from different body sites and fully characterized, and this number continues to grow (Bzhalava et al., 2015; Smelov et al., 2017). Based on the nucleotide sequences of the major capsid protein L1, HPVs are classified into genera, species, and types (Bernard et al., 2010). HPV types are organized into five major genera: alpha, beta, gamma, mu, and nu (de Villiers et al., 2004). The genera alpha, beta, and gamma include the majority of the known HPVs. The alpha HPV types have been extensively studied, because of their clear association with human carcinogenesis (Tommasino, 2014). The high-risk (HR) HPV group includes at least 12 HPV types (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59), which are the etiological agents of anogenital cancers and a subset of head and neck cancers, particularly oropharyngeal cancer (Bouvard et al., 2009; Haedicke and Iftner, 2013). The genus alpha also includes the low-risk HPV types (HPV6 and 11) that are associated with benign genital lesions and with laryngeal disease in children (Giuliano et al., 2008b; Goon et al., 2008).

The genus beta includes approximately 50 different HPV types, fully characterized, that are subdivided into five species (beta HPV species 1–5). The majority of the beta HPV types belong to species beta-1 and beta-2 and are widely present in the skin of healthy individuals. Only 7 HPV types have been classified into the species beta-3 ($n = 4$), beta-4 ($n = 1$), and beta-5 ($n = 2$). HPV types of genus beta can induce warts and have been associated with certain forms of non-melanoma skin carcinoma (NMSC) (Orth, 2006). The first beta HPVs, HPV5 and 8, were isolated from the skin of patients with epidermodysplasia verruciformis (EV), a rare autosomal recessive hereditary skin disorder that confers high susceptibility to beta HPV infection and cutaneous squamous cell carcinoma development at sun-exposed regions (Pfister, 2003). Several studies showed that beta HPV types are associated with NMSC development in non-EV individuals (Andersson et al., 2008; Berkhout et al.,

* Corresponding authors.
E-mail addresses: icb@iarc.fr (M. Tommasino), gheitt@iarc.fr (T. Gheit).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.virol.2018.04.017
Received 12 December 2017; Received in revised form 23 April 2018; Accepted 26 April 2018

2000; Bouwes Bavinck et al., 2010; Casabonne et al., 2007; Cornet et al., 2012; de Jong-Tieben et al., 1995; Harwood et al., 2000; Iannacone et al., 2014; Iftner et al., 2003; Karagas et al., 2006; Waterboer et al., 2008). Patients with a history of NMSC show elevated positivity for markers of beta HPV infection compared with healthy individuals (Ally et al., 2013; Asgari et al., 2008; Iannacone et al., 2012). Recent studies reported the presence of beta HPV types at additional anatomical sites other than the skin, such as the oral mucosal epithelium, eyebrow hairs, penile and external genital samples, and the anal canal (Arroyo et al., 2013; Barzon et al., 2011; Donà et al., 2016; Pierce Campbell et al., 2016; Smelov et al., 2017).

Species beta-3 HPV types appear to have a dual tropism, being present in the skin and the mucosal epithelia (Forslund et al., 2013; Hampras et al., 2017). Interestingly, studies in *in vitro* and *in vivo* experimental models have highlighted some biological similarities between beta-3 HPV and mucosal HR HPV types (Cornet et al., 2012; Viarisio et al., 2016). In addition, Viarisio et al. (2016) showed that beta-3-HPV49 transgenic mice were highly susceptible to upper digestive tract carcinogenesis upon initiation with 4-nitroquinoline 1-oxide.

HPVs from the gamma, mu, and nu genera induce cutaneous papillomas or warts (de Villiers et al., 2004) and have been poorly investigated so far. To date, approximately 80 different gamma HPV types have been isolated from the skin and genital tract (retrieved from GenBank, September 2017).

In addition to the fully characterized HPV types, a substantial number of partial genomic sequences of putative novel HPV types have been deposited to GenBank, indicating that many more HPV types exist. So far, the molecular biology techniques for the isolation of novel HPV types have been based mainly on the use of degenerate and/or consensus primers, followed by cloning and Sanger sequencing (Chouhy et al., 2010; Forslund et al., 1999). However, considering the large number of recently characterized HPV genomes, degenerate primers may be improved in order to discover novel HPV types. In particular, this strategy may lead to the expansion of species that so far include a very small number of HPV types, such as species beta-3 ($n = 4$), beta-4 ($n = 1$), and beta-5 ($n = 2$).

In this study, we used novel and well-validated consensus and degenerate primers to amplify genomic HPV sequences from human DNA isolated from oral and skin specimens. Analysis of the PCR products by next-generation sequencing (NGS) resulted in the identification of 105 putative new PV types.

2. Materials and methods

2.1. Sample collection and DNA extraction

Skin swabs and oral rinses from two different ongoing studies aiming to determine the prevalence of viral DNA and its associations with disease were used in the present analysis (Hampras et al., 2014, 2015; Nunes et al., 2016; Pierce Campbell et al., 2013, 2016).

Skin swab specimens ($n = 119$) were randomly selected from the VIRUSCAN Study, an ongoing five-year (2014–2019) prospective cohort study conducted at Moffitt Cancer Center and the University of South Florida (R01CA177586-01; “Prospective study of cutaneous viral infections and non-melanoma skin cancer”). An area of approximately 5 × 5 cm of the top of the sun-exposed forearm was sprayed with 0.9% saline solution. A cotton-tipped Dacron swab (Digene, Gaithersburg, MD, USA) was then rubbed back and forth a few times to collect exfoliated skin cells. Individual swabs were placed in a separate vial and preserved in Digene Standard Transport Medium.

In addition, 62 oral rinses were randomly selected from the HPV Infection in Men (HIM) study, a large, multinational (Brazil, Mexico, and the USA) prospective cohort study of the natural history of HPV infection in men. The HIM study methods have previously been described in detail (Giuliano et al., 2008a, 2009, 2011; Nyitray et al., 2011). A further 85 oral samples were selected from a pilot study that aimed to estimate the prevalence of *Helicobacter pylori* in oral gargles from a Latvian population. The study was approved (No. 8-A/15) by the Ethics Committee of Riga East University Hospital Support Foundation.

After DNA extraction, all samples were analyzed at the International Agency for Research on Cancer (Lyon, France) for viral DNA from HPV.

Table 1
Sequences of the oligonucleotides and composition of the different protocols. i = inosine; W = A or T; D = A or G or T; K = T or G; Y = C or T; M = A or C; R = A or G; V = A or C or G; H = A or C or T.

Primer mix	Primer sequence (5–3′)
Beta-3-1	
B3L1FW3	AGGACATCCATACTTTGAGGTTCGAG
B3L1FW4	TAGGACATCCATATTTTGATGTGAGAG
B3L1FW5	GATGTTAGAGACACTGGAGATTCAACA
B3L1FW6	GATGTTAGAGACACTGGGGATTCAACA
B3L1FW7	GATGTTAGAGACACTGTGGATCAAACA
B3L1RW	ATAATAGTATTTCTTAATTCTAATGGAGG
B3L1RW4	ATAACTGAATTGATTAATTCTAATGGAGG
B3L1RW5	ATAACTGTATTTACTAATTCTAAAGGTGG
B3L1RW6	TACAGTATTTACCAGTTCCAAAGGTGG
B3L1RW7	ATTACAGTATTAACTAATTCTAAAGGTGG
B3L1RW8	ATTACAGTATTTACTAATTCTAAAGGTGG
Beta-3-2	
B3L1FW1	GTAGGACATCCATAYTTTGAKGTKiGAG
B3L1FW2	TTGATGTTAGAGACACTGiDGATYMAACA
B3L1RW1	ATAAiWGWATTKYTTAATTCTAATGGAGG
B3L1RW2	ATTACAGTATTiACKARTTCYAAAGGTGG
CUT	
CUT1Fw	TRCCiGAYCCiAATAARTTTG
CUT1AFw	TRCCiGAYCCiAACAGRTTTG
CUT1BFw	TRCCiGAYCCiAATAGRTTTG
CUT1CFw	TRCCiGAYCCiAACAARTTTG
CUT1BRv	ARGAYGGiGAYATGGTiGA
FAP	
FAP59	TAACWGTiGGiCAYCCWTATT
FAP64	CCWATATCWVHCATiTCiCCATC
FAPM1	
FAP59.1	TAACAGTDGGiCAYCCWTWTT
FAP59.2	TAACAGTDGGiCAYCCWTAYT
FAP64.1	CCDATATCWVHCATATCiCCATC
FAP59	TAACWGTiGGiCAYCCWTATT
FAP64	CCWATATCWVHCATiTCiCCATC
FAPM2	
FAP59.2	TAACAGTDGGiCAYCCWTAYT
FAP64.1	CCDATATCWVHCATATCiCCATC

2.2. PCR protocols

The following PCR protocols using different sets of primers were run (Table 1): (i) CUT primers, as previously described (Chouhy et al., 2010); (ii) FA-type (FAP) primers, as previously described (Forslund et al., 1999); (iii) a new set of FAP primers, i.e. FAP59.1, FAP59.2, and FAP64.1 (Fig. 1; Table 1); these primers were used to generate two different primer mixtures (FAPM1 and FAPM2); the PCR conditions were the same as for the original FAP protocol; (iv) a set of 11 beta-3 specific primers (henceforth referred to as beta-3-1) (Table 1); and (v) a set of 4 broad-spectrum beta-3 degenerate primers (henceforth referred to as beta-3-2). The beta-3-1 and beta-3-2 primers were synthesized by MWG Biotech (Ebersberg, Germany) and mixed to obtain a 10 × solution containing 2 μM of each primer. PCR was performed with the Qiagen Multiplex PCR kit (Hilden, Germany) according to the manufacturer's instructions. The use of these primers enables the amplification of a region in the L1 gene of approximately 450 bp.

2.3. Validation of the new set of primers

To evaluate the sensitivity of the novel HPV PCR protocols (beta-3-1, beta-3-2, FAPM1, and FAPM2), we used an artificial mixture containing cloned HPV genomes at different relative concentrations (10-

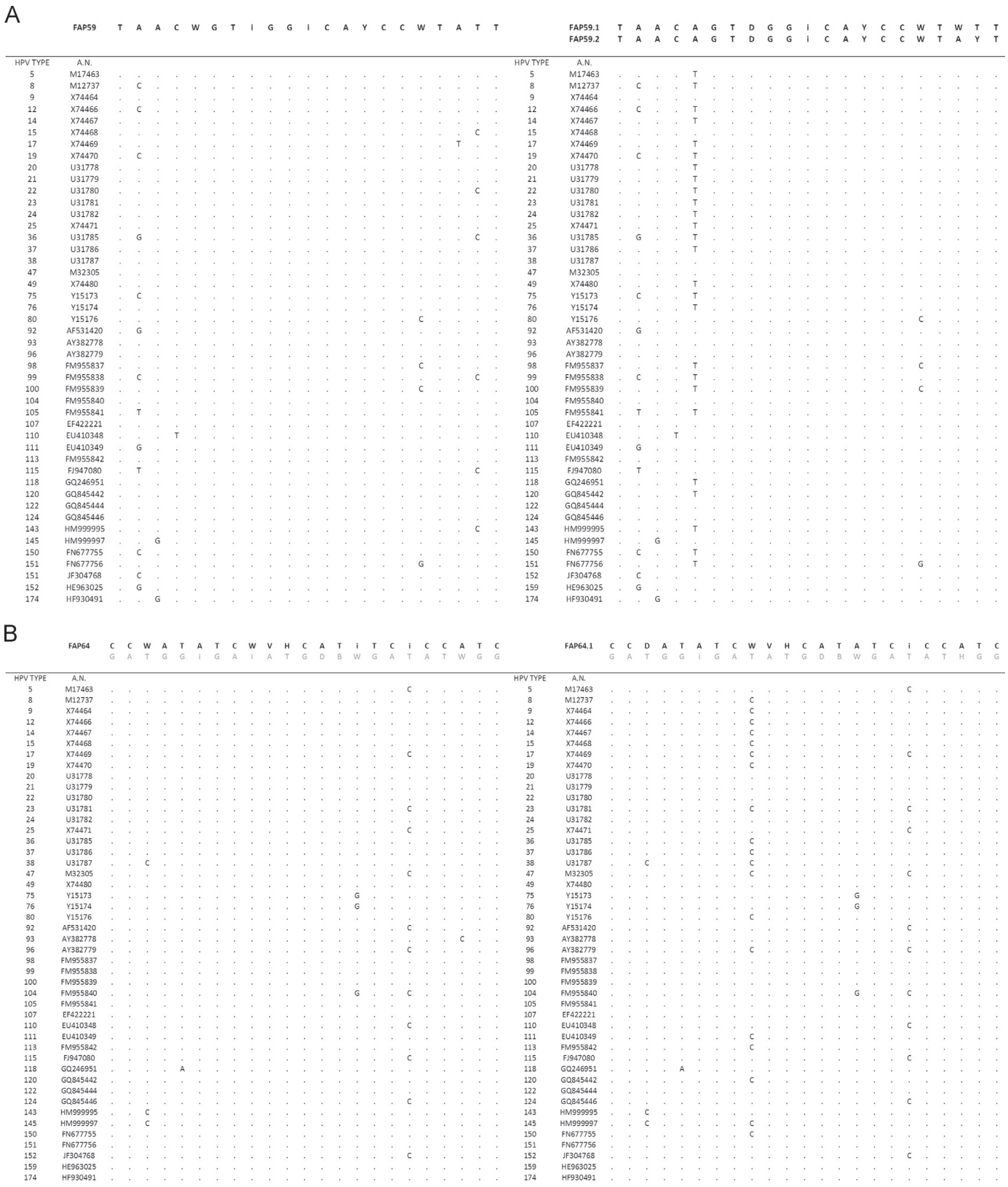


Fig. 1. Construction of improved FAP degenerate primers by alignment of 46 beta HPV type L1 regions with MUSCLE 3.8.

Table 2
Description of the PCR protocols and NGS pools.

PCR pools	PCR protocols	Specimens	N	NGS pools
1	Beta-3-1	Skin swab	41	1
2	Beta-3-2		9	
3	FAP		52	2
4	FAPM1		54	3
5	CUT		57	4
6	FAPM2	Oral gargle	43	5
7	FAPM1		56	6
8	CUT		55	7
9	Beta-3-1		9	8
10	Beta-3-2		4	
11	FAP		11	
12	FAPM1		11	
13	FAPM2		12	

fold dilution series starting from 10,000 to 0 copies of the viral genome) and mixed with human genomic DNA. PCR products were analyzed by electrophoresis on a 2% agarose gel.

2.4. NGS analysis

The PCR products were purified on a 2% agarose gel using the QIAquick gel extraction kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. One additional purification step was performed to remove any remaining contaminants using the Agencourt AMPure XP PCR purification kit with a beads ratio of 1.8 × (Beckman Coulter). Purified PCR products were divided into eight different pools. Each pool included approximately 50 different amplicons generated from different PCR protocols (Table 2).

Libraries were prepared using the Nextera XT DNA Library preparation kit (Illumina, San Diego, CA, USA). Illumina MiSeq dual-indexed adapters (Illumina, San Diego, CA, USA) were added to each of the PCR pools.

NGS was performed using the Illumina MiSeq kit v3 (600 cycles) on the Illumina MiSeq system. In order to enrich the diversity of the libraries, 10% of PhiX (Illumina, San Diego, California, USA) was added to the NGS reaction.

2.5. Bioinformatics analysis

Quality control was conducted using FastQC (Andrews, 2010) (v0.11.5) and MultiQC (Käller and Ewels, 2016) (v1.0). Trim Galore (v0.4.4) (Krueger, 2015) was used to remove remaining adapter sequences and trim low-quality ends of reads. The merging of forward and reverse reads, the de-replication step, the de novo chimeric sequence identification, and the clustering steps were carried out using VSEARCH (Mahé and Rognes, 2016) (v2.4.0). MegaBlast in the Blast package (v2.6.0+) (Altschul et al., 1990) was launched against the nucleotide collection (nr/nt, March 2017) database in a local server to enable the identification of the previously constructed clusters.

Another level of clustering was applied for the reads having the same best MegaBlast results inside each pool (based on the E-value). Each cluster of reads was then processed using the CAP3 program in order to assemble contigs (Huang and Madan, 1999).

A reference species phylogenetic tree was constructed based on the full-L1 ORF nucleotide sequences of 458 available PV genomes retrieved from the PaVE database (https://pave.niaid.nih.gov/) (Van Doorslaer et al., 2013) in January 2018. The sequences were aligned at the nucleotide level using the MUSCLE algorithm, with the default parameters (Edgar, 2004), in MEGA7 (Kumar et al., 2016). The final full-length L1-ORF alignment encompassed 458 full L1-ORF nucleotide sequences, 2259 positions, and 627 distinct alignment patterns. MEGA7 was used to test the best substitution model and for the phylogenetic inference. The codon positions included were 1st + 2nd + 3rd + noncoding. Based on the alignment using MUSCLE, all positions with < 95% site coverage were eliminated (partial deletions), to enable the inclusion of taxa with some missing data. There was a total of 1383 positions in the final dataset.

A discrete gamma distribution was used to model evolutionary rate differences among sites (five categories; +G, parameter = 1.0326). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 2.5307% sites).

The initial trees for the heuristic search were obtained automatically by applying the neighbor-joining (NJ)/BioNJ algorithm to a matrix of pairwise distances estimated using the maximum composite likelihood (MCL) approach and then by selecting the topology with the highest log likelihood value (−389774.5274).

Phylogenetic inference was performed with MEGA7 using the general time reversible (GTR) model of nucleotide substitution and 500 bootstrap replicates (Nei and Kumar, 2000).

PaPaRa (v2.5) (Berger and Stamatakis, 2011) was used to align the sequences reconstructed using the CAP3 algorithm with respect to the reference multiple sequence alignment. Subsequently, the evolutionary placement algorithm (EPA) in RAxML (v8.2.11) (Berger et al., 2011; Stamatakis, 2014) was run to place the sequences into the reference species phylogenetic tree. The EPA was run using the same nucleotide substitution model used to infer the reference phylogenetic tree. A script was developed in-house to parse the output format (Matsen et al., 2012) of the EPA.

In addition, a blastn local alignment query of the contigs was used to align them against a comprehensive database of reference PVs present in the PaVE database (n = 330 genomes). This approach mimics locally the L1 taxonomic tool of the PaVE database.

All the results in this study are based on the identification of the sequences using the EPA in RAxML (henceforth referred to as RAxML-EPA). Only the longest sequence was considered for RAxML-EPA classification when several singlets or contigs were available. Krona (Ondov et al., 2011) was used for the graphical representation of the data.

3. Results

3.1. Design and validation of novel HPV PCR primers

As a first step, we generated new consensus primers considering all known beta HPV types. HPV beta-3 species primers were designed by aligning the L1 open reading frame (ORF) from all four beta-3 HPV types (HPV49, 75, 76, and 115) using the ClustalW2 multiple sequence alignment tool (Chenna et al., 2003). Two different sets of primers were generated: (i) a set of 11 specific primers, termed beta-3-1, and (ii) a set of 4 degenerate primers, termed beta-3-2. The composition of the two primer mixtures is shown in Table 1.

As a second approach, we generated additional degenerate primers based on the well-validated FAP primers (Forslund et al., 1999). The FAP primers were developed in 1999 by aligning 77 L1 ORF sequences from different genera that included at that time only a limited number of beta (n = 22) and gamma (n = 5) HPV type sequences, obtained from the 1996 and 1997 HPV Sequence Database Compendia (Myers et al., 1996, 1997).

Forty-six L1 sequences representative of the beta HPV types known to date (Van Doorslaer et al., 2017) were aligned against FAP primer sequences, using the MUSCLE (3.8) multiple sequence alignment tool (Edgar, 2004). Subsequently, three improved broad-spectrum FAP primers, with an increased specificity for beta HPV types, were generated (Fig. 1): FAP59.1, FAP59.2, and FAP64.1. These were mixed in different combinations, generating two different mixtures: FAPM1 and FAPM2 (Table 1). The beta-3 protocol enabled the detection of beta-3 HPVs with a limit of detection of 10 copies. The detection limit using the FAPM1 mixture was 10 copies for HPV types that belong to species beta-2, beta-3, and beta-4 and 1000 copies for beta-5. Using the FAPM2 protocol, the detection limit was 10 copies for HPV types that belong to

beta-3, beta-4, and beta-5; however, a lower sensitivity was observed for species beta-2 (10,000 copies) (data not shown).

3.2. NGS data analysis: characterization and taxonomic classification

Randomly selected DNA extracted from skin swabs (n = 119) and oral gargles (n = 147) obtained from healthy individuals was amplified using the PCR protocols described above and the original FAP and CUT protocols (Table 2) (Chouhy et al., 2010; Forslund et al., 1999). PCR products were mixed to obtain 8 different pools (Table 2) and sequenced using the Illumina MiSeq sequencing platform.

A total of 50,017,076 paired-end raw reads were obtained from the NGS analysis. After quality trimming, de-replication, and chimeric PCR sequence removal, 47.3% (23,647,656) of the reads were considered for further analysis. Approximately 67% (16,043,298 reads) were related to PV sequences. Each read was matched against National Center for Biotechnology Information (NCBI) database sequences (nr/nt, March 2017) using the MegaBlast algorithm and assigned to its closest PV type, before contig construction.

Analysis of the data with RAxML-EPA revealed that the reads generated from the sequencing of the 119 skin DNA samples were assigned to a total of 265 different PV types (Fig. 2A; Table S1), which belong mainly to the alpha (33.4%) and beta (29.5%) genera, thus representing the PV distribution in skin. In addition, a substantial fraction of reads (12.9%) was assigned to taxonomically unclassified PV sequences (hereafter called "unclassified PVs"): bovine papillomavirus type 19 (BPV19), equine papillomavirus type 8 (EcPV8), *Myotis ricketti* papillomavirus 1 (MrPV1), *Pudu puda* papillomavirus type 1 (PpuPV1), and *Sparus aurata* papillomavirus type 1 (SaPV1). Moreover, 9% of the reads were assigned to the gamma genus.

The FAPM1 protocol enabled the detection of 107 PVs (8 alpha, 37 beta, 60 gamma, and 2 mu), and the CUT and FAP protocols enabled the detection of 118 PVs (11 alpha, 36 beta, 68 gamma, 2 mu, and 1 nu) and 87 PVs (3 alpha, 34 beta, 49 gamma, and 1 mu), respectively. The combined beta-3-1 and beta-3-2 protocols generated a majority of reads assigned to a non-human alpha PV type: *Colobus guereza* monkey papillomavirus type 1 (CgPV1). Two reads were assigned to HPV16. Five beta HPVs were detected using these combined protocols (797,800 reads), of which 3 were assigned to species beta-3. Only 2 non-referenced gamma HPV types were detected (HPV-mDysk1 – KX781280 and HPV-mDysk6 – KX781285).

The reads generated from the sequencing of the 147 oral DNA samples were assigned to a total of 161 different PV types. PV types that belong to the genus beta were most common (29.5%), followed by genus gamma (19.6%) and genus alpha (7.8%) (Fig. 2B; Table S2). In addition, a substantial fraction of reads (36.9%) was assigned to taxonomically unclassified PVs: EcPV8, *Miniopterus schreibersii* papillomavirus type 1 (MscPV1), PpuPV1, and SaPV1 (Fig. 2B; Table S2).

The FAPM1 and FAPM2 protocols enabled the detection of 55 PVs (4 alpha, 30 beta, and 21 gamma) and 42 PVs (5 alpha, 21 beta, and 16 gamma), respectively. Forty-six PVs (6 alpha, 17 beta, and 23 gamma) were detected using the CUT protocol (Fig. 2B; Table S2).

Substantial numbers of reads identified in both skin (745,860 reads) and oral (163,448 reads) samples were related to taxonomically classified non-human PVs (i.e. PVs not belonging to the genera alpha, beta, gamma, mu, and nu) (Tables S1 and S2; Fig. 2).

3.3. Subdivision of the NGS reads into known and putative novel PVs

The NGS sequences were divided into two groups, on the basis of the initial MegaBlast results: (i) L1 sequences with ≥ 90% similarity with a known PV (i.e. known PV types) and (ii) L1 sequences with < 90% similarity with any known PV (i.e. putative novel PV types). This subdivision was followed by contig construction and sequences identification using CAP3 and RAxML-EPA, respectively.

Regarding the sequences that share ≥ 90% of identity with known PVs, a total of 8,002,617 reads were generated. The majority were from the genus beta (2,358,670 reads), followed by alpha (1,992,264 reads) and gamma (1,002,061 reads) (Fig. 3A). A substantial proportion of the reads (1,678,061 reads) was assigned to the "unclassified PVs" category, mainly represented by SaPV1 (KX643372.1). The beta-3-1 and beta-3-2 protocols generated a total of 2,588,649 reads (pool 1, Table 2), with a majority (56.6%) of alpha PV sequences, followed by beta HPV sequences (30.8%) (Fig. 3A). The FAP protocol (pool 2) generated 985,675 reads in skin samples, of which 40.8% belonged to the genus beta and 14.5% to gamma (Table 2; Fig. 3A). The FAPM1 protocol (pool 3) enabled the detection in skin samples of 861,810 reads, comprising alpha (23.3%), beta (13.6%), gamma (14.3%), and mu (7.6%) PV-related sequences (Table 2; Fig. 3A). In oral samples (pool 6), when the same PCR protocol was used, generating 244,587 reads, a different distribution of alpha, beta, and gamma PVs was observed, with 0.1%, 53.6%, and 12.3%, respectively (Table 2; Fig. 3A).

The use of the CUT protocol on skin samples (pool 4) generated 884,923 reads, from the alpha (13%), beta (19.5%), and gamma (23.3%) genera. When the same protocol was used on oral samples (pool 7), generating 78,060 reads, the proportion of alpha (2.1%), beta (11%), and gamma (17.2%) PV-related sequences was different (Table 2; Fig. 3A). The highest proportion of reads (43.4%) generated from this pool corresponded to an unclassified PV (SaPV1).

The FAPM2 protocol (pool 5), used in oral samples, generated 466,004 reads, with a distribution of 9.3% alpha, 39.6% beta, and 32.5% gamma PV-related sequences (Table 2; Fig. 3A).

In addition, products from five PCR protocols (pool 8, Table 2) were pooled and analyzed by NGS. This pool generated 1,892,909 reads, of which 24.5% were representative of beta, 8.8% alpha, and 17.6% gamma PVs. The highest proportion of reads (44.3%) was representative of unclassified PVs (Tables S2).

All the reads correspond to 296 known PV types, including 30 alpha PVs, of which 14 were found in skin samples, 8 in oral samples, and 8 in both tissues. Fifty-four beta HPVs were identified, of which 13 were from the skin, 3 from the oral cavity, and 38 from both tissues. Regarding the genus gamma, 123 known HPVs were identified, of which 70 were isolated from the skin, 8 from the oral cavity, and 45 from both anatomical sites. Three mu HPVs were found (1 in the skin and 2 in both skin and oral samples), and only 1 nu HPV was found (in the skin). Six unclassified PV types were identified, of which 2 were isolated from the skin, 1 from the oral cavity, and 3 from both sites (data not shown).

In addition, 11.3% of the reads (n = 909,308) corresponded to 79 sequences of diverse PVs that do not belong to any of the five PV genera (alpha, beta, gamma, mu, and nu) that contain HPVs; 34 of these 79 sequences were isolated from the skin, 11 from the oral cavity, and 34 from both sites (data not shown).

Regarding the putative novel PVs, we identified 19,032 reads with < 90% similarity with known PVs. The majority of these reads were related to beta (35.6%) and gamma (23.2%) HPV types (Fig. 3B; Table S3). The beta-3-1 and beta-3-2 protocols enabled the identification in pool 1 of 22 reads (26.8%) that are representative of 2 putative new beta-3-related sequences (Fig. 3B; Table S3). In the same pool, 54 reads (65.8%) were assigned to an unclassified PV. However, in the same cluster, a smaller contig was assigned to *Psipapillomavirus* (Table S3). The remaining reads (n = 6, 7.3%) were assigned to *Dyophipapillomavirus 1*, but were matched against HPV115 using the PaVE classification. The FAP protocol enabled the detection in pool 2 of putative new beta (40 reads, 1.2%) and gamma (2228 reads, 69.2%) HPV types. Of the 116 reads that were assigned to unclassified PV using RAxML-EPA, 2 were related to HPV MTS2 (gamma-7) according to the PaVE classification (Dutta et al., 2017). Finally, 833 reads were identified as *Taupapillomavirus 3*, 4 reads as *Deltapapillomavirus 5*, and 3 reads as *Dyorhopapillomavirus 1* (Table S3).

The FAPM1 protocol enabled the detection in pool 3 of sequences representative of putative new beta (294 reads, 70.2%), gamma (48

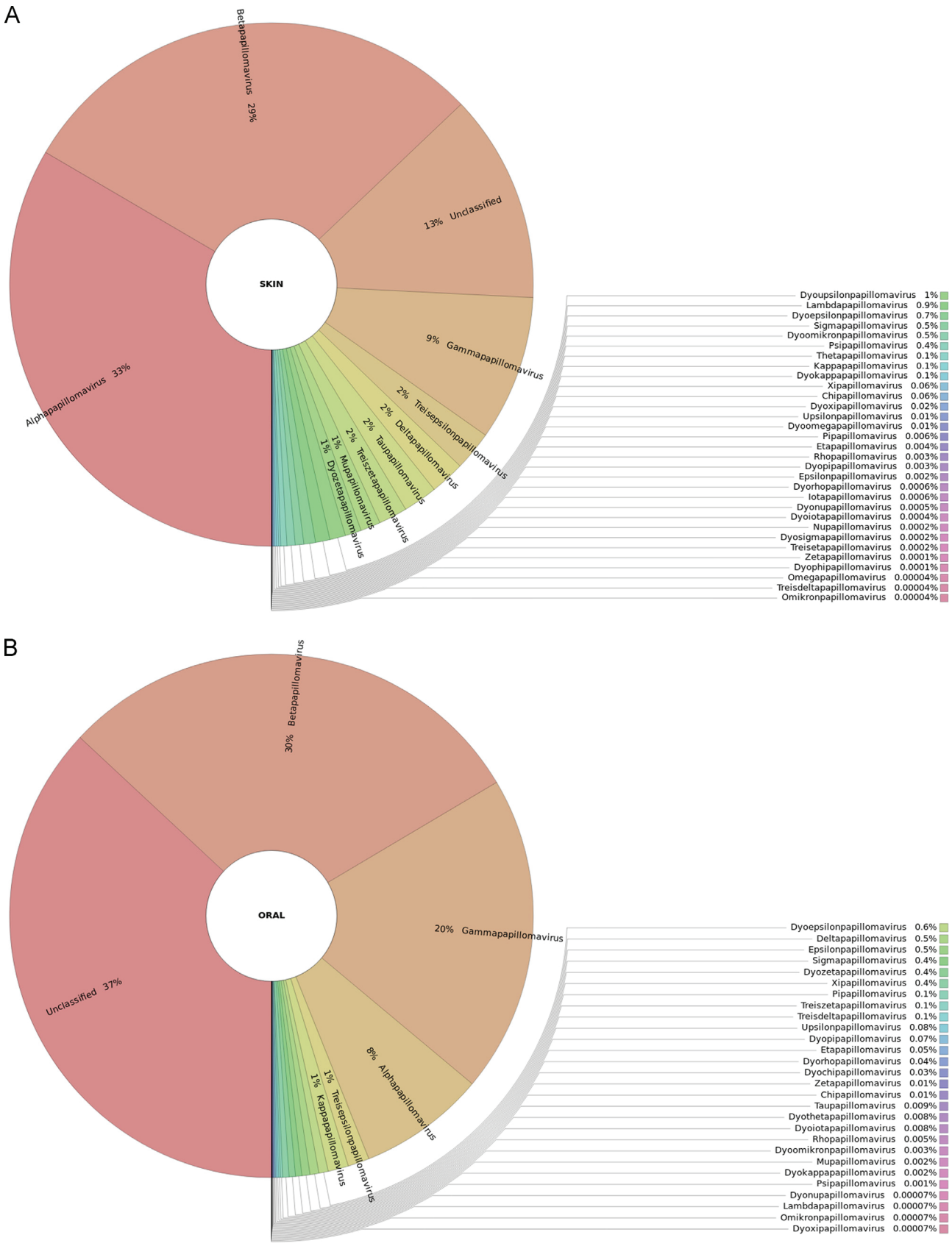


Fig. 2. Graphical representation of the unnormalized abundance of PV genera and species in terms of number of reads: (A) skin samples, (B) oral samples.

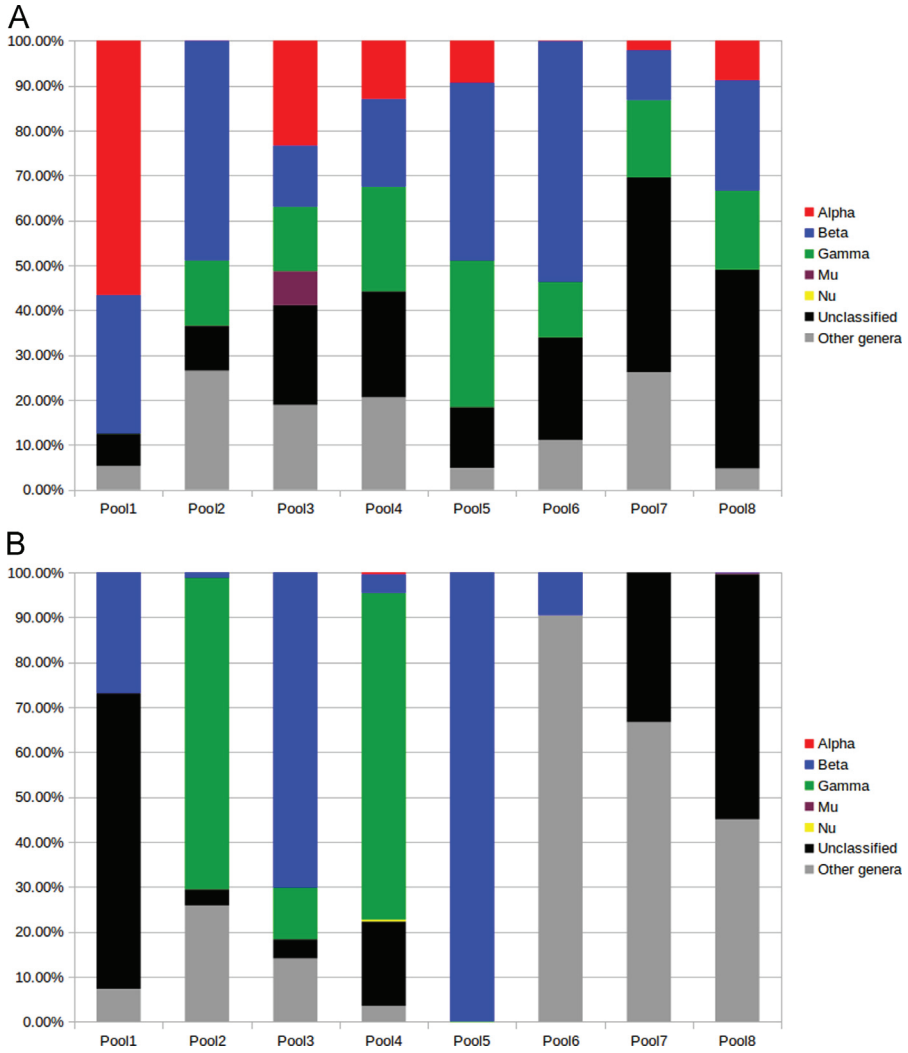


Fig. 3. (A) Distribution of the known PVs detected in the different NGS pools, in terms of percentage of reads within each pool; (B) distribution of the putative new PVs detected in the different NGS pools, in terms of percentage of reads within each pool (RAxML-EPA).

reads, 11.5%), delta-2 (52 reads, 12.4%), and lambda-3 (7 reads, 1.7%) PV types. In oral samples, the same protocol enabled the identification of 21 reads, of which 9.5% were found to be related to putative new beta-1 HPV types, 23.8% to *Sigmapapillomavirus 1*, and 66.7% to *Dyoiotapapillomavirus 2* using RAxML-EPA (Fig. 3B; Table S3). No putative new gamma HPV types were identified.

The use of the CUT protocol on skin samples (pool 4) revealed the presence of 2126 reads (72.7%) representative of putative new gamma HPV types. A smaller fraction (4.2%) was representative of putative new beta HPV types. This protocol also revealed the presence of 12 nu and 12 alpha (assigned to species alpha-2 and alpha-3) PV-related reads. RAxML-EPA indicated that 2 reads corresponded to *Lambdapapillomavirus 3*, whereas the same reads got their best initial MegaBlast match against canine papillomavirus 6 (CPV6). Eight other putative non-human PVs were also found (Table S3).

In oral samples (pool 5), when the FAPM2 protocol was used, 6295 reads (99.9%) were representative of putative new beta HPV types, and 0.1% were representative of putative new gamma HPV types (Fig. 3B; Table S3). The CUT protocol (pool 7) enabled the identification in oral samples of only one putative new non-human PV (*Chipapapillomavirus 2*), as well as an unclassified type using RAxML-EPA, but all such reads were assigned to species beta-1 using the PaVE database (Table S3).

Regarding pool 8, only 0.07% and 0.3% of the reads were assigned to beta and mu HPVs, respectively. The remaining reads were related to an unclassified PV (3308 reads), to *Treisdeltapapillomavirus 1* (2713 reads), and to other non-human PV genera (*Treisepsilonpapillomavirus*, *Treisdeltapapillomavirus*, and *Treiszetapapilloamvirus*; 16 reads).

In summary, all the reads corresponded to 105 putative novel PV types, including 29 beta HPVs, of which 21 were found in skin and 8 in oral samples. Thirty-two gamma HPV types were identified, of which 30 were found in skin and 2 in oral samples. Only 2 putative new alpha HPVs were found in the skin. One mu HPV was found in skin samples. Twenty-four diverse PVs that do not belong to any of the five PV genera that contain HPVs were identified, of which 17 were found in skin and 7 in oral samples. Moreover, 17 unclassified PVs were isolated from skin ($n = 15$) and oral ($n = 2$) samples. However, these reads were found to correspond to beta ($n = 9$) and gamma ($n = 8$) HPVs when the PaVE algorithm was used.

4. Discussion

Since the discovery of the first HPV type four decades ago (Orth et al., 1978), 127 alpha, 93 beta, and 135 gamma HPVs have been described in the Papillomavirus Episteme database (Van Doorslaer et al., 2017). Alpha, beta, and gamma are the most representative genera (Doorbar et al., 2012).

To identify new HPV types, FAP and CUT primers combined with cloning and Sanger sequencing-based strategies have been used successfully in the past (Chouhy et al., 2010; Forslund et al., 1999). However, this approach is quite laborious and time-consuming, and

enables the identification of the most represented amplicons only. In particular, this strategy is ineffective in the context of multiple infections. With the advent of new molecular tools (e.g. NGS), the discovery of new HPVs has accelerated over the past few years (Bzhalava et al., 2014; Kocjan et al., 2015). Several studies have shown the capability of NGS in detecting low-copy HPV infections, especially in multiple infections (Arroyo et al., 2013; Barzon et al., 2011; Ekström et al., 2011; Johansson et al., 2013).

Here, we developed a strategy that combines the use of specific or degenerate primers targeting the L1 region of a broad spectrum of HPVs with NGS, for the detection of new HPV types, especially from the genus beta. This strategy incorporates the selective enrichment of PV sequences before NGS is performed. Approximately two thirds of the reads were related to PV sequences. Similar approaches have been reported previously (Arroyo Mühr et al., 2015; Ekström et al., 2013, 2011).

The growing interest in the beta genus arises from evidence that a number of beta HPV types may be involved in pre-malignant and malignant skin lesions (Pfister et al., 2003; Tommasino, 2017). Interestingly, species beta-3 HPV types have been detected in the skin and mucosal epithelia (Forslund et al., 2013; Hampras et al., 2017). Functional studies in *in vitro* and *in vivo* experimental models have highlighted some biological similarities between beta-3 and mucosal HR HPV types. HPV49 shows transforming activity in primary human keratinocytes, and shares some features with HPV16 (Cornet et al., 2012; Viarisio et al., 2016). One of our objectives was to expand the biologically relevant species beta-3, which includes only 4 HPV types, by using beta-3 consensus and degenerate primers.

Combining PCR with novel sets of HPV primers and NGS, we showed the presence of a total of 105 putative new PVs. This procedure also demonstrated the presence of 296 known PV types. Our study showed the presence of a substantial number of beta and gamma HPV types in the oral cavity, which supports the hypothesis of a possible mucosal tropism. However, environmental contamination of the oral cavity cannot be excluded. Furthermore, several other sequences related to unclassified and non-human PVs were identified in skin and oral samples. Environmental contamination may explain the presence of non-human PVs in skin and oral samples. However, cross-species transmission of PVs between animals and humans may also be a consideration (Bravo and Félez-Sánchez, 2015; Gottschling et al., 2011), even though PVs are typically considered to be highly host-restricted (with a few exceptions). Sequences related to bovine PVs have been found in horses and other equids, suggesting interspecies transmission events (Lunardi et al., 2013; Trewby et al., 2014). Other studies also reported cases of cross-species transmission of PVs between bat species (García-Pérez et al., 2014), between rhesus and cynomolgus macaques (Chen et al., 2009), and between humans and cats (Anis et al., 2010; O'Neill et al., 2011); however, additional studies are needed to confirm the latter.

In addition, the notion of "non-human" PV genera needs to be interpreted with caution as they may also include some HPVs. Similarly, alpha and beta genera include few non-human primate PVs (Bernard et al., 2010; Rector and Van Ranst, 2013).

All the results in this study are based on the identification of the sequences using the RAxML-EPA classification. A total of 105 putative new PVs (including 29 beta, 32 gamma, 2 alpha, and 1 mu PVs) were found. In addition, 24 diverse PVs that do not belong to any of the five PV genera that contain HPVs were identified. Interestingly, 17 of the 105 putative new PVs (16.2%) were assigned to taxonomically unclassified PVs. These PVs may not belong to any of the known genera that contain human or animal PVs, and thus may be representative of putative new genera.

The taxonomic assignment performed in this study must be interpreted cautiously, because only small portions of putative new PV genomes have been obtained. In addition, the results obtained using the blastn algorithm refer exclusively to the fraction of the sequence that is aligned by the algorithm. The percentage of similarity indicated by the initial MegaBlast results must also be interpreted with caution, because the definition of novelty for a PV is based on the full L1 ORF length.

In this study, the different protocols were run on different human specimens, and showed different efficacies in detecting putative new PVs, as well as known PVs. The beta-3-1 and beta-3-2 protocols enabled the identification of 4 new beta-3-related sequences in skin samples (using the RAxML-EPA classification), which may potentially expand the beta-3 group to 8 PV types. In vitro experiments are needed to provide insight into the biological properties of these PV types, and to investigate whether these types share biological features with HPV49 (Cornet et al., 2012; Viarisio et al., 2016). The CUT primers enabled the detection of a broad range of PV types in skin and oral samples, including alpha PV types, as previously reported (Chouhy et al., 2010). In contrast, the original FAP protocol was much less likely to identify PVs belonging to the genus alpha. The FAPM1 and FAPM2 protocols enabled the detection of the largest number of putative new PVs in oral samples, whereas the CUT primers enabled the detection of the largest number of putative new PVs in skin samples. Interestingly, the FAPM1 and CUT protocols showed good performance in the detection in skin samples of new PV types that belong to non-human PV genera.

Together, the different protocols enabled the identification of a substantial number ($n = 62$) of putative new beta and gamma HPV types, as well as putative non-human PVs ($n = 24$), in both skin and oral samples.

The gamma HPV types constitute a large group of HPVs that are not yet clearly associated with human disease. However, HPV197, a member of species gamma-24, has recently been detected in human skin cancer specimens (Arroyo Mühr et al., 2015; Grace and Munger, 2017). To date, only 3 HPV types have been classified into the species gamma-24. The use of consensus or degenerate gamma-24 primers might facilitate the discovery of new related PV types, if any exist. Some of the putative new beta or gamma HPV types may also show transforming activity.

In summary, the present study describes a robust strategy based on the use of specific or degenerate primers and NGS technology to detect putative novel PVs. Although the identification of novel PV types or species can only be definitively confirmed by sequencing the whole L1 ORF, initial studies have confirmed the validity of our new protocol as a first step for the isolation and full characterization of novel HPV genomes (e.g. HPV ICB1) (Brancaccio et al., 2017).

The discovery of novel HPV types remains of paramount importance, because new associations between HPV infections and human diseases may be established.

Acknowledgments

We are grateful to Dr. Karen Müller and Jessica Cox for editing. This study was supported in part by the European Commission project HPV-AHEAD (FP7-HEALTH-2011-282562), by the grant VIRUSCAN R01 (no. R01CA177586-01), and by a grant from "Fondation ARC" (no. PJA 20151203192).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2018.04.017.

References

Ally, M.S., Tang, J.Y., Arron, S.T., 2013. Cutaneous human papillomavirus infection and basal cell carcinoma of the skin. J. Investig. Dermatol. 133. http://dx.doi.org/10.1038/jid.2013.46.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

Andersson, K., Waterboer, T., Kirnbauer, R., Slupetzky, K., Iftner, T., de Villiers, E.-M.,

Forslund, O., Pawlita, M., Dillner, J., 2008. Seroreactivity to cutaneous human papillomaviruses among patients with nonmelanoma skin cancer or benign skin lesions. Cancer Epidemiol. Biomark. Prev. 17, 189–195. http://dx.doi.org/10.1158/1055-9965.EPI-07-0405.

Andrews,S., 2010. FastQC: a quality control tool for high throughput sequence data.

Anis, E.A., O'Neill, S.H., Newkirk, K.M., Brahmbhatt, R.A., Abd-Eldaim, M., Frank, L.A., Kania, S.A., 2010. Molecular characterization of the L1 gene of papillomaviruses in epithelial lesions of cats and comparative analysis with corresponding gene sequences of human and feline papillomaviruses. Am. J. Vet. Res. 71, 1457–1461. http://dx.doi.org/10.2460/ajvr.71.12.1457.

Arroyo, L.S., Smelov, V., Bzhalava, D., Eklund, C., Hultin, E., Dillner, J., 2013. Next generation sequencing for human papillomavirus genotyping. J. Clin. Virol. 58, 437–442. http://dx.doi.org/10.1016/j.jcv.2013.07.013.

Arroyo Mühr, L.S., Hultin, E., Bzhalava, D., Eklund, C., Lagheden, C., Ekström, J., Johansson, H., Forslund, O., Dillner, J., 2015. Human papillomavirus type 197 is commonly present in skin tumors. Int. J. Cancer 136, 2546–2555. http://dx.doi.org/10.1002/ijc.29325.

Asgari, M.M., Kiviat, N.B., Critchlow, C.W., Stern, J.E., Argenyi, Z.B., Raugi, G.J., Berg, D., Odland, P.B., Hawes, S.E., de Villiers, E.-M., 2008. Detection of human papillomavirus DNA in cutaneous squamous cell carcinoma among immunocompetent individuals. J. Investig. Dermatol. 128, 1409–1417. http://dx.doi.org/10.1038/sj.jid.5701227.

Barzon, L., Militello, V., Lavezzo, E., Franchin, E., Peta, E., Squarzon, L., Trevisan, M., Pagni, S., Dal Bello, F., Toppo, S., Palù, G., 2011. Human papillomavirus genotyping by 454 next generation sequencing technology. J. Clin. Virol. 52, 93–97. http://dx.doi.org/10.1016/j.jcv.2011.07.006.

Berger, S.A., Krompass, D., Stamatakis, A., 2011. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. Syst. Biol. 60, 291–302. http://dx.doi.org/10.1093/sysbio/syr010.

Berger, S.A., Stamatakis, A., 2011. Aligning short reads to reference alignments and trees (https://doi.org/). Bioinformatics 27, 2068–2075. http://dx.doi.org/10.1093/bioinformatics/btr320.

Berkhout, R.J., Bouwes Bavinck, J.N., ter Schegget, J., 2000. Persistence of human papillomavirus DNA in benign and (pre)malignant skin lesions from renal transplant recipients. J. Clin. Microbiol. 38, 2087–2096.

Bernard, H.-U., Burk, R.D., Chen, Z., van Doorslaer, K., zur Hausen, H., de Villiers, E.-M., 2010. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. Virology 401, 70–79. http://dx.doi.org/10.1016/j.virol.2010.02.002.

Bouvard, V., Baan, R., Straif, K., Grosse, Y., Secretan, B., El Ghissassi, F., Benbrahim-Tallaa, L., Guha, N., Freeman, C., Galichet, L., Cogliano, V., WHO International Agency for Research on Cancer Monograph Working Group, 2009. A review of human carcinogens–Part B: biological agents. Lancet Oncol. 10, 321–322.

Bouwes Bavinck, J.N., Neale, R.E., Abeni, D., Euvrard, S., Green, A.C., Harwood, C.A., de Koning, M.N.C., Naldi, L., Nindl, I., Pawlita, M., Pfister, H., Proby, C.M., Quint, W.G.V., ter Schegget, J., Waterboer, T., Weissenborn, S., Feltkamp, M.C.W., EPI-HPV-UV-CA group, 2010. Multicenter study of the association between betapapillomavirus infection and cutaneous squamous cell carcinoma. Cancer Res. 70, 9777–9786. http://dx.doi.org/10.1158/0008-5472.CAN-10-0352.

Brancaccio, R.N., Robitaille, A., Dutta, S., Rollison, D.E., Fischer, N., Grundhoff, A., Tommasino, M., Gheit, T., 2017. Complete genome sequence of a novel human gammapapillomavirus isolated from skin. Genome Announc. 5. http://dx.doi.org/10.1128/genomeA.00833-17.

Bravo, I.G., Félez-Sánchez, M., 2015. Papillomaviruses: viral evolution, cancer and evolutionary medicine. Evol. Med. Public Health 2015, 32–51. http://dx.doi.org/10.1093/emph/eov003.

Bzhalava, D., Eklund, C., Dillner, J., 2015. International standardization and classification of human papillomavirus types. Virology 476, 341–344. http://dx.doi.org/10.1016/j.virol.2014.12.028.

Bzhalava, D., Guan, P., Franceschi, S., Dillner, J., Clifford, G., 2013. A systematic review of the prevalence of mucosal and cutaneous human papillomavirus types. Virol. Spec. Issue.: Papillomavirus Epistem. 445, 224–231. http://dx.doi.org/10.1016/j.virol.2013.07.015.

Bzhalava, D., Mühr, L.S.A., Lagheden, C., Ekström, J., Forslund, O., Dillner, J., Hultin, E., 2014. Deep sequencing extends the diversity of human papillomaviruses in human skin. Sci. Rep. 4, 5807. http://dx.doi.org/10.1038/srep05807.

Casabonne, D., Michael, K.M., Waterboer, T., Pawlita, M., Forslund, O., Burk, R.D., Travis, R.C., Key, T.J., Newton, R., 2007. A prospective pilot study of antibodies against human papillomaviruses and cutaneous squamous cell carcinoma nested in the Oxford component of the European Prospective investigation into cancer and Nutrition. Int. J. Cancer 121, 1862–1868. http://dx.doi.org/10.1002/ijc.22885.

Chen, Z., van Doorslaer, K., DeSalle, R., Wood, C.E., Kaplan, J.R., Wagner, J.D., Burk, R.D., 2009. Genomic diversity and interspecies host infection of alpha12 Macaca fascicularis papillomaviruses (MfPVs). Virology 393, 304–310. http://dx.doi.org/10.1016/j.virol.2009.07.012.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D., 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31, 3497–3500.

Chouhy, D., Gorosito, M., Sánchez, A., Serra, E.C., Bergero, A., Fernandez Bussy, R., Giri, A.A., 2010. New generic primer system targeting mucosal/genital and cutaneous human papillomaviruses leads to the characterization of HPV 115, a novel Beta-papillomavirus species 3. Virology 397, 205–216. http://dx.doi.org/10.1016/j.virol.2009.11.020.

Cornet, I., Bouvard, V., Campo, M.S., Thomas, M., Banks, L., Gissmann, L., Lamartine, J., Sylla, B.S., Accardi, R., Tommasino, M., 2012. Comparative analysis of transforming properties of E6 and E7 from different beta human papillomavirus types. J. Virol. 86, 2366–2370. http://dx.doi.org/10.1128/JVI.06579-11.

de Jong-Tieben, L.M., Berkhout, R.J., Smits, H.L., Bouwes Bavinck, J.N., Vermeer, B.J., van der Woude, F.J., ter Schegget, J., 1995. High frequency of detection of epidermodysplasia verruciformis-associated human papillomavirus DNA in biopsies from malignant and premalignant skin lesions from renal transplant recipients. J. Investig. Dermatol. 105, 367–371.

de Villiers, E.-M., Fauquet, C., Broker, T.R., Bernard, H.-U., zur Hausen, H., 2004. Classification of papillomaviruses. Virology 324, 17–27. http://dx.doi.org/10.1016/j.virol.2004.03.033.

Donà, M.G., Gheit, T., Vescio, M.F., Latini, A., Moretto, D., Benevolo, M., Cristaudo, A., Tommasino, M., Giuliani, M., 2016. Incidence, clearance and duration of cutaneous beta and gamma human papillomavirus anal infection. J. Infect. 73, 380–383. http://dx.doi.org/10.1016/j.jinf.2016.07.006.

Doorbar, J., Quint, W., Banks, L., Bravo, I.G., Stoler, M., Broker, T.R., Stanley, M.A., 2012. The biology and life-cycle of human papillomaviruses. Vaccine 30 (Suppl 5), F55–70. http://dx.doi.org/10.1016/j.vaccine.2012.06.083.

Dutta, S., Robitaille, A., Olivier, M., Rollison, D.E., Tommasino, M., Gheit, T., 2017. Genome sequence of a novel human gammapapillomavirus isolated from skin. Genome Announc. 5. http://dx.doi.org/10.1128/genomeA.00439-17.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. http://dx.doi.org/10.1093/nar/gkh340.

Ekström, J., Bzhalava, D., Svenback, D., Forslund, O., Dillner, J., 2011. High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions. Int. J. Cancer 129, 2643–2650. http://dx.doi.org/10.1002/ijc.26204.

Ekström, J., Mühr, L.S.A., Bzhalava, D., Söderlund-Strand, A., Hultin, E., Nordin, P., Stenquist, B., Paoli, J., Forslund, O., Dillner, J., 2013. Diversity of human papillomaviruses in skin lesions. Virology 447, 300–311. http://dx.doi.org/10.1016/j.virol.2013.09.010.

Forslund, O., Antonsson, A., Nordin, P., Stenquist, B., Hansson, B.G., 1999. A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. J. Gen. Virol. 80 (Pt 9), 2437–2443. http://dx.doi.org/10.1099/0022-1317-80-9-2437.

Forslund, O., Johansson, H., Madsen, K.G., Kofoed, K., 2013. The nasal mucosa contains a large spectrum of human papillomavirus types from the betapapillomavirus and gammapapillomavirus genera. J. Infect. Dis. 208, 1335–1341. http://dx.doi.org/10.1093/infdis/jit326.

García-Pérez, R., Ibáñez, C., Godínez, J.M., Aréchiga, N., Garin, I., Pérez-Suárez, G., de Paz, O., Juste, J., Echevarría, J.E., Bravo, I.G., 2014. Novel papillomaviruses in free-ranging Iberian bats: no virus-host co-evolution, no strict host specificity, and hints for recombination. Genome Biol. Evol. 6, 94–104. http://dx.doi.org/10.1093/gbe/evt211.

Giuliano, A., Lazcano, E., Villa, L., Flores, R., Salmeron, J., Lee, J.-H., Papenfuss, M., Abrahamsen, M., Baggio, M., Silva, R., Quiterio, M., 2009. Circumcision and sexual behavior: factors independently associated with human papillomavirus (HPV) detection among men in The HIM study. Int. J. Cancer J. Int. Cancer 124, 1251–1257. http://dx.doi.org/10.1002/ijc.24097.

Giuliano, A.R., Lazcano-Ponce, E., Villa, L.L., Flores, R., Salmeron, J., Lee, J.-H., Papenfuss, M.R., Abrahamsen, M., Jolles, E., Nielson, C.M., Baggio, M.L., Silva, R., Quiterio, M., 2008a. The human papillomavirus infection in men study: human papillomavirus prevalence and type distribution among men residing in Brazil, Mexico, and the United States. Cancer Epidemiol. Biomark. Prev. 17, 2036–2043. http://dx.doi.org/10.1158/1055-9965.EPI-08-0151.

Giuliano, A.R., Lee, J.-H., Fulp, W., Villa, L.L., Lazcano, E., Papenfuss, M.R., Abrahamsen, M., Salmeron, J., Anic, G.M., Rollison, D.E., Smith, D., 2011. Incidence and clearance of genital human papillomavirus infection in men (HIM): a cohort study. Lancet 377, 932–940. http://dx.doi.org/10.1016/S0140-6736(10)62342-2.

Giuliano, A.R., Tortolero-Luna, G., Ferrer, E., Burchell, A.N., de Sanjose, S., Kjaer, S.K., Muñoz, N., Schiffman, M., Bosch, F.X., 2008b. Epidemiology of human papillomavirus infection in men, in cancers other than cervical and in benign conditions. Vaccine 26, K17–K28. http://dx.doi.org/10.1016/j.vaccine.2008.06.021.

Goon, P., Sonnex, C., Jani, P., Stanley, M., Sudhoff, H., 2008. Recurrent respiratory papillomatosis: an overview of current thinking and treatment. Eur. Arch. Otorhinolaryngol. 265, 147–151. http://dx.doi.org/10.1007/s00405-007-0546-z.

Gottschling, M., Göker, M., Stamatakis, A., Bininda-Emonds, O.R.P., Nindl, I., Bravo, I.G., 2011. Quantifying the phylodynamic forces driving papillomavirus evolution. Mol. Biol. Evol. 28, 2101–2113. http://dx.doi.org/10.1093/molbev/msr030.

Grace, M., Munger, K., 2017. Proteomic analysis of the gamma human papillomavirus type 197 E6 and E7 associated cellular proteins. Virology 500, 71–81. http://dx.doi.org/10.1016/j.virol.2016.10.010.

Haedicke, J., Iftner, T., 2013. Human papillomaviruses and cancer. Radiother. Oncol. 108, 397–402. http://dx.doi.org/10.1016/j.radonc.2013.06.004.

Hampras, S.S., Giuliano, A.R., Lin, H.-Y., Fisher, K.J., Abrahamsen, M.E., McKay-Chopin, S., Gheit, T., Tommasino, M., Rollison, D.E., 2015. Natural history of polyomaviruses in men: the HPV infection in men (HIM) study. J. Infect. Dis. 211, 1437–1446. http://dx.doi.org/10.1093/infdis/jiu626.

Hampras, S.S., Giuliano, A.R., Lin, H.-Y., Fisher, K.J., Abrahamsen, M.E., Sirak, B.A., Iannacone, M.R., Gheit, T., Tommasino, M., Rollison, D.E., 2014. Natural history of cutaneous human papillomavirus (HPV) infection in men: the HIM study. PLoS One 9. http://dx.doi.org/10.1371/journal.pone.0104843.

Hampras, S.S., Rollison, D.E., Giuliano, A.R., McKay-Chopin, S., Minoni, L., Sereday, K., Gheit, T., Tommasino, M., 2017. Prevalence and concordance of cutaneous beta human papillomavirus infection at mucosal and cutaneous sites. J. Infect. Dis. 216, 92–96. http://dx.doi.org/10.1093/infdis/jix245.

Harwood, C.A., Surentheran, T., McGregor, J.M., Spink, P.J., Leigh, I.M., Breuer, J., Proby, C.M., 2000. Human papillomavirus infection and non-melanoma skin cancer

in immunosuppressed and immunocompetent individuals. J. Med. Virol. 61, 289–297.
Huang, X., Madan, A., 1999. CAP3: a DNA sequence assembly program. Genome Res. 9, 868–877.
Iannacone, M.R., Gheit, T., Pfister, H., Giuliano, A.R., Messina, J.L., Fenske, N.A., Cherpelis, B.S., Sondak, V.K., Roetzheim, R.G., Silling, S., Pawlita, M., Tommasino, M., Rollison, D.E., 2014. Case-control study of genus-beta human papillomaviruses in plucked eyebrow hairs and cutaneous squamous cell carcinoma. Int. J. Cancer 134, 2231–2244. http://dx.doi.org/10.1002/ijc.28552.
Iannacone, M.R., Gheit, T., Waterboer, T., Giuliano, A.R., Messina, J.L., Fenske, N.A., Cherpelis, B.S., Sondak, V.K., Roetzheim, R.G., Michael, K.M., Tommasino, M., Pawlita, M., Rollison, D.E., 2012. Case-control study of cutaneous human papillomaviruses in squamous cell carcinoma of the skin. Cancer Epidemiol. Biomark. 21, 1303–1313. http://dx.doi.org/10.1158/1055-9965.EPI-12-0032.
Iftner, A., Klug, S.J., Garbe, C., Blum, A., Stancu, A., Wilczynski, S.P., Iftner, T., 2003. The prevalence of human papillomavirus genotypes in nonmelanoma skin cancers of nonimmunosuppressed individuals identifies high-risk genital types as possible risk factors. Cancer Res. 63, 7515–7519.
Johansson, H., Bzhalava, D., Ekström, J., Hultin, E., Dillner, J., Forslund, O., 2013. Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types. Virology 440, 1–7. http://dx.doi.org/10.1016/j.virol.2013.01.023.
Käller, M., Ewels, P., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics.
Karagas, M.R., Nelson, H.H., Sehr, P., Waterboer, T., Stukel, T.A., Andrew, A., Green, A.C., Bavinck, J.N.B., Perry, A., Spencer, S., Rees, J.R., Mott, L.A., Pawlita, M., 2006. Human papillomavirus infection and incidence of squamous cell and basal cell carcinomas of the skin. J. Natl. Cancer Inst. 98, 389–395. http://dx.doi.org/10.1093/jnci/djj092.
Kocjan, B.J., Bzhalava, D., Forslund, O., Dillner, J., Poljak, M., 2015. Molecular methods for identification and characterization of novel papillomaviruses. Clin. Microbiol. Infect. 21, 808–816. http://dx.doi.org/10.1016/j.cmi.2015.05.011.
Krueger F., 2015. Trim Galore!: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.
Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874. http://dx.doi.org/10.1093/molbev/msw054.
Lunardi, M., de Alcântara, B.K., Otonel, R.A.A., Rodrigues, W.B., Alfieri, A.F., Alfieri, A.A., 2013. Bovine papillomavirus type 13 DNA in equine sarcoids. J. Clin. Microbiol. 51, 2167–2171. http://dx.doi.org/10.1128/JCM.00371-13.
Mahé, F., Rognes, T., 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ.
Matsen, F.A., Hoffman, N.G., Gallagher, A., Stamatakis, A., 2012. A format for phylogenetic placements. PLoS One 7, e31009. http://dx.doi.org/10.1371/journal.pone.0031009.
Myers, G., Baker, G.M.C., Münger, K., Sverdrup, F., McBride, A. &.Bernard, H.U., 1997. Alignments. In Human Papillomaviruses 1997. HPV Sequence Database II-L1–23–73.
Myers, G., Baker, G.M.C., Münger, K., Sverdrup, F., McBride, A. &.Bernard, H.U., 1996. Alignments. In Human Papillomaviruses 1996. HPV Sequence Database II-L1–1–67.
Nei, Kumar, 2000. Molecular Evolution and Phylogenetics. Oxford University Press.
Nunes, E.M., Sudenga, S.L., Gheit, T., Tommasino, M., Baggio, M.L., Ferreira, S., Galan, L., Silva, R.C., Pierce Campbell, C.M., Lazcano-Ponce, E., Giuliano, A.R., Villa, L.L., Sichero, L., 2016. Diversity of beta-papillomavirus at anogenital and oral anatomic sites of men: the HIM Study. Virology 495, 33–41. http://dx.doi.org/10.1016/j.virol.2016.04.031.
Nyitray, A.G., Carvalho da Silva, R.J., Baggio, M.L., Lu, B., Smith, D., Abrahamsen, M., Papenfuss, M., Villa, L.L., Lazcano-Ponce, E., Giuliano, A.R., 2011. Age-specific prevalence of and risk factors for anal human papillomavirus (HPV) among men who have sex with women and men who have sex with men: the HPV in men (HIM) study. J. Infect. Dis. 203, 49–57. http://dx.doi.org/10.1093/infdis/jiq021.
Ondov, B.D., Bergman, N.H., Phillippy, A.M., 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinform. 12, 385. http://dx.doi.org/10.1186/1471-2105-12-385.
O'Neill, S.H., Newkirk, K.M., Anis, E.A., Brahmbhatt, R., Frank, L.A., Kania, S.A., 2011. Detection of human papillomavirus DNA in feline premalignant and invasive squamous cell carcinoma. Vet. Dermatol. 22, 68–74. http://dx.doi.org/10.1111/j.1365-3164.2010.00912.x.
Orth, G., 2006. Genetics of epidermodysplasia verruciformis: insights into host defense against papillomaviruses. Semin. Immunol. 18, 362–374. http://dx.doi.org/10.1016/j.smim.2006.07.008.
Orth, G., Jablonska, S., Favre, M., Croissant, O., Jarzabek-Chorzelska, M., Rzesa, G., 1978. Characterization of two types of human papillomaviruses in lesions of epidermodysplasia verruciformis. Proc. Natl. Acad. Sci. USA 75, 1537–1541.
Pfister, H., 2003. Chapter 8: human papillomavirus and skin cancer. J. Natl. Cancer Inst. Monogr. 52–56.
Pfister, H., Fuchs, P.G., Majewski, S., Jablonska, S., Pniewska, I., Malejczyk, M., 2003. High prevalence of epidermodysplasia verruciformis-associated human papillomavirus DNA in actinic keratoses of the immunocompetent population. Arch. Dermatol. Res. 295, 273–279. http://dx.doi.org/10.1007/s00403-003-0435-2.
Pierce Campbell, C.M., Gheit, T., Tommasino, M., Lin, H.-Y., Torres, B.N., Messina, J.L., Stoler, M.H., Rollison, D.E., Sirak, B.A., Abrahamsen, M., Carvalho da Silva, R.J., Sichero, L., Villa, L.L., Lazcano-Ponce, E., Giuliano, A.R., 2016. Cutaneous beta human papillomaviruses and the development of male external genital lesions: a case-control study nested within the HIM Study. Virology 497, 314–322. http://dx.doi.org/10.1016/j.virol.2016.08.002.
Pierce Campbell, C.M., Messina, J.L., Stoler, M.H., Jukic, D.M., Tommasino, M., Gheit, T., Rollison, D.E., Sichero, L., Sirak, B.A., Ingles, D.J., Abrahamsen, M., Lu, B., Villa, L.L., Lazcano-Ponce, E., Giuliano, A.R., 2013. Cutaneous human papillomavirus types detected on the surface of male external genital lesions: a case series within the HPV Infection in Men Study. J. Clin. Virol. 58, 652–659. http://dx.doi.org/10.1016/j.jcv.2013.10.011.
Rector, A., Van Ranst, M., 2013. Animal papillomaviruses. Virology 445, 213–223. http://dx.doi.org/10.1016/j.virol.2013.05.007.
Smelov, V., Hanisch, R., McKay-Chopin, S., Sokolova, O., Eklund, C., Komyakov, B., Gheit, T., Tommasino, M., 2017. Prevalence of cutaneous beta and gamma human papillomaviruses in the anal canal of men who have sex with women. Papillomavirus Res. 3, 66–72. http://dx.doi.org/10.1016/j.pvr.2017.02.002.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. http://dx.doi.org/10.1093/bioinformatics/btu033.
Tommasino, M., 2017. The biology of beta human papillomaviruses. Virus Res. 231, 128–138. http://dx.doi.org/10.1016/j.virusres.2016.11.013.
Tommasino, M., 2014. The human papillomavirus family and its role in carcinogenesis. Semin. Cancer Biol. 26, 13–21. http://dx.doi.org/10.1016/j.semcancer.2013.11.002.
Trewby, H., Ayele, G., Borzacchiello, G., Brandt, S., Campo, M.S., Del Fava, C., Marais, J., Leonardi, L., Vanselow, B., Biek, R., Nasir, L., 2014. Analysis of the long control region of bovine papillomavirus type 1 associated with sarcoids in equine hosts indicates multiple cross-species transmission events and phylogeographical structure. J. Gen. Virol. 95, 2748–2756. http://dx.doi.org/10.1099/vir.0.066589-0.
Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., Sun, Q., Kaur, R., Huyen, Y., McBride, A.A., 2017. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Res. 45, D499–D506. http://dx.doi.org/10.1093/nar/gkw879.
Van Doorslaer, K., Tan, Q., Xirasagar, S., Bandaru, S., Gopalan, V., Mohamoud, Y., Huyen, Y., McBride, A.A., 2013. The papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. Nucleic Acids Res. 41, D571–D578. http://dx.doi.org/10.1093/nar/gks984.
Viarisio, D., Müller-Decker, K., Zanna, P., Kloz, U., Aengeneyndt, B., Accardi, R., Flechtenmacher, C., Gissmann, L., Tommasino, M., 2016. Novel ß-HPV49 transgenic mouse model of upper digestive tract cancer. Cancer Res. 76, 4216–4225. http://dx.doi.org/10.1158/0008-5472.CAN-16-0370.
Waterboer, T., Abeni, D., Sampogna, F., Rother, A., Masini, C., Sehr, P., Michael, K.M., Pawlita, M., 2008. Serological association of beta and gamma human papillomaviruses with squamous cell carcinoma of the skin. Br. J. Dermatol. 159, 457–459. http://dx.doi.org/10.1111/j.1365-2133.2008.08621.x.

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Isolation of a Novel Beta-2 Human Papillomavirus from Skin

Rosario N. Brancaccio,[a] Alexis Robitaille,[a] Sankhadeep Dutta,[a]* Dana E. Rollison,[b] Massimo Tommasino,[a] Tarik Gheit[a]

[a]Infections and Cancer Biology Group, International Agency for Research on Cancer, Lyon, France
[b]Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, Florida, USA

ABSTRACT We report the complete genome characterization of a novel human papillomavirus (HPV) (ICB2) isolated from a skin swab. The L1 region of HPV ICB2 shares 87.9% nucleotide similarity with its closest relative, HPV37, and thus constitutes a novel human betapapillomavirus.

Human papillomaviruses (HPVs) are double-stranded circular DNA viruses with a genome of approximately 8 kb and belong to the *Papillomaviridae* family. HPVs infect basal keratinocytes of the mucosal and cutaneous epithelia. Based on the nucleotide sequences of the major capsid protein L1, HPVs are classified into five major genera, *Alphapapillomavirus*, *Betapapillomavirus*, *Gammapapillomavirus*, *Mupapillomavirus*, and *Nupapillomavirus* (1–3). The mucosal high-risk HPV types, which belong to the genus *Alphapapillomavirus*, are the etiological agents of anogenital cancers and of a subset of head and neck cancers (4). Moreover, an etiological role of cutaneotrophic HPVs from the genus *Betapapillomavirus* in association with exposure to UV radiation in the development of nonmelanoma skin cancer is also suggested (5–7).

Here, we report the complete genome sequence of a novel HPV type (HPV ICB2; 7,441 bp) isolated from a human forearm skin swab.

A partial L1 region sequence of HPV ICB2 (99 bp) was previously obtained from DNA extracted from the skin swab using broad-spectrum primers in combination with next-generation sequencing (8). Multiply primed rolling-circle amplification (RCA) was performed on the corresponding skin swab DNA according to the manufacturer's instructions (illustra TempliPhi 100 amplification kit; GE Healthcare, USA). To obtain the complete viral genome, first, long-range PCR was performed on the RCA product using PrimeSTAR GXL DNA polymerase (TaKaRa Bio), outward-directed primers specific for HPV ICB2 (forward primer, 5′-CAGACAGAACACATCTTTTGATCC-3′; and reverse primer, 5′-TCGTCCCGTGACCCACCCTGA-3′).

The resulting amplicon of approximately 8 kb was then cloned in pCR-XL-2 TOPO vector using the TOPO XL-2 complete PCR cloning kit (Invitrogen, Carlsbad, CA). The sequence of the whole genome was obtained by Sanger sequencing using a primer-walking strategy (GATC Biotech, Germany). This sequencing service uses cycle sequencing technology (dideoxy chain termination/cycle sequencing) on an ABI 3730XL sequencing machine. The viral genome was covered at least twice in order to identify and correct sequencing errors. Thirty-one sequences were generated and aligned to reconstruct the whole genome using the CAP3 sequence assembly program (9), with default parameters.

The clone has been submitted to the International Human Papillomavirus Reference Center in Stockholm (www.hpvcenter.se) for assignment of HPV type number.

The L1 open reading frame (ORF) of HPV ICB2 showed 87.9% nucleotide identity with its closest relative, HPV37, which belongs to the species beta-2 of the genus *Betapapillomavirus*. HPV ICB2 thus constitutes a novel human betapapillomavirus by sharing less than 90% nucleotide sequence identity with the closest HPV type in the L1 ORF (3).

Citation Brancaccio RN, Robitaille A, Dutta S, Rollison DE, Tommasino M, Gheit T. 2019. Isolation of a novel beta-2 human papillomavirus from skin. Microbiol Resour Announc 8:e01628-18. https://doi.org/10.1128/MRA.01628-18.

Editor Jelle Matthijnssens, KU Leuven

Address correspondence to Tarik Gheit, gheitt@iarc.fr.

* Present address: Sankhadeep Dutta, Department of Oncogene Regulation, Chittaranjan National Cancer Institute, Kolkata, India.

Received 5 December 2018
Accepted 5 February 2019
Published 28 February 2019

The G+C content of ICB2 is 40.7%. The virus has the typical genome organization of other cutaneotrophic HPVs; it is composed of five early (E1, E2, E4, E6, and E7) and two late (L1 and L2) ORFs, and no E5 was identified.

The long control region (LCR) is 382 bp. This region contains two polyadenylation sites (AATAAA) for L1 and L2 transcripts and four consensus palindromic E2-binding sites, as follows: ACCG-N_4-CGGT ($n = 2$), ACC-N_5-GGT ($n = 1$), and ACC-N_1-GGT ($n = 1$). A putative TATA box domain (TATAAGA) for the downstream early promoter was also identified.

The two conserved zinc-binding domains of the viral E6 protein [CxxC$(x)_{29}$CxxC and CxxC$(x)_{30}$CxxC] are present and are separated by 36 amino acids (5).

A zinc-binding domain [CxxC$(x)_{29}$CxxC] and one LxCxE motif are located in the E7 protein (5). An ATP-binding site (GPPDTGKS) for ATP-dependent helicase activity was identified in the carboxy terminus of the E1 protein. In conclusion, we identified and fully characterized a new HPV belonging to species beta-2, HPV ICB2. This finding contributes to the expansion of our knowledge about the impressive diversity of the *Betapapillomavirus* genus.

Data availability. The complete genome sequence of HPV ICB2 is available in GenBank under accession number MK080568.

ACKNOWLEDGMENTS

We are grateful to Karen Müller for editing the manuscript.

This work was partially supported by grants from the U.S. National Cancer Institute (R01-CA177586-01A1) to D.E.R. and from Fondation ARC (PJA 20151203192) to M.T.

REFERENCES

1. Tommasino M. 2014. The human papillomavirus family and its role in carcinogenesis. Semin Cancer Biol 26:13–21. https://doi.org/10.1016/j.semcancer.2013.11.002.
2. Bernard H-U, Burk RD, Chen Z, van Doorslaer K, Zur Hausen H, de Villiers E-M. 2010. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. Virology 401:70–79. https://doi.org/10.1016/j.virol.2010.02.002.
3. de Villiers E-M. 2013. Cross-roads in the classification of papillomaviruses. Virology 445:2–10. https://doi.org/10.1016/j.virol.2013.04.023.
4. Haedicke J, Iftner T. 2013. Human papillomaviruses and cancer. Radiother Oncol 108:397–402. https://doi.org/10.1016/j.radonc.2013.06.004.
5. Tommasino M. 2017. The biology of beta human papillomaviruses. Virus Res 231:128–138. https://doi.org/10.1016/j.virusres.2016.11.013.
6. Accardi R, Gheit T. 2014. Cutaneous HPV and skin cancer. Presse Med 43:e435–e443. https://doi.org/10.1016/j.lpm.2014.08.008.
7. Viarisio D, Gissmann L, Tommasino M. 2017. Human papillomaviruses and carcinogenesis: well-established and novel models. Curr Opin Virol 26:56–62. https://doi.org/10.1016/j.coviro.2017.07.014.
8. Brancaccio RN, Robitaille A, Dutta S, Cuenin C, Santare D, Skenders G, Leja M, Fischer N, Giuliano AR, Rollison DE, Grundhoff A, Tommasino M, Gheit T. 2018. Generation of a novel next-generation sequencing-based method for the isolation of new human papillomavirus types. Virology 520:1–10. https://doi.org/10.1016/j.virol.2018.04.017.
9. Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res 9:868–877. https://doi.org/10.1101/gr.9.9.868.

OXFORD UNIVERSITY PRESS | NAR Genomics and Bioinformatics

PVAmpliconFinder: a workflow for the identification of human papillomaviruses from high throughput amplicon sequencing

Journal:	*NAR Genomics and Bioinformatics*
Manuscript ID	Draft
Manuscript Type:	Bioinformatics Methods Article
Date Submitted by the Author:	n/a
Complete List of Authors:	ROBITAILLE, Alexis; International Agency for Research on Cancer, Infection and Cancer Biology Brancaccio, Rosario; International Agency for Research on Cancer, Infection and Cancer Biology Dutta, Sankhadeep; International Agency for Research on Cancer, Infection and Cancer Biology; Chittaranjan National Cancer Institute, Department of Oncogene Regulation Rollison, Dana; Moffitt Cancer Center, Department of Cancer Epidemiology Leja, Marcis; University of Latvia, Institute of Clinical and Preventive Medicine Fischer, Nicole; German Centre for Infection Research Association, German Center for Infection Research; University Medical Center Hamburg-Eppendorf, Institute for Medical Microbiology, Virology and Hygiene Grundhoff, Adam; German Centre for Infection Research Association, German Center for Infection Research; Heinrich-Pette-Institut Leibniz-Institut fur Experimentelle Virologie, Leibniz Institut for Experimental Virology Gheit, Tarik; International Agency for Research on Cancer, Infection and Cancer Biology Tommasino, Massimo; International Agency for Research on Cancer, Infection and Cancer Biology Olivier, Magali; International Agency for Research on Cancer, Molecular Mechanisms and Biomarkers Group
Keywords:	virus discovery, papillomavirus, workflow, Amplicon sequencing, phylogeny

PVAmpliconFinder: a workflow for the identification of human papillomaviruses from high-throughput amplicon sequencing

Alexis Robitaille[1,*], Rosario N. Brancaccio[1], Sankhadeep Dutta[1#], Dana E. Rollison[2], Marcis Leja[3], Nicole Fischer[4,5], Adam Grundhoff[4,6], Tarik Gheit[1], Massimo Tommasino[1,*], Magali Olivier[1,*]

Affiliations

[1]International Agency for Research on Cancer, Lyon, France
[2]Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, Florida, USA
[3]Institute of Clinical and Preventive Medicine, University of Latvia, Riga, Latvia
[4]German Center for Infection Research, Hamburg, Borstel, Lübeck, Riems, Germany
[5]Institute for Medical Microbiology, Virology and Hygiene, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[6]Heinrich Pette Institut, Leibniz Institut for Experimental Virology, Hamburg, Germany

* To whom correspondence should be addressed: robitaillea@students.iarc.fr
Correspondence may also be addressed to olivierm@iarc.fr; tommasinom@iarc.fr

#Present Address: Department of Oncogene Regulation, Chittaranjan National Cancer Institute, Kolkata, India.

Keywords: Amplicon sequencing, virus discovery, papillomavirus, workflow, phylogeny

Abstract

The detection of known human papillomaviruses (PVs) from targeted wet-lab approaches has traditionally used PCR-based methods coupled with Sanger sequencing. With the introduction of next-generation sequencing (NGS), these approaches can be revisited to integrate the sequencing power of NGS. Although computational tools have been developed for metagenomic approaches to search for known or novel viruses in NGS data, no appropriate tool is available for the classification and identification of novel viral sequences from data produced by amplicon-based methods. We have developed **PVAmpliconFinder**, a data analysis workflow designed to rapidly identify and classify known and potentially new *Papillomaviridae* sequences from NGS amplicon sequencing with degenerate PV primers. Here, we describe the features of **PVAmpliconFinder** and its implementation using biological data obtained from amplicon sequencing of human skin swab specimens and oral rinses from healthy individuals. **PVAmpliconFinder** identified putative new HPV sequences, including one that was validated by wet-lab experiments. **PVAmpliconFinder** can be easily modified and applied to other viral families. **PVAmpliconFinder** addresses a gap by providing a solution for the analysis of NGS amplicon sequencing, increasingly used in clinical research. The **PVAmpliconFinder** workflow, along with its source code, is freely available on the GitHub platform: https://github.com/IARCbioinfo/PVAmpliconFinder.

1 Introduction

Papillomaviruses (PVs) are widely distributed across vertebrates. PVs are classified into genera, species, and types based on the nucleotide sequence identity of the major capsid protein L1. Human PVs (HPVs) have a tropism for the skin and mucosal epithelia of different anatomical sites and are organized into five major genera: alpha, beta, gamma, mu, and nu (1, 2). HPV infection is responsible for various diseases, including several types of cancer (3, 4). To date, more than 200 HPVs have been fully characterized (1, 5). Recent studies have provided evidence that many more HPV types exist (6, 7). Thus, it is important to comprehensively describe the family of HPV types and evaluate their role in human diseases.

Traditionally, single-step or nested PCR amplification using consensus or degenerate primers has been used for the identification and characterization of novel HPVs (8–10). This approach is time-consuming and laborious and has limitations in terms of sensitivity, especially in samples with low viral DNA load or in the case of co-infections with multiple HPV types. More recently, several PCR-based strategies using degenerate primers have been combined with the use of next-generation sequencing (NGS) to characterize PV virome composition or to search for new viruses (11–13). We have recently developed a novel approach that enabled the description of 105 putative new PV types in skin and oral samples (7). This approach required the development of a specific bioinformatics workflow, because no existing tools were adapted to our protocol design. Several bioinformatics tools have been developed to analyze NGS data for the detection of viruses, but most of them are designed to analyze the virome composition of known viruses in clinical settings, or to discover new viruses from DNA or RNA shotgun sequencing (14–21).

Here, we describe a new bioinformatics workflow, PVAmpliconFinder, specifically designed to rapidly identify and classify known and potentially novel viruses from the *Papillomaviridae* family from amplicon NGS using degenerate PV primers. PVAmpliconFinder is based on alignment similarity metrics, but also considers molecular evolution time for improved identification and taxonomic classification of novel PVs. The final output of the tool includes a list of fully characterized putative new *Papillomaviridae* sequences together with a graphical representation of the relative abundance and diversity of HPV sequence diversity in the tested samples.

2 Materials and Methods

Details of the workflow can be found in Supplementary Data 2. Briefly, PVAmpliconFinder takes paired-end FastQ files as input and applies common data preprocessing steps for quality control and filtering (Figure 1A). Then, data complexity is reduced before the identification of the PV-related sequences (Figure 1B). Groups of sequences are defined based on similarity between identified sequences and available PV sequences in the NCBI database (Figure 1C). *De novo* assembly is then performed to reconstruct the full amplified region covered by several primer systems (Figure 1D). Finally, the reconstructed sequences are taxonomically classified based on two independent methodologies, which are alignment-based and homology-based, respectively, before the generation of diverse output reports (Figure 1E and F).

3. Results

We applied the PVAmpliconFinder workflow (Figure 1) to the data obtained from amplicon sequencing of human skin swab specimens and oral rinses from healthy individuals, aiming to identify new PVs (the detailed protocol is in Supplementary Data 4). Different sets of degenerate primers targeting the L1 region of HPVs (7) were used to amplify 8 DNA sample pools. The 8 DNA sample pools were subjected to paired-end sequencing on the Illumina MiSeq system, generating about 2.65 million raw reads in total (331,359 raw reads on average per sample pool) (Table 1). PVAmpliconFinder was run with an info file describing the characteristics of each sample pool to enable the output of data stratified by tissue type and primer system (Supplementary Table S1).

3.1 Preprocessing and complexity reduction analysis

The first step of the analysis, consisting of quality trimming, had a small impact on the total numbers of reads, removing less than 2% of the reads in the 8 DNA sample pools (Table 1). Merging the paired reads (step 2) reduced by at least two-fold the total number of sequences but extended their length. Although more than 90% of the reads were merged for most samples, about 40% of the reads were not successfully merged at this step for DNA sample pool 6 (Table 1). A quality check of this sample pool with the FastQC report generated in step 2 enabled the identification of primer contamination in about 10% of the reads, explaining a sub-optimal reconstruction of the full insert (data not shown).
The following step, de-replication, consisted of collapsing identical sequences into a single template but keeping the information on the number of reads used to form the final template. For the 8 DNA sample pools, the different amplicons were highly represented, as shown by the substantial decrease in the number of unique sequences remaining after this step (about 5% of the total number of sequences after merging of the mate reads) (Table 1). Less than 1% of the sequences were identified as potentially chimeric (Table 1). Then, a *de novo* clustering of highly related sequences was performed to correct for sequencing and/or polymerase errors present at low frequency at each position. A user-defined threshold had been set to 98% of identity for two sequences to cluster together. This clustering step drastically reduced the number of unique sequences retained, decreasing the number of sequences from about 8% to 1% of the overall sequences considered in the preceding step (Table 1). Overall, for the entire run, about 28.5% (756,506/2,650,877) of the total raw reads were retained for the MegaBlast step (Supplementary Table S5A).

3.2 Identification of PV-related sequences and definition of groups

To identify the sequences in an unbiased manner, the sequences were aligned against the complete NCBI “nt” nucleotide sequence database, which includes all sequences from all species (Figure 1C). Subsequently, groups of sequences were defined based on two characteristics: the best MegaBlast subject sequence for each query, and the percentage of similarity of each sequence with its corresponding best subject sequence (Figure 1C).

3.2.1 Identification of PV-related sequences

On average, more than 90% of the centroid-clustered unique sequences of the 5 pools from skin swab specimens (S1-S5) matched against a *Papillomaviridae* family sequence, highlighting the specificity of the amplification using partially degenerate primers (Table 1). This represented a mean of 99.5% of *Papillomaviridae*-related reads among all reads submitted to MegaBlast from those 5 skin sample pools (Supplementary Table S5A).

For the 3 pools from oral rinses (S6-S8), about 86.5% of the centroid-clustered unique sequences had their best match against a *Papillomaviridae* family sequence (Table 1), representing a mean of 18.5% of *Papillomaviridae*-related reads among all reads submitted to MegaBlast from those 3 oral sample pools (Supplementary Table S5A).
A total of 549,280 reads (72.6%) of the sequences subjected to MegaBlast matched against *Papillomaviridae* family sequences (Supplementary Table S5A).

3.2.2 Definition of groups

When all the PV-related sequences identified above were grouped based on the best match and percentage of similarity, a total of 139 groups of PV sequences were found in the overall NGS run, including 136 known PVs or putative known PV variants (presenting less than 10% of dissimilarity with an already characterized PV) and 3 putative new PVs (Table 1). The known PV sequences corresponded to 549,273 raw reads, and the putative new PV sequences were supported by 7 raw reads (Supplementary Tables S5 E and G).

3.3 *De novo* assembly of grouped sequences

The grouped sequences for each sample pool were then *de novo* assembled to extend the sequence lengths in order to cover the full L1 region targeted by the different primer systems used in the PCRs (Figure 1D).

3.4 Taxonomic classification of PV sequences

The taxonomic classification of each PV sequence was then assigned to the extended sequences using two methods, one based on the taxonomic classification of the best subject match (using the e-value computed by BlastN) when aligned against a comprehensive database of PV sequences, and the other based on molecular evolution using the Randomized Axelerated Maximum Likelihood-Evolutionary Placement Algorithm (RaxML-EPA) (Figure 1E). For details, see Supplementary Data 2.
The results of the classification for DNA sample pool S5 (skin samples pool; CUT primer) are described in Table 2. In this sample pool, 2 putative new PV sequences represented by 5 reads and 39 putative known PV sequences represented by 60,892 reads were identified (Tables 1 and 2; Supplementary Tables S5 E and G).
One of the putative new PV sequences in this pool was represented by 3 reads (PV_2). The MegaBlast algorithm (using the full “nt” database) aligned it against “Gammapapillomavirus 13 isolate Gamma13_HIVGc158, complete genome” (MF588722.1) with 81.25% of identity. Of note, although the Gamma13_HIVGc158 is a complete genome, this sequence is not reported in the Papillomavirus Episteme (PaVE) database. The BlastN algorithm (using the PaVE database) aligned this sequence against HPV-mEV03c45 (MF588721), an unreferenced Gamma PV genome, with 78.69% of identity. RaxML-EPA found the best position of this putative new sequence in the reference tree close to HPV213 (MF509818), also a potential Gamma PV, but with pending approval of its classification by the International Committee on Taxonomy of Viruses (ICTV) (Tables 2 and 3). Although the three methodologies agreed on classifying this sequence as a putative Gamma PV, the two alignment methods did not perfectly align the putative new PV sequence (less than 85% similarity against known PVs).
Among the 39 putative known PV sequences identified in this sample pool, one was represented by about 7% of the total reads (4,211 raw reads out of 60,892 reads) (Table 2; Supplementary Table S6: Sequence identifier “69VIRUSput”). The MegaBlast algorithm

aligned this sequence to a partial cds (342 bp) of a major capsid protein L1 gene (isolate GC12_1; FJ969907.1) with nearly 99% of identity. In comparison, the BlastN alignment against the PaVE database aligned this sequence against a Gamma-10 referenced PV genome (HPV130; GU117630), with a percentage of identity below 10% (86.12%). When aligning the isolate GC12_1 partial cds and the HPV130 full genome with the MegaBlast algorithm, the two sequences presented 86.01% of identity on 98% coverage. Finally, RaxML-EPA found homology with EdPV2 (MH376689), an unclassified *Erethizon dorsatum* PV species (Table 2; Supplementary Table S6). EdPV2 was proposed to represent a new genus in the family *Papillomaviridae* (22). From these results, this 352 bp sequence may represent a novel PV type, although it remains to be fully characterized.

3.5 Relative unnormalized abundance of *Papillomaviridae*-related sequence: differences based on the methodology

The relative unnormalized abundance of *Papillomaviridae*-related sequences identified by MegaBlast, BlastN, and RaxML-EPA for all samples is shown in Figures 2, 3, and 4, respectively, and Supplementary Tables S2, S3, and S4 provide the detailed taxonomic assignation based on MegaBlast, BlastN, and RaxML-EPA, respectively. Beta-3 species were the most represented species identified by the three methods, with 42% of beta-3-related sequences identified by MegaBlast, and 62% identified by both BlastN and RaxML-EPA (Supplementary Tables S2, S3, and S4). The second most represented group was the “unclassified” sequences for MegaBlast (28% of the sequences), due to the incomplete taxonomic classification of a proportion of *Papillomaviridae*-related sequences present in the NCBI database. The third most represented genus based on MegaBlast was the gamma genus, with about 24% of the sequence, followed by the alpha genus (2%) and a small proportion of *Lambdapapillomavirus* (0.03%) due to the identification of a feline PV partial cds sequence (EF535004.1) in sample pools 1 and 2 (Supplementary Tables S2 and S6).

The second most represented group based on BlastN and RaxML-EPA was the unreferenced PVs, with a major subset putatively classified as unreferenced *Gammapapillomavirus* sequences (about 17%) and a small subset as unreferenced *Betapapillomavirus* sequences (about 1%). Of note, unreferenced sequences represented about 40% of the total entries available in the PaVE database version used (version of May 23, 2019). The third and fourth most represented genera were the referenced gamma and alpha PVs by both BlastN and RaxML-EPA (Figures 3 and 4). BlastN could not classify 0.008% of the sequences, due to a best subject sequence associated with an e-value under the threshold defined as 1e-1 (Supplementary Table S3). RaxML-EPA also classified 0.8% of the sequences as “Unclassified” because those sequences presented homology to a newly described *Erethizon dorsatum* PV (EdPV2; MH376689), not yet classified by the ICTV, and potentially the first representative genome of a new PV genus (22). Interestingly, the 46 reads that were unclassified by BlastN (due to the e-value threshold) were classified as *Taupapillomavirus* by RaxML-EPA, with homology to *Felis catus* PV type 4 and 5 (Supplementary Table S4).

3.6 Discovery and characterization of putative new PV-related sequences

Overall, from the entire run, a total of 3 putative new sequences belonging to the *Papillomaviridae* family were identified by the algorithm (Table 3). Based on MegaBlast, “PV_1” is close to an unreferenced Gamma-12 complete genome, also present in the PaVE database (MF588716). However, it shows a higher percentage of identity with HPV-

mSK197 (MH777339) based on BlastN alignment against the PaVE database. RaxML-EPA was in agreement with BlastN results, finding homology with the unreferenced *Gammapapillomavirus* HPV-mSK197. “PV_2” presented similarity (based on MegaBlast) with an unreferenced Gamma-13 complete genome (MF588722), which is absent from the PaVE database. BlastN found similarity with HPV-mEV03c45 (MF588721), an unreferenced *Gammapapillomavirus* genome, and RaxML-EPA found homology to HPV213 (MF509818), a referenced but unofficially classified *Gammapapillomavirus* genome. “PV_3” presented similarity with an unclassified partial cds of the isolate GC04 (FJ969896), but presented a higher similarity with the unreferenced HPV-mSK014 (MH777162) when aligned using BlastN. RaxML-EPA also found homology with the same HPV-mSK014 unreferenced *Gammapapillomavirus* genome. The sequence sizes ranged from 160 to 372 nucleotides, and all sequences presented more than 15% of dissimilarity with non-referenced PV sequences based on MegaBlast. All were amplified from skin DNA samples, using FAP and CUT primers (8, 9).

A previous analysis of the same data had led to the characterization of the full genome sequence of a novel Gamma-8 PV (Table 3, “37VIRUSput“) (23). In the current analysis, this sequence appeared in the putative known PV sequence, because it is now included in the NCBI database (MF356498.1) as well as in the PaVE database. However, this sequence is still assigned to an unclassified group by the BlastN algorithm because the taxonomy has not yet been updated in the PaVE database (Supplementary Table S6, “37VIRUSput”). The official number of this novel PV, named “HPV isolate ICB1” in the NCBI database, is HPV224.

3.7 Performances

The PVAmpliconFinder execution time on this dataset was less than 150 minutes when using an indexed NCBI database (Table 4). The most time-consuming step was the MegaBlast search against the full “nt” NCBI database (more than 95% of total time). When using a non-indexed NCBI “nt” database, the MegaBlast computational time was reduced to less than 5 minutes (Supplementary Table S7). For most of the steps, parallelization at the sample level was implemented to reduce the total computation time.

4. Discussion

We developed PVAmpliconFinder, a complete workflow enabling the discovery and identification of viral sequences related to the *Papillomaviridae* family from targeted amplicon sequencing by NGS. PVAmpliconFinder is an easy single-line command workflow that takes FastQ files as input files and generates tabular and graphical output files that describe the nature and abundance of PV-related sequences present in a complex mixture of host, phage, bacterial, and viral DNA. The data output discriminates between putative new and previously known *Papillomaviridae*-related sequences. Furthermore, it includes sequencing metrics and sequence details, enabling the design of subsequent laboratory experiments for confirming the *in silico* findings (Supplementary Data 3).

In contrast to read-subtraction methods, PVAmpliconFinder performs an alignment step against the entire NCBI database. This is a deliberate choice because removing host sequences may remove potentially new viral sequences that present some similarity to the host. Indeed, viruses are the fastest mutating DNA element on Earth (24), so the chance of random sequence similarity between a large host genome and a small viral sequence is high. Moreover, the use of degenerate primer leads to the amplification of more diverse

pieces of DNA and finding the best match against a *Papillomaviridae* sequence when aligning against a multi-organism database provides more robust results.

Several steps of the workflow are specifically tailored to deal with the specificity of NGS amplicon sequencing: the merging of the read pairs, enabling the reconstruction of the full insert; the de-replication step, to reduce data complexity and keep only one copy of identical sequences; and the elimination of chimeric sequences (PCR-derived sequences should be represented by at least two copies during the de-replication step; thus, single copies are probably sequences without biological significance). The number of de-replicated sequences corresponding to each template is saved in memory by the program to compute an unnormalized abundance. A step of clustering of highly related sequences is applied to correct for PCR amplification and sequencing errors. Because 2% of dissimilarity from any known L1 gene is enough to define a new PV variant (25), the tool uses a 98% identity threshold for clustering by default. When searching for new PV types (at least 10% of dissimilarity on the L1 gene), this threshold is a good compromise between sensitivity and specificity, because the potential loss of precision at the variant taxonomic level may be counterbalanced by an increased specificity of the reconstructed sequence.

To identify sequences in an unbiased manner, the sequences are aligned against the entire “nt” NCBI database. Although this step is time-consuming due to the large size of the database, it reduces the false-positive discovery rate. Indeed, querying a database with reduced diversity (such as a virus database) using the e-value as a threshold could increase the chances of getting a hit even if the subject sequence has a low identity with the queried sequence. Considering only the sequences that have their best match against a *Papillomaviridae* family sequence produces an unbiased result.

PVAmpliconFinder includes a grouping step to separate sequences that are putative new PVs from those that are already known PVs, using the threshold of 10% of dissimilarity. This grouping is done before the *de novo* assembly and classification steps because, although they are partially degenerate, the primers favor the amplification of known PV sequences. Because the tool is focused on the discovery of new PVs, it is important to separate potential new sequences at the earliest possible stage. A *de novo* assembly step is performed because of the possibility of using several primer sets that have different hybridization positions along the L1 gene. The objective is to reconstruct the longest possible sequence for each potential PV sequence.

PVAmpliconFinder uses an advanced identification and taxonomic classification of the sequences using both sequence similarity and homology. For the sequence similarity, the BlastN algorithm is used against the PaVE database (5). This database is the most complete PV database. It includes PV sequences validated by full genome resequencing, but also several “non-referenced” genomes that are not classified taxonomically. Currently, non-referenced PV genomes in the PaVE database represent more than 37% of the overall available PV genomes (244/649), and this percentage continues to increase (26, 27). PVAmpliconFinder presents the results based on the initial MegaBlast step and those obtained based on BlastN alignment against the PaVE database, but a huge number of sequences remain unclassified using the former approach because they match against incomplete L1 cds. Moreover, pairwise alignment with a low percentage of similarity raises a concern about the pertinence of the results obtained. This is especially true for the 3 putative new sequences identified in the application example reported here, because all sequences had at least 15% of dissimilarity against their best match. To circumvent this limitation, we use a complementary approach in parallel based on a molecular evolution method: RaxML-EPA (28). A multiple sequence alignment is used to infer evolutionary time and to reconstruct a phylogenetic reference tree of selected species. Then, the

Parsimony-based Phylogeny-Aware Read alignment (PaPaRa) algorithm is used to find the best position of the sequence into the reference multiple sequence alignment (29). RaxML-EPA is subsequently used to find the best position of those sequences in the reference tree. The accuracy of the PaPaRa alignment is critical for the correct positioning of the query sequence into the reference tree.

Some limitations of the PVAmpliconFinder workflow are due to the inherent limitations of the methods implemented. Evolutionary based methods such as RaxML suffer from long-branch attraction errors. Long-branch attraction is an error where distant lineages are inferred to be close relatives because both have undergone a large number of changes. This is what is suspected to happen for the classification by EPA of the *Erethizon dorsatum* sequences identified in our experiment. They are inferred to be close to EdPV2 (MH376689), a recently referenced but unclassified *Erethizon dorsatum* PV (22), presenting large differences from other known PVs on its L1 gene, and thought to represent a new genus in the family *Papillomaviridae*. Although this led to an incomplete classification, these sequences may represent new species or virus features. Finally, PVAmpliconFinder does not control for potential contamination. Cross-contamination between samples during library preparation, amplification, and sequencing, or environmental contamination are difficult to detect using *in silico* methods. Low-abundance sequences may truly be present in the samples but may also come from cross-contamination from another sample. PVAmpliconFinder will report sequences represented by only 2 reads. These low-abundance sequences should be considered with caution. Defining an empirical abundance threshold could be considered. Environmental contamination may explain the presence of non-human PV in human samples. However, cross-contamination between species has recently been described (30, 31) and thus cannot be excluded.

While there is an increasing use of NGS amplicon sequencing in the clinical research setting, only few bioinformatics methods are available for the sensitive detection of HPV, and they are often restricted to a panel of already well characterized PV types (32). The use of degenerated primers and PVAmpliconFinder may thus provide a solution for the detection and discovery of a broad range of HPV types.

In summary, we have developed the first bioinformatics tool for the identification of novel viruses of the *Papillomaviridae* family from amplicon sequencing data. This tool addresses a gap because no other tool exists for the analysis of this type of data. PVAmpliconFinder uses an advanced identification and taxonomic classification of the viral sequences extracted, which combines methodologies based on sequence similarity and homology. PVAmpliconFinder produces several tabular and graphical outputs that provide the necessary information to select the most promising putative new PV sequences that may be validated by further wet-lab approaches. Furthermore, PVAmpliconFinder can be easily modified and applied to other viral families, because this would only require a change in the interrogated databases and the reconstruction of a reference tree for the viral family considered. As no other tool exist for the analysis of NGS amplicon sequencing data of PV, PVAmpliconFinder addresses a gap with potential application in clinical research settings.

5. Data availability

The PVAmpliconFinder workflow, along with its source code, is freely available on the GitHub platform: https://github.com/IARCbioinfo/PVAmpliconFinder. Raw sequencing files have been deposited in the NCBI database under the BioProject accession number PRJNA555194.

6. Supplementary Data

Supplementary Data are available at NAR online.

7. Funding

The work performed in the groups is partially supported by grants from Institut National de la Santé et de la Recherche Médicale [ENV201610]; Fondation ARC pour la Recherche sur le Cancer [JA 20151203192]; Deutsche Krebshilfe [no. 110259], and the National Cancer Institute of the National Institutes of Health [grant 1R01-CA17758].

8. Acknowledgements

We thank all members of the Infections and Cancer Biology Group for their constant support, Dr. Matthieu Foll for his useful advice, and Karen Müller for editing the manuscript.

9. Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer (IARC)/World Health Organization (WHO), the authors alone are responsible for the views expressed in this article, and they do not necessarily represent the decisions, policy, or views of IARC/WHO.

References

1. Bzhalava,D., Eklund,C. and Dillner,J. (2015) International standardization and classification of human papillomavirus types. *Virology*, **476**, 341–344.

2. de Villiers,E.-M. (2013) Cross-roads in the classification of papillomaviruses. *Virology*, **445**, 2–10.

3. Tommasino,M. (2014) The human papillomavirus family and its role in carcinogenesis. *Seminars in Cancer Biology*, **26**, 13–21.

4. Bouvard,V., Baan,R., Straif,K., Grosse,Y., Secretan,B., Ghissassi,F.E., Benbrahim-Tallaa,L., Guha,N., Freeman,C., Galichet,L., *et al.* (2009) A review of human carcinogens—Part B: biological agents. *The Lancet Oncology*, **10**, 321–322.

5. Van Doorslaer,K., Tan,Q., Xirasagar,S., Bandaru,S., Gopalan,V., Mohamoud,Y., Huyen,Y. and McBride,A.A. (2013) The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res*, **41**, D571–D578.

6. Pastrana,D.V., Peretti,A., Welch,N.L., Borgogna,C., Olivero,C., Badolato,R., Notarangelo,L.D., Gariglio,M., FitzGerald,P.C., McIntosh,C.E., *et al.* (2018) Metagenomic Discovery of 83 New Human Papillomavirus Types in Patients with Immunodeficiency. *mSphere*, **3**.

7. Brancaccio,R.N., Robitaille,A., Dutta,S., Cuenin,C., Santare,D., Skenders,G., Leja,M., Fischer,N., Giuliano,A.R., Rollison,D.E., *et al.* (2018) Generation of a novel next-generation sequencing-based method for the isolation of new human papillomavirus types. *Virology*, **520**, 1–10.

8. Chouhy,D., Gorosito,M., Sánchez,A., Serra,E.C., Bergero,A., Fernandez Bussy,R. and Giri,A.A. (2010) New generic primer system targeting mucosal/genital and cutaneous human papillomaviruses leads to the characterization of HPV 115, a novel Beta-papillomavirus species 3. *Virology*, **397**, 205–216.

9. Forslund,O., Antonsson,A., Nordin,P., Stenquist,B. and Göran Hansson,B. (1999) A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. *Journal of General Virology*, **80**, 2437–2443.

10. Forslund,O., Ly,H. and Higgins,G. (2003) Improved detection of cutaneous human papillomavirus DNA by single tube nested 'hanging droplet' PCR. *Journal of Virological Methods*, **110**, 129–136.

11. Kocjan,B.J., Bzhalava,D., Forslund,O., Dillner,J. and Poljak,M. (2015) Molecular methods for identification and characterization of novel papillomaviruses. *Clinical Microbiology and Infection*, **21**, 808–816.

12. Ekström,J., Mühr,L.S.A., Bzhalava,D., Söderlund-Strand,A., Hultin,E., Nordin,P., Stenquist,B., Paoli,J., Forslund,O. and Dillner,J. (2013) Diversity of human papillomaviruses in skin lesions. *Virology*, **447**, 300–311.

13. Mühr,L.S.A., Hultin,E., Bzhalava,D., Eklund,C., Lagheden,C., Ekström,J., Johansson,H., Forslund,O. and Dillner,J. Human papillomavirus type 197 is commonly present in skin tumors. *International Journal of Cancer*, **136**, 2546–2555.

14. Borozan,I., Wilson,S., Blanchette,P., Laflamme,P., Watt,S.N., Krzyzanowski,P.M., Sircoulomb,F., Rottapel,R., Branton,P.E. and Ferretti,V. (2012) CaPSID: A bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. *BMC Bioinformatics*, **13**, 206.

15. Zhao,G., Wu,G., Lim,E.S., Droit,L., Krishnamurthy,S., Barouch,D.H., Virgin,H.W. and Wang,D. (2017) VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology*, **503**, 21–30.

16. Takeuchi,F., Sekizuka,T., Yamashita,A., Ogasawara,Y., Mizuta,K. and Kuroda,M. (2014) MePIC, Metagenomic Pathogen Identification for Clinical Specimens. *Jpn J Infect Dis*, **67**, 62–65.

17. Hong,C., Manimaran,S., Shen,Y., Perez-Rogers,J.F., Byrd,A.L., Castro-Nallar,E., Crandall,K.A. and Johnson,W.E. (2014) PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, **2**, 33.

18. Wang,Q., Jia,P. and Zhao,Z. (2015) VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Medicine*, **7**, 2.

19. Zheng,Y., Gao,S., Padmanabhan,C., Li,R., Galvez,M., Gutierrez,D., Fuentes,S., Ling,K.-S., Kreuze,J. and Fei,Z. (2017) VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology*, **500**, 130–138.

20. Chen,Y., Yao,H., Thompson,E.J., Tannir,N.M., Weinstein,J.N. and Su,X. (2013) VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics*, **29**, 266–267.

21. Lagström,S., Umu,S.U., Lepistö,M., Ellonen,P., Meisal,R., Christiansen,I.K., Ambur,O.H. and Rounge,T.B. (2019) TaME-seq: An efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration. *Scientific Reports*, **9**, 524.

22. Vanmechelen,B., Maes,R.K., Sledge,D.G., Lockwood,S.L., Schwartz,S.L. and Maes,P. (2018) Genomic characterization of Erethizon dorsatum papillomavirus 2, a new papillomavirus species marked by its exceptional genome size. *Journal of General Virology*, **99**, 1699–1704.

23. Brancaccio,R.N., Robitaille,A., Dutta,S., Rollison,D.E., Fischer,N., Grundhoff,A., Tommasino,M. and Gheit,T. (2017) Complete Genome Sequence of a Novel Human Gammapapillomavirus Isolated from Skin. *Genome Announc*, **5**.

24. Virgin,H.W. (2014) The Virome in Mammalian Physiology and Disease. *Cell*, **157**, 142–150.

25. Bernard,H.-U., Burk,R.D., Chen,Z., van Doorslaer,K., Hausen,H. zur and de Villiers,E.-M. (2010) Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology*, **401**, 70–79.

26. Simmonds,P., Adams,M.J., Benkő,M., Breitbart,M., Brister,J.R., Carstens,E.B., Davison,A.J., Delwart,E., Gorbalenya,A.E., Harrach,B., *et al.* (2017) Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, **15**, 161–168.

27. Tirosh,O., Conlan,S., Deming,C., Lee-Lin,S.-Q., Huang,X., Su,H.C., Freeman,A.F., Segre,J.A. and Kong,H.H. (2018) Expanded skin virome in DOCK8-deficient patients. *Nature Medicine*, **24**, 1815.

28. Berger,S.A., Krompass,D. and Stamatakis,A. (2011) Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Syst Biol*, **60**, 291–302.

29. Berger,S.A. and Stamatakis,A. PaPaRa 2.0: A Vectorized Algorithm for Probabilistic Phylogeny-Aware Alignment Extension.

30. Bravo,I.G. and Félez-Sánchez,M. (2015) Papillomaviruses: Viral evolution, cancer and evolutionary medicine. *Evol Med Public Health*, **2015**, 32–51.

31. Gottschling,M., Göker,M., Stamatakis,A., Bininda-Emonds,O.R.P., Nindl,I. and Bravo,I.G. (2011) Quantifying the phylodynamic forces driving papillomavirus evolution. *Mol. Biol. Evol.*, **28**, 2101–2113.

32. Schmitt,M., Depuydt,C., Benoy,I., Bogers,J., Antoine,J., Arbyn,M. and Pawlita,M. (2013) Multiple Human Papillomavirus Infections with High Viral Loads Are Associated with Cervical Lesions but Do Not Differentiate Grades of Cervical Abnormalities. *Journal of Clinical Microbiology*, **51**, 1458–1464.

Figure Legends

Figure 1: Workflow of PVAmpliconFinder
Figure 2: Graphical representation of the unnormalized abundance of PV genera and species in terms of number of reads based on MegaBlast alignment
Figure 3: Graphical representation of the unnormalized abundance of PV genera and species in terms of number of reads based on BlastN alignment
Figure 4: Graphical representation of the unnormalized abundance of PV genera and species in terms of number of reads based on RaxML-EPA

Appendix. Supporting information

Supplementary Tables

Supplementary Table S1: Info file
Supplementary Table S2: Taxonomic classification of the reads identified in the overall NGS experiment by MegaBlast alignment
Supplementary Table S3: Taxonomic classification of the reads identified in the overall NGS experiment by BlastN alignment
Supplementary Table S4: Taxonomic classification of the reads identified in the overall NGS experiment by RaxML-EPA
Supplementary Table S5: NGS metrics, summary classification of putative known and putative new virus based on the three methodologies
Supplementary Table S6: Putative known *Papillomaviridae*-related sequences detected in the NGS experiment
Supplementary Table S7: Performances using non-indexed NCBI database

Supplementary Data

Supplementary Data 1: Info file description
Supplementary Data 2: Details of the workflow steps
Supplementary Data 3: Description of output files format
Supplementary Data 4: Sample collection, preparation, and sequencing

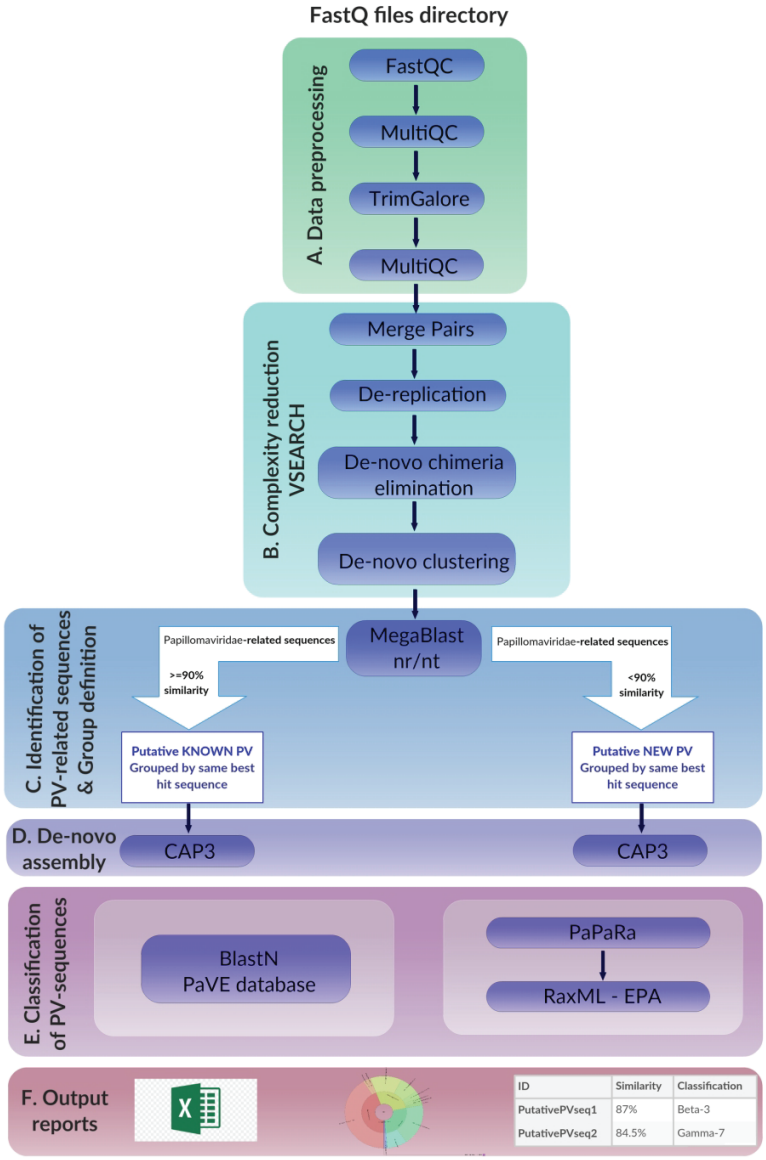


Figure 1: Workflow of PVAmpliconFinder

141x215mm (300 x 300 DPI)

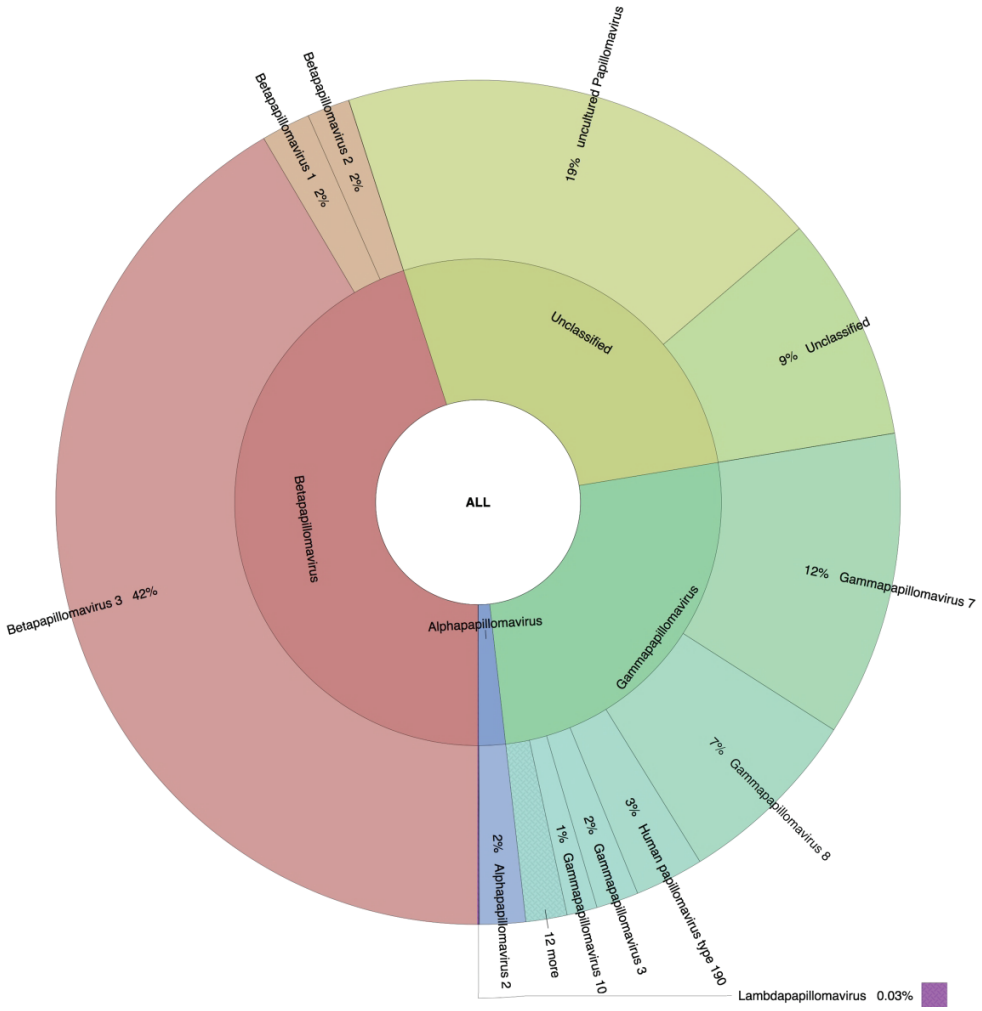


Figure 2: Graphical representation of the unnormalized abundance of PV genera and species in terms of number of reads based on MegaBlast alignment

456x471mm (300 x 300 DPI)

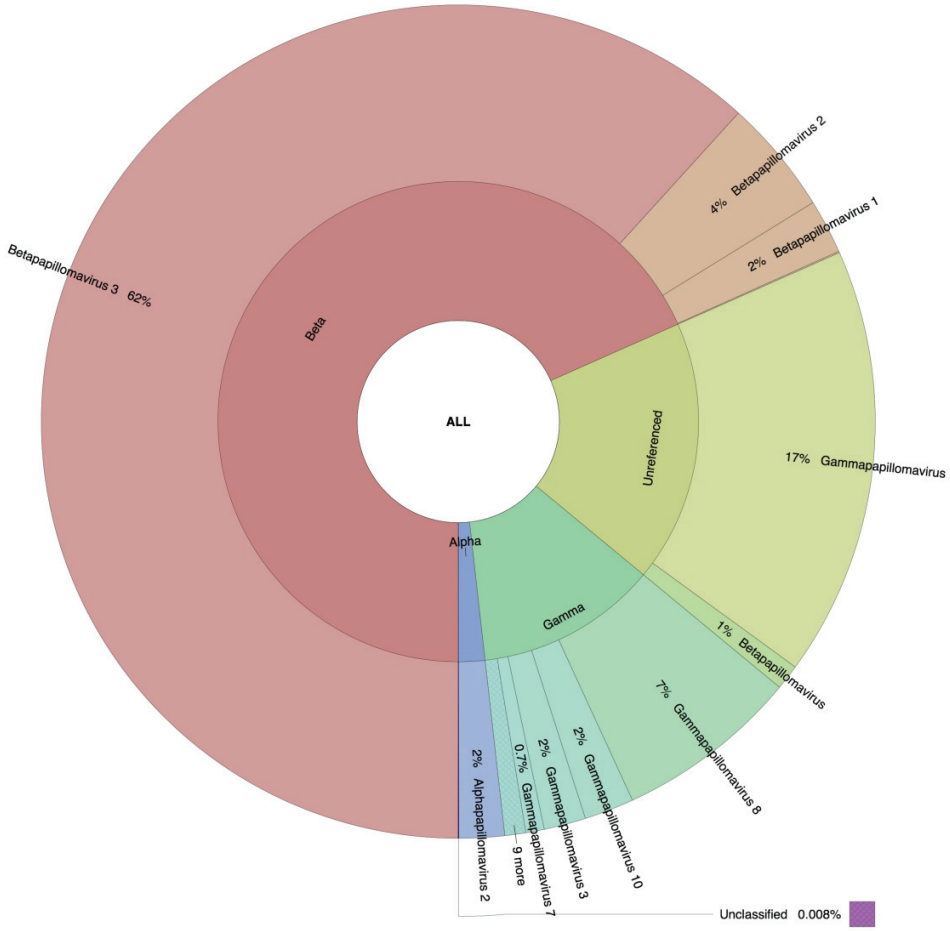


Figure 3: Graphical representation of the unnormalized abundance of PV genera and species in terms of number of reads based on BlastN alignment

462x458mm (300 x 300 DPI)

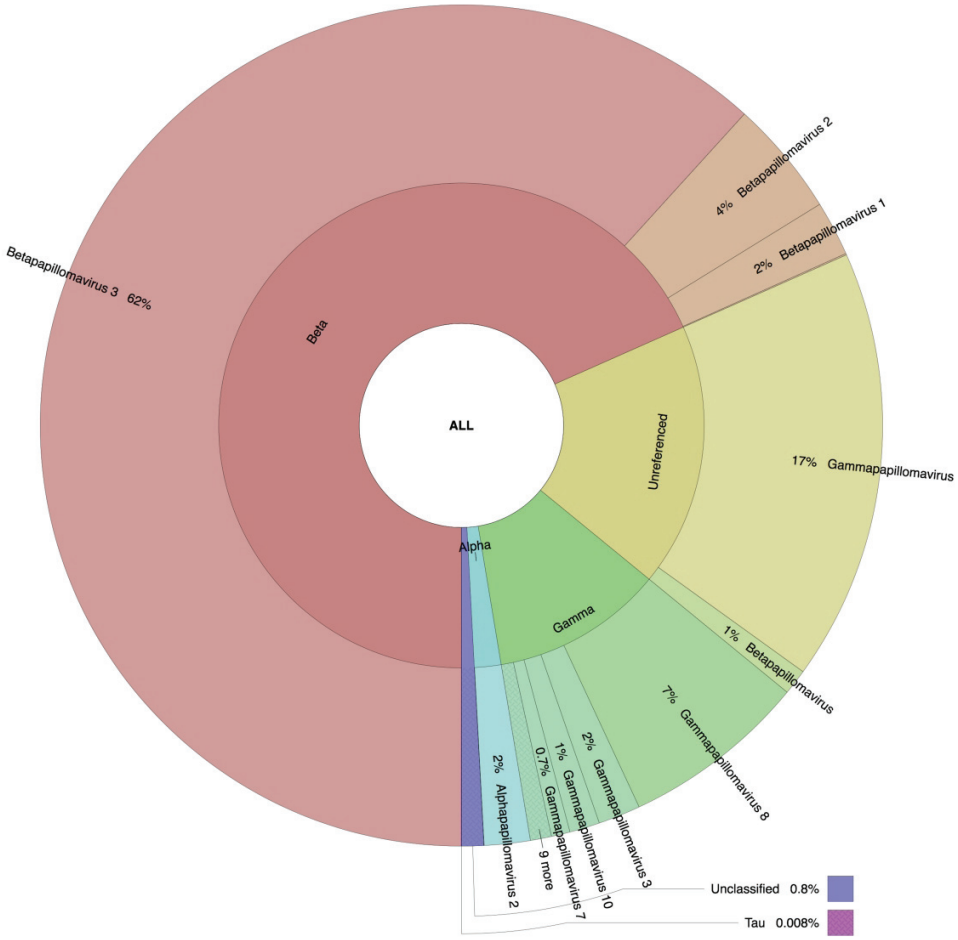


Figure 4: Graphical representation of the unnormalized abundance of PV genera and species in terms of number of reads based on RaxML-EPA algorithm

457x463mm (300 x 300 DPI)

Table 1: Summary of the number of individual sequences considered at each step of the workflow

Step	Total sequencing raw reads	TrimGalore		Merging		Dereplication		Chimeric identification		Clustering		Papillomaviridae best hit (eval<=1e-5)		Defined group (same best hit)			
														Putative new (>10% dissimilarity)		Putative known (<10% dissimilarity)	
Samples	N paired-end reads	N paired-end reads	%	N sequences	%	N sequences	%	N sequences	%	N sequences	%	N sequences	%	N sequences	%	N sequences	%
S1	564435	564064	99.93	551266	97.73	22551	4.09	22498	99.76	79	0.35	61	77.22	0	0	5	8.20
S2	62148	61708	99.29	58031	94.04	3281	5.65	3268	99.60	162	4.96	138	85.19	0	0	6	4.35
S3	316297	315999	99.91	307400	97.28	15562	5.06	15562	100.00	51	0.33	49	96.08	1	2.04	18	36.73
S4	109441	109326	99.89	106406	97.33	4842	4.55	4822	99.59	62	1.29	62	100.00	0	0	28	45.16
S5	309779	309390	99.87	294563	95.21	14101	4.79	14091	99.93	140	0.99	129	92.14	2	1.55	39	30.23
S6	554415	551742	99.52	331648	60.11	13820	4.17	13738	99.41	1162	8.46	910	78.31	0	0	16	1.76
S7	470655	467944	99.42	421764	90.13	28729	6.81	28659	99.76	609	2.12	513	84.24	0	0	16	3.12
S8	263707	263270	99.83	244177	92.75	13293	5.44	13283	99.92	194	1.46	188	96.91	0	0	8	4.26
Total num	2650877	2643443	99.72	2315255	87.58	116179	5.02	115921	99.78	2459	2.12	2050	83.37	3	0.45	136	16.73

Table 2: Taxonomic classification of Papillomaviridae-related reads from Sample 5

	Putative New			Putative known		
N total sequences	N=2 (5 reads)			N=39 (60,892 reads)		
N sequence shown	N=1 (3 reads)			N=1 (4,211 reads)		
	Megablast	**BlastN**	**RaxML**	**Megablast**	**BlastN**	**RaxML**
Alpha	-	-	-	-	-	-
Beta	-	-	-	-	-	-
Gamma	3 (100%)	3 (100%)	3 (100%)	-	4,211 (100.00%)	-
Unclassified	-	-	-	4,211 (100.00%)	-	4,211 (100.00%)
Lambda	-	-	-	-	-	-
Tau	-	-	-	-	-	-

Feuille1

Table 3: Putative new *Papillomaviridae*-related sequences identified by PVAmpliconFinder

VIRUSname	PV_1	PV_2	PV_3	37VIRUSput
%dissimilarity	16.67	18.75	20.49	0.85
Abundance	0.0025	0.0049	0.0033	3.2918
N°reads	2	3	2	698
GInum	gi|1273499301|gb|MF588716.1|	gi|1273499348|gb|MF588722.1|	gi|270048224|gb|FJ969896.1|	gi|1214938671|gb|MF356498.1|
AlignmentPosition_MegaBlast	3-78	1-352	207-327	1-236
VIRUS_closest_MegaBlast	Gammapapillomavirus 12 isolate Gamma12_EV07c367, complete genome	Gammapapillomavirus 13 isolate Gamma13_HIVGc158, complete genome	Human papillomavirus isolate GC04 major capsid protein L1 gene, partial cds	Human papillomavirus isolate ICB1, complete genome
Pool	pool3-skin-pathogen_S3_L001	pool5-skin-pathogen_S5_L001	pool5-skin-pathogen_S5_L001	pool4-skin-pathogen_S4_L001
Tissu	skin	skin	skin	skin
Primer	FAP	CUT	CUT	FAPM1
Length	160	353	372	262
AlignmentPositionBlastN_start:stop(length)	1-160(160)	1-352(352)	3-370(368)	1-255(255)
VIRUS_closest_Blast	HPV-mSK197(92.5%)	HPV-mEV03c45(78.69%)	HPV-mSK014(94.57%)	HPV224(97.25%)
BlastN_Classification	Gammapapillomavirus	Gammapapillomavirus	Gammapapillomavirus	Gammapapillomavirus
RaxML_closest_PV	HPV-mSK197	HPV213	HPV-mSK014	HPV224
RaxML_Classification	Gammapapillomavirus	Gammapapillomavirus	Gammapapillomavirus	Gammapapillomavirus
Sequence(s)	TAACAGTGGGCCACCCTTATTTCAGTGTTA AGAATGAAGGCACACAAGCCATAGTAGTTC CAAAGGTTTCAGGAGACCAGTTTAGAGTTT TCAGATTAAGACTCCCAGATCCTAACAAAT TTGCTTTAATAGACCCATCTATATATAATCC AGAAAGAGA	GCCGGATCCGAATAAGTTTGCATTGATAGA TCAGGACATTTATAATCCAGAAACGGAGAG ATTAGTTTGGAGAGTTAAAGGCTTAGAGGT TGACAGAGGTGGTCCTCTAGGTATTGGAG CTGTAGGTCATCCTTTATTAAATAAATATGG AGATACAGAAAATCCTTTGGGAAGACCCAT TCCAGAACAAGATGATAATAGAGTTAATTTA TCGTTTGAACCAAAACAAACTCAAATTCTTA TTGTTGGTTGTGCACCTCCTATAGGACAAC ATTGGGACGTTACAACACCTTGTAATAAAC AGAATGCAGGCGAATGTCCACCTATAGCAT	TGCCGGATCCGAATAAGTTTGCAATTGCAG ATACTTGCTTGTATAATCCTGAAAAGGAGC GCTTGGTATGGCAGTTAGTGGGTTTAGAAG TTGACAGAGGTGGTCCTTTAGGAATTGGAG CCACCGGTCACCCATATTTCAATAAATATG TAGATACAGAAAATCCAGTAGCATATCCTC CAAAGCAAGAAGAAGCAGCATTAGATAGCA GGCAAGATATGTCCTTTGACCCTAAACAAG TACAAATGATAATTGTGGGCTGTGCACCTC CAACAGGAGAATATTGGGACACAACTAAAT TTTGTGAATCTCATAAAAGTAGCCCAGGAG	TAACAGTTGGGCATCCTTATTTCAATAAAAT TCAAGACACAGAAAACCCAAATAAATATGT GCCTAAAGCGGGCGATGAAAACAGATTAAA TATTAGTGTTGATCCAAAACAGGTACAGCT ACTTATTGTGGGCTGTGTGCCTGCAACAGG AGAACATTGGGATATTGCAAGGCCATGTGA TGATGAGCAAAATGCTGGTGACTGTCCTCC TATCCAGCTTTTAAATACTGTAATTCAGGAT GGCGATATGAGAGATATCGG

Table 4: Performance metrics of PVAmpliconFinder

Step	Total sequencing raw reads	TrimGalore	Merging	Dereplication	Chimeric identification	Clustering	MegaBlast	De-novo assembly	BlastN	RaxML	End
Time	+0'00''	+0'26"	+1'13"	+0'06"	+0'02''	+0'16"	+145'33''	+0'32''	+0'29"	+4'30"	+0'01"
Cumulative time	+0'00''	+0'26"	+1'39"	+1'45"	+1'47''	+2'03"	+147'36"	+148'08"	+148'40"	+149'10"	+149'11"

Supplementary Data 1: Info file description

This file should be a simple tabular text file (.txt or .csv) containing in the first column a character string corresponding to the FastQ file name upstream of the “R1” or “R2” tag (one line for one sample). The second column should be named “primer”, and should contain the information about the primer set used to amplify the L1 region (e.g. CUT or FAP) of the corresponding sample. The third and last column should be named “tissue” and should describe the sample source (e.g. skin or oral swabs). This information will be used during the creation of the output files and will help to distinguish the virome composition and the new target coming from different tissue types and amplified by different primer sets. If this option remains empty while the program is called, all the samples will be consider as coming from the same tissue type and being amplified with the same degenerate primers.

Supplementary Data 2: Details of the workflow steps

2.1 Input data type and format

The PVAmpliconFinder workflow is designed for the analysis of sequencing reads generated from paired-end sequencing of DNA amplified using degenerate primers targeting specifically the L1 sequence of papillomaviruses (1–3). These primers enable the amplification of a region in the L1 gene out of a region of approximately 450 bp. The input data are FastQ files that can be uncompressed or compressed. The files will be automatically uncompressed if the detected format is a common compression format such as .zip, .gz, or .tar.gz. FastQ files from the forward and reverse reads of the same sample should have the same name, with only “R1” and “R2” differentiating the two files.

2.2 Input parameters

Three mandatory input parameters must be set: the path to the input directory that contains the FastQ files; a tag corresponding to the suffix of the FastQ file names to be selected for the analysis in the input directory; and the path to the output directory where the output files will be written. The following optional input arguments can also be set: [1] the name of the identifier of the NCBI “nt” database to be used (the Blast database should be present in the environment, and the default value is “nt”); [2] the number of threads to be used for the analysis (the default value is 2); [3] the directory path of an info file containing information on sample type and primer used (see Supplementary Table 1 for an example, and Supplementary Data 1 for how to format the file); [4] the threshold for the percentage of identity to be used in the *de novo* centroid-based clustering (the default value is 98).

2.3 Data preprocessing

The preprocessing of the FastQ files includes an initial quality control (QC) of the raw FastQ files using FastQC (4) and the aggregation of the FastQC reports using MultiQC (5) (Figure 1A). FastQ files are then trimmed for adapter sequences and sequences of amplification primers if required, using TrimGalore (6). This step also discards low-quality bases, sequences of less than 32 bp, poly-A sequences, and reads with low average quality score. FastQC and MultiQC are run on the trimmed FastQ files for a final QC (Figure 1A).

2.4 Complexity reduction and removal of artifacts

The step’s aim is to eliminate the redundant sequences generated during the different PCR steps preceding sequencing and to correct sequencing and/or polymerase errors. Four modules from the existing tool VSEARCH (7) are used to perform three different steps, as described below (Figure 1B).

2.4.1 Merging of reads

The ”fastq_mergepairs” module merges Read 1 and Read 2 pairs and reconstructs the full amplicon (around 450 bp);

2.4.2 De-replication of reads

The “derep_fulllength” module de-replicates reads by keeping only one template of several identical sequences. This step is particularly important because duplicates are generated during the PCR amplification steps used to amplify the L1 region as well as for the pre-sequencing processing of samples.

2.4.3 Chimera detection

The “uchime_denovo” module is then run to identify and remove chimeric DNA sequences that often form during PCR amplification, especially when sequencing a unique region. The option –minuniquesize is used with 2 as default value to account for the fact that at

this step each of the sequences is expected to be represented by at least 2 raw sequencing reads, corresponding to a minimum of one PCR cycle.

2.4.4 Reduction of amplification artifacts

The "cluster_size" module consists of *de novo,* centroid-based clustering of the sequences sharing more than a user-defined level of identity: 98% is the default value. This unique sequence will be used for downstream analysis. 2% of dissimilarity from any known L1 gene is enough to define a new PV variant (8). When searching for new PV types (at least 10% of dissimilarity on the L1 gene), 98% of identity enables a good clustering to balance between sensitivity and specificity.

2.5 Identification of PV-related sequences

All sequences identified by the preceding metagenomic analysis are subject to a MegaBlast alignment against the full "nt" nucleotide collection from the NCBI database (default parameters) (9). All sequences that have their best hit against any sequence belonging to the *Papillomaviridae* family with an e-value smaller than or equal to 1e-5 are kept for the next steps of the workflow (Figure 1C). *Papillomaviridae*-related sequences are identified using a lineages file created using the "ncbitax2lin" tool (10) (https://github.com/zyxue/ncbitax2lin).

2.6 Classification of PV sequences

This step uses two different approaches based on two different tools, BlastN and Randomized Axelerated Maximum Likelihood-Evolutionary Placement Algorithm (RaxML-EPA) (11), and the results of both approaches are returned. With RaxML-EPA, a method based on molecular evolution, a full taxonomic classification of the putative new sequences is obtained based on the homology of each sequence to its closest taxon. In both approaches, PV sequences are first grouped based on both the best MegaBlast subject sequence for each query and the percentage of similarity of this sequence with its corresponding best subject sequence. Then, a *de novo* assembly of sequences formed by this "two-features" grouping is performed with CAP3 (12) to reconstruct the full PCR amplicon because the different primers systems used are not targeting exactly the same L1 region. Finally, a taxonomic classification is performed on the reconstructed sequences (Figure 1C and Figure 1D), as detailed below.

2.6.1 Definition of groups and *de novo* assembly

For each sample, the sequences that have their best MegaBlast hit against a sequence belonging to the *Papillomaviridae* family are kept for the analysis. These sequences are grouped if their best hit is the same subject sequence. Subsequently, the grouped sequences are split into two groups: [1] putative known PVs, corresponding to sequences that present less than 10% of dissimilarity on their aligned portion with a known PV; [2] putative new PVs, corresponding to sequences that present more than 10% of dissimilarity on their aligned portion with a known PV. A *de novo* assembly is then performed for each group with CAP3 with default parameters (12) for contigs reconstruction (Figure 1C and Figure 1D).

2.6.2 BlastN-based taxonomical classification

Each contig sequence reconstructed during the previous step is then classified based on the taxonomic classification of its best alignment (BlastN best match) against the full L1 gene nucleotide sequence database available in the Papillomavirus Episteme (PaVE) database, the most comprehensive database of PVs (13) (Figure 1E). This step mimics the L1 taxonomic tool of the PaVE database (L1 Taxonomic tool, 1). The PaVE database provides full papillomavirus genome sequences with complete taxonomic classification (referenced PV), as well as full genomes with incomplete taxonomic classification

(unreferenced PV). Referenced genomes correspond to genomes validated and fully characterized by the re-sequencing of the entire genome. Unreferenced genomes are mostly genomes identified through metagenomics approaches and submitted to PaVE but not validated for accuracy or novelty of the PV (14).

2.6.3 RaxML-EPA-based taxonomical classification

A reference phylogenetic tree (reference tree; RT) was constructed based on the full-L1 ORF nucleotide sequences of 597 available PV genomes retrieved from the PaVE database (https://pave.niaid.nih.gov/) in June 2019 (13). The sequences were aligned at the nucleotide level using the MUSCLE algorithm, with the default parameters (15) in MEGA7 (16). The final full-length L1-ORF alignment encompassing 597 full L1-ORF nucleotide sequences, 2913 positions, and 468 distinct alignment patterns constitutes the reference multiple sequence alignment (MSA). MEGA7 was used to test the best substitution model and for the phylogenetic inference. The codon positions included were 1st + 2nd + 3rd + non-coding. Based on the alignment using MUSCLE, all positions with <95% site coverage were eliminated (partial deletions), to enable the inclusion of taxa with some missing data. There were a total of 1383 positions in the final dataset.

A discrete gamma distribution was used to model evolutionary rate differences among sites (five categories; +G, parameter = 0.658). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 0.019% sites). The initial trees for the heuristic search were obtained automatically by applying the neighbor-joining (NJ)/BioNJ algorithm to a matrix of pairwise distances estimated using the maximum composite likelihood (MCL) approach and then by selecting the topology with the highest log likelihood value (-468961.607). The final tree selected constitutes the RT.

Phylogenetic inference was performed with MEGA7 using the general time-reversible (GTR) model of nucleotide substitution and 500 bootstrap replicates (17).

The Parsimony-based Phylogeny-Aware Read alignment (PaPaRa) program (18) algorithm is used to align each contig sequence, reconstructed during the previous *de novo* assembly step, against the MSA (19) (Figure 1E). Subsequently, the evolutionary placement algorithm (EPA) (20) in RaxML (11) is run to place the sequences into the RT (Figure 1E), based on PaPaRa multiple alignment. The EPA is run using the same nucleotide substitution model used to infer the reference phylogenetic tree. A script was developed in-house to parse the output format of the EPA (21) to extract, for each reconstructed sequence, its closest related taxon in the phylogenetic tree, and use this taxon to assign a taxonomic classification.

2.7 Output reports

Several output reports are generated as Excel files, fasta files, or graphical images from the different steps of the workflow. They describe summary sequencing statistics, the sequences of known or putative new PVs, the relative unnormalized abundance of PV types, and the taxonomic classification of all identified PV sequences. The use of an info file providing sample characteristics enables the output of statistics stratified by these characteristics (Figure 1F, Supplementary Table 1). The detailed list of files and file contents is available in Supplementary Data 3.

2.8 Performance testing

The performance of the bioinformatics workflow has been estimated on a computer with an Intel® Core™ i7-6700 processor CPU @ 3.40GHz × 8, 64 bits, 62.9 GB RAM, 256 GB SSD, in the Linux environment (Ubuntu 16.04 LTS).

Supplementary Data 3: Description of output files format

The output files generated by PVAmpliconFinder are:

- an Excel file named “Table_Summary_MegaBlast” (Supplementary Table 5) that contains several tables providing sequencing metrics by sample, primer set, or tissue type, and a classification of putative new and known PVs found in the samples, based on MegaBlast, RaxML, and BlastN results.

- several Excel file(s) containing a full taxonomic classification of the species present in the samples based on MegaBlast, BlastN, or RaxML-EPA results, with unnormalized relative abundance estimated as number of reads (Supplementary Tables 2, 3, and 4, respectively). Several tables are created if several tissue types have been specified in the info file, for the three methodologies applied to classify the PV sequences, and the information about primer used to detect the species is present (if also specified in the info file).

- a KRONA (22) graphical representation of the unnormalized abundance of PV genera and species, taxonomically classified based on MegaBlast, BlastN, and RaxML-EPA results, in terms of number of reads (Figure 2, 3 and 4, respectively). If an info file was provided as input, a graphical representation is produced for each tissue type, as well as an overall representation mixing the different tissue types.

- an Excel file named “Table_putative_known_PV”, containing the putative known PV sequences detected in the different samples (Supplementary Table 6). This file contain information such as: a unique identification for the sequence or the cluster of sequences corresponding to a putative known PV; the percentage of dissimilarity on the aligned portion of the sequence returned by MegaBlast (if several sequences in the cluster, the percentage of dissimilarity of the longest sequence is reported); the relative unnormalized abundance of the sequence(s) into the overall reads generated for the sample; the absolute number of reads used to generate the sequence(s); the GI number from MegaBlast rent; the closest PV species given by the BlastN against the PaVE database; the taxonomic classification at the genus level given by the BlastN against the PaVE database; the closest PV species given by the RaxML-EPA algorithm; the taxonomic classification at the genus level given by the RaxML-EPA algorithm; and the nucleotide sequence(s).

- an Excel file named “Table_putative_new_PV”, containing the putative new PV sequences detected in the different samples (Table 3). This file contains the same information as the “Table_putative_known_PV” Excel file described above.

- fasta files of the putative known and putative new PVs, named “Putative_known_PV.fa” and “Putative_new_PV.fa”, respectively. If a cluster contains several sequence, the sequences are attributed a unique incremental number after the unique name of the cluster (e.g. if there are 3 sequences in the cluster named “PV_1”: >PV_1.1, >PV_1.2, >PV_1.3).

Supplementary Data 4: Sample collection, preparation, and sequencing

Skin swab specimens (n = 25) were randomly selected baseline samples from the VIRUSCAN Study, an ongoing five-year (2014–2019) prospective cohort study conducted at Moffitt Cancer Center and the University of South Florida (R01CA177586-01; “Prospective study of cutaneous viral infections and non-melanoma skin cancer”).
In addition, oral rinses (n = 22) were randomly selected from a pilot study that aimed to estimate the prevalence of *Helicobacter pylori* in oral gargles from a Latvian population. The study was approved (No. 8-A/15) by the Ethics Committee of Riga East University Hospital Support Foundation.
After DNA extraction, all samples were analyzed at the International Agency for Research on Cancer (Lyon, France). The PCR protocols use different sets of primers as described in (23). The use of these primers enables the amplification of a region in the L1 gene of approximately 450 bp. Each NGS pool included approximately 5 different samples generated from different PCR protocols.
Libraries were prepared using the NEBNext Ultra DNA library prep kit and MiSeq reagent kit version 2 (Illumina). Paired-end NGS sequencing was performed using an Illumina MiSeq (600 cycles), and final mean read size was 227 bp. The SRA accession number of the data is PRJNA555194.

References

1. Chouhy,D., Gorosito,M., Sánchez,A., Serra,E.C., Bergero,A., Fernandez Bussy,R. and Giri,A.A. (2010) New generic primer system targeting mucosal/genital and cutaneous human papillomaviruses leads to the characterization of HPV 115, a novel Beta-papillomavirus species 3. *Virology*, **397**, 205–216.

2. Forslund,O., Antonsson,A., Nordin,P., Stenquist,B. and Göran Hansson,B. (1999) A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. *Journal of General Virology*, **80**, 2437–2443.

3. Forslund,O., Ly,H. and Higgins,G. (2003) Improved detection of cutaneous human papillomavirus DNA by single tube nested 'hanging droplet' PCR. *Journal of Virological Methods*, **110**, 129–136.

4. Andrews,S. (2010) FastQC: a quality control tool for high throughput sequence data.

5. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.

6. Krueger,F. (2015) Trim Galore! : A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries.

7. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahé,F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.

8. Bernard,H.-U., Burk,R.D., Chen,Z., van Doorslaer,K., Hausen,H. zur and de Villiers,E.-M. (2010) Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology*, **401**, 70–79.

9. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

10. Mahmoudabadi,G. and Phillips,R. (2018) A comprehensive and quantitative exploration of thousands of viral genomes. *eLife*, **7**, e31955.

11. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

12. Huang,X. and Madan,A. (1999) CAP3: A DNA Sequence Assembly Program. *Genome Res*, **9**, 868–877.

13. Van Doorslaer,K., Tan,Q., Xirasagar,S., Bandaru,S., Gopalan,V., Mohamoud,Y., Huyen,Y. and McBride,A.A. (2013) The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res*, **41**, D571–D578.

14. Simmonds,P., Adams,M.J., Benkő,M., Breitbart,M., Brister,J.R., Carstens,E.B., Davison,A.J., Delwart,E., Gorbalenya,A.E., Harrach,B., *et al.* (2017) Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, **15**, 161–168.

15. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–1797.

16. Kumar,S., Stecher,G. and Tamura,K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*, **33**, 1870–1874.

17. Nei,M. and Kumar,S. (2000) Molecular Evolution and Phylogenetics Oxford University Press.

18. Berger,S.A. and Stamatakis,A. PaPaRa 2.0: A Vectorized Algorithm for Probabilistic Phylogeny-Aware Alignment Extension.

19. Berger,S.A. and Stamatakis,A. (2011) Aligning short reads to reference alignments and trees. *Bioinformatics*, **27**, 2068–2075.

20. Berger,S.A., Krompass,D. and Stamatakis,A. (2011) Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Syst Biol*, **60**, 291–302.

21. Matsen,F.A., Hoffman,N.G., Gallagher,A. and Stamatakis,A. (2012) A Format for Phylogenetic Placements. *PLOS ONE*, **7**, e31009.

22. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.

23. Brancaccio,R.N., Robitaille,A., Dutta,S., Cuenin,C., Santare,D., Skenders,G., Leja,M., Fischer,N., Giuliano,A.R., Rollison,D.E., *et al.* (2018) Generation of a novel next-generation sequencing-based method for the isolation of new human papillomavirus types. *Virology*, **520**, 1–10., D571–D578.

Detection of human papillomaviruses in paired healthy skin and actinic keratosis by next generation sequencing

[1]Luisa Galati, [1]Rosario Brancaccio, [1]Alexis Robitaille, [1]Cyrille Cuenin, [2]Fabiola Luzi, [3,4]Gianna Fiorucci, [3]Maria Vincenza Chiantore, [5]Nadia Marascio, [5]Giovanni Matera, [5]Maria Carla Liberto, [6]Maria Gabriella Donà, [3]Paola Di Bonito, [1]Tarik Gheit, [1*]Massimo Tommasino

[1]International Agency for Research on Cancer-World Health Organization, Lyon, France

[2]Plastic and Reconstructive Surgery, San Gallicano Dermatologic Institute IRCCS, Rome, Italy

[3]Department of Infectious Diseases, EVOR unit, Istituto Superiore di Sanità, Rome, Italy

[4] Institute of Molecular Biology and Pathology, Consiglio Nazionale delle Ricerche, Rome, Italy

[5]"Magna Graecia" University, Catanzaro, Italy

[6]STI/HIV Unit, San Gallicano Dermatologic Institute IRCCS, Rome, Italy

*Correspondence: Infections and Cancer Biology Group, International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France. Tel. +33-4-72738190, E-mail: icb@iarc.fr

ABSTRACT

Actinic keratosis (AK) arises on photo-damaged skin and is considered to be the precursor lesion of cutaneous squamous cell carcinoma (cSCC). Many findings support the involvement of β human papillomaviruses (HPVs) in cSCC, while very little is known on γ HPV types. The objective of this study was to characterize the spectrum of PV types in healthy skin (HS) and AK samples of the same immunocompetent individuals using next generation sequencing (NGS). Viral DNA of 244 AK and 242 HS specimens were amplified by PCR using two different sets of primers (FAP59/64 and FAPM1). Purified amplicons were pooled and sequenced using NGS. The study resulted in the identification of a large number of known β and γ PV types. In addition, 27 putative novel β and 16 γ and 4 unclassified PVswere isolated. Only HPV types of species γ-1 (e.g. HPV4) appeared to be strongly enriched in AK versus HS. The NGS analysis revealed that a large spectrum of known and novel PVs is present in HS and AK. The evidence that species γ-1 HPV types appears to be enriched in AK in comparison to HS warrants further biological and epidemiological studies to evaluate their role in development of skin (pre)cancerous lesions.

INTRODUCTION

Cutaneous squamous cell carcinoma (cSCC) arises from progression of the precursor lesion, actinic keratosis (AK), which develops on photo-damaged skin (Hasche et al., 2018). Ultraviolet (UV) radiation exposure is the main risk factor in the development of AK and cSCC (Brash et al.,1991) (Werner et al., 2015). Skin lesion development is also positively associated with fair skin, advanced age and immunosuppression (Didona et al., 2018). The concept that impairment of the immune system favors cSCC development supports the involvement of an infectious agent, such as the epitheliotropic human papillomaviruses (HPVs). HPVs are circular double-stranded DNA viruses infecting mucosal and cutaneous epithelia. To date, more than 200 HPV genotypes have been fully characterized and classified into five genera (α, β, γ, mu and nu papillomaviruses) according to the nucleotide sequences of the ORF encoding for the major capsid protein L1 (Bzhalava et al., 2015) (https://pave.niaid.nih.gov/). A subgroup of α-genus HPV types, referred to as mucosal high-risk (HR) HPV types, has been clearly associated with human carcinogenesis (Egawa et al., 2015; Schiffman et al., 2016). Twelve HR HPV types, namely 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 have been classified as Group 1, carcinogens to humans, by the International Agency for Research on Cancer (IARC) (Bouvard et al., 2009). In addition to the mucosal HR HPV types, epidemiological and biological studies support the role of β-genus HPV types in the development of cSCC, together with UV radiation (Wang and Roden, 2013). The first β HPV types, HPV5 and 8, were identified in skin lesions of *epidermodysplasia verruciformis* (EV) patients, who are highly susceptible to β HPV infection and UV-induced cSCC (Orth, 1987; Patel et al. 2010). Accordingly, IARC has classified β HPV 5 and HPV 8 as "possibly carcinogenic" agents (Group 2B) in EV patients (Bouvard et al., 2009). Since their isolation, additional 52 β HPV types have been characterized so far, which are subdivided into 5 species, β1-5 (http://www.nordicehealth.se/hpvcenter/reference_clones/), and are abundantly present on the skin of healthy individuals (Antonsson et al., 2003a; Antonsson et al., 2003b; Foulongne et al., 2012; Wylie et al., 2014). In addition to EV patients, β HPV types appear to be involved in cSCC development also in immunocompromised individuals, such as organ transplant recipients (OTR), as well as in elderly general population (Nunes et al., 2018; Harwood et al., 2017; Quint et al., 2015; Bouwes Bavinck et al., 2018). In contrast to α HR HPV types, the presence of β HPVs does not appear to be required for the maintenance of the malignant phenotype (Viarisio et al., 2018). Studies in *in vivo* experimental models provide

evidence for a "hit-and-run" mechanism of β HPVs involvement in UV-induced skin carcinogenesis (Viarisio et al., 2018; Viarisio et al., 2017; Hasche et al., 2018; Viarisio et al., 2011). Accordingly β HPV prevalence and viral load decrease during carcinogenesis process in humans, being significantly higher in AK than in cSCC (Weissenborn et al. 2005; Rollison et al., 2019). Other cutaneous HPV types that are frequently detected in skin are the ones that belong to γ genus. They represent the largest clade within the *Papillomaviridae* family. Almost 100 γ HPV types subdivided into 27 species have been fully characterized so far. No clear association of γ HPVs with malignant lesions has been demonstrated, although biological studies showed that E6 and E7 proteins from some γ HPVs display *in vitro* transforming activities (Grace and Munger, 2017).

To gain new insights on the presence of a broad spectrum of β and γ HPV types in healthy skin (HS) and actinic keratosis (AK) of the same individual, we used different PCR protocols (Forslund et al., 1999; Forslund et al., 2003; Brancaccio et al., 2018) combined with Next Generation Sequencing (NGS). The results revealed the presence of a large spectrum of β and γ HPV types. Interestingly, species γ-1 HPV types appear to be more represented in AK than in HS.

RESULTS

Amplification of HPV DNA by two PCR protocols

Skin scrapings from 244 patients, for a total of 244 AK and 242 HS samples of the corresponding patients were processed for total DNA extraction, and subsequent PCR amplifications using FAP59/FAP64 and an improved version (FAPM1) of the original FAP primers, targeting part of L1 ORF (Forslund et al. 1999) (Forslund et al. 2003) (Bolatti et al. 2018). HPV DNA was detected in 75.2% (182/242) and in 85.1% (206/242) of HS samples, using FAP59/FAP64 and FAPM1 protocols, respectively. A PCR product of the expected size was detected in 71.2% (175/244) and in 50% (122/244) of the AK samples by using the primer sets FAP59/FAP64 and the novel FAPM1, respectively.

NGS data and known PVs sequence analyses

PCR amplicons generated by the use of the two different sets of primers on HS and AK DNA samples were pooled as shown in method section and sequenced using the NGS platform MiSeq Illumina. The NGS analysis generated a total of 1,209,249 reads. A total of 1,208,356 of the reads were considered for further analysis after quality trimming, and chimeric PCR sequence removal. All of them, were identified as related to PVs sequences (>99% of reads). Each read was matched against the National Center for Biotechnology Information (NCBI) sequences database by means of BLAST algorithm and assigned to its closest PV types.

The different PV sequences were analyzed following the official taxonomic HPV classification based on the similarity in L1 ORF (Bzhalava et al., 2015).

Data analysis obtained using RAxML-EPA, a method that offers an accurate classification of short PV fragments, reported that the 1,208,356 reads analyzed comprised 1,204,447 (99.7%) reads from known PVs (≥90% of identity with L1 ORF of any known PV). The majority of the reads (81.1%, 976,693 reads) corresponded to β PVs, followed by γ (17.3%, 208,932 reads) and α types (0.01%, 121 reads) (**Table 1**). According to RaxML-EPA analysis of known PV sequences, the major number of reads were related to human PVs (n= 1,181,306), while the remaining were closely related to non-human PVs (total non-human reads: 23,141) i.e. *Macaca fascicularis PV type 2* (MfPV2) belonging to β-6 genus (3,769 reads), *Macaca mulatta papillomavirus type 5* (MmPV5) (671 reads) that is classified into the γ genus, and *Erethizon dorsatum papillomavirus 2* (EdPV2) (18,701 reads), a new PV still unclassified (**Table S1**). In summary, 1,204,447 reads are representative of 1786 PVs sequences. As a specific PV sequence can be represented more than one time among the different pools, or different PV sequences can be assigned to the same PV type, thus 1786 PV sequences corresponded to 195 distinct PV types (**Table 1 and Table S1**). Of the 195 PV types, 93 resulted to be officially recognized, namely 2 sequences from α-2 species, 49 sequences from β 1-6 and 42 sequences spreading into 18 γ species (**Figure 1**). The remaining sequences corresponded to 12 unclassified-β and 89 unclassified-γ PVs. Only one sequence remained unclassified and was assigned by RAxML-EPA analysis to a divergent and unclassified EdPV2 sequence (**Tables 1 and S1**).

Known PV sequences in HS and AK

We next compared the distribution of the different PVs sequences in AK and HS. The distribution of all known HPV types detected in HS and AK is shown in **Table S1 and Figure 1.** Regarding α HPV types, the small number of reads (n=121) generated exclusively by the FAPM1 protocol corresponded to sequences of the two closely related cutaneous HPVs 3 and 28. However, most of the reads were from HPV28, which was equally distributed in HS and AK (**Table S1**). Reads of β HPV sequences were approximately equally represented in HS and AK (485,918 and 490,775 respectively), with the exception of β-4 species, represented by HPV type 92 only. For this the number of reads was more abundant in AK than HS (1440 vs 198 reads) (**Figure 1A, Table S1**).

Regarding the γ HPV types, reads for the different species were differently detected in HS (85,568 reads) and AK (123,364 reads) samples, being in some cases higher in AK than HS (i.e. γ-1, γ-3, γ-7, γ-8, γ-9, γ-11, γ-15, γ-17) and vice versa in other cases (i.e. γ-12, γ-13, γ-21, γ-24) (**Figure1B**). Moreover, for the majority of the species only a small number of reads were detected in HS and AK (i.e γ -4, γ-20 and γ -25). Interestingly, for γ-1 species 600 fold difference in number of reads was observed in AK versus HS (13,248 and 22 reads, respectively) (**Figure 1B and Table S1**). The majority of these γ-1 reads corresponded to HPV4 (13,207 reads) (**Table S1**).

Putative novel PVs

Finally, 3,909 (0.3%) reads generated a total of 47 putative novel PV types, since the fragment sequence showed less than 90% similarity to L1 ORF of any known PVs. As per the RaxML-EPA classification, the majority of reads were closely related to human PVs (3,827 reads). Of the 3,827 reads, a substantial number of reads were closely related to β-HPVs (3,457 reads), and γ-HPVs (370 reads). Whereas, for the non-human PVs, out of 82 reads, 74 reads were from the unclassified PVs category (**Table S2**).

Among the unknown PV sequences, 26 (55.3%) putative novel sequences were found in HS and 21 (44.7%) in AK specimens, respectively (**Tables 1 and S2**). The FAPM1 primers detected a slightly higher number of putative novel HPV sequences than FAP59/64, i.e. 27 and 20, respectively (**Table 1**).

Using RAxML-EPA classification, 15 putative novel β PVs and 11 putative novel γ PVs were isolated from HS samples, whereas 12 novel β PVs and 5 γ PVs were isolated from AK samples.

The remaining 4 putative novel PVs, isolated from AK samples, remained unclassified (**Table 1**).

The FAPM1 protocol allowed the isolation of a total of 12 novel γ PVs in AK and HS samples, while the FAP protocol allowed the isolation of 4 γ PVs only in HS samples (**Tables 1 and S2**).

Putative new PV types in AK samples were related to HPV5, 21 (belonging to species β-1), HPV15, 22, 23, 120 (species β-2) and HPV130 (species γ-10) (**Table S2**). In HS samples, the new PV sequences were related to species β-1 (HPV5, 21, 24), β-2 (HPV22, 23, 38), γ-10 (HPV133) and γ-27 (HPV201).

DISCUSSION

Cutaneous HPV types spread over all five HPV genera and are abundantly present in normal skin. Since several lines of evidence support the role of β HPV types in favoring the UV-induced skin carcinogenesis, epidemiological studies focused mostly only on β-HPV detection in pre-malignant and malignant skin lesions. In contrast to β HPV types, the biology and epidemiology of γ HPV types have been poorly investigated so far. In a recent study, we have determined the prevalence of 46 β and 52 γ HPV types in HS and AK of the same individuals who have been included in this study (Donà et al., 2019). Dona' et al. reported that the prevalence of most of the β and γ HPV types decreased from HS to AK, suggesting that cutaneous HPVs may play a role at early phase of AK lesion development and can be lost once the lesion is fully established (Donà et al., 2019). To have a more accurate scenario on the HPV types present in HS and AK, we have re-analyzed the same cohort performing a broad spectrum analysis of cutaneous HPV types by NGS. Our data confirmed previous findings that β1 and β2 are the most represented species in both HS and AK, followed by β3, β4 and β5 (Hampras et al., 2017; Donà et al. 2019). It is not yet clear why the β4 and β5 HPV types are poorly present in the skin. One possible hypothesis is that these HPVs have a low efficiency in persisting in the host skin. Alternatively, β3-5 HPV types may preferentially infect other anatomical sites than the skin. In support of this hypothesis, it has been shown that β3 are more prevalent in mucosal epithelia than in the skin (Forslund et al., 2013; Hampras et al., 2017). In agreement with the epidemiological data, functional studies revealed that β3 HPV types 49 and 76 share some biological properties with the mucosal HR HPV16 in vitro and in vivo experimental models (Cornet et al., 2012; Viarisio et al., 2016; White et al. 2014).

The γ genus is the largest clade within the Papillomaviridae family and the improvement of sequencing methods has led to the identification of many novel γ types over the last years (Dutta et al., 2017; Bolatti et al., 2018; Pastrana et al. 2018). The γ PVs can be found in common warts, in skin tumors and AK samples, as well as in normal skin (Ekström et al., 2011; Ekström et al., 2013; Hošnjak et al., 2015; Donà et al., 2019) .

In our study, the comparison of HS and AK samples collected from the same individuals (n=244) revealed that the number of NGS reads for γ PVs were differently represented in HS (85,568 reads) vs AK (123,364 reads) samples. These results are consistent with prevalence studies that reported a high value of γ PVs in AK (Bolatti et al., 2018). Our NGS-based analysis revealed that almost all γ species were represented in HS and AK, except for γ-2, γ-5, γ-6, γ-14, γ-18 and γ-23. In addition to this, a relevant number of γ species that are not yet classified by the HPV reference center was also found.

Interestingly, although most of the β and γ HPV types were equally represented in HS and AK samples, γ-1 HPV4 was strongly enriched in AK samples versus HS. Similar results were observed in our recent study where HPV detection was performed by a highly specific genotyping assay (Donà et al., 2019). In this study, using the same samples, the number of reads that correspond to HPV4 was indeed higher in AK (13,207 reads) in comparison to HS (20 reads). These findings suggest a possible link between HPV4 infection and AK development. Alternatively, this specific γ HPV type might have some biological differences with respect to the other γ HPV types, for instance it could benefit from the tissue alterations occurring in AK for completion of its life cycle. Additional work is required to further evaluate these two hypothesis. So far, it has been reported that HPV4 is associated with the development of mosaic warts (Cubie, 2013; Doorbar et al. 2015). Regarding HPV4 biological properties, it has been shown that its E7 is able to degrade pRb (Wang et al., 2010), as the mucosal HR HPV E7s.

In the present study we identify 195 known HPV types and in addition to this, using different PCR protocols combining with NGS, we identified 47 putative novel PVs. The analysis of these putative novel PVs revealed that they are related to 27 β, 16 γ and 4 unclassified PVs. Of which, 1 β PV, 1 γ and 4 unclassified PVs were non-human PVs. The presence of non-human PV sequences in human skin may be explained by environmental contamination, or alternately it may result from human viruses closely related to animal PVs. Interestingly, our study led to the

identification of 15 putative novel β-2 HPV types phylogenetically related to HPV38, which displays *in vitro* and *in vivo* transforming properties. Also HPv38 has been found significantly associated with the risk of cSCC in a recent meta-analysis (Chahoud et al., 2016).

In summary, using a robust strategy based on the use of specific or degenerate primers and NGS technology this study expanded our knowledge and efficiently depicted the PV population in AK and HS sample. Moreover, it allowed the detection of putative novel PVs, although the identification of novel PV types or species can only be definitively confirmed by sequencing the whole L1 ORF. Finally, it showed that some γ HPV types (e.g., HPV4) are enriched in AK vs. HS, and might thus play a role in skin carcinogenesis, thus deserving further *in vivo* and *in vitro* investigations.

MATERIALS AND METHODS

Patient selection, sample collection and DNA extraction

Skin scraping samples (HS and AK) from a previous study aimed to determine the prevalence of cutaneous HPVs in AK lesions by using a sensitive Luminex based-beads multiplex assay were used in the present analysis (Donà et al., 2019). Skin samples were collected from 244 immunocompetent patients (142 men and 102 women in age range 48-94 years) with a diagnosis of AK attending the dermatology outpatient clinic of the National Institute for Health, Migration and Poverty (NIHMP) in Rome (Italy). A total of 488 individual samples were collected by scraping the lesions and, separately, the healthy skin of the glabellar region with a sterile spatula. The majority of the AK lesions were in the head region (n=221) while others were located in the limbs (n=5) and other anatomical sites (n=18). In the present analysis, two HS samples were excluded due to the shortage of the residual sample. Samples were stored at -80°C until treatment with proteinase K for 4h at 50°C in 10 mM Tris-HCl pH 8.0, 50 mM NaCl, 5 mM EDTA, 1 mM DTT, 0.5% SDS (0.4 ml/sample). Nucleic acids, extracted by magnetic silica using the automated system NucliSENS EasyMag (Biomérieux, France) according to the manufacturer's directions, were analyzed at IARC (Lyon, France) by NGS. Written informed consent was obtained from all enrolled patients. The study was approved by the Ethical Commettes of both NIHMP (2014) and San Gallicano Dermatologic Institute (CE943/17).

PCR amplification and amplicon purification

Extracted DNA was amplified using two different sets of primers; the consensus primer pair FAP (FAP59\FAP64) targeting the 5'end of the L1 ORF as previously reported (Forslund et al. 1999), and a new set of degenerated FAP primers (FAPM1 primer mix) as previously described by Brancaccio et al. (Brancaccio et al., 2018). Both FAP and FAPM1 primers target a region of the L1 ORF yielding an amplicon of about 480 bp. PCR amplicons were visualized by electrophoresis on a 2% agarose gel and purified using QIAquick gel extraction purification kit according to the manufacturer's instructions (QIAGEN, Hilden, Germany).

Library preparation and NGS

Purified PCR amplicons were divided into twelve different pools as described in **Table 2**. Each pool was obtained using 2 µl of each purified PCR product. Before library preparation, one additional purification step was performed in each pool to remove any residual contaminants using the Agencourt AMPure XP PCR purification kit with a beads ratio of 1.8 X (Beckman Coulter) according to the manufacturer's instructions.

Twelve libraries were prepared using the Nextera ^{TM} DNA Flex Library preparation kit (Illumina, San Diego, CA, US). Illumina MiSeq dual-indexed adapters (Illumina, San Diego, CA, US) were added to each of the PCR pools. The library sizes were checked using the Bioanalyzer 2100 Expert (Agilent) using high sensitivity DNA assay. NGS analysis was performed on 4 nM of DNA pooled library using an Illumina MiSeq instrument (2 X 150 paired-end reads with the Illumina MiSeq kit v3). In order to enrich the diversity of the libraries, 10% of PhiX (Illumina, San Diego, CA, US) was added to the NGS reaction.

Bioinformatic analysis of NGS sequences

The bioinformatic workflow includes common data preprocessing steps for quality control and filtering. Then, data complexity is reduced before the identification of the PV-related sequences. Groups of sequences are defined based on similarity between identified sequences and available PVs sequences in the NCBI database. De-novo assembly is then performed to reconstruct the full amplified region covered by several primers systems. Finally, the reconstructed sequences are taxonomically classified based on two independent methodologies: alignment-based, and

homology-based, respectively, before generation of diverse output reports. Details of the bioinformatic pipeline named "PVAmpliconFinder" and parameters used can be found in (Robitaille A. et al., 2019), and the code of the tool is freely available at (https://github.com/IARCbioinfo/PVAmpliconFinder).

All the results in this study are based on the identification of the sequences following the homology-based classification using the EPA in RAxML (Stamatakis, 2014; Berger and Stamatakis, 2011) (henceforth referred to as RAxML-EPA). Only the longest sequence was considered for RAxML-EPA classification when several singlets or contigs were available.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We are grateful to all members of our laboratories for their cooperation, Nicole Suty for her help with preparation, and Dr Karen Müller for editing this manuscript.

The study was supported by Fondation ARC pour la recherche sur le cancer (no. PJA 20151203192) (https://www.fondation-arc.org/espace-chercheur) and the Institut National de la Santé et de la Recherche Médicale (no. ENV201610) (https://www.eva2.inserm.fr/EVA/jsp/AppelsOffres/CANCER/) to MT.

The authors alone are responsible for the views expressed in this article, and they do not necessarily represent the views, decisions, or policies of the institutions with which they are affiliated.

REFERENCES

Antonsson A, Erfurt C, Hazard K, Holmgren V, Simon M, Kataoka A, et al. Prevalence and type spectrum of human papillomaviruses in healthy skin samples collected in three continents. J. Gen. Virol. 2003a;84(7):1881–6

Antonsson A, Karanfilovska S, Lindqvist P, Hansson B. General Acquisition of Human Papillomavirus Infections of Skin Occurs in Early Infancy. J. Clin. Microbiol. 2003b;41(6):2509–14

Bolatti EM, Hošnjak L, Chouhy D, Re-Louhau MF, Casal PE, Bottai H, et al. High prevalence of Gammapapillomaviruses (Gamma-PVs) in pre-malignant cutaneous lesions of

immunocompetent individuals using a new broad-spectrum primer system, and identification of HPV210, a novel Gamma-PV type. Virology. 2018;525(June):182–91
Bouvard V, Baan R, Straif K, Grosse Y, Secretan B, Ghissassi F El, et al. A review of human carcinogens—Part B: biological agents. Lancet Oncol. 2009;10(4):321–2
Bouwes Bavinck JN, Feltkamp MCW, Green AC, Fiocco M, Euvrard S, Harwood CA, et al. Human papillomavirus and posttransplantation cutaneous squamous cell carcinoma: A multicenter, prospective cohort study. Am. J. Transplant. 2018;18(5):1220–30
Brancaccio RN, Robitaille A, Dutta S, Cuenin C, Santare D, Skenders G, et al. Generation of a novel next-generation sequencing-based method for the isolation of new human papillomavirus types. Virology. Academic Press; 2018;520:1–10
Brash DE, Rudolph JA, Simon JA, Lin A, Mckenna GJ, Badent HP, et al. A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma (UV light/tumor suppressor genes). Genetics. 1991;88(November):10124–8
Bzhalava D, Eklund C, Dillner J. International standardization and classification of human papillomavirus types. Virology. Elsevier; 2015;476:341–4
Chahoud J, Semaan A, Chen Y, Cao M, Rieber AG, Rady P, et al. Association between β-genus human papillomavirus and cutaneous squamous cell carcinoma in immunocompetent individuals-a meta-analysis. JAMA Dermatology. 2016;152(12):1354–64
Cornet I, Bouvard V, Campo MS, Thomas M, Banks L, Gissmann L, et al. Comparative Analysis of Transforming Properties of E6 and E7 from Different Beta Human Papillomavirus Types. J. Virol. 2012;86(4):2366–70
Cubie HA. Diseases associated with human papillomavirus infection. Virology. United States; 2013;445(1–2):21–34
Didona D, Paolino G, Bottoni U, Cantisani C. Non Melanoma Skin Cancer Pathogenesis Overview. Biomedicines. 2018;6(1):6
Donà MG, Chiantore MV, Gheit T, Fiorucci G, Vescio MF, La Rosa G, et al. Comprehensive analysis of β- and γ-human papillomaviruses in actinic keratosis and apparently healthy skin of elderly patients. Br. J. Dermatol. England; 2019
Doorbar J, Egawa N, Griffin H, Kranjec C, Murakami I. Human papillomavirus molecular biology and disease association. Rev. Med. Virol. England; 2015;25 Suppl 1:2–23
Dutta S, Robitaille A, Olivier M, Rollison DE, Tommasino M, Gheit T. Genome Sequence of a Novel Human Gammapapillomavirus Isolated from Skin. Genome Announc. 2017;5(23):10–1
Egawa N, Egawa K, Griffin H, Doorbar J. Human papillomaviruses; Epithelial tropisms, and the development of neoplasia. Viruses. 2015;7(7):3863–90
Ekström J, Bzhalava D, Svenback D, Forslund O, Dillner J. High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions. Int. J. Cancer. 2011;129(11):2643–50
Ekström J, Mühr LSA, Bzhalava D, Söderlund-Strand A, Hultin E, Nordin P, et al. Diversity of human papillomaviruses in skin lesions. Virology. Academic Press; 2013;447(1–2):300–11
Forslund O, Antonsson A, Nordin P, Stenquist B, Hansson BG. A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. J. Gen. Virol. 1999;80(9):2437–43
Forslund O, Johansson H, Madsen KG, Kofoed K. The nasal mucosa contains a large spectrum of human papillomavirus types from the Betapapillomavirus and Gammapapillomavirus genera. J. Infect. Dis. 2013;208(8):1335–41
Forslund O, Ly H, Higgins G. Improved detection of cutaneous human papillomavirus DNA by single tube nested 'hanging droplet' PCR. J. Virol. Methods. 2003;110(2):129–36

Foulongne V, Sauvage V, Hebert C, Dereure O, Cheval J, Gouilh MA, et al. Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. PLoS One. 2012;7(6):e38499
Grace M, Munger K. Proteomic analysis of the gamma human papillomavirus type 197 E6 and E7 associated cellular proteins. Virology. United States; 2017;500:71–81
Hampras SS, Rollison DE, Giuliano AR, McKay-Chopin S, Minoni L, Sereday K, et al. Prevalence and Concordance of Cutaneous Beta Human Papillomavirus Infection at Mucosal and Cutaneous Sites. J. Infect. Dis. 2017;216(1):92–6
Harwood CA, Arron ST, Proby CM, Asgari MM, Bouwes Bavinck JN, Green AC, et al. Organ transplantation and cutaneous squamous cell carcinoma: progress, pitfalls and priorities in immunosuppression-associated keratinocyte carcinoma. Br. J. Dermatol. 2017;177(5):1150–1
Hasche D, Vinzón SE, Rösl F. Cutaneous papillomaviruses and non-melanoma skin cancer: Causal agents or innocent bystanders? Front. Microbiol. 2018;9:1–19
Hošnjak L, Kocjan BJ, Pirš B, Seme K, Poljak M. Characterization of two novel Gammapapillomaviruses, HPV179 and HPV184, isolated from common warts of a renal-transplant recipient. PLoS One. 2015;10(3)
Nunes E, Talpe-Nunes V, Sichero L. Epidemiology and biology of cutaneous human papillomavirus. Clinics. 2018;73(Suppl 1):1–9
Orth G. Epidermodysplasia Verruciformis. In: Salzman NP, Howley PM, editors. The Papovaviridae: The Papillomaviruses. Boston, MA: Springer US; 1987. p. 199–243
Pastrana D V., Peretti A, Welch NL, Borgogna C, Olivero C, Badolato R, et al. Metagenomic Discovery of 83 New Human Papillomavirus Types in Patients with Immunodeficiency. mSphere. 2018;3(6):1–14
Patel T, Morrison LK, Rady P, Tyring S. Epidermodysplasia verruciformis and susceptibility to HPV. Dis. Markers. 2010;29(3–4):199–206
Quint KD, Genders RE, De Koning MNC, Borgogna C, Gariglio M, Bavinck JNB, et al. Human Beta-papillomavirus infection and keratinocyte carcinomas. J. Pathol. 2015;235(2):342–54
Rollison DE, Viarisio D, Amorrortu RP, Gheit T, Tommasino M. An Emerging Issue in Oncogenic Virology: the Role of Beta Human Papillomavirus Types in the Development of Cutaneous Squamous Cell Carcinoma. Sullivan CS, editor. J. Virol. 2019;93(7)
Schiffman M, Doorbar J, Wentzensen N, de Sanjosé S, Fakhry C, Monk BJ, et al. Carcinogenic human papillomavirus infection. Nat. Rev. Dis. Prim. 2016;2:16086
Viarisio D, Gissmann L, Tommasino M. Human papillomaviruses and carcinogenesis: well-established and novel models. Curr. Opin. Virol. Elsevier; 2017;26:56–62
Viarisio D, Mueller-Decker K, Kloz U, Aengeneyndt B, Kopp-Schneider A, Gröne H-J, et al. E6 and E7 from Beta Hpv38 Cooperate with Ultraviolet Light in the Development of Actinic Keratosis-Like Lesions and Squamous Cell Carcinoma in Mice. Lambert P, editor. PLoS Pathog. 2011;7(7):e1002125
Viarisio D, Müller-Decker K, Accardi R, Robitaille A, Dürst M, Beer K, et al. Beta HPV38 oncoproteins act with a hit-and-run mechanism in ultraviolet radiation-induced skin carcinogenesis in mice. PLoS Pathog. 2018;14(1):1–20
Viarisio D, Müller-Decker K, Zanna P, Kloz U, Aengeneyndt B, Accardi R, et al. Novel β-HPV49 transgenic mouse model of upper digestive tract cancer. Cancer Res. 2016;76(14):4216–25
Wang JW, Roden RBS. L2, the minor capsid protein of papillomavirus. Virology. 2013;445(1–2):175–86

Wang J, Zhou D, Prabhu A, Schlegel R, Yuan H. The canine papillomavirus and gamma HPV E7 proteins use an alternative domain to bind and destabilize the retinoblastoma protein. PLoS Pathog. United States; 2010;6(9):e1001089
Weissenborn SJ, Nindl I, Purdie K, Harwood C, Proby C, Breuer J, et al. Human papillomavirus-DNA loads in actinic keratoses exceed those in non-melanoma skin cancers. J. Invest. Dermatol. United States; 2005;125(1):93–7
Werner RN, Stockfleth E, Connolly SM, Correia O, Erdmann R, Foley P, et al. Evidence- and consensus-based (S3) Guidelines for the Treatment of Actinic Keratosis - International League of Dermatological Societies in cooperation with the European Dermatology Forum - Short version. J. Eur. Acad. Dermatology Venereol. 2015;29(11):2069–79
White EA, Walther J, Javanbakht H, Howley PM. Genus beta human papillomavirus E6 proteins vary in their effects on the transactivation of p53 target genes. J. Virol. 2014;88(15):8201–12
Wylie KM, Mihindukulasuriya KA, Zhou Y, Sodergren E, Storch GA, Weinstock GM. Metagenomic analysis of double-stranded DNA viruses in healthy adults. BMC Med. 2014;12(1):1–10

Table 1. Known and putative novel PVs sequences in healthy skin (HS) and actinic keratosis (AK) samples. The number of sequences and corresponding reads are reported for alpha, beta, gamma and unclassified PVs, stratified according to the primer set, by RAxML-EPA taxonomic classification.

	KNOWN PVs					UNKNOWN PVs				
		HS		AK			HS		AK	
PV genus	Known PVs sequences N (reads)	FAP59/64 PV sequences (N reads)	FAPM1 PV sequences (N reads)	FAP59/64 PV sequences (N reads)	FAPM1 PV sequences (N reads)	Putative new PV sequences N (reads)	FAP59/64 Unique PV sequences (N reads)	FAPM1 Unique PV sequences (N reads)	FAP68/64 Unique PV sequences (N reads)	FAPM1 Unique PV sequences (N reads)
alpha	**2 (121)**	0 (0)	2 (54)	0 (0)	1 (67)	**0**	0 (0)	0 (0)	0 (0)	0 (0)
beta	**61 (976,693)**	54 (311,187)	60 (174,731)	54 (253,407)	57 (237,368)	**27 (3,459)**	9 (1878)	6 (206)	6 (675)	6 (700)
gamma	**131 (208,932)**	67 (46,262)	91 (39,306)	85 (61,128)	87 (62,236)	**16 (376)**	4 (153)	7 (117)	0 (0)	5 (106)
unclassified PV	**1 (18,701)**	1 (1,510)	1 (8,678)	1 (2,568)	1 (5,945)	**4 (74)**	0 (0)	0 (0)	1 (14)	3 (60)
Total	**195 (1,204,447)**	122 (358,959)	154 (222,769)	140 (317,103)	146 (305, 616)	**47 (3,909)**	13 (2,031)	13 (323)	7 (689)	14 (866)

Table S1. Known PV types in healthy skin (HS) and actinic keratosis (AK) samples according to RAxML-EPA taxonomic classification.

PV types (n=195) and corresponding NGS reads are reported for PV species stratified according to the PCR primer sets and skin specimen.

Table 2. Description of the NGS pools. All the PCR products (n=685) were stratified in 12 NGS pools according to the type of skin sample and PCR protocol applied. The paired samples both positive in HS and AK (1-3, 2-4, 5-7, 6-8) and the unpaired samples (9, 10, 11 and 12) are reported.

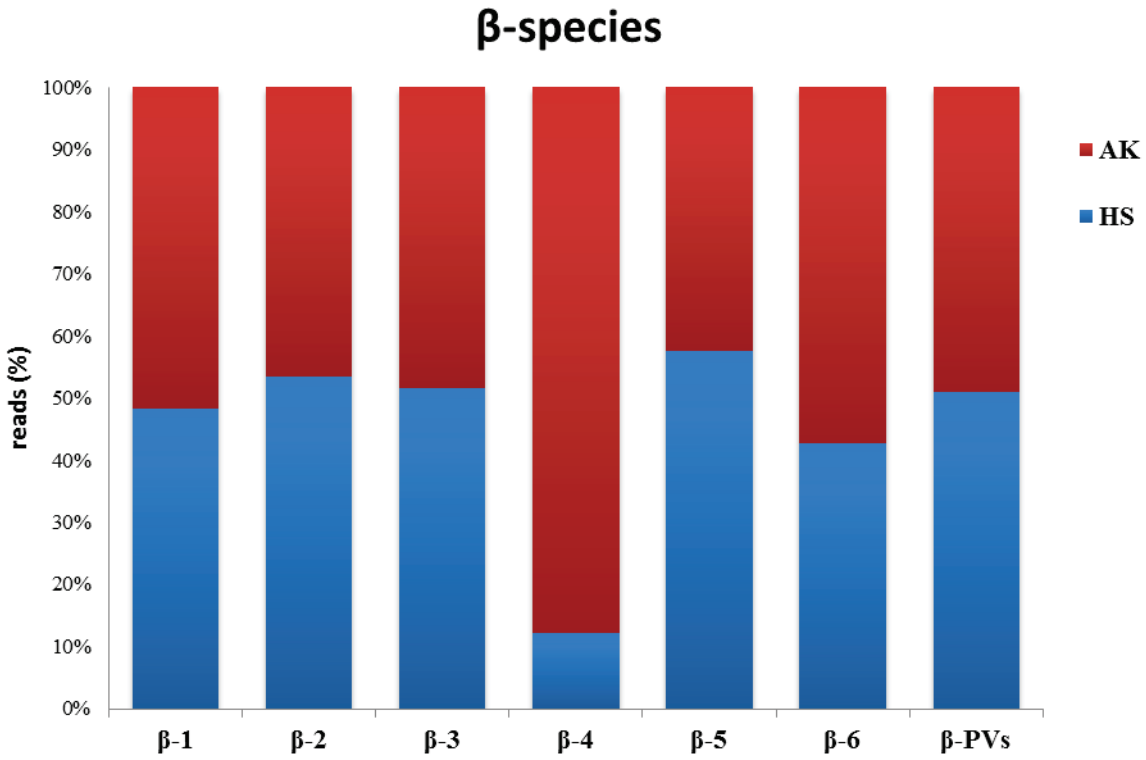
NGS pool	PCR protocol	Specimen (AK or HS)*	Total number	
1	FAP59/64	AK	71	
3	FAP59/64	HS	71	
2	FAP59/64	AK	70	
4	FAP59/64	HS	70	**Paired PV-positive**
5	FAPM1	AK	53	**samples**
7	FAPM1	HS	53	
6	FAPM1	AK	53	
8	FAPM1	HS	53	
9	FAP59/64	AK	34	
10	FAPM1	AK	16	**Unpaired PV-positive**
11	FAP59/64	HS	41	**samples**
12	FAPM1	HS	100	
*AK, actinic keratosis; HS, healthy skin			685	

Table S2. Putative new PVs in healthy skin (HS) and actinic keratosis (AK) samples according to RAxML-EPA taxonomic classification.
The 47 putative new PVs and corresponding NGS reads are listed according to the PCR primer sets and skin specimens.

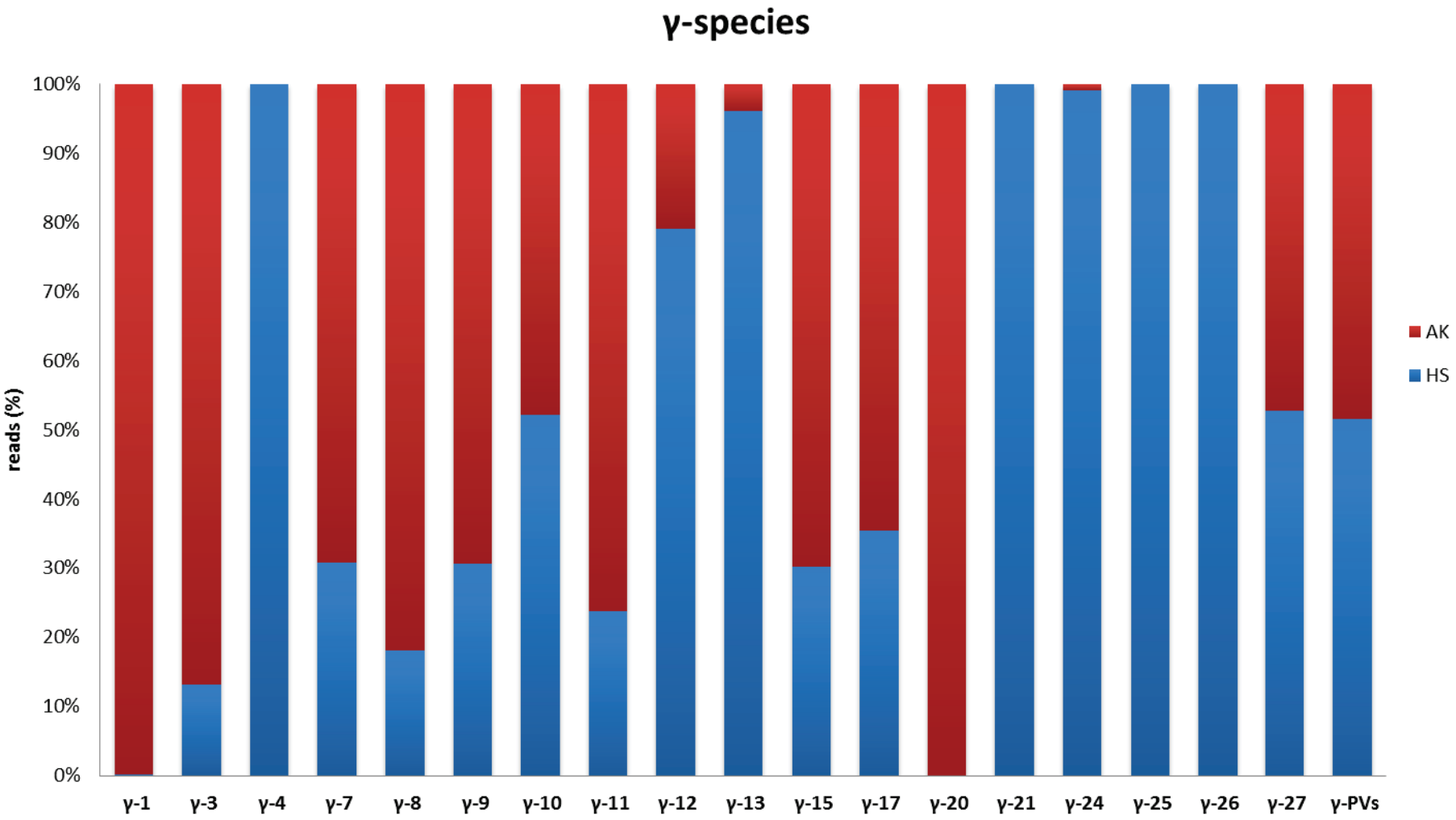
Figure Legend

Figure 1. Detected PV species in HS and AK samples. Proportions (%) of beta (1A) and gamma (1B) species according to RaxML-EPA classification in both healthy skin (HS) and actinic keratosis (AK) samples are shown.

1A


β-species
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
reads (%)
AK
HS
β-1
β-2
β-3
β-4
β-5
β-6
β-PVs

1B


γ-species
reads (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
γ-1
γ-3
γ-4
γ-7
γ-8
γ-9
γ-10
γ-11
γ-12
γ-13
γ-15
γ-17
γ-20
γ-21
γ-24
γ-25
γ-26
γ-27
γ-PVs
AK
HS

Discussion and conclusion

The discovery of novel HPV types remains of paramount importance, as new associations between HPV infections and human diseases may be established. In this thesis, we developed a novel strategy for the detection of new HPV types, in particular from the genus beta. This strategy combines the use of specific or degenerate primers targeting the L1 region of a broad spectrum of HPVs with NGS, and an automated bioinformatic workflow, PVAmpliconFinder. PVAmpliconFinder is an easy single-line command workflow that takes FastQ files as input files and generates tabular and graphical output files that describe the nature and abundance of PV-related sequences present in a complex mixture of host, phage, bacterial, and viral DNA. The data output discriminates between putative new and previously known *Papillomaviridae*-related sequences. Furthermore, it includes sequencing metrics and sequence details, enabling the design of subsequent laboratory experiments to confirm the *in silico* findings.

The beta genus

The growing interest in the beta genus arises from evidence that a number of beta HPV types may be involved in pre-malignant and malignant skin lesions (254). Moreover, the hypothesis has been raised that beta HPV infection is only required at the first stages of skin cancer initiation and is not required for cancer development and progression (the hit-and-run mechanism) (65). Thus, there is a clear need to have highly sensitive methods, allowing the detection of even minute amounts of sequences in a DNA sample in which the viral load is low. For this purpose, the novel strategy incorporates the selective enrichment of PV sequences before NGS is performed.

The choice of the bioinformatic strategy

Several choices were made during the design of the bioinformatic workflow. The first was to identify sequences in an unbiased manner, as the sequences are aligned against the entire “nt” NCBI database. Although this step is time consuming due to

the large size of the database, it reduces the false-positive discovery rate. Indeed, querying a database with reduced diversity (such as a virus database) using the e-value as a threshold could increase the chances of getting a hit even if the subject sequence has a low identity with the queried sequence. Considering only the sequences that have their best match against a *Papillomaviridae* family sequence produces an unbiased result. The second choice was not to combine similar sequences from the different samples. Indeed, a specific HPV can be identified in two independent samples during a single *in silico* analysis, with the exact same sequence. The choice to keep the results separated for each sample is due to the fact that the perfect identity between the sequences coming from the 2 independent samples may not be conserved once the full L1 is reconstructed. Thus, merging the results inter-sample would report only one of the sequences as the putative novel HPV type, but there could be two independent putative novel types. The third choice was to combine alignment methods with phylogeny-based methods. This double classification comes from the fact that pairwise alignment with a low percentage of similarity raises a concern about the pertinence of the results obtained. This is especially true for the putative new sequences having a greater than 15% dissimilarity against their best match. The maximum size of the amplicon is 450 nucleotides, thus à 15% dissimilarity corresponds to nearly 70 mismatches along the sequence. To circumvent this limitation, we used a complementary parallel approach based on a molecular evolution method, RaxML-EPA. Lastly, the fourth choice was the use of a greedy algorithm, CAP3, for the *de novo* reconstruction of sequence clusters. Due to the possibility to use several primer sets that have different hybridization positions along the L1 gene, it is important to reconstruct the longest possible sequence for each potential PV sequence. As the sequences are already clustered based on their best BLAST hit and their percentage of similarity to this hit, and because the input sequence length cannot exceed a few hundred bases, local optimal assembly is sufficient.

Limits of the bioinformatic methods

The main limit of the methods is that the taxonomic assignment performed must be interpreted with caution, because only small portions of putative new PV genomes have been obtained. In addition, the results obtained using the blastn algorithm refer

exclusively to the fraction of the sequence that is aligned by the algorithm. The percentage of similarity indicated by the initial MegaBlast results must also be interpreted with caution, because the definition of novelty for a PV is based on the full L1 ORF length. Thus, the alignment position for the MegaBlast and the BlastN are specified in the output files. The second limit is that evolutionary-based methods such as RaxML suffer from long-branch attraction. Long-branch attraction is an error where distant lineages are inferred to be close relatives because both have undergone a large number of changes. This is what is suspected of happening for the EPA classification of the *Erethizon dorsatum* papillomavirus (EdPV) sequences, for example, as EdPV2 is a novel PV genome presenting large differences from other known PVs on its L1 gene and is thought to represent a new genus in the family *Papillomaviridae*. This is also suspected of occurring when sequences are classified close to the *Sparus aurata* papillomavirus 1 (SaPV1), which has unique characteristics, such as an intron within the L1 gene (3). Although this led to an incomplete classification, these sequences may represent new species or virus features. The third limit is the absence of control for potential contamination. Cross-contamination between samples during library preparation, amplification and sequencing, or environmental contamination, are difficult to detect using *in silico* methods. Low-abundance sequences may truly be present in the samples but may also come from cross-contamination from another sample. PVAmpliconFinder will report sequences represented by only 2 reads. These low-abundance sequences should be considered with caution and defining an empirical abundance threshold could be considered. This environmental contamination may explain the presence of non-human PVs in human skin and oral samples. However, cross-species transmission of PVs between animals and humans may also be a consideration (24, 71). In addition, the notion of “non-human” PV genera needs to be interpreted with caution as they may also include some HPVs.

The issue of the taxonomic definition

PVAmpliconFinder presents the results based on the initial MegaBlast step and those obtained based on BlastN alignment against the PaVE database, but a huge number of sequences remain unclassified using the former approach because they match against incomplete L1 cds. The PaVE database includes PV sequences validated by

full genome resequencing, but also several “non-referenced” genomes that are not classified taxonomically. Currently, non-referenced PV genomes in the PaVE database account for more than 37% of the overall available PV genomes (244/649). The choice to include non-referenced genomes thus comes from their huge proportion. However, as this *in silico* reconstructed genome can produce chimeric sequences, comparison using the 10% dissimilarity threshold to this genome can lead to false assumptions if it appears that the non-reference genome is a chimera or is full of artifact bases occurring during the bioinformatics step. Questions regarding taxonomic classification of genomes reconstructed only from metagenomic experiments are currently being discussed by the International Committee on Taxonomy of Viruses (ICTV) (255, 256).

Future prospects

PVAmpliconFinder has been developed for classification of *Papillomaviridae*-related sequences. Nonetheless, it can be easily modified and applied to other viral families, as this would only require a change in the interrogated databases and the reconstruction of a reference tree for the viral family being considered. The only limitation would be the number of sequences in the viral family of interest. Indeed, construction of the reference phylogenetic tree, and the use of RaxML-EPA, is time consuming and the computational time will increase with the number of sequences.
Once the partial L1 gene sequence of a putative novel HPV type has been characterized, the full genome reconstruction can be realized. To reach this goal, the use of the primer-walking strategy is one solution, but this method is laborious and time consuming. Another option is to use long-read technology, such as the MinION using a Nanopore technology. Although this novel sequencing technology is error-prone, the advantage of the long reads avoids potential chimeric reconstruction.

Conclusions

In summary, the present thesis describes a robust strategy based on the use of specific or degenerate primers and NGS technology to detect putative novel PVs. Although the identification of novel PV types or species can only be definitively confirmed by sequencing the entire L1 ORF, initial studies have confirmed the validity

of our new protocol as a first step in the isolation and full characterization of novel HPV genomes. While there is increasing use of NGS amplicon sequencing in clinical research settings, only few bioinformatics methods are available for the sensitive detection of HPV, and they are often restricted to a panel of already well-characterized PV types (257). The use of degenerate primers and PVAmpliconFinder may thus provide a solution for the detection and discovery of a broad range of HPV types.

List of figures

List of tables

Bibliography

1. Davis RA, Semken HA, Jr. Fossils of uncertain affinity from the upper devonian of iowa. Science. 1975;187(4173):251-4. PMID: 17838783
2. Lucey BP, Nelson-Rees WA, Hutchins GM. Henrietta Lacks, HeLa cells, and cell culture contamination. Archives of pathology & laboratory medicine. 2009;133(9):1463-7. PMID: 19722756
3. Lopez-Bueno A, Mavian C, Labella AM, Castro D, et al. Concurrence of Iridovirus, Polyomavirus, and a Unique Member of a New Group of Fish Papillomaviruses in Lymphocystis Disease-Affected Gilthead Sea Bream. J Virol. 2016;90(19):8768-79. PMID: 27440877
4. Isegawa N, Ohta M, Shirasawa H, Tokita H, et al. Nucleotide-sequence of a canine oral papillomavirus containing a long noncoding region. Int J Oncol. 1995;7(1):155-9. PMID: 21552821
5. Bravo IG, de Sanjose S, Gottschling M. The clinical importance of understanding the evolution of papillomaviruses. Trends in microbiology. 2010;18(10):432-8. PMID: 20739182
6. Bernard HU, Burk RD, Chen Z, van Doorslaer K, et al. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. Virology. 2010;401(1):70-9. PMID: 20206957
7. Agalliu I, Gapstur S, Chen Z, Wang T, et al. Associations of Oral alpha-, beta-, and gamma-Human Papillomavirus Types With Risk of Incident Head and Neck Cancer. JAMA oncology. 2016. PMID: 26794505
8. Guan J, Bywaters SM, Brendle SA, Ashley RE, et al. Cryoelectron Microscopy Maps of Human Papillomavirus 16 Reveal L2 Densities and Heparin Binding Site. Structure (London, England : 1993). 2017;25(2):253-63. PMID: 28065506
9. Harari A, Chen Z, Burk RD. Human papillomavirus genomics: past, present and future. Current problems in dermatology. 2014;45:1-18. PMID: 24643174
10. Chen HC, Schiffman M, Lin CY, Pan MH, et al. Persistence of type-specific human papillomavirus infection and increased long-term risk of cervical cancer. J Natl Cancer Inst. 2011;103(18):1387-96. PMID: 21900119
11. Van Doorslaer K, McBride AA. Molecular archeological evidence in support of the repeated loss of a papillomavirus gene. Scientific reports. 2016;6:33028. PMID: 27604338
12. Longworth MS, Laimins LA. Pathogenesis of human papillomaviruses in differentiating epithelia. Microbiology and molecular biology reviews : MMBR. 2004;68(2):362-72. PMID: 15187189
13. McBride AA. The papillomavirus E2 proteins. Virology. 2013;445(1-2):57-79. PMID: 23849793
14. Zobel T, Iftner T, Stubenrauch F. The papillomavirus E8-E2C protein represses DNA replication from extrachromosomal origins. Mol Cell Biol. 2003;23(22):8352-62. PMID: 14585992
15. Lace MJ, Anson JR, Thomas GS, Turek LP, et al. The E8--E2 gene product of human papillomavirus type 16 represses early transcription and replication but is dispensable for viral plasmid persistence in keratinocytes. J Virol. 2008;82(21):10841-53. PMID: 18753207
16. Ammermann I, Bruckner M, Matthes F, Iftner T, et al. Inhibition of transcription and DNA replication by the papillomavirus E8-E2C protein is mediated by interaction with corepressor molecules. J Virol. 2008;82(11):5127-36. PMID: 18353941
17. Tommasino M. The human papillomavirus family and its role in carcinogenesis. Semin Cancer Biol. 2014;26:13-21. PMID: 24316445
18. Gheit T. Mucosal and Cutaneous Human Papillomavirus Infections and Cancer Biology. Frontiers in oncology. 2019;9:355. PMID: 31134154
19. de Villiers EM, Fauquet C, Broker TR, Bernard HU, et al. Classification of papillomaviruses. Virology. 2004;324(1):17-27. PMID: 15183049

20. de Sanjose S, Quint WG, Alemany L, Geraets DT, et al. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. The Lancet Oncology. 2010;11(11):1048-56. PMID: 20952254
21. Dickens P, Srivastava G, Loke SL, Larkin S. Human papillomavirus 6, 11, and 16 in laryngeal papillomas. J Pathol. 1991;165(3):243-6. PMID: 1662265
22. Brianti P, De Flammineis E, Mercuri SR. Review of HPV-related diseases and cancers. The new microbiologica. 2017;40(2):80-5. PMID: 28368072
23. Egawa N, Egawa K, Griffin H, Doorbar J. Human Papillomaviruses; Epithelial Tropisms, and the Development of Neoplasia. Viruses. 2015;7(7):3863-90. PMID: 26193301
24. Bravo IG, Felez-Sanchez M. Papillomaviruses: Viral evolution, cancer and evolutionary medicine. Evolution, medicine, and public health. 2015;2015(1):32-51. PMID: 25634317
25. Viarisio D, Mueller-Decker K, Kloz U, Aengeneyndt B, et al. E6 and E7 from beta HPV38 cooperate with ultraviolet light in the development of actinic keratosis-like lesions and squamous cell carcinoma in mice. PLoS Pathog. 2011;7(7):e1002125. PMID: 21779166
26. Saidj D, Cros MP, Hernandez-Vargas H, Guarino F, et al. Oncoprotein E7 from beta human papillomavirus 38 induces formation of an inhibitory complex for a subset of p53-regulated promoters. J Virol. 2013;87(22):12139-50. PMID: 24006445
27. Schmitt A, Rochat A, Zeltner R, Borenstein L, et al. The primary target cells of the high-risk cottontail rabbit papillomavirus colocalize with hair follicle stem cells. J Virol. 1996;70(3):1912-22. PMID: 8627717
28. Egawa K. Do human papillomaviruses target epidermal stem cells? Dermatology (Basel, Switzerland). 2003;207(3):251-4. PMID: 14571065
29. Kines RC, Thompson CD, Lowy DR, Schiller JT, et al. The initial steps leading to papillomavirus infection occur on the basement membrane prior to cell surface binding. Proc Natl Acad Sci U S A. 2009;106(48):20458-63. PMID: 19920181
30. Day PM, Lowy DR, Schiller JT. Papillomaviruses infect cells via a clathrin-dependent pathway. Virology. 2003;307(1):1-11. PMID: 12667809
31. Joyce JG, Tung JS, Przysiecki CT, Cook JC, et al. The L1 major capsid protein of human papillomavirus type 11 recombinant virus-like particles interacts with heparin and cell-surface glycosaminoglycans on human keratinocytes. J Biol Chem. 1999;274(9):5810-22. PMID: 10026203
32. Handisurya A, Day PM, Thompson CD, Buck CB, et al. Murine skin and vaginal mucosa are similarly susceptible to infection by pseudovirions of different papillomavirus classifications and species. Virology. 2012;433(2):385-94. PMID: 22985477
33. Horvath CA, Boulet GA, Renoux VM, Delvenne PO, et al. Mechanisms of cell entry by human papillomaviruses: an overview. Virol J. 2010;7:11. PMID: 20089191
34. Moody CA, Laimins LA. Human papillomavirus oncoproteins: pathways to transformation. Nat Rev Cancer. 2010;10(8):550-60. PMID: 20592731
35. Bodily J, Laimins LA. Persistence of human papillomavirus infection: keys to malignant progression. Trends in microbiology. 2011;19(1):33-9. PMID: 21050765
36. Maglennon GA, McIntosh P, Doorbar J. Persistence of viral DNA in the epithelial basal layer suggests a model for papillomavirus latency following immune regression. Virology. 2011;414(2):153-63. PMID: 21492895
37. Doorbar J. The papillomavirus life cycle. J Clin Virol. 2005;32 Suppl 1:S7-15. PMID: 15753007
38. Bedell MA, Hudson JB, Golub TR, Turyk ME, et al. Amplification of human papillomavirus genomes in vitro is dependent on epithelial differentiation. J Virol. 1991;65(5):2254-60. PMID: 1850010
39. Florin L, Sapp C, Streeck RE, Sapp M. Assembly and translocation of papillomavirus capsid proteins. J Virol. 2002;76(19):10009-14. PMID: 12208977
40. Roden RBS, Stern PL. Opportunities and challenges for human papillomavirus vaccination in cancer. Nat Rev Cancer. 2018;18(4):240-54. PMID: 29497146

41. Antonsson A, Forslund O, Ekberg H, Sterner G, et al. The ubiquity and impressive genomic diversity of human skin papillomaviruses suggest a commensalic nature of these viruses. J Virol. 2000;74(24):11636-41. PMID: 11090162
42. Burchell AN, Winer RL, de Sanjose S, Franco EL. Chapter 6: Epidemiology and transmission dynamics of genital HPV infection. Vaccine. 2006;24 Suppl 3:S3/52-61. PMID: 16950018
43. Blake DR, Middleman AB. Human Papillomavirus Vaccine Update. Pediatric clinics of North America. 2017;64(2):321-9. PMID: 28292448
44. Burk RD, Chen Z, Van Doorslaer K. Human papillomaviruses: genetic basis of carcinogenicity. Public health genomics. 2009;12(5-6):281-90. PMID: 19684441
45. Bouvard V, Baan R, Straif K, Grosse Y, et al. A review of human carcinogens--Part B: biological agents. The Lancet Oncology. 2009;10(4):321-2. PMID: 19350698
46. Munoz N, Bosch FX, de Sanjose S, Herrero R, et al. Epidemiologic classification of human papillomavirus types associated with cervical cancer. The New England journal of medicine. 2003;348(6):518-27. PMID: 12571259
47. Smith JS, Lindsay L, Hoots B, Keys J, et al. Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update. Int J Cancer. 2007;121(3):621-32. PMID: 17405118
48. Li N, Franceschi S, Howell-Jones R, Snijders PJ, et al. Human papillomavirus type distribution in 30,848 invasive cervical cancers worldwide: Variation by geographical region, histological type and year of publication. Int J Cancer. 2011;128(4):927-35. PMID: 20473886
49. Kreimer AR, Clifford GM, Boyle P, Franceschi S. Human papillomavirus types in head and neck squamous cell carcinomas worldwide: a systematic review. Cancer Epidemiol Biomarkers Prev. 2005;14(2):467-75. PMID: 15734974
50. Reusser NM, Downing C, Guidry J, Tyring SK. HPV Carcinomas in Immunocompromised Patients. Journal of clinical medicine. 2015;4(2):260-81. PMID: 26239127
51. Smith JS. Prevalence of human papillomavirus infection in adolescent girls before reported sexual debut. J Infect Dis. 2014;210(6):835-6. PMID: 24740632
52. Houlihan CF, de Sanjose S, Baisley K, Changalucha J, et al. Prevalence of human papillomavirus in adolescent girls before reported sexual debut. J Infect Dis. 2014;210(6):837-45. PMID: 24740630
53. Castellsague X, Drudis T, Canadas MP, Gonce A, et al. Human Papillomavirus (HPV) infection in pregnant women and mother-to-child transmission of genital HPV genotypes: a prospective study in Spain. BMC Infect Dis. 2009;9:74. PMID: 19473489
54. Cheng S, Schmidt-Grimminger DC, Murant T, Broker TR, et al. Differentiation-dependent up-regulation of the human papillomavirus E7 gene reactivates cellular DNA replication in suprabasal differentiated keratinocytes. Genes & development. 1995;9(19):2335-49. PMID: 7557386
55. Sherman L, Jackman A, Itzhaki H, Stoppler MC, et al. Inhibition of serum- and calcium-induced differentiation of human keratinocytes by HPV16 E6 oncoprotein: role of p53 inactivation. Virology. 1997;237(2):296-306. PMID: 9356341
56. DiMaio D, Petti LM. The E5 proteins. Virology. 2013;445(1-2):99-114. PMID: 23731971
57. Bravo IG, Alonso A. Mucosal human papillomaviruses encode four different E5 proteins whose chemistry and phylogeny correlate with malignant or benign growth. J Virol. 2004;78(24):13613-26. PMID: 15564472
58. Ruesch MN, Stubenrauch F, Laimins LA. Activation of papillomavirus late gene transcription and genome amplification upon differentiation in semisolid medium is coincident with expression of involucrin and transglutaminase but not keratin-10. J Virol. 1998;72(6):5016-24. PMID: 9573271
59. Borgogna C, Landini MM, Lanfredini S, Doorbar J, et al. Characterization of skin lesions induced by skin-tropic alpha- and beta-papillomaviruses in a patient with epidermodysplasia verruciformis. Br J Dermatol. 2014;171(6):1550-4. PMID: 24902472

60. Antonsson A, Karanfilovska S, Lindqvist PG, Hansson BG. General acquisition of human papillomavirus infections of skin occurs in early infancy. J Clin Microbiol. 2003;41(6):2509-14. PMID: 12791874
61. Michael KM, Waterboer T, Sehr P, Rother A, et al. Seroprevalence of 34 human papillomavirus types in the German general population. PLoS Pathog. 2008;4(6):e1000091. PMID: 18566657
62. Grce M, Mravak-Stipetic M. Human papillomavirus-associated diseases. Clin Dermatol. 2014;32(2):253-8. PMID: 24559561
63. Gottschling M, Goker M, Kohler A, Lehmann MD, et al. Cutaneotropic human beta-/gamma-papillomaviruses are rarely shared between family members. J Invest Dermatol. 2009;129(10):2427-34. PMID: 19516265
64. Viarisio D, Muller-Decker K, Hassel JC, Alvarez JC, et al. The BRAF Inhibitor Vemurafenib Enhances UV-Induced Skin Carcinogenesis in Beta HPV38 E6 and E7 Transgenic Mice. J Invest Dermatol. 2017;137(1):261-4. PMID: 27650607
65. Viarisio D, Muller-Decker K, Accardi R, Robitaille A, et al. Beta HPV38 oncoproteins act with a hit-and-run mechanism in ultraviolet radiation-induced skin carcinogenesis in mice. PLoS Pathog. 2018;14(1):e1006783. PMID: 29324843
66. Chahoud J, Semaan A, Chen Y, Cao M, et al. Association Between beta-Genus Human Papillomavirus and Cutaneous Squamous Cell Carcinoma in Immunocompetent Individuals-A Meta-analysis. JAMA dermatology. 2016;152(12):1354-64. PMID: 26720285
67. Hasche D, Vinzon SE, Rosl F. Cutaneous Papillomaviruses and Non-melanoma Skin Cancer: Causal Agents or Innocent Bystanders? Frontiers in microbiology. 2018;9:874. PMID: 29770129
68. Pfister H, Fuchs PG, Majewski S, Jablonska S, et al. High prevalence of epidermodysplasia verruciformis-associated human papillomavirus DNA in actinic keratoses of the immunocompetent population. Arch Dermatol Res. 2003;295(7):273-9. PMID: 14618345
69. Weissenborn SJ, Nindl I, Purdie K, Harwood C, et al. Human papillomavirus-DNA loads in actinic keratoses exceed those in non-melanoma skin cancers. J Invest Dermatol. 2005;125(1):93-7. PMID: 15982308
70. Gottschling M, Stamatakis A, Nindl I, Stockfleth E, et al. Multiple evolutionary mechanisms drive papillomavirus diversification. Molecular biology and evolution. 2007;24(5):1242-58. PMID: 17344207
71. Gottschling M, Goker M, Stamatakis A, Bininda-Emonds OR, et al. Quantifying the phylodynamic forces driving papillomavirus evolution. Molecular biology and evolution. 2011;28(7):2101-13. PMID: 21285031
72. Zhao KN, Liu WJ, Frazer IH. Codon usage bias and A+T content variation in human papillomavirus genomes. Virus Res. 2003;98(2):95-104. PMID: 14659556
73. Tindle RW. Immune evasion in human papillomavirus-associated cervical cancer. Nat Rev Cancer. 2002;2(1):59-65. PMID: 11902586
74. Zhou J, Liu WJ, Peng SW, Sun XY, et al. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. J Virol. 1999;73(6):4972-82. PMID: 10233959
75. Johannsen E, Lambert PF. Epigenetics of human papillomaviruses. Virology. 2013;445(1-2):205-12. PMID: 23953230
76. Sanjuan R, Nebot MR, Chirico N, Mansky LM, et al. Viral mutation rates. J Virol. 2010;84(19):9733-48. PMID: 20660197
77. Shah SD, Doorbar J, Goldstein RA. Analysis of host-parasite incongruence in papillomavirus evolution using importance sampling. Molecular biology and evolution. 2010;27(6):1301-14. PMID: 20093429
78. Garcia-Vallve S, Iglesias-Rozas JR, Alonso A, Bravo IG. Different papillomaviruses have different repertoires of transcription factor binding sites: convergence and divergence in the upstream regulatory region. BMC evolutionary biology. 2006;6:20. PMID: 16526953

79. Vartanian JP, Guetard D, Henry M, Wain-Hobson S. Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. Science. 2008;320(5873):230-3. PMID: 18403710
80. Protic-Sabljic M, Tuteja N, Munson PJ, Hauser J, et al. UV light-induced cyclobutane pyrimidine dimers are mutagenic in mammalian cells. Mol Cell Biol. 1986;6(10):3349-56. PMID: 3540589
81. Dasgupta S, Zabielski J, Simonsson M, Burnett S. Rolling-circle replication of a high-copy BPV-1 plasmid. Journal of molecular biology. 1992;228(1):1-6. PMID: 1333015
82. Sakakibara N, Chen D, McBride AA. Papillomaviruses use recombination-dependent replication to vegetatively amplify their genomes in differentiated cells. PLoS Pathog. 2013;9(7):e1003321. PMID: 23853576
83. Jiang M, Xi LF, Edelstein ZR, Galloway DA, et al. Identification of recombinant human papillomavirus type 16 variants. Virology. 2009;394(1):8-11. PMID: 19758676
84. Narechania A, Chen Z, DeSalle R, Burk RD. Phylogenetic incongruence among oncogenic genital alpha human papillomaviruses. J Virol. 2005;79(24):15503-10. PMID: 16306621
85. Robles-Sikisaka R, Rivera R, Nollens HH, St Leger J, et al. Evidence of recombination and positive selection in cetacean papillomaviruses. Virology. 2012;427(2):189-97. PMID: 22386054
86. Woolford L, Rector A, Van Ranst M, Ducki A, et al. A novel virus detected in papillomas and carcinomas of the endangered western barred bandicoot (Perameles bougainville) exhibits genomic features of both the Papillomaviridae and Polyomaviridae. J Virol. 2007;81(24):13280-90. PMID: 17898069
87. Guerrero E, Daniel RW, Bosch FX, Castellsague X, et al. Comparison of ViraPap, Southern hybridization, and polymerase chain reaction methods for human papillomavirus identification in an epidemiological investigation of cervical cancer. J Clin Microbiol. 1992;30(11):2951-9. PMID: 1333485
88. Kuypers JM, Critchlow CW, Gravitt PE, Vernon DA, et al. Comparison of dot filter hybridization, Southern transfer hybridization, and polymerase chain reaction amplification for diagnosis of anal human papillomavirus infection. J Clin Microbiol. 1993;31(4):1003-6. PMID: 8385147
89. Morris BJ, Rose BR, Flanagan JL, McKinnon KJ, et al. Automated polymerase chain reaction for papillomavirus screening of cervicovaginal lavages: comparison with dot-blot hybridization in a sexually transmitted diseases clinic population. J Med Virol. 1990;32(1):22-30. PMID: 2173735
90. Schiffman MH, Bauer HM, Lorincz AT, Manos MM, et al. Comparison of Southern blot hybridization and polymerase chain reaction methods for the detection of human papillomavirus DNA. J Clin Microbiol. 1991;29(3):573-7. PMID: 1645370
91. Ward P, Parry GN, Yule R, Coleman DV, et al. Comparison between the polymerase chain reaction and slot blot hybridization for the detection of HPV sequences in cervical scrapes. Cytopathology : official journal of the British Society for Clinical Cytology. 1990;1(1):19-23. PMID: 1966674
92. Van den Brule AJ, Snijders P, Meijer CJ, Walboomers JM. PCR-based detection of genital HPV genotypes: an update and future perspectives. Papillomavirus Report. 1993(4):95-9. PMID:
93. Manos MM, Ting Y, Wright DK, Lewis J, et al. The use of polymerase chain reaction amplification for the detection of genital human papillomaviruses. Cancer Cells. 1989(7):209-14. PMID:
94. Snijders PJ, van den Brule AJ, Schrijnemakers HF, Snow G, et al. The use of general primers in the polymerase chain reaction permits the detection of a broad spectrum of human papillomavirus genotypes. J Gen Virol. 1990;71 (Pt 1):173-81. PMID: 2154534
95. Guerrero E, Shah KV. Polymerase chain reaction in HPV diagnosis. Papillomavirus Report. 1991(2):115-8. PMID:
96. Smits HL, Tieben LM, Tjong AHSP, Jebbink MF, et al. Detection and typing of human papillomaviruses present in fixed and stained archival cervical smears by a consensus

polymerase chain reaction and direct sequence analysis allow the identification of a broad spectrum of human papillomavirus types. J Gen Virol. 1992;73 (Pt 12):3263-8. PMID: 1335027
97. Ylitalo N, Bergstrom T, Gyllensten U. Detection of genital human papillomavirus by single-tube nested PCR and type-specific oligonucleotide hybridization. J Clin Microbiol. 1995;33(7):1822-8. PMID: 7665652
98. Clewley JP, Arnold C. MEGALIGN. The multiple alignment module of LASERGENE. Methods Mol Biol. 1997;70:119-29. PMID: 9089607
99. Gravitt PE, Peyton CL, Alessi TQ, Wheeler CM, et al. Improved amplification of genital human papillomaviruses. J Clin Microbiol. 2000;38(1):357-61. PMID: 10618116
100. de Roda Husman AM, Walboomers JM, van den Brule AJ, Meijer CJ, et al. The use of general primers GP5 and GP6 elongated at their 3' ends with adjacent highly conserved sequences improves human papillomavirus detection by PCR. J Gen Virol. 1995;76 (Pt 4):1057-62. PMID: 9049358
101. van den Brule AJ, Snijders PJ, Raaphorst PM, Schrijnemakers HF, et al. General primer polymerase chain reaction in combination with sequence analysis for identification of potentially novel human papillomavirus genotypes in cervical lesions. J Clin Microbiol. 1992;30(7):1716-21. PMID: 1321168
102. Schmitt M, Dondog B, Waterboer T, Pawlita M. Homogeneous amplification of genital human alpha papillomaviruses by PCR using novel broad-spectrum GP5+ and GP6+ primers. J Clin Microbiol. 2008;46(3):1050-9. PMID: 18199790
103. Schmitt M, Bravo IG, Snijders PJ, Gissmann L, et al. Bead-based multiplex genotyping of human papillomaviruses. J Clin Microbiol. 2006;44(2):504-12. PMID: 16455905
104. Garcia DA, Cid-Arregui A, Schmitt M, Castillo M, et al. Highly Sensitive Detection and Genotyping of HPV by PCR Multiplex and Luminex Technology in a Cohort of Colombian Women with Abnormal Cytology. The open virology journal. 2011;5:70-9. PMID: 21769306
105. Jablonska S, Majewski S. Epidermodysplasia verruciformis: immunological and clinical aspects. Current topics in microbiology and immunology. 1994;186:157-75. PMID: 8205840
106. Shamanin V, Delius H, de Villiers EM. Development of a broad spectrum PCR assay for papillomaviruses and its application in screening lung cancer biopsies. J Gen Virol. 1994;75 (Pt 5):1149-56. PMID: 8176375
107. de Villiers EM, Lavergne D, McLaren K, Benton EC. Prevailing papillomavirus types in non-melanoma carcinomas of the skin in renal allograft recipients. Int J Cancer. 1997;73(3):356-61. PMID: 9359482
108. Berkhout RJ, Tieben LM, Smits HL, Bavinck JN, et al. Nested PCR approach for detection and typing of epidermodysplasia verruciformis-associated human papillomavirus types in cutaneous cancers from renal transplant recipients. J Clin Microbiol. 1995;33(3):690-5. PMID: 7751378
109. Boxman IL, Berkhout RJ, Mulder LH, Wolkers MC, et al. Detection of human papillomavirus DNA in plucked hairs from renal transplant recipients and healthy volunteers. J Invest Dermatol. 1997;108(5):712-5. PMID: 9129220
110. Surentheran T, Harwood CA, Spink PJ, Sinclair AL, et al. Detection and typing of human papillomaviruses in mucosal and cutaneous biopsies from immunosuppressed and immunocompetent patients and patients with epidermodysplasia verruciformis: a unified diagnostic approach. Journal of clinical pathology. 1998;51(8):606-10. PMID: 9828820
111. Forslund O, Antonsson A, Nordin P, Stenquist B, et al. A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. J Gen Virol. 1999;80 (Pt 9):2437-43. PMID: 10501499
112. Myers G. HPV Sequence Database. In: Myers G, Baker CC, Münger K, Sverdrup F, McBride AA, Bernard HU, editors. Alignments In Human Papillomaviruses. Los Alamos1997. p. II-L1–23–73.

113. Forslund O, Ly H, Reid C, Higgins G. A broad spectrum of human papillomavirus types is present in the skin of Australian patients with non-melanoma skin cancers and solar keratosis. Br J Dermatol. 2003;149(1):64-73. PMID: 12890196
114. Antonsson A, Hansson BG. Healthy skin of many animal species harbors papillomaviruses which are closely related to their human counterparts. J Virol. 2002;76(24):12537-42. PMID: 12438579
115. Forslund O, Ly H, Higgins G. Improved detection of cutaneous human papillomavirus DNA by single tube nested 'hanging droplet' PCR. J Virol Methods. 2003;110(2):129-36. PMID: 12798239
116. Chan SY, Delius H, Halpern AL, Bernard HU. Analysis of genomic sequences of 95 papillomavirus types: uniting typing, phylogeny, and taxonomy. J Virol. 1995;69(5):3074-83. PMID: 7707535
117. Walsh EE, Falsey AR, Swinburne IA, Formica MA. Reverse transcription polymerase chain reaction (RT-PCR) for diagnosis of respiratory syncytial virus infection in adults: use of a single-tube "hanging droplet" nested PCR. J Med Virol. 2001;63(3):259-63. PMID: 11170067
118. Alotaibi L, Provost N, Gagnon S, Franco EL, et al. Diversity of cutaneous human papillomavirus types in individuals with and without skin lesion. J Clin Virol. 2006;36(2):133-40. PMID: 16678481
119. Nordin P, Hansson BG, Hansson C, Blohme I, et al. Human papilloma virus in skin, mouth and uterine cervix in female renal transplant recipients with or without a history of cutaneous squamous cell carcinoma. Acta dermato-venereologica. 2007;87(3):219-22. PMID: 17533486
120. Forslund O, DeAngelis PM, Beigi M, Schjolberg AR, et al. Identification of human papillomavirus in keratoacanthomas. Journal of cutaneous pathology. 2003;30(7):423-9. PMID: 12859739
121. Li J, Pan Y, Xu Z, Wang Q, et al. Improved detection of human papillomavirus harbored in healthy skin with FAP6085/64 primers. J Virol Methods. 2013;193(2):633-8. PMID: 23871757
122. Gheit T, Billoud G, de Koning MN, Gemignani F, et al. Development of a sensitive and specific multiplex PCR method combined with DNA microarray primer extension to detect Betapapillomavirus types. J Clin Microbiol. 2007;45(8):2537-44. PMID: 17581938
123. Ekstrom J, Bzhalava D, Svenback D, Forslund O, et al. High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions. Int J Cancer. 2011;129(11):2643-50. PMID: 21630257
124. Ekstrom J, Muhr LS, Bzhalava D, Soderlund-Strand A, et al. Diversity of human papillomaviruses in skin lesions. Virology. 2013;447(1-2):300-11. PMID: 24210127
125. Munday JS, Tucker RS, Kiupel M, Harvey CJ. Multiple oral carcinomas associated with a novel papillomavirus in a dog. Journal of veterinary diagnostic investigation : official publication of the American Association of Veterinary Laboratory Diagnosticians, Inc. 2015;27(2):221-5. PMID: 25613043
126. Kocjan BJ, Hosnjak L, Racnik J, Zadravec M, et al. Complete Genome Sequence of Phodopus sungorus Papillomavirus Type 1 (PsPV1), a Novel Member of the Pipapillomavirus Genus, Isolated from a Siberian Hamster. Genome announcements. 2014;2(2). PMID: 24723726
127. Stevens H, Heylen E, De Keyser K, Maes R, et al. Complete Genome Sequence of the Crocuta crocuta Papillomavirus Type 1 (CcrPV1) from a Spotted Hyena, the First Papillomavirus Characterized in a Member of the Hyaenidae. Genome announcements. 2013;1(1). PMID: 23405364
128. Ure AE, Elfadl AK, Khalafalla AI, Gameel AA, et al. Characterization of the complete genomes of Camelus dromedarius papillomavirus types 1 and 2. J Gen Virol. 2011;92(Pt 8):1769-77. PMID: 21471319
129. Claus MP, Lunardi M, Alfieri AF, Ferracin LM, et al. Identification of unreported putative new bovine papillomavirus types in Brazilian cattle herds. Veterinary microbiology. 2008;132(3-4):396-401. PMID: 18617336

130. Kocjan BJ, Bzhalava D, Forslund O, Dillner J, et al. Molecular methods for identification and characterization of novel papillomaviruses. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2015;21(9):808-16. PMID: 26003284
131. Chouhy D, Gorosito M, Sanchez A, Serra EC, et al. New generic primer system targeting mucosal/genital and cutaneous human papillomaviruses leads to the characterization of HPV 115, a novel Beta-papillomavirus species 3. Virology. 2010;397(1):205-16. PMID: 19948351
132. Bolatti EM, Hosnjak L, Chouhy D, Re-Louhau MF, et al. High prevalence of Gammapapillomaviruses (Gamma-PVs) in pre-malignant cutaneous lesions of immunocompetent individuals using a new broad-spectrum primer system, and identification of HPV210, a novel Gamma-PV type. Virology. 2018;525:182-91. PMID: 30292127
133. Brancaccio RN, Robitaille A, Dutta S, Cuenin C, et al. Generation of a novel next-generation sequencing-based method for the isolation of new human papillomavirus types. Virology. 2018;520:1-10. PMID: 29747121
134. Baines JE, McGovern RM, Persing D, Gostout BS. Consensus-degenerate hybrid oligonucleotide primers (CODEHOP) for the detection of novel papillomaviruses and their application to esophageal and tonsillar carcinomas. J Virol Methods. 2005;123(1):81-7. PMID: 15582702
135. Rose TM, Henikoff JG, Henikoff S. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Res. 2003;31(13):3763-6. PMID: 12824413
136. Chouhy D, Bolatti EM, Piccirilli G, Sanchez A, et al. Identification of human papillomavirus type 156, the prototype of a new human gammapapillomavirus species, by a generic and highly sensitive PCR strategy for long DNA fragments. J Gen Virol. 2013;94(Pt 3):524-33. PMID: 23136368
137. Rector A, Tachezy R, Van Ranst M. A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification. J Virol. 2004;78(10):4993-8. PMID: 15113879
138. Marincevic-Zuniga Y, Gustavsson I, Gyllensten U. Multiply-primed rolling circle amplification of human papillomavirus using sequence-specific primers. Virology. 2012;432(1):57-62. PMID: 22739442
139. Johne R, Muller H, Rector A, van Ranst M, et al. Rolling-circle amplification of viral DNA genomes using phi29 polymerase. Trends in microbiology. 2009;17(5):205-11. PMID: 19375325
140. Bzhalava D, Muhr LS, Lagheden C, Ekstrom J, et al. Deep sequencing extends the diversity of human papillomaviruses in human skin. Scientific reports. 2014;4:5807. PMID: 25055967
141. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-7. PMID: 271968
142. Lander ES, Linton LM, Birren B, Nusbaum C, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921. PMID: 11237011
143. Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008;26(10):1135-45. PMID: 18846087
144. Bahassi el M, Stambrook PJ. Next-generation sequencing technologies: breaking the sound barrier of human genetics. Mutagenesis. 2014;29(5):303-10. PMID: 25150023
145. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics (Oxford, England). 2012;28(24):3169-77. PMID: 23060614
146. Arroyo Muhr LS, Hultin E, Bzhalava D, Eklund C, et al. Human papillomavirus type 197 is commonly present in skin tumors. Int J Cancer. 2015;136(11):2546-55. PMID: 25388227
147. Bzhalava D, Johansson H, Ekstrom J, Faust H, et al. Unbiased approach for virus detection in skin lesions. PLoS One. 2013;8(6):e65953. PMID: 23840382
148. Phan TG, Vo NP, Aronen M, Jartti L, et al. Novel human gammapapillomavirus species in a nasal swab. Genome announcements. 2013;1(2):e0002213. PMID: 23516180

149. Canuti M, Deijs M, Jazaeri Farsani SM, Holwerda M, et al. Metagenomic analysis of a sample from a patient with respiratory tract infection reveals the presence of a gamma-papillomavirus. Frontiers in microbiology. 2014;5:347. PMID: 25071755
150. Iwamori S. [Advantages in the introduction of POS]. [Kango kyoiku] Japanese journal of nurses' education. 1985;26(7):438-41. PMID: 3874987
151. Johansson H, Bzhalava D, Ekstrom J, Hultin E, et al. Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types. Virology. 2013;440(1):1-7. PMID: 23522725
152. da Silva FR, Cibulski SP, Daudt C, Weber MN, et al. Novel Bovine Papillomavirus Type Discovered by Rolling-Circle Amplification Coupled with Next-Generation Sequencing. PLoS One. 2016;11(9):e0162345. PMID: 27606703
153. Yan L, Liu K, Sintim HO. Convenient detection of HPV virus in a clinical sample using concurrent rolling circle and junction probe amplifications. Chemical communications (Cambridge, England). 2014;50(54):7147-9. PMID: 24852020
154. Pastrana DV, Peretti A, Welch NL, Borgogna C, et al. Metagenomic Discovery of 83 New Human Papillomavirus Types in Patients with Immunodeficiency. mSphere. 2018;3(6). PMID: 30541782
155. Bankevich A, Nurk S, Antipov D, Gurevich AA, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology. 2012;19(5):455-77. PMID: 22506599
156. Tirosh O, Conlan S, Deming C, Lee-Lin SQ, et al. Expanded skin virome in DOCK8-deficient patients. Nature medicine. 2018;24(12):1815-21. PMID: 30397357
157. Zhang Q, Dove CG, Hor JL, Murdock HM, et al. DOCK8 regulates lymphocyte shape integrity for skin antiviral immunity. J Exp Med. 2014;211(13):2549-66. PMID: 25422492
158. Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, et al. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. Nucleic Acids Res. 2013;41(Database issue):D571-8. PMID: 23093593
159. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome biology. 2016;17(1):132. PMID: 27323842
160. Jiang P, Singh M. SPICi: a fast clustering algorithm for large biological networks. Bioinformatics (Oxford, England). 2010;26(8):1105-11. PMID: 20185405
161. Yang X, Charlebois P, Gnerre S, Coole MG, et al. De novo assembly of highly diverse viral populations. BMC genomics. 2012;13:475. PMID: 22974120
162. Berthet N, Descorps-Declere S, Nkili-Meyong AA, Nakoune E, et al. Improved assembly procedure of viral RNA genomes amplified with Phi29 polymerase from new generation sequencing data. Biological research. 2016;49(1):39. PMID: 27605096
163. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience. 2019;8(9). PMID: 31494669
164. Yang Y, Smith SA. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC genomics. 2013;14:328. PMID: 23672450
165. Binga EK, Lasken RS, Neufeld JD. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. The ISME journal. 2008;2(3):233-41. PMID: 18256705
166. Yilmaz S, Allgaier M, Hugenholtz P. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nature methods. 2010;7(12):943-4. PMID: 21116242
167. Lasken RS. Genomic DNA amplification by the multiple displacement amplification (MDA) method. Biochemical Society transactions. 2009;37(Pt 2):450-3. PMID: 19290880
168. Chevreux B, Wetter T, Suhai S, editors. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. The German Conference on Bioinformatics; 1999: Computer Science and Biology.
169. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome research. 2001;11(10):1725-9. PMID: 11591649

170. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2010;26(5):589-95. PMID: 20080505
171. Li R, Zhu H, Ruan J, Qian W, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome research. 2010;20(2):265-72. PMID: 20019144
172. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494-512. PMID: 23845962
173. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics (Oxford, England). 2012;28(11):1420-8. PMID: 22495754
174. Meisel JS, Nasko DJ, Brubach B, Cepeda-Espinoza V, et al. Current progress and future opportunities in applications of bioinformatics for biodefense and pathogen detection: report from the Winter Mid-Atlantic Microbiome Meet-up, College Park, MD, January 10, 2018. Microbiome. 2018;6(1):197. PMID: 30396371
175. Woese CR, Stackebrandt E, Macke TJ, Fox GE. A phylogenetic definition of the major eubacterial taxa. Systematic and applied microbiology. 1985;6:143-51. PMID: 11542017
176. Woese CR. Bacterial evolution. Microbiological reviews. 1987;51(2):221-71. PMID: 2439888
177. Delwart EL. Viral metagenomics. Reviews in medical virology. 2007;17(2):115-31. PMID: 17295196
178. Baker BJ, Dick GJ. Omic approaches in microbial ecology: Charting the unknown. Microbe. 2013(8):353-60. PMID:
179. Nix WA, Oberste MS, Pallansch MA. Sensitive, seminested PCR amplification of VP1 sequences for direct identification of all enterovirus serotypes from original clinical specimens. J Clin Microbiol. 2006;44(8):2698-704. PMID: 16891480
180. Pehler-Harrington K, Khanna M, Waters CR, Henrickson KJ. Rapid detection and identification of human adenovirus species by adenoplex, a multiplex PCR-enzyme hybridization assay. J Clin Microbiol. 2004;42(9):4072-6. PMID: 15364992
181. Berry D, Ben Mahfoudh K, Wagner M, Loy A. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Applied and environmental microbiology. 2011;77(21):7846-9. PMID: 21890669
182. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008;36(16):e105. PMID: 18660515
183. Gomez-Alvarez V, Teal TK, Schmidt TM. Systematic artifacts in metagenomes from complex microbial communities. The ISME journal. 2009;3(11):1314-7. PMID: 19587772
184. Huse SM, Huber JA, Morrison HG, Sogin ML, et al. Accuracy and quality of massively parallel DNA pyrosequencing. Genome biology. 2007;8(7):R143. PMID: 17659080
185. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environmental microbiology. 2010;12(1):118-23. PMID: 19725865
186. Nakamura K, Oshima T, Morimoto T, Ikeda S, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011;39(13):e90. PMID: 21576222
187. Taub MA, Corrada Bravo H, Irizarry RA. Overcoming bias and systematic errors in next generation sequencing data. Genome medicine. 2010;2(12):87. PMID: 21144010
188. Zhou Q, Su X, Ning K. Assessment of quality control approaches for metagenomic data analysis. Scientific reports. 2014;4:6957. PMID: 25376098
189. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research. 1998;8(3):186-94. PMID: 9521922
190. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 2012;7(2):e30619. PMID: 22312429
191. Andrews S. FastQC: a quality control tool for high throughput sequence data 2010 [Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

192. Almeida OGG, De Martinis ECP. Bioinformatics tools to assess metagenomic data for applied microbiology. Applied microbiology and biotechnology. 2019;103(1):69-82. PMID: 30362076
193. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17(1):3. PMID:
194. Krueger F. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files 2015 [Available from: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
195. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT. Individual genome assembly from complex community short-read metagenomic datasets. The ISME journal. 2012;6(4):898-901. PMID: 22030673
196. Mende DR, Waller AS, Sunagawa S, Jarvelin AI, et al. Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS One. 2012;7(2):e31386. PMID: 22384016
197. Junemann S, Kleinbolting N, Jaenicke S, Henke C, et al. Bioinformatics for NGS-based metagenomics and the application to biogas research. Journal of biotechnology. 2017;261:10-23. PMID: 28823476
198. van der Walt AJ, van Goethem MW, Ramond JB, Makhalanyane TP, et al. Assembling metagenomes, one community at a time. BMC genomics. 2017;18(1):521. PMID: 28693474
199. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95(6):315-27. PMID: 20211242
200. Chapman BE, O'Sullivan WJ, Scopes RK, Reed GH. Magnetic resonance studies on manganese-nucleotide complexes of phosphoglycerate kinase. Biochemistry. 1977;16(5):1005-10. PMID: 321006
201. Huang X. A contig assembly program based on sensitive detection of fragment overlaps. Genomics. 1992;14(1):18-25. PMID: 1427824
202. Bang-Jensen J, Gutin G, Yeo A. When the greedy algorithm fails. Discrete Optimization. 2004;1(2):121-7. PMID:
203. Peng Y, Leung HC, Yiu SM, Chin FY. Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics (Oxford, England). 2011;27(13):i94-101. PMID: 21685107
204. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40(20):e155. PMID: 22821567
205. Boisvert S, Laviolette F, Corbeil J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. Journal of computational biology : a journal of computational molecular cell biology. 2010;17(11):1519-33. PMID: 20958248
206. Boisvert S, Raymond F, Godzaridis E, Laviolette F, et al. Ray Meta: scalable de novo metagenome assembly and profiling. Genome biology. 2012;13(12):R122. PMID: 23259615
207. Treangen TJ, Koren S, Sommer DD, Liu B, et al. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome biology. 2013;14(1):R2. PMID: 23320958
208. Li D, Liu CM, Luo R, Sadakane K, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics (Oxford, England). 2015;31(10):1674-6. PMID: 25609793
209. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome research. 2017;27(5):824-34. PMID: 28298430
210. Sczyrba A, Hofmann P, Belmann P, Koslicki D, et al. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nature methods. 2017;14(11):1063-71. PMID: 28967888
211. Meyer F, Paarmann D, D'Souza M, Olson R, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC bioinformatics. 2008;9:386. PMID: 18803844
212. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38(20):e191. PMID: 20805240

213. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England). 2010;26(19):2460-1. PMID: 20709691
214. Wilke A, Harrison T, Wilkening J, Field D, et al. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. BMC bioinformatics. 2012;13:141. PMID: 22720753
215. Kent WJ. BLAT--the BLAST-like alignment tool. Genome research. 2002;12(4):656-64. PMID: 11932250
216. Keegan KP, Glass EM, Meyer F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Methods Mol Biol. 2016;1399:207-33. PMID: 26791506
217. Altschul SF, Gish W, Miller W, Myers EW, et al. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403-10. PMID: 2231712
218. Huson DH, Beier S, Flade I, Gorska A, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS computational biology. 2016;12(6):e1004957. PMID: 27327495
219. Burge S, Kelly E, Lonsdale D, Mutowo-Muellenet P, et al. Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. Database : the journal of biological databases and curation. 2012;2012:bar068. PMID: 22301074
220. Gerlach W, Stoye J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Res. 2011;39(14):e91. PMID: 21586583
221. Krause L, Diaz NN, Goesmann A, Kelley S, et al. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. 2008;36(7):2230-9. PMID: 18285365
222. Finn RD, Bateman A, Clements J, Coggill P, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):D222-30. PMID: 24288371
223. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology. 2014;15(3):R46. PMID: 24580807
224. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome research. 2016;26(12):1721-9. PMID: 27852649
225. Gregor I, Droge J, Schirmer M, Quince C, et al. PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. PeerJ. 2016;4:e1603. PMID: 26870609
226. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nature methods. 2009;6(9):673-6. PMID: 19648916
227. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome research. 2014;24(7):1180-92. PMID: 24899342
228. Takeuchi F, Sekizuka T, Yamashita A, Ogasawara Y, et al. MePIC, metagenomic pathogen identification for clinical specimens. Japanese journal of infectious diseases. 2014;67(1):62-5. PMID: 24451106
229. Hong C, Manimaran S, Shen Y, Perez-Rogers JF, et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome. 2014;2:33. PMID: 25225611
230. Francis OE, Bendall M, Manimaran S, Hong C, et al. Pathoscope: species identification and strain attribution with unassembled sequencing data. Genome research. 2013;23(10):1721-9. PMID: 23843222
231. Zaharia M, Bolosky WJ, Curtis K, Fox A, et al. Faster and More Accurate Sequence Alignment with SNAP. ArXiv. 2011:abs/1111.5572. PMID:
232. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics (Oxford, England). 2012;28(1):125-6. PMID: 22039206
233. Simpson JT, Wong K, Jackman SD, Schein JE, et al. ABySS: a parallel assembler for short read sequence data. Genome research. 2009;19(6):1117-23. PMID: 19251739

234. Treangen TJ, Sommer DD, Angly FE, Koren S, et al. Next generation sequence assembly with AMOS. Current protocols in bioinformatics. 2011;Chapter 11:Unit 11.8. PMID: 21400694
235. Roux S, Tournayre J, Mahul A, Debroas D, et al. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC bioinformatics. 2014;15:76. PMID: 24646187
236. Wommack KE, Bhavsar J, Polson SW, Chen J, et al. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Standards in genomic sciences. 2012;6(3):427-39. PMID: 23407591
237. Punta M, Coggill PC, Eberhardt RY, Mistry J, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40(Database issue):D290-301. PMID: 22127870
238. Borozan I, Wilson S, Blanchette P, Laflamme P, et al. CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. BMC bioinformatics. 2012;13:206. PMID: 22901030
239. Ho T, Tzanetakis IE. Development of a virus detection and discovery pipeline using next generation sequencing. Virology. 2014;471-473:54-60. PMID: 25461531
240. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357-9. PMID: 22388286
241. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research. 2008;18(5):821-9. PMID: 18349386
242. Zheng Y, Gao S, Padmanabhan C, Li R, et al. VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. Virology. 2017;500:130-8. PMID: 27825033
243. Chen Y, Yao H, Thompson EJ, Tannir NM, et al. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics (Oxford, England). 2013;29(2):266-7. PMID: 23162058
244. Wang Q, Jia P, Zhao Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome medicine. 2015;7(1):2. PMID: 25699093
245. Zhao G, Krishnamurthy S, Cai Z, Popov VL, et al. Identification of novel viruses using VirusHunter--an automated data analysis pipeline. PLoS One. 2013;8(10):e78470. PMID: 24167629
246. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England). 2006;22(13):1658-9. PMID: 16731699
247. Naeem R, Rashid M, Pain A. READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. Bioinformatics (Oxford, England). 2013;29(3):391-2. PMID: 23193222
248. Bhaduri A, Qu K, Lee CS, Ungewickell A, et al. Rapid identification of non-human sequences in high-throughput sequencing datasets. Bioinformatics (Oxford, England). 2012;28(8):1174-5. PMID: 22377895
249. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009;10(3):R25. PMID: 19261174
250. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nature biotechnology. 2011;29(5):393-6. PMID: 21552235
251. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research. 2008;18(11):1851-8. PMID: 18714091
252. Zhao G, Wu G, Lim ES, Droit L, et al. VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. Virology. 2017;503:21-30. PMID: 28110145
253. Lagstrom S, Umu SU, Lepisto M, Ellonen P, et al. TaME-seq: An efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration. Scientific reports. 2019;9(1):524. PMID: 30679491

254. Tommasino M. The biology of beta human papillomaviruses. Virus Res. 2017;231:128-38. PMID: 27856220
255. Adams MJ, Lefkowitz EJ, King AM, Harrach B, et al. 50 years of the International Committee on Taxonomy of Viruses: progress and prospects. Arch Virol. 2017;162(5):1441-6. PMID: 28078475
256. Simmonds P, Adams MJ, Benko M, Breitbart M, et al. Consensus statement: Virus taxonomy in the age of metagenomics. Nature reviews Microbiology. 2017;15(3):161-8. PMID: 28134265
257. Schmitt M, Depuydt C, Benoy I, Bogers J, et al. Multiple human papillomavirus infections with high viral loads are associated with cervical lesions but do not differentiate grades of cervical abnormalities. J Clin Microbiol. 2013;51(5):1458-64. PMID: 23447632

Publications from collaborations

9) Ainouze M, Rochefort P, Parroche P, Roblot G, Tout I, et al. (2018) Human papillomavirus type 16 antagonizes IRF6 regulation of IL-1β. PLOS Pathogens 14(8): e1007158. https://doi.org/10.1371/journal.ppat.1007158

10) Gheit T., Dutta S., Oliver J., Robitaille A., Hampras S., Combes J.-D., McKay-Chopin S., et al. (2017) Isolation and characterization of a novel putative human polyomavirus. Virology, 506 , pp. 45-54. https://doi.org/10.1016/j.virol.2017.03.007

11) Gupta P, Shahzad N, Harold A, Shuda M, Venuti V, et al. (2019) Merkel Cell Polyomavirus downregulates N-myc downstream regulated gene-1 (NDRG1) leading to cellular proliferation and migration. Journal of Virology (Accepted)

12) Olivier M, Bouaoun L, Villar S, Robitaille A, Cahais V, et al. (2019) Molecular features of premenopausal breast cancers in Latin American women: Pilot results from the PRECAMA study. PLOS ONE 14(1): e0210372. https://doi.org/10.1371/journal.pone.0210372

13) Romero M. (2019) Human Papillomavirus Type 38 alters wild-type p53 activity to promote cell proliferation via the down-regulation of Integrin Alpha-1 expression" (reference number: NMICROBIOL-19071774A), Nature Microbiology (Submitted)

14) Savelli B, Li Q, Webber M, Jemmat A, **Robitaille A**, Zamocky M, ..., Dunand C. (2019). RedoxiBase: A database for ROS homeostasis regulated proteins. *Redox biology*, 101247. https://doi.org/10.1016/j.redox.2019.101247

15) Vargas-Ayala RC., Jay A, Manara F, Maroui MA, Hernandez-Vargas H, Diederichs A, ..., Cros MP. (2019). Interplay between the epigenetic enzyme lysine (K)-specific demethylase 2B and Epstein-Barr virus infection. *Journal of virology*, *93*(13), e00273-19. https://doi.org/10.1128/JVI.00273-19

16) Viarisio D, Müller-Decker K, Accardi R, **Robitaille A,** Dürst M, Beer K, ... & Voegele C. (2018). Beta HPV38 oncoproteins act with a hit-and-run mechanism in ultraviolet radiation-induced skin carcinogenesis in mice. *PLoS pathogens*, *14*(1), e1006783. https://doi.org/10.1371/journal.ppat.1006783

17) Viarisio D, Robitaille A, Müller-Decker K, Flechtenmacher C, Gissmann L, Tommasino M (2019) Cancer susceptibility of beta HPV49 E6 and E7 transgenic mice to 4-nitroquinoline 1-oxide treatment correlates with mutational signatures of tobacco exposure, Virology, Volume 538, Pages 53-60, ISSN 0042-6822, https://doi.org/10.1016/j.virol.2019.09.010.

18) Zhivagui M, Ng AW, Ardin M, Churchwell MI, Pandey M., Renard C, ... & Zavadil J. (2019). Experimental and pan-cancer genome analyses reveal widespread contribution of

acrylamide exposure to carcinogenesis in humans. *Genome research*, *29*(4), 521-531. https://doi.org/10.1101/gr.242453.118

19) Zhivagui M, Ardin M, Ng AW, Churchwell M, Pandey M, Villar S, ... & Zavadil J. (2018). Experimental analysis of exome-scale mutational signature of glycidamide, the reactive metabolite of acrylamide. *BioRxiv*, 254664.

PLOS PATHOGENS

RESEARCH ARTICLE

Human papillomavirus type 16 antagonizes IRF6 regulation of IL-1β

Michelle Ainouze[1,2,3,4,5‡], Pauline Rochefort[1,2,3,4,5‡], Peggy Parroche[1,2,3,4,5‡], Guillaume Roblot[1,2,3,4,5], Issam Tout[1,2,3,4,5], François Briat[1,2,3,4,5], Claudia Zannetti[1,2,3,4,5], Marie Marotel[1,2,3,4,5], Nadege Goutagny[6], Philip Auron[7], Alexandra Traverse-Glehen[5,6], Aude Lunel-Potencier[5], Francois Golfier[5], Murielle Masson[8], Alexis Robitaille[9], Massimo Tommasino[9], Christine Carreira[9], Thierry Walzer[1,2,3,4,5], Thomas Henry[1,2,3,4,5], Katia Zanier[8], Gilles Trave[8], Uzma Ayesha Hasan[1,2,3,4,5]*

1 Centre International de recherche en Infectiologie, CIRI, Inserm, U1111, Lyon, France, **2** Université Claude Bernard Lyon 1, Lyon, France, **3** CNRS, UMR5308, Lyon, France, **4** École Normale Supérieure de Lyon, Univ Lyon, France, **5** Hospices Civils de Lyon, France, **6** Cancer Research Centre of Lyon, INSERM U1052-CNRS UMR5286, Lyon, France, **7** Duquesne University, Pittsburgh, Pennsylvania, United States of America, **8** IGBMC, UMR 7104-U964, ILLKIRCH, France, **9** IARC, Lyon, France

‡ These authors share first authorship on this work.
* uzma.hasan@inserm.fr

Citation: Ainouze M, Rochefort P, Parroche P, Roblot G, Tout I, Briat F, et al. (2018) Human papillomavirus type 16 antagonizes IRF6 regulation of IL-1β. PLoS Pathog 14(8): e1007158. https://doi.org/10.1371/journal.ppat.1007158

Editor: Paul Francis Lambert, University of Wisconsin Madison School of Medicine and Public Health, UNITED STATES

Received: January 12, 2018

Accepted: June 15, 2018

Published: August 8, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Funding for reagents for this stufy was provided by LNCC: https://www.ligue-cancer.net/ and awarded to UAH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Human papillomavirus type 16 (HPV16) and other oncoviruses have been shown to block innate immune responses and to persist in the host. However, to avoid viral persistence, the immune response attempts to clear the infection. IL-1β is a powerful cytokine produced when viral motifs are sensed by innate receptors that are members of the inflammasome family. Whether oncoviruses such as HPV16 can activate the inflammasome pathway remains unknown. Here, we show that infection of human keratinocytes with HPV16 induced the secretion of IL-1β. Yet, upon expression of the viral early genes, IL-1β transcription was blocked. We went on to show that expression of the viral oncoprotein E6 in human keratinocytes inhibited IRF6 transcription which we revealed regulated IL-1β promoter activity. Preventing E6 expression using siRNA, or using E6 mutants that prevented degradation of p53, showed that p53 regulated IRF6 transcription. HPV16 abrogation of p53 binding to the IRF6 promoter was shown by ChIP in tissues from patients with cervical cancer. Thus E6 inhibition of IRF6 is an escape strategy used by HPV16 to block the production IL-1β. Our findings reveal a struggle between oncoviral persistence and host immunity; which is centered on IL-1β regulation.

Author summary

Oncoviruses block innate immune responses to persist in the host. However, to avoid viral persistence, the immune response attempts to clear the infection. IL-1β is a pro-inflammatory cytokine produced by the inflammasome pathway. Whether oncoviruses such as human papillomavirus (HPV) can activate the inflammasome remains to be explored. We demonstrated that keratinocytes, the host cell type for papillomaviruses,

when infected with HPV16 induced IL-1β transcription and secretion. Yet, upon expression of the viral oncoprotein E6, IL-1β transcription was blocked. E6 expression inhibited IRF6 transcriptional regulation of the IL-1β promoter. Preventing E6 expression, or its ability to degrade p53, restored the ability of IRF6 to bind to the IL-1β promoter. HPV16 abrogation of p53, IRF6 and IL-1β expression was fully confirmed in cervical cancer cells and tissues from patients. These data highlight the equilibrium between the host innate immune rheostat and viral immune escape.

Introduction

The innate immune system is the first line of defense in response to danger signals from microbial invasion or tissue injury. Viruses are sensed by several immune receptors that activate signaling pathways leading to cytokine production. Many oncogenic viruses can deregulate several immune-related pathways which guarantee a persistent infection. High-Risk Human Papilloma Viruses (HR HPV) are the etiological factor of cervical as well as certain head and neck cancers and is responsible for 20% of all human cancers linked to infection [1]. Persistence and progression of the disease are achieved by deregulating both cellular and immune defense mechanisms. Among the HR types, HPV16 is the most prevalent type in premalignant and malignant cervical lesions [2]. HPV16 viral oncoproteins E6 and E7 can target many cellular proteins such as binding and degrading the tumor suppressors' p53, and pRb, respectively. In parallel E6 and E7 are able to deregulate several innate immune-related pathways that block cytokine and chemokine production, antigen presentation, and adherence molecules [3]. Recently Lau et al., showed that E7 from HPV18 suppresses the cGAS pathway by inhibiting the adapter protein STING [4]. Similarly, some antiviral genes induced by interferons such as IFIT1, MX1 and the innate sensors RIG-I, TLR3 and TLR9 are also inhibited by HPV [5,6]. Indeed, Niebler et al., and Karim et al., have shown that HPV is capable of blocking IL-1β [7,8]. On the flip side, host cells have strategies to thwart viral immune escape.

IL-1β is crucial in host-defenses towards infection and injury. Our current understanding is that regulation of IL-1β is controlled by two checkpoints: 1. The activation and translocation of the nuclear factor-κB (NF-κB) which initiates the transcription of the pro-IL-1β gene. 2. Post-translational regulation of pro-IL-1β into its cleaved form by the inflammasome cytosolic multi-protein complex. The inflammasome complex consists of an innate pathogen recognition receptors such as the nucleotide-binding domain and leucine-rich repeat pyrin domain 3 (NLRP3) or absent in melanoma 2 (AIM2). Upon viral recognition the inflammasome sensor recruits the apoptosis-associated, speck-like protein containing a carboxy-terminal CARD (ASC). Caspase-1 is activated within the inflammasome multiprotein complex through interaction with ASC that bridges NLRP3 or AIM2. The activation of caspase-1 is associated with pyroptosis, a form of programmed cell death distinct from apoptosis, as well as the cleavage of the proinflammatory cytokines IL-1β and IL-18. Once released, mature IL-1β and IL-18 signal to their target cells, thus allowing the expansion of innate and adaptive immune responses. NLRP3 inflammasomes are activated by viruses such as adenovirus [9], vaccinia virus [10], and hepatitis C virus (HCV) [11]. The AIM2 inflammasome has been shown to detect vaccinia virus [12] and murine CMV [13]. Whether the inflammasome plays a protective role against HPV16 remains to be investigated.

Here we demonstrate that HPV16 induces the secretion of IL-1β from human keratinocytes. IL-1β produced from HPV16 infected keratinocytes blocked gene viral transcription. However, inhibition was lost after 8h due to the ability of the viral oncoprotein E6 (16E6) to

inhibit IL-1β transcription. A 16E6 protein binding domain essential for p53 degradation played a crucial role in regulating IL-1β transcription. 16E6 blocked the p53 transcriptional regulation of Interferon Regulatory Factor 6 (IRF6), which we found was essential for IL-1β promoter activity. The identification of this inhibitory transcriptional loop represents an undiscovered mechanism of oncoviral immune hijacking in the infected host cell.

Results

HPV16 induces the transcription and secretion of IL-1β

We first determined whether inflammasome activation could be achieved in normal human keratinocytes, the host of HPV infection. Addition of poly dA:dT (an AIM2 activator) or Nigericin (an NLRP3 agonist) led to the secretion of IL-1β (S1 Fig). Of note, the induction of pro-IL-1β did not require the first check point signal (S1A Fig). Pro-IL-1β is constitutively expressed in human keratinocytes and has been previously described by Sand et al., and Zepter et al [14,15]. We next tested whether HPV16 induced IL-1β gene expression in human keratinocytes. To do this, we generated HPV16 Quasivirions (16QsV) that closely resemble the natural virus as well control Pseudovirions (PsV). 16QsV are viral particles that contain the full viral genome of HPV16 encaspidated by the viral late proteins L1 and L2 (L1/L2). PsV are viral particles that contain GFP DNA encaspidated by L1/L2 [6]. Infection in keratinocytes with 16QsV up to 4h led to an increase of IL-1β transcripts (Fig 1A). However, post 8h infection, IL-1β transcription decreased (Fig 1A). The level of IL-1β gene expression inversely correlated to viral gene transcription (Fig 1B). Furthermore, primary keratinocytes infected with 16QsV induced IL-1β or IL-18 secretion at 4h but not at 24h (Fig 1C). 16QsV induction of IL-1β depended on caspase-1 activity (S1B Fig). Pyroptosis was also induced by 16QsV as measured by lactate dehydrogenase activity (S1C Fig). We did not observe IL-1β secretion when PsV or when extracts of the late proteins L1/L2 was added to keratinocytes (Fig 1C). These data suggest that 16QsV can induce caspase -1 dependent IL-1β, IL-18 as well as pyroptosis during the early phases of infection.

IL-1β production is blocked by the viral oncoproteins 16E6 and E7

IL-1β has been shown to block HBV replication in human hepatocytes [16]. Therefore we evaluated whether IL-1β could inhibit HPV16 viral gene transcription. Primary keratinocytes were infected with 16QsV or PsV ± recombinant IL-1β. We observed that IL-1β blocked 16QsV viral expression as measured by E1 transcripts (Fig 1D). This effect was reversed when we blocked the IL-1 receptor using Anakinra (Fig 1E). The viral oncoproteins E6 and E7 inhibit several innate immune pathways such as TLR9, STING and IRF signaling [4,6,17]. Based on these reports we hypothesized that E6 and E7 were responsible for the inhibition of IL-1β. To test this, human primary keratinocytes were transduced with recombinant retrovirus expressing HPV16 E6 and E7 (16E6E7) or with the empty vector control (pLXSN). 16E6E7 blocked both AIM2 and NLRP3-mediated secretion of IL-1β (Fig 2A). Furthermore, knock down of the viral oncoproteins using siRNA targeting16E6E7 restored the ability of cells to produce IL-1β (Fig 2B). In the epidermis, keratinocytes are the first cells to be encountered by external stimuli to induce IL-1β which in turn stimulates IL-8 secretion by human dermal fibroblasts [18]. We established an IL-8 bioassay in which addition of recombinant IL-1β induced IL-8 promoter activity of the luciferase gene in HEK293 cells (Fig 2C). Specificity of the assay was controlled using IL-1R inhibitor (Anakinra) (Fig 2C). Supernatants that were derived from AIM2 stimulated primary human keratinocytes induced the expression of the IL-8 luciferase gene. However, supernatants derived from AIM2 stimulated 16E6E7 cells failed to induce IL-8 transcription (Fig 2D). Furthermore, knock down of the viral oncoproteins using siRNA for

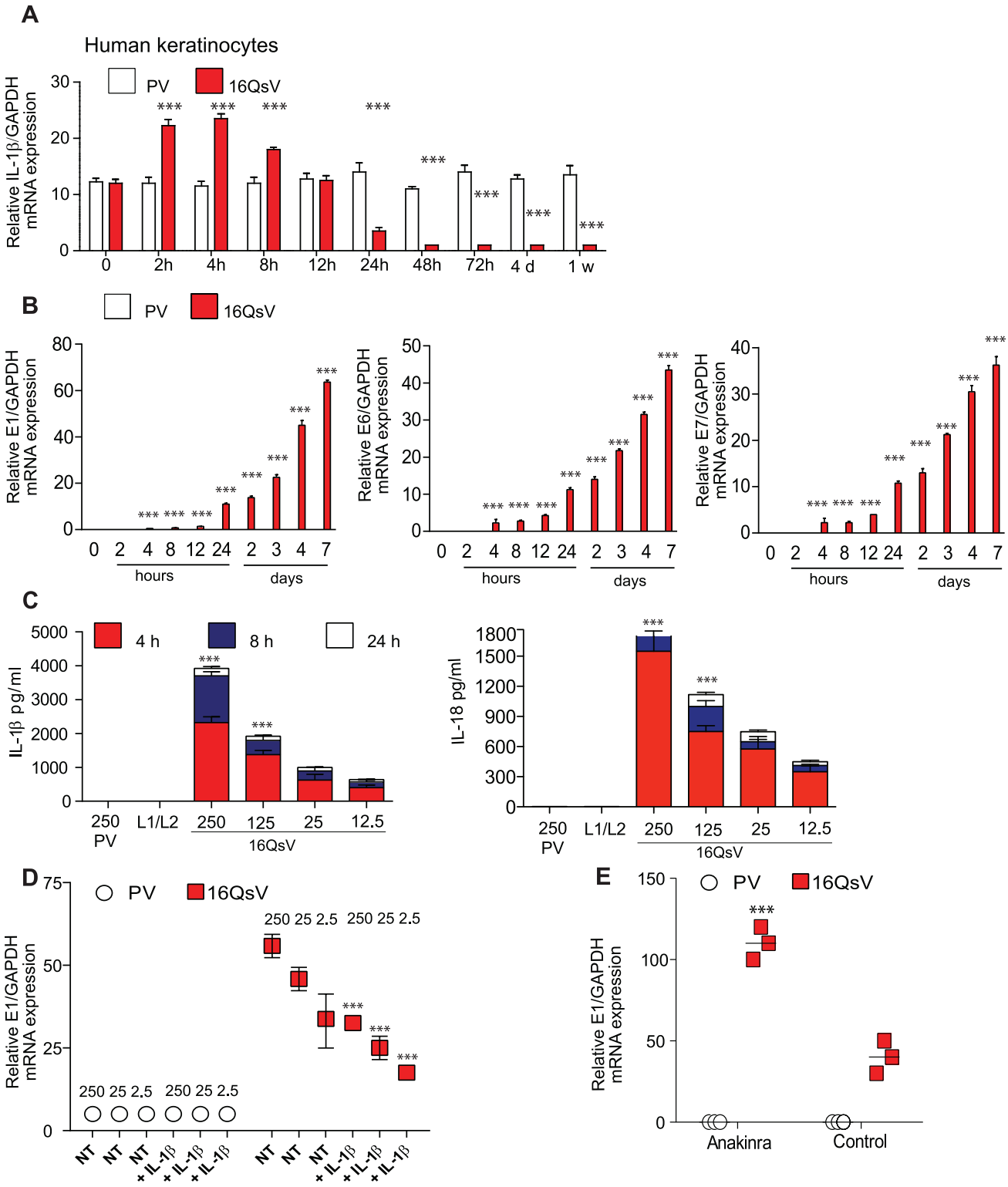


Fig 1. HPV16 induces transient IL-1β secretion by keratinocytes. (A) Human primary keratinocytes were treated as indicated with 16QsV or PsV (at 200 viral genome equivalents (v.g.e) per cell). IL-1β transcripts were determined by RT-qPCR. n = 5. (B) As in A, E1, E6 and E7 mRNA relative levels were determined by RT-qPCR. n = 4. (C) Human keratinocytes were treated at 4, 8 and 24 h with 16QsV at different v.g.e per cell. Supernatants were harvested and IL-1β or IL-18 production was measured by ELISA. PsV or L1/L2 fractions were added as controls. n = 5. (D) Human keratinocytes were treated with 16QsV

or PsV at decreasing v.g.e per cell for 24 h ± recombinant IL-1β (200pg/ml). Cells were harvested and E7 mRNA levels were measured by RT-qPCR. n = 5. (E) Human keratinocytes were treated with 16QsV or PsV (200 v.g.e) for 24 h ± IL-1R inhibitor (Anakinra). Cells were harvested and E1 mRNA levels were measured by RTqPCR. qPCR. n = 5. Data are representative of n independent experiments performed. Shown are the mean ± SEM with ***, $P < 0.0001$, based on a two way ANOVA test.

https://doi.org/10.1371/journal.ppat.1007158.g001

16E6E7 restored the ability of a cervical cancer-derived cell line (SiHa HPV16+) to produce IL-1β in response to Nigericin, poly dA:dT and 16QsV (Fig 2E). Thus 16E6E7 oncoproteins block IL-1β secretion. We corroborated our findings using supernatants from cervical cancer cell lines that were stimulated with the NLRP3 ligand. We observed that supernatants from the cervical cell line C33A (HPV-) stimulated with nigericin induced IL-8 luciferase activity (Fig 2F). Furthermore, Anakinra blocked IL-8 gene induction from supernatants derived from C33A cells stimulated with the NLRP3 ligand (Fig 2F). In contrast, supernatants taken from SiHa and CaSki cells (HPV16+) that were stimulated with nigericin failed to induce IL-8 promoter activity (Fig 2F). In summary, we have demonstrated the ability of HPV16 E6 and/or E7 to block IL-1β paracrine induction of IL-8 transcription.

HPV16E6E7 abrogates mRNA expression of pro-IL-1β

We hypothesized that the loss of IL-1β production might be due to the ability of 16E6E7 to block NLPR3 and AIM2 transcription. Neither AIM2 nor NLRP3 transcript levels were altered in human primary keratinocytes transduced with 16E6E7, compared to the pLXSN control (S2A Fig). HPV16 E6 and E7 interact with p53 and retinoblastoma (pRb), respectively, and promote their degradation via the proteasome pathway [19]. Therefore, we next determined whether a similar mechanism affected NLRP3 or AIM2 protein expression in 16E6E7-expressing keratinocytes. Human NLRP3-CFP, AIM2-CFP or p53 constructs were co-transfected with 16E6E7 or pLXSN in human primary keratinocytes and their expression was examined by immunoblotting. We did not observe any alteration in AIM2 or NLRP3 protein levels. As expected we found that p53 was degraded by 16E6E7 (S2B Fig). As we did not detect any change at the receptor level, we next focused our attention on the downstream signaling molecules that are shared between NLRP3 and AIM2. Inflammasome activation requires ASC dependent caspase-1 maturation of pro-IL-1β [12]. Neither ASC nor caspase-1 transcript levels were altered in 16E6E7 compared to pLXSN transduced cells (S2C Fig). In addition cleavage of pro-caspase-1 was detected in 16E6E7 transduced cells stimulated with NLRP3 or AIM2 ligands (S2D Fig). We observed that levels of the pro-form of IL-1β were already reduced in 16E6E7 compared to LXSN transduced cells. These data indicated that the synthesis of IL-1β was affected by the viral oncoproteins before AIM2 or NLRP3 stimulation (Fig 3A–3C). The same loss of pro-IL-1β was observed in cervical cancer cell lines positive for HPV16 (Fig 3D). All these observations showed that 16E6E7 exerts an inhibitory effect on the synthesis of the pro-form of IL-1β. While Niebler et al., previously reported the ability of 16E6 to degrade pro-IL-1β via the proteasome [8], under our experimental conditions the addition of a specific proteasome inhibitor on 16E6E7 expressing keratinocytes did not restore the pro-IL-1β protein (Fig 3E). As expected, p53 levels increased in the presence of 16E6E7 confirming the specificity of the proteasome inhibitor (Fig 3E). Protein levels for 16E6 were controlled by western blot (Fig 3E). Indeed an alternative hypothesis was that 16E6E7 proteins can alter IL-1β mRNA, as shown by Karim et al, and Niebler et al., [7,8]. We observed that 16E6E7 blocked the level of IL-1β transcripts compared to normal cells (Fig 3F). Little or no IL-1β mRNA was detected in CaSki or SiHa compared to C33A cells (Fig 3G). These data indicated that 16E6E7 in human keratinocytes as well as in cervical cancer cells supresses mRNA expression of IL-1β.

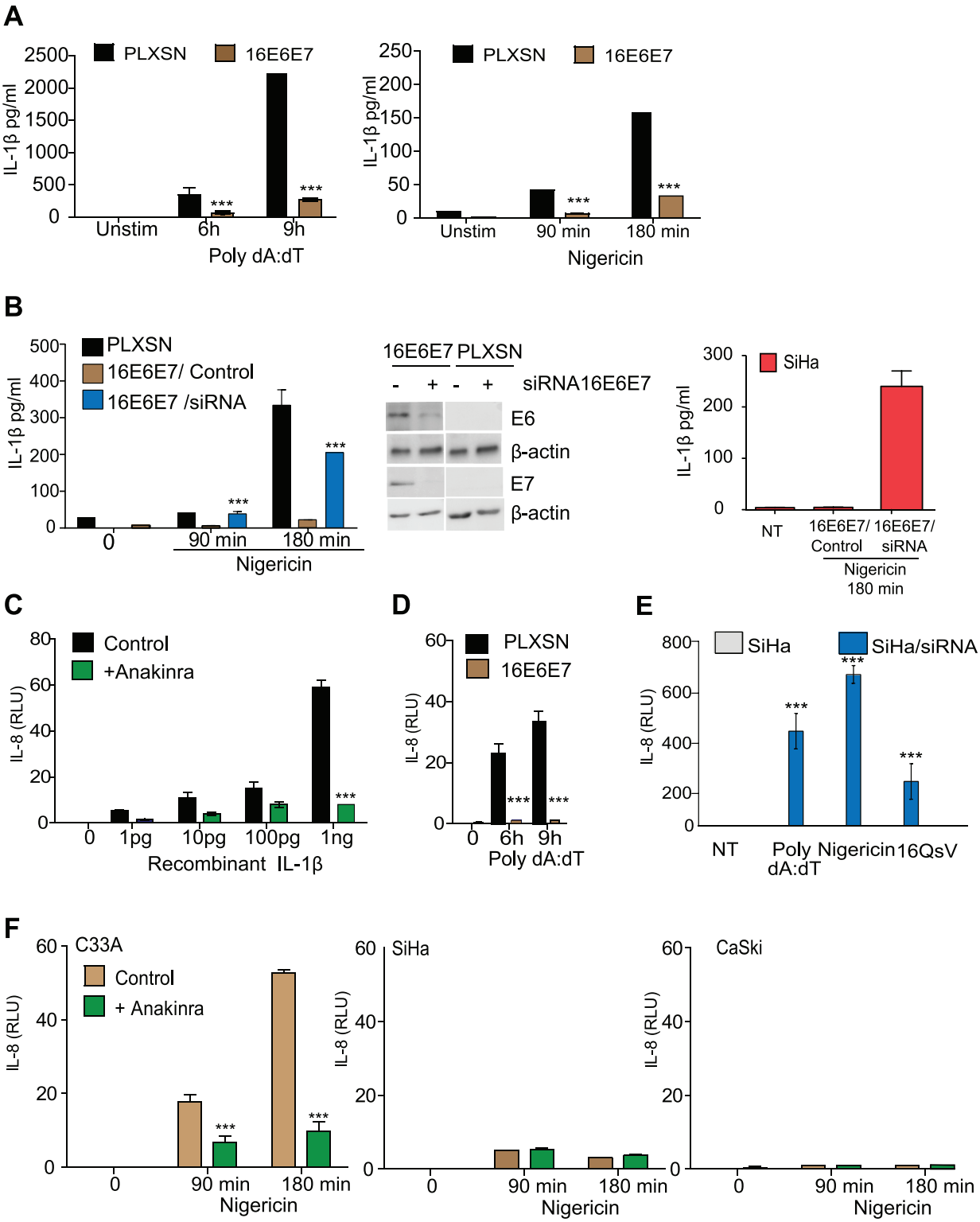


Fig 2. 16E6E7 block IL-1β production in primary human keratinocytes and in cervical cancer derived cells lines. (A) Analysis of the IL-1β production by ELISA in human keratinocytes transduced with pLXSN or 16E6E7 stimulated with nigericin or poly dA:dT. n = 10. (B) Human keratinocytes transduced with pLXSN or 16E6E7 transfected with a siRNA targeting 16E6E7 (+) or the scramble control (-). Cells were stimulated

with the NLRP3 ligand nigericin and IL-1β secretion was measured by ELISA. Middle, western blot of E6 or E7 siRNA efficacy on 16E6E7 or PLXSN transduced cells. n = 4. Left SiHa cell were treated with a siRNA targeting 16E6E7 (+) or the scramble control (-). Cells were stimulated with the NLRP3 ligand nigericin and IL-1β secretion was measured by ELISA. n = 4. (C) IL-8 bioassay: HEK293T cells transiently expressing the IL-8 promoter linked to luciferase gene were treated with increasing concentrations of recombinant IL-1β ± Anakinra. Twenty four h post treatment cells were harvested and luciferase activity was measured. n = 4. (D) IL-8 bioassay using supernatants from human keratinocytes transduced with pLXSN or 16E6E7± AIM2 ligand poly dA:dT. n = 4. (E) Cervical cancer cells (SiHa) were transfected with a siRNA targeting 16E6E7 or the scramble control. The cells were stimulated with the NLRP3 ligand nigericin, AIM2 ligand poly dA:dT or 16QsV (200 v.g.e per cell) and IL-1β was measured by ELISA. n = 4. (F) IL-8 bioassay using supernatants from cervical cancer cell lines ± nigericin. n = 6. Data are representative of n independent experiments performed in triplicate. Shown are the mean ± SEM with ***, $P < 0.0001$, based on a two way ANOVA test. For immunoblotting data, 1 out of n = 3 experiments is shown.

https://doi.org/10.1371/journal.ppat.1007158.g002

16E6 as well as E6 from other high-risk HPV types block IL-1β transcription

HPV16 may use E6 and/or E7 to directly inhibit IL-1β transcription. To determine whether HPV16 E6 or E7 proteins influence IL-1β transcription, the IL-1β promoter linked to the luciferase reporter gene was co-transfected ±16E6E7, 16E6 or E7 into spontaneously immortalized human keratinocytes (NIKs). NIKs already expressed high protein levels of endogenous pro-IL-1β. Indeed high basal luciferase activity was detected in these cells after transient transfection. However, 16E6E7 inhibited IL-1β luciferase activity even with low DNA concentrations (Fig 4A left), indicating that 16E6E7 can block the transcription of the IL-1β.. Furthermore 16E6, and to a lesser extent 16E7, inhibited IL-1β promoter activity (Fig 4A and 4B). Knock down of 16E6 restored pro-IL-1β (Fig 4C). We also compared the efficiency of E6 from other high-risk (HR) human papillomavirus types and one low risk (LR) type in repressing IL-1β transcriptional activity. HR types 18E6 and 31E6 inhibited IL-1β transcription, although less efficiently than 16E6 (S3A Fig). LR HPV6E6 did not affect IL-1β promoter activity (S3A Fig). These data demonstrated that E6 from HPV16 as well as other HR types strongly inhibit IL-1β transcription.

The inhibition of IL-1β transcription by 16E6 involves an ISRE *cis* element on the IL-1β promoter

We next made deletions in the promoter to determine which region is required by 16E6 to inhibit IL-1β transcription. (Fig 4D). WT and IL-1β deletion constructs were co-transfected with 16E6. We restored IL-1β promoter activity with deletion 2 in the presence of 16E6 (Fig 4E). The deletion contains an area called LILRE was previously characterized by Unlu and colleagues [20]. The LILRE element has a high degree of inter-species conservation and plays and important role in IL-1β regulation (Fig 4F). Within the LILRE region, Unlu et al., showed the involvement of three different protein binding sites [20], an Spi-1 cis site (ETS); an IRF8-binding site (ISRE) and a Stat1 cis site (GAS) [20].

We hypothesized that 16E6 requires the regulatory LILRE site to inhibit IL-1β transcription. To test this, primary human keratinocytes were co-transfected ± 16E6 or pLXSN with WT, del-LILRE (deletion of the LILRE site) and constructs that contained point mutations (m) for ISRE, ETS or GAS on the IL-1β promoter. Luciferase activity was restored with the delLILRE promoter indicating that this site contains a region required for IL-1β inhibition by 16E6 (Fig 4G). Luciferase activity remained suppressed in cells transfected with ETS mutant, suggesting that this *cis* element was not involved in the down-regulation of IL-1β transcription by 16E6 (Fig 4G). Luciferase activity was partially rescued in cells transfected with the mGAS promoter (Fig 4G). However, a complete rescue was observed in cells that were transfected with the mISRE promoter in the presence of 16E6. These results suggested that IL-1β down regulation by 16E6 principally involves the ISRE site on the IL-1β promoter.

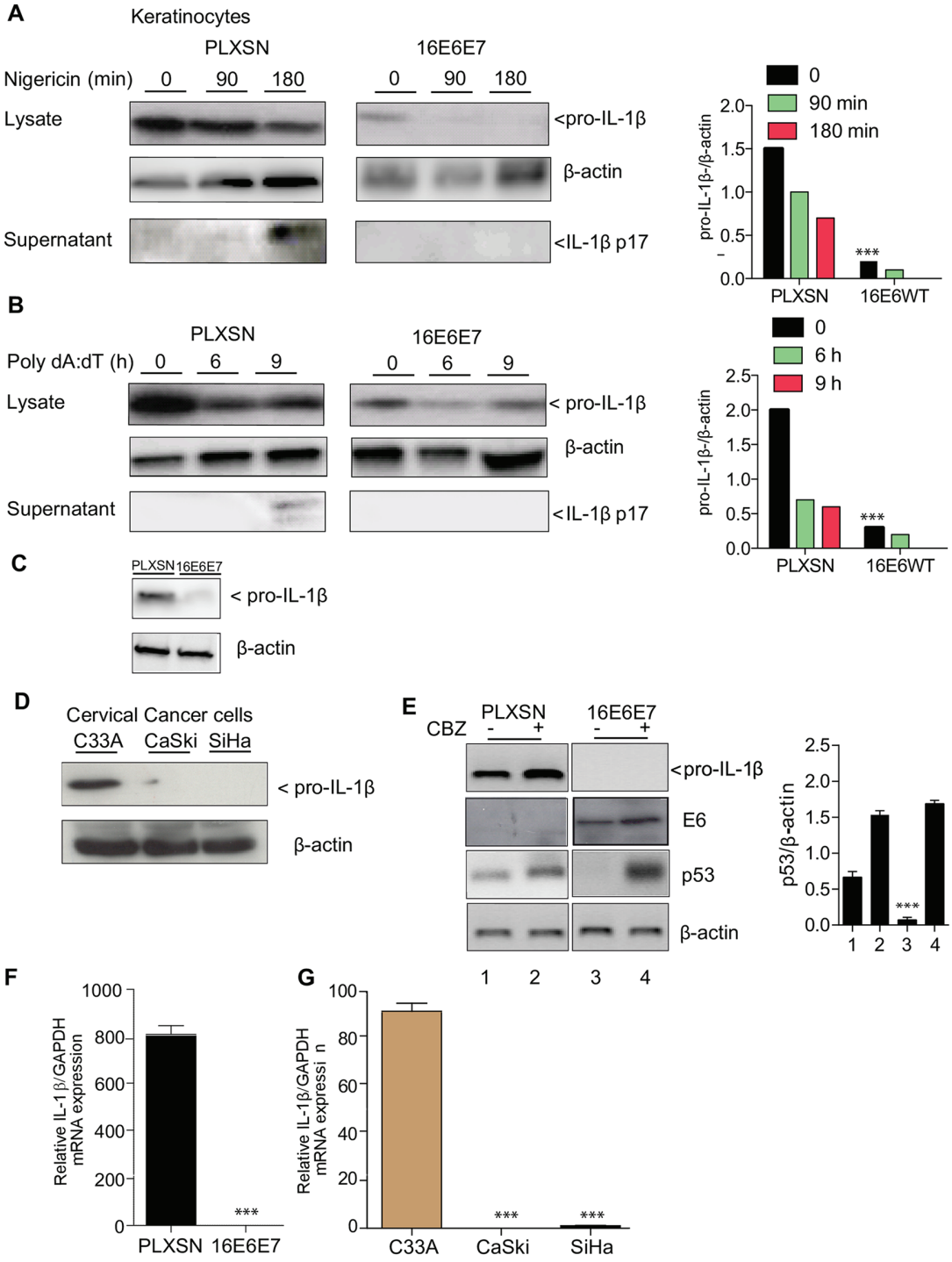


Fig 3. HPV16 oncoproteins inhibit pro-IL-1β levels. (A) Human keratinocytes transduced with pLXSN or 16E6E7 were stimulated with AIM2 and (B) NLPR3 ligands and both pro-IL-1β and IL-1β from cell lysates or supernatants were analysed by immunoblotting. β-actin was used as a loading control. Densitometry analysis was performed n = 3. (C) Immunoblotting of pro-IL-1β in human keratinocytes transduced with pLXSN or 16E6E7. (D) Cervical cancer cell lines were lysed and immunoblotting for pro-IL-1β was performed. n = 4 (E) Human keratinocytes transduced with pLXSN or 16E6E7 were treated for 24 h with N-CBZ-Leu-Leu-Leu-al. Cells were harvested and p53, E6 as well as pro-IL-1β levels were determined by immunoblotting. Right, p53 densitometry levels were normalized to β-actin. Below, immunoblot analysis of the 16E6 protein. n = 3. (F) RNA was extracted from Human keratinocytes transduced with pLXSN or 16E6E7 and IL-1β transcripts relative expression was determined by RT-

qPCR. n = 5. (G) RNA was extracted from patient derived cervical cancer cell lines and IL-1β transcripts were determined by RT-qPCR. n = 6. Panels A-E. Data are representative of n independent experiments performed in triplicate. Shown are the mean ± SEM with ***, $P < 0.0001$, based on a two way ANOVA test. Panel F $P < 0.0001$, based on a one way ANOVA test. Panel G Student unpaired T test was performed comparing C33A to CaSki or SiHa. For immunoblotting data, 1 out of 3 experiments is shown.

https://doi.org/10.1371/journal.ppat.1007158.g003

16E6 expression inhibits the binding of IRF6 and not IRF8 on the IL-1β promoter

The observation that an ISRE site is required for IL-1β suppression by 16E6 prompted us to determine which transcription is involved in this event. IRF8 is required for the development of monocytes, macrophages, dendritic cells (DCs), basophils, and eosinophils, while it inhibits the generation of neutrophils [21], yet nothing has been described for its role in keratinocytes. We observed no difference in gene or protein expression of IRF8 in primary human keratinocytes vs. 16E6 or E7 transduced cells (S3B Fig). Furthermore, by ChIP we observed in human macrophages IRF8 binding on the ISRE element, however in human keratinocytes we failed to demonstrate binding (S2C Fig). We concluded that IRF8 did not regulate the IL-1β promoter in human keratinocytes. In contrast to most IRFs, IRF6 has no identified function in innate immunity but is essential for normal keratinocyte epidermal development and differentiation [22]. We hypothesized that IRF6 might be involved in IL-1β transcription. To test this we co-transfected the IL-1β promoter with IRF8, IRF6 or pUNO expression vectors in HEK293 cells. As expected, IRF8 induced a significant increase in IL-1β luciferase activity when compared to pUNO transfected cells (Fig 5A). We also observed for the first time that IRF6 expression also increased IL-1β promoter activity in a dose dependent manner (Fig 5A). Oligo pull-down assays revealed IRF6 as well as IRF8 specific binding to the ISRE site on the IL-1β promoter (Fig 5B).

Having established that IRF6 binds to the IL-1β promoter and induces IL-1β transcription, we hypothesized that 16E6 might alter IRF6 expression. Indeed, IRF6 expression in human keratinocytes was decreased in cells expressing 16E6 (Fig 5C and 5D). Furthermore, immunofluorescence detection of IRF6 in primary keratinocytes was localized in the nucleus but shifted into the cytoplasm in 16E6 cells (Fig 5E). ImageJ analysis of IRF6 fluorescence showed that both cytoplasmic and nuclear levels were reduced in keratinocytes expressing 16E6 (Fig 5E). Furthermore, both mRNA and protein levels for IRF6 were lower in CaSki (HPV16+) versus NIKs (Fig 5F and 5G). siRNA targeting of 16E6 reversed the effect, and IRF6 levels were resorted (Fig 5H). We also observed that IRF6 protein levels and mRNA levels were reduced when epithelial cells were treated with increasing amounts of 16QsV (Fig 5I and 5J). The decrease of IRF6 mRNA levels was inversely proportional to viral DNA expression of E7 (Fig 5J). ChIP assays revealed that IRF6 bound less to the ISRE element when cells were infected with 16QsV (Fig 5K).

In summary, we confirmed that IRF8 is required to induce IL-1β expression in monocytes, yet in human keratinocytes IRF6 regulates IL-1β transcription. Furthermore, IRF6 binding to the ISRE site on the IL-1β promoter is inhibited by 16E6 expression in primary human keratinocytes.

16E6 mutations reveal that E6 degradation of p53 is required to inhibit IRF6 transcription

The HPV16 oncoprotein E6 interacts with numerous proteins by hijacking several host cellular networks. To gain further insight into the mechanistic role of 16E6 on IL-1β transcription, we co-transfected the IL-1β promoter with plasmid constructs that contain point mutations that

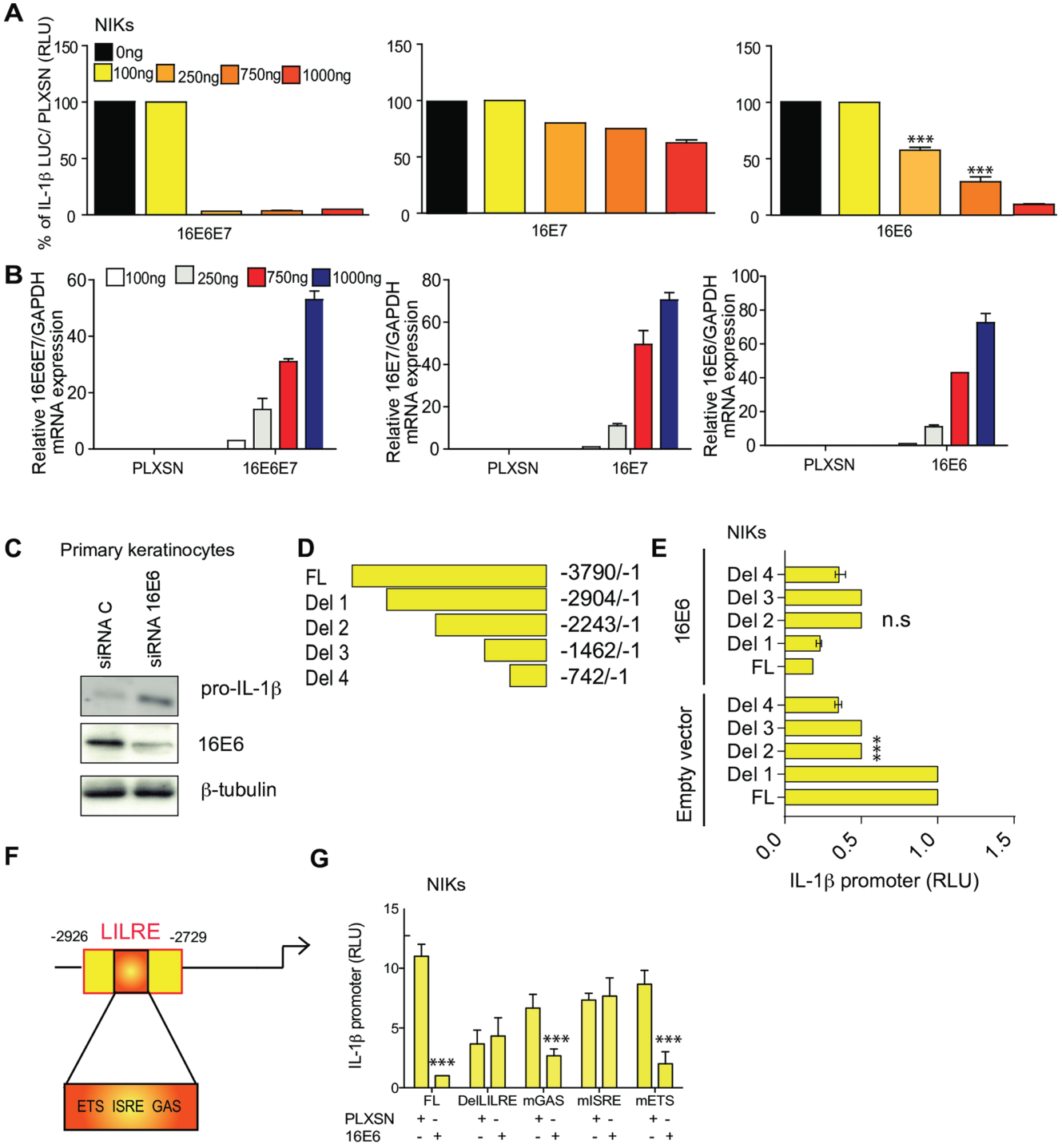


Fig 4. HPV16 E6 down-regulates the IL-1β promoter in cervical cells via the ISRE site. (A) NIKs were co-transfected with the IL-1β promoter with increasing concentrations of pLXSN HPV16E6E7 or 16E6 or 16E7 as indicated. After 48 h, cells were harvested and luciferase activity was measured. n = 5. (B) Relative expression of 16E6E7, 16E6 or 16E7 were measured by RT-qPCR. n = 5. (C) Primary human keratinocytes transduced with 16E6 and treated with a scramble or siRNA against 16E6. Protein levels of pro-IL-1β and loading control β-tubulin were evaluated by immunoblotting. n = 4. (D) Schematic representation of IL-1β promoter luciferase deletion mutations. (E) WT and deleted IL-1β promoter constructs were transiently transfected into NIKs expressing pLXSN or 16E6. After 48 h, cells were harvested and luciferase activity was measured. n = 4. (F) Schematic representation of the IL-1β LILRE site. (G) WT and deleted or mutated IL-1β promoter constructs were transiently transfected into NIKs expressing pLXSN or 16E6. After 48 h, cells were harvested and luciferase activity was measured. n = 4. Data are representative of n independent experiments performed in triplicate. Panel A and B shown are the mean ± SEM with ***, $P < 0.0001$, based on a two way ANOVA test. Panel F, is based on an one way ANOVA test and panel G a paired T test. For immunoblotting data, 1 out of 4 experiments is shown.

https://doi.org/10.1371/journal.ppat.1007158.g004

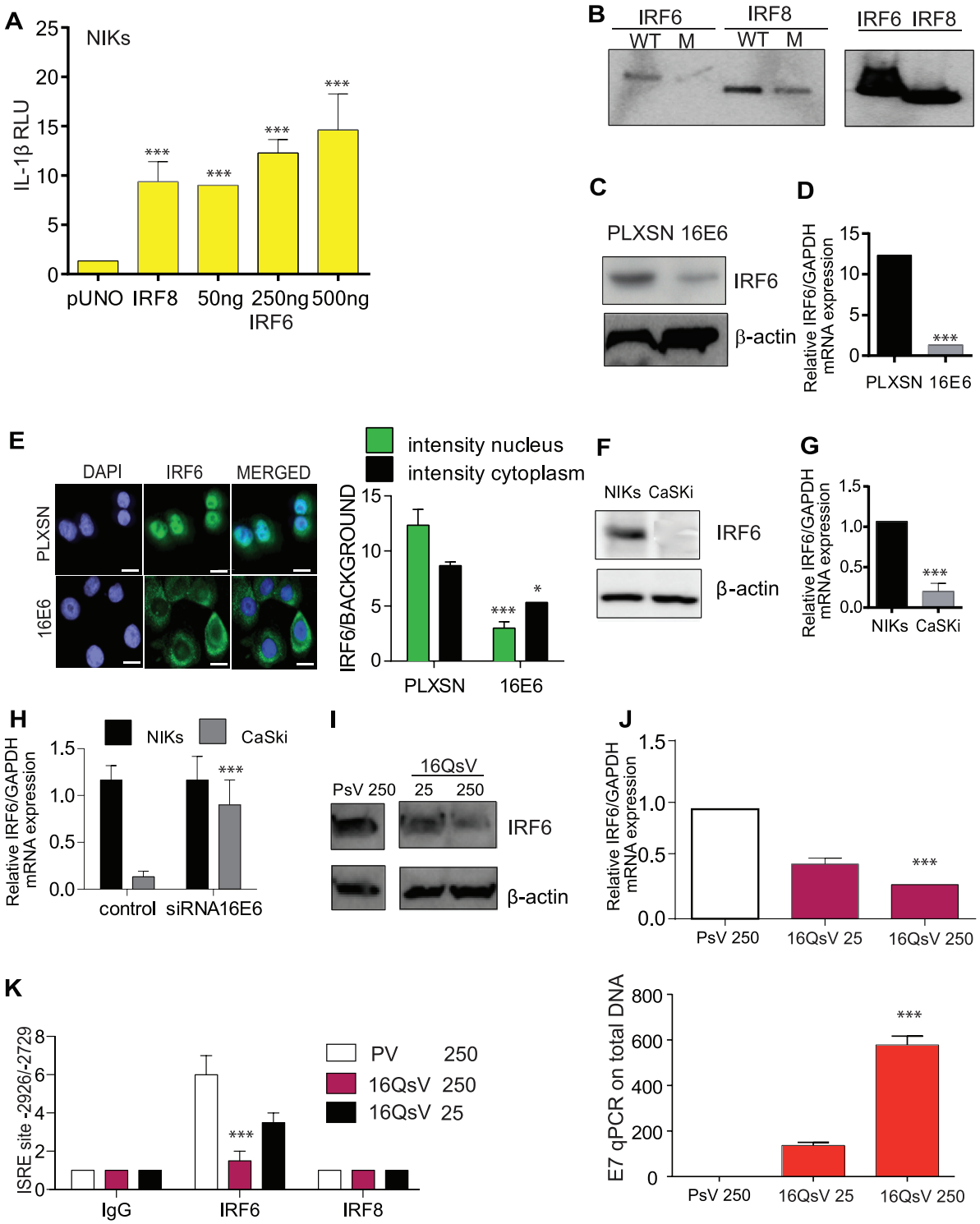


Fig 5. IRF6 and not IRF8 is recruited to the IL-1β promoter which is blocked by HPV16E6. (A) HEK293 cells were co-transfected with IL-1β promoter luciferase construct along with the empty vector pUNO, IRF8 or IRF6 plasmid at the indicated concentration. Post 48 h cells were lysed and luciferase activity measured. n = 4. (B) Oligo pulldown assay for WT or the mutated ISRE site using protein lysates from HEK293 cells transfected with IRF6 or IRF8. Bound proteins were assessed by immunoblotting for IRF8 or IRF6. Input controls (10%). n = 4. (C) Immunoblot analysis of IRF6 protein levels in in pLXSN and 16E6 transduced human primary keratinocytes. n = 4. (D) IRF6 relative levels were measured in

pLXSN, 16E6 and 16E7 transduced human primary keratinocytes by RT-qPCR. n = 4. (E) Immunofluorescent staining of IRF6 in human keratinocytes transduced with pLXSN or HPV16E6. Left, semi-quantative analysis of IRF6 was examined by calculating immunofluorescent intensity. The mean and S.E.M of five fields were plotted. n = 4. (F) Immunoblot analysis of IRF6 protein levels in C33A and NIKs. n = 4. (G) IRF6 mRNA levels detected by RT-qPCR in NIKs and CaSki cells. n = 4. (H) NIKs and CaSki cells were co-transfected with IL-1β promoter luciferase construct ± siRNA for 16E6. Post 48 h cells were lysed and luciferase activity measured. (I) C33A cells were treated with control PsV or 16QsV at different v.g.e per cell for 24 h and IRF6 protein levels were examined by immunoblot. n = 3. (J) C33A cells were treated with control PsV or HPV16 at different v.g.e for 24h and IRF6 mRNA levels were examined by RT-qPCR and (below) viral DNA expression of E7 vs β2-microgloubulin. n = 3. (K), ChIP using IgG, IRF6 or IRF8 antibodies was performed for the ISRE site on C33A cells infected with HPV16 or PsV for 24 h. n = 3. Data are representative of n independent experiments performed in triplicate. Panels A, E, J and K are shown as the mean ± SEM with ***, $P < 0.0001$, * P, < 0,01, based on a two way ANOVA test. Panels D and G are shown as the mean ± SEM with ***, $P < 0.0001$ based on an unpaired T test. For immunoblotting data, 1 out of 4 experiments is shown. For immunoblotting data, 1 out of 4 experiments is shown.

https://doi.org/10.1371/journal.ppat.1007158.g005

alter E6 binding to cellular host proteins [23,24,25,26,27,28] (S4A and S4B Fig and Fig 6A). We then co-transfected increasing amounts of 16E6 WT or mutations with the IL-1β promoter (Fig 6A and 6B). IL-1β luciferase activity was restored with the 16E6F47RdelPBM mutant and partial restored with delPBM and 4C/4S K11E. These data indicated that the 16E6F47R mutation, which fully disrupts its ability to degrade p53 [23,28] can no longer block IL-1β transcription. These data suggest that p53 also regulates IL-1β transcription (Fig 6A and 6B). We next explored the role of p53 on IL-1β transcription. We suppressed p53 expression in primary human keratinocytes using the CRISPR/CAS9 technology (Fig 6C). Suppression of p53 led to a decrease in IL-1β and IRF6 transcription (Fig 6C). Blocking 16E6 mediated E6AP proteosome degradation of p53 using a siRNA for E6AP restored p53 protein levels as well as IL-1β and IRF6 mRNA expression (S4C Fig and Fig 6D). Over expression of p53 restored IL-1β promoter activity in the presence of 16E6 (Fig 6E). In addition, overexpression of p53 or IRF6 expression in keratinocytes transduced with 16E6 also reconstituted pro-IL-1β protein levels (Fig 6E). Taken together, these data show that 16E6 degradation of p53 is required to inhibit IL-1β transcription.

So far we have shown that both IRF6 and/or p53 regulate IL-1β transcription and that both proteins are blocked by 16E6. Whether both proteins independently or dependently control IL-1β transcription remained to be determined. IRF6 transcription was no longer inhibited when cells transiently expressed 16E6 mutations that altered p53 degradation (Fig 7A). Based on these data we hypothesized that p53 regulates IRF6 transcription. Indeed, using the gene card software, we identified a *p53* cis element on the IRF6 promoter. We, therefore, performed ChIP experiments in human primary keratinocytes ±16E6 to determine if p53 was able to bind to the IRF6 promoter (Fig 7B). We observed that p53 bound to the *cis* element on the IRF6 promoter in human keratinocytes (Fig 7C and 7D). Occupation of this site was reduced in 16E6 expressing cells (Fig 7C and 7D). In summary we demonstrated the existence of a negative feedback loop in which 16E6 degradation of p53 prevented the transcription of IRF6 and the subsequent transcription of IL-1β.

IRF6 transcriptional regulation by p53 is lost in cervical neoplasia

Our next approach was to validate our *in vitro* findings in patients with cervical cancer. (HPV16 +). Cervical cancer and matched normal tissue biopsies were taken from 6 patients and snap frozen. After analysis and HPV typing, sections were stained by immunofluorescence for IL-1β as well as p53. Basal cells of the normal epidermis showed strong cytoplasmic staining for IL-1β and nuclear staining for p53 (Fig 8A). No staining for IL-1β and p53 was observed in tumour cells (representative staining in Fig 8A). Quantification of the cytoplasmic staining clearly showed that IL-1β expression was strongly down-regulated in cancerous compared to normal tissue (Fig 8A). We next wanted to determine if IL-1β and IRF6 transcripts

A

	p53	E6AP	PDZ
WT	degrades	binds	binds
F47RdelPBM	likely not to bind or degrade	binds	no binding
delPBM	likely not to degrade	binds	no binding
4C/4S K11E	impaired binding and degrading	binds	binds

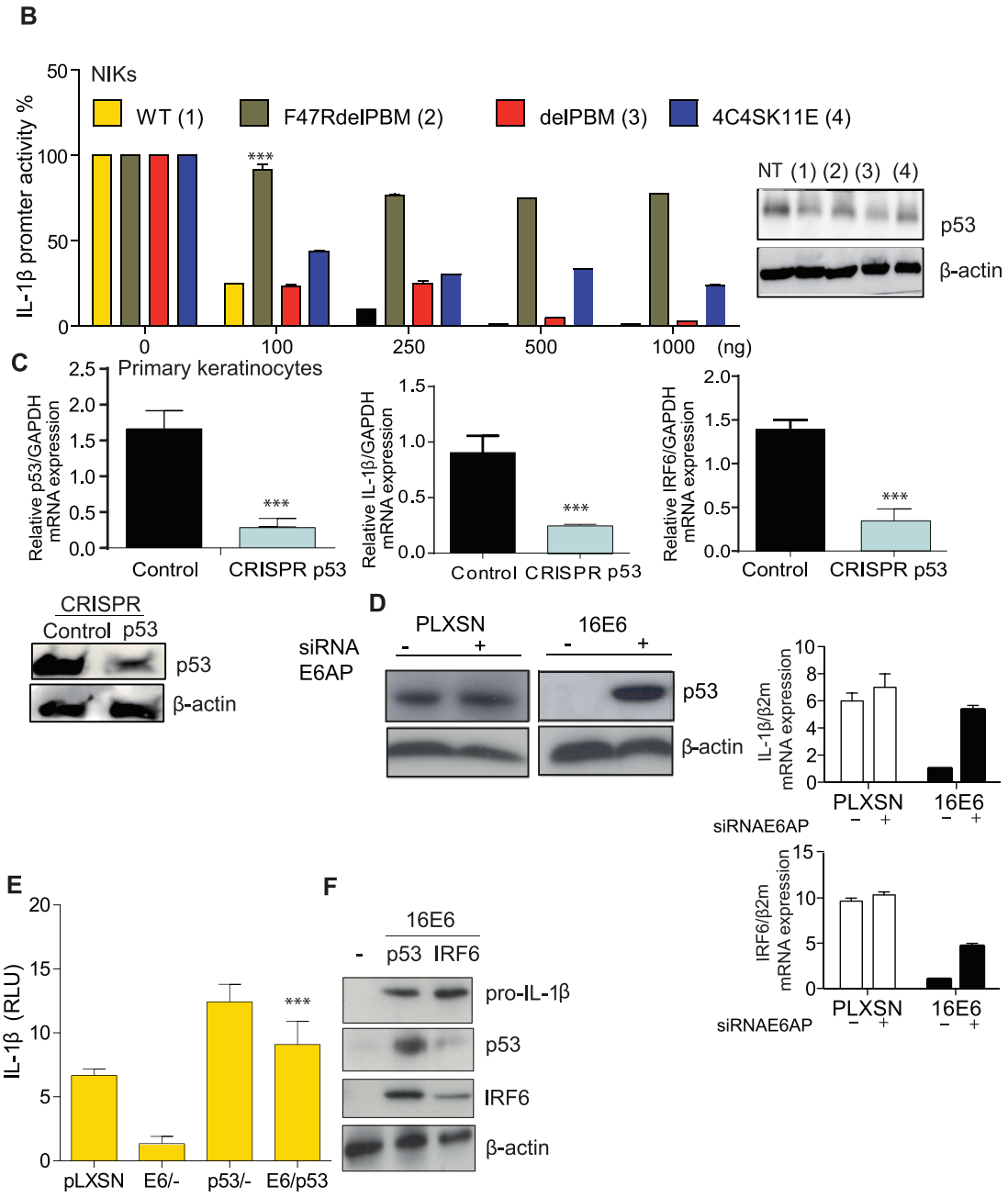


Fig 6. Loss of p53 inhibition by 16E6 restores IRF6 activity. (A) Table defining the 16E6 mutations that alter p53, E6AP or PDZ binding sites. (B) NIKs were co-transfected with IL-1β promoter luciferase construct along with the HPV16E6 WT and mutated constructs at the indicated concentration. n = 5. (C) Human primary keratinocytes were transfected with cas9/sgRNA for p53 or control and 36h later cells were examined for mRNA levels for p53, IL-1β or IRF6. n = 4. (D) siRNAE6AP (+) or siRNA scramble

control (-) was transfected into PLXSN or 16E6 transduced human keratinocytes for 48 h, cells were harvested for protein and RNA. Western blot analysis for p53 and β-actin. Left top, RT-qPCR for IL-1β and, left below IRF6. n = 3, (E) NIKs were co-transfected with the IL-1β promoter and pLXSN, E6, p53 or E6 with p53. Luciferase activity was measured 48 h post infection. n = 4. (F) 16E6 transduced primary keratinocytes were transfected with vector control (-), p53 or IRF6 expression vectors. Twenty-four hours later cells were harvested and pro-IL-1β, p53 or IRF6 levels were examined by immunoblotting. n = 4. Data are representative of n independent experiments performed in triplicate. Shown are the mean ± SEM with ***, $P < 0.0001$, based on an one or two way (applicable to > 2 conditions) ANOVA test. For immunoblotting data, 1 out of 4 experiments is shown.

https://doi.org/10.1371/journal.ppat.1007158.g006

were down regulated in cervical cancer patients. To do this we used a larger cohort from normal (29 patients) and cervical tumor biopsies (29 patients) and RNA was extracted. RT-qPCR of IRF6 and IL-1β transcripts revealed that both genes were reduced in tumor tissues compared to normal biopsies (Fig 8B).

Cervical intraepithelial neoplasia (CIN) is the premalignant abnormal growth of squamous cells on the surface of the cervix. Most cases of CIN remain stable, or are eliminated by the host's immune system without intervention. We next explored if IL-1β and IRF6 transcription were altered in patients during the progression of CIN positive for HPV16. We obtained Formalin-Fixed Paraffin-Embedded biopsies from normal cervical tissues (n = 4) as well as HPV16-positive CINI (n = 8), II (n = 8) and III (n = 5; Fig 8C). Immunohistochemical staining of normal cervical tissue revealed high nuclear expression of IRF6 in the basal layers; which decreased as CIN status increased (Fig 8C and 8D). We observed a decrease in both IRF6 and

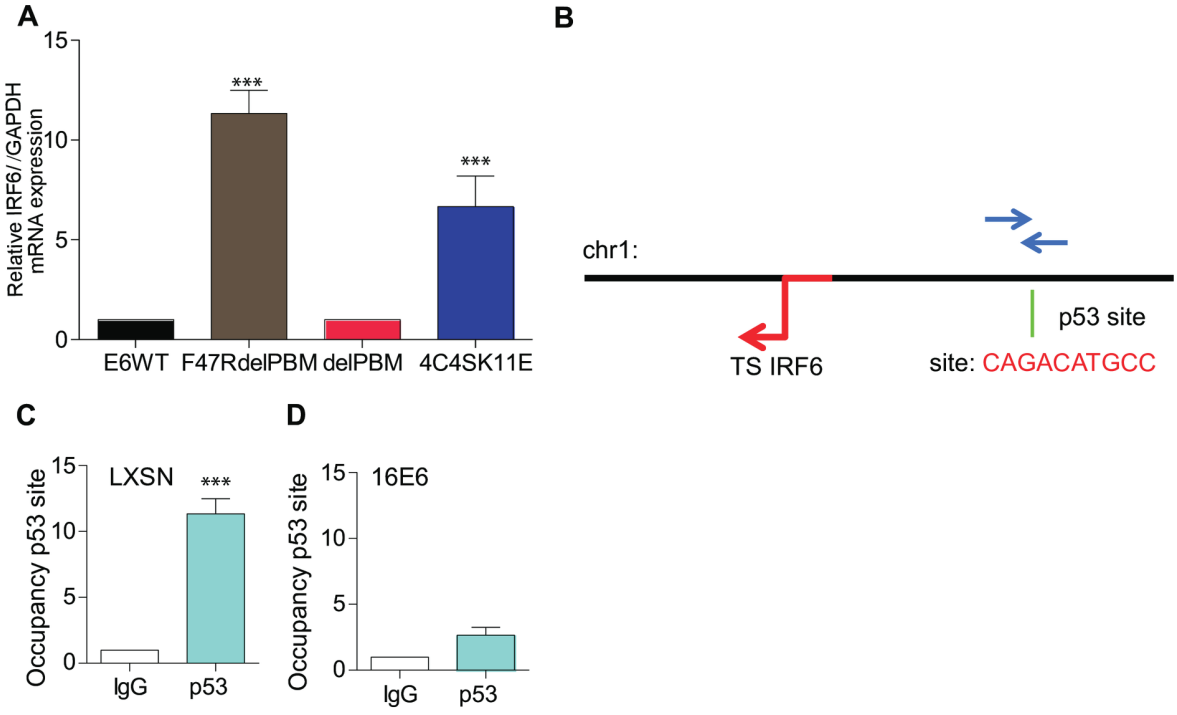


Fig 7. p53 is a transcription factor required for IRF6 transcription in human keratinocytes. (A) HPV16E6WT or mutations were transfected into NIKs and IRF6 mRNA levels were measured by RT-qPCR. Data were normalised to β2-microglpbulin and GAPDH housekeeping genes. n = 4. (B) Schematic diagram of the p53 binding site and sequence on the IRF6 promoter. The red arrow indicates the transcription start site along Chr1. The green line indicated where the p53 cis element is located (written in red). The blue arrows indicate the primer amplication over 200bp. (C, D) ChIP assay of p53 binding on the IRF6 promoter in human primary cells (LXSN) as well as in 16E6 transduced cells. n = 3. Data are the representative of n independent experiments performed in triplicate. Shown are the mean ± SEM with ***, $P < 0.0001$, based on an one way ANOVA test.

https://doi.org/10.1371/journal.ppat.1007158.g007

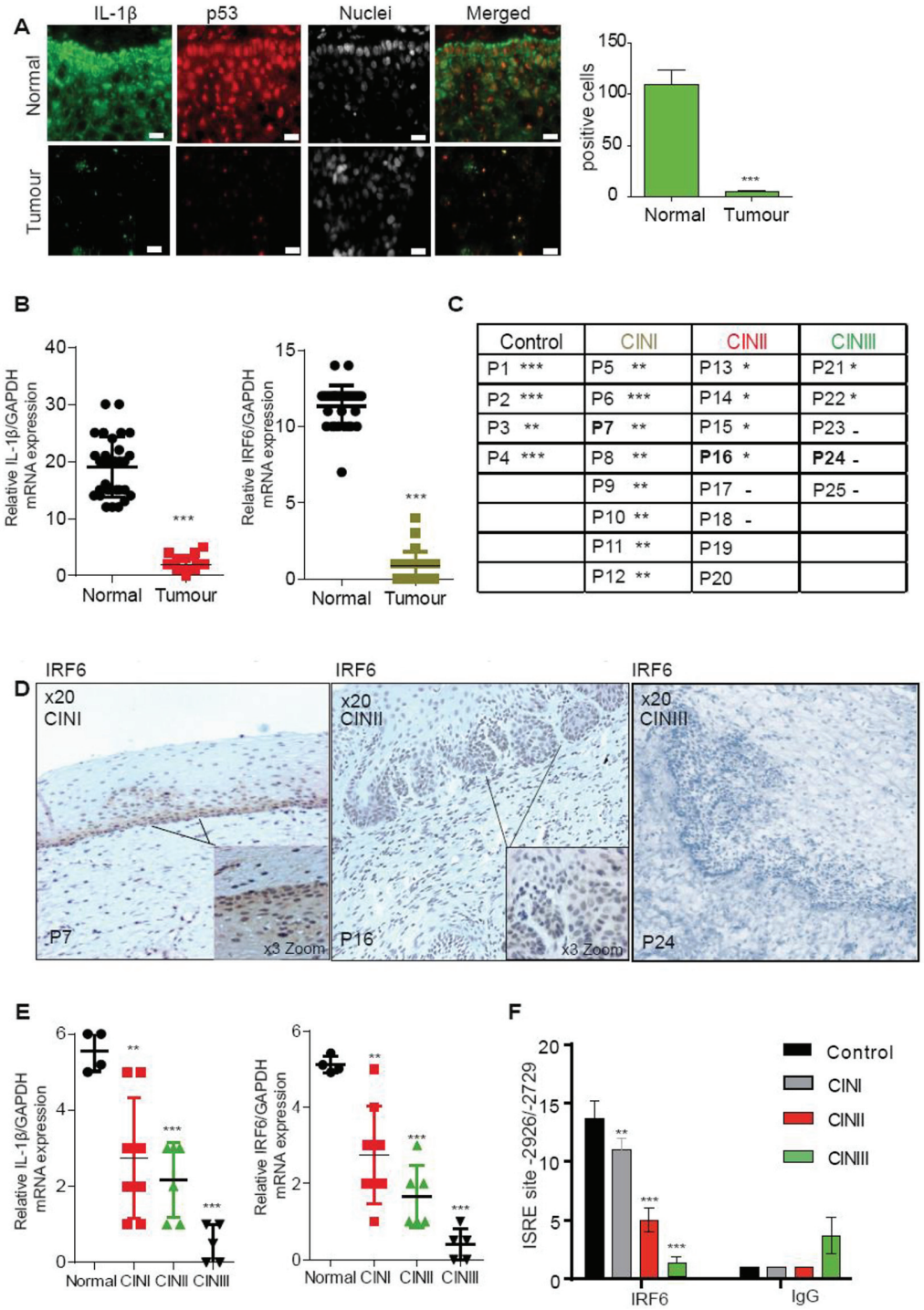


Fig 8. HPV16-positive cervical cancer lesions contain less IRF6 and IL-1β. (A) Immunofluorescence of normal cervical issue and HPV16+ cervical cancer biopsies. IL-1β (green), p53 (red) with trace indicating the basal layer and nucleus (white). Normal (HPV−) and a neoplastic biopsy (HPV16+) from one representative patient out of six with similar results are shown. Bars represent a scale of 10 μm. For each stained biopsy, six fields were examined IL-1β staining was counted manually and the percentage scored out of 100 cells. n = 4. (B) RNA was extracted from normal (29) and cervical cancer biopsies (29). IL-1β relative and IRF6 mRNA levels were measured by RT-qPCR. n = 4. (C) Table of immunohistochemical scoring IRF6 in patients at different stages of cervical neoplasia. Scoring, *** strong, ** medium, * low and—no staining (4 normal, 8 CINI, 8 CINII, 5 CINIII). n = 2. All tissue staining data were examined by two pathologists. (D),

Immunohistochemical staining of IRF6 in cervical tissue in patients with CINI, II or III. n = 2. (E) RT-qPCR of IL-1β and IRF6 mRNA expression levels in normal vs neoplastic cervical tissue. n = 3. **A p53 site is required to bind to the IRF6 promoter but is lost in cervical cancer tissues**. (F) ChIP analysis was performed on normal and cervical neoplastic tissue for IRF6 binding on ISRE site on the IL-1β promoter. n = 3.

https://doi.org/10.1371/journal.ppat.1007158.g008

IL-1β mRNA during disease progression, (Fig 8E). ChIP experiments using chromatin extracted from patient tissue revealed that IRF6 binding to the ISRE site on the IL-1β promoter decreased during CIN severity (Fig 8F), indicating that loss of IRF6 inversely correlates with cervical neoplasia progression. Furthermore, p53 binding was observed in normal cervical biopsies, but binding was reduced in patients with cervical tumors (Fig 9). These data strongly suggest that the p53/IRF6 regulation of IL-1β transcription is lost during CIN disease stages that could lead to cervical cancer.

Discussion

We showed that human keratinocytes produce IL-1β when exposed to 16QsV. Furthermore, addition of recombinant IL-1β on 16QsV infected keratinocytes led to a block in viral gene transcription. Viral gene transcription was restored in the presence of an antagonist for the IL-1 receptor. More importantly we delineated that IL-1β gene transcription increased when exposed to 16QsV. These data show that HPV16 stimulates IL-1β secretion that has an anti-viral effect on infected cells. IL-1β depends on inflammasome activation; we have data showing that 16QsV was not sensed by NLPR3 or AIM2 (S5A Fig). Bone marrow derived macrophages from NLRP3 and AIM2 knock out mice were still able to produce IL-1β in the presence of 16QsV (S5A Fig). Therefore we still need to elucidate which innate-inflammasome sensor can detect 16QsV. However IL-1β gene expression began to decrease post 8h infection with 16QsV. These data implicate that HPV16 has developed an escape mechanism to block IL-1β production. Our findings are summarised in Fig 10.

Characterizing how HPV blocks immune surveillance is central in understanding the events involved in the establishment of head and neck as well as cervical cancers. In this study we showed the loss of IL-1β transcription was mediated mainly by oncoprotein 16E6. Our data

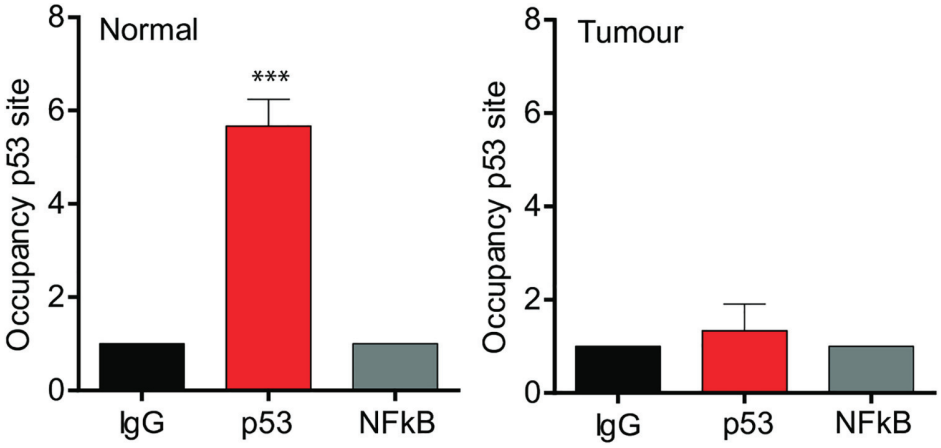


Fig 9. Four normal and tumor biopsies from (A) were used to perform ChIP analysis for p53 binding onto the IRF6 promoter. n = 3. Data are representative of n independent experiments performed in triplicate. Shown are the mean ± SEM with ***, $P < 0.0001$, **, $P < 0.0005$ based on a paired Student's t test (when patient matched) or unpaired when not patient matched.

https://doi.org/10.1371/journal.ppat.1007158.g009

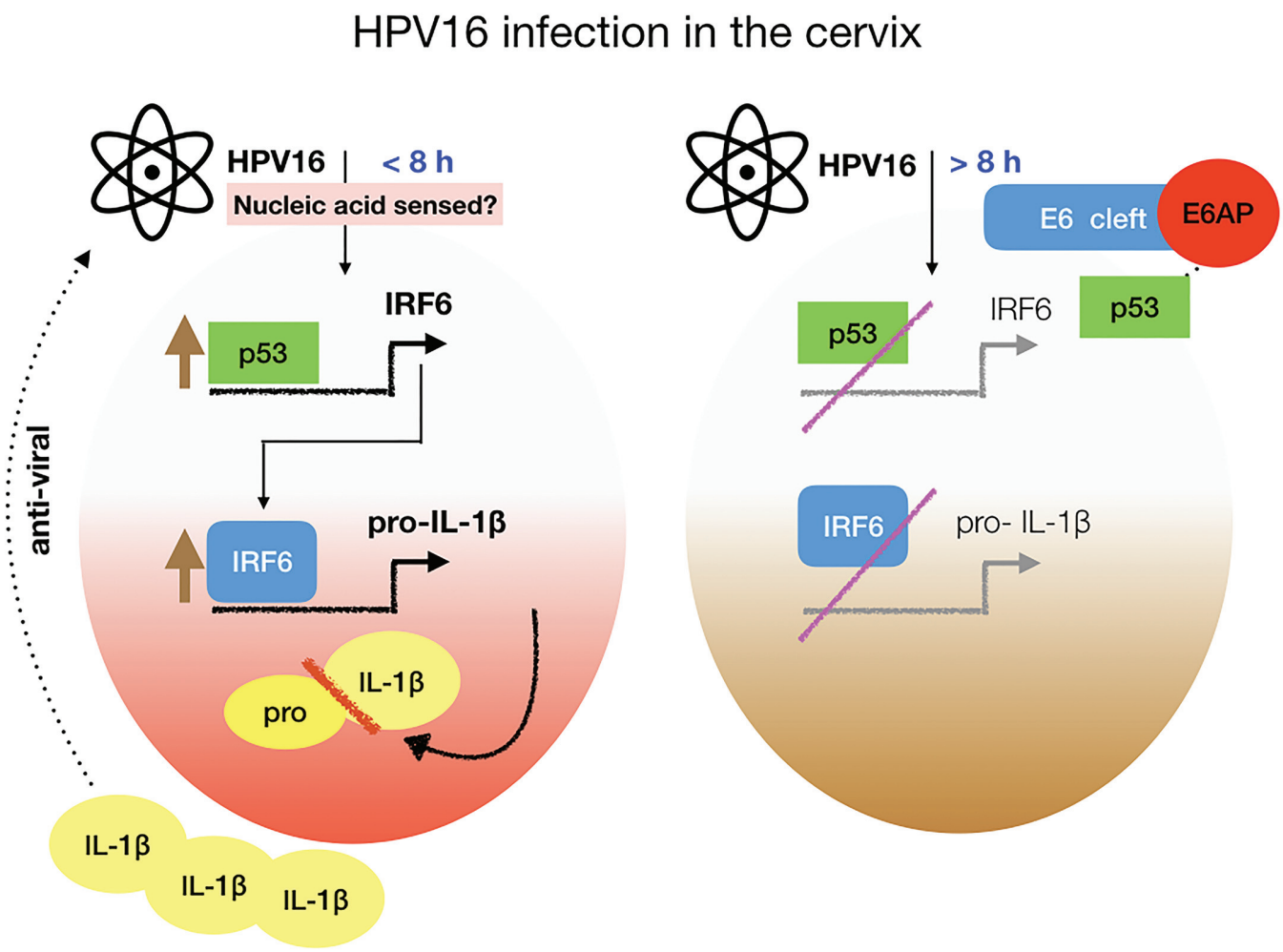


Fig 10. Model representing the induction and inhibition of IL-1β by HPV16. Infection of the basal keratinocytes with HPV16 induces inflammasome dependent IL-1β production sensed by an unknown innate receptor. p53 transcriptional regulation of IRF6 is increased, which we show drives IL-1β transcription. The pro form of IL-1β is cleaved by caspase 1 (red bar). The active form of IL-1β can block the increase in viral copies. However, when the oncoprotein E6 is expressed this drives p53 degradation by E6AP preventing IRF6 and consequently IL-1β transcription. This mechanism of viral inhibition of innate responses may contribute to HPV16 persistence in the host.

https://doi.org/10.1371/journal.ppat.1007158.g010

are in line with Karim et al., showing that IL-1β mRNA levels were decreased in epithelial cells expressing 16E6E7 [29]. The addition of 16QsV, or expression of E6 alone, blocked IRF6 expression and binding to the ISRE site on the IL-1β promoter. IRF6 has previously been shown to play an important role in the embryonal development of the craniofacial region. Mutations in this gene have been found in two human syndromes: Van der Woude and Popliteal Pterygium Syndrome, which are characterized by the cleft palate, lip pits, skin webbings, syndactyly, genital deformities and oral adhesions. In contrast to most IRFs shown to be essential in IFN gene regulation, IRF6 had no identified function in innate immune gene activation. Other IRF family members have been shown to be hijacked during HPV mediated carcinogenesis, such as IRF1 [30]. We demonstrated that mutation of the ISRE site on the IL-1β promoter prevented 16E6 to inhibit IL-1β promotor activity. Gene silencing of the viral oncoproteins 16E6E7 or 16E6, restored IRF6 and IL-1β expression in human keratinocytes. This was shown

by calculating the percentage of 16E6 inhibition against the cells that are induced with the PLXSN vector alone (S5B Fig). Furthermore we showed that p53 regulated IRF6 transcription.

Using 16E6 mutations that cannot, partially or fully degrade p53 allowed us to correlate the degradation of p53 by 16E6 led to the loss of IRF6 transcription (Fig 10). p53 has also been shown to amplify intracellular IFN responses. IFN-stimulated genes (ISG) promoters do not contain p53 consensus binding sites. However Munoz-Fontela et al., identified IFN regulatory factor 9 (IRF9), a component of the ISG factor 3 (ISGF3) complex, as a p53 target gene. ISGF3 directly induces the expression of ISRE-containing genes and could represent a mechanistic link between p53 and ISG induction [31]. Several additional IFN-stimulated mediators of ISG expression, including IFN regulatory factor 5 (IRF5), immune-stimulated gene 15 (ISG15) and the Toll-like receptor 3 (TLR3), have been identified as direct p53 target genes. Therefore IRFs and p53 play a central role in regulating innate immune responses.

To our knowledge, this is the first description of p53-IRF6 axis mediating differential regulation of an immune gene. Our ChIP experiments showed that lack of p53 protein due to 16E6 prevented its recruitment to the IRF6 promoter in cervical cancer patients. Based on our findings we hypothesized that loss of IRF6 and IL-1β expression favours cervical cancer development. These data were corroborated in cervical neoplasia and tumours. In cohorts of cervical neoplastic patients we observed a decrease in both IL-1β and IRF6 mRNA levels. Rotondo et al., evaluated the gene expression changes involved in neoplastic progression of cervical intraepithelial neoplasia compared to normal keratinocytes [32]. Microarray analysis revealed that IRF6 was one of the 24 genes significantly down regulated during CIN progression [32]. Furthermore two independent studies showed that IRF6 gene mutations were associated to head and neck squamous cell carcinomas [33] [34]. However these scientific findings conflict with two other data sets. Our analysis of the data set by den Boon et al., showed that IL-1β was not affected during cervical cancer progression [35]. Also neither IRF6 nor IL-1β mRNA levels were suppressed when analysing the data set from the TCGA cervical carcinoma cohort [36]. One should consider that neither studies were hypothesis driven nor were the data sets designed to examine the mechanism of HPV16E6 regulation on p53/IRF6/IL-1β. We validated that IRF6 and IL-1β expression were altered by the viral oncoprotein 16E6 using several read-outs and models. Furthermore, we showed that IL-8 gene transcription depends on IL-1β stimulation. An increase in local cervical IL-8 levels correlates with HPV viral clearance [37]. Experts in HPV incidence have discussed that infection of the cervical epithelium is a prerequisite for the development of cervical cancer and the local immune response is an important determinant of progression and disease outcome [38]. The transiency of most HPV infections and the observed regression of certain cervical intraepithelial neoplasia lesions to normal epithelium suggest a change in local immune responses, which may be caused by differences in host genomics. We observed that loss of IL-1β production in cervical cancer cells led to a loss of paracrine IL-8 transcription. Furthermore, IL-1β down regulation in HPV induced carcinogenesis is underlined by the fact that specific polymorphisms in IL-1β have been demonstrated to be associated with cervical carcinoma risk [38].

The work of Niebler et al., showed that 16E6 alters IL-1β by proteosome degradation of the pro-form [8]. We did not observe the same findings using our cellular models. This could be due to the fact that the primary keratinocytes used by Niebler et al., were from neonatal foreskin, whereas our model used keratinocytes from adult female skin (see Method and materials). Yet, Niebler et al., also showed in Fig 6 of their article a drop in mRNA IL-1β levels in CIN patients [8]. These data fall in line with our findings.

We propose that inflammasome activation of IL-1β secretion favors' HPV viral clearance. Loss of IRF6 and IL-1β function during cervical neoplastic stages reflects a prognostic read out towards cancer development. Thus, interfering with the regulation of IL-1β with synthetic

agonists that target p53 and IRF6 levels may provide a novel therapeutic strategy for cervical cancer patients.

Methods and materials

Cell culture

Cervical cancer cell lines C33A (HPV negative cat: HTB-31), SiHa (HPV16 positive cat: HTB-35), CaSki (HPV16 positive cat: CRL-1550), HeLa (HPV18 positive cat: CCL-2) and Human embryonic kidney 293 (HEK293 cat: CRL-1573) cells were purchased from American Type Culture Collection (Manassas, VA) and cultured in DMEM medium (Life technologies), supplemented with 10% foetal bovine serum (FBS), L-glutamine, pyruvate and 0.1% ciprofloxacin (Euromedex). HEK293TT cells were a kind gift from the lab of Dr. Pawlita (DKFZ, Germany). Cells were cultured with hygromycin using the same culture medium as HEK293. When preparing HEK293TT cells for transfection cells were grown without hygromycin and antibiotics. Cells were cultured at 37˚C with 5% CO_2. Immortalized near-diploid human keratinocyte cell line (NIKS, kind gift from Professor John Doorbar, University of Cambridge, UK) and Human Primary Keratinocytes produced by the lab of Massimo Tommasino were from Adult female, or femaile skin keraintocytes were purchased from American Type Culture Collection Cat: PCS-200-011). Cells were cultured as previously described [6]. Human Primary Keratinocytes were cultivated at low passages numbers for a period of 3 weeks (called keratinocytes after 1 passage). High-titer retroviral supernatants ($>5 \times 10^6$ IU/ml) were generated as previously described [39]. The 16QsV and PV production, infection, and viral genome expression quantification of HPV16 are described below.

Agonists and antagonists

NLRP3 ligand Nigericin was used at 1μg/mL (Sigma), AIM2 ligand poly(dA:dT) was used at 1μg/well (Invivogen) and transfected using lipofectamine 2000 (Invitrogen). ANAKINRA (Biovitrum) was used at 200μg/ml.

Oligo pull-down

Oligo pulldown was performed as previously described [40] with cellular extracts as stated in the figure legend and oligo probes as listed in Table 1. IRF8 and IRF6 antibodies were purchased from Cell Signaling.

Table 1. Oligo sequences.

Deletions IL-1β	*OLIGO PULL DOWN IL-1β*
FWD 1 CTAGCTAGCTCTAGACCAGGGA	FWD Biotin' TTTGACATAAGAGGTTTCACTTCC
FWD 2 CTAGCTAGCTAAGAGGTTTCACT	REV GGAAGTGAAACCTCTTATGTCAAA
FWD 3 CTAGCTAGCCTCCAGCCTGGGG	*IRF6 qPCR*
FWD 4 CTAGCTAGCCCTGAATGTACATGCC	FWD GGCATAGCCCTCAACAAGAA
FWD 5 CTAGCTAGCTTAGGCAGAGCTCAT	REV CACCCCTTCCTGGTACTTCC
REV 1 GAAGATCTAAGAGGTTGGTA	*IRF8 qPCR*
REV 2 GAAGATCTAAGAGGTTTG	FWD ACGAGGTTACGCTGTGCTTT
	REV GACATCTCGGCAGGGCTATG
	p53 qPCR
	FWD GGTTTGTAATGCAGGGCTGAGG
	REV GGGTATGGTGGTGTATGCCTGT

https://doi.org/10.1371/journal.ppat.1007158.t001

ChIP

ChIP assays were performed using the Shearing Optimization kit and the OneDay ChIP kit (Diagenode). For C33A cells or primary keratinocytes, cell sonication cycles last 15s with 5s on and 2 s off at 20% of amplitude and were repeated four times. For tissue, immunoprecipitation was performed overnight on a rotating wheel at 4˚C. 2.5 μl/reaction of DNA solution was used for qPCR. The primers used to amplify IL-1β, or IRF6 binding regions are available on request. ChIP on the tissue was performed according to the protocol from Epigenome Network of Excellence for tissue preparation after the Red ChIP kit from diagenode was used to prepare chromatin and the 1-d ChIP kit for the immunoprecipitation. Immunoprecipitation was performed overnight on a rotating wheel at 4˚C. 2.5 μl/reaction of DNA solution was used for qPCR.

Plasmid constructs

The constructs pLXSN empty, pLXSN-16E6E7, pLXSN-HPV16E6, pLXSN-HPV16E7 and pLXSN-HPV18E6E7 were obtained from M.Tommasino (IARC, Lyon, France) (6). The pGL3 Luc vector was purchased from Promega. The constructs The full-length IL-1β-Luc, LILRE (IL-1 response element) and mutants were obtained from Philip E.Auron (University of Pittsburgh, Pittsburgh, PA 15261, USA). IL-1β deletions were cloned using the primers listed in Table 1. Nine E6 mutations were obtained from Dr Gilles Trave (CNRS, Illkirch, France); and previously described. These mutations were cloned into the pX5 plasmid. The retroviral pBabe-puro encoding HPV16 and 6 E6 and or E7 have been previously described [41]. The constructs pLXSN-HPV16 E6, HPV18 and HPV38 E6 and HPV6 E6 were a gift from D. Galloway (Fred Hutchinson Cancer Research Center, Seattle, WA). The plasmids used for HPV16 structural genes and control PsV production, the target HPV16 genome, and GFP (for PsV control) were kindly donated from the laboratories of Martin Muller and Angel Alonso (DKFZ, Germany). pUNO, human IRF6 and IRF8 constructs were purchased from Invivogen. The p53 plasmid was obtained from Addgene. siRNA for 16E6E7 and E6 was purchased from Dharmacon and Sigma respectively. siRNA for E6AP [42]CRISPR for p53 was purchased from Santa Cruz.

Viral production 16QsV and PsV [43],[44]

16QsV are viral particles that contain the full viral genome of HPV16 encaspidated by the viral late proteins L1 and L2. PsV contain GFP DNA encaspidated by L1 and L2.

293TT cells at 75% confluency the day of transfection. The transfection mix consisted of 13μg of the L1-L2 expression vector and ~ the same amount of HPV16DNA or GFP control vector were prepared in a separate tube, a mix 85μl of Lipofectamine with 2ml OptiMEM. Both mixtures were incubated separately at RT for 10′-30′, then combined and incubated for at least another 20 minutes. The resulting lipid/DNA complexes were directly added to the pre-plated cells. The cells were incubated with the transfection mix for 4–6 h then split 1:2 or 1:3 and left overnight. The next day cells were detached, spun down and the supernatant discarded. *Cell lysis and Capsid Maturation*: Using a 5ml plastic pipet, cells were suspended in 0.5ml in DPBS-Mg and transferred to a siliconized 2ml tube, screw-capped (Nalgene tubes for freezing cells). For 100 million cells 1 ml of lysis buffer was prepared and incubated for 1-2h at 37C then with inversion for a further 16 h at least at 37C.

Salt extraction

The next day optiprep gradients prepared were diffused for 4 hours. The lysate was then layered on top of the gradient. The tubes were spun for using 13.2ml tubes SW40.1 Ti 14 h at 16

C. The L1 band is a visible as a slight grey layer a little over a third of the gradient. Using a large needle and a 5ml syringe we removed the 60% cushion layer, then we tool a 1.0ml syringe and 26 gauge needle to extract 250μl fractions (6–8 fractions). Each fraction was placed into a screw cap tube (not freezing tubes). *Screen fractions by SDS PAGE*: A mini gel of 10% were used to screen fractions for the presence of the L1 protein (55kDa) fractions with significant amounts of L1 were pooled, aliquoted and frozen; the protein yield can be estimated through BSA standards or BCA assay. *Analysis of virions- Encapsidated DNA*: Fifty μl of fractions were run on a 0,8% agarose gel. Supercoiled DNA from the HPV genome, linear human DNA with nucleases and exonuclease treatment captured by L1 and L2 will run at 8Kb. *nuclease should cut up all the human genomic DNA, then any tailed DNA that gets incorporated into the capsid were cut off with the exonucleases. *Capsid protein levels*: Capsid protein levels (20μl fractions) were measured on 10% SDS-PAGE and silver staining with serially diluted BSA as concentration standard or by western blotting for L1. Viral genome equivalents were measured by qPCR on the viral DNA of infected HEK293T cells using W-12 cell lysates as a standard (kind gift from Dr Franck. Stubenrauch, Forschungssektion Experimentelle Virologie, Tubingen, Germany).

Ethics statement

Our cohort of normal, CIN and tumor samples was provided by the hospital in Lyon Sud, Lyon, France. Samples were obtained with written informed consent from each patient with the procedure approved by the local Ethics Committee, Comités de Protection des Personnes. All, normal, CIN or tumor biopsies were from females aged between 30–50 years. Where available the same normal patient-matched samples were provided (HPV negative genotyped using multiplex PCR with HPV type-specific primers). Biopsies were either snap frozen or FPPE.

Genotyping

CIN and Tumor samples were genotyped using multiplex PCR with HPV type-specific primers for amplification of viral DNA and array primer extension for typing [41].

Infectivity and viral gene transcription assay

NIKs or primary keratinocytes were infected with packaged viruses as stated in the figure legends at 37C. Cells were removed, and RNA extracted for RT-PCR for E1, E6 and E7 transcripts (mRNA) or DNA to measure viral DNA expression for E7 [6].

Immunofluorescence and immunohistochemistry

Keratinocytes transduced with pLXSN or HPV16E6 were fixed as previously described [45]. Sections of 5-μm thickness were cut and either stained for immunofluorescence using the TSA system (PerkinElmer). The p53 antibody was purchased from Cell Signaling and the anti-IL1β 3ZD (kindly provided by Dr. Trinchieri, NCI). The IRF6 antibody (F12) was purchased from Santa Cruz. Cells or tissues were washed, the coverslips were mounted onto slides using a 1/10 dilution of 4′,6′-diamidino-2-phenylindole (nuclear stain; Invitrogen) in fluoromount (Southern Biotechnology Associates), and protein expression was detected by direct fluorescence microscopy. Photographs were taken at magnification x40 using the Zeiss confocal 710 microscope. Semi-quantitative analysis of IRF6 levels was estimated using the ImageJ software. Immunohistochemistry staining for IRF6 was performed as previously described [6].

ELISA

NIK, primary keratinocytes, HPV16E6 and E7 induced keratinocytes were seeded into a six-well plate with 2.5 x10^5 cells per well with 4x10^3 NIH 3T3 feeders. Two days later the feeders were removed, and the medium was replaced. After two hours; keratinocytes were stimulated either with 20μM of Nigericin (Sigma) or transfected with 1ug/mL of poly (dA:dT) (Invivogen) using lipofectamine 2000 (Invitrogen). After the indicated period (see figure legend), the supernatant was harvested and quantified for IL-1β by ELISA (Bender Med System) or IL-18 [46].

Luciferase assay

Twenty-four hours before transfection HEK293 cells were plated at 20% of confluency in 96 well plates with 180μl of complete medium per well. Cells were transfected using GeneJuice Transfection Reagent (Novagen) following the manufacturer's instructions. Cells were transiently co-transfected with HPV constructs as indicated with pGL3-LILRE, mutants or pGL3-XTLuc. A Renilla plasmid with a CMV promoter was used to normalize transfection efficiency. Twenty-four hours after transfection cells were lysed at room temperature in passive lysis buffer (Promega) for 20 minutes. Luciferase buffer was composed of MgSO4 (2,67mM), EDTA pH8 (0.1 mM), DTT (33.3 mM), ATP (0.53 mM), acetyl-CoA (207 μg/ml), luciferin (0.13 mg/ml), Magnesium carbonate hydroxide (0,265 mM) and tricine (20 mM). Renilla buffer was made by diluting coelenterazine. Luciferase and renilla activity from transfected cells were measured using a luminoskan Ascent (Thermo). A single read program with an integration time of 1000 ms was used. Firefly luciferase (*Photinuspyralis*) activity of individual cell lysates was normalized against renilla (*Renillareniformis*) activity to correct for transfection efficiency in each reaction.

IL-8 bioassay

Supernatants from stimulated cells were added onto HEK 293 cells transfected with IL-8 luciferase promoter, and a Renilla plasmid with a CMV promoter was used to normalise transfection efficiency [47]. Twenty-four post stimulation cells were processed as listed above.

Protein/RNA extraction

Cells were preserved in RP1 lysis buffer complemented with β-mercaptoethanol (1%) until RNA and total proteins extraction using the NucleoSpin RNA/protein extraction kit (Macherey-Nagel). Supernatants from stimulated cells where concentrated using MeOH/chloroform. All RNA samples were treated with DNAse before reverse transcription was performed.

Western blot analysis

Eighteen μg of total cellular protein were incubated during 5 minutes at 95˚C. The protein samples were separated by electrophoresis using Novex 4–20% Tris-Glycine gels (Life Technologies) for 1 hour at 100V. Proteins then were transferred on a PVDF membrane (PerkinElmer) during 1 hour at 100V. After blocking with PBS 0.1% tween and 5% milk for 1 hour, membranes were probed with the following primary antibodies: anti-caspase 1 P10 (SantaCruz Biotechnology), anti-IL1β 3ZD (kindly provided by Dr Trinchieri, NCI), anti-ASC (Santa Cruz Biotechnology), 16E6 (provided by the lab of Dr Trave (GBMC, France) and 16E7 (Santa Cruz, France) over night at 4˚C. β-actin (Sigma) primary antibodies were added for 2h at RT. After three PBS 0.1% tween washes, secondary antibodies are added for two hours at RT. Anti-Rabbit and anti-mouse HRP conjugate secondary antibodies were provided by Promega.

Proteins were revealed with Lumiglo chemiluminescent substrate system (Kpl). Western blots were developed using the intelligent dark box (Fuji film).

RT qPCR

We retro transcribed (RT) 1–1.5 μg of RNA extracted from cells using first strand RT-PCR kit with oligodT primers (Fermentas). The RT reaction was diluted according to detection sensitivity. One μl of the diluted samples was added to a 20 μl PCR mixture containing 0.4 μl of primers forward and reverse (10 μM) and 10 μl of Master Mix. Mx300P real-time PCR system (Stratagene, La Jolla, CA) were used to performed qPCR with Mesa Green qPCR Master Mix Plus (Eurogentec) on CaSki, C33A and SiHa cells. Primer sequences designed to detect gene expression of AIM2, NLRP3, ASC, IL-1β, house-keeping β2-microglubulin and GADPDH are listed as previously described [46]. As relative levels of house-keeping genes between samples did not alter, data were plotted against GAPDH. Primers for IRF6, IRF8, and p53 are listed in Table 1.

Statisical tests

Where appropriate, anova, unpaired or paired T test were performed using prism software version 6 (Graph Pad) Statistical studies were validated by Omran Allatif (Statistician CIRI, Lyon, France[46,48]).

Supporting information

S1 Fig. IL-1β production by primary human keratinocytes. A: IL-1β was measured by ELISA in human keratinocytes (pLXSN) in response to NLRP3 or AIM2 ligands. n = 10. B IL-1β was measured by ELISA in human keratinocytes in response to PV or 16QsV at 250 v.g.e ± Glybride (inhibits ATP mediated proton pump) or ± Caspase-1 inhibitor. C IL-1β was measured a 4h and 8h by ELISA in human keratinocytes (pLXSN) in response to NLRP3, AIM2 ligands or 16QsV (left Y axis) or LDH release (right Y axis) using the Pierce ™ LDH kit (Thermofisher). n = 4. Shown are the mean ± SEM with ***, $P < 0.0001$, based on a two way ANOVA test.
(TIF)

S2 Fig. 16E6E7 has no effect on the inflammasome activation of caspase-1. (A) RNA was extracted from Human keratinocytes ± 16E6E7 and NLRP3 or AIM2 relative expression was determined by RT-qPCR. n = 5. (B) Immunoblot analysis of keratinocytes transduced with LXSN or 16E6E7 were transfected with NLRP3-CFP or AIM2-CFP. Membranes were probed for GFP, p53 or β-actin n = 5. (C) RNA was extracted from human keratinocytes ± 16E6E7 and ASC or caspase-1 relative expression was determined by RT-qPCR. n = 5. (D) Human keratinocytes ± HPV16E6E7 were stimulated with AIM2 and NLPR3 ligands and both pro-or mature caspase-1 were analysed in cell lysates or in the supernatant by immunoblotting. β-actin was used as a loading control. n = 3.
(TIF)

S3 Fig. Other HPV HR types but not LR blocks IL-1β promoter activity. (A) NIKs were co-transfected with the IL-1β promoter with pLXSN, 16E6, 18E6, 31E6 or 6E6 as indicated. After 48 h, cells were harvested and luciferase activity was measured. n = 5. **IRF8 is not involved in IL-1β transcription in human keratinocytes**. (B) IRF8 relative levels were measured in pLXSN, 16E6 and 16E7 transduced human primary keratinocytes by RT-qPCR. n = 4. Immunoblot analysis of IRF8 protein levels in in pLXSN, 16E6 and 16E7 transduced human primary keratinocytes. n = 4. (C) ChIP assay of IRF8 binding on the IL-1β promoter in human primary

cells (LXSN) as well as in human macrophages. n = 4.
(TIF)

S4 Fig. Mutations in 16E6 restore IL-1β activity. (A) Table describing 16E6 mutations. NIKs were transfected with 16E6Wt and mutations were co-transfected with IL-1β promoter luciferase construct. Forty-eight hours post transfection cells were lysed and luciferase activity measured. n = 4. (B) NIKs were transfected with WT and mutations for 16E6. Forty-eight hours post transfection proteins were probed using 16E6 antibody. n = 3. (C) Western blot to control E6AP knock down by control and SiRNA E6AP, using β-actin as a loading control. n = 4. Data are representative of n independent experiments; graphs shown are the mean ± SEM from triplicate values.
(TIF)

S5 Fig. (A) 16QsV activates IL-1β production independently of AIM2 and NLRP3. Bone marrow derived macrophages from C56BL/6 WT, AIM2-/-, ASC -/-, Caspase 1 -/- (from Thomas Henry, France) and NLRP3 mice (From Virginie Petrilli, France) were isolated and cultivated as previously described [49]. (B) Percentage of IL-1β promoter inhibition with PLXSN cells vs 16E6 transfected with the IL-1β point mutation or LILRE deletion.
(TIF)

Author Contributions

Conceptualization: Peggy Parroche, Uzma Ayesha Hasan.

Data curation: François Briat, Claudia Zannetti, Marie Marotel, Nadege Goutagny, Christine Carreira, Uzma Ayesha Hasan.

Formal analysis: Peggy Parroche, Alexis Robitaille, Gilles Trave, Uzma Ayesha Hasan.

Funding acquisition: Uzma Ayesha Hasan.

Investigation: Peggy Parroche, Thomas Henry, Uzma Ayesha Hasan.

Methodology: Michelle Ainouze, Pauline Rochefort, Peggy Parroche, Guillaume Roblot, Issam Tout, François Briat, Massimo Tommasino, Thomas Henry, Uzma Ayesha Hasan.

Project administration: Uzma Ayesha Hasan.

Resources: Philip Auron, Alexandra Traverse-Glehen, Aude Lunel-Potencier, Francois Golfier, Murielle Masson, Katia Zanier, Gilles Trave, Uzma Ayesha Hasan.

Software: Uzma Ayesha Hasan.

Supervision: Uzma Ayesha Hasan.

Validation: Uzma Ayesha Hasan.

Visualization: Uzma Ayesha Hasan.

Writing – original draft: Peggy Parroche, Uzma Ayesha Hasan.

Writing – review & editing: Pauline Rochefort, Massimo Tommasino, Thierry Walzer, Gilles Trave, Uzma Ayesha Hasan.

References

1. de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, et al. (2012) Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. Lancet Oncol 13: 607–615. https://doi.org/10.1016/S1470-2045(12)70137-7 PMID: 22575588

2. Bouvard V, Baan R, Straif K, Grosse Y, Secretan B, et al. (2009) A review of human carcinogens—Part B: biological agents. Lancet Oncol 10: 321–322. PMID: 19350698
3. Amador-Molina A, Hernandez-Valencia JF, Lamoyi E, Contreras-Paredes A, Lizano M (2013) Role of innate immunity against human papillomavirus (HPV) infections and effect of adjuvants in promoting specific immune response. Viruses 5: 2624–2642. https://doi.org/10.3390/v5112624 PMID: 24169630
4. Lau L, Gray EE, Brunette RL, Stetson DB (2015) DNA tumor virus oncogenes antagonize the cGAS-STING DNA-sensing pathway. Science 350: 568–571. https://doi.org/10.1126/science.aab3291 PMID: 26405230
5. Reiser J, Hurst J, Voges M, Krauss P, Munch P, et al. (2011) High-risk human papillomaviruses repress constitutive kappa interferon transcription via E6 to prevent pathogen recognition receptor and antiviral-gene expression. J Virol 85: 11372–11380. https://doi.org/10.1128/JVI.05279-11 PMID: 21849431
6. Hasan UA, Zannetti C, Parroche P, Goutagny N, Malfroy M, et al. (2013) The human papillomavirus type 16 E7 oncoprotein induces a transcriptional repressor complex on the Toll-like receptor 9 promoter. J Exp Med 210: 1369–1387. https://doi.org/10.1084/jem.20122394 PMID: 23752229
7. Karim R, Tummers B, Meyers C, Biryukov JL, Alam S, et al. (2013) Human papillomavirus (HPV) upregulates the cellular deubiquitinase UCHL1 to suppress the keratinocyte's innate immune response. PLoS Pathog 9: e1003384. https://doi.org/10.1371/journal.ppat.1003384 PMID: 23717208
8. Niebler M, Qian X, Hofler D, Kogosov V, Kaewprag J, et al. (2013) Post-translational control of IL-1beta via the human papillomavirus type 16 E6 oncoprotein: a novel mechanism of innate immune escape mediated by the E3-ubiquitin ligase E6-AP and p53. PLoS Pathog 9: e1003536. https://doi.org/10.1371/journal.ppat.1003536 PMID: 23935506
9. Muruve DA, Petrilli V, Zaiss AK, White LR, Clark SA, et al. (2008) The inflammasome recognizes cytosolic microbial and host DNA and triggers an innate immune response. Nature 452: 103–107. https://doi.org/10.1038/nature06664 PMID: 18288107
10. Delaloye J, Roger T, Steiner-Tardivel QG, Le Roy D, Knaup Reymond M, et al. (2009) Innate immune sensing of modified vaccinia virus Ankara (MVA) is mediated by TLR2-TLR6, MDA-5 and the NALP3 inflammasome. PLoS Pathog 5: e1000480. https://doi.org/10.1371/journal.ppat.1000480 PMID: 19543380
11. Burdette D, Haskett A, Presser L, McRae S, Iqbal J, et al. (2011) Hepatitis C virus activates interleukin-1beta via caspase-1-inflammasome complex. J Gen Virol 93: 235–246. https://doi.org/10.1099/vir.0.034033-0 PMID: 21994322
12. Hornung V, Ablasser A, Charrel-Dennis M, Bauernfeind F, Horvath G, et al. (2009) AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. Nature 458: 514–518. https://doi.org/10.1038/nature07725 PMID: 19158675
13. Rathinam VA, Jiang Z, Waggoner SN, Sharma S, Cole LE, et al. (2010) The AIM2 inflammasome is essential for host defense against cytosolic bacteria and DNA viruses. Nat Immunol 11: 395–402. https://doi.org/10.1038/ni.1864 PMID: 20351692
14. Zepter K, Haffner A, Soohoo LF, De Luca D, Tang HP, et al. (1997) Induction of biologically active IL-1 beta-converting enzyme and mature IL-1 beta in human keratinocytes by inflammatory and immunologic stimuli. J Immunol 159: 6203–6208. PMID: 9550423
15. Sand J, Haertel E, Biedermann T, Contassot E, Reichmann E, et al. (2018) Expression of inflammasome proteins and inflammasome activation occurs in human, but not in murine keratinocytes. Cell Death Dis 9: 24. https://doi.org/10.1038/s41419-017-0009-4 PMID: 29348630
16. Watashi K, Liang G, Iwamoto M, Marusawa H, Uchida N, et al. (2013) Interleukin-1 and tumor necrosis factor-alpha trigger restriction of hepatitis B virus infection via a cytidine deaminase activation-induced cytidine deaminase (AID). J Biol Chem 288: 31715–31727. https://doi.org/10.1074/jbc.M113.501122 PMID: 24025329
17. Ronco LV, Karpova AY, Vidal M, Howley PM (1998) Human papillomavirus 16 E6 oncoprotein binds to interferon regulatory factor-3 and inhibits its transcriptional activity. Genes Dev 12: 2061–2072. PMID: 9649509
18. Boxman IL, Ruwhof C, Boerman OC, Lowik CW, Ponec M (1996) Role of fibroblasts in the regulation of proinflammatory interleukin IL-1, IL-6 and IL-8 levels induced by keratinocyte-derived IL-1. Arch Dermatol Res 288: 391–398. PMID: 8818187
19. Thomas M, Pim D, Banks L (1999) The role of the E6-p53 interaction in the molecular pathogenesis of HPV. Oncogene 18: 7690–7700. https://doi.org/10.1038/sj.onc.1202953 PMID: 10618709
20. Unlu S, Kumar A, Waterman WR, Tsukada J, Wang KZ, et al. (2007) Phosphorylation of IRF8 in a pre-associated complex with Spi-1/PU.1 and non-phosphorylated Stat1 is critical for LPS induction of the IL1B gene. Mol Immunol 44: 3364–3379. https://doi.org/10.1016/j.molimm.2007.02.016 PMID: 17386941

21. Scott CL, Soen B, Martens L, Skrypek N, Saelens W, et al. (2013) The transcription factor Zeb2 regulates development of conventional and plasmacytoid DCs by repressing Id2. J Exp Med 213: 897–911.
22. Richardson RJ, Dixon J, Malhotra S, Hardman MJ, Knowles L, et al. (2006) Irf6 is a key determinant of the keratinocyte proliferation-differentiation switch. Nat Genet 38: 1329–1334. https://doi.org/10.1038/ng1894 PMID: 17041603
23. Martinez-Zapien D, Ruiz FX, Poirson J, Mitschler A, Ramirez J, et al. (2016) Structure of the E6/E6AP/p53 complex required for HPV-mediated degradation of p53. Nature 529: 541–545. https://doi.org/10.1038/nature16481 PMID: 26789255
24. Zanier K, Charbonnier S, Sidi AO, McEwen AG, Ferrario MG, et al. (2013) Structural basis for hijacking of cellular LxxLL motifs by papillomavirus E6 oncoproteins. Science 339: 694–698. https://doi.org/10.1126/science.1229934 PMID: 23393263
25. Cooper B, Schneider S, Bohl J, Jiang Y, Beaudet A, et al. (2003) Requirement of E6AP and the features of human papillomavirus E6 necessary to support degradation of p53. Virology 306: 87–99. PMID: 12620801
26. Ganti K, Massimi P, Manzo-Merino J, Tomaic V, Pim D, et al. (2016) Interaction of the Human Papillomavirus E6 Oncoprotein with Sorting Nexin 27 Modulates Endocytic Cargo Transport Pathways. PLoS Pathog 12: e1005854. https://doi.org/10.1371/journal.ppat.1005854 PMID: 27649450
27. Nguyen ML, Nguyen MM, Lee D, Griep AE, Lambert PF (2003) The PDZ ligand domain of the human papillomavirus type 16 E6 protein is required for E6's induction of epithelial hyperplasia in vivo. J Virol 77: 6957–6964. https://doi.org/10.1128/JVI.77.12.6957-6964.2003 PMID: 12768014
28. Zanier K, ould M'hamed ould Sidi A, Boulade-Ladame C, Rybin V, Chappelle A, et al. (2012) Solution structure analysis of the HPV16 E6 oncoprotein reveals a self-association mechanism required for E6-mediated degradation of p53. Structure 20: 604–617. https://doi.org/10.1016/j.str.2012.02.001 PMID: 22483108
29. Karim R, Meyers C, Backendorf C, Ludigs K, Offringa R, et al. (2011) Human papillomavirus deregulates the response of a cellular network comprising of chemotactic and proinflammatory genes. PLoS One 6: e17848. https://doi.org/10.1371/journal.pone.0017848 PMID: 21423754
30. Um SJ, Rhyu JW, Kim EJ, Jeon KC, Hwang ES, et al. (2002) Abrogation of IRF-1 response by high-risk HPV E7 protein in vivo. Cancer Lett 179: 205–212. PMID: 11888675
31. Munoz-Fontela C, Macip S, Martinez-Sobrido L, Brown L, Ashour J, et al. (2008) Transcriptional role of p53 in interferon-mediated antiviral immunity. J Exp Med 205: 1929–1938. https://doi.org/10.1084/jem.20080383 PMID: 18663127
32. Rotondo JC, Bosi S, Bassi C, Ferracin M, Lanza G, et al. (2015) Gene expression changes in progression of cervical neoplasia revealed by microarray analysis of cervical neoplastic keratinocytes. J Cell Physiol 230: 806–812. https://doi.org/10.1002/jcp.24808 PMID: 25205602
33. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, et al. (2011) The mutational landscape of head and neck squamous cell carcinoma. Science 333: 1157–1160. https://doi.org/10.1126/science.1208130 PMID: 21798893
34. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, et al. (2011) Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science 333: 1154–1157. https://doi.org/10.1126/science.1206923 PMID: 21798897
35. den Boon JA, Pyeon D, Wang SS, Horswill M, Schiffman M, et al. (2015) Molecular transitions from papillomavirus infection to cervical precancer and cancer: Role of stromal estrogen receptor signaling. Proc Natl Acad Sci U S A 112: E3255–3264. https://doi.org/10.1073/pnas.1509322112 PMID: 26056290
36. (2017) Integrated genomic and molecular characterization of cervical cancer. Nature 543: 378–384. https://doi.org/10.1038/nature21386 PMID: 28112728
37. Scott ME, Shvetsov YB, Thompson PJ, Hernandez BY, Zhu X, et al. (2013) Cervical cytokines and clearance of incident human papillomavirus infection: Hawaii HPV cohort study. Int J Cancer 133: 1187–1196. https://doi.org/10.1002/ijc.28119 PMID: 23436563
38. Mehta AM, Mooij M, Brankovic I, Ouburg S, Morre SA, et al. (2017) Cervical Carcinogenesis and Immune Response Gene Polymorphisms: A Review. J Immunol Res 2017: 8913860. https://doi.org/10.1155/2017/8913860 PMID: 28280748
39. Mansour M, Touka M, Hasan U, Bellopede A, Smet A, et al. (2007) E7 properties of mucosal human papillomavirus types 26, 53 and 66 correlate with their intermediate risk for cervical cancer development. Virology 367: 1–9. https://doi.org/10.1016/j.virol.2007.05.005 PMID: 17568647
40. Lopez-Rovira T, Chalaux E, Massague J, Rosa JL, Ventura F (2001) Direct binding of Smad1 and Smad4 to two distinct motifs mediates bone morphogenetic protein-specific transcriptional activation of Id1 gene. J Biol Chem 277: 3176–3185. https://doi.org/10.1074/jbc.M106826200 PMID: 11700304

41. Hasan UA, Bates E, Takeshita F, Biliato A, Accardi R, et al. (2007) TLR9 expression and function is abolished by the cervical cancer-associated human papillomavirus type 16. J Immunol 178: 3186–3197. PMID: 17312167

42. Kelley ML, Keiger KE, Lee CJ, Huibregtse JM (2005) The global transcriptional effects of the human papillomavirus E6 protein in cervical carcinoma cell lines are mediated by the E6AP ubiquitin ligase. J Virol 79: 3737–3747. https://doi.org/10.1128/JVI.79.6.3737-3747.2005 PMID: 15731267

43. Buck C.B. P DV, Lowy D.R., Schiller J.T. (2005) Generation of HPV pseudovirions using transfection and their use in neutralization assays. Methods Mol Med 119: 445–462. https://doi.org/10.1385/1-59259-982-6:445 PMID: 16350417

44. Pyeon D L P, Ahlquist P. (2005) Production of infectious human papillomavirus independently of viral replication and epithelial cell differentiation. Proc Natl Acad Sci U S A 102: 9311–9316. https://doi.org/10.1073/pnas.0504020102 PMID: 15958530

45. Hasan U (2013) Human papillomavirus (HPV) deregulation of Toll-like receptor 9. Oncoimmunology 3: e27257.

46. Zannetti C, Roblot G, Charrier E, Ainouze M, Tout I, et al. (2016) Characterization of the Inflammasome in Human Kupffer Cells in Response to Synthetic Agonists and Pathogens. J Immunol 197: 356–367. https://doi.org/10.4049/jimmunol.1502301 PMID: 27226092

47. Hasan UA, Dollet S, Vlach J (2004) Differential induction of gene promoter constructs by constitutively active human TLRs. Biochem Biophys Res Commun 321: 124–131. https://doi.org/10.1016/j.bbrc.2004.06.134 PMID: 15358224

48. Parroche P, Roblot G, Le Calvez-Kelm F, Tout I, Marotel M, et al. (2016) TLR9 re-expression in cancer cells extends the S-phase and stabilizes p16(INK4a) protein expression. Oncogenesis 5: e244. https://doi.org/10.1038/oncsis.2016.49 PMID: 27454079

49. Broz P, Newton K, Lamkanfi M, Mariathasan S, Dixit VM, et al. (2010) Redundant roles for inflammasome receptors NLRP3 and NLRC4 in host defense against Salmonella. J Exp Med 207: 1745–1755. https://doi.org/10.1084/jem.20100257 PMID: 20603313

Virology 506 (2017) 45–54

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Isolation and characterization of a novel putative human polyomavirus

Tarik Gheit[a], Sankhadeep Dutta[a], Javier Oliver[a], Alexis Robitaille[a], Shalaka Hampras[b], Jean-Damien Combes[a], Sandrine McKay-Chopin[a], Florence Le Calvez-Kelm[a], Neil Fenske[c,d], Basil Cherpelis[c,d], Anna R. Giuliano[e], Silvia Franceschi[a], James McKay[a], Dana E. Rollison[b], Massimo Tommasino[a,*]

[a] International Agency for Research on Cancer, World Health Organization, Lyon 69372, France
[b] Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL, USA
[c] Department of Dermatology & Cutaneous Surgery, University of South Florida, Morsani College of Medicine, Tampa, FL, USA
[d] Department of Cutaneous Oncology, Moffitt Cancer Center, Tampa, FL, USA
[e] Center for Infection Research in Cancer, Moffitt Cancer Center, Tampa, FL, USA

ARTICLE INFO

Keywords:
New polyomavirus
Human
Skin

ABSTRACT

The small double-stranded DNA polyomaviruses (PyVs) form a family of 73 species, whose natural hosts are primarily mammals and birds. So far, 13 PyVs have been isolated in humans, and some of them have clearly been associated with several diseases, including cancer. In this study, we describe the isolation of a novel PyV in human skin using a sensitive degenerate PCR protocol combined with next-generation sequencing. The new virus, named Lyon IARC PyV (LIPyV), has a circular genome of 5269 nucleotides. Phylogenetic analyses showed that LIPyV is related to the raccoon PyV identified in neuroglial tumours in free-ranging raccoons.

Analysis of human specimens from cancer-free individuals showed that 9 skin swabs (9/445; 2.0%), 3 oral gargles (3/140; 2.1%), and one eyebrow hair sample (1/439; 0.2%) tested positive for LIPyV.

Future biological and epidemiological studies are needed to confirm the human tropism and provide insights into its biological properties.

1. Introduction

Members of the polyomaviruses (PyVs) are non-enveloped double-stranded DNA viruses with a genome of approximately 5000 nucleotides. The organization of the viral genome is highly conserved throughout the PyV family and comprises early and late coding regions and the viral non-coding control region (NCCR) of approximately 500 bp. The early region encodes for two regulatory proteins: small T-antigen (ST-Ag) and large T antigen (LT-Ag). The late region encodes for three viral proteins that are necessary for formation of the capsid: the major capsid protein VP1 and two minor capsid proteins, VP2 and VP3. The NCCR contains the origin of DNA replication, regulatory elements, and transcription promoters (Moens et al., 2008).

With the development of high-performance molecular biology tools, many of the PyVs have been isolated during the past decade, mainly from mammals, birds, and fish (Johne et al., 2011; Peretti et al., 2015). Based on the observed distance between LT-Ag coding sequences, the International Committee on Taxonomy of Viruses (ICTV) Polyomaviridae Study Group has classified the different species of PyVs into four genera: alpha-, beta-, gamma- and delta-PyV) (Calvignac-Spencer et al., 2016).

To date, a total of 13 PyVs have been isolated from humans: BKPyV (Gardner et al., 1971), JCPyV (Padgett et al., 1971), KIPyV (Allander et al., 2007), WUPyV (Gaynor et al., 2007), Merkel cell PyV (MCPyV) (Feng et al., 2008), human PyV 6 (HPyV6) (Schowalter et al., 2010), human PyV 7 (HPyV7) (Schowalter et al., 2010), trichodysplasia spinulosa-associated PyV (TSPyV) (van der Meijden et al., 2010), human PyV 9 (HPyV9) (Scuda et al., 2011), Malawi PyV (MWPyV) (Siebrasse et al., 2012), Saint Louis PyV (STLPyV) (Lim et al., 2013), human PyV 12 (HPyV12) (Korup et al., 2013) and New Jersey PyV (NJPyV) (Mishra et al., 2014). PyVs are widely spread in the human population. Many PyV infections occur early in life, and in most cases it remains asymptomatic (Nickeleit et al., 2015). Serological studies have shown that up to 90% of the human population has been exposed to HPyV, with several HPyV infections occurring during childhood (Egli et al., 2009; Kean et al., 2009; Sroller et al., 2016).

Four HPyVs have been clearly associated with human diseases, many occurring more frequently in immunocompromised individuals. JCPyV has been associated with progressive multifocal leukoencephalopathy, a fatal brain disease, in immunocompromised individuals (Jiang et al., 2009; Koralnik, 2006), and BKPyV has been associated

* Correspondence to: Infections and Cancer Biology Group, International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France.
E-mail address: icb@iarc.fr (M. Tommasino).

http://dx.doi.org/10.1016/j.virol.2017.03.007
Received 27 January 2017; Received in revised form 7 March 2017; Accepted 16 March 2017
Available online 22 March 2017

with nephropathy and hemorrhagic cystitis (Azzi et al., 1994; Coleman et al., 1978), particularly among kidney transplant patients; MCPyV has been isolated from Merkel cell carcinomas of the skin (Feng et al., 2008), a cancer with higher incidence in immunocompromised individuals, and TSVPyV has been associated with a rare cutaneous condition, trichodysplasia spinulosa, in an immunocompromised patient (van der Meijden et al., 2010). A possible association between HPyV7 and non-neoplastic diseases in immunosuppressed individuals has also been reported recently (Ho et al., 2015; Toptan et al., 2016). In addition, HPyV7 has been found in human thymic epithelial tumours, but a causal association has not been established (Rennspiess et al., 2015). The remaining PyVs that have been isolated from humans, KIPyV, WUPyV, HPyV6, HPyV9, HPyV12, MWPyV, STLPyV and NJPyV, have not so far been associated with any human diseases.

The oncogenic potential of HPyVs has been extensively studied in experimental animal models, where these viruses induce a wide range of tumours. The inoculation of JCPyV in a small rodent model and in non-human primates leads to the development of brain tumours (Miller et al., 1984; Varakis et al., 1978; Walker et al., 1973; Zu Rhein et al., 1979). Transgenic mice expressing the early region of HPyVs have been used to investigate the carcinogenesis induced by MCPyV (Shuda et al., 2015; Verhaegen et al., 2015), JCPyV (Shollar et al., 2004), and BKPyV (Dalrymple et al., 1990). In addition, simian virus 40 (SV40), BKPyV, and JCPyV have been shown to display transforming activity in *in vitro* experimental models (Moens et al., 2008). It is still unclear whether other HPyVs exist. Here, we report the characterization of a new PyV isolated from human skin swabs. We found it to be phylogenetically related to the raccoon PyV (RacPyV) associated with brain tumours in free-ranging raccoons, and gave it the provisional name of Lyon IARC PyV (LIPyV).

2. Materials and methods

2.1. Human specimens

Skin swabs, eyebrow hairs and oral gargles from three different ongoing studies aiming to determine the prevalence of human papillomaviruses (HPVs) and HPyVs were used in the present analysis (Franceschi et al., 2015; Hampras et al., 2015, 2014; Nunes et al., 2016; Pierce Campbell et al., 2016, 2013). Skin swabs and eyebrow hairs were collected at baseline from 448 subjects participating in the VIRUSCAN study, an ongoing five-year (2014–2019) prospective cohort study being conducted at Moffitt Cancer Center and the University of South Florida (R01CA177586-01; “Prospective study of cutaneous viral infections and non-melanoma skin cancer”). In addition, 25 cutaneous skin swabs were randomly selected from the HPV Infection in Men (HIM) study, a large, multi-national prospective cohort study of the natural history of HPV infection in men. The 25 skin swabs were collected from men in Tampa, Florida, USA. The HIM study methods have been described in detail previously and are similar to those used in the VIRUSCAN Study (Giuliano et al., 2011). An area of approximately 5×5 cm of the top of the sun-exposed forearm was sprayed with 0.9% saline solution. A cotton-tipped Dacron swab (Digene, Gaithersburg, MD, USA) was then rubbed back and forth a few times to collect exfoliated skin cells. Individual swabs were placed in a separate vial and preserved in Digene Standard Transport Medium (STM). Three or four eyebrow hairs were plucked from each eyebrow using disposable tweezers. The eyebrow hairs with attached follicles were snap-frozen in liquid nitrogen and stored at −80 °C until further use.

We used 140 oral gargles that were collected for the Study of Natural History of HPV Infection and Precancerous Lesions in the Tonsils (SPLIT), which is an ongoing study on the detection of HPV infection and precancerous lesions in age-stratified immunocompetent individuals who underwent tonsillectomy for benign diseases in selected university hospitals across France (Combes et al., In press; Franceschi et al., 2015).

After DNA extraction, all samples were analysed at the International Agency for Research on Cancer (Lyon, France) for HPVs and all known HPyVs.

2.2. Design of degenerate primers and PCR conditions for PyV screening

Complete HPyV sequences were obtained from GenBank and were used for alignment of the early region genes. A pair of degenerate primers was developed based on the more conserved parts of LT-Ag of several PyV genomes. The accession numbers of the GenBank sequences that were used as references, with the corresponding HPyV types given in parentheses, are EU37584 (MCPyV), NC_001538 (BKPyV), NC_001669 (SV40), EF520287 (KIPyV), NC_009539 (WUPyV), and NC_001699 (JCPyV). Two oligonucleotides (forward primer, 5′-CAW GCT GTR TIT AGT AAT A-3′ and reverse primer, 5′-RWT TAT TMA CHC CIT TAC-3′), allowing the amplification of a region of approximately 240 bp, were synthesized by MWG Biotech (Ebersberg, Germany). The polymerase chain reaction (PCR) mix contained 1x PCR buffer, 200 µmol/L of each dNTP, 0.2 µmol/L of each primer, and 0.625 U of HotStarTaq DNA polymerase in a final volume of 25 µL (Qiagen). Forty-five amplification cycles were run in the GeneAmp PCR System 2400 with a 94 °C denaturation step (1 min), a 48 °C annealing step (1 min), and a 72 °C extension step (1 min), including an initial denaturation step of 15 min and a final extension step of 10 min, resulting in a 240-bp product.

2.3. Next-generation sequencing

The libraries were prepared using 50 ng of the PCR products with DNA NEBNext Fast DNA Library Prep Set for Ion Torrent (New England Biolabs, Ipswich, MA, USA) following the manufacturer’s protocol, and sequenced with the Ion Torrent PGM sequencer (Life Technologies) at 100x coverage using the Ion OneTouch 200 Template Kit v2 DL and the Ion PGM Sequencing 200 Kit v2 with the 314 or 316 chip kits (all produced by Life Technologies), following the manufacturer’s instructions. The data analysis was conducted using Geneious version 6.0.1 (http://www.geneious.com) (Kearse et al., 2012).

2.4. Luminex assay for high throughput screening of LIPyV

As described previously, LIPyV DNA from eyebrow hairs, and skin swabs was detected using a highly sensitive and specific assay which combines multiplex PCR and bead-based Luminex technology (Schmitt et al., 2006, 2010). The following PCR primers and Luminex probe were used: forward primer, 5′-CAA GCC TTG CTG CAG CAT TCC TAG-3′ and reverse primer, 5′-ATC TTT GTT TTG TCC TCT AGA ACC CT-3′; and probe, 5′-ATC TAT CTT GGG GGC AAT-3′. Briefly, PCR products were denatured and hybridized to the beads coupled with specific probes for LIPyV. Results were expressed as the median fluorescence intensity (MFI) of at least 100 beads per bead set. For each probe, MFI values with no respective PCR product added to the hybridization mixture were considered background values. The cut-off was computed by adding 5 MFI to 1.1x the median background value.

2.5. Rolling circle amplification

DNA was extracted and purified from skin swabs as described previously (Schowalter et al., 2010). The DNA was amplified by multiply primed rolling circle amplification (RCA) using the Illustra TempliPhi 100 Amplification Kit according to the manufacturer’s recommendations (GE Healthcare, Piscataway, NJ), with supplementation of 450 µM dNTPs as described by Rector et al. (2004).

2.6. Long-range PCR

Long-range PCR was performed for amplification of the entire

genome using the Takara LA Taq HS polymerase, following the manufacturer's instructions (Takara Bio Inc.). The following primers were used at a final concentration of 0.5 μM each: forward primer, 5′-TAA ATT TTG AGT TGG GTT GTG CAC AAG AT-3′ and reverse primer, 5′-ATC TAT CTT GGG GGC AAT TAA TAT TTA ATG-3′.

2.7. Proofreading PCR

PCR using the proofreading Pfu ultra hot start DNA polymerase (Agilent Technologies, Santa Clara, CA, USA) was performed according to the manufacturer's instructions.

2.8. Cell culture and transient transfection

First, HEK293 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100U/ML penicillin G, 100 μg/ML streptomycin, 2 mM L-glutamine (Invitrogen Life Technologies), and 1 mM sodium pyruvate (Sigma-Aldrich). Then, 1.5×10^5 cells were transiently transfected with 1.0 μg of pcDNA3 expression vector (Invitrogen) containing the early region of LIPyV by using the X-tremeGENE 9 reagent (Roche) according to the manufacturer's protocols. At 48 h after transfection, cells were collected for isolation of total RNA.

2.9. Reverse transcription and qPCR

Total RNA was extracted using the NucleoSpin RNA kit (Macherey-Nagel). The obtained RNA was reverse-transcribed to cDNA with the RevertAid H Minus M-MuLV Reverse Transcriptase kit (ThermoFisher Scientific) according to the manufacturer's instructions. The LIPyV cDNA was amplified by PCR using the following pair of specific primers: forward primer, 5′-AGA ATA TGG TAA TAT ACC ATT AAT GAA GAA TG-3′ and reverse primer 5′-GTG ATC AGA TTG TGA TTT TGC TGA G-3′. The amplicon was purified using the QIAquick gel extraction kit (Qiagen) and sequenced by GATC (GATC Biotech, Germany).

2.10. Phylogenetic analyses

Phylogenetic trees were constructed from the alignments of the nucleotide sequences of LT-Ag and VP1, and the amino acid sequences of LT-Ag from the following 47 avian and mammalian PyVs: budgerigar PyV (NC_004764), crow PyV (NC_007922), finch PyV (NC_007923), goose hemorrhagic PyV (NC_004800), TSPyV (NC_014361), Bornean orangutan PyV1 (NC_013439), chimpanzee PyV (NC_014743), murine PyV (NC_001515), hamster PyV (NC_001663), HPyV9 (NC_015150), African green monkey PyV (NC_004763), SV40 (NC_001669), BKPyV (NC_001538), JCPyV (NC_001699), simian virus 12 (NC_007611), California sea lion PyV1 (NC_013796), bovine PyV (NC_001442), murine pneumotropic virus (NC_001505), squirrel monkey PyV (NC_009951), HPyV6 (NC_014406), HPyV7 (NC_014407), KIPyV (NC_009238), WUPyV (NC_009539), MWPyV_MA095 (JQ898291), MWPyV_WD976 (JQ898292), STLPyV_MA138 (NC_020106), STL PyV_WD972 (JX463184), raccoon PyV_R45 (JQ178241), raccoon PyV_Rac17 (KU533635), equine PyV (NC_017982), *Artibeus planirostris* PyV3_A504 (JQ958890), *Myotis* PyV (NC_011310), *Mastomys* PyV (AB588640), dolphin PyV1 (KC594077), vervet monkey PyV1 (NC_019844), *Otomops* PyV2 (NC_020066), *Chaerephon* PyV1 (NC_020065), bat PyV_B0454 (JQ958888), *Eidolon* PyV1 (NC_020068), *Pan troglodytes verus* PyV1a (HQ385746), *Pan troglodytes verus* PyV2a (HQ385748), MCPyV (NC_010277), gorilla PyV1 (HQ385752), *Cardioderma* PyV1 (NC_020067), *Otomops* PyV1 (NC_020071), bat PyV (JQ958889) and LIPyV (KY404016).

The sequences were aligned using the MUSCLE algorithm with default parameters (Edgar, 2004a), implemented in MEGA7 (Kumar et al., 2016). MEGA7 was used to test substitution models, and for all the following phylogenetic analysis. Based on the alignment with MUSCLE, all positions with less than 95% site coverage were eliminated (partial deletion), to allow the inclusion of taxa with some missing data. Codon positions included were 1st+2nd+3rd+ non-coding. There were a total of 1011 positions in the final dataset for VP1 nucleotide sequences, 1674 for LT-Ag nucleotide sequences, and 554 for LT-Ag amino acid sequences. A discrete gamma distribution was used to model evolutionary rate differences among sites (five categories; +G, parameter=0.8203 and 1.1058, respectively, for VP1 and LT-Ag nucleotides sequences, and 1.1420 for LT-Ag amino acid sequences). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 8.6157% sites and 10.0832% sites, respectively, for VP1 and LT-Ag nucleotides sequences, and 8.9514% for LT-Ag amino acid sequences).

The initial trees for the heuristic search were obtained automatically by applying the Neighbour-Joining (NJ)/BioNJ algorithm to the three different matrixes of pairwise distances estimated using the maximum composite likelihood (MCL) approach for the VP1 and LT-Ag nucleotide sequences, and estimated using a Jones–Thornton–Taylor (JTT) model for the LT-Ag amino acid sequences. The initial trees were obtained using the NJ/BioNJ algorithm to have a first representation of the relationships between the sequences according to their genetic distance, because this algorithm produces a single tree.

In the next step, the statistical method used was the maximum likelihood (ML) algorithm, applied with the goal of comparing the initial trees with other trees generated by the ML search, using the likelihood criterion. During this ML run, the parameter values were optimized to converge to the true parameter value, aiming to find the smallest possible variance among all estimates with the same expected value.

For the nucleotide sequences of both VP1 and LT-Ag, the evolutionary history was inferred by using the ML method based on the general time-reversible model (a nucleotide substitution model), whereas for the amino acid sequences of LT-Ag, the evolutionary history was inferred by using the ML method based on the Le_Gascuel_2008 model (Le and Gascuel, 2008) (an amino acid substitution model).

For all trees, the topology was then optimized using the nearest-neighbour-interchange heuristic (NNI) to improve the likelihood. This heuristic specifies a neighbour relation between two unrooted trees and then swaps their subtrees in an attempt to obtain a tree that has a higher likelihood.

The final trees that are kept are the trees with the highest log likelihood. Five-hundred ML bootstrap replicates were performed and the support for each node annotated onto the ML tree for each of the phylogenetic trees.

2.11. Nucleotide sequence accession number

The sequence of LIPyV was submitted to Gen Bank and was assigned accession number KY404016.

3. Results

3.1. Viral discovery and sequencing of a new polyomavirus

To identify new human PyVs, 25 skin swabs from the HIM study were tested for PyV using a sensitive degenerate PCR that amplifies a region of approximately 240 bp in LT-Ag. Electrophoretic analysis of the PCR products revealed the presence of amplicons of the expected size in 6 skin swab samples (6/25; 24%). The purified PCR products were pooled and sequenced using Ion Torrent technology (Life Sciences). Approximately 1900 reads were obtained, generating 37 contigs. Nucleotide sequence analysis (BLASTn) revealed that a group of 105 reads of approximately 200 bp shared the highest nucleotide sequence similarity (76%) to RacPyV strains (accession numbers

KU533635 and JQ178241), thus representing a potential new PyV. The re-analysis by PCR of the 25 skin swabs using specific primers showed that only one skin swab (1/25; 4%) tested positive for the new PyV sequence.

Multiply primed RCA (Johne et al., 2009) was performed on the DNA extracted from the skin swab of a woman aged 65 years, who had previously tested positive for the new PyV sequence by PCR. To obtain the complete viral genome, first, long-range PCR was performed using outward-directed primers specific for the putative new PyV and the RCA product as template, generating an amplicon of approximately 5 kb. Then, by a primer-walking strategy (GATC Biotech, Germany), a sequence of 5269 bp representing a whole circular genome of the PyV was obtained. The sequence was validated twice using a proofreading polymerase followed by Sanger sequencing.

BLASTn analysis of the whole viral genome confirmed the RacPyV as the closest relative among all known PyVs. Moreover, using MUSCLE (Edgar, 2004b), the new PyV showed the highest nucleotide sequence identity (~65%) to RacPyV strains (accession numbers KU533635 and JQ178241). Because PyV sequences that share less than 81% whole-genome nucleotide sequence identity to members of known species are traditionally considered to be distinct viral species (Buck et al., 2016; Johne et al., 2011), this new species of PyV was given the provisional name of Lyon IARC PyV (LIPyV).

3.2. Genome characterization

The genome of LIPyV is circular and 5269 bp in length (accession number KY404016), encoding open reading frames (ORFs) for all of the major PyV proteins. Analysis of the complete nucleotide sequence showed that the LIPyV genome shares the features of other known PyVs with an early region consisting of ST-Ag and LT-Ag and a late region coding for the VP1, VP2, and VP3 structural proteins. A NCCR (nucleotide positions 1–401) sharing the characteristics of the ori regions of most of the mammalian polyomaviruses was found (Fig. 1**A**). This region contains six LT-Ag binding sites (An et al., 2012; Pipas, 1992): four GAGGC, one reverse complement GCCTC, and the sixth with the sequence 5′-GTGGC-3′.

The early gene expression region (nucleotide positions 2455–5269) has a length of 2815 bp and contains ORFs encoding for ST-Ag and LT-Ag. To accurately determine the splice donor and splice acceptor giving rise to LT-Ag (Fig. 1**B**), the entire early region was cloned into an expression vector (pcDNA3), and 293 cells were transfected for 48 h. Reverse transcription PCR (RT-PCR) was performed using the primers spanning a region of 900 bp that most likely harbours the splice sites. An amplicon of 488 bp was obtained and sequenced in order to identify the splice junction, resulting in the identification of a LT ORF of 2403 bp that encodes for a LT-Ag of 800 amino acids. In addition, the RT-PCR experiment enabled the identification of two additional mRNA transcripts that correspond to (i) a splicing of 98 bp within the transcript that gives rise to ST-Ag of 179 amino acids (Shuda et al., 2009), and (ii) a splicing of 222 bp that yields an ORF of 148 amino acids, with the first 138 amino acids shared with ST-Ag. We named this putative protein 148T-Ag (Figs. 1**B** and 2). Experiments with the entire viral genome may reveal additional alternative mRNA splice variants; however, we do not yet have indications about whether LIPyV could efficiently replicate in *in vitro* experimental models.

As for MCPyV (Carter et al., 2013) or TSPyV (van der Meijden et al., 2015), we identified a putative alternate T antigen ORF (ALTO) overprinted in the +1 frame of the second exon of LIPyV LT (nucleotide positions 4035–4619). This ORF encodes for 194 amino acids (Fig. 1**A**).

The structure of ST-Ag, LT-Ag and 148T-Ag is shown in Fig. 2 and Table 1. The late region of LIPyV includes ORFs that encode for the VP1 (nucleotide positions 1120–2427), VP2 (nucleotide positions 402–1145), and VP3 (nucleotide positions 630–1145) capsid proteins. The start codon for the VP3 ORF is located within the VP2 ORF, and the start codon for the VP1 ORF overlaps the C-terminal region of the VP2 ORF (Fig. 1**A**).

3.3. Phylogenetic relationship among polyomaviruses

To investigate the evolutionary history of LIPyV, we constructed ML phylogenetic trees using MEGA7 (Kumar et al., 2016) based on an alignment of the nucleotide sequences of VP1 and LT-Ag (Fig. 3**A** and **B**), and on the alignment of the amino acid sequences of LT-Ag (Fig. 3**C**) of 47 mammalian or avian PyVs. The phylogenies of VP1 and

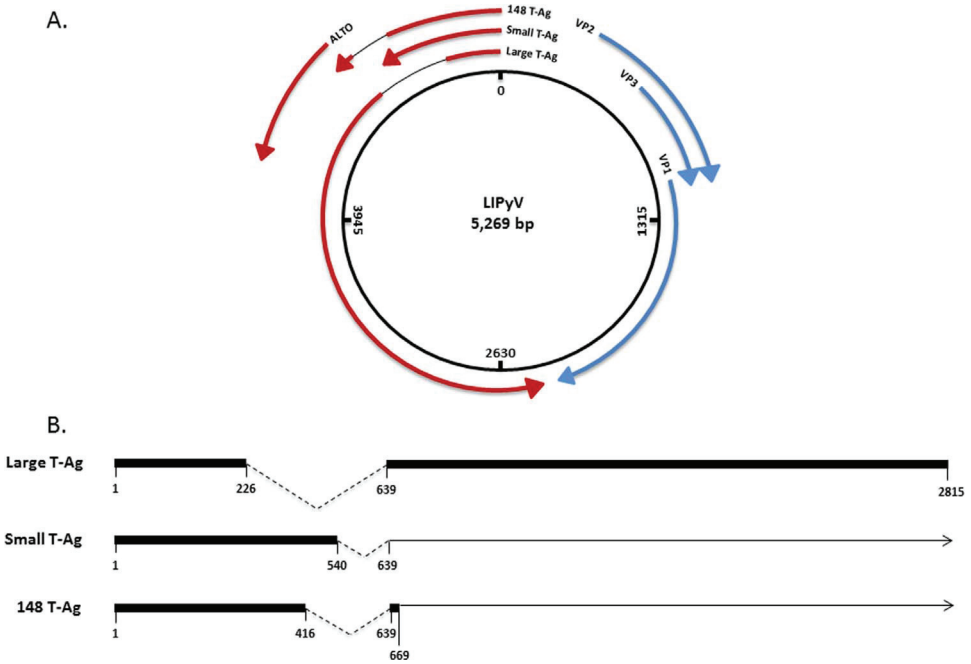


Fig. 1. Genome organization of LIPyV. (A) The viral genome of 5269 bp comprises early and late coding regions that encode for two regulatory proteins (small T-antigen and large T-antigen), the 148 T-antigen, the putative alternate T antigen (ALTO), the major capsid protein VP1, and two minor capsid proteins, VP2 and VP3. These regions are separated by a non-coding control region (NCCR) of 401 bp. (B) Transcript mapping of small, large, and 148 T-antigens.

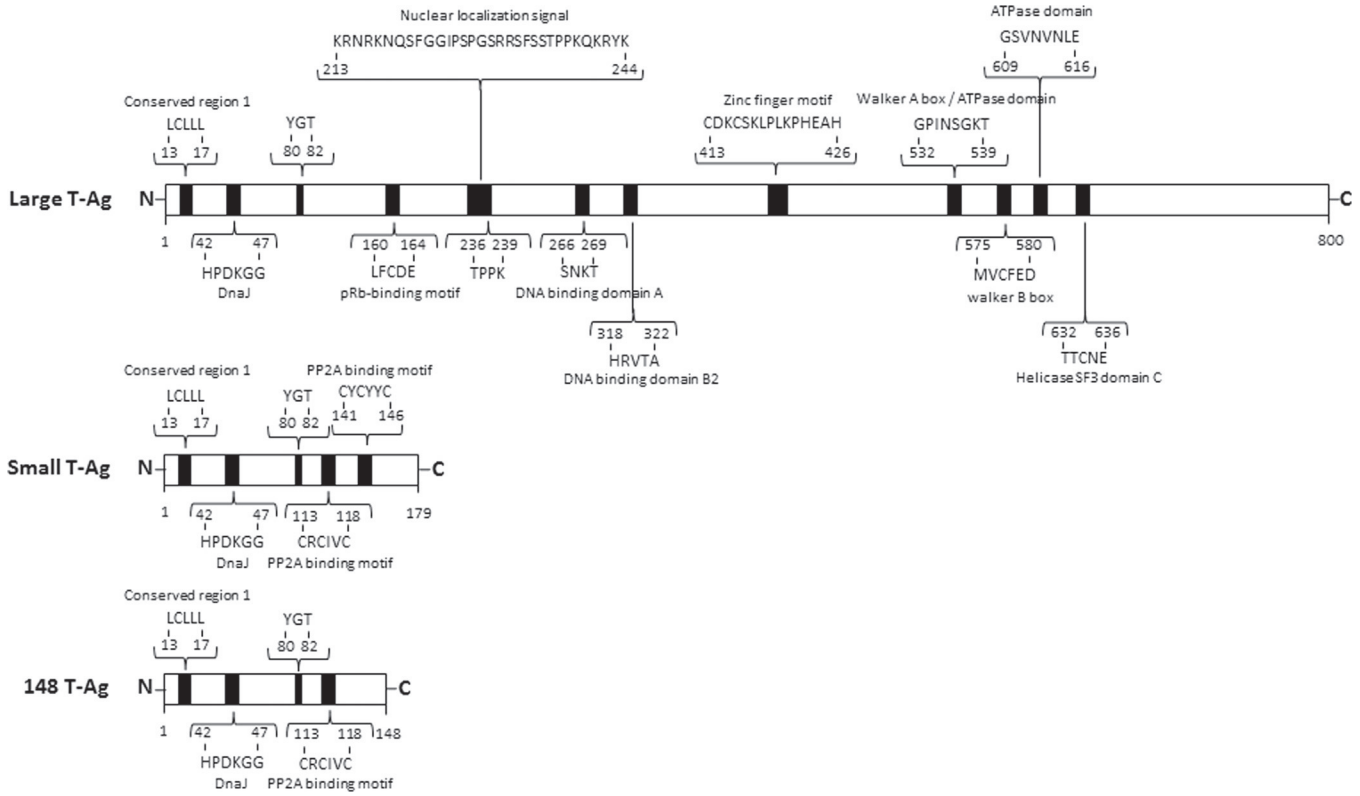


Fig. 2. Structure of the large, small, and 148 T-antigens.

LT-Ag showed that LIPyV is closely related to RacPyVs. In all trees, LIPyV and RacPyVs clustered significantly with different mammalian PyVs that include one HPyV (MCPyV) and several PyVs isolated from bats (*Otomops, Cardioderma, Eidolon*) or primates (gorilla, *Pan troglodytes verus*, vervet monkey).

3.4. LIPyV prevalence in human specimens

The prevalence of LIPyV in humans was evaluated using a highly specific and sensitive Luminex-based assay (Schmitt et al., 2006, 2010). The analysis was performed on skin swabs and eyebrow hairs

Table 1
Description of the LIPyV T-antigens.

T-antigens	Amino acid motifs	Names	Amino acid positions	References
Large T-antigen				
	LCLLL	CR1	13–17	Pipas (1992)
	HPDKGG	DnaJ	42–47	
	YGT	YGT	80–82	Houben et al. (2015)
	LFCDE	pRB-binding motif	160–164	DeCaprio et al. (1988)
	KRNRKNQSFGGIPSPGSRRSFSSTPPKQKRYK	Nuclear localization signal*	213–244	Kosugi et al. (2009a) Kosugi et al. (2009b)
	TPPK	threonine–proline–proline–lysine	236–239	DeCaprio and Garcea (2013)
	SNKT	DNA-binding domain A	266–269	Johne et al. (2006)
	HRVTA	DNA-binding domain B2	318–322	Schuurman et al. (1990) Simmons et al. (1990)
	CDKCSKLPLKPHEAH	Zinc-finger motif	413–427	Ehlers and Moens (2014)
	GPINSGKT	Walker A box/ATPase domain	532–539	Pipas (1992)
	GSVNVNLE	ATPase domain	609–616	
	MVCFED	Walker B box	575–580	van der Meijden et al. (2010)
	TTCNE	Helicase superfamily 3 motif C	632–636	
Small T-antigen				
	LCLL	CR1	13–17	Pipas (1992)
	HPDKGG	DnaJ	42–47	
	YGT	YGT	80–82	Houben et al. (2015)
	CRCIVC	PP2A binding site	113–118	Pipas (1992)
	CYCYYC	PP2A binding site	141–146	
148 T-antigen				
	LCLL	CR1	13–17	Pipas (1992)
	HPDKGG	DnaJ	42–47	
	YGT	YGT	80–82	Houben et al. (2015)
	CRCIVC	PP2A binding site	113–118	Pipas (1992)

* A putative nuclear localization signal (NLS) has been predicted using an NLS-prediction algorithm, cNLS Mapper (http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi).

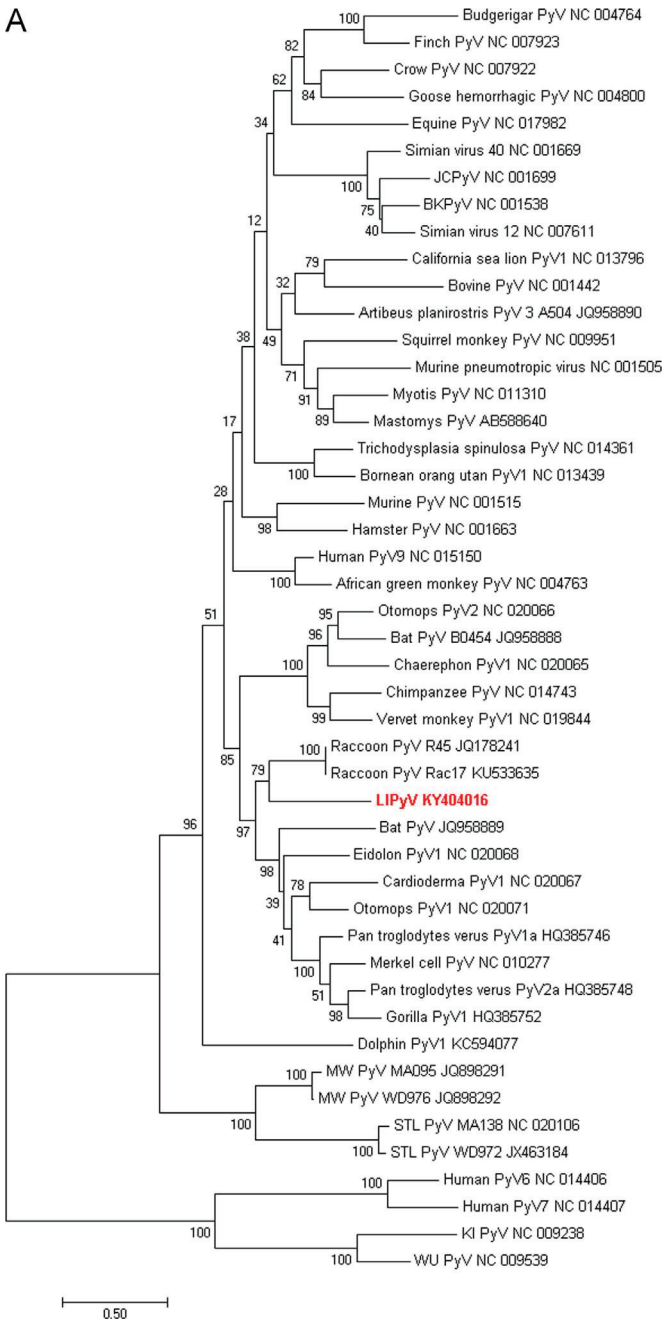


Fig. 3. Maximum-likelihood phylogenetic unrooted trees produced from different regions of the LIPyV genome. The VP1 open reading frame (A), the large T-antigen (LT-Ag) open reading frame (B), and the amino acid sequence of LT-Ag (C) are compared separately. Scale bar shows substitution rate per site.

collected in the USA from 448 skin cancer screening patients participating in the VIRUSCAN study prospective cohort. The mean age at study enrolment was 69.4 years. The study population included 54.7% women, 96.4% Whites, and 93.4% non-Hispanics. Skin swabs and eyebrow hair samples were obtained at baseline from 445 and 439 individuals, respectively. Based on the Luminex analysis, 9 skin swabs (9/445; 2.0%) and 1 eyebrow hair sample (1/439; 0.2%) tested positive for LIPyV. The LIPyV-positive eyebrow hair sample was obtained from a Hispanic White man aged 79 years. The LIPyV-positive skin swabs were obtained from 4 non-Hispanic White men, 1 Hispanic White man, and 4 non-Hispanic White women, aged 62–81 years. Interestingly, the eyebrow hair follicles and the skin swab collected from one individual both tested positive for LIPyV.

In addition, 140 oral gargles from the SPLIT study were collected in France from 59 women and 81 men, aged 18–67 (mean age=29.3 years). A total of 3 oral gargles (3/140; 2.1%) tested positive for LIPyV. Two women (aged 34 and 51 years) and one man (aged 26 years) were positive for LIPyV.

4. Discussion

Polyomaviridae is a growing family that infects fish, birds, rodents, humans, and non-human primates (Johne et al., 2011; Peretti et al., 2015). With the advent of new molecular tools, the discovery of new PyVs has accelerated over the past decade. However, although a large number of PyVs have been detected in different animals, the discovery of new HPyVs has been less frequent. The latest HPyV (NJPyV) was discovered in 2014 (Mishra et al., 2014). Although most of the HPyV

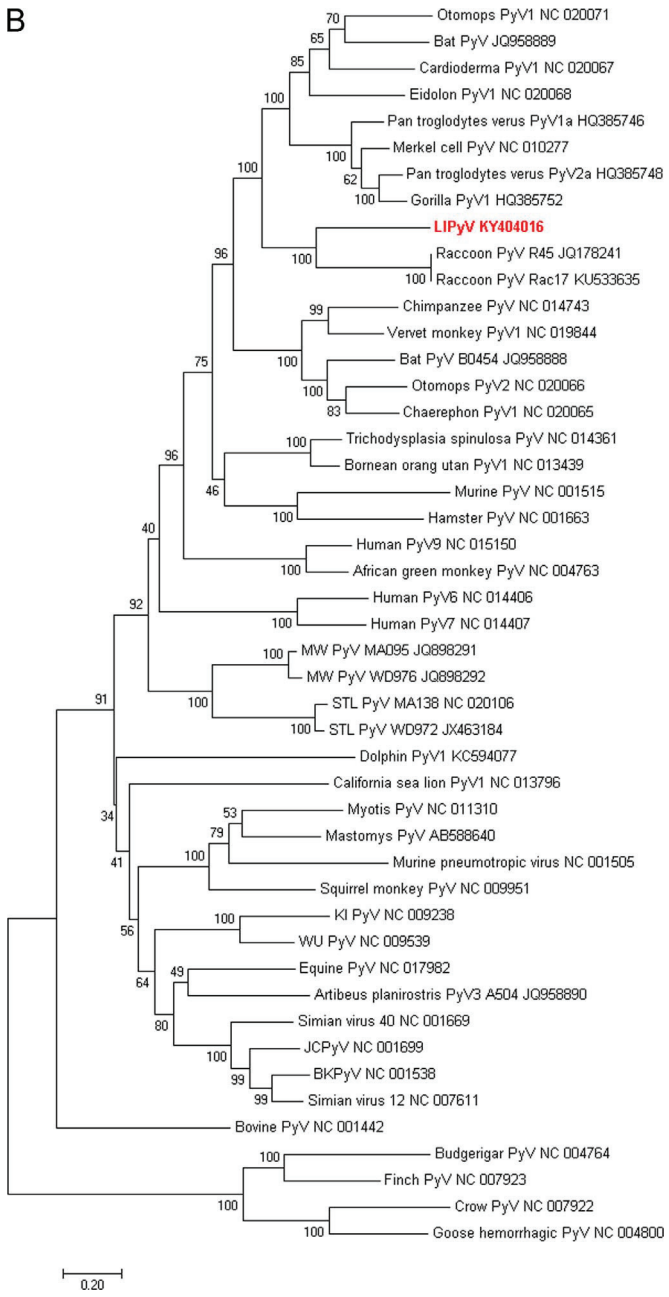


Fig. 3. (*continued*)

infections are asymptomatic, a few HPyVs may induce diseases or cancer, notably in immunocompromised individuals (Feng et al., 2008; Jiang et al., 2009; van der Meijden et al., 2010).

In the present study, the use of degenerate PCR primers combined with high-throughput sequencing enabled the discovery of a new PyV in human skin specimens. The analysis of the LIPyV genome showed that its organization shares most of the features of other known PyVs, and contains conserved domains that play roles in PyV-induced cell transformation. The N-terminal region of LIPyV LT-Ag contains a LXCXE motif that has the ability to bind pRB family members (Chestukhin et al., 2002; Moens et al., 2007). The pRB binding motif is always preserved after the integration of MCPyV LT into the host genome; in addition, it has been shown to be required in promoting growth of Merkel cell carcinoma cells (Houben et al., 2012; Shuda et al., 2008). LIPyV LT-Ag also has an ATPase domain that contains two highly conserved motifs, GPXXXGKT and GXXXVNLE, that are necessary for complex formation with p53 (Pipas, 1992). LIPyV ST-Ag contains two PP2A binding sites. Several ST-Ag PyVs (BKPyV, JCPyV, MuPyV, SV40, and MCPyV) have the ability to interact and inhibit PP2A phosphatase activity. SV40 ST-Ag alters PP2A activity by interacting with the PP2A scaffolding A subunit; the loss of this interaction impairs the tumorigenic activity of ST-Ag (Cho et al., 2007; Guergnon et al., 2011; Kwun et al., 2015; Sablina et al., 2008; Yu et al., 2001). However, *in vitro* experimental studies are required to demonstrate the ability of LIPyV to replicate in human cells, and to characterize the biological activity of its viral proteins; such studies will make it possible to predict the potential role of this newly discovered PyV in human transformation.

LIPyV shares approximately 65% sequence identity with RacPyV, a PyV that has been found in brain tumours from raccoons (Dela Cruz et al., 2013). This proximity with the RacPyV strains has been confirmed by a phylogenetic analysis based on the LT-Ag and VP1 ORFs and suggests an oncogenic potential of LIPyV. In the present study, human specimens collected from different anatomical sites in

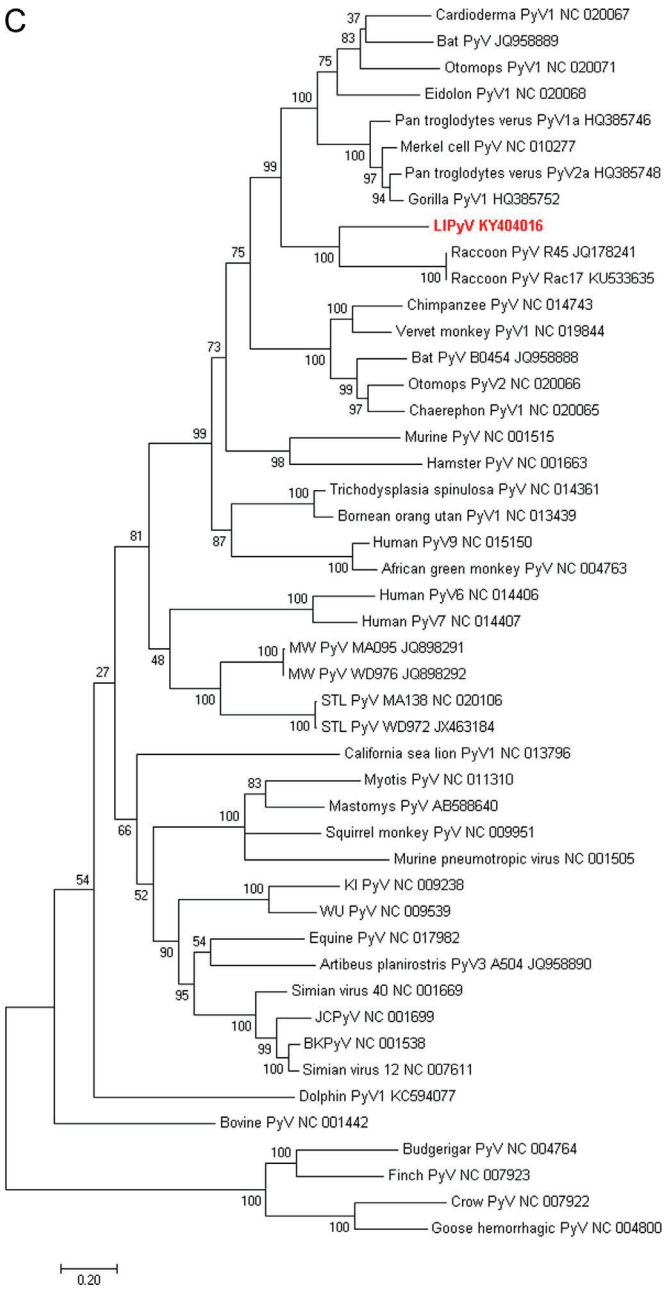


Fig. 3. (*continued*)

individuals from subjects in the USA and France were tested for LIPyV, showing a relatively low prevalence of approximately 2% in oral gargles and skin swabs. Interestingly, the eyebrow hair follicles and the skin swab collected from one individual were both positive for LIPyV, which may suggest a possible replication and shedding of this virus in the human host. Moreover, this prevalence is comparable to those observed for other human PyVs in stool, blood, cerebrospinal fluid, urine, or respiratory specimens (Li et al., 2015; Lim et al., 2013; Rockett et al., 2013; Siebrasse et al., 2012).

It is also possible that LIPyV has a tropism for other anatomical sites that still need to be elucidated. In addition, we cannot exclude the possibility that LIPyV is an animal virus and its presence in the human body may represent an environmental contamination. However, the detection of LIPyV DNA in eyebrow hair follicles does not support this hypothesis. Serological studies aiming to determine the presence of antibodies against LIPyV in human sera will further clarify this issue and will conclusively demonstrate whether LIPyV represents the 14th HPyV.

Acknowledgments

We are grateful to Mrs Nicole Suty for her help with manuscript preparation and Dr Karen Müller for editing. This study was supported in part by the European Commission, Grant HPV-AHEAD (FP7-HEALTH-2011-282562), by the Grant VIRUSCAN R01 (No. R01CA177586-01), and by a grant from Fondation ARC (No. PJA 20151203192).

References

Allander, T., Andreasson, K., Gupta, S., Bjerkner, A., Bogdanovic, G., Persson, M.A., Dalianis, T., Ramqvist, T., Andersson, B., 2007. Identification of a third human polyomavirus. J. Virol. 81, 4130–4136.

An, P., Saenz Robles, M.T., Pipas, J.M., 2012. Large T antigens of polyomaviruses: amazing molecular machines. Annu. Rev. Microbiol. 66, 213–236.

Azzi, A., Fanci, R., Bosi, A., Ciappi, S., Zakrzewska, K., de Santis, R., Laszlo, D., Guidi, S., Saccardi, R., Vannucchi, A.M., et al., 1994. Monitoring of polyomavirus BK viruria in bone marrow transplantation patients by DNA hybridization assay and by polymerase chain reaction: an approach to assess the relationship between BK

viruria and hemorrhagic cystitis. Bone Marrow Transplant. 14, 235–240.
Buck, C.B., Van Doorslaer, K., Peretti, A., Geoghegan, E.M., Tisza, M.J., An, P., Katz, J.P., Pipas, J.M., McBride, A.A., Camus, A.C., McDermott, A.J., Dill, J.A., Delwart, E., Ng, T.F., Farkas, K., Austin, C., Kraberger, S., Davison, W., Pastrana, D.V., Varsani, A., 2016. The ancient evolutionary history of polyomaviruses. PLoS Pathog. 12, e1005574.
Calvignac-Spencer, S., Feltkamp, M.C., Daugherty, M.D., Moens, U., Ramqvist, T., Johne, R., Ehlers, B., 2016. A taxonomy update for the family Polyomaviridae. Arch. Virol. 161, 1739–1750.
Carter, J.J., Daugherty, M.D., Qi, X., Bheda-Malge, A., Wipf, G.C., Robinson, K., Roman, A., Malik, H.S., Galloway, D.A., 2013. Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. In: Proceedings of the National Academy of Sciences of the United States of America, vol. 110. pp. 12744–12749.
Chestukhin, A., Litovchick, L., Rudich, K., DeCaprio, J.A., 2002. Nucleocytoplasmic shuttling of p130/RBL2: novel regulatory mechanism. Mol. Cell. Biol. 22, 453–468.
Cho, U.S., Morrone, S., Sablina, A.A., Arroyo, J.D., Hahn, W.C., Xu, W., 2007. Structural basis of PP2A inhibition by small t antigen. PLoS Biol. 5, e202.
Coleman, D.V., Mackenzie, E.F., Gardner, S.D., Poulding, J.M., Amer, B., Russell, W.J., 1978. Human polyomavirus (BK) infection and ureteric stenosis in renal allograft recipients. J. Clin. Pathol. 31, 338–347.
Combes, J.D., Dalstein, V., Gheit, T., Clifford, G.M., Tommasino, M., Clavel, C., Lacau St Guily, J., Franceschi, S., SPLIT study group. 2017. Prevalence of human papillomavirus in tonsil brushings and gargles in cancer-free patients: The SPLIT studyOral. Oncol. 66, 52–57.
Dalrymple, S.A., Beemon, K.L., 1990. BK virus T antigens induce kidney carcinomas and thymoproliferative disorders in transgenic mice. J. Virol. 64, 1182–1191.
DeCaprio, J.A., Garcea, R.L., 2013. A cornucopia of human polyomaviruses. Nature reviews. Microbiology 11, 264–276.
DeCaprio, J.A., Ludlow, J.W., Figge, J., Shew, J.Y., Huang, C.M., Lee, W.H., Marsilio, E., Paucha, E., Livingston, D.M., 1988. SV40 large tumor antigen forms a specific complex with the product of the retinoblastoma susceptibility gene. Cell 54, 275–283.
Dela Cruz, F.N., Jr., Giannitti, F., Li, L., Woods, L.W., Del Valle, L., Delwart, E., Pesavento, P.A., 2013. Novel polyomavirus associated with brain tumors in free-ranging raccoons, western United States. Emerg. Infect. Dis. 19, 77–84.
Edgar, R.C., 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform. 5, 113.
Edgar, R.C., 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Egli, A., Infanti, L., Dumoulin, A., Buser, A., Samaridis, J., Stebler, C., Gosert, R., Hirsch, H.H., 2009. Prevalence of polyomavirus BK and JC infection and replication in 400 healthy blood donors. J. Infect. Dis. 199, 837–846.
Ehlers, B., Moens, U., 2014. Genome analysis of non-human primate polyomaviruses. Infect. Genet. Evol.: J. Mol. Epidemiol. Evolut. Genet. Infect. Dis. 26, 283–294.
Feng, H., Shuda, M., Chang, Y., Moore, P.S., 2008. Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma 319. Science, New York, N.Y, 1096–1100.
Franceschi, S., Combes, J.D., Dalstein, V., Caudroy, S., Clifford, G., Gheit, T., Tommasino, M., Clavel, C., Lacau St Guily, J., Birembaut, P., 2015. Deep brush-based cytology in tonsils resected for benign diseases. Int. J. Cancer 137, 2994–2999.
Gardner, S.D., Field, A.M., Coleman, D.V., Hulme, B., 1971. New Human Papovavirus (B.K.) Isolated from Urine after Renal Transplantation 1. Lancet, London, England, 1253–1257.
Gaynor, A.M., Nissen, M.D., Whiley, D.M., Mackay, I.M., Lambert, S.B., Wu, G., Brennan, D.C., Storch, G.A., Sloots, T.P., Wang, D., 2007. Identification of a novel polyomavirus from patients with acute respiratory tract infections. PLoS Pathog. 3, e64.
Giuliano, A.R., Lee, J.H., Fulp, W., Villa, L.L., Lazcano, E., Papenfuss, M.R., Abrahamsen, M., Salmeron, J., Anic, G.M., Rollison, D.E., Smith, D., 2011. Incidence and Clearance of Genital Human Papillomavirus Infection in Men (HIM): a Cohort Study 377. Lancet, London, England, 932–940.
Guergnon, J., Godet, A.N., Galioot, A., Falanga, P.B., Colle, J.H., Cayla, X., Garcia, A., 2011. PP2A targeting by viral proteins: a widespread biological strategy from DNA/RNA tumor viruses to HIV-1. Biochim. Biophys. Acta 1812, 1498–1507.
Hampras, S.S., Giuliano, A.R., Lin, H.Y., Fisher, K.J., Abrahamsen, M.E., McKay-Chopin, S., Gheit, T., Tommasino, M., Rollison, D.E., 2015. Natural history of polyomaviruses in men: the HPV infection in men (HIM) study. J. Infect. Dis. 211, 1437–1446.
Hampras, S.S., Giuliano, A.R., Lin, H.Y., Fisher, K.J., Abrahamsen, M.E., Sirak, B.A., Iannacone, M.R., Gheit, T., Tommasino, M., Rollison, D.E., 2014. Natural history of cutaneous human papillomavirus (HPV) infection in men: the HIM study. PLoS One 9, e104843.
Ho, J., Jedrych, J.J., Feng, H., Natalie, A.A., Grandinetti, L., Mirvish, E., Crespo, M.M., Yadav, D., Fasanella, K.E., Proksell, S., Kuan, S.F., Pastrana, D.V., Buck, C.B., Shuda, Y., Moore, P.S., Chang, Y., 2015. Human polyomavirus 7-associated pruritic rash and viremia in transplant recipients. J. Infect. Dis. 211, 1560–1565.
Houben, R., Angermeyer, S., Haferkamp, S., Aue, A., Goebeler, M., Schrama, D., Hesbacher, S., 2015. Characterization of functional domains in the Merkel cell polyoma virus Large T antigen. Int. J. Cancer 136, E290–E300.
Houben, R., Adam, C., Baeurle, A., Hesbacher, S., Grimm, J., Angermeyer, S., Henzel, K., Hauser, S., Elling, R., Brocker, E.B., Gaubatz, S., Becker, J.C., Schrama, D., 2012. An intact retinoblastoma protein-binding site in Merkel cell polyomavirus large T antigen is required for promoting growth of Merkel cell carcinoma cells. Int. J. Cancer 130, 847–856.
Jiang, M., Abend, J.R., Johnson, S.F., Imperiale, M.J., 2009. The role of polyomaviruses in human disease. Virology 384, 266–273.
Johne, R., Wittig, W., Fernandez-de-Luco, D., Hofle, U., Muller, H., 2006. Characterization of two novel polyomaviruses of birds by using multiply primed rolling-circle amplification of their genomes. J. Virol. 80, 3523–3531.
Johne, R., Muller, H., Rector, A., van Ranst, M., Stevens, H., 2009. Rolling-circle amplification of viral DNA genomes using phi29 polymerase. Trends Microbiol. 17, 205–211.
Johne, R., Buck, C.B., Allander, T., Atwood, W.J., Garcea, R.L., Imperiale, M.J., Major, E.O., Ramqvist, T., Norkin, L.C., 2011. Taxonomical developments in the family Polyomaviridae. Arch. Virol. 156, 1627–1634.
Kean, J.M., Rao, S., Wang, M., Garcea, R.L., 2009. Seroepidemiology of Human polyomaviruses. PLoS Pathog. 5, e1000363.
Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: an Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data 28. Bioinformatics, Oxford, England, 1647–1649.
Koralnik, I.J., 2006. Progressive multifocal leukoencephalopathy revisited: has the disease outgrown its name? Ann. Neurol. 60, 162–173.
Korup, S., Rietscher, J., Calvignac-Spencer, S., Trusch, F., Hofmann, J., Moens, U., Sauer, I., Voigt, S., Schmuck, R., Ehlers, B., 2013. Identification of a novel human polyomavirus in organs of the gastrointestinal tract. PLoS One 8, e58021.
Kosugi, S., Hasebe, M., Tomita, M., Yanagawa, H., 2009b. Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. In: Proceedings of the National Academy of Sciences of the United States of America, vol. 106, pp. 10171–10176.
Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H., Miyamoto-Sato, E., Tomita, M., Yanagawa, H., 2009a. Six classes of nuclear localization signals specific to different binding grooves of importin alpha. J. Biol. Chem. 284, 478–485.
Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.
Kwun, H.J., Shuda, M., Camacho, C.J., Gamper, A.M., Thant, M., Chang, Y., Moore, P.S., 2015. Restricted protein phosphatase 2A targeting by Merkel cell polyomavirus small T antigen. J. Virol. 89, 4191–4200.
Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25, 1307–1320.
Li, K., Zhang, C., Zhao, R., Xue, Y., Yang, J., Peng, J., Jin, Q., 2015. The prevalence of STL polyomavirus in stool samples from Chinese children. J. Clin. Virol.: Off. Publ. Pan Am. Soc. Clin. Virol. 66, 19–23.
Lim, E.S., Reyes, A., Antonio, M., Saha, D., Ikumapayi, U.N., Adeyemi, M., Stine, O.C., Skelton, R., Brennan, D.C., Mkakosya, R.S., Manary, M.J., Gordon, J.I., Wang, D., 2013. Discovery of STL polyomavirus, a polyomavirus of ancestral recombinant origin that encodes a unique T antigen by alternative splicing. Virology 436, 295–303.
Miller, N.R., McKeever, P.E., London, W., Padgett, B.L., Walker, D.L., Wallen, W.C., 1984. Brain tumors of owl monkeys inoculated with JC virus contain the JC virus genome. J. Virol. 49, 848–856.
Mishra, N., Pereira, M., Rhodes, R.H., An, P., Pipas, J.M., Jain, K., Kapoor, A., Briese, T., Faust, P.L., Lipkin, W.I., 2014. Identification of a novel polyomavirus in a pancreatic transplant recipient with retinal blindness and vasculitic myopathy. J. Infect. Dis. 210, 1595–1599.
Moens, U., Johannessen, M., 2008. Human polyomaviruses and cancer: expanding repertoire. J. Dtsch. Dermatol. Ges.=J. Ger. Soc. Dermatol.: JDDG 6, 704–708.
Moens, U., Van Ghelue, M., Johannessen, M., 2007. Oncogenic potentials of the human polyomavirus regulatory proteins. Cell. Mol. Life Sci. : CMLS 64, 1656–1678.
Nickeleit, V., Singh, H.K., 2015. Polyomaviruses and disease: is there more to know than viremia and viruria? Curr. Opin. Organ Transplant. 20, 348–358.
Nunes, E.M., Sudenga, S.L., Gheit, T., Tommasino, M., Baggio, M.L., Ferreira, S., Galan, L., Silva, R.C., Pierce Campbell, C.M., Lazcano-Ponce, E., Giuliano, A.R., Villa, L.L., Sichero, L., 2016. Diversity of beta-papillomavirus at anogenital and oral anatomic sites of men: the HIM Study. Virology 495, 33–41.
Padgett, B.L., Walker, D.L., ZuRhein, G.M., Eckroade, R.J., Dessel, B.H., 1971. Cultivation of Papova-like Virus from Human Brain with Progressive Multifocal Leucoencephalopathy 1. Lancet, London, England, 1257–1260.
Peretti, A., FitzGerald, P.C., Bliskovsky, V., Pastrana, D.V., Buck, C.B., 2015. Genome Sequence of a Fish-Associated Polyomavirus, Black Sea Bass (Centropristis striata) Polyomavirus 1. Genome Announc., 3.
Pierce Campbell, C.M., Messina, J.L., Stoler, M.H., Jukic, D.M., Tommasino, M., Gheit, T., Rollison, D.E., Sichero, L., Sirak, B.A., Ingles, D.J., Abrahamsen, M., Lu, B., Villa, L.L., Lazcano-Ponce, E., Giuliano, A.R., 2013. Cutaneous human papillomavirus types detected on the surface of male external genital lesions: a case series within the HPV Infection in Men Study. J. Clin. Virol. : Off. Publ. Pan Am. Soc. Clin. Virol. 58, 652–659.
Pierce Campbell, C.M., Gheit, T., Tommasino, M., Lin, H.Y., Torres, B.N., Messina, J.L., Stoler, M.H., Rollison, D.E., Sirak, B.A., Abrahamsen, M., Carvalho da Silva, R.J., Sichero, L., Villa, L.L., Lazcano-Ponce, E., Giuliano, A.R., 2016. Cutaneous beta human papillomaviruses and the development of male external genital lesions: a case-control study nested within the HIM Study. Virology 497, 314–322.
Pipas, J.M., 1992. Common and unique features of T antigens encoded by the polyomavirus group. J. Virol. 66, 3979–3985.
Rector, A., Tachezy, R., Van Ranst, M., 2004. A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification. J. Virol. 78, 4993–4998.
Rennspiess, D., Pujari, S., Keijzers, M., Abdul-Hamid, M.A., Hochstenbag, M., Dingemans, A.M., Kurz, A.K., Speel, E.J., Haugg, A., Pastrana, D.V., Buck, C.B., De Baets, M.H., Zur Hausen, A., 2015. Detection of human polyomavirus 7 in human thymic epithelial tumors. J. Thorac. Oncol. : Off. Publ. Int. Assoc. Study Lung Cancer

10, 360–366.
Rockett, R.J., Sloots, T.P., Bowes, S., O'Neill, N., Ye, S., Robson, J., Whiley, D.M., Lambert, S.B., Wang, D., Nissen, M.D., Bialasiewicz, S., 2013. Detection of novel polyomaviruses, TSPyV, HPyV6, HPyV7, HPyV9 and MWPyV in feces, urine, blood, respiratory swabs and cerebrospinal fluid. PLoS One 8, e62764.
Sablina, A.A., Hahn, W.C., 2008. SV40 small T antigen and PP2A phosphatase in cell transformation. Cancer Metastasis Rev. 27, 137–146.
Schmitt, M., Bravo, I.G., Snijders, P.J., Gissmann, L., Pawlita, M., Waterboer, T., 2006. Bead-based multiplex genotyping of human papillomaviruses. J. Clin. Microbiol. 44, 504–512.
Schmitt, M., Dondog, B., Waterboer, T., Pawlita, M., Tommasino, M., Gheit, T., 2010. Abundance of multiple high-risk human papillomavirus (HPV) infections found in cervical cells analyzed by use of an ultrasensitive HPV genotyping assay. J. Clin. Microbiol. 48, 143–149.
Schowalter, R.M., Pastrana, D.V., Pumphrey, K.A., Moyer, A.L., Buck, C.B., 2010. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. Cell host Microbe 7, 509–515.
Schuurman, R., Sol, C., van der Noordaa, J., 1990. The complete nucleotide sequence of bovine polyomavirus. J. Gen. Virol. 71 (Pt 8), 1723–1735.
Scuda, N., Hofmann, J., Calvignac-Spencer, S., Ruprecht, K., Liman, P., Kuhn, J., Hengel, H., Ehlers, B., 2011. A novel human polyomavirus closely related to the african green monkey-derived lymphotropic polyomavirus. J. Virol. 85, 4586–4590.
Shollar, D., Del Valle, L., Khalili, K., Otte, J., Gordon, J., 2004. JCV T-antigen interacts with the neurofibromatosis type 2 gene product in a transgenic mouse model of malignant peripheral nerve sheath tumors. Oncogene 23, 5459–5467.
Shuda, M., Feng, H., Kwun, H.J., Rosen, S.T., Gjoerup, O., Moore, P.S., Chang, Y., 2008. T antigen mutations are a human tumor-specific signature for Merkel cell polyomavirus. In: Proceedings of the National Academy of Sciences of the United States of America, vol. 105, pp. 16272–16277.
Shuda, M., Guastafierro, A., Geng, X., Shuda, Y., Ostrowski, S.M., Lukianov, S., Jenkins, F.J., Honda, K., Maricich, S.M., Moore, P.S., Chang, Y., 2015. Merkel cell Polyomavirus small T antigen induces cancer and embryonic Merkel cell proliferation in a transgenic mouse model. PLoS One 10, e0142329.
Shuda, M., Arora, R., Kwun, H.J., Feng, H., Sarid, R., Fernandez-Figueras, M.T., Tolstov, Y., Gjoerup, O., Mansukhani, M.M., Swerdlow, S.H., Chaudhary, P.M., Kirkwood, J.M., Nalesnik, M.A., Kant, J.A., Weiss, L.M., Moore, P.S., Chang, Y., 2009. Human Merkel cell polyomavirus infection I. MCV T antigen expression in Merkel cell carcinoma, lymphoid tissues and lymphoid tumors. Int. J. Cancer 125, 1243–1249.
Siebrasse, E.A., Reyes, A., Lim, E.S., Zhao, G., Mkakosya, R.S., Manary, M.J., Gordon, J.I., Wang, D., 2012. Identification of MW polyomavirus, a novel polyomavirus in human stool. J. Virol. 86, 10321–10326.
Simmons, D.T., Loeber, G., Tegtmeyer, P., 1990. Four major sequence elements of simian virus 40 large T antigen coordinate its specific and nonspecific DNA binding. J. Virol. 64, 1973–1983.
Sroller, V., Hamsikova, E., Ludvikova, V., Musil, J., Nemeckova, S., Salakova, M., 2016. Seroprevalence rates of HPyV6, HPyV7, TSPyV, HPyV9, MWPyV and KIPyV polyomaviruses among the healthy blood donors. J. Med. Virol. 88, 1254–1261.
Toptan, T., Yousem, S.A., Ho, J., Matsushima, Y., Stabile, L.P., Fernandez-Figueras, M.T., Bhargava, R., Ryo, A., Moore, P.S., Chang, Y., 2016. Survey for human polyomaviruses in cancer. JCI Insight, 1.
van der Meijden, E., Kazem, S., Dargel, C.A., van Vuren, N., Hensbergen, P.J., Feltkamp, M.C., 2015. Characterization of T antigens, including middle T and alternative T, expressed by the human polyomavirus associated with trichodysplasia spinulosa. J. Virol. 89, 9427–9439.
van der Meijden, E., Janssens, R.W., Lauber, C., Bouwes Bavinck, J.N., Gorbalenya, A.E., Feltkamp, M.C., 2010. Discovery of a new human polyomavirus associated with trichodysplasia spinulosa in an immunocompromized patient. PLoS Pathog. 6, e1001024.
Varakis, J., ZuRhein, G.M., Padgett, B.L., Walker, D.L., 1978. Induction of peripheral neuroblastomas in Syrian hamsters after injection as neonates with JC virus, a human polyoma virus. Cancer Res. 38, 1718–1722.
Verhaegen, M.E., Mangelberger, D., Harms, P.W., Vozheiko, T.D., Weick, J.W., Wilbert, D.M., Saunders, T.L., Ermilov, A.N., Bichakjian, C.K., Johnson, T.M., Imperiale, M.J., Dlugosz, A.A., 2015. Merkel cell polyomavirus small T antigen is oncogenic in transgenic mice. J. Investig. Dermatol. 135, 1415–1424.
Walker, D.L., Padgett, B.L., ZuRhein, G.M., Albert, A.E., Marsh, R.F., 1973. Human Papovavirus (JC): Induction of Brain Tumors in Hamsters 181. Science, New York, N.Y, 674–676.
Yu, J., Boyapati, A., Rundell, K., 2001. Critical role for SV40 small-t antigen in human cell transformation. Virology 290, 192–198.
Zu Rhein, G.M., Varakis, J.N., 1979. Perinatal induction of medulloblastomas in Syrian golden hamsters by a human polyoma virus (JC). Natl. Cancer Inst. Monogr., 205–208.

Merkel Cell Polyomavirus downregulates N-myc downstream regulated gene-1 (NDRG1) leading to cellular proliferation and migration

Purnima Gupta[a], Naveed Shahzad[a*], Alexis Harold[b], Masahiro Shuda[b], Assunta Venuti[a], Maria Carmen Romero-Medina[a], Laura Pacini[a*], Lise Brault[a*], Alexis Robitaille[a], Valerio Taverniti[a], Hector Hernandez-Vargas[c*], Geoffroy Durand[d], Florence Le Calvez-Kelm[d], Tarik Gheit[a], Rosita Accardi[a*], Massimo Tommasino[a#]

[a]Infections and Cancer Biology Group, International Agency for Research on Cancer, Lyon, France

[b]Cancer Virology Program, University of Pittsburgh Medical Center Hillman Cancer Center, Pittsburgh, Pennsylvania, USA

[c]Epigenetics Group, International Agency for Research on Cancer, Lyon, France

[d]Genetic Cancer Susceptibility Group, International Agency for Research on Cancer, Lyon, France

Running Head: MCPyV targets NDRG1 to induce cellular proliferation

[#]Address correspondence to Massimo Tommasino, tommasino@iarc.fr

*Present address:

Naveed Shahzad, School of Biological Sciences, University of the Punjab, Lahore, Pakistan

Hector Hernandez-Vargas, Cancer Research Centre of Lyon (CRCL), Inserm U 1052, CNRS UMR 5286, Centre Léon Bérard, Université de Lyon, 28 Rue Laennec, 69008 Lyon, France

Laura Pacini, Division of Molecular Pathology, The Institute of Cancer Research, London, SW3 6JB, United Kingdom

Lise Brault, INSERM, Aix Marseille Univ, CNRS, Institut Paoli-Calmettes, CRCM (Signaling, Hematopoiesis and Mechanism of Oncogenesis), Marseille, France

Rosita Accardi, Epigenetics Group, International Agency for Research on Cancer, Lyon, France

P.G. and N.S. contributed equally to this work.

Key words: Merkel Cell Polyomavirus, NDRG1, Keratinocytes, Cellular proliferation, Gene expression profile

Abstract

Merkel cell polyomavirus (MCPyV) is the first human polyomavirus etiologically associated with Merkel cell carcinoma (MCC), a rare and aggressive form of skin cancer. Similar to other polyomaviruses, MCPyV encodes early T antigen genes, a viral oncogene required for MCC tumor growth. To identify the unique oncogenic properties of MCPyV, we analysed the gene expression profiles in human spontaneously immortalized keratinocytes (NIKs) expressing the early genes from six distinct human polyomaviruses (PyVs), including MCPyV. A comparison of the gene expression profiles revealed 28 genes specifically deregulated by MCPyV. In particular, the MCPyV early gene downregulated the expression of the tumor suppressor gene N-myc downstream regulated gene-1 (NDRG1) in NIKs-MCPyV and hTERT-MCPyV human keratinocytes (HK) as compared to their controls. In MCPyV-positive MCC cells, the expression of NDRG1 was downregulated by the MCPyV early gene, as T antigen knockdown rescued the level of NDRG1. In addition, NDRG1 overexpression in hTERT-MCPyV HK or MCC cells resulted in decrease of cells in S phase and cell proliferation inhibition. Moreover, a decrease in wound healing capacity in hTERT-MCPyV HK was observed. Further analysis revealed that NDRG1 exerts its biological effect in Merkel cell lines by regulating the expression of CDK2 and cyclinD1 proteins. Overall NDRG1 plays an important role in MCPyV induced cellular proliferation.

Importance

Merkel Cell Carcinoma was first described in 1972 as a neuroendocrine tumor of skin most of which in 2008, was reported to be caused by a PyV named Merkel Cell Polyomavirus (MCPyV), the first PyV linked to human cancer. Thereafter, numerous studies have been conducted to understand the etiology of this virus induced carcinogenesis. However, it is still a new field and much work is needed to understand the molecular pathogenesis of MCC. In the current work, we sought to identify the host genes specifically deregulated by MCPyV as opposed to other PyVs in order to better understand the relevance of the genes analyzed on the biological impact and progression of the disease. These findings open newer avenues for targeted drug therapies thereby providing hope for management of patients suffering from this highly aggressive cancer.

Introduction

For nearly 40 years, BK polyomavirus (BKPyV) and JC polyomavirus (JCPyV) have been the only known human polyomaviruses (PyVs). During the last decade, 11 new human PyVs, including KI polyomavirus (KIPyV), WU polyomavirus (WUPyV), Merkel cell polyomavirus (MCPyV), Human polyomavirus 6 (HPyV6), Human polyomavirus 7 (HPyV7), Human polyomavirus 9 (HPyV9), New Jersey Polyomavirus (NJPyV), Trichodysplasia spinuolsa-associated polyomavirus (TSPyV), Malawi polyomavirus (MWPyV), HPyV12, and St Louis polyomavirus (STLPyV) have been discovered (1). To add to this list, a putative human PyV named IARC-Lyon PyV (LIPyV) was recently isolated from human skin (2). Although there is little information known about the pathogenesis of these novel human PyVs, some of them have been linked to human diseases: BKV-associated nephropathy, JCV-associated progressive multifocal leukoencephalopathy (PML), WU-PyV-associated bronchitis, HPyV6/HPyV7-associated dermatosis and TSPyV-associated trichodysplasia spinulosa. MCPyV is responsible for an aggressive type of skin cancer called Merkel cell carcinoma (MCC) (3, 4), as its presence has been confirmed in 80% of MCC cases (5). MCPyV is clonally integrated into the host cancer cell genome and harbours mutations that impair viral replication activity, which are important events in MCPyV-mediated MCC (4).

PyVs are small (40-50 nanometres in diameter), non-enveloped, double-stranded (ds) DNA viruses that have a capsid with icosahedral symmetry (6, 7). All PyVs harbour a circular genome with a size ranging from 4.5 to 5.4 kbp, which can be divided into two oppositely oriented protein encoding regions: the early T antigen gene and the late VP gene, separated by a non-coding control segment (8). Alternative splicing of all PyVs early genes gives rise to viral proteins large T-antigen (LT) and small T-antigen (ST) while late region encodes viral capsid forming proteins VP1, VP2, and VP3 (9, 10). MCPyV-positive MCCs express the early T antigen gene and require the early gene products for tumor cell survival and proliferation (11).

Although LT and ST of almost all known human PyVs have displayed transforming properties in various experimental models, the oncogenic potential varies among these proteins (12). Unlike other human PyVs where LT acts as a major transforming protein, MCPyV ST has shown strong transforming activities (13), indicating that in some aspects MCPyV behaves differently from other human PyVs (14). In addition, the proven implication of only one PyV (MCPyV) in human carcinogenesis raises the question, why MCPyV alone is carcinogenic and not the other human PyVs. Even though some efforts were made to explain the structural and biological differences between MCPyV and other human PyVs (15, 16), the mechanistic aspect which renders MCPyV to be carcinogenic is not very well defined.

Virus-mediated host genetic alteration is one of the major contributing factors in virus-induced cancers. The aim of this study is to investigate the effect of human PyVs early genes on host global gene expression with a focus on identification of the genes uniquely altered by the expression of MCPyV early genes that could confer the transforming abilities to MCPyV. Of the numerous genes identified in this study, N-myc downstream regulated gene-1 (NDRG1) was downregulated by MCPyV T antigen. Interestingly, previous studies have shown that NDRG1 functions as a metastatic suppressor (17-20) and a transcriptional repressor and is involved in cell cycle inhibition (21, 22), cellular differentiation (20) and in apoptosis (23). However, nothing is known about the role of this protein and the underlying mechanism by which it affects the cell transforming ability of MCPyV. Taken together, our work identifies significant changes in gene expression upon MCPyV oncogene expression which may explain the difference in the carcinogenic potential of MCPyV compared to other human PyVs. In addition, altered expression of NDRG1 by MCPyV provides mechanistic insight in host-cell signaling deregulation in MCPyV-mediated cellular proliferation.

Results

Early gene products from different human PyVs differently affect cellular gene expression

To determine whether some human PyVs differ in their ability to alter cellular gene expression, we generated normal immortalized keratinocyte (NIK) lines expressing the entire early region of five human PyVs, namely, BKV, JCV, KIV, MCV, and WUV. In addition, SV40 was included in the study due to its well characterized transforming activities in experimental models. NIKs were transduced with recombinant retroviruses containing the early regions of the six PyVs and the relative and absolute expression levels of sT and LT were found to be nearly the same as shown in our previously published data (24). The expression profiles of 24000 annotated genes were determined by Illumina microarray (Illumina HT-12 v4). The microarray data were first subjected to quality control where 3 different methods namely unsupervised hierarchical clustering of the duplicates, scatterplots, and boxplots of gene expression data, were applied. All samples analysis was performed in duplicates except for MCPyVs with its corresponding negative control for which a quadruplicate was analyzed. Unsupervised hierarchical clustering was performed on genes that passed the filtering criteria revealed that all the samples clustered separately with their duplicates and quadruplicates clustering next to each other (Figure 1A).

Subsequently, we compared the expression profile data of each PyV with the expression profile of the negative control, i.e. NIKs transduced with empty retrovirus (pLXSN). The expressions of genes are mentioned in ratios of the values obtained relative to the control condition after

normalization of the data. For comparison between these classes, genes were considered differentially expressed when they displayed a difference of at least 1.5-fold increase or decrease in expression pattern in both replicates, with a p and false discover rate (FDR) values < 0.001. Using these selection criteria, we identified numerous genes deregulated by each PyV upon comparison with negative control (Figure 1B). Notably, most of the genes were downregulated in each class comparison except for WUV, for which the upregulated genes were higher than the downregulated ones. However, SV40 scored a maximum for the deregulation of genes (n=967) while MCPyV scored for 325 genes on list (n=325).

A comparison of the genes deregulated by MCPyV in this study with the published datasets from 3 different studies (25-27) revealed a total of 73 genes to be commonly deregulated in our and at least one of the previous studies (Figure 1C). Only 1 gene was found unanimously deregulated in all studies included in the comparative analysis, HIST1H1C. The encoded protein is involved in cell senescence, DNA repair and cell cycle. A comprehensive list of the 73 genes was prepared which is included as a Supplemental Table 1. Strikingly, the pathway analysis of the 73 genes showed genes regulating mostly the cell senescence, DNA repair, cell cycle and signal transduction pathways which are included in Supplemental Table 2.

Next, we focused on the expression profile induced by MCPyV, since it is the only PyV clearly associated with human carcinogenesis. The MCPyV-deregulated genes (n=325) were subjected to Biocarta pathways analysis. Although a significant number of altered genes by MCPyV are functionally related to many pathways, the cell cycle regulation and MAP kinase pathways ranked the highest with a significant number of variable genes (Supplemental Table 3). Our observation revealed essential genes involved in cell cycle regulation, particularly at the G1/S phase, are modulated by the expression of MCPyV early genes (Figure 1D). These genes include cyclins, cyclin dependent kinases (CDKs) and cyclin dependent kinase inhibitors (28). These results further strengthen the notion that cell cycle deregulation could be one of the major driving factors in MCPyV-mediated carcinogenesis.

Comparative analyses of deregulated cellular gene expression mediated by the 6 PyVs

Next, we determined whether the MCPyV displays unique features in deregulating cellular geneexpression in comparison to the other PyVs. For this purpose, single class comparisons between 6 PyVs early genes expressed in NIKs and NIKs/ pLXSN control, were followed by Venn intersections of the 6 datasets using R scripts. This led to the identification of a total of 23 genes, namely: C12orf24, C1orf116, C9orf41, CCNA1, CDR2L, CTSH, DLK2, ECM1, FOXQ1, INPP4B, KIAA0101, KIF13B, KLF6, LIPG, MXRA5, NDRG1, PTPRE, PYGB, S100A16, SH3KBP1, SLC1A3, TRIB1, and UGT1A6, specifically altered by MCPyV while 60 other

deregulated genes were common to all PyVs (Figure 2A). In addition, 97, 44, 25, 285, and 398 genes were specifically deregulated by BKV, JCV, KIV, WUV and SV40, respectively. To further elucidate the MCPyV signature genes, BRB-ArrayTools were used whereby a single class comparison between MCPyV and pLXSN were made after subtracting the background genes and the genes resulting from the class comparisons of BKV, JCV, KIV, WUV, SV40, and the negative control, pLXSN. Interestingly, this method also showed that MCPyV exclusively and significantly altered 28 genes in comparison to other PyVs (Figure 2B). Notably, 23 of 28 MCPyV-deregulated genes identified with this analysis, were also found by Venn Diagram intersections. However, an increase of 5 significant genes namely, SPRR2E, CTSC, ANXA2, PTGS1, and DUSP10 was seen in the latter approach. Further to understand whether these 23 genes are specific to MCPyV mediated deregulation, we did a comparative analysis between the published datasets for SV40 (29, 30), BKV (31, 32) and JCV (33, 34) and found 2 genes CCNA1 and LIPG commonly deregulated for SV40 and 3 genes KIAA0101, MXRA5 and SLC1A3 commonly deregulated for BKV (Figure 2C and 2D). However, no gene was found to be commonly deregulated for JCV.

We also evaluated whether the products of these MCPyV-deregulated genes are involved in crucial cellular pathways by using GeneOntology software. The analysis showed that 19 of these genes were involved in the response to stimuli and in the regulation of processes linked to cellular transformation (Figure 2E). Importantly, as shown in Supplemental Table 4, some of these genes have been found deregulated in different types of human cancers. Reactome Pathway analysis using 23 genes revealed glucuronidation, biological oxidation and Tp53 and G1-S mediated transcription to be the top 5 pathways regulated by them (Figure 2F). Together, these results show that the products of the MCPyV early gene have a unique property to deregulate cellular gene expression in comparison to other PyVs.

Validation of the role of MCPyV early proteins in altering cellular gene expression

In order to confirm the microarray data, we performed quantitative RT-PCR. Of the 28 MCPyV-deregulated genes, we selected 5 (NDRG1, KLF6, TRIB1, INPP4B and ANX2A) based on their biological functions as major tumor suppressors as described in Supplemental Table 4. Quantitative RT-PCR confirmed that all 5 genes were downregulated in NIKs in the presence of the viral genes (Figure 3A). Similar findings were obtained when MCPyV early genes were expressed in hTERT- human keratinocytes. NDRG1, KLF6, TRIB1, and INPP4B, but not ANXA2, were significantly downregulated in the presence of viral early genes (Figure 3B). To corroborate our findings, we also determined whether silencing the expression of MCPyV large and/or small antigens (LT and ST, respectively) in the MCPyV-positive MCC cell line (MKL-1)

could influence the expression of the 5 selected genes. Figure 3C shows that silencing the expression of ST alone or both ST and LT (PAN) resulted in a significant increase of NDRG1, KLF6, and INPP4B. There were no observed changes in their transcript levels in the MCPyV-negative cell line, UISO, when transduced with ST or PAN shRNAs (Figure 3D).

When comparing the data obtained in the different cell lines, it indicated NDRG1 is the most consistent MCPyV-deregulated gene in keratinocytes and in the Merkel cell carcinoma-derived cell line. Moreover, NDRG1 is participating in the second most important pathway, being the transcriptional regulation of cell death genes by Tp53, as revealed by the Reactome Pathway analysis (Figure 2F). To further confirm these observations, we determined the NDRG1 protein levels in the three experimental models described above. Silencing the early gene expression by PAN shRNA in MKL-1 or expression of MCPyV early genes in NIKs as well as in hTERT-HK resulted in a rescue or decrease in NDRG1 protein levels respectively, (Figure 3E). Together, this data show that MCPyV ST and LT can downregulate the NDRG1 mRNA and protein levels.

Ectopic expression of NDRG1 in cells expressing early genes of MCPyV or MCC cell lines inhibits cellular proliferation and migration

Next, we aimed to understand the biological significance of MCPyV-mediated NDRG1 downregulation. hTERT-HK, previously transduced with recombinant retroviruses containing the MCPyV early gene (MCPyV-hTERT HK), were transfected with the NDRG1 pBABE vector (Figure 4A). Ectopic expression of NDRG1 decreased the number of colonies by approximately 50% (Figure 4B and 4C). To corroborate these findings, MCC cell lines, MKL-1 and MKL-2 were transduced with lentiviruses expressing NDRG1 under the control of a doxycycline-inducible promoter. Induction of NDRG1 expression by doxycycline resulted in a decrease in cellular proliferation of the MCC cell lines. At day 12, MKL-1 and MKL-2 showed a 31.9% and 34.4% decrease compared to doxycycline treated controls respectively ($p<0.05$) (Figures 4D and E). Interestingly, the cellular morphology of the two MCC cell lines showed distinct features. MKL-1 cell lines, upon NDRG1 induction, form smaller cellular aggregates compared to the controls whereas MKL-2 cells showed larger clumps (Figure 4F). Even though NDRG1 decreases the overall cellular proliferation in MCC cell lines, these results indicate NDRG1 differentially impacts the cellular physiology in the two cell lines.

As NDRG1 is also implicated as a metastasis suppressor (35), we evaluated whether NDRG1 influences the cell migration in our experimental models. MCPyV-hTERT HK were transiently transfected with the NDRG1 overexpression vector and a wound healing assay was performed every 24 h for a period of 2 days. We observed nearly complete wound closure (94.9%) in MCPyV-hTERT HK control cells after 48 h whereas we only observed 26.1%, wound closure in

NDRG1 overexpressing MCPyV-hTERT HK (Figure 4G and 4H). Together, these results show that NDRG1 plays an important role in cellular proliferation and migration.

NDRG1 overexpression differentially regulates cell cycle in MCC cell lines

Analysis of the cell cycle profile by flow cytometry showed that NDRG1 overexpression resulted in a modest, but reproducible decrease of cells in S/G2 phase (22.4% decrease; Figure 5A and 5B). In addition, an increase of sub-G0 population were observed in the same cells in presence of ectopic NDRG1 levels (2.1-fold, $p<0.01$ as compared to control; Figure 5A and 5C). Interestingly, there were not many differences observed in the cell cycle profile of the MCC cell lines (Figure 5D). This can be attributed to how they are slow cycling cells with doubling times of nearly 3 days. However, the BrdU incorporation assay also did not show many changes in MKL-1 cell lines overexpressing NDRG1 (Figure 5D, left panel and Figure 5E) whereas MKL-2 showed a significant decrease (49.1% decrease over control, $p<0.05$, Figure 5D, right panel and Figure 5E). This data highlights a different mechanism in the NDRG1-mediated inhibition of cellular proliferation in hTERT-HK and Merkel cancer-derived cell lines.

NDRG1 regulates expression of key cell cycle regulators, CDK2 and cyclinD1

After studying the role of NDRG1 in cell cycle regulation, we aimed to investigate the expression of cell cycle regulators in the presence of NDRG1. First, we determined by immunoblotting whether the expression of MCPyV early genes in hTERT HK influences the protein levels of positive regulators of the cell cycle, namely cyclin D1 and cyclin-dependent-kinase 2 (CDK2). Figure 6A shows the expression of viral genes resulted in a significant increase of cyclin D1, and CDK2 protein levels (6.6- and 3.5-fold respectively compared to control, $p<0.001$, 0.01 respectively). Overexpression of NDRG1 in the same cells partially reduced the levels of these cellular proteins (43.6% and 49.6% respectively; Figure 6B). However, as seen in the BrdU incorporation assay, MCC cell lines, MKL-1 and MKL-2 transduced with lentiviruses expressing NDRG1 in the presence of doxycycline showed differential regulation of cyclin D1 and CDK2 expression. A significant decrease in cyclin D1 expression along with a modest decrease in CDK2 expression was observed for MKL-2 whereas no changes in expression of the two proteins were observed for MKL-1 cell lines (Figure 6C). This justifies the observed discrepancies in the BrdU incorporation observed for the MCC cell lines.

Furthermore, to understand whether the MCPyV early genes expression has any impact on the regulation of NDRG1, CDK2 and cyclin D1 in different MCC cell lines, we knocked-down the early genes in MKL-1, MKL-2, MS-1 and CVG-1 and checked their protein expression by western blot. The immunoblot shows an increase in NDRG1 expression in all the four cell lines

upon transduction with PAN shRNA compared to controls (Figure 6D-G). However MKL-2, MS-1 and CVG-1 showed decrease in cyclin D1 and CDK2 levels upon ST and LT knockdown (Figure 6D-G). Thus, while in the less transformed hTERT-HK NDRG1 overexpression resulted in a reduction in the levels of positive cell cycle regulators, this phenomenon is only conserved in some Merkel cancer-derived cell lines, possibly due to a different status of their cellular transformation. In any case, silencing the expression of the viral oncogenes rescues the NDRG1 protein levels in all analyzed Merkel cancer-derived cell lines.

To further characterize the role of cyclin D1 in proliferation of cells expressing the MCPyV early genes, we silenced cyclin D1 expression in MCPyV-hTERT HK by siRNA (Figure 6H). Although we did not observe significant changes in total number of colonies of the mock and siRNA cyclin D1 cells, silencing of cyclin D1 expression resulted in a strong reduction of colony size (Figure 6I and J), highlighting the inhibition of cellular proliferation. These results further indicated and validated our previous observation in Figure 5D, that NDRG1 might involve multiple downstream effector molecules.

Discussion

In this study, we performed a comparative gene expression profiling of cells expressing the early genes of 6 PyVs namely BKV, JCV, KIV, MCPyV, SV40, and WUV, with the aim of identifying the unique features of MCPyV that endows it with oncogenic characteristics. Our results showed that in comparison to other 5 PyVs, MCPyV uniquely deregulated 28 genes with 13 genes showing upregulation and 15 genes displaying downregulation. We hypothesized specific deregulation of these 28 genes could be due to the direct oncogenic potential of the MCPyV early protein activities, as the transforming nature of these oncoproteins have been previously described (13). Also, comparisons between our datasets and datasets from different publications revealed differences in the deregulated gene list. This may be due to the experimental models, cell types, type of transfection, technology used to appreciate the gene expression changes, which are not exactly the same compared to ours, leading to a "not so perfect" overlap. The same can also be considered true for gene list comparisons we found for SV40, BKV and JCV. In fact, MCPyV distinctly downregulated certain genes, including tumor suppressor genes such as NDRG1, INPP4B, KLF6, TRIB1 and ANXA2. The reduced expression of these tumor suppressor genes has been reported in several cancer types including lung, prostate, and breast cancer (36-38). Strikingly, the genes which were specifically upregulated by MCPyV also included oncogenes such as FOXQ1, DUSP10, and CTSH. The unique ability of MCPyV to suppress tumor suppressor genes and activate oncogenes could explain its high

oncogenic potential in comparison to other human PyVs, as viral-mediated cancers are thought to be the result of accumulation of genetic alterations induced by their oncoproteins.

The genetic alterations disturbing cell cycle regulation are one of the leading causes for cancer development. In fact, cell cycle progression is a firmly controlled process where cyclins, cyclins dependent kinases (CDKs), and CDK interacting protein/kinase inhibitory proteins (cip/kip family) coordinate to ensure the proper transition of the cell cycle across cell cycle checkpoints (28). Certain oncoviruses have evolved diverse strategies to deregulate the cell cycle progression because the loss of proper cell cycle control is one of the major driving forces in cellular transformation (39). In fact, dsDNA viruses, such as human papilloma viruses (HPVs) and PyVs, depend on cell cycle machinery for replication. Therefore, they need to push the cell into S-phase of the cell cycle. Although the PyV early region gene products have been reported to target cellular proteins implicated in cell cycle regulation, there is little known about cell cycle regulation for cells infected with MCPyV (1).

Our results showed the significant modulation in the expression of genes related to particular cell cycle genes involved in the G1/S checkpoint, showing cell cycle regulation is a highly influenced pathway in MCPyV infected cells. Several CDK inhibitors, including cyclin-dependent kinase inhibitor 1A (CDKN1A), cyclin-dependent kinase inhibitor 2B (CDKN2B), cyclin-dependent kinase inhibitor 2C (CDKN2C), and cyclin-dependent kinase inhibitor 2D (CDKN2D) are known to prevent the progression of the cell cycle, are strongly downregulated by MCPyV (Supplemental Table 3 and Figure 1D). Interestingly, the regulation of CDKN1A by KLF6 has been reported (37), which we also found to be strongly downregulated by the MCPyV. In addition to some other CDKs, CDK4 is strongly upregulated in the presence of MCPyV early genes (Figure 1D). Furthermore, there was an observable slight upregulation of some cyclins, including CCNE1 (cyclin E1), CCND3 (cyclin D3), CCND2 (cyclin D2), and core cell cycle regulation gene CDC25, with a marginal suppression of Retinoblastoma 1 (RB1) in the presence of MCPyV early genes. All these indicators of cell cycle modulations suggest that MCPyV favor S-phase progression. However, the impact of MCPyV on cell cycle deregulation needs to be further elucidated. There is evidence that MCPyV early proteins promote the cell growth as illustrated by the reported robust cell death and cell cycle arrest associated to the inhibition of early gene expression, using PAN shRNA in MCC cell lines (13). The requirement of an intact pRb binding site by MCPyV to induce cellular growth shed light on the fact that MCPyV targets the cell cycle to increase its replication (40). Moreover, a MCPyV-infected MCC cell line showed impaired cell cycle arrest following exposure to UV radiation. Subsequently linked to LT these results suggest that the presence of the virus affects the normal cell cycle in cells exposed to UV (41).

Our results further indicated that among the genes deregulated specifically by MCPyV, NDRG1 was reproducibly downregulated in all model systems. NDRG1 is known to be a tumor suppressor and metastasis suppressor in a variety of cancer cells like brain, breast, colon and rectum, pancreas, prostate and esophagus whereas it is also known to promote tumorigenesis in some other cancer forms of kidney, liver, mouth, skin and uterine cervix (42). The mechanism by which NDRG1 exerts its effect is largely dependent on the cellular context (42). In MCF-7 breast and EJ bladder cancer cell lines, NDRG1 overexpression was shown to decrease cellular proliferation (21). Similarly, in pancreatic cancers, NDRG1 overexpression led to an inhibition of tumor growth and an increase in apoptosis (43). Keeping in line with these observations, our results also showed that NDRG1 overexpression served to induce cell arrest in MCPyV early gene-transduced cells as evident from the reduced number of cells in S/G2 phase compared to control. These results further imply MCPyV downregulates NDRG1 to aid cell cycle progression thereby promoting cell survival.

Our investigation further revealed the wound healing capacity to be severely compromised in Merkel positive cells overexpressing NDRG1. However, majority of the works with NDRG1 fail to unravel the underlying mechanism responsible for these diverse biological effects. One of the reports suggests that NDRG1 expression results in ATF3 mediated expression of the KAI gene to inhibit metastasis in prostate cancer (44). Another study highlights the upregulation of PTEN, SMAD4, and NEDD4L as important contributors of anti-tumor effect of NDRG1 (45). However, a number of studies support the fact that NDRG1 via interacting with either the Wnt receptor (46) or GSK3β (47) leads to regulation of β-catenin distribution and activity affecting cell proliferation and migration. We show here that NDRG1 overexpression has important effects on the expression of CDK2 and cyclinD1 which are important regulators of cell cycle check point.

Even though the works in MCPyV-hTERT-HK showed clear roles of NDRG1 in limiting cellular proliferation via inhibiting cell cycle progression, NDRG1 overexpression in MCC cell lines shows that differential mechanisms may exist by which NDRG1 mediates its effects in MKL-1 and MKL-2 cell lines. This may be partly attributed to the fact that cell lines may have accumulated changes over period owing to continuous passage leading to differential outcomes. However, as seen with 3 of 4 MCC cell lines, MKL-2, MS-1 and CVG-1 there was a decrease in expression of CDK2 and cyclin D1 upon early gene knockdown along with increase in NDRG1 expression indicating a redundant role of NDRG1 in regulating these cell cycle regulators. Further studies are needed to understand what features distinguish one MCC cell line from others leading to the observed discrepancies in terms of NDRG1. This is in line with previous observation where expression of a protein survivin was differentially regulated in MS-1 as opposed to three other cell lines under study including MKL-1 (48).

Recently, it was shown that NDRG1 is important in HCV assembly by regulating the biogenesis of lipid droplets, considered to be the main site for virus assembly (49). In yet another study, it has been described that miRNAs encoded by EBV can downregulate NDRG1, to promote EBV-mediated epithelial carcinogenesis (50). This might be another aspect of regulation of viral loads for which the viruses have developed strategies to downregulate this versatile tumor suppressor. While this feature of NDRG1 was beyond the scope of present investigation, it will be interesting to have a deeper understanding of the role of NDRG1 in MCPyV viral assembly as little is known in this area. Moreover, the mechanism of NDRG1 downregulation by viral early genes needs to be investigated as a clear decrease of the NDRG1 expression was seen at both transcriptional and translational levels.

In summary, we showed that MCPyV deregulates the cell cycle by specifically modulating genes associated with cell cycle regulation and MAPK pathways, and that NDRG1 is a key player in cell arrest and migration by mediating its effect in downregulating cyclinD1 and CDK2 in Merkel cell carcinoma.

Materials and Methods

Expression vectors

All expression vectors for early genes of BKV, JCV, KIV, SV40, WUV were prepared as mentioned before (24). Full-length MCPyV early genes were a kind gift from Dr. D.A. Galloway (Fred Hutchinson Cancer Research Center, Seattle, USA). NDRG1 of 1185bp length was subcloned in retroviral vector pBABE-Hyg or lentiviral vector pLenti TRE empty EF-puro (51). To knockdown ST alone or both LT and ST mRNA expression in MCC cells (Figure 3 C and D), pLKO sh sT1 (ST) and pLKO sh pan-T1 (PAN) were used together with a control pLKO shCtrl construct (Scr) (13). In Figure 6, pan T antigen knockdown (PAN) was performed by pLenti e7SK-shpanT-puro and pLenti e7SK-Ctrl-puro (Scr), in which shRNA sequences identical to pLKO sh pan-T1 and pLKO shCtrl were cloned under e7SK promoter (51). Lentivirus was produced as described (13).

MCPyV-positive MCC cells (MKL-1, MKL-2, CVG-1 and MS-1) were cultured in RPMI 1640 medium (Invitrogen Life Technologies, Cergy-Pontoise, France) supplemented with 10% fetal bovine serum (PAA, Pasching, Austria), 100 U/ml penicillin and 0.1 mg/ml streptomycin (Pen Strep; GIBCO, Invitrogen), 2 mM L-glutamine (PAA), and 1 mM sodium pyruvate (PAA) (11, 13). NIH3T3, 293FT and Phoenix cells were cultured following the previously described protocol (13, 52) Naturally immortalized keratinocytes (NIKS) and hTERT-HK were grown

420 together with NIH3T3 feeder cells in FAD medium, containing Ham's F-12 (PAA), DMEM
421 (GIBCO), 2% fetal calf serum (PAA), 100 U/ml penicillin and 0.1mg/ml streptomycin (Pen
422 Strep; GIBCO, Invitrogen), adenine (SIGMA), 10 ng/ml human epidermal growth factor (R&D),
423 5 mg/ml insulin (Sigma-Aldrich), 400 µg/ml hydrocortisone (SIGMA), 10 mg/ml ciprofloxacin
424 hydrochloride (EUROMEDEX) and 20 mg of cholera toxin (List Biological Laboratories). All
425 cells were cultured at 37°C with 5% CO_2.

426 **Retroviral and lentiviral infections**

427 Retroviral transduction of keratinocytes with the early genes from PyVs cloned in pLXSN was
428 performed as previously described (24, 52). After viral transduction, keratinocytes were selected
429 in medium containing 1 mg/ml G418 (PAA). Lentiviral transductions were performed as
430 previously described (13). Briefly, MCC cells infected with indicated lentiviruses were selected
431 for 6 days with 1 µg/mL puromycin. After puromycin selection of infected cells, fresh cell
432 culture medium was added to recover cells, and cells were harvested for analysis as indicated.
433 For NDRG1 experiments, 0.5 µg/mL of doxycycline was added to puromycin-selected cells.

434 **Transfection**

435 Cells were plated in 6 well plates and transfected with control or cyclin D1 siRNA (Dharmacon)
436 using Lipofectamine 2000 following the manufacturers protocol.

437 **Image acquisition and processing**

438 The images were acquired directly in 6 well plates using Nikon Eclipse Ti wide-field inverted
439 microscope. The images thus captured were analyzed using ImageJ software.

440 **mRNA extraction and quality control**

441 For microarray, total RNA was extracted from NIKs cell expressing early genes of 6 PyVs, by
442 using Absolutely RNA miniprep kit (Stratagene) according to the manufacturer protocol. RNA
443 integrity and quantification were characterized by measuring the 28s/18s rRNA ratio and RIN
444 (RNA Integrity Number) using the Agilent 2100 bioanalyzer instrument and the RNA 6000
445 Nano kit.

446 **Genome expression profiling**

447 Genome-wide gene expression profiling analysis was performed using Illumina Human HT-12
448 v4 Expression Bead Chips, providing a coverage of more than 24,000 annotated genes (47,231
449 probes corresponding to 1 to 3 probes per gene) including well characterized genes and splice

variants derived from the National Center for Biotechnology Information Reference Sequence (NCBI). Using the Illumina Total Prep RNA Amplification Kit (Ambion®), 500 ng of extracted RNAs was converted to cDNAs and subsequent biotin labelled single-stranded cRNAs. The distribution of homogeneous *in vitro* transcription products (cRNAs) was checked with the Agilent bioanalyzer instrument and the RNA 6000 Nano kit. 750 ng of biotin labelled cRNAs was then hybridized overnight to Human HT-12 Expression Bead Chips. Subsequent steps included washing, streptavadin-Cy3 staining and scanning of the arrays on an Illumina Bead Array Reader. Fluorescence emission by Cy3 was quantitatively detected for downstream analysis. The Illumina Genome Studio V2010.2 was used to obtain the signal values (AVG-Signal), with no normalization and no background subtraction.

Data analysis

Quality of the bead array data was verified using the internal controls present on the HumanHT-12 bead chip and was visualized as a control summary plot. Non-normalized raw data was then imported to BRB-Array Tools version 4.3.0 (developed by Dr. Richard Simon and the BRB-ArrayTools Development Team) for downstream analyses. Background subtraction, color correction and SSN normalization was performed using the lumi R package plugin (53). Quality of the data was further checked by generating boxplots of total gene expression data, principal component analysis, unsupervised clustering using centered correlation and average linkage of replicates, and consistency between duplicated probes. For class comparisons, genes were considered differentially expressed between groups when mean expression was at least 1.5-fold different (up or down), with a corrected p value and false discovery rate (FDR) below 0.001. The test was based on comparing the differences in mean log-intensities between classes relative to the variation expected in the mean differences. Technical replicates were averaged before class comparison.

Gene ontology analyses were performed with the WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) using the whole human genome as reference. In addition, gene set comparison was done in BRB-Array Tools for Gene Ontology categories, and biological pathways (BioCarta and KEGG). The gene set comparison tool analyzes pre-defined gene sets for differential expression among pre-defined classes. The significance values are based on testing the null hypothesis that the list of genes that belong to a given GO category is a random selection from the project gene list, against the alternative hypothesis that it contains more genes differentially expressed between the classes being compared. Pathway analysis was performed using the Reactome pathway database (54).

RT-PCR and Quantitative PCR

Quantitative real-time PCR (qRT-PCR) was performed as previously described (55). Briefly, total RNA was extracted from cells using NucleoSpin RNA (Macherey-Nagel) and reverse transcribed to cDNA by using RevertAid H-Minus M-MuLV Reverse Transcriptase (MBI; Fermentas) according to the manufacturer's protocol. The primer sequences used for qRT- PCR are listed in Supplemental Table 5.

Immunoblotting and antibodies

Whole cell lysates were prepared and sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) and immunoblotting (IB) were performed according to previously described protocols (55, 56). Protein concentrations were measured with bicinchoninic acid (BCA) assay reagent; 30-40 μg of protein extracts were used for SDS-PAGE and immunoblot analyses, and IB was performed according to the previously described method. The antibodies used for IB were β-actin (MP Biomedicals), NDRG1 (Cell Signaling), cyclinD1 (Cell Signaling), CDK2 (Pharmingen), MMP-7 (Santa Cruz).

Cell cycle analysis

Analysis of cellular DNA content was ascertained by propidium iodide (PI) staining to determine the proportion of cells in different phases of cell cycle. Cells were harvested by trypsinization and fixed using ice-cold 70% ethanol at 4°C for 30 mins. Cells were pelleted and washed with phosphate-buffered saline (PBS) 3 times. Cells were resuspended in 500μl of PBS with 20 μg/ml PI and 10μg/ml RNase A and incubated at room temperature for 30 mins. Analysis was performed using FACSCanto flow cytometer.

BrdU (10μM) was added to MCC cells expressing with or without NDRG1 for 1 hour before harvesting. Cells were fixed in ,10% buffered formalin for 10 min denatured with 2N HCl for 30 min, permebialized in 0.3% Ttiron X/PBS for 10 min at room temperature. Cells were neutralized and incubated with anti-BrdU antibody (1:2000, Cell Signaling) in 1% BSA/PBS overnight at 4°C. Cells were washed with 1% BSA/PBS once and incubated with secondary anti-mouse IgG Alexa Fluor 488 (1:1000), Invitrogenin 1% BSA/PBS for 1 hour at room temperature. Cells were washed, suspended in PBS containing ribonuclease A (100 μg/ml), propidium iodide (50 μg/ml), and 0.05% Triton X, incubated for 30 min at 37°C in the dark, and then analyzed with BD Accuri C6 flow cytometer. (Beckton Dickinson)

Colony Formation Assay

After 24 h of transient transfection, hygromycin was added to the medium and cells selected for 72h. Thereafter cells were split at 1:10, 1:100 or 1:1000 and allowed to grow for several days in

hygromycin selection (57). Cells were washed in phosphate-buffered saline (PBS), fixed and stained with crystal violet in 20% methanol and the number of colonies was counted.

Wound Healing Assay

Transfected cells were wounded using a pipette tip in a continuous straight line. The images were obtained under a light microscope using an objective with 5x magnification at 0, 24 and 48 h postscratch (57).

Cell Proliferation Assay

At day 6 post-transduction, 2.5×10^4 cells were seeded in a 96-well plate (day 0). Cell proliferation was measured using WST-8 (Wako) at days 1, 3, 5, 8, 10, and 12. OD values were normalized by values from day 1. The WST-8 formazan product was measured at 440 nm with a reference filter at 600 nm.

Statistical analyses

The Student's *t* test was applied to check the statistical significance of the obtained data. p values < 0.05 and $>$ than 0.01 are indicated with *, p values < 0.01 and $>$ than 0.001 are indicated with ** and p values < 0.0001 are indicated with ***. Error bars in the graphs represent the standard deviation (SD).

Data availability statement

The GSE137328 dataset analyzed during the current study is available in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/).

Acknowledgments

The work reported here was undertaken during the tenure of an IARC Postdoctoral Fellowship from the International Agency on Cancer. N.S was supported by a Ph.D. fellowship from the Higher Education Commission (HEC) of Pakistan. Shuda was supported by the Hillman Foundation, the Pennsylvania Tobacco Settlement grant, P50CA121973 University of Pittsburgh Skin Cancer SPORE and NIH Cancer Center Support grant P30 CA047904.

Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

References

1. Spurgeon ME, Lambert PF. 2013. Merkel cell polyomavirus: a newly discovered human virus with oncogenic potential. Virology 435:118-30.
2. Gheit T, Dutta S, Oliver J, Robitaille A, Hampras S, Combes JD, McKay-Chopin S, Le Calvez-Kelm F, Fenske N, Cherpelis B, Giuliano AR, Franceschi S, McKay J, Rollison DE, Tommasino M. 2017. Isolation and characterization of a novel putative human polyomavirus, p 45-54, Virology, vol 506. 2017 Elsevier Inc, United States.
3. Lemos B, Nghiem P. 2007. Merkel cell carcinoma: more deaths but still no pathway to blame. J Invest Dermatol 127:2100-3.
4. Chang Y, Moore PS. 2012. Merkel cell carcinoma: a virus-induced human cancer. Annu Rev Pathol 7:123-44.
5. Feng H, Shuda M, Chang Y, Moore PS. 2008. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science 319:1096-100.
6. Croul S, Otte J, Khalili K. 2003. Brain tumors and polyomaviruses. J Neurovirol 9:173-82.
7. Fields BN, Knipe DM, Howley PM. 2007. Fields virology. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia.
8. Gu R, Zhang Z, DeCerbo JN, Carmichael GG. 2009. Gene regulation by sense-antisense overlap of polyadenylation signals. Rna 15:1154-63.
9. Touze A, Gaitan J, Arnold F, Cazal R, Fleury MJ, Combelas N, Sizaret PY, Guyetant S, Maruani A, Baay M, Tognon M, Coursaget P. 2010. Generation of Merkel cell polyomavirus (MCV)-like particles and their application to detection of MCV antibodies. J Clin Microbiol 48:1767-70.
10. Tolstov YL, Pastrana DV, Feng H, Becker JC, Jenkins FJ, Moschos S, Chang Y, Buck CB, Moore PS. 2009. Human Merkel cell polyomavirus infection II. MCV is a common human infection that can be detected by conformational capsid epitope immunoassays. Int J Cancer 125:1250-6.
11. Houben R, Shuda M, Weinkam R, Schrama D, Feng H, Chang Y, Moore PS, Becker JC. 2010. Merkel cell polyomavirus-infected Merkel cell carcinoma cells require expression of viral T antigens. J Virol 84:7064-72.
12. Delbue S, Comar M, Ferrante P. 2012. Review on the relationship between human polyomaviruses-associated tumors and host immune system. Clin Dev Immunol 2012:542092.
13. Shuda M, Kwun HJ, Feng H, Chang Y, Moore PS. 2011. Human Merkel cell polyomavirus small T antigen is an oncoprotein targeting the 4E-BP1 translation regulator. J Clin Invest 121:3623-34.
14. Arora R, Chang Y, Moore PS. 2012. MCV and Merkel cell carcinoma: a molecular success story. Curr Opin Virol 2:489-98.
15. Erickson KD, Garcea RL, Tsai B. 2009. Ganglioside GT1b is a putative host cell receptor for the Merkel cell polyomavirus. J Virol 83:10275-9.
16. Neu U, Hengel H, Blaum BS, Schowalter RM, Macejak D, Gilbert M, Wakarchuk WW, Imamura A, Ando H, Kiso M, Arnberg N, Garcea RL, Peters T, Buck CB, Stehle T. 2012. Structures of Merkel cell polyomavirus VP1 complexes define a sialic acid binding site required for infection. PLoS Pathog 8:e1002738.

17. Stafford LJ, Vaidya KS, Welch DR. 2008. Metastasis suppressors genes in cancer. Int J Biochem Cell Biol 40:874-91.
18. Bandyopadhyay S, Pai SK, Hirota S, Hosobe S, Tsukada T, Miura K, Takano Y, Saito K, Commes T, Piquemal D, Watabe M, Gross S, Wang Y, Huggenvik J, Watabe K. 2004. PTEN up-regulates the tumor metastasis suppressor gene Drg-1 in prostate and breast cancer. Cancer Res 64:7655-60.
19. Bandyopadhyay S, Wang Y, Zhan R, Pai SK, Watabe M, Iiizumi M, Furuta E, Mohinta S, Liu W, Hirota S, Hosobe S, Tsukada T, Miura K, Takano Y, Saito K, Commes T, Piquemal D, Hai T, Watabe K. 2006. The tumor metastasis suppressor gene Drg-1 down-regulates the expression of activating transcription factor 3 in prostate cancer. Cancer Res 66:11983-90.
20. Guan RJ, Ford HL, Fu Y, Li Y, Shaw LM, Pardee AB. 2000. Drg-1 as a differentiation-related, putative metastatic suppressor gene in human colon cancer. Cancer Res 60:749-55.
21. Kurdistani SK, Arizti P, Reimer CL, Sugrue MM, Aaronson SA, Lee SW. 1998. Inhibition of tumor cell growth by RTP/rit42 and its responsiveness to p53 and DNA damage. Cancer Res 58:4439-44.
22. Kim KT, Ongusaha PP, Hong YK, Kurdistani SK, Nakamura M, Lu KP, Lee SW. 2004. Function of Drg1/Rit42 in p53-dependent mitotic spindle checkpoint. J Biol Chem 279:38597-602.
23. Stein S, Thomas EK, Herzog B, Westfall MD, Rocheleau JV, Jackson RS, 2nd, Wang M, Liang P. 2004. NDRG1 is necessary for p53-dependent apoptosis. J Biol Chem 279:48930-40.
24. Shahzad N, Shuda M, Gheit T, Kwun HJ, Cornet I, Saidj D, Zannetti C, Hasan U, Chang Y, Moore PS, Accardi R, Tommasino M. 2013. The T antigen locus of Merkel cell polyomavirus downregulates human Toll-like receptor 9 expression. J Virol 87:13009-19.
25. Berrios C, Padi M, Keibler MA, Park DE, Molla V, Cheng J, Lee SM, Stephanopoulos G, Quackenbush J, DeCaprio JA. 2016. Merkel Cell Polyomavirus Small T Antigen Promotes Pro-Glycolytic Metabolic Perturbations Required for Transformation, p e1006020, PLoS Pathog, vol 12, United States.
26. Masterson L, Thibodeau BJ, Fortier LE, Geddes TJ, Pruetz BL, Malhotra R, Keidan R, Wilson GD. 2014. Gene expression differences predict treatment outcome of merkel cell carcinoma patients, p 596459, J Skin Cancer, vol 2014, United States.
27. Daily K, Coxon A, Williams JS, Lee CR, Coit DG, Busam KJ, Brownell I. 2015. Assessment of cancer cell line representativeness using microarrays for Merkel cell carcinoma. J Invest Dermatol 135:1138-1146.
28. Suryadinata R, Sadowski M, Sarcevic B. 2010. Control of cell cycle progression by phosphorylation of cyclin-dependent kinase (CDK) substrates. Biosci Rep 30:243-55.
29. Deeb KK, Michalowska AM, Yoon CY, Krummey SM, Hoenerhoff MJ, Kavanaugh C, Li MC, Demayo FJ, Linnoila I, Deng CX, Lee EY, Medina D, Shih JH, Green JE, Ali-Seyed M, Laycock N, Karanam S, Xiao W, Blair ET, Moreno CS. 2007. Identification of an integrated SV40 T/t-antigen cancer signature in aggressive human breast, prostate, and lung carcinomas with poor prognosis.
30. Ali-Seyed M, Laycock N, Karanam S, Xiao W, Blair ET, Moreno CS. 2006. Cross-platform expression profiling demonstrates that SV40 small tumor antigen activates Notch, Hedgehog, and Wnt signaling in human cells. BMC Cancer 6:54.

636 31. Jia L, Fu W, Jia R, Wu L, Li X, Jia Q, Zhang H. 2018. Identification of potential key
637 protein interaction networks of BK virus nephropathy in patients receiving kidney
638 transplantation. Sci Rep 8:5017.
639 32. Abend JR, Low JA, Imperiale MJ. 2010. Global effects of BKV infection on gene
640 expression in human primary kidney epithelial cells. Virology 397:73-9.
641 33. Ferenczy MW, Johnson KR, Steinberg SM, Marshall LJ, Monaco MC, Beschloss AM,
642 Jensen PN, Major EO, Abend JR, Low JA, Imperiale MJ. 2013. Clonal immortalized
643 human glial cell lines support varying levels of JC virus infection due to differences in
644 cellular gene expression.
645 34. Radhakrishnan S, Otte J, Enam S, Del Valle L, Khalili K, Gordon J. 2003. JC virus-
646 induced changes in cellular gene expression in primary human astrocytes, p 10638-44, J
647 Virol, vol 77, United States.
648 35. Fang BA, Kovacevic Z, Park KC, Kalinowski DS, Jansson PJ, Lane DJ, Sahni S,
649 Richardson DR. 2014. Molecular functions of the iron-regulated metastasis suppressor,
650 NDRG1, and its potential as a molecular target for cancer therapy. Biochim Biophys Acta
651 1845:1-19.
652 36. Hodgson MC, Shao LJ, Frolov A, Li R, Peterson LE, Ayala G, Ittmann MM, Weigel NL,
653 Agoulnik IU. 2011. Decreased expression and androgen regulation of the tumor
654 suppressor gene INPP4B in prostate cancer. Cancer Res 71:572-82.
655 37. Calderon MR, Verway M, An BS, DiFeo A, Bismar TA, Ann DK, Martignetti JA,
656 Shalom-Barak T, White JH. 2012. Ligand-dependent corepressor (LCoR) recruitment by
657 Kruppel-like factor 6 (KLF6) regulates expression of the cyclin-dependent kinase
658 inhibitor CDKN1A gene. J Biol Chem 287:8662-74.
659 38. Lai LC, Su YY, Chen KC, Tsai MH, Sher YP, Lu TP, Lee CY, Chuang EY. 2011. Down-
660 regulation of NDRG1 promotes migration of cancer cells during reoxygenation. PLoS
661 One 6:e24375.
662 39. Nakanishi M, Shimada M, Niida H. 2006. Genetic instability in cancer cells by impaired
663 cell cycle checkpoints. Cancer Sci 97:984-9.
664 40. Houben R, Adam C, Baeurle A, Hesbacher S, Grimm J, Angermeyer S, Henzel K,
665 Hauser S, Elling R, Brocker EB, Gaubatz S, Becker JC, Schrama D. 2012. An intact
666 retinoblastoma protein-binding site in Merkel cell polyomavirus large T antigen is
667 required for promoting growth of Merkel cell carcinoma cells. Int J Cancer 130:847-56.
668 41. Demetriou SK, Ona-Vu K, Sullivan EM, Dong TK, Hsu SW, Oh DH. 2012. Defective
669 DNA repair and cell cycle arrest in cells expressing Merkel cell polyomavirus T antigen.
670 Int J Cancer 131:1818-27.
671 42. Shtutman M, Zhurinsky J, Simcha I, Albanese C, D'Amico M, Pestell R, Ben-Ze'ev A.
672 1999. The cyclin D1 gene is a target of the beta-catenin/LEF-1 pathway. Proc Natl Acad
673 Sci U S A 96:5522-7.
674 43. Angst E, Dawson DW, Stroka D, Gloor B, Park J, Candinas D, Reber HA, Hines OJ, Eibl
675 G. 2011. N-myc downstream regulated gene-1 expression correlates with reduced
676 pancreatic cancer growth and increased apoptosis in vitro and in vivo. Surgery 149:614-
677 24.
678 44. Liu W, Iiizumi-Gairani M, Okuda H, Kobayashi A, Watabe M, Pai SK, Pandey PR, Xing
679 F, Fukuda K, Modur V, Hirota S, Suzuki K, Chiba T, Endo M, Sugai T, Watabe K. 2011.
680 KAI1 gene is engaged in NDRG1 gene-mediated metastasis suppression through the
681 ATF3-NFkappaB complex in human prostate cancer. J Biol Chem 286:18949-59.

45. Kovacevic Z, Chikhani S, Lui GY, Sivagurunathan S, Richardson DR. 2013. The iron-regulated metastasis suppressor NDRG1 targets NEDD4L, PTEN, and SMAD4 and inhibits the PI3K and Ras signaling pathways. Antioxid Redox Signal 18:874-87.
46. Liu W, Xing F, Iiizumi-Gairani M, Okuda H, Watabe M, Pai SK, Pandey PR, Hirota S, Kobayashi A, Mo YY, Fukuda K, Li Y, Watabe K. 2012. N-myc downstream regulated gene 1 modulates Wnt-beta-catenin signalling and pleiotropically suppresses metastasis. EMBO Mol Med 4:93-108.
47. Lu WJ, Chua MS, Wei W, So SK. 2015. NDRG1 promotes growth of hepatocellular carcinoma cells by directly interacting with GSK-3beta and Nur77 to prevent beta-catenin degradation. Oncotarget 6:29847-59.
48. Arora R, Shuda M, Guastafierro A, Feng H, Toptan T, Tolstov Y, Normolle D, Vollmer LL, Vogt A, Domling A, Brodsky JL, Chang Y, Moore PS. 2012. Survivin is a therapeutic target in Merkel cell carcinoma. Sci Transl Med 4:133ra56.
49. Schweitzer CJ, Zhang F, Boyer A, Valdez K, Cam M, Liang TJ. 2018. N-Myc Downstream-Regulated Gene 1 Restricts Hepatitis C Virus Propagation by Regulating Lipid Droplet Biogenesis and Viral Assembly. J Virol 92.
50. Kanda T, Miyata M, Kano M, Kondo S, Yoshizaki T, Iizasa H. 2015. Clustered microRNAs of the Epstein-Barr virus cooperatively downregulate an epithelial cell-specific metastasis suppressor. J Virol 89:2684-97.
51. Velasquez C, Amako Y, Harold A, Toptan T, Chang Y, Shuda M. 2018. Characterization of a Merkel Cell Polyomavirus-Positive Merkel Cell Carcinoma Cell Line CVG-1. Front Microbiol 9:713.
52. Caldeira S, Zehbe I, Accardi R, Malanchi I, Dong W, Giarre M, de Villiers EM, Filotico R, Boukamp P, Tommasino M. 2003. The E6 and E7 proteins of the cutaneous human papillomavirus type 38 display transforming properties. J Virol 77:2195-206.
53. Du P, Kibbe WA, Lin SM. 2008. lumi: a pipeline for processing Illumina microarray. Bioinformatics 24:1547-8.
54. Jupe S, Fabregat A, Hermjakob H. 2015. Expression data analysis with Reactome, p 8 20 1-9, Curr Protoc Bioinformatics, vol 49. Inc., United States.
55. Accardi R, Scalise M, Gheit T, Hussain I, Yue J, Carreira C, Collino A, Indiveri C, Gissmann L, Sylla BS, Tommasino M. 2011. IkappaB kinase beta promotes cell survival by antagonizing p53 functions through DeltaNp73alpha phosphorylation and stabilization. Mol Cell Biol 31:2210-26.
56. Accardi R, Dong W, Smet A, Cui R, Hautefeuille A, Gabet AS, Sylla BS, Gissmann L, Hainaut P, Tommasino M. 2006. Skin human papillomavirus type 38 alters p53 functions by accumulation of deltaNp73. EMBO Rep 7:334-40.
57. Knight LM, Stakaityte G, Wood JJ, Abdul-Sada H, Griffiths DA, Howell GJ, Wheat R, Blair GE, Steven NM, Macdonald A, Blackbourn DJ, Whitehouse A. 2015. Merkel cell polyomavirus small T antigen mediates microtubule destabilization to promote cell motility and migration. J Virol 89:35-47.

Legends of Figures

Fig. 1. The ability of deregulating the cellular genome expression varies among different PyVs. (A) Schematic presentation of the unsupervised clustering of replicates after removing the background. The dendrogram (upper panel) shows the clustering of duplicates and quadruplicates using centered correlation and average linkages. Heat map (lower panel) shows the differential expression of genes in all samples. Each row indicates the expression of a specific gene across all the samples while each column represents the sample in which gene expression was measured. The colour scale at the bottom reveals the relative expression level of the genes among all the samples. Blue and red colors represent down- and upregulation respectively. (B) The histogram shows the total number of differentially expressed (either downregulated or upregulated) genes, upon each class comparison. Each PyV representing one class was compared with the negative control pLXSN and the resulting deregulated genes at 1.5-fold change with p and FDR <0.001 for each class are represented in the graph. The numbers on the top of each bar shows the total up- and downregulated genes by early genes of each PyV. (C) Venn diagram represents the common and differentially expressed genes by MCPyV dataset from this publication and publications by Berrios et al (25), Masterson et al (26), and Daily et al (27). The number 1 in the middle showing the gene HIST1C1 that was commonly deregulated in the 4 mentioned datasets. (D) Cluster analysis of differentially expresses genes involved in cell cycle regulation. The heat maps obtained from the Biorcarta shows the differential expression of 28 genes involved in cell cycle at G1/S check point (left panel) or 23 genes related to cyclins and cell cycle regulation (right panel) between MCPyV and pLXSN. Colour intensities reflect the fold change relative to the control cells. Blue and brown colors show the down and upregulation respectively.

Fig. 2. MCPyV specifically deregulates certain cellular genes. (A) Venn diagram represents the common and differentially expressed genes by 6 PyVs: MCV, JCV, KIV, WUV, BKV, and SV40. Single class comparisons were made between all 6 PyVs and pLXSN control and then using R scripts performed Venn intersections of the 6 datasets. The numbers at the extremities represent the specifically deregulated genes by each PyV while number 60 in the middle are the genes that are commonly deregulated by the mentioned 6 PyVs. The 23 genes showing the MCPyV specific signature are highlighted in red colour. (B) Heatmap shows the relative expression of 28 genes in other 5 human PyVs, which were uniquely and specifically deregulated by MCPyV early genes. The 28 genes enlisted are uniquely and significantly deregulated by MCPyV with 15 genes downregulated and 13 upregulated. The numbers at the Y-axis shows the number of genes while the X-axis represents the number of samples. The colour bar at the bottom represents the fold change scale varying between -2.4 (blue, down regulated) to 2.3 (red,

upregulated). (C and D) The 23 genes of the MCPyV specific signature compared to SV40 (C) and BKV (D). (E) Bar diagram shows the number of genes involved in the biological (left panel) and molecular (right panel) functions. Using GeneOntology software, the 28 genes representing the specific signature of MCPyV were analyzed for the involvement in various biological processes. Each bar represents one biological category and numbers on the top of each bar shows the genes out of 28, involved in the respective functional category. The number of genes is reported on the Y-axis while X-axis represents the categories of biological functions. (F) The Reactome pathway analysis showing the top 5 pathway regulated by the 23 genes. FDR, False Discovery Rate.

Fig. 3. Expression of differently expressed genes across different Merkel cell lines Total RNA was extracted from the NIKS (A) or hTERT HK (B) stably expressing the early genes of MCPyV and converted into cDNA as described in Material and Methods. Expression of indicated genes were analysed in these samples. (C-D) The knockdown of both LT and sT (Pan) or sT alone in positive MCC cells, MKL-1 and MCPyV negative (UISO) was achieved by transduction with lentiviral-based shRNA as described in Material and Methods. Scrambled shRNA (Scr) was used as negative control. Cells were collected and processed for total RNA. After reverse transcription, mRNA levels of the indicated genes were determined by qRT-PCR and normalized to the levels of the housekeeping GAPDH gene. (E) Total protein lysates isolated from Merkel positive cell line MKL-1 transduced with PAN shRNA or NIKs or hTERT HK stably expressing MCPyV early genes were subjected to immunoblotting and expression of NDRG1 was checked in these samples. The results (±S.D) are representative of atleast two independent experiments performed in duplicates.

Fig. 4. Effect of NDRG1 overexpression on cellular activities in hTERT-HK cells expressing early genes of MCPyV or MCC cell lines. hTERT-HK cells expressing early region of MCPyV was transiently transfected with pBABE empty vector or pBABE-NDRG1 vector. (A) Expression of NDRG1 in the transfected cells. (B-C) Cells transfected with NDRG1 or not were plated in 6-well plates at ratio 1:10, 1:100 or 1:1000 after selection with hygromycin for 7-8 days as described in Materials and Methods section. Representative image (B) and the number of colonies are shown in bar graph (C). (D-F) MCC cell lines MKL-1 and MKL-2 were transduced with empty or NDRG1 expressing doxycycline-inducible lentiviral constructs in presence of 0.5μg/ml. NDRG1 expression in MKL-1 and MKL-2 was confirmed by immunoblot (D). Cell proliferation was monitored by WST8 cell proliferation assay reagent (Dojindo). First, a fold-increase of cell proliferation for the assay data point (day12) was determined by dividing the OD value of the data point by that of day 1. To calculate the relative cell proliferation activity in the presence of NDRG1 expression, the fold increase value of NDRG1-induced cells was divided by

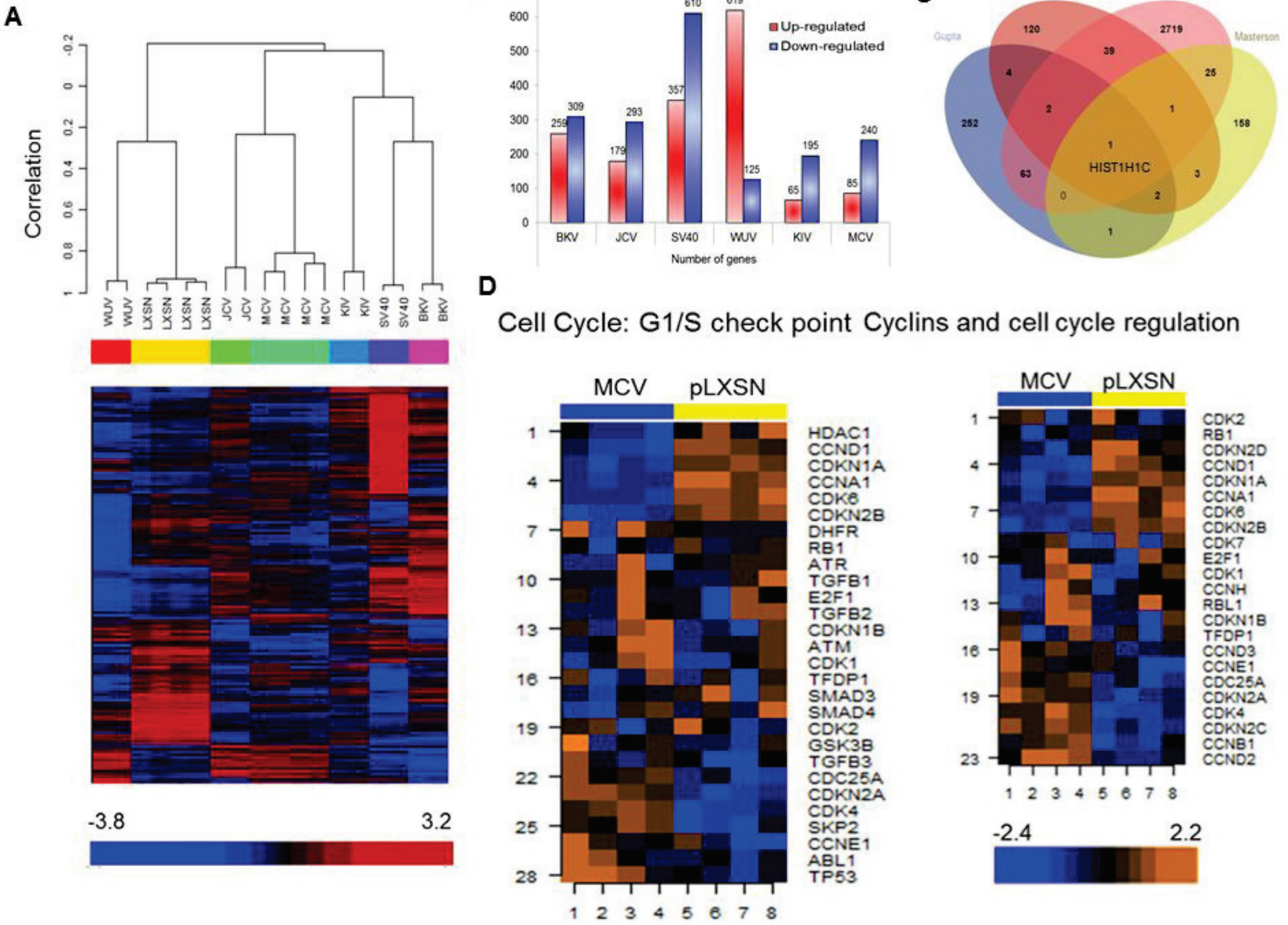
that of empty vector control cells (E). Representative image of MCC cell lines expressing or not NDRG1 5-days post doxycycline treatment (F). (G-H) As before MCPyV keratinocytes were seeded in 6 well plates and transiently transfected with pBABE empty or pBABE-NDRG1 vector. After 48 h of transfection, a scratch was introduced using a pipette tip and imaged every 24 h as mentioned in Materials and Methods. Migration of cells in hTERT keratinocytes was observed for 48 h and representative image (G) and bar graph (H) showing the wound healing or closure of wound are expressed as ratio over 0 h control. Results were obtained by three independent experiments. Error bars indicate standard deviation. Statistical significance was determined by student *t*-test.

Fig. 5. Cells engaged in DNA synthesis are reduced in MCPyV-hTERT-HK and MKL-2 but not MKL-1 cells expressing NDRG1.

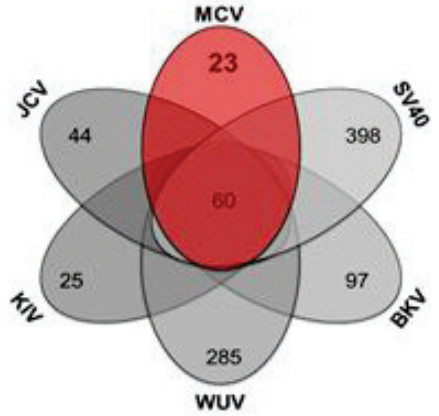
(A-C) hTERT-HK cells expressing early region of MCPyV was transiently transfected with pBABE empty vector or pBABE-NDRG1 vector. Cell cycle profiles (A) of cells overexpressing NDRG1 or not were fixed and stained with PI. The percentage of cells in S/G2 (B) or percentage of cells in sub-G0 population (C) are represented as bar graph. (D-E) Representative results depicting cell cycle profile (top) and BrdU incorporation in S phase of cell cycle (bottom). MKL-1 or MKL-2 cells treated with doxycycline for 8 days were labelled with 10mM BrdU for 1h. Incorporated BrdU and cellular DNA was stained by anti-BrdU antibody and propidium iodide. Cells were analyzed by flowcytometry (D). Quantitation of BrdU incorporation (E). The results (±S.D) are representative of three independent experiments. Statistical analysis was performed using the Student's t test.

Fig. 6. Interrelation between products of early genes of MCPyV, NDRG1 and cell cycle regulatory proteins CDK2 and cyclin D1 (A) Total protein lysates from cells stably expressing early genes of MCPyV (MCPyV) or not (pLXSN) were prepared and immunoblotted (A) for the indicated proteins mostly known to be involved in cell cycle regulation. (B) Protein lysates from cells expressing early genes of MCPyV and overexpressing NDRG1 were subjected to Western Blotting and probed with the indicated antibodies. (C) Lysates from MKL-1 and MKL-2 cells overexpressing NDRG1 were probed for indicated proteins. (D-G) The knockdown of both LT and sT (PAN) MCC positive cells (MKL-1, MKL-2, MS-1 and CVG-1) was achieved by transduction with lentiviral-based shRNA as described in Material and Methods. Scrambled shRNA (Scr) was used as negative control. Immunoblot analysis for NDRG1, β-catenin, CDK2, cyclinD1 and LT was performed in MKL-1 (D), MKL-2 (E), MS-1 (F) and CVG-1 (G). β-actin was used as a loading control. (H-J) Total protein lysates from cells stably expressing early genes of MCPyV transfected with cyclinD1 siRNA or control siRNA were prepared and

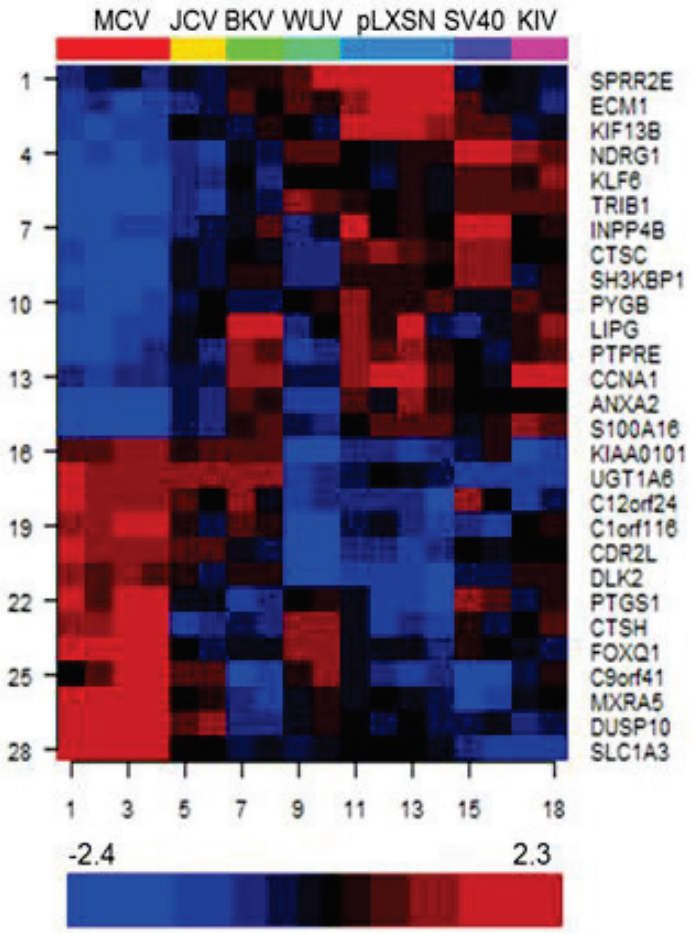
828 immunoblotted (H) for expression of cyclinD1. (I-J) Cells transfected with NDRG1 or not were
829 plated in 6-well plates at 1:100 ratio for 7-8 days as described in Materials and Methods section.
830 Representative image showing colonies and number of cells per colony (I) and the number of
831 cells per colony are shown in bar graph (J).


A
Correlation
WUV WUV LXSN LXSN LXSN LXSN JCV JCV MCV MCV MCV MCV KIV KIV SV40 SV40 BKV BKV
-3.8
3.2
B
Up-regulated
Down-regulated
BKV JCV SV40 WUV KIV MCV
Number of genes
C
Gupta
Daily
Masterson
HIST1H1C
D
Cell Cycle: G1/S check point Cyclins and cell cycle regulation
MCV pLXSN
HDAC1 CCND1 CDKN1A CCNA1 CDK6 CDKN2B DHFR RB1 ATR TGFB1 E2F1 TGFB2 CDKN1B ATM CDK1 TFDP1 SMAD3 SMAD4 CDK2 GSK3B TGFB3 CDC25A CDKN2A CDK4 SKP2 CCNE1 ABL1 TP53
MCV pLXSN
CDK2 RB1 CDKN2D CCND1 CDKN1A CCNA1 CDK6 CDKN2B CDK7 E2F1 CDK1 CCNH RBL1 CDKN1B TFDP1 CCND3 CCNE1 CDC25A CDKN2A CDK4 CDKN2C CCNB1 CCND2
-2.4
2.2

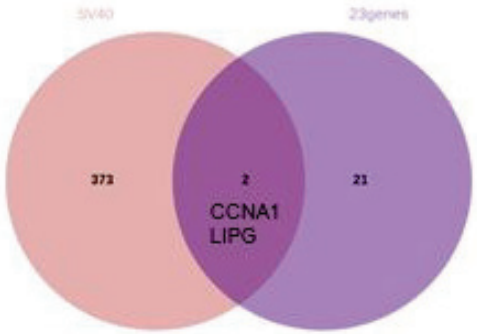
A



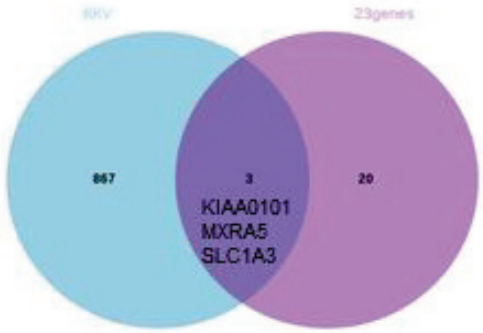
B



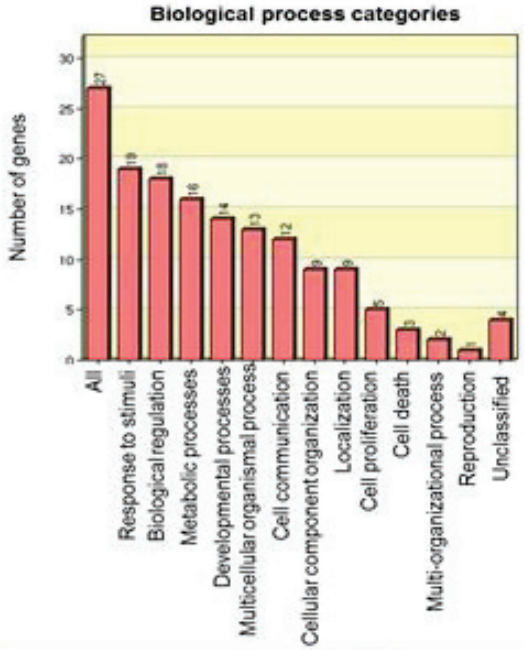
C

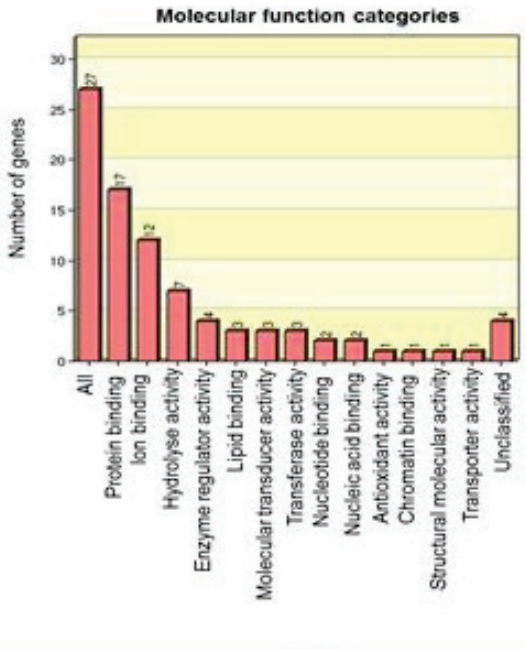


D



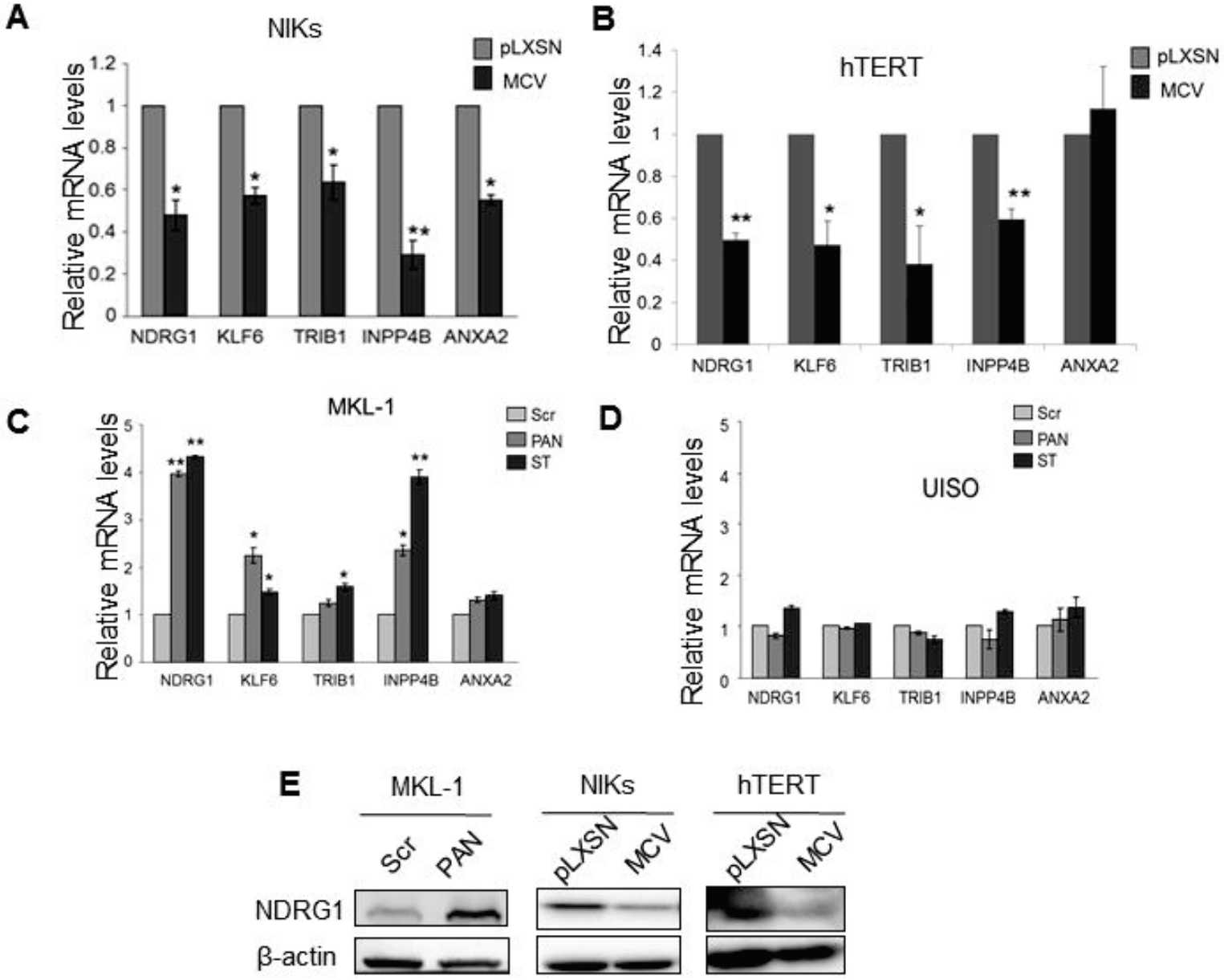
E

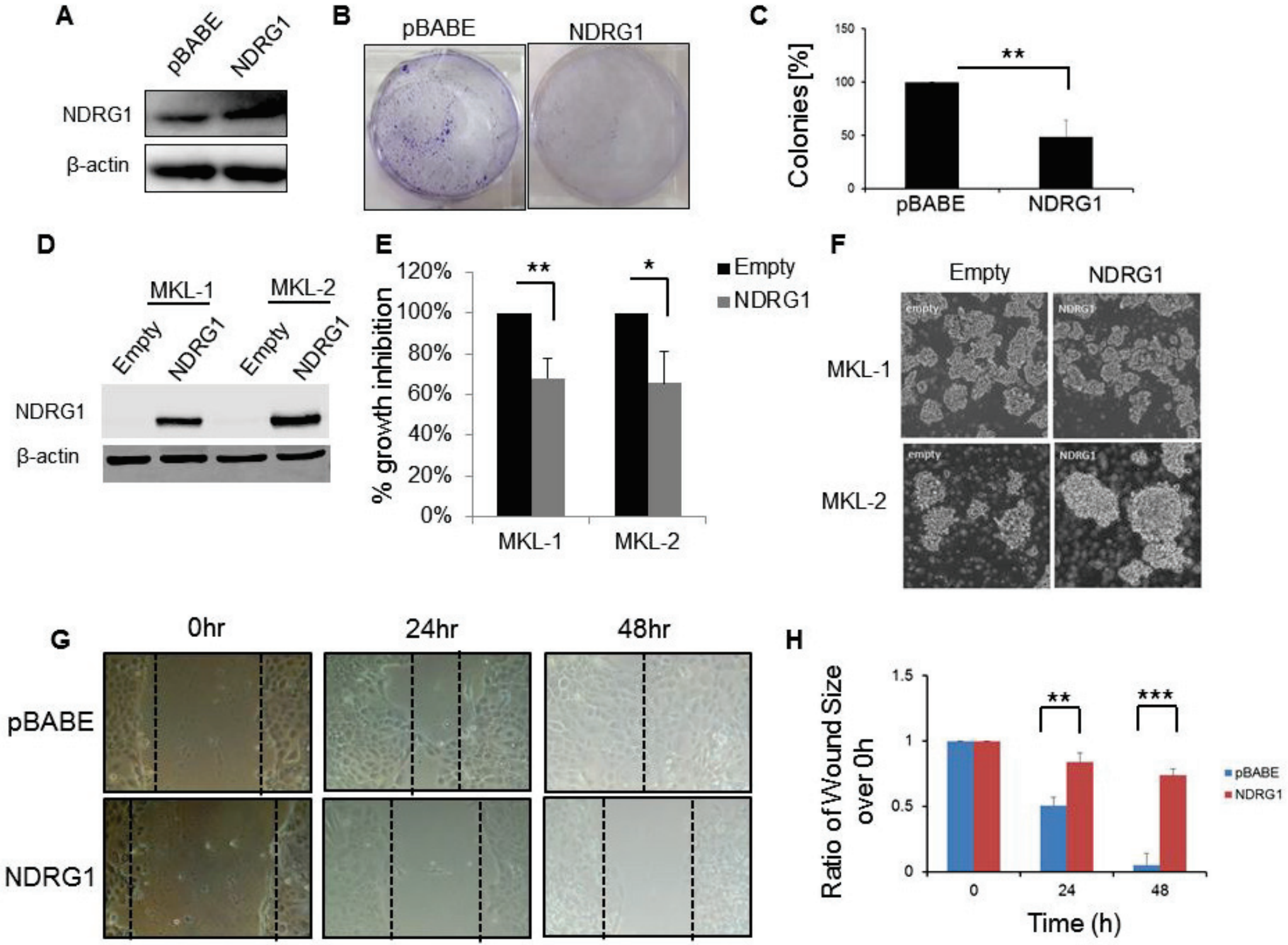


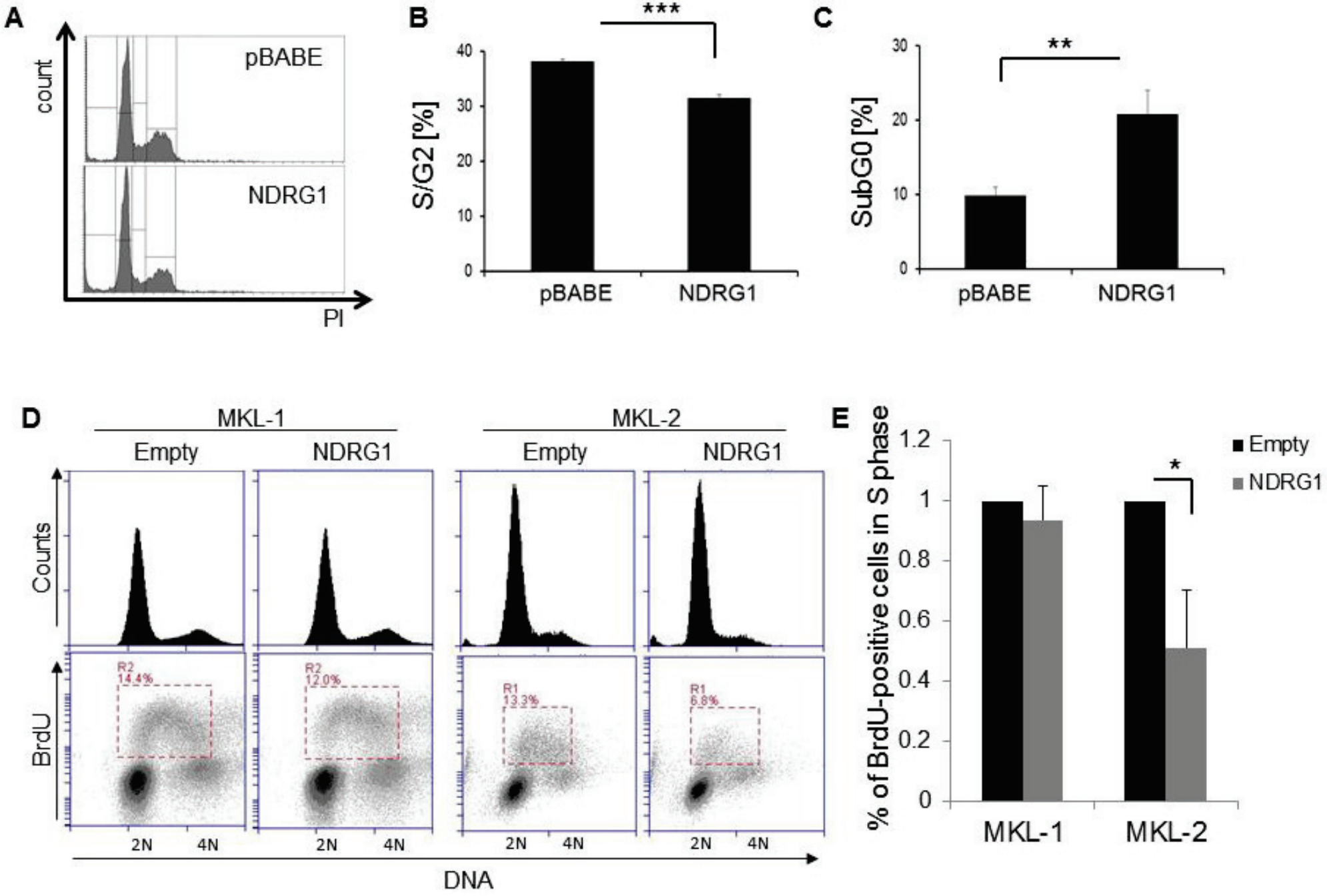


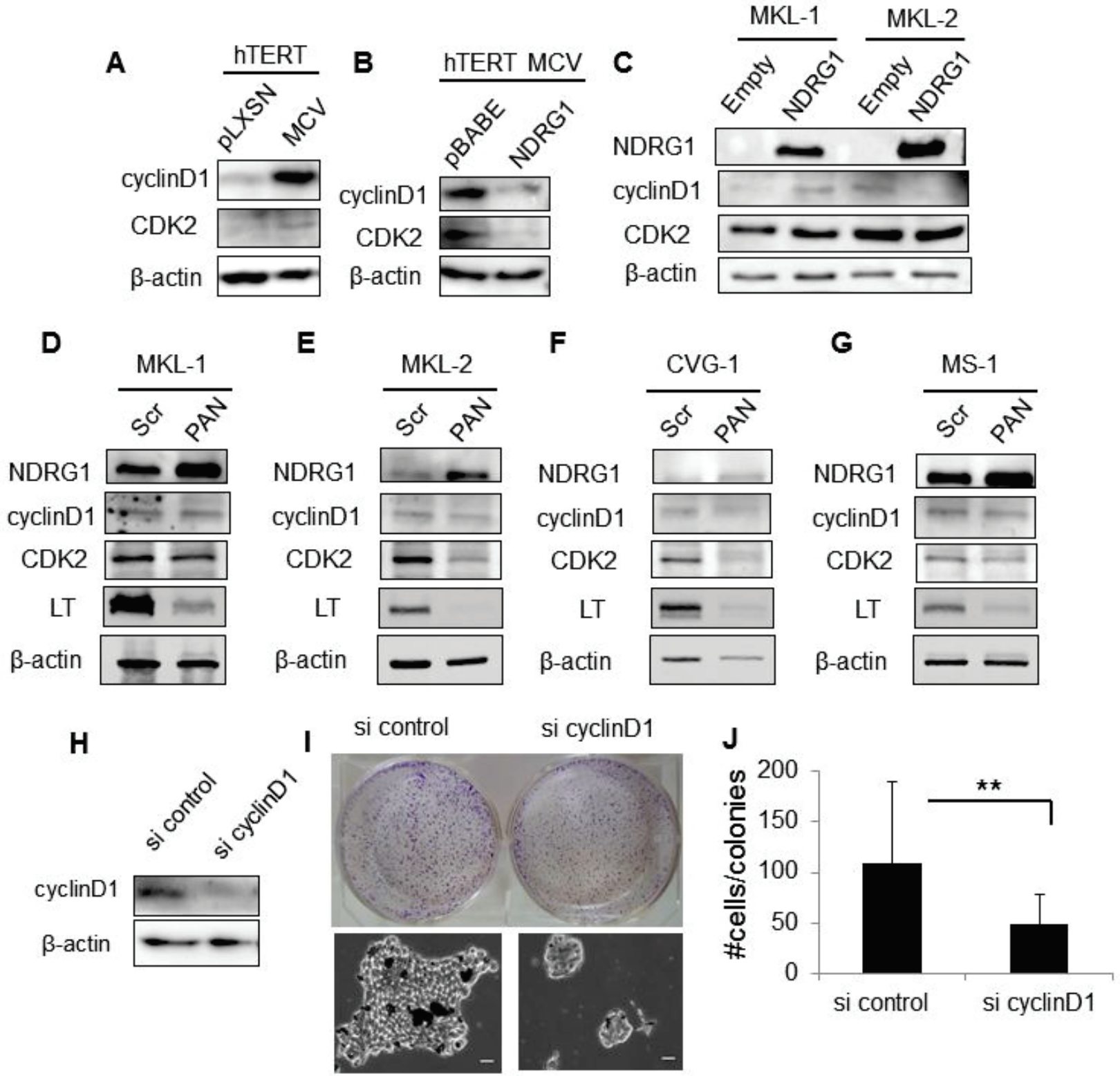
F

Pathway name	Entities				Reactions	
	found	ratio	p-value	FDR*	found	ratio
Glucuronidation	4 / 55	0.004	9.59e-06	0.002	6 / 11	9.01e-04
TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain	2 / 28	0.002	0.002	0.189	2 / 19	0.002
Phase II - Conjugation of compounds	4 / 258	0.018	0.003	0.19	6 / 72	0.006
G1/S-Specific Transcription	2 / 43	0.003	0.005	0.19	2 / 28	0.002
Biological oxidations	5 / 545	0.038	0.009	0.19	8 / 188	0.015


A
NIKs
pLXSN
MCV
Relative mRNA levels
NDRG1
KLF6
TRIB1
INPP4B
ANXA2
B
hTERT
pLXSN
MCV
Relative mRNA levels
NDRG1
KLF6
TRIB1
INPP4B
ANXA2
C
MKL-1
Scr
PAN
ST
Relative mRNA levels
NDRG1
KLF6
TRIB1
INPP4B
ANXA2
D
UISO
Scr
PAN
ST
Relative mRNA levels
NDRG1
KLF6
TRIB1
INPP4B
ANXA2
E
MKL-1
NIKs
hTERT
Scr
PAN
pLXSN
MCV
pLXSN
MCV
NDRG1
β-actin


A
pBABE
NDRG1
NDRG1
β-actin
B
pBABE
NDRG1
C
Colonies [%]
150
100
50
0
**
pBABE
NDRG1
D
MKL-1
MKL-2
Empty
NDRG1
Empty
NDRG1
NDRG1
β-actin
E
% growth inhibition
120%
100%
80%
60%
40%
20%
0%
**
*
Empty
NDRG1
MKL-1
MKL-2
F
Empty
NDRG1
MKL-1
MKL-2
empty
NDRG1
empty
NDRG1
G
0hr
24hr
48hr
pBABE
NDRG1
H
Ratio of Wound Size over 0h
1.5
1
0.5
0
**
***
pBABE
NDRG1
0
24
48
Time (h)


A
count
pBABE
NDRG1
PI
B
***
S/G2 [%]
40
30
20
10
0
pBABE
NDRG1
C
**
SubG0 [%]
30
20
10
0
pBABE
NDRG1
D
MKL-1
MKL-2
Empty
NDRG1
Empty
NDRG1
Counts
BrdU
R2
14.4%
R2
12.0%
R1
13.3%
R1
6.8%
2N
4N
2N
4N
2N
4N
2N
4N
DNA
E
% of BrdU-positive cells in S phase
1.2
1
0.8
0.6
0.4
0.2
0
*
Empty
NDRG1
MKL-1
MKL-2


A
hTERT
pLXSN
MCV
cyclinD1
CDK2
β-actin
B
hTERT MCV
pBABE
NDRG1
cyclinD1
CDK2
β-actin
C
MKL-1
MKL-2
Empty
NDRG1
Empty
NDRG1
NDRG1
cyclinD1
CDK2
β-actin
D
MKL-1
Scr
PAN
NDRG1
cyclinD1
CDK2
LT
β-actin
E
MKL-2
Scr
PAN
NDRG1
cyclinD1
CDK2
LT
β-actin
F
CVG-1
Scr
PAN
NDRG1
cyclinD1
CDK2
LT
β-actin
G
MS-1
Scr
PAN
NDRG1
cyclinD1
CDK2
LT
β-actin
H
si control
si cyclinD1
cyclinD1
β-actin
I
si control
si cyclinD1
J
#cells/colonies
200
150
100
50
0
**
si control
si cyclinD1

PLOS ONE

RESEARCH ARTICLE

Molecular features of premenopausal breast cancers in Latin American women: Pilot results from the PRECAMA study

Magali Olivier[1]*, Liacine Bouaoun[2], Stephanie Villar[1], Alexis Robitaille[1], Vincent Cahais[1], Adriana Heguy[3], Graham Byrnes[2], Florence Le Calvez-Kelm[4], Gabriela Torres-Mejía[5], Isabel Alvarado-Cabrero[6], Fazlollah Shahram Imani-Razavi[7], Gloria Inés Sánchez[8], Roberto Jaramillo[9], Carolina Porras[10], Ana Cecilia Rodriguez[10], Maria Luisa Garmendia[11], José Luis Soto[12], Isabelle Romieu[13], Peggy Porter[14], Jamie Guenthoer[15], Sabina Rinaldi[13], on behalf of the PRECAMA team[¶]

1 Section of Mechanisms of Carcinogenesis, International Agency for Research on Cancer, Lyon, France, **2** Section of Environment and Radiation, International Agency for Research on Cancer, Lyon, France, **3** Department of Pathology and Genome Technology Center, New York University Langone Medical Center, New York, United States of America, **4** Genetic Cancer Susceptibility Group, International Agency for Research on Cancer, Lyon, France, **5** Center for Population Health Research, National Institute of Public Health, Cuernavaca, Mexico, **6** Department of Pathology, Hospital de Oncología, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, Mexico City, Mexico, **7** Department of Pathology, UMAE Hospital de Gineco Obstetricia No. 4 "Luis Castelazo Ayala", Instituto Mexicano del Seguro Social, Mexico City, Mexico, **8** Group Infection and Cancer, School of Medicine, University of Antioquia, Medellín, Colombia, **9** Hemato Oncologos, Cali, Colombia, **10** Agencia Costarricense de Investigaciones Biomédicas (ACIB)-Fundación INCIENSA, Costa Rica, **11** Instituto de Nutrición y de Tecnología de los Alimentos, Santiago, Chile, **12** National Institute of Cancer, Santiago, Chile, **13** Section of Nutrition and Metabolism, International Agency for Research on Cancer, Lyon, France, **14** Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, United States of America, **15** Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, United States of America

¶ Complete membership of the PRECAMA team can be found in the Acknowledgments
* olivierm@iarc.fr

OPEN ACCESS

Citation: Olivier M, Bouaoun L, Villar S, Robitaille A, Cahais V, Heguy A, et al. (2019) Molecular features of premenopausal breast cancers in Latin American women: Pilot results from the PRECAMA study. PLoS ONE 14(1): e0210372. https://doi.org/10.1371/journal.pone.0210372

Editor: Obul Reddy Bandapalli, German Cancer Research Center (DKFZ), GERMANY

Received: August 31, 2018

Accepted: December 20, 2018

Published: January 17, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The study is funded by the International Agency for Research on Cancer (IARC), the Union for International Cancer Control (UICC), the Pan American Health Organization (PAHO), the Ibero-American Programme for the Development of Science and Technology (CYTED), COLCIENCIAS (grant n°1115-569-348899), CODI-Universidad de Antioquia (grant CPT-1229).

Abstract

Background

In Latin America (LA), there is a high incidence rate of breast cancer (BC) in premenopausal women, and the genomic features of these BC remain unknown. Here, we aim to characterize the molecular features of BC in young LA women within the framework of the PRECAMA study, a multicenter population-based case–control study of BC in premenopausal women.

Methods

Pathological tumor tissues were collected from incident cases from four LA countries. Immunohistochemistry (IHC) was performed centrally for ER, PR, HER2, Ki67, EGFR, CK5/6, and p53 protein markers. Targeted deep sequencing was done on genomic DNA extracted from formalin-fixed, paraffin-embedded tumor tissues and their paired blood samples to screen for somatic mutations in eight genes frequently mutated in BC. A subset of samples was analyzed by exome sequencing to identify somatic mutational signatures.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: BC, breast cancer; COSMIC, Catalogue of Somatic Mutations in Cancer; EGFR, epidermal growth factor receptor; ER, estrogen receptor; FFPE, formalin-fixed, paraffin-embedded; HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; LA, Latin America; postBC, postmenopausal breast cancer; PR, progesterone receptor; preBC, premenopausal breast cancer; SNVs, single nucleotide variants; TCGA, The Cancer Genome Atlas; TN, triple-negative; TNBC, triple-negative breast cancer..

Results

The majority of cases were positive for ER or PR (168/233; 72%), and 21% were triple-negative (TN), mainly of basal type. Most tumors were positive for Ki67 (189/233; 81%). In 126 sequenced cases, *TP53* and *PIK3CA* were the most frequently mutated genes (32.5% and 21.4%, respectively), followed by *AKT1* (9.5%). *TP53* mutations were more frequent in HER2-enriched and TN IHC subtypes, whereas *PIK3CA/AKT1* mutations were more frequent in ER-positive tumors, as expected. Interestingly, a higher proportion of G:C>T:A mutations was observed in *TP53* in PRECAMA cases compared with TCGA and METABRIC BC series (27% vs 14%). Exome-wide mutational patterns in 10 TN cases revealed alterations in signal transduction pathways and major contributions of mutational signatures caused by altered DNA repair pathways.

Conclusions

These pilot results on PRECAMA tumors give a preview of the molecular features of premenopausal BC in LA. Although the overall mutation burden was as expected from data in other populations, mutational patterns observed in *TP53* and exome-wide suggested possible differences in mutagenic processes giving rise to these tumors compared with other populations. Further -omics analyses of a larger number of cases in the near future will enable the investigation of relationships between these molecular features and risk factors.

Introduction

Breast cancer (BC) incidence is increasing sharply in countries in economic transition, with a large number of cases in premenopausal women. In Latin America (LA), the proportion of BC in women younger than 45 years is nearly twice the proportion in developed countries, a difference that is only partly explained by population age structure [1]. Behavioral, reproductive, and lifestyle factors typical of Western populations are becoming more prevalent in LA and may play a role in the increased BC incidence in this population, but the reason for the sharp increase in incidence in premenopausal women in LA remains to be established [2].

BC is a heterogeneous disease in terms of biology and outcome. It is clinically classified into four subtypes (luminal A, luminal B, HER2-positive, and triple-negative [TN]), based on the expression of the estrogen receptor (ER), the progesterone receptor (PR), the human epidermal growth factor receptor 2 (HER2), and the proliferation marker Ki67 [3]. More sophisticated classifications based on genomic and transcriptional analyses provide a better description of the tumor biology and outcome [4, 5]. The two most frequently somatically mutated genes in BC are *TP53* and *PIK3CA* [6]. Mutations in *PIK3CA*, which render cells dependent on *PI3K* pathway signaling, are the most common genetic abnormality identified in hormone receptor-positive BC, whereas mutations in the tumor suppressor gene *TP53* are more prevalent in the HER2-enriched and TN subtypes [6–8].

Genomic analyses can also provide information related to tumor etiology. Indeed, somatic mutational signatures can reveal the contribution of specific mutational processes to the development of cancer. For example, *TP53* mutation patterns specific to exposure to exogenous mutagens have been reported in several cancer types [9], and at the genome-wide level, more than 30 mutational signatures have been described in cancer tissues and some have been

linked to endogenous mechanisms of mutagenesis or to exposure to human carcinogens [10, 11].

Although BC genomic subtypes have been associated with different patient outcomes, how specific genomic alterations relate to risk factors or etiology remains largely unknown. Moreover, knowledge of the genomic features of premenopausal BC (preBC), particularly in countries in economic transition, is limited. The PRECAMA study was initiated to investigate the molecular, pathological, and risk factor patterns of preBC in LA (http://precama.iarc.fr/). It is the largest case–control study conducted in four countries in LA that systematically collects extensive information on lifestyle and risk factors as well as different biological samples (tumor tissues, blood fractions, and urine) according to standardized procedures. PRECAMA is thus a powerful framework for investigating relationships between BC tumor biology and etiology.

Here, we investigate the tumor genomic features of preBC in women in LA using the first set of samples collected within the framework of the PRECAMA study.

Materials and methods

Study population

The present study included 126 cases recruited between August 2012 and November 2015 in the context of the PRECAMA case–control study (http://precama.iarc.fr). Subjects included in PRECAMA are women diagnosed with BC at age 20–45 years and recruited at major general or cancer-dedicated hospitals in Chile, Colombia, Costa Rica, and Mexico that cover populations with a wide range of socioeconomic status. Women who had a positive biopsy for BC were recruited before any treatment. Women were invited to a home or hospital visit, during which a trained nurse presented the informed consent, collected biological samples and anthropometric measurements (height, weight, and hip and waist circumferences), and administered a standardized questionnaire on clinical, reproductive, and lifestyle risk factors. All participants gave written informed consent before enrollment, and the study protocols were approved by the institutional review boards of Chile (Oncologic Institute Foundation Arturo Lopez Pérez and National Cancer Institute), Colombia (Cancer Institute Las Americas and University of Antioquia), Costa Rica (Costa Rican Institute of Clinical Research [ICIC] and Center for Strategic Development and Information in Health and Social Security [CENDEISSS] of the Costa Rican Social Security Fund [CCSS]), Mexico (National Institute of Public Health and the Mexican Social Security Institute), and the International Agency for Research on Cancer (IARC).

Biological specimens

Each study site applied common standardized protocols for specimen collection. The protocols were previously developed and extensively used by IARC [12, 13], and were subsequently fine-tuned based on a detailed review of the conditions at each center. Blood samples were obtained at recruitment by venipuncture using vacutainers, and buffy coats were prepared and stored at −80˚C less than 6 hours after the blood draw. Buffy coats were shipped to IARC for genomic DNA extraction. Tumor samples were formalin-fixed and paraffin-embedded (FFPE) according to standard operating procedures. Paraffin blocks and hematoxylin and eosin sections were stored at the local pathology service facilities. Sections from tumor tissues were sent to Fred Hutchinson Cancer Research Center for centralized immunohistochemistry (IHC) analyses and tumor DNA extraction.

Pathology review and IHC analyses

Histology sections from tumor biopsies obtained before any treatment were reviewed for histological diagnosis and grade, lymphovascular invasion, and stromal and lymphocyte response. IHC was conducted for ER (SP1, LabVision, Fremont, CA), PR (PgR 636, Dako, Denmark), HER2 (AO485, Dako, Denmark), epidermal growth factor receptor (EGFR) (31G7, Invitrogen, Camarillo, CA), CK5/6 (D5/16 B4, Dako, Denmark), p53 (Pab 1801, Calbiochem, La Jolla, CA), and Ki67 (MIB-1, Dako, Denmark) according to standardized and optimized protocols that included antigen retrieval when required. BCs were classified into subtypes according to ER, PR, and HER2 IHC results. Triple-negative (ER-, PR-, HER2-) BCs were additionally subtyped using EGFR and CK5/6 staining to define basal-like cancers. ER and PR positivity were defined as staining score >1%, and Ki67 positivity as staining >14%, as recommended by the St Gallen International Breast Cancer Conference [3].

DNA extraction and sequencing

Tumor genomic DNA was extracted from 3–9 sections of 6 μm using the QIAamp DNA FFPE Tissue Kit (Qiagen) following the manufacturer's recommended protocol, with the following modification. The tissue was incubated in ATL buffer and proteinase K overnight at 56˚C with agitation, with the addition of 20 μL of proteinase K after the first 4 hours. Matched constitutive genomic DNA from cases was isolated from buffy coats at IARC with the Autopure LS system (Qiagen) using the "frozen buffy coat" protocol and following the manufacturer's instructions. DNA was quantified by PicoGreen (ThermoFisher Scientific).

For **targeted sequencing**, exonic regions of the selected gene panel (*AKT1*, *CDH1*, *ERBB2*, *NOTCH1*, *PIK3CA*, *PTEN*, *RB1*, and *TP53*) were amplified from 80 ng of genomic DNA using GeneRead DNAseq Mix-n-Match Panel V2 (Qiagen) following the manufacturer's instructions. Libraries were prepared with NEBNext reagents (New England BioLabs) following the manufacturer's instructions. Libraries were quantified by PicoGreen (ThermoFisher Scientific) and pooled in equal quantities, and the library pool was quantified by the Qubit fluorometer (ThermoFisher Scientific) and quality checked with the Bioanalyzer (Agilent Technologies). Then, 800 pM of the library pool was used for sequencing on a Ion Proton sequencer (Life Technologies) according to the manufacturer's instructions, aiming at a minimum of 100X coverage for blood DNA and 1000X coverage for tumor DNA. Tumor samples were processed in duplicate to control for artefactual mutations from FFPE fixation (see bioinformatics analyses below).

For **whole-exome sequencing**, exonic regions and splice junctions of tumor–blood DNA sample pairs were captured using the SeqCap EZ MedExome kit (Roche Diagnostics France) following the manufacturer's instructions. This assay captures exonic regions covering 47 Mb of protein-coding bases. Libraries were prepared with the KAPA Hyper Prep Kit (Roche Diagnostics France) following the manufacturer's instructions, and sequenced by 150-base paired-end massively parallel sequencing on an Illumina HiSeq 4000 sequencer at the New York University Langone Medical Center according to the manufacturer's instructions.

Bioinformatics analyses

Data from the Ion Proton sequencer were processed with the Ion Torrent built-in pipeline (TorrentSuite V4) to generate BAM files, and variant calling was done with the built-in ITVC in the somatic mode and with a minimum allele frequency threshold of 4%. Variants were annotated with Annovar and filtered to eliminate known single nucleotide polymorphisms (SNPs) (variants present in the Exome Aggregation Consortium [ExAC] or 1000 Genomes [1000G] databases at a frequency >0.001) and sequencing artefacts using the MutSpec Galaxy

package developed in-house [14]. Further manual checks of BAM files using IGV were done when appropriate. All non-synonymous mutations found in the targeted regions and present in both duplicates of tumor samples but not in any blood samples of the 126 cases were retained for analysis.

Exome data from the HiSeq 4000 sequencer were analyzed with a pipeline developed in-house and based on standard tools for quality control and processing (FastQC 0.11.3, AdapterRemoval 2.1.7, BWA-MEM 0.7.15, Qualimap 2, GATK 3.5, and Picard 1.131). Somatic variant calling was done on tumor–blood sample pairs with Strelka [15] using the default parameters. Variant annotation and filtering was done as described above with MutSpec [14], and only somatic indels and single nucleotide variants (SNVs) in coding regions were retained and analyzed. Pathway analysis of mutated genes was done with ConsensusPathDB (r32) using the KEGG, Biocarta, Reactome, and WikiPathways databases and a minimum of 3 overlapping genes and q-value <0.05 as settings [16]. To define cancer genes, we used the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (v82) [17], and genes identified as drivers for BC in the IntOGen database (r2014.12) [18].

Public data on somatic mutations in breast cancer

Data from The Cancer Genome Atlas (TCGA) breast and METABRIC genomic studies [19, 20] and from the IARC TP53 Database [21] were used as comparison datasets. Gene-specific mutation files (*AKT1*, *CDH1*, *ERBB2*, *NOTCH1*, *PIK3CA*, *PTEN*, *RB1*, and *TP53*) and related clinical files for the TCGA and METABRIC studies were retrieved from cBioPortal [22, 23] in February 2017. MAF files from exome sequencing data of TCGA BC cases were retrieved on 26 March 2015 via a https protocol at https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/. Gene-specific data from TCGA and METABRIC were combined, including only cases with documented age and ER, PR, and HER2 status, and stratified by age (younger than 45 or older than 55 years). For TCGA exome data, only data with documented age and ER, PR, and HER2 status were selected, resulting in a dataset of 453 samples, including 96 preBC and 357 postBC. Version R18 of the somatic dataset of the IARC TP53 Database was used to select for mutations reported in primary BC in women age 45 or younger and in studies using Sanger sequencing. Finally, another independent dataset, named hereafter 560BC, was assembled from public data obtained from whole-genome sequencing of 560 BC cases [6]. For this dataset, mutation data were retrieved from COSMIC and clinical data were retrieved from the original publication. Only cases with documented ER, PR, and HER2 status and diagnosed at 45 years or younger were included ($N = 123$).

Statistical analyses

Associations between study variables were tested using Fisher's exact test. For mutational signature analyses, we used PRECAMA exome data ($N = 12$ samples) and TCGA exome data ($N = 453$). Mutations were classified into 96 types, corresponding to the 6 possible base substitutions (C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G, and T:A>G:C) and the 16 possible pairs of nucleotides immediately flanking 5′ and 3′. Mutational signatures in these samples were then extracted using the non-negative matrix factorization (NMF) algorithm implemented in a NMF R package [24, 25]. NMF decomposition identifies signatures and estimates their contributions to each sample. Six signatures were identified using the cophenetic correlation coefficient as a measure of stability of the signatures. We calculated the cosine similarity between the 6 extracted signatures and those published in COSMIC and in other original reports [17, 26], as described elsewhere [27].

We wished to identify possible systematic differences of signature contributions between IHC subtypes, by menopausal status, and by study source (TCGA vs PRECAMA). Because of the small number of PRECAMA samples, we used 2000 permutations of samples to obtain an empirical distribution of the Kruskal–Wallis rank-sum statistic. This permutation test was applied to test for possible association of each signature with a) menopausal status; b) IHC subtype; c) menopausal status stratified by IHC subtype; d) menopausal status adjusted for subtype by linear model; and e) study source, with partial adjustment for subtype (TN vs others) and also restricted to preBC samples.

All statistical analyses were performed using R statistical software version 3.3.2. The statistical significance level was set to 0.05, without adjustment for multiple comparisons.

Results

IHC subtypes in PRECAMA tumors

In the first consecutive cases recruited in PRECAMA, for which tumor pathological evaluation has been completed ($N = 229$), most BC cases (72%) were ER-positive and 16% were HER2--positive (Table 1). Using ER/PR/HER2 IHC subtyping, the majority of cases were luminal A (58%), followed by TN (21%), luminal B (11%), and HER2-enriched (5%) (Table 1). TN tumors were predominately basal-like (94%) (**Table A in S1 File**). Proliferation status was assessed by Ki67 IHC staining. More than 80% (189/233) of cases had high Ki67 staining with a median percentage of 31.6 (not shown). Overall, Ki67-positivity (staining >14%) was significantly associated with IHC subtypes (p-value = 2×10^{-4}; Fisher's exact test). In particular, the proportion of Ki67-positive cases was significantly lower in luminal A cases than in TN cases (72% vs 98%, p-value = 2.8×10^{-5}).

Table 1. Sample classification by IHC results.

IHC result	Number of samples *N* (%)	Ki67-positive *N* (%)
ER		
Negative	65 (28%)	63 (97%)
Positive	168 (72%)	126 (75%)
PR		
Negative	71 (30%)	67 (94%)
Positive	162 (70%)	122 (75%)
HER2		
Negative	183 (79%)	144 (79%)
Equivocal	13 (6%)	12 (92%)
Positive	37 (16%)	33 (90%)
Total	233	189 (81%)
IHC SUBTYPE*		
Luminal A	134 (58%)	96 (72%)
Luminal B	26 (11%)	23 (88%)
HER2-enriched	11 (5%)	10 (91%)
Triple-negative	48 (21%)	47 (98%)
Of basal type	*45 (94%)*	*42 (93%)*
Undetermined**	14 (6%)	13 (93%)

* Tumor subtype definitions: luminal A: ER+/HER2-; luminal B: ER+/HER2+; HER2-enriched: ER-/HER2+; triple-negative: ER-/PR-/HER2-; TN of basal type: EGFR + and/or CK5/6+.

** 14 cases were not assigned a subtype: 13 cases had equivocal HER2 results and no confirmatory FISH; 1 case was ER-/PR+/HER2- with a weak PR positivity.

https://doi.org/10.1371/journal.pone.0210372.t001

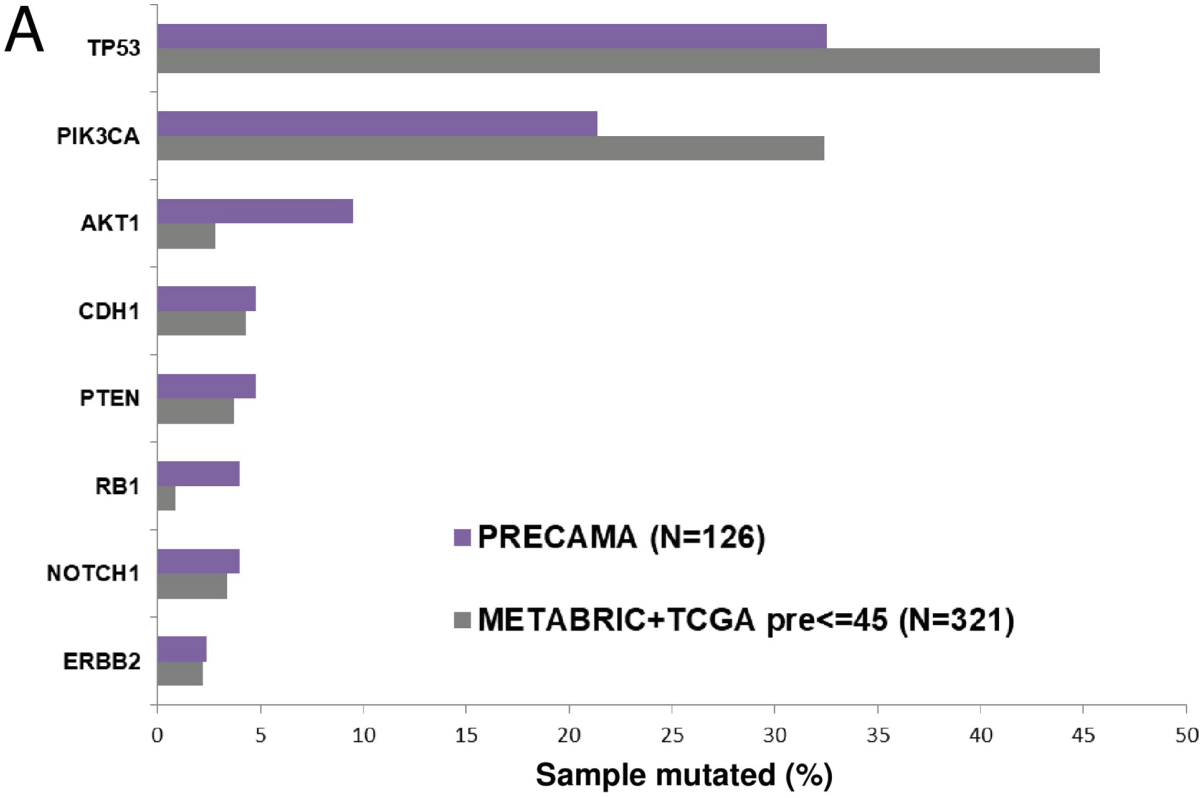
Somatic mutations in premenopausal BC

Tumor genomic DNA was extracted from FFPE tissue sections prepared at each collecting center according to a standardized protocol. More than 250 ng of DNA was obtained for 75% of the samples, with a median yield of 994 ng. A limiting amount of DNA (<100 ng) was obtained for 12% (21/172) of the samples. Targeted deep sequencing of a panel of 8 genes frequently mutated in BC (*AKT1*, *CDH1*, *ERBB2*, *NOTCH1*, *PIK3CA*, *PTEN*, *RB1*, and *TP53*) was successfully performed on 126 cases for which more than 250 ng of tumor genomic DNA was available. Tumor DNA and patient-matched blood DNA were sequenced at minimum coverages of 1000X and 100X, respectively. To control for potential artefacts due to formalin fixation, FFPE tumor samples were sequenced in duplicates and only mutations detected in both duplicates were considered (see Materials and Methods). Potentially deleterious somatic mutations (affecting splicing, indels, nonsense, stop-loss, and non-synonymous substitutions) in the 8-gene panel were found in 63.5% (80/126) of samples. *TP53* was the most frequently mutated gene (32.5%), followed by *PIK3CA* (21.4%) and *AKT1* (9.5%), whereas other genes were mutated in less than 5% of samples. This distribution differed from that observed in preBC from the TCGA/METABRIC datasets (**Fig 1A**). Indeed, there were fewer cases with *TP53* or *PIK3CA* mutations and more cases with *AKT1* and *RB1* mutations in PRECAMA versus TCGA/METABRIC cases (p-values at best = 0.03). These differences may be explained in part by a different distribution of subtypes between PRECAMA and TCGA/METABRIC cases (p-value = 8.2×10^{-6}), because a higher proportion of luminal A cases (known to carry frequent *AKT1* and infrequent *TP53* mutations) and a lower proportion of TN cases (known to carry frequent *TP53* mutations) was observed in PRECAMA compared with TCGA/METABRIC (p-value = 0.04). Interestingly, whereas the *PIK3CA*/*AKT1* pathway was mutated at the expected rates in the luminal A PRECAMA tumors, *AKT1* mutations were more frequent relative to *PIK3CA* mutations in PRECAMA compared with TCGA/METABRIC luminal A cases (14% *AKT1* and 23% *PIK3CA* mutations in PRECAMA vs 4% *AKT1* and 54% *PIK3CA* mutations in TCGA/METABRIC), although this was not statistically significant.

PIK3CA and *AKT1* mutations were at classical hotspots (p.H1047R, p.E542K, and p.E545K for *PIK3CA* and p.E17K for *AKT1*), and *TP53* mutations were mostly missense substitutions that spread across the coding sequence (**Table A in S1 File**). The relationship between IHC subtypes and mutated genes was as expected from previous studies (**Fig 1B**). *TP53*, *RB1*, or *PTEN* mutated samples had higher proportions of the TN subtype, whereas *AKT1* or *CDH1* mutated samples had higher proportions of the luminal A subtype. The majority of samples with no mutation in the tested genes were of luminal A subtype (34/46; 74%).

Twenty one tumors had mutations in more than one gene (**Table A in S1 File** filtered for genes_mutated >1). One case was of HER2-enriched subtype and had mutations in *TP53* and *ERBB2*. Three cases were of luminal B subtype and had mutations in *PIK3CA* and *TP53* or *CDH1*. Seven cases were of TN subtype and had mutations in *TP53* combined with *RB1* (3 cases), *PTEN* (2 cases), *PIK3CA* (1 case), or *NOTCH1* (1 case). Ten cases were of luminal A subtype and had mutations in *TP53* and *PIK3CA* (4 cases), in *TP53* and *AKT1* (3 cases), or in other gene combinations. Details of mutations are provided in **Table A in S1 File**.

In a subset of 12 samples (2 luminal A and 10 TN cases selected randomly) analyzed with the 8-gene panel, we also performed whole-exome sequencing. With a median coverage of 200X in tumor DNA and 80X in blood DNA and more than 99.5% of mapped reads (see **Table B in S1 File**), we identified 2634 somatic mutations in coding regions, including 2128 non-synonymous SNVs and indels (see **Table C in S1 File**). All mutations found by targeted sequencing in the 8-gene panel were confirmed in the exome analysis. There was an average of 3.9 non-synonymous SNVs and indels per MB, with 2 samples carrying more than 6 mutations


A
TP53
PIK3CA
AKT1
CDH1
PTEN
RB1
NOTCH1
ERBB2
PRECAMA (N=126)
METABRIC+TCGA pre<=45 (N=321)
0
5
10
15
20
25
30
35
40
45
50
Sample mutated (%)

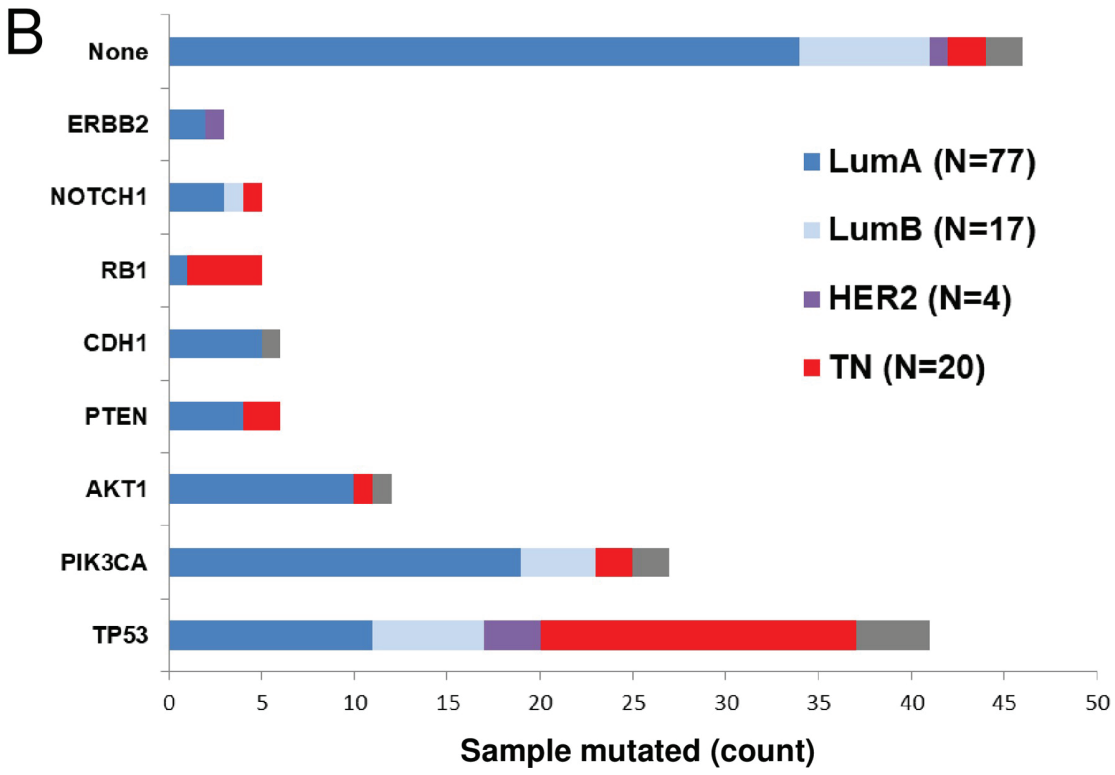

B
None
ERBB2
NOTCH1
RB1
CDH1
PTEN
AKT1
PIK3CA
TP53
LumA (N=77)
LumB (N=17)
HER2 (N=4)
TN (N=20)
0
5
10
15
20
25
30
35
40
45
50
Sample mutated (count)

Fig 1. Occurrences of mutations in 8 BC genes. (A) Gene mutation frequencies in PRECAMA samples are compared with those observed in a dataset of premenopausal women selected from the TCGA and METABRIC BC series [19, 20]. (B) IHC subtype distributions of samples according to their mutation status. Luminal A: ER+/HER2-; luminal B: ER+/HER2+; HER2-enriched: ER-/HER2+; triple-negative: ER-/PR-/HER2-.

https://doi.org/10.1371/journal.pone.0210372.g001

per MB (**Fig 2A, top panel**). The top mutated genes included four cancer genes (*TP53*, *RB1*, *PIK3CA*, and *AHNAK*) and several large genes, such as mucin genes and the *TTN* gene (**Fig 2A, middle panel**). *AHNAK* has recently been described as a novel tumor suppressor gene in BC, especially in the TN subtype, acting via different signaling pathways, such as AKT/MAPK or TGFβ [28, 29]. The *AHNAK* mutations, like the *RB1* mutations, were all in TN cases. However, the impact of *AHNAK* mutations on protein function is unknown, because *AHNAK* is a large gene and 75% of the mutations were predicted as benign by PolyPhen-2 [30]. In TN cases ($N = 10$), there were 92 cancer genes mutated, dominated by *TP53*, which was mutated in all samples, and with 14 other cancer genes mutated in more than one sample (**Fig 2B**). Pathway enrichment analysis of potential driver mutations in these TN samples (**Table D in S1 File**) showed enrichment for several growth factor signaling pathways and for pathways involved in insulin receptor signaling, telomere maintenance, transmembrane transport of small molecules, and G1 checkpoint or O-glycan biosynthesis (**Fig 2C** and **Table E in S1 File**).

Mutation patterns and signatures in premenopausal BC

To study the underlying mutational processes involved in the development of preBC tumors in the studied populations, we analyzed somatic mutation patterns in the *TP53* gene and at the exome-wide level. **Fig 3A** shows the distribution of *TP53* mutation types in PRECAMA tumors and in tumors from young women (< = 45 years) from other datasets. There was a higher proportion of G:C>T:A mutations in PRECAMA compared with the IARC TP53 Database (p-value = 0.004) or TCGA/METABRIC (p-value = 0.05) datasets. In fact, G:C>T:A was the most frequent type, followed by indels, in PRECAMA, whereas G:C>A:T at CpG was the most frequent type in the other datasets. The overall distribution of *TP53* mutation types was not significantly associated with IHC subtypes in PRECAMA samples (p-value = 0.06), although cases with indels were more frequently of the TN subtype (**Fig 3B**). *TP53* indels were truncating mutations (predicted to result in loss of p53 protein expression) in 6/10 cases, and 5/6 of these truncating mutations were indeed associated with null p53 IHC staining (see **Table A in S1 File**). Therefore, although the presence of frequent *TP53* truncating mutations in the TN subtype was similar to previous reports [31], the high frequency of G:C>T:A mutations in PRECAMA was unexpected. To validate this result, we used another dataset from a whole-genome sequencing study [6], and PRECAMA samples did show a significantly higher frequency of G:C>T:A mutations (**S2B Fig**) (p-value = 0.02). Mutational signatures at the exome-wide level were analyzed using a dataset including the 12 PRECAMA samples and 453 BC samples from TCGA (including both preBC and postBC cases; see Materials and Methods). We identified 6 signatures that matched with previously reported signatures (**Fig 4A and Table F in S1 File**). The estimated contribution of each signature to the mutation load in PRECAMA samples (**Fig 4B**) showed that 5/6 signatures had a contribution above 20% in at least one sample. Sig.A had the highest median contribution in these samples (24.3%). Sig.A matched with COSMIC signature-3, which has been established as a biomarker of homologous recombination defects through genetic and epigenetic inactivation of the *BRCA1/2* pathway, a distinctive feature of basal-like tumors [6, 10, 32]. Sig.B, which contributed in 6/12 samples, matched with COSMIC signature-26, proposed to be linked to defective DNA repair and previously reported in BC. Sig.C, which contributed in 6/12 samples, matched with several

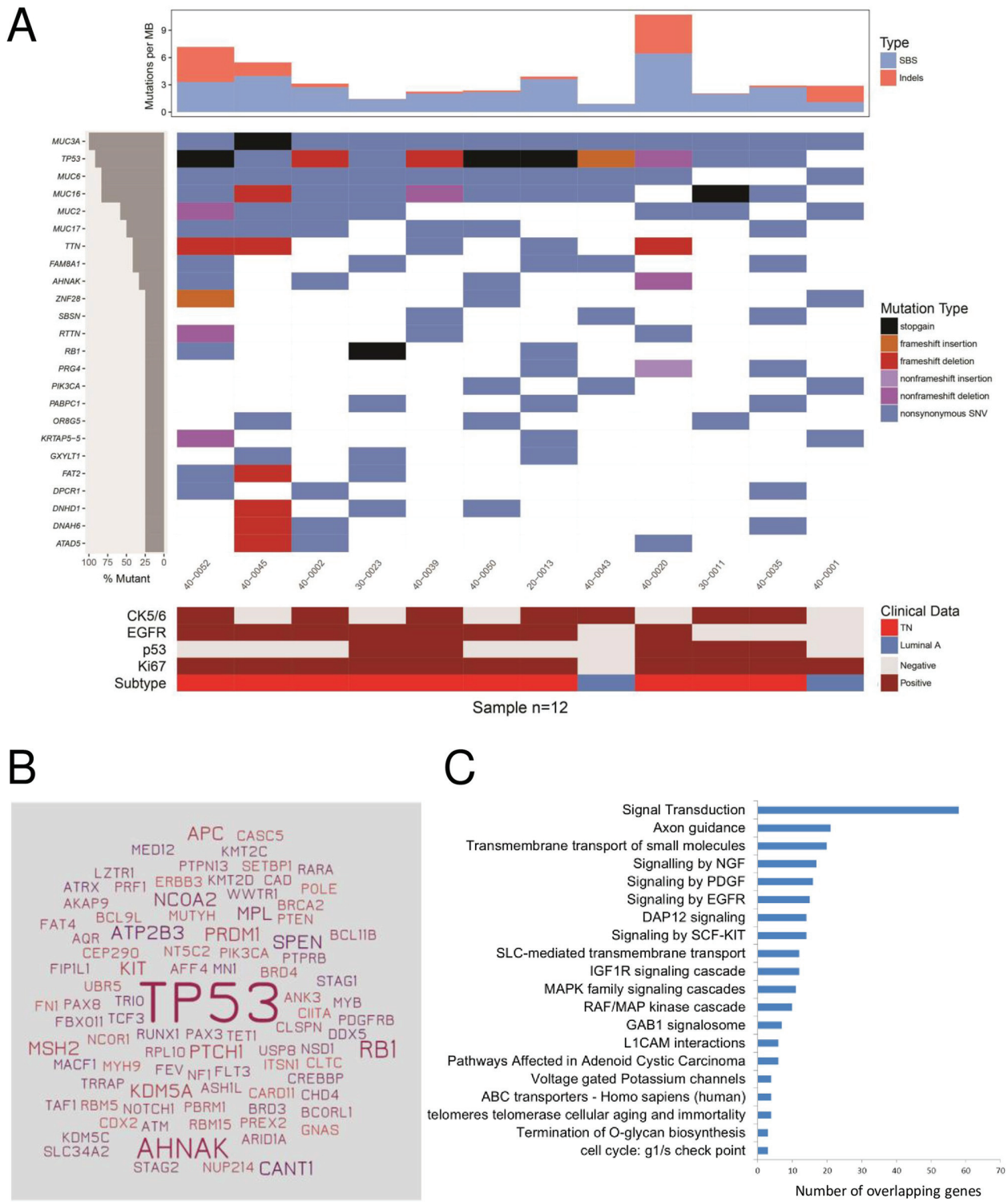


Fig 2. Whole-exome sequencing results in 12 PRECAMA samples. Only coding non-silent somatic mutations are considered. (A) Mutation rates (top panel), top mutated genes and their mutation types (middle panel), and IHC features (lower panel), sorted by top mutated genes. Luminal A: ER +/HER2-; triple-negative: ER-/PR-/HER2-. (B) All cancer genes somatically mutated in the 10 TN samples are depicted; the size of gene names is proportional to the number of samples mutated for each gene. (C) Pathways enriched (q-value <0.05) in the list of genes mutated in TN cases with allele frequency >20% and predicted deleterious/probably deleterious by PolyPhen-2 (N = 333 genes). Number of overlapping genes in each pathway is shown.

https://doi.org/10.1371/journal.pone.0210372.g002

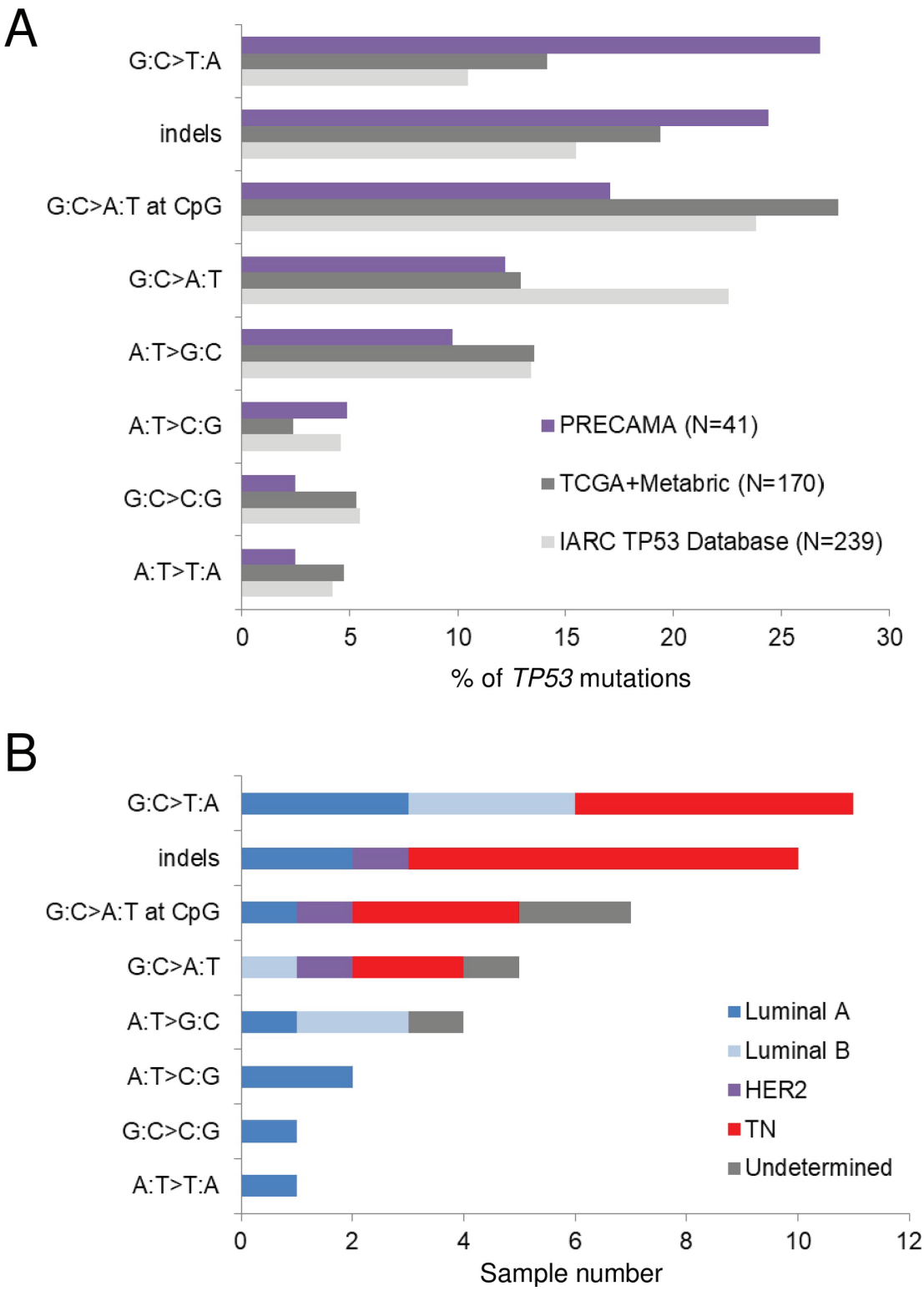


Fig 3. Distribution of *TP53* mutation types in preBC. (A) Distribution of mutation types in PRECAMA samples is compared with those observed in women 45 years old or younger selected from the TCGA and METABRIC BC series [19, 20] or the IARC TP53 Database [21]. (B) IHC subtype distributions of PRECAMA samples in each mutation type category. Luminal A: ER+/HER2-; luminal B: ER+/HER2+; HER2-enriched: ER-/HER2+; triple-negative: ER-/PR-/HER2-.

https://doi.org/10.1371/journal.pone.0210372.g003

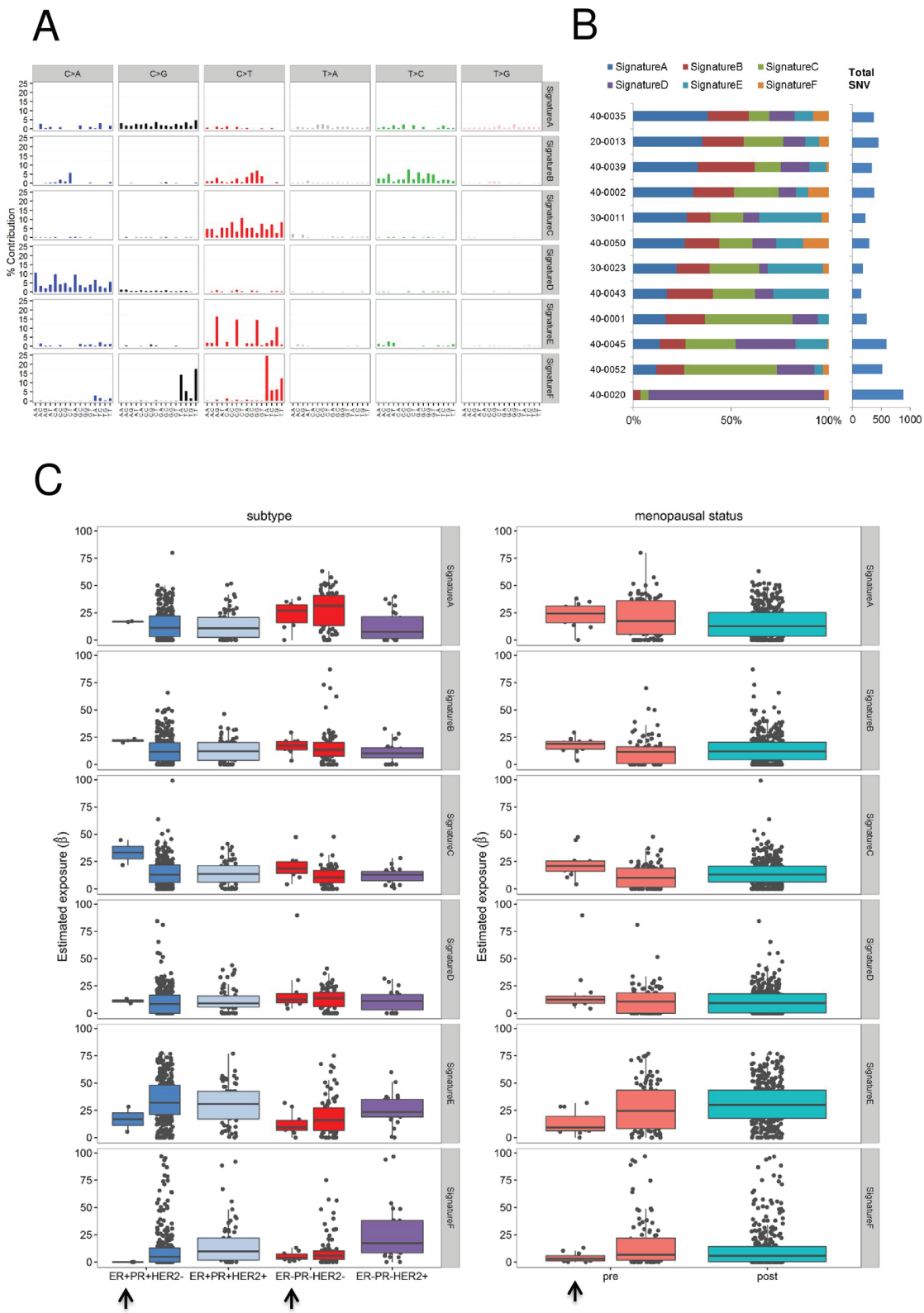


Fig 4. Mutational signatures identified in TCGA and PRECAMA samples, and their relationship with tumor subtype and patient menopausal status. (A) The 6 mutational signatures identified in 453 TCGA samples (including preBC and postBC) plus 12

PRECAMA samples. The 6 types of base substitutions are color-coded and further stratified by their adjacent 5′ and 3′ sequence context. Sig.A matches with COSMIC signature-3; Sig.B matches with COSMIC signature-26; Sig.C matches with COSMIC signatures-11/19/23/30 and experimental signatures of MNU and MNNG; Sig.D matches with COSMIC signature-18; Sig.E matches with COSMIC signature-1; and Sig.F matches with COSMIC signatures-2/13 (see **Table A in S1 File**). (B) Percentage contributions of the 6 mutational signatures to the SNVs found in PRECAMA samples. (C) Percentage contributions of the 6 mutational signatures in the PRECAMA and TCGA samples stratified by tumor IHC subtypes (left graphs) and by menopausal status (right graphs). PRECAMA samples are indicated with arrows.

https://doi.org/10.1371/journal.pone.0210372.g004

signatures characterized by C>T mutations outside CpG sites, including experimental signatures induced by alkylating agents (MNNG and MNU) in rodent systems [26, 27], COSMIC signature-11, observed in recurrent brain tumors of patients treated with MNNG [33], and COSMIC signature-30, of unknown origin but previously observed in some BC. Sig.D, which matched with COSMIC signature-18, was mainly observed in one sample, where it contributed to 90% of the mutation load and where the overall mutation load was the highest. The origin of this signature in BC remains to be established, but it has recently been associated with germline mutation in the repair enzyme *MUTYH* in colorectal and adrenocortical carcinomas [34, 35]. Interestingly, the sample in which Sig.D dominated carried a truncating somatic mutation in *MUTYH* (see **Table A in S1 File**). Finally, Sig.E, characterized by C>T mutations at a CpG site and matching with COSMIC signature-1, known to be due to spontaneous deamination of 5-methylcytosine (also referred as the "age" signature), had a contribution of at least 20% in only 3 samples.

As shown in **Fig 4C**, we explored possible systematic differences of signature contributions between IHC subtypes, and by menopausal status or study source (TCGA vs PRECAMA). Sig.F, which matched with COSMIC signature-2 and COSMIC signature-13, linked to mutagenesis by APOBEC, was more prevalent in HER2-enriched subtype cases and underrepresented in TN cases (p-value $< 5 \times 10^{-4}$, permutation test), as reported previously [36]. This is consistent with the fact that we did not find a strong contribution of the APOBEC signature in the PRECAMA samples (median contribution: 2.9%) because we analyzed only TN and luminal A cases. COSMIC signature-3 (Sig.A) was enriched in TN cases (p-value $< 5 \times 10^{-4}$) and preBC (p-value = 0.03). This signature was the predominant one in PRECAMA TN cases (median contribution: 26.8%). The contribution of the "age" signature (Sig.E) was lower in TN cases than in all other subtypes (p-value $< 5 \times 10^{-4}$) and also lower in preBC compared with postBC (p-value = 0.006). It was the lowest in PRECAMA samples (13.4% vs 30.5%, p-value = 1×10^{-3}). The contribution of Sig.D was slightly higher in TN cases compared with all other subtypes (p-value = 0.01) and was the main contributor to the mutation load in one PRECAMA sample (**Fig 4B**). Because this sample carried a somatic mutation in *MUTYH*, and a recent study found germline mutations in *MUTYH* in young women with BC [37], it will be interesting to further study the role of *MUTYH* alteration in TN and preBC. In stratified analyses by IHC subtypes, the contributions of signatures in PRECAMA TN cases were similar to those observed in TCGA TN samples, except for Sig.C (contribution was higher in PRECAMA than in TCGA TN samples, p-value = 0.006; median contributions: 18.6% vs 10.4%). Because Sig.C matched with several signatures, including signatures linked to exposure to alkylating agents not expected in these treatment-naive samples, its origin remains to be established. There was no effect of menopausal status on the contributions of signatures when taking into account IHC subtype using linear models (all p-values > 0.16; permutation tests of linear regression model).

Discussion

The results obtained in this pilot phase of the PRECAMA study demonstrate the feasibility of advanced genomic analyses of the tumor and blood samples collected at multiple sites in LA.

They provide a preview of the molecular features of preBC in that population, with interesting mutational patterns that deserve further study.

Indeed, more than 92% of samples processed for IHC analyses were successfully scored for 7 markers (only 20/253 were excluded due to absence of invasive tumor or insufficient tissue for testing), and 80% of samples processed for DNA extraction yielded DNA quantities and quality compatible with genomic analyses (136/172 samples yielded more than 200 ng of DNA). With a target of 1200 cases recruited for the full study (with Guatemala and Brazil joining the study), this will be the largest series of preBC in Latin American women with genomic characterization of the tumors.

The IHC analyses showed a majority of ER-positive cases and a proportion of TN subtype similar to previous reports in Hispanic women [38]. The overall prevalence of ER-negative tumors in PRECAMA was substantiated by sequencing results on the 8-gene panel analyzed here. Indeed, *TP53* mutations, which are strongly associated with ER-negative status [7, 39], were found in 32.5% of the cases, consistent with an overall 28% of ER-negative cases. Also, the frequency of *AKT1* mutations, typical of ER-positive cases [20, 40], was higher in PRECAMA than in the comparative dataset of young women. Continued enrollment will enable us to determine more precise estimates of subtype distribution in PRECAMA and to explore potential differences in tumor subtype distributions between countries.

Although the overall tumor characteristics were more similar to those described in postBC than in preBC from other series, IHC staining with Ki67 showed high levels of staining in these preBC samples, even in luminal A cases (72% positive cases), which is consistent with previous reports on preBC [41, 42]. Liao et al. (2015) [43] recently compared the molecular features of preBC versus postBC from the TCGA and METABRIC datasets using multi-omic data integration. They reported no difference in gene expression between preBC and postBC in ER-negative cases but significant differences in ER-positive cases, with activation of integrin signaling and EGFR pathways and TGFβ as the top upstream regulator in preBC. It would therefore be important in future studies to assess whether activation of these pathways drives the level of proliferation reflected by high Ki67 positivity in ER-positive preBC, because they may be potential clinical targets.

The characteristics of the mutations found by target sequencing of the 8-gene panel were similar to those observed in other series of BC, with classical hotpots found in *AKT1* and *PIK3CA*, a majority of missense mutations found in *TP53*, a higher proportion of truncating *TP53* mutations in TN cases compared with other subtypes, and an expected distribution of mutated genes within IHC subtypes. However, an interesting difference in the distribution of *TP53* single base substitutions was observed. The most frequent *TP53* mutation type was G:C>T:A, which represented 27% of all *TP53* mutations. This proportion of G:C>T:A mutations was 1.5–3.3 times those observed in the comparative datasets used here, matching figures reported in lung cancers linked to exposure to polycyclic aromatic hydrocarbons [44, 45]. This pattern is therefore unexpected in BC. These G:C>T:A mutations do not exhibit a strand bias, do not cluster at any hotspot, and seemed similarly distributed within IHC subtypes or country of origin, although the numbers are still too low to enable any conclusion to be drawn. Because these results may suggest a specific, as-yet unknown, mutational process at the origin of *TP53* mutations, it will be important to confirm them in the full PRECAMA study.

Exome-wide mutation profiling of a subset of basal-like TN tumors confirmed that *TP53* and *RB1* were the only cancer genes recurrently affected by deleterious mutations (>2 samples). These results are concordant with previous reports on TNBC of basal-like type that showed a predominance of *TP53* mutations and of *TP53* and *RB1* pathway alterations [40, 46]. These reports also suggest activation of the PIK3CA/AKT pathway, based on gene copy number analyses (*PIK3CA* gene amplification, *PTEN* gene deletion) and protein phosphorylation

assays [40]. Here, we found one activating *PIK3CA* mutation in 10 TNBC samples, which is in the range of previous reports (9%). However, because we limited our analyses to SNVs and small indels, we could not further assess the functionality of the *PIK3CA* pathway. Pathway analysis of potentially functional mutations across all genes showed enrichment of signal transduction pathways including EGFR, PDGF, and IGF1R, and mutational signatures showed a large contribution of DNA repair defects to the mutation load. These overall results on TN cases are consistent with our previous analyses of another series of TN cases from Mexico, in which transcriptomics analyses showed an overexpression of growth-promoting signals (including *EGFR*, *PDGFR*, and *PIK3CA*), a repression of cell cycle control pathways (*TP53* and *RB1*), and a deregulation of DNA repair pathways [47].

Our exploratory analysis of exome-wide mutational signatures in relation to IHC subtype and menopausal status in the TCGA and PRECAMA samples showed that the contributions of mutational signatures are determined by the tumor subtype but not the menopausal status, and that PRECAMA TN cases showed contributions similar to TCGA TN samples for 5/6 signatures identified in the analyzed set.

Some limitations of the results presented should be noted. First, the prevalences of IHC subtypes are based on still-limited numbers and may therefore not be representative of the distribution at the population level. Second, confirmation of HER2 status by FISH could not be done in this pilot phase, and therefore the prevalence of the luminal B or HER2-enriched subtypes may be under- or over-estimated. Third, the exome analyses have been performed on a limited number of cases to establish the feasibility of these assays. Results on this small set did show feasibility and enabled us to identify both similarities and differences in genomic alterations compared with other series of BC. Analysis of the full series will determine whether any specific genomic feature may characterize preBC in women in LA.

Conclusions

These pilot results on PRECAMA tumors give a preview of the molecular features of preBC in LA. Although the overall mutation burden was as expected from data in other populations, mutational patterns observed in *TP53* suggested possible differences in mutagenic processes giving rise to these tumors compared with other populations. Further -omics analyses of a larger number of PRECAMA cases in the near future will enable the investigation of relationships between these molecular features and etiological factors.

Supporting information

S1 File. Tables A-F Table A: Demographics and molecular characteristics of cases analyzed by next-generation sequencing; Table B: Whole-exome sequencing data metrics; Table C: Mutations in coding regions from whole-exome sequencing and mutation calling with Strelka; Table D: List of mutated genes in TN cases with mutations present at an allele frequency >20% and predicted to affect protein function (splice, truncating, and non-synonymous predicted deleterious/probably deleterious by PolyPhen-2) ($N = 333$); Table E: Pathway analysis of 333 altered genes in TN samples; Table F: Cosine similarity values for the comparisons between each of the 6 extracted signatures and 37 published signatures.
(XLSX)

S1 Fig. Distribution of IHC subtypes in PRECAMA samples and in preBC extracted from METABRIC and TCGA studies. Comparison of the distribution of IHC subtypes observed in PRECAMA and in preBC from a dataset extracted from METABRIC and TCGA (see Materials and Methods). Luminal A: ER+/HER2-; luminal B: ER+/HER2+; HER2-enriched: ER-/HER2

+; triple-negative: ER-/PR-/HER2-.
(PDF)

S2 Fig. Mutation characteristics and distribution of IHC subtypes in PRECAMA samples compared with an independent dataset of preBC. Data on 123 preBC cases with receptor status information reported in Nik-Zainal et al. (2016) were retrieved from supplementary materials (clinical information) or from COSMIC (mutation data) [6]. (A) Occurrences of mutations in the 8 BC genes analyzed in PRECAMA. (B) Distribution of *TP53* mutation types in preBC cases. (C). Comparison of the distribution of IHC subtypes observed in preBC in the two datasets. Luminal A: ER+/HER2-; luminal B: ER+/HER2+; HER2-enriched: ER-/HER2+; triple-negative: ER-/PR-/HER2-.
(PDF)

Acknowledgments

The PRECAMA team includes International Agency for Research on Cancer (Coordinating Center) investigators and staff: Carine Biessy, Magali Olivier, Sabina Rinaldi, Isabelle Romieu; investigators and staff in Chile: Eva Bustamante, Fancy Gaete, Maria Luisa Garmendia, Jose Soto; investigators and staff in Colombia: Alberto Angel, Carlos Andres Ossa, William H. Arias, Gabriel Bedoya, Mauricio Borrero, Alicia Cock-Rada, Israel Díaz-Yunez, Carolina Echeverri, Fernando Herazo, Angel Hernández, Roberto Jaramillo, Edgar Navarro, Yorlany Rodas Cortes, Gloria I. Sanchez; investigators and staff in Costa Rica: Bernal Cortes, Paula Gonzalez, Diego Guillen, Viviana Loría, Rebecca Ocampo, Carolina Porras, Ana Cecilia Rodriguez; investigators and staff in Mexico: Gabriela Torres-Meija, Angelica Angeles-Llerenas, Jenny Tejeda; and investigators and staff in Seattle: Peggy Porter. Ana Cecilia Rodriguez was part of the Proyecto Guanacaste when this work was initiated.

The authors are grateful for the substantial support provided by the research nurses and health workers, as well as by Tracy Lignini, Mathilde His, Dacia Cristin, Cecile Le Duc, Jordi de Battle, Talita Duarte-Salles, and Ana Cristiana Ocampo.

The authors also wish to thank the women participating in the project for their time and commitment.

Author Contributions

Conceptualization: Magali Olivier, Gloria Inés Sánchez, Isabelle Romieu, Peggy Porter, Sabina Rinaldi.

Data curation: Vincent Cahais.

Formal analysis: Magali Olivier, Liacine Bouaoun, Alexis Robitaille, Graham Byrnes, Peggy Porter.

Methodology: Magali Olivier, Stephanie Villar, Adriana Heguy, Florence Le Calvez-Kelm, Peggy Porter, Jamie Guenthoer.

Resources: Gabriela Torres-Mejía, Isabel Alvarado-Cabrero, Fazlollah Shahram Imani-Razavi, Roberto Jaramillo, Carolina Porras, Ana Cecilia Rodriguez, Maria Luisa Garmendia, José Luis Soto.

Writing – original draft: Magali Olivier.

Writing – review & editing: Magali Olivier, Graham Byrnes, Florence Le Calvez-Kelm, Gabriela Torres-Mejía, Isabel Alvarado-Cabrero, Isabelle Romieu, Peggy Porter, Sabina Rinaldi.

References

1. Amadou A, Torres-Mejia G, Hainaut P, Romieu I. Breast cancer in Latin America: global burden, patterns, and risk factors. Salud Publica Mex. 2014; 56(5):547–554. PMID: 25604300
2. Porter PL. Global trends in breast cancer incidence and mortality. Salud Publica Mex. 2009; 51 Suppl 2: s141–146.
3. Goldhirsch A, Wood WC, Coates AS, Gelber RD, Thurlimann B, Senn HJ, Panel m. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Annals of oncology: official journal of the European Society for Medical Oncology. 2011; 22(8):1736–1747.
4. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486(7403):346–352. https://doi.org/10.1038/nature10983 PMID: 22522925
5. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA et al. Molecular portraits of human breast tumours. Nature. 2000; 406(6797):747–752. https://doi.org/10.1038/35021093 PMID: 10963602
6. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016; 534(7605):47–54. https://doi.org/10.1038/nature17676 PMID: 27135926
7. Langerod A, Zhao H, Borgan O, Nesland JM, Bukholm IR, Ikdahl T, Karesen R, Borresen-Dale AL, Jeffrey SS. TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. Breast cancer research: BCR. 2007; 9(3):R30. https://doi.org/10.1186/bcr1675 PMID: 17504517
8. Huo D, Hu H, Rhie SK, Gamazon ER, Cherniack AD, Liu J, Yoshimatsu TF, Pitt JJ, Hoadley KA, Troester M et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. JAMA oncology. 2017.
9. Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harbor perspectives in biology. 2010; 2(1):a001008. https://doi.org/10.1101/cshperspect.a001008 PMID: 20182602
10. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L et al. Signatures of mutational processes in human cancer. Nature. 2013; 500(7463):415–421. https://doi.org/10.1038/nature12477 PMID: 23945592
11. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell reports. 2013; 3(1):246–259. https://doi.org/10.1016/j.celrep.2012.12.008 PMID: 23318258
12. Gormally E, Caboux E, Vineis P, Hainaut P. Circulating free DNA in plasma or serum as biomarker of carcinogenesis: practical aspects and biological significance. Mutation research. 2007; 635(2–3):105–117. https://doi.org/10.1016/j.mrrev.2006.11.002 PMID: 17257890
13. Vaught JB, Caboux E, Hainaut P. International efforts to develop biospecimen best practices. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2010; 19(4):912–915.
14. Ardin M, Cahais V, Castells X, Bouaoun L, Byrnes G, Herceg Z, Zavadil J, Olivier M. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. BMC bioinformatics. 2016; 17:170. https://doi.org/10.1186/s12859-016-1011-z PMID: 27091472
15. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012; 28(14):1811–1817. https://doi.org/10.1093/bioinformatics/bts271 PMID: 22581179
16. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic acids research. 2013; 41(Database issue):D793–800. https://doi.org/10.1093/nar/gks1055 PMID: 23143270
17. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nature reviews Cancer. 2004; 4(3):177–183. https://doi.org/10.1038/nrc1299 PMID: 14993899
18. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods. 2013; 10(11):1081–1082. https://doi.org/10.1038/nmeth.2642 PMID: 24037244
19. Pereira B, Chin SF, Rueda OM, Vollan HK, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut SJ et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nature communications. 2016; 7:11479. https://doi.org/10.1038/ncomms11479 PMID: 27161491

20. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012; 486(7403):400–404. https://doi.org/10.1038/nature11017 PMID: 22722201

21. Bouaoun L, Sonkin D, Ardin M, Hollstein M, Byrnes G, Zavadil J, Olivier M. TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. Human mutation. 2016; 37 (9):865–876. https://doi.org/10.1002/humu.23035 PMID: 27328919

22. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012; 2(5):401–404. https://doi.org/10.1158/2159-8290.CD-12-0095 PMID: 22588877

23. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013; 6(269):pl1.

24. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004; 101(12):4164–4169. https://doi.org/10.1073/pnas.0308531101 PMID: 15016911

25. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC bioinformatics. 2010; 11(1):367.

26. Westcott PMK, Halliwill KD, To MD, Rashid M, Rust AG, Keane TM, Delrosario R, Jen K-Y, Gurley KE, Kemp CJ et al. The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. Nature. 2015; 517(7535):489–492. https://doi.org/10.1038/nature13898 PMID: 25363767

27. Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallee MP, McKay J, Nedelko T, Muehlbauer KR, Marusawa H et al. Modelling mutational landscapes of human cancers in vitro. Scientific reports. 2014; 4:4482. https://doi.org/10.1038/srep04482 PMID: 24670820

28. Lee IH, Sohn M, Lim HJ, Yoon S, Oh H, Shin S, Shin JH, Oh SH, Kim J, Lee DK et al. Ahnak functions as a tumor suppressor via modulation of TGFbeta/Smad signaling pathway. Oncogene. 2014; 33 (38):4675–4684. https://doi.org/10.1038/onc.2014.69 PMID: 24662814

29. Chen B, Wang J, Dai D, Zhou Q, Guo X, Tian Z, Huang X, Yang L, Tang H, Xie X. AHNAK suppresses tumour proliferation and invasion by targeting multiple pathways in triple-negative breast cancer. J Exp Clin Cancer Res. 2017; 36(1):65. https://doi.org/10.1186/s13046-017-0522-4 PMID: 28494797

30. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20.

31. Holstege H, Horlings HM, Velds A, Langerod A, Borresen-Dale AL, van de Vijver MJ, Nederlof PM, Jonkers J. BRCA1-mutated and basal-like breast cancers have similar aCGH profiles and a high incidence of protein truncating TP53 mutations. BMC cancer. 2010; 10:654. https://doi.org/10.1186/1471-2407-10-654 PMID: 21118481

32. Polak P, Kim J, Braunstein LZ, Karlic R, Haradhavala NJ, Tiao G, Rosebrock D, Livitz D, Kubler K, Mouw KW et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nature genetics. 2017.

33. Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY, Fouse SD, Yamamoto S, Ueda H, Tatsuno K et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. Science (New York, NY). 2014; 343(6167):189–193.

34. Pilati C, Shinde J, Alexandrov LB, Assie G, Andre T, Helias-Rodzewicz Z, Ducoudray R, Le Corre D, Zucman-Rossi J, Emile JF et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. J Pathol. 2017; 242(1):10–15. https://doi.org/10.1002/path.4880 PMID: 28127763

35. Viel A, Bruselles A, Meccia E, Fornasarig M, Quaia M, Canzonieri V, Policicchio E, Urso ED, Agostini M, Genuardi M et al. A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. EBioMedicine. 2017; 20:39–49. https://doi.org/10.1016/j.ebiom.2017.04.022 PMID: 28551381

36. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nature genetics. 2013; 45(9):970–976. https://doi.org/10.1038/ng.2702 PMID: 23852170

37. Rummel SK, Lovejoy L, Shriver CD, Ellsworth RE. Contribution of germline mutations in cancer predisposition genes to tumor etiology in young women diagnosed with invasive breast cancer. Breast cancer research and treatment. 2017.

38. Lara-Medina F, Perez-Sanchez V, Saavedra-Perez D, Blake-Cerda M, Arce C, Motola-Kuba D, Villarreal-Garza C, Gonzalez-Angulo AM, Bargallo E, Aguilar JL et al. Triple-negative breast cancer in Hispanic patients: high prevalence, poor prognosis, and association with menopausal status, body mass

index, and parity. Cancer. 2011; 117(16):3658–3669. https://doi.org/10.1002/cncr.25961 PMID: 21387260

39. Olivier M, Langerod A, Carrieri P, Bergh J, Klaar S, Eyfjord J, Theillet C, Rodriguez C, Lidereau R, Bieche I et al. The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. Clinical cancer research: an official journal of the American Association for Cancer Research. 2006; 12 (4):1157–1167.

40. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61–70. https://doi.org/10.1038/nature11412 PMID: 23000897

41. Yamashita H. Tumor biology in estrogen receptor-positive, human epidermal growth factor receptor type 2-negative breast cancer: Mind the menopausal status. World J Clin Oncol. 2015; 6(6):220–224. https://doi.org/10.5306/wjco.v6.i6.220 PMID: 26677435

42. Regan MM, Pagani O, Francis PA, Fleming GF, Walley BA, Kammler R, Dell'Orto P, Russo L, Szoke J, Doimi F et al. Predictive value and clinical utility of centrally assessed ER, PgR, and Ki-67 to select adjuvant endocrine therapy for premenopausal women with hormone receptor-positive, HER2-negative early breast cancer: TEXT and SOFT trials. Breast cancer research and treatment. 2015; 154(2):275–286. https://doi.org/10.1007/s10549-015-3612-z PMID: 26493064

43. Liao S, Hartmaier RJ, McGuire KP, Puhalla SL, Luthra S, Chandran UR, Ma T, Bhargava R, Modugno F, Davidson NE et al. The molecular landscape of premenopausal breast cancer. Breast cancer research: BCR. 2015; 17:104. https://doi.org/10.1186/s13058-015-0618-8 PMID: 26251034

44. DeMarini DM, Landi S, Tian D, Hanley NM, Li X, Hu F, Roop BC, Mass MJ, Keohavong P, Gao W et al. Lung tumor KRAS and TP53 mutations in nonsmokers reflect exposure to PAH-rich coal combustion emissions. Cancer research. 2001; 61(18):6679–6681. PMID: 11559534

45. Olivier M, Hussain SP, Caron de Fromentel C, Hainaut P, Harris CC. TP53 mutation spectra and load: a tool for generating hypotheses on the etiology of cancer. IARC scientific publications. 2004;(157):247–270. PMID: 15055300

46. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012; 486 (7403):395–399. https://doi.org/10.1038/nature10933 PMID: 22495314

47. Vaca-Paniagua F, Alvarez-Gomez RM, Maldonado-Martinez HA, Perez-Plasencia C, Fragoso-Ontiveros V, Lasa-Gonsebatt F, Herrera LA, Cantu D, Bargallo-Rocha E, Mohar A et al. Revealing the Molecular Portrait of Triple Negative Breast Tumors in an Understudied Population through Omics Analysis of Formalin-Fixed and Paraffin-Embedded Tissues. PloS one. 2015; 10(5):e0126762. https://doi.org/10.1371/journal.pone.0126762 PMID: 25961742

Human Papillomavirus Type 38 alters wild-type p53 activity to promote cell proliferation via the down-regulation of Integrin Alpha-1 expression

[a]Maria Carmen Romero-Medina, [a]Assunta Venuti, [b]Roberto Ferrari, Djamel Saidj, [a*]Giusi Melita, [a]Alexis Robitaille, [a*]Maria Grazia Ceraolo, [a*]Laura Pacini, [a]Cecilia Sirand, [c]Daniele Viarisio, [a]Valerio Taverniti, [a]Purnima Gupta, [d]Mariafrancesca Scalise, [d]Cesare Indiveri, [a]Rosita Accardi, [a#]Massimo Tommasino

[a]International Agency for Research on Cancer (IARC), World Health Organization, 150 Cours Albert Thomas, 69372 Lyon Cedex 08, France; [c]Centre for Genomic Regulation, Barcelona, Spain; [c]Deutsches Krebsforschungszentrum (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany; [d]Unit of Biochemistry and Molecular Biotechnology, Department DiBEST (Biologia, Ecologia, Scienze della Terra), University of Calabria, via P. Bucci 4C, 87036, Arcavacata di Rende, Italy

Running Head: HPV38 E6 and E7 Inhibit ITG-alpha expression

[#]Corresponding author: e-mail tommasinom@iarc.fr.
*Present addresses: Giusi Melita, Laboratory of Cell and Gene Therapy "Stefano Verri", Tettamanti Research Center, ASST-Monza, San Gerardo Hospital, Monza, Italy; Maria Grazia Ceraolo, Tumor Immunology Unit, Division of Immunology, Transplantation and Infectious Diseases, San Raffaele Scientific Institute, Milano, Italy; Laura Pacini, Division of Molecular Pathology, The Institute of Cancer Research, London, 24 SW3 6JB, United Kingdom
Daniele Viarisio, Freelance

KEYWORDS: Integrin 1-alpha, HPV38, primary keratinocytes, p53

Abstract

Beta human papillomaviruses (HPVs) appear to cooperate with UV in the development of cutaneous squamous cell carcinoma (cSCC). Accordingly, beta HPV E6 and E7 oncoproteins display transforming activities *in vitro* and *in vivo* models. Here, we show that beta HPV38 alters the tumour suppressor functions of wild-type (WT) p53, promoting cellular proliferation. WT p53 is accumulated in human keratinocytes (HK) expressing HPV38 E6 and E7. A proportion of this WT p53 form is phosphorylated at S392 by double-stranded (ds) RNA-dependent protein kinase (PKR) and form a complex with DNMT1 acting as a transcriptional repressor, which is recruited to the Integrin alpha 1 promoter (ITGA-1). Ectopic expression of ITGA-1 in HPV38 E6/E7 HK promotes EGFR degradation, inhibition of cellular proliferation and cellular death.
ITGA-1 expression was also inhibited in the skin of HPV38 transgenic mice. These animals have high susceptibility to UV-induced skin carcinogenesis, and after long-term UV irradiation, they accumulate many DNA exome mutations, including in p53 and ITGA-1 genes.
In summary, our study shows that beta HPV38 can convert p53 functions from a tumour suppressor to oncoprotein via the formation of a transcriptionally repressive complex and that inactivation of ITGA-1 plays a key role in skin carcinogenesis.

INTRODUCTION

HPV phylogenetic tree classifies the different genotypes into 5 genera. Mucosal HPV types belonging to the genus alpha are referred to as high-risk (HR) HPV types and are the etiological agents of several human cancers. In addition, beta HPV types, together with the ultra-violet (UV) radiations, appear to be involved in cSCC development [1-4]. Approximately 50 beta HPV types have been fully characterized so far. They are subdivided in five species (Beta 1-5). Beta-1 and Beta-2 species contributes to the major bulk and are abundantly detected in the skin [5]. Epidemiological studies revealed that individuals with history of cSCC are more frequently positive for viral infection markers, such as beta HPV DNA in the skin and/or major capsid protein L1 antibodies than the general population [6-11].
Studies in *in vitro* experimental models have demonstrated the transforming properties of the early gene products, E6 and E7, from some beta HPV types, e.g. HPV8 and HPV38 [12-19]. Moreover, we previously reported that, beta-2 HPV38 E6 and E7 are able to immortalize primary human HK [20] similarly to the mucosal HR HPV types, via the inactivation of the tumour suppressor gene products, p53 and retinoblastoma [20,21]. However, in comparison to the mucosal HR HPV types that promote p53 degradation via the proteasome pathway, HPV38 leads to accumulation of p53, possibly with an altered transcriptional activity [21,22].

RESULTS

38HK proliferation is dependent on WT p53

We have previously shown that HPV38 E6 and E7 induce accumulation of p53 in immortalized human keratinocytes (38HK) (Caldeira et al 2003, J. Virol; Gabet et al. 2008 FASEB J). To clarify the role of this p53 form in 38HK, we have deleted p53 gene using CRISP/CAS9 technology.

Figures 1A and B shows that after 72 hours post-transfection the decrease of p53 levels correlate with a significant decrease of cellular proliferation. Similar results were obtained by inhibiting p53 functions by Pfithrin (Fig. S1A). Several isoforms of p53 have been identified with truncations at the N- or C-terminus and altered transcriptional functions and oncogenic properties[37]. Therefore, we next evaluate whether the decrease of cellular proliferation observed in 38HK after deletion of the endogenous p53 gene is linked to the loss of the full-length (FL) p53 form. To evaluate this possibility, we generate a retroviral vector that expresses a N-terminus HA-tagged p53 gene (Δ-CRISP), in which the third base of several codons was mutated (Fig. 1C). This mutated p53 gene encode a WT protein but is not targeted by the guide RNA that was designed to delete the endogenous p53 gene. In addition, two loxP elements were located immediately upstream and downstream of the Δ-CRISP p53 gene in order to modulate its expression via the Cre recombinase. Cre recombinase gene fused to a triple-mutant form of the human estrogen receptor that gains access to the nuclear compartment only after exposure to 4-hydroxytamoxifen (TMX) but not to the natural ligand 17β-estradiol was cloned in a second retroviral vector (Fig. 1C). 38HK were sequentially transduced with two recombinant retroviruses and subsequently the endogenous p53 gene was deleted by CRISP/CAS9. We observed that the modified 38HK line expressing ectopic levels of Δ-CRISP p53 gene had a higher proliferation rate than 38HK (Fig.1D). Importantly, after TMX addition and loss of HA-p53, the proliferation of these cells was rapidly reduced, while no a significant effect was observed when TMX was added to 38HK (Fig. 1D). To corroborate these findings, we have transduced 38HK with recombinant retroviruses that allow the synthesis of FL p53fused to HA-Tag at the N- or C-terminus. Both HA-p53 fusion proteins were detected by immunoblotting (Fig. 1F) and able to stimulate the proliferation of 38HK (Fig. 1G and S1B). Together, these findings show that HPV38 E6 and E7 induce a change in the biological properties of full-length WT p53 from tumour suppressor to pro-proliferation factor.

p53 and DNMT1 form a complex that is recruited to ITGA-1 promoter in 38HK

To valid the ChIP-seq data, we performed an analysis of ITGA-1 promoter to identify putative responsive elements (REs) using TFBind and JASPAR softwares and . The analyses revealed the presence of several p53 REs in a region upstream of the transcriptional start site, spanning from the -936 to -835 nucleotides (Fig. 3A). To evaluate whether these putative p53 REs have the ability to interact with p53, we performed an electromobility shift assay (EMSA) using oligos encompassing the WT or mutated REs. RE2 showed a stronger signal for p53 binding that was highly reduced upon mutation of the p53-binding motif (Fig. 3B, lines 2 and 5) or by competition with WT unlabeled probe, but less efficiently with mutated unlabeled probe (Fig. 3B, lines 7-9). These data were corroborated by chromatin immunoprecipitation (ChIP) experiments that showed a significant enrichment on RE2 compared to the negative control (Fig. 3C). Oligo pull-down experiments using biotinylated DNA probes, which contain a region of the ITGA-1 promoter encompassing the RE2, revealed that p53 was efficiently precipitated by the RE2 together with the epigenetic enzyme DNMT1, known to be associated with gene expression silencing (Fig. 3D). DNMT1 recruitment to ITGA-1 promoter was also confirmed by ChIP experiments (Fig. 3E).

Inhibition of p53 functions, by using the chemical inhibitor, pifithrin, severely impaired p53 and DNMT1 interaction with ITGA-1 promoter (Fig. S2A), indicating that two cellular proteins are

part of the same complex. Indeed, ChIP-reChIP experiments confirmed their interaction and their recruitment to p53RE2 of ITGA-1 promoter (Fig. 3F). Similarly to what has been observed with p53 inhibition by pifithrin, silencing the expression of DNMT1 by small interfering RNA (siRNA) significantly affected the recruitment of p53 (Fig. S2B). Together, these data show that p53 and DNMT1 form a complex and their interaction appears to be important for the binding to the p53 RE2 of the ITGA-1 promoter.

ITGA-1 expression is inhibited in 38HK in comparison to primary HK

Next we compared the mRNA levels of ITGA-1 in primary HK and 38HK. We observed that the expression of ITGA-1 was significantly down-regulated by the viral oncoproteins (Fig. 4A). Therefore, we focused our study on ITGA-1 and further validated its down-regulation in 38HK by using a TaqMan PCR. Also with this assay, we observed a statistically significant decrease in ITGA-1 mRNA levels in comparison with the HK (Fig. S3A). To evaluate whether the decrease of ITGA-1 mRNA levels is a direct consequence of the viral gene expression, and it is not due to the immortalization of the 38HK, we used, as an experimental model, human primary keratinocytes expressing the human telomerase reverse transcriptase gene (hTERT), which extends the lifespan of primary cells. ITGA-1 mRNA levels were also reduced in hTERT/HPKs expressing HPV38 E6 and E7 genes in comparison with the mock cells (Fig. S3B). The inhibition of ITGA-1 gene expression appears to be mainly associated with E6 protein (Fig. S3B). In addition, ectopic expression of HA-tagged full-length p53 at the N- or C-terminus in 38HK further repressed ITGA-1 expression (Fig. 4D and E). Next we evaluate whether disruption of p53/DNMT1 complex could lead to an activation of ITGA-1 expression. We observed that the knock-down of p53 by CRISPR/Cas9 in in 38HK resulted in an increase of ITGA-1 mRNA and protein levels (Fig. 4B and C).

We have shown above that silencing of DNMT1 expression resulted in a loss of recruitment of the p53/DNMT1 complex on the ITGA-1 promoter Fig. S2B). Accordingly, ITGA-1 mRNA levels increased upon inhibition of DNMT1 expression (Fig. 4F). Similarly, treatment with 5-Aza-2′-deoxycytidine (Aza), 2′-deoxycytidine analogue, a global demethylating agent, resulted in activation of ITGA-1 expression (Figure 4G). This event also coincided with the acetylation in histone 3 at K9 (H3K9), which is associated with transcriptional activation (Fig. 4H).

Finally, since the 38HK contains also high levels of ΔNp73α that inhibits the expression of p53 regulated genes [22], we evaluated the impact of its depletion on ITGA-1 expression. No significant changes in ITGA-1 mRNA levels were detected in 38HK transfected with ΔNp73α antisense or sense oligonucleotides (Fig. S3D).These findings show that p53/DNMT1 inhibits the ITGA-1 expression.

Full-length p53 phosphorylated at S392 plays a key role in the inhibition of ITGA-1 expression

We have previously shown that the accumulated p53 in 38HK is phosphorylated only at two serines (15 and 392) [22]. We next examined by DNA pull-down assay the binding of p53 Ser15 and/or Ser392 phosphorylated forms to ITGA-1 promoter. Fig. 5A shows that the major p53 form, able to bind the p53 RE2 of ITGA-1 promoter, is phosphorylated at Ser392. It has been previously

shown that the double-stranded (ds) RNA-dependent protein kinase (PKR) directly interacts with p53 and phosphorylates Ser392 [33,34]. Immunoblotting showed that HPV38 E6 and E7 activate PKR (Fig. 5B). In addition, blocking the PKR activity by the use of a chemical inhibitor, 2-aminopurine (2-AP) resulted always in a small, but significant reduction, of the levels of Ser392 p53 form, indicating that PKR is not the only cellular kinase involved in p53 phosphorylation in 38HK (Fig. 5C). Despite the small reduction in Ser392 p53 protein levels upon treatment with 2-AP, a considerable increase of ITGA-1 mRNA levels and proteins (Fig. 5D and E) was seen. In addition, ChIP assay in 38HK cells treated or not-treated with 2-AP demonstrated that the recruitment of both p53 and DNMT1 cellular proteins was affected by chemical inhibition of PKR phosphorylation (Fig. 5F and G).

To corroborate these data that indicate a cross-talk between p53 and PKR in 38HK, we performed reciprocal immunoprecipitation using PKR or p53 antibodies to assess the possible interaction between the two cellular proteins. Fig. 6A shows that PKR/p53 complex was immunoprecipitated by both antibodies. Importantly, the Ser392 p53 form was found associated with PKR (Fig. 6B). 2AP treatment resulted in a strong decrease on Ser392 phosphorylation of the PKR-associated p53 form (Fig. 6B). No significant changes were observed in the total form of p53 co-precipitated with PKR, suggesting that the PKR-mediated p53 phosphorylation does not affect the interaction between the two proteins (Fig. 6B). ChIP/re-ChIP assay using an antibody specific for the T446-phosphorylated PKR form, showed that the p53/p446PKR complex is able to bind the p53 RE2 of the ITGA-1 promoter (Fig. 6C). To further characterize the p53/PKR complex, we first fractionated the nuclear extracts of 38HK exposed or not exposed to 2AP by sucrose density gradient ultracentrifugation. Subsequently, p53 complex was immunoprecipitated in each sucrose gradient fraction (Fig. 6D). A trimeric complex containing p53/PKR/DNMT1 was found in some fractions of the sucrose gradient (Fig. 6D). 2AP treatment, in agreement with the data shown in Fig. 6B, did not influence the p53/PKR interaction, while DNMT1 is lost from the complex.

Together, these findings provide evidence that the full-length p53 form phosphorylated at S392 by PKR interacts with DNMT1 and inhibits ITGA-1 expression.

Ectopic overexpression of ITGA-1 induces cell death and inhibits cell proliferation

Interestingly, it was previously shown that ITGA-1 is implicated in a negative regulation of the epidermal growth factor receptor (EGFR) signalling and cellular proliferation [31]. Moreover, ITGA-1 down-regulation has been associated to poor patient outcome and drug resistance in ovarian cancer [32]. Next, to understand the biological significance of ITGA-1 down-regulation in 38HK, we investigated the impact of ITGA-1 overexpression in 38 HK proliferation using a colony formation assay. Cells were transfected with a construct expressing ITGA-1 and the zeocin-resistant gene and cultured under antibiotic selection. We observed a significant decrease in colony formation in 38HK expressing ectopic levels of ITGA-1 in comparison to the mock cells (Fig. 7A). In addition, analysis of the cell cycle profile by flow cytometry showed that ITGA-1 overexpression significantly increased the sub-G0 cell population, which is a sign of cellular death (Fig. 7B).

It has been previously reported that ITGA-1 negatively regulates EGFR signalling by promoting EGFR de-phosphorylation, with consequent inhibition of cellular proliferation [31]. Therefore, we

evaluated the status of EGFR signalling in 38HK after ITGA-1 ectopic expression by determining the levels of cyclin D1, which is positively regulated by activation of EGFR signalling [35,36]. In accordance with the inhibition of cellular proliferation, cyclin D1 levels decreased upon ectopic expression of ITGA-1 (Fig. 7C). Surprisingly, we also observed a reduction in the EGFR protein levels upon ITGA-1 over-expression. However, no significant changes were observed in EGFR mRNA levels upon ITGA-1-overexpression in HK38 and mock cells, suggesting that EGFR destabilization in presence of ITGA-1 is mediated post-translationally (Fig. 7D).

These findings indicate that HPV38 E6 and E7 inhibit the expression of ITGA-1 to promote cellular proliferation, which in part appears to be mediated by EGFR signalling.

HPV38 E6 and E7 expression in the skin of transgenic mice inhibits the ITGA-1 transcription

To corroborate our findings in *in vitro* experimental models, we interrogated whether HPV38 E6 and E7 has the ability to alter ITGA-1 expression in the skin keratinocytes in mice. We previously developed a transgenic (Tg) mouse model that expresses HPV38 E6 and E7 in the keratinocytes of the skin basal layer under the control of keratin 14 promoter (K14) [38]. After isolation of skin keratinocytes from WT and HPV38 E6/E7 transgenic mice, the ITGA-1 mRNA level was determined by quantitative RT-PCR. Fig. 8A showed that the viral proteins inhibit ITGA-1 expression in skin keratinocytes, confirming the *in vitro* findings.

We have recently shown that the HPV38 E6/E7 transgenic mice are highly susceptible to UV-induced DNA mutations and skin cancer development in comparison to WT animals [39]. By whole exome sequencing, we observed that these animals, upon long-term UV exposure, accumulate mutations in crucial cancer-linked genes, including p53 [39]. Therefore, we next determined whether p53 mutations in the core DNA binding domain detected in 3 different cSCC may result in loss of the inhibition of ITGA-1 expression. Quantitative RT-PCR experiments showed that p53 mutations correlated with an increase of ITGA-1 expression in at least two malignant lesions (Fig. 8A). However, we determined that in 2 cSCC, ITGA-1 gene contains deleterious non-synonymous mutations (Fig. 8B). Although the third cSCC expresses high levels of WT ITGA-1, it contains a non-synonymous, but not deleterious, mutation in a region of EGFR gene encoding the tyrosine kinase domain (Fig.8A and 8B). Mutations in this EGFR domain have been identified in human cancer and result in activation of EGFR signaling [40-42].

In summary, these results confirm the ability of HPV38 to inhibit ITGA-1 expression and highlight the importance of ITGA-1 inactivation in UV-induced cSCC development.

DISCUSSIONS

Oncogenic virus-mediated cellular transformation is intimately linked to the inactivation of p53. Indeed, this tumour suppressor was discovered due to its interaction with simian virus 40 (SV40) large T antigen. In addition, it is now well demonstrated that the E6 oncoprotein from the mucosal HR HPV types interacts with p53, promoting its degradation via the proteasomal pathways [43]. In non-virus related cancers, p53 is inactivated by DNA mutation, frequently occurring in the DNA binding motif. As a consequence, mutated p53 loses its normal transcriptional functions as tumour suppressor [44]. Importantly, a vast number of studies showed that p53 mutations, in addition to the disruption of its tumour suppressor function, can also confer oncogenic gain-of-function (GOF) activities [45,46]. Several findings support the model that GOF p53 mutations induce conformational changes allowing mutated p53 to interact with other cellular proteins, including products of tumour suppressor genes or oncogenes as well as specific promoter responsive elements [44]. Cellular response to a broad spectrum of stresses leads to post-translational modification of WT p53 that can be phosphorylated, acetylated, and ubiquitinated at specific serine, threonine, and lysine residues respectively [47,48]. Similarly, also the mutated p53 forms are post-translationally modified at specific residues, with consequent acquisition of more aggressive oncogenic functions. For instance, it has been shown that S392 is one of the most frequently phosphorylated residues in the p53 mutated forms [49,50].

Although p53 gene is highly mutated, approximately 50% of human cancers retain the WT p53 gene and its tumour suppressor functions can be altered by additional mechanisms, e.g. by overexpression of truncated N-terminus isoforms of p53 and p73 that act as dominant negative mutants of p53 [51]. Other plausible models of alteration of WT p53 tumour suppression functions can rely on specific patterns of post-translational modifications and interactions with cellular proteins. In this study, we describe that expression of HPV38 E6 and E7 in HK promotes the formation of a transcriptionally repressive p53/DNMT1 complex on ITGA-1 promoter. Inhibition of ITGA-1 results in activation of EGFR signaling and cellular proliferation. Paradoxically, ectopic levels of WT p53 in 38HK, further repressed the ITGA-1 expression and increased cellular proliferation. Thus, in HPV38-transformed HK, WT p53 acquires oncogenic properties. This conclusion is further corroborated, by the fact that HK38 are addicted to WT p53 for their cellular proliferation. Indeed, deletion of p53 gene by CRISPR/Cas strongly inhibited cellular growth. Interestingly, a previous study has described the formation of a p53/DNMT1 complex with transcriptional repressive function in non-virus related experimental model [52]. Therefore, it is plausible to hypothesize that HPV38 oncoproteins exploit cellular mechanisms that can be generated in other contexts. Our data shows that p53 phosphorylation at S392 by PKR is essential for the interaction of p53 with DNMT1. Although PKR has been initially considered to be a tumour suppressor, it is now well demonstrated that it also exert oncogenic functions, being overexpressed and activated in many types of cancers, including several hemapoietic malignancies [53]. Based on these findings, we could hypothesize that HPV38 E6 and E7 generate a specific scenario in the infected cells, in which PKR acts as an oncoprotein.

Our experiments in *in vitro* and *in vivo* experimental models also support the key role of ITGA-1 inactivation in cellular transformation. In agreement with these findings, many independent studies

have shown that alteration of integrin network is a frequent event in the development of several types of cancers, including cSCC [23].

Our previous studies in Tg mouse models demonstrated that expression of E6 and E7 oncoproteins in the skin of the animals strongly increases the susceptibility to UV-induced skin carcinogenesis [38]. The viral oncoproteins appear to be necessary only at an early stage of carcinogenesis and, after accumulation of a large number of UV-induced mutations, they are dispensable for the maintenance of the malignant phenotype [39]. Findings presented in this article further support the concept that beta HPV types act with a "hit-and-run" mechanism in promoting cSCC development. Our data show that HPV38 E6 and E7 inhibit the ITGA-1 expression in mouse skin, but upon UV irradiation and accumulation of DNA mutations, ITGA-1 mRNA levels are elevated and its gene is inactivated by deleterious mutations.

In conclusion, here we described a novel virus-mediated mechanism that converts WT p53 in an oncoprotein. This WT p53 form acquires the properties to interact with PKR and DNMT1 and to repress cellular gene expression. It will be important to evaluate whether similar mechanisms occur in cancers cells of different origin, offering the possibility to develop novel anti-cancer therapeutic strategies.

METHODS

Cell cultures and treatments

The experiments were carried out in HK isolated from neonatal foreskin and in human keratinocyte cell line expressing the human telomerase reverse transcriptase gene (hTERT), in order to prolong the life span of the cells. HK stably expressing HPV38 E6 or E7 as well as p53HA-Tag 38HK were generated by retroviral transduction [20]. 38HK, p53HA-Tag 38HK and hTERT cell line were cultured together with NIH 3T3 feeder layers in FAD medium containing: 3 parts Ham's F12, 1 part DMEM, 2.5% fetal calf serum, insulin (5 μg/ml), epidermal growth factor (10 ng/ml), cholera toxin (8.4 ng/ml), adenine (24 μg/ml), hydrocortisone (0.4 μg/ml) and 1% of pen/strep preparation. Feeder layers were prepared by treating NIH 3T3 with mitomycin for 2h. NIH 3T3 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum and 1% pen/strep preparation.

Transient transfection experiments were performed using Lipofectamine 2000 transfection reagent (Invitrogen) or TransIT®-Keratinocytes transfection reagents (Mirus) according to manufacture protocols.

Cells were incubated for 6 hours in media containing Cyclic Pifithrin-α hydrobromide (PFT) at 20 μM concentrations (Sigma); others were incubated for 24 hours in media containing 30 μM 5-aza-2'-deoxycytidine (Sigma).

2-Aminopurin (Sigma) was prepared in phosphate-buffered saline:glacial acetic acid (200:1). Cells were treated for 4 hours at 10mM final concentration; phosphate-buffered saline:glacial acetic acid (200:1) was used as a mock control.

For fluorescence-activated cell sorter (FACS) staining, cells were collected, washed twice in PBS, and fixed in 70% of ethanol for 30 minutes in ice. Samples were stained with propidium iodide (PI)

at a final concentration of 5µg/ml. Subsequently, cells were analyzed by FACS CANTO (Becton Dickinson).

For the colony formation assay, cells were transfected using TransIT®-Keratinocyte Transfection Reagent (Mirus) according to the manufacturer protocols. Cells were transfected with pcDNA 3.1/Zeo (mock) or pcDNA 3.1/Zeo expressing ITGA-1 (2.5 µg) (a gift from A. Pozzi, Vanderbilt University). After 24 hours the cells were splitted for selection in zeocin (Invivogen). They were diluted 10, 100, 1000 times and were allowed to grow for 4 days. After this period the colonies were fixed and stained as described in [54].

For determination of cell growth curves, 1.5-2.5x10^5 cells were seeded into 6 well plates. After 24, 48 and 72 hours cells were collected and stained with trypan blue (1:1) (Bio rad). Samples were counted in duplicate with TC20 automated cell counter (Bio rad).

Microscope images were taken under light microscope 48 hours post-transfection.

Cell viability was determined by MTS assay. Briefly, 20µL of CellTiter 96® AQueous One Solution Cell Proliferation Assay (Promega) was added to 1.5x10^4 cells in 96 well plates and incubated for 2 hours at 37°C for 24, 48 and 72 hours. Absorbance at 490nm was read by Multiskan GO (Thermo Scientific) in duplicate. Blank absorbance was subtracted.

Transgenic mice

DNA and mRNA were extracted from normal and cancer mouse tissues as previously described [38]. A detailed description of the HPV38 E6/E7 transgenic mouse line can be found here https://mito.dkfz.de/mito/Animal%20line/10954. A detailed description of the UV-induced skin carcinogenesis protocol can be found here https://mito.dkfz.de/mito/Tumor%20model/10474.

Ethics statement

The animal facility of the German Cancer Research Center has been officially approved by responsible authority (Regional Council of Karlsruhe, Schlossplatz 4–6, 76131 Karlsruhe, Germany), official approval file number 35–9185.64. Housing conditions are thus in accordance with the German Animal Welfare Act (TierSchG) and EU Directive 425 2010/63/EU. Regular inspections of the facility are conducted by the Veterinary Authority of Heidelberg (Bergheimer Str. 69, 69115 Heidelberg, Germany). All experiments were in accordance with the institutional guidelines (designated veterinarian according to article 25 of Directive 2010/63/EU and Animal-Welfare Body according to article 27 of Directive 2010/63/EU) and were officially approved by Regional Council of Karlsruhe (File No 35–9185.81/G-64/13 and 35–9185.81/G-200/15).

Gene silencing

Gene silencing of DNMT1 was achieved using synthetic siRNA (Table 1). siRNA or scrambled RNA at a concentration of 250 nM was transfected using TransIT®-Keratinocytes transfection reagents (Mirus) according to the standard protocol. Cells were collected after 72h.

Plasmids for CRISPR/Cas9 were obtained from the Addgene plasmid repository. All single-guide RNAs were designed by Thermo Fisher Scientific. The target sequence information is shown in Table S1. The CRISPR/Cas9 vectors were generated according to manufacturer protocols and then transiently transfected into keratinocytes. Purification of the cells carrying the CRISPR/Cas9 vectors was performed 48 hours after transfection according to the manufacturer's protocol

(GeneArt CRISPR Nuclease Vector Kit; Life Technologies).

Reverse transcription and quantitative PCR

For the experiments in vitro models, total RNA was extracted using the NucleoSpin RNA II Kit (MACHEREY NAGEL). The RNA obtained was reverse-transcribed to cDNA using the RevertAid H minus First strand cDNA Kit (Life Technologies) according to the manufacturer's protocols. Real-time quantitative PCR (qPCR) was performed using the MESA GREEN qPCR MasterMix Plus for SYBR Assay (Eurogentec) with the primers listed on Table 2.

For the experiments in mice, total RNA was isolated from dorsal skin of WT (n = 4), K14 HPV38 E6/E7 Tg animals (n = 3), histologically confirmed pre-malignant (AK) and SCC from three independent mice. cDNA was synthesized from 1 μg of total RNA using M-MLV reverse transcriptase (Invitrogen, Darmstadt, Germany), and a mix of random hexamers were used as primers. Quantitative reverse transcription PCR (RT-qPCR) was performed using LightCycler 480 SYBR Green I Master (Roche) with specific mouse primers (Table 2).

TaqMan assay was performed with ITGA-1 TaqMan gene expression assay probe (Hs00235006_m1; Life technologies) following manufacturer's instructions. Reactions were run in triplicate and expression was normalized to GAPDH (Hs99999905_m1; ThermoFisher).

Immunoblotting

Cells were lysed using IP buffer (TrisHCl 20 mM pH 7.5, NaCl 200 mM, EDTA 1 mM, NP-40 0.5%) supplemented with Complete Protease Inhibitor mixture (Roche). Samples were resolved by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) and transferred to polyvinylidene difluoride (PVDF) membranes (PERKIN ELMER). Membranes were blocked in 5% non-fat milk and incubated overnight at 4°C with the appropriate primary antibody. Membranes were probed with the following primary antibodies: β-actin (clone C4; MP Biomedicals), p53 (DO-1) (sc-126; Santa Cruz Biotechnology), DNMT1 (clone 60B1220.1; MAB0079; Abnova), PKR (3072; Cell Signaling Technology), phosphorylated PKR Thr 446 (PA5-37704; Thermo Fischer Scientific), EGF receptor (4267; Cell Signaling Technology), Cyclin D1 (2978; Cell Signaling Technology), HA-Tag (3F10; Roche) and ITGA-1 (106267; Abcam). Images were produced using the ChemiDoc XRS imaging system (Bio-Rad).

Chromatin immunoprecipitation

ChIP was performed using the Shearing ChIP and OneDay ChIP kits (Diagenode) according to the manufacturer's instructions. Briefly, cells were sonicated to obtain DNA fragments of 200–500 bp. Sheared chromatin was immunoprecipitated with isotype control IgG or the indicated antibodies: p53 (DO-1) (sc-126; Santa Cruz Biotechnology), DNMT1 (clone 60B1220; Abnova) and phospho PKR T446 (ab32036). For ChIP/Re-ChIP experiments, bead-bound protein-DNA complexes obtained after the first ChIP were incubated with 10 mM Reverse dithiothreitol (DTT) for 30 min at 37°C with shaking at 400 rpm. Supernatant was collected after centrifugation at 12,000g for 1 min. Pelleted beads were incubated again with 10 mM DTT for 20 min at 37ºC and centrifuged at 12,000g for 1 min. Ten percent of the combined supernatants was kept as the input for the second

ChIP, which was performed according to the OneDay ChIP kit (Diagenode) manufacturer's protocol.
For histones chromatin immunoprecipitation, chromatin shearing kit Low SDS and Auto iDeal ChIP-seq kit for Histones (Diagenode) were used together with SX-8G IP-Star Compact Automated system and histone H3K9ac antibody (Euromedex).
The eluted DNA was used as a template for qPCR (Table 2).

Oligonucleotide pulldown assay

Cells were lysed and sonicated in HKMG buffer (10 mM HEPES [pH 7.9], 100 mM KCl, 5 mM MgCl2, 10% glycerol, 1 mM DTT, 0.5% Nonidet P-40) containing protease and phosphatase inhibitors. After centrifugation at 12,000g for 10 min, protein extracts were precleared with streptavidin-agarose beads. The ITGA-1 promoter was used as a template to amplify the p53RE. PCR amplification was performed using a biotinylated forward primer and a nonbiotinylated reverse primer listed on table 2. Amplicons were extracted from agarose gel by using a MinElute gel extraction kit (Qiagen) and quantified. Then, 2 mg of cellular protein extracts were incubated with 1 µg of biotin-ITGA-1 promoter probes and 10 µg of poly(dI-dC)·poly(dI-dC) for 16 h at 4°C. DNA-bound proteins were collected with streptavidin-agarose beads for 1 h and washed five times with HKMG buffer. DNA-bound proteins were then analyzed by IB.

Electro mobility shift assay

Nuclear extracts from cells were prepared as previously described [55]**.** Briefly, 3×10^6 cells were collected, washed in PBS 1×, and resuspended in hypotonic buffer A (10 mM HEPES, pH 7.9; 1.5 mM MgCl2; 10 mM KCl; 0.5 mM DTT; 0.2 mM PMSF). The cell suspensions were then incubated on ice and homogenized by 15 passages through a 25-gauge needle. Cytoplasm fractions were collected by centrifugation at 12,000 rpm for 1 min at 4°C. Nuclei were washed in buffer A, centrifuged, and dissolved in hypertonic buffer B (20 mM HEPES, pH 7.9; 25% glycerol; 0.42 M NaCl; 1.5 mM $MgCl_2$; 0.2 mM EDTA; 0.5 mM DTT). The nuclear extracts were collected by centrifugation at 12,000 rpm for 2 min at 4°C. Protein concentration was estimated using an assay kit (Bio-Rad Laboratoires, Richmond, CA). Five µg of the extracts were incubated with 0.5 pmoles of biotin-labelled DNA probe listed on Table 2 and poly (dI-dC) in binding buffer (10 mM Tris, 100 mM NaCl, 1 mM EDTA, 1 mM DTT, 5% glycerol, pH 7.5) in a final volume of 15 µl. Binding reactions were incubated for 20 min at room temperature. The dye solution was then added and samples were loaded into a 5% polyacrylamide gel in 0.5X TBE buffer for running. The gels were then transferred to BM-Nylon (+) blotting membrane (Roche) and developed by using the "Chemiluminescent Nucleic Acid Detection Module" provided in the nonradioactive "LightShift chemiluminescent EMSA Kit" (Thermo Scientific). Specificity of the protein–DNA complex was verified by a competition experiment where the nuclear extracts were incubated with an excess of unlabelled DNA.

Sucrose gradient protein complex isolation

Sucrose density gradients were prepared from 10% to 50% sucrose (10mM NaCl, 2mM TrisHCl and 0.5mM $MgCl_2$). Cells treated with 2-Aminopurin or Acetic acid (200:1) 10mM for 4 hours were processed for nuclear extraction as previously described and a total of 1mg of protein was

added to the sucrose gradient. 20ug of nuclear extract was kept as input control. After 16 hours centrifugation at 35300rpm and 4°C, 21 fractions of a total volume of 500uL were collected.

Protein immunoprecipitation

Immunoprecipitations were performed as previously described [56]. Briefly, 5 µL of the indicated antibodies (p53DO-1, PKR or IgG) pre-adsorbed on 50uL of protein A/G plus agarose beads (SC-2003) suspended in PBS-1%Np40 for 2h at 4°C. Suspended beads were also incubated with total protein lysate or sucrose protein fractions 5 to 21 for pre-clearing in rotation at 4°C for 2 hours. After overnight incubation with the extracts, beads were resolved by SDS–PAGE and transferred to PVDF membranes (PERKIN ELMER). Immunoblotting was performed with DNMT1, PKR and p53 antibodies and immunoreactivity was revealed by means of secondary antibodies specific for IP (Abcam). Immunoreactive proteins were visualized by means of the ECL method (Millipore).

Exome analysis

The exome analysis was performed as described in [39]. SIFT missense predictions for genomes annotator was used to predicts whether the amino acid substitution affects protein function [57].

Statistical analysis

Statistical significance was determined using the Student t-test with Prism7 (Graphpad). The levels of statistical significance for each experiment (*p<0.05, **p<0.01,***p<0.001,****p<0.0001 or not significant) are indicated in the corresponding figures. The error bars in the graphs represent the standard deviation.

Acknowledgements

We are grateful to all members of our laboratories for their cooperation. We are also grateful to Nicole Suty for her help with preparation. Raw data of all experiments were provided to the editor for review purposes. The study was supported by a grant from Fondation ARC pour la recherche sur le cancer (no. PJA 20151203192) (https://www.fondation-arc.org/espace-chercheur) and the Institut National de la Santé et de la Recherche Médicale (no. ENV201610) (https://www.eva2.inserm.fr/EVA/jsp/AppelsOffres/CANCER/) to MT.

Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

Competing interests

The authors declare no competing interests.

Author contributions

M.T and R.A. conceived the project. M.T. designed, and interpreted the experiments and wrote the first draft of the manuscript. M.R. and A.V. contributed to the design, execution, and analyses of

the experiments, in addition to revising the initial submission. M.S. and R.A. conducted preliminary experiments. D.V. conducted the *in vivo* experiments. M.G.C., L.P., G.M. and C.S. performed some experiments. P.G. helped to perform FACs analyses. V.T. helped for sucrose gradient experiment. A.R. performed in silico analyses. C.I. helped to develop protocols and contributed with reagents.

References

1 Rollison, D. E., Viarisio, D., Amorrortu, R. P., Gheit, T. & Tommasino, M. An Emerging Issue in Oncogenic Virology: the Role of Beta Human Papillomavirus Types in the Development of Cutaneous Squamous Cell Carcinoma. *J Virol* **93**, doi:10.1128/JVI.01003-18 (2019).

2 Howley, P. M. & Pfister, H. J. Beta genus papillomaviruses and skin cancer. *Virology* **479-480**, 290-296, doi:10.1016/j.virol.2015.02.004 (2015).

3 Hasche, D., Vinzon, S. E. & Rosl, F. Cutaneous Papillomaviruses and Non-melanoma Skin Cancer: Causal Agents or Innocent Bystanders? *Frontiers in microbiology* **9**, 874, doi:10.3389/fmicb.2018.00874 (2018).

4 Harwood, C. A., Toland, A. E., Proby, C. M., Euvrard, S., Hofbauer, G. F. L. *et al.* The pathogenesis of cutaneous squamous cell carcinoma in organ transplant recipients. *Br J Dermatol* **177**, 1217-1224, doi:10.1111/bjd.15956 (2017).

5 Tommasino, M. The biology of beta human papillomaviruses. *Virus Res* **231**, 128-138, doi:10.1016/j.virusres.2016.11.013 (2017).

6 Farzan, S. F., Waterboer, T., Gui, J., Nelson, H. H., Li, Z. *et al.* Cutaneous alpha, beta and gamma human papillomaviruses in relation to squamous cell carcinoma of the skin: a population-based study. *Int J Cancer* **133**, 1713-1720, doi:10.1002/ijc.28176 (2013).

7 Iannacone, M. R., Gheit, T., Pfister, H., Giuliano, A. R., Messina, J. L. *et al.* Case-control study of genus-beta human papillomaviruses in plucked eyebrow hairs and cutaneous squamous cell carcinoma. *Int J Cancer* **134**, 2231-2244, doi:10.1002/ijc.28552 (2014).

8 Waterboer, T., Abeni, D., Sampogna, F., Rother, A., Masini, C. *et al.* Serological association of beta and gamma human papillomaviruses with squamous cell carcinoma of the skin. *Br J Dermatol* **159**, 457-459, doi:10.1111/j.1365-2133.2008.08621.x (2008).

9 Karagas, M. R., Waterboer, T., Li, Z., Nelson, H. H., Michael, K. M. *et al.* Genus beta human papillomaviruses and incidence of basal cell and squamous cell carcinomas of skin: population based case-control study. *BMJ (Clinical research ed.)* **341**, c2986, doi:10.1136/bmj.c2986 (2010).

10 Bouwes Bavinck, J. N., Feltkamp, M. C. W., Green, A. C., Fiocco, M., Euvrard, S. *et al.* Human papillomavirus and posttransplantation cutaneous squamous cell carcinoma: A multicenter, prospective cohort study. *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons* **18**, 1220-1230, doi:10.1111/ajt.14537 (2018).

11 Chahoud, J., Semaan, A., Chen, Y., Cao, M., Rieber, A. G. *et al.* Association Between beta-Genus Human Papillomavirus and Cutaneous Squamous Cell Carcinoma in

Immunocompetent Individuals-A Meta-analysis. *JAMA dermatology*, doi:10.1001/jamadermatol.2015.4530 (2015).

12 Jackson, S., Harwood, C., Thomas, M., Banks, L. & Storey, A. Role of Bak in UV-induced apoptosis in skin cancer and abrogation by HPV E6 proteins. *Genes & development* **14**, 3065-3073 (2000).

13 Muench, P., Probst, S., Schuetz, J., Leiprecht, N., Busch, M. *et al.* Cutaneous papillomavirus E6 proteins must interact with p300 and block p53-mediated apoptosis for cellular immortalization and tumorigenesis. *Cancer Res* **70**, 6913-6924, doi:10.1158/0008-5472.CAN-10-1307 (2010).

14 Wallace, N. A., Robinson, K., Howie, H. L. & Galloway, D. A. beta-HPV 5 and 8 E6 disrupt homology dependent double strand break repair by attenuating BRCA1 and BRCA2 expression and foci formation. *PLoS Pathog* **11**, e1004687, doi:10.1371/journal.ppat.1004687 (2015).

15 Wallace, N. A., Robinson, K. & Galloway, D. A. Beta human papillomavirus E6 expression inhibits stabilization of p53 and increases tolerance of genomic instability. *J Virol* **88**, 6112-6127, doi:10.1128/jvi.03808-13 (2014).

16 Wallace, N. A., Robinson, K., Howie, H. L. & Galloway, D. A. HPV 5 and 8 E6 abrogate ATR activity resulting in increased persistence of UVB induced DNA damage. *PLoS Pathog* **8**, e1002807, doi:10.1371/journal.ppat.1002807 (2012).

17 Holloway, A. & Storey, A. A conserved C-terminal sequence of high-risk cutaneous beta-human papillomavirus E6 proteins alters localization and signalling of beta1-integrin to promote cell migration. *J Gen Virol* **95**, 123-134, doi:10.1099/vir.0.057695-0 (2014).

18 Heuser, S., Hufbauer, M., Steiger, J., Marshall, J., Sterner-Kock, A. *et al.* The fibronectin/alpha3beta1 integrin axis serves as molecular basis for keratinocyte invasion induced by betaHPV. *Oncogene* **35**, 4529-4539, doi:10.1038/onc.2015.512 (2016).

19 Meyers, J. M., Uberoi, A., Grace, M., Lambert, P. F. & Munger, K. Cutaneous HPV8 and MmuPV1 E6 Proteins Target the NOTCH and TGF-beta Tumor Suppressors to Inhibit Differentiation and Sustain Keratinocyte Proliferation. *PLoS Pathog* **13**, e1006171, doi:10.1371/journal.ppat.1006171 (2017).

20 Caldeira, S., Zehbe, I., Accardi, R., Malanchi, I., Dong, W. *et al.* The E6 and E7 proteins of the cutaneous human papillomavirus type 38 display transforming properties. *J Virol* **77**, 2195-2206 (2003).

21 Accardi, R., Scalise, M., Gheit, T., Hussain, I., Yue, J. *et al.* IkappaB kinase beta promotes cell survival by antagonizing p53 functions through DeltaNp73alpha phosphorylation and stabilization. *Mol Cell Biol* **31**, 2210-2226, doi:10.1128/MCB.00964-10 (2011).

22 Accardi, R., Dong, W., Smet, A., Cui, R., Hautefeuille, A. *et al.* Skin human papillomavirus type 38 alters p53 functions by accumulation of deltaNp73. *EMBO Rep* **7**, 334-340, doi:10.1038/sj.embor.7400615 (2006).

23 Cooper, J. & Giancotti, F. G. Integrin Signaling in Cancer: Mechanotransduction, Stemness, Epithelial Plasticity, and Therapeutic Resistance. *Cancer cell* **35**, 347-367, doi:10.1016/j.ccell.2019.01.007 (2019).

24 Barczyk, M., Carracedo, S. & Gullberg, D. Integrins. *Cell Tissue Res* **339**, 269-280, doi:10.1007/s00441-009-0834-6 (2010).
25 Hodivala, K. J., Pei, X. F., Liu, Q. Y., Jones, P. H., Rytina, E. R. *et al.* Integrin expression and function in HPV 16-immortalised human keratinocytes in the presence or absence of v-Ha-ras. Comparison with cervical intraepithelial neoplasia. *Oncogene* **9**, 943-948 (1994).
26 Woappi, Y., Hosseinipour, M., Creek, K. E. & Pirisi, L. Stem Cell Properties of Normal Human Keratinocytes Determine Transformation Responses to Human Papillomavirus 16 DNA. *J Virol* **92**, doi:10.1128/jvi.00331-18 (2018).
27 Cooper, B., Brimer, N. & Vande Pol, S. B. Human papillomavirus E6 regulates the cytoskeleton dynamics of keratinocytes through targeted degradation of p53. *J Virol* **81**, 12675-12679, doi:10.1128/jvi.01083-07 (2007).
28 Oldak, M., Smola, H., Aumailley, M., Rivero, F., Pfister, H. *et al.* The human papillomavirus type 8 E2 protein suppresses beta4-integrin expression in primary human keratinocytes. *J Virol* **78**, 10738-10746, doi:10.1128/jvi.78.19.10738-10746.2004 (2004).
29 Holloway, A., Simmonds, M., Azad, A., Fox, J. L. & Storey, A. Resistance to UV-induced apoptosis by beta-HPV5 E6 involves targeting of activated BAK for proteolysis by recruitment of the HERC1 ubiquitin ligase. *Int J Cancer* **136**, 2831-2843, doi:10.1002/ijc.29350 (2015).
30 Oldak, M., Maksym, R. B., Sperling, T., Yaniv, M., Smola, H. *et al.* Human papillomavirus type 8 E2 protein unravels JunB/Fra-1 as an activator of the beta4-integrin gene in human keratinocytes. *J Virol* **84**, 1376-1386, doi:10.1128/jvi.01220-09 (2010).
31 Mattila, E., Pellinen, T., Nevo, J., Vuoriluoto, K., Arjonen, A. *et al.* Negative regulation of EGFR signalling through integrin-alpha1beta1-mediated activation of protein tyrosine phosphatase TCPTP. *Nature cell biology* **7**, 78-85, doi:10.1038/ncb1209 (2005).
32 Wei, L., Yin, F., Zhan, W. & Li, L. ITGA1 and cell adhesion-mediated drug resistance in ovarian cancer. *Int J Clin Exp Pathol* **10**, 5522-5529 (2017).
33 Cuddihy, A. R., Wong, A. H., Tam, N. W., Li, S. & Koromilas, A. E. The double-stranded RNA activated protein kinase PKR physically associates with the tumor suppressor p53 protein and phosphorylates human p53 on serine 392 in vitro. *Oncogene* **18**, 2690-2702, doi:10.1038/sj.onc.1202620 (1999).
34 Yoon, C. H., Lee, E. S., Lim, D. S. & Bae, Y. S. PKR, a p53 target gene, plays a crucial role in the tumor-suppressor function of p53. *Proc Natl Acad Sci U S A* **106**, 7852-7857, doi:10.1073/pnas.0812148106 (2009).
35 Lenferink, A. E., Busse, D., Flanagan, W. M., Yakes, F. M. & Arteaga, C. L. ErbB2/neu kinase modulates cellular p27(Kip1) and cyclin D1 through multiple signaling pathways. *Cancer Res* **61**, 6583-6591 (2001).
36 Wee, P. & Wang, Z. Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers* **9**, doi:10.3390/cancers9050052 (2017).
37 Vieler, M. & Sanyal, S. p53 Isoforms and Their Implications in Cancer. *Cancers* **10**, doi:10.3390/cancers10090288 (2018).
38 Viarisio, D., Mueller-Decker, K., Kloz, U., Aengeneyndt, B., Kopp-Schneider, A. *et al.* E6 and E7 from beta HPV38 cooperate with ultraviolet light in the development of actinic

keratosis-like lesions and squamous cell carcinoma in mice. *PLoS Pathog* **7**, e1002125, doi:10.1371/journal.ppat.1002125 (2011).

39 Viarisio, D., Muller-Decker, K., Accardi, R., Robitaille, A., Durst, M. *et al.* Beta HPV38 oncoproteins act with a hit-and-run mechanism in ultraviolet radiation-induced skin carcinogenesis in mice. *PLoS Pathog* **14**, e1006783, doi:10.1371/journal.ppat.1006783 (2018).

40 Lopes, G. L., Vattimo, E. F. & Castro Junior, G. Identifying activating mutations in the EGFR gene: prognostic and therapeutic implications in non-small cell lung cancer. *Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia* **41**, 365-375, doi:10.1590/s1806-37132015000004531 (2015).

41 Gazdar, A. F. Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors. *Oncogene* **28 Suppl 1**, S24-31, doi:10.1038/onc.2009.198 (2009).

42 Uribe, P. & Gonzalez, S. Epidermal growth factor receptor (EGFR) and squamous cell carcinoma of the skin: molecular bases for EGFR-targeted therapy. *Pathology, research and practice* **207**, 337-342, doi:10.1016/j.prp.2011.03.002 (2011).

43 Tommasino, M. The human papillomavirus family and its role in carcinogenesis. *Semin Cancer Biol* **26**, 13-21, doi:10.1016/j.semcancer.2013.11.002 (2014).

44 Yamamoto, S. & Iwakuma, T. Regulators of Oncogenic Mutant TP53 Gain of Function. *Cancers* **11**, doi:10.3390/cancers11010004 (2018).

45 Weisz, L., Oren, M. & Rotter, V. Transcription regulation by mutant p53. *Oncogene* **26**, 2202-2211, doi:10.1038/sj.onc.1210294 (2007).

46 Kim, M. P., Zhang, Y. & Lozano, G. Mutant p53: Multiple Mechanisms Define Biologic Activity in Cancer. *Frontiers in oncology* **5**, 249, doi:10.3389/fonc.2015.00249 (2015).

47 Bode, A. M. & Dong, Z. Post-translational modification of p53 in tumorigenesis. *Nat Rev Cancer* **4**, 793-805, doi:10.1038/nrc1455 (2004).

48 Lane, D. & Levine, A. p53 Research: the past thirty years and the next thirty years. *Cold Spring Harbor perspectives in biology* **2**, a000893, doi:10.1101/cshperspect.a000893 (2010).

49 Ullrich, S. J., Sakaguchi, K., Lees-Miller, S. P., Fiscella, M., Mercer, W. E. *et al.* Phosphorylation at Ser-15 and Ser-392 in mutant p53 molecules from human tumors is altered compared to wild-type p53. *Proc Natl Acad Sci U S A* **90**, 5954-5958, doi:10.1073/pnas.90.13.5954 (1993).

50 Minamoto, T., Buschmann, T., Habelhah, H., Matusevich, E., Tahara, H. *et al.* Distinct pattern of p53 phosphorylation in human tumors. *Oncogene* **20**, 3341-3347, doi:10.1038/sj.onc.1204458 (2001).

51 De Laurenzi, V. & Melino, G. Evolution of functions within the p53/p63/p73 family. *Annals of the New York Academy of Sciences* **926**, 90-100, doi:10.1111/j.1749-6632.2000.tb05602.x (2000).

52 Esteve, P. O., Chin, H. G. & Pradhan, S. Human maintenance DNA (cytosine-5)-methyltransferase and p53 modulate expression of p53-repressed promoters. *Proc Natl Acad Sci U S A* **102**, 1000-1005, doi:10.1073/pnas.0407729102 (2005).

53 Watanabe, T., Imamura, T. & Hiasa, Y. Roles of protein kinase R in cancer: Potential as a therapeutic target. *Cancer science* **109**, 919-925, doi:10.1111/cas.13551 (2018).
54 Giarre, M., Caldeira, S., Malanchi, I., Ciccolini, F., Leao, M. J. *et al.* Induction of pRb degradation by the human papillomavirus type 16 E7 protein is essential to efficiently overcome p16INK4a-imposed G1 cell cycle Arrest. *J Virol* **75**, 4705-4712, doi:10.1128/JVI.75.10.4705-4712.2001 (2001).
55 Venuti, A., Musarra-Pizzo, M., Pennisi, R., Tankov, S., Medici, M. A. *et al.* HSV-1\EGFP stimulates miR-146a expression in a NF-kappaB-dependent manner in monocytic THP-1 cells. *Scientific reports* **9**, 5157, doi:10.1038/s41598-019-41530-5 (2019).
56 Venuti, A., Pastori, C., Pennisi, R., Riva, A., Sciortino, M. T. *et al.* Class B beta-arrestin2-dependent CCR5 signalosome retention with natural antibodies to CCR5. *Scientific reports* **6**, 39382, doi:10.1038/srep39382 (2016).
57 Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat Protoc* **11**, 1-9, doi:10.1038/nprot.2015.123 (2016).

Figures and figures legends

Figure 1. ITGA-1 expression is downregulated in HPV38 E6 E7 expressing cells.

(A) Human primary keratinocytes were transduced with pLXSN HPV38 E6/E7 or pLSXN. Total RNA levels were measured by RT-qPCR and normalized to GAPDH. Error bars represent standard deviations from three biological replicates of two different donors (n=6). ***, P<0.001 and **** p<0.0001. (B) Total RNA levels of HKs expressing or not HPV38 E6 and E7 were analyzed by TaqMan PCR. Commercial probes for ITGA-1 and GAPDH were used. Results were normalized to GAPDH. Data are the mean of three independent experiments for two different donors (n=6). ****, p<0.0001. (C) TaqMan assay was also performed as previously described in human primary keratinocytes previously retrovirally transduced with hTERT gene and expressing E6 and/or E7 from HPV 38 (n=3). Results were normalized to GAPDH. ****, p<0.0001.

Figure 2. p53 and DNMT1 form a complex that is recruited to ITGA-1 promoter.

(A) Schematic representation of ITGA-1 promoter sequence. p53 responsive elements were predicted by TFBind and JASPAR softwares.

(B) Electromobility Shift Assay performed with 38HK nuclear protein extracts and biotinylated probes containing p53REs WT or mutated sequences. Probes were incubated and cross-linked with protein extracts. Unlabeled WT or mutant p53RE2 probes were used as control. Representative images are shown as an example of two different experiments. (C) 38HK cells were cross-linked and chromatin was processed for ChIP using p53 antibody. Results were analyzed by qPCR with primers spanning p53RE1, p53RE2, p53RE3 or intergenic region of chromosome 22 (nc). Error bars represent standard deviation of three independent experiments performed in triplicate. **, p<0.01. (D) Cell lysate was incubated with biotinylated wild type probes containing p53 REs of the ITGA-1 promoter. Incubation without probe was used as control. DNA-associated proteins were recovered by precipitation with streptavidin beads and analyzed by immunoblotting (IB). A representative image is shown as an example of three independent experiments. (E) Chromatin from 38HK was processed for ChIP experiments using p53 or DNMT1 antibodies. Results were obtained by qPCR with primers spanning p53RE2 or the intergenic region of chromosome 22 (nc). Error bars indicate standard deviation from three independent experiments performed in duplicate. **, p<0.01, ***, p<0.001.

Figure 3. p53/DNMT1 interaction on ITGA-1 promoter.

(A) 38HK were cultured in medium containing Cyclic Pifithrin-α hydrobromide or DMSO as a control. Chromatin was processed for ChIP using p53 or DNMT1 antibodies. Results were obtained by qPCR using primers spanning p53RE2. Data are the mean of two independent experiments performed in triplicate. *, p<0.05, **, p<0.01. (B) Chromatin was processed for ChIP-reChIP assay in which p53-immunoprecipitated DNA was re-immunoprecipitated by DNMT1. Enrichment on p53RE2 or intergenic region of chromosome 22 (nc) were obtained by qPCR. Data are the mean of three independent experiments performed in triplicate. **, p<0.01. (C) 38HK cells were transfected with DNMT1 siRNA or siRNA control (Scramble). After 72 hours, ChIP assay was performed with p53 or DNMT1 antibodies. Results were obtained by qPCR using p53RE2 primers. Error bars represent standard deviation from three independent experiments. *,p<0.05, **,

p<0.01.

Figure 3. p53/DNMT1 complex inhibits ITGA-1 expression.
(A)Total RNA levels of 38HK treated with Cyclic Pifithrin-α hydrobromide or DMSO for 6 hours were analyzed by RT-qPCR and normalized to GAPDH. Histogram represents the mean of at least three independent experiments. ****, p<0.0001. (B and C) ITGA-1 and p53 mRNA and protein levels from 38HK expressing wild type p53 (Scramble) or with CRISPR/Cas9-mediated p53 deletion (CRISPRp53) were measured by RT-qPCR and IB. Data show the mean of four independent experiments. *, p<0.05. (D) 38HK cells were transfected with control siRNA (Scramble) or with DNMT1 siRNA (siDNMT1). After 72 hours, cells were collected for RNA extraction and RT-qPCR analysis (n=3). *, p<0.05, ***, p<0.001.(E) ITGA-1 expression was evaluated by RT-qPCR after 24 hours of 5-Aza-2'- deoxycytidine (Aza) or DMSO treatment at a final concentration of 30μM. Error bars represent standard deviation of three independent experiments. **, p<0.01. (F) H3K9ac at the ITGA-1 promoter was evaluated by ChIP assay upon 5-Aza-2'- deoxycytidine (Aza) or DMSO treatment as previously described (n=4). Results were obtained by qPCR using primers for p53RE2. *, p<0.05. (G) 38HK cells were transfected with sense (control) and antisense oligonucleotides against ΔNp73α. After 24 hours, cells were collected processed for RT-qPCR (left) or IB (right) (n=3). ns, not significant.

Figure 5. Full-length WT p53 displays oncogenic properties in 38HK
(A and B) 38HK N-HA-p53 or p53-C-HA cells were generated by retroviral transduction with WT-p53 tagged at N-terminus or C-terminus. As a control, 38HK cells were transduced with the corresponding empty plasmid. Protein extracts and total mRNA levels were processed for immunoblotting (A) and RT-qPCR analysis (B). A representative image is shown as an example of three independent experiments. ITGA-1 mRNA levels (B) was normalized to GAPDH. Error bars indicate standard deviation of three independent experiments. ***, p<0.001. (C and D) Cells expressing N-HA-p53 or p53-C-HA were seeded into 6 wells plates or 96 well. After 24, 48 and 72 hours cells were collected, stained with trypan blue and counted (C). Cells in 96 well were incubated with 20uL of MTS solution for 2 hours (D). Absorbance was obtained at 490nm wavelength. Results are the mean of three and two independent experiments respectively performed in duplicate. ***, p<0.001, ****, p<0.0001. (E and F) 38HK cells were seeded into 6 well plate and transfected with Scramble or CRISPRp53 plasmid for p53 knock-out. Immunoblot is an example of two independent experiments (E). After 24, 48 and 72 hours cells were collected, stained with trypan blue and counted (F). Data are the mean of two independent experiments performed in duplicate. *, p<0.05, **, p<0.01,***, p<0.001. (G) Cells treated with pifithrin for 24 hours were stained with trypan blue and counted. Histogram represents the mean of four independent experiments. *, p<0.05,**,p<0.01.

Figure 6. Full-length p53 phosphorylated at S392 plays a key role in ITGA-1 inhibition
(A) Protein extracts from 38HK were processed for oligonucleotide pulldown as previously described. The image is an example of three independent experiments.
(B) HK and 38HK were processed for protein extraction and IB. After p446PKR antibody

incubation, membrane was stripped and incubated with total PKR antibody.
(C) 38HK were treated with PKR inhibitor, 2-Aminopurine (2AP), or Acetic Acid: PBS solution (1:200), as a control, for 4 hours at 10mM final concentration. p-p53 Ser392 and p53 band intensities were quantified and normalized to total p53 (central panel) or β-Actin (right panel), respectively. Membranes were first incubated with p446PKR, then stripped and incubated with total PKR. Data are the mean of three independent experiments. *, $p<0.05$, **, $p<0.01$.
(D and E) 38HK were treated with 2AP) and ITGA-1 mRNA (D) and protein levels were determined by RT-PCR and IB respectively. The values in (D) are the mean of three independent experiments (**, $p<0.01$). The image in (E) is representative of three independent experiments. (F and G) ChIP assay using p53 or DNMT1 antibodies was performed in 38HK treated with or Acetic Acid: PBS solution (1:200) (CTR) or 2-Aminopurin (2AP). Results are the mean of two independent experiments performed in duplicate by qPCR using p53RE2 primers. *, $p<0.05$.

Figure 7 PKR inhibition reduces p53 phosphorylation and regulates ITGA-1 expression.
(A) Protein extracts from 38HK were processed for immunoprecipitation. Agarose beads were conjugated with total PKR antibody (top) or p53 antibody (bottom). Conjugated beads were incubated with protein lysate overnight. IgG was used as a control. Results were obtained by IB. Pictures were cropped due to the presence of irrelevant samples to this work. They are an example of two independent experiments. (B) Total protein extracts from cells treated with PKR inhibitor, 2-Aminopurin (2AP), or Acetic Acid: PBS solution (1:200), as a mock control (-) were obtained for PKR immunoprecipitation as previously described (top). Input was run in a different gel to improve picture quality. (C) 38HK cells were crosslinked and chromatin was extracted for ChIP-reChIP assay by p53 immunoprecipitation followed by p446PKR immunoprecipitation. Data from three independent experiments performed in duplicate were analysed by qPCR. P53RE2 and Chr22 intergenic sequence (nc) primers were used. *, $p<0.05$. (D) 38HK were treated with PKR inhibitor, 2-Aminopurine (2AP), or Acetic Acid: PBS solution (1:200), as a control, for 4 hours at 10mM final concentration. Nuclear extracts were used for 50% to 10% sucrose gradient protein complex isolation. Fractions obtained were immunoprecipitated with p53 antibody. Results were analysed by IB.

Figure 8. ITGA-1 ectopic overexpression impairs cell growth and induces cell death.
(A) 38HK cells were transfected with ITGA-1 cDNA (ITGA-1) or empty plasmid control (pcDNA). After zeocin selection, 38HK were fixed with crystal violet and total colony number was counted per well. Results are the mean of three independent experiments. ****,$p<0.0001$. (B) 38HK transfected with ITGA-1 cDNA (ITGA-1) or empty control (pcDNA) plasmid were fixed and stained with Propidium Iodide for Flow cytometry analysis. Histogram (right) represents the mean of sub-G0 population of three independent experiments. **, $p<0.01$. (C and D) Total protein and mRNA extracts from 38HK transfected cells were analysed by IB and RT-qPCR. Band intensity was quantified and normalized to β-Actin (n=4). ITGA-1, EGFR and Cyclin D1 (Cyc D1) mRNA levels were normalized to GAPDH (n=4). *, $p<0.05$, **$p<0.01$, ***$p<0.001$, ns, not significant.

Figure 9. ITGA-1 is downregulated and mutated in transgenic mice expressing HPV38

E6/E7.

(A) Skin keratinocytes were isolated from WT animals (n=4) and K14 HPV38 E6/E7 Tg mice (n=3). After 30 weeks UV radiation, squamous cell carcinoma samples (cSCC1-3) were isolated from HPV38 E/E7 Tg mice. Total RNA extraction was performed and ITGA-1 mRNA levels were determined by quantitative RT-PCR by normalizing to GAPDH. Whole exome sequencing DNA sequencing of the same mice was also performed. Mutational analysis of p53, ITGA-1 and EGFR genes was performed as described in materials and methods.

(B) Genomic position of the exonic mutations and the corresponding amino acid change are represented for ITGA-1 and EGFR genes. WT exons are represented as blue boxes while mutated exons are represented in red boxes. Text boxes describe the squamous cell carcinoma sample, genomic position of nucleotide change on the GRCm38/mm10 mouse reference genome, the type of mutation and the corresponding amino acid change.

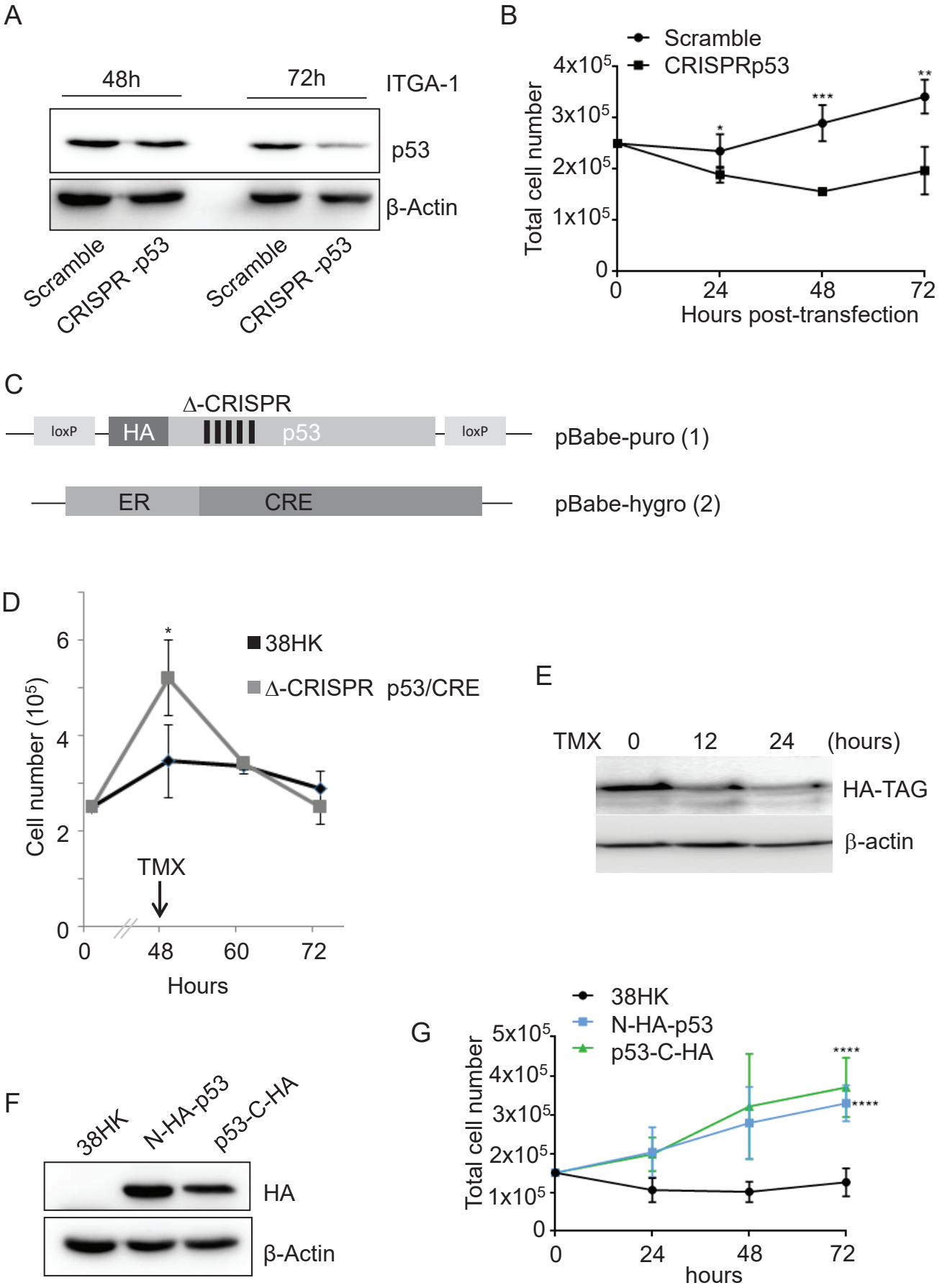


Figure 1, Romero et al.

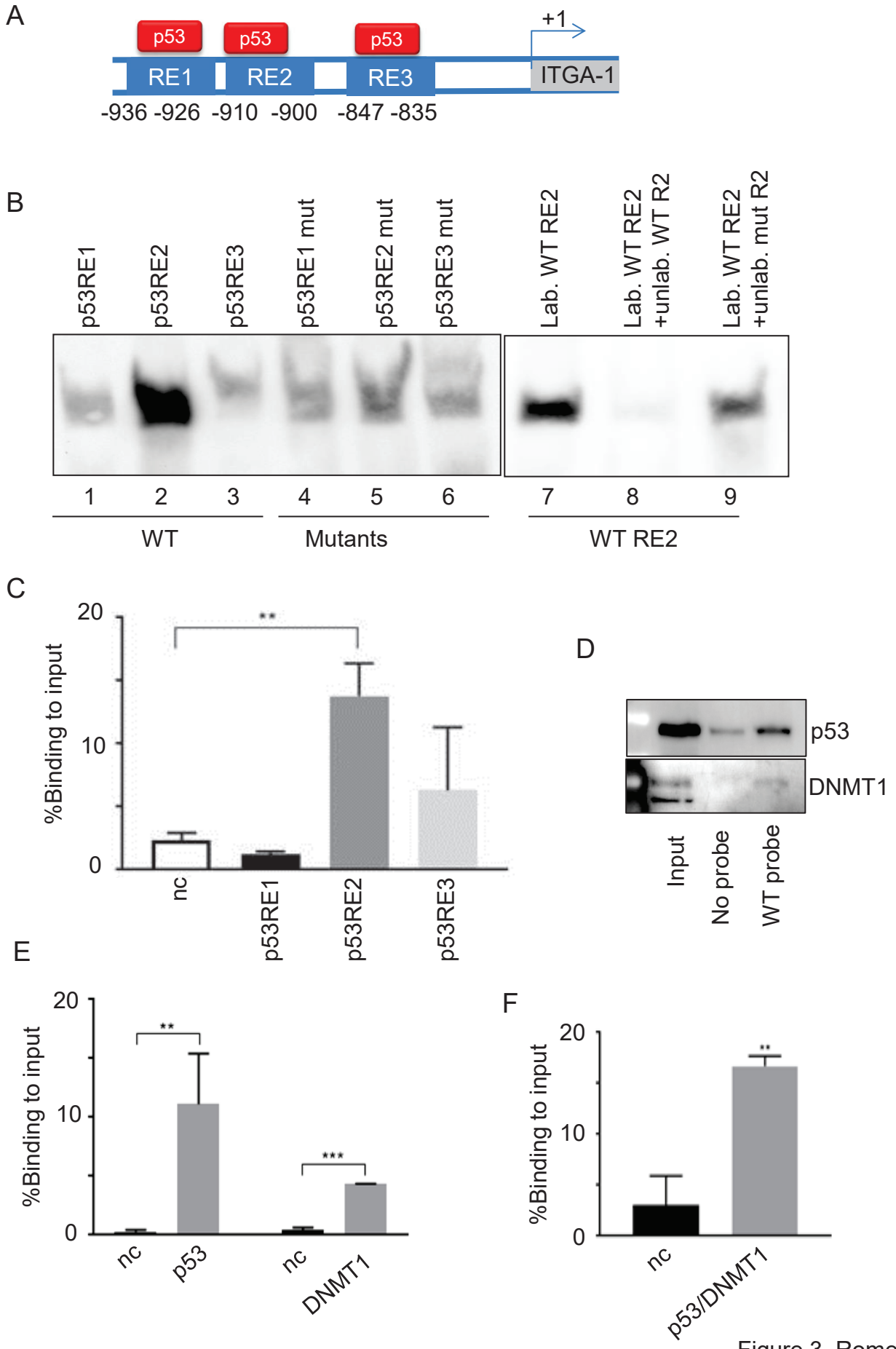


Figure 3, Romero et al.

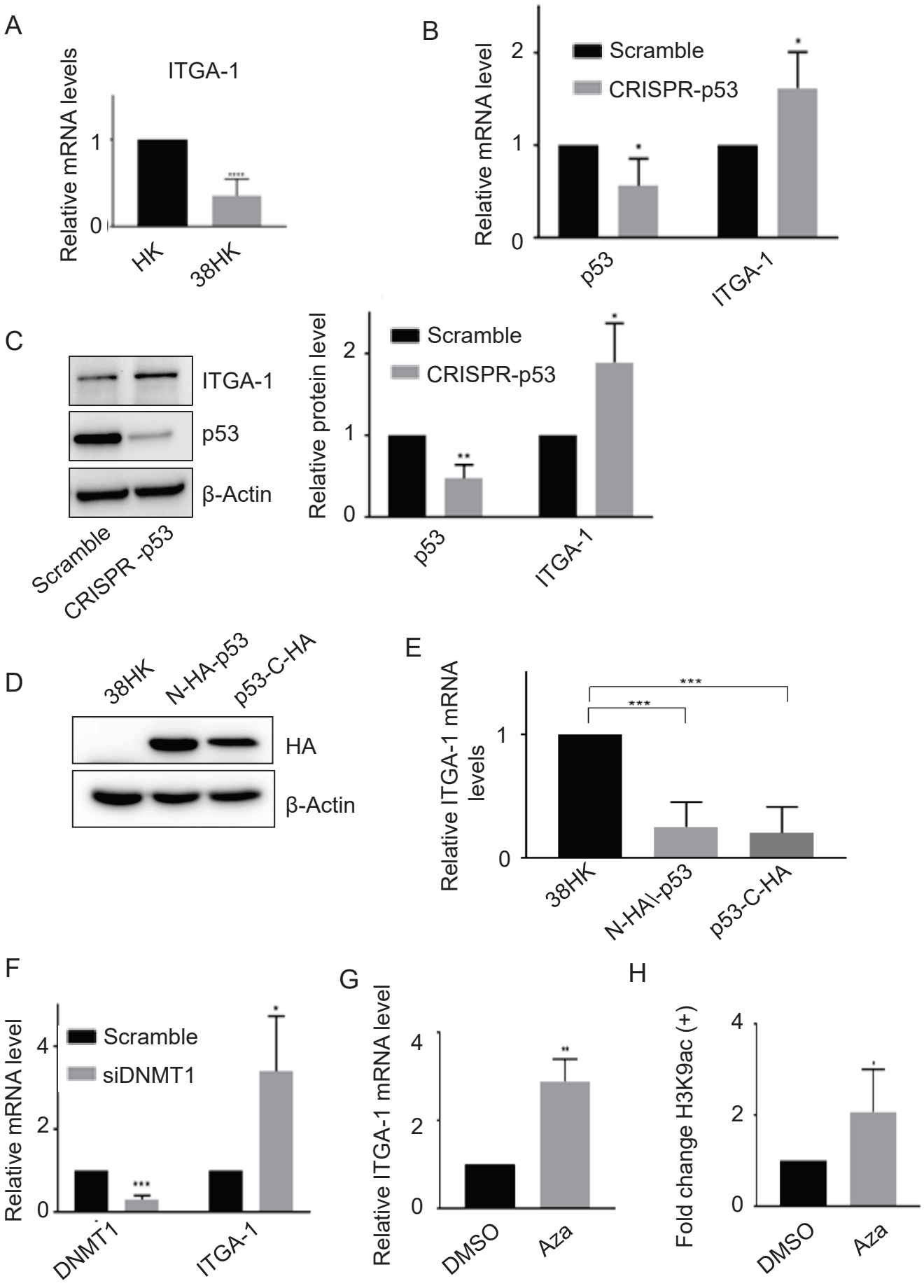


Figure 4, Romero et al.

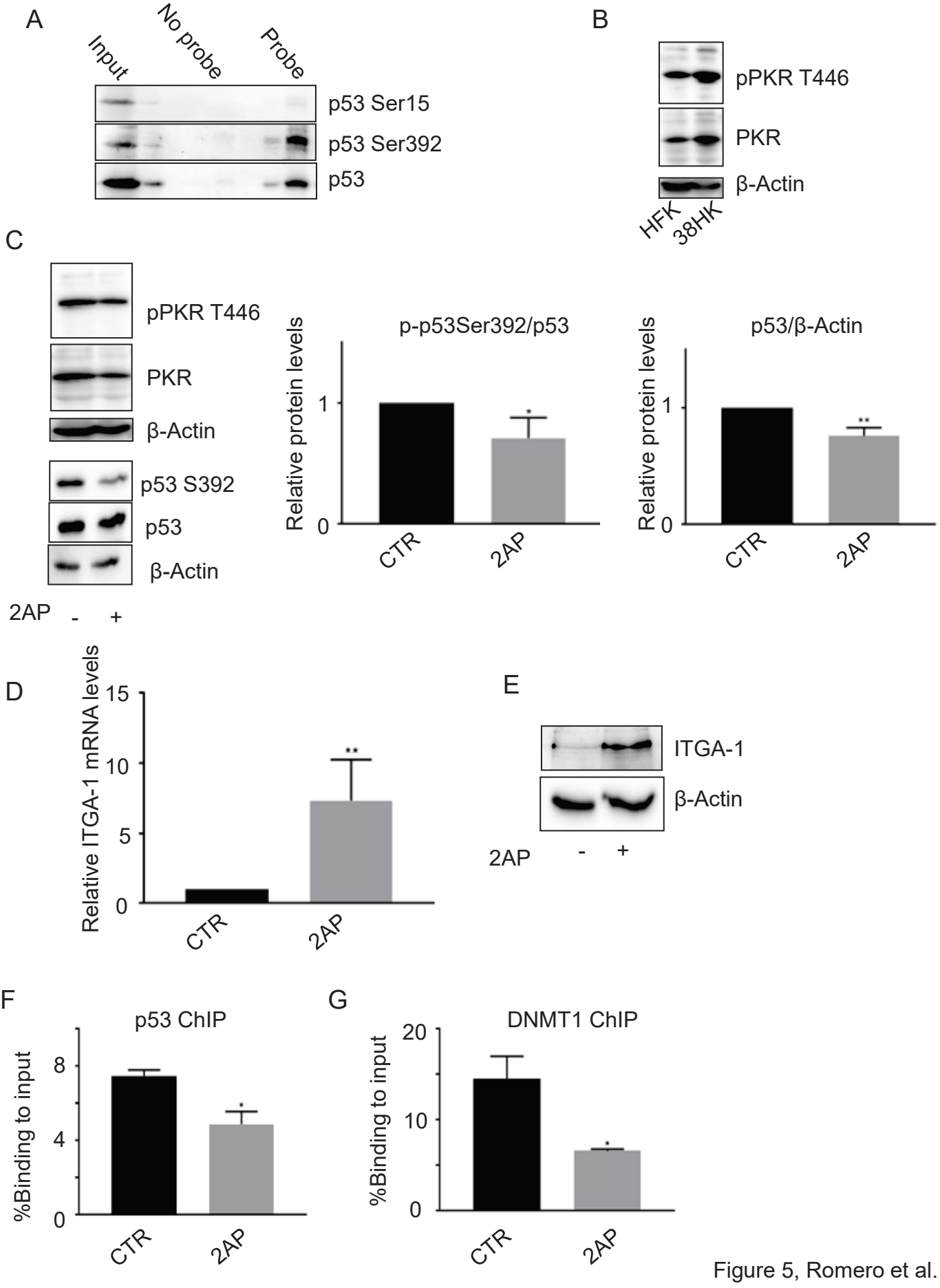


Figure 5, Romero et al.

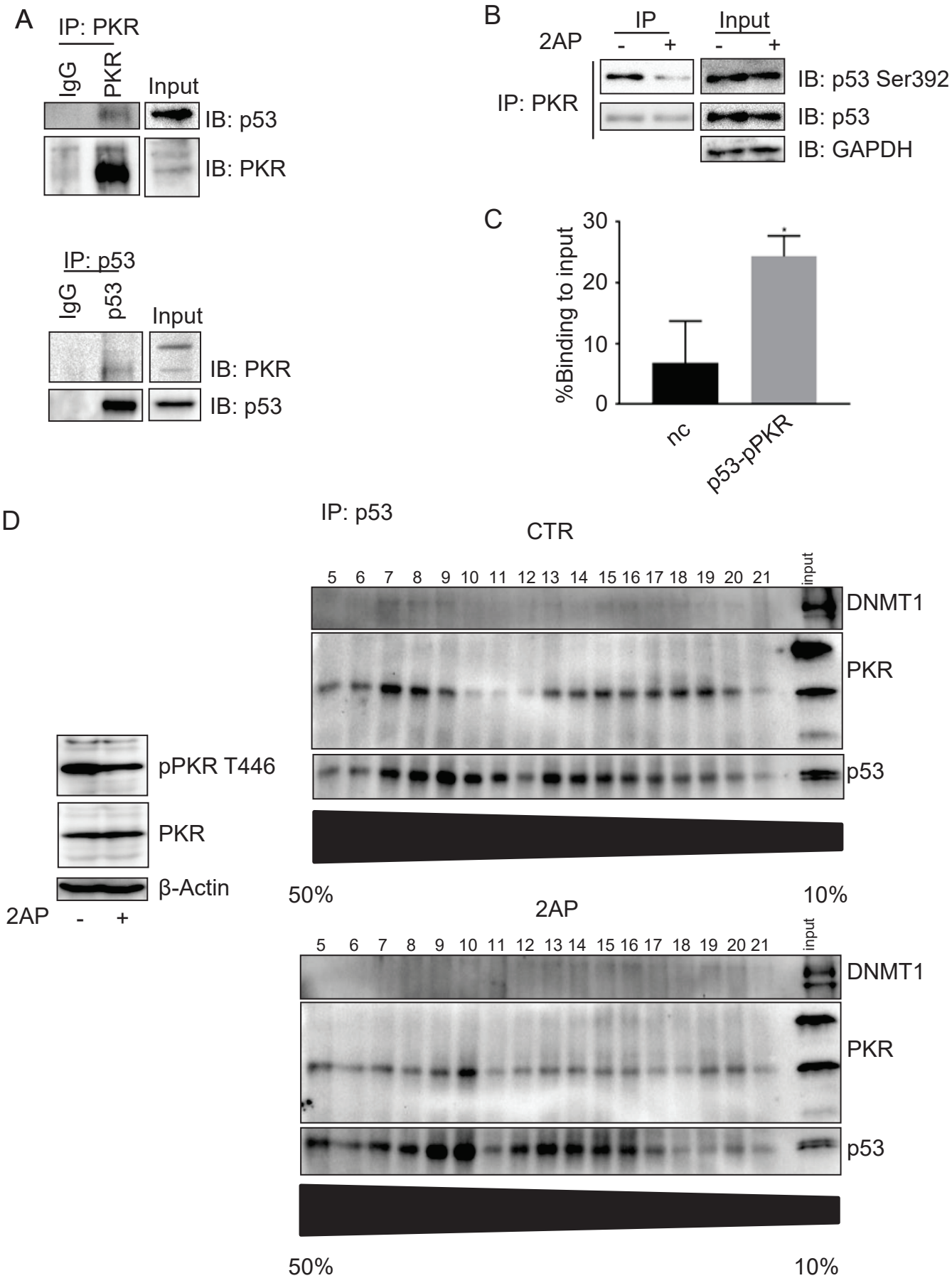


Figure 6, Romero et al.

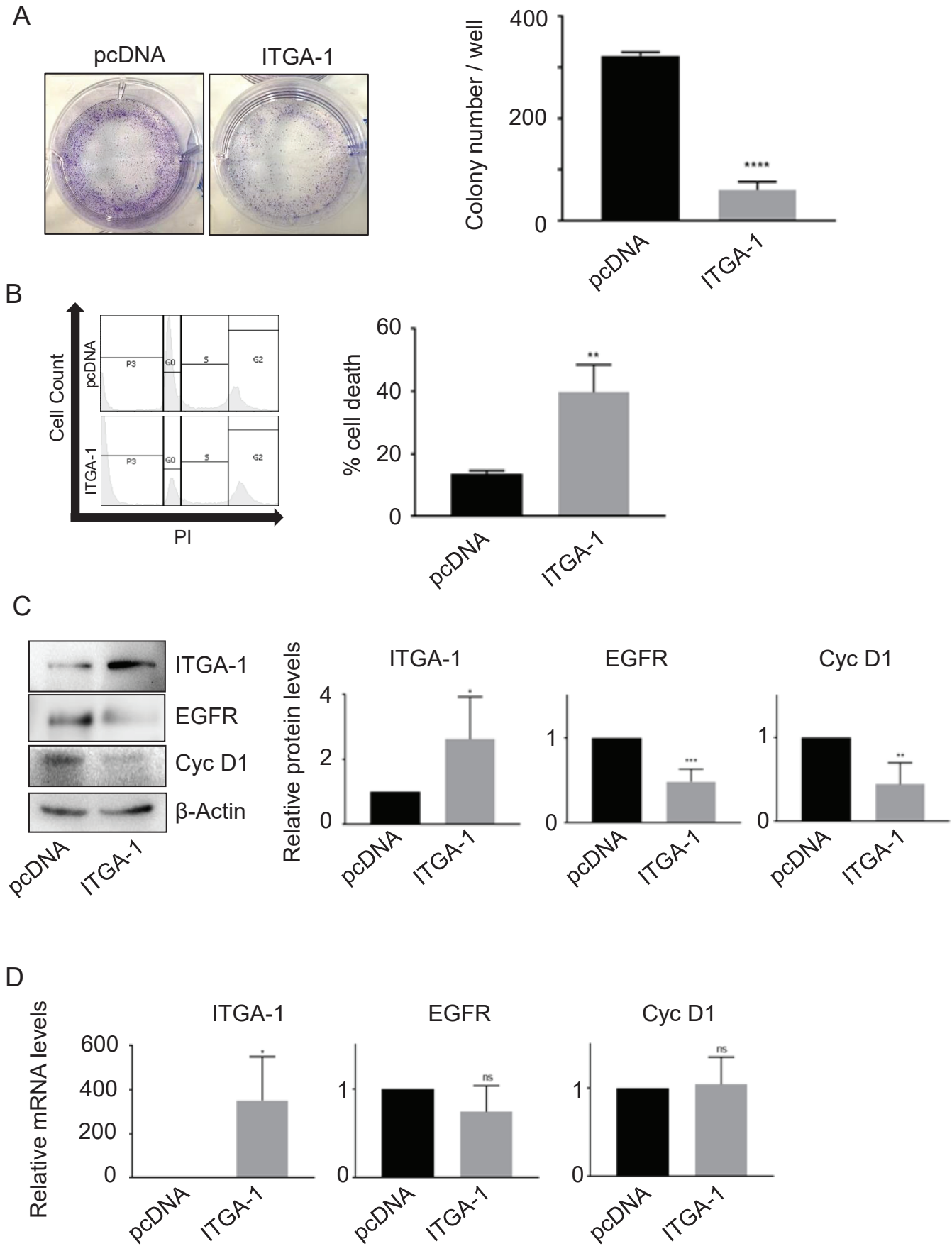


Figure 7, Romero et al.

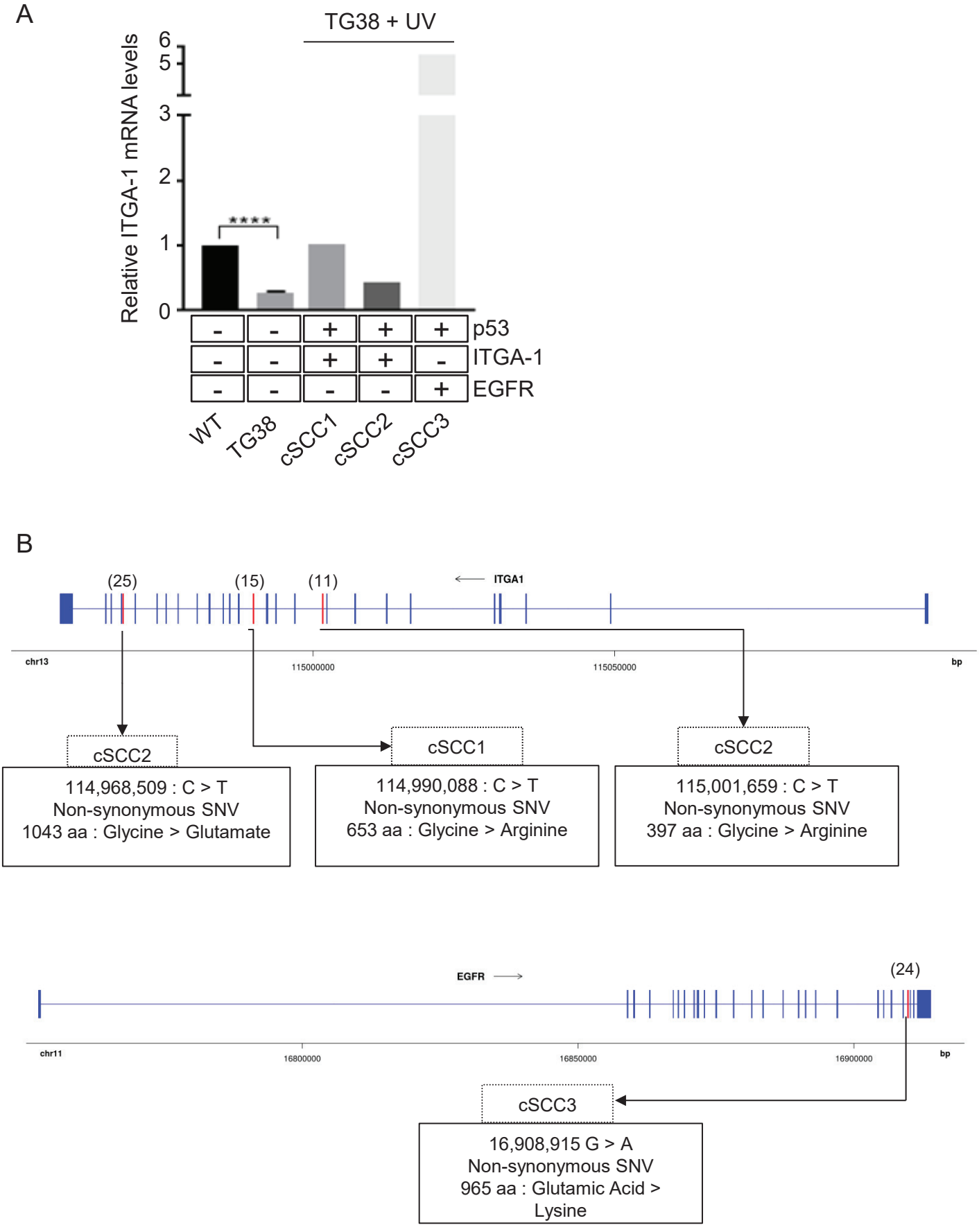
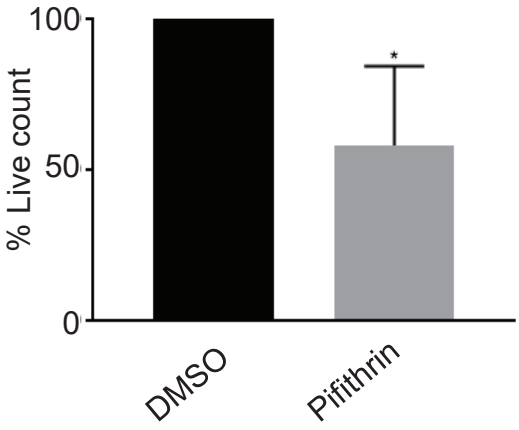


Figure 8, Romero et al.

A



B

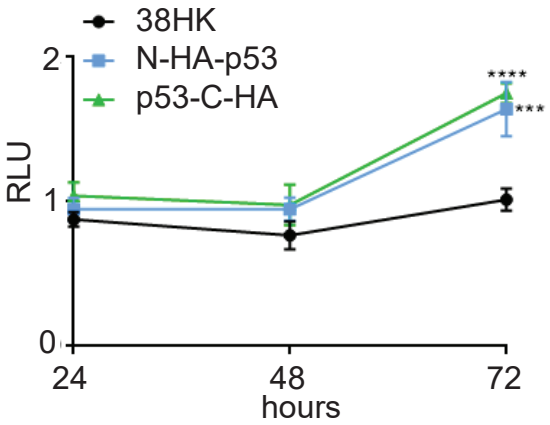


Figure S1, Romero et al.

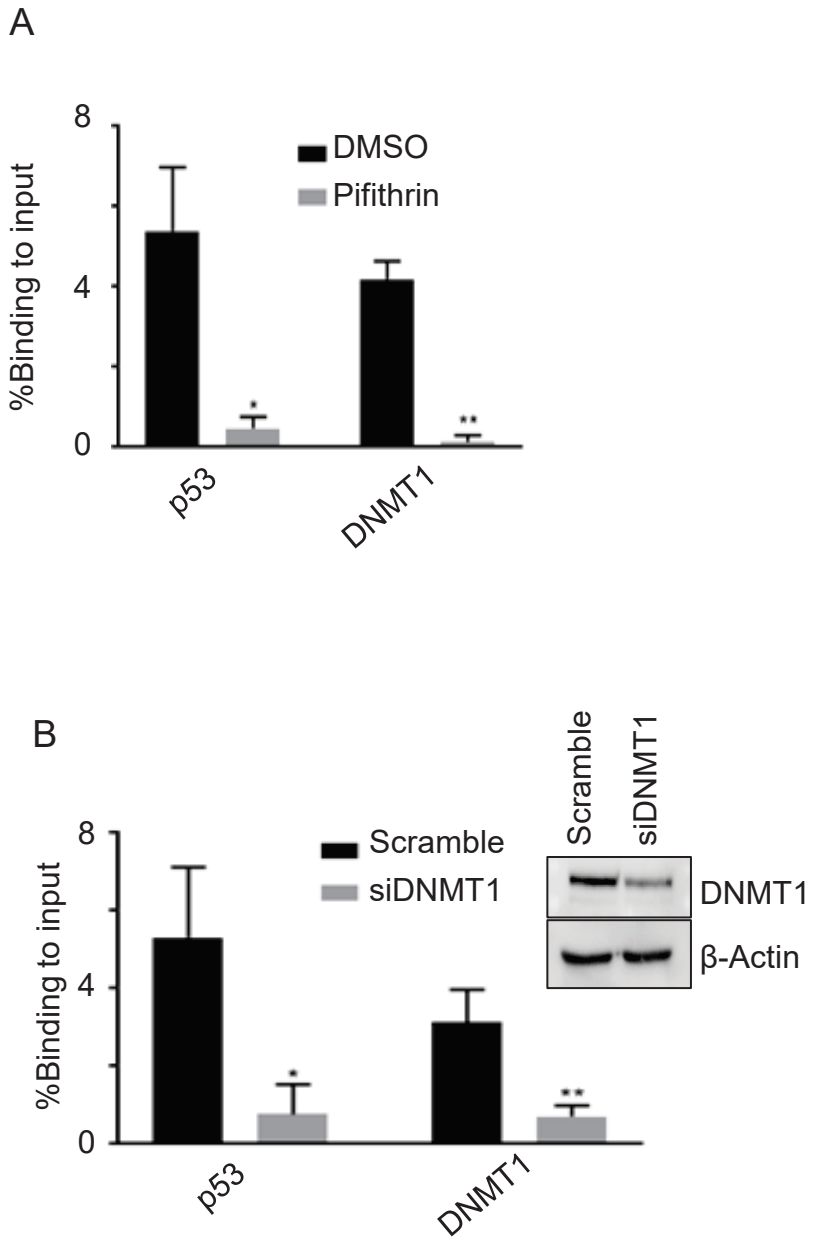


Figure S2, Romero et al.

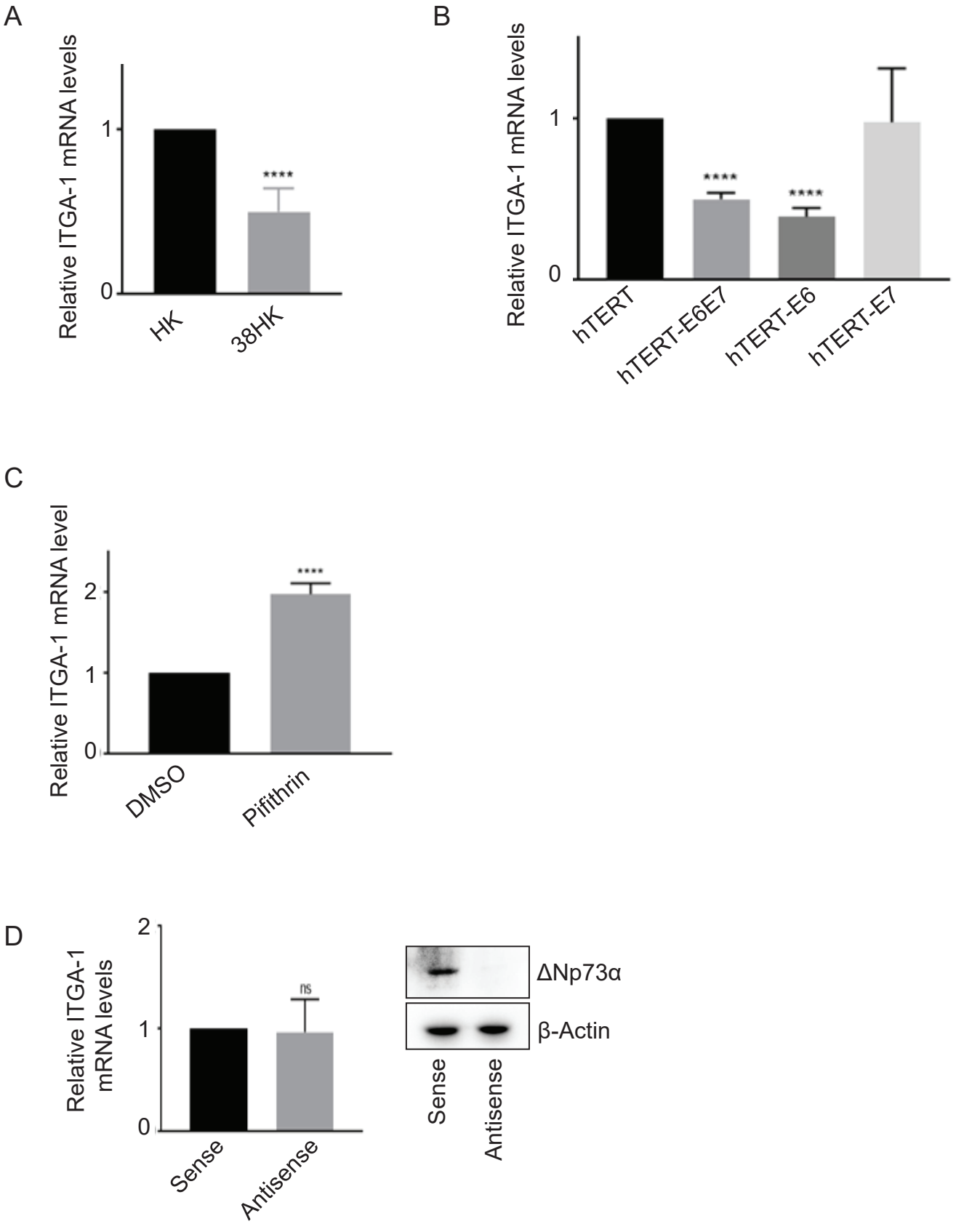


Figure S3, Romero et al.

Redox Biology 26 (2019) 101247

Contents lists available at ScienceDirect

Redox Biology

journal homepage: www.elsevier.com/locate/redox

RedoxiBase: A database for ROS homeostasis regulated proteins

Bruno Savelli[a,**], Qiang Li[a,b], Mark Webber[a], Achraf Mohamed Jemmat[a,c], Alexis Robitaille[a,d], Marcel Zamocky[e,f], Catherine Mathé[a], Christophe Dunand[a,*]

[a] Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, 24 chemin de Borde Rouge, Auzeville, BP42617, 31326, Castanet-Tolosan, France
[b] Citrus Research Institute, Southwest University/Chinese Academy of Agricultural Sciences, Beibei, Chongqing, 400712, China
[c] Institute for Botany and Molecular Genetics, Bioeconomy Science Center, RWTH Aachen University, Aachen, Germany
[d] International Agency for Research on Cancer, Lyon, France
[e] Department of Molecular Evolution & Development University Vienna, Althanstrasse 14, A-1090, Vienna, Austria
[f] Laboratory of Phylogenomic Ecology, Institute of Molecular Biology, Slovak Academy of Sciences, Dubravska cesta 21, SK-84551, Bratislava, Slovakia

ARTICLE INFO

Keywords:
Catalase
Peroxidases
Oxido-reductases
Multigenic family
ROS homeostasis

ABSTRACT

We present a new database, specifically devoted to ROS homeostasis regulated proteins. This database replaced our previous database, the PeroxiBase, which was focused only on various peroxidase families. The addition of 20 new protein families related with ROS homeostasis justifies the new name for this more complex and comprehensive database as RedoxiBase.

Besides enlarging the focus of the database, new analysis tools and functionalities have been developed and integrated through the web interface, with which the users can now directly access to orthologous sequences and see the chromosomal localization of sequences when available.

OrthoMCL tool, completed with a post-treatment process, provides precise predictions of orthologous gene groups for the sequences present in this database. In order to explore and analyse orthogroups results, taxonomic visualization of organisms containing sequence of a specific orthogroup as well as chromosomal distribution of the orthogroup with one or two organisms have been included.

1. Introduction

Reactive Oxygen Species (ROS) are represented by reactive molecules and free radicals derived from molecular oxygen: hydrogen peroxide, organic peroxides, superoxide, hydroxy radical, hydroxyl ion, singlet oxygen, and nitric oxide. They are produced at elevated concentrations during several essential biological processes such as respiration in most of living organisms, photosynthesis and photorespiration in chloroplastic organisms. They can also be released in a control manner during various developmental processes and stress responses. In particular, ROS can be produced as a part of innate immunity in Metazoans [1]. Although they can be deleterious, they are also necessary. To manage this ambivalent situation, each living being possesses a large battery of proteins which can produce or scavenge ROS in order to control their homeostasis. Among these proteins, haem or non-haem peroxidases were already centralized in a dedicated database namely the PeroxiBase [2].

In order to have a more integrative and phylogenomic overview on ROS-regulated proteins, new classes, families and superfamilies have been added to cover most of the proteins able to regulate ROS level. Then, the RedoxiBase, which includes all the data and the tools already present in the former PeroxiBase, was created. In the new database all living kingdoms are represented. The PeroxiBase served as a reference in the field of peroxidase families, the new enhanced version of this database should become a similar reference for all ROS regulation proteins. It is cross-referenced in UniProt [3] since 2006 and, more recently, in the Arabidopsis database TAIR [4].

Several databases centralize entries of all (InterPro [5]) or particular protein families (PLantCAZyme [6], CAZy [7], MEROPS [8], ThYme [9] and CaspBase, a curated database dedicated to the caspase family [10], or specific to a species such as GFDP which includes 6551 genes of poplar from 145 families [11]. Regarding the oxidase families, two independent databases are currently present in the web. Namely, PREX [12] is dedicated to only one type of non-haem peroxidases and fPOXDB [13] a fungal-specific database. They both bring structural and sequence information complementary to those found in our previous database PeroxiBase but they are merely devoted to subfamily assignment. Lastly, the antioxidant protein database AOD [14], was

* Corresponding author.
** Corresponding author.
E-mail addresses: savelli@lrsv.ups-tlse.fr (B. Savelli), dunand@lrsv.ups-tlse.fr (C. Dunand).

https://doi.org/10.1016/j.redox.2019.101247
Received 22 May 2019; Accepted 5 June 2019
Available online 06 June 2019

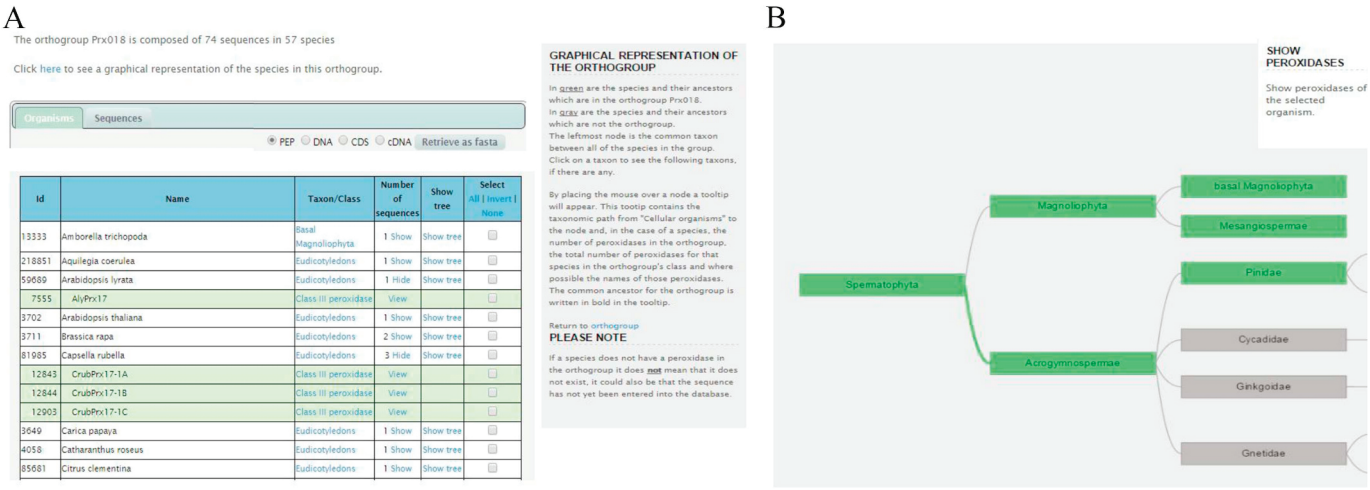


Fig. 1. Orthogroup pipeline results. A. List of the organisms containing sequences belonging to the selected orthogroup. B. Visualization of the taxonomic distribution within an othogroup. Green boxes stand for organisms containing sequences belonging to the selected orthogroup. Gray boxes stand for organisms lacking sequences belonging to the selected orthogroup. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

developed to understand the biological function of important antioxidant proteins but is was not maintained anymore.

Despite these different repositories, the (updated) RedoxiBase is still unique, since it is the only specialized collection of public sequences deduced from expert annotations with manual curation leading to re-annotation. Indeed, whole automatic genome annotation generates numbers of errors, notably with gene merging, splicing problems or tandem duplications [15]. These problems are exacerbated in the case of multigenic families like most proteins already included in our database. The guarantee of a high-quality sequence input is a prerequisite for performing reliable analyses, especially phylogeny. Efforts to provide only expert annotation derived sequences, in opposition to automated ones, exist elsewhere, but are still rather marginal.

Since its creation in 2004, the PeroxiBase has been a very active database with new sequences and new organisms daily added together with constant update of the interface with new tools and functionalities. Then, the RedoxiBase will take advantage of this existing dynamics to go further and pursue increase of available contents and features. Despite the explosion of genomic projects producing huge amounts of novel sequences that remain unexploited [16], the database will keep its initial interest to centralize high quality annotation for peroxidases and ROS-related proteins whereas it has only slightly evolved for semi-automatic annotation.

2. Description of tools and functions

2.1. Data available for each entry and tools

In April 2019, the database contains more than 15 000 sequences distributed over 2599 organisms. This brings an important biodiversity aspect and can grow further with availability of genomes from novel organisms. In addition to protein, cDNA, CDS, genomic, 2000 bp upstream and downstream sequences, the gene structure information (intron/exon structure), in Genbank format, is displayed along with a schematic representation.

The main challenge concerning large multigenic families is to obtain a comprehensive and reliable image of their evolution. To help establishing an evolutionary scenario, our interface provides many tools either to analyse the database entries or to compare them with input sequences. A regular BLAST including usual options (such as the nature of query and subject sequences and the choice of organism(s)) allows the users to search for sequences similar to their query in the database. Peroxiscan is a tool that provides the user with a prediction of a particular family or superfamily after testing the query sequence against pre-defined specific profiles [17]. CIWOG [18] and GECA [19] are tools that search for common introns in genes families based on intron position and protein sequence similarity around it. They return a graphical representation and comparison of several gene structures and highlight the conservation between sequences. The visualization of the alternative splicing, common in Metazoans, need to be developed. For multiple alignments, ClustalW and MAFFT are available directly online following multicriteria or BLAST searches, and a connection to the French phylogeny web site (http://www.phylogeny.fr) allows for further phylogenetic analysis. *Cis*-regulatory element analysis can be further performed with upstream and downstream sequences using PLACE [20] and MEME [21]. In addition, two major tools have been included for evolutionary and comparative genomic analyses and are described below.

2.2. New tool for evolutionary analysis: orthogroup

An orthogroup is defined as a group of peroxidases or ROS-related proteins that share a common ancestor. They are therefore either orthologs or paralogs. To perform clustering analysis and visualization, a specific pipeline, thereafter called ortho-pipeline, has been developed. This pipeline is based on OrthoMCL [22] and includes a post-treatment to reduce the false positives and negatives usually obtained with OrthoMCL. The originality and the relevance of our ortho-pipeline is to provide orthogroup classification even for partial sequences, based on sequence similarities.

Few new pages (Fig. 1A) were created on the web interface in order to visualize and analyse the taxonomic distribution of the orthogroups within different organisms. Graphical representation (Fig. 1B) of the orthogroup is available directly from one entry or from the tab “Browse the database by orthogroup” and “Analysis from input/Orthogroup search”. The green displayed the species and their ancestors, which

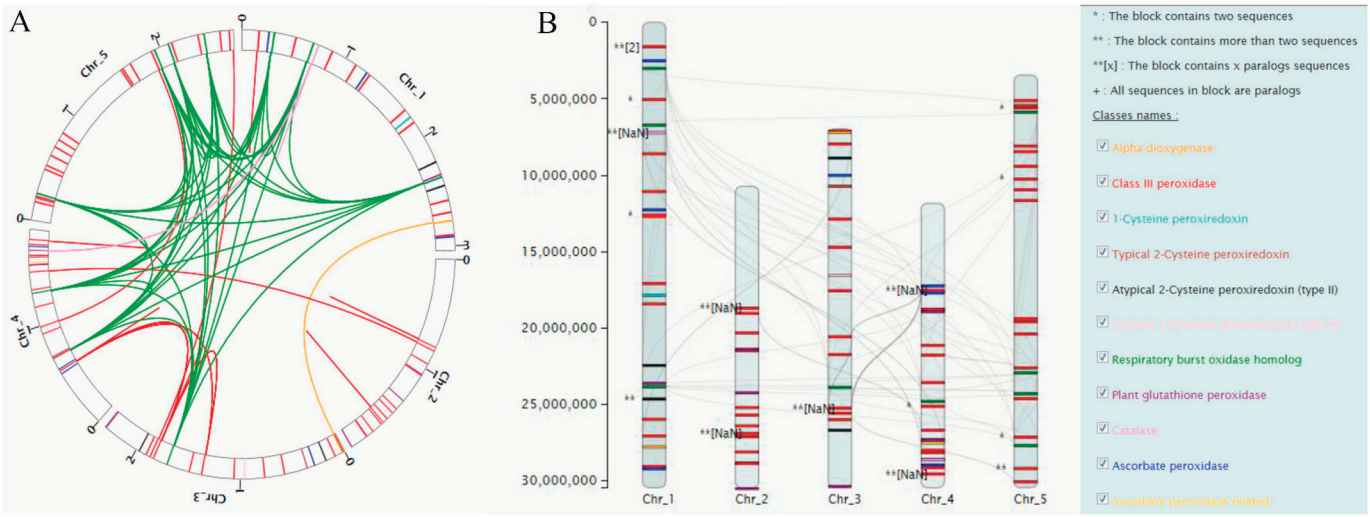


Fig. 2. Orthogroup pipeline visualization within one species. A. Circos-like visualization. B. Chromosome Map visualization. Sequences belonging to the same orthogroup are linked. Each class is represented with one colour. Chromosome and gene loci on chromosomes are on scale. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

possess sequences from the visualized orthogroup, while gray showed species that do not have sequences from the visualized orthogroup. The lack of sequence inside a visualized orthogroup can result from the absence of data or to the loss of sequence in a given species.

2.3. New tools for comparative genomics: Circos and chromodraw

As we are convinced that the information resulting from the orthoMCL-pipeline can play a major role to elucidate evolutionary history, an additional pipeline with chromosomal localization was developed: Circos-like visualization [23] and Chromosome Map (mapchart like [24]), allowing large scale genomic analysis, have been included. Standardised name for each chromosome, the location of each peroxidase or ROS-related protein on their respective chromosome (if available) and the paralogy/orthology relathionship obtained from OrthMCL pipeline were included in the final output (Figs. 2 and 3).

2.4. New web interface and new code

As described above, the availability of a set of tools – some developed by our team - directly executable through the database website, facilitates evolutionary analysis. In addition, to improve the management of the database, as well as the speed of script execution and the database querying, the web application has been implemented in an open-source PHP framework (Codeigniter). This framework uses the Model-View-Controller concept and allows faster development, best security, better maintenance of the code and a reusability of applications developed in the laboratory with the same framework. Since 2008, the database is hosted by the GenoToul bioinformatics facility (http://bioinfo.genopole-toulouse.prd.fr). Recently, a new powerful computing cluster is available and can be used for local phylogenetic and clustering analysis.

3. Discussion and future prospects

With the accumulation of available genomes, the number of sequences included in the database was largely increased (from 6026 in 2008 [17] to 10710 in 2012 [2] and 15136 in 2019). Although, the numbers of organisms within each kingdom are in the same range, the RedoxiBase (formerly PeroxiBase) is still mainly composed of sequences originated from Viridiplantae (64%) and from fungi (22%). This is mainly due to the larger size of the red-ox proteins families found in plants and fungi which are subjected to large duplication events. Then, a particular effort needs to be done to increase the representation of ROS-related proteins from other kingdoms (mainly Protista and Animalia) and within them from exotic and poorly represented organisms. Special attention must be paid to genes from those species threatened with global extinction as reported recently by IPBES (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services Paris 2019). Regularly updating RedoxiBase with manually annotated sequences will allow to perform robust evolutionary analyses also for concatenated sequences.

The quality of the annotation, which is our main concern since the creation of the database, has been maintained, but manual annotation does not allow an efficient coverage of all the available sequences. The semi-automatic protocols developed will facilitate the upload of peroxidase-encoding sequences from already annotated proteomes while maintaining our high-quality standard. In addition, the annotation procedure relying on Scipio which has already demonstrated its effectiveness for gene prediction based on homology with closely related already annotated organisms [25], will be improved. Indeed, a new strategy that will take advantage of our specific profiles defined with controlled batches of sequences need to be developed for the prediction in more divergent genomes.

Many red-ox proteins families included in the RedoxiBase belong to multigenic families and result from tandem, segmental and chromosomal duplication events, which complicates global phylogenetic analysis and the understanding of their evolutionary history. The visualization of inter- or intra-species sequence orthogroup belonging and their chromosomal localization is very helpful in this context. This requires the availability of genomic localization for larger number of organisms. In addition, we have recently developed ExpressWeb, an online tool to perform gene clustering using personal or selected expressed value sets in order to construct co-expression gene networks [26]. ExpressWeb is available directly from the RedoxiBase and a current priority is to set up a pipeline to load publicly available expression data in order to perform expression clustering with our favorite genes.

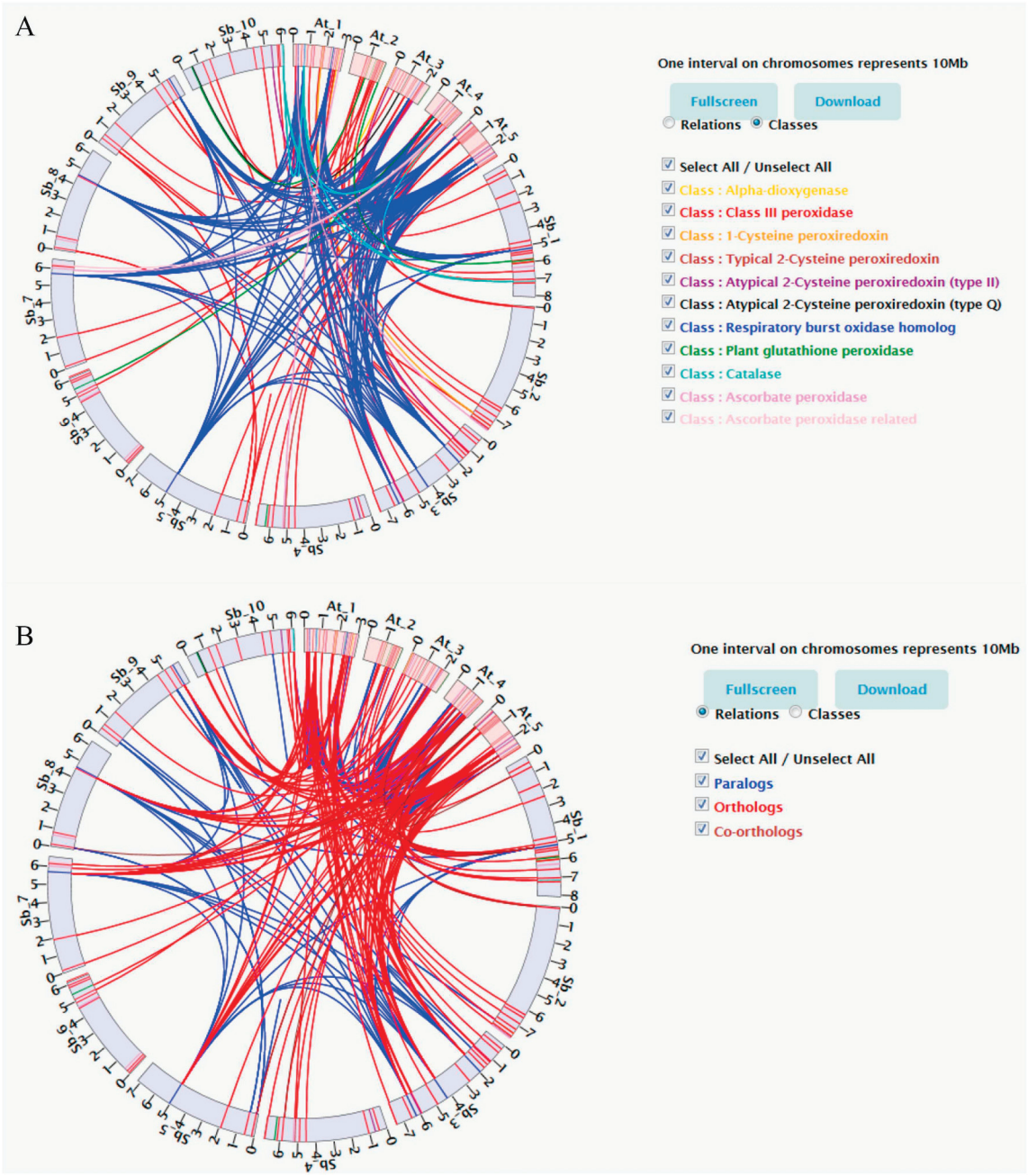


Fig. 3. Orthogroup pipeline visualization between two species. A. Circos-like visualization of relation based on class belonging. B. Circos-like visualization of orthology/paralogy relations. Each class is represented with one colour. Chromosome and gene loci on chromosomes are on scale. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Acknowledgments

The authors are thankful to the Paul Sabatier-Toulouse 3 University and to the *Centre National de la Recherche Scientifique* (CNRS) for granting their work. We thank all the past contributors and curators of the PeroxiBase, and Sylvain Picard and Raphael Taris for their contributions to the development of the RedoxiBase. The RedoxiBase is hosted by the Toulouse Midi-Pyrénées bioinformatics platform. We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing help and/or computing and/or storage resources. This work has been done in the Plant Science Research Laboratory (LRSV). MZ was supported by the Austrian Science Fund (FWF, project P 31707-B32) and by the Slovak Grant Agency VEGA (grant 2/0061/18).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.redox.2019.101247.

References

[1] C. Kohchi, H. Inagawa, T. Nishizawa, G. Soma, ROS and innate immunity, Anticancer Res. 29 (2009) 817–821.
[2] N. Fawal, Q. Li, B. Savelli, M. Brette, G. Passaia, M. Fabre, C. Mathé, C. Dunand, PeroxiBase: a database for large-scale evolutionary analysis of peroxidases, Nucleic Acids Res. 41 (2013) D441–D444.
[3] U. Consortium, Reorganizing the protein space at the universal protein resource (UniProt), Nucleic Acids Res. 40 (2012) D71–D75.
[4] P. Lamesch, T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller,

K. Dreher, D.L. Alexander, M. Garcia-Hernandez, et al., The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools, Nucleic Acids Res. 40 (2012) D1202–D1210.
[5] R. Finn, T. Attwood, P. Babbitt, A. Bateman, P. Bork, A. Bridge, H. Chang, Z. Dosztanyi, S. El-Gebali, M. Fraser, et al., InterPro in 2017-beyond protein family and domain annotations, Nucleic Acids Res. 45 (2017) D190–D199.
[6] A. Ekstrom, R. Taujale, N. McGinn, Y. Yin, PlantCAZyme: a Database for Plant Carbohydrate-Active Enzymes. Database (Oxford), 2014, (2014).
[7] B.L. Cantarel, P.M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, B. Henrissat, The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics, Nucleic Acids Res. 37 (2009) D233–D238.
[8] N.D. Rawlings, A.J. Barrett, A. Bateman, MEROPS: the database of proteolytic enzymes, their substrates and inhibitors, Nucleic Acids Res. 40 (2012) D343–D350.
[9] D.C. Cantu, Y. Chen, M.L. Lemons, P.J. Reilly, ThYme: a database for thioester-active enzymes, Nucleic Acids Res. 39 (2011) D342–D346.
[10] R.D. Grinshpon, A. Williford, J. Titus-McQuillan, A. Clay Clark, The CaspBase: a curated database for evolutionary biochemical studies of caspase functional divergence and ancestral sequence inference, Protein Sci. 27 (2018) 1857–1870.
[11] H. Wang, H. Yan, H. Liu, R. Liu, J. Chen, Y. Xiang, GFDP: The Gene Family Database in Poplar, Database 2018 (2018) 1–8.
[12] L. Soito, C. Williamson, S.T. Knutson, J.S. Fetrow, L.B. Poole, K.J. Nelson, PREX: PeroxiRedoxin classification indEX, a database of subfamily assignments across the diverse peroxiredoxin family, Nucleic Acids Res. 39 (2011) D332–D337.
[13] R.A. Ohm, R. Riley, A. Salamov, B. Min, I.G. Choi, I.V. Grigoriev, Genomics of wood-degrading fungi, Fungal Genet. Biol. 72 (2014) 82–90.
[14] P. Feng, H. Ding, H. Lin, W. Chen, AOD: the antioxidant protein database, Sci. Rep. 7 (2017) 7449.
[15] N. Fawal, Q. Li, C. Mathé, C. Dunand, Automatic multigenic family annotation: risks and solutions, Trends Genet. 30 (2014) 323–325.
[16] S. Mukherjee, D. Stamatis, J. Bertsch, G. Ovchinnikova, H.Y. Katta, A. Mojica, I.A. Chen, N.C. Kyrpides, T. Reddy, Genomes OnLine database (GOLD) v.7: updates and new features, Nucleic Acids Res. 47 (2019) D649–D659.
[17] D. Koua, L. Cerutti, L. Falquet, C.J.A. Sigrist, G. Theiler, N. Hulo, C. Dunand, PeroxiBase: a database with new tools for peroxidase family classification, Nucleic Acids Res. 37 (2009) D261–D266.
[18] M.D. Wilkerson, Y.B. Ru, V.P. Brendel, Common introns within orthologous genes: software and application to plants, Briefings Bioinf. 10 (2009) 631–644.
[19] N. Fawal, B. Savelli, C. Dunand, C. Mathé, GECA: a fast tool for gene evolution and conservation analysis in eukaryotic protein families, Bioinformatics 28 (2012) 1398–1399.
[20] K. Higo, Y. Ugawa, M. Iwamoto, T. Korenaga, Plant cis-acting regulatory DNA elements (PLACE) database: 1999, Nucleic Acids Res. 27 (1999) 297–300.
[21] T.L. Bailey, J. Johnson, C.E. Grant, W.S. Noble, The MEME suite, Nucleic Acids Res. 43 (2015) W39–W49.
[22] L. Li, C.J. Stoeckert, D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes, Genome Res. 13 (2003) 2178–2189.
[23] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S.J. Jones, M.A. Marra, Circos: an information aesthetic for comparative genomics, Genome Res. 19 (2009) 1639–1645.
[24] R.E. Voorrips, MapChart: software for the graphical presentation of linkage maps and QTLs, J. Hered. 93 (2002) 77–78.
[25] O. Keller, F. Odronitz, M. Stanke, M. Kollmar, S. Waack, Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species, BMC Bioinf. 9 (2008).
[26] B. Savelli, S. Picard, C. Roux, C. Dunand, ExpressWeb: A Web Application for Clustering and Visualization of Expression Data, bioRxiv (2019) 625939.

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

VIRUS-CELL INTERACTIONS

Interplay between the Epigenetic Enzyme Lysine (K)-Specific Demethylase 2B and Epstein-Barr Virus Infection

Romina C. Vargas-Ayala,[a] Antonin Jay,[a] Francesca Manara,[a] Mohamed Ali Maroui,[b,c] Hector Hernandez-Vargas,[d] Audrey Diederichs,[d] Alexis Robitaille,[a] Cecilia Sirand,[a] Maria Grazia Ceraolo,[a,e] Maria Carmen Romero-Medina,[a] Marie Pierre Cros,[a] Cyrille Cuenin,[a] Geoffroy Durand,[a] Florence Le Calvez-Kelm,[a] Lucia Mundo,[f] Lorenzo Leoncini,[f] Evelyne Manet,[b,c] Zdenko Herceg,[a] Henri Gruffat,[b,c] Rosita Accardi[a]

[a]International Agency for Research on Cancer, World Health Organization, Lyon, France
[b]CIRI, Centre International de Recherche en Infectiologie (Oncogenic Herpesviruses Team), Université de Lyon, Lyon, France
[c]INSERM, U1111, CNRS, UMR5308, ENS de Lyon, Université Claude Bernard Lyon 1, Lyon, France
[d]Lyon Cancer Research Center (CRCL), INSERM U1052, Centre Léon Bérard, Lyon, France
[e]Tumor Immunology Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milan, Italy
[f]Department of Medical Biotechnology, Section of Pathology, University of Siena, Siena, Italy

ABSTRACT The histone modifier lysine (K)-specific demethylase 2B (KDM2B) plays a role in the differentiation of hematopoietic cells, and its expression appears to be deregulated in certain cancers of hematological and lymphoid origins. We have previously found that the *KDM2B* gene is differentially methylated in cell lines derived from Epstein-Barr virus (EBV)-associated endemic Burkitt lymphoma (eBL) compared with that in EBV-negative sporadic Burkitt lymphoma-derived cells. However, whether KDM2B plays a role in eBL development has not been previously investigated. Oncogenic viruses have been shown to hijack the host cell epigenome to complete their life cycle and to promote the transformation process by perturbing cell chromatin organization. Here, we investigated whether EBV alters KDM2B levels to enable its life cycle and promote B-cell transformation. We show that infection of B cells with EBV leads to downregulation of KDM2B levels. We also show that LMP1, one of the main EBV transforming proteins, induces increased DNMT1 recruitment to the *KDM2B* gene and augments its methylation. By altering KDM2B levels and performing chromatin immunoprecipitation in EBV-infected B cells, we show that KDM2B is recruited to the EBV gene promoters and inhibits their expression. Furthermore, forced KDM2B expression in immortalized B cells led to altered mRNA levels of some differentiation-related genes. Our data show that EBV deregulates KDM2B levels through an epigenetic mechanism and provide evidence for a role of KDM2B in regulating virus and host cell gene expression, warranting further investigations to assess the role of KDM2B in the process of EBV-mediated lymphomagenesis.

IMPORTANCE In Africa, Epstein-Barr virus infection is associated with endemic Burkitt lymphoma, a pediatric cancer. The molecular events leading to its development are poorly understood compared with those leading to sporadic Burkitt lymphoma. In a previous study, by analyzing the DNA methylation changes in endemic compared with sporadic Burkitt lymphoma cell lines, we identified several differential methylated genomic positions in the proximity of genes with a potential role in cancer, and among them was the *KDM2B* gene. *KDM2B* encodes a histone H3 demethylase already shown to be involved in some hematological disorders. However, whether KDM2B plays a role in the development of Epstein-Barr virus-mediated lymphoma has not been investigated before. In this study, we show that Epstein-Barr virus deregulates KDM2B expression and describe the underlying mechanisms. We

Citation Vargas-Ayala RC, Jay A, Manara F, Maroui MA, Hernandez-Vargas H, Diederichs A, Robitaille A, Sirand C, Ceraolo MG, Romero-Medina MC, Cros MP, Cuenin C, Durand G, Le Calvez-Kelm F, Mundo L, Leoncini L, Manet E, Herceg Z, Gruffat H, Accardi R. 2019. Interplay between the epigenetic enzyme lysine (K)-specific demethylase 2B and Epstein-Barr virus infection. J Virol 93:e00273-19. https://doi.org/10.1128/JVI.00273-19.

Editor Richard M. Longnecker, Northwestern University

Address correspondence to Henri Gruffat, henri.gruffat@ens-lyon.fr, or Rosita Accardi, accardir@iarc.fr.

R.C.V.-A. and A.J. are co-first authors of this article.

Received 18 February 2019
Accepted 4 April 2019
Accepted manuscript posted online 17 April 2019
Published 14 June 2019

also reveal a role of the demethylase in controlling viral and B-cell gene expression, thus highlighting a novel interaction between the virus and the cellular epigenome.

KEYWORDS Burkitt lymphomas, EBV, epigenetic, KDM2B

Epstein-Barr virus (EBV) is a human gammaherpesvirus that infects more than 95% of the adult population worldwide. After infection, EBV establishes a lifelong latency, often with no adverse health consequences. Despite its ubiquity, EBV infection is also associated with many human cancer types, among which is endemic Burkitt lymphoma (eBL), the most common childhood cancer in equatorial Africa (1). Although this malignancy was associated with EBV infection more than 50 years ago, the exact mechanism by which the virus contributes to the eBL pathogenic process is still not fully understood.

Many studies have highlighted a key role of epigenetic deregulations in cell transformation and cancer development. Increasing evidence indicates that different viruses may abrogate cellular defense systems by hijacking epigenetic mechanisms to deregulate the host cell gene expression program and modulate their own life cycle (2, 3). Our recent study of the methylome profiles of sporadic Burkitt lymphoma (sBL)- versus eBL-derived cell lines revealed an EBV infection-specific pattern of methylation, with aberrant methylation being detected in genes with a known role in lymphomagenesis, such as *ID3*, which is often found to be mutated in sBL (4). We therefore hypothesized that a virus-driven mechanism is responsible for modifying the epigenome of B cells to facilitate the lymphomagenic process, circumventing the need for mutations in lymphoma driver genes. Among the genes differentially methylated in eBL compared with sBL, we identified the lysine (K)-specific demethylase 2B (*KDM2B*) gene, which encodes a histone H3 demethylase known to target specific sites, such as trimethylated lysine 4 (H3K4me3) and dimethylated lysine 36 (H3K36me2). KDM2B sets the stage for DNA methylation and gene silencing by recruiting polycomb-1 proteins to unmethylated CpG regions (5) and plays a key role in somatic cell reprogramming (6). It also represses the transcription of rRNA genes, thus inhibiting cell growth and proliferation (7). KDM2B has been identified as a putative tumor suppressor by retroviral insertion analysis in mice (8). Low levels of KDM2B expression have been found in aggressive brain tumors, suggesting its potential role in cancer development. Moreover, KDM2B is involved in hematopoietic cell development and plays opposite roles in tumors of hematopoietic and lymphoid origins (9). Although high levels of KDM2B expression have been observed in different hematological malignancies, its depletion from hematopoietic cells has been reported to activate the cell cycle and reduce the activity of interferon and lymphoid-specific transcription factors, thereby contributing to myeloid transformation (9). However, whether KDM2B affects the EBV life cycle has not been determined, and its role in eBL has not been assessed. Here, using *in vitro* EBV infection models, we aimed to assess whether EBV can alter the expression of KDM2B by inducing methylation of its gene. Finally, we investigated how this event affects EBV infection and B-cell homeostasis. Overall, our data highlight a novel cross talk between EBV and the cellular epigenome and identify KDM2B to be a master regulator of EBV gene expression, in addition to B-cell gene expression, suggesting a role for EBV-mediated KDM2B deregulation in the lymphomagenic process.

(This article was submitted to an online preprint archive [10].)

RESULTS

KDM2B is epigenetically silenced in EBV(+) BL-derived cell lines. Our previous comparative analysis of the whole-genome methylation profiles of a set of EBV-positive [EBV(+)] and EBV-negative [EBV(−)] Burkitt lymphoma (BL)-derived cell lines (4) led to the identification of two CpGs (CpG15695155 and CpG21423404) flanking a CpG island named CpG127 (Fig. 1A) in an intragenic putative regulatory region of *KDM2B* (as shown by the accumulation of the H3K27 acetylation [H3K27Ac] marker) (Fig. 1A). CpG15695155 and CpG21423404 were highly methylated in EBV(+) BL-derived cells

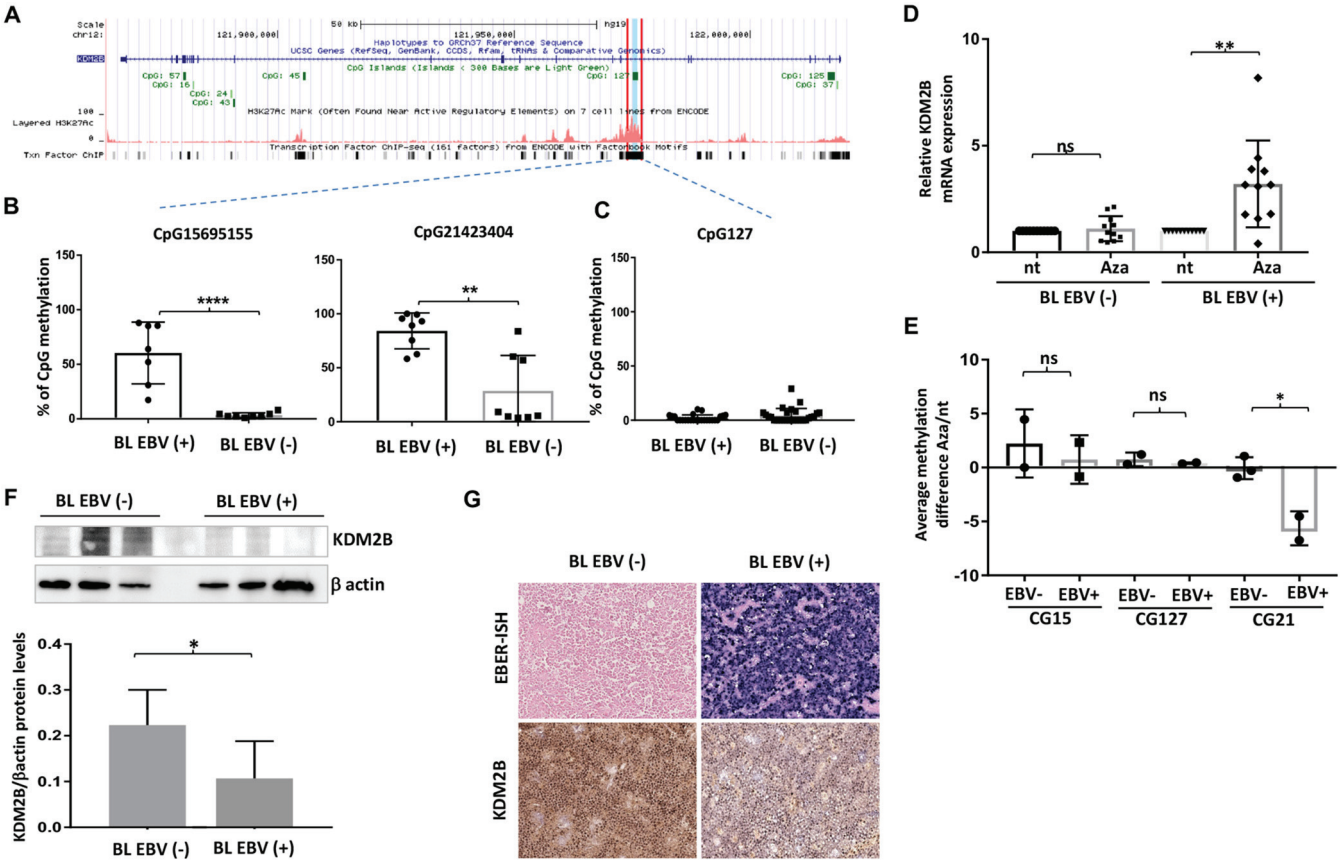


FIG 1 The *KDM2B* gene is methylated and silenced in EBV(+) BL cell lines and specimens. (A) Schematic diagram of the *KDM2B* gene (modified from the UCSC Genome Browser). Red lines show CpG15695155 and CpG21423404, and CpG island 127 is in light blue. ChIP data (obtained with the lymphoma cell line GM12878) for the distribution of the H3K27Ac marker within the selected region are also shown. ChIP, chromatin immunoprecipitation. (B) The histograms show the average percentage of methylation measured by pyrosequencing of CpG15695155 and CpG21423404 in the DNA of 10 EBV(+) and 9 EBV(−) BL cell lines (****, $P < 0.0001$; **, $P < 0.01$). (C) The histogram shows the average percentage of methylation at 17 positions within CpG island 127 measured in 2 EBV(+) and 2 EBV(−) BL cell lines. The difference between the two groups was not significant. (D) Three EBV(+) and 3 EBV(−) BL cell lines were cultured in the presence of dimethyl sulfoxide (DMSO; nontreated [nt]) or 5-aza-2′-deoxycytidine (Aza; 10 μM) for 48 h. The KDM2B mRNA expression level was evaluated by RT-qPCR. The pooled results of 4 independent experiments are presented in the histogram (**, $P < 0.01$; ns, not significant). KDM2B mRNA levels in Aza-treated cells were measured relative to the levels in DMSO-treated control cells. (E) Aza-treated EBV(−) and EBV(+) BL cells from 2 independent experiments were processed for DNA extraction and analyzed by pyrosequencing for the methylation level at CpG15695155 (CG15) and CpG21423404 (CG21) as well as for the average methylation at 17 positions within CpG island 127 (CG127). The average difference in the percentage of methylation between Aza- and DMSO-treated cells is indicated in the histogram (*, $P < 0.05$; ns, not significant). (F) Three EBV(+) and 3 EBV(−) BL cell lines were cultured and analyzed by immunoblotting for KDM2B expression levels. The histogram shows the average KDM2B expression levels normalized to the β-actin signal, measured in 4 independent experiments by Image Lab software (Bio-Rad) in EBV(+) versus EBV(−) BL cells (*, $P < 0.05$). (G) KDM2B levels in EBV(+) and EBV(−) BL samples were analyzed by immunohistochemistry. The same samples were analyzed for EBER expression by ISH for EBER (EBER-ISH) as described in Materials and Methods. The images shown are representative of the KDM2B staining obtained in 11 EBV(+) and 11 EBV(−) BL specimens.

compared with EBV(−) BL-derived cells. Here, to validate these data we performed direct pyrosequencing on DNA extracted from 10 EBV(+) BL-derived cell lines and 9 EBV(−) BL-derived cell lines (Table 1). The samples for which the pyrosequencing gave results technically suitable for analysis are displayed in the histogram in Fig. 1B. Pyrosequencing analysis confirmed that the *KDM2B* gene is hypermethylated at CpG15695155 and CpG21423404 in EBV(+) BL cell lines compared with EBV(−) BL cell lines (Fig. 1B). In contrast, we did not observe high methylation levels or differences between EBV(+) and EBV(−) BL cell lines when analyzing 17 positions within the CpG island 127 (Fig. 1C). Next, we assessed whether the high DNA methylation level of the *KDM2B* gene would affect its expression level. Treatment of 3 EBV(−) BL and 3 EBV(+) BL cell lines with the demethylating agent 5-aza-2′-deoxycytidine (Aza) for 48 h led to a significant rescue of KDM2B expression in EBV(+) BL cells, whereas this treatment had no noticeable effect on KDM2B mRNA expression in EBV(−) BL cells (Fig. 1D). Pyrosequencing analysis of DNA from EBV(+) and EBV(−) BL cell lines exposed to Aza or to

TABLE 1 Description of BL-derived cell lines used in the present study[a]

BL case identifier	Diagnosis	EBV infection	Cytogenetic information	Clinical data	Ethnic origin
BL103	BL	EBV(−)	t(8;14)		Caucasian
BL70	BL	EBV(−)	t(8;14)	Dec	Caucasian
BL56	BL	EBV(−)	t(8;14)	Dec	
BL58	BL	EBV(−)	t(8;14)	Dec	Caucasian
BL53	BL	EBV(−)	t(8;14)	Dec	Caucasian
BL102	BL	EBV(−)	t(8;22)	Dec	Caucasian
BL104	BL	EBV(−)	t(8;22)		Caucasian
BL2	BL	EBV(−)	t(8;22)	Dec	Caucasian
BL41	BL	EBV(−)	t(8;14)		Caucasian
BL110	BL	EBV(+)	t(8;14)		Caucasian
BL135	BL	EBV(+)	t(8;14)		African
BL65	BL	EBV(+)	t(8;14)		African
BL116	BL	EBV(+)	t(8;14)		African
BL60	BL	EBV(+)	t(8;22)	Dec	African
BL79	BL	EBV(+)	t(8;14)		African
BL112	BL	EBV(+)	t(8;14)		Caucasian
I100		EBV(+)			
I373		EBV(+)			
I176B		EBV(+)			

[a]The cells were obtained from the IARC Biobank. Abbreviations: BL, Burkitt lymphoma; EBV, Epstein-Barr virus; Dec, deceased.

dimethyl sulfoxide (DMSO) for 48 h revealed a moderate but significant reduction in the methylation level at CpG21423404 in Aza-exposed EBV(+) BL cells, whereas methylation at all the other positions analyzed remained unchanged (Fig. 1E). We then determined whether the different methylation patterns observed in EBV(+) and EBV(−) BL cells affected KDM2B protein expression. We analyzed KDM2B protein expression in 3 EBV(+) and 3 EBV(−) BL-derived cell lines and observed a significantly lower expression level of KDM2B protein in EBV-infected cells (Fig. 1F). Moreover, we analyzed 11 EBV(+) BL and 11 EBV(−) BL samples (Table 2) by immunohistochemistry for the KDM2B protein expression level; 8 of the 11 EBV(+) BL samples showed weak KDM2B staining, and 10 of the 11 EBV(−) BL samples showed a strong signal for KDM2B immunohistochemistry, suggesting that EBV infection induces reduced expression of

TABLE 2 BL case main features[a]

				Detection of the following:			
BL subtype	Age (yr)	Sex	Site of biopsy	EBER	*MYC* by FISH B.A.	*MYC-IGH* by FISH	*MYC-IGK* or *IGL* by FISH
sBL	40	F	Inguinal lymph node	−	+	n.p.	n.p.
sBL	18	M	Ileum	−	+	n.p.	n.p.
sBL	38	F	Lymph node	−	+	n.p.	n.p.
sBL	40	F	Bone marrow	−	+	n.p.	n.p.
sBL	20	M	Lymph node	−	−	−	n.p.
sBL	15	M	Lymph node	−	+	n.p.	n.p.
sBL	25	F	Ileum	−	+	n.p.	n.p.
sBL	45	M	Stomach	−	+	+	n.p.
sBL	16	M	Abdomen mass	−	+	n.p.	n.p.
sBL	20	M	Lymph node	−	−	−	n.p.
sBL	14	M	Lymph node	+	+	n.p.	n.p.
eBL	9	F	Abdomen mass	+	+	n.p.	n.p.
eBL	5	F	Soft tissues	+	+	n.p.	n.p.
eBL	6	F	Maxilla	+	+	n.p.	n.p.
eBL	10	F	Ileum	+	+	n.p.	n.p.
eBL	3	F	Lymph node	+	+	n.p.	n.p.
eBL	4	F	Lymph node	+	+	+	n.p.
eBL	10	M	Maxilla	+	+	−	+ (*IGL*), − (*IGK*)
eBL	7	M	Abdomen	−	+	n.p.	n.p.
eBL	12	F	Maxilla	−	+	n.p.	n.p.
eBL	3	M	Stomach	−	+	n.p.	n.p.
eBL	5	M	Ileum	−	+	n.p.	n.p.

[a]Abbreviations: F, female; M, male; n.p., not performed; EBER, Epstein-Barr virus-encoded small RNA; B.A., break apart; eBL, endemic Burkitt lymphoma; sBL, sporadic Burkitt lymphoma, FISH, fluorescence *in situ* hybridization.

the KDM2B protein *in vivo* (Fig. 1G). Of note, the EBV(+) samples with stronger KDM2B staining had fewer Epstein-Barr virus-encoded small RNA (EBER)-expressing cells, as determined by EBER *in situ* hybridization (ISH) (data not shown). In conclusion, these data show that two specific CpG sites in the regulatory region of *KDM2B* are hypermethylated in EBV(+) BL cell lines compared with EBV(−) BL cell lines, confirming our previous whole-genome methylation profiling data (4). Moreover, KDM2B expression also appears to be reduced in eBL specimens, which is probably mediated by DNA methylation at CpG21423404. These data suggest that EBV may regulate KDM2B expression by inducing the methylation of a specific position within a regulatory region of its gene.

EBV infection of B cells downregulates KDM2B expression. To assess the ability of EBV to deregulate KDM2B expression, primary B cells isolated from 3 independent donors were infected with EBV. Cells were collected at different times after infection and analyzed for the expression level of KDM2B mRNA by reverse transcription-quantitative PCR (RT-qPCR). It appeared that soon after infection, the expression level of KDM2B was drastically reduced (Fig. 2A). This downregulation was also seen at the protein level (Fig. 2B). The reduced expression of KDM2B was maintained during the immortalization process and stayed low in lymphoblastoid cell lines (LCL). This result suggests that EBV infection plays a role in the regulation of KDM2B mRNA expression. To exclude the possibility that this result could be due to the activation of the primary B cells independently of EBV infection, as a consequence of the engagement of the membrane B-cell receptors by the virus, we stimulated primary B cells from 2 independent healthy donors with CD40 ligand (CD40L) and interleukin 4 (IL-4) with or without EBV. Analysis of the EBNA1 expression by RT-qPCR showed that the infection worked efficiently (Fig. 2C). Both treatment with CD40L and IL-4 and infection with EBV similarly activated the B cells, as shown by the strong induction of expression of CCL22 (Fig. 2D), a cytokine that is known to be produced during primary B-cell activation and EBV infection (11). Importantly, significant downregulation of KDM2B mRNA expression was observed only when the cells were infected with EBV (Fig. 2E). Moreover, infecting the primary B cells with UV-inactivated EBV led to reduced downregulation of KDM2B compared with that in the cells infected with the untreated virus (Fig. 2F), further indicating that this event is independent of B-cell activation and requires active expression of the viral genes. To assess whether downregulation of KDM2B depends on differences between the proliferation status of the cells, we measured KDM2B expression in Akata2000 cells, an EBV(+) BL-derived cell line, and in Akata31 cells, which were derived by expansion of a clone of Akata2000 cells that had lost the virus. Analysis of the viral DNA in both cell lines confirmed that Akata31 cells had few or no EBV genomes compared with Akata2000 cells (Fig. 2G). Interestingly, Akata2000 cells displayed very low levels of KDM2B mRNA compared with Akata31 cells, indicating that KDM2B expression varies according to the amount of virus (Fig. 2G) and independently of the cell type. Similarly, reactivation of EBV by exposing Raji cells to 12-*O*-tetradecanoylphorbol-13-acetate (TPA) and sodium butyrate (NaB) treatment led to reduced expression of KDM2B (Fig. 2H). Finally, at 48 h after EBV infection, KDM2B mRNA and protein expression levels were reduced in RPMI-8226 (RPMI) cells (Fig. 2I), confirming the results obtained in EBV-infected primary B cells and further proving that downregulation of KDM2B expression is a direct effect of the virus and not a side effect of the activation of the cells or of the immortalization process.

To determine whether the reduced expression of KDM2B in EBV-immortalized cells could be due to an increase in DNA methylation, LCL were treated with the demethylating agent Aza and analyzed by RT-qPCR for the KDM2B expression level. Aza-treated LCL showed increased expression of KDM2B mRNA compared with their untreated counterparts, indicating that downregulation of KDM2B in EBV-infected cells is mediated by DNA methylation (Fig. 3A). Because incorporation of Aza into DNA impedes its methylation by DNA methyltransferases (DNMTs) (12), we hypothesized that EBV could contribute to KDM2B silencing by increasing the recruitment of DNMT1 to the *KDM2B*

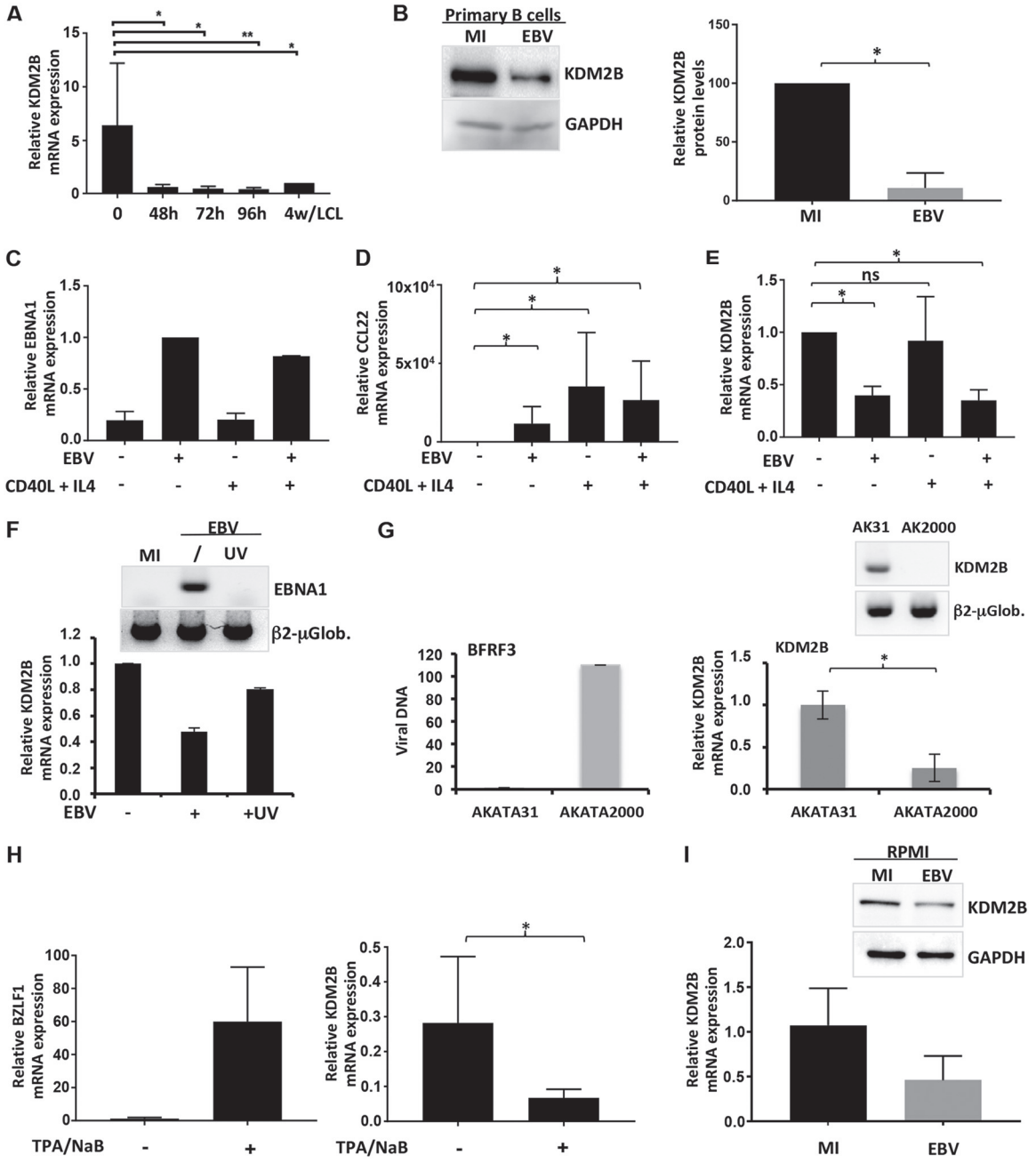


FIG 2 EBV-dependent silencing of KDM2B expression *in vitro*. (A) Primary B cells from 3 independent donors were infected with EBV and collected to make dry pellets at the indicated time points. Some of the infected cells were left in culture for 4 weeks (4w) to generate LCL. Cell pellets were processed and analyzed for the expression level of KDM2B mRNA by RT-qPCR. The difference between the levels of KDM2B in primary B cells (time point 0) and in EBV-infected cells was significant (*, $P < 0.05$; **, $P < 0.01$). The levels of KDM2B were measured relative to its levels in LCL (which was used as the calibrator, for which the value was 1). (B) KDM2B protein expression levels in primary B cells mock infected (MI) or infected with EBV for 48 h were analyzed by Western blotting (left). The KDM2B protein signal was normalized to the levels of GAPDH. The histogram (right) shows the average from 3 independent experiments. (C to E) Primary B cells from 2 independent healthy donors were activated by treating them for 24 h with CD40L and IL-4, as described in Materials and Methods, and/or infected with EBV for 48 h. Cells were then processed and analyzed for the expression levels of EBNA1 (C), CCL22 (D), and KDM2B (E) mRNA by RT-qPCR. The results shown in the histogram are the average from 2 independent experiments (*, $P < 0.05$; ns, not significant). EBNA1 mRNA levels were measured relative to its levels in EBV-infected untreated cells. CCL22 and KDM2B mRNA levels were measured relative to their levels in mock-infected untreated cells. (F) Primary B cells were infected with 2 aliquots of the same EBV batch, one of which was UV inactivated before infection. At 48 h after infection, cells were collected and EBNA1 and KDM2B expression levels were assessed by RT-PCR and qPCR, respectively. The levels of KDM2B mRNA were measured relative to its levels in mock-infected cells. (G) Akata2000 and Akata31 cells were collected, processed for DNA/RNA extraction, and analyzed for the presence of the EBV genome by PCR (left) as well as for KDM2B expression levels by RT-qPCR (lower right) or RT-PCR (upper right). Viral DNA (left) indicates the relative amount of the *BFRF3* gene present in Akata2000 versus Akata31 cells that was analyzed by real-time PCR and normalized to the amount of GAPDH in the extracted DNA. The levels of KDM2B mRNA were measured relative to its levels in mock-infected cells (*, $P < 0.05$). (H)

(Continued on next page)

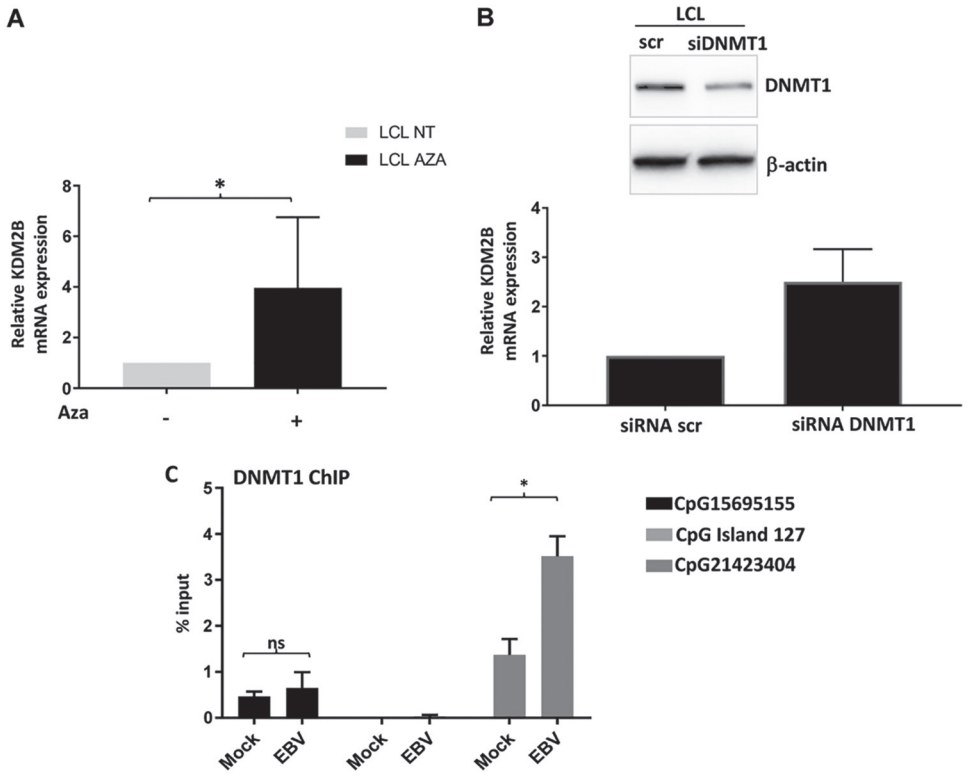


FIG 3 Methylation-dependent silencing of KDM2B expression. (A) LCL were cultured for 96 h in the presence of DMSO (nontreated [NT]) or 5-aza-2′-deoxycytidine (Aza). Cells were then collected, and the KDM2B expression levels were analyzed by RT-qPCR. The histogram shows the average KDM2B mRNA levels measured in 3 independent experiments (*, $P < 0.05$). KDM2B mRNA levels in Aza-treated LCL were measured relative to its levels in DMSO-treated cells. (B) LCL were transfected with stabilized siRNA targeting DNMT1 (siDNMT1) in two independent experiments. At 4 days after transfection, cells were collected and analyzed for the protein levels of DNMT1 (top) and the mRNA levels of KDM2B (bottom). KDM2B mRNA levels in DNMT1-targeted siRNA-treated cells were measured relative to its levels in cells treated with scrambled (scr) siRNA. (C) RPMI cells infected with EBV or mock infected were fixed and processed for ChIP with a DNMT1 antibody or an IgG antibody as a negative control. The eluted DNA was analyzed by qPCR with primers flanking CpG15695155, CpG21423404, and CpG island 127 (primer sequences are described in Table 3) (*, $P < 0.05$; ns, not significant).

gene. Indeed, depletion of DNMT1 in LCL by transfection of DNMT1 small interfering RNA (siRNA) led to an increased KDM2B mRNA level (Fig. 3B). Moreover, chromatin immunoprecipitation (ChIP) experiments showed increased recruitment of DNMT1 to the CpG site CpG21423404 of the *KDM2B* gene in EBV-infected RPMI cells (Fig. 3C).

Taken together, these results indicate that infection of B cells with EBV leads to reduced expression of KDM2B, mediated by the recruitment of DNMT1 to its gene promoter.

The oncogenic viral protein LMP1 induces the silencing of KDM2B. Next, we assessed whether LMP1, the main EBV oncoprotein, plays a role in the deregulation of KDM2B expression. To do this, we generated RPMI cells stably expressing LMP1 (RPMI-LMP1 cells). As a negative control, the cells were transduced with the empty retroviral vector (pLNSX) (Fig. 4A). As revealed by RT-qPCR analysis, the expression level of KDM2B was reduced in the RPMI cells expressing LMP1 (Fig. 4A). Thus, LMP1 appears to play a role in the EBV-mediated downregulation of KDM2B expression. In addition, treating RPMI-LMP1 cells with Aza led to a rescue of KDM2B mRNA levels (Fig. 4B). In

FIG 2 Legend (Continued)
Raji cells untreated or treated with TPA (50 ng/ml)-NaB (3 mM) for 48 h were analyzed for expression levels of BZLF1 and KDM2B mRNA by RT-qPCR (*, $P < 0.05$). (I) RPMI cells untreated or infected with EBV were collected, processed for RNA and protein extraction, and analyzed for the levels of the KDM2B protein (Western blotting, top) and mRNA (RT-qPCR, bottom). The levels of KDM2B mRNA were measured relative to its levels in mock-infected cells.

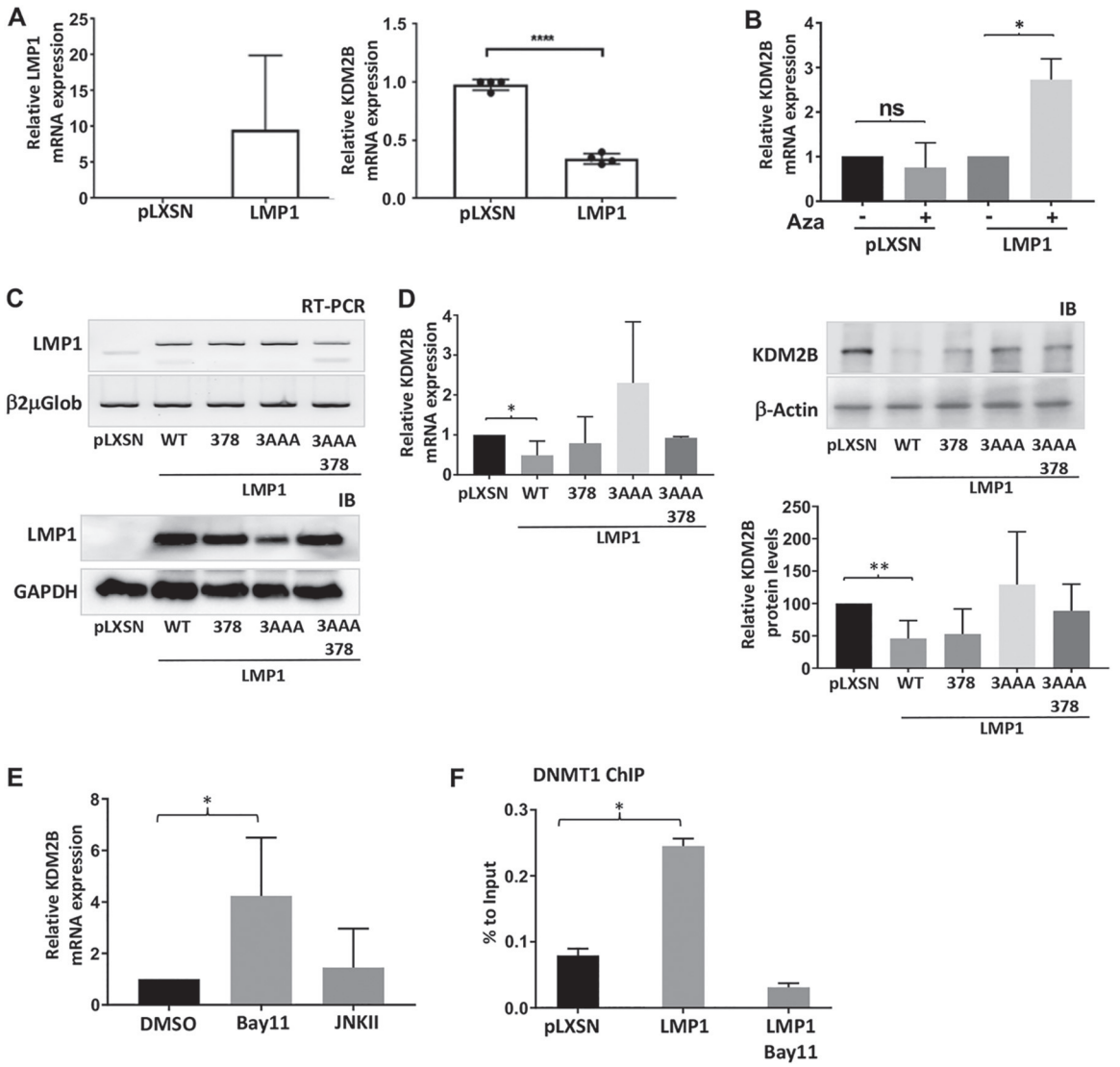


FIG 4 LMP1 mediates downregulation of KDM2B. (A) RPMI cells were stably transduced with pLXSN or with pLXSN-LMP1 (LMP1) in four independent transduction experiments. Cells were then collected, and the expression levels of LMP1 (left) and KDM2B (right) were analyzed by RT-qPCR (****, $P < 0.0001$). (B) RPMI-pLXSN or RPMI-LMP1 cells were cultured in the presence of Aza (+) or DMSO (−) for 48 h, and the KDM2B mRNA expression level was analyzed by RT-qPCR. The histogram shows the average from 2 independent experiments (*, $P < 0.05$; ns, not significant). (C and D) Louckes cells were stably transduced with pLXSN, pLXSN-LMP1 (LMP1), or pLXSN-LMP1 mutants (3AAA, 378, and 3AAA/378). Cells were collected and processed for RNA and protein analysis. mRNA expression and protein levels of LMP1 were detected by RT-PCR (C, top) and immunoblotting (C, bottom). KDM2B mRNA and protein levels were also shown by RT-qPCR and immunoblotting, respectively (D, left and right). In panel D, the difference between KDM2B mRNA and protein levels in Louckes cells with pLXSN and in Louckes cells stably expressing wild-type LMP1 was significant (*, $P < 0.05$; **, $P < 0.01$). The levels of KDM2B mRNA in cells expressing wild-type (WT) LMP1 and LMP1 mutants were measured relative to its level in cells expressing pLXSN. (E) RPMI-LMP1 cells were treated for 2 h with BAY11-7082 (Bay11) (10 μM) or with JNK inhibitor II (JNKII) (10 μM) in three independent experiments. KDM2B mRNA levels were analyzed by RT-qPCR (*, $P < 0.05$). (F) RPMI pLXSN and RPMI-LMP1 cells, the latter of which were untreated or treated with BAY11-7082 (10 μM) for 2 h, were fixed to perform ChIP with DNMT1 or IgG antibodies. The eluted DNA was analyzed by qPCR with primers designed to surround CpG21423404. The histogram shows the average percentage of recruitment of DNMT1 in 4 independent experiments (*, $P < 0.05$).

contrast, no change in KDM2B mRNA expression was observed in the control cells after treatment with Aza (Fig. 4B).

LMP1 activates both the NF-κB and Jun N-terminal protein kinase (JNK) pathways through its CTAR1 and CTAR2 domains, respectively. Therefore, to gain insights into the mechanism by which LMP1 deregulates KDM2B expression, we generated RPMI and Louckes cells expressing LMP1 mutants harboring mutations in CTAR1 (3AAA mutant), CTAR2 (378 mutant), or both (3AAA/378 double mutant) and therefore having a

hampered ability to activate the NF-κB pathway, the JNK pathway, or both (13). After selection, cells were analyzed for LMP1 expression by RT-qPCR and immunoblotting. All the cells generated expressed similar levels of the LMP1 transcript and protein (Fig. 4C), with the exception of the LMP1 3AAA mutant, which appeared to be present at lower protein levels than the wild-type LMP1 and the other LMP1 mutated molecules. This is probably due to a reduced protein stability of the LMP1 3AAA mutant caused by the mutation in the CTAR1 in the context of a wild-type CTAR2 domain; however, the 3AAA/378 double mutant and the 378 mutant showed mRNA and protein levels similar to those of wild-type LMP1. We then compared the ability of the different LMP1 mutants to deregulate KDM2B expression. Whereas wild-type LMP1 efficiently downregulated KDM2B mRNA and protein expression (Fig. 4D), both the 3AAA mutant and the double mutant were unable to downregulate KDM2B mRNA expression, and the 378 mutant partially maintained the ability to inhibit KDM2B expression (Fig. 4D), indicating that the CTAR1 domain and, in part, the CTAR2 domain play a role in the LMP1-mediated regulation process of KDM2B expression. The same results were observed in RPMI cells stably expressing LMP1 and its CTAR mutants (data not shown). To further evaluate the impact of the NF-κB and JNK pathways on LMP1-mediated KDM2B deregulation, we treated RPMI-LMP1 cells with specific chemical inhibitors: BAY11-7082 and JNK inhibitor II, respectively. Whereas the JNK inhibitor had no effect, treating RPMI-LMP1 cells with the NF-κB inhibitor BAY11-7082 led to a rescue of KDM2B mRNA expression (Fig. 4E).

Previous studies showed that LMP1 activates and induces the recruitment of DNMT proteins to the promoters of cancer-related genes (14–16). In line with these findings, our ChIP experiments showed an increase in the recruitment of DNMT1 to the *KDM2B* gene at the CpG21423404 position in LMP1-expressing cells compared with control cells (Fig. 4F). Treating RPMI-LMP1 cells with BAY11-7082 significantly reduced the amount of DNMT1 recruited to the *KDM2B* gene. Taken together, these data indicate that EBV induces silencing of KDM2B expression mainly via the ability of its main transforming protein, LMP1, to activate the NF-κB signaling pathway.

LMP1 is not the only EBV protein able to induce downregulation of KDM2B expression. LMP1 is only rarely expressed in EBV(+) BL samples. Therefore, our observation that KDM2B is downregulated and silenced in EBV(+) BL cell lines and specimens led us to investigate whether EBV genes other than LMP1 could play a role in this event. Infection of primary B cells with a recombinant EBV lacking the *LMP1* gene (EBVΔLMP1) still led to decreased expression of KDM2B mRNA (Fig. 5A). Downregulation of KDM2B occurred at 12 h after infection of primary B cells with EBV; as expected at this early time point, we could not yet detect LMP1 expression, and only the EBNAs were efficiently expressed (Fig. 5B). Moreover, when comparing KDM2B mRNA expression levels in different EBV(+) BL-derived cell lines, in an LCL, and in primary B cells, we observed a reduced level of KDM2B transcript in all the BL cells, independently of whether they were in latency phase I or III (Fig. 5C). As expected, although BL cells in phase I expressed EBNA1 but had little or no LMP1 expression, BL cells in phase III and the LCL efficiently expressed both genes (Fig. 5C). Taken together, these experiments indicate that KDM2B deregulation can also occur in the absence of LMP1 expression; therefore, other EBV proteins could be involved in this event in BL cells. Transient transfection of Louckes and RPMI cells with different constructs expressing a panel of EBV genes (Fig. 5D and E) indicated that, in addition to LMP1, different latent viral proteins could downregulate KDM2B mRNA. To rule out the possibility that this result was the consequence of the activation of the NF-κB pathway as a side effect of the transfection of the cells, we checked the levels of IκBα phosphorylation. As expected, based on its known function (17), among the analyzed EBV proteins, LMP1 was the most efficient in inducing NF-κB activation (Fig. 5F), indicating that the other viral proteins may use other mechanisms to downregulate KDM2B mRNA expression. MicroRNA-146-5p (miRNA-146-5p), which is known to target KDM2B expression, has been reported to deregulate KDM2B mRNA levels during human papillomavirus (HPV) infection (18). A previous study showed that EBNA2 induces the expression of miRNA-

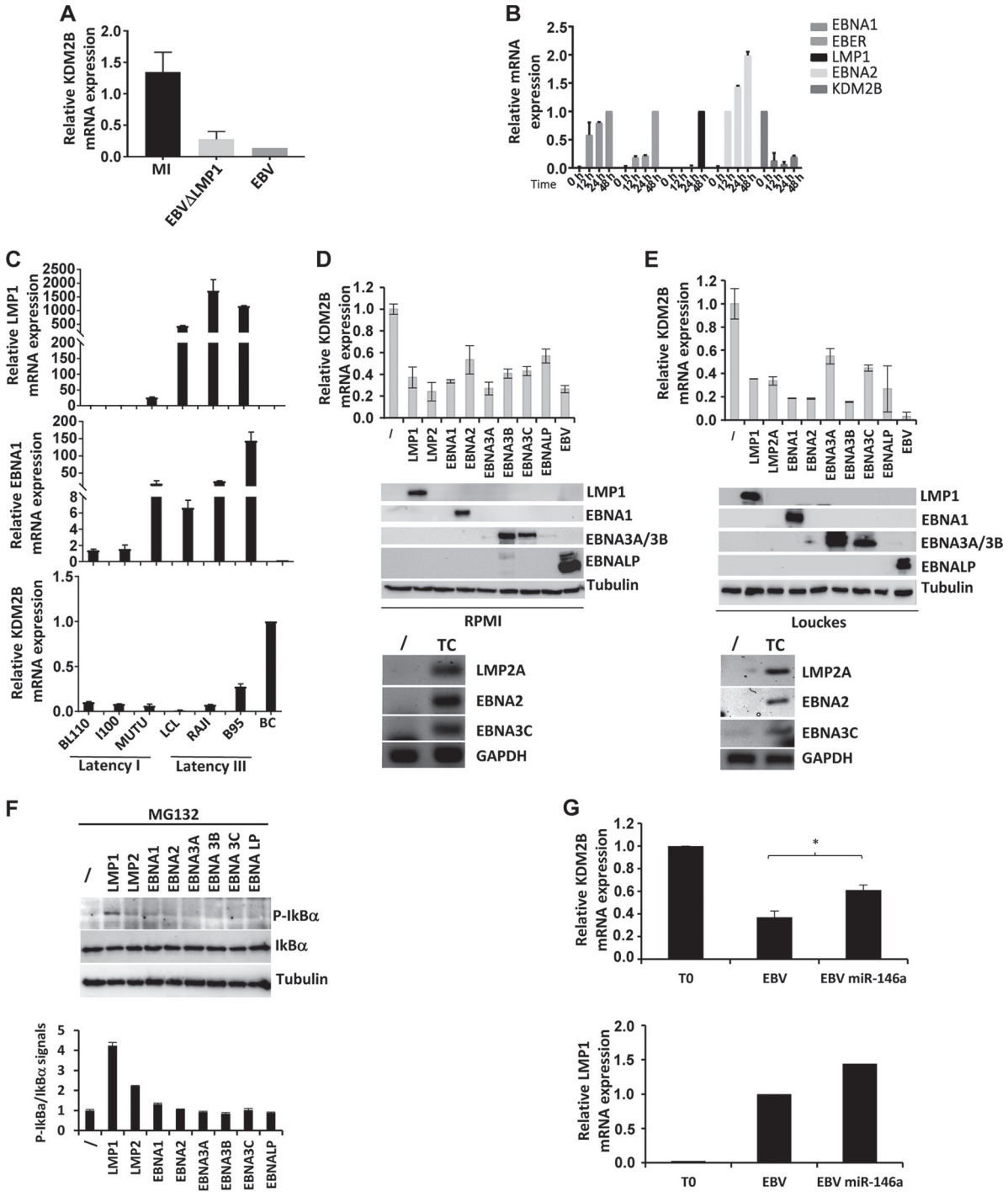


FIG 5 EBV also uses LMP1-independent mechanisms to downregulate KDM2B. (A) B cells from two donors were infected with EBV or EBVΔLMP1 or mock infected (MI) and collected at 48 h after infection. Reverse-transcribed RNA samples were analyzed by qPCR for the KDM2B expression level. (B) B cells from two donors were infected with EBV and collected at 12, 24, and 48 h after infection. Cells were processed for RNA extraction and analyzed by qPCR for the expression levels of EBNA1, EBNA2, EBER, LMP1, and KDM2B transcripts. The expression levels of viral gene transcripts were measured relative to their levels in B cells collected 12 h (EBNA2) or 48 h (EBNA1, EBER, LMP1) postinfection. The levels of KDM2B mRNA were measured relative to its levels in mock-infected cells. (C) Different EBV(+) cell lines in latency phase I (BL110, I100, MUTU) or in latency phase III (LCL, Raji, B95) and primary B cells (BC) were collected, processed for RNA extraction and reverse transcription, and analyzed by qPCR for LMP1 (top), EBNA1 (middle), and KDM2B (bottom) mRNA levels. (D and E) RPMI (D) and Louckes (E) cells were transiently transfected with different constructs carrying individual EBV genes (TC) in three independent experiments. At 48 h after transfection, cells were processed for RNA and protein extraction and analyzed for the KDM2B expression level by RT-qPCR analysis. The levels of KDM2B mRNA were measured relative to its levels in mock-infected cells. Western blotting (LMP1, EBNA1, EBNA3A/3B, and EBNALP) or RT-PCR analysis (LMP2A, EBNA2, and EBNA3C) was performed to measure the expression of the different viral latent proteins. (F) RPMI cells were transfected with different EBV gene-carrying constructs. At 44 h after transfection, cells were exposed to the proteasome inhibitor MG132 for 4 h and then collected, processed for total protein extraction, and analyzed for the indicated proteins by immunoblotting. The histogram

(Continued on next page)

146-5p (19). Therefore, we tested whether miRNA-146-5p could contribute to KDM2B downregulation during EBV infection. To do this, we treated primary B cells with a miRNA-146-p5-specific inhibitor before EBV infection. As shown in Fig. 5G, this treatment partially rescued KDM2B expression in the miRNA-146-p5 inhibitor-treated cells. These results show that EBV may be able to reduce intracellular levels of KDM2B by a redundancy of mechanisms, further indicating that this event could be important for the virus.

KDM2B regulates expression of viral genes in EBV-infected B cells. To evaluate the biological relevance of EBV-mediated KDM2B downregulation, we hypothesized that the epigenetic enzyme could regulate EBV transcription, similar to what was observed for other histone modifiers and chromatin-interacting proteins (e.g., EZH2, CTCF, KMT5B) (20–22). To assess whether ectopic expression of KDM2B in B cells could have an impact on EBV infection, we transfected increasing concentrations of a KDM2B construct into Louckes cells, an EBV(−) BL-derived cell line. One day after transfection, cells were infected with EBV-green fluorescent protein (GFP) and monitored for the infection efficiency 24 h later. Figure 6A shows efficient overexpression of ectopic KDM2B at both the protein and mRNA levels at 24 h after transfection. The infection efficiency of Louckes cells transfected with the KDM2B expression vector or the control empty vector (pCDNA) was indistinguishable, as revealed by TaqMan PCR, showing that the increased KDM2B expression level did not alter the viral genome copy number per cell (Fig. 6B). In contrast, the GFP mean fluorescence intensity (MFI) decreased in the presence of enhanced KDM2B expression (Fig. 6C), indicating that KDM2B could affect viral gene expression. Indeed, RT-qPCR analysis of the expression levels of different EBV transcripts (LMP1, EBNA1, and BZLF1) at 24 h after infection showed a significant and dose-dependent reduction in their mRNA levels in the presence of an increasing amount of KDM2B (Fig. 6D). Taken together, these data suggest that KDM2B plays a role in regulating EBV gene expression.

The results presented above, obtained by overexpressing KDM2B, could be due to a generalized chromatin demethylation as a consequence of high intracellular levels of the demethylase. Therefore, next we depleted endogenous KDM2B in Louckes cells by transfecting an siRNA targeting KDM2B mRNA. RT-qPCR and Western blot analysis showed that KDM2B was efficiently downregulated 24 h after transfection with the specific siRNA (Fig. 6E). Cells were then infected with EBV-GFP and monitored for the infection efficiency, as described above. At 24 h after infection, neither the percentage of GFP-positive cells nor the genome copy number had changed significantly between the control cells (cells transfected with scrambled siRNA [siSCR]) and the cells transfected with the siRNA directed against KDM2B (Fig. 6F and data not shown), indicating that the loss of KDM2B does not affect the efficiency of infection. However, efficient depletion of KDM2B by siRNA led to a significant increase in EBV transcripts 24 h after infection (Fig. 6G). This result indicates that a reduced intracellular level of KDM2B promotes the expression of the viral genes, further confirming the ability of KDM2B to repress EBV gene expression at early stages of infection.

KDM2B inhibits viral gene expression in latently infected EBV-immortalized B cells. To test whether the activity of KDM2B in regulating EBV gene expression is required for the maintenance of virus latency, similar to what has been reported for other epigenetic enzymes (20), we aimed to overexpress KDM2B ectopically and examine its impact on viral gene expression in EBV-immortalized B cells. LCL displayed detectable KDM2B mRNA and protein levels, although they were lower than those in primary B cells (Fig. 2A and B and data not shown). Therefore, KDM2B was overex-

FIG 5 Legend (Continued)
shows the average phosphorylated IκBα (P-IκBα) signal normalized to the levels of total IκBα. (G) B cells from different donors were untreated or treated with a miRNA-146a-5p inhibitor for 24 h before infection with EBV. At 48 h after infection, cells were collected and processed for RNA extraction and reverse transcription. cDNA samples were analyzed by qPCR for LMP1 and KDM2B expression levels. The levels of KDM2B mRNA were measured relative to its levels in mock-infected cells. The histograms show the average expression levels measured in 2 independent experiments (*, $P < 0.05$). T0, time zero.

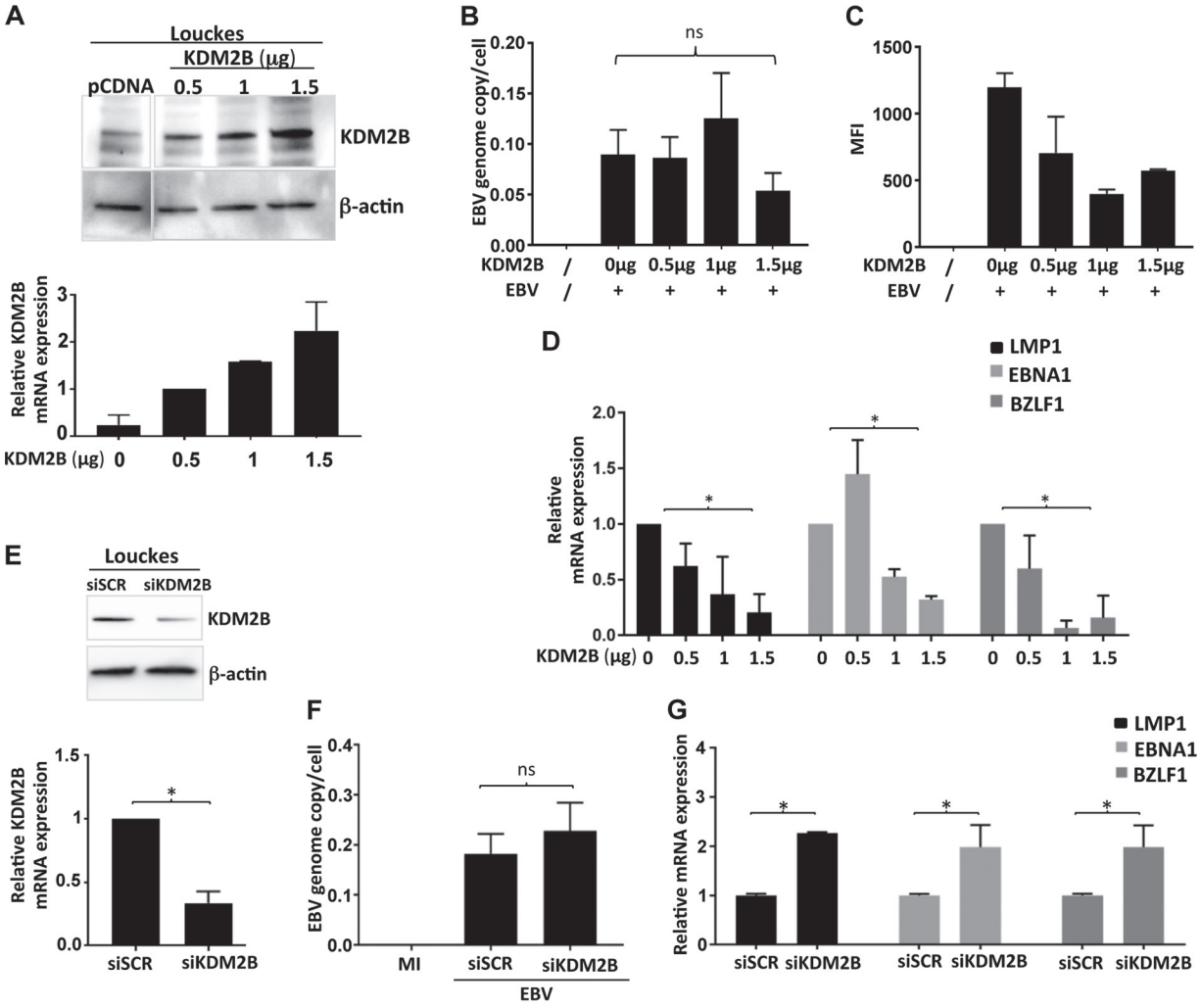


FIG 6 KDM2B deregulation alters EBV gene expression. (A) Louckes cells were transiently transfected with increasing concentrations of the KDM2B expression vector in three independent experiments, collected at 24 h after transfection, and processed for protein and RNA extraction to assess KDM2B levels by immunoblotting (top) or qPCR (bottom). (B) Cells were then infected with EBV, and at 24 h after infection they were collected and processed for FACS analysis and for RNA/DNA extraction. DNA samples were used to measure EBV genome copy number by TaqMan PCR (ns, not significant). (C) Live cells were analyzed by FACS to measure the GFP mean fluorescence intensity (MFI). (D) cDNA samples were analyzed for the expression level of EBV early and late genes by qPCR. The levels of EBV genes transcripts were normalized to the mRNA levels of the housekeeping gene β-globin and calculated relative to their levels in cells transfected with the empty vector (KDM2B, 0 μg). The values shown in the histogram are the average from 3 independent experiments (*, $P < 0.05$). (E to G) Louckes cells were transfected with KDM2B siRNA and scrambled siRNA as a control in three independent experiments. (E) At 24 h after transfection, half of the cells were collected and analyzed for the expression levels of KDM2B protein (top) and mRNA (bottom), and half of the cells were infected with EBV. At 24 h after infection, cells were collected and processed for RNA/DNA extraction. The number of EBV genome copies per cell was determined by TaqMan PCR on the DNA template (MI, mock infected; Louckes cell DNA was also included as a negative control) (F), and qPCR of the cDNA samples enabled the assessment of the mRNA expression levels of different viral genes, calculated as explained in the legend to panel D (G). siSCR, scrambled siRNA; siKDM2B, siRNA targeting KDM2B mRNA.

pressed in LCL by ectopic expression of a KDM2B construct. At 48 h after transfection, LCL were collected and processed for total protein and DNA/RNA extraction. Western blotting and RT-qPCR showed substantial increases of KDM2B at the protein and mRNA levels, respectively (Fig. 7A). LCL overexpressing KDM2B carried a similar number of EBV genome copies compared with the same cells transfected with the control pCDNA vector (Fig. 7B), as revealed by TaqMan PCR analysis. We then analyzed the mRNA expression level of different EBV transcripts by RT-qPCR. Ectopic expression of KDM2B led to a significant reduction in the mRNA level of all the analyzed EBV genes (Fig. 7C). In contrast, depletion of KDM2B from LCL by transfecting KDM2B siRNA (Fig. 7D) led to a significant increase in the expression of the viral genes compared with the expression level of the same genes in LCL transfected with scrambled siRNA (Fig. 7F). Similar to

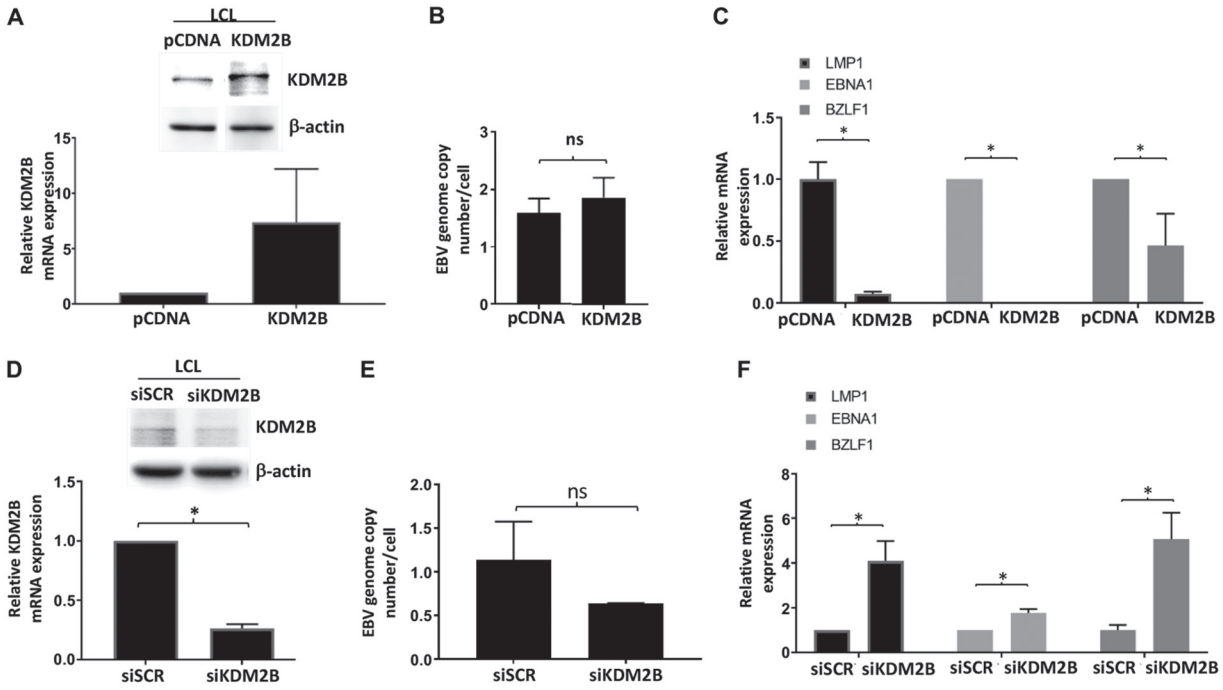


FIG 7 KDM2B regulates EBV gene expression in latently infected cells. (A to C) LCL were transfected with 1.5 μg of pCDNA3-KDM2B or with pCDNA as a control in three independent experiments. Cells were collected at 24 h after transfection and processed for RNA/DNA and total protein extraction. (A) KDM2B mRNA and protein levels were shown by qPCR (bottom) and immunoblotting (top; the lines corresponding to conditions with 0.5 μg and 1 μg of KDM2B, originally present in the Western blot, are not shown because they were excluded from all the analyses). (B) DNA samples were analyzed by TaqMan PCR to assess the number of EBV genome copies per cell (ns, not significant). (C) The mRNA expression levels of EBV latent and early genes were assessed by qPCR (*, $P < 0.05$). The mRNA levels of viral genes in LCL overexpressing KDM2B were measured relative to their levels in pCDNA-transfected control cells and normalized on the mRNA levels of β-globin. (D to F) (D) LCL transfected in three independent experiments with KDM2B siRNA (siKDM2B) or scrambled siRNA (siSCR) as a control were collected at 48 h after transfection and analyzed for the levels of KDM2B protein (top) and transcript (bottom). (E) DNA samples were analyzed by TaqMan PCR to assess the number of EBV genome copies per cell. (F) The mRNA expression levels of EBV latent and early genes were assessed by qPCR (*, $P < 0.05$). Viral transcript levels in KDM2B siRNA-treated cells were measured relative to their levels in siSCR-treated cells and normalized to the mRNA levels of β-globin.

what we observed in EBV-infected Louckes cells, removal of KDM2B did not significantly affect the EBV genome copy number (Fig. 7E). Taken together, these data indicate that KDM2B controls viral gene expression in latently infected cells.

Next, we determined whether KDM2B could directly bind to EBV gene promoters to regulate their expression. To do this, purified primary B cells infected or mock infected with EBV for 48 h, as well as the corresponding LCL, were formaldehyde fixed and processed for KDM2B ChIP (Fig. 8A). ChIP analysis showed that KDM2B can be recruited to the EBV genome and, more precisely, to the Qp promoter at 48 h after infection. We also found KDM2B recruited to the Cp promoter and to a region of the BZLF1 promoter proximal to the start site (Zp 0) in EBV-immortalized B cells. However, we did not detect KDM2B recruitment when we analyzed a region of the BZLF1 promoter 600 bp upstream of the start site (Zp +600) in the same cells (Fig. 8A, top), indicating that KDM2B is specifically recruited to a specific region of the viral genome. As expected, we did not detect KDM2B recruitment to the promoter of the housekeeping gene GAPDH (glyceraldehyde-3-phosphate dehydrogenase); however, we observed efficient KDM2B binding to a sequence located downstream of the transcription start site of the ribosomal DNA (rDNA) repeated units 8 (H8) (Fig. 8A, bottom), which was previously reported to be targeted by the epigenetic enzyme (23). These data indicate that KDM2B is directly recruited to specific EBV promoter regions. To further assess the ability of KDM2B to be recruited to the EBV genome, B cells were untreated or EBV infected, collected, and processed for immunofluorescent *in situ* hybridization (immuno-FISH) experiments. EBV DNA molecules were detected by FISH with a labeled probe directed against the BamHI W EBV genomic repeated region; KDM2B was concomitantly de-

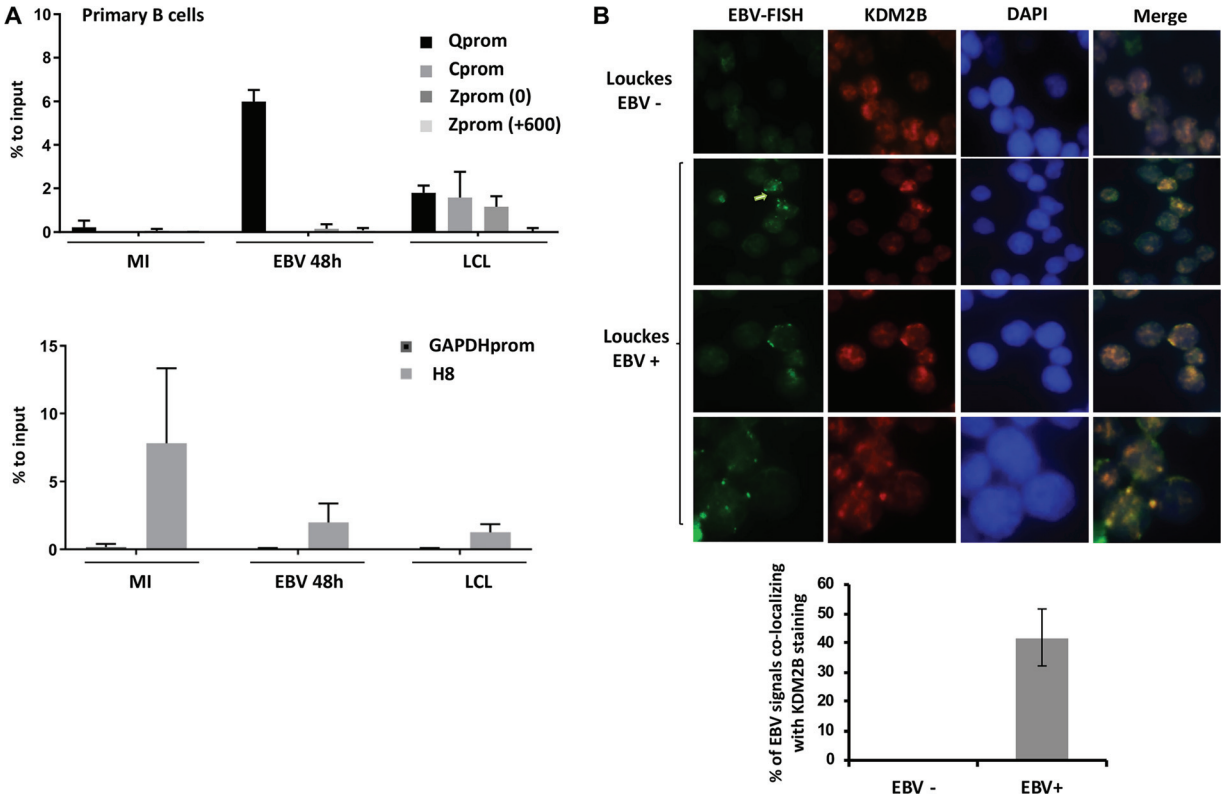


FIG 8 KDM2B is recruited to the EBV genome. (A) Primary B cells from three donors were infected with EBV for 48 h or left immortalized until they generated LCL. Mock-infected (MI) B cells, B cells infected with EBV for 48 h, and LCL were formaldehyde fixed and processed for ChIP using a KDM2B antibody and an IgG antibody. The eluted DNA was analyzed by qPCR with primers designed on different EBV gene promoters. As negative controls, the GAPDH promoter was also amplified, and the recruitment of KDM2B to its known cellular gene target, H8, was also assessed. Qprom, Cprom, and Zprom, Qp, Cp, and Zp EBV promoters, respectively. (B) Louckes cells untreated or infected with EBV (the B95-8 strain) were fixed on glass slides and processed for immuno-FISH. EBV DNA was detected by FISH, and KDM2B was detected by immunofluorescence. Overlapping EBV and KDM2B signals are shown in the merge fields as yellow dots. The histogram shows the percentage of merged dots, estimated by counting the number of EBV signals that colocalized with the KDM2B patches in at least 3 different fields from 3 independent stainings. DAPI, 4′,6-diamidino-2-phenylindole.

tected by immunofluorescence by using an anti-KDM2B antibody. In EBV-infected B cells, KDM2B patches partially overlapped or were found in close proximity to the viral DNA (on average, 40% of the green dots colocalized with the red dots) (Fig. 8B), suggesting that the epigenetic enzyme is recruited to or is close to the viral genome at early stages of infection. Taken together, our data demonstrate a role of KDM2B in controlling EBV gene expression.

Deregulation of KDM2B mRNA level in B cells alters their expression profile. KDM2B has been shown to play a role in cell differentiation, cell growth, and the proliferation of hematopoietic cells (9). We therefore assessed whether altered KDM2B expression could also have an impact on the B-cell phenotype in our experimental models. Louckes cells and LCL overexpressing KDM2B did not show an altered proliferation ability, nor did they show an altered cell cycle or apoptotic profiles (Fig. 9A to D). Therefore, to gain further insights into the impact of KDM2B deregulation on B cells, we performed an analysis with an RNA expression chip array in Louckes cells expressing increasing levels of KDM2B (Fig. 9E). Cellular genes whose expression was significantly altered by the enhanced levels of KDM2B were identified by bioinformatics analysis (Fig. 9E). In line with the known function of KDM2B as a demethylase of H3K4me3, the set of genes altered by KDM2B was enriched in H3K4me1 (Fig. 9F). Among the genes that were significantly deregulated in the presence of altered KDM2B expression, some play a role in immunity (Fig. 9G). Our analysis revealed that KDM2B overexpression was associated with deregulated expression of genes involved in the tumor necrosis factor receptor 2 (TNFR2) pathway (Fig. 9G). Interestingly, this pathway is important for the

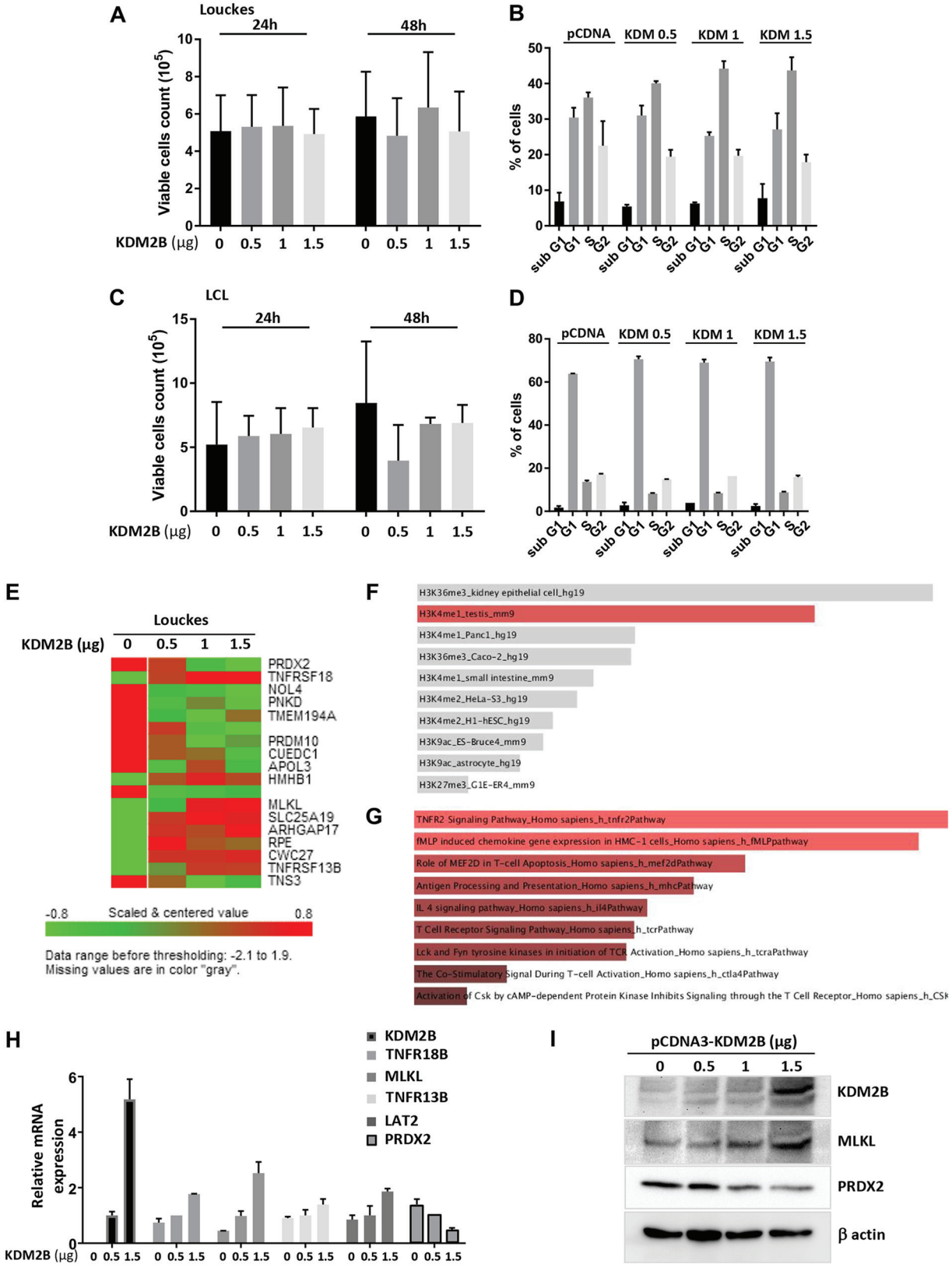


FIG 9 KDM2B deregulation alters cellular gene transcription. (A to D) Louckes cells (A and B) and LCL (C and D) were transfected with increasing concentrations of the KDM2B expression vector. (A and C) At 24 and 48 h after transfection, viable cells were counted. (B and D) Transfected cells were ethanol fixed and processed for cell cycle analysis by flow cytometry. (E) Louckes cells, generated as described in the legend to Fig. 5A, were collected and processed for RNA extraction and RNA expression profiling, as described in Materials and Methods. The differential expression analysis was conducted using BRB-ArrayTools software. (F and G) Genes differentially regulated in Louckes cells overexpressing KDM2B compared

(Continued on next page)

transition of B cells from germinal center B cells to resting memory B cells (24). Moreover, the TNFR2 pathway mediates specific tumor necrosis factor (TNF) effects and is an important mediator of the cell antiviral response (9). Deregulated expression of genes in this pathway was confirmed and validated by RT-qPCR (Fig. 9H) and by Western blot analysis for two of them (MLKL and PRDX2), which were, respectively, induced and downregulated during KDM2B overexpression in Louckes cells (Fig. 9I). Taken together, these data indicate that KDM2B plays a role in B-cell differentiation and in regulation of the TNF pathway, which is often altered during the lymphomagenic process.

DISCUSSION

In this study, we show that a regulatory region within the gene encoding the KDM2B protein is more methylated in EBV(+) BL than in EBV(−) BL, confirming data from our previous study, which aimed to characterize the whole epigenetic profile of a set of BL-derived cell lines of eBL or sBL origins. This region is found to be methylated in certain cancer-derived cell lines (ENCODE DNA methylation tracks, CpG methylation by Methyl 450K bead arrays from ENCODE/HAIB, UCSC Genome Browser). Furthermore, KDM2B levels appear to be deregulated in cancers of hematological origin (9). We therefore assessed whether increased methylation of the *KDM2B* gene in eBL is mediated by EBV to deregulate intracellular levels of the histone modifier and promote EBV-mediated lymphomagenesis. Indeed, KDMs have been shown to have altered expression in cancer (25). For instance, a KDM2B paralogue, KDM2A, behaves as a tumor suppressor in hematopoietic stem cells, in which it antagonizes mixed-lineage leukemia-associated leukemogenesis by erasing H3K36me2 markers (26). Similarly, altered levels of KDM2B could modify the chromatin structure and the pattern of expression of B cells and favor their transformation.

eBL specimens and derived cell lines showed low KDM2B protein expression, underscoring a potentially important role of KDM2B downregulation in the lymphomagenic process *in vivo*. Treating eBL cell lines with a chemical that blocks DNA methylation led to a rescue of the KDM2B mRNA level, indicating that DNA methylation contributes to the silencing of KDM2B expression in these cells. However, the same treatment left the level of the KDM2B transcript in EBV(−) sBL-derived cell lines unchanged. These events are similar to what we have previously described for another gene, *ID3* (4). *ID3*, which plays a key role in lymphomagenesis and which is often found mutated in the sBL variant (27, 28), was found to be silenced by hypermethylation in eBL-derived cell lines (4). EBV infection therefore plays a direct role in eBL pathogenesis by altering the cellular epigenome and deregulating the expression of genes with a key role in lymphomagenesis. Indeed, *in vitro* infection of primary B cells with EBV led to rapid downregulation of KDM2B expression. Low KDM2B mRNA levels at early stages of EBV infection of primary B cells were also observed in a data set from RNA expression profiling performed in an independent study (E. Manet, unpublished data). Our data also show that downregulation of KDM2B by EBV is not due to a specific B-cell activation status but is, rather, a direct result of infection by an actively transcribing virus. We also show that downregulation of KDM2B expression in EBV-infected cells is mediated by DNMT1 recruitment to its gene and by its DNA methylation. It has already been reported that EBV can alter DNMT1 activity through its viral proteins LMP1 and LMP2A (29). In particular, Tsai and colleagues showed that LMP1 activates DNMT1 activity (15) and that this event requires activation of the JNK/AP1 pathway.

Here, we show that cells stably expressing LMP1 display lower levels of KDM2B. The

FIG 9 Legend (Continued)
with their expression in Louckes cells transfected with an empty pCDNA vector were analyzed for their enrichment in specific pathways by using the Enrichr web tool. Enrichment results for the Epigenomics Roadmap HM ChIP sequencing and BioCarta 2016 databases (top and bottom, respectively) are shown. hESC, human embryonic stem cells; fMLP, *N*-formyl-methionyl-leucyl-phenylalanine; mhc, major histocompatibility complex; TCR, T-cell receptor; MEF2D, myocyte enhancer factor 2D. (H and I) Differential expression of a set of genes found to be deregulated in the expression profile array was validated by qPCR (H) and immunoblotting (I). The levels of the different transcripts were measured relative to their levels in Louckes cells transfected with 0.5 μg of the KDM2B construct.

ability of LMP1 to downregulate KDM2B depends mostly on its NF-κB activity, as shown by using an LMP1 molecule mutated on its NF-κB-activating domain and by blocking the NF-κB pathway via a specific chemical inhibitor. A LMP1 mutant lacking the JNK activation pathway also downregulates KDM2B, but to a lesser extent than wild-type LMP1. This indicates a partial contribution of that pathway to LMP1-mediated KDM2B downregulation. However, exposing the cells to a specific JNK inhibitor had no effect on the KDM2B mRNA expression level. Moreover, our ChIP experiments showed that expression of LMP1 in the EBV(−) RPMI cells is able to trigger DNMT1 recruitment to the *KDM2B* gene. The finding that the recruitment of DNMT1 to the *KDM2B* gene can be hampered by treating LMP1-RPMI cells with the IκBα kinase inhibitor further confirms its dependence on the ability of LMP1 to induce the NF-κB pathway. Recently published data showed that HPV16 E6/E7 transforming proteins inhibit the expression of miRNA-146-5p, known to target the KDM2B transcript, which results in an increase in the KDM2B expression level in HPV16-infected cells. In contrast to E6/E7, the EBV transforming protein LMP1 induces miRNA-146-5p via its ability to activate NF-κB (18). EBNA2 also induces miRNA-146-5p (19). This event could contribute to the reduction in the KDM2B mRNA level in EBV-infected cells. Our data show that blocking miRNA-146-5p in primary B cells before infection leads to a partial rescue of KDM2B levels compared with those in untreated cells. It is therefore possible that the virus uses alternative mechanisms to target KDM2B: (i) increasing its DNMT1-mediated gene methylation associated with LMP1 protein expression and (ii) controlling KDM2B mRNA levels by a specific miRNA associated with the expression of LMP1 and/or EBNA2. Further studies are needed to assess the contributions of deregulated miRNA-146-5p levels to the events described here. Our data also indicate that proteins other than LMP1 and EBNA2 appear to be able to mediate downregulation of KDM2B. EBV may need to tightly control KDM2B levels and activity to regulate expression of its own genes as well as that of cellular genes. Indeed, a recent study from Gillman and colleagues showed that EBNA3C interacts with KDM2B by its TFGC motif (HD motif) (30). Mutation of the latter motif impairs the ability of EBNA3C to bind to KDM2B and repress its target genes (30).

Previous studies by us and others showed that EBV can modulate the level and the activity of different epigenetic enzymes (20, 21, 31), which in turn play a role in regulating viral gene expression. One example is EZH2, whose intracellular expression level is induced in B cells by EBV infection in an LMP1-dependent manner (32). EZH2, in turn, is recruited to the viral genome, where it participates in the establishment and maintenance of EBV latency by methylating H3K27 in proximity to the BZLF1 and BRFL1 promoters (22). Our observation that EBV infection alters KDM2B expression prompted us to assess whether the epigenetic enzyme could regulate EBV infection and/or the EBV life cycle. Altering KDM2B expression in B cells before EBV infection or in latently EBV-infected B cells led to the deregulated expression of all analyzed viral genes. This was consistent with the recruitment of KDM2B to their respective promoters, as observed by ChIP experiments. In line with its ability to demethylate the active chromatin marker and repress transcription, KDM2B depletion in EBV-infected B cells led to increased viral gene expression, whereas its forced ectopic expression had the opposite effect and caused a strong reduction in the levels of viral transcripts. Taken together, our data indicate that KDM2B plays a role in controlling EBV gene expression, for instance, during the establishment of latency. Recruitment of KDM2B to the viral episome during the first step of infection could be necessary for the repression of viral gene transcription, especially at the end of the pre-latent phase, when the EBV lytic genes are silenced to allow the virus to persist in resting peripheral B cells (32). In contrast, KDM2B could work as a host restriction factor; therefore, its downregulation during the early stages of EBV infection would allow efficient viral gene expression and replication during the pre-latent phase. Future studies will be needed to investigate the exact role of EBV-mediated KDM2B deregulation in EBV life cycle control.

It is known that in order to escape the immune surveillance of the host and establish a chronic infection, EBV has evolved different mechanisms to maintain B cells in a status

of long-lived circulating memory B cells and prevent them from differentiating into antibody-secreting plasma cells. A recent study showed that EBNA3A and EBNA3C block the terminal differentiation of memory B cells to plasma cells by epigenetically repressing the gene encoding BLIMP-1, a master regulator in B-cell differentiation (33). Notably, a previous study showed that KDM2B plays a key role in the differentiation of hematological stem cells (9). Therefore, EBV-mediated deregulation of KDM2B expression could contribute to the mechanism that prevents the cells from undergoing terminal differentiation. This hypothesis is supported by our data from an RNA profile analysis conducted on cells transfected with increasing levels of a KDM2B-expressing construct. Cells harboring high KDM2B levels showed a deregulated expression of genes enriched in the TNFR2 pathway, which is known to play a role in the differentiation from B cells to plasma cells (24). TNFR2 also regulates the interferon pathway, an important mediator of the antiviral response. Ablation of KDM2B in hematopoietic cells has previously been shown to downregulate the interferon response (9). The downregulation of the KDM2B expression level upon EBV infection could therefore contribute to the viral escape from immune system surveillance.

Taken together, our data show a novel interplay between EBV infection and the host epigenome. EBV alters KDM2B expression via an epigenetic mechanism involving LMP1 and other latent viral proteins. The histone modifier, in turn, plays a role in regulating the expression of viral and host cell genes. These data, in addition to the observed deregulated KDM2B levels in BL-derived cell lines, indicate that altered expression of this epigenetic enzyme could contribute to B-cell transformation. Future studies aimed at investigating the functional importance of *KDM2B* gene methylation and downregulated expression during the EBV-mediated lymphomagenic process are warranted.

MATERIALS AND METHODS

Case selection, immunophenotype, and FISH. We studied 22 morphologically and immunophenotypically typical BL cases (11 sporadic and 11 endemic). All cases were diagnosed according to the updated World Health Organization (WHO) classification of tumors of hematopoietic and lymphoid tissues (34). The cases were retrieved from the archives of Siena University Hospital (Siena, Italy; $n = 2$) and Nairobi University (Nairobi, Kenya; $n = 20$). Before enrolling the cases in this study, they were reevaluated by an expert hematopathologist (L.L.), and the diagnosis was confirmed by morphology on histological slides stained with hematoxylin and eosin (H&E) or Giemsa and by immunophenotyping. The main clinical features of our samples are summarized in Table 2. The study was conducted at the University of Siena in Italy according to the principles of the Helsinki declaration after approval of the local review board. All the procedures were carried out automatically on representative paraffin sections from each case by Bench Mark Ultra (Ventana, Monza, Italy) using extended antigen retrieval and with diaminobenzidine (DAB) as the chromogen.

Cell culture and treatment. Peripheral B cells were purified from blood samples using the RosetteSep human enrichment kit (catalog number 15064; Stem Cell Technologies). LCL were generated in this study by infection of primary B cells from different donors. The myeloma-derived RPMI-8226 cells (http://web.expasy.org/cellosaurus/CVCL_0014) and the BL-derived cell lines, including the BL EBV(−) Louckes cell line (http://web.expasy.org/cellosaurus/CVCL_8259), were obtained from the International Agency for Research on Cancer (IARC) Biobank. The Akata2000 and Akata31 cell lines (35) were also used in the present study. Primary and immortalized B cells were cultured in RPMI 1640 medium (Gibco, Invitrogen Life Technologies, Cergy-Pontoise, France) supplemented with 10% fetal bovine serum, 100 U/ml penicillin G, 100 mg/ml streptomycin, 2 mM L-glutamine, and 1 mM sodium pyruvate (PAA; Pasching, Austria) or in advanced RPMI 1640 medium (catalog number 12633012; Life Technologies). EBV (the B95-8 strain) particles produced by culturing $HEK293_{EBVgfp}$ cells were used to infect B cells. EBV infection of B cells was performed either using a recombinant EBV-GFP genome or using the EBV mutant strain EBVΔLMP1 lacking the entire LMP1. The percentage of GFP-positive cells was assessed by fluorescence-activated cell sorting (FACS; FACSCanto system; Becton, Dickinson) and spanned from 10% to 15% at 24 to 48 h postinfection in Louckes and RPMI cells and 60% to 80% when measured at 48 h postinfection in primary B cells. Analysis of the cell cycle and apoptosis (sub-G_1 phase) was performed by ethanol fixing the cells and staining their DNA with propidium iodide at a final concentration of 5 μg/ml. Subsequently, cells were analyzed by FACS. Inactivation of EBV was performed with UV light (6×10 mJ). EBV reactivation in BL cells was done by treatment of the cells with TPA (50 ng/ml) in association with NaB (3 mM). Anti-miRNA treatment was done by the addition of Hsa-miR-146a-5p miRCURY LNA miRNA power inhibitor (250 nM; Qiagen) directly to the cell culture medium 24 h before infection with EBV. To induce their activation, primary B cells were seeded at a density of 0.5×10^6 cells in 6-well dishes and treated with 100 ng/ml recombinant human CD40 ligand (hCD40L; catalog number 6245-CL; R&D Systems) and 20 ng/ml of recombinant human IL-4 (R&D Systems) for 48 h. Cells were then collected and processed for further analysis. To block DNA methylation, cells were treated with 5-aza-2′-deoxycytidine (≥97%; catalog number A3656; Sigma-Aldrich) at a final concentration of 10 μM for 48

TABLE 3 Primers used for qPCR and ChIP-qPCR

Primer use and primer	Sequence	
	Forward	Reverse
qPCR		
LMP1	CCAGTCCAGTCACTCATAACG	CCTACATAAGCCTCTCACACT
EBNA1	GGTCGTGGACGTGGAGAAAA	GGTGGAGACCCGGATGATG
DNMT1	GAG GAA GCT GCT AAG GAC TAG TTC	ACT CCA CAA TTT GAT CAC TAA ATC
β2 microglobulin	CTCACGTCATCCAGCAGAGA	CGGCAGGCATACTCATCTTT
Beta globin	GCATCTGACTCCTGAGGAGA	AGCACACACACCAGCACATT
CCL22	ACTGCACTCCTGGTTGTCCT	CGGCACAGATCTCCTTATCCC
PRDX2	GTG TCC TTC GCC AGA TCA CT	ACG TTG GGC TTA ATC GTG TC
Actin	CTG GGA GTG GGT GGA GGC	TCA ACT GGT CTC AAG TCA GTG
GAPDH	GCCAAAAGGGTCATCATC	TGCCAGTGAGCTTCCCGTTC
KDM2B	CCC AGC ATC TGA AGG AGA AG	GTT GGA GGA ATC AGC CAA AA
LAT2	ACTCCTCTCTCTCCTGCAGA	CGAGGATAGTAGGGGCAAGG
GNA15	AGAATCGCTTGAACCCAGGA	ATTTCGAACTCCTGGCCTCA
TNFRSF13B	AACTCGGGAAGGTACCAAGG	GAAGACTTGGCCGGACTTTG
TNFRSF18	CTCTTGAAACCCGAGCATGG	ACTCGGAACAGCACTCCTC
LTA	ACTCCTCTCTCTCCTGCAGA	AGGAAGAGACGTTCAGGTGG
MLKL	AGGTCTAGGCCACACTTGTC	TGCAGGTCATGGGCTTCTAA
BZLF1	AATGCCGGGCCAAGTTTAAGCA	TTGGGCACATCTGCTTCAACAGGA
BDLF1	CGCAGACATGCTCGATGTA	TAGTGGTGCCCCAGGTATG
EBER	CCCTAGTGGTTTCGGACACA	ACTTGCAAATGCTCTAGGCG
BDRF1	CGGAGTGGCTCAGTCTAAGG	AGGTGGGCTGACACAGAC
ChIP-qPCR		
CpG island 127	TGACCTCTGCAGCTTCCTCT	GATGATCTGCCGCCAACTT
Z prom (0)	TAGCCTCGAGGCCATGCATATTTCAACT	GCCAAGCTTCAAGGTGCAATGTTTAGTG
Z prom (+600)	AGGTATGTTCCTGCCAAAGC	GTTCATGGACAGGTCCTGTG
H8 rDNA	AGTCGGGTTGCTTGGGAATGC	CCCTTACGGTACTTGTTGACT
BZLF1 prom	GGAGAAGCACCTCAACCTG	CTCCTTACCGATTCTGGCTG
EBNA Cp	AGT TGG TGT AAA CAC GCC GT	TCCACCTCTAAGGTCCCACG
Globin prom	AGGACAGGTACGGCTGTCATC	TTTATGCCCAGCCCTGGCTC
GAPDH prom	CGTGCCCAGTTGAACCAGG	AGGAGGAGCAGAGAGCGAAG
EBNA Qp	GGCTCACGAAGCGAGAC	GTCGTCACCCAATTTCTGTC
KDM2B cg21423404	ACCTGACACACCTCAACTCC	TTGTGGTTTGGGAGAAGGGT
KDM2B cg15695155	CTTGCCCCTTCCCACTAGAG	CCCTCTTCCCCAAACCATG

or 96 h. To inhibit the different pathways, cells were treated with the IκBα kinase inhibitor BAY11-7082 (Calbiochem) at a final concentration of 10 μM or with the JNK inhibitor II SP600125 (catalog number 420119; VWR International), used at a final concentration of 10 μM. Cells were preincubated with the different inhibitors for 1.5 h and 2 h.

qPCR. Total RNA was extracted using an AllPrep DNA/RNA minikit (Qiagen). RNA reverse transcription to cDNA was carried out by the use of RevertAid H Minus Moloney murine leukemia virus reverse transcriptase (Thermo Fisher Scientific), according to the manufacturer's protocol. Quantitative PCR (qPCR) was performed using a MesaGreen qPCR MasterMix Plus for SYBR assay (Eurogentec). For each primer set, the qPCR was performed in duplicate and the mRNA levels obtained were normalized to the average mRNA levels of three housekeeping genes (β-globin, β-actin, and GAPDH) measured in the same samples or to the mRNA level of β_2-microglobulin only. For each PCR, a sample in which the DNA template was replaced with PCR-grade water was included as a negative control. To measure the EBV genome copy number per cell, total DNA was extracted using an AllPrep DNA/RNA minikit (Qiagen) and measured by use of a NanoDrop spectrophotometer. Similar amounts of DNA were used as a template for TaqMan PCR, performed according to the protocol described by Accardi et al. (36). The PCR primer sequences are indicated in Table 3. All the primers used for the first time in the present study were assessed for their efficiency (90% to 110%).

miRNA-146-5p analysis. To analyze cellular miRNA, total RNA extracted by the AllPrep DNA/RNA minikit (Qiagen) was reverse transcribed using an miRCURY LNA miRNA PCR system (miRCURY LNA RT kit; catalog number 339340) according to the manufacturer's protocol. The cDNA was then analyzed for the levels of miRNA-146-5p using a specific miRNA qPCR primer (catalog number 339306; Hsa-miR-146a-5p miRCURY LNA miRNA PCR assay) and a PCR kit (catalog number 339345; miRCURY LNA SYBR green PCR kits) on a Bio-Rad qPCR machine (CFX96 Touch real-time PCR).

KDM2B overexpression and gene expression silencing. The KDM2B coding region was cloned into a pCDNA3 vector in frame with a hemagglutinin (HA) tag at the N terminus. LCL (1×10^7) and Louckes cells (5×10^6) were transfected with increasing concentrations of HA-KDM2B pCDNA3 (0.5, 1.0, and 1.5 μg) or with the pCDNA3 vector as a control by electroporation using a Neon transfection system (10-μl tips; pulse voltage, 1,350 V; pulse width, 30 ms; pulse number, 1). At 24 h after transfection, the cells were collected and processed for RNA/DNA extraction. Gene silencing of KDM2B was performed

TABLE 4 Primers used for pyrosequencing

Pyrosequencing primers	Sequence		
	Forward	Reverse[a]	Sequencing primer
KDM2B cg15695155	GGAGTGGGGTAGAGTTGAA	CCTACATACTACTAAACCCCC	AGGTTTGGT
			GAGTTTTAGGTGG
			GGATGGGTAGTT
			AGGGAAGGAATG
			AGTGGAGATAATG
CpG.127	AAATACAACAACCCTCCTACC	AAATACAACAACCCTCCTACC	GGGTGGTTGGGATAG
			TTTGGTTGGTTTGTT
			TTTTTTTAAGTATAT
			TTAAGTTTTTTTA
KDM2B cg21423404	GATAAGTATAGGGAGGTTTGTGA	CTATAAAACCATTTCCAACCC	GTAGGTGGTGATT
			GGTTATTAGAGT
	GGGTTGGAAATGGTTTTATAG	CCTCCCTAATAACTAAAACTACA	GTTGTTTATATG
			TTAATATAATGGT
KDM2B cg15695155	GAGTTTTAGGTGGTAYGGATG	GAYGGATAGGGAGGAGTTAGT	GGATGGGTAGTT
KDM2B cg00031896	GTAAAGGAGGAAATTAGGATTA	TATGTTTAAAGGAGGTTGTATG	TGGTTTGGTTAT
KDM2B cg21423404	AGGAGGAGTTTAGAGGTTATAGT	AGTTATTGTAGGGGTAGATTTTAG	GTAGGTGGTGATT
KDM2B cg12251659	GGAGGGAGTTYGGGAGGTAT	TTGGAGGGTYGAGTTGTAGG	AGTGYGTTTTTGTA

[a]The reverse primer was labeled with biotin at the 5′ end.

using KDM2B (human) unique 27-mer siRNA duplexes (catalog number HSS150072; Thermo Fisher Scientific). LCL (1×10^7) and Louckes cells (5×10^6) were transfected with the siRNA (final concentration, 250 nM) by electroporation using the Neon transfection system (10-μl tips; pulse voltage, 1,350 V; pulse width, 30 ms; pulse number, 1). At 48 h after transfection, the cells were collected and processed for RNA/DNA extraction. The levels of silencing were evaluated by qPCR using KDM2B-specific primers, indicated in Table 3.

Immunoblotting and antibodies. Whole-cell lysate extracts were obtained using lysis buffer, as previously described (37). The cell extracts were then fractionated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and processed for immunoblotting using standard techniques. The following antibodies were used for immunoblotting: KDM2B (catalog number 09-864 from Merck Millipore and catalog number ab5199 from Abcam), mouse monoclonal anti-human MLKL (catalog number SC-293201; clone 3B2; CliniSciences), LMP1 (S12 monoclonal antibody), β-actin (clone C4; MP Biomedicals), and GAPDH. Images were produced using a ChemiDoc XRS imaging system (Bio-Rad).

Immuno-FISH. Fifty thousand cells were resuspended in 5 μl of phosphate-buffered saline (PBS). The cells were gently spread on a microscope glass slide, air dried, and fixed in 4% paraformaldehyde–PBS for 10 min at room temperature. Slides containing fixed cells were washed 3 times in PBS for 5 min, permeabilized with PBS–0.5% Triton X-100 (Sigma-Aldrich) for 15 min, and then washed twice with PBS–0.05% Tween. The slides were then soaked in methanol–0.3% H_2O_2 for 30 min and incubated for 1 h with antibody diluent (catalog number S3022; Dako) and then for 30 min with Image-iT FX signal enhancer. The slides were incubated overnight at 4°C with an anti-KDM2B antibody (catalog number ab5199; Abcam) diluted to a concentration of 1 μg/ml, followed by incubation with a secondary antibody (anti-goat immunoglobulin; 5 μl/ml; Elite kit; Vector). To amplify the signal, the slides were incubated for 30 min at 37°C with ABC kit reagents according to the manufacturer's protocol. EBV DNA staining by FISH was performed as previously described (36), using a biotinylated probe to the EBV DNA genomic region BWRF1 (A300P.0100 DS-Dish-Probes). The stained cells were visualized with a fluorescence microscope with an incubator (Nikon Eclipse).

Immunohistochemistry and ISH for EBER. Immunohistochemistry analysis for KDM2B (dilution, 1:200; catalog number ab5199; Abcam) was performed by an automated staining system (Ventana BenchMark Ultra; Roche Diagnostics, Monza, Italy) on formalin-fixed, paraffin-embedded 4-μm-thick sections. An UltraView universal detection kit (Ventana) using a horseradish peroxidase multimer and DAB (as the chromogen) was used. ISH for EBER was carried out in each sample on 4-μm-thick sections, as previously reported (38). A control slide, prepared from a paraffin-embedded tissue block containing a metastatic nasopharyngeal carcinoma in a lymph node, was used as a positive control.

Chromatin immunoprecipitation. ChIP was performed with Diagenode Shearing ChIP and OneDay ChIP kits according to the manufacturer's protocols. The following antibodies were used: KDM2B (catalog number ab5199; Abcam), DNMT1 (catalog number MAB0079; Abnova), and IgG (Diagenode). The eluted DNA was used as a template for qPCR with primers designed on the *KDM2B* gene. The primers used for quantitative ChIP are listed in Table 3. The value of binding obtained for each antibody was calibrated on the input sample and normalized to the values for IgG.

Bisulfite modification and pyrosequencing. Samples for pyrosequencing were processed as previously described (4, 39). The primers are indicated in Table 4.

Whole-genome expression analysis. Differential expression analysis was performed using human HT-12 expression BeadChips (Illumina) as previously described (4, 36). Probes with *P* values of <0.01, a false-discovery rate (FDR) of <0.05, and a fold change in expression of at least 1.5 were considered differentially expressed.

Statistical analysis. Statistical significance was determined by Student's *t* test. The *P* value of each experiment is indicated in the corresponding figure legend. Error bars in the graphs represent the standard deviation.

ACKNOWLEDGMENTS

We are grateful to all members of the Epigenetics Group and the Infections and Cancer Biology Group at IARC for their support. We are thankful to Elizabeth Page and Latifa Bouanzi for helping with manuscript preparation, to Ester Sorrentino for helping with the immunohistochemistry, and to Karen Müller for editing the manuscript.

Finally, we thank our funders: La Ligue Contre le Cancer (LNCC) Rhone (GR-IARC-2014-04-07-03 to R.A. and OPE-2017-0009 to H.G.), Oncostarter-(CLARA) (GR-IARC-2014-05-15-02 to R.A.), IARC Junior Award 2016 (AFEES-2016 to R.A.), and Plan Cancer-INSERM (to R.A.).

R.A. designed, analyzed, and interpreted the experiments and wrote the first draft of the manuscript. H.G. and R.C.V.-A. contributed to the design, execution, and analysis of the experiments, in addition to revising the initial submission. R.A., H.G., Z.H., A.J., R.C.V.-A., L.M., L.L., F.M., H.H.-V., A.D., M.P.C., A.R., M.G.C., M.C.R.-M., C.C., G.D., C.S., M.A.M., F.L.C.-K., and E.M. contributed to the execution of the experiments, prepared the figures and tables, and reviewed the manuscript; in addition, they all contributed to revising the initial submission and the subsequent versions of the article.

We declare that we have no competing interests.

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article, and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

REFERENCES

1. Magrath I. 2012. Epidemiology: clues to the pathogenesis of Burkitt lymphoma. Br J Haematol 156:744–756. https://doi.org/10.1111/j.1365-2141.2011.09013.x.
2. Paschos K, Allday MJ. 2010. Epigenetic reprogramming of host genes in viral and microbial pathogenesis. Trends Microbiol 18:439–447. https://doi.org/10.1016/j.tim.2010.07.003.
3. Herceg Z, Paliwal A. 2011. Epigenetic mechanisms in hepatocellular carcinoma: how environmental factors influence the epigenome. Mutat Res 727:55–61. https://doi.org/10.1016/j.mrrev.2011.04.001.
4. Hernandez-Vargas H, Gruffat H, Cros MP, Diederichs A, Sirand C, Vargas-Ayala RC, Jay A, Durand G, Le Calvez-Kelm F, Herceg Z, Manet E, Wild CP, Tommasino M, Accardi R. 2017. Viral driven epigenetic events alter the expression of cancer-related genes in Epstein-Barr-virus naturally infected Burkitt lymphoma cell lines. Sci Rep 7:5852. https://doi.org/10.1038/s41598-017-05713-2.
5. He J, Shen L, Wan M, Taranova O, Wu H, Zhang Y. 2013. Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes. Nat Cell Biol 15:373–384. https://doi.org/10.1038/ncb2702.
6. Zhou Z, Yang X, He J, Liu J, Wu F, Yu S, Liu Y, Lin R, Liu H, Cui Y, Zhou C, Wang X, Wu J, Cao S, Guo L, Lin L, Wang T, Peng X, Qiang B, Hutchins AP, Pei D, Chen J. 2017. Kdm2b regulates somatic reprogramming through variant PRC1 complex-dependent function. Cell Rep 21:2160–2170. https://doi.org/10.1016/j.celrep.2017.10.091.
7. He J, Kallin EM, Tsukada Y, Zhang Y. 2008. The H3K36 demethylase Jhdm1b/Kdm2b regulates cell proliferation and senescence through P15(Ink4b). Nat Struct Mol Biol 15:1169–1175. https://doi.org/10.1038/nsmb.1499.
8. Suzuki T, Minehata K, Akagi K, Jenkins NA, Copeland NG. 2006. Tumor suppressor gene identification using retroviral insertional mutagenesis in Blm-deficient mice. EMBO J 25:3422–3431. https://doi.org/10.1038/sj.emboj.7601215.
9. Andricovich J, Kai Y, Peng W, Foudi A, Tzatsos A. 2016. Histone demethylase KDM2B regulates lineage commitment in normal and malignant hematopoiesis. J Clin Invest 126:905–920. https://doi.org/10.1172/JCI84014.
10. Vargas-Ayala RC, Jay A, Hernandez-Vargas H, Diederichs A, Robitaille A, Sirand C, Ceraolo MG, Romero M, Cros MP, Cuenin C, Durand G, Le Calvez-Kelm F, Mundo L, Maroui MA, Leoncini L, Manet E, Herceg Z, Gruffat H, Accardi R. 2019. Interplay between the epigenetic enzyme lysine (K)-specific demethylase 2B and Epstein-Barr virus infection. bioRxiv https://doi.org/10.1101/367094.
11. Guo Y, Walsh AM, Fearon U, Smith MD, Wechalekar MD, Yin X, Cole S, Orr C, McGarry T, Canavan M, Kelly S, Lin TA, Liu X, Proudman SM, Veale DJ, Pitzalis C, Nagpal S. 2017. CD40L-dependent pathway is active at various stages of rheumatoid arthritis disease progression. J Immunol 198:4490–4501. https://doi.org/10.4049/jimmunol.1601988.
12. Christman JK. 2002. 5-Azacytidine and 5-aza-2′-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. Oncogene 21:5483–5495. https://doi.org/10.1038/sj.onc.1205699.
13. Eliopoulos AG, Gallagher NJ, Blake SM, Dawson CW, Young LS. 1999. Activation of the P38 mitogen-activated protein kinase pathway by Epstein-Barr virus-encoded latent membrane protein 1 coregulates interleukin-6 and interleukin-8 production. J Biol Chem 274:16085–16096. https://doi.org/10.1074/jbc.274.23.16085.
14. Tsai CN, Tsai CL, Tse KP, Chang HY, Chang YS. 2002. The Epstein-Barr virus oncogene product, latent membrane protein 1, induces the downregulation of E-cadherin gene expression via activation of DNA methyltransferases. Proc Natl Acad Sci U S A 99:10084–10089. https://doi.org/10.1073/pnas.152059399.
15. Tsai CL, Li HP, Lu YJ, Hsueh C, Liang Y, Chen CL, Tsao SW, Tse KP, Yu JS, Chang YS. 2006. Activation of DNA methyltransferase 1 by EBV LMP1 involves C-Jun NH(2)-terminal kinase signaling. Cancer Res 66:11668–11676. https://doi.org/10.1158/0008-5472.CAN-06-2194.
16. Hino R, Uozaki H, Murakami N, Ushiku T, Shinozaki A, Ishikawa S, Morikawa T, Nakaya T, Sakatani T, Takada K, Fukayama M. 2009. Activation of DNA methyltransferase 1 by EBV latent membrane protein 2A leads to promoter hypermethylation of PTEN gene in gastric carcinoma. Cancer Res 69:2766–2774. https://doi.org/10.1158/0008-5472.CAN-08-3070.
17. Huen DS, Henderson SA, Croom-Carter D, Rowe M. 1995. The Epstein-Barr virus latent membrane protein-1 (LMP1) mediates activation of NF-kappa B and cell surface phenotype via two effector regions in its carboxy-terminal cytoplasmic domain. Oncogene 10:549–560.
18. Peta E, Sinigaglia A, Masi G, Di Camillo B, Grassi A, Trevisan M, Messa L,

Loregian A, Manfrin E, Brunelli M, Martignoni G, Palu G, Barzon L. 2018. HPV16 E6 and E7 upregulate the histone lysine demethylase KDM2B through the C-MYC/Mir-146a-5p Axys. Oncogene 37:1654–1668. https://doi.org/10.1038/s41388-017-0083-1.
19. Rosato P, Anastasiadou E, Garg N, Lenze D, Boccellato F, Vincenti S, Severa M, Coccia EM, Bigi R, Cirone M, Ferretti E, Campese AF, Hummel M, Frati L, Presutti C, Faggioni A, Trivedi P. 2012. Differential regulation of Mir-21 and Mir-146a by Epstein-Barr virus-encoded EBNA2. Leukemia 26:2343–2352. https://doi.org/10.1038/leu.2012.108.
20. Tempera I, Lieberman PM. 2014. Epigenetic regulation of EBV persistence and oncogenesis. Semin Cancer Biol 26:22–29. https://doi.org/10.1016/j.semcancer.2014.01.003.
21. Poreba E, Broniarczyk JK, Gozdzicka-Jozefiak A. 2011. Epigenetic mechanisms in virus-induced tumorigenesis. Clin Epigenetics 2:233–247. https://doi.org/10.1007/s13148-011-0026-6.
22. Murata T, Kondo Y, Sugimoto A, Kawashima D, Saito S, Isomura H, Kanda T, Tsurumi T. 2012. Epigenetic histone modification of Epstein-Barr virus BZLF1 promoter during latency and reactivation in Raji cells. J Virol 86:4752–4761. https://doi.org/10.1128/JVI.06768-11.
23. Galbiati A, Penzo M, Bacalini MG, Onofrillo C, Guerrieri AN, Garagnani P, Franceschi C, Trere D, Montanaro L. 2017. Epigenetic up-regulation of ribosome biogenesis and more aggressive phenotype triggered by the lack of the histone demethylase JHDM1B in mammary epithelial cells. Oncotarget 8:37091–37103. https://doi.org/10.18632/oncotarget.16181.
24. Ticha O, Moos L, Wajant H, Bekeredjian-Ding I. 2017. Expression of tumor necrosis factor receptor 2 characterizes TLR9-driven formation of interleukin-10-producing B cells. Front Immunol 8:1951. https://doi.org/10.3389/fimmu.2017.01951.
25. Thinnes CC, England KS, Kawamura A, Chowdhury R, Schofield CJ, Hopkinson RJ. 2014. Targeting histone lysine demethylases—progress, challenges, and the future. Biochim Biophys Acta 1839:1416–1432. https://doi.org/10.1016/j.bbagrm.2014.05.009.
26. Zhu L, Li Q, Wong SH, Huang M, Klein BJ, Shen J, Ikenouye L, Onishi M, Schneidawind D, Buechele C, Hansen L, Duque-Afonso J, Zhu F, Martin GM, Gozani O, Majeti R, Kutateladze TG, Cleary ML. 2016. ASH1L links histone H3 lysine 36 dimethylation to MLL leukemia. Cancer Discov 6:770–783. https://doi.org/10.1158/2159-8290.CD-16-0058.
27. Richter J, Schlesner M, Hoffmann S, Kreuz M, Leich E, Burkhardt B, Rosolowski M, Ammerpohl O, Wagener R, Bernhart SH, Lenze D, Szczepanowski M, Paulsen M, Lipinski S, Russell RB, Adam-Klages S, Apic G, Claviez A, Hasenclever D, Hovestadt V, Hornig N, Korbel JO, Kube D, Langenberger D, Lawerenz C, Lisfeld J, Meyer K, Picelli S, Pischimarov J, Radlwimmer B, Rausch T, Rohde M, Schilhabel M, Scholtysik R, Spang R, Trautmann H, Zenz T, Borkhardt A, Drexler HG, Moller P, Macleod RA, Pott C, Schreiber S, Trumper L, Loeffler M, Stadler PF, Lichter P, Eils R, Kuppers R, Hummel M, et al. 2012. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. Nat Genet 44:1316–1320. https://doi.org/10.1038/ng.2469.
28. Love C, Sun Z, Jima D, Li G, Zhang J, Miles R, Richards KL, Dunphy CH, Choi WW, Srivastava G, Lugar PL, Rizzieri DA, Lagoo AS, Bernal-Mizrachi L, Mann KP, Flowers CR, Naresh KN, Evens AM, Chadburn A, Gordon LI, Czader MB, Gill JI, Hsi ED, Greenough A, Moffitt AB, McKinney M, Banerjee A, Grubor V, Levy S, Dunson DB, Dave SS. 2012. The genetic landscape of mutations in Burkitt lymphoma. Nat Genet 44:1321–1325. https://doi.org/10.1038/ng.2468.
29. Scott RS. 2017. Epstein-Barr virus: a master epigenetic manipulator. Curr Opin Virol 26:74–80. https://doi.org/10.1016/j.coviro.2017.07.017.
30. Gillman ACT, Parker G, Allday MJ, Bazot Q. 2018. Epstein-Barr virus nuclear antigen 3C inhibits expression of COBLL1 and the ADAM28-ADAMDEC1 locus via interaction with the histone lysine demethylase KDM2B. J Virol 92:e01362-18. https://doi.org/10.1128/JVI.01362-18.
31. Accardi R, Fathallah I, Gruffat H, Mariggio G, Le Calvez-Kelm F, Voegele C, Bartosch B, Hernandez-Vargas H, McKay J, Sylla BS, Manet E, Tommasino M. 2013. Epstein-Barr virus transforming protein LMP-1 alters B cells gene expression by promoting accumulation of the oncoprotein deltanp73alpha. PLoS Pathog 9:E1003186. https://doi.org/10.1371/journal.ppat.1003186.
32. Thorley-Lawson DA. 2015. EBV persistence—introducing the virus. Curr Top Microbiol Immunol 390:151–209. https://doi.org/10.1007/978-3-319-22822-8_8.
33. Styles CT, Bazot Q, Parker GA, White RE, Paschos K, Allday MJ. 2017. EBV epigenetically suppresses the B cell-to-plasma cell differentiation pathway while establishing long-term latency. PLoS Biol 15:E2001992. https://doi.org/10.1371/journal.pbio.2001992.
34. Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H, Thiele J, Arber DA, Hasserjian RP, Le Beau MM, Orazi A, Siebert R (ed). 2017. WHO classification of tumours of haematopoietic and lymphoid tissues, 4th ed. World Health Organization, Geneva, Switzerland.
35. Bryant H, Farrell PJ. 2002. Signal transduction and transcription factor modification during reactivation of Epstein-Barr virus from latency. J Virol 76:10290–10298. https://doi.org/10.1128/JVI.76.20.10290-10298.2002.
36. Accardi R, Gruffat H, Sirand C, Fusil F, Gheit T, Hernandez-Vargas H, Le Calvez-Kelm F, Traverse-Glehen A, Cosset FL, Manet E, Wild CP, Tommasino M. 2015. The mycotoxin aflatoxin B1 stimulates Epstein-Barr virus-induced B-cell transformation in in vitro and in vivo experimental models. Carcinogenesis 36:1440–1451. https://doi.org/10.1093/carcin/bgv142.
37. Giarre M, Caldeira S, Malanchi I, Ciccolini F, Leao MJ, Tommasino M. 2001. Induction of Prb degradation by the human papillomavirus type 16 E7 protein is essential to efficiently overcome P16ink4a-imposed G_1 cell cycle arrest. J Virol 75:4705–4712. https://doi.org/10.1128/JVI.75.10.4705-4712.2001.
38. Mundo L, Ambrosio MR, Picciolini M, Lo Bello G, Gazaneo S, Del Porro L, Lazzi S, Navari M, Onyango N, Granai M, Bellan C, De Falco G, Gibellini D, Piccaluga PP, Leoncini L. 2017. Unveiling another missing piece in EBV-driven lymphomagenesis: EBV-encoded microRNAs expression in EBER-negative Burkitt lymphoma cases. Front Microbiol 8:229. https://doi.org/10.3389/fmicb.2017.00229.
39. Hernandez-Vargas H, Lambert MP, Le Calvez-Kelm F, Gouysse G, McKay-Chopin S, Tavtigian SV, Scoazec JY, Herceg Z. 2010. Hepatocellular carcinoma displays distinct DNA methylation signatures with potential as clinical predictors. PLoS One 5:E9749. https://doi.org/10.1371/journal.pone.0009749.

PLOS | PATHOGENS

RESEARCH ARTICLE

Beta HPV38 oncoproteins act with a hit-and-run mechanism in ultraviolet radiation-induced skin carcinogenesis in mice

Daniele Viarisio[1], Karin Müller-Decker[1], Rosita Accardi[2], Alexis Robitaille[2], Matthias Dürst[3], Katrin Beer[3], Lars Jansen[3], Christa Flechtenmacher[4], Matthias Bozza[1], Richard Harbottle[1], Catherine Voegele[2], Maude Ardin[2], Jiri Zavadil[2], Sandra Caldeira[1¤], Lutz Gissmann[1,5], Massimo Tommasino[2]*

1 Deutsches Krebsforschungszentrum, Heidelberg, Germany, **2** International Agency for Research on Cancer, World Health Organization, Lyon, France, **3** Department of Gynecology, Jena University Hospital - Friedrich Schiller University, Jena, Germany, **4** Department of Pathology, University Hospital of Heidelberg, Heidelberg, Germany, **5** Department of Botany and Microbiology (honorary member), King Saud University, Riyadh, Saudi Arabia

¤ Current address: European Commission, Joint Research Centre (JRC), Ispra, Italy
* tommasinom@iarc.fr

Citation: Viarisio D, Müller-Decker K, Accardi R, Robitaille A, Dürst M, Beer K, et al. (2018) Beta HPV38 oncoproteins act with a hit-and-run mechanism in ultraviolet radiation-induced skin carcinogenesis in mice. PLoS Pathog 14(1): e1006783. https://doi.org/10.1371/journal.ppat.1006783

Editor: Paul Francis Lambert, University of Wisconsin Madison School of Medicine and Public Health, UNITED STATES

Received: July 3, 2017

Accepted: November 30, 2017

Published: January 11, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The study was supported by a grant from Deutsche Krebshilfe (no. 110259) to LG and MT (https://www.krebshilfe.de/) and by a grant from Fondation ARC pour la recherche sur le cancer (no. PJA 20151203192) (https://www.fondation-arc.org/espace-chercheur) and Institut National de la

Abstract

Cutaneous beta human papillomavirus (HPV) types are suspected to be involved, together with ultraviolet (UV) radiation, in the development of non-melanoma skin cancer (NMSC). Studies in *in vitro* and *in vivo* experimental models have highlighted the transforming properties of beta HPV E6 and E7 oncoproteins. However, epidemiological findings indicate that beta HPV types may be required only at an initial stage of carcinogenesis, and may become dispensable after full establishment of NMSC. Here, we further investigate the potential role of beta HPVs in NMSC using a Cre-loxP-based transgenic (Tg) mouse model that expresses beta HPV38 E6 and E7 oncogenes in the basal layer of the skin epidermis and is highly susceptible to UV-induced carcinogenesis. Using whole-exome sequencing, we show that, in contrast to WT animals, when exposed to chronic UV irradiation K14 HPV38 E6/E7 Tg mice accumulate a large number of UV-induced DNA mutations, which increase proportionally with the severity of the skin lesions. The mutation pattern detected in the Tg skin lesions closely resembles that detected in human NMSC, with the highest mutation rate in p53 and Notch genes. Using the Cre-lox recombination system, we observed that deletion of the viral oncogenes after development of UV-induced skin lesions did not affect the tumour growth. Together, these findings support the concept that beta HPV types act only at an initial stage of carcinogenesis, by potentiating the deleterious effects of UV radiation.

Author summary

Many epidemiological and biological findings support the hypothesis that beta HPV types cooperate with UV radiation in the induction of NMSC, the most common form of human cancer. We have previously shown that K14 HPV38 E6/E7 Tg mice, when exposed

Santé Et de la Recherche Médicale (no. ENV201610) to MT (https://www.eva2.inserm.fr/EVA/jsp/AppelsOffres/CANCER/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

to long-term UV radiation, developed NMSC, whereas WT animals subjected to identical treatments did not develop any type of skin lesions. Here, we show that the high skin cancer susceptibility of these Tg animals tightly correlates with their tendency to accumulate UV-induced mutations in genes that are frequently mutated in human NMSC. Importantly, deletion of the HPV38 E6 and E7 genes in existing skin lesions did not affect the further growth of the cancer cells. Together, these findings support the model that beta HPV infection is a co-factor in skin carcinogenesis, facilitating the accumulation of the UV-induced DNA mutations.

Introduction

Non-melanoma skin cancer (NMSC) is the most common cancer in adult Caucasian populations [1]. The cutaneous human papillomavirus (HPV) types belonging to genus beta are suspected, together with ultraviolet (UV) radiation, to be involved in NMSC [2,3]. The first two beta HPV types, 5 and 8, were isolated from skin lesions of patients with a disorder called epidermodysplasia verruciformis (EV). EV patients are highly susceptible to beta HPV infection in the skin and develop cutaneous squamous cell carcinoma (cSCC) at anatomical sites exposed to sunlight [4]. The fact that organ transplant recipients, due to their immunosuppressed status, have an elevated risk of beta HPV infection and development of cSSC provided evidence for the role of beta HPV types in skin carcinogenesis also in non-EV individuals [5,6]. Finally, many epidemiological studies support the link between these viruses and cSCC in the general population [2,3,7]. These studies showed that, compared with the general population, patients with a history of cSCC are more frequently positive for viral DNA in the skin and/or for antibodies against the major capsid protein L1.

Molecular analysis showed that not all cancer cells contain a copy of the beta HPV genome and that the copy number of the beta HPV genome is higher in pre-malignant actinic keratosis (AK), a precursor lesion of SCC, than in SCC [8]. Thus, these data suggest that beta HPV types may act at an initial stage of skin carcinogenesis and that after full transformation of the infected cells, viral DNA can be lost. This model is consistent with the fact that additional carcinogens are involved in skin carcinogenesis. Considering that UV radiation is the key risk factor for cSSC development [9–11], the most plausible hypothesis is that beta HPV types exacerbate the accumulation of a large number of UV-induced somatic mutations, facilitating cellular transformation. Subsequently, the expression of the viral oncogenes may become irrelevant for the maintenance of the malignant phenotype.

Several studies in human keratinocytes, the natural host of beta HPV types, showed that E6 and E7 from some beta HPV types target key pathways linked to DNA repair, apoptosis, and cellular transformation [3]. Several transgenic (Tg) models for beta HPV have been generated [12–16], some of which have highlighted the synergism between viral oncogene expression in the skin epithelium and UV radiation in promoting cSCC [3]. Tg mice expressing beta HPV38 E6 and E7 in the basal layer of the epidermis under the control of the cytokeratin K14 promoter (K14) did not spontaneously develop any lesions during their life span. Upon long-term exposure to UV radiation (30 weeks), they developed first skin lesions closely resembling human AK and subsequently cSCCs. In contrast, wild-type (WT) mice developed neither pre-malignant lesions nor cSCCs when exposed to the same dose of UV radiation [15]. However, it is still unknown whether the high susceptibility of the K14 HPV38 E6/E7 Tg animals to UV-induced skin carcinogenesis is linked to the accumulation of mutations facilitated by the viral oncoproteins, which may become dispensable after cSCC development. In this study, we

addressed this open question on the synergism between UV radiation and beta HPV38 E6 and E7 oncoproteins using the Tg mouse model. We showed that viral oncoproteins act at an initial stage of UV-induced skin carcinogenesis, facilitating the accumulation of a large number of somatic mutations in crucial genes that are associated with cSCC development in humans. In addition, silencing of the expression of the viral genes in established skin lesions does not affect further tumour growth.

Results

Expression of HPV38 E6 and E7 in mouse skin facilitates the accumulation of UV-induced DNA mutations

We have previously shown that HPV38 E6/E7 expression in mouse skin strongly increases susceptibility to UV-induced carcinogenesis [15]. To evaluate whether the development of skin lesions present in K14 HPV38 E6/E7 Tg mice of chronic UV irradiation correlated with the number of accumulated DNA mutations, we used whole-exome sequencing of WT and Tg samples.

For this analysis, we selected normal skin from WT mice not exposed or exposed to UV radiation for 30 weeks ($n = 2$) and histologically confirmed skin specimens from three independent K14 HPV38 E6/E7 Tg mice UV-irradiated for 30 weeks, i.e., (i) normal skin, (ii) pre-malignant skin lesions and (iii) cSCC. For the pre-malignant lesions, the histological analyses revealed that they have the classic features observed in humans of the precancerous condition of AK, including slight atypia, parakeratosis, and acanthosis (S1 Fig) [15]. Exome sequencing (Illumina Hi-Seq) of collected samples generated an average coverage of 141.71× ± 11.9 (mean ± standard deviation).

The genomic sequence of the WT mouse not exposed to UV radiation was used as a control sample in paired analysis. Only 10 mutations were detected in the skin of the UV-irradiated WT mouse. Similarly, less than 10 mutations were detected in the Tg mouse not exposed to UV irradiation. In both cases, all the mutations were in genes not directly linked to carcinogenesis (S1 Table).

In UV-irradiated Tg animals, the mutational load varied across our cohort of well-differentiated cSCC exomes, averaging 3541 somatic variants (range, 3261–4027) or 68.58 ± 7.64 variants per Mb. The exome of the pre-malignant samples had substantially fewer variants, with an average of 1337 somatic variants (range, 937–2026) or 23.14 ± 14.70 variants per Mb. The exome of the chronically UV-exposed normal skin of Tg mice harboured an average of 15 somatic variants (range, 11–20) or 0.29 ± 0.08 variants per Mb (S2 Table). Thus, the number of somatic mutations was proportional to the severity of the skin lesion; the average number in SCCs was approximately double that in the pre-malignant lesions (Fig 1A).

The vast majority of the somatic mutations detected in SCCs were C:G > T:A mutations, mutations that are also prevalent in the UV-induced mutational signature (Fig 1B and 1C). We applied the non-negative matrix factorization (NMF) method to extract the mutational signatures composed of 96 single base substitution (SBS) types considering the sequence context (one base upstream and one base downstream) (S2 Fig). The extracted signature was compared with known mutational signatures by the cosine similarity method [17,18]. The value of the similarity obtained for the new B signature is 0.86 for COSMIC signature 27 (UV signature) (S2 Fig), indicating the clear prevalence of the impact of UV radiation on the etiology of these cSCCs.

To assess the biological significance of the somatic mutations detected in the skin lesions of the K14 HPV38 E6/E7 Tg mice, we determined whether they were detected in the previously compiled lists of epi-driver and epi-modifier genes [19–23], as well as genes identified in the Cancer Gene Census [24]. As shown in Fig 2, three classes of genes were found to be

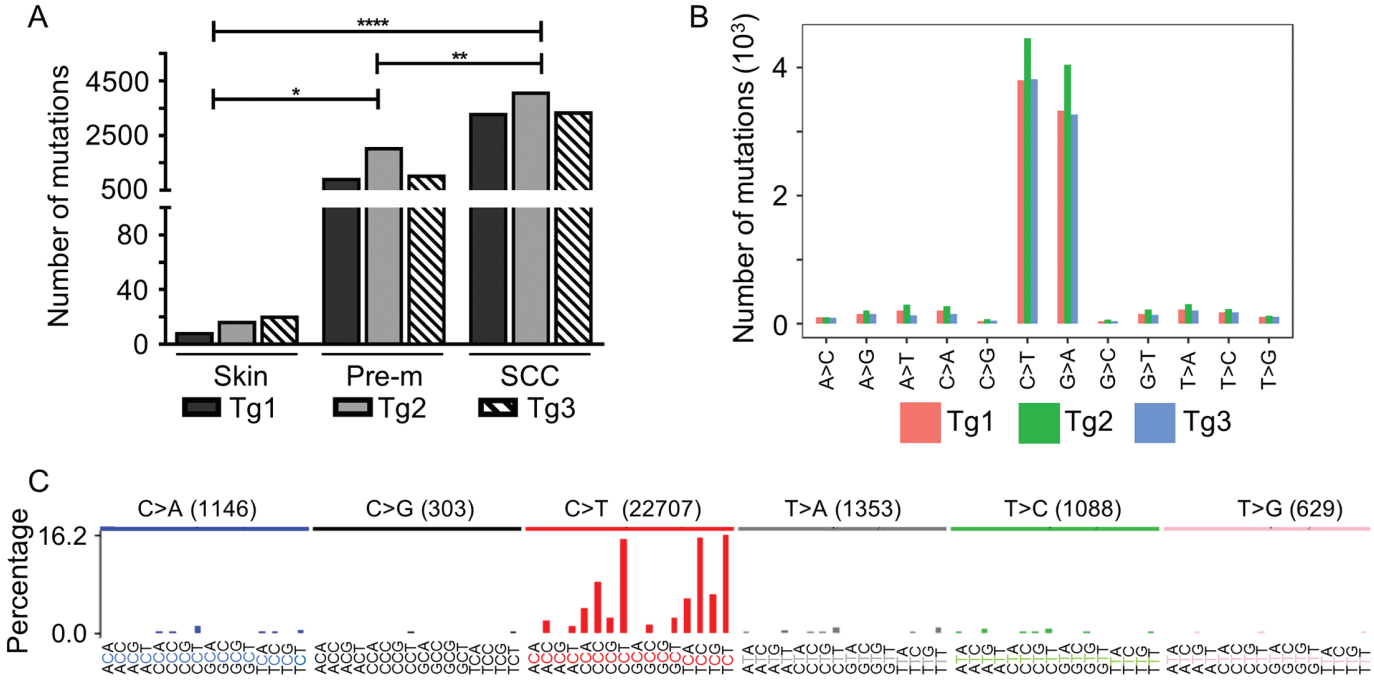


Fig 1. HPV38 E6 and E7 induce an increased steady-state level of UV-induced mutations in mouse skin keratinocytes. (A) UV-induced cSCCs in K14 HPV38 E6/E7 Tg mice have a vast number of somatic mutations. SCCs display a very high mutational load, with each Tg animal (Tg1–3) harbouring almost 3 times the number of variants compared with pre-malignant lesions (Pre-m). All differences in number of DNA mutations among the tree types of specimens were statistically significant: * ≤0.05; ** ≤0.01; **** ≤0.0001. (B) cSCCs of K14 HPV38 E6/E7 Tg mice display the classic UV-induced mutation signature with a very high number of C:G > T:A mutations. This type of mutation represents the majority of the SNV type in SCC samples of the three Tg animals. (C) Mutation spectrum of pooled SCC samples from the three mice. This spectrum displays the high prevalence of C:G > T:A mutations, especially in the 5′-T_N-3′ and 5′-C_N-3′ context. The *y* axis represents the percentage of mutations, and the *x* axis the trinucleotide sequence context.

https://doi.org/10.1371/journal.ppat.1006783.g001

recurrently mutated in pre-malignant and malignant skin lesions of K14 HPV38 E6/E7 Tg animals, suggesting a selective process for the enrichment of mutations in these groups of genes.

Pathway analyses confirmed that the mutations detected in mouse cSCC affect key pathways intimately linked to cellular transformation (S3 Table).

A comparison of somatic mutations detected in our experimental Tg mouse model and in human cSCC [25] revealed that a large number of epi-driver, epi-modifier, and Cancer Gene Census genes were recurrently mutated in murine and human cSCC (Fig 3A).

A recent study identified the top human genes mutated in cSCC [26]. Interestingly, most of these genes are also found to be mutated in the UV-induced skin lesions of the K14 HPV38 E6/E7 Tg animals (Fig 3B). In agreement with previous findings on human cSCC [25], Trp53 showed up as the most mutated gene in the murine Tg-derived cSCC (Figs 2A and 3B). Here, p53 mutations appear to be an early event in skin carcinogenesis, because they were detected in one sample of normal skin as well as in all pre-malignant lesions and cSCCs. In agreement with our data, it was reported that p53 mutations can be detected in keratinocytes of UV-exposed normal skin [27,28]. However, all mutations were identified in the p53 DNA-binding domain (S4 Table), supporting their key role in the process of carcinogenesis. Consistent with the fact that in keratinocytes the Notch signalling pathway promotes cell-cycle exit and differentiation [29,30], *NOTCH1* and *NOTCH2* have been found to be mutated in human cSCC [25]. In our Tg mouse model, mutated *NOTCH1* and/or *NOTCH2* were also detected in all three cSCCs, but never in pre-malignant lesions (Fig 3B).

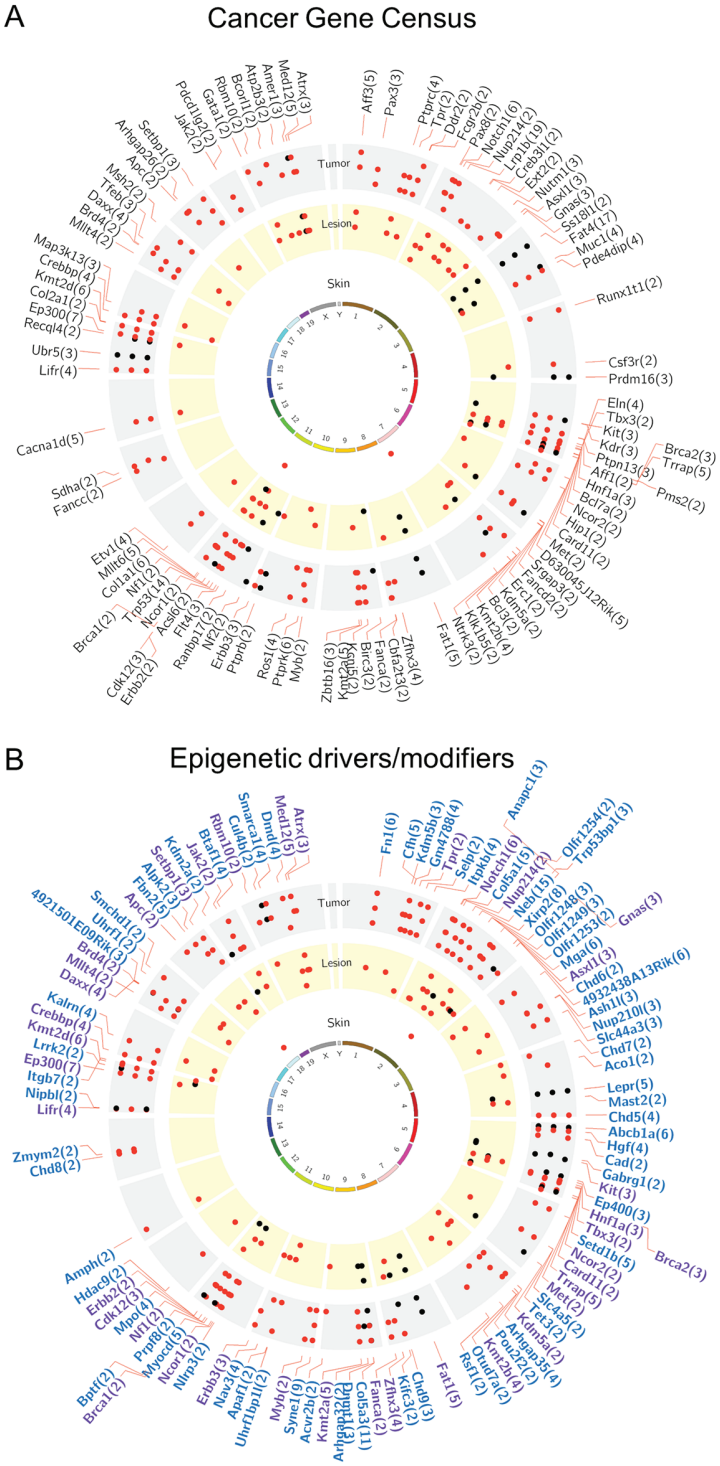


Fig 2. Cancer-related genes recurrently mutated in cSCCs of K14 HPV38 E6/E7 Tg mice. (A) Circos presentation of mutations occurring in the same genes between the different mice. From the centre to the outside, the skin samples (white), the lesion samples (yellow), and the SCC samples (grey) are displayed for $n = 3$ mice each. Each track (three per colour) corresponds to one animal. Red dots represent C:G > T:A mutations, and black dots represent the other types of mutations. For Circos A, only the mutations that occur in genes present in the Cancer Gene Census list from the COSMIC database are displayed, with the number

of recurrent mutations in these genes in parentheses. (B) For the epigenetic drivers/modifiers, only the mutations that occur in the epi-driver or the epi-modifier gene lists are displayed. Blue gene names correspond to genes that are only involved in epigenetic processes, and purple gene names correspond to genes that are involved in epigenetic processes and that are present in the Cancer Gene Census list. The total number of recurrent mutations occurring in each of these genes is also displayed in parentheses.

https://doi.org/10.1371/journal.ppat.1006783.g002

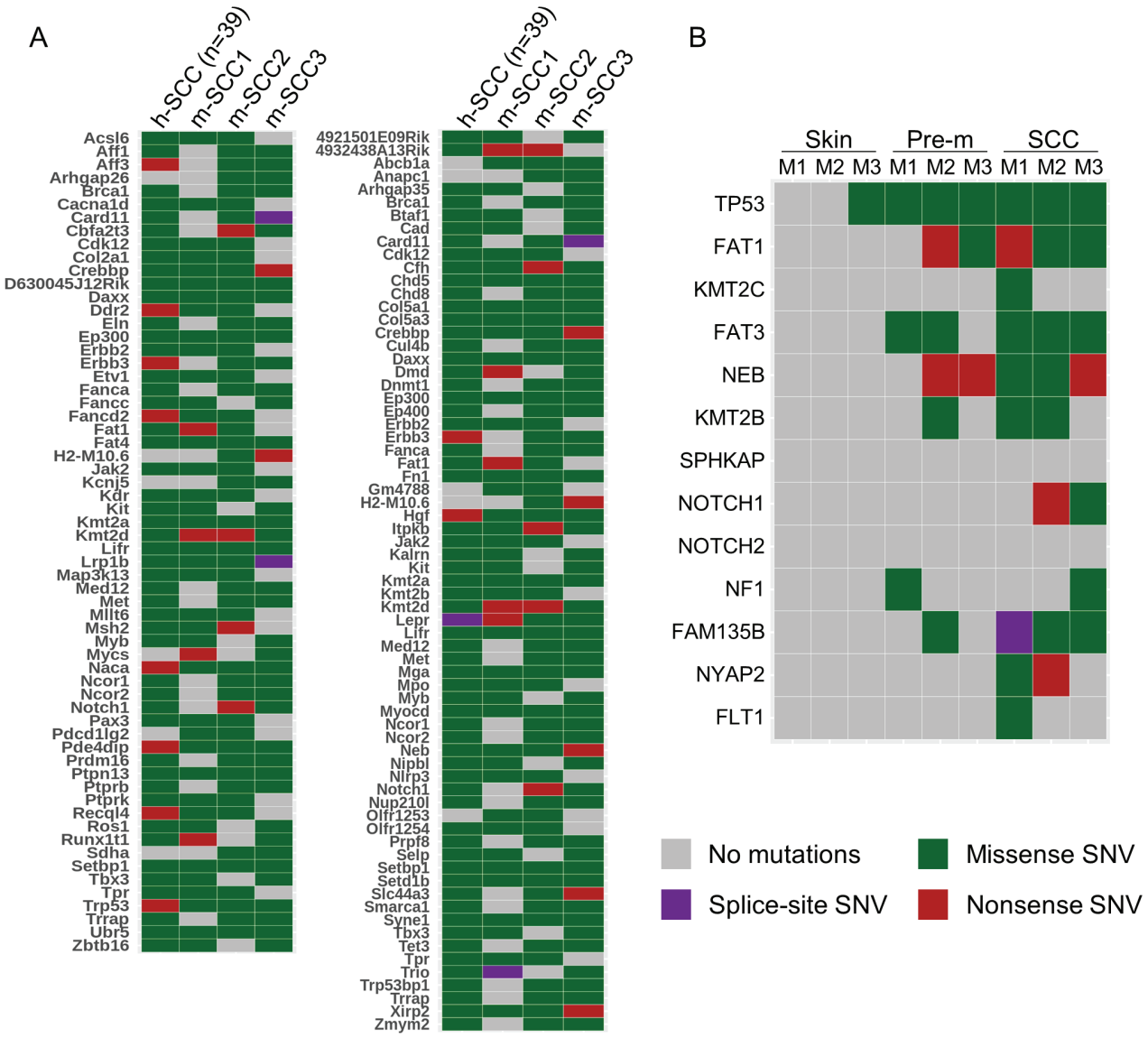


Fig 3. Several genes mutated in human skin lesions are also mutated in the UV-induced skin lesions of cSCCs of K14 HPV38 E6/E7 Tg mice. (A) Heatmap of significantly mutated genes, corresponding to genes recurrently mutated in at least two mouse SCC samples and reported in the Cancer Gene Census list from the COSMIC database (left panel) or having an impact on epigenetic regulation processes (right panel). The types of mutation represented by colours are chosen according to the most prevalent mutation type in each sample. The data for the human samples displayed in the first column are derived from a previous publication on cutaneous SCC. (B) Heatmap of mutations in genes in normal skin, pre-malignant lesions, and cSCC from different mice (M1–3) reported as significantly mutated in human cSCC.

https://doi.org/10.1371/journal.ppat.1006783.g003

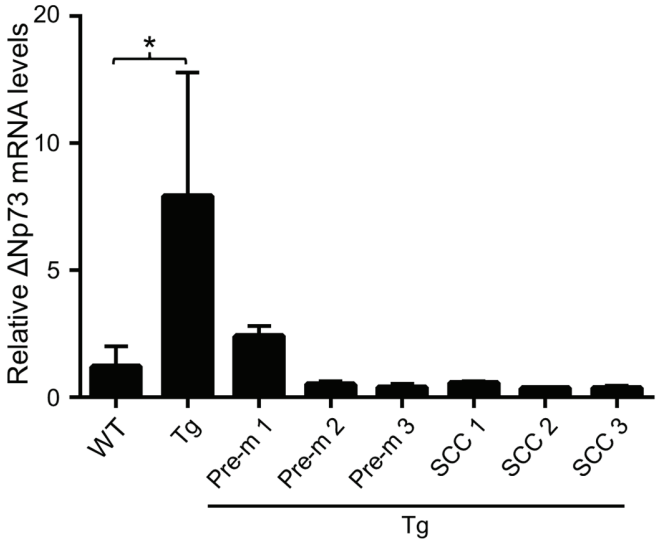


Fig 4. ΔNp73α mRNA levels are high in the skin of HPV38 E6/E7 Tg mice, but are decreased in the UV-induced skin lesions harbouring p53 mutations. Total RNA was extracted from the skin of WT ($n = 4$) or K14 HPV38 E6/E7 Tg animals ($n = 5$) as well as histologically confirmed pre-malignant (pre-m) and SCC from three independent mice and harbouring mutated p53. ΔN73α levels were measured by quantitative RT-PCR. The data shown are the mean of two independent experiments. The differences in ΔN73α mRNA levels between WT and K14 HPV38 E6/E7 Tg animals were statistically significant: * <0.05.

https://doi.org/10.1371/journal.ppat.1006783.g004

Our previous data showed that HPV 38 E6 and E7 expression in human keratinocytes resulted in accumulation of TAp53, which is recruited to the internal promoter located in intron 3 of p53 gene, with resulting transcriptional activation of ΔNp73α [31,32]. Fig 4 shows that also in the mouse skin, expression of the viral genes leads to increased ΔNp73α transcription. In contrast, in histologically confirmed pre-malignant and SCC lesions, p53 mutation correlates with a strong decrease in ΔNp73α mRNA levels (Fig 4).

In conclusion, our findings show that the expression of HPV38 E6 and E7 oncogenes in mouse skin increases susceptibility to UV-induced cSCC by facilitating the accumulation of somatic mutations that have been clearly associated with skin cancer development in humans.

HPV38 E6 and E7 play a role at initial stages of UV-induced skin carcinogenesis but are not required for cancer maintenance

Many studies support the role of beta HPV types, together with UV radiation, in the development of skin SCC [2,3]. However, in contrast to the mucosal high-risk HPV types such HPV16 that are required in all steps of cervical carcinogenesis, beta HPV types appear to have a role in the initial steps of carcinogenesis. To test this hypothesis, we constructed our K14 HPV38 E6/E7 Tg mice as a conditional expression model with two loxP elements, located immediately upstream and downstream of the viral genes [15]. Originally, we crossed the K14 HPV38 E6/E7 Tg mice with K14 Cre-ERT2 Tg animals overexpressing the Cre recombinase gene fused to a triple-mutant form of the human estrogen receptor that gains access to the nuclear compartment only after exposure to 4-hydroxytamoxifen (TMX) but not to the natural ligand 17β-estradiol, in order to silence E6/E7 expression by Cre-mediated deletion of the floxed viral genes at different times of the chronic UV irradiation, i.e., different stages of SCC development. Although the expression of the viral genes could be efficiently silenced upon administration of TMX to 5-week-old K14 Cre-ERT2 HPV38 E6/E7 compound mice, in the compound

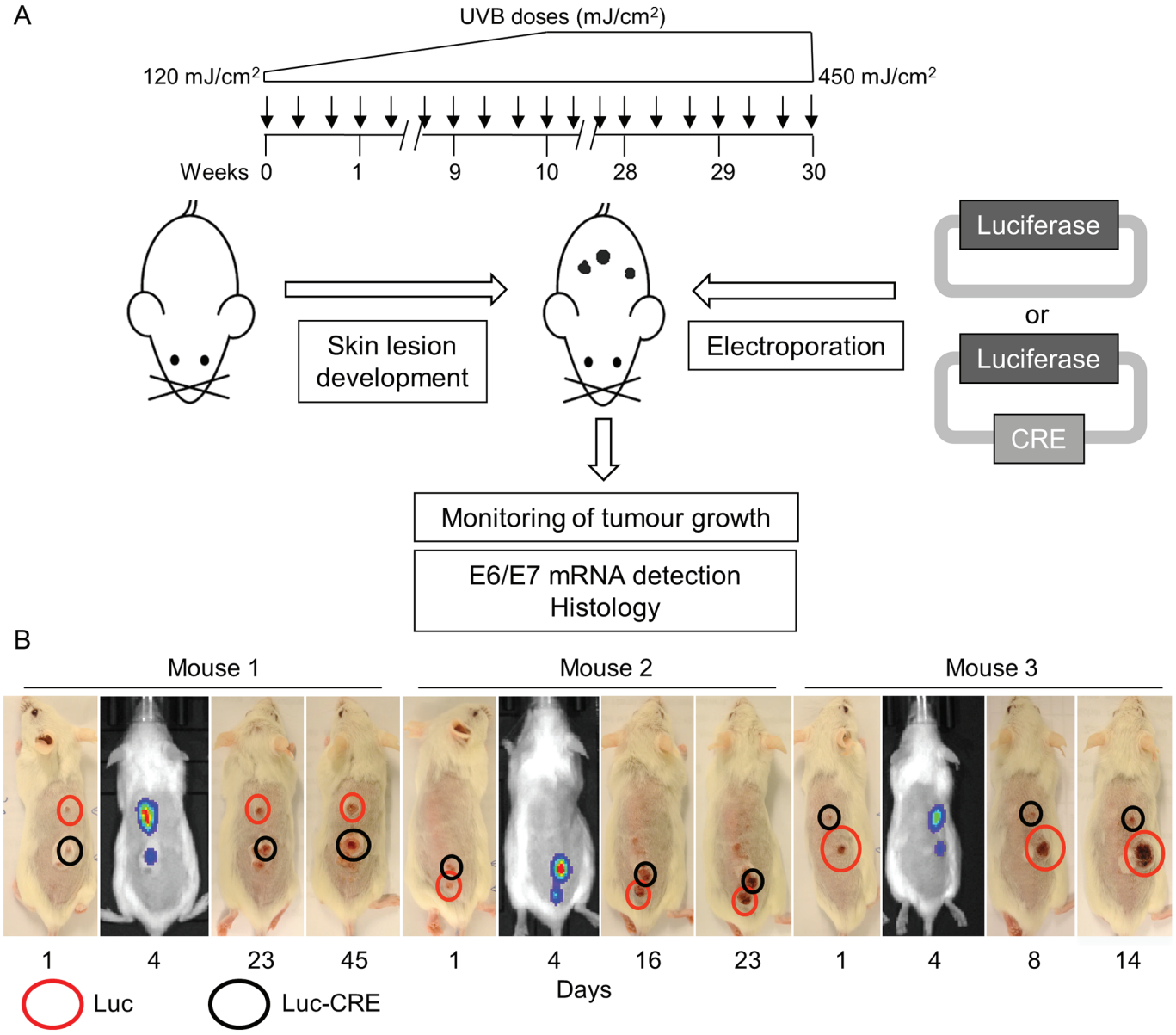


Fig 5. Luciferase expression vectors can be efficiently electroporated into skin lesions of K14 HPV38 E6/E7 Tg mice. (A) Schematic diagram of the electroporation procedure of skin lesions of K14 HPV38 E6/E7 Tg mice. (B) Luciferase activity is detected in lesions electroporated with the control vector (Luc) as well as in lesions electroporated with the plasmid coding for the Cre recombinase and luciferase genes (CRE-Luc). Mice and tumour growth are closely monitored at regular intervals.

https://doi.org/10.1371/journal.ppat.1006783.g005

mice a strong decrease in viral gene expression was observed during the 30 weeks of UV irradiation in the absence of TXM treatment (S3 Fig). The loss of HPV38 E6 and E7 genes in long-term experiments was most likely due to a basal, non-specific Cre recombinase activity in the nucleus of mouse skin keratinocytes. None of the K14 Cre-ERT2 HPV38 E6/E7 Tg compound lines developed cSCC after 30 weeks of UV irradiation, further highlighting the importance of the viral proteins in UV-induced carcinogenesis.

Therefore, we developed a different strategy to evaluate the requirement of HPV38 E6 and E7 genes for cancer maintenance (Fig 5A). K14 HPV38 E6/E7 Tg mice were exposed to long-

term UV irradiation, and after the appearance of well-defined skin lesions, after about 22–25 weeks of irradiation, two different DNA vectors were delivered by electroporation into the abnormal tissues. Because of the small size of the electroporated skin lesions, we could not perform any biopsy; therefore, we did not have any histological information about whether they correspond to pre-malignant or malignant lesions. Results obtained in several independent experiments showed that the lesions that occurred after 22–25 weeks of UV irradiation correspond to pre-malignant lesions or an early stage of cSCC [15,16]. Both vectors contain a scaffold/matrix attachment region (S/MAR) that keeps the plasmid in an episomal state, avoiding any integration-mediated toxicity, and ensures robust and persistent gene expression [33]. The vector codes for luciferase and Cre recombinase genes (Cre-Luc) separated by the P2A cleavage site, whereas the control vector expresses only a luciferase gene (Luc). Luciferase was used to monitor the efficiency of transfection by non-invasive *in vivo* imaging, and Cre was used to induce the excision of the viral genes. A total of 23 lesions on 14 mice were transfected either with the Luc vector ($n = 9$) or with the Cre-Luc vector ($n = 14$). When possible, the same mouse was injected with both vectors, each on a different lesion. Three representative mice are shown in Fig 5B. Luciferase activity was detected in the animals' skin in each of the electroporated areas independently of the vector type.

After electroporation, the animals were irradiated until the end of the 30-week UV irradiation protocol and closely monitored for several weeks to evaluate the progression of the skin lesions. No significant difference in tumour growth was observed in animals transfected with the Luc or Cre-Luc vectors (Fig 6A). Histological analyses confirmed that 100% percent of the Luc-injected lesions and 93% of the Cre-Luc injected lesions (13 out of 14) evolved into invasive cSCC; a morphological examination revealed no major differences between the two groups of tumours (Fig 6B). Detection of the viral RNA transcripts by RNA-RNA *in situ* hybridization confirmed that electroporation of skin lesions with the Cre-Luc vector, but not with the Luc vector, resulted in the loss of E6/E7 expression in large islands of cancer tissue (Fig 6B).

In conclusion, our findings show that after the accumulation of UV-induced DNA mutations and the development of skin lesions, the expression of the HPV38 E6/E7 genes is dispensable for the maintenance of the malignant phenotype of skin cancer cells.

Discussion

Although the HPV family includes more than 200 types, to date only the mucosal high-risk (HR) HPV types have been clearly associated with human carcinogenesis. These viruses are the etiological agents of cervical cancers as well as a subset of other genital and oropharyngeal cancers [34]. Beta HPV types have been proposed to be associated with cSCC. They were initially linked to cSCC in EV patients, but now many epidemiological and biological studies support the role of beta HPV types in skin carcinogenesis also in non-EV individuals [3].

We have previously shown in a Tg mouse model that expression of beta HPV38 E6 and E7 in the skin strongly increases the risk of cSCC development upon UV irradiation [15]. Here, we showed that the higher susceptibility of K14 HPV38 E6/E7 Tg mice to UV-induced skin carcinogenesis tightly correlates with the accumulation of a high number of mutations in the keratinocyte genome. Remarkably, exposure of WT animals to the same doses of UV radiation did not lead to accumulation of DNA mutations and development of cSCC. These data suggest that the HPV38 oncoproteins can negatively affect the DNA repair machinery and/or immune pathways that lead to the elimination of damaged cells. We have recently shown that K14 HPV38 E6/E7 Tg mice are hampered in the production of interleukin 18 (IL-18) during their exposure to UV radiation [16]. Upon UV irradiation and activation of the inflammasome,

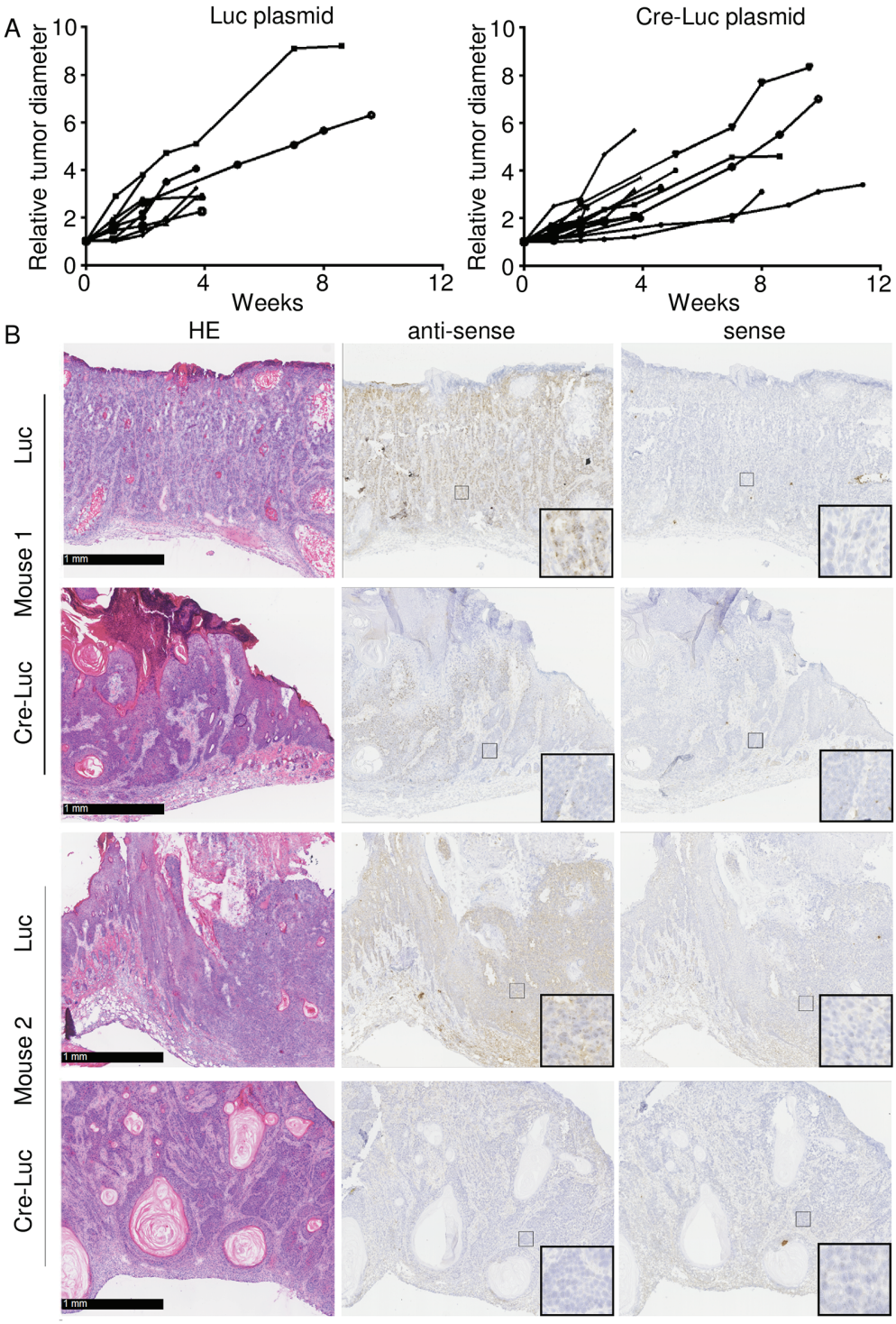


Fig 6. HPV38 E6 and E7 expression is not required for the viability of cancer cells in K14 HPV38 E6/E7 Tg mice. (A) Electroporated lesions were kept under control and the diameter was recorded weekly. On the day of injection, the lesion diameter varied between 1.2 mm and 2.5 mm for the lesions injected with the Luc plasmid, and between 1.3 mm and 2.6 mm for the lesions injected with the Cre-Luc plasmid. To standardize the measurement, each lesion diameter was set to an arbitrary value of 1 on the day of injection, and the following measurements were adjusted accordingly. The difference in tumour growth between the lesions injected with

the Luc plasmid and the lesions injected with the Cre-Luc plasmid was not significant according to an unpaired two-sample Student's *t*-test ($p = 0.3108$, $t = 1.052$; $df = 14$). The test was run on data from the fourth week, because afterwards the number of living animals was substantially reduced. (B) Representative images of SCC sections from two different HPV38 E6/E7 Tg mice. Sections were taken from tumours initially electroporated with pS/MARt-Luc plasmid (Luc) or with pS/MARt-Luc-P2A-Cre plasmid (Cre-Luc). The morphological analysis revealed no substantial differences between the specimens; the tumours were all classified as invasive cSCC, with deep penetration into the dermis or into the muscular fibres, and clear and diffuse atypia. The loss of the viral mRNA in the tumours injected with the Cre-Luc plasmid was confirmed by *in situ* RNA hybridization using a complementary (antisense) riboprobe, while the staining with a sense probe confirmed the specificity of the signal.

https://doi.org/10.1371/journal.ppat.1006783.g006

keratinocytes secrete high levels of cytokines from the IL-1 family, including IL-18, thus inducing a broad spectrum of processes, such as infiltration and activation of inflammatory leukocytes, immunosuppression, DNA repair, and apoptosis [35–38]. Thus, it is likely that the high susceptibility to UV-induced DNA mutations and skin carcinogenesis of K14 HPV38 E6/E7 Tg mice may be linked to the negative impact of HPV38 on IL-18 production.

Analysis of the mutational profile revealed that a large number of genes encoding for epi-drivers or epi-modifiers and proteins known to be associated with carcinogenesis (Cancer Gene Census) harbour missense or nonsense mutations. Most importantly, the gene mutation profile found in murine cSCC shows remarkable similarities to the mutational profile found in human cSCC. In particular, mutations in p53 appear to be an early event in murine and human skin carcinogenesis. We have previously shown that beta HPV38 E7 alters the p53/73 network by inducing accumulation of p53/p73 antagonist ΔNp73α [31,32]. In human keratinocytes expressing beta HPV38 E6 and E7, ΔNp73α forms a transcriptional inhibitory complex, which binds a subset of p53-regulated promoters, preventing their activation in the presence of cellular stress [39]. Because the major role of p53 is to safeguard genome integrity, the high cancer susceptibility of K14 HPV38 E6/E7 Tg mice along with the high numbers of accumulated UV-induced DNA mutations can be explained, at least in part, by the properties of the beta HPV oncoproteins. However, once p53, and likely other cellular genes, are irreversibly inactivated by DNA mutations induced by UV radiation, the progression and maintenance of the skin carcinogenic process could become independent of the expression of viral genes. In agreement with this view, ΔNp73α mRNA levels decrease strongly in UV-induced skin lesions of K14 HPV38 E6/E7 Tg animals after accumulation of p53 mutations. In addition, we observed that the deletion of the HPV38 E6 and E7 genes does not affect further growth of the tumour. In contrast, in K14 Cre-ERT2 HPV38 E6/E7 Tg the loss of the viral genes at early stages of the irradiation protocol prevents the development of UV-induced skin lesions, underlining the key function of HPV38 E6 and E7 in UV-mediated carcinogenesis.

These findings in the K14 HPV38 E6/E7 Tg mouse model are in agreement with the studies on human skin lesions, supporting an early role of beta HPV types in skin carcinogenesis. Indeed, the copy numbers of the beta HPV genome appear to be higher in the pre-malignant lesion, AK, than in cSCC [8]. In addition, not all cancer cells contain a copy of a beta HPV genome [8]. Thus, the mechanisms of carcinogenesis induced by beta HPV types appear to be substantially different from those of the mucosal HR HPV types. In the case of the mucosal HR HPV types, the viral oncoproteins are the major drivers of cancer development (e.g. in the cervix) that, in addition, are required throughout the entire carcinogenic process (Fig 7). In contrast, UV-induced damage is the main carcinogen of cSCC. Here, however, beta HPV oncoproteins can facilitate the accumulation of UV-induced DNA damage but they are dispensable after full development of a malignant lesion (Fig 7).

Why do different HPV types display different biological properties? Cutaneous and mucosal HPV types infect cells at distinct anatomical sites exposed to different environmental

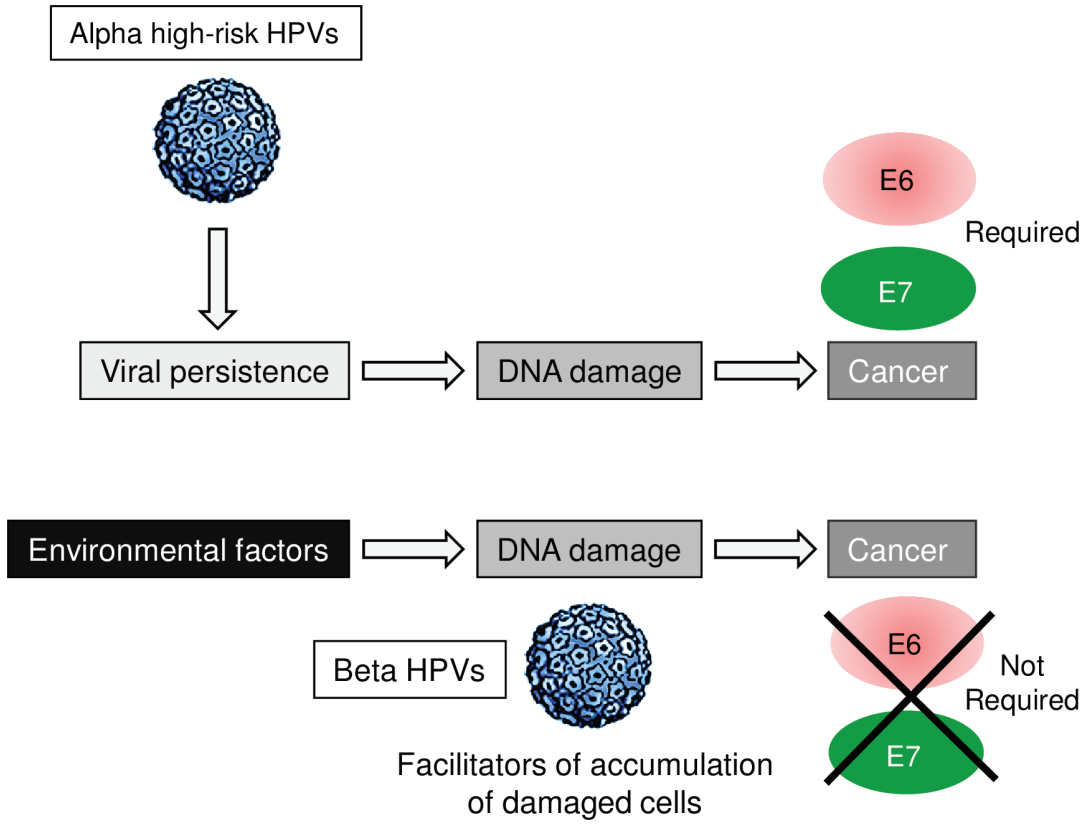


Fig 7. Schematic representation of well-known and hypothetical models of virus-associated carcinogenesis.

https://doi.org/10.1371/journal.ppat.1006783.g007

stresses. Thus, it is not surprising that they have evolved with divergent biological properties. All HPV types rely on the DNA replication machinery of the host cell. Therefore, they must have developed several mechanisms to maintain the infected cell in a proliferative state to guarantee efficient viral genome replication. Exposure of skin keratinocytes to UV radiation leads to accumulation of DNA damage, which in turn induces cell-cycle arrest or apoptosis to allow repair or elimination, respectively, of the damaged cell. The cutaneous HPV types appear to be able to circumvent this adverse effect of UV radiation on keratinocyte proliferation, promoting the accumulation of damaged cells in the skin and, consequently, carcinogenesis.

Our previous findings showed that different HPV38 E6/E7 expression levels in independent Tg lines influence the rate of SCC development [15]. Thus, it plausible to hypothesize that also in humans, the viral gene expression levels may have an impact on UV-induced skin carcinogenesis. Limited data are available on beta E6 and E7 gene expression in normal skin and premalignant and malignant skin lesions (reviewed in [2,3]). There is no information on the different spliced forms of beta HPV genes and how they could determine a different efficiency in protein synthesis. Thus, additional studies are required in humans to corroborate the findings obtained in the Tg mouse model on the hit-and-run mechanism of HPV38 in UV-induced carcinogenesis.

In conclusion, our findings in a Tg mouse model highlight a novel mechanism of infection-associated carcinogenesis, in which the virus is not the driving force but synergizes with UV radiation in promoting cSCC.

Methods

Tg mice

The transgenic animal model FVB/NTgN(38E6E7)187DKFZ (https://mito.dkfz.de/mito/Animal%20line/10954) has been previously described [15]. UVB irradiation was performed under sevoflurane anaesthesia, and every effort was made to minimize suffering.

Ethics statement

The animal facility of the German Cancer Research Center has been officially approved by responsible authority (Regional Council of Karlsruhe, Schlossplatz 4–6, 76131 Karlsruhe, Germany), official approval file number 35–9185.64. Housing conditions are thus in accordance with the German Animal Welfare Act (TierSchG) and EU Directive 425 2010/63/EU. Regular inspections of the facility are conducted by the Veterinary Authority of Heidelberg (Bergheimer Str. 69, 69115 Heidelberg, Germany). All experiments were in accordance with the institutional guidelines (designated veterinarian according to article 25 of Directive 2010/63/EU and Animal-Welfare Body according to article 27 of Directive 2010/63/EU) and were officially approved by Regional Council of Karlsruhe (File No 35–9185.81/G-64/13 and 35–9185.81/G-200/15).

Plasmid construction

To generate the Luc and the Luc-Cre vectors, the pS/MARt-GFP DNA vector was first digested with the restriction enzymes NheI and BglII to linearize the vector and eliminate the transgene GFP. The InFusion system provided by Clonetech was used to introduce the luciferase gene alone or in combination with the Cre recombinase gene to generate the vector pS/MARt-Luc or the vector pS/MARt-Luc-P2A-Cre, respectively.

UVB treatments

UVB irradiation was performed with a Bio-Spectra system (Vilber Lourmat, Marne La Vallee, France) at a wavelength of 312 nm as previously described [15]. Briefly, animals were anesthetized with 3% Sevorane (Abbott, Wiesbaden, Germany) in an inhalation anesthetizer (Provet, Lyssach, Switzerland) and placed in a covered compartment with an upper square opening (3×2 cm) at a distance of 40 cm from the UVB lamp.

To study UV-induced carcinogenesis, 7-week-old female FVB/N WT or K14 HPV38 E6/E7 Tg animals were shaved on the dorsal skin with electric clippers and irradiated 3 times a week for 10 weeks with increasing doses of UVB, starting from 120 mJ/cm^2 to a final dose of 450 mJ/cm^2, with a constant weekly increase to allow skin thickening. For the following 20 weeks, mice were irradiated 3 times a week with 450 mJ/cm^2. The UV irradiation protocol was based on the data described in [40] and to mimic the situation in humans. For instance, the maximum dose of the UV irradiation protocol, 450 mJ/cm^2, corresponds to 50 minutes of solar exposure in July in Paris. The tumour incidence (tumour bearers/group) was recorded weekly. Tumours were identified first macroscopically and by histological diagnosis. After 30 weeks, or earlier if the tumour reached the ethically allowed maximal size, the animals were sacrificed and H&E-stained sections of dorsal skin were used for histological diagnosis.

Excision of floxed viral transgenes

To study the effect of the loss of the viral genes on skin cancer development, 7-week-old K14 HPV38 E6/E7 Tg mice (n = 14) were shaved on the dorsal skin and treated for 30 weeks with

increasing doses of UVB as previously described [15]. As soon as skin lesions (maximum diameter 2.6 mm) became evident, 46 μg of pS/MARt-Luc or 50 μg of pS/MARt-Luc-P2A-Cre dissolved in isotonic saline solution was injected directly into the lesions. To facilitate the uptake of the injected DNA, an electric field was applied to the area of the injection site using a Tweezertrodes connected to a BTX ECM 630 generator (Harvard Apparatus, Holliston, MA, USA). A first high-voltage electric pulse (1400 V/cm, 100 μs, 2 times), to induce temporary gaps in the keratinocytes cell membrane, was followed by a low-voltage electric field (140 V/cm, 400 ms, 2 times), to facilitate the migration of the DNA into the cells. At 72 h after the DNA injection, the mice were injected intraperitoneally with 150 mg/kg of luciferin in sterile water, and the luciferase activity was then assessed using an IVIS Lumina III imaging system (Perkin Elmer, Rodgau, Germany). When possible, a single mouse received both plasmids at the same time, each on a different lesion. The UV irradiation continued until week 30, according to the protocol [15]. The lesions were then closely monitored and the animals were sacrificed in accordance with an ethical protocol to avoid animal suffering. Skin lesions were collected for histological examination and detection for HPV38 E6/E7 RNA by *in situ* hybridization.

Total RNA isolation and reverse transcription PCR analyses

Total RNA was isolated from dorsal skin of WT ($n = 4$) or K14 HPV38 E6/E7 Tg animals ($n = 5$) as well as histologically confirmed pre-malignant (pre-m) and SCC from three independent mice. cDNA was synthesized from 1 μg of total RNA using M-MLV reverse transcriptase (Invitrogen, Darmstadt, Germany), and a mix of random hexamers were used as primers. Quantitative reverse transcription PCR (RT-qPCR) was performed in a 20 μl mixture containing 1 μl of 1:10 diluted cDNA and Mesa green quantitative PCR (qPCR) Master Mix (Eurogentec, Angers, France) with specific mouse ΔNp73α primers (5′-GCCAAAAGGGTCATCATC-3′ and 5′-TGCCAGTGAGCTTCCCGTTC-3′) or mouse GAPDH primers to amplify a housekeeping gene as internal control (5′-GTGACCCCATGAGACACCTC-3′ and 5′-GTATGTC CAGGTGGCCGAC–3′), using an Applied Biosystems 7300 machine (Applied Biosystems, Darmstadt, Germany). The fluorescence threshold value was calculated using the SDS analysis software from Applied Biosystems.

In situ hybridization

Once the tumours reached the maximum ethically allowed size, the mice were killed and the lesions isolated. Half of the lesion was embedded in OCT medium and slowly cooled down to −80˚C. Sense and antisense riboprobes were generated from linearized plasmid DNA containing full-length HPV38E6E7 cDNA using the Digoxigenin RNA labelling Mix from Roche. RNA-RNA *in situ* hybridization was performed as previously described[41]. In brief, serial 5 μm cryo-sections were mounted on Superfrost Plus slides (Thermo Scientific), fixed in 4% paraformaldehyde in 2× SSPE, digested with proteinase K (0.5 μg/ml), and pre-hybridized at 42˚C for 2–4 h. Hybridization was performed overnight at 42˚C in 50% formamide, 2× SSPE, 10% dextran sulfate, 10 mM Tris-HCl pH 7.5, 1× Denhardt's solution, 500 μg/ml tRNA, 100 μg/ml herring sperm DNA, 0.1% SDS, and 10 μg/ml DIG-labelled riboprobe. After hybridization, slides were washed once in 50% formamide, 2× SSPE; 0.1% SDS for 30 min at 50˚C, treated with RNaseA (50 μg/ml in 2× SSC, 0.1% SDS), and washed again in 50% formamide, 0.5× SSPE, 0.1% SDS for 30 min at 37˚C. Hybridization signals were visualized using Biotin Tyramide (TSA Biotin System, PerkinElmer) according to the manufacturer's protocol.

Statistical analysis

Tumour growth values of lesions injected with the pS/MARt-Luc or pS/MARt-Luc-P2A-Cre vector were compared with the two-sample *t*-test. The statistical analysis was performed with GraphPad Prism (version 6, GraphPad Software Inc., La Jolla, CA, USA).

Exome analysis

The quality of the raw reads was estimated with FastQC software (version 0.11.5, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped to the GRCm38 Mouse reference genome (ftp://hgdownload.cse.ucsc.edu/goldenPath/mm10/) using Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net/) version 0.7.15 and producing a BAM file. The following GATK Best Practice Recommendations were applied to the BAM files to improve variant detection quality. Picard (version 2.4.1, https://broadinstitute.github.io/picard/) SortSAM was used to sort and index BAM files, and the AddOrReplaceReadGroups tool was used to replace all read groups with a single new read group. The duplicate reads were marked with the MarkDuplicates tool from Picard, and the newly produced BAM file was indexed with the BuildBamIndex tool. GATK (version 3.6.0, https://software.broadinstitute.org/gatk/download/) RealignerTargetCreator was used to determine the position concerned by local realignment, and IndelRealigner was used to perform local realignment around these sites. The GATK BaseRecalibrator tool was used to detect systematic errors in base quality scores. Dbsnp and dbindel (version 142) for the mm10 reference genome was downloaded from the Sanger website (ftp://ftp-mouse.sanger.ac.uk/REL-1505-SNPs_Indels/) and considered as input. Lastly, the index of the output BAM file was created with Picard BuildBamIndex, and GATK PrintReads was used to write out sequence read data.

The quality of the alignment was estimated with Qualimap (version 2.0.2, http://qualimap.bioinfo.cipf.es/). Then the variant calling was done with Mutect (version 1.1.7, http://archive.broadinstitute.org/cancer/cga/mutect), by using a skin sample from a WT mouse not exposed to UV as the "normal sample" for paired analysis. Only somatic mutations passing Mutect internal filters were considered for the analysis. The VCF files are annotated with Annovar by using the MutSpec Annot Tool in Galaxy [42]. Variants were then filtered based on SegDup databases from UCSC (version from 4 May 2014, http://hgdownload.cse.ucsc.edu/goldenPath/mm10/database/genomicSuperDups.txt.gz), as well as Tandem Repeat and Repeat Masker (version from 9 February 2012, http://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/). House-made scripts were then used to keep only SNPs that have a functional impact and fall in exonic or splicing regions. Non-negative matrix factorization mutational signatures were inferred with MutSpec-NMF tools, as previously reported.

The pathway analysis was performed using the EnrichR web application (http://amp.pharm.mssm.edu/Enrichr/; citations*2). The input gene list was made by merging the mutations detected in the pre-malignant lesions ($n = 3$) or cSCCs ($n = 3$) of the K14 HPV38 E6/E7 Tg animals. The analysis included only genes harbouring mutations that are likely to alter the biological properties of the encoded products, i.e., 3111 genes in the pre-malignant lesions and 6372 genes in the cSCCs. The gene lists were then loaded into the EnrichR software, and the result from the KEGG database (version 2016) was considered. Only pathways with a significant adjusted *p*-value are shown in S1 Table. The list of pathways is ranked by combined score (combined score is computed by taking the log of the *p*-value from the Fisher exact test and multiplying it by the *z*-score of the deviation from the expected rank).

Comparison with epigenetic driver/modifier genes and Cancer Gene Census list

The list of epigenetic driver and modifier genes was constructed on the basis of genes reported in different publications [19–23]. The Cancer Gene Census list was downloaded from the COSMIC website (12 November 2016, http://cancer.sanger.ac.uk/census) and is based on a previous publication [24].

The comparison of the mouse data with the human data [25,26] was done with Bioconductor (release 3.4, https://www.bioconductor.org/) in R (version 3.3.2, "Sincere Pumpkin Patch"). The module BioMart[43,44], version 2.3 enables the conversion of nearly 87.86% of human gene names from the Chitsazzadeh et al. publication [26] to their corresponding mouse gene names.

Supporting information

S1 Fig. Representative images of H&E-stained sections from WT or Tg mice from which the genomic DNA was extracted for exome sequencing. (A, B) Normal skin from WT (A) and K14 HPV38 E6/E7 Tg (B) mice UV-irradiated for 30 and 28 weeks, respectively. Both specimens show a clearly intact epithelium composed of a few layers of keratinocytes. (C, D) Pre-cancerous lesions from K14 HPV38 E6/E7 Tg mice UV-irradiated for 26 (C) and 28 (D) weeks, respectively. In both lesions, the keratinocytes present acanthosis, diffused intraepithelial atypia, and a high number of mitosis; an intact basal membrane is evident. Enlargements of the most affected areas are displayed. (E, F). Cancerous lesions (SCC) from K14 HPV38 E6/E7 Tg mice UV-irradiated for 26 (E) and 28 (F) weeks, respectively. Both sections are characterized by the presence of polymorphic tumour cells with big nuclei, diffused presence of horn pearls, and hyperkeratinization. The enlargements show tumour invasion of the subcutaneous fat (E) or of muscle fibres (F). The stained sections were first scanned with no enlargement and then zoomed in via software analysis.
(TIF)

S2 Fig. Mutational signature detected in skin keratinocytes of UV-irradiated K14 HPV38 E6/E7 Tg mice. (A) Mutational signature obtained after applying the NMF method to all 9 samples (3 normal skin, 3 pre-malignant lesions, and 3 SCCs). (B) The B signature shows a strong identity with the UV signature (cosine similarity of 0.86). (C) The SCC and pre-malignant samples of the different mice are the main contributors to inference of the B signature.
(TIF)

S3 Fig. Modulation of HPV38 E6 and E7 expression in skin keratinocytes of K14 Cre-ERT2 HPV38 E6/E7 Tg mice. Total RNA was extracted from dorsal skin keratinocytes from K14 HPV38 E6/E7 Tg mice (10-week-old, $n = 3$; 30-week-old, $n = 3$), and from Cre-ERT2 HPV38 E6/E7 Tg mice, treated (10-week-old, $n = 9$; 30-week-old, $n = 4$) or not (10-week-old, $n = 7$; 30-week-old, $n = 4$)) with 4-hydroxytamoxifen (TMX). HPV38 E6 mRNA quantification was performed by quantitative RT-PCR. The relative quantification + SD is shown. The following differences are statistically significant according to *t*-test analysis: 10-week-old K14 HPV38 E6/E7 Tg vs 10-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg, $p < 0.05$; 10-week-old K14 HPV38 E6/E7 Tg vs 30-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg, $p < 0.0001$; 10-week-old K14 HPV38 E6/E7 Tg vs 10-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg + TMX, $p < 0.0001$; 10-week-old K14 HPV38 E6/E7 Tg vs 30-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg + TMX, $p < 0.0001$; 10-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg vs 10-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg + TMX, $p < 0.0001$; 10-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg vs 30-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg, $p < 0.01$; 30-week-old K14 HPV38 E6/E7 Tg vs

30-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg, $p < 0.05$; 30-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg vs 30-week-old K14 Cre-ERT2 HPV38 E6/E7 Tg + TMX, $p < 0.01$.
(TIF)

S1 Table. Mutated genes in animals without skin lesions. Cellular pathways were linked to the different gene products using the information at http://www.genecards.org.
(DOCX)

S2 Table. Global view of the somatic mutations and coverage of the sequencing of skin samples from different mice (M1–3).
(DOCX)

S3 Table. Pathway analyses. The pathway deregulated in the pre-malignant lesions (A) or cSCC (B). The gene list used as input is the consensus of the genes mutated in the different pre-malignant samples. Only the significant pathways (adjusted p-value > 0.05) are shown.
(DOCX)

S4 Table. Trp53 nonsynonymous mutations in the DNA-binding domain detected in normal skin, pre-malignant lesions, and cSCCs.
(DOCX)

Acknowledgments

We are grateful to Dr Christopher P. Wild for his constant support and to all members of our laboratories at DKFZ and IARC for their cooperation. We thank the High-Throughput Sequencing unit of the Genomics & Proteomics Core Facility, German Cancer Research Center (DKFZ), for providing excellent sequencing services. We are also grateful to Nicole Suty for her help with preparation, and Dr Karen Müller for editing this manuscript.

Author Contributions

Conceptualization: Daniele Viarisio, Lutz Gissmann, Massimo Tommasino.

Data curation: Alexis Robitaille, Christa Flechtenmacher, Catherine Voegele, Maude Ardin, Lutz Gissmann, Massimo Tommasino.

Formal analysis: Daniele Viarisio, Karin Müller-Decker, Rosita Accardi, Alexis Robitaille, Christa Flechtenmacher, Catherine Voegele, Maude Ardin.

Funding acquisition: Daniele Viarisio, Lutz Gissmann, Massimo Tommasino.

Investigation: Daniele Viarisio, Karin Müller-Decker, Rosita Accardi.

Methodology: Daniele Viarisio, Karin Müller-Decker, Rosita Accardi, Katrin Beer, Lars Jansen, Matthias Bozza, Richard Harbottle.

Project administration: Daniele Viarisio, Massimo Tommasino.

Supervision: Lutz Gissmann, Massimo Tommasino.

Writing – original draft: Daniele Viarisio, Lutz Gissmann, Massimo Tommasino.

Writing – review & editing: Karin Müller-Decker, Matthias Dürst, Richard Harbottle, Jiri Zavadil, Sandra Caldeira, Lutz Gissmann, Massimo Tommasino.

References

1. Xiang F, Lucas R, Hales S, Neale R. Incidence of nonmelanoma skin cancer in relation to ambient UV radiation in white populations, 1978–2012: empirical relationships. JAMA dermatology. 2014; 150 (10):1063–71. https://doi.org/10.1001/jamadermatol.2014.762 PMID: 25103031
2. Howley PM, Pfister HJ. Beta genus papillomaviruses and skin cancer. Virology. 2015; 479–480:290–6. https://doi.org/10.1016/j.virol.2015.02.004 PMID: 25724416
3. Tommasino M. The biology of beta human papillomaviruses. Virus Res. 2017; 231:128–38. https://doi.org/10.1016/j.virusres.2016.11.013 PMID: 27856220
4. Orth G. Host defenses against human papillomaviruses: lessons from epidermodysplasia verruciformis. Current topics in microbiology and immunology. 2008; 321:59–83. PMID: 18727487
5. Boyle J, MacKie RM, Briggs JD, Junor BJ, Aitchison TC. Cancer, warts, and sunshine in renal transplant patients. A case-control study. Lancet (London, England). 1984; 1(8379):702–5.
6. Kiviat NB. Papillomaviruses in non-melanoma skin cancer: epidemiological aspects. Semin Cancer Biol. 1999; 9(6):397–403. https://doi.org/10.1006/scbi.1999.0143 PMID: 10712886
7. Chahoud J, Semaan A, Chen Y, Cao M, Rieber AG, Rady P, et al. Association between beta-genus human papillomavirus and cutaneous squamous cell carcinoma in immunocompetent individuals-a meta-analysis. JAMA dermatology. 2016; 152(12):1354–64. https://doi.org/10.1001/jamadermatol.2015.4530 PMID: 26720285
8. Weissenborn SJ, Nindl I, Purdie K, Harwood C, Proby C, Breuer J, et al. Human papillomavirus-DNA loads in actinic keratoses exceed those in non-melanoma skin cancers. J Invest Dermatol. 2005; 125(1):93–7. https://doi.org/10.1111/j.0022-202X.2005.23733.x PMID: 15982308
9. Preston DS, Stern RS. Nonmelanoma cancers of the skin. The New England journal of medicine. 1992; 327(23):1649–62. https://doi.org/10.1056/NEJM199212033272307 PMID: 1435901
10. Armstrong BK, Kricker A. The epidemiology of UV induced skin cancer. Journal of photochemistry and photobiology B, Biology. 2001; 63(1–3):8–18. PMID: 11684447
11. Ananthaswamy HN, Loughlin SM, Cox P, Evans RL, Ullrich SE, Kripke ML. Sunlight and skin cancer: inhibition of p53 mutations in UV-irradiated mouse skin by sunscreens. Nature medicine. 1997; 3 (5):510–4. PMID: 9142118
12. Michel A, Kopp-Schneider A, Zentgraf H, Gruber AD, de Villiers EM. E6/E7 expression of human papillomavirus type 20 (HPV-20) and HPV-27 influences proliferation and differentiation of the skin in UV-irradiated SKH-hr1 transgenic mice. J Virol. 2006; 80(22):11153–64. https://doi.org/10.1128/JVI.00954-06 PMID: 16971438
13. Dong W, Kloz U, Accardi R, Caldeira S, Tong WM, Wang ZQ, et al. Skin hyperproliferation and susceptibility to chemical carcinogenesis in transgenic mice expressing E6 and E7 of human papillomavirus type 38. J Virol. 2005; 79(23):14899–908. https://doi.org/10.1128/JVI.79.23.14899-14908.2005 PMID: 16282489
14. Schaper ID, Marcuzzi GP, Weissenborn SJ, Kasper HU, Dries V, Smyth N, et al. Development of skin tumors in mice transgenic for early genes of human papillomavirus type 8. Cancer Res. 2005; 65 (4):1394–400. https://doi.org/10.1158/0008-5472.CAN-04-3263 PMID: 15735026
15. Viarisio D, Mueller-Decker K, Kloz U, Aengeneyndt B, Kopp-Schneider A, Grone HJ, et al. E6 and E7 from beta HPV38 cooperate with ultraviolet light in the development of actinic keratosis-like lesions and squamous cell carcinoma in mice. PLoS Pathog. 2011; 7(7):e1002125. https://doi.org/10.1371/journal.ppat.1002125 PMID: 21779166
16. Viarisio D, Muller-Decker K, Zanna P, Kloz U, Aengeneyndt B, Accardi R, et al. Novel ss-HPV49 transgenic mouse model of upper digestive tract cancer. Cancer Res. 2016; 76(14):4216–25. https://doi.org/10.1158/0008-5472.CAN-16-0370 PMID: 27216183
17. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500(7463):415–21. https://doi.org/10.1038/nature12477 PMID: 23945592
18. Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallee MP, et al. Modelling mutational landscapes of human cancers in vitro. Scientific reports. 2014; 4:4482. https://doi.org/10.1038/srep04482 PMID: 24670820
19. Timp W, Feinberg AP. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nat Rev Cancer. 2013; 13(7):497–510. https://doi.org/10.1038/nrc3486 PMID: 23760024
20. Shen H, Laird PW. Interplay between the cancer genome and epigenome. Cell. 2013; 153(1):38–55. https://doi.org/10.1016/j.cell.2013.03.008 PMID: 23540689

21. Sturm D, Bender S, Jones DT, Lichter P, Grill J, Becher O, et al. Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. Nat Rev Cancer. 2014; 14(2):92–107. https://doi.org/10.1038/nrc3655 PMID: 24457416

22. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr., Kinzler KW. Cancer genome landscapes. Science. 2013; 339(6127):1546–58. https://doi.org/10.1126/science.1235122 PMID: 23539594

23. Gonzalez-Perez A, Jene-Sanz A, Lopez-Bigas N. The mutational landscape of chromatin regulatory factors across 4,623 tumor samples. Genome biology. 2013; 14(9):r106. https://doi.org/10.1186/gb-2013-14-9-r106 PMID: 24063517

24. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004; 4(3):177–83. https://doi.org/10.1038/nrc1299 PMID: 14993899

25. Pickering CR, Zhou JH, Lee JJ, Drummond JA, Peng SA, Saade RE, et al. Mutational landscape of aggressive cutaneous squamous cell carcinoma. Clin Cancer Res. 2014; 20(24):6582–92. https://doi.org/10.1158/1078-0432.CCR-14-1768 PMID: 25303977

26. Chitsazzadeh V, Coarfa C, Drummond JA, Nguyen T, Joseph A, Chilukuri S, et al. Cross-species identification of genomic drivers of squamous cell carcinoma development across preneoplastic intermediates. Nature communications. 2016; 7:12601. https://doi.org/10.1038/ncomms12601 PMID: 27574101

27. Jonason AS, Kunala S, Price GJ, Restifo RJ, Spinelli HM, Persing JA, et al. Frequent clones of p53-mutated keratinocytes in normal human skin. Proc Natl Acad Sci U S A. 1996; 93(24):14025–9. PMID: 8943054

28. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015; 348 (6237):880–6. PMID: 25999502

29. Lowell S, Jones P, Le Roux I, Dunne J, Watt FM. Stimulation of human epidermal differentiation by delta-notch signalling at the boundaries of stem-cell clusters. Current biology: CB. 2000; 10(9):491–500. PMID: 10801437

30. Rangarajan A, Talora C, Okuyama R, Nicolas M, Mammucari C, Oh H, et al. Notch signaling is a direct determinant of keratinocyte growth arrest and entry into differentiation. The EMBO journal. 2001; 20 (13):3427–36. https://doi.org/10.1093/emboj/20.13.3427 PMID: 11432830

31. Accardi R, Dong W, Smet A, Cui R, Hautefeuille A, Gabet AS, et al. Skin human papillomavirus type 38 alters p53 functions by accumulation of deltaNp73. EMBO Rep. 2006; 7(3):334–40. https://doi.org/10.1038/sj.embor.7400615 PMID: 16397624

32. Accardi R, Scalise M, Gheit T, Hussain I, Yue J, Carreira C, et al. IkappaB kinase beta promotes cell survival by antagonizing p53 functions through DeltaNp73alpha phosphorylation and stabilization. Mol Cell Biol. 2011; 31(11):2210–26. https://doi.org/10.1128/MCB.00964-10 PMID: 21482671

33. Piechaczek C, Fetzer C, Baiker A, Bode J, Lipps HJ. A vector based on the SV40 origin of replication and chromosomal S/MARs replicates episomally in CHO cells. Nucleic Acids Res. 1999; 27(2):426–8. PMID: 9862961

34. Tommasino M. The human papillomavirus family and its role in carcinogenesis. Semin Cancer Biol. 2014; 26:13–21. https://doi.org/10.1016/j.semcancer.2013.11.002 PMID: 24316445

35. Latz E, Xiao TS, Stutz A. Activation and regulation of the inflammasomes. Nature reviews Immunology. 2013; 13(6):397–411. https://doi.org/10.1038/nri3452 PMID: 23702978

36. Nasti TH, Timares L. Inflammasome activation of IL-1 family mediators in response to cutaneous photodamage. Photochemistry and photobiology. 2012; 88(5):1111–25. https://doi.org/10.1111/j.1751-1097.2012.01182.x PMID: 22631445

37. Schwarz A, Maeda A, Stander S, van Steeg H, Schwarz T. IL-18 reduces ultraviolet radiation-induced DNA damage and thereby affects photoimmunosuppression. J Immunol. 2006; 176(5):2896–901. PMID: 16493047

38. Schwarz T, Schwarz A. DNA repair and cytokine responses. The journal of investigative dermatology Symposium proceedings. 2009; 14(1):63–6. https://doi.org/10.1038/jidsymp.2009.3 PMID: 19675557

39. Saidj D, Cros MP, Hernandez-Vargas H, Guarino F, Sylla BS, Tommasino M, et al. Oncoprotein E7 from beta human papillomavirus 38 induces formation of an inhibitory complex for a subset of p53-regulated promoters. J Virol. 2013; 87(22):12139–50. https://doi.org/10.1128/JVI.01047-13 PMID: 24006445

40. Jeanmougin M, Civatte J. [Dosimetry of solar ultraviolet radiation. Daily and monthly changes in Paris]. Annales de dermatologie et de venereologie. 1987; 114(5):671–6. PMID: 3631842

41. Durst M, Glitz D, Schneider A, zur Hausen H. Human papillomavirus type 16 (HPV 16) gene expression and DNA replication in cervical neoplasia: analysis by in situ hybridization. Virology. 1992; 189(1):132–40. PMID: 1318602

42. Ardin M, Cahais V, Castells X, Bouaoun L, Byrnes G, Herceg Z, et al. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. BMC bioinformatics. 2016; 17:170. https://doi.org/10.1186/s12859-016-1011-z PMID: 27091472

43. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009; 4(8):1184–91. https://doi.org/10.1038/nprot.2009.97 PMID: 19617889

44. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics (Oxford, England). 2005; 21(16):3439–40.

Virology 538 (2019) 53–60

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/virology

Cancer susceptibility of beta HPV49 E6 and E7 transgenic mice to 4-nitroquinoline 1-oxide treatment correlates with mutational signatures of tobacco exposure

Daniele Viarisio[a], Alexis Robitaille[b], Karin Müller-Decker[a], Christa Flechtenmacher[c], Lutz Gissmann[a,d], Massimo Tommasino[b,*]

[a] Deutsches Krebsforschungszentrum (DKFZ), Im Neuenheimer Feld 280, 69120, Heidelberg, Germany
[b] International Agency for Research on Cancer (IARC), World Health Organization, 150 Cours Albert Thomas, 69372, Lyon Cedex 08, France
[c] Department of Pathology, University Hospital of Heidelberg, Im Neuenheimer Feld 220, 69120, Heidelberg, Germany
[d] Department of Botany and Microbiology (honorary MMember), King Saud University, Riyadh, Saudi Arabia

ARTICLE INFO

Keywords:
Beta HPV types
Upper-digestive tract cancer
4-Nitroquinoline 1-oxide
Mutational signatures of tobacco exposure

ABSTRACT

We have previously showed that a transgenic (Tg) mouse model with cytokeratin 14 promoter (K14)-driven expression of E6 and E7 from beta-3 HPV49 in the basal layer of the epidermis and of the mucosal epithelia of the digestive tract (K14 HPV49 E6/E7 Tg mice) are highly susceptible to upper digestive tract carcinogenesis upon exposure to 4-nitroquinoline 1-oxide (4NQO). Using whole-exome sequencing, we show that in K14 HPV49 E6/E7 Tg mice, development of 4NQO-induced cancers tightly correlates with the accumulation of somatic mutations in cancer-related genes. The mutational signature in 4NQO-treated mice was similar to the signature observed in humans exposed to tobacco smoking and tobacco chewing. Similar results were obtained with K14 Tg animals expressing mucosal high-risk HPV16 E6 and E7 oncogenes. Thus, beta-3 HPV49 share some functional similarities with HPV16 in Tg animals.

1. Introduction

Human papillomaviruses (HPV) are a large family of double-stranded DNA viruses that infect the mucosal and cutaneous epithelia. They are classified in a phylogenetic tree in genera, based on nucleotide sequence homology of the major capsid protein L1 (Van Doorslaer et al., 2013). Genus alpha HPV types have been most studied so far, because a subgroup, the mucosal alpha high-risk (HR) HPV types, is responsible for the development of cervical cancer and a subset of other anogenital and oropharyngeal cancers. The products of two early genes, E6 and E7, are the key oncoproteins of HPV (Tommasino, 2014). In addition to alpha HPV types, genus beta HPVs also appear to be associated with human carcinogenesis. Beta HPV types are subdivided into five different species (beta-1–5), of which beta-1 and beta-2 are the largest subgroups (Van Doorslaer et al., 2011). They are abundantly present on the skin, and many findings support their role, together with ultraviolet (UV) radiation, in the development of cutaneous squamous cell carcinoma (cSCC) (Rollison et al., 2019). Accordingly, mechanistic studies have well demonstrated the transforming properties of E6 and E7 from a few beta-1 and beta-2 HPV types (e.g. HPV8 and HPV38) in *in vitro* and *in vivo* experimental models (reviewed in refs. (Tommasino, 2017; Hasche et al., 2018). In particular, these viral oncoproteins are able to promote proliferation and to circumvent cellular stresses induced by UV radiation. These findings indicate that in the context of the natural infection, beta HPV E6/E7 expression keeps cells alive despite the accumulation of UV-induced DNA mutations. As a consequence, beta HPV-infected keratinocytes may acquire a high probability of progressing towards cellular transformation. Thus, beta HPVs act as facilitators of the accumulation of UV-induced DNA mutations, but they are not the main drivers. In agreement with this model, findings indicate that beta HPV types are necessary at an early stage of carcinogenesis and are dispensable for the maintenance of the cancer phenotype (Rollison et al., 2019).

In addition to the skin, beta HPV types can be found at other anatomical sites, including the mucosal epithelia (Bottalico et al., 2011; Forslund et al., 2013; Hampras et al., 2017; Pierce Campbell et al., 2013; Torres et al., 2015). In particular, beta-3 HPV types, i.e. HPV types 49, 75, 76, and 115, appear to preferentially infect the mucosal

* Corresponding author. Infections and Cancer Biology Group, International Agency for Research on Cancer, World Health Organization, 150 Cours Albert Thomas, 69372, Lyon Cedex 08, France.
E-mail address: tommasino@iarc.fr (M. Tommasino).

https://doi.org/10.1016/j.virol.2019.09.010
Received 18 June 2019; Received in revised form 20 September 2019; Accepted 23 September 2019
Available online 24 September 2019
0042-6822/

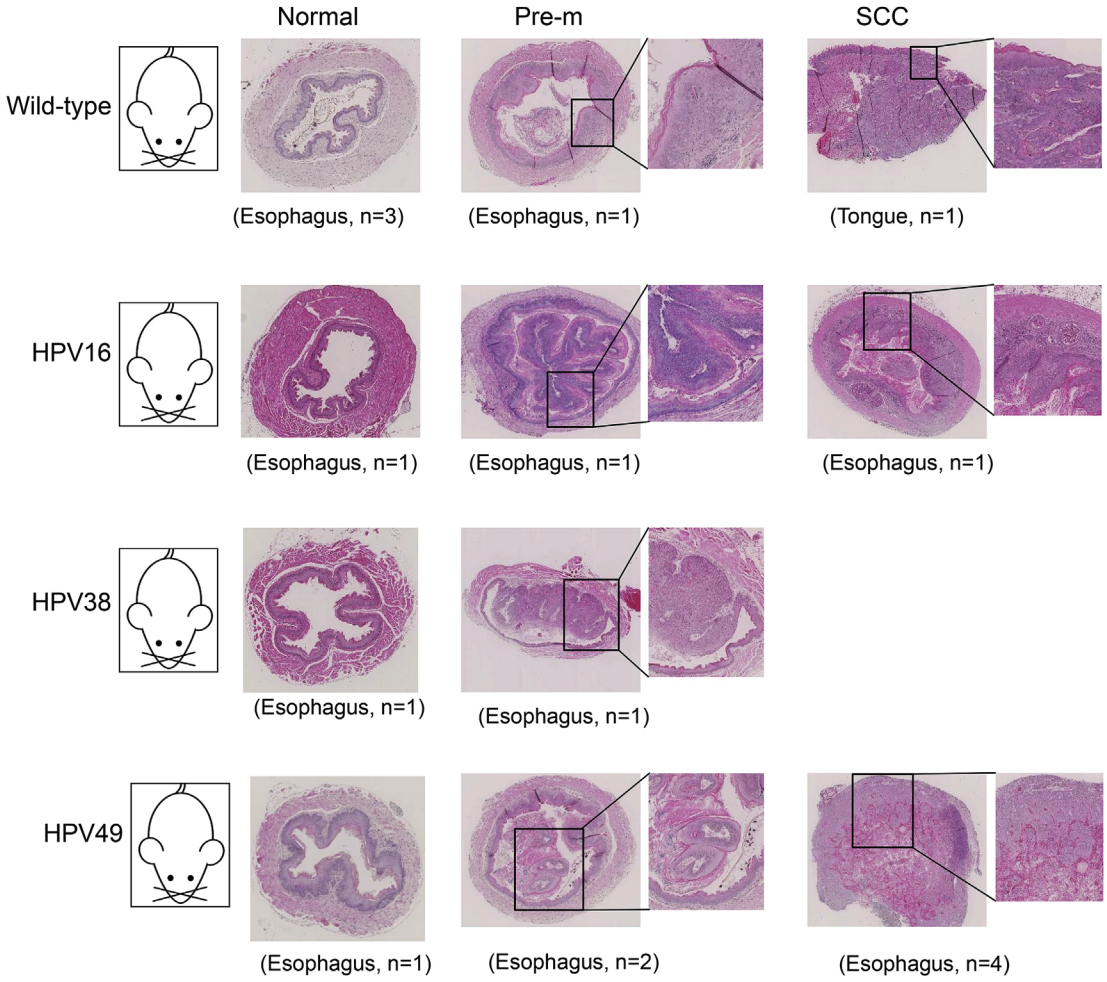


Fig. 1. Representative images of H&E-stained sections from WT or Tg mice from which the genomic DNA was extracted for whole-exome sequencing. Normal tissue, pre-malignant lesions (Pre-m), and squamous cell carcinoma (SCC) from the indicated mice exposed to long-term 4NQO treatment were collected for DNA extraction. Tissues were also processed for histological analyses. The text in brackets indicates the anatomical site, followed by the number of specimens from independent animals used for the whole-exome sequencing. From left to right, sections of unaffected esophagus, sections of esophagus affected by Pre-m lesions characterized by medium grade dysplasia (WT, HPV16, HPV49 E6/E7 Tg mice) or high grade dysplasia (HPV38 E6/E7 Tg mouse), and sections of invasive SCC from the tongue of a WT mouse and from the esophagus of HPV16 and HPV49 E6/E7 Tg mice. The stained sections were first scanned with a 5 × enlargement and then zoomed in via software analysis.

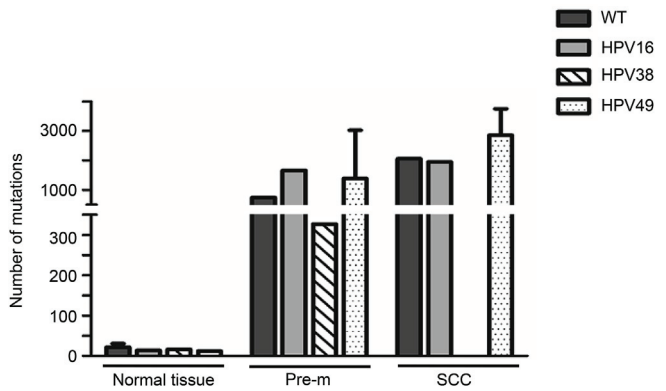


Fig. 2. The number of 4NQO-induced DNA mutations varies in the WT and different K14 HPV E6/E7 Tg animals. After whole-exome sequencing, the numbers of somatic mutations (SNPs and indels) were determined that have a functional impact and fall in exonic or splicing regions, and have an allelic fraction of 5% or more. The differences in mutation numbers between the different animal models are statistically significant: normal versus Pre-m, $P = 0.004$; Normal versus SCC, $P = 0.002$; Pre-m versus SCC, $P = 0.05$. Bars indicate standard deviations.

epithelia compared with the skin (Forslund et al., 2013; Hampras et al., 2017).

Functional studies showed that E6 and E7 from beta-3 HPV49 and the mucosal HR HPV16 share some functional similarities (Cornet et al., 2012; Viarisio et al., 2016). Similarly to what has been observed in HPV16 E6/E7 transgenic (Tg) mice (Strati et al., 2006), K14-driven expression of HPV49 E6 and E7 in mouse epithelia resulted in elevated susceptibility to upper digestive tract carcinogenesis upon initiation with the tobacco-mimicking and DNA-damaging agent 4-nitroquinoline 1-oxide (4NQO) (Viarisio et al., 2016; Ikenaga et al., 1975). However, these Tg mice did not show an increased susceptibility to chronic UV irradiation compared with the wild-type (WT) animals. Vice versa, beta-2 HPV38 E6 and E7 expression in K14 HPV38 E6/E7 Tg mice strongly cooperates with UV radiation in the development of cSCC, but the mice were little affected by 4NQO treatment (Viarisio et al., 2011, 2016). In the case of K14 beta-2 HPV38 E6/E7 Tg mice, the high cSCC incidence upon long-term UV exposure tightly correlates with their tendency to accumulate the classic UV-induced DNA mutational profile (Viarisio et al., 2018), further supporting the model described above for the role of the beta-1 and beta-2 HPVs as facilitators of UV-mediated skin carcinogenesis. A similar scenario could be hypothesized in the cooperation of beta-3 HPV49 E6 and E7 and 4NQO in promoting upper

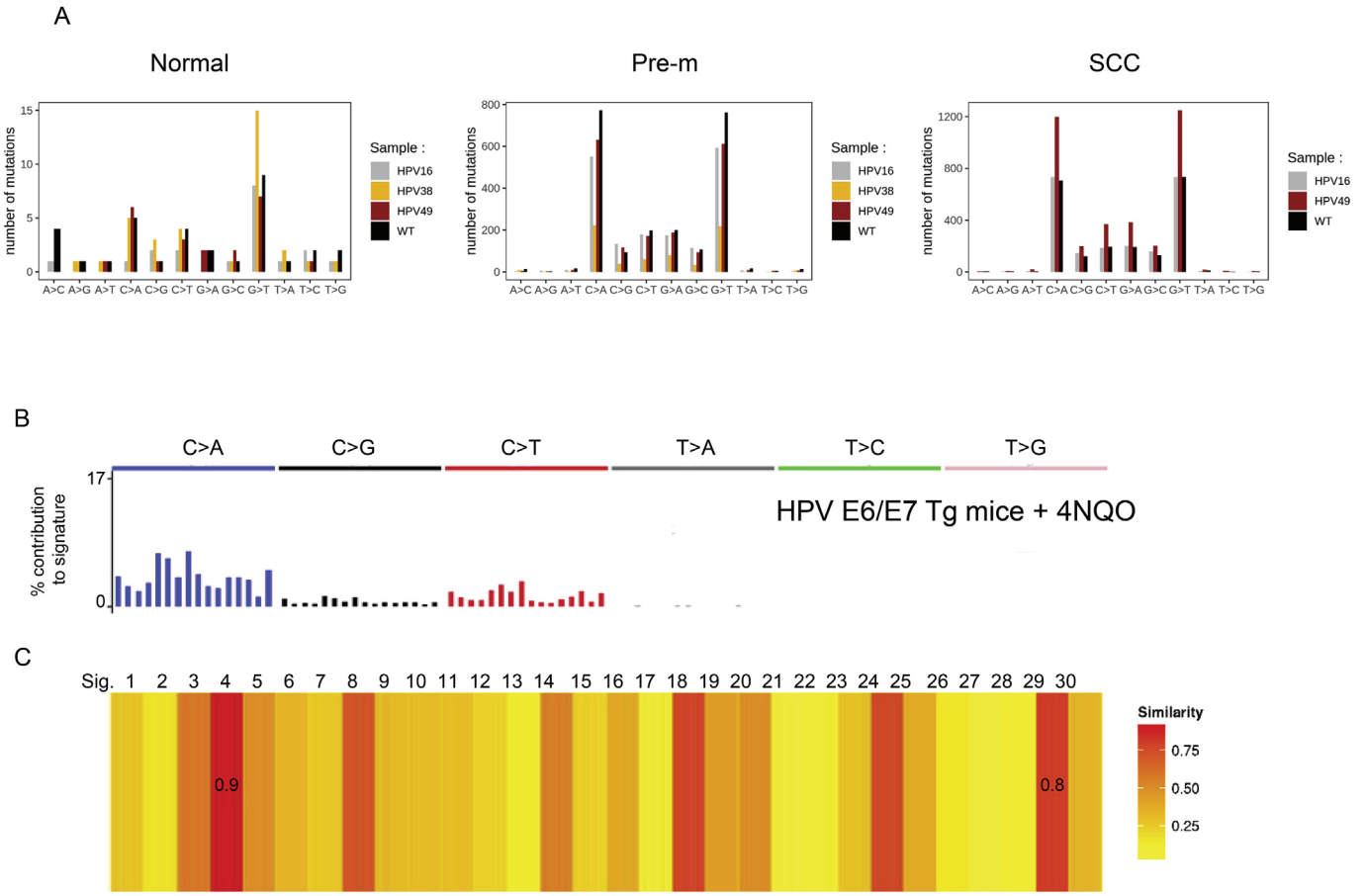


Fig. 3. 4NQO-induced DNA mutations increase with the severity of upper digestive tract lesions in the different K14 HPV E6/E7 Tg animals. (**A**) Different types of DNA mutations detected in upper digestive tract normal tissue, pre-malignant lesions (Pre-m), and squamous cell carcinoma (SCC) in Tg animals exposed to long-term 4NQO treatment. (**B**) SCCs of K14 HPV49 E6/E7 Tg mice treated with 4NQO display a clear tobacco-induced mutational signature with a very high number of C:G > A:T mutations. This type of mutation makes up the majority of the single-nucleotide variant types in Pre-m and SCC samples from the WT and Tg animals. The *y* axis shows the percentage contribution of those mutations to signatures, and the *x* axis shows the trinucleotide sequence context. (**C**) Heatmap presenting the similarity of the 4NQO-induced mutational signature to the 30 mutational signature available in COSMIC database version 2; the 4NQO-induced mutational signature presents a cosine similarity closer to signature 4 (tobacco smoking; 0.9) than to signature 29 (tobacco chewing; 0.81).

digestive tract carcinogenesis in mice. However, no information is available for the genome integrity of K14 beta-3 HPV49 E6/E7 Tg mice upon exposure to 4NQO treatment.

In this study, we perform whole-exome sequencing of upper-digestive tract lesions of different K14 HPV E6/E7 Tg animals and show that HPV49 E6 and E7 strongly increase the accumulation of 4NQO-induced DNA mutations.

2. Materials and methods

2.1. Animal models and ethics statement

All Tg animal models used in this study have been previously described (Viarisio et al., 2011, 2016, 2018) and https://mito.dkfz.de/mito/Animal%20line/10954, https://mito.dkfz.de/mito/Animal%20line/11244, and https://mito.dkfz.de/mito/Animal%20line/11245).

The animal facility of the German Cancer Research Center has been officially approved by the responsible authority (Regional Council of Karlsruhe, Schlossplatz 4–6, 76131 Karlsruhe, Germany) (file no. 35–9185.64). Housing conditions are thus in accordance with the German Animal Welfare Act (TierSchG) and EU Directive 425 2010/63/EU. Regular inspections of the facility are conducted by the Veterinary Authority of Heidelberg (Bergheimer Str. 69, 69115 Heidelberg, Germany). All experiments were in accordance with the institutional guidelines (designated veterinarian according to article 25 of Directive, 2010/63/EU and Animal Welfare Body according to article 27 of Directive, 2010/63/EU) and were officially approved by the Regional Council of Karlsruhe (file no. 35–9185.81/G-164/12).

2.2. 4NQO treatment

Experimental groups of 6-week-old female WT or K14 HPV E6/E7 Tg mice of type 16, 38, and 49 were treated as described previously (Viarisio et al., 2016) and https://mito.dkfz.de/mito/Tumor%20model/10635). Biopsies were taken from the upper digestive tract (tongue and esophagus) of both control and treated animals, used for DNA extraction (DNeasy Blood and Tissue Kit, Qiagen, Hilden, Germany) or fixed in 4% formaldehyde in phosphate-buffered saline for 24 h at room temperature, and embedded in paraffin. Sections of 5 µm were then stained with hematoxylin and eosin (H&E). The whole-exome sequencing was performed in the High-Throughput Sequencing unit of the Genomics & Proteomics Core Facility of the German Cancer Research Center (DKFZ) using Agilent SureSelect Whole Exome Kit.

The histological diagnosis was carried out in a blinded manner by a certified pathologist (CF).

2.3. Exome analysis

The quality of the raw reads was estimated with FastQC software

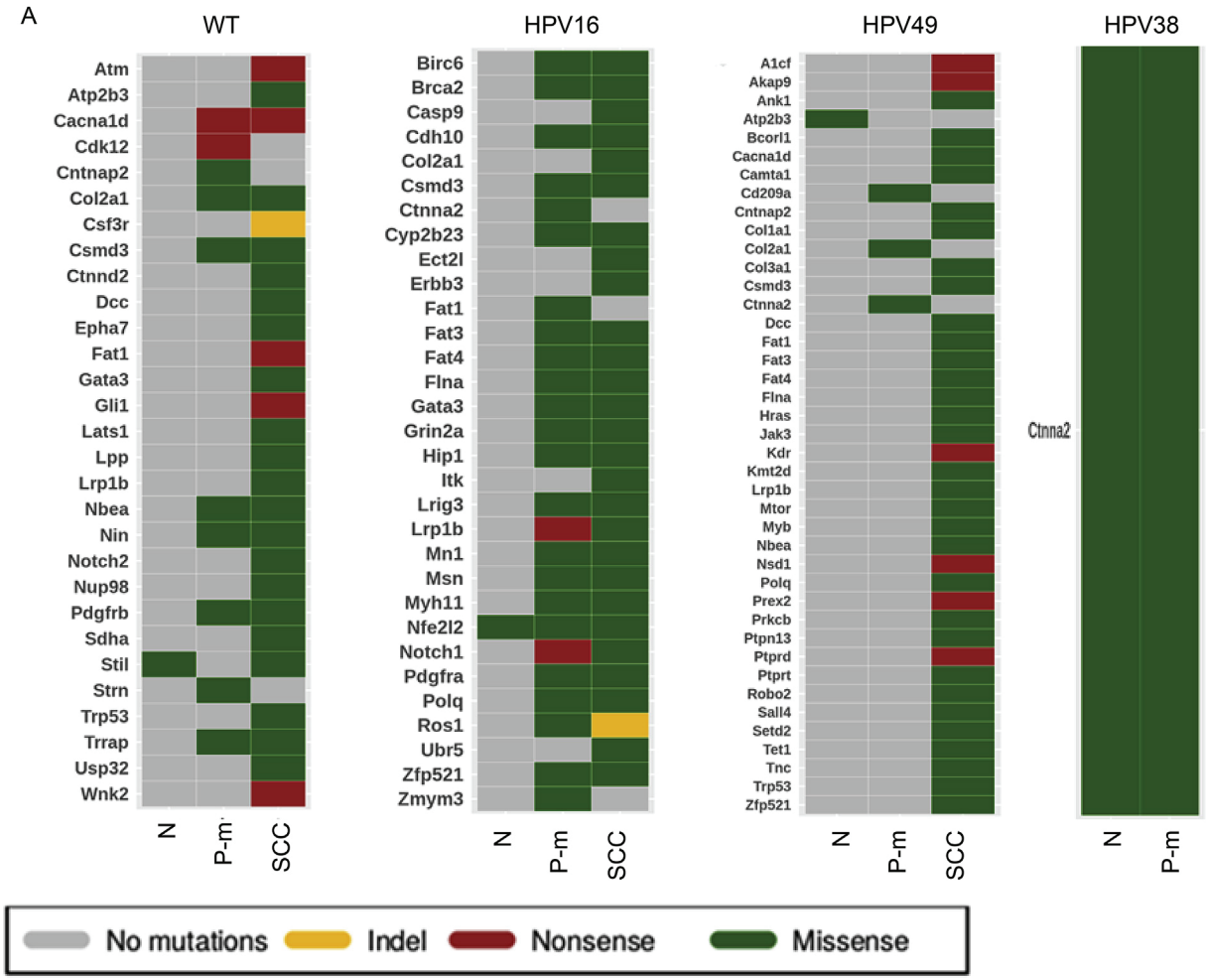


Fig. 4. Analyses of mutated genes in normal tissue and lesions from 4NQO-exposed Tg animals. (**A**) Heat-map of significantly mutated genes, corresponding to genes mutated in the corresponding sample and reported in the Cancer Gene Census list from the COSMIC database. The types of mutations indicated by colors are chosen according to the most prevalent mutation type in each sample. For categories with more than one sample (WT mice Normal: $n = 3$, HPV49 Tg mice Pre-m: $n = 2$, and HPV49 Tg mice SCC: $n = 4$), only the genes mutated in more than 50% of the samples are considered, and the type of mutation is defined as the most prevalent type among the samples. (**B**) Heatmap of significantly mutated genes, corresponding to genes mutated in the corresponding sample and reported to have an impact on epigenetic regulation processes. The types of mutations indicated by colors are chosen according to the most prevalent mutation type in each sample. For categories with more than one sample (WT mice Normal: $n = 3$, HPV49 Tg mice Pre-m: $n = 2$, and HPV49 Tg mice SCC: $n = 4$), only the genes mutated in more than 50% of the samples are considered, and the type of mutation is defined as the most prevalent type among the samples. (**C**) Heatmap of mutations in top genes mutated in human esophageal SCCs and their corresponding gene names in 4NQO-exposed animals. The types of mutation indicated by colors are chosen according to the most prevalent mutation type in each sample.

(version 0.11.5, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped to the GRCm38 (mm10) mouse reference genome (ftp://hgdownload.cse.ucsc.edu/goldenPath/mm10/) using Burrows-Wheeler Aligner (version 0.7.15, http://bio-bwa.sourceforge.net/) and producing a BAM file. The following GATK Best Practice Recommendations were applied to the BAM files to improve variant detection quality. Picard (version 2.4.1, https://broadinstitute.github.io/picard/) SortSam was used to sort and index BAM files, and the AddOrReplaceReadGroups tool was used to replace all read groups with a single new read group. The duplicate reads were marked with the MarkDuplicates tool, and the newly produced BAM file was indexed with the BuildBamIndex tool. GATK (version 3.6.0, https://software.broadinstitute.org/gatk/download/) RealignerTargetCreator was used to determine the position concerned by local realignment, and IndelRealigner was used to perform local realignment around these sites. The GATK BaseRecalibrator tool was used to detect systematic errors in base quality scores. dbSNP and dbindel (version 142) for the GRCm38 (mm10) reference genome was downloaded from the Sanger website (ftp://ftp-mouse.sanger.ac.uk/REL-1505-SNPs_Indels/) and considered as input. Lastly, the index of the output BAM file was created with Picard BuildBamIndex, and GATK PrintReads was used to write out sequence read data.

The quality of the alignment was estimated with QualiMap (version 2.0.2, http://qualimap.bioinfo.cipf.es/). Then, the variant calling was done with MuTect2 (https://software.broadinstitute.org/gatk/documentation/tooldocs/3.6-0/org_broadinstitute_gatk_tools_walkers_cancer_m2_MuTect2.php) by using a skin sample from a WT mouse not exposed to 4NQO as the "normal sample" for paired analysis. Only somatic mutations passing the MuTect2 internal filters were considered for the analysis. The VCF files are annotated with Annovar by using the MutSpec-Annot tool in Galaxy (Ardin et al., 2016). Variants were then filtered based on SegDup databases from UCSC (version from 4 May 2014, http://hgdownload.cse.ucsc.edu/goldenPath/mm10/database/genomicSuperDups.txt.gz), as well as Tandem Repeat and RepeatMasker (version from 9 February 2012, http://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/) (Tables S1–17). House-made scripts were then used to keep only SNPs that have a functional impact and fall in exonic or splicing regions. NMF mutational signatures were

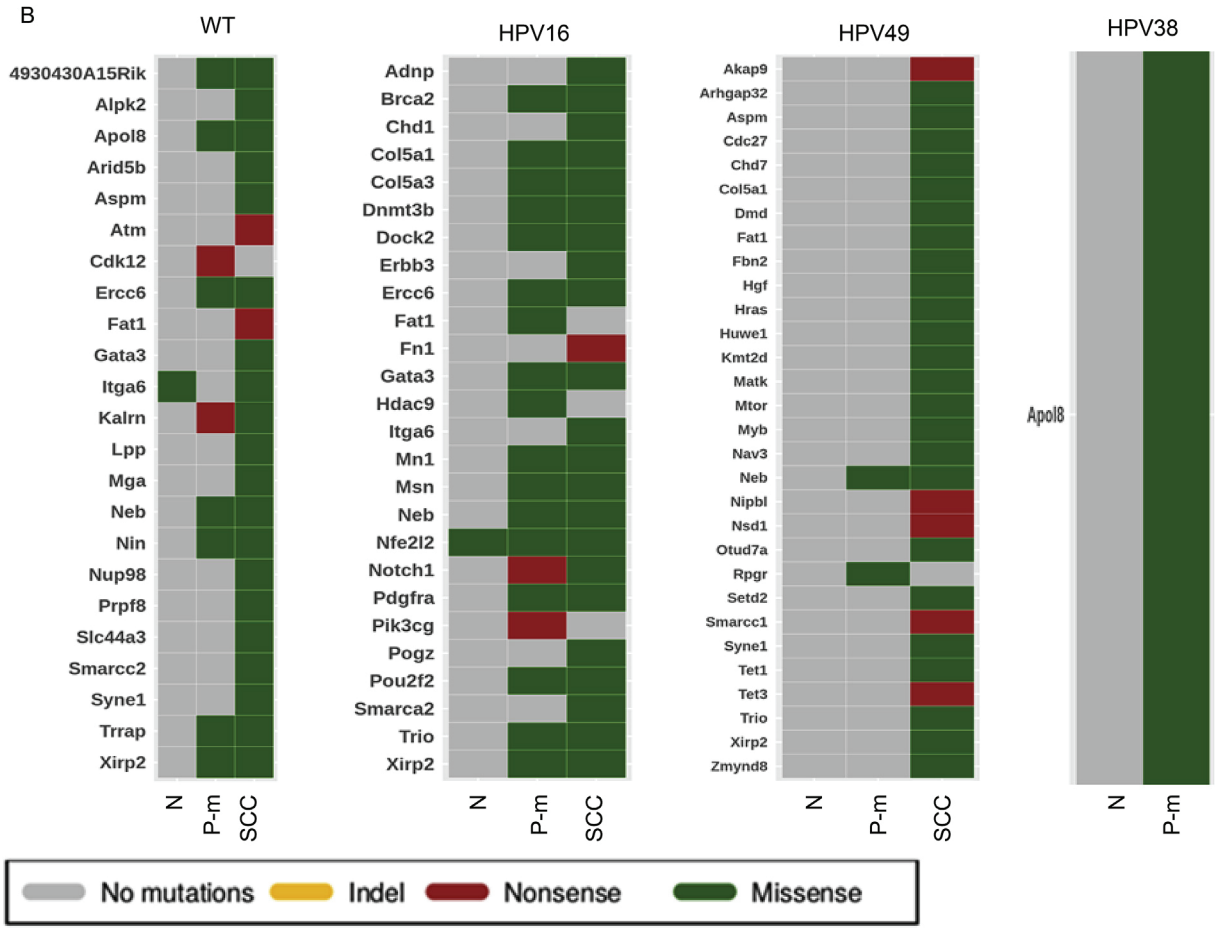


Fig. 4. (*continued*)

inferred with MutSpec-NMF tools, as previously reported.

The raw sequencing data has been deposited in Sequence Read Archive (NCBI) database, under the accession number PRJNA557836.

2.4. Comparison with epigenetic driver/modifier genes and Cancer Gene Census list

The list of epigenetic driver and modifier genes was constructed on the basis of genes reported in different publications (Gonzalez-Perez et al., 2013; Shen and Laird, 2013; Sturm et al., 2014; Timp and Feinberg, 2013; Vogelstein et al., 2013).

The Cancer Gene Census list was downloaded from the COSMIC website (March 2019, http://cancer.sanger.ac.uk/census).

The comparison of the mouse data with the human data was done with Bioconductor (release 3.6, https://www.bioconductor.org/) in R (version 3.4.4, codename "Someone to Lean On"). The top mutated genes in human esophagus SCC were retrieved from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). The module BioMart (Durinck et al., 2005, 2009) (version 2.34.2) enables the conversion of 35 of the 40 (87.5%) top mutated human gene names to their corresponding mouse gene names using the Ensembl database (version 95). The heatmaps were generated considering the genes mutated in more than 50% of the analyzed samples, i.e. WT mice Normal: $n = 3$, HPV49 Tg mice Pre-m: $n = 2$, and HPV49 Tg mice SCC: $n = 4$).

2.5. Comparison with human cancer mutated genes

Data from human head and neck (HNC) cancers were retrieved from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). Sample were selected for the following anatomical divisions: tonsil, oropharynx or base of the tongue. A gene was included in the analysis if mutated in at least one individual of the cohort considered. The module BioMart (Durinck et al., 2005, 2009) (version 2.34.2) enables the conversion of 6056 of the 7030 (86%) mutated mice gene names.

3. Results

3.1. Whole-exome sequencing analyses of upper digestive tract lesions from 4NQO-treated mice

To evaluate whether the high susceptibility of K14 beta-3 HPV49 E6/E7 Tg mice to 4NQO-induced cancers can be explained by their tendency to accumulate DNA mutations, we performed whole-exome sequencing (Illumina HiSeq). WT animals were included in the experiment as a comparative model. In addition, we selected a few specimens of 4NQO-treated K14 HPV16 or HPV38 E6/E7 Tg animals, which showed, respectively, high and low susceptibility to 4NQO-mediated carcinogenesis (Strati et al., 2006; Viarisio et al., 2016). As shown in Fig. 1, histologically confirmed specimens were selected from 4NQO-treated animals from two independent experiments (Viarisio et al., 2016) (Fig. 1). In the 4NQO-treated WT animals, only one SCC was detected and included in the whole-exome analysis, whereas 4NQO-treated K14 HPV38 E6/E7 Tg mice did not develop any SCC (Viarisio et al., 2016).

For the analysis of the DNA mutations in 4NQO-treated animals, the genomic sequence of the WT mouse not exposed to any type of treatment was determined in an independent experiment (Viarisio et al., 2018) and was used as a control sample in paired analysis. Exome

C

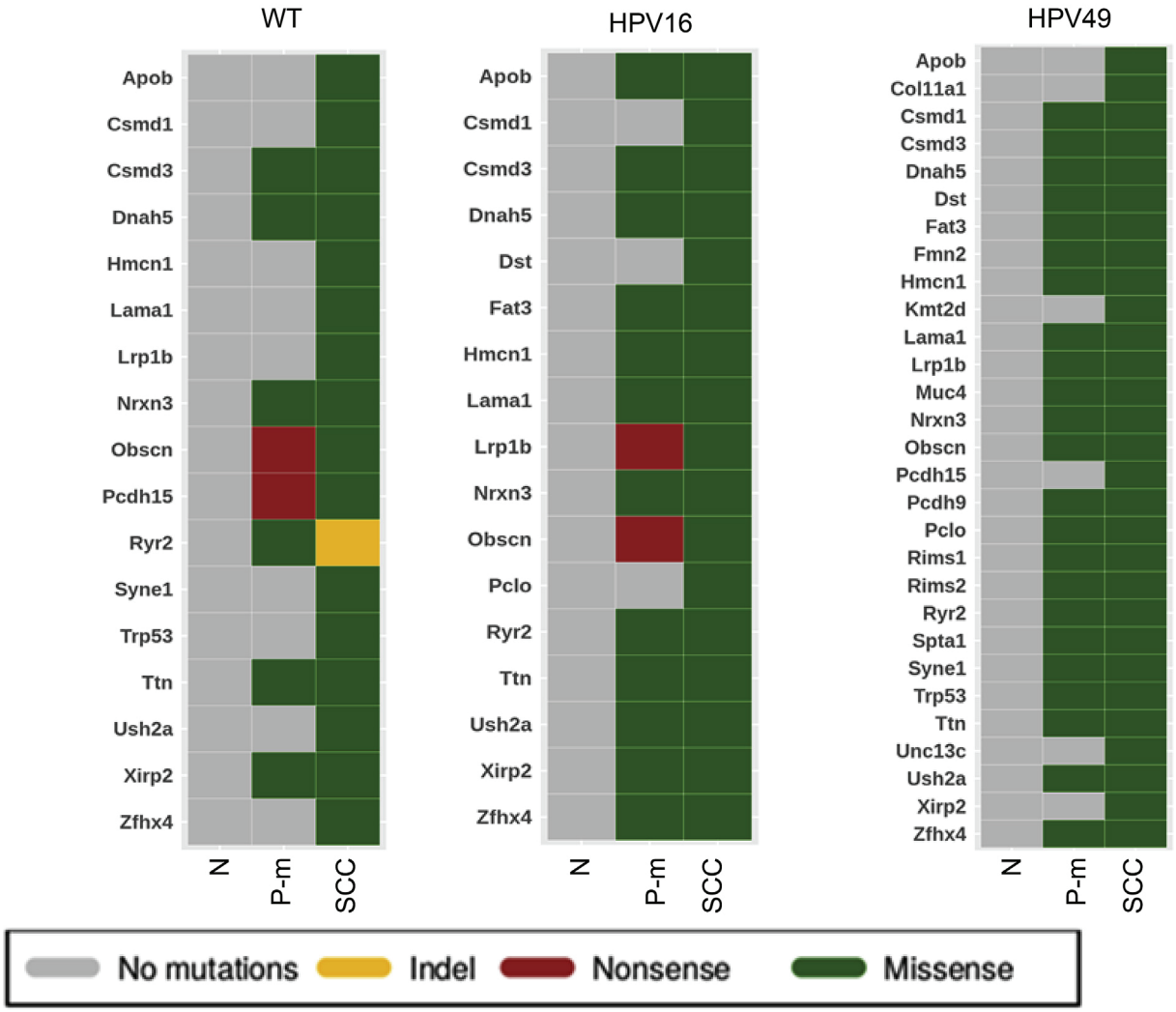


Fig. 4. (*continued*)

Table 1
Number and type of mutations in WT and Tg animals upon 4NQO treatments. Each specimen processed for whole exome sequencing was collected from different mice (n:17).

Mouse Number	Mouse Type	Tissue type	Histo-pathology	Mutect2[a] Somatic mutations (SNP; indels)	Number of mutated genes	Number of cancer gene census mutated (n = 801)	Number of cancer gene Epidriver/EpiModifier (n = 637)	Number of shared mutated genes in animals and human ESCA[b] (n = 40)
1	WT	Esophagus	Skin	**18** (18; 0)	**18**	**1**	**1**	**0**
2	WT	Esophagus	Skin	**13** (11; 2)	**13**	**1**	**0**	**0**
3	WT	Esophagus	Skin	**32** (30; 2)	**32**	**0**	**2**	**0**
4	WT	Esophagus	Pre-M	**739** (726; 13)	**694**	**37**	**27**	**10**
5	WT	Tongue	SCC	**2056** (2036; 20)	**1824**	**107**	**86**	**23**
6	HPV16	Esophagus	Skin	**14** (12; 2)	**14**	**1**	**1**	**0**
7	HPV16	Esophagus	Pre-M	**1657** (1640; 17)	**1505**	**73**	**58**	**17**
8	HPV16	Esophagus	SCC	**1943** (1931; 12)	**1723**	**93**	**78**	**23**
9	HPV38	Esophagus	Skin	**16** (14; 2)	**16**	**3**	**1**	**0**
10	HPV38	Esophagus	Pre-M	**324** (315; 9)	**316**	**14**	**14**	**5**
11	HPV49	Tongue	Skin	**12** (12; 0)	**12**	**2**	**0**	**0**
12	HPV49	Esophagus	Pre-M	**235** (234; 1)	**228**	**10**	**5**	**2**
13	HPV49	Esophagus	Pre-M	**2543** (2522; 21)	**2216**	**110**	**97**	**23**
14	HPV49	Esophagus	SCC	**3413** (3397; 16)	**2795**	**150**	**109**	**26**
15	HPV49	Esophagus	SCC	**3623** (3590; 33)	**2994**	**145**	**134**	**25**
16	HPV49	Esophagus	SCC	**2709** (2692; 17)	**2318**	**121**	**98**	**25**
17	HPV49	Esophagus	SCC	**1648** (1637; 11)	**1484**	**79**	**69**	**21**

[a] Mutect2 filtered mutations (see Methods for filtering parameters).
[b] ESCA: Esophageal Carcinoma.

sequencing of the collected samples generated an average coverage of 154.66 × ± 15.82 × (mean ± standard deviation). Normal tissue from all four types of animals contained a relatively low number of somatic mutations (18 ± 6.61) (mean ± standard deviation) (Fig. 2). In contrast, the number of somatic mutations increased according to the severity of the lesions (Fig. 2). In 4NQO-exposed Tg animals, the mutational load varied across our cohort of pre-malignant lesions, averaging 1102 somatic variants (range, 235–2547) or 4.67 ± 3.73 variants per Mb. The exome of the well-differentiated SCCs had a substantially higher number of variants, with an average of 2571 somatic variants (range, 1648–3638) or 11.89 ± 3.18 variants per Mb.

In conclusion, whole-exome sequencing revealed that K14 HPV16 and HPV49 E6/E7 Tg mice have a high susceptibility to the accumulation of DNA mutations induced by 4NQO treatment.

3.2. Characterization of DNA mutations in 4NQO-treated animals

Most of the somatic mutations detected in pre-malignant lesions and SCCs were C:G > A:T mutations (Fig. 3A) (Tables S1–17). The application of the non-negative matrix factorization (NMF) method enabled the extraction of the mutational signatures composed of 96 single base substitution types, considering the trinucleotide sequence context (one base upstream and one base downstream) (Fig. 3B). Next, we compared the mutational signature of 4NQO-treated K14 HPV49 E6/E7 Tg mice mice with the 30 mutational signatures available in COSMIC database version 2, by the cosine similarity method (Alexandrov et al., 2013; Olivier et al., 2014). The value of the cosine similarity obtained for the new signature is 0.9 for COSMIC signature 4 (tobacco smoking) and 0.8 for COSMIC signature 29 (tobacco chewing) (Fig. 3C).

To evaluate whether the somatic mutations detected in specimens from the 4NQO-treated animals have some biological relevance in the development of pre-malignant and malignant lesions, we compared the list of mutated genes in our animal models with one identified in the Cancer Gene Census (Futreal et al., 2004; Sondka et al., 2018). Cancer genes were found to be mutated in pre-malignant and malignant lesions from 4NQO-treated mice (Fig. 4A). In addition, the number of mutated cancer genes gradually increased in SCCs from WT, K14 HPV16 E6/E7 Tg, and K14 HPV49 E6/E7 Tg animals. Only one cancer gene was found mutated in the pre-malignant lesion of K14 HPV38 E6/E7 Tg animals.

Similar results were obtained when we analyzed the DNA mutations in epi-driver and epi-modifier genes (Gonzalez-Perez et al., 2013; Shen and Laird, 2013; Sturm et al., 2014; Timp and Feinberg, 2013; Vogelstein et al., 2013) (Fig. 4B). K14 HPV49 E6/E7 Tg animals showed higher number of mutated genes upon 4NQO exposure in comparison WT and the other HPV Tg mice (Table 1).

We have previously shown in K14 HPV38 E6/E7 Tg animals that the viral proteins acts an early stage of UV-induced skin carcinogenesis facilitating the accumulation of DNA mutations, but they are dispensable for the cancer cell growth after full development of cSCC (Viarisio et al., 2011; 2018). To evaluate whether HPV38 and HPV49 cooperates with UV and 4NQO, respectively, to alter a similar pattern of cellular genes/pathways in mouse carcinogenesis, we compared the DNA mutations of SCC from 4NQO-exposed HPV49 Tg mice and UV-exposed HPV38 Tg mice. We identified a large number of common mutations (n = 3705) in malignant lesions of both animal models, leading to alteration of similar cellular pathways (Fig. S1).

Finally, we compared the pattern of DNA mutations detected in lesions from 4NQO-exposed Tg animals with the pattern of mutations found in esophageal SCC in humans. As shown in Fig. 4C and Table 1, 29 of the 35 (83%) top genes mutated in human SCCs were found mutated in lesions from HPV49 E6/E7 Tg animals. In contrast, a lower number of these human genes were mutated in lesions from WT and K14 HPV16 E6/E7 Tg animals (Fig. 4C). In addition, we compared the pattern of mutations detected in SCC of 4NQO-exposed K14 HPV49 E6/E7 Tg animals with the pattern of mutations detected in human HNCSCC associated with tobacco, HPV infection or tobacco/HPV infection. The analysis revealed that a large proportion of genes mutated in human SCC were also detected in the SCC of the Tg animals (Fig. S2A). Accordingly, the pathway analysis showed that similar alterations occurred in human and mouse SCC (Fig. S2B).

In conclusion, HPV49 E6 and E7 expression in upper digestive tract epithelia favors the accumulation of 4NQO-induced DNA mutations that resemble the signature of tobacco exposure.

4. Discussion

In a previous study, we showed that beta-3 HPV49 E6 and E7 expression driven by K14 promoter in a Tg mouse model strongly cooperates with the carcinogen 4NQO in promoting cancer in the upper digestive tract (Viarisio et al., 2016). A similar scenario has been observed in a Tg mouse model for the mucosal HR HPV16 (Strati et al., 2006). The synergism between 4NQO and viral oncogene expression in promoting carcinogenesis appeared to be beta-HPV-type specific, because beta-2 HPV38 E6 and E7 weakly cooperated with 4NQO in the same Tg model, promoting only papillomas but never cSCC (Viarisio et al., 2016). The opposite situation was observed when K14 HPV38 and K14 HPV49 E6/E7 Tg mice were exposed to another protocol of carcinogenesis using UV radiation. Only HPV38 E6 and E7 expression in K14 HPV38 E6/E7 Tg mice was found to cooperate with UV irradiation in the development of cSCC (Viarisio et al., 2011, 2016, 2018). These different abilities of HPV38 and HPV49 E6 and E7 in the Tg mouse models may be explained by the different tissue tropism or intrinsic properties of the mouse tissue. However as regards to different tissue tropism, it is possible that these viruses, in order to efficiently complete their life cycle, may have developed specific mechanisms to counteract the anti-proliferative events induced by environmental factors at distinct anatomical sites. Interestingly, compelling lines of evidence from epidemiological and functional studies support the model that beta-1 and beta-2 HPV types play a role at an initial stage of skin carcinogenesis, facilitating the accumulation of UV-induced DNA mutations that, in turn, render cancer cell proliferation independent of the expression of viral genes. In line with this model, beta HPV DNA is not detected in all cancer cells, and the viral load decreases with the progression of the severity of the skin lesion (Correa et al., 2017; Dona et al., 2019; Weissenborn et al., 2005). Thus, specific beta HPV types may act with a hit-and-run mechanism in UV-induced cSCC development (Rollison et al., 2019). Based on the findings presented here, it is possible to speculate that a similar synergistic model could exist for other HPVs and environmental factors at different anatomical sites. Importantly, it could be possible that oral HPV infections may act with a hit-and-run mechanism in the development of a subset of HNC cancers.

A limitation of our study is the relatively small number of specimens that were subjected to whole-exome sequencing, especially for the K14 HPV16 and HPV38 E6/E7 Tg animals. However, the high susceptibility of K14 HPV49 E6/E7 animals to 4NQO-induced mutations was consistently observed in all mice in two independent experiments with 4NQO-exposed animals. Further epidemiological and biological studies are needed to evaluate the possible synergism of beta-3 HPV types and tobacco exposure in promoting any pathological condition in humans.

Conflicts of interest

The authors declare no competing financial interests.

Acknowledgments

We are grateful to all members of our laboratories at DKFZ and IARC for their cooperation, Nicole Suty for her help with preparation, and Dr Karen Müller for editing this manuscript.

The study was supported by a grant from Deutsche Krebshilfe, Germany (no. 110259) to LG and MT and Fondation ARC, France pour

la recherche sur le cancer (no. PJA 20151203192) (https://www.fondation-arc.org/espace-chercheur) and the Institut National de la Santé et de la Recherche Médicale, France (no. ENV201610) (https://www.eva2.inserm.fr/EVA/jsp/AppelsOffres/CANCER/) to MT.

We also thank the High-Throughput Sequencing unit of the Genomics & Proteomics Core Facility, German Cancer Research Center (DKFZ), for providing excellent sequencing services.

The authors alone are responsible for the views expressed in this article, and they do not necessarily represent the views, decisions, or policies of the institutions with which they are affiliated.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.virol.2019.09.010.

References

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., Boyault, S., Burkhardt, B., Butler, A.P., Caldas, C., Davies, H.R., Desmedt, C., Eils, R., Eyfjord, J.E., Foekens, J.A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Jager, N., Jones, D.T., Jones, D., Knappskog, S., Kool, M., Lakhani, S.R., Lopez-Otin, C., Martin, S., Munshi, N.C., Nakamura, H., Northcott, P.A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J.V., Puente, X.S., Raine, K., Ramakrishna, M., Richardson, A.L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T.N., Span, P.N., Teague, J.W., Totoki, Y., Tutt, A.N., Valdes-Mas, R., van Buuren, M.M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L.R., Australian Pancreatic Cancer Genome, I, Consortium, I.B.C., Consortium, I.M.-S., PedBrain, I., Zucman-Rossi, J., Futreal, P.A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S.M., Siebert, R., Campo, E., Shibata, T., Pfister, S.M., Campbell, P.J., Stratton, M.R., 2013. Signatures of mutational processes in human cancer. Nature 500, 415–421.

Ardin, M., Cahais, V., Castells, X., Bouaoun, L., Byrnes, G., Herceg, Z., Zavadil, J., Olivier, M., 2016. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. BMC Bioinf. 17, 170.

Bottalico, D., Chen, Z., Dunne, A., Ostoloza, J., McKinney, S., Sun, C., Schlecht, N.F., Fatahzadeh, M., Herrero, R., Schiffman, M., Burk, R.D., 2011. The oral cavity contains abundant known and novel human papillomaviruses from the Betapapillomavirus and Gammapapillomavirus genera. J. Infect. Dis. 204, 787–792.

Cornet, I., Bouvard, V., Campo, M.S., Thomas, M., Banks, L., Gissmann, L., Lamartine, J., Sylla, B.S., Accardi, R., Tommasino, M., 2012. Comparative analysis of transforming properties of E6 and E7 from different beta human papillomavirus types. J. Virol. 86, 2366–2370.

Correa, R.M., Vladimirsky, S., Heideman, D.A., Coringrato, M., Abeldano, A., Olivares, L., Del Aguila, R., Alonio, L.V., Snijders, P.J., Picconi, M.A., 2017. Cutaneous human papillomavirus genotypes in different kinds of skin lesions in Argentina. J. Med. Virol. 89, 352–357.

Dona, M.G., Chiantore, M.V., Gheit, T., Fiorucci, G., Vescio, M.F., La Rosa, G., Accardi, L., Costanzo, G., Giuliani, M., Romeo, G., Rezza, G., Tommasino, M., Luzi, F., Di Bonito, P., 2019. Comprehensive analysis of beta and gamma Human Papillomaviruses in actinic keratosis and apparently healthy skin of elderly patients. Br. J. Dermatol. 18, 620–622.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., Huber, W., 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21, 3439–3440.

Durinck, S., Spellman, P.T., Birney, E., Huber, W., 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 4, 1184–1191.

Forslund, O., Johansson, H., Madsen, K.G., Kofoed, K., 2013. The nasal mucosa contains a large spectrum of human papillomavirus types from the Betapapillomavirus and Gammapapillomavirus genera. J. Infect. Dis. 208, 1335–1341.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.R., 2004. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183.

Gonzalez-Perez, A., Jene-Sanz, A., Lopez-Bigas, N., 2013. The mutational landscape of chromatin regulatory factors across 4,623 tumor samples. Genome Biol. 14, r106.

Hampras, S.S., Rollison, D.E., Giuliano, A.R., McKay-Chopin, S., Minoni, L., Sereday, K., Gheit, T., Tommasino, M., 2017. Prevalence and concordance of cutaneous beta human papillomavirus infection at mucosal and cutaneous sites. J. Infect. Dis. 216, 92–96.

Hasche, D., Vinzon, S.E., Rosl, F., 2018. Cutaneous papillomaviruses and non-melanoma skin cancer: causal agents or innocent bystanders? Front. Microbiol. 9, 874.

Ikenaga, M., Ishii, Y., Tada, M., Kakunaga, T., Takebe, H., 1975. Excision-repair of 4-nitroquinolin-1-oxide damage responsible for killing, mutation, and cancer. Basic Life Sci. 763–771 5b.

Olivier, M., Weninger, A., Ardin, M., Huskova, H., Castells, X., Vallee, M.P., McKay, J., Nedelko, T., Muehlbauer, K.R., Marusawa, H., Alexander, J., Hazelwood, L., Byrnes, G., Hollstein, M., Zavadil, J., 2014. Modelling mutational landscapes of human cancers in vitro. Sci. Rep. 4, 4482.

Pierce Campbell, C.M., Messina, J.L., Stoler, M.H., Jukic, D.M., Tommasino, M., Gheit, T., Rollison, D.E., Sichero, L., Sirak, B.A., Ingles, D.J., Abrahamsen, M., Lu, B., Villa, L.L., Lazcano-Ponce, E., Giuliano, A.R., 2013. Cutaneous human papillomavirus types detected on the surface of male external genital lesions: a case series within the HPV Infection in Men Study. J. Clin. Virol. 58, 652–659.

Rollison, D.E., Viarisio, D., Amorrortu, R.P., Gheit, T., Tommasino, M., 2019. An emerging issue in oncogenic virology: the role of beta human papillomavirus types in the development of cutaneous squamous cell carcinoma. J. Virol. 93.

Shen, H., Laird, P.W., 2013. Interplay between the cancer genome and epigenome. Cell 153, 38–55.

Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., Forbes, S.A., 2018. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer 18, 696–705.

Strati, K., Pitot, H.C., Lambert, P.F., 2006. Identification of biomarkers that distinguish human papillomavirus (HPV)-positive versus HPV-negative head and neck cancers in a mouse model. Proc. Natl. Acad. Sci. U. S. A. 103, 14152–14157.

Sturm, D., Bender, S., Jones, D.T., Lichter, P., Grill, J., Becher, O., Hawkins, C., Majewski, J., Jones, C., Costello, J.F., Iavarone, A., Aldape, K., Brennan, C.W., Jabado, N., Pfister, S.M., 2014. Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. Nat. Rev. Cancer 14, 92–107.

Timp, W., Feinberg, A.P., 2013. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nat. Rev. Cancer 13, 497–510.

Tommasino, M., 2014. The human papillomavirus family and its role in carcinogenesis. Semin. Cancer Biol. 26, 13–21.

Tommasino, M., 2017. The biology of beta human papillomaviruses. Virus Res. 231, 128–138.

Torres, M., Gheit, T., McKay-Chopin, S., Rodriguez, C., Romero, J.D., Filotico, R., Dona, M.G., Ortiz, M., Tommasino, M., 2015. Prevalence of beta and gamma human papillomaviruses in the anal canal of men who have sex with men is influenced by HIV status. J. Clin. Virol. 67, 47–51.

Van Doorslaer, K., Bernard, H.U., Chen, Z., de Villiers, E.M., zur Hausen, H., Burk, R.D., 2011. Papillomaviruses: evolution, Linnaean taxonomy and current nomenclature. Trends Microbiol. 19, 49–50 author reply 50-41.

Van Doorslaer, K., Tan, Q., Xirasagar, S., Bandaru, S., Gopalan, V., Mohamoud, Y., Huyen, Y., McBride, A.A., 2013. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. Nucleic Acids Res. 41, D571–D578.

Viarisio, D., Mueller-Decker, K., Kloz, U., Aengeneyndt, B., Kopp-Schneider, A., Grone, H.J., Gheit, T., Flechtenmacher, C., Gissmann, L., Tommasino, M., 2011. E6 and E7 from beta HPV38 cooperate with ultraviolet light in the development of actinic keratosis-like lesions and squamous cell carcinoma in mice. PLoS Pathog. 7, e1002125.

Viarisio, D., Muller-Decker, K., Accardi, R., Robitaille, A., Durst, M., Beer, K., Jansen, L., Flechtenmacher, C., Bozza, M., Harbottle, R., Voegele, C., Ardin, M., Zavadil, J., Caldeira, S., Gissmann, L., Tommasino, M., 2018. Beta HPV38 oncoproteins act with a hit-and-run mechanism in ultraviolet radiation-induced skin carcinogenesis in mice. PLoS Pathog. 14, e1006783.

Viarisio, D., Muller-Decker, K., Zanna, P., Kloz, U., Aengeneyndt, B., Accardi, R., Flechtenmacher, C., Gissmann, L., Tommasino, M., 2016. Novel ss-HPV49 transgenic mouse model of upper digestive tract cancer. Cancer Res. 76, 4216–4225.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz Jr., L.A., Kinzler, K.W., 2013. Cancer genome landscapes. Science 339, 1546–1558.

Weissenborn, S.J., Nindl, I., Purdie, K., Harwood, C., Proby, C., Breuer, J., Majewski, S., Pfister, H., Wieland, U., 2005. Human papillomavirus-DNA loads in actinic keratoses exceed those in non-melanoma skin cancers. J. Investig. Dermatol. 125, 93–97.

Experimental and pan-cancer genome analyses reveal widespread contribution of acrylamide exposure to carcinogenesis in humans

Maria Zhivagui,[1] Alvin W.T. Ng,[2,3,4] Maude Ardin,[1] Mona I. Churchwell,[5] Manuraj Pandey,[1] Claire Renard,[1] Stephanie Villar,[1] Vincent Cahais,[6] Alexis Robitaille,[7] Liacine Bouaoun,[8] Adriana Heguy,[9] Kathryn Z. Guyton,[10] Martha R. Stampfer,[11] James McKay,[12] Monica Hollstein,[1,13,14] Magali Olivier,[1] Steven G. Rozen,[2,3,4] Frederick A. Beland,[5] Michael Korenjak,[1] and Jiri Zavadil[1]

[1]Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, Lyon 69008, France; [2]Centre for Computational Biology, Duke–NUS Medical School, Singapore 169857, Singapore; [3]Program in Cancer and Stem Cell Biology, Duke–NUS Medical School, 169857, Singapore; [4]NUS Graduate School for Integrative Sciences and Engineering, Singapore 117456, Singapore; [5]Division of Biochemical Toxicology, National Center for Toxicological Research, Jefferson, Arkansas 72079, USA; [6]Epigenetics Group, International Agency for Research on Cancer, Lyon 69008, France; [7]Infections and Cancer Biology Group, International Agency for Research on Cancer, Lyon 69008, France; [8]Environment and Radiation Section, International Agency for Research on Cancer, Lyon 69008, France; [9]Department of Pathology and Genome Technology Center, New York University, Langone Medical Center, New York, New York 10016, USA; [10]IARC Monographs Group, International Agency for Research on Cancer, Lyon 69008, France; [11]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; [12]Genetic Cancer Susceptibility Group, International Agency for Research on Cancer, Lyon 69008, France; [13]Deutsches Krebsforschungszentrum, 69120 Heidelberg, Germany; [14]Faculty of Medicine and Health, University of Leeds, LIGHT Laboratories, Leeds LS2 9JT, United Kingdom

Humans are frequently exposed to acrylamide, a probable human carcinogen found in commonplace sources such as most heated starchy foods or tobacco smoke. Prior evidence has shown that acrylamide causes cancer in rodents, yet epidemiological studies conducted to date are limited and, thus far, have yielded inconclusive data on association of human cancers with acrylamide exposure. In this study, we experimentally identify a novel and unique mutational signature imprinted by acrylamide through the effects of its reactive metabolite glycidamide. We next show that the glycidamide mutational signature is found in a full one-third of approximately 1600 tumor genomes corresponding to 19 human tumor types from 14 organs. The highest enrichment of the glycidamide signature was observed in the cancers of the lung (88% of the interrogated tumors), liver (73%), kidney (>70%), bile duct (57%), cervix (50%), and, to a lesser extent, additional cancer types. Overall, our study reveals an unexpectedly extensive contribution of acrylamide-associated mutagenesis to human cancers.

[Supplemental material is available for this article.]

Cancer can be caused by lifestyle factors, environmental or occupational exposures involving chemicals, their complex mixtures, and physical and biological agents. Many human carcinogens show shared key characteristics (Smith et al. 2016), and different carcinogens may have a spectrum of such characteristics and operate through distinct mechanisms to produce genetic alterations. Recognizable somatic alteration patterns characterize carcinogens that are mutagenic. Single-base substitution (SBS) mutational signatures can be expressed in simple mathematical terms that enable them to be extracted from thousands of cancer genomes (Alexandrov et al. 2013a, 2018). Several of the identified mutational signatures have been attributed to specific external exposures or endogenous factors through epidemiological and/or experimental studies (Alexandrov et al. 2018). The majority of the signatures remain of unknown origin, and additional, yet unrecognized, signatures are likely to be extracted from rapidly accumulating cancer genome data. Well-controlled experimental exposure systems can help identify the causes of the orphan mutational signatures and define new carcinogen-generated patterns (for review, see Hollstein et al. 2017; Zhivagui et al. 2017).

Various diet-related and iatrogenic exposures contribute to human cancer burden, involving, for instance, food contaminants (aflatoxin B1 [AFB1]) or alternative medicines (aristolochic acid [AA]) with well-documented mutagenic properties; AFB1 induces predominantly C:G > A:T and AA generates T:A > A:T transversions. These characteristic mutations, arising in preferred sequence contexts, allowed unequivocal association of exposure to AFB1 or AA with specific subtypes of hepatobiliary or urological cancers (Poon et al. 2013; Meier et al. 2014; Scelo et al. 2014; Jelaković

Corresponding authors: zavadilj@iarc.fr; korenjakm@iarc.fr
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.242453.118. Freely available online through the *Genome Research* Open Access option.

et al. 2015; Hoang et al. 2016; Chawanthayatham et al. 2017; Huang et al. 2017; Ng et al. 2017; Zhang et al. 2017).

Among dietary compounds with carcinogenic potential, acrylamide (ACR) is of interest because of its ubiquitous presence. Important sources of exposure to ACR include tobacco smoke (Mojska et al. 2016), coffee (Takatsuki et al. 2003), and a spectrum of occupational settings (IARC 1994). ACR forms in carbohydrate-rich foods (e.g., potatoes and cereals) heated at high temperatures, because of Maillard reactions involving reducing sugars and the amino acid asparagine (Tareke et al. 2002). There is sufficient evidence that ACR is carcinogenic in rodents (Beland et al. 2013, 2015), and it was classified by the International Agency for Research on Cancer (IARC) as a probable carcinogen (Group 2A) (IARC 1994). The associations of dietary ACR exposure with renal, endometrial, and ovarian cancers have been explored in epidemiological studies (Hogervorst et al. 2008; Virk-Baker et al. 2014; Pelucchi et al. 2015). However, accurate ACR exposure assessment by questionnaires has been difficult, whereas more direct measures of molecular markers, such as hemoglobin adduct levels, may not yield conclusive findings on past exposures (Olesen et al. 2008; Wilson et al. 2009; Xie et al. 2013; Obón-Santacana et al. 2016a, b,c). Thus, innovative well-controlled exposure model systems can improve our understanding of the ACR exposure–associated effects and risk.

Oxidation of ACR by cytochrome P450 produces the highly reactive electrophilic epoxide glycidamide (GA) (Segerbäck et al. 1995; Sumner et al. 1999; Ghanayem et al. 2005). The *Hras* mutation loads in neoplasms of mice exposed to ACR or GA were higher upon exposure to GA (Von Tungeln et al. 2012), and more mutations in the *cII* reporter gene of Big Blue mouse embryonic fibroblasts were obtained by GA treatment in comparison to ACR (Besaratinia and Pfeifer 2003, 2004). In vivo and in vitro reporter gene mutagenesis studies showed an increased association of ACR and GA exposure with T:A > C:G transitions and T:A > A:T and C:G > G:C transversions (Besaratinia and Pfeifer 2003, 2004; Von Tungeln et al. 2009, 2012; Ishii et al. 2015; Manjanatha et al. 2015). In addition, GA exposure induces C:G > A:T transversions (Besaratinia and Pfeifer 2004). However, these ACR- and GA-specific patterns were based on limited mutation counts and do not allow translating adequately the reported mutation types into genome-wide patterns.

Massively parallel sequencing allows studying a large number of mutations in a single sample, thus significantly enhancing the power of mutation analysis in experimental models. Analogously to human cancer genome projects, genome-scale mutational signatures can be extracted from highly controlled carcinogen exposure experiments using mammalian cells and animal models, in combination with advanced computational methods (Olivier et al. 2014; Nik-Zainal et al. 2015; Huang et al. 2017). By integrating massively parallel sequencing and DNA adduct analysis in a mammalian cell clonal expansion model (Olivier et al. 2014; Nik-Zainal et al. 2015; Huskova et al. 2017) and by computational interrogation of the Pan-Cancer Analysis of Whole Genomes (PCAWG) data, we aimed to systematically investigate the mutational signatures of ACR and GA and to determine the contribution of ACR/GA to human carcinogenesis.

Results

Human *TP53* mutations generated by ACR or GA treatment

Primary Hupki MEF cultures from three different embryos (Prim_1, Prim_2, and Prim_3) exposed to ACR or GA at the predetermined cytotoxic and genotoxic conditions yielded multiple immortalized clones (Methods) (Supplemental Fig. S1) suitable for massively parallel sequencing (Olivier et al. 2014). Sanger sequencing of *TP53* in the clones derived from ACR exposure (ACR clones) and GA exposure (GA clones) and spontaneous immortalization (Spont), showed that ACR clones obtained from the Prim_2 MEFs showed loss of heterozygosity in the *TP53* codon 72 involving a loss of the proline allele (ACR_1 clone), and also loss of the arginine allele resulting in a hemizygous ACR_2 clone (Table 1). No *TP53* mutations were observed in the Spont clones. The detection of *TP53* mutations in three out of seven ACR clones and in one out of five GA clones (Table 1) provided a sound rationale for extended sequencing at the exome scale.

Analysis of mutation spectra

Whole-exome sequencing (WES) of all Spont as well as exposed clones revealed that the total number of acquired SBS did not differ markedly between the ACR and Spont clones. The Spont clones harbored on average 190 (median = 151, range = 141–277) SBSs, whereas the ACR clones had on average 208 (median = 173, range = 151–262) SBSs. In contrast, the total number of SBSs was considerably increased in the GA clones, with an average of 485 SBSs (median = 448, range = 370–592) (Supplemental Tables S1, S2). This finding reveals stronger mutagenic properties of GA in the MEFs.

Principal component analysis (PCA) performed on the resulting SBS spectra unambiguously separated the GA clones from other experimental conditions (Fig. 1A). The ACR-exposed samples showed a diffuse pattern across the six SBS classes, whereas the Spont clones showed an enrichment of C:G > G:C SBS in the 5′-GCC-3′ context, also present across the exposed cultures (Supplemental Fig. S2). This background mutation type appears related to the culture conditions used for the MEF immortalization assay, and its consistent formation has been observed previously (Olivier et al. 2014; Nik-Zainal et al. 2015). No significant transcription strand bias (TSB) was observed for any mutation class in the Spont or ACR clones (Supplemental Fig. S3). In the clones derived from the GA-treated primary MEF cultures, we observed an enrichment of T:A > A:T and C:G > A:T transversions and T:A > C:G transitions (Supplemental Fig. S2B), marked by significant TSB (Supplemental Fig. S3). The GA-associated clones showed lower numbers (25 per clone) of small insertions/deletions (indels) in comparison to the ACR (44 per clone) or Spont clones (39 per clone) (see Supplemental Tables S1, S3). Thus, higher SBS counts owing to GA treatment may selectively promote the senescence bypass and the selection, with a decreased functional contribution of indels, whereas an inverse scenario is plausible for the Spont and ACR clones, consistent with a previous report based on the Big Blue mouse embryonic fibroblasts and *cII* transgene (Besaratinia and Pfeifer 2005).

Variant allele frequency (VAF) analysis performed for GA clones detected a large proportion of acquired mutations manifesting at VAF between 25% and 75% (Supplemental Fig. S4C). Upon grouping of substitutions into bins of high (67%–100%), medium (34%–66%), and low (0%–33%) VAF, the predominant GA-specific mutation types (T:A > A:T, T:A > C:G, and C:G > A:T) started manifesting at high VAF and became increasingly enriched in the medium and low VAF intervals. The background 5′-N[T>G]T-3′ SBS, corresponding to COSMIC signature 17 arising in cultured mouse cells including MEFs (Behjati et al. 2014; Nik-Zainal et al. 2015; Milholland et al. 2017), displayed minor, although not statistically

Table 1. Summary of cell lines, treatment conditions, and *TP53* mutation status

Sample ID	Embryo	Exposure	Conc. (mM)	Exposure duration (h)	Coding DNA change[a]	Genomic DNA change[b]	aa change	Codon 72 (rs1042522)[c]
Prim_1	E210	-	-	-				Pro/Pro
Prim_2	E213	-	-	-				Arg/Pro
Prim_3	E214	-	-	-				Pro/Pro
Spont_1	E213	-	-	-				Arg/Pro
Spont_2	E214	-	-	-				Pro/Pro
Spont_3	E214	-	-	-				Pro/Pro
ACR_S9_1	E213	ACR	5	24				Arg/Pro
ACR_S9_2	E213	ACR	5	24				Arg/Pro
ACR_1	E213	ACR	10	24	c.881delA	g.7577057delT	p.E294fs	Arg/-
ACR_2	E213	ACR	10	24	c.818G>T	g.7577120C>A	p.R273L	Pro/-
ACR_3	E214	ACR	10	24	c.740A>T; c.839G>C	g.7577541T>A; g.7577099C>G	p.N247I; p.R280T	Pro/Pro
ACR_4	E214	ACR	10	24				Pro/Pro
ACR_5	E214	ACR	10	24				Pro/Pro
GA_1	E210	GA	3	24				Pro/Pro
GA_2	E210	GA	3	24				Pro/Pro
GA_3	E210	GA	3	24	c.309-310CC>TA	g.7579377-7579378GG>TA	[p.Y103Y; p.Q104K]	Pro/Pro
GA_4	E214	GA	3	24				Pro/Pro
GA_5	E214	GA	3	24				Pro/Pro

(*TP53*) human *TP53* gene; (Prim) primary cells; (Spont) spontaneously immortalized clones; (ACR) acrylamide-exposure derived clones; (GA) glycidamide-exposure derived clones. Each exposure condition was carried out in two biological replicates (embryos). (S9) human S9 fraction; (Pro) proline; (Arg) arginine; (Arg/-) or (Pro/-) loss of allele; (fs) frameshift; (aa) amino acid.
[a]NM_000546.4 coding sequence.
[b]hg19 genomic coordinates.
[c]Human polymorphic site (rs1042522).

significant, lower-VAF enrichment ($P = 0.25$, assessed by χ^2 test) (Supplemental Fig. S5). These observations suggest early effects of the GA exposure, reproducible contribution of the induced mutations to senescence bypass, and their clonal propagation during the immortalization stage.

Mutational signature of GA

Three distinct mutational signatures were extracted from all MEF clones, termed signatures A, B, and C. Signatures A and C were enriched in the Spont and ACR clones, whereas the more robust signature B was selectively enriched in the GA clones (Fig. 1B; Supplemental Fig. S6). The TSB analysis in the GA clones revealed significant enrichment of the prominent mutation types C:G > A:T, T:A > A:T, and T:A > C:G (using the pyrimidine-based mutation class convention) on the transcribed strand ($P < 0.05$, χ^2 test), consistent with the less efficient transcription-coupled nucleotide excision repair because of adduct formation on purines (Fig. 1C; Supplemental Fig. S3). In signature C and to a lesser extent in signatures A and B, we observed an admixture of a pattern identical to the COSMIC signature 17 (T:A > G:C in the 5′-NTT-3′ trinucleotide context), present in human cancers (notably esophageal and gastric adenocarcinomas) but also seen in AFB1-driven mouse liver cancers (Huang et al. 2017), in murine small cell lung carcinoma initiated by loss of *Trp53* and *Rb1* (McFadden et al. 2014), and in primary MEF-derived clones (Olivier et al. 2014; Nik-Zainal et al. 2015). This signature has been linked to cell culture conditions (Behjati et al. 2014; Milholland et al. 2017) and may be linked to oxidative stress effects on the free dGTP pool (Tomkova et al. 2018). To further refine the putative GA mutational signature from signature B, we used extended-input nonnegative matrix factorization (NMF) by combining the MEF clone data with signature 17–rich esophageal adenocarcinoma data from the International Cancer Genome Consortium (ICGC) ESAD-UK study (Secrier et al. 2016), as well as with The Cancer Genome Atlas (TCGA) esophageal adenocarcinoma (ESCA) and gastric carcinoma (STAD) samples enriched for or lacking signature 17 (see Methods) (Supplemental Methods; Supplemental Figs. S6, S7). This considerably reduced (average = 47%, median = 48%) the signature 17–specific T > G peaks in signature B associated with GA treatment and resulted in a cleaner pattern (Fig. 1D; Supplemental Fig. S6). The refined GA signature retains the strand-biased enrichment of the T:A > A:T transversions and T:A > C:G transitions in the 5′-CTG-3′ and 5′-CTT-3′ trinucleotide contexts, as well as the C:G > A:T component (Fig. 1D; Supplemental Fig. S8A; Supplemental Table S4).

Quantitative DNA adduct analysis supports the GA mutational signature

Following metabolic activation, ACR induces GA-DNA adducts at the N7 and N3 positions of guanine and adenine, respectively. Analysis using liquid chromatography–tandem mass spectrometry (LC-MS/MS) revealed the absence of these adducts in the untreated samples, as well as in MEFs exposed to ACR in the absence of S9 fraction (with levels below the limit of detection [LOD]). This suggests a lack of Cyp2e1 activity normally required for the metabolism of ACR to GA in the MEFs. Upon addition of human S9 fraction, N7-(2-carbamoyl-2-hydroxyethyl)-guanine (N7-GA-Gua) levels increased to 11 adducts/10^8 nucleotides (twice the LOD levels), suggesting limited metabolic activation of ACR despite the enzymatic activity of the S9 fraction (Fig. 1E,F). In contrast, cells exposed to GA showed high DNA adduct levels, with N7-GA-Gua and N3-(2-carbamoyl-2-hydroxyethyl)-adenine (N3-GA-Ade) observed at 49,000 adducts/10^8 nucleotides and 350 adducts/10^8 nucleotides, respectively, after subtracting the trace amount of contamination from the internal standard (Fig. 1E,F). These observed DNA adducts provide a possible mechanistic basis for the mutation types, the TSB, and the mutational signature arising upon treatment with GA, the reactive metabolite of ACR.

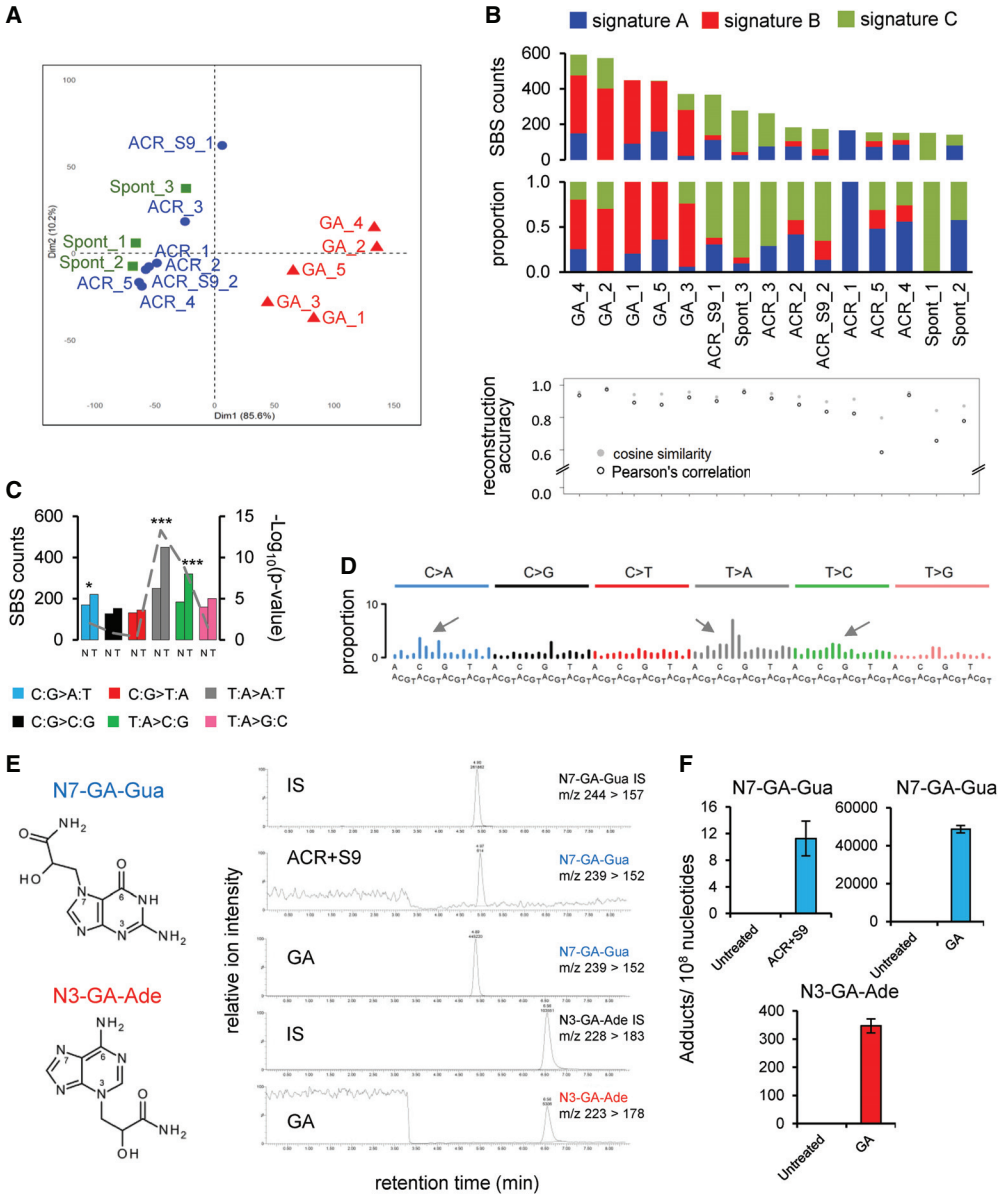


Figure 1. Analysis of the mutation patterns derived from experimental exome sequencing data. (*A*) Principle component analysis (PCA) of WES data. PCA was computed using as input the mutation count matrix of the clones that immortalized spontaneously (Spont) or were derived from exposure to acrylamide (ACR) or glycidamide (GA). Each sample is plotted considering the value of the first and second principal components (Dim1 and Dim2). The percentage of variance explained by each component is indicated within brackets on each axis. Spont and ACR- and GA-exposed samples are represented by differently colored symbols. (*B*) Mutational signatures (sig A, sig B, and sig C), identified by NMF, and their contribution to each sample (*x*-axis), assigned either by absolute SBS counts or by proportion (bar graphs). The reconstruction accuracy of the identified mutational signatures in individual samples is shown in the *bottom* dot plot (*y*-axis value of 1 = 100% accuracy). (*C*) Transcription strand bias analysis for the six mutation types in GA-exposed clones. For each mutation type, the number of mutations occurring on the transcribed (T) and nontranscribed (N) strand is shown on the *y*-axis. (***) $P<10^{-8}$, (*) $P<10^{-2}$. (*D*) Extraction of GA signature, with arrows pointing at the enriched SBS classes. The contribution of signature 17 (T:A > G:C in 5′-NTT-3′ context), present in all clones, was decreased by performing NMF on human-*TP53* knock-in (Hupki) MEF samples pooled with primary tumor samples with high levels of signature 17 (see Methods and Supplemental Methods). (*E*) DNA adducts analysis as determined by LC-MS/MS. (*F*) Levels of N7-GA-Gua adduct in ACR + S9- and GA-treated cells and N3-GA-Ade DNA adduct level in GA-treated cells compared with untreated cells yielding no adducts. The data are presented as the number of adducts in 10^8 nucleotides in replicated experiments ($n \geq 2$).

Comparison of the GA signature with PCAWG mutational signatures

We next performed cosine similarity comparison of the putative GA signature with the recently updated PCAWG SBS mutational signatures (Alexandrov et al. 2018) and with known T:A > A:T-rich experimental signatures (Fig. 2A; Supplemental Figs. S7, S9). The highest cosine similarity value (84%) corresponded to PCAWG SBS25 (Fig. 2A). However, unlike the GA signature, neither SBS25 nor any other signatures show TSB for the three

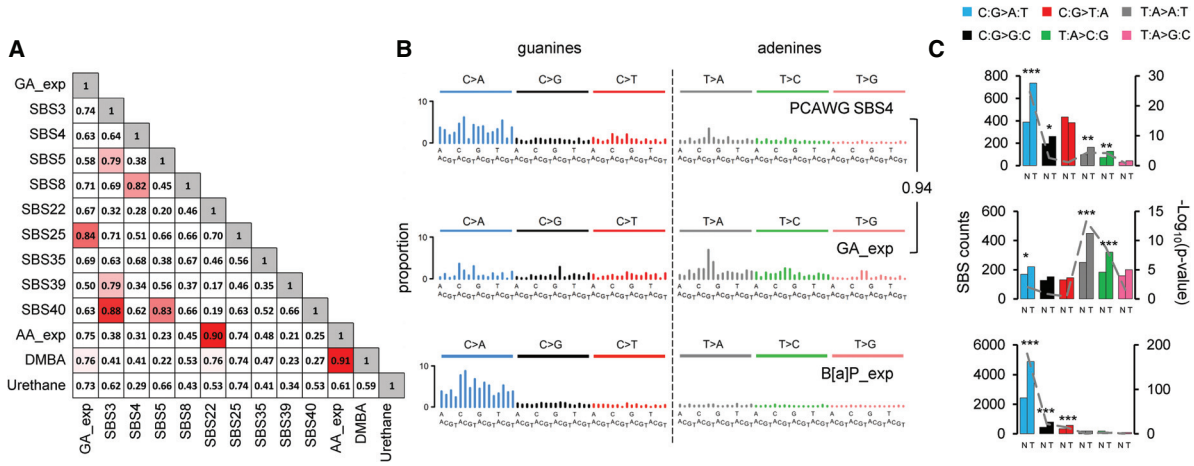


Figure 2. Comparison of GA signature to known signatures. (*A*) Cosine similarity matrix comparing GA mutational signature with the human PCAWG data (SBS3, -4, -5, -8, -22, -25, -35, -39, and -40) and other A > T-rich mutational signatures from experimental exposure assays using specific carcinogens (7,12-dimethylbenz[*a*]anthracene [DMBA], urethane, and aristolochic acid [AA]). (*B*) Comparison of PCAWG SBS4 with two experimentally derived signatures: B[a]P_exp = benzo[*a*]pyrene mutational signature extracted from HMECs; GA_exp = GA mutational signature extracted from MEF cells. Cosine similarity between the T > N (adenine) components of SBS4 and GA signature is shown on the *right*. (*C*) Transcription strand bias analysis for the six mutation types underlying the signatures in panel *B*. For each mutation type (using the pyrimidine convention), the number of mutations occurring on the transcribed (T) and nontranscribed (N) strand is shown on the *left y*-axis. The significance is expressed as $-\log_{10}$(*P*-value) indicated on the *right y*-axis. (***) $P<10^{-8}$, (**) $P<10^{-4}$, (*) $P<10^{-2}$.

mutation classes (C:G > A:T, T:A > A:T, and T:A > C:G). Thus, the mutation patterns with a three-class strand bias generated by the GA treatment render the resulting mutational signature unique and novel.

GA signature in the human pan-cancer genomes

The initial visual comparison with PCAWG signatures indicated similarity between the GA signature and signature SBS4 of tobacco smoking (Supplemental Fig. S8; Alexandrov et al. 2018), in keeping with the established presence of ACR in tobacco smoke. This was further corroborated by the cosine similarity of 94% between the adenine (T > N) components of SBS4 and the GA signature (Fig. 2B). We thus hypothesized that SBS4 reflects the coexposure to benzo[*a*]pyrene (B[a]P; generating the predominant, strand-biased C > N/guanine mutations) and to GA (generating strand-biased T > N/adenine mutations) (Fig. 2B,C; Supplemental Fig. S8). To obtain experimental evidence, we modeled a B[a]P mutational signature by whole-genome sequencing (WGS) of cell clones derived from B[a]P-exposed normal human mammary epithelial cells (HMECs) (Stampfer and Bartley 1985, 1988). This yielded a robust pattern characterized by predominant strand-biased guanine (mainly C:G > A:T) mutation levels and negligibly mutated adenines (T > N) (Fig. 2B; Supplemental Figs. S8, S10; Supplemental Table S4). Next, we interrogated the PCAWG data for the presence of the experimentally defined, 192-class (strand-biased) GA and B[a]P signatures in 1584 tumors of 19 cancer types from 14 organ sites (Fig. 3; Supplemental Table S5). The stringency of the process was controlled by determining the *P*-value and the false-discovery rate (FDR) for the signature presence test and the reconstruction accuracy (Supplemental Table S6) and by modeling false-positive rates (FPRs) and FDRs of the experimental signature detection using 2000 synthetic tumors as described in the Methods and in Supplemental Tables S7 through S10. In the subset of PCAWG-7 cancers known to carry SBS4 signature (adenocarcinomas and squamous cell carcinomas of the lung, hepatocellular carcinomas of the liver and head, and neck squamous cell carcinomas), we compared the GA and B[a]P signatures to estimated levels of SBS4 and found that in the lung and head and neck cancers, a combination of the GA and B[a]P signatures accounted for very similar numbers of mutations as SBS4, suggesting that SBS4 represents combined and highly correlated exposure to GA and B[a]P (Fig. 3A). In contrast, we found more variability in the assignment of mutation numbers to GA and B[a]P versus SBS4 in liver cancers (Fig. 3), which may reflect a weaker relationship between GA and B[a]P exposure because of generally more complex exposure history in the liver. Successful reconstruction of SBS4 by the experimental 192-class (strand-biased) GA and B[a]P signatures in the lung and liver human tumors enabled correct assignment of the GA signature in a subset of 24 lung adenocarcinomas, 42 lung SCCs, and 239 liver tumors with a subset of 184 GA-positive HCCs lacking the B[a]P signature mutations (Fig. 3B; Supplemental Table S11). Moreover, we identified the GA signature in additional 15 cancer types without SBS4, including clear cell renal cell carcinoma (78 GA-positive of 111 analyzed tumors), papillary renal cell carcinoma (26 GA-positive out of 32), biliary adenocarcinoma (20 GA-positive out of 35), colorectal adenocarcinomas (24 GA-positive out of 60), stomach adenocarcinoma (17 GA-positive out of 75), bladder transitional cell carcinoma (six GA-positive out of 23), and uterine adenocarcinoma (10 GA-positive out of 51) (Fig. 3B,C). The signature assignments results for the 537 individual GA-positive PCAWG tumors are summarized by cancer type in the Supplemental Table S11.

Discussion

ACR and GA exposures induce an almost identical set of tumors in both mice and rats, providing a substantial argument for a GA-mediated tumorigenic effect of ACR (Beland et al. 2015). This is supported by further mechanistic studies showing that lung tissue from mice exposed to ACR and GA displays comparable DNA adduct patterns, as well as similar mutation frequencies in the *cII* transgene (Manjanatha et al. 2015). Similar observations were

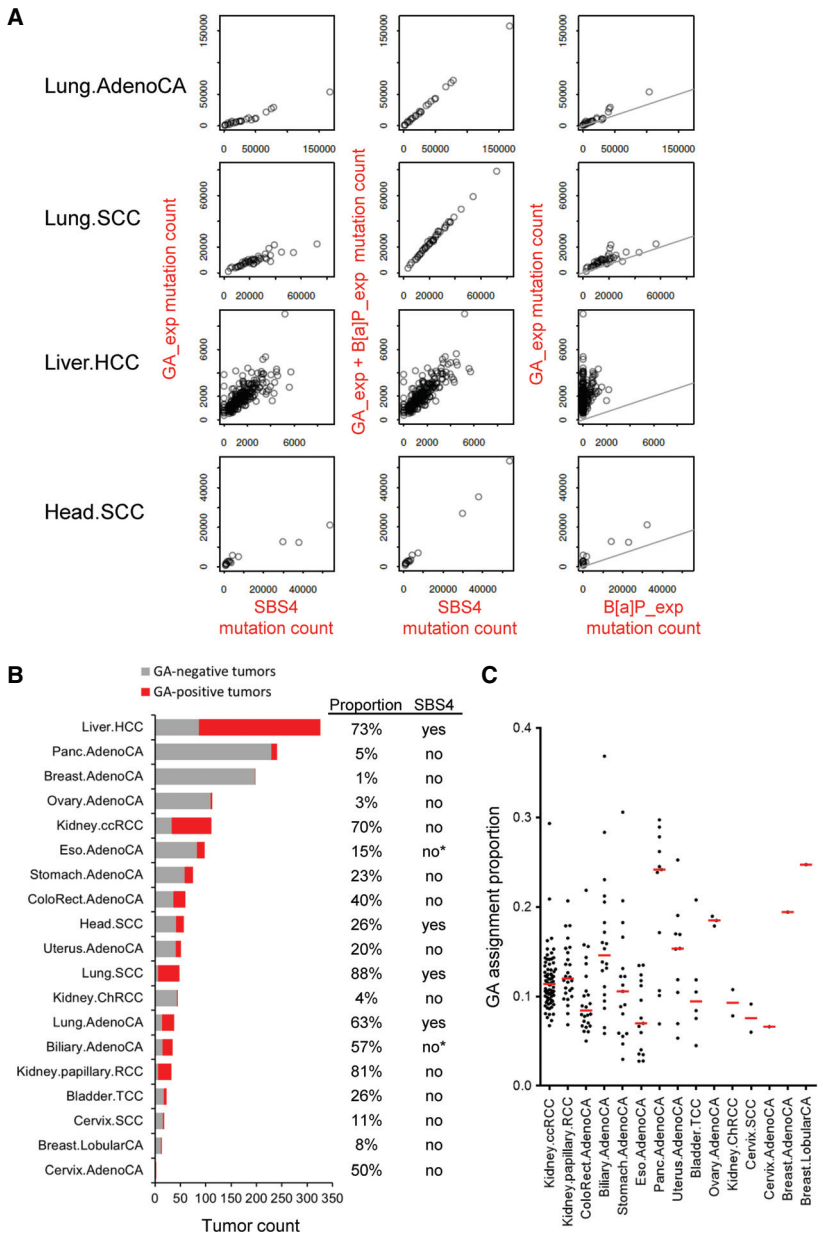


Figure 3. Identification of experimental GA signature in the human cancer PCAWG data sets. (*A*) Scatter plots of the experimental GA_exp and B[a]P_exp mutational signature assignments by mSigAct show reconstruction of tobacco-smoking signature SBS4 assignments in cancer types with SBS4 present. (Lung.AdenoCA) Lung adenocarcinoma, (Lung.SCC) lung squamous cell carcinoma, (Liver.HCC) liver hepatocellular carcinoma (Head.SCC) head squamous cell carcinoma. The combination of GA_exp and B[a]P_exp mutation counts reconstructed SBS4 mutation counts in Lung.AdenoCA and Lung.SCC and, to an extent, in Head.SCC. In liver HCCs, GA counts alone partially reconstructed SBS4 mutation counts and indicate GA_exp-positive and B[a]P_exp-negative tumors (third row, *right* scatter plot). The lines in GA versus B[a]P scatter plots have a slope of 0.3, reflecting the 3:1 ratio of B[a]P:GA mutation counts that reconstruct SBS4. (*B*) Summary of GA mutation assignment analysis of 1584 individual tumors of 19 cancer types from the PCAWG data sets. Assignments were performed using mSigAct (positivity was determined by the signature.presence.test tool at FDR < 0.05) with the PCAWG annotations of signature present in each subtype, in addition to the GA and B[a]P signatures. The tumor types manifesting or lacking SBS4 signature of tobacco smoking are labeled accordingly in the column SBS4. Asterisk denotes borderline SBS4 presence in PCAWG Billiary.AdenoCA (two of 173, 1.16%) and Eso.AdenoCA (two of 347, 0.06%). Proportion indicates percentage of GA-positive tumors within each listed cancer type. (*C*) The dot plot shows the proportion of mutations assigned to GA signature among other identified signatures (see Supplemental Material) in individual tumors of cancer types not showing the direct effects of tobacco smoking (i.e., lacking signature SBS4). Red horizontal lines denote median values (*y*-axis, 1 = 100%).

made in the context of in vitro mutagenicity of ACR in human and mouse cells, suggesting the key role for the epoxide metabolite GA to form premutagenic DNA adducts (Besaratinia and Pfeifer 2004). Thus, in keeping with the established ACR/GA carcinogenicity in rodents (IARC 1994; Olstørn et al. 2007; Von Tungeln et al. 2012; Beland et al. 2015), our findings provide new information on the characteristic mutagenic effects of GA and their contribution to tumor development.

The observation that ACR itself is not efficiently metabolized by MEFs is consistent with similar differences reported by previous animal carcinogenicity studies. In neonatal B6C3F1 mice, GA, but not ACR, induces hepatocellular carcinomas, likely because of the inability of neonatal mice to efficiently metabolize ACR (Von Tungeln et al. 2012). Moreover, unlike ACR, GA induces tumors in the small intestine in a dose-dependent manner upon perinatal exposure (Olstørn et al. 2007). Similar differences between GA and ACR mutagenicity, possibly because of limited metabolization of ACR, were observed in vitro (Besaratinia and Pfeifer 2004). We addressed the lack of ACR activation by the addition of human S9 fraction, yet the assessment of DNA adducts suggested limited metabolic activation of ACR with adduct levels substantially lower compared with the direct GA exposure. This may explain the mutagenicity differences observed between GA and ACR. A consistent minor contribution of the GA mutational signature was detected in the majority of ACR clones, whereas it was mostly absent in the Spont clones, suggesting subtle metabolic activation of ACR in the MEFs resulting in low levels of GA. However, a robust mutational signature in the experimental setting was generated exclusively by exposing the cells directly to GA.

Single reporter gene studies had previously linked ACR and GA exposure to multiple different mutation types. Thanks to the larger number of mutations obtained by exome sequencing, we were able to attribute to the GA exposure a particular mutational signature characterized by three strand-biased mutation classes (C:G > A:T, T:A > A:T, and T:A > C:G). The identification of the N7-GA-Gua and N3-GA-Ade DNA adducts originating from the metabolic conversion of ACR (Segerbäck et al. 1995; da Costa et al. 2003; Besaratinia and Pfeifer

2005), underlines the relationship between DNA adduct profiles and the mutational signature of GA. N3-GA-Ade and N7-GA-Gua are depurinating adducts resulting in apurinic/apyrimidinic sites. During replication, these lead to misincorporation of deoxyadenine, leading to the respective T:A > A:T and C:G > A:T transversions observed in the GA signature. The T:A > C:G transitions enriched in the GA signature correspond to the miscoding N1-GA-Ade adduct, the most commonly identified adenine adduct in vitro (Randall et al. 1987; da Costa et al. 2003; Besaratinia and Pfeifer 2005; Ishii et al. 2015). The levels of the guanine adduct were especially high in the GA-exposed MEF cells, whereas the associated C:G > A:T transversions in the resulting postsenescence clones were less represented. This could reflect differences in DNA repair efficiency concerning the individual guanine and adenine adduct species or the fact that the resulting clones are derived from single cells that selectively immortalized but do not accurately represent the bulk exposed primary cell population in which the GA-DNA adduct levels were measured after exposure. It is also plausible that the excessive and possibly highly cytotoxic N7-GA-Gua adduct burden leads to negative selection of a large number of affected cells.

The established animal models (Beland et al. 2013, 2015) of ACR- and GA-mediated tumorigenesis provide a suitable starting point for a comparison of the mutational signatures obtained from the mouse and in vitro. Next, genome-scale sequencing of human tumors and adduct analysis of normal tissues collected in well-designed molecular epidemiological studies focusing on ACR intake are warranted to provide further evidence that the GA signature mutations identified in various cancer types indeed correlate with the exposure to ACR.

The GA signature has not been identified among the currently known computationally extracted PCAWG signatures (Fig. 2A; Alexandrov et al. 2018). Here we show that a new pattern can be identified in a large subset of pan-cancer tumors when experimentally modeled signatures are combined with sophisticated computational signature reconstruction methods while considering the extended features, such as TSB supported by premutagenic adduct analysis. Such integrated approaches can thus lead to future identification of yet unrecognized carcinogen signatures that may be eluding the solely computation-based analyses of the pan-cancer data.

The quest for understanding the contribution of ACR to cancer development is reflected by recent accumulation of mechanistic data on the compound's mutagenicity and carcinogenicity in experimental models. The possibly carcinogenic effects of ACR in humans were recommended for re-evaluation by the Advisory Group to the Monographs Program of the International Agency for Research on Cancer (Straif et al. 2014). Our findings related to the reconstruction of signature SBS4 by the experimental signatures of GA and B[a]P, together with the detection of the GA signature in lung and liver cancer, are relevant given the established high content of ACR in tobacco smoke. Compared with the GA effects, experimental B[a]P exposure generates very few T > N (adenine) mutations. However, we cannot exclude a possibility that in the human tissues directly exposed to tobacco smoke the adenine residues can be targeted by carcinogens such as B[a]P derivatives or nitrosamines.

A subset of 184 liver tumor samples identified in this study harbored the GA signature but no features of the B[a]P signature or SBS4 (Fig. 3B; Supplemental Material). Furthermore, we found 217 GA-positive, SBS4-negative tumors of additional 15 cancer types (Fig. 3B,C). The numerous GA-positive, SBS4-negative tumors are of particular interest as they likely reflect dietary and/or occupational exposures to ACR unrelated to tobacco smoking. Overall, our findings offer new insights into the thus-far tenuous association of ACR with human carcinogenesis.

Methods

Source and authentication of primary cells

Primary human-p53 knock-in (Hupki) MEFs were isolated from 13.5-d-old *Trp53*$^{tm/Holl}$ mouse embryos from the Central Animal Laboratory of the Deutsches Krebsforschungszentrum as described previously (Liu et al. 2004). The mice had been tested for specific pathogen-free (SPF) status. The derived primary cells were genotyped for the human *TP53* codon 72 polymorphism (Table 1) to authenticate the embryo of origin. Cells from three different embryos (E210, E213, and E214) were used for the exposure experiments (Table 1). All subsequent cell cultures were routinely tested at all stages for the absence of mycoplasma.

Cell culture, exposure, and immortalization

The primary MEF cells were expanded in advanced DMEM supplemented with 15% fetal calf serum, 1% penicillin/streptomycin, 1% pyruvate, 1% glutamine, and 0.1% β-mercaptoethanol. The cells were then seeded in six-well plates and, at passage 2, were exposed for 24 h to 5 mM ACR (A4058, Sigma-Aldrich) in the presence of 2% human S9 fraction (Life Technologies) complemented with NADPH (Sigma-Aldrich) or the absence of S9 to 10 mM ACR or 3 mM GA (04704, Sigma-Aldrich), or to vehicle (PBS). Exposed and untreated control primary cells were cultured until they bypassed senescence and immortalized clonal cell populations could be isolated (Todaro and Green 1963). The HMEC cultures used in this study for WGS were generated from primary HMECs (passage 4) exposed to B[a]P and propagated in M87A medium to passage 13, as described previously (Stampfer and Bartley 1985, 1988; Garbe et al. 2009; Severson et al. 2014).

MTT assay for cell metabolic activity and viability

Cells were seeded in 96-well plates and treated as indicated. Cell viability was measured 48 h after treatment cessation using the CellTiter 96 AQ_{ueous} One Solution Cell Proliferation Assay (Promega). Plates were incubated for 4 h at 37°C, and absorbance was measured at 492 nm using the Apollo 11 LB913 plate reader. The MTT assay was performed in triplicate for each experimental condition.

Phospho-H2AFX immunofluorescence

Immunofluorescence staining of phosphorylated histone H2AFX (γH2AFX) was performed using phospho-histone H2A.X (Ser139) (20E3) Rabbit monoclonal antibody (9718, Cell Signaling Technology). Briefly, primary MEFs were seeded on coverslips in 12-well plates and, the following day, treated as indicated in duplicate for 24 h. Four hours after treatment cessation, the cells were fixed with 4% formaldehyde for 15 min at room temperature. Following blocking in 5% normal goat serum (31872, Life Technologies) for 60 min, they were incubated with the γH2AFX-antibody (1:500 in 1% BSA) overnight at 4°C. Subsequent incubation with a fluorochrome-conjugated secondary antibody (4412, Cell Signaling Technology) was performed for 60 min at room temperature. Coverslips were mounted in Vectashield mounting medium with DAPI (Eurobio). Immunofluorescence images were captured using a Nikon Eclipse Ti.

DNA adduct analysis

GA-DNA adducts (N7-GA-Gua and N3-GA-Ade) were quantified by LC-MS/MS with stable isotope dilution as previously described (da Costa et al. 2003). The DNA was isolated from the cells using standard digestion with Proteinase K, followed by phenol-chloroform extraction and ethanol precipitation. The DNA was subsequently treated with RNase A and T1, extracted with phenol-chloroform, and reprecipitated with ethanol. N7-GA-Gua and N3-GA-Ade were released by neutral thermal hydrolysis for 15 min, using Eppendorf Thermomixer R (Eppendorf North America) set to 99°C. The samples were filtered through Amicon 3K molecular-weight cutoff filters (Merck Millipore) to separate the adducts from the intact DNA. The LC-MS/MS used for quantification consisted of an Acquity UPLC system (Waters) and a Xevo TQ-S triple quadrupole mass spectrometer (Waters). The following MRM transitions were monitored with a cone voltage of 50 V and a collision energy of 20 eV: N3-GA-Ade, m/z 223→178; [15N5]N3-GA-Ade (internal standard), m/z 228→183; N7-GA-Gua, m/z 239→152; and [15N5]N7-GA-Gua (internal standard), m/z 244→157 (da Costa et al. 2003).

TP53 genotyping

Exons 4 to 8 of the knocked-in human *TP53* gene (NC_000017.11) were sequenced using standard protocols. Sanger sequencing of PCR products was performed at BIOfidal, using the Applied Biosystems 3730xl genetic analyzer. The amplicon and sequencing primers are listed in the Supplemental Methods. Sequences were analyzed using the CodonCode Aligner version 7.1 software.

Library preparation and WES

Refer to the online Supplemental Methods for details on the standard procedures for library preparation and WES, sequencing data preprocessing, read alignment, and the calling of the SBS and indel variants in the MEF and HMEC cell lines.

Bioinformatics and extraction of experimental mutational signatures

Refer to the Supplemental Methods for detailed information on PCA, assessment of sequencing-related artifacts and damage, and computation of the TSB and its significance. The TSB was considered statistically significant at P-value ≤ 0.05. To analyze the mutation spectra and treatment-specific mutational signatures, filtered mutations were classified into 96 types corresponding to the six possible base substitutions (C:G > A:T, C:G > G:C, C:G > T:A, T:A > A:T, T:A > C:G, T:A > G:C) and the 16 combinations of flanking nucleotides immediately 5′ and 3′ of the mutated base. Mutation patterns were then deconvolved into mutational signatures using NMF (Brunet et al. 2004; Alexandrov et al. 2013b) embedded in the MutSpec suite (Ardin et al. 2016). For details on estimates of the optimal number of signatures to extract, see the Supplemental Methods. The reconstruction error calculation evaluated the accuracy with which the deciphered mutational signatures describe the original mutation spectra of each sample by applying Pearson's correlation and cosine similarity.

The GA mutational signature was further polished by using an extended input including samples from ICGC (ESAD-UK study) with high level of signature 17 (>65% contribution as determined by independent NMF analysis), and with samples from the TCGA esophageal adenocarcinoma (ESCA) and gastric cancer (STAD) collection (exon data, to address comparable coverage of the genome). The samples used for this procedure are listed in the Supplemental Methods, and the results are summarized in Supplemental Figures S6 and S7.

Cosine similarity analysis was used to evaluate the concordance between the identified T:A > A:T-rich mutational signature of GA with the newly characterized SBS mutational signatures from the PCAWG (pan-cancer whole genome) data (Alexandrov et al. 2018). Cosine similarity values of more than 0.5 were found for PCAWG SBS3, SBS4, SBS5, SBS8, SBS22, SBS25, SBS35, SBS39, and SBS40 and the experimentally derived mutational signature of AA (Olivier et al. 2014; Ardin et al. 2016), 7,12-dimethylbenz [*a*]anthracene (DMBA) (McCreery et al. 2015; Nassar et al. 2015), and urethane (Westcott et al. 2014).

The experimental B[a]P signature was generated by WGS (using Illumina HiSeq X Ten by GENEWIZ) of finite lifespan poststasis clones derived from primary HMECs treated with B[a]P, as previously described (Stampfer and Bartley 1985, 1988; Severson et al. 2014). Following read alignment to NCBI GRCh38 genome build, mutations were called in the two poststasis samples with MuTect2 or Strelka2.8 using a primary HMEC sample as a comparison. Only mutations called by both algorithms were retained, and additional criteria were applied to filter out mutations with a match in public SNP databases (dbSNP150, and/or AF > 0.001 in either 1000 Genomes, gnomAD or NHLBI-ESP), with an allele frequency above zero in the primary sample, with coverage lower than 10 reads, or mutations overlapping tandem repeats. Finally, a cut-off was applied on VAF, and only mutations with a VAF equal or higher than 20% were retained, being 54,587 unique mutations. The NMF procedure to extract the experimental B[a]P signature used input extended with SBS data from the TCGA lung cancer collection (15 Lung.AdenoCA positive [>50%] for tobacco-smoking SBS4, 15 Lung.AdenoCA negative for SBS4, 15 Lung.SCC positive [>50%] for SBS4 and 15 Lung.SCC negative for SBS4). See the Supplemental Methods for sample details. The recovered signatures showed the strongest enrichment of the C > A-based signature B (Supplemental Fig. S10) in the B[a]P-treated HMEC clones. We next calculated the reconstruction error to evaluate the accuracy with which the extracted B[a]P_exp signature describes the original mutation spectra of each sample by applying Pearson's correlation and cosine similarity (Supplemental Fig. S10).

Identification of the experimental signatures in PCAWG data

We used the mutational signature activity (mSigAct v0.10.R) software (Ng et al. 2017) to test for the presence of the experimental mutational signatures of GA and B[a]P in the human primary tumor data from PCAWG study. mSigAct conducts a statistical test for optimal reconstructions of the observed human tumor mutation spectrum with and without the GA mutational signature, in addition to a set of other mutational signatures from the PCAWG study. The 192-class strand-biased versions of the GA and B[a]P mutational signatures (Supplemental Fig. S8; Supplemental Table S4) were used to detect tumors with the experimentally defined signatures present, at high stringency achieved also by incorporating the same TSB information in the 192-class reconstructions of each tumor. To generate a 192-class reconstructed spectrum, the assignment of mutation counts for each 192-class signature is determined by mSigAct and multiplied with the 192-class versions of the PCAWG, GA, and B[a]P mutational signatures. The 192-class versions of each signature and spectrum is equivalent to the 96-class versions when the mutation counts on each strand are summed and then represented in the pyrimidine mutations (C > A, C > G, C > T, T > A, T > C, T > G). Specifically, B[a]P was added to cancer types with tobacco-smoking SBS4 signature previously found in the PCAWG signature set, and a combination of B[a]P and GA signatures was used in these cancers to reconstruct SBS4.

For other signatures and cancers without evidence of SBS4 present, only GA was used to reconstruct the tumor spectra. This was followed by computing the likelihood ratio test between the original spectrum and the reconstructed tumor. A total of 1673 tumor samples from the PCAWG repository from 20 cancer subtypes were interrogated. We excluded hypermutated and recently identified AA signature–containing tumors (Ng et al. 2017) as the presence of strong T>A signature adversely affected the reconstruction process. A set of active mutational signatures were obtained from the PCAWG annotations of each cancer subtype, with flat signatures (SBS3, SBS8) removed to improve the sparsity of the mutation assignments. Final assignments of mutations to each mutational signature were performed by using the 96-class mutational signatures. Further fine-tuning was conducted using parameters for a negative binomial model, and the FDR was adjusted for mutational signature presence (FDR < 0.05).

The proportion matrices of the strand-biased and NMF versions of the experimental GA signature, the GA signature normalized to the human genome trinucleotide frequency to allow for human PCAWG data screening, and the strand-biased and NMF versions of the whole-genome B[a]P signature are available in Supplemental Table S4. The statistics underlying the assignment of GA_exp to PCAWG cancer data sets (*P*-values for "signature.presence.test" and cosine similarity between the reconstruction and spectra) are summarized in Supplemental Table S5.

FPR and FDR estimation for GA signature detection in synthetic tumors

To determine how often false positives arise when detecting the GA signature with mSigAct and to accurately estimate the FDR of the detection of GA signature, we performed a deeper validation analysis. We generated 2000 synthetic tumors with signatures from the PCAWG-7 data set and assignments sampled from the assignments to each signature in the PCAWG-7 data set, which represented the tumor types in which we found GA signature present, with similar signatures and mutation burdens associated with each signature. The synthetic tumors had the same frequency of observing a particular signature for a cancer type, similar to the PCAWG-7 tumors. One hundred tumors per 20 tumor types (included in the main analysis and listed in Supplemental Table S9) have been generated, with 1015 of the tumors harboring GA signature and 985 with GA signature absent. By using the synthetic tumor set and mSigAct to assign GA signature, we established the true-positive rates (TPRs), FPRs and FDRs (calculated by using the raw synthetic tumor counts and the formula FP/(TP + FP)). The results are shown as a short summary (Supplemental Table S7), raw tumor counts (Supplemental Table S8), per cancer type distribution (Supplemental Table S9) and a full listing of TPRs, FPRs, and FDRs (Supplemental Table S10).

Data access

Aligned WES reads from the primary MEF cells and clones arising from ACR- and GA-treated cultures and immortalized spontaneously, as well as Sanger sequencing files, have been submitted to the NCBI BioProject database (BioProject; https://www.ncbi.nlm.nih.gov/bioproject) under accession number PRJNA238303 (for the individual BioSample accession numbers, refer to Supplemental Tables S12, S13). The WES data reported here are a new extension of the BioProject PRJNA238303 dedicated to systematic identification of mutational signatures of carcinogenic agents (Olivier et al. 2014).

Acknowledgments

We thank the New York University Genome Technology Center, funded in part by the NIH/NCI Cancer Center support grant P30CA016087, and GENEWIZ, for expert assistance with Illumina sequencing. The study was supported by funding obtained from INCa-INSERM (Plan Cancer 2015 grant to J.Z.), NIH/NIEHS (1R03ES025023-01A1 grant to M.O.), and the Singapore National Medical Research Council (NMRC/CIRG/1422/2015 grant to S.G.R.) and the Singapore Ministry of Health via the Duke-NUS Signature Research Programmes to S.G.R. M.R.S. was supported by the U.S. Department of Energy under contract no. DE-AC02-05CH11231. M.P. was supported by the European Commission FP7 Marie Curie Actions-People-COFUND Fellowship. The views expressed in this article do not necessarily represent those of the U.S. Food and Drug Administration.

Author contributions: M.Z., M.K., and J.Z. drafted the manuscript and prepared figures. M.I.C. and F.A.B. performed DNA adduct analyses. M.Z., M.P., S.V., and M.K. performed laboratory studies under the supervision of M.R.S., J.M., A.H., M.H., and J.Z.; M.Z., A.W.T.N, M.A., C.R., V.C., A.R., L.B., M.O., and S.R.G. performed computational analyses and prepared relevant display items. M.Z., A.W.T.N., A.H., K.G., J.M., M.O., F.A.B., M.K., and J.Z. edited the manuscript, and M.Z., K.G., M.H., M.K., and J.Z. designed the study.

References

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, et al. 2013a. Signatures of mutational processes in human cancer. *Nature* **500:** 415–421. doi:10.1038/nature12477

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013b. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3:** 246–259. doi:10.1016/j.celrep.2012.12.008

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Boot A, Covington KR, Gordenin DA, Bergstrom E, Lopez-Bigas N, et al. 2018. The repertoire of mutational signatures in human cancer. bioRxiv doi:10.1101/322859

Ardin M, Cahais V, Castells X, Bouaoun L, Byrnes G, Herceg Z, Zavadil J, Olivier M. 2016. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics* **17:** 170. doi:10.1186/s12859-016-1011-z

Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, Martincorena I, Petljak M, Alexandrov LB, Gundem G, et al. 2014. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513:** 422–425. doi:10.1038/nature13448

Beland FA, Mellick PW, Olson GR, Mendoza MCB, Marques MM, Doerge DR. 2013. Carcinogenicity of acrylamide in B6C3F_1 mice and F344/N rats from a 2-year drinking water exposure. *Food Chem Toxicol* **51:** 149–159. doi:10.1016/j.fct.2012.09.017

Beland FA, Olson GR, Mendoza MCB, Marques MM, Doerge DR. 2015. Carcinogenicity of glycidamide in B6C3F_1 mice and F344/N rats from a two-year drinking water exposure. *Food Chem Toxicol* **86:** 104–115. doi:10.1016/j.fct.2015.09.017

Besaratinia A, Pfeifer GP. 2003. Weak yet distinct mutagenicity of acrylamide in mammalian cells. *J Natl Cancer Inst* **95:** 889–896. doi:10.1093/jnci/95.12.889

Besaratinia A, Pfeifer GP. 2004. Genotoxicity of acrylamide and glycidamide. *J Natl Cancer Inst* **96:** 1023–1029. doi:10.1093/jnci/djh186

Besaratinia A, Pfeifer GP. 2005. DNA adduction and mutagenic properties of acrylamide. *Mutat Res* **580:** 31–40. doi:10.1016/j.mrgentox.2004.10.011

Brunet JP, Tamayo P, Golub TR, Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci* **101:** 4164–4169. doi:10.1073/pnas.0308531101

Chawanthayatham S, Valentine CC 3rd, Fedeles BI, Fox EJ, Loeb LA, Levine SS, Slocum SL, Wogan GN, Croy RG, Essigmann JM. 2017. Mutational spectra of aflatoxin B_1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc Natl Acad Sci* **114:** E3101–E3109. doi:10.1073/pnas.1700759114

da Costa GG, Churchwell MI, Hamilton LP, Von Tungeln LS, Beland FA, Marques MM, Doerge DR. 2003. DNA adduct formation from

acrylamide via conversion to glycidamide in adult and neonatal mice. *Chem Res Toxicol* **16:** 1328–1337. doi:10.1021/tx034108e

Garbe JC, Bhattacharya S, Merchant B, Bassett E, Swisshelm K, Feiler HS, Wyrobek AJ, Stampfer MR. 2009. Molecular distinctions between stasis and telomere attrition senescence barriers shown by long-term culture of normal human mammary epithelial cells. *Cancer Res* **69:** 7557–7568. doi:10.1158/0008-5472.CAN-09-0270

Ghanayem BI, McDaniel LP, Churchwell MI, Twaddle NC, Snyder R, Fennell TR, Doerge DR. 2005. Role of CYP2E1 in the epoxidation of acrylamide to glycidamide and formation of DNA and hemoglobin adducts. *Toxicol Sci* **88:** 311–318. doi:10.1093/toxsci/kfi307

Hoang ML, Chen CH, Chen PC, Roberts NJ, Dickman KG, Yun BH, Turesky RJ, Pu YS, Vogelstein B, Papadopoulos N, et al. 2016. Aristolochic acid in the etiology of renal cell carcinoma. *Cancer Epidemiol Biomarkers Prev* **25:** 1600–1608. doi:10.1158/1055-9965.EPI-16-0219

Hogervorst JG, Schouten LJ, Konings EJ, Goldbohm RA, Brandt P. 2008. Dietary acrylamide intake and the risk of renal cell, bladder, and prostate cancer. *Am J Clin Nutr* **87:** 1428–1438. doi:10.1093/ajcn/87.5.1428

Hollstein M, Alexandrov LB, Wild CP, Ardin M, Zavadil J. 2017. Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. *Oncogene* **36:** 158–167. doi:10.1038/onc.2016.192

Huang MN, Yu W, Teoh WW, Ardin M, Jusakul A, Ng AWT, Boot A, Abedi-Ardekani B, Villar S, Myint SS, et al. 2017. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res* **27:** 1475–1486. doi:10.1101/gr.220038.116

Huskova H, Ardin M, Weninger A, Vargova K, Barrin S, Villar S, Olivier M, Stopka T, Herceg Z, Hollstein M, et al. 2017. Modeling cancer driver events *in vitro* using barrier bypass-clonal expansion assays and massively parallel sequencing. *Oncogene* **36:** 6041–6048. doi:10.1038/onc.2017.215

International Agency for Research on Cancer (IARC). 1994. *Some industrial chemicals: IARC monographs on the evaluation of carcinogenesis risks to humans*, Vol. 60. World Health Organization Press, Geneva.

Ishii Y, Matsushita K, Kuroda K, Yokoo Y, Kijima A, Takasu S, Kodama Y, Nishikawa A, Umemura T. 2015. Acrylamide induces specific DNA adduct formation and gene mutations in a carcinogenic target site, the mouse lung. *Mutagenesis* **30:** 227–235. doi:10.1093/mutage/geu062

Jelaković B, Castells X, Tomić K, Ardin M, Karanović S, Zavadil J. 2015. Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid. *Int J Cancer* **136:** 2967–2972. doi:10.1002/ijc.29338

Liu Z, Hergenhahn M, Schmeiser HH, Wogan GN, Hong A, Hollstein M. 2004. Human tumor p53 mutations are selected for in mouse embryonic fibroblasts harboring a humanized p53 gene. *Proc Natl Acad Sci* **101:** 2963–2968. doi:10.1073/pnas.0308607101

Manjanatha MG, Guo LW, Shelton SD, Doerge DR. 2015. Acrylamide-induced carcinogenicity in mouse lung involves mutagenicity: *cII* gene mutations in the lung of big blue mice exposed to acrylamide and glycidamide for up to 4 weeks. *Environ Mol Mutagen* **56:** 446–456. doi:10.1002/em.21939

McCreery MQ, Halliwill KD, Chin D, Delrosario R, Hirst G, Vuong P, Jen KY, Hewinson J, Adams DJ, Balmain A. 2015. Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. *Nat Med* **21:** 1514–1520. doi:10.1038/nm.3979

McFadden DG, Papagiannakopoulos T, Taylor-Weiner A, Stewart C, Carter SL, Cibulskis K, Bhutkar A, McKenna A, Dooley A, Vernon A, et al. 2014. Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell* **156:** 1298–1311. doi:10.1016/j.cell.2014.02.031

Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, Raine K, Maddison M, Anderson E, Stratton MR, et al. 2014. *C. elegans* whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res* **24:** 1624–1636. doi:10.1101/gr.175547.114

Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. 2017. Differences between germline and somatic mutation rates in humans and mice. *Nat Commun* **8:** 15183. doi:10.1038/ncomms15183

Mojska H, Gielecińska I, Cendrowski A. 2016. Acrylamide content in cigarette mainstream smoke and estimation of exposure to acrylamide from tobacco smoke in Poland. *Ann Agric Environ Med* **23:** 456–461. doi:10.5604/12321966.1219187

Nassar D, Latil M, Boeckx B, Lambrechts D, Blanpain C. 2015. Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nat Med* **21:** 946–954. doi:10.1038/nm.3878

Ng AWT, Poon SL, Huang MN, Lim JQ, Boot A, Yu W, Suzuki Y, Thangaraju S, Ng CCY, Tan P, et al. 2017. Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. *Sci Transl Med* **9:** eaan6446. doi:10.1126/scitranslmed.aan6446

Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM, Weninger A, Hollstein M, Stratton MR, Phillips DH. 2015. The genome as a record of environmental exposure. *Mutagenesis* **30:** 763–770. doi:10.1093/mutage/gev073

Obón-Santacana M, Freisling H, Peeters PH, Lujan-Barroso L, Ferrari P, Boutron-Ruault MC, Mesrine S, Baglietto L, Turzanski-Fortner R, Katzke VA, et al. 2016a. Acrylamide and glycidamide hemoglobin adduct levels and endometrial cancer risk: a nested case-control study in nonsmoking postmenopausal women from the EPIC cohort. *Int J Cancer* **138:** 1129–1138. doi:10.1002/ijc.29853

Obón-Santacana M, Lujan-Barroso L, Freisling H, Cadeau C, Fagherazzi G, Boutron-Ruault M-C, Kaaks R, Fortner RT, Boeing H, Ramón Quirós J, et al. 2016b. Dietary and lifestyle determinants of acrylamide and glycidamide hemoglobin adducts in non-smoking postmenopausal women from the EPIC cohort. *Eur J Nutr* **56:** 1157–1168. doi:10.1007/s00394-016-1165-5

Obón-Santacana M, Lujan-Barroso L, Travis RC, Freisling H, Ferrari P, Severi G, Baglietto L, Boutron-Ruault MC, Fortner RT, Ose J, et al. 2016c. Acrylamide and glycidamide hemoglobin adducts and epithelial ovarian cancer: a nested case–control study in nonsmoking postmenopausal women from the EPIC cohort. *Cancer Epidemiol Biomarkers Prev* **25:** 127–134. doi:10.1158/1055-9965.EPI-15-0822

Olesen PT, Olsen A, Frandsen H, Frederiksen K, Overvad K, Tjønneland A. 2008. Acrylamide exposure and incidence of breast cancer among postmenopausal women in the Danish Diet, Cancer and Health Study. *Int J Cancer* **122:** 2094–2100. doi:10.1002/ijc.23359

Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallée MP, McKay J, Nedelko T, Muehlbauer KR, Marusawa H, et al. 2014. Modelling mutational landscapes of human cancers *in vitro*. *Sci Rep* **4:** 4482. doi:10.1038/srep04482

Olstørn HBA, Paulsen JE, Alexander J. 2007. Effects of perinatal exposure to acrylamide and glycidamide on intestinal tumorigenesis in *Min*/+ mice and their wild-type litter mates. *Anticancer Res* **27:** 3855–3864.

Pelucchi C, Bosetti C, Galeone C, La Vecchia C. 2015. Dietary acrylamide and cancer risk: an updated meta-analysis. *Int J Cancer* **136:** 2912–2922. doi:10.1002/ijc.29339

Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P, Weng WH, Siew EY, Liu Y, Heng HL, et al. 2013. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci Transl Med* **5:** 197ra101–197ra101. doi:10.1126/scitranslmed.3006086

Randall SK, Eritja R, Kaplan BE, Petruska J, Goodman MF. 1987. Nucleotide insertion kinetics opposite abasic lesions in DNA. *J Biol Chem* **262:** 6864–6870.

Scelo G, Riazalhosseini Y, Greger L, Letourneau L, Gonzàlez-Porta M, Wozniak MB, Bourgey M, Harnden P, Egevad L, Jackson SM, et al. 2014. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun* **5:** 5135. doi:10.1038/ncomms6135

Secrier M, Li X, de Silva N, Eldridge MD, Contino G, Bornschein J, MacRae S, Grehan N, O'Donovan M, Miremadi A, et al. 2016. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat Genet* **48:** 1131–1141. doi:10.1038/ng.3659

Segerbäck D, Calleman CJ, Schroeder JL, Costa LG, Faustman EM. 1995. Formation of *N*-7-(2-carbamoyl-2-hydroxyethyl) guanine in DNA of the mouse and the rat following intraperitoneal administration of [^{14}C] acrylamide. *Carcinogenesis* **16:** 1161–1165. doi:10.1093/carcin/16.5.1161

Severson PL, Vrba L, Stampfer MR, Futscher BW. 2014. Exome-wide mutation profile in benzo[*a*]pyrene-derived post-stasis and immortal human mammary epithelial cells. *Mut Res Genet Toxicol Environ Mutagen* **775–776:** 48–54. doi:10.1016/j.mrgentox.2014.10.011

Smith MT, Guyton KZ, Gibbons CF, Fritz JM, Portier CJ, Rusyn I, DeMarini DM, Caldwell JC, Kavlock RJ, Lambert PF, et al. 2016. Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. *Environ Health Perspect* **124:** 713–721. doi:10.1289/ehp.1509912

Stampfer MR, Bartley JC. 1985. Induction of transformation and continuous cell lines from normal human mammary epithelial cells after exposure to benzo[*a*]pyrene. *Proc Natl Acad Sci* **82:** 2394–2398. doi:10.1073/pnas.82.8.2394

Stampfer MR, Bartley JC. 1988. Human mammary epithelial cells in culture: differentiation and transformation. *Cancer Treat Res* **40:** 1–24. doi:10.1007/978-1-4613-1733-3_1

Straif K, Loomis D, Guyton K, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Mattock H. 2014. Future priorities for the IARC Monographs. *Lancet Oncol* **15:** 683–684. doi:10.1016/S1470-2045(14)70168-8

Sumner SC, Fennell TR, Moore TA, Chanas B, Gonzalez F, Ghanayem BI. 1999. Role of cytochrome P450 2E1 in the metabolism of acrylamide and acrylonitrile in mice. *Chem Res Toxicol* **12:** 1110–1116. doi:10.1021/tx990040k

Takatsuki S, Nemoto S, Sasaki K, Maitani T. 2003. Determination of acrylamide in processed foods by LC/MS using column switching. *Shokuhin Eiseigaku Zasshi* **44:** 89–95. doi:10.3358/shokueishi.44.89

Tareke E, Rydberg P, Karlsson P, Eriksson S, Törnqvist M. 2002. Analysis of acrylamide, a carcinogen formed in heated foodstuffs. *J Agric Food Chem* **50:** 4998–5006. doi:10.1021/jf020302f

Todaro GJ, Green H. 1963. Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. *J Cell Biol* **17:** 299–313. doi:10.1083/jcb.17.2.299

Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. 2018. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol* **19:** 129. doi:10.1186/s13059-018-1509-y

Virk-Baker MK, Nagy TR, Barnes S, Groopman J. 2014. Dietary acrylamide and human cancer: a systematic review of literature. *Nutr Cancer* **66:** 774–790. doi:10.1080/01635581.2014.916323

Von Tungeln LS, Churchwell MI, Doerge DR, Shaddock JG, McGarrity LJ, Heflich RH, da Costa GG, Marques MM, Beland FA. 2009. DNA adduct formation and induction of micronuclei and mutations in B6C3F$_1$/*Tk* mice treated neonatally with acrylamide or glycidamide. *Int J Cancer* **124:** 2006–2015. doi:10.1002/ijc.24165

Von Tungeln LS, Doerge DR, da Costa GG, Marques M M, Witt WM, Koturbash I, Pogribny IP, Beland FA. 2012. Tumorigenicity of acrylamide and its metabolite glycidamide in the neonatal mouse bioassay. *Int J Cancer* **131:** 2008–2015. doi:10.1002/ijc.27493

Westcott PMK, Halliwill KD, To MD, Rashid M, Rust AG, Keane TM, Delrosario R, Jen K-Y, Gurley KE, Kemp CJ, et al. 2014. The mutational landscapes of genetic and chemical models of *Kras*-driven lung cancer. *Nature* **517:** 489–492. doi:10.1038/nature13898

Wilson KM, Bälter K, Adami HO, Grönberg H, Vikström AC, Paulsson B, Törnqvist M, Mucci LA. 2009. Acrylamide exposure measured by food frequency questionnaire and hemoglobin adduct levels and prostate cancer risk in the Cancer of the Prostate in Sweden Study. *Int J Cancer* **124:** 2384–2390. doi:10.1002/ijc.24175

Xie J, Terry KL, Poole EM, Wilson KM, Rosner BA, Willett WC, Vesper HW, Tworoger SS. 2013. Acrylamide hemoglobin adduct levels and ovarian cancer risk: a nested case–control study. *Cancer Epidemiol Biomarkers Prev* **22:** 653–660. doi:10.1158/1055-9965.EPI-12-1387

Zhang W, He H, Zang M, Wu Q, Zhao H, Lu LL, Ma P, Zheng H, Wang N, Zhang Y, et al. 2017. Genetic features of aflatoxin-associated hepatocellular carcinoma. *Gastroenterology* **153:** 249–262.e2. doi:10.1053/j.gastro.2017.03.024

Zhivagui M, Korenjak M, Zavadil J. 2017. Modelling mutation spectra of human carcinogens using experimental systems. *Basic Clin Pharmacol Toxicol* **121:** 16–22. doi:10.1111/bcpt.12690

Received July 31, 2018; accepted in revised form February 1, 2019.

Title

Experimental analysis of exome-scale mutational signature of glycidamide, the reactive metabolite of acrylamide

Authors

Maria Zhivagui[1], Maude Ardin[1], Alvin W. T. Ng[2,3,4], Mona I. Churchwell[5], Manuraj Pandey[1], Stephanie Villar[1], Vincent Cahais[6], Alexis Robitaille[7], Liacine Bouaoun[8], Adriana Heguy[9], Kathryn Guyton[10], Martha R. Stampfer[11], James McKay[12], Monica Hollstein[1,13,14], Magali Olivier[1], Steven G. Rozen[2,3], Frederick A. Beland[5], Michael Korenjak[1] and Jiri Zavadil[1]

Affiliations

[1] Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, Lyon 69008, France

[2] Centre for Computational Biology, Duke-NUS Medical School, Singapore 169857, Singapore

[3] Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, 169857, Singapore

[4] NUS Graduate School for Integrative Sciences and Engineering, 117456, Singapore

[5] Division of Biochemical Toxicology, National Center for Toxicological Research, Jefferson, AR 72079, USA

[6] Epigenetics Group, International Agency for Research on Cancer, Lyon 69008, France

[7] Infections and Cancer Biology Group, International Agency for Research on Cancer, Lyon 69008, France

[8] Environment and Radiation Section, International Agency for Research on Cancer, Lyon 69008, France

[9] Department of Pathology and Genome Technology Center, New York University, Langone Medical Center, New York, NY 10016, USA

[10] IARC Monographs Section, International Agency for Research on Cancer, Lyon 69008, France

[11] Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

[12] Genetic Cancer Susceptibility Group, International Agency for Research on Cancer, Lyon 69008, France

[13] Deutsches Krebsforschungszentrum, 69120 Heidelberg, Germany

[14] Faculty of Medicine and Health, University of Leeds, LIGHT Laboratories, Leeds LS2 9JT, United Kingdom

Keywords: Acrylamide, glycidamide, DNA adducts, massively parallel sequencing, mutational signatures

Correspondence:

ZavadilJ@iarc.fr and/or KorenjakM@iarc.fr

Abstract

Acrylamide, a probable human carcinogen, is ubiquitously present in the human environment, with sources including heated starchy foods, coffee and cigarette smoke. Humans are also exposed to acrylamide occupationally. Acrylamide is genotoxic, inducing gene mutations and chromosomal aberrations in various experimental settings. Covalent haemoglobin adducts were reported in acrylamide-exposed humans and DNA adducts in experimental systems. The carcinogenicity of acrylamide has been attributed to the effects of glycidamide, its reactive and mutagenic metabolite capable of inducing rodent tumors at various anatomical sites. In order to characterize the pre-mutagenic DNA lesions and global mutation spectra induced by acrylamide and glycidamide, we combined DNA-adduct and whole-exome sequencing analyses in an established exposure-clonal immortalization system based on mouse embryonic fibroblasts. Sequencing and computational analysis revealed a unique mutational signature of glycidamide, characterized by predominant T:A>A:T transversions, followed by T:A>C:G and C:G>A:T mutations exhibiting specific trinucleotide contexts and significant transcription strand bias. Computational interrogation of human cancer genome sequencing data indicated that a combination of the glycidamide signature and an experimental benzo[*a*]pyrene signature are nearly equivalent to the COSMIC tobacco-smoking related signature 4 in lung adenocarcinomas and squamous cell carcinomas. We found a more variable relationship between the glycidamide- and benzo[*a*]pyrene-signatures and COSMIC signature 4 in liver cancer, indicating more complex exposures in the liver. Our study demonstrates that the controlled experimental characterization of specific genetic damage associated with glycidamide exposure facilitates identifying corresponding patterns in cancer genome data, thereby underscoring how mutation signature laboratory experimentation contributes to the elucidation of cancer causation.

A 40-word summary

Innovative experimental approaches identify a novel mutational signature of glycidamide, a metabolite of the probable human carcinogen acrylamide. The results may elucidate the cancer risks associated with exposure to acrylamide, commonly found in tobacco smoke, thermally processed foods and beverages.

Introduction

Cancer can be caused by chemicals, complex mixtures, occupational exposures, physical agents, and biological agents, as well as lifestyle factors. Many human carcinogens show a number of characteristics that are shared among carcinogenic agents (1). Different human carcinogens may exhibit a spectrum of these key characteristics, and operate through separate mechanisms to generate patterns of genetic alterations. Recognizable patterns of genetic alterations or mutational signatures characterize carcinogens that are genotoxic. Recent work shows that these DNA sequence changes can be expressed in simple mathematical terms that enable mutational signatures to be extracted from thousands of cancer genome sequencing data sets (2). Several of the over 30 identified mutational signatures have been attributed to specific external exposures or endogenous factors through epidemiological and experimental studies (2). However, about 40% of the current signatures remain of unknown origin, and additional, thus far unrecognized, signatures are likely to be defined in rapidly accumulating cancer genome data. Well-controlled experimental exposure systems can thus help identify the underlying causes of known orphan mutational signatures as well as define new patterns generated by candidate carcinogens (reviewed in (3,4)).

Various diet-related exposures contribute to the human cancer burden. Examples include contaminants in food or alternative medicines, such as aflatoxin B1 (AFB1) or aristolochic acid (AA). The mutagenicity of these compounds is well-documented; AFB1 induces predominantly C:G>A:T base substitutions and AA causes T:A>A:T transversions. The characteristic mutations coupled with information on the preferred sequence contexts in which they are likely to arise allowed unequivocal association of exposure to AFB1 or AA with specific subtypes of hepatobiliary or urological cancers, respectively (5-13).

Among dietary compounds with carcinogenic potential, acrylamide is of special interest due to extensive human exposure. Important sources of exposure to acrylamide include tobacco smoke (14), coffee (15), and a broad spectrum of occupational settings (16). Dietary sources of acrylamide comprise carbohydrate-rich food products that have been subject to heating at high temperatures. This is due to Maillard reactions, which involve reducing sugars and the amino acid asparagine, present in potatoes and cereals (17). There is sufficient evidence that acrylamide is carcinogenic in experimental animals (18,19) and it has been classified as a probable carcinogen (Group 2A) by the International Agency for Research on Cancer in 1994 (16). The association of dietary acrylamide exposure with renal, endometrial and ovarian cancers has been explored in recent epidemiological studies (20,21). However, accurate acrylamide exposure assessment in epidemiological studies based on questionnaires has been difficult, and more direct measures of molecular markers, such as hemoglobin adduct levels, may not yield conclusive findings on past exposures (22-

27). An improved understanding of its mechanism of action using well-controlled experimental systems is critical for understanding the potential carcinogenic risk associated with exposure.

Acrylamide undergoes oxidation by cytochrome P450, producing the reactive metabolite glycidamide that is highly efficient in DNA binding due to its electrophilic epoxide structure (28-30). The *Hras* mutation load in neoplasms of mice exposed to acrylamide or glycidamide was found to be considerably higher in mice treated with glycidamide (31). This finding is corroborated by a considerably higher mutation frequency in the *cII* reporter gene of Big Blue mouse embryonic fibroblasts treated with glycidamide in comparison to acrylamide (32,33). Mutation analysis in different experimental *in vivo* and *in vitro* models using reporter genes showed an increased association of acrylamide and glycidamide exposure with T:A>C:G transitions, as well as T:A>A:T and C:G>G:C transversion mutations (31-36), whereas glycidamide exposure was also characterized by C:G>A:T transversions (33). However, these proposed acrylamide- and glycidamide-specific mutation patterns were based on limited mutation counts in reporter genes and thus do not reflect the complexity of genome-wide distributions and profiles. Based on the limited data available thus far, it is not possible to translate adequately the reported mutation types (T:A>C:G, T:A>A:T, C:G>G:C, C:G>A:T) to global alteration patterns.

The advent of massively parallel sequencing has created the opportunity to study a large number of mutations in a single sample, thus significantly enhancing the power of mutation analysis in experimental models and enabling reliable identification of specific sequence contexts for the induced alterations. Analogously to human cancer genome projects, genome-scale mutational signatures can be extracted from highly controlled carcinogen exposure experiments using mammalian cell and animal models coupled with advanced mathematical approaches (2,3,37,38).

Here we report the systematic assessment of acrylamide and glycidamide mutagenicity based on DNA adduct formation and mutation profile analysis using massively parallel sequencing in a cell model amenable to the analysis of carcinogen-induced mutation patterns and their impact on the resulting cell phenotype (3,37-39). We identify a specific and robust mutational signature attributable to glycidamide, and by computationally interrogating human cancer genome-wide mutation data, we characterize glycidamide signature-positive tumors, thereby highlighting a potential contribution of acrylamide/glycidamide exposure to carcinogenesis in humans.

Materials and methods

Source and authentication of primary cells

Primary Human-p53 knock-in mouse embryonic fibroblasts (Hupki MEFs) were isolated from 13.5-day old *Trp53*$^{tm/HolI}$ mouse embryos from the Central Animal Laboratory of the Deutsches Krebsforschungszentrum, Heidelberg, as described previously (40). The mice had been tested for Specific Pathogen-Free (SPF) status. The derived primary cells were genotyped for the human *TP53* codon 72 polymorphism (Table 1) to authenticate the embryo of origin. Cells from three different embryos (E210, E213 and E214) were used for the exposure experiments (Table 1). All subsequent cell cultures were routinely tested at all stages for the absence of mycoplasma.

Cell culture, exposure and immortalization

The primary MEF cells were expanded in Advanced DMEM supplemented with 15% fetal calf serum, 1% penicillin/streptomycin, 1% pyruvate, 1% glutamine, and 0.1% β-mercapto-ethanol. The cells were then seeded in six-well plates and, at passage 2, exposed for 24 hours to acrylamide (A4058, Sigma), glycidamide (04704, Sigma), or vehicle (PBS). Acrylamide exposure was carried out in the absence or presence of 2% human S9 fraction (Life Technologies) complemented with NADPH (Sigma). Exposed and control primary cells were cultivated until they bypassed senescence and immortalized clonal cell populations could be isolated (41). The human mammary epithelial cell (HMEC) cultures utilized in this study for whole-genome sequencing (WGS) were generated from benzo[*a*]pyrene (B[a]P) exposed HMEC described previously (42,43).

MTT assay for cell metabolic activity and viability

Cells were seeded in 96-well plates and treated as indicated. Cell viability was measured 48 hours after treatment cessation using CellTiter 96® Aqueous One solution Cell Proliferation Assay (Promega). Plates were incubated for 4 hours at 37°C and absorbance was measured at 492 nm using the APOLLO 11 LB913 plate reader. The MTT assay was performed in triplicates for each experimental condition.

γH2Ax Immunofluorescence

Immunofluorescence staining was carried out using an antibody specific for Ser139-phosphorylated H2Ax (γH2Ax) (9718, Cell Signaling Technology). Primary MEFs were seeded on coverslips in 12 well-plates. The cells were incubated in with γH2Ax-antibody (1:500 in 1% BSA) at 4°C overnight. Subsequent incubation with a fluorochrome-conjugated secondary antibody (4412, Cell Signaling Technology) was carried out for 60 minutes at

room temperature. Coverslips were mounted in Vectashield mounting medium with DAPI (Eurobio). Immunofluorescence images were captured using a Nikon Eclipse Ti.

DNA adduct analysis

Glycidamide-DNA adducts (N7-(2-carbamoy-2-hydroxyethyl)-guanine (N7-GA-Gua) and N3-(2-carbamoy-2-hydroxyethyl)-adenine (N3-GA-Ade)) were quantified by liquid chromatography-mass spectrometry (LC-MS/MS) with stable isotope dilution as previously described (44) (see Supplementary Materials and Methods for details). The LC-MS/MS used for quantification consisted of an Acquity UPLC system (Waters) and a Xevo TQ-S triple quadrupole mass spectrometer (Waters). The same MRM transitions as previously described (44) were monitored with a cone voltage of 50V and collision energy of 20eV for each adduct transition and its corresponding labeled isotope transition.

***TP53* genotyping**

Exons 4 to 8 of the knocked-in human *TP53* gene (NC_000017.11) were sequenced using standard protocols. Sanger sequencing of PCR products was performed at Biofidal (Lyon, France). *TP53* primer sequences are listed in Supplementary Materials and Methods. Resulting sequences were analyzed using the CodonCode Aligner software.

Library preparation and whole-exome sequencing (WES)

Library preparation was carried out using the Kapa Hyper Plus library preparation kit (Kapa Biosystems) according the manufacturer's instructions. Exome capture was performed using the SureSelect XT Mouse All Exon Kit (Agilent Technologies). Eighteen exome-captured libraries were sequenced in the paired-end 150 base-pair run mode using the Illumina HiSeq4000 sequencer.

Processing of WES data

Fastq files were analyzed for data amount and quality using FastQC (0.11.3) and were processed with an in-house pipeline for adapter trimming and alignment to the mm10 genome (release GRCm38). These components of the pipeline are publicly available at https://github.com/IARCbioinfo/alignment-nf. The resulting alignment files had a mean depth-of-coverage of 135 and 175 for acrylamide and glycidamide samples, respectively. All alignment files can be accessed from the NCBI Sequence Read Archive (SRA) data portal under the BioProject accession number PRJNA238303. Two somatic variant callers were employed with default parameters in order to detect single base substitutions (SBS) and small insertions/deletions (indels) (MuTect 1.1.6-4 and Strelka 1.015) in exposed clones, using primary cells as normal samples. Each immortalized clone was compared to primary

MEFs from three different embryos (conditions Prim_1, Prim_2, and Prim_3). The overlap of the variant calling outcome with respect to the different primary MEFs showed concordance close to 80% (Suppl. Fig. S1) with MuTect exhibiting more stringent calling performance. Thus, mutation data obtained from the MuTect variant caller were further processed with the MutSpec suite ((45); https://github.com/IARCbioinfo/mutspec). For more details, see Supplementary Materials and Methods and the summary of sequencing metrics (Suppl. Table S1 – not available in the preprint version), the list of identified MuTect SBS variants (Suppl. Table S2 – not available in the preprint version) and indels (Suppl. Table S3 – not available in the preprint version).

Bioinformatics and statistical analyses

The FactoMiner R package (R package version 3.3.2; https://cran.r-project.org/web/packages/FactoMineR) was used to perform the principal component analysis (PCA). To perform the transcription strand bias (SB) analyses, *p*-values were calculated using Pearson's $\chi 2$ test. As multiple comparisons were assessed, the *p*-value was adjusted by applying a false discovery rate (FDR). Statistical analyses were carried out using the stats R package. The SB was considered statistically significant at *p*-value ≤ 0.05. To analyze samples mutation spectra and treatment-specific mutational signatures, filtered mutations were classified into 96 types corresponding to the six possible base substitutions (C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G, T:A>G:C) and the 16 combinations of flanking nucleotides immediately 5' and 3' of the mutated base. Mutation patterns were then deconvoluted into mutational signatures using the non-negative matrix factorization (NMF) algorithm (46,47). The reconstruction error calculation evaluated the accuracy with which the deciphered mutational signatures describe the original mutation spectra of each sample by applying Pearson correlation and cosine similarity.

In order to clean up the profile of the glycidamide mutational signature from the residual signature 17 signal and to increase the stability of NMF decomposition, we supplied the NMF input by adding samples with a high level of signature 17 (over 65% contribution as determined by independent NMF analysis, see Supplementary Materials and Methods).

Cosine similarity analysis was used to evaluate the concordance of the newly identified T:A>A:T-rich mutational signature of glycidamide with the previously reported mutational signatures characterized by a predominant T:A>A:T content. These comprised COSMIC signatures 22 (AA), 25 and 27 (both of unknown etiology(2)), the experimentally derived mutational signature of AA (37,45), 7,12-dimethylbenz[*a*]anthracene (DMBA) (48,49), and urethane (50).

We employed the mutational signature activity (mSigAct) software's sparse signature assignment function (sparse.assign.activity) (13) to assess the presence of the experimental

mutational signatures of glycidamide and benzo[*a*]pyrene in whole-genome somatic mutation data from 38 lung adenocarcinomas, 48 lung squamous carcinomas, and 320 liver cancers from the ICGC Pan-Cancer Analysis of Whole Genomes (PCAWG) study. We excluded 244 hyper-mutated microsatellite unstable and aristolochic acid signature-containing liver tumors as the presence of high numbers of T>A mutations adversely prevented assessment of the possible presence of the glycidamide signature. A set of 11 active COSMIC mutational signatures were identified in the remaining tumor samples (excluding COSMIC signature 4).

We defined a 'pure' experimental C>N benzo[*a*]pyrene signature by WGS (using Illumina HiSeq4000 by Genewiz, NJ, USA) of finite lifespan post-stasis clones derived from primary human mammary epithelial cells (HMEC) treated with B[a]P as previously described (42,43,51). The read alignment to NCBI GRCh38 genome build, variant calling, filtering and annotation were consistent with the MutSpec pipeline described above (45). Proportion matrices of the experimental GA-signature, the GA-signature normalized to the human genome trinucleotide frequency to allow for human PCAWG data screening, and the whole-genome B[a]P signature are available in Suppl. Table S4 (not accessible in the preprint version).

Results

Acrylamide and glycidamide induce cytotoxic and genotoxic responses in Hupki MEFs

Upon exposure of primary Hupki MEFs to a range of concentrations of acrylamide (ACR) (in the absence or presence of the S9 fraction) and its metabolite, glycidamide (GA), we observed a dose-dependent cytotoxic effect on the cells for either compound (Fig. 1A). This analysis informed the selection of two conditions for the ACR exposure to be used in the subsequent exposure/immortalization experiments, 10 mM ACR for 24 hours in the absence of human S9 fraction, and 5 mM ACR for 24 hours in the presence of S9 fraction, which elicited 50% (range 30-70%) decrease in cell viability. The IC50 condition for GA was used for subsequent mutagenesis analysis, corresponding to a 24-hour treatment with 3 mM of the compound. The genotoxic effects of either ACR or GA manifested by a marked increase in γH2Ax staining in the exposed cell populations, in comparison to the mock-treated control cells (Fig. 1B).

Immortalized MEF cells accumulate *TP53* mutations following acrylamide or glycidamide treatment

Primary MEF cultures from three different embryos (Prim_1, Prim_2, and Prim_3) were exposed to ACR or GA using the established conditions and multiple immortalized clones were derived. MEF senescence and immortalization phases were evident from the growth

curves generated for each culture (Suppl. Fig. S2). Subsequently, the clones derived from ACR exposure (ACR clones) and GA exposure (GA clones) and spontaneous immortalization (Spont), were pre-screened for *TP53* mutations by Sanger sequencing, to assess the mutagenic process prior to exome-scale analysis. In the context of ACR treatment, clones obtained from the Prim_2 MEFs that were heterozygous for the polymorphic site in codon 72 showed a loss of heterozygosity involving a loss of the proline allele in the ACR_1 clone whereas the arginine allele was lost in ACR_2, giving rise to a hemizygous clone (Table 1). No *TP53* mutations were observed in any of the three Spont clones, whereas 3 out of 7 ACR clones and 1 of 5 GA clones carried non-synonymous *TP53* mutations (Table 1). The detected mutations indicated specific selection for mutations in the *TP53* gene during cell immortalization and confirmed the clonal nature of MEF immortalization.

Analysis of mutation spectra

Whole-exome sequencing of all spontaneously immortalized and exposed clones and subsequent extraction of acquired variants revealed that the total number of acquired SBS did not differ markedly between the ACR and Spont clones. The Spont clones harbored on average 190 (median = 151, range = 141-277) SBS, whereas the ACR clones had on average 208 (median = 173, range = 151-262) SBS. In contrast, the total number of SBS was considerably increased in the GA clones, with an average of 485 SBS (median = 448, range = 370-592) (Suppl. Table S1 and S2 – not available in the preprint version). This finding suggests markedly stronger mutagenic properties of GA in the MEFs. To estimate the extent of sequencing-related damage in our samples, we determined the GIV score of each sample as described in Materials and Methods and in (52). No detectable damage for any of the mutation types was observed in our dataset (data not shown). The ACR exposed samples exhibited an overall diffuse pattern across the six different SBS types (Suppl. Fig. S3). The Spont clones showed an enrichment of C:G>G:C SBS in the 5'-GCC-3' context, which was also present at varying levels in the exposed cultures. This particular mutation type appears to be related to the culture conditions used for the immortalization assay, as its presence has previously been noted upon spontaneous as well as exposure-driven MEF immortalization (37). No significant transcription strand bias was observed for any of the mutation classes in the Spont or ACR clones (Suppl. Fig. S4). In the five clones derived from the GA-treated primary MEF cultures, we observed an enrichment of acquired T:A>A:T and C:G>A:T transversions and T:A>C:G transitions (Suppl. Fig. S3B), marked by significant transcription strand bias (Suppl. Fig. S4).

PCA performed on the resulting 6-class SBS spectra unambiguously separated the GA clones from the remaining experimental conditions (Fig. 2A). The analysis of indels

(listed in Suppl. Table S3 – not available in the preprint version) showed lower numbers of these alterations in the GA-associated clones compared to the ACR or Spont clones (Fig. 2B). This suggests that a higher accumulation of SBS may selectively promote the senescence bypass and selection of the GA clones, with a decreased functional contribution of indels, while an inverse scenario is plausible in case of the Spont and ACR clones, reminiscent of a previous report based on the Big Blue mouse embryonic fibroblasts and c*II* transgene (53).

Variant allele frequency analysis

Variant allele frequency (VAF) analysis was carried out for GA clones. Overall, a significant proportion of acquired mutations was present at allelic frequencies between 25-75% (Suppl. Fig. S5). Upon grouping of substitutions into bins of high (67-100%), medium (34-66%) and low (0-33%) VAF, the predominant GA-specific mutation types (T:A>A:T, T:A>C:G and C:G>A:T) started manifesting at high VAF, whereas the 5'-NTT-3' alterations, corresponding to the COSMIC signature 17 previously reported to arise in cultured mouse cells including MEFs (38,54,55) showed lower VAF, therefore a later appearance in the cultures (Suppl. Fig. S6). This observation suggests the early effects of the GA exposure and the reproducible contribution of the induced mutations to the senescence bypass and their clonal propagation during the immortalization stage.

Mutational signature analysis

Using NMF, we extracted the mutational signatures from all the MEF clones. Using computed statistics for estimating the number of signatures, three signatures were identified as an optimal number, with signatures A and C enriched in the Spont and ACR clones, and signature B selectively enriched in the GA clones (Fig. 2C,D). Reconstruction of the observed mutation spectra supports the robustness of the signature analysis with strong Pearson's correlation and cosine similarity in GA-derived clones (Fig. 2D). In signature C and also to a lesser extent in signatures A and B, we observed an admixture of a pattern identical to the orphan COSMIC signature 17 (T:A>G:C in a 5'-NTT-3' trinucleotide context), described in various human cancers (most notably esophageal adenocarcinoma), but also seen in aflatoxin B1-driven mouse liver cancers (11), as well as primary MEF-derived clones (37,38). In *in vitro* contexts, this signature has been linked to cell culture conditions and associated oxidative stress (54,55). To refine further the obtained experimental signatures, we developed a signature 'baiting' approach that combined the MEF clones data with signature 17-rich data from esophageal adenocarcinomas from the ICGC ESAD-UK study for new NMF analysis (56). This resulted in considerable reduction (average = 47%, median = 48%) of the signature 17-specific most prominent T>G peaks and a more refined pattern

for signature B, associated primarily with GA treatment (Fig. 3A and Suppl. Fig. S7). This putative GA signature retains the predominant enrichment for the T:A>A:T transversions and T:A>C:G transitions in the 5'-CTG-3' and 5'-CTT-3' trinucleotide contexts, and the C:G>A:T component. Moreover, these mutation types were marked by significant transcription strand bias (Fig. 3B and Suppl. Fig. S4), exhibiting higher accumulation of mutations on the non-transcribed strand consistent with the decreased efficiency of the transcription-coupled nucleotide excision repair due to adduct formation.

DNA adduct analysis

Following metabolic activation, acrylamide induces well-characterized glycidamide DNA adducts at the N7- and N3-positions of guanine and adenine, respectively. LC-MS/MS-based adduct quantification revealed the absence of these adducts in the spontaneously immortalized control samples as well as in MEFs exposed to acrylamide in the absence of S9 fraction (levels below the limit of detection). This suggests the lack of CYP2E1 activity, which is required for the metabolism of acrylamide to glycidamide, in the MEFs. Upon addition of human S9 fraction, N7-GA-Gua levels increased to 11adducts/10^8 nucleotides, suggesting limited metabolic activation of acrylamide due to the presence of enzymatic activity in the S9 fraction (Fig. 3C and Suppl. Fig. S8). Glycidamide-exposed cells exhibited significantly increased DNA adduct levels, with both N7-GA-Gua and N3-GA-Ade observed at very high average levels, 49 000 adducts/10^8 nucleotides and 350 adducts/10^8 nucleotides, respectively, after subtracting the trace amount of contamination from the internal standard (Fig. 3C and Suppl. Fig. S8).

Comparison of the glycidamide signature to known signatures characterized by prominent T:A>A:T profiles

We next performed cosine similarity analysis of the putative GA signature and all known T:A>A:T-rich signatures extracted from primary cancers as well as experimental systems (Fig. 3D and Suppl. Fig. S9). The best match was 84% pattern similarity with COSMIC signature 25 (derived from four Hodgkin lymphoma cell lines) (Fig. 3D). However, unlike the GA signature, COSMIC signature 25 exhibits strand bias for only T:A>A:T mutations and no transcription strand bias for the T:A>C:G mutations. Thus, the mutation patterns and strand bias on all three main mutation types generated by GA treatment (Fig. 3A,B) appear specific and novel.

Glycidamide signature screening in human tumor data from the ICGC PCAWG

The initial mSigAct test performed on PCAWG data from lung and liver tumors indicated a marked presence of the GA signature. This observation was in keeping with the presence of

acrylamide in tobacco smoke and was further corroborated by a cosine similarity of 94% between the adenine (T>N) components of COSMIC signature 4 (tobacco smoking) and the GA signature (Fig. 4A). We thus hypothesized that COSMIC signature 4 reflects co-exposure to B[a]P (generating C>N/guanine mutations with transcription strand bias) and to GA (generating T>N/adenine mutations with transcription strand bias) (Fig. 4A,B). To provide further experimental evidence, we generated a 'pure' B[a]P mutational signature by whole-genome sequencing of cell clones derived from B[a]P-exposed normal human mammary epithelial cells (HMEC). This yielded a robust signature characterized by predominant strand biased guanine (mainly C>A) mutation levels and negligibly mutated adenines (T>N) (Fig. 4A,B). Next, we used mSigAct to interrogate the PCAWG tumor samples for the level of exposure to the experimentally defined GA and B[a]P signatures (alongside other COSMIC mutational signatures) in 48 lung squamous carcinomas, 38 lung adenocarcinomas, and 320 liver cancers. We compared these to estimated levels of exposure to COSMIC signature 4, and found that in the lung cancers, a combination of the GA and B[a]P signatures accounted for very similar numbers of mutations as COSMIC signature 4, thus further supporting the hypothesis that COSMIC signature 4 represents combined and highly correlated exposure to GA and B[a]P (Fig. 4C). Compared to lung cancers, we found more variability in the assignment of mutation numbers to GA and B[a]P versus COSMIC signature 4 in liver cancers (Fig. 4C), which may reflect a decreased relationship between GA and B[a]P exposure due to generally more complex exposure history in the liver. The successful reconstruction of COSMIC signature 4 by the experimental GA- and B[a]P- signatures in the lung and liver human tumors enabled correct assignment of the GA-signature in a subset of 29 lung adenocarcinomas, 46 lung SCC and 26 liver tumors (Fig. 4D). The SBS counts corresponding to GA-mutational signature ranged between 300 up to 43,000 mutations/per sample in lung tumors, and between 190 to 23,000 mutations/per sample in liver tumors (Fig. 4D and Suppl. Table S5 – not available in the preprint version). These findings indicate exposure to glycidamide linked to tobacco smoking – when concomitant with B[a]P-signature, or through diet or occupation – in the absence of B[a]P signature (samples Liver-HCC::SP112224; Liver-HCC::SP49551; Liver-HCC::SP50105; Liver-HCC::SP98861; Liver-HCC::SP50183, see Suppl. Fig. S10 and Suppl. Table S5 – not available in the preprint version).

Discussion

In this study we report the identification of an exome-wide mutational signature for glycidamide, a metabolite of the probable human carcinogen acrylamide. The newly identified signature is based on massively parallel sequencing performed in a well-controlled

experimental carcinogen exposure-clonal immortalization model, revealing characteristic mutagenic effects of glycidamide. The glycidamide mutational signature presented here and the results of statistical assessment of its presence in multiple human tumor types may help clarify the thus-far tenuous association of acrylamide with human cancer.

In concordance with its *in vivo* carcinogenicity in rodents (16,19,31,57), our findings in the established MEF carcinogen exposure and immortalization system suggest that characteristic mutagenic effects may play a role during acrylamide/glycidamide-driven tumor development. In contrast to glycidamide, acrylamide exposure led neither to an increased number of SBS nor did it induce characteristic mutation types in the MEF exposure system. Despite the absence of a mutagenic effect of acrylamide in our experiments, acrylamide and glycidamide exposures induce an almost identical set of tumors in both mice and rats, providing a substantial argument for a glycidamide-mediated tumorigenic effect of acrylamide (19). This is further supported by mechanistic studies showing that lung tissue from mice exposed to acrylamide and glycidamide displays comparable DNA adduct patterns as well as similar mutation frequencies in the *cII* transgene (36). Similar observations had been made in the context of *in vitro* mutagenicity of acrylamide in human and mouse cells, suggesting the key role for epoxide metabolite glycidamide to form pre-mutagenic DNA adducts (33).

As shown by our adduct analysis, acrylamide is not efficiently metabolized by MEFs. This finding is in keeping with the results from previous animal carcinogenicity studies. In fact, glycidamide induces hepatocellular carcinomas in neonatal B6C3F1 mice, whereas administration of acrylamide does not increase the tumor incidence. This has been attributed to the inability of neonatal mice to efficiently metabolize acrylamide (31). Moreover, in contrast to acrylamide treatment, glycidamide induces tumors of the small intestine in a dose-dependent manner upon perinatal exposure (57) and similar observations were made for glycidamide mutagenicity *in vitro* (33). We compensated for the lack of proper acrylamide metabolic activation by the addition of human S9 fraction, and the assessment of DNA adducts indeed suggests acrylamide metabolic activation upon addition of S9. However, the adduct levels are substantially lower compared to glycidamide exposure, which may account for the observed differences in mutagenicity. Interestingly, a consistent minor contribution of the glycidamide mutational signature was detected in the majority of ACR clones, whereas it was absent in the Spont clones. This raises the possibility that partial metabolic activation of acrylamide in the MEF system resulted in low levels of glycidamide. However, a clear mutational signature in the employed experimental setting was achieved only by exposing the cells directly to glycidamide.

Single reporter gene studies had previously linked acrylamide and glycidamide exposure to multiple different mutation types. Thanks to the larger number of mutations

captured by exome sequencing, we were able to attribute to the glycidamide exposure a particular mutational signature characterized by strand-biased C:G>A:T and T:A>A:T transversions, and T:A>C:A transitions towards the non-transcribed strand suggesting a formation of DNA-adducts. The presence of N7-GA-Gua and N3-GA-Ade, two well-characterized glycidamide DNA adducts originating from the metabolic conversion of acrylamide (30,44,53), shows a remarkable relationship between DNA adduct profiles and the putative mutational signature of glycidamide. N3-GA-Ade and N7-GA-Gua are depurinating adducts. They can result in apurinic/apyrimidinic sites, which, during replication, induce the mis-incorporation of deoxyadenine, leading to the observed T:A>A:T and C:G>A:T transversions of the glycidamide signature, respectively. The third mutation type specifically enriched in the glycidamide signature, T:A>C:G transitions, has been ascribed to the N1-GA-Ade adduct, a miscoding adduct and the most commonly identified adenine adduct *in vitro* (35,44,53,58). Levels of the guanine adduct were especially high in the exposed MEF cells, whereas the associated C:G>A:T transversions in the resulting post-senescence clones were less represented. This could reflect differences in DNA repair efficiency concerning individual GA-DNA adduct species, or the fact that the resulting clones are derived from single cells whereas the GA-DNA adducts were measured on average in the bulk primary cell population. A mechanism of negative selection of cells with high N7-GA-Gua adduct burden is also plausible.

We observed consistent presence of COSMIC signature 17 in the data generated from the untreated and treated MEF clones. The etiology of signature 17 remains unknown. While some candidate causal factors have been proposed in esophageal adenocarcinoma and gastric cancers (e.g., inflammatory conditions due to acid reflux, *H. pylori*) (56) and in cultured mouse cell systems (54,55), further studies are required to establish why signature 17 tends to arise *in vitro* in immortalized clones derived from mouse embryonic fibroblasts as observed in our study and also previous work (38).

Genome-scale sequencing of tumor tissues will be needed to verify, *in vivo*, the glycidamide mutational signature identified in this study. The established animal models (18,19) of acrylamide- and glycidamide-mediated tumorigenesis provide a suitable starting point, and it would be interesting to compare mutational signatures derived from these models with the *in vitro* results. The identified glycidamide signature with its extended features of transcription strand bias for the major mutation types differs from the currently known COSMIC signatures (Fig. 3D). In addition, we show that in the cancer genome sequencing data sets from the ICGC PCAWG effort, the putative glycidamide-mutational signature can be identified in a subset of tumors of the lung and liver (sites of possible acrylamide exposure due to tobacco smoking), based on combining experimentally derived signatures with sophisticated computational signature reconstruction approaches (Fig. 4).

The continued interest in understanding the contribution of acrylamide and its electrophilic metabolite glycidamide to cancer development reflects recent accumulation of new mechanistic data on the animal carcinogenicity of the compounds. The possible carcinogenic effects in humans have been recommended for re-evaluation by the Advisory Group to the Monographs Program of the International Agency for Research on Cancer (59). Our findings related to the reconstruction of COSMIC signature 4 using the experimental GA-signature and B[a]P signature, together with the presence of the GA signature in the lung and liver cancer data are relevant given the established high contents of acrylamide in tobacco smoke. Despite the absence of prominent T>N (adenine) mutations in the experimental B[a]P exposure setting, we cannot exclude a possibility that in the human lung cells the adenine residues can be additionally targeted by other tobacco carcinogens such as benzo[*a*]pyrene derivatives or nitrosamines. Importantly, five liver tumor samples identified in this study harbored the GA signature but the major features of signature 4 as represented by the experimental B[a]P signature were absent (Suppl. Fig. S10, Suppl. Table S5 – not available in the preprint version). These tumors are thus of particular interest as they could reflect dietary or occupational exposure to acrylamide.

The presented mutational signature of glycidamide and its potential use for screening of cancer genome sequencing data may provide a basis for relevant assessment of cancer risk through new carefully designed molecular cancer epidemiology studies. Future validation analyses involving e.g. GA-DNA adduct monitoring in non-tumor tissue of cancer patients or in animal exposure models are warranted to provide additional evidence that the predominant T>N mutations in the cancers identified in this study indeed originate from exposure to acrylamide and its reactive metabolite glycidamide.

Acknowledgments

The views expressed in this manuscript do not necessarily represent those of the U.S. Food and Drug Administration. The study was supported by funding obtained from INCa-INSERM (Plan Cancer 2015 grant to J.Z.), NIH/NIEHS (1R03ES025023-01A1 grant to M.O.), and the Singapore National Medical Research Council (NMRC/CIRG/1422/2015 grant to S.G.R.) and the Singapore Ministry of Health via the Duke-NUS Signature Research Programmes to S.G.R.. M.R.S. was supported by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank the NYU Genome Technology Center, funded in part by the NIH/NCI Cancer Center Support Grant P30CA016087, and GENEWIZ, South Plainfield, NJ, USA, for expert assistance with Illumina sequencing.

References

1. Smith, M.T., *et al.* (2016) Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. *Environ Health Perspect*, **124**, 713-21.
2. Alexandrov, L.B., *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415-421.
3. Zhivagui, M., *et al.* (2017) Modelling Mutation Spectra of Human Carcinogens Using Experimental Systems. *Basic Clin Pharmacol Toxicol*, **121 Suppl 3**, 16-22.
4. Hollstein, M., *et al.* (2017) Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. *Oncogene*, **36**, 158-167.
5. Poon, S.L., *et al.* (2013) Genome-Wide Mutational Signatures of Aristolochic Acid and Its Application as a Screening Tool. *Science Translational Medicine*, **5**, 197ra101-197ra101.
6. Meier, B., *et al.* (2014) C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res*, **24**, 1624-36.
7. Scelo, G., *et al.* (2014) Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun*, **5**, 5135.
8. Jelakovic, B., *et al.* (2015) Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid. *Int J Cancer*, **136**, 2967-72.
9. Hoang, M.L., *et al.* (2016) Aristolochic Acid in the Etiology of Renal Cell Carcinoma. *Cancer Epidemiology, Biomarkers & Prevention*, **25**, 1600-1608.
10. Chawanthayatham, S., *et al.* (2017) Mutational spectra of aflatoxin B1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc Natl Acad Sci U S A*, **114**, E3101-E3109.
11. Huang, M.N., *et al.* (2017) Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res*, **27**, 1475-1486.
12. Zhang, W., *et al.* (2017) Genetic Features of Aflatoxin-Associated Hepatocellular Carcinoma. *Gastroenterology*, **153**, 249-262 e2.
13. Ng, A.W.T., *et al.* (2017) Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. *Sci Transl Med*, **9**.
14. Mojska, H., *et al.* (2016) Acrylamide content in cigarette mainstream smoke and estimation of exposure to acrylamide from tobacco smoke in Poland. *Annals of agricultural and environmental medicine: AAEM*, **23**, 456-461.
15. Takatsuki, S., *et al.* (2003) Determination of acrylamide in processed foods by LC/MS using column switching. *Shokuhin Eiseigaku Zasshi. Journal of the Food Hygienic Society of Japan*, **44**, 89-95.
16. IARC Monograph vol. 60 (1994) *Some industrial chemicals. Lyon, 15 - 22 February 1994*, Lyon.
17. Tareke, E., *et al.* (2002) Analysis of Acrylamide, a Carcinogen Formed in Heated Foodstuffs. *Journal of Agricultural and Food Chemistry*, **50**, 4998-5006.
18. Beland, F.A., *et al.* (2013) Carcinogenicity of acrylamide in B6C3F(1) mice and F344/N rats from a 2-year drinking water exposure. *Food and Chemical Toxicology,* **51**, 149-159.
19. Beland, F.A., *et al.* (2015) Carcinogenicity of glycidamide in B6C3F1 mice and F344/N rats from a two-year drinking water exposure. *Food and Chemical Toxicology*, **86**, 104-115.
20. Hogervorst, J.G., *et al.* (2008) Dietary acrylamide intake and the risk of renal cell, bladder, and prostate cancer. *The American Journal of Clinical Nutrition*, **87**, 1428-1438.
21. Virk-Baker, M.K., *et al.* (2014) Dietary Acrylamide and Human Cancer: A Systematic Review of Literature. *Nutrition and Cancer*, **66**, 774-790.

22. Olesen, P.T., *et al.* (2008) Acrylamide exposure and incidence of breast cancer among postmenopausal women in the Danish Diet, Cancer and Health Study. *International Journal of Cancer*, **122**, 2094-2100.
23. Wilson, K.M., *et al.* (2009) Acrylamide exposure measured by food frequency questionnaire and hemoglobin adduct levels and prostate cancer risk in the Cancer of the Prostate in Sweden Study. *International Journal of Cancer*, **124**, 2384-2390.
24. Xie, J., *et al.* (2013) Acrylamide Hemoglobin Adduct Levels and Ovarian Cancer Risk: A Nested Case-Control Study. *Cancer Epidemiology Biomarkers & Prevention*, **22**, 653-660.
25. Obón-Santacana, M., *et al.* (2016) Acrylamide and glycidamide hemoglobin adduct levels and endometrial cancer risk: A nested case-control study in nonsmoking postmenopausal women from the EPIC cohort. *International Journal of Cancer*, **138**, 1129-1138.
26. Obón-Santacana, M., *et al.* (2016) Acrylamide and Glycidamide Hemoglobin Adducts and Epithelial Ovarian Cancer: A Nested Case-Control Study in Nonsmoking Postmenopausal Women from the EPIC Cohort. *Cancer Epidemiology, Biomarkers & Prevention*, **25**, 127-134.
27. Obón-Santacana, M., *et al.* (2016) Dietary and lifestyle determinants of acrylamide and glycidamide hemoglobin adducts in non-smoking postmenopausal women from the EPIC cohort. *European Journal of Nutrition*.
28. Sumner, S.C., *et al.* (1999) Role of cytochrome P450 2E1 in the metabolism of acrylamide and acrylonitrile in mice. *Chemical Research in Toxicology*, **12**, 1110-1116.
29. Ghanayem, B.I., *et al.* (2005) Role of CYP2E1 in the epoxidation of acrylamide to glycidamide and formation of DNA and hemoglobin adducts. *Toxicological Sciences*, **88**, 311-318.
30. Segerbäck, D., *et al.* (1995) Formation of N-7-(2-carbamoyl-2-hydroxyethyl) guanine in DNA of the mouse and the rat following intraperitoneal administration of [14C] acrylamide. *Carcinogenesis*, **16**, 1161-1165.
31. Von Tungeln, L.S., *et al.* (2012) Tumorigenicity of acrylamide and its metabolite glycidamide in the neonatal mouse bioassay. *International Journal of Cancer*, **131**, 2008-2015.
32. Besaratinia, A., *et al.* (2003) Weak yet distinct mutagenicity of acrylamide in mammalian cells. *Journal of the National Cancer Institute*, **95**, 889-896.
33. Besaratinia, A., *et al.* (2004) Genotoxicity of acrylamide and glycidamide. *Journal of the National Cancer Institute*, **96**, 1023-1029.
34. Von Tungeln, L.S., *et al.* (2009) DNA adduct formation and induction of micronuclei and mutations in B6C3F1/Tk mice treated neonatally with acrylamide or glycidamide. *International Journal of Cancer*, **124**, 2006-2015.
35. Ishii, Y., *et al.* (2015) Acrylamide induces specific DNA adduct formation and gene mutations in a carcinogenic target site, the mouse lung. *Mutagenesis*, **30**, 227-235.
36. Manjanatha, M.G., *et al.* (2015) Acrylamide-induced carcinogenicity in mouse lung involves mutagenicity: c*II* gene mutations in the lung of big blue mice exposed to acrylamide and glycidamide for up to 4 weeks. *Environ Mol Mutagen*, **56**, 446-56.
37. Olivier, M., *et al.* (2014) Modelling mutational landscapes of human cancers in vitro. *Scientific Reports*, **4**.
38. Nik-Zainal, S., *et al.* (2015) The genome as a record of environmental exposure. *Mutagenesis*, **30**, 763-70.
39. Huskova, H., *et al.* (2017) Modeling cancer driver events in vitro using barrier bypass-clonal expansion assays and massively parallel sequencing. *Oncogene*, **36**, 6041-6048.
40. Liu, Z., *et al.* (2004) Human tumor p53 mutations are selected for in mouse embryonic fibroblasts harboring a humanized p53 gene. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 2963-2968.

41. Todaro, G.J., *et al.* (1963) Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. *The Journal of Cell Biology*, **17**, 299-313.
42. Severson, P.L., *et al.* (2014) Exome-wide mutation profile in benzo[a]pyrene-derived post-stasis and immortal human mammary epithelial cells. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, **775-776**, 48-54.
43. Stampfer, M.R., *et al.* (1985) Induction of transformation and continuous cell lines from normal human mammary epithelial cells after exposure to benzo[a]pyrene. *Proc Natl Acad Sci U S A*, **82**, 2394-8.
44. Gamboa da Costa, G., *et al.* (2003) DNA adduct formation from acrylamide via conversion to glycidamide in adult and neonatal mice. *Chemical Research in Toxicology*, **16**, 1328-1337.
45. Ardin, M., *et al.* (2016) MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics*, **17**, 170.
46. Brunet, J.-P., *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4164-4169.
47. Alexandrov, Ludmil B., *et al.* (2013) Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, **3**, 246-259.
48. McCreery, M.Q., *et al.* (2015) Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. *Nature Medicine*, **21**, 1514-1520.
49. Nassar, D., *et al.* (2015) Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nature Medicine*, **21**, 946-954.
50. Westcott, P.M.K., *et al.* (2014) The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. *Nature*, **517**, 489-492.
51. Stampfer, M.R., *et al.* (1988) Human mammary epithelial cells in culture: differentiation and transformation. *Cancer Treat Res*, **40**, 1-24.
52. Chen, L., *et al.* (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, **355**, 752-756.
53. Besaratinia, A., *et al.* (2005) DNA adduction and mutagenic properties of acrylamide. *Mutation Research*, **580**, 31-40.
54. Behjati, S., *et al.* (2014) Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, **513**, 422-425.
55. Milholland, B., *et al.* (2017) Differences between germline and somatic mutation rates in humans and mice. *Nat Commun*, **8**, 15183.
56. Secrier, M., *et al.* (2016) Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nature Genetics*, **48**, 1131-1141.
57. Olstørn, H.B.A., *et al.* (2007) Effects of perinatal exposure to acrylamide and glycidamide on intestinal tumorigenesis in Min/+ mice and their wild-type litter mates. *Anticancer Research*, **27**, 3855-3864.
58. Randall, S.K., *et al.* (1987) Nucleotide insertion kinetics opposite abasic lesions in DNA. *Journal of Biological Chemistry*, **262**, 6864-6870.
59. Straif, K., *et al.* (2014) Future priorities for the IARC Monographs. *The Lancet Oncology*, **15**, 683-684.

Figure legends

Figure 1: Acrylamide- and glycidamide-induced cytotoxicity and genotoxicity *in vitro*. (**A**) Cell viability, following 24-hour treatment of primary MEFs with the indicated concentrations of

acrylamide (top panel), in the absence (diamonds) and presence (circles) of human S9 fraction, and glycidamide (bottom panel), as determined by MTT assay. Absorbance was measured 48 hours after treatment cessation and was normalized to untreated cells. The results are expressed as mean percent ±SD of three replicates. (**B**) DNA damage assessment by immunofluorescence with an antibody specific for Ser139-phosphorylated histone H2Ax (ɣH2Ax). Primary MEFs were treated with acrylamide or glycidamide for 24 hours prior to immunofluorescence. Compound concentrations used were based on 20-70% viability reduction in the MTT assay: 10 mM acrylamide, 5 mM acrylamide in the presence of S9 fraction and 3 mM glycidamide. ACR: acrylamide; GA: glycidamide.

Figure 2: Analysis of the mutation patterns derived from exome sequencing data from immortalized Hupki MEF clones. (**A**) Principle component analysis (PCA) of WES data. PCA was computed using as input the mutation count matrix of the clones that immortalized spontaneously (Spont) or were derived from exposure to acrylamide (ACR) or glycidamide (GA). Each sample is plotted considering the value of the first and second principal components (Dim1 and Dim2). The percentage of variance explained by each component is indicated within brackets on each axis. Spont, ACR- and GA-exposed samples are represented by differently colored symbols. (**B**) Representation of small insertions and deletions (indels) counts within the immortalized clones as determined by the Strelka variant caller. (**C**) Mutational signatures identified by non-negative matrix factorization (NMF) in the 15 Hupki MEF-derived clones (sig A, sig B, and sig C). X-axis represents the trinucleotide sequence context. Y-axis represents the frequency distribution of the mutations. The predominant trinucleotide context for T:A > A:T mutations is indicated in sig B (5'-CTG-3'). The trinucleotide contexts for C:G > G:C (5'-GCC-3') and T:A > G:C mutations (5'-NTT-3') are highlighted in sig C. (**D**) Contribution of the identified signatures to each sample (X-axis), assigned either by absolute SBS counts or by proportion (bar graphs). The reconstruction accuracy of the identified mutational signatures in individual samples is shown in the bottom scatter plot (Y-axis value of 1 = 100% accuracy).

Figure 3: (**A**) Refinement of GA signature. The contribution of signature 17 (T:A>G:C in 5'-NTT-3' context), present in all clones, was decreased by performing NMF on Hupki samples pooled with primary tumor samples with high levels of signature 17 (see Methods). (**B**) Transcription strand bias analysis for the six mutation types in GA-exposed clones. For each mutation type, the number of mutations occurring on the transcribed (T) and non-transcribed (N) strand is shown on the Y-axis. *** $p < 10^{-8}$; * $p < 10^{-2}$. (**C**) DNA adducts analysis as determined by LC-MS/MS. Levels of N7-GA-Gua adduct in ACR+S9 and GA treated MEFs and N3-GA-Ade DNA adduct level in GA treated MEFs. The data are presented as the number of adducts in 10^8 nucleotides. $n \geq 2$. (**D**) Cosine similarity matrix comparing the

putative glycidamide mutational signature with other A>T rich mutational signatures from COSMIC (signatures 22, 25, and 27) and from experimental exposure assays using specific carcinogens (7,12-dimethylbenz[a]anthracene (DMBA), urethane, and aristolochic acid (AA)).

Figure 4: GA signature in human primary cancer genome PCAWG data. (**A**) Comparison of COSMIC signature 4 with two experimentally derived signatures (B[a]P_Exp = signature in clones from benzo[*a*]pyrene treated HMEC cells; GA_Exp = signature in clones from glycidamide-treated MEF cells). Cosine similarity between the T>N (adenine) components of signature 4 and GA signature is shown to the right. (**B**) Transcription strand bias analysis for the six mutation types underlying the signatures in panel A). For each mutation type, the number of mutations occurring on the transcribed (T) and non-transcribed (N) strand is shown on the left Y-axis. The significance is expressed as –log10(p-value) indicated on the right Y-axis. *** $p < 10^{-8}$; ** $p < 10^{-4}$; * $p < 10^{-2}$. (**C**) Scatter plots show reconstruction of COSMIC signature 4 using B[a]P- and glycidamide- experimental mutational signatures in lung adenocarcinoma, lung squamous cell carcinoma and hepatocellular carcinoma from the PCAWG data set. (**D**) mSigAct analysis identifies the assignment and the contributions of mutational signatures (including the experimental signature_GA_Exp (red) and signature_B[a]P_Exp (blue)) to the mutation burden of a total of 101 PCAWG lung and liver tumors identified as positive for the GA signature signal.

Table 1: Summary of cell lines, treatment conditions and *TP53*[1] mutation status.

Sample ID	Embryo	Exposure	Conc. (mM)	Exposure duration (hrs)	coding DNA change[2]	genomic DNA change[3]	aa change	Codon 72 (rs1042522)[4]
Prim_1	E210	-	-	-				Pro/Pro
Prim_2	E213	-	-	-				Arg/Pro
Prim_3	E214	-	-	-				Pro/Pro
Spont_1	E213	-	-	-				Arg/Pro
Spont_2	E214	-	-	-				Pro/Pro
Spont_3	E214	-	-	-				Pro/Pro
ACR_S9_1	E213	ACR	5	24				Arg/Pro
ACR_S9_2	E213	ACR	5	24				Arg/Pro
ACR_1	E213	ACR	10	24	c.881delA	g.7577057delT	p.E294fs	Arg/-
ACR_2	E213	ACR	10	24	c.818G>T	g.7577120C>A	p.R273L	Pro/-
ACR_3	E214	ACR	10	24	c.740A>T; c.839G>C	g.7577541T>A; g.7577099C>G	p.N247I; p.R280T	Pro/Pro
ACR_4	E214	ACR	10	24				Pro/Pro
ACR_5	E214	ACR	10	24				Pro/Pro
GA_1	E210	GA	3	24				Pro/Pro
GA_2	E210	GA	3	24				Pro/Pro
GA_3	E210	GA	3	24	c.309-310CC>TA	g.7579377-7579378GG>TA	[p.Y103Y; p.Q104K]	Pro/Pro
GA_4	E214	GA	3	24				Pro/Pro
GA_5	E214	GA	3	24				Pro/Pro

[1] human TP53 gene; [2] NM_000546.4 coding sequence; [3] hg19 genomic coordinates; [4] human polymorphic site (rs1042522)

Prim = Primary cells; Spont = spontaneously immortalized clones; ACR = acrylamide-exposure derived clones; GA = glycidamide-exposure derived clones. Each exposure condition was carried out in two biological replicates (embryos). S9 = human S9 fraction; Pro = proline; Arg = arginine; Arg/- or Pro/- = loss of allele; fs = frameshift; aa = amino acid.

Figure 1

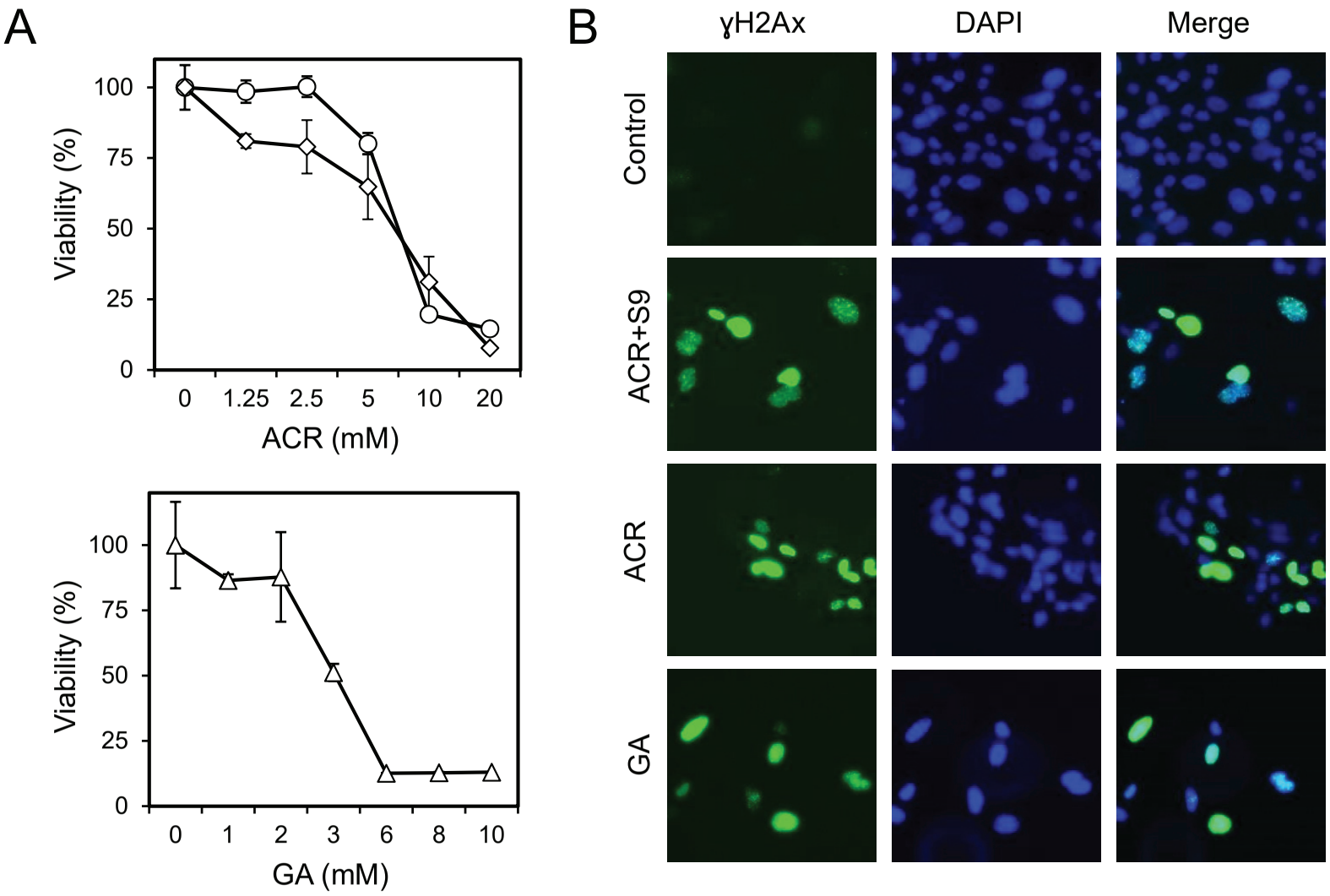

A
Viability (%)
100
75
50
25
0
0
1.25
2.5
5
10
20
ACR (mM)
Viability (%)
100
75
50
25
0
0
1
2
3
6
8
10
GA (mM)
B
γH2Ax
DAPI
Merge
Control
ACR+S9
ACR
GA

Figure 2

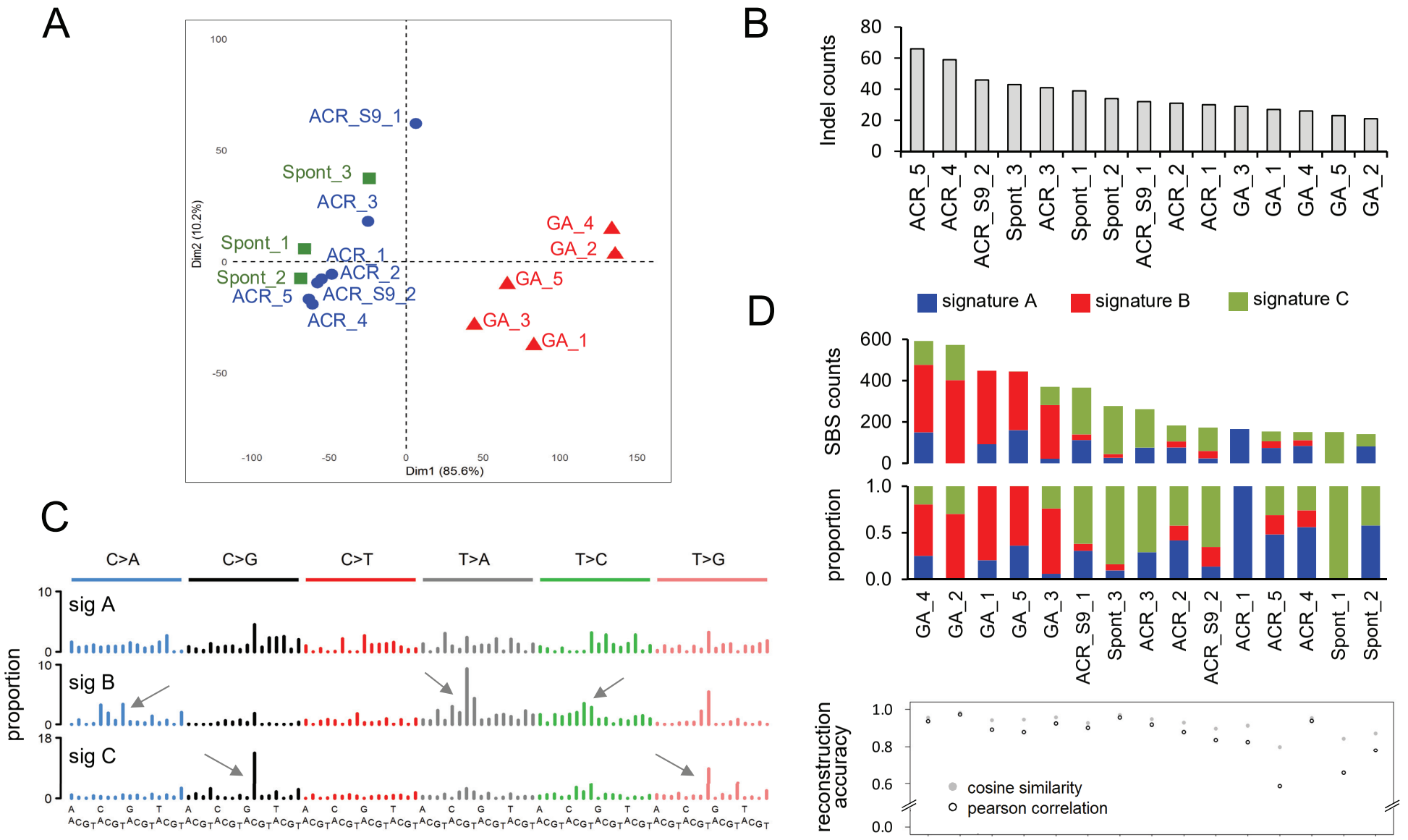
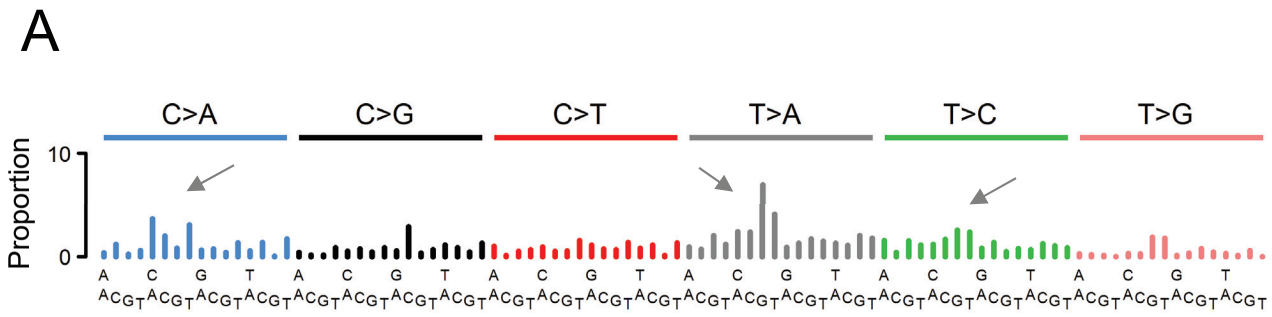
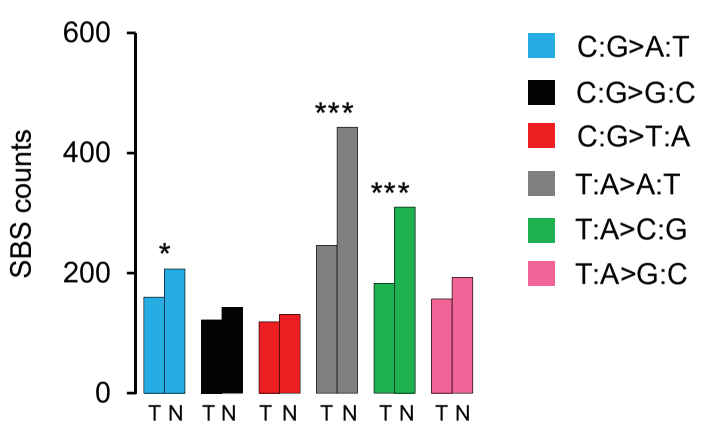
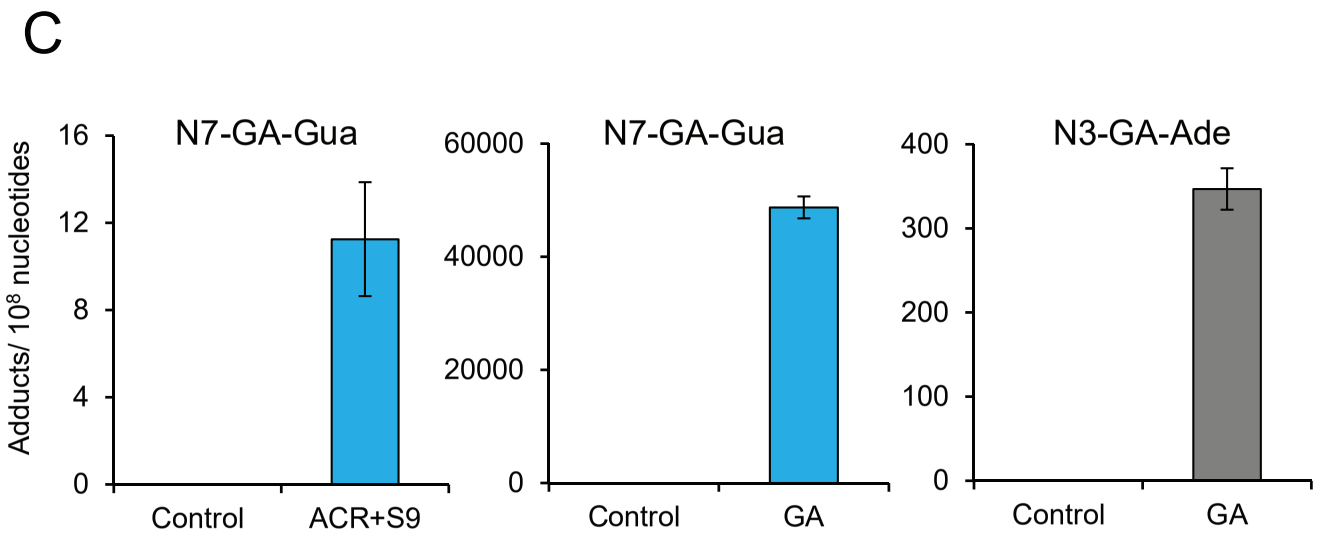

A
ACR_S9_1
Spont_3
ACR_3
GA_4
GA_2
Spont_1
ACR_1
ACR_2
GA_5
Spont_2
ACR_S9_2
ACR_5
ACR_4
GA_3
GA_1
Dim2 (10.2%)
Dim1 (85.6%)
B
Indel counts
ACR_5
ACR_4
ACR_S9_2
Spont_3
ACR_3
Spont_1
Spont_2
ACR_S9_1
ACR_2
ACR_1
GA_3
GA_1
GA_4
GA_5
GA_2
C
C>A
C>G
C>T
T>A
T>C
T>G
sig A
sig B
sig C
proportion
D
signature A
signature B
signature C
SBS counts
proportion
GA_4
GA_2
GA_1
GA_5
GA_3
ACR_S9_1
Spont_3
ACR_3
ACR_2
ACR_S9_2
ACR_1
ACR_5
ACR_4
Spont_1
Spont_2
reconstruction accuracy
cosine similarity
pearson correlation

Figure 3


A
C>A
C>G
C>T
T>A
T>C
T>G
Proportion
10
0


B
SBS counts
600
400
200
0
*
***
***
T N
C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C


C
Adducts/ 10^8 nucleotides
N7-GA-Gua
N7-GA-Gua
N3-GA-Ade
Control
ACR+S9
Control
GA
Control
GA

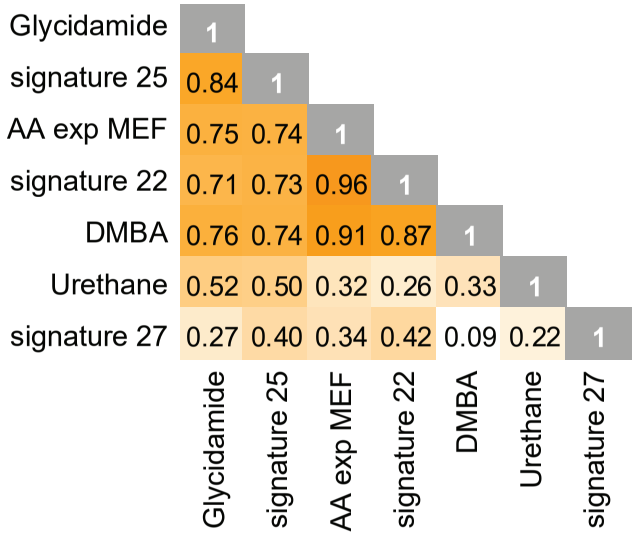

D
Glycidamide
signature 25
AA exp MEF
signature 22
DMBA
Urethane
signature 27
1
0.84 1
0.75 0.74 1
0.71 0.73 0.96 1
0.76 0.74 0.91 0.87 1
0.52 0.50 0.32 0.26 0.33 1
0.27 0.40 0.34 0.42 0.09 0.22 1

Figure 4