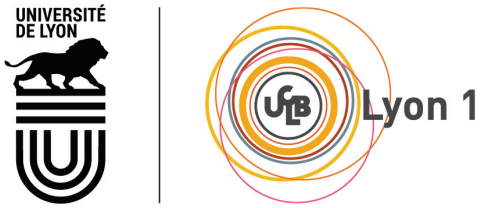
Investigating cancer aetiology through the analysis of somatic mutation signatures

Maude Ardin



N° d'ordre NNT : 2016LYSE1236

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale ED340
Biologie Moléculaire, Intégrative et Cellulaire de Lyon (BMIC)

Spécialité de doctorat : Bioinformatique

Soutenue publiquement le 30/11/2016, par :
Maude ARDIN

Investigating cancer aetiology through the analysis of somatic mutation signatures

---

Devant le jury composé de :

M. DUMONTET Charles	Professeur Centre de Recherche en Cancérologie de Lyon (CRCL)	Président
M. DELEUZE Jean-François	Docteur CEA / Centre National de Genotypage	Rapporteur
M. SICHEL François	Professeur Université de Caen Basse-Normandie	Rapporteur
M. EILS Roland	Professeur Deutsches Krebsforschungszentrum (DKFZ)	Examinateur
M. COX David	Docteur Centre Léon Bérard	Examinateur
M^{me} FERVERS Béatrice	Professeur Centre Léon Bérard	Examinatrice
M. HERCEG Zdenko	Docteur Centre international de Recherche sur le Cancer	Directeur de thèse
M. ZAVADIL Jiri	Docteur Centre international de Recherche sur le Cancer	Co-superviseur
M^{me} OLIVIER Magali	Docteur Centre international de Recherche sur le Cancer	Co-superviseur

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directrice Générale des Services	Mme Dominique MARCHAND

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur G.RODE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. X. PERROT
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y. VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

LABORATORY

Groupe Mécanismes Moléculaire et Biomarqueurs (MMB)
Centre international de Recherche sur le Cancer (CIRC)
150 Cours Albert Thomas
69372 Lyon cedex 08
France

Résumé en français

Analyse des empreintes mutationnelles pour la recherche sur l'étiologie des cancers humains

Les cellules cancéreuses sont caractérisées par des altérations de l'ADN causées par des facteurs exogènes, comme l'exposition à des agents environnementaux tels que le tabac ou les UV, ou par des mécanismes endogènes tels que les erreurs de polymérase lors de la réplication de l'ADN. L'analyse des causes et des conséquences de ces altérations permet de mieux comprendre les facteurs et mécanismes à l'origine du développement d'un cancer. Les technologies de séquençages à haut débit offrent l'opportunité d'étudier la nature précise de ces altérations à l'échelle du génome et permettent de révéler des signatures mutationnelles distinctes et spécifiques de cancérigènes, fournissant ainsi des hypothèses sur l'étiologie des cancers.

L'objectif de ma thèse a consisté à développer des méthodes et des outils bioinformatiques accessibles et conviviaux permettant de faciliter l'analyse et l'interprétation des signatures mutationnelles à partir de données de séquençage à haut débit. L'application de ces outils et méthodes à des séries originales de tumeurs humaines et de systèmes expérimentaux de mutagénèse et carcinogénèse a permis de mieux caractériser la signature mutationnelle de l'acide aristolochique (AA) ainsi que d'autres cancérigènes d'intérêt.

Mots-clés: Bioinformatique, séquençage de nouvelle génération (NGS), signatures mutationnelles, acide aristolochique, Galaxy, cancérigènes

Résumé en anglais

Investigating cancer aetiology through the analysis of somatic mutation signatures

Cellular genomes accumulate alterations following exposures to exogenous factors, like environmental agents such as tobacco smoking or UV, or to endogenous mechanisms such as DNA replication errors. Analysing the causes and consequences of these changes allows better understanding of the mechanisms underlying cancer development and progression. Next-generation sequencing (NGS) technologies provide the opportunity to study the nature of the resulting alterations on a genome-wide scale and started to reveal distinct mutational signatures specific to past carcinogenic exposures providing clues on cancer aetiology.

The aim of my thesis was to develop user-friendly bioinformatic tools and methods for facilitating the analysis and interpretation of carcinogen-specific mutational signatures from NGS data. Applying these tools and methods to human tumours and experimental models of mutagenesis led to improved characterisation of the mutational signature of aristolochic acid (AA), as well as other carcinogens of interest.

Mots-clés: Bioinformatic, next-generation sequencing (NGS), mutational signatures, aristolochic acid, Galaxy, carcinogens

ACKNOWLEDGEMENTS

First of all I would like to thank my PhD advisors Dr Jiri Zavadil and Dr Magali Olivier for trusting me during my Master 2 internship and allowing me to go on and realise this thesis.

Special thanks are addressed to Jiri for his availability, continuous support and insightful discussions and comments concerning the different projects I was working on during this thesis. His competencies, scientific rigour and advices taught me a lot and helped me to move on.

I would like to express my sincere gratitude to Magali for putting up with my keyboard noise while I was coding in Perl for hours, for her patience, motivation and availability for answering my scientific questions. Her guidance and advices have been extremely precious during all the time of my PhD internship and during the painful moment of writing. You have been a tremendous mentor for me, allowing me to grow as a research scientist.

I am also very grateful to Dr Zdenko Herceg for agreeing to be my thesis director and for his encouragement and support during this three years.

I would like to thank my thesis committee members, Professor Charles Dumontet and Dr Philipp Bucher, for all their guidance and precious advice throughout this process.

I also thank my rapporteurs, Professor François Sichel and Dr Jean-François Deleuze, who agreed to evaluate my work, as well as all the members of the jury for their presence in my thesis defence.

I would especially like to thank Vincent Cahais, for all his help and support in the implementation of tools within the Galaxy platform, and Dr Graham Byrnes and Mr Liacine Bouaoun for the help in all the statistical analyses required during this thesis. My thanks also go to all bioinformaticians and many other colleagues at the Agency for their help, assistance and support.

I thank all the members of the MCA section for providing a great and motivating work environment, for their help and all the fun we have had in the last 3 years.

I would like to express my special thanks to all the people who reviewed my thesis for their valuable advice.

I am also very grateful to my Parents for the love and unconditional support they have been providing me since I started university. In particular, I would like to thank my Mum for her constant encouragement during these three years of PhD even if she thought I was simply doing an internship.

My sisters for supporting me in everything and for encouraging me throughout this experience.

I especially thank Aurélien for being supportive during the good and, especially the bad times during my PhD.

Thanks to all my friends for providing the support and friendship I needed.

Thank you all very much for your support without which I could not have survived this invaluable formative experience.

Résumé

Le cancer est une maladie génétique conduisant à une prolifération cellulaire anormale due à l'acquisition d'altérations du génome qui sont acquises au cours de la vie d'un individu et qui diffèrent entre cancers. Ces altérations peuvent être dues soit à des facteurs endogènes comme par exemple des erreurs lors de la réplication de l'ADN, soit à des facteurs externes comme l'exposition à des agents cancérigènes (tabac, UV). L'avènement du séquençage à haut débit de l'ADN a permis d'étudier de façon systématique à l'échelle du génome entier d'une tumeur, la nature précise de ces altérations. De récentes méthodes statistiques d'analyse des données de séquençage à haut débit ont permis de définir des signatures mutationnelles reflétant les facteurs et mécanismes à l'origine de ces altérations au niveau de l'ADN. L'étude des signatures mutationnelles permet donc d'exprimer des hypothèses sur l'étiologie des cancers.

L'identification des mécanismes moléculaires impliqués dans le développement tumoral est essentielle pour la mise en place des stratégies de prévention limitant l'exposition aux agents cancérigènes. Afin de faciliter l'identification de ces signatures mutationnelles, l'utilisation de systèmes expérimentaux simples et robustes permettant de tester des agents cancérigènes d'intérêt revêt toute son importance pour mettre en lumière les mécanismes moléculaires conduisant au développement tumoral.

Néanmoins, l'analyse des signatures mutationnelles à partir des données de séquençage à haut débit requiert des savoir-faire bioinformatiques complexes. En effet, l'analyse des données de séquençage brutes requiert l'utilisation de nombreux programmes et bases de données impliquant la gestion de grands volumes de données hétérogènes.

L'objectif général de ma thèse a été de développer des méthodes et outils bioinformatiques accessibles et conviviaux permettant de faciliter l'analyse et l'interprétation des signatures mutationnelles spécifiques de cancérigènes d'intérêt à partir de données de séquençage de nouvelle génération. L'application de ces méthodes et outils à des séries originales de tumeurs humaines et de systèmes expérimentaux de mutagénèse et carcinogénèse a permis de mieux caractériser la signature mutationnelle de l'acide aristolochique (AA) ainsi que d'autres cancérigènes d'intérêt.

Dans le but de combler un vide dans la disponibilité d'outils permettant l'analyse des signatures mutationnelles facilement accessibles à une large communauté de scientifiques, plusieurs méthodes et outils ont été intégrés dans la plateforme web Galaxy. Cette plateforme permet de développer des pipelines d'analyse reproductibles dans une interface graphique conviviale, rendant ainsi accessibles des analyses bioinformatiques complexes accessibles aux chercheurs ayant des compétences limitées en programmation. La suite d'outils développée s'appelle MutSpec et est dédiée à l'analyse des signatures mutationnelles à partir de données de séquençage de nouvelle génération provenant de tumeurs humaines ou de systèmes expérimentaux. Les outils MutSpec permettent d'apporter des annotations fonctionnelles à une liste de mutations, de filtrer les polymorphismes connus, d'extraire des signatures mutationnelles et de calculer des statistiques décrivant les spécificités de ces signatures et enfin, de comparer ces signatures avec celles déjà connues et publiées. Les outils MutSpec offrent une alternative conviviale pour faciliter l'identification de signatures mutationnelles spécifiques de cancérigènes humains et l'interprétation de leur impact au niveau du développement tumoral par une plus grande communauté de chercheurs.

Les outils MutSpec ont été utilisés pour analyser des données publiques et des données issues du séquençage de l'exome (WES) générées dans le groupe de recherche MMB dans le but de mieux caractériser la signature mutationnelle de l'AA ainsi que d'autres cancérigènes d'intérêt.

L'un de ces projets porte sur l'analyse de données issues d'un modèle *in vitro* de cellules de souris de fibroblastes primaires (MEF) exposées à des agents cancérigènes humains d'intérêt : l'AA, le benzo(a)pyrène (BaP), l'agent alkylant N-methyl-N'-nitro-N-nitrosoguanidine (MNNG) ou les rayons ultraviolets (UVC). Les clones immortels obtenus, porteurs de mutations spécifiques de l'exposition, permettent l'identification de la signature mutationnelle spécifique de l'agent cancérigène utilisé. Ce modèle unique de mutagénèse offre une approche simple et efficace pour étudier les signatures mutationnelles spécifiques de cancérigènes humains suspectés d'être impliqués dans l'étiologie de certains cancers en apportant des preuves moléculaires.

Un autre projet du groupe MMB porte sur la détection de la signature mutationnelle de l'AA dans des populations à risque situées dans la région des Balkans. Dans cette

région, la population est exposée suite à l'ingestion de pain fabriqué à partir de farine contaminée par des graines du genre *Aristolochia* et développe de sévères néphropathies, dites endémiques des Balkans (BEN), caractérisées par une fibrose tubulo-interstitielle se compliquant fréquemment par des tumeurs urothéliales (haut appareil urinaire). Récemment, l'utilisation du séquençage de nouvelle génération a permis de définir une signature spécifique dans des tumeurs urothéliales de patients exposés à l'AA. Cette signature est caractérisée par un enrichissement de mutations de types A:T>T:A localisées sur le brin non-transcrit et enrichie dans le contexte de séquence 5'-CpApG-3'.

Nous avons utilisé les caractéristiques uniques de cette signature mutationnelle pour implémenter une méthode de séquençage de l'exome à très faible couverture, à environ 10x alors qu'un séquençage classique se fait à une couverture de 100x, dans des échantillons de tumeurs urothéliales archivées fixés au formol et inclus en paraffine (FFPE). Notre méthode a permis la détection de la signature dans des échantillons tumoraux à partir d'une faible quantité d'ADN, rendant cette méthode fiable pour la détection d'une exposition à l'AA dans des tissus archivés ainsi que pour l'étude de ces conséquences sur les tumeurs humaines.

Nous avons ensuite étudié l'implication potentielle de l'AA dans le développement d'un autre type de tumeurs, en particulier des tumeurs du rein chez des patients issus de la région BEN. Le séquençage de l'exome de ces patients a révélé la présence de la signature mutationnelle de l'AA, fournissant les premières évidences moléculaires de l'implication de l'AA dans l'étiologie de ce type de cancer dans cette région endémique. Le rôle potentiel de l'AA dans le développement des tumeurs du rein doit donc continuer à être étudié dans les régions où la population a un fort risque d'exposition accidentelle.

Considérant le nombre grandissant de types de tumeurs liées à une exposition à l'AA ainsi que la découverte d'autres signatures mutationnelles caractérisées par des mutations de types A:T>T:A, il nous est apparu important d'évaluer la spécificité de la signature mutationnelle de l'AA. Afin d'étudier toutes les variations des mutations de types A:T>T:A dans les cancers humains, nous avons procédé à une fouille approfondie des bases de données publiques de dépôt de génome du cancer. Cette analyse a permis de révéler que l'enrichissement des mutations de types A:T>T:A par rapport aux autres types de mutations était spécifique à une exposition à l'AA, mais que le contexte de séquence préférentiel,

5'-CpApG-3', était aussi ciblé par d'autres agents cancérigènes qui restent à identifier.

Les outils MutSpec ont rapidement été adoptés par les membres du groupe ainsi que par un nombre croissant de scientifiques du Centre international de Recherche sur le Cancer (CIRC) et d'autres organismes de recherche. Les outils disponibles gratuitement depuis le dépôt de Galaxy (`https://toolshed.g2.bx.psu.edu/`) ont été téléchargés plus d'une trentaine de fois et ont reçu des commentaires positifs des utilisateurs. Les outils MutSpec ont montré toute leur utilité pour révéler les spectres mutationnels spécifiques d'agents cancérigènes provenant de tumeurs humaines ou de données *in vitro* issues du séquençage de nouvelle génération.

List of abbreviations

AA	Aristolochic acid
AAN	Aristolochic acid nephropathy
AFB1	Aflatoxin B1
B[a]P	Benzo[a]pyrene
BAM	Binary alignment map
BEN	Balkan endemic nephropathy
BER	Base-excision repair
CHN	Chinese herb nephropathy
CNV	Copy number variation
COSMIC	Catalogue Of Somatic Mutations In Cancer
CRT	Cyclic reversible termination
DNA	Deoxyribonucleic acid
GATK	Genome Analysis Toolkit
GDC	Genomic Data Commons
HCC	Hepatocellular carcinoma
HPC	High-performance computing
IARC	International Agency for Research on Cancer
ICGC	International Cancer Genome Consortium
Indel	Insertion / Deletion
MEF	Mouse embryonic fibroblasts
MMB	Molecular mechanisms and biomarkers Group

MMR	Mismatch repair
NCI	National Cancer Institute
NER	Nucleotide excision repair
NGS	Next-generation sequencing
NHGRI	National Human Genome Research Institute
NTP	National Toxicology Program
QC	Quality control
RCC	Renal cell carcinoma
RNA	Ribonucleic acid
SBS	Single base substitution
SMRT	Single molecule real-time sequencing
SNA	Single nucleotide addition
SNV	Single nucleotide variant
TCGA	The Cancer Genome Atlas
TCR	Transcription-coupled repair
UTUC	Upper-tract urothelial carcinoma
VCF	Variant calling format
WES	Whole-exome sequencing
WGS	Whole genome sequencing
ZMW	Zero-mode waveguides

Table of Contents

INTRODUCTION

Chapter 1

CANCER GENOMICS

Contents

1.1 The hallmarks of cancer

Human cancer is a genetic disease involving dynamic changes in the genome leading to an abnormal cellular proliferation of normal tissue. Cancer is a multi-stage process in which normal cells progressively acquire additional traits that contribute to the transformation of a normal cell into a cancer cell. These advanced properties had been summarised as the "hallmarks of cancer" [1]. Six major distinctive and complementary changes in the physiology of a cell, in order to become malignant, were described as follows:

- Sustained proliferative signalling:
  Normal tissue controls the production and release of growth-promoting signals regulating the progression through the cell cycle ensuring a homeostasis of cell number

and maintenance of normal tissue architecture and function. Cancer cells deregulate these signals by acquiring the capability to sustain proliferative signalling by producing their own growth factor ligands resulting in autocrine proliferative stimulation. Alternatively, the number of receptor signalling can be increased in the cancer cells surface rendering them hyper-responsive to otherwise-limiting amounts of growth factor ligand. By deregulating these signals they become autonomous units in control of their own fate.

- Evading growth suppressors:
  Cancer cells also have to evade the effect of tumour suppressor genes that negatively regulate cell proliferation. The two canonical tumour suppressor genes: *RB* (retinoblastoma-associated) and *TP53* have crucial importance in regulating cell proliferation. They cooperate in governing the decision of cells to proliferate or activate senescence and apoptotic programs, and are often mutated in cancer.

- Activating invasion and metastasis:
  Cancer cells develop alterations in morphology as well as in the attachment to other cells. One of the best characterised alteration involves the loss of a key cell adhesion molecule, E-cadherin, which forms junctions with adjacent epithelial cells. The increased expression of E-cadherin in tumour cells is frequently observed and provides a strong support as key suppressor for invasion and metastasis.

- Enabling replicative immortality:
  Cancer cells require unlimited replicative potential in order to generate tumours. This capability is in complete contrast with the behaviour of normal cells that are able to pass through only a limited number of successive cell growth and division cycles. Two distinct barriers to proliferation have been associated with these capabilities: senescence, an entrance into a non-proliferative but viable state, and crisis, which involves cell death. Multiple evidences indicate that telomeres, protecting the ends of chromosomes, are centrally involved in the capability for unlimited proliferation [2]. In normal cells, telomeres shorten progressively through cell growth and the length of telomeres indicate how many successive cycles a cell can go through before they are largely eroded and thus lose their protective functions triggering entrance into crisis. The formation of tumour cells has been attributed to their

ability to maintain long telomeres through the division cycles.

- Inducing angiogenesis:
  Tumours require maintenance of nutrients and oxygen as well as the ability to evacuate metabolic wastes like normal tissues. These needs are addressed by the process of angiogenesis that is turn on only transiently in normal cells. However during tumour progression, angiogenesis is activated causing continual growth of new vessels that help sustaining the expansion of neoplastic growth.

- Resisting cell death:
  Tumour cells establish a variety of strategies to limit or circumvent apoptosis that serves as a natural barrier to cancer development. The most common one is the loss of *TP53* tumour suppressor gene, which eliminates the activation of apoptosis circuit. Alternatively, tumours may increase expression of anti-apoptotic regulators or survival signals by down-regulating pro-apoptotic factors.

More recently, four additional emerging characteristics have been proposed [3].

- Avoiding immune destruction:
  Immune surveillance theory proposes that cells and tissues are constantly monitored by the immune system that is responsible for recognising and eliminating the vast majority of cancer cells and nascent tumours. According to this theory, tumours that appear have somehow managed to escape detection by the immune system. In fact, cancer cells evade immune destruction by limiting the extent of immunological killing.

- Tumour-promoting inflammation:
  Inflammation contributes to cancer development by supplying molecules into the tumour micro-environment. The molecules include growth factors that sustain proliferative signalling, oxygen species that are actively mutagenic for nearby cancer cells and proangiogenic factors that facilitate angiogenesis. Inflammation enhances tumorigenesis and progression through the acquisition of hallmark capabilities.

- Genome instability and mutation:
  The acquisition of the hallmarks described above mainly depends on the development of genomic instability in the genome of cancer cells, which generates random

mutations. The ability of the genome maintenance systems to detect and repair defects in the DNA ensures that rates of spontaneous mutations are kept very low during each cell generation. In contrast, cancer cells often increase the rates of mutations through an increased sensitivity to mutagenic agents and/or through a failure in the genomic maintenance machinery. With the advent of next-generation sequencing (NGS) technologies studies of the entire cancer genome have revealed distinct patterns of DNA mutations in different tumour types and helped clarify the causal role of particular mutations in tumour pathogenesis.

- Deregulating cellular energetics:

  The chronic and often uncontrolled proliferation of cancer cells involves major adjustments of cellular energy metabolism in order to support continuous cell growth and proliferation, replacing the metabolic program that operates in normal tissues.

These 10 hallmarks (summarised in Figure 1) reflect the functional capabilities that allow a cancer cell to survive, proliferate and disseminate. These hallmarks are acquired differently and at different times across the different cancer types, and also across individuals.

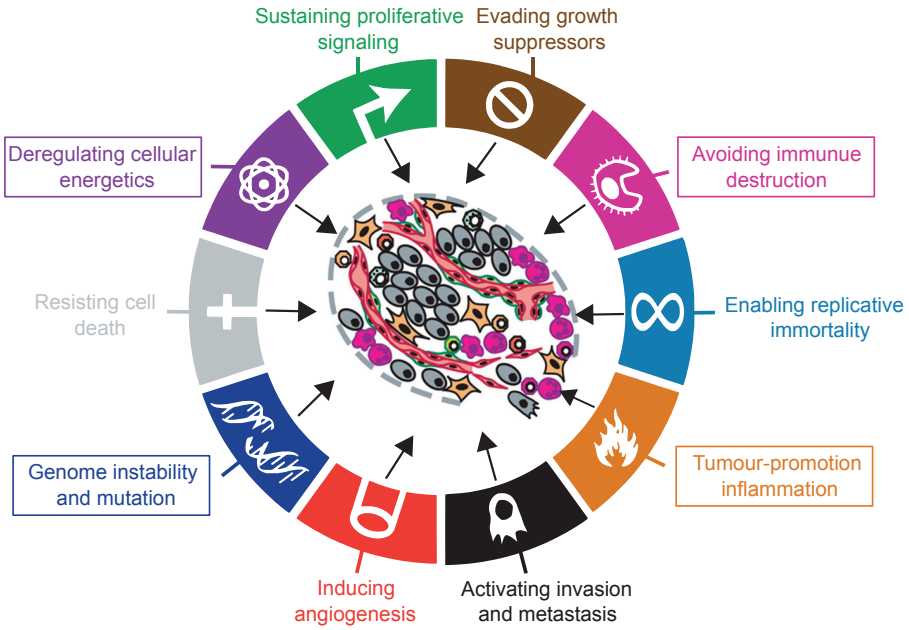


Figure 1: The hallmarks of cancer. Biological capacities a cell must acquire to become a tumour cell. Modified from [1] and [3].

1.2 Genomic alterations

The formation of a cancer is not the consequence of a single DNA mutation but results from the accumulation of multiple DNA alterations over time. These alterations may be due to endogenous mechanisms (inherent genomic instability, deficiency in DNA repair) or to exogenous mechanisms (exposure to carcinogens) [4, 5]. Tobacco smoking, occupational and environmental exposures (e.g. asbestos, ultraviolet light) are examples of such external carcinogens. The effect of both random events and carcinogens may lead to the successive acquisition of the biological capacities of the cancer cell as described in Figure 1.

Mutations include several types of distinct DNA sequence changes reflecting damage on the DNA by various mechanisms. These damages can be visualised graphically on a genome scale, integrating various types of information (see Figure 2).

The most common DNA alteration involves small changes in the genetic code that affect a single base pair. They are called point mutations or single nuleotide variants (SNV) or single base substitutions (SBS).

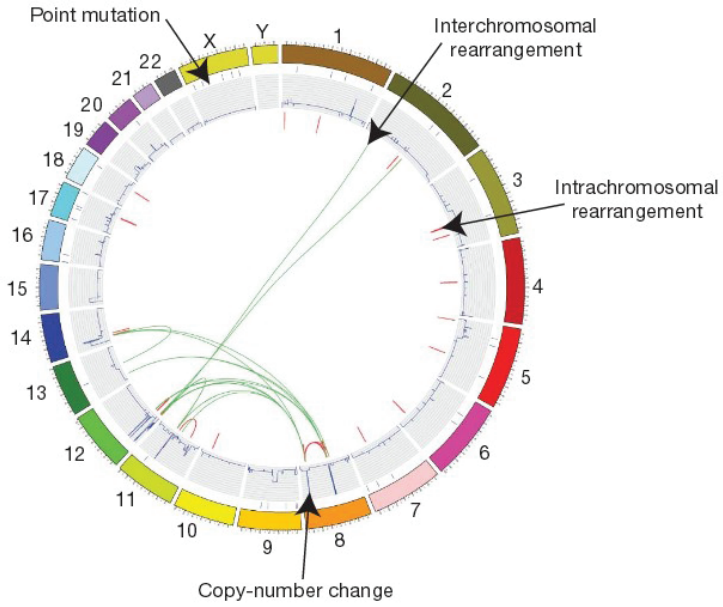


Figure 2: Visualisation of the different types of genomic alterations present in cancer genome. Circos plot are used to depict chromosomes, point mutations, copy-number and rearrangements from the outer circle to the inner circle. Each alteration is represented relative to its position on chromosomes. Adopted from [6].

In principle, there are six possible types of SBS: C:G>A:T, C:G>G:C, C:T>T:A, T:A>A:T, T:A>C:G and T:A>G:C. The single base substitutions are conventionally reported as the pyrimidine of the mutated Watson-Crick base pair. For example a C:G>A:T substitution will be referred to as a C>A substitution, although it covers also G>T. These point mutations can have varying effects on the resulting protein depending where the mutations falls in the coding sequence.

Among SBS a majority are non-synonymous, replacing one amino acid by another and altering the protein function. Point mutations may also introduce stop codon, leading to a truncated non-functional protein, or fall in genes regulatory regions disrupting the transcriptional activity of the gene.

Besides point mutations, alterations on the DNA can be due to insertion or deletion of one or more nucleotides, generally referred to as indels, which may alter the reading frame leading to shortened abnormal or non-functional proteins.

Another form of DNA alteration is copy number variation (CNV) where instead of having the two copies present in a normal diploid genome, there are zero, one or up to several hundred copies. More complex DNA alterations are rearrangements in which DNA has been broken and then rejoined to a DNA segment from elsewhere in the genome. Finally, epigenetic changes which alter chromatin structure and gene expression are present at the DNA sequence level by changes in the methylation status of some cytosine residues. Hyper-methylation is associated with the activation of important genes, for example tumour suppressors in cancer. In contrast, hypo-methylation is implicated in the over-expression of oncogenes.

Cancer is a complex process in which cells acquire DNA alterations that differ from the progenitor fertilised egg. These alterations are called somatic mutations in order to distinguish them from germline mutations that are inherited from parents and transmitted to offspring.

1.3 Somatic mutations drive cancer development

As described previously, cancer development is a complex process in which cells accumulate somatic mutations over a person lifespan and become malignant. All these mutations, however, are not necessarily involved in the development of cancer and it

is likely that some make no contribution at all. To reflect this concept, mutations are classified according to their consequences on cancer development (Figure 3).

The vast majority of somatic mutations found in a cancer genome are "passenger" mutations. They do not confer growth advantage, are not selected for functionality and therefore are not likely to contribute to cancer development. These mutations are present in the cancer genome because somatic mutations without functional consequences often occur during cell division. Passenger mutations are present in all the cells of the final cancer and are believed to reflect the mutational processes operative throughout the lineage of the cancer cells. In contrast, a small minority of these alterations are "driver" mutations occurring in cancer genes (driver genes). They confer growth advantages to the cancer cells and are positively selected during the evolution of the cancer [6].

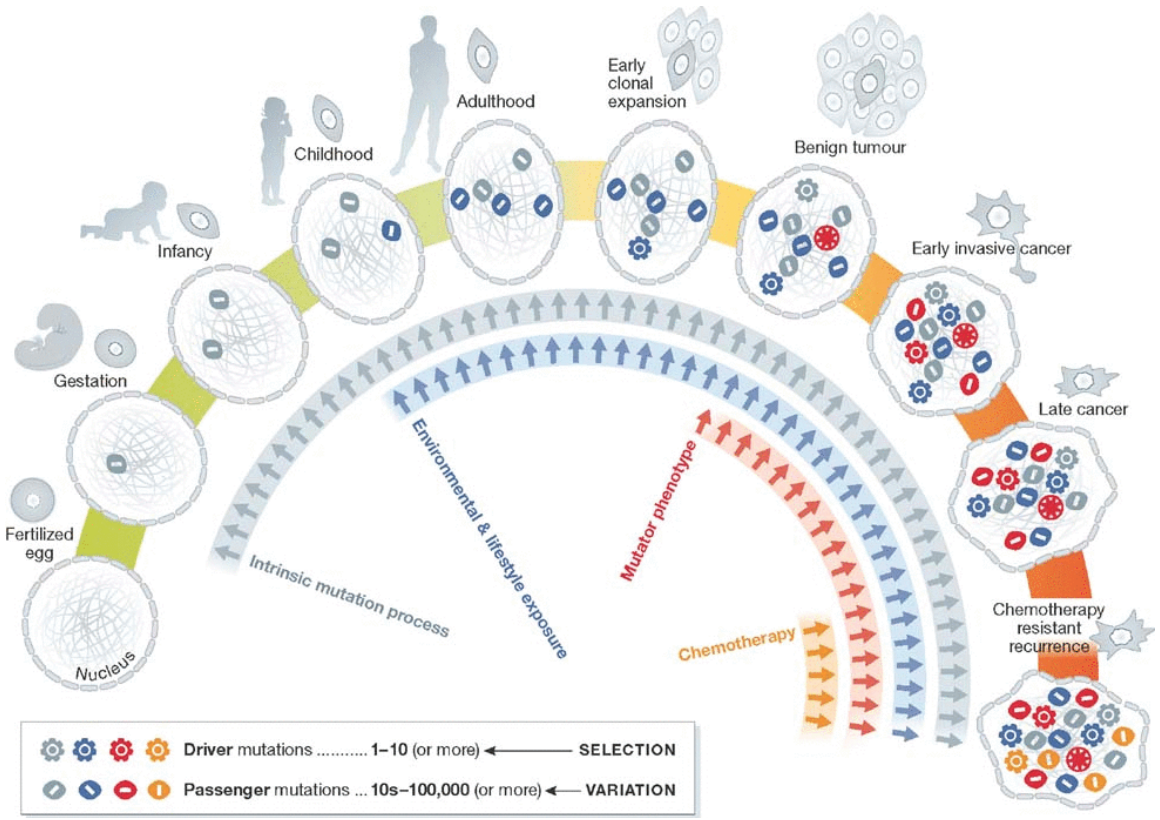


Figure 3: Cellular lineage of cancer cell. Coloured symbols represent the progressive accumulation of somatic mutations between the fertilised egg and a fully malignant cancer cell. Adopted from [7].

Among the driver mutations there is an important subclass of mutations, that confers resistance to chemotherapeutic treatments. These mutations are typically found in recurrences of cancer that initially responded to treatment but became resistant over time.

Cancer genome analysis aims to identify cancer genes carrying driver mutations. In the last decade, over 600 genes have been identified in the genome that can provide a growth advantage to the cells when mutated. Thus fewer than 1% of all human genes appear to contribute to cancer development. Analysing the biological consequences of driver mutations offers a unique opportunity to identify and study unexplored mechanisms of mutagenesis. This knowledge can then be translated into cancer prevention strategies, such as reduction of exposures to newly discovered environmental mutagens.

1.4 The advent of NGS technologies

The rapid evolution of sequencing technologies have revolutionised biomedical sciences by addressing the need to generate inexpensive, reproducible and high-throughput DNA sequence data. Decreasing cost per megabase generated an increase in the number and diversity of genomes sequenced. Over the past decade NGS technologies have continued to evolve and increased the sequencing capacity. NGS platforms provide vast quantities of data, making genomics a Big data science [8]. Big data is defined as any large amount of structure or unstructured data that can be mined for information [9]. The total amount of data produced in genomic is doubling approximately every seven months due to the dramatically decrease of cost of NGS technologies (Figure 4). Compared to the "old" Sanger sequencing technology, the error rates are higher (between 0.1 and 15%) and the read lengths generally shorter (35-700 bp), requiring a careful examination of the results especially for variant discovery and clinical applications.

NGS technologies rely on a combination of template preparation, sequencing, imaging and data analysis [10].

Until 2016, there were two main sequencing approaches: sequencing by ligation and sequencing by synthesis. In both approaches, DNA is clonally amplified on a solid surface resulting in thousands of identical copies of the same DNA fragment in a defined area, ensuring that the signal can be distinguished from the background noise.

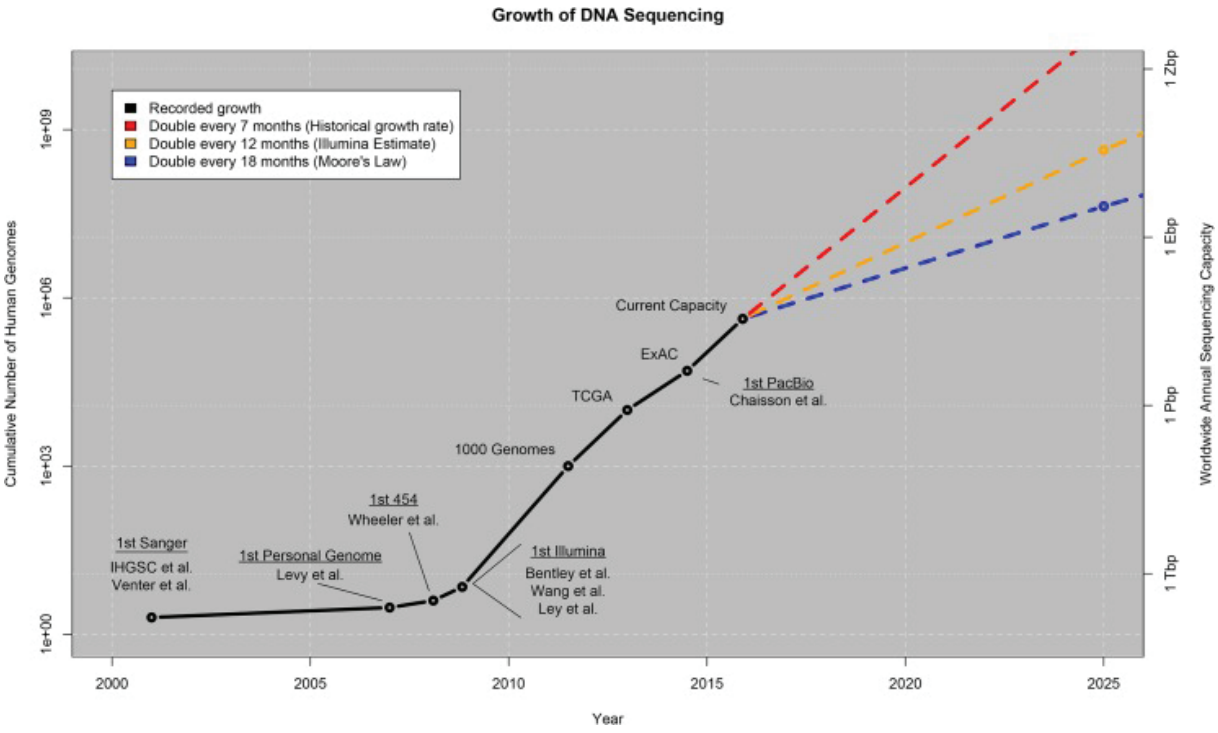


Figure 4: Growth of DNA sequencing. The plot represents the growth of DNA sequencing in total number of genomes sequenced based on literature record (left axis) as well as worldwide annual sequencing capacity (right axis). The ratio is expected to increase sharply as whole genome sequencing (WGS) will take over on whole exome sequencing (WES). Adopted from [11].

Sequencing by ligation was only used by Life Technologies SOLiD platform, discontinued in 2016. This approach involves the hybridisation and ligation of a fluorescent labelled probe which emission spectrum is measured indicating the identity of the base. A new cycle begins after complete removal of the fluorophore and regeneration of the ligation site. Even though this technology allows a greater sensitivity (around 99.99%) by measuring the fluorescence twice for each base, the runtime on the order of several days made that the SOLiD system was only used by a small niche of researchers, this leading to its shut-down in 2016.

Currently, all platforms rely on sequencing by synthesis approach, which implies DNA-polymerase-dependent methods, namely cyclic reversible termination (CRT) and single nucleotide addition (SNA). In these approaches a polymerase is used and the signal, a fluorophore or change in ionic concentration, identifies the incorporation of a nucleotide into an elongation strand.

Illumina systems rely on CRT approach that comprises nucleotide incorporation, fluorescence imaging and cleavage. For its part, SNA relies on the incorporation of a single base into an elongation strand, resulting in a biolumiscence signal (case of Roche 454 pyrosequencing instrument) or the detection of the H^+ ions released at each dNTP incorporation (case of Ion Torrent instrument).

Sequencing platforms vary with differences in throughput, cost, error profile and read structure (Table 1). Illumina platforms dominate the sequencing industry due to their technological maturity as a technology (relying on CRT approach), a high level of cross-platform compatibility and their wide range of applications from targeted to WGS, epigenomics and transcriptomics applications (ChIP-seq and RNA-seq) and DNA methylation sequencing.

Although the accuracy rate of Illumina platforms is high (>99.5%) there is an under-representation of AT-rich and GC-rich regions as well as an increased rate of substitutions errors [12]. Roche 454 and Ion Torrent platforms offer a higher read lengths (up to 700 bp and 400 bp respectively) providing an advantage for applications focusing on repetitive elements or complex DNA rearrangements. Ion Torrent platform offers the fastest run time, between 2 and 7 hours, making it well suited for gene-panel sequencing in clinical applications.

Current technologies have limitations regarding sequencing complex regions of the genome with long repetitive elements, detection of copy number alterations and structural variations. Long-read sequencing technologies deliver reads of several kilobases allowing to span complex or repetitive regions with a single continuous read. There are two main types of long-read technologies: single molecule real-time (SMRT) sequencing and synthetic approaches.

The SMRT approach used by Pacific Biosciences (PacBio) or Oxford Nanopore Technologies, relies on the direct detection of the DNA composition of a native single-stranded DNA molecule.

Table 1: Sequencing platforms comparison. Modified from [10, 13]

Platform	NGS chemistry	Read length (bp)	Gb / run	Runtime	Advantages	Disadvantages
Targeted sequencing						
SOLiD*	SBL	50-75 (SE)	80-320 Gb	6-10 d	Two-bases encoding	Long run time
Ion PGM	SNA	200-400 (SE)	0.6-1 Gb	4-7 h	Run time	No PE
Ion Proton	SNA	Up to 200 (SE)	Up to 10 Gb	2-4 h	Run time	No PE
MiSeq	CRT	75-300 (PE)	3.3-15 Gb	21-56 h	Low errors rate	Long run time
WES / WGS						
Hiseq 2500	CRT	50-100 (PE)	180-500 Gb	2.5-5 d	Most widely used platform	Substitutions errors
Hiseq 4000	CRT	50-150 (PE)	105-750 Gb	1-3.5 d	Most widely used platform	Substitutions errors
Roche 454*	SNA	75-300 (SE/PE)	0.35-0.7 Gb	10-24 h	Read length	Homopolymer errors
Long reads sequencing						
Pacific Biosciences	SMRT	20 Kb	0.5-1 Gb	4 h	Read length	High error rates
Oxford Nanopore	SMRT	Up to 200 Kb	Up to 1.5 Gb	0.5-6 h	USB-based device	High indel errors

bp: base pair; CRT: cyclic reversible termination; d: days; Gb: gigabase pairs; h: hours; Kb: kilobase pairs; PE: paired-end sequencing; SBL: sequencing by ligation; SE: single-end sequencing; SLR: synthetic long reads; SMRT: single-molecule real-time sequencing; SNA: single nucleotide addition

*: Discontinued in 2016

PacBio instrument uses a special flow cell with thousands of zero-mode waveguides (ZMW) wells for measuring a fluorescent signal corresponding to the incorporation of a single nucleotide by a DNA polymerase. Oxford Nanopore Technologies measures an electric signal during the translocation of the DNA through a protein pore. As the DNA translocates, one base at a time, a voltage blockage specific of each dNTP is measured.

Synthetic long reads rely on short-read sequencing to construct *in silico* long reads. There are currently two platforms available for generating synthetic long-reads: the Illumina synthetic long-read sequencing platform and the 10X Genomics emulsion-based system. Synthetic long reads approaches rely on existing sequencing platform affording researchers the ability to simply purchase a kit for long-read sequencing. The emulsion approach allows to use as little as 1 ng of starting material which can be beneficial for situations in which the DNA is precious and available in limited quantity.

Long-read sequencing technologies are ideal for *de novo* genome assembly, for revealing complex genomic structures, DNA methylation and for full-length transcript sequencing.

WGS is becoming one of the most widely used applications in NGS as it offers the most comprehensive view of genomic information and associated biological implications. The recent diversification of NGS platforms has created new opportunities. Sequencing a whole genome sample with Illumina HiSeq 4000 costs around $2,000, with the possibility to multiplex up to 8 genomes (for a coverage of 30x) in a single run performed in 1-3.5 days. In comparison, between 1998 and 2008, it cost 3 billion US dollars and took 10 years to sequence the first human genome.

WES and targeted sequencing are also very useful in research. By reducing the amount of material used, more samples can be sequenced within one sequencing run and the amount of data to analyse is better suited for clinical applications. NGS technologies have opened new opportunities and challenges to biologists, clinicians and bioinformaticians in terms of data acquisition, storage, distribution and analysis [11], and led to revolutionary discoveries in cancer biology and other biological sciences [14].

1.5 Cancer genomic resources

The development of NGS technologies has led to an "explosion" of data generated in the last few years from cancer genomes of a large number of cancer types. These data have

been made publicly accessible, promoting the advancement of scientific discovery. The combination of these data from several omics techniques provides a unique encyclopedia of common cancers [15].

There are three major repositories storing and managing all these data:

- The Catalogue Of Somatic Mutations In Cancer (COSMIC) which is maintained by Sanger Institute, with manually curated scientific literature.
- The Cancer Genome Atlas (TCGA) which stores all US cancer genomic projects.
- The International Cancer Genome Consortium (ICGC) which is a worldwide effort to characterise cancers of importance across the globe.

1.5.1 Catalogue Of Somatic Mutations In Cancer - COSMIC

The Catalogue Of Somatic Mutations In Cancer (COSMIC) accessible at `http://cancer.sanger.ac.uk/cosmic/signatures`, is the largest and most comprehensive public database of somatic mutations with their annotations on human cancer [16]. The database started in 2004 with the aim to store somatic mutation published in the scientific literature and to display the data and metadata related to human cancer [17]. The data present in the database comes from manually curated scientific literature, allowing very precise definitions of disease types and patient details, and also from curated data produced by large scale genome projects such as TCGA and ICGC for reference. The database is updated quarterly and the v78 release (September 2016) contains data from 1,235,846 samples, 28,366 whole genomes, including 4,067,689 coding mutations, 1,271,436 copy number variants and 23,489 publications. The data can be queried on COSMIC website directly using keywords or downloaded by registered users. COSMIC contains huge manually curated and regularly updated datasets making it an invaluable resource for cancer studies.

1.5.2 The Cancer Genome Atlas - TCGA

The Cancer Genome Atlas (TCGA) (`http://cancergenome.nih.gov/`) is managed by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) funded by the US government [18]. TCGA started in 2006 with the aim to

produce a comprehensive genomic characterisation of at least 20 different cancer types. Currently, the project has generated a comprehensive multi-dimensional map of the key genomic changes in 33 types of cancer from more than 11,000 patients. The data have been made publicly available on TCGA data portal (https://tcga-data.nci.nih.gov) and have been widely used by the research community. More than 500 papers have been published by TCGA collaborators and data users. TCGA is coming to a close in early 2017 and a new repository, the Genomic Data Commons (GDC) Data Portal accessible at https://gdc.nci.nih.gov/, will continue to provide an access to TCGA data as well as others NCI cancer genome programs such as the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and the Cancer Genome Characterization Initiative (CGCI). All the datasets available in GDC are harmonised using the latest bioinformatics analyses pipelines allowing a better comparison and reproducibility when using data from different datasets.

1.5.3 International Cancer Genome Consortium - ICGC

The International Cancer Genome Consortium (ICGC) accessible at http://icgc.org/, is a collaborative worldwide effort for coordinating large-scale cancer genomics research projects from 50 different cancer types that have a clinical and societal importance across the globe [19]. The project was launched in 2008 with the aim of elucidating comprehensively the cancer genome changes at the genomic, epigenomic and transcriptomic levels. The data are available for the entire research community on ICGC data portal (https://dcc.icgc.org/) with minimal restrictions in order to accelerate research into the causes and control of cancer. ICGC projects are updated every 5 months and the v22 release (August 2016) comprises data from more than 19,000 cancer donors spanning 70 projects and 21 tumour sites, including 46,429,997 somatic mutations. The catalogue is expected to grow exponentially and will have an immediate relevance in the cancer research community.

1.5.4 cBioportal for Cancer Genomics

The cBioPortal for Cancer Genomics [20, 21] offers an open-access web resource for exploring, visualising, and analysing multidimensional cancer genomics data generated by the large-scale projects described above and others.

The portal is accessible at http://www.cbioportal.org/, and offers an interactive interface for visualising the molecular profiles and clinical characteristics of samples across different tumour types. For example, users can input a list of genes and visualise the frequency of mutations across selected tumour types (Figure 5). The cBioPortal facilitates the analysis of multidimensional cancer genomics data to researchers and clinicians without bioinformatics expertise, accelerating the translation of these rich data sets into biological insights and clinical applications. Currently, data from more than 20,000 tumour samples from 126 cancer studies are available on the portal.

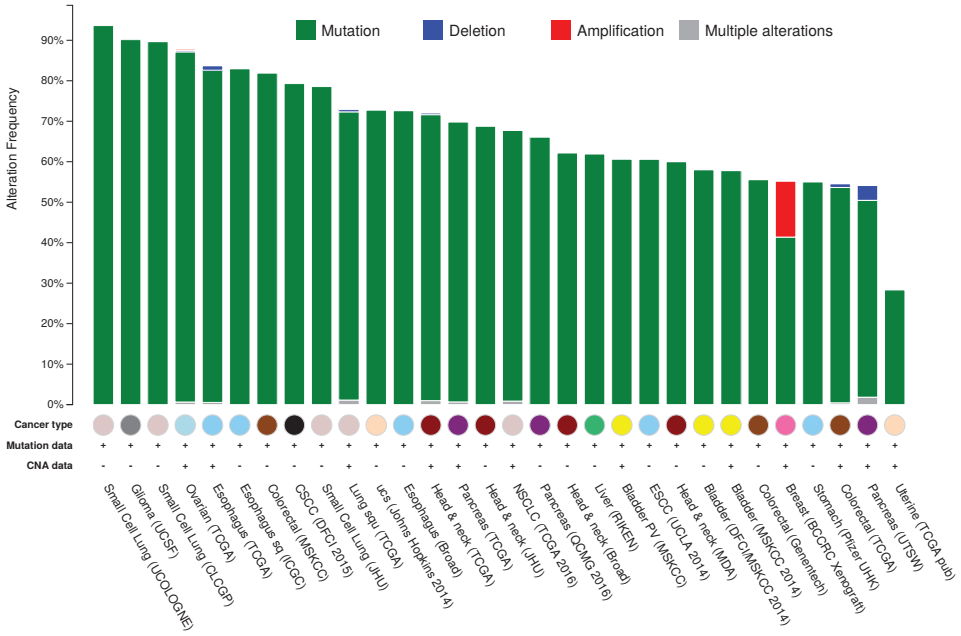


Figure 5: Frequency of *TP53* alterations in 30 cancers types.

Chapter 2

ENVIRONMENT AND CANCER

Contents

2.1 Environmental carcinogenesis

Cancer is influenced by complex series of interactions between genetic and environmental factors that cause damages on the genome. Environmental factors are non-genetic factors that contribute to cancer risk. They are broadly defined as compounds that humans are exposed to through lifestyle factors (such as tobacco smoking, diet, physical activity), radiation, infectious agents (viruses) and exposures to various other carcinogens. Endogenous processes such as oxidative agents, DNA replication errors, mutations in DNA polymerases or DNA repair genes, are also important in carcinogenesis.

Several pathways of repair mechanisms continuously monitor the genome for repairing DNA damages. Base-excision repair (BER) removes small base lesions that could cause the insertion of incorrect base during replication. Transcription-coupled repair (TCR), a particular mechanism of nucleotide excision repair (NER) is responsible for removing adducts covalently bound to DNA. Mismatch repair (MMR) system recognises and repairs misincorporated bases during DNA replication. The good efficiency of DNA repair

pathways limits the number of DNA lesions present at the time of replication and thus limits the number of potential mutations to arise [22].

The existence of polymorphisms in genes involved in DNA repair pathways, resulting in less efficiency to clear the damages, has been associated with an elevated risk in developing some cancers [23]. More is to be learned about the interactions between genetic make-up and individual susceptibility and responses to carcinogenic compounds as it is unclear to what extent many polymorphisms contribute to cancer development.

It has been estimated that exposure to environmental chemical carcinogens may contribute to more than 90% to the causation of cancers and only a small proportion (10%) can be explained by genetics alone [24]. These contributions are summarised in Figure 6.

Exposure to environmental carcinogens is, in theory, avoidable. However, there remains a great challenge for molecular epidemiology to document gene-environment interactions in order to develop new screening tests that can be used for cancer prevention strategies, which could lead to a reduction in cancer mortality worldwide [25].

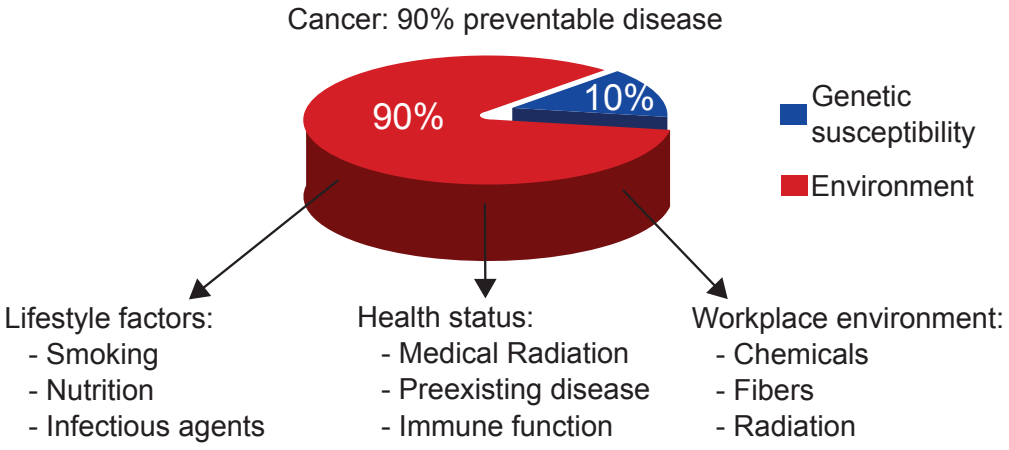


Figure 6: Environmental factors in human cancers.

Molecular epidemiology approaches use biomarkers to study risk factors in populations in order to provide clues on the interactions between exposure and susceptibility factors to determine cancer risk [26].

Many carcinogens leave characteristic fingerprints on the DNA when they bind covalently on it, forming DNA adducts. When these adducts escape repair machinery, they

can cause mutations, thus increasing the probability to develop a cancer. DNA adducts provide a molecular link between specific environmental exposures and cancer risk.

The use of animal models for studying the interactions between genotype and environmental exposures provided the first insights into the causes and mechanisms that influence cancer development [27]. In such experiments, mice or rats were exposed to test agents in order to assess the carcinogenicity of the tested compounds. Besides providing knowledge on the causes and mechanisms that lead to cancer development, animal models also served to identify chemicals potentially carcinogenic to humans.

The U.S. National Toxicology Program (NTP), an inter-agency animal testing programme overssen by NIH/NIEHS, started in 1978 with the mission to identify and evaluate potential chemical agents of public health concern [28]. Researchers were aimed to provide scientific evidence of the direct or indirect implication of chemicals in human diseases, for helping to decrease or to eliminate human exposure to those chemicals. To date, the programme conducted and reported extensive rodent toxicology and carcinogenesis studies for over 600 compounds that were linked to human cancer through epidemiological studies. The programme demonstrated the invaluable resources that are animal models for testing potential human carcinogens and for providing evidence of carcinogenicity that can be used by regulatory organism.

Interactions between health risks and environmental and lifestyle exposures is essential for building preventive health policy and taking health decisions. Since 1971, the International Agency for Research on Cancer (IARC) Monographs Programme reviews the scientific literature in order to evaluate and classify environmental agents that can increase the risk of human cancers [29]. These evaluations are done by a group of independent scientific experts, called "Working Group", that critically review scientific literature for evidence on the carcinogenicity of the studied agent in humans [30].

The agents are evaluated regarding evidence of carcinogenicity in humans using epidemiological studies, evidence of carcinogenicity in animals based on experimental studies and evidence on mechanisms based on molecular epidemiology and biological studies. For more than four decades, the IARC Monographs Programme has been evaluating more than 900 agents classified into five categories [31] that are summarised in Table 2.

The evaluations provided by the IARC Monographs Programme represent unique eval-

uations and classifications of agents and exposures implicated in human cancers, widely used by governments and national health agencies for implementing public health policies to prevent potential exposure to carcinogens.

Table 2: Classification of IARC Monographs Programme (volumes 1-116).

Category	Scientific evidence	Agents
Group 1: Carcinogenic to humans	Sufficient evidence in humans	118
Group 2A: Probably carcinogenic to humans	Limited evidence in humans and animal models	80
Group 2B: Possibly carcinogenic to humans	Limited evidence in humans and less than sufficient in animal models	282
Group 3: not classifiable as to its carcinogenicity to humans	Inadequate evidence in humans and animal models	502
Group 4: probably not carcinogenic to humans	Lack of evidence in humans and animals	1

2.2 Pattern of somatic mutations and environmental exposures

Mutational spectra in tumours provide valuable molecular evidence of the importance of environmental carcinogenesis. Indeed they reveal the activities of both endogenous and exogenous mechanisms reflecting the history of the evolutionary processes that underlie cancer development and can provide new insights into cancer aetiology. Being able to analyse and understand these spectrums at a single gene or a genome-wide scale offers a unique opportunity to identify and study unknown mechanisms of mutagenesis [32].

2.2.1 Single gene approach

Before the advent of NGS, the mutational records were explored through the analysis of single commonly mutated genes, notably the *TP53* tumour suppressor gene [33, 34, 35]. The coding region of *TP53* has been sequenced to identify somatic mutations in over 100,000 tumour samples [36]. Usually no more than one substitution was found per

cancer sample, resulting in the need to compile mutations across multiple samples from the same cancer type. The IARC TP53 Database, accessible at http://p53.iarc.fr/, compiles from the literature all the available information on *TP53* variations related to cancer. Studies investigating tumorigenesis in the skin provided the first evidence that environmental factors determine somatic mutation patterns [37]. Exposure to ultraviolet light, the leading cause of skin cancer, generates C:G>T:A transitions predominantly occurring in dipyrimidines sequence contexts. Next, tobacco smoking has also been documented to leave specific patterns characterised by accumulation of C:G>A:T mutations, the predominant alterations observed in lung tumours [38, 39].

These studies demonstrated the first comprehensive evidence in clinical samples that exposure to mutagenic carcinogens leave specific molecular fingerprints on the genomes of cancer cells [33, 40].

2.2.2 Genome-wide approach

Examination of *TP53* mutations offered insights for strong mutagenic activities that generate the majority of mutations in samples from a particular cancer type and was extremely useful for establishing link between environmental carcinogens and cancer [33, 39]. However, for a cancer type, the mutational spectra are not uniquely induced by a single carcinogen, and multiple mutational processes typically operate within a tumour. The final catalogue of observed mutations is a composite mixture of patterns originating from different mutagenic processes (Figure 7).

Combining mutations from multiple samples from the same cancer type may indicate strong mutational processes, but provides little information about the quantity and individual characteristics of other mutational activities [41].

Taking advantage of the development of high-throughput NGS technologies that allow rapid and cost-effective examination of genome-wide sequencing data from thousands of cancer patients combined with advanced mathematical approaches enable to decipher the different mutational processes operative within a tumour type [42].

This approach allowed to identify 30 distinct mutational signatures in 40 cancer types [43]. Some of these signatures have been linked to endogenous mechanisms (spontaneous deamination of 5-methylcytosine, aberrant activity of APOBEC family deaminases) and others are suspected to be caused by exogenous carcinogens.

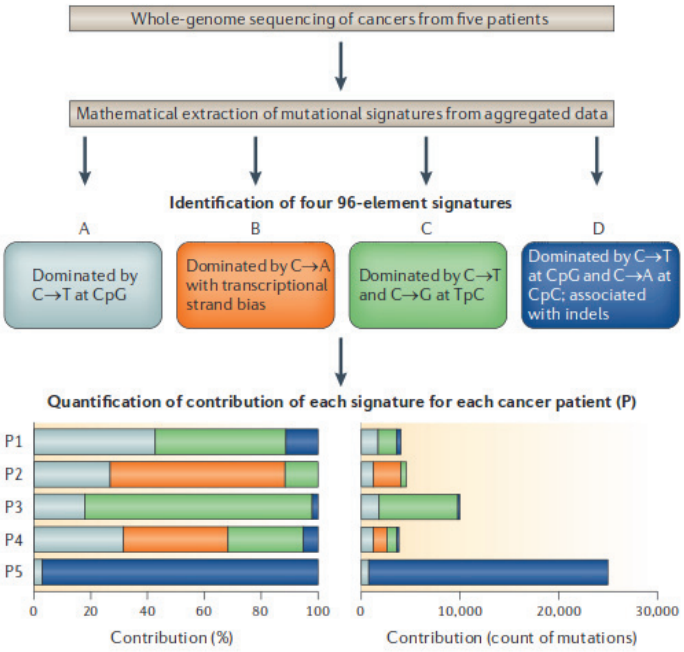


Figure 7: Extraction of mutational signatures operative in tumour genomes. Simulated example to illustrate the mutational processes operative in a set of cancer genomes. Four mutational signatures were extracted and the number of mutations caused by each signature in each of the genomes was estimated. Adopted from [44]

2.2.3 Examples of carcinogen-specific mutational signatures

Three well-documented examples of association between an environmental factor and a specific mutational spectrum observed from single-gene and genome-wide studies are discussed below and summarised in Table 3.

Aflatoxin B1 (AFB1), the most toxic type of aflatoxins, a metabolite produced by moulds growing in grains is an important carcinogen increasing the risk of hepatocellular carcinoma (HCC) in sub-Saharan Africa and East Asia regions. Analysis of *TP53* mutations first revealed C:G>A:T transversions as a predominant mutation type clustering at codon 249 in HCC tumours from these regions [45, 46]. Recently, the mutational signature extracted from WES of 242 HCC tumours from African or Asian migrants confirmed the enrichment of C:G>A:T transversions at a exome-wide scale in 11 patients exposed to AFB1. Additionally, the analysis also revealed a preferred sequence context, 5'-GpCpC-3', and a strong transcriptional strand bias [47].

Cigarette smoke is known to contain a multitude of toxic and carcinogenic compounds. A majority of lung cancers arises in former or current smokers patients. The mutational signature derived from hundreds of cancer patients is characterised by an increased number of mutations and more particularly C:G>A:T transversions located on the non-transcribed strand which are present only in patients with known tobacco smoking history [48, 42]. The preferred sequence contexts of those C:G>T:A transversions are not fully established, perhaps reflecting the chemical complexity of tobacco smoke. This signature is similar to the one obtained from the main suspected compound, benzo[a]pyrene (B[a]P), based on *in vitro* assays [49, 50] but other compounds, which remain to be identified contribute to this as well.

Ultraviolet (UV) light intense exposure has been implicated in the development of melanoma cancers. Analysis of the cancer genome of skin tumours revealed a preponderance of singe-base C:G>T:A substitutions and CC:GG>TT:AA dinucleotide substitutions both occurring at dipyrimidine sites. The effect is so important that CC:GG>TT:AA double substitutions can constitute up to 25% of the total mutation burden in skin cancers related to UV exposures [42]. This specific mutational signature can be used as a clear evidence of UV-related DNA damage.

Table 3: Mutational spectrum of human carcinogens.

Exposure	Carcinogen compound	Mutation type	Sequence context
Dietary contaminants	Aflatoxins	C:G>A:T transcriptional strand bias	5'-GpCpC-3'
Tobacco smoke	Benzo[a]pyrene	C:G>A:T transcriptional strand bias	5'-C/TpCpC-3'
Sunlight	UV	C:G>T:A CC:GG>TT:AA some transcriptional strand bias	5'-C/TpCpN-3'

2.3 Molecular epidemiology of aristolochic acid

Aristolochic acids (AA) are natural compounds found in plants of the genus *Aristolochia* that have been used for centuries in traditional Chinese remedies for treating various health problems such as weight loss, menstrual symptoms, and rheumatism [51]. *Aristolochia* family contains more than 500 species present all around the world.

In 1992, more than 100 Belgium women developed chronic tubulointerstitial renal diseases associated with upper-tract urothelial carcinoma (UTUC), an investigation of this group of patients revealed that they all followed a slimming regimen that included *Aristolochia fangchi*, a herbaceous plant used in traditional Chinese medicine [52]. This syndrome was classified as Chinese herb nephropathy (CHN) but exhibited clinical and pathological similarities with another disease known since the 50's as Balkan Endemic Nephropathy (BEN) [53].

BEN affects specific villages along the Danube river (Croatia, Bosnia-Herzegovina, Serbia, Bulgaria, and Romania) [54]. A hypothesis formulated to explain the similarities between CHN and BEN was that the population living near the Danube river is exposed through contamination of wheat seeds used to prepare home-made bread by *Aristolochia clematitis* [55] a species of the *Aristolochia* family.

Aristolochia plants contain two major forms of AAs: AAI and AAII [56]. They both form aristolactam adducts with deoxyadenosine AL-dA and deoxyguanosine AL-dG (Figure 8) [57] and are found in the proximal tubules of the renal cortex [58].

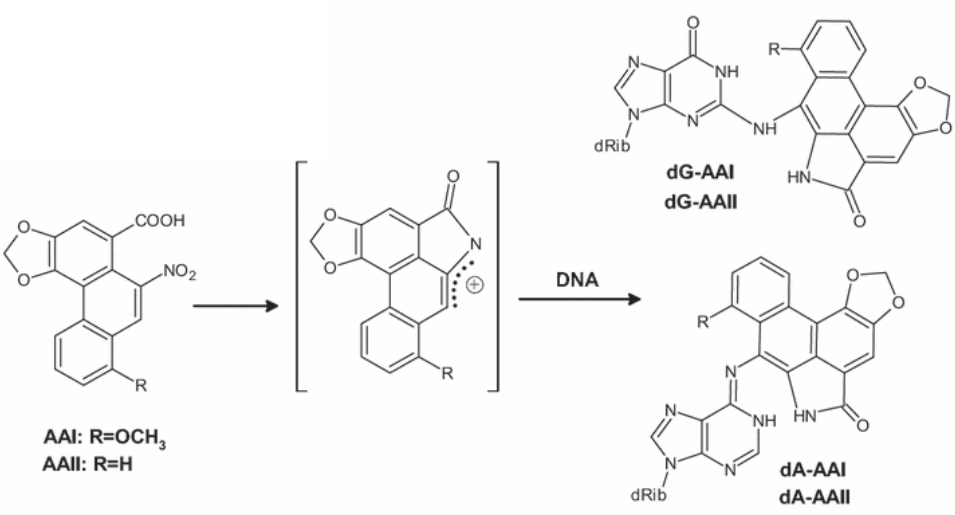


Figure 8: Aristolochic acid I and II adducts formation. Adopted from [59]

The presence of adducts in the proximal tubules of the renal cortex of exposed patients measured by ^{32}P-postlabelling or quantitative mass spectroscopy was proven to be a robust biomarker of exposure to AA [58]. AL-dG adducts are effectively repaired by DNA repair mechanisms while dA-AL-I adducts may persist in target tissues for more than 20 years after the exposure has stopped.

The link between BEN and AAs was confirmed when DNA-adduct analysis of the kidney cortex of exposed patients detected high levels of aristolactam (AL)-DNA adducts [60]. The terms CHN and BEN were replaced by aristolochic acid nephropathy (AAN) for covering both clinical conditions linked to exposure to AA [61].

As a consequence, in 2001 all preparations containing *Aristolochia* plants were banned in Europe and North America and in 2003 in Asian regions [62]. In 2012, IARC classified AA as a Group 1 human carcinogen [63]. However the species nomenclature is complex and the substitution of one plant by another, is common making the regulation of the use complicated.

The accumulation of AL-DNA adducts in target tissues followed by DNA synthesis increases the average number of mutations and generates predominantly A:T>T:A transversions, described first in human *TP53* knock-in mouse cells treated with AA [64]. This specific pattern of mutations in *TP53* was confirmed in human UTUC tumours where the dietary exposure to AA was established [54]. In cancers not associated with AA exposure, such A:T>T:A transversions are infrequent and are found in only 5.3% of all human cancers [65].

Two studies took advantage of NGS technologies to examine the genome-wide mutational spectrum in AA-associated UTUC of Taiwanese patients where the exposure was documented. They characterised a specific mutational signature with a high prevalence of A>T transversions within the 5'-Pyr-A-Pur-3' sequence context (enriched for 5'-CpApG-3' context) preferentially located on the non-transcribed DNA strand [66, 67]. This signature was also found in UTUC of AAN patients originating from the BEN region, where the exposure was confirmed by the presence of AL-DNA adducts in the renal cortex of those patients [68].

Recently, using high-throughput NGS technology, an AA-like mutational signature was found in several tumour types. A deep sequencing study of clear cell renal cell carcinomas (ccRCC) across Europe described this signature in a subset of Romanian cases [69] that

was later confirmed by the presence of dA-AL-I adducts in the kidney cortex of those patients [70]. Other studies also revealed the implication of AA in the aetiology of renal cancers from patients originating from the BEN region [71] and in ccRCC of Taiwanese patients [72].

Using mutational signature analysis, AA exposure was suspected to be a contributing factor causing hepatobiliary cancers in Chinese patients [73] and in Japanese and American Asian populations [74] (Figure 9). Furthermore, an AA-like mutational signature was also found in bladder cancers from Taiwanese patients [75]. DNA adduct analysis in those patients with the AA-like signature is required to confirm the source of exposure.

Figure 9: AA-like mutational spectrum and strand bias. The top panel represents the canonical AA mutational spectrum found in urothelial tumours [66, 67]. The middle and bottom panels represent an AA-like mutational spectrum found in 16 intrahepatic cholangiocarcinoma (IHCC) [73] and in 26 liver hepatocellular carcinoma (LIHC) tumours studied by other groups [74]. The bar graphs on the right show strand bias ratios (NT: non-transcribed strand, T: transcribed strand). Asterisks indicate X^2 test P-values significance (**: P<10e-70; P=0 for the canonical spectrum).

These studies suggest that AA exposure may be implicated in a wide range of different cancer types and millions of people are potentially at risk of AA exposure and consequence developing AA-related cancers [76]. Continuing to improve our knowledge about AA exposure would provide opportunities for prevention and early detection through systematic screening program of patients with known or suspected AA exposure in order to manage a potentially devastating global disease.

Chapter 3

BIOINFORMATICS

Contents

3.1 The concepts of bioinformatics

The term "bioinformatics" was used for the first time by Hesper and Hogeweg in 1970 and was referring to the study of informatic processes in biotic systems [77]. In the early 2000, bioinformatics really became popular with the completion of the human genome project. Nowadays, the term refers to an interdisciplinary field dedicated to the development and use of computational methods and tools for understanding biological research questions. Bioinformatics is present in many biological fields including genomics, evolutionary biology, protein structure prediction and data mining.

There is no full-consensus definition within the scientific community of what being a bioinformatician means (Figure 10). For some people, the term refers to any scientists performing data analyses using a computer, including biostatistics and biomathematics.

For others, a bioinformatician is a scientist who develops and uses computer programs to solve biological problems or to design and maintain databases for interpreting the biology, i.e. in summary: a person who understands both biology and information technology. While computational biologists primarily use computer tools to solve biological problems, they may be knowledgeable in biology but less so in programming languages [78].

On a lighter note, a bioinformatician possessing deep knowledge in all programming, statistics and biology would be seen as a kind of superman.

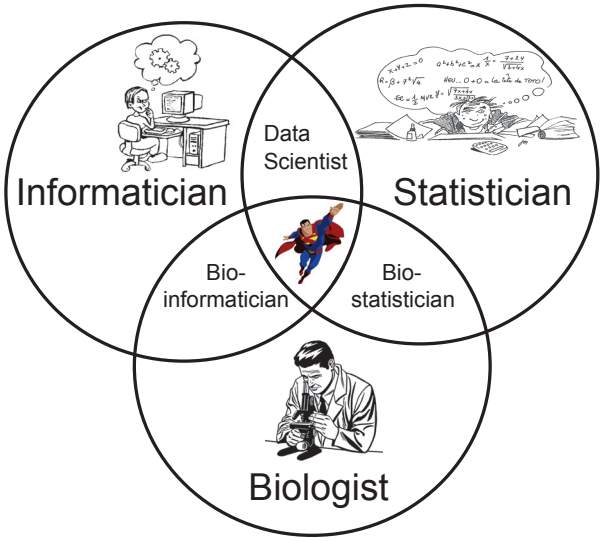


Figure 10: Chart describing the overlap between biology and computer science. Modified from [79].

3.2 Next-generation sequencing analysis pipelines

One biological field in which bioinformatics is ubiquitously applied is genomics, mainly due to the breakthroughs in understanding genome biology in the last two decades, and more recently thanks to the advances in DNA and RNA sequencing technologies that generate volumes of data which analysis relies heavily on bioinformatics tools. In DNA sequencing, WES using NGS technologies has been widely used these last years because it focused on alterations in the coding sequence. Compared to WGS, WES produces more manageable amounts of data and is cheaper. However, WGS is gradually replacing WES because the sequencing cost continues to decrease while the cost of the exome capture reagents (synthesised oligonucleotides) remains stable. Furthermore WGS offers a broader and more uniform coverage of the genome, allowing for a more reliable detection of the copy-number as well as the structural rearrangements.

The complete NGS data analysis workflow for WES and WGS is complex. It includes several steps that rely on a variety of different programs and databases available either commercially (such as Illumina tools) or as open-source softwares allowing more flexibility and customised configurations. While some steps of the analysis in the workflow are well-established with gold standard software, other steps are more challenging and rely on tools still under development. The first steps of analysis are becoming more standard. For example, leaders in the field such as the Broad Institute of MIT and Harvard provide "best practice" guidelines for performing analyses from the raw reads generated by the sequencing machine to obtain a list of highly-confident variants. Broad Institute has developed the Genome Analysis Toolkit (GATK) [80], a suite of tools dedicated to the analysis of high-throughput sequencing data. More downstream tools such as for variant calling are more specific of each application or study design. Each step of a typical NGS workflow is summarised in Figure 11 and detailed further below.

Quality control of raw data Genome-wide sequencing relies on the mechanical or enzymatic break-up of genomic DNA into a library of small fragments that are then sequenced on a given platform depending on the type of analysis (genome, exome or targeted-sequencing). The sequencing machine produces information on a complex set of short DNA fragments, called sequencing reads. These constitute primary raw data output. The first step is to assess the quality control (QC) of the raw sequence data (FASTQ file) coming from the sequencing machine before starting the alignment to the reference genome.

One commonly used QC tool is FastQC, developed by the Babraham Institute (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). The tool provides a quick overview of the number of bases sequenced, the read length distribution, the Phred score (base quality score) distribution along the reads through tables and graphs, and it also detects over-represented sequences indicating primer or adaptor contamination. FastQC summary indicates whether problems were encountered during the sequencing before continuing the analysis, and if pre-processing steps such as base trimming or read filtering are necessary.

The base quality of the reads tends to deteriorate toward the ends of reads requiring to trim these low quality bases for improving the downstream steps of analysis.

Tools such as FASTX-Toolkit (`http://hannonlab.cshl.edu/fastx_toolkit/`), Cutadapt [81] or Trimmomatic [82] can be used for this task.

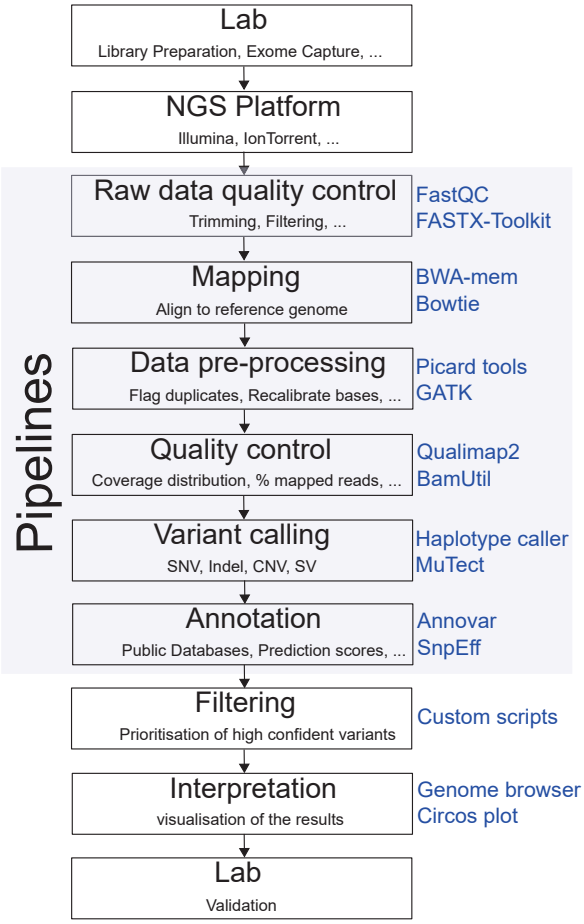


Figure 11: Typical workflow for NGS sequencing projects. Processing steps are shown in boxes and example of bioinformatics tools are shown on the right end side. Modified from [83].

Mapping Once the FASTQ files are checked and preprocessed, the next step is to map (align) the reads on a reference genome, which will produce a file in a binary sequence alignment map (BAM) format. Several alignment programs have been developed to efficiently map the millions of short reads. Depending on the type of sequencing different softwares are used. For WGS or WES, the reads need to be mapped to an entire genome with the two alignment programs widely used for this. The first one is bowtie2 developed at Johns Hopkins University [84] and the second one is BWA [85] developed by the Wellcome Trust Sanger Institute. For targeted sequencing, the reads need to be aligned

to a smaller proportion of the genome considering a high coverage. For the moment the sequencing providers propose proprietary solutions, which are instrument-dependent and not publicly available.

Data pre-processing and quality control Before moving to the variants discovery it is important to remove technical issues in the data by further processing the BAM files generated after the alignment.

Current NGS technologies rely on PCR amplification steps during the library preparations. After sequencing, identical multiple reads originating from only one template might be sequenced, creating PCR duplicates. The duplicate reads may bias the coverage metrics for complex templates (this does not apply to targeted sequencing) and they thus need to be flagged and not be taken into account in downstream analyses. In fact, for variant calling it is important that a duplicate reads are not be considered as an additional evidence for the detection of a variant. A popular tool for this step is called *Mark Duplicates* from the Picard tools suite (`https://broadinstitute.github.io/picard/`) developed by the Broad Institute.

Mapping softwares are often unable to map reads containing short insertions or deletions (Indels). It is thus important to correct these alignment artefacts for not introducing false positive SNVs. The realignment of the reads around those Indels may be performed with both GATK and Picard tools.

Finally, for improving variant calling accuracy it is important to assign accurate confidence scores to each base sequenced. These scores are given by the sequencing machines and indicate the per-base estimates of error. Unfortunately the machines are subject to various sources of systematic technical error and the results of these scores need to be adjusted, with tools such as *Base Recalibrator* from the GATK suite.

The quality of the alignment can be checked with tools such as Qualimap2 [86], BAMStats (`http://bamstats.sourceforge.net/`) or BamUtil (`http://genome.sph.umich.edu/wiki/BamUtil`). Qualimap2 and BAMstats propose a graphical user interface for facilitating the quality control assessment while BamUtil can only be used with a command-line interface. All the tools calculate statistics on the number of bases sequenced, the proportion of reads mapped, on the genome or on a targeted region, the coverage distribution across the reference genome, the overall mapping quality and the proportion of duplicate reads flagged. The descriptive statistics produced by Qualimap2 and BAMstats

are summarised in a HTML page with tables and graphs, whereas BamUtil produces a simple text file containing the different statistics.

Variant calling The recalibrated BAM files are the input for the most important part of NGS data analysis: the variant discovery. The greatest difficulty for this step is to a detect true genetic variation from sequencing and mapping errors. The choice of software depends on three parameters: the sequencing technology used, the type of variants to identify (SBS, Indels, copy-number, structural variants) and the type of study (calling germline variants or somatic variants). For targeted sequencing technologies, due to the design of the capture and regarding the type of study and variants to detect, there are few variant callers available and generally the one provided with the sequencing machine is used.

For genome-wide variant detection, there is a large panel of existing tools dedicated to each particular analysis. Some of them are summarised in Table 4 and discussed in the paragraphs below.

Table 4: Computational software for detecting genomic alterations. Modified from [87].

Program	Function	Application	Ref
BreakDancer	Structural variant	Germline	[88]
deFuse	Gene fusion detection	Germline	[89]
GATK (Haplotype caller)	SNV and Indel	Germline	[80]
ExomeCNV	CNV	Somatic	[90]
MuTect	SNV	Somatic	[91]
PatternCNV	CNV	Germline	[92]
Pindel	Indel	Germline	[93]
SAMtools	SNV and Indel	Germline	[94]
ScanIndel	Indel	Germline	[95]
Strelka	SNV and Indel	Somatic	[96]
VarScan2	SNV and Indel	Germline and Somatic	[97]

SNV: single nucleotide variant; Indel: small insertion and deletion; CNV: copy-number variation

Germline callers are used in inherited diseases studies for the characterisation of germline mutations, variants that differ from a given reference genome. Among the variety of software available we can cite SAMtools [94] for the detection of SBS and the tool *Haplotype caller* from the GATK suite for the detection of both SBS and Indels.

Somatic callers are used in cancer studies for identifying somatic mutations, variants that differ from the reference genome but that are also only present in the tumour samples (not in matching "normal"/germline samples). It is thus important to use a tool that would consider both the tumour and the germline samples to call the variants. One tool used at IARC is MuTect, a method developed by the Broad Institute for detecting somatic mutations with very low allele fractions followed by high-quality filtering to ensure high-specificity variants [91]. MuTect uses both the tumour and germline BAM files for first determining tumour variants that differ from the reference genome and then to make sure that the variant is not present in the germline of the individual. Only SBS variants are reported by MuTect and other tools were required for calling Indels or copy-number. In November 2015, a second version of MuTect was released and integrated within the GATK suite. This version, MuTect2, is able to detect both SBS and indels facilitating the detection of high-quality variants from NGS data. Another tool widely used in cancer research is VarScan2 [97], able to detect SBS and Indels from NGS data. Even though its performance for detecting low-allele fraction variants is lower compared to MuTect, VarScan2 outperforms MuTect for detecting variants with a high-coverage and high-allele frequency [98].

The output produced by the different variant callers is a text file in variant calling format (VCF) [99]. This generic format contains a header describing the different information present in the VCF (metadata), a line with the table column names, and then one line for each variant called. Having a same output format for all the different callers available allows a better comparison of their results.

There is not yet a gold standard to perform variant calling and despite the variety of variant caller tools available, identifying somatic alterations remains complex. In order to address this problem, the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) have launched the ICGC-TCGA DREAM Somatic Mutation Calling (SMC) Challenge in 2013 [100].

ICGC-TCGA DREAM Challenge is an open, collaborative international competition, crowd-sourcing initiative for identifying and improving current methods for the detection of somatic alterations (SBS, Indels, copy-number alterations and structural variations) in WGS data (https://www.synapse.org/#!Synapse:syn312572). Twenty-one teams evaluated different versions and the impact of parameter changes of various variant callers. Their results highlighted characteristic error profiles for different algorithms and a large variability in calling accuracy simply by changing the parameters of the softwares [101]. The Challenge is still ongoing but they suggested to combine the results of multiple variant callers to accurately detect somatic mutations.

Annotations The VCF file contains only little information, limited to the genomic coordinates of the variants, some quality scores and genotype information. In order to be able to study the consequences of the mutations found, it is therefore necessary to add functional annotations on the variants. Currently, there are three popular softwares dedicated to the annotation and prediction of the effects of genetic mutations: ANNOVAR [102], SnpEff [103] and Variant Effect Predictor (VEP) [104]. All softwares are freely available, use various up-to-date databases and can be used as command-line tools or via a web interface. ANNOVAR and VEP are based on Perl programming language while SnpEff is based on Java and is compatible with the GATK suite of tools. However, the outputs of the tools are different: ANNOVAR produces a tab separated file while SnpEff and VEP produce a VCF file using the "INFO" column for reporting their annotations.

Within our research group, we currently use ANNOVAR software to efficiently annotate and predict the effects of genetic mutations. ANNOVAR provides three different types of annotations and includes databases to be used for both human and mouse genomes. The complete list of available databases are listed on Annovar web page (http://annovar.openbioinformatics.org/en/latest/).

The first annotation is gene-based. It identifies whether the mutation (SNV or Indels) causes a protein coding change and annotates the amino acid affected using RefSeq genes, UCSC known genes and/or ENSEMBL genes resources.

The second annotation is region-based. It identifies variants in specific genomic regions, such as segmental duplication regions, conserved regions among 44 species, predicted transcription factor binding sites or ChIP-Seq peaks.

The last annotation is filter-based. It identifies variants reported in specific databases, for example, the presence of the variant in the Single Nucleotide Polymorphism Database (dbSNP), its allele frequency in the 1000 Genome Project, the Exome Sequencing Project (NHLBI-ESP) or in the Exome Aggregation Consortium (ExAC) and its presence in COSMIC [16]. Furthermore, it reports the prediction of the impact on protein function using, among others, SIFT [105], Polyphen2 [106], MutationTaster [107] prediction scores. Finally, users can also provide their own databases not currently provided by ANNOVAR to have additional customised annotations.

Filtering The annotations described above are then used to filter lists of variants in order to keep only those with a potential functional impact on the studied disease. The filtering process depends entirely on the project and generally requires in-house dedicated scripts.

For example, when analysing mutation patterns, all proven somatic variants may be kept as they all reflect mutagenic events while functional studies may want to retain only variants with functional impact on the proteins (removing synonymous variants or selecting deleterious variations based on the SIFT or Polyphen2 algorithms).

Interpretation The filtered list of variants can now be used for inferring the biological relevance of the obtained results. Using visual representation of the data helps to further prioritise and select potential candidate genes important for the studied disease.

Providing summary graphs of the characteristics of the identified mutations (distribution of mutations within their trinucleotide sequence context, functional impact, ...) helps to provide a global view of the alterations present in the studied samples.

Viewing the aligned reads using a genome browser such as the Integrative Genomics Viewer (IGV) [108] or GenomeView [109], allows to browse and check for potential alignment issues at specific positions where variants have been called or for cancer studies, if they are also present in the normal corresponding sample (supported by a number of reads below the threshold set by the variant caller).

Furthermore, the UCSC genome browser [110] may be used to display the identified mutations combined with annotations facilitating the biological interpretation of the data. Visualisation of the identified genomic alterations provides a global picture of the genomic alterations present in the set of samples analysed. A tool such as Circos [111] uses a circu-

lar ideogram to display positions and relationships between any kind of genomic elements and intervals (SBS, Indels, structural rearrangement, gene expression, ...) facilitating the comparison of results from multiple samples. The visualisation can help to further filter the list of mutations and can be followed by functional interpretation of the genes mutated using signalling, regulatory or metabolic pathways analyses.

3.3 Computing infrastructures

The analysis of NGS data requires important computing resources and advanced bioinformatic skills. In order to perform such analysis requiring a lot of computing resources and power, researchers needs to use high-performance computing (HPC) clusters. And these analyses can be facilitated using the Galaxy platform.

3.3.1 High-performance computing

HPC clusters contain the same elements as regular desktop computers (operating system, disk, memory and processors) but they have more of them. The interest of using a HPC cluster is that each individual computer of the cluster can perform intensive computational tasks in parallel that would take a lot of time with a single computer.

Each "individual computer" in the cluster is called a computing node and there is one master machine managing them, called "head node" (HN). IARC HPC cluster is composed of one HN and twelve computing nodes (named cn01 to cn12) linked to a main data storage of 60TB and a backup storage of 100 TB. (Figure 12). Each node contains 96GB of RAM shared by 2 CPUs of 12 cores, totalling 24 processes per node.

The HN and computing nodes need to communicate between each others to work in an efficient way. The communication is made through the network and dedicated softwares exist for managing the communication between the nodes. IARC HPC cluster uses the Load Sharing Facility (LSF) management system also called "job scheduler" to schedule and manage transfer of information between the head and its nodes.

Users access the HPC cluster using their regular desktop computer via a secure shell (ssh) protocol, and run analyses, call jobs, via the HN using a language specific to the job scheduler. The amount of resources, memory and processors, required for the script to run correctly is also specified.

Once the job is finished, the result is written on the HN and users can retrieve them on their own computer.

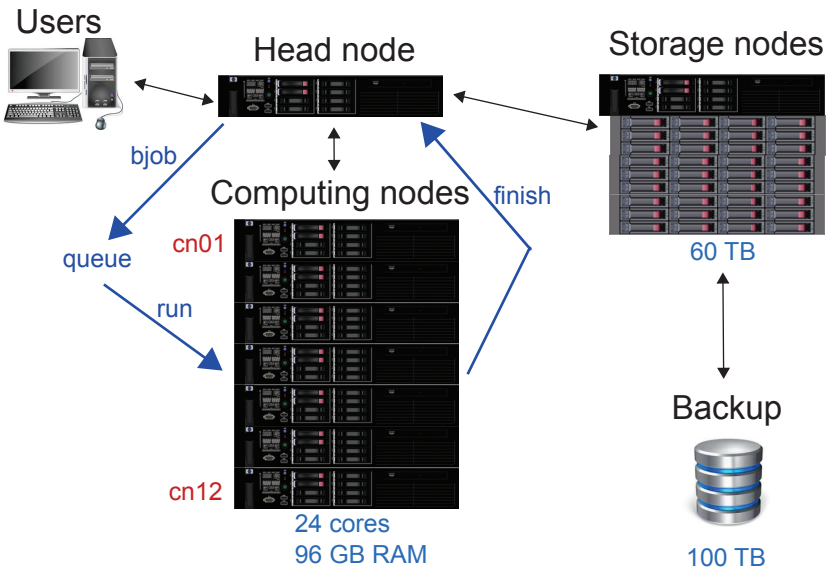


Figure 12: Architecture of the IARC HPC cluster. Users connect to the head node (HN) from any IARC computer. In blue is shown a typical job workflow: the job is launched from the HN, added to a queue until the computing resources are available and once finished the result is returned to the HN.

3.3.2 Galaxy platform

The web-based platform Galaxy provides a graphical environment for the integration of complex bioinformatic tools or scripts written in any programming language [112]. Galaxy allows to analyse, reproduce and share complete workflows within a graphical user interface, allowing researchers to perform complex bioinformatic analyses without having to use command lines or scripts (Figure 13). A strength of Galaxy is the creation of Workflow which allows to reproduce identically a multi-step analysis in a simple and easy way.

The platform is accessible as a free public web server (`https://usegalaxy.org/`) that includes many bioinformatic tools commonly used in genomics research, but can also be installed locally in order to address specific needs with the integration of in-house tools. Integrating a tool in Galaxy is quite simple, you just need to create a repository containing

the developed scripts and all the dependencies required by the tool to run properly.

These tools can then be shared freely with the community using the Galaxy ToolShed (`http://toolshed.g2.bx.psu.edu`) that hosts all Galaxy repositories as well as exported Galaxy Workflows.

Figure 13: Web interface of IARC Galaxy. Top menu in red contains the access to the different sections of Galaxy (Analyse data, Workflow, Help and User). The left menu in green lists all the tools available. The right menu in yellow contains all the analyses performed. The middle panel in purple displays the tools menu and the results.

At IARC a local instance of Galaxy was installed on a server with 50 GB of RAM shared by 2 CPUs of 6 cores and 14 TB of disk space. The IARC Galaxy instance enables scientists of the Agency to analyse NGS data from the quality control of the raw data generated by the sequencing machine to complex analyses, such as looking for mutational signatures, within a user-friendly environment.

OBJECTIVES

OBJECTIVES

Somatic mutations accumulating in cancer genomes reflect endogenous and exogenous mutagenic events, including carcinogen exposures, and are thus informative of cancer's life history and possible aetiology. The analysis of somatic mutation patterns at the genome-wide level with high-throughput sequencing identified at least 30 distinct mutational signatures in 40 cancer types [43]. Although some of the signatures have been linked to known mechanisms of mutagenesis, the origin of most of these signatures remains to be elucidated.

To identify mutational signatures that characterise human cancers and understand their origin, the Molecular Mechanisms and Biomarkers (MMB) group at IARC has developed an integrated approach using massively parallel sequencing and involving the systematic screening of the mutagenic effects of cancer-risk compounds in *in vitro* assays and *in vivo* bioassays, as well as the systematic analysis of cancer data extracted from public repositories or generated from series of patients with documented exposure to the studied compound (Figure 14).

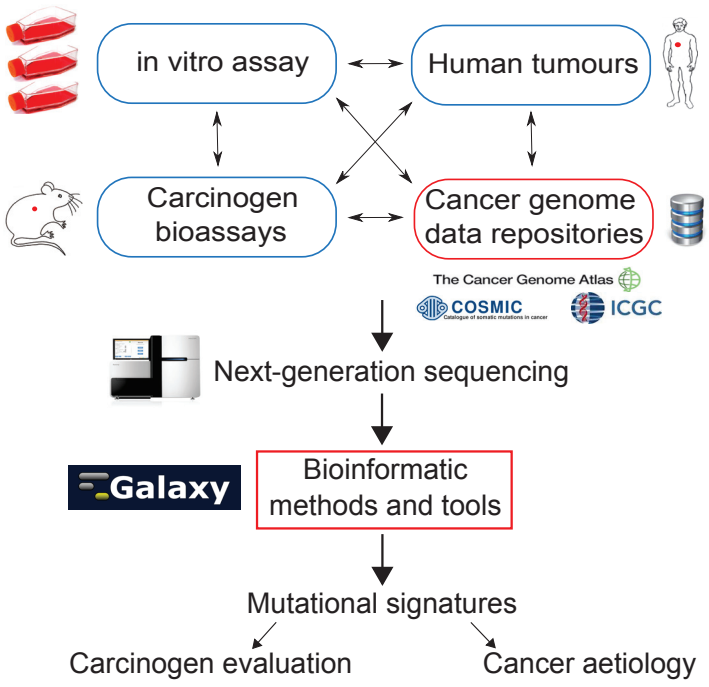


Figure 14: Overall approach for the identification of mutational signatures induced by human carcinogens.

Within this integrative framework, the overall aim of my thesis was to develop bioinformatic tools and methods facilitating the analysis and interpretation of carcinogen-specific mutational signatures from NGS data, and to apply these methods to specific NGS datasets in order to characterise the mutational signatures of aristolochic acid (AA) and other carcinogens.

The specific objectives of my thesis are summarised below.

1. Develop user-friendly bioinformatic tools and methods for extracting and identifying mutational signatures from mutation lists obtained from rodent and human cancer genomes (**Paper 1, Ardin M et al.**).

2. Characterise carcinogen-specific mutational signatures from a multi-system approach involving *in vitro* assays using mouse embryonic fibroblasts (MEF) and human cancer genomes from exposed individuals (**Paper 2, Olivier M et al.**).

3. Investigate carcinogenesis environmental due to exposure to AA

   (a) Establish a low-coverage WES and analysis approach for detecting patients exposed to AA using archived DNA material (**Paper 3, Castells X, Karanović S, Ardin M et al.**).

   (b) Investigate the implication of AA in the aetiology of new tumour types in the Balkan endemic nephropathy (BEN) regions using WES on archived tumour DNAs (**Paper 4, Jelaković B et al.**).

   (c) Characterise the specificity of AA mutational signature using a systematic data mining approach of public genomic data and cancer genomes of exposed individuals (**Paper 5, Ardin M et al., in preparation**).

RESULTS

RESULTS

The articles presented in this chapter refer to the specific objectives of this thesis and are preceded by a brief introduction.
The other articles in which I co-authored are presented in the Appendix.

Objective 1

Paper 1: MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes

Maude Ardin, Vincent Cahais, Xavier Castells, Liacine Bouaoun, Graham Byrnes, Zdenko Herceg, Jiri Zavadil and Magali Olivier

BMC Bioinformatics, 2016

Aim Develop user-friendly tools for the processing of mutation data and the identification of mutational signatures in the genome of various species.

Approach Within the web-based platform Galaxy, implement a suite of public or in-house tools and scripts allowing the identification of mutational signatures. Integrated the workflows in a pipeline named MutSpec (for Mutation Spectra), designed to work with rodent and human data.

Graphical summary

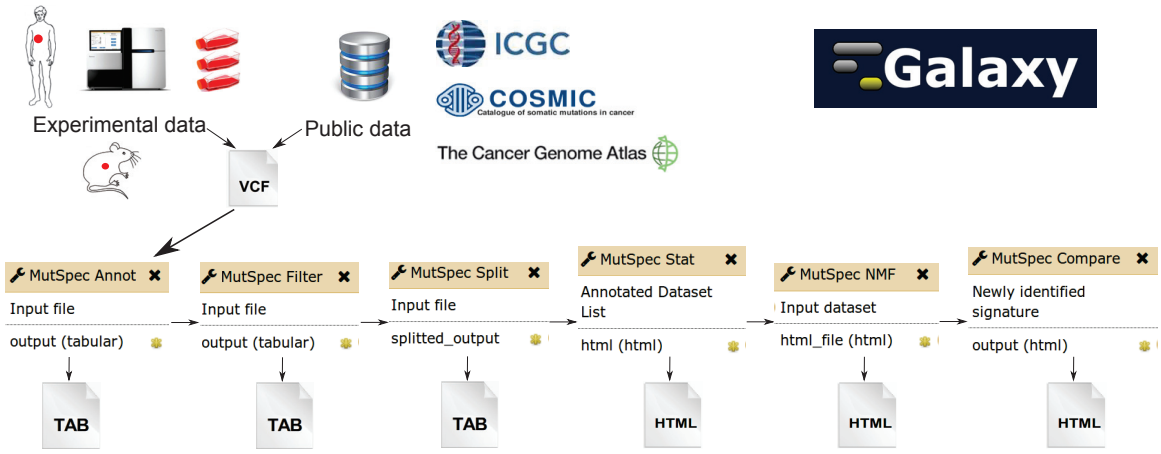


Figure 15: MutSpec pipeline. MutSpec integrates a suite of tools that perform variant annotation (MutSpec-Annot), filtering (Mutspec-Filter), compute various statistics describing mutation spectra features (MutSpec-Stat) and extract mutational signatures considering 96 parameters (six types of single base substitutions (SBS) within their immediate 5' and 3' sequence context) (MutSpec-NMF). A last tool compares the obtained signatures with published ones (MutSpec-Compare) to help interpreting the obtained signatures.

Novelty and highlights

- MutSpec tools implementation in the user-friendly Galaxy platform is unique in facilitating mutational signature analysis to non-bioinformaticians.
- The tools accept as input a list of mutations obtained from various variant callers.
- MutSpec tools are compatible with human as well as rodent (mouse and rat) genome data.
- MutSpec provides a comprehensive workflow from mutation annotation to mutation signature comparison with known signatures.

Ardin et al. BMC Bioinformatics (2016) 17:170
DOI 10.1186/s12859-016-1011-z

BMC Bioinformatics

SOFTWARE Open Access

MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes

Maude Ardin[1], Vincent Cahais[2], Xavier Castells[1], Liacine Bouaoun[3], Graham Byrnes[3], Zdenko Herceg[2], Jiri Zavadil[1] and Magali Olivier[1*]

Abstract

Background: The nature of somatic mutations observed in human tumors at single gene or genome-wide levels can reveal information on past carcinogenic exposures and mutational processes contributing to tumor development. While large amounts of sequencing data are being generated, the associated analysis and interpretation of mutation patterns that may reveal clues about the natural history of cancer present complex and challenging tasks that require advanced bioinformatics skills. To make such analyses accessible to a wider community of researchers with no programming expertise, we have developed within the web-based user-friendly platform Galaxy a first-of-its-kind package called MutSpec.

Results: MutSpec includes a set of tools that perform variant annotation and use advanced statistics for the identification of mutation signatures present in cancer genomes and for comparing the obtained signatures with those published in the COSMIC database and other sources. MutSpec offers an accessible framework for building reproducible analysis pipelines, integrating existing methods and scripts developed in-house with publicly available R packages. MutSpec may be used to analyse data from whole-exome, whole-genome or targeted sequencing experiments performed on human or mouse genomes. Results are provided in various formats including rich graphical outputs. An example is presented to illustrate the package functionalities, the straightforward workflow analysis and the richness of the statistics and publication-grade graphics produced by the tool.

Conclusions: MutSpec offers an easy-to-use graphical interface embedded in the popular Galaxy platform that can be used by researchers with limited programming or bioinformatics expertise to analyse mutation signatures present in cancer genomes. MutSpec can thus effectively assist in the discovery of complex mutational processes resulting from exogenous and endogenous carcinogenic insults.

Keywords: Galaxy, Mutation spectra, Mutation signatures, Single base substitutions

Background

DNA mutations accumulate during the natural history of tumors, reflecting the insults from endogenous and exogenous mutagenic processes as well as the selection of cancer-driving events. The nature of somatic mutations observed in single genes or on a genome-wide scale in human tumors can thus reveal information on past carcinogenic exposures and provide clues on cancer etiology [1, 2]. Current efforts in the systematic sequencing of tumor genomes generate large amounts of data on mutation patterns that characterise human cancers. Recent analyses of these data have revealed over 30 somatic mutation signatures [1, 3–6]. While suspected mutational processes have been proposed for some of these signatures, the majority have not yet been attributed to any specific mechanism and their origins thus remain unexplained. Experimental systems developed for modelling in vitro and in vivo genome-wide mutational processes have been reported recently [7–13]. These assays have the potential to generate direct evidence for the identification

* Correspondence: olivierm@iarc.fr
[1]Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, F69372 Lyon, France
Full list of author information is available at the end of the article

of carcinogens or mutagenic processes underlying the mutation signatures observed in human tumors. The analysis of experimental and human-derived data requires advanced bioinformatics skills and thus remains limited to a small research community. Tools that would allow streamlined analyses of mutation spectra from genome-wide sequencing data and be accessible to a wider community could speed up research in this area.

Galaxy is a web-based platform that allows the integration of complex programs or scripts built in any language and accessible in a single web interface [14–16]. Tools can be built so that users with no programming skills can perform complex analyses through a user-friendly graphical interface.

Here we present a set of Galaxy tools named MutSpec that offer an accessible framework for advanced analyses of mutation spectra and signatures present in human cancers or experimental systems. MutSpec expands on existing approaches and methods and integrates scripts developed in-house with publicly available R packages to offer a user-friendly interface accessible to biologists with no or limited bioinformatics skills. MutSpec is expected to accelerate the interpretation of mutation patterns observed in human cancers by facilitating their analysis by a wider community and should thus contribute to the identification of new human carcinogens and to a better understanding of how these carcinogens impact the genome.

Implementation

Overview and code sources

MutSpec is an implementation of Perl and R scripts or packages into several Galaxy tools designed to (1) annotate genome variations (MutSpec-Annot), (2) filter and parse list of variants (MutSpec-Filter and MutSpec-Split respectively), (3) compute various statistics describing mutation spectra features (MutSpec-Stat), (4) extract mutation signatures defined by the six types of single base substitutions (SBS) in their trinucleotide sequence context (MutSpec-NMF), (5) compare the obtained signatures with published ones (MutSpec-Compare). The tools are designed to work in a logical sequence, using as input the outputs of each preceding tool. A typical analysis workflow is shown in Fig. 1. The public packages used and the Perl scripts that support each tool are described in Table 1. The tools produce simple tab-delimited text files or content-rich html pages with graphical representations of the data and hyperlinks to underlying data. All figures and tables that are produced by the different tools can be downloaded as individual files. Format requirements and details on the produced outputs are described below.

Data import and formatting

The first tool to be run is MutSpec-Annot. This tool will retrieve different types of structural and functional annotations that will be used by the other MutSpec tools. MutSpec-Annot accepts variant call format (VCF, version 4.1) files as well as tab-delimited text files that may be obtained from popular sources such as the International Cancer Genome Consortium (ICGC) or The Cancer Genome Atlas (TCGA) data portals and Catalogue of Somatic mutations in Cancer (COSMIC) database. The minimal information required is, for each variant, the chromosome number, the start genomic position, the reference allele and the alternate allele. The columns containing this information should have a header. There are different supported names for these header columns (case-sensitive names) that correspond to formats of data retrieved from popular variant callers or public databases (details provided in the tool interface). These four columns may be in any order, and other columns may be present. The additional columns will be kept in the output file after the retrieved annotations. Galaxy automatically recognises file formats, but if the format of the imported files need to be corrected, it can be easily done by editing the file attribute or by using the tool Convert (a default tool available in Galaxy).

The output of MutSpec-Annot is a tab-delimited text file. This file may be used as input of the tools MutSpec-Filter and MutSpec-Split that both require a tab-delimited text file as input. These later tools are optional as the imported data may not need to be filtered or parsed (see an example analysis further below). The next tool to use is MutSpec-Stat that requires a dataset list (also named 'collection') as input, a specific feature of Galaxy. MutSpec-Stat is designed to calculate statistics on mutation features for each individual samples as well as on the sample pool. MutSpec-Stat generates an Excel file (see Additional file 1 as example), with results of individual samples in individual datasheets, and several html pages showing summary results for each sample with links to data downloads. MutSpec-Stat output can be used directly with the tool MutSpec-NMF for extracting mutation signatures. MutSpec-NMF also accepts a tab-delimited matrix formatted as specified in the tool interface. MutSpec-NMF output can be directly used as input of MutSpec-Compare. MutSpec-Stat, –NMF and -Compare produce results in graphical and tabular formats that can be displayed as html pages or downloaded as tab-delimited text files or images (Fig. 1).

Annotations, filtering and databases

MutSpec-Annot uses the ANNOVAR software [17] to provide various functional annotations of variants, as well as Perl scripts developed in-house to retrieve strand orientation of transcripts and sequence context of variants. ANNOVAR includes several databases and annotation types for mouse and human genomes (required databases and corresponding genomes are listed in Table 2), among

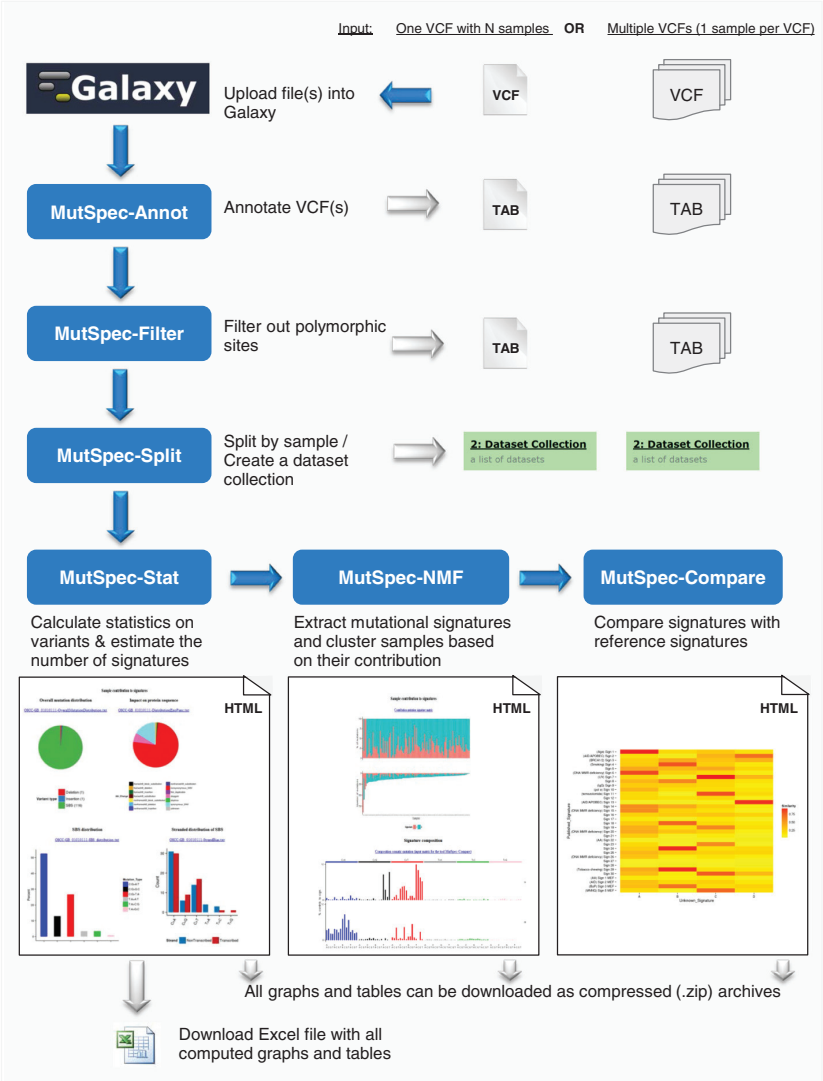


Fig. 1 Overview of MutSpec tools and workflow. List of variants identified in a set of cancer samples may be imported as a single VCF files that contains all samples (identified by a sample ID) or as multiple VCFs (one for each sample). The first tool to use is MutSpec_Annot for annotating variants with structural and functional information. These annotations may be used to filter out variants that are known polymorphisms or located in segmental duplication regions with the MutSpec-Filter tool. If a single VCF file containing several samples is uploaded, the MutSpec-Split tool should be used to split data by sample using the sample ID. This tool generates automatically a dataset collection. If multiple VCFs are uploaded, MutSpec-Split should not be run but the annotated VCF should be grouped in a dataset collection. MutSpec-Stat can then be run on the dataset collection to generate various statistics on variants characteristics. These statistics may be visualized as graphs on html pages or downloaded as a single Excel file. The report generated by MutSpec-Stat can then be used as input of MutSpec-NMF to extract mutation signatures present in the sample set. MutSpec-NMF generates plots showing the identified signatures and the contribution of each signature to the mutation load of each sample. Finally, MutSpec-Compare can be used to calculate cosine similarity values between the obtained signatures and a set of reference signatures (published or user-defined). These results are shown as a heatmap

which some are optional and some are required for MutSpec tools to function properly. MutSpec has been validated for hg19 and mm9 genome builds. Other genomes and ANNOVAR databases may be installed to retrieve additional annotations or study other species based on user preferences. Database updates are regularly provided by ANNOVAR, users should thus check ANNOVAR website for these updates and install them as needed (we created an install file, listAVDB.txt that can be modified by the

Table 1 Algorithms and code sources used in MutSpec

Packages	Version	MutSpec-				Source
		Annot	Stat	NMF	Compare	
Annovar[a]	June 2015	X	-	-	-	[17]
Statistics::R[a]	0.33	-	X	-	-	http://search.cpan.org/~gmpassos/Statistics-R-0.02/lib/Statistics/R.pm#AUTHOR
Spreadsheet::WriteExcel[a]	2.40	-	X	-	-	http://search.cpan.org/~jmcnamara/Spreadsheet-WriteExcel-2.40/lib/Spreadsheet/WriteExcel.pm
ggplot2	1.0.1	-	X	X	X	http://ggplot2.org/
gplots	2.17.0	-	X	-	-	http://cran.r-project.org/web/packages/gplots/gplots.pdf
gtable	0.1.2	-	X	-	-	http://cran.r-project.org/web/packages/gtable/gtable.pdf
reshape	1.4.1	-	X	X	X	http://cran.r-project.org/web/packages/reshape/reshape.pdf
scales	0.2.5	-	X	X	-	http://cran.r-project.org/web/packages/scales/scales.pdf
gridExtra	0.9.1	-	X	X	-	http://cran.r-project.org/web/packages/gridExtra/gridExtra.pdf
NMF	0.20.6	-	X	X	-	[18]
getopt	1.20.0	-	-	X	-	http://cran.r-project.org/web/packages/getopt/getopt.pdf
lsa	0.73.1	-	-	-	X	http://cran.r-project.org/web/packages/lsa/lsa.pdf

[a]These packages are developed in Perl while all other packages are developed in R

Galaxy administrator to specify related databases and reference genomes to be used).

Once variants are annotated, the MutSpec-Filter tool may be used to filter out variants that are likely neutral polymorphisms or that are contained in duplicated regions of the genome. For the human genome, there are currently three databases available for polymorphisms: dbSNP (http://www.ncbi.nlm.nih.gov/SNP/), the 1000 genomes project (http://www.1000genomes.org/) and the Exome Sequencing Project (ESP, http://evs.gs.washington.edu/EVS/) databases. Users may filter against all three or any of these databases. Filtering against dbSNP database will remove all variants with an rs number (SNP ID). It is important to note that ANNOVAR provides two versions of dbSNP, dbSNP138 that includes all variants present in dbSNP and dbSNPNonFlagged that includes only variants that are frequent in human populations (>1 %) and that are not flagged as "clinically associated" in dbSNP database. We prefer to use dbSNPNonFlagged but users should decide. Another caution about filtering with dbSNP concerns the fact that rs numbers in dbSNP database may correspond to several variants. Although ANNOVAR will only identify exact match by taking into account not only position but also nucleotide change, annotations about a specific variant may not be accurate (more details in: "Assigning dbSNP Identifiers" at http://annovar.openbioinformatics.org/en/latest/articles/dbSNP). With 1000 genomes and ESP filters, the tool will remove variants according to a predefined standard frequency. To use different frequencies, it is recommended to use other tools proposed in Galaxy. For the mouse genome, there is currently only one SNP database available, dbSNP.

Table 2 List of databases and reference genomes

Reference genome	Related databases	Used in
hg19	refGene	MutSpec-Stat
	genomicSuperDups	MutSpec-Filter
	snp138	MutSpec-Filter
	snp138NonFlagged	MutSpec-Filter
	1000 genome (ALL)	MutSpec-Filter
	esp6500siv2 (ALL)	MutSpec-Filter
mm9	refGene	MutSpec-Stat
	genomicSuperDups	MutSpec-Filter
	snp128	MutSpec-Filter

Statistics on variant features

MutSpec-Stat provides various statistics on the characteristics of mutations observed in a human or mouse sample or group of samples (see Table 3). Briefly, summary statistics include counts and distributions of overall mutation types (six types of SBS and indels) and their functional impact (based on RefSeq annotations); the distribution of SBS in different genomic regions or by chromosome; counts and distributions of SBS in their trinucleotide sequence context (96 mutation types) and calculated on the genome sequence or only on transcribed sequences (stranded analysis). The stranded analysis calculates the strand bias for both the 6 and 96 types of SBS. Statistical tests are applied for the stranded analysis (the significance of the differences between the mutational frequencies on the non-transcribed and the transcribed strand is assessed using a chi-squared test followed by the Benjamini-Hochberg procedure for multiple testing corrections), and for the chromosomal distribution of mutations (a Pearson correlation coefficient is calculated to assess the correlation between

Table 3 Analyses performed by the tool MutSpec-Stat

Analysis	Table	Graph	Statistics
Overall mutation distribution	-	X	-
Impact on protein sequence	X	X	-
SBS type distribution	X	X	-
Stranded analysis of SBS type distribution	X	X	Chi-squared test
SBS distribution by functional region	X	-	-
Strand bias by functional region	X	-	-
SBS distribution per chromosome	X	-	Pearson Correlation
Trinucleotide sequence context of SBS on the genomic sequence	X	X	-
Stranded analysis of trinucleotide sequence context of SBS	X	X	-

SBS counts and chromosome size). An option in MutSpec-Stat is to compute statistics that can be used to estimate the number of mutation signatures present in the analysed dataset (an NMF R package is used [18], see next section). Another option available is the computation of statistics on the pooled samples.

The output of MutSpec-Stat is an html page that contains links to summary results for each individual sample and to an Excel file (called "Report") that contains sample datasheets with all results displayed in various formats (tables, heatmaps, bar graphs, matrices, WebLogo). Each datasheet is named after the sample ID. It is of note that because Excel has a limitation on datasheet names, sample identifiers must be within a limit of 31 characters. All individual tables and graphs can also be downloaded as individual files in a compressed archive.

Extraction of mutation signatures

MutSpec-NMF extracts the minimal set of mutation signatures that optimally explains the proportion of each mutation type (96 types represented by the 6 base substitutions in their trinucleotide sequence context) found in each sample and then estimates the contribution of each signature to each sample. MutSpec-NMF uses the non-negative matrix factorization algorithm from Brunet et al. [19] implemented in the NMF R package developed by Gaujoux and Seoighe [18]. The Brunet algorithm has been successfully used to extract mutation signatures from somatic mutation data in human cancers [3, 8]. Here we use the default algorithm of Brunet with the Kullback–Leibler divergence penalty and a number of iterations set to 200 in order to achieve stability of the results. The aim of the method is to reduce the dimension of the original data, with the caveat that the factorisation rank needs to be specified. It is thus necessary to first estimate the factorisation rank (number of expected signatures) for the analysed dataset. This can be done with the option available in MutSpec-Stat that performs NMF with different rank values (2 to 8 by default) and compute some quality measures of the results, including the cophenic coefficient and the rss curve. The NMF R package cited above is also used to perform these analyses. The best rank value indicated by the quality measures may be used as the number of expected signatures in MutSpec-NMF as suggested by the authors of the NMF package [18]. The calculation of these statistics is optional in MutSpec-Stat because running NMF on a large dataset requires intensive computations (for estimating the rank value a total of 50 runs are performed for each value while 200 runs are performed for the full analysis). For reducing the computation time it is recommended to use all available central processing unit (CPU) on the machine where Galaxy is installed (to be checked with the Galaxy administrator). The input matrices for NMF are extracted from the output of MutSpec-Stat, so that users can select a MutSpec-Stat report as input. Users may alternatively use matrices imported from other analyses as long as they are in the required format (the tool works with a tab-delimited text file or a MutSpec-Stat report as input). For example, matrices obtained from different MutSpec-Stat analyses may be combined to run NMF on groups of samples that have been analysed separately (see example in Additional file 2).

Results are shown graphically as bar charts representing the obtained signatures or showing the contribution of each signature to the mutation load of each sample (Fig. 1). This package also performs unsupervised hierarchical clustering of samples based on mutation signature contributions. It should be noted that mutation signatures are based on 96 SBS types, the current standard in the field that is used in the COSMIC database. It does not allow deriving other types of signatures as it is designed to produce signatures comparable to the ones compiled in the COSMIC database and to produce preformatted graphs.

Comparison of obtained signatures with published signatures

MutSpec-Compare computes the similarity between the signatures identified by MutSpec-NMF and a set of published signatures using the cosine similarity method

implemented in the LSAfun R package [20]. This method measures the similarity between two vectors of an inner product space by calculating the cosine of the angle between them. The resulting values range between 0 and 1, corresponding respectively to an absence or a complete similarity. Results are displayed graphically as a heatmap and provided as a tabular matrix. A cosine value above 0.9 can be considered as a good match. For the reference signatures, user may select the matrix provided with the tool or their own matrix. The matrix provided includes 30 signatures published in the COSMIC database (v72) [21] plus four experimental signatures (methylnitronitrosoguanidine, aristolochic acid, benzo(a) pyrene, activation-induced cytidine deaminase) previously published in Olivier et al. [8]. As this tool requires two text-tabulated matrices, users may also input matrices produced by other tools as long as they are in the required format.

Results and discussion

To illustrate MutSpec functionalities and show an example of analysis workflow, we have analysed a public dataset reporting mutation data on 106 cases of oral squamous cell carcinomas (OSCC) from India [22]. The aetiology of OSCC is linked to several risk factors, including tobacco smoking, tobacco chewing, alcohol drinking, HPV infection and UV radiations. These risk factors vary between different geographical regions, with tobacco chewing being prevalent in the Indian population while the association with tobacco smoking and alcohol drinking play major roles in Western countries. Tobacco and UV are strong mutagens that create specific types of DNA damage; one can thus expect to identify various mutation signatures reflecting exposure to these mutagens in the selected dataset. Screenshots of the following steps of analyses are provided in Additional file 3 to illustrate MutSpec functionalities.

OSCC data retrieval and annotation

Somatic mutation data were retrieved from the ICGC data portal (ORCA-IN dataset, downloaded on June 2015). The dataset was available as a tab-delimited text file containing a single list of mutations for the 106 samples. This list was uploaded in a Galaxy history. ICGC formatted files are supported by MutSpec-Annot so no further formatting was needed.

The first step was to annotate the file with MutSpec-Annot. Variants in OSCC-IN dataset were mapped to the genome build hg19, we thus selected "hg19" as reference genome. Another option to specify is the length of the sequence context to retrieve; here we chose 1 as we are only interested in the trinucleotide sequence context (one base on each side of the variant base). One output file, OSCC-IN_annotated was thus created and appeared in the history. We then proceeded to the next step, which is to filter out potential polymorphisms with MutSpec-Filter. We filtered against all databases but dbSNP. Although the data analysed have been curated and thus should have been already filtered, 673 (5 %) variants were removed. Then, we used MutSpec-split to parse the file into individual sample files using the sample ID column. MutSpec-split automatically creates a collection of files, so we obtained a dataset collection of 106 files.

Mutation spectra in OSCC

The dataset collection created by MutSpec-Split was then used as input for MutSpec-Stat to generate statistics on mutation spectra for each sample and to compute the mutation matrix to be used for extracting mutation signatures. We ran the tool with the "pool sample" option in order to obtain statistics for the pooled samples. The reference genome should be specified again at this step. Finally, we also selected the option that calculates statistics for estimating the number of signatures present in the dataset. A summary of the results are shown in Fig. 2 for the sample pool (see detailed results in Additional file 1). The overall mutation pattern shows that the majority of variants are non-synonymous SBS (Fig. 2a), and that the most frequent SBS types are C:G > A:T followed by C:G > T:A (Fig. 2b). The trinucleotide sequence context distribution of these mutations show specific patterns, with C > A occurring preferentially within 5'-GCN-3' motifs and C > T within CpG sites (Fig. 2c). The third most frequent SBS are C > G. Both C > G and C > T occur preferentially within 5'-TCN-3' motifs, suggesting the presence of APOBEC-induced mutations [23]. Based on the cophenic and rss statistics calculated for estimating the NMF factorization value, 4 signatures may be present in this dataset as it is the first value for which the cophenetic coefficient starts decreasing and where the rss curve presents an inflection point (Fig. 2d).

Extraction and identification of mutation signatures in oral cancers

To analyse mutation signatures present in the dataset, we ran MutSpec-NMF using the output report of MutSpec-Stat and factorisation value was set to 4. We then compared the obtained signatures with published signatures using the tool MutSpec-Compare. Fig. 3a shows the 4 signatures obtained. Signature A matched best with signature 1 (Age), signature B with signatures 29 (tobacco chewing) and 24 (aflatoxin), signature C with signature 7 (UV) and signature D with signature 13 (APOBEC) (Fig. 3b). MutSpec-NMF also produces a graph showing the total number of SBS per sample and the proportion contributing to the 4 signatures (Fig. 3c). On this graph, one sample is standing out with the largest number of

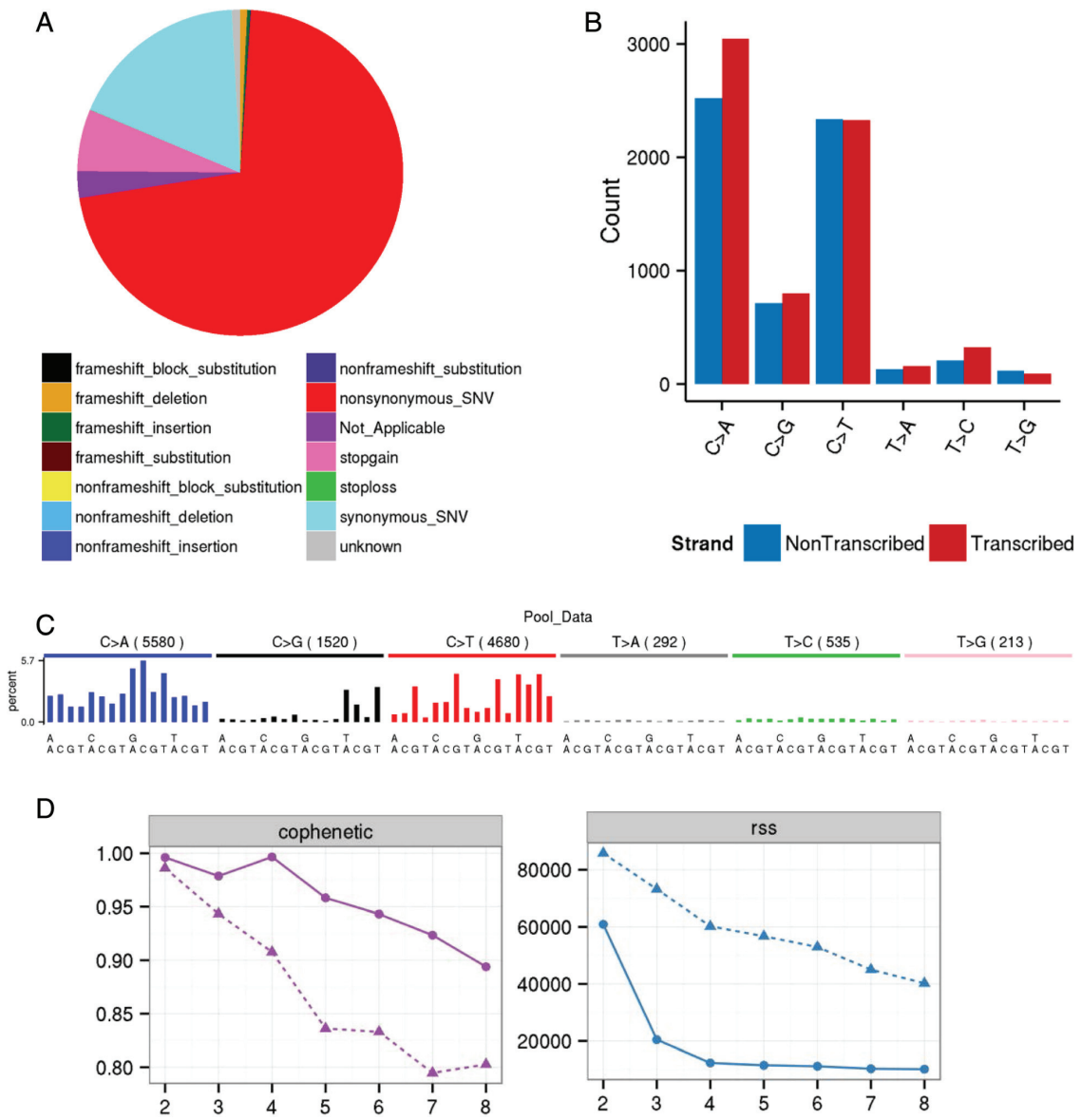
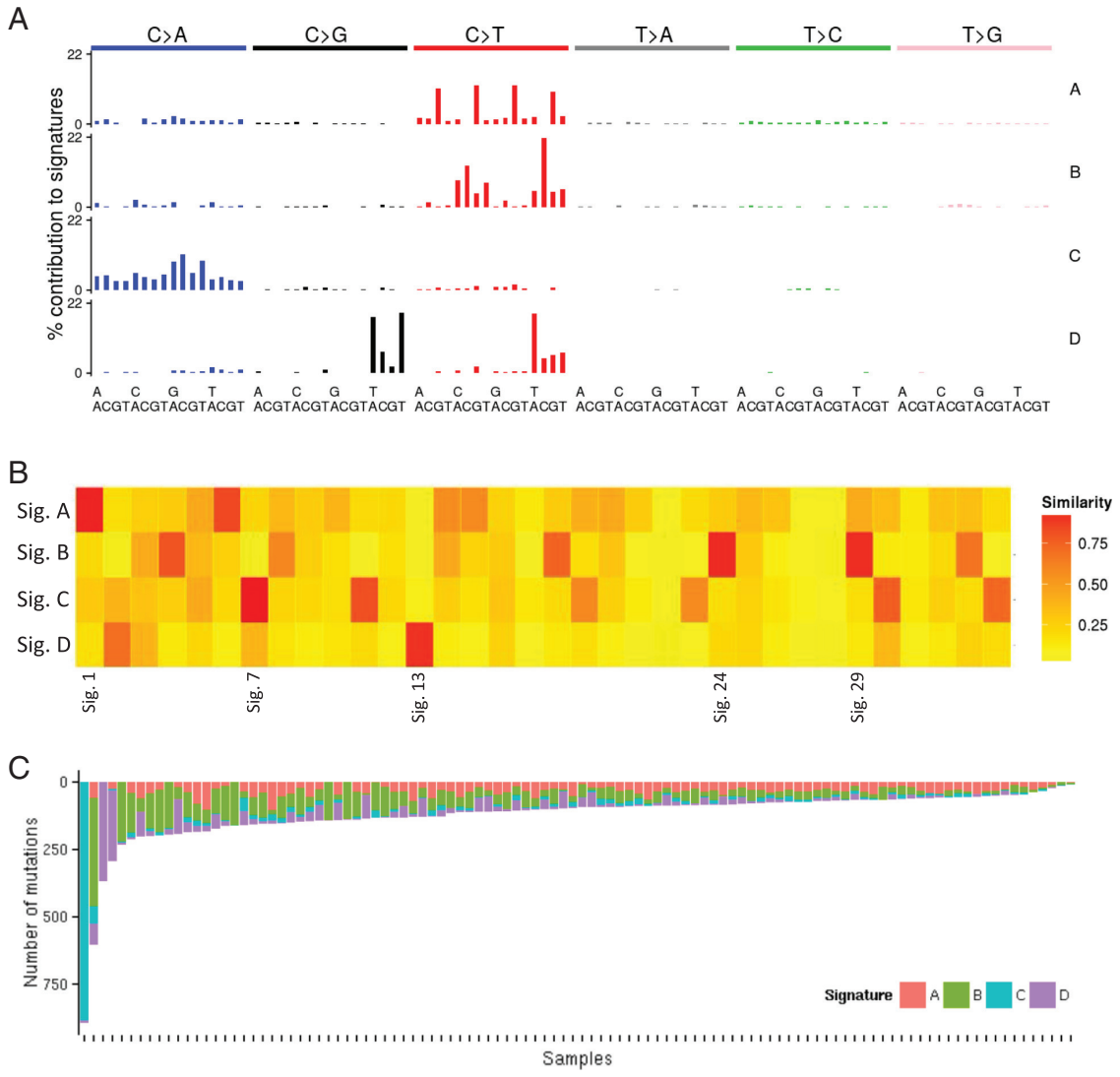


Fig. 2 Mutation spectra in OSCC from Indian patients. Results for the pool of 106 samples are shown. **a** Distribution of variants (N = 13059) according to their functional impact on protein sequences. **b** Stranded analysis of the 6 types of SBS showing counts for SBS with transcript annotations (N = 12789). **c** Distribution of SBS according to their trinucleotide sequence context (SBS counts are indicated in parenthesis). **d** Plots of the cophenic and rss analyses using a range of factorisation values (2 to 8). The solid lines represent the results obtained with the original data while the dotted lines represent the results obtained with randomized data (original data are shuffled)

SBS and a close to 100 % contribution to the UV signature (sig 7). Finally, NMF clusters samples based on their signatures composition. From these data, MutSpec-NMF produces a summary analysis that shows the number of samples by cluster and the average contributions of each signature in each cluster (Fig. 3d). In the 106 OSCC samples analysed here, we found one sample likely to be related to UV exposure (high number of SBS corresponding to the previously reported UV signature, sig.7) while a majority of samples (N = 47) had a predominant signature related to tobacco chewing and/or aflatoxin. Another large set of samples (N = 41) had the age signature as the predominant signature, and in a small number of samples (N = 17) the APOBEC signature was the most



D

	Sig. A	Sig. B	Sig. C	Sig. D	Number of samples
Cluster 1	52.77	24.81	11.58	10.84	41
Cluster 2	20.53	61.06	8.46	9.96	47
Cluster 3	0.00	0.00	98.88	1.12	1
Cluster 4	20.59	18.32	7.26	53.83	17

Fig. 3 Mutation signatures in Indian OSCC and their suspected origin. Summary results of MutSpec-NMF and MutSpec-Compare analyses obtained on the 106 OSCC samples. **a** Mutation signatures obtained by NMF with a factorisation value of four. **b** Comparison of the four OSCC signatures (vertical axis) with 34 reference signatures (horizontal axis) using the cosine similarity method. The heatmap is color-coded according to the cosine value that ranges from 0 to 1. Only reference signatures with a significant match (cosine > 0.9) are labelled. **c** Number of mutations contributing to each of the four signatures identified, for each sample analyzed. **d** Average contributions of the four identified signatures to the mutation load of clustered samples and number of samples by cluster

prominent. Because the cases analysed are from India where tobacco chewing is one major risk factor for oral cancer, it is more likely that the signature B found here is related to tobacco chewing and not aflatoxin. These two signatures (24 and 29) are in fact very close (they share several predominant C > A in specific contexts due to similar mechanisms of the suspected underlying carcinogens) and thus difficult to separate by NMF [5]. The fact that a majority of samples were found to carry this signature confirms the major role of tobacco chewing in the etiology of OSCC in India.

It should be noted that the NMF algorithm used in MutSpec is not expected to give identical results to that used by Alexandrov et al. which uses a complex sequence of pre-filtering and bootstrapping in addition to the NMF algorithm itself. However, certain features of the algorithm used by Alexandrov et al. do not scale with sample size (in particular setting all counts of less than 5 to zero) and are therefore not well adapted to the small sample sizes found in experimental studies or small to medium scale analyses for which MutSpec is designed.

Performance

MutSpec may be used to analyse data from whole-exome, whole-genome or targeted sequencing experiments performed on human or mouse genomes. The tool manages CPU usage to optimize performance in term of analysis time. For example, to annotate a file with less than 5000 variants it will use one CPU, while for a file with more than 100,000 variants it will use the maximal capacity allowed by the Galaxy server administrator. For such a file, using 24 CPU will take about 7 min, using 8 cores will take about 14 min to annotate while it would take 4 h with only one CPU. Computation of the statistics for estimating the NMF factorisation value (option in MutSpec-Stat) and for running MutSpec-NMF is also time consuming when large number of samples are analysed. Here the tool will use the maximum allowed CPU capacity.

There is no limit in the number of samples or mutation per sample that can be analysed. However, the capacity to open Excel files with large number of datasheets (over 500) will depend on user's computer settings and performance (ie, a file with 530 samples takes 18 s to open up on a computer with 2 GB of RAM). This limitation can be overcome by adapting the design of the analysis workflow to limit the number of samples included in one Excel file. All graphical outputs generated by the tools can be downloaded as high resolution images suitable for publication (300 dpi). MutSpec toolbox is well documented with short descriptions of input and output formats and options for each tool. While the methods for defining and extracting mutation signatures implemented in MutSpec correspond to current standards well accepted in the field, we will provide package updates and upgrades reflecting progresses in the field, such as new format or definition for mutation signatures.

Conclusions

MutSpec offers an easy-to-use framework for variant annotation and statistical analyses of mutation patterns from genome-wide data obtained from deep sequencing experiments. It is based on the Galaxy open-source framework that offers a powerful management system for reproducible bioinformatics analyses. MutSpec accepts the standard VCF format as input as well as any list of variants in tab-delimited text format and implements established methods. The example analysis presented here illustrates the straightforward workflow and the richness of the statistics and publication-grade graphics produced by the tool. MutSpec is versatile as data from both human and mouse and from different genome builds can be analysed for easy comparison of human and experimental data. To our knowledge, MutSpec is the only tool available as a graphical interface to researchers with no computer programing skills or higher level of bioinformatics expertise in the analysis of mutation signatures present in cancer genomes. Given the positive feedbacks from test users, we believe that MutSpec can be a very useful tool for a large community in the field of genomics, namely investigators interested in interpreting the mutational processes that contribute to human carcinogenesis.

Availability and requirements

Project name: MutSpec

Availability: MutSpec package in the Galaxy toolshed at https://toolshed.g2.bx.psu.edu/

Operating system(s): Linux.

Programming language: Perl (version 5.18.1), R (version 3.1.2), XML, HTML

Other requirements: Galaxy, ANNOVAR

License: GPLv2

Additional files

Additional file 1: Example of Excel file generated by MutSpec-Stat tool. The data in this file correspond to the example analysis of OSCC described herein. (XLS 37674 kb)

Additional file 2: Example of NMF analysis with combined matrices from different analyses. Matrices from two different analyses may be combined in a single matrix to analyse samples from analysis 1 and 2 together. This matrix should contain a header with sample IDs and have 96 rows describing the 6 SBS types in their sequence context. The matrix should be formatted as tab-delimited text to be accepted as input of MutSpec-NMF. (PPT 348 kb)

Additional file 3: Screenshots of MutSpec tools inputs and outputs in Galaxy. (PPT 2468 kb)

Abbreviations

COSMIC: Catalogue of Somatic Mutations in Cancer; CPU: Central Processing Unit; ESP: Exome Sequencing Project; ICGC: International Cancer Genome Consortium; NMF: Non-negative Matrix Factorisation; OSCC: Oral Squamous

Cell Carcinomas; SBS: Single Base Substitution; SNP: Single-Nucleotide Polymorphism; TCGA: The Cancer Genome Atlas; VCF: Variant Call Format.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MA designed and developed the software, and wrote the draft manuscript. MO designed and supervised the software development, ran the example analysis and wrote the draft manuscript. VC implemented the various scripts into Galaxy. GB and LB provided statistical expertise on the computed statistics. JZ, XC and ZH participated in the design of the software and revised the manuscript. All authors have read and approved the final manuscript.

Acknowledgements
This project was supported by the International Agency for Research on Cancer budget. We thank Estelle Chanudet and Catherine Voegele for testing the software and providing useful feedbacks.

Author details
[1]Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, F69372 Lyon, France. [2]Epigenetic Group, International Agency for Research on Cancer, F69372 Lyon, France. [3]Biostatistics Group, International Agency for Research on Cancer, F69372 Lyon, France.

Received: 27 August 2015 Accepted: 4 April 2016
Published online: 18 April 2016

References

1. 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.
2. Olivier M, Hussain SP, de FC C, Hainaut P, Harris CC. TP53 mutation spectra and load: a tool for generating hypotheses on the etiology of cancer. IARC Sci Publ. 2004;157:247–70.
3. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013;3(1):246–59.
4. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014;24:52–60.
5. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014;15(9):585–98.
6. Nik-Zainal S, Alexandrov LB, Wedge DC, Van LP, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012; 149(5):979–93.
7. Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, et al. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. Genome Res. 2014;24(10):1624–36.
8. Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallee MP, et al. Modelling mutational landscapes of human cancers in vitro. Sci Rep. 2014;4:4482.
9. Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. Sci Transl Med. 2013;5(197):197ra101.
10. Severson PL, Vrba L, Stampfer MR, Futscher BW. Exome-wide mutation profile in benzo[a]pyrene-derived post-stasis and immortal human mammary epithelial cells. Mutat Res Genet Toxicol Environ Mutagen. 2014;775-776:48–54.
11. Segovia R, Tam AS, Stirling PC. Dissecting genetic and environmental mutation signatures with model organisms. Trends Genet. 2015;31(8):465–74.
12. Westcott PM, Halliwill KD, To MD, Rashid M, Rust AG, Keane TM, et al. The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. Nature. 2015;517(7535):489–92.
13. Nassar D, Latil M, Boeckx B, Lambrechts D, Blanpain C. Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. Nat Med. 2015;21(8):946–54.
14. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.
15. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15(10):1451–5.
16. Blankenberg D, Von KG, Coraor N, Ananda G, Lazarus R, Mangan M, et al. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. 2010;Chapter 19:Unit-21.
17. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
18. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010;11:367.
19. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004;101(12):4164–9.
20. Gunther F, Dudschig C, Kaup B. LSAfun–An R package for computations based on Latent Semantic Analysis. Behav Res Methods. 2015;47(4):930–44.
21. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43(Database issue):D805–11.
22. Maitra A, Biswas NK, Amin K, Kowtal P, Kumar S, Das S, et al. Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. Nat Commun. 2013;4:2873.
23. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat Genet. 2013;45(9):977–83.

Objective 2

Paper 2: Modelling mutational landscapes of human cancers *in vitro*

Magali Olivier, Annette Weninger, **Maude Ardin**, Hana Huskova, Xavier Castells, Maxime P. Vallée, James McKay, Tatiana Nedelko, Karl-Rudolf Muehlbauer, Hiroyuki Marusawa, John Alexander, Lee Hazelwood, Graham Byrnes, Monica Hollstein and Jiri Zavadil

Scientific Reports, 2014

Aim Devise an experimental model of *in vitro* mutagenesis to investigate mutational signatures of human carcinogens by exome sequencing.

Approach Exposing primary murine embryonic fibroblasts (MEF) cells to specific human carcinogens produces immortalised clones carrying mutations specific of the exposure that can be studied with WES and dedicated bioinformatic tools.

Graphical summary

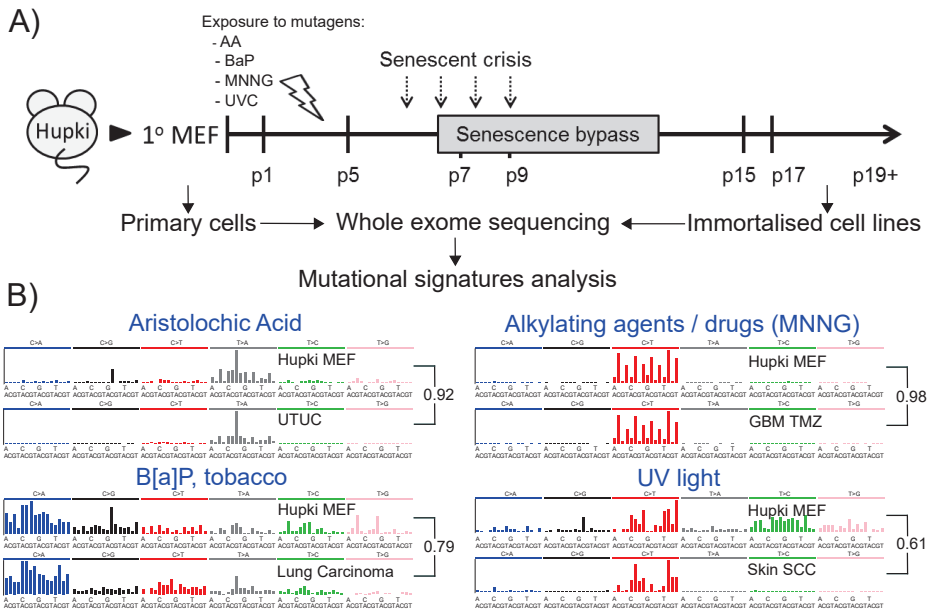


Figure 16: Overview of MEF system and carcinogens mutational signatures. A) Protocol used for MEFs immortalisation following carcinogens exposure to obtain cell lines for WES. B) Mutational signatures of four human carcinogens identified in MEFs and human tumours.
B[a]P: Benzo[a]pyrene, GBM TMZ: glioblastoma treated with temozolomide, MNNG: N-methyl-N'-nitro-N-nitrosoguanidine, Skin SCC: skin squamous cell carcinomas, UTUC: upper tract urothelial carcinoma

Novelty and highlights

- The MEF clonal immortalisation system is able to reproduce mutational signatures observed in human tumours.
- This system is a cost-effective approach for investigating mutational signatures of carcinogens.
- The assay can provide mechanistic evidence on the involvement of potential carcinogens in human tumour development.

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
EXPERIMENTAL MODELS OF DISEASE
CANCER MODELS
CANCER GENOMICS

Received
10 January 2014

Accepted
5 March 2014

Published
27 March 2014

Correspondence and requests for materials should be addressed to M.H. (M.Hollstein@leeds.ac.uk) or J.Z. (zavadilj@iarc.fr)

Modelling mutational landscapes of human cancers *in vitro*

Magali Olivier[1], Annette Weninger[2], Maude Ardin[1], Hana Huskova[1], Xavier Castells[1], Maxime P. Vallée[3], James McKay[3], Tatiana Nedelko[2], Karl-Rudolf Muehlbauer[2], Hiroyuki Marusawa[4], John Alexander[5], Lee Hazelwood[5], Graham Byrnes[6], Monica Hollstein[2,5] & Jiri Zavadil[1]

[1]Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, 69008 Lyon, France, [2]German Cancer Research Center (Deutsches Krebsforschungszentrum), D69120 Heidelberg, Germany, [3]Genetic Cancer Susceptibility Group, International Agency for Research on Cancer, 69008 Lyon, France, [4]Department of Gastroenterology and Hepatology, Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan, [5]Faculty of Medicine and Health, University of Leeds, Leeds LS2 9JT, United Kingdom, [6]Biostatistics Group, International Agency for Research on Cancer, 69008 Lyon, France.

Experimental models that recapitulate mutational landscapes of human cancers are needed to decipher the rapidly expanding data on human somatic mutations. We demonstrate that mutation patterns in immortalised cell lines derived from primary murine embryonic fibroblasts (MEFs) exposed *in vitro* to carcinogens recapitulate key features of mutational signatures observed in human cancers. In experiments with several cancer-causing agents we obtained high genome-wide concordance between human tumour mutation data and *in vitro* data with respect to predominant substitution types, strand bias and sequence context. Moreover, we found signature mutations in well-studied human cancer driver genes. To explore endogenous mutagenesis, we used MEFs ectopically expressing activation-induced cytidine deaminase (AID) and observed an excess of AID signature mutations in immortalised cell lines compared to their non-transgenic counterparts. MEF immortalisation is thus a simple and powerful strategy for modelling cancer mutation landscapes that facilitates the interpretation of human tumour genome-wide sequencing data.

High-throughput sequencing has shown that cancer genomes are riddled with somatic alterations, with numerous tumour types harbouring hundreds to thousands of mutations. Bioinformatic analyses of this massive catalogue in search of recurring mutations have substantially contributed to the identification of most of the genes functionally impaired in cancer development[1]. The genome-wide human cancer sequencing data also provide a powerful resource for investigating the nature of mutagenic insults that give rise to mutations in human population[2,3]. However, what is lacking in order to make optimal use of this resource for such a purpose is a suitable database of experimentally induced mutations to test inferences that come from inspecting human mutation patterns. It is well known that mutagenic factors, whether chemical or enzymatic, mutate DNA in characteristic ways, thereby revealing clues to their identities. This principle was elegantly demonstrated decades before the advent of new sequencing technologies in a wide variety of assays[4,5]. These assays, however, had one feature in common that limited their scope. Mutations were typically scored in a single gene (allowing clonal selection), or at best, in a discrete number of specific genes. Although data from such experiments have been fundamentally important to biology, the tests were not designed to recapitulate or interpret the more complex mutation profiles generated from genome-wide data. Genome-wide sequencing of cells exposed to sources of mutation in a controlled fashion will now allow more comprehensive experimental investigations of the mutagenic activities of human carcinogens.

This study aims to determine whether a simple experimental system using *in vitro* immortalisation of normal mammalian cells would generate genome-wide mutation data relevant to human tumours. Immortalisation of primary cells, notably murine embryonic fibroblasts (MEFs), has been used extensively as a powerful *in vitro* model for exploring genetic control of cellular homeostasis and its disruption in disease[6]. Recent research in this area has shown that various molecular pathways that control cellular senescence and become circumvented *in vitro* to allow cell immortalisation are cancer gene pathways, including oncogenes and tumour suppressor genes known to be mutated in human cancer[7–9]. Encouragingly, we showed in previous work that when carcinogens are applied to Hupki MEFs (MEFs carrying normal human p53 sequences embedded in the *Trp53* gene, human p53 knock-in) prior to senescence, emerging clonal cell lines harbour *TP53* gene signature mutations characteristic of

the carcinogens and consistent with *TP53* mutations detected in human cancers from exposed patient cohorts[10–12].

In the present study, we expanded this approach by assessing exome-wide mutation patterns in the Hupki MEF immortalisation assay (**Supplementary Fig. 1**). We sequenced the exome of immortalised MEF lines established from primary cultures exposed to well-known carcinogens and compared the mutation profiles obtained in these assays to those observed in genome-wide data from human tumours with related aetiologies. Mutation signatures derived from these assays were also compared to currently known signatures in human cancers[2]. Since the MEF *in vitro* immortalisation process has parallels with the conversion of normal cells to tumour cells *in vivo*, we also investigated alterations in specific driver genes that could provide mechanistic clues to molecular events governing senescence bypass and immortalisation. Finally, in a proof-of-concept experiment devised to explore an endogenous process proposed to contribute to the human mutation load, we examined the effect of activation-induced cytidine deaminase transgene (AID-Tg) expression on the pattern of base substitutions that accumulate during MEF immortalisation[13].

Results

Genome-wide mutation spectra from immortalised MEF cell lines. Genomic DNA isolated from primary MEFs and from immortalised cell lines derived from MEF cultures exposed to aristolochic acid (AA), ultraviolet light subclass C (UVC), the alkylating agent N-methyl-N'-nitro-N-nitrosoguanidine (MNNG), the tobacco mutagen benzo(a)pyrene (BaP), or unexposed cultures (see **Supplementary Table 1**), were subjected to genome-wide mutation profiling by whole-exome sequencing (WES) (see **Supplementary Fig. 1** for assay overview and Methods for WES data processing and analysis). Two cell lines per exposure category were investigated. As shown in **Fig. 1**, the patterns of mutations found in the MEF cell lines were in marked concordance with those observed in human tumours with aetiologies related to the mutagens tested, and were as expected from previous knowledge on the mutagenic properties of these particular exposures. In the cell lines derived from AA-exposed cultures, the most frequent type of mutation was A:T > T:A as in urinary tract urothelial cancers (UTUC) from AA exposed patients (**Fig. 1a**), and a significant strand-bias towards the non-transcribed strand was observed for A > T (**Table 1**), in keeping with previous reports on human UTUC from AA-exposed patients[14,15]. In cell lines from BaP-exposed cultures, the most frequent type of mutation was C:G > A:T with a strand bias towards the non-transcribed strand for G > T as is seen in lung cancers (Lung_Ca) from heavy smokers (**Fig. 1b** and **Table 1**). In cell lines from MNNG-exposed cells, the most frequent mutation type was C:G > T:A with no significant strand-bias (**Fig. 1c** and **Table 1**), consistent with the alkylating properties of this agent and with the pattern observed in brain tumours from patients treated with the alkylating agent temozolomide. In the cell lines from UVC-exposed cultures, the most frequent type of mutation was C:G > T:A, as in skin squamous cell carcinomas (Skin_SCC), and with a strand-bias of borderline significance (**Fig. 1d** and **Table 1**).

In addition to exogenous exposures, we assessed the effect of an endogenous mutagenic process by analysing immortalised MEFs harbouring a transgene expressing activation-induced cytidine deaminase (AID) (see Methods). In these cell lines, referred to as HxAID-Tg, a predominance of C:G > T:A transitions was observed (**Fig. 1e**), as expected from experimental studies on the mutagenic properties of AID[13,16]. Finally, four immortalised MEF cell lines from untreated cultures were analysed to determine underlying mutagenesis in this model. Interestingly, the most predominant mutation type was C:G > G:C (**Fig. 1f**) in all four cell lines, as has been observed previously in the *Trp53* gene of immortalised MEFs[17].

The sequence context of mutations is an important feature of mutation patterns because many mutagenic agents and processes exhibit a preferred base context. We analysed the 5′ and 3′ base context of mutations in all conditions described above. As shown in **Fig. 1**, a previously described preferred sequence context for each specific exposure was recapitulated in the MEF assay. Indeed, A:T > T:A mutations occurred predominantly within a 5′-C**A**G-3′ motif, as in the selected human set (**Fig. 1g**) and as reported in other published series[14,15,18–20]. For BaP exposure, C > A mutations occurred most frequently in 5′-C**C**N-3′ triplets (corresponding to 5′-N**G**G-3′ for the complementary G > T), as in the human lung tumour dataset (**Fig. 1h**). In cell lines from MNNG cultures, C > T transitions with a C or T in 3′ and any base in 5′ were the most frequent (corresponding to 5′-(G/A)**G**N-3′ for the complementary G > A mutations), observed also in recurrent glioblastomas of temozolomide-treated patients (**Fig. 1i**). These mutations occurred mainly outside CpG sites as expected. In the cell lines derived from UVC treated MEF cultures, C > T changes within a 5′-(C/T)**C**N-3′ motif were the major events as seen in human skin SCC (**Fig. 1j**). This context is expected from the published literature on UV mutagenesis, which describes the highly characteristic alterations at pyrimidine dimers induced by UV exposure. Interestingly, the frequent C > G mutations found in the spontaneous lines showed a preferred sequence context for 5′-G**C**C-3′ a signature that was also present, although much less prominently, in most of the other cell lines (**Fig. 1l**). Finally, in the spontaneously immortalised lines from HxAID-Tg MEFs (**Fig. 1k**) the predominant C > T changes were most frequently observed in a 5′-G**C**(A/C/T)-3′ sequence context, followed by 5′-A**C**(A/C/T)-3′. These findings match the preferred contexts previously demonstrated for AID activity[21,22]. The most frequent single base substitutions (SBS) observed in human cancers and in mammalian evolution are C > T transitions at 5′-N**C**G-3′ sites (CpGs). These mutations occur following spontaneous deamination of 5-methyl-cytosines and result in C > T transitions[23]. In the cell lines analysed here, CpGs accounted for 25–30% of the C > T mutations in the cell lines derived from AA, BaP and untreated MEF cultures, but for less than 15% of C > T mutations in the MNNG, UVC or HxAID-Tg cell lines, which is consistent with the treatment-specific sequence context of C > T transitions in these latter cell lines.

Two types of statistical analyses were then applied to these data (**Fig. 2**). Firstly, principal component analysis (PCA) was performed to assess whether global mutation patterns obtained in the MEF immortalisation assays can distinguish between cell lines obtained from different treatments/conditions. Using percent frequency values of the six mutation types in their triplet sequence context (amounting to a total of 96 variables), the two first components were able to discriminate the replicate cell lines according to each specific treatment condition (**Fig. 2a**). When including the human cancer datasets in the exposure model, we observed a good concordance with the mouse datasets for most conditions, with the exception of the UVC treatment which showed a broader confidence interval (**Fig. 2b**). Secondly, the method used by Alexandrov *et al.*[24], to extract signatures was adapted and applied to the 14 cell line data (see Methods). Although this method is optimized for large datasets, it could identify six signatures that corresponded to the six experimental conditions (**Supplementary Fig. 2**) and were concordant with the mutation patterns shown in **Fig. 1**. The comparison of the MEF experimental signatures with the 27 human-cancer derived signatures reported by Alexandrov *et al.*[2], showed high similarity between the MNNG signature and Signature 11 (temolozomide), similarity between the BaP signature and Signature 4 (smoking), and between AID signature and Signature 19 (not identified) (**Fig. 2c**). No similarity was found for the AA signature (patients with AA-associated tumours were not analysed by Alexandrov *et al.*), or for the signature observed in the spontaneous immortalised cell lines.

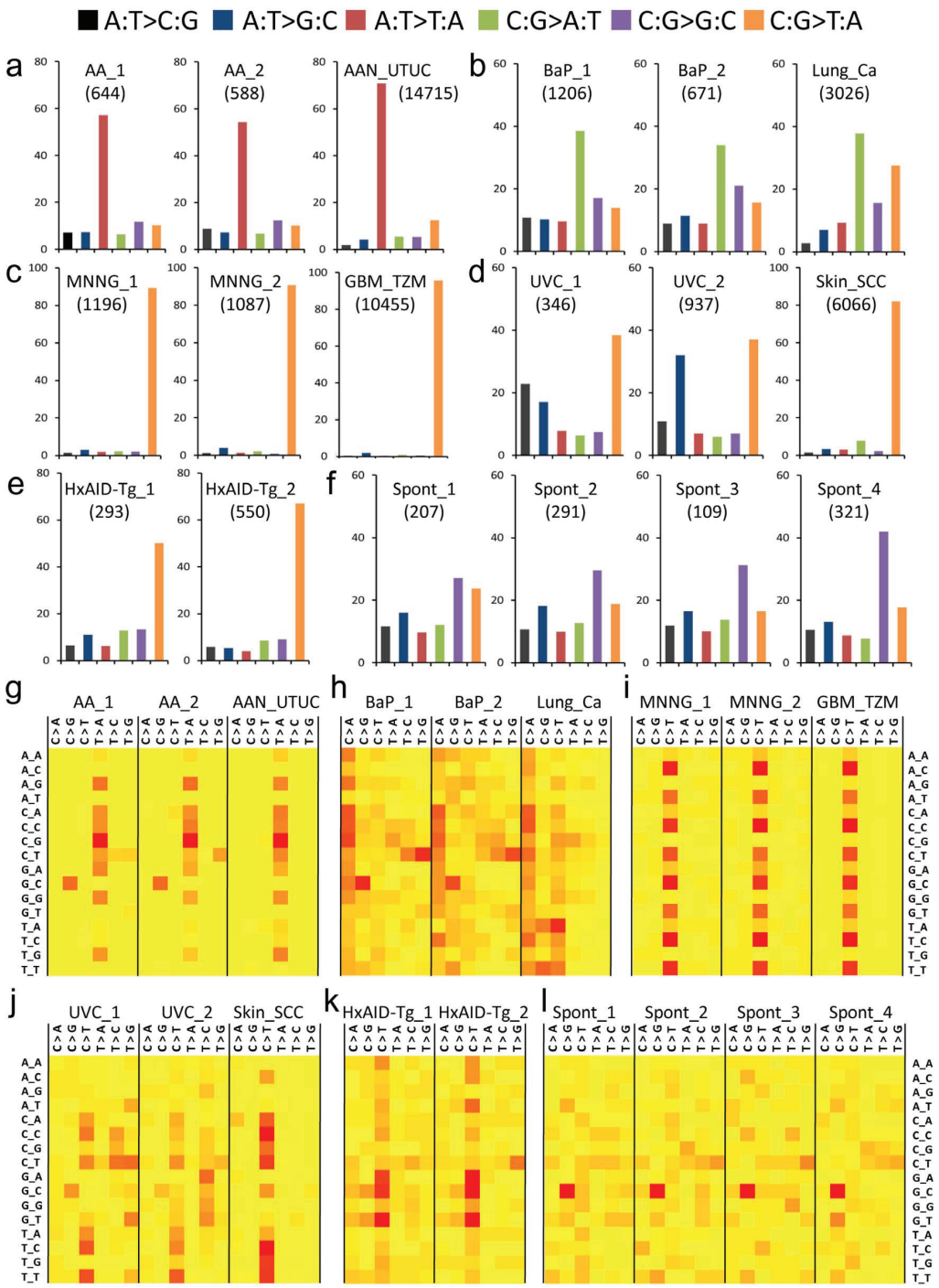


Figure 1 | Mutation patterns derived from exome data obtained from MEF immortalised cell lines. Mutation type distributions (a–f) and sequence context (g–l) of single base substitutions. For each treatment condition, data are shown for two independent immortalised cell lines and for a set of human tumours related to the tested condition. In (a–f), the percentage of each substitution type is shown with the total number of mutations indicated in parentheses. In (g–l), heat maps of mutation sequence context are shown. The percentage of each substitution type within a triplet sequence context is colour-coded according to the percent values. Highly abundant mutations are represented in red and low abundance mutations are in yellow. (a,g) Aristolochic-acid treatment (two left panels) and upper urinary tract human tumours (right panel) from patients exposed to AA. (b,h) Benzo(a)pyrene treatment (two left panels) and lung adenocarcinomas (right panel) from heavy smokers. (c,i) N-methyl-N'-nitro-N-nitrosoguanidine treatment (two left panels) and human recurrent glioblastoma treated with temozolomide (right panel). (d,j) UVC treatment (two left panels) and human skin squamous cell carcinomas (right panel) (COSMIC v65). (e,k) AID transgene (two panels). (f,l) Data from four independent cell lines obtained by spontaneous immortalisation of the Hupki MEF primary cells (Spont, no treatment).

Table 1 | Significance of the mutation strand bias for all mutation types in each experimental condition (ratio of the number of mutation on the non-transcribed to transcribed strand and FDR q-values for significance)

SBS type	Condition	AA	BaP	MNNG	UVC	HxAID-Tg	Spont
	Ratio	1.0	1.1	1.1	1.0	1.0	1.4
A:T > C:G	FDR-q	1.000	0.881	1.000	1.000	1.000	0.277
	Significance	-	-	-	-	-	-
	Ratio	1.3	1.0	1.0	1.0	0.8	0.9
A:T > G:C	FDR-q	0.591	1.000	1.000	1.000	0.805	1.000
	Significance	-	-	-	-	-	-
	Ratio	2.1	1.5	0.5	1.1	1.1	1.6
A:T > T:A	FDR-q	<10–18	0.038	0.277	1.000	1.000	0.147
	Significance	***	*	-	-	-	-
	Ratio	0.9	0.6	0.7	0.8	1.0	1.0
C:G > A:T	FDR-q	1.000	<10–9	0.707	0.796	1.000	1.000
	Significance	-	***	-	-	-	-
	Ratio	1.0	0.9	0.7	0.8	1.1	1.0
C:G > G:C	FDR-q	1.000	0.805	0.796	0.796	0.948	1.000
	Significance	-	-	-	-	-	-
	Ratio	2.2	1.1	1.0	1.3	1.0	1.1
C:G > T:A	FDR-q	<10–3	0.805	0.796	0.026	1.000	0.796
	Significance	***	-	-	*	-	-

Significance: * $q < 0.05$; ** $q < 0.01$; *** $q < 0.001$.

These results show that mutation patterns obtained in the MEF immortalisation assay are specific to the exposure and can reveal carcinogen-specific signatures that are relevant to human cancers.

Driver gene mutation status in immortalised MEF cell lines. The cell lines from carcinogen exposure experiments chosen here for WES studies harbour *TP53* mutations that arose during immortalisation of the primary cells. To investigate whether other cancer driver genes were recurrently affected during the senescence bypass/immortalisation process, we analysed the mutation status of other established or putative cancer drivers, including all those defined as oncogenes or tumour suppressor genes according to the "20/20 rule" formulated by Vogelstein *et al.*[1], as well as genes encoding regulators of the epigenome that have been described as a newly emerging class of cancer driver genes[1,25,26]. Non-synonymous and truncating mutations found in these selected driver genes are detailed in **Supplementary Dataset 1** and graphically summarized in **Supplementary Dataset 2**. A number of genes in these functional classes were found altered by mutations characteristic of the exposure that cells underwent prior to immortalisation. Although most genes were mutated only in one line, the *Ep400*, *Dnmt1*, *Kdm6b*, *Kmt2d*, *Arid1b* and *Arid2* genes were mutated in at least two lines. *Ep400* and *Kmt2d* in particular were mutated in four cell lines. *Ep400*, a regulator of cellular senescence within the p53-p21 axis[27], was affected by a truncating mutation in one line, and *Kmt2d* carried three mutations in important functional domains. The most unequivocal driver gene mutations were two activating Ras missense mutations, highly recurrent in human cancers: the (c.A182T/p.Q61L) *Hras1* mutation in one AA cell line corresponding to the *HRAS* mutation previously associated with exposure to AA in humans and animal models[14,18,28–30], and an activating mutation in the *Kras* oncogene (c.A182G/p.Q61R) identified in one of the UVC lines. Overall, these observations suggest that the MEF immortalisation assay captures and selects for driver gene mutations relevant to cancer biology, and are in keeping with extensive literature on the impact of cancer-related genes on senescence bypass, immortalisation and transformation of MEFs[9,31].

Discussion

In this report we show that mutations acquired in MEFs during establishment in culture and studied at the exome level reveal patterns relevant to human cancers. While the MEF immortalisation assay protocol has been shown previously to recapitulate *TP53* mutation patterns in the context of specific carcinogen exposures[10,11], we demonstrate here that this assay is highly suitable as a selection strategy to obtain a cell population harbouring a suite of base substitutions relevant to exome-wide mutation data derived from human cancers. In principle, one single immortalised cell line provides information to identify a mutation signature, whereas many cell lines would be necessary when interrogating a single gene such as *TP53*. In practice, of course, WES on multiple cell line replicates per exposure or condition is warranted and will be called for in extended studies in the future to generate highly robust mutation signatures. The scope of overlap between mutation patterns in human datasets and immortalised MEF lines includes: (a) the global distribution of mutation types, (b) the accumulation of mutations on the non-transcribed strand (strand bias) for treatments with carcinogens known to elicit transcription-coupled DNA repair, and (c) the sequence context of the dominant mutation type. Thus, using four carcinogens with well-known mutagenic properties, the predominant mutation signatures we found with this model for the four tested carcinogens were the ones expected for these mutagenic agents. Although we analysed only two cell lines for each carcinogen, the expected signatures were evident in single cell lines and were highly reproducible between the two cell lines. The human tumour datasets used for comparison with our *in vitro* data were selected from publicly available data and our selection was based on whether the suspected aetiologies of the tumour sets were linked to the carcinogen tested in MEFs. The most striking matching condition was the AA treatment. In both human and *in vitro* MEF data, over 50% of mutations were A > T transversions, with a significant strand bias of 2 : 1 and a sequence context dominated by 5′-CAG-3′. The aetiology of the tumours included in the human set has been clearly associated with the AA exposure[14]. Since AA is a potent carcinogen that mainly causes A > T transversions, enrichment of these somatic mutations in exposed individuals is likely to reflect the insult of AA exposure. The AA signature has not been found in any other cancer type so far[32]. In the case of the *in vitro* WES data from cell lines arising from MNNG-exposed cultures, the human set chosen for comparison consisted of patients treated with the drug temozolomide, which, like MNNG, is an alkylating agent. The global mutation type distribution was strikingly similar between the mouse and human data and very distinct from primary tumours of the same type but not exposed to temozolomide (**Supplementary Fig. 3**). The MNNG signature was

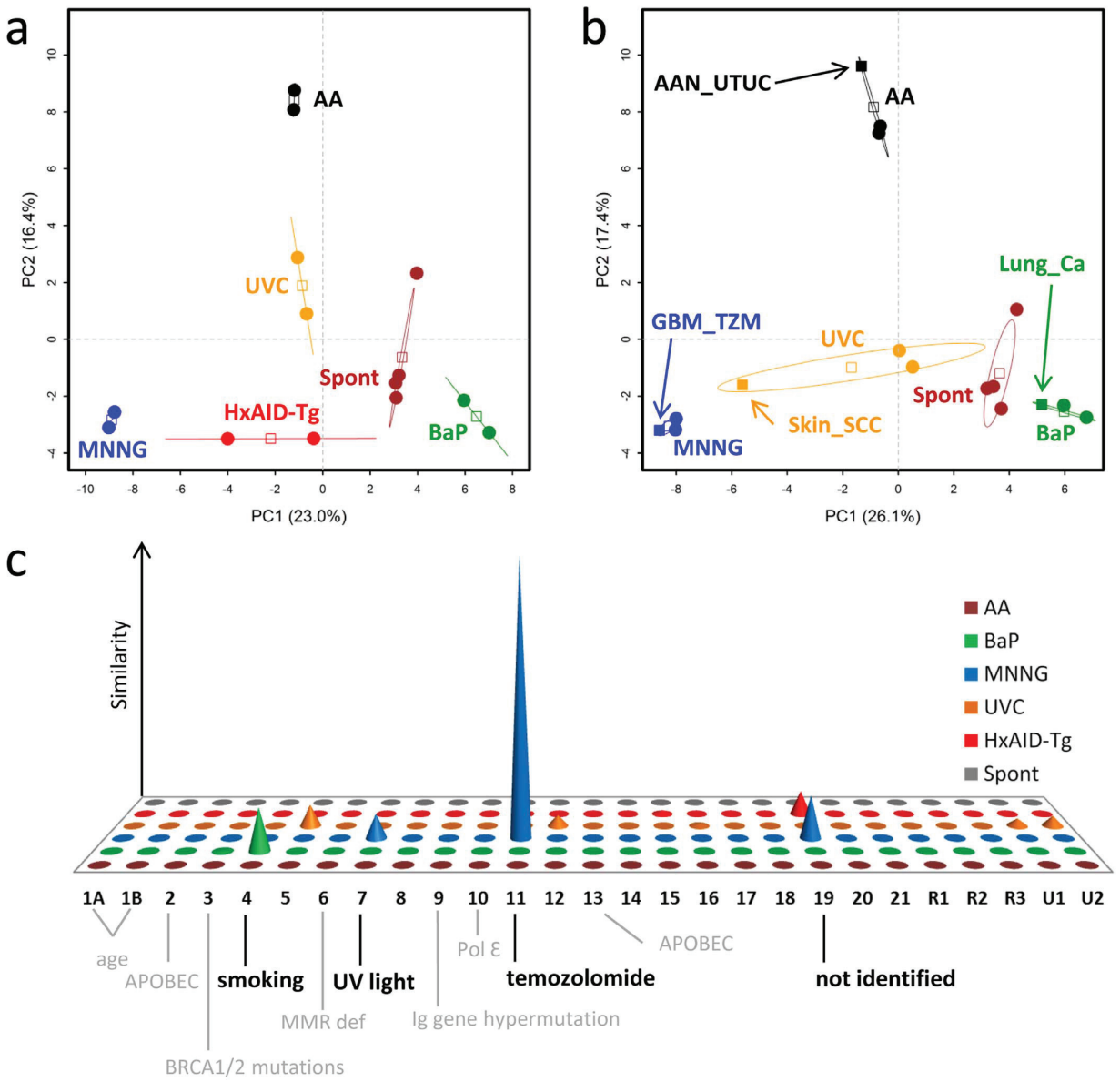


Figure 2 | Analysis of mutation signatures derived from exome data obtained from MEF immortalised cell lines. (a) Principal component analysis (PCA) of WES data using mutation signatures. PCA was computed using as input the frequency matrices of sequence context mutations (96 variables) from cell lines immortalised following exposure of primary Hupki MEFs to a carcinogen (AA, BaP, MNNG or UVC), from Hupki MEFs carrying the AID transgene (HxAID-Tg) or from Hupki MEF-derived cell lines that immortalised spontaneously (Spont). Each sample is plotted considering the value of the first and second principal components (PC1 and PC2). The percentage of variance explained by each component is indicated within brackets in each axis. A 95% confidence ellipsis is drawn for each experimental condition and the empty squares indicate the respective centre of gravity. Cells and samples are represented by round and squared solid symbols, respectively. (b) Same as in (a) but with the human tumour datasets (same as shown in Figure 1) added to the input and labelled by arrows. HxAID-Tg samples are omitted in (b) as no corresponding relevant tumour data were identified. AAN_UTUC, aristolochic acid nephropathy-related upper urinary tract urothelial carcinoma; GBM_TZM, glioblastoma after temozolomide treatment; Lung_Ca, lung carcinoma; Skin_SCC, skin squamous cell carcinoma. (c) Graphical representation of the similarity distance of each of the six MEF signatures (front-back axis) to each of the 27 human cancer signatures[2] (horizontal axis, 1A through U2). The vertical axis measures the similarity of signatures between the two systems, expressed as negative log(tan(angle), see Methods. Negative values below the x-z plane correspond to angles >45° and represent dissimilarity, and are thus not shown.

also very similar to the temozolomide signature derived from another set of temozolomide-exposed patients reported previously[2]. This signature is thus highly specific for alkylating agents such as MNNG and temozolomide. With respect to other exposures, it is clear that human tumour development typically involves various mutational mixtures and selection processes, resulting in a complex picture of mutation signatures. These considerations may explain why the tumour data from lungs of heavy smokers differ from the

in vitro data from BaP exposure with respect to the less prominent mutation types. Tobacco smoke contains a highly complex mixture of carcinogens. Nevertheless, the BaP signature derived from the *in vitro* assay exhibited similarity with the human tumour-derived smoking signature reported previously[2], suggesting that BaP and possibly other smoke components that have similar mutagenic properties constitute one of the main carcinogenic insults responsible for the smoking signature observed in human tumours. The dataset from tumours associated with UV exposure was the most distant from the signature obtained *in vitro*, although the expected C > T mutations within a 5′-(C/T)CN-3′ context were prominent in both sets. There are several possible explanations, such as the technical aspects of the experimental procedure *in vitro*, or the mutagenic activities of sunlight compared to UVC alone. These results are reflected in the principal component analysis that showed the closest relationship between human and MEF data for AA and BaP and a more distant relationship for UVC. In addition to exogenous exposures, the spontaneous decay of DNA is a well-known cause of the human mutation load[33] and the deregulation of endogenous cellular enzymes that accelerate the accumulation of sequence changes is becoming of increasing interest to cancer biologists. It is a considerable challenge, however, to determine the relative contributions of different DNA metabolism pathways to genetic alterations observed in human cancers, and to understand the factors that may result in the deregulation of normal processes governing DNA integrity. Recently the APOBEC/AID families of cytidine deaminases have come under scrutiny because of their potential roles in cancer as endogenous sources of mutation in various cancer types[2,34,35]. AID, which is normally expressed in B-lymphocytes, has been proposed as a possible source of mutagenic activity in the development of various inflammation-associated cancer types when expressed inappropriately[36,37]. An early investigation on mutation patterns produced by AID in a single reporter gene showed a strong C > T mutation signature as anticipated[16], and there are now many studies exploring the impact of ectopic AID expression on cancer development[13]. Here we compared the sequence changes during immortalisation of MEFs harbouring a constitutively active AID transgene with MEFs that did not carry the transgene. The AID signature mutation was easily captured by this strategy, providing a proof-of-concept demonstration of the applicability of this approach to investigating endogenous mutagenesis. Interestingly, the *in vitro* AID signature showed some similarity with one of the signatures found in pilocytic astrocytoma[2,34,35], but was not represented in other cancer types and had no similarity to two previously reported APOBEC signatures[2,34,35]. The full role of AID in shaping mutation patterns in humans remains to be investigated.

Surprisingly, the analysis of spontaneously immortalised cell lines from untreated cultures showed a strikingly high frequency of C > G mutations in the 5′-GCC-3′ sequence context, also present (albeit at lower and variable frequency) in all cell lines (**Fig. 1** and **Supplementary Fig. 2**). A high frequency of C > G mutations in the *p53* gene was observed previously in both Hupki MEFs (with human *TP53* sequences) and MEFs with the murine *Trp53* gene[17]. Our WES results show that this phenomenon is global and may be due to specific mutagenic pressures inherent to the experimental conditions. Gene Ontology Biological Process analysis of genes affected by non-synonymous C > G mutations in the spontaneously immortalised cell lines identified 11 genes involved in regulation of apoptosis/programmed cell death (GO 0042981; GO 0043067) as high scoring categories (enrichment p-value < 0.05, Fisher's exact test, **Supplementary Table 2**), a finding consistent with the cultures overcoming senescence, and with individual cells acquiring immortalised properties. This signature did not show any similarity to those reported by Alexandrov *et al.*[2]. Although C > G mutations have been associated with two signatures linked to APOBEC activity in human cancers[2], they occur in a different sequence context of 5′-TCA-3′. A recent genome-wide analysis of gingivo-buccal oral SCC from Indian patients reported a high frequency of C > G mutations (although the sequence context was not reported) in three tobacco users carrying a high mutation load in their tumours[38]. These authors proposed that the unexpectedly high numbers of C > G mutations in their sample set may be caused by oxidative damage. The DNA lesion 8-oxoguanine caused by exposure to reactive oxygen species[39,40] can lead to this transversion. The elevated numbers of C > G substitutions in spontaneously immortalising MEFs may be a cell culture artefact caused by high oxygen levels of standard incubation conditions, and culturing the cells at physiological levels of oxygen can test this premise. Further investigation of the origin of these C > G substitutions in the MEF *in vitro* assay is warranted.

Although the number of cell lines analysed in the present study is limited, we identified several recurrently mutated genes among oncogenes and tumour suppressor genes classified as cancer drivers, or regulators of the epigenome, an emerging new class of potential driver genes. We note, however, that most of these mutations are likely to be passenger events occurring in the MEF immortalisation/transformation process, analogous to observations in human cancers. Interestingly, while mutations in all categories of genes but histone genes were observed in the carcinogen-exposed lines, the HxAID-Tg cell lines accumulated mutations mainly in histone genes (**Supplementary Datasets 1** and **2**). A study to explore the reasons for this observation will require larger numbers of immortalised cell lines both with and without the AID transgene. The p53 status of emerging immortalised cells may also influence the subset of target genes subsequently mutated and selected for, but again, to address this speculation properly, an extensive set of cell line replicates will be needed.

Exome-wide analysis of MEFs thus joins epigenetic profiling and senescence bypass screens in the modern assembly of *in vitro* tools to elucidate cancer biology[9,41]. Analysis of more cell lines and detailed functional analyses of the specific mutations will be important in order to distinguish driver from passenger mutations in WES-MEF studies in a robust, statistically sound manner. Analysis of indels will be considered in future studies with more cell line replicates as the number of indels called in the current sample set was too small to derive meaningful interpretations of how indels might contribute to particular mutational signatures. The analyses of indels might also provide a more complete picture of mutations in tumour suppressor genes as this type of cancer gene is more often altered by indels. The present study is limited to a small number of cell lines, and these have acquired typical human tumour *TP53* mutations during immortalisation. The *p53* gene mutation is the most common specific alteration known to drive senescence bypass and immortalisation of MEFs[42,43]. It will be interesting to investigate to what extent the p53 status influences global mutation patterns and the subset of mutated driver genes by comparing WES data from cell lines retaining the wild-type *p53* gene sequence with cell lines that have acquired *p53* mutations typical of human tumours.

In summary, the present study demonstrates the potential of the MEF immortalisation assay to reveal mutation signatures of human carcinogens. Although the use of mouse cells can be seen as a limitation because of differences in metabolism and DNA repair between humans and mice, *in vitro* cell models offer various strategies to accommodate or even exploit these distinctions, such as the addition of human liver microsomes to the culture, or breeding of mice with transgenic or knock-in strains expressing human genes to investigate various parameters relevant to a particular cancer risk factor. The ability of the MEF immortalisation model to recapitulate human carcinogen mutation signatures observed from whole-genome analysis of human tumours suggests that the model can provide important clues about the involvement of potential carcinogens in instances where aetiological and mechanistic evidence is deficient.

Methods

Hupki MEF cell lines. This study included immortalised MEF cell lines derived from primary cultures exposed to carcinogens that were reported previously (**Supplementary Table 1** and references therein). They were generated following a procedure referred to in the literature as the 3T3 protocol[44], with minor adaptations. Briefly, fibroblasts from 13.5-day old Trp53$^{tm/Holl}$ mouse embryos harbouring a knock-in humanised version of the *p53* gene (Hupki MEFs) were seeded into six-well plates, exposed to cancer agents or solvent during early passages, and maintained in culture with occasional passaging until cultures emerged from senescence. Immortalised cultures were passaged at low density for several passages thereafter, prior to screening for the presence of a heterozygous or homo/hemizygous *TP53* mutation, and their designation as established cell lines. Cell lines chosen for the present analysis had acquired a dysfunctional *TP53* mutation during immortalisation (**Supplementary Table 1**). Acquisition of *Trp53* gene mutations frequently occurs during senescence bypass and establishment in culture[10,17,42,43] providing a convenient way to assess the identity and clonal origin of the immortalised cultures.

Cell lines with the AID transgene were established for the present study by crossing Hupki mice[45,46] with AID transgenic mice[47]. From the interbred colony, we harvested MEFs from embryos homozygous for the (non-mutated) knock-in *TP53* allele and either with or without the transgene (referred to as HxAID-Tg and MEFs respectively). T12.5 flasks (6 per MEF genotype) were seeded with 5×10^4 cells and cultured until the cells emerged from senescence, regained uniform morphology, and could sustain repeated passaging at >1 : 10 dilution. Genomic DNAs from two immortalised cell lines per condition (independent biological duplicates) were prepared for WES analysis.

DNA preparation and WES. Genomic DNA (gDNA) was extracted from cells using DNeasy blood and tissue kit (QIAGEN) and checked for purity, concentration, and integrity by OD260/280 ratio using NanoDrop Instruments (NanoDrop Technologies, Wilmington, DE, USA) and agarose gel electrophoresis. DNA was sheared by fragmentation by Covaris (Covaris, Inc., Woburn, MA USA) or Bioruptor (Diagenode, Inc., Denville, NJ, USA) and purified using Agencourt AMPure XP beads (Beckman Coulter, Fullerton, CA, USA). DNA samples were then tested for size distribution and concentration using an Agilent Bioanalyzer 2100 or Tapestation 2200 and by OD260/280 ratio. Fragment ends were repaired and Illumina libraries were generated using NEBNext reagents (New England Biolabs, Ipswich, MA, USA). Libraries were then subjected to exome enrichment using SureSelect XT Mouse All Exon Kit (Agilent Technologies, Wilmington, DE USA) following manufacturer's instructions. Enrichment was verified by qPCR and the quality, quantity and fragment size distribution of DNA determined by an Agilent Bioanalyzer or Tapestation. The libraries were sequenced in paired-end 100 nucleotide (nt) reads using the Illumina HiSeq2500 platform according to manufacturer's protocols.

MEF whole-exome data processing. All FASTQ files were analysed with FastQC to check sample homogeneity and quality. The FASTQ sequences were next aligned to the mm9 mouse reference genome with Burrows-Wheeler Aligner (BWA, version 0.7.5a) and the resulting SAM file was sorted and compressed in BAM format using Picard SortSam (version 1.98). Duplicate reads in the resulting BAM files were flagged with Picard MarkDuplicates. Local realignment around indels was performed in three steps: firstly, creation of a table of possible indels using GATK (version 2.7-2) RealignerTargetCreator, secondly, realignment of reads around those targets with GATK IndelRealigner, and lastly a correction of mate pair information was done using Picard FixMateInformation. The base quality score recalibration required two steps: first to generate a recalibration table with GATK BaseRecalibrator, then to print reads based on the previous table with GATK PrintReads. An average of 58.8 million reads (100 bp) were sequenced per sample, of which 98% were mapped, 77% on target with a mean coverage of 61 (see **Supplementary Table 3** for detailed metrics of sequencing quality and coverage). The recalibrated BAM files were used to call variants with MuTect software (version 1.1.4) using default parameters (including reads quality >20 and calls made only if the position has at least 14 reads in the tumour sample and at least 8 reads in the normal sample). As MuTect is tuned to perform normal/tumour comparison, primary cell cultures were used as "normal" samples and immortalised cell lines as "tumour" samples. Each immortalised cell line was compared to two primary MEF cultures and only the overlapping calls were taken into consideration to maximize the chance of robust variant calls and to exclude potential polymorphisms.

Human genome-wide sequencing datasets. Publicly available somatic mutation data obtained from whole-genome or whole-exome sequencing of human tumours were retrieved from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database or original papers (selecting only SBS): a set of 14,715 substitutions reported in 19 UTUC samples from AA-exposed patients[14]; a set of 6,066 substitutions observed in seven primary skin squamous cell carcinomas (COSMIC v67); a set of 3,026 mutations observed in 10 primary lung adenocarcinomas from heavy smokers[48]; a set of 10,455 mutations observed in eight glioblastomas recurring in patients treated with temozolomide[49]; a set of 288 mutations observed in eight primary astrocytomas[49]; and a set of 378 mutations observed in four primary glioblastomas[50].

Annotations of mutation data. For all datasets (MEF and human sets), the chromosome number, genomic coordinates, reference and mutated nucleotides were extracted for each variants. Variants were annotated with AnnoVar (version 2013aug23) using refGene, knownGene, ensGene, cytoBand, genomicSuperDups, and dbSNP128 databases for the mm9 mouse genome build. The human sets were annotated using additional databases: gwasCatalog, 1000 Genomes Project, NHLBI GO Exome Sequencing Project (ESP), COSMIC, dbSNP137 (hg19 build), and PolyPhen and SIFT databases for predicting the functional impact of mutations. Gene strand orientation was retrieved from the UCSC Genome Browser database using a Perl script developed by Heng Li at the Sanger Institute. Mutations were included in the analyses only if they could be successfully annotated. Variants present in the dbSNP128 polymorphism database were excluded. The comprehensive lists of all SBS identified in all MEF conditions and SBS from human tumour datasets are available as **Supplementary Dataset 3.**

Functional annotation analysis. A comprehensive list of established cancer driver genes (oncogenes and tumour suppressor genes) and candidate drivers coding for modifiers of DNA, histones and regulators of chromatin structure was assembled from literature and somatic mutation database mining[1,25,26]. Selected gene classes were annotated with functional domain information obtained from the UniProt and ENSEMBL databases. The comprehensive list of functional gene classes was matched against genes with mutations found in all MEF cell lines. Non-synonymously mutated genes were further selected, considering both exposure-specific alterations and any other mutation type. Human orthologues of the selected genes were examined in the COSMIC database for frequency of mutations in human tumours. For oncogenes, positions corresponding to non-synonymous mutations in MEFs were identified in human orthologues and investigated for mutation status in the COSMIC database. Gene Ontology Biological Process analyses were performed with the NIH DAVID web tool using default settings.

Statistical analyses. Statistical analyses were performed using the free R software (R Core Team, 2013) v3.0.2 or Excel. For the strand bias analyses, the statistical difference in the number of SBS between the non-transcribed and the transcribed strand was evaluated through the Pearson's $\chi 2$ test, using the prop.test function available in the stats R package. The test evaluated, for each experimental condition, whether the proportions of SBS in the non-transcribed strand differed from 0.5, which is the expected value by chance. As multiple conditions were assessed in parallel, a false discovery rate (FDR) correction was applied using the p.adjust function from the stats R package.

For analyses of mutation signatures, mutations were classified into 96 types determined by the six possible substitutions (A:T > C:G, A:T > G:C, A:T > T:A, C:G > A:T, C:G > G:C, C:G > T:A) and the 16 combinations of flanking (5′ and 3′) nucleotides. First, a PCA analysis was performed using as input the 96 variables. A 95% confidence interval was computed, including either only MEF samples or MEF/human data, to define the limits on the PCA plot for each experimental condition. Such analysis was performed based on the available functions in the FactoMiner package available in the Bioconductor repository (R package version 1.25. http://CRAN.R-project.org/package=FactoMineR). Second, the catalogue of experimental mutations defined by their 96 types was decomposed into signatures using the non-negative matrix factorisation algorithm of Brunet with the Kullback-Leibler divergence penalty[24,51]. The number of signatures was pre-set to six (the expected number of signatures based on the number of conditions) but the process was otherwise unsupervised: no information regarding exposures was used for the extraction of the signatures. To evaluate the similarity between the signatures from the cell lines and from human tumours by Alexandrov *et al.*[2,24], each signature was represented as a vector in 96-dimensional space. The tangent of the angle between each pair of vectors was taken as the distance metric: the tangent transformation serving to expand the scale to compensate for the geometry of high-dimensional space. This distance was used to compute the grid of distance from each of the six MEF signatures to each of the 27 human signatures and converted for presentation to a similarity matrix by taking the negative log of the distance with negative values (angles >45°) suppressed.

1. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
2. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
3. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
4. Hollstein, M., McCann, J., Angelosanto, F. A. & Nichols, W. W. Short-term tests for carcinogens and mutagens. *Mutat. Res.* **65**, 133–226 (1979).
5. Miller, J. H. Carcinogens induce targeted mutations in Escherichia coli. *Cell* **31**, 5–7 (1982).
6. Hahn, W. C. & Weinberg, R. A. Modelling the molecular circuitry of cancer. *Nat. Rev. Cancer* **2**, 331–341 (2002).
7. Collado, M. & Serrano, M. Senescence in tumours: evidence from mice and humans. *Nat. Rev. Cancer* **10**, 51–57 (2010).
8. Fridman, A. L. & Tainsky, M. A. Critical pathways in cellular senescence and immortalization revealed by gene expression profiling. *Oncogene* **27**, 5975–5987 (2008).
9. Odell, A., Askham, J., Whibley, C. & Hollstein, M. How to become immortal: let MEFs count the ways. *Aging (Albany. NY)* **2**, 160–165 (2010).
10. Liu, Z. *et al.* Human tumor p53 mutations are selected for in mouse embryonic fibroblasts harboring a humanized p53 gene. *Proc. Natl. Acad. Sci. U. S. A* **101**, 2963–2968 (2004).

11. Liu, Z. *et al.* p53 mutations in benzo(a)pyrene-exposed human p53 knock-in murine fibroblasts correlate with p53 mutations in human lung tumors. *Cancer Res.* **65**, 2583–2587 (2005).
12. vom Brocke, J., Schmeiser, H. H., Reinbold, M. & Hollstein, M. MEF immortalization to investigate the ins and outs of mutagenesis. *Carcinogenesis* **27**, 2141–2147 (2006).
13. Schmitz, K. M. & Petersen-Mahrt, S. K. AIDing the immune system-DIAbolic in cancer. *Semin. Immunol.* **24**, 241–245 (2012).
14. Hoang, M. L. *et al.* Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci. Transl. Med.* **5**, 197ra102 (2013).
15. Poon, S. L. *et al.* Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
16. Yoshikawa, K. *et al.* AID enzyme-induced hypermutation in an actively transcribed gene in fibroblasts. *Science* **296**, 2033–2036 (2002).
17. Whibley, C. *et al.* Wild-type and Hupki (human p53 knock-in) murine embryonic fibroblasts: p53/ARF pathway disruption in spontaneous escape from senescence. *J Biol. Chem.* **285**, 11326–11335 (2010).
18. Chen, C. H. *et al.* Aristolochic acid-associated urothelial cancer in Taiwan. *Proc. Natl. Acad. Sci. U. S. A* **109**, 8241–8246 (2012).
19. Moriya, M. *et al.* TP53 Mutational signature for aristolochic acid: an environmental carcinogen. *Int. J. Cancer* **129**, 1532–1536 (2011).
20. Sidorenko, V. S. *et al.* Lack of recognition by global-genome nucleotide excision repair accounts for the high mutagenicity and persistence of aristolactam-DNA adducts. *Nucleic Acids Res.* **40**, 2494–2505 (2012).
21. Bransteitter, R., Pham, P., Calabrese, P. & Goodman, M. F. Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase. *J. Biol. Chem.* **279**, 51612–51621 (2004).
22. Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424**, 103–107 (2003).
23. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560–561 (1980).
24. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
25. Plass, C. *et al.* Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nat. Rev. Genet.* **14**, 765–780 (2013).
26. Timp, W. & Feinberg, A. P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Rev. Cancer* **13**, 497–510 (2013).
27. Chan, H. M., Narita, M., Lowe, S. W. & Livingston, D. M. The p400 E1A-associated protein is a novel component of the p53 -->p21 senescence pathway. *Genes Dev.* **19**, 196–201 (2005).
28. Schmeiser, H. H. *et al.* Aristolochic acid activates ras genes in rat tumors at deoxyadenosine residues. *Cancer Res.* **50**, 5464–5469 (1990).
29. Schmeiser, H. H., Scherf, H. R. & Wiessler, M. Activating mutations at codon 61 of the c-Ha-ras gene in thin-tissue sections of tumors induced by aristolochic acid in rats and mice. *Cancer Lett.* **59**, 139–143 (1991).
30. Wang, Y. *et al.* ACB-PCR measurement of H-ras codon 61 CAA-->CTA mutation provides an early indication of aristolochic acid I carcinogenic effect in tumor target tissues. *Environ. Mol. Mutagen.* **53**, 495–504 (2012).
31. Lundberg, A. S., Hahn, W. C., Gupta, P. & Weinberg, R. A. Genes involved in senescence and immortalization. *Curr. Opin. Cell Biol.* **12**, 705–709 (2000).
32. Hollstein, M., Moriya, M., Grollman, A. P. & Olivier, M. Analysis of TP53 mutation spectra reveals the fingerprint of the potent environmental carcinogen, aristolochic acid. *Mutat. Res.* **753**, 41–49 (2013).
33. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
34. Burns, M. B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366–370 (2013).
35. Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* **45**, 977–983 (2013).
36. Chiba, T., Marusawa, H. & Ushijima, T. Inflammation-associated cancer development in digestive organs: mechanisms and roles for genetic and epigenetic modulation. *Gastroenterology* **143**, 550–563 (2012).
37. Shimizu, T., Marusawa, H., Endo, Y. & Chiba, T. Inflammation-mediated genomic instability: roles of activation-induced cytidine deaminase in carcinogenesis. *Cancer Sci.* **103**, 1201–1206 (2012).
38. Maitra, A. *et al.* Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. *Nat. Commun.* **4**, 2873 (2013).
39. Kino, K. & Sugiyama, H. GC-->CG transversion mutation might be caused by 8-oxoguanine oxidation product. *Nucleic Acids Symp. Ser.* 139–140 (2000).
40. Paz-Elizur, T. *et al.* DNA repair activity for oxidative damage and risk of lung cancer. *J. Natl. Cancer Inst.* **95**, 1312–1319 (2003).
41. Tommasi, S. *et al.* Mammalian cells acquire epigenetic hallmarks of human cancer during immortalization. *Nucleic Acids Res.* **41**, 182–195 (2013).
42. Harvey, D. M. & Levine, A. J. p53 alteration is a common event in the spontaneous immortalization of primary BALB/c murine embryo fibroblasts. *Genes Dev.* **5**, 2375–2385 (1991).
43. Kamijo, T. *et al.* Functional and physical interactions of the ARF tumor suppressor with p53 and Mdm2. *Proc. Natl. Acad. Sci. U. S. A* **95**, 8292–8297 (1998).
44. Todaro, G. J. & Green H. Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. *J. Cell Biol.* **17**, 299–313 (1963).
45. Luo, J. L. *et al.* Knock-in mice with a chimeric human/murine p53 gene develop normally and show wild-type p53 responses to DNA damaging agents: a new biomedical research tool. *Oncogene* **20**, 320–328 (2001).
46. Reinbold, M. *et al.* Common tumour p53 mutations in immortalized cells from Hupki mice heterozygous at codon 72. *Oncogene* **27**, 2788–2794 (2008).
47. Okazaki, I. M. *et al.* Constitutive expression of AID leads to tumorigenesis. *J. Exp. Med.* **197**, 1173–1181 (2003).
48. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
49. Johnson, B. E. *et al.* Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189–193 (2014).
50. Yost, S. E. *et al.* High-resolution mutational profiling suggests the genetic validity of glioblastoma patient-derived pre-clinical models. *PLoS. One* **8**, e56185 (2013).
51. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A* **101**, 4164–4169 (2004).

Acknowledgments

We gratefully acknowledge core financial support for whole-exome sequencing from the Deutsches Krebsforschungszentrum (DKFZ). We thank the Centre Léon Bérard in Lyon, France, for providing computational capacity. We thank Dr Zdenko Herceg for comments on the manuscript scientific contents and Sylvie Nouveau for carefully reading the manuscript.

Author contributions

M.O., M.H., J.Z. designed the study. A.W., T.N., K.-R.M., H.M., M.H. conducted cell culture and animal experiments. H.M. generated the AID-Tg and HxAID-Tg mice. M.V., J.McK., J.A., L.H. performed WES data processing and primary analyses. G.B. and X.C. performed the statistical analyses. M.O., M.A., X.C., H.H., H.M., G.B., M.H., J.Z. analysed and interpreted data. M.H. initiated the study. M.O., M.H., J.Z. wrote the manuscript. All authors approved the present submission of the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Olivier, M. *et al.* Modelling mutational landscapes of human cancers *in vitro*. *Sci. Rep.* **4**, 4482; DOI:10.1038/srep04482 (2014).

Objective 3-a

Paper 3: Low-coverage exome sequencing screen in formalin-fixed paraffin-embedded tumors reveals evidence of exposure to carcinogenic aristolochic acid

Xavier Castells*, Sandra Karanović*, **Maude Ardin***, Karla Tomić, Evanguelos Xylinas, Geoffroy Durand, Stephanie Villar, Nathalie Fore5, Florence Le Calvez-Kelm, Catherine Voegele, Krešimir Karlović, Maja Mišić, Damir Dittrich, Igor Dolgalev, James McKay5, Shahrokh F. Shariat, Viktoria S. Sidorenko, Andrea Fernandes, Adriana Heguy, Kathleen G. Dickman, Magali Olivier, Arthur P. Grollman, Bojan Jelaković, and Jiri Zavadil

Cancer Epidemiology,Biomarkers & Prevention, 2015

* Equally contributing authors

Aim Devise a method for detecting the aristolochic acid (AA) mutational signature in archived tumour samples.

Approach Design a low-coverage whole-exome sequencing (LC-WES) approach, using approximately 10x coverage, optimised for DNA tumour samples of limited quantity and quality extracted from formalin-fixed paraffin-embedded (FFPE) urothelial carcinoma tissues.

Graphical summary

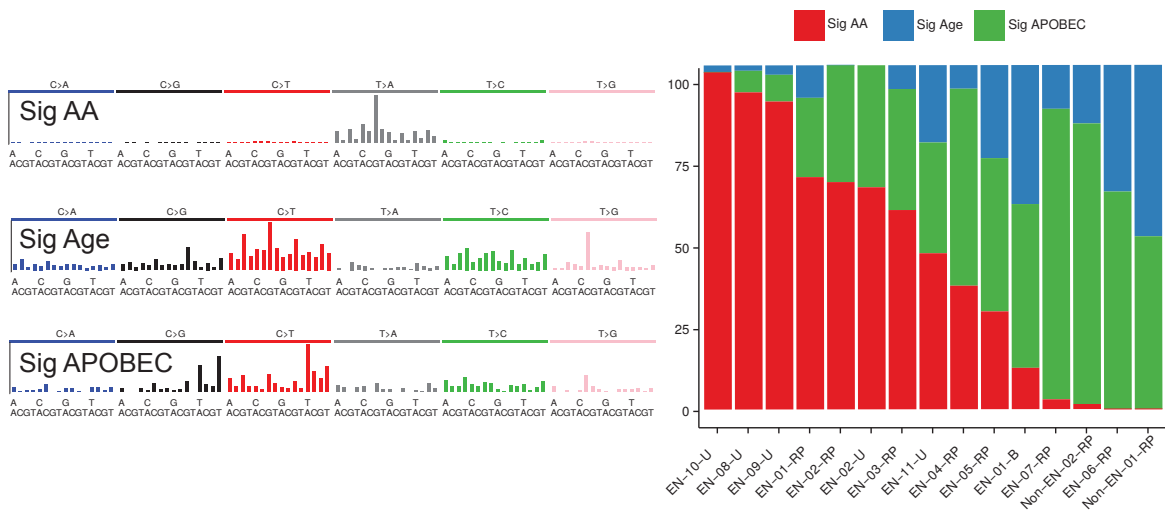


Figure 17: Mutational signatures in urological tumours analysed by LC-WES. Aristolochic acid signature (Sig AA) was found in 11 patients along side with the age (Sig Age) and APOBEC enzyme activity (Sig APOBEC) signatures. Bar graphs in the right panel show the contribution (expressed in %) of each signature to the mutation load in each individual tumour sample.

Novelty and highlights

- Low coverage sequencing, at around 10x (as opposed to conventional 50-100x), can successfully detect AA mutational signature in archived samples.
- LC-WES provides molecular evidence of AA exposure using limited DNA material.
- Cost-effective molecular epidemiology methodology for investigating AA exposure in human cancers.

Published OnlineFirst September 17, 2015; DOI: 10.1158/1055-9965.EPI-15-0553

Research Article

Cancer Epidemiology, Biomarkers & Prevention

Low-Coverage Exome Sequencing Screen in Formalin-Fixed Paraffin-Embedded Tumors Reveals Evidence of Exposure to Carcinogenic Aristolochic Acid

Xavier Castells[1], Sandra Karanović[2], Maude Ardin[1], Karla Tomić[3], Evanguelos Xylinas[4], Geoffroy Durand[5], Stephanie Villar[1], Nathalie Forey[5], Florence Le Calvez-Kelm[5], Catherine Voegele[5], Krešimir Karlović[3], Maja Mišić[3], Damir Dittrich[3], Igor Dolgalev[6], James McKay[5], Shahrokh F. Shariat[4], Viktoria S. Sidorenko[7], Andrea Fernandes[7], Adriana Heguy[6], Kathleen G. Dickman[7,8], Magali Olivier[1], Arthur P. Grollman[7,8], Bojan Jelaković[2], and Jiri Zavadil[1]

Abstract

Background: Dietary exposure to cytotoxic and carcinogenic aristolochic acid (AA) causes severe nephropathy typically associated with urologic cancers. Monitoring of AA exposure uses biomarkers such as aristolactam-DNA adducts, detected by mass spectrometry in the kidney cortex, or the somatic A>T transversion pattern characteristic of exposure to AA, as revealed by previous DNA-sequencing studies using fresh-frozen tumors.

Methods: Here, we report a low-coverage whole-exome sequencing method (LC-WES) optimized for multisample detection of the AA mutational signature, and demonstrate its utility in 17 formalin-fixed paraffin-embedded urothelial tumors obtained from 15 patients with endemic nephropathy, an environmental form of AA nephropathy.

Results: LC-WES identified the AA signature, alongside signatures of age and APOBEC enzyme activity, in 15 samples sequenced at the average per-base coverage of approximately 10×. Analysis at 3 to 9× coverage revealed the signature in 91% of the positive samples. The exome-wide distribution of the predominant A>T transversions exhibited a stochastic pattern, whereas 83 cancer driver genes were enriched for recurrent nonsynonymous A>T mutations. In two patients, pairs of tumors from different parts of the urinary tract, including the bladder, harbored overlapping mutation patterns, suggesting tumor dissemination via cell seeding.

Conclusions: LC-WES analysis of archived tumor tissues is a reliable method applicable to investigations of both the exposure to AA and its biologic effects in human carcinomas.

Impact: By detecting cancers associated with AA exposure in high-risk populations, LC-WES can support future molecular epidemiology studies and provide evidence-base for relevant preventive measures. *Cancer Epidemiol Biomarkers Prev; 24(12); 1873–81.*

[1]Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, Lyon, France. [2]School of Medicine, University of Zagreb, Department of Nephrology, Hypertension, Dialysis, and Transplantation, University Hospital Center Zagreb, Zagreb, Croatia. [3]General Hospital "Dr. Josip Benčević," Slavonski Brod, Croatia. [4]Department of Urology, Weill Cornell Medical College, New York, New York. [5]Genetic Cancer Susceptibility Group, International Agency for Research on Cancer, Lyon, France. [6]OCS Genome Technology Center, New York University Langone Medical Center, New York, New York. [7]Department of Pharmacological Sciences, Stony Brook University, Stony Brook, New York. [8]Department of Medicine, Stony Brook University, Stony Brook, New York.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Current address for E. Xylinas: Department of Urology, Cochin Hospital, Paris Descartes University, Paris, France; and current address for S.F. Shariat; Department of Urology, Medical University of Vienna, Vienna General Hospital, Vienna, Austria.

X. Castells, S. Karanović, and M. Ardin contributed equally to this article.

Corresponding Author: Jiri Zavadil, International Agency for Research on Cancer, 150 cours Albert Thomas, Room 709, Lyon 69008, France. Phone: 33-4-7273-8362; Fax: 33-4-7273-8322; E-mail: zavadilj@iarc.fr

doi: 10.1158/1055-9965.EPI-15-0553

Introduction

The International Agency for Research on Cancer (IARC) classified aristolochic acid (AA) as a group 1 carcinogen (1). Exposure to AA, following intake of *Aristolochia* herbaceous plants as traditional medicines or due to consumption of bread from flour contaminated by *Aristolochia* seeds, can lead to AA nephropathy (AAN). AAN is a progressive tubulointerstitial nephropathy with high risk of developing upper tract urothelial carcinoma (UTUC; refs. 2–5). In addition, recent studies proposed AA as a factor contributing to the development of hepatocellular (6–8), renal cell (9, 10) and urinary bladder carcinomas (11), and intrahepatic cholangiocarcinoma (12). Given this growing spectrum of AA-associated tumor types, AA exposure detection methods for screening of disease-risk populations are of key importance.

Following metabolic activation of AA, aristolactam (AL)-DNA adducts accumulate in the proximal tubules of the renal cortex and are used as biomarker of exposure (4, 13, 14). AL-DNA adducts may persist for over 20 years after the exposure had ceased (15, 16) and can be measured by ^{32}P-postlabeling (14, 17) or by ultra-performance-liquid chromatography-electrospray ionization-multistage scan mass spectrometry (UPLC-ESI-MS/MS^n), both

Published OnlineFirst September 17, 2015; DOI: 10.1158/1055-9965.EPI-15-0553

applicable to formalin-fixed paraffin-embedded (FFPE) tissues (16, 18, 19). However, the ^{32}P-postlabeling method lacks specificity, and access to the UPLC-ESI-MS/MS^n methodology and its optimization for biomaterial of low quantity are limiting factors.

DNA sequencing established a characteristic AA mutational signature marked by accumulation of A>T transversions within the 5′-Pyr-A-Pur-3′ sequence context (enriched for 5′-CpApG-3′), preferentially located on the nontranscribed strand (8–10, 20, 21). In cancers not associated with AA, such A>T transversions are infrequent (22, 23).

We exploited the unique features of the AA mutational signature to devise a sensitive method for AA exposure detection, based on low-coverage whole-exome sequencing (LC-WES, at approximately 10× in contrast with the conventional 100× coverage), optimized for analysis of tumor-specific DNA of limited quantity and integrity extracted from archived FFPE tissues. The studied urothelial tumor samples originated from a well-characterized population residing in the endemic nephropathy (EN) regions of Croatia and Bosnia and Herzegovina (13), with EN being thus far the only recognized environmental form of AAN (4, 24). For the first time, we report in the urothelial tumors of EN patients the genome-wide signatures of AA, age, and APOBEC cytidine deaminase activity, thereby extending previous mutational analyses of this population based solely on the mutations of the TP53 tumor-suppressor gene (4, 13, 25). In addition, we demonstrate the ability of LC-WES to elucidate the impact of the AA mutation spectra on key homeostatic biologic pathways and to reveal possible mechanisms of tumor dissemination along the urinary tract.

Materials and Methods

Patients and tumor samples

Exposure to AA was investigated in 15 patients with urothelial tumors, diagnosed with EN following established criteria (13, 26). As controls, UTUC samples were obtained from 4 patients from a metropolitan area of the United States, unlikely exposed to AA. All specimens were FFPE-converted in the histopathologic laboratories of the participating centers. The involved anatomical sites were renal pelvis, ureter, and bladder (ICD-10 codes C65, C66, and C67, respectively). Clinicopathologic features and *Aristolochia* exposure history are listed in Supplementary Table S1. The study protocols included patients' informed consent and were approved by the IARC Ethics Committee and the Institutional Review Boards of the participating institutions.

DNA isolation from paraffin sections

Hematoxylin and eosin preparations of the paraffin block sections were used to identify tumor tissue free of necrotic areas. The tumor cell areas were measured by ImageJ software (27). Ten micromolar sections, cut with the Leica RM 2145 microtome (Leica Microsystems), were used to macrodissect the tumor-enriched areas and isolate genomic DNA yielding 1 to 2 μg (5 to 10 ng/mm^2) per sample. Before DNA isolation, slides were deparaffinized for 5 minutes in 100% xylene, kept for 5 minutes in absolute ethanol, 5 minutes in 85% ethanol, 5 minutes in 75% ethanol and stored in milliQ water. DNA isolation was done using the QIAamp DNA FFPE Tissue Kit (Qiagen) following the manufacturer's protocol. DNA yields and concentrations were measured using the Picogreen assay (LifeTechnologies) and Fluoroskan Ascent FL microplate fluorometer (Thermo Fisher Scientific). DNA purity was evaluated by the NanoDrop 8000 spectrophotometer (Thermo Fisher Scientific), and DNA integrity assessed by 0.8% agarose gel electrophoresis.

AL-DNA adduct analysis and *TP53* resequencing

DNA was isolated from the renal cortex and tumor tissues by standard phenol-chloroform extraction techniques. The level of AL-DNA adducts in the renal cortex DNA (10–20 μg) was determined using ^{32}P-postlabeling PAGE, as previously described (13). The *TP53*-specific mutations were identified using the AmpliChip p53 Research Test (Roche Molecular Diagnostics), sensitively detecting all single base-pair substitutions and single-base deletions (13).

WES library preparation, exome capture, and sequencing

Two hundred and fifty (250) ng of genomic DNA was sheared by the adaptive focused acoustics method (Covaris, Inc.) to obtain approximately 300 bp fragments, with water temperature of 4°C, one cycle at 175 Watt peak power, duty factor 10 and 200 cycles per burst. Resulting fragment size was assessed using the 2100 Bioanalyzer and High Sensitivity DNA Kit (Agilent Technologies). The sheared DNA was converted into libraries using the Kapa LTP Library Preparation Kit (Kapa Biosystems). Briefly, the fragmented DNA was subjected to end repair reaction followed by poly-A-tailing and adapter ligation, excess adapters removed by Agencourt AMPure XP beads (Beckman Coulter). Eight cycles of PCR were performed to amplify the libraries with correct adapters on both ends. Four libraries (250 ng each) were pooled per exome capture with the Nimblegen SeqCap EZ Exome reagent. Exome-enriched mixes were PCR-amplified in 10 cycles, post-enrichment libraries pooled in 420 μL of water to a final concentration of 6 pmol/L. This volume was divided and loaded in two lanes of the rapid run mode flow cell for cluster generation and sequencing on the HiSeq2500, in a paired-end 50 bp cycle run. Multiplexing 16 samples per run resulted in the target coverage of approximately 10×.

Four additional EN UTUC samples and two UTUC samples from the metropolitan United States were analyzed in a validation assay using the SOLiD 5500XL sequencer (Life Technologies). See Supplementary Methods for details.

Raw HiSeq2500 sequencing data were deposited to the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) repository (ID SRP042035) to become available from the NCBI's dbGaP database. The annotated list of single-base substitutions (HiSeq2500 data) is provided in Supplementary Table S2.

Sequencing data analysis

FastQ reads were aligned to the human genome (hg19) using Burrows–Wheeler Aligner. Realignment and base quality score recalibration was done by the Genome Analysis Toolkit (GATK) and the duplicate-read removal by Picard. GATK HaplotypeCaller was used to call variants subsequently annotated on the RefSeq Gene transcript contents by ANNOVAR (28). Polymorphisms present in normal population and removed from our data originated from these collections: 1,000 genomes (1,000 g, http://www.1000genomes.org/), Exome Sequencing Project (ESP, http://exome.gs.washington.edu/) and the SNP database build 137 (dbSNP, http://www.ncbi.nlm.nih.gov/SNP/). We removed variants with frequency above 0.1% in either the 1,000 g or ESP databases, or annotated in the dbSNP database, or present in a

Published OnlineFirst September 17, 2015; DOI: 10.1158/1055-9965.EPI-15-0553

custom germline variant catalog built from 560 cases from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). Variants mapping to repetitive sequences contained in the genomic segmental duplication database (29) alongside variants with ≥90% homology with multiple regions were excluded. R functions were developed to compute the mutation type distributions and strand bias. The strand bias significance was determined by the Pearson χ^2 test. These tertiary analysis parameters were computed in two separate coverage ranges, ≥3× with no defined maximum, and between 3 and 9× to emulate ultra-low coverage.

Mutational signature analysis using nonnegative matrix factorization

Nonnegative matrix factorization (NMF) decomposes mutational patterns based on factorization of one matrix ($n \times m$) in two matrices W ($n \times r$) and H ($r \times m$) with the constraint that all three matrices must be composed of nonnegative elements (30). The r is the rank of factors to be extracted from the input matrix, corresponding to the number of signatures. The input matrix contained one column per patient (only HiSeq2500 data considered) and in rows the frequency of mutations types in 96 possible two-base sequence contexts. The R package NMF (31) was used to extract mutational signatures. The correlation between the extracted signatures and previously published ones (7, 21, 22, 32) and/or available in COSMIC (23) was computed as the inner product of the two signatures (vectors) divided by the product of their norms.

Functional analysis of tumor-specific nonsynonymous mutations

To examine the biologic impact of the gene mutants in the AA signature-positive samples, analysis was performed using the DAVID tool (33), with two input gene lists: (i) genes harboring nonsynonymous (missense, stop-gain, or stop-loss) SBS and (ii) genes nonsynonymously mutated in the EN dataset and in AA signature-positive samples from at least one of the two published datasets on UTUC in Taiwanese patients (8, 21). The list was further narrowed by classifying the mutated genes as established oncogenes or tumor suppressors listed by the Gene Set Enrichment Analysis (GSEA) database (34) and/or a cancer driver genes defined by recent seminal studies (35–39).

Results and Discussion

Low-coverage detection of AA exposure signature in urothelial tumors of EN patients

We applied HiSeq2500 LC-WES to genomic DNAs isolated from FFPE urothelial tumors from 11 EN cases, two of whom had concurrent UTUC and bladder carcinoma, and from two U.S. patients providing non-EN control samples (13 patients and 15 tumors in total, see Supplementary Table S1). Features of AA signature had been described earlier, as follows: mutational load of ≥40 SBS or ≥10 A>T in exonic positions, high proportion of A>T (>35% of all SBS types or as the predominant type) with a strand bias of ≥1.25, and ≥33% enrichment of A>T in the 5′-(C/T) pApG-3′ sequence context (8, 21). We used analogous criteria for the AA signature (≥50 SBS/sample of which ≥15% are A>T SBS, of which ≥20% are in the 5′-CpApG-3′ context), applying more stringent statistical analysis of the strand bias ratio combined with a cutoff of ≥1.5 (9, 11). Under these criteria the AA signature was readily observed in 10 of the 13 analyzed EN tumor samples, with 33% to 77% of A>T transversions per sample (Fig. 1), a nontranscribed strand bias of 2.0 to 3.3 and the 5′-C_G-3′ context enrichment above 19% (mean 24.6%, SD = 4.9, range of 19.1%–27.4%; Table 1). In contrast, A>T mutations and their enrichment in the 5′-CpApG-3′context are generally low in cancers of non-AA etiology, based on our analysis of 7,160 tumors of 52 cancer types in the COSMIC database (average 5.8% A>T, range 0%–12.1%, of which 10% are in the 5′-CpApG-3′ context). Similarly, the average percentage of A>T in the 5′-CpApG-3′ context in TCGA urothelial carcinoma data (only bladder data available) is 10.8% (0%–50%), whereas the mean percentage of all A>T mutations is low (average of 3.9%, range 0.8%–8.3%).

A weaker signature marked by 18.7% A>T, strand bias of 2.1 and the 5′-CpApG-3′ context proportion of 12.5% was observed in the bladder tumor sample (EN-01-B) of a patient with a concurrent AA signature-positive UTUC (EN-01-RP, see Table 1 and Supplementary Table S1). Two EN samples (EN-06 and EN-07) and the two non-EN controls were found negative for the AA signature, with A>T transversions present at 4% to 8%. In the case of EN-07 (bladder carcinoma with no history of UTUC), despite the presence of AL-DNA adducts in the patient's renal cortex, the mutation profile (Fig. 1) suggested AA-unrelated etiology. Among the AA signature-positive samples, we detected an average of 1,142 (range, 349–2,707) mutations per tumor [∼18 SBS/sample per exome megabase (Mb)] whereas the mutation rate in the control and negative samples (including the weaker AA-signature bladder cancer) was on average 357 (range, 258–440) mutations per sample (∼6 per sample/exome Mb). As shown in Table 1, the predominant A>T transversions substantially contributed to the high SBS counts.

Thus, LC-WES analysis of the EN UTUC generates results consistent with previous reports on the highly mutagenic potential of AA (8, 21, 40), and our results justify the use of exome sequencing for reliable detection of exposure to AA in archived FFPE material.

LC-WES identifies AA signature at ultra-low coverage

We next investigated whether the AA signature can be identified at ultra-low coverage. Upon considering 3 to 9 non-duplicate per-base reads, mutation counts in the AA-associated samples decreased to 233 per tumor on average (∼4 per sample/exome Mb) and to average 67 per tumor (1/sample/exome Mb) in the negative samples and the weakly positive bladder tumor (EN-01-B). The 10 tumors shown in Fig. 1 (two top rows) still exhibited the AA signature at ultra-low coverage (Supplementary Fig. S1), with the strand bias ratios between 1.7 and 4.7, and retained prominent enrichment of the 5′-CpApG-3′context (>25%). Thus, the specific and unique features of the AA signature can be reliably detected in FFPE tumor samples by superficial coverage sequencing.

These results open an attractive opportunity for retrospective analyses of archived pathologic specimens from the regions of AA exposure risk. In comparison with the ^{32}P-postlabeling and mass spectrometry adduct detection techniques, the LC-WES approach is based on a commodity technology that generates genome-wide information. LC-WES is also very sensitive, using low input DNA amounts (250 ng compared with 5–10 μg required for adduct analysis). Finally, it can indicate exposure to AA when neither AL-DNA adducts nor mutations in *TP53* are detected, as we demonstrate for the AA signature-positive cases EN-01, EN-03, EN-04, and EN-11 (Fig. 1; Supplementary Table S1).

Published OnlineFirst September 17, 2015; DOI: 10.1158/1055-9965.EPI-15-0553

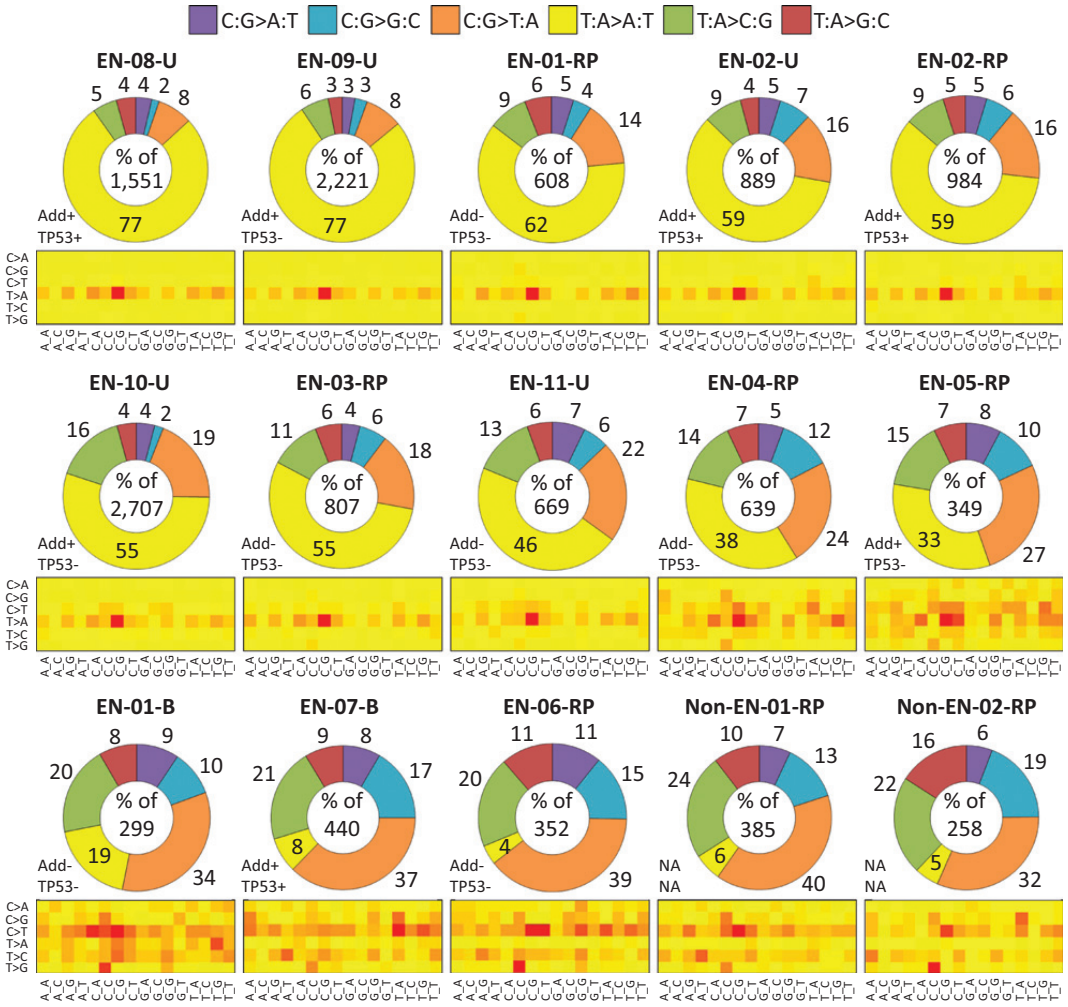


Figure 1.
SBS alterations in urothelial tumors analyzed by LC-WES. The distribution of six SBS types and their trinucleotide context are shown for variants detected at ≥3× per-base coverage. The doughnut charts correspond to individual samples (sample ID on top), ordered from high to low percentage of A>T. Total SBS counts per sample are provided in the center of each graph. The numbers outside the chart sections denote each mutation type percentage. The suffixes -B, -RP, and -U stand for bladder, renal pelvis, and ureter, respectively. Add±, sample positive or negative for aristolactam-DNA adducts; TP53±, mutated (+) or wild-type (−) TP53 gene. The heatmaps summarize relative frequencies of the six mutation types (C>A stands for C:G>A:T, etc.) across the 16 possible trinucleotide contexts listed at the bottom. Red, high frequency; yellow, low frequency.

AA-associated urothelial tumors harbor three major mutational signatures

NMF extracts individual mutational signatures from complex alteration patterns observed in primary tumors, reflecting thus the specific effects of etiologic factors (7, 9, 22, 41). NMF was used to describe the AA signature in human UTUC, bladder, liver, and renal carcinomas (7, 9, 11) and in experimental *in vitro* system designed to model mutational signatures of carcinogens (32). Here, in the EN urothelial tumors, the NMF approach identified three distinct signatures, the AA-specific signature (Signature 22; ref. 42), the signature related to age (C:G>T:A in the 5′-XpCpG-3′ context, Signature 1; ref. 22), and the Signatures 2 and/or 13 associated with the cytidine deaminase activity of the APOBEC enzymes (Fig. 2; ref. 22). All three signatures are currently listed in the COSMIC database (23). Furthermore, NMF aided in classifying the EN-01-B bladder tumor as positive due to nonnegligible sample contribution to the AA signature (18%), in contrast with the negative samples (EN-06-RP and EN-07-B, with contributions of 0% for both) and non-EN controls (each 0% contribution, see Table 1 and Fig. 2B). The identified EN UTUC AA signature

Published OnlineFirst September 17, 2015; DOI: 10.1158/1055-9965.EPI-15-0553

Table 1. Summary of the SBS detected at ≥ 3x per-base coverage and of the AA-signature analysis results

Case ID	Total SBS	SBS per Mbp	A > T SBS	A > T per Mbp	A>T (%)	CAG context (%)	SB A > T ratio (NTr/Tr)	SB[a] *P* value	SB FDR *Q* value	Contribution to Sig 22 (AA)	AA signature
EN-01-RP	608	9.1	376	5.6	61.8	26.1	3.3 (272/82)	0	0	380 (63%)	Yes
EN-01-B	299	4.5	56	0.8	18.7	12.5	2.1 (37/18)	0.015	1	53 (18%)	Yes[b]
EN-02-RP	984	14.7	585	8.7	59.5	26.7	3.1 (424/136)	0	0	514 (52%)	Yes
EN-02-U	889	13.3	529	7.9	59.5	26.1	3.3 (384/118)	0	0	457 (51%)	Yes
EN-03-RP	807	12.0	443	6.6	54.9	25.0	2.0 (278/140)	2.1E−11	1.7E−09	402 (50%)	Yes
EN-04-RP	639	9.5	241	3.6	37.7	19.5	2.3 (157/68)	4.4E−09	3.6E−07	164 (26%)	Yes
EN-05-RP	349	5.2	115	1.7	33.0	19.1	2.8 (79/28)	1.3E−06	1.1E−04	87 (25%)	Yes
EN-06-RP	352	5.3	15	0.2	4.3	0.0	5.5 (11/2)	0.027	1	0 (0%)	–
EN-07-B	440	6.6	35	0.5	8.0	11.4	1.2 (15/13)	0.850	1	0 (0%)	–
EN-08-U	1551	23.1	1195	17.8	77.0	23.4	2.8 (840/303)	0	0	1391 (90%)	Yes
EN-09-U	2221	33.1	1701	25.4	76.6	27.4	2.7 (1185/431)	0	0	1931 (87%)	Yes
EN-10-U	2707	40.4	1486	22.2	54.9	26.2	2.8 (1052/377)	0	0	2706 (100%)	Yes
EN-11-U	669	10.0	309	4.6	46.2	26.9	2.5 (211/85)	3.7E-13	0	297 (44%)	Yes
Non-EN-01-RP	385	5.7	24	0.4	6.2	0.0	0.6 (9/14)	0.400	1	0 (0%)	–
Non-EN-02-RP	258	3.9	14	0.2	5.4	14.3	5.0 (10/2)	0.043	1	0 (0%)	–

NOTE: Suffixes -B, -RP, and -U indicate bladder, renal pelvis, and ureter, respectively.

Abbreviations: SBS, single base substitution; Mbp, megabase pair; A > T, A > T transversion(s); SB A > T, strand bias, the ratio of A > T transversions on the non-transcribed versus transcribed strand (the number of respective transversions is shown in brackets); CAG context (%), percentage of A > T transversions in the most frequent context reported for the AA signature; NTr/Tr, ratio of A>T variants on nontranscribed versus transcribed strand; SB *P* value and FDR *q* value, measures of significance of the strand bias, see Materials and Methods. Contribution to Sig 22 (AA) (Sig = NMF-determined mutational signature) shown as the number of SBS and the corresponding percentage (in brackets) of the total SBS per sample. AA signature, positivity for AA signature considering the co-occurrence of ≥ 50 SBS and ≥ 15% A > T, ≥ 20% CAG context, strand bias (SB) and its significance (SB *P* and/or *q* value) and mutation load contribution to Sig 22 (AA), as described previously (9).

[a]Zero values correspond to a *P* value below 2×10^{-16}.

[b]Lower percentage in the 5′-CpApG-3′ context and nonsignificant *q* value; supported by the NMF analysis.

correlated highly (>90%) with the COSMIC Signature 22 (23), derived from AA-associated primary UTUC tumors from Taiwanese patients (8, 21), and with the AA signature modeled *in vitro* (Fig. 2C; ref. 32). The other EN tumor signatures matched their COSMIC counterparts with 72% similarity (age) and 70% and 64% similarity (APOBEC, Signature 2 and 13, respectively; Fig. 2C).

Validation of the LC-WES performance on a distinct sequencing platform

To validate the LC-WES performance using another sequencing chemistry and platform, we analyzed four additional EN UTUCs positive for AL-DNA adducts and p53 A>T mutations (EN-12, EN-13, EN-14, and EN-15), and two control UTUCs from U.S. patients (Non–EN-03 and Non–EN-04), on the SOLiD 5500xl sequencer. At the average 14.5× coverage, we observed the AA signature in all EN samples, although in samples EN-13-RP and EN-14-RP, the A>T transversion was the second most abundant mutation type following C>T (see Supplementary Fig. S2A). The signature remained detectable at ultra-low coverage (∼4.6×), when considering only the 3 to 9 read interval (Supplementary Fig. S2B).

Chromosomal distribution of the AA-specific mutations and recurrently mutated cancer driver genes

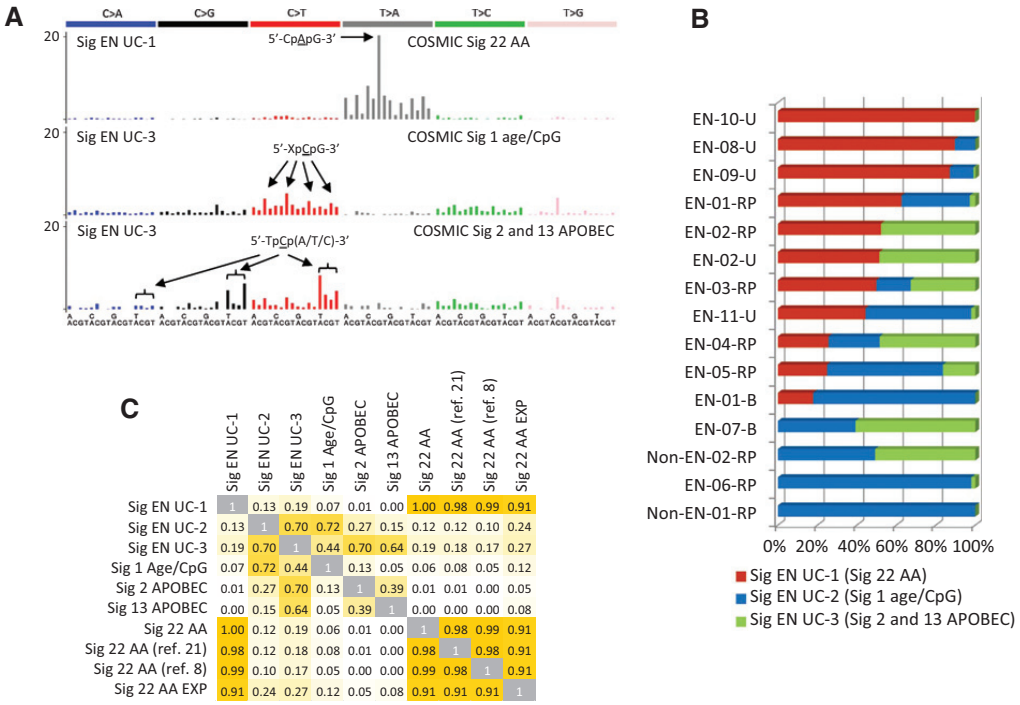
In the A>T enriched samples, A>T transversions were randomly distributed along the sequenced regions, with linear correlation between A>T SBS counts and chromosome size ($R^2 = 0.9$; Fig. 3A). Similar correlation was maintained in the minimum coverage interval of 3 to 9× (data not shown). This result was confirmed by the analysis of the Taiwan UTUCs (8, 21) in which a similar, although less linear trend was observed ($R^2 = 0.61$–0.64). These findings suggest a stochastic A>T mutation distribution within the gene/transcription units represented by the exome.

Despite this apparently random pattern, we identified 83 cancer driver genes carrying protein sequence-altering A>T SBS, that were recurrently mutated across the three datasets of the AA signature-positive tumor samples (this study, $n = 10$, and the two previously reported Taiwanese sets of $n = 18$ (21) and $n = 9$ (8). These findings are summarized in Fig. 3B and in Supplementary Table S4. The recurrently mutated genes included numerous known drivers and chromatin-associated factors such as *TP53*, *ARID1B*, *ATRX*, *CREBBP*, *CHD2*, *CHD5*, *CHD8*, *FAT1*, *KDM6A*, *MLL2* (*KMT2D*), *SETBP1*, *TRRAP*. *TP53* was the most frequently mutated gene [17/37 (46%) samples] with all its mutations being A>T transversions. Fifteen samples exhibited mutations in the histone methyl-transferase *KMT2D* (*MLL2*), with varying SBS types, suggesting that secondary mutation processes possibly linked due to high mutational loads and increased genomic instability. Further systematic investigations should be undertaken to establish possible recurrent alterations in particular genes and pathways in UTUC across studies of different populations/geographical areas. For instance, data in Fig. 3 and in Supplementary Table S3 indicate that *TP53*, *CREBBP*, and *LRRK2* are mutated mostly in the Taiwanese samples whereas mutations in the *AHNAK*, *ATRX*, *SMCHD1*, and *XIRP2* genes are enriched in the EN UTUC samples. Other factors contributing to these differences merit further investigations, including varying modes of AA exposure (low-dose chronic intake in the EN regions as compared with higher-dose, (sub)acute exposures resulting from the use of traditional herbal medicines in Asia) and disease susceptibility due to the patients' genetic background.

Biologic impact of the AA-signature

Using NIH DAVID, we performed Gene Ontology (GO) and KEGG pathway analyses of the genes harboring nonsynonymous A>T mutations in the AA-signature–positive samples analyzed by HiSeq2500 ($n = 10$). We identified gene targets from the

Published OnlineFirst September 17, 2015; DOI: 10.1158/1055-9965.EPI-15-0553



C

	Sig EN UC-1	Sig EN UC-2	Sig EN UC-3	Sig 1 Age/CpG	Sig 2 APOBEC	Sig 13 APOBEC	Sig 22 AA	Sig 22 AA (ref. 21)	Sig 22 AA (ref. 8)	Sig 22 AA EXP
Sig EN UC-1	1	0.13	0.19	0.07	0.01	0.00	1.00	0.98	0.99	0.91
Sig EN UC-2	0.13	1	0.70	0.72	0.27	0.15	0.12	0.12	0.10	0.24
Sig EN UC-3	0.19	0.70	1	0.44	0.70	0.64	0.19	0.18	0.17	0.27
Sig 1 Age/CpG	0.07	0.72	0.44	1	0.13	0.05	0.06	0.08	0.05	0.12
Sig 2 APOBEC	0.01	0.27	0.70	0.13	1	0.39	0.01	0.01	0.00	0.05
Sig 13 APOBEC	0.00	0.15	0.64	0.05	0.39	1	0.00	0.00	0.00	0.08
Sig 22 AA	1.00	0.12	0.19	0.06	0.01	0.00	1	0.98	0.99	0.91
Sig 22 AA (ref. 21)	0.98	0.12	0.18	0.08	0.01	0.00	0.98	1	0.98	0.91
Sig 22 AA (ref. 8)	0.99	0.10	0.17	0.05	0.00	0.00	0.99	0.98	1	0.91
Sig 22 AA EXP	0.91	0.24	0.27	0.12	0.05	0.08	0.91	0.91	0.91	1

Figure 2.
Mutational signatures determined by NMF. Results are shown for urothelial carcinoma samples sequenced on HiSeq2500. A, contribution of each mutation type to signatures of AA, age/CpG, and APOBEC. The *x*-axis represents the trinucleotide sequence contexts, with the 5′-flank base in the first row and the 3′-flank in the second. B, contributions of the studied urothelial tumors to the individual signatures shown in A. C, correlation between NMF-identified EN sample (EN UC) signatures and previously described COSMIC signatures 1, 2/13, and 22 (22), signature 22 identified in Taiwan UTUC samples (8, 21), and signature 22 AA Exp, modeled experimentally *in vitro* (32).

functional classes of cell adhesion, cell-matrix contact, cell migration, cell-cycle, cell signaling (MAPKKK/RAS and PI3K cascades, mTOR pathway), pathways of WNT, insulin and ERBB signaling, nucleotide excision repair, and the DNA-dependent ATPase and helicase activity, chromatin modification and histone binding related to gene-expression regulation, with dozens to hundreds of mutated genes per category (Supplementary Table S3). This observation suggests massive deregulation and/or destabilization of key homeostatic pathways by the high A>T mutation loads.

Next, for the 83 recurrently mutated cancer genes (Fig. 3B; Supplementary Table S4), we observed enrichment of GO and KEGG categories related to regulation of transcription, chromatin/histone modification, and categories of DNA damage response and DNA repair (Supplementary Table S5). These included numerous previously established cancer driver genes (*TP53*, *AHNAK*, *ARID1B*, *ATRX*, *BLM*, *CHD2*, *CHD5*, *CHD8*, *CHD9*, *CHEK2*, *CLTC*, *ERBB4*, *FN1*, *HUWE1 IARS2*, *KALRN*, *LRRK2*, *MLL2*, *NEB*, *RXRA*, *SMCHD1*, *SPEG*, *STAG2*, *SYNE1*, *TRIO*; refs. 35–39). Thus, the LC-WES analysis of AA-exposed urothelial tumors and associated data mining can reveal biologic information contents, particularly upon meta-analysis with data from different populations characterized by identical etiology and tumor types.

Overlapping mutation patterns in distinct tumors from same patients

Two EN patients had synchronous urothelial tumors in distinct anatomical sites (renal pelvis and bladder, samples EN-01-RP and EN-01-B; and renal pelvis and ureter, samples EN-02-RP and EN-02-U). By using LC-WES, we investigated the common genetic origins of these synchronous tumor pairs. In patient EN-01, the overlapping SBS were enriched for C>T mutations (42%) followed by A>G (20%), and only 7.7% of the overlapping SBS were A>T transversions affecting the coding sequence of mere 3 non-cancer genes (*VWA3B*, *KDM3B*, and *ACIN1*). However, the A>T SBS were enriched among the mutations unique to the renal pelvis and to the bladder tumor (77% and 28%, respectively, Supplementary Fig. S3A), suggestive of a common precursor carrying mainly non–A>T driver mutations, giving rise to two tumor progenies subsequently accumulating distinct patterns of A>T alterations in either anatomical site. The distinct AA signature in the bladder tumor is in keeping with a recent study of Asian bladder cancer patients in whose tumors the AA signature manifested without the involvement of upper tract or a history of renal disease (11). In contrast, the tumors in the renal pelvis and ureter of patient EN-02 shared the majority of mutations contributing to a prominent AA signature, suggesting a

Published OnlineFirst September 17, 2015; DOI: 10.1158/1055-9965.EPI-15-0553

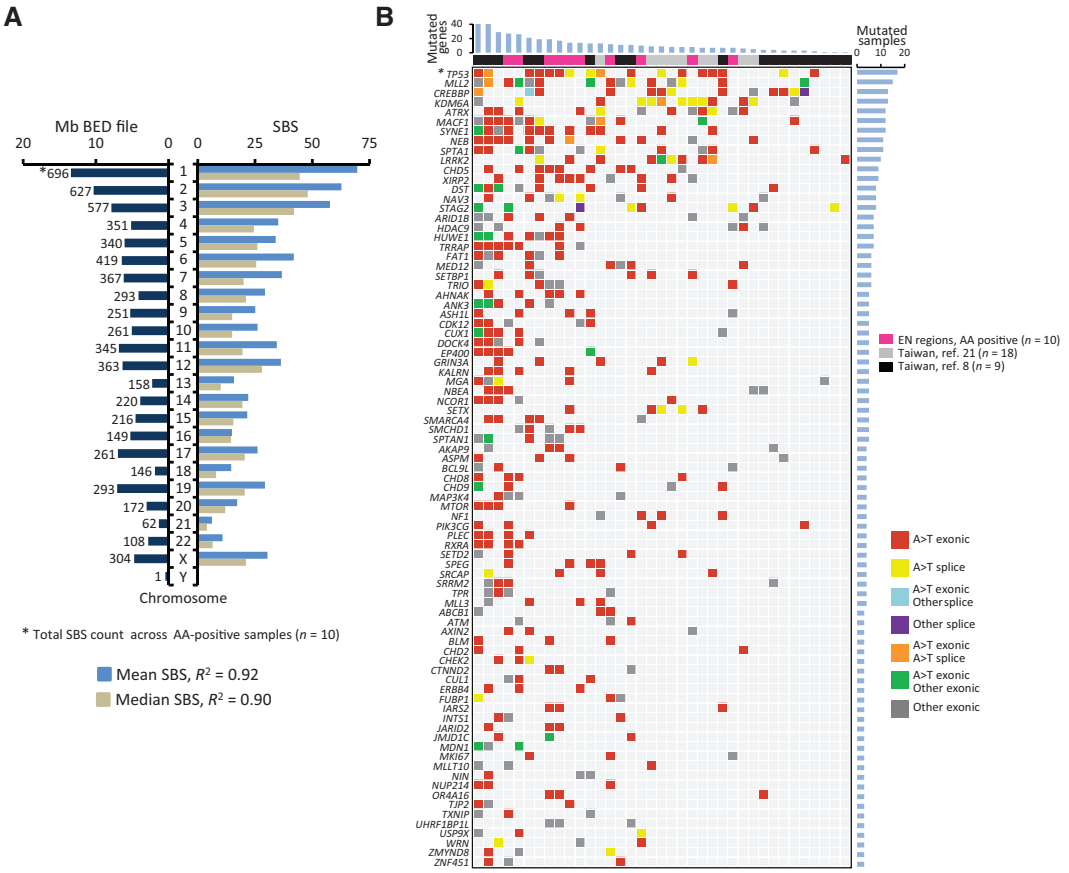


Figure 3.
Distribution of A>T mutations and their impact on cancer driver genes. A, correlation (squared Pearson product–moment correlation coefficient R^2) between the mean (light blue) and median (gray) values of A>T SBS counts per chromosome and the chromosome size in Mb (dark blue, left side). Variants based on ≥3 unique reads were considered. B, meta-analysis of recurrently mutated genes in AA-associated UTUC. Genes with nonsynonymous SBS variants identified in this study were compared with gene mutants found by two previously published AAN-UTUC datasets from Taiwan (8, 21). *TP53 mutations combine results from the AmpliChip and LC-WES analyses. The list of recurrently mutated genes was narrowed down to cancer driver genes only, as described in Materials and Methods. See also Supplementary Table S4 for detailed annotation of these mutations.

common precursor carrying mostly A>T alterations (Supplementary Fig. S3B). This genetic relationship between same-patient tumors suggests cell seeding along the tract as the basis for tumor dissemination. However, further investigations of a larger multiple-tumor case series and with the use of deep sequencing is needed to further elucidate the exact mechanisms of multifocal and recurrent tumorigenesis in the urinary tract of AA-exposed patients.

In summary, we report successful detection of the genome-wide AA signature in urothelial tumors of EN patients, using archived FFPE specimens and a customized low-coverage exome sequencing. The described technique is a cost-effective screening tool potentially applicable to molecular epidemiology studies aiming at identifying cancers associated with AA exposure. This ability of the LC-WES and its applicability to archived biomaterial may be exploited in future systematic studies on AAN and associated cancers, in support of established or future disease prevention programs.

Disclosure of Potential Conflicts of Interest

S.F. Shariat has received honoraria from speakers bureau from Astellas, Takeda, Ipsen, Janssen, Sanofi, Olympus, Wolff, Pierre Fabre, and Sanochenia, has ownership interest (including patents) in prostate and bladder cancer biomarkers, and is a consultant/advisory board member for Astellas, Ipsen, Sanochemia, Olympus, Wolff, and Janssen. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: A.P. Grollman, B. Jelaković, J. Zavadil
Development of methodology: X. Castells, M. Ardin, G. Durand, N. Forey, F. Le Calvez-Kelm, I. Dolgalev, J. McKay, A. Heguy, J. Zavadil
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): X. Castells, S. Karanović, K. Tomić, E. Xylinas,

Published OnlineFirst September 17, 2015; DOI: 10.1158/1055-9965.EPI-15-0553

G. Durand, N. Forey, F. Le Calvez-Kelm, D. Dittrich, S.F. Shariat, V.S. Sidorenko, A. Heguy, K.G. Dickman, A.P. Grollman, B. Jelaković, M. Mišić
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): X. Castells, M. Ardin, K. Tomić, C. Voegele, I. Dolgalev, J. McKay, S.F. Shariat, M. Olivier, J. Zavadil
Writing, review, and/or revision of the manuscript: X. Castells, S. Karanović, M. Ardin, K. Tomić, E. Xylinas, F. Le Calvez-Kelm, C. Voegele, S.F. Shariat, K.G. Dickman, M. Olivier, A.P. Grollman, B. Jelaković, J. Zavadil
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): X. Castells, S. Karanović, K. Tomić, S. Villar, K. Karlović, D. Dittrich, I. Dolgalev, A. Fernandes, K.G. Dickman, M. Ardin, M. Mišić
Study supervision: B. Jelaković, J. Zavadil

Acknowledgments

The authors thank the Genetic Platform, IARC, and the Genome Technology Center, NYU Langone Medical Center for expert assistance with exome sequencing; the authors thank Christine Carreira, Dr. Behnoush Abedi-Ardekani, and Dr. Elisabetta Kuhn for assistance with the biospecimen evaluation and processing. The authors thank the staff of the General Hospital Slavonski Brod, the University Hospital Centre Zagreb and of the Weill Medical College of Cornell University, New York, for expert assistance and support. The authors are grateful to Drs. Monica Hollstein and Michael Korenjak for critical comments on the article.

Grant Support

IARC Regular Budget; grant P01 ES04068 (to A.P. Grollman) from the NIH/NIEHS; R03-TW007042 grant (to A.P. Grollman) from the NIH Fogarty International Center; grant (to A.P. Grollman) from Henry and Marsha Laufer; grant no. 108-0000000-0329 (to B. Jelaković) from the Croatian Ministry of Science, and grant 04/38 (to B. Jelaković) from the Croatian Foundation for Science; and grant NIH/NCI P30 CA016087-33, which partially supports the NYU Genome Technology Center.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received May 18, 2015; revised August 11, 2015; accepted September 8, 2015; published OnlineFirst September 17, 2015.

References

1. IARC monographs on the evaluation of carcinogenic risks to humans volume 100A; A review of human carcinogens: pharmaceuticals. Lyon: International Agency for Research on Cancer; 2012.
2. Debelle FD, Vanherweghem JL, Nortier JL. Aristolochic acid nephropathy: a worldwide problem. Kidney Int 2008;74:158–69.
3. Grollman AP. Aristolochic acid nephropathy: Harbinger of a global iatrogenic disease. Environ Mol Mutagen 2013;54:1–7.
4. Grollman AP, Shibutani S, Moriya M, Miller F, Wu L, Moll U, et al. Aristolochic acid and the etiology of endemic (Balkan) nephropathy. Proc Natl Acad Sci U S A 2007;104:12129–34.
5. Nortier JL, Martinez MC, Schmeiser HH, Arlt VM, Bieler CA, Petein M, et al. Urothelial carcinoma associated with the use of a Chinese herb (Aristolochia fangchi). N Engl J Med 2000;342:1686–92.
6. Huang J, Deng Q, Wang Q, Li KY, Dai JH, Li N, et al. Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. Nat Genet 2012;44:1117–21.
7. Poon SL, McPherson JR, Tan P, Teh BT, Rozen SG. Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. Genome Med 2014;6:24.
8. Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. Sci Translat Med 2013;5:197ra01.
9. Jelakovic B, Castells X, Tomic K, Ardin M, Karanovic S, Zavadil J. Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid. Int J Cancer 2015;136:2967–72.
10. Scelo G, Riazalhosseini Y, Greger L, Letourneau L, Gonzalez-Porta M, Wozniak MB, et al. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. Nat Commun 2014;5:5135.
11. Poon S, Huang M, Choo Y, McPherson J, Yu W, Heng H, et al. Mutation signatures implicate aristolochic acid in bladder cancer development. Genome Med 2015;7:38.
12. Zou S, Li J, Zhou H, Frech C, Jiang X, Chu JS, et al. Mutational landscape of intrahepatic cholangiocarcinoma. Nat Commun 2014;5:5696.
13. Jelakovic B, Karanovic S, Vukovic-Lela I, Miller F, Edwards KL, Nikolic J, et al. Aristolactam-DNA adducts are a biomarker of environmental exposure to aristolochic acid. Kidney Int 2012;81:559–67.
14. Schmeiser HH, Bieler CA, Wiessler M, van Ypersele de Strihou C, Cosyns JP. Detection of DNA adducts formed by aristolochic acid in renal tissue from patients with Chinese herbs nephropathy. Cancer Res 1996;56:2025–8.
15. Schmeiser HH, Nortier JL, Singh R, da Costa GG, Sennesael J, Cassuto-Viguier E, et al. Exceptionally long-term persistence of DNA adducts formed by carcinogenic aristolochic acid I in renal tissue from patients with aristolochic acid nephropathy. Int J Cancer 2014;135:502–7.
16. Yun BH, Yao L, Jelakovic B, Nikolic J, Dickman KG, Grollman AP, et al. Formalin-fixed paraffin-embedded tissue as a source for quantitation of carcinogen DNA adducts: aristolochic acid as a prototype carcinogen. Carcinogenesis 2014;35:2055–61.
17. Dong H, Suzuki N, Torres MC, Bonala RR, Johnson F, Grollman AP, et al. Quantitative determination of aristolochic acid-derived DNA adducts in rats using ^{32}P-postlabeling/polyacrylamide gel electrophoresis analysis. Drug Metab Dispos 2006;34:1122–7.
18. Yun BH, Rosenquist TA, Sidorenko V, Iden CR, Chen CH, Pu YS, et al. Biomonitoring of aristolactam-DNA adducts in human tissues using ultra-performance liquid chromatography/ion-trap mass spectrometry. Chem Res Toxicol 2012;25:1119–31.
19. Yun BH, Rosenquist TA, Nikolic J, Dragicevic D, Tomic K, Jelakovic B, et al. Human formalin-fixed paraffin-embedded tissues: an untapped specimen for biomonitoring of carcinogen DNA adducts by mass spectrometry. Anal Chem 2013;85:4251–8.
20. Hollstein M, Moriya M, Grollman AP, Olivier M. Analysis of TP53 mutation spectra reveals the fingerprint of the potent environmental carcinogen, aristolochic acid. Mutat Res 2013;753:41–9.
21. Hoang ML, Chen CH, Sidorenko VS, He J, Dickman KG, Yun BH, et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. Sci Translat Med 2013;5:197ra02.
22. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415–21.
23. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015;43:D805–11.
24. Hranjec T, Kovac A, Kos J, Mao W, Chen JJ, Grollman AP, et al. Endemic nephropathy: the case for chronic poisoning by aristolochia. Croat Med J 2005;46:116–25.
25. Moriya M, Slade N, Brdar B, Medverec Z, Tomic K, Jelakovic B, et al. TP53 Mutational signature for aristolochic acid: an environmental carcinogen. Int J Cancer 2011;129:1532–6.
26. Jelakovic B, Nikolic J, Radovanovic Z, Nortier J, Cosyns JP, Grollman AP, et al. Consensus statement on screening, diagnosis, classification and treatment of endemic (Balkan) nephropathy. Nephrol Dial Transplant 2014;29:2020–7.
27. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods 2012;9:671–5.
28. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.
29. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. Science 2002;297:1003–7.

Published OnlineFirst September 17, 2015; DOI: 10.1158/1055-9965.EPI-15-0553

30. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788–91.
31. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 2010;11:367.
32. Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallee MP, et al. Modelling mutational landscapes of human cancers *in vitro*. Sci Rep 2014;4:4482.
33. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
34. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
35. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505:495–501.
36. Plass C, Pfister SM, Lindroth AM, Bogatyrova O, Claus R, Lichter P. Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. Nat Rev Genet 2013;14:765–80.
37. Shen H, Laird PW. Interplay between the cancer genome and epigenome. Cell 2013;153:38–55.
38. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci Rep 2013;3:2650.
39. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science 2013;339:1546–58.
40. Clyne M. Bladder cancer: aristolochic acid—one of the most potent carcinogens known to man. Nat Rev Urol 2013;10:552.
41. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Reports 2013;3:246–59.
42. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet 2014;15:585–98.

Published OnlineFirst September 17, 2015; DOI: 10.1158/1055-9965.EPI-15-0553

Cancer Epidemiology, Biomarkers & Prevention

AACR American Association for Cancer Research

Low-Coverage Exome Sequencing Screen in Formalin-Fixed Paraffin-Embedded Tumors Reveals Evidence of Exposure to Carcinogenic Aristolochic Acid

Xavier Castells, Sandra Karanovic, Maude Ardin, et al.

Cancer Epidemiol Biomarkers Prev 2015;24:1873-1881. Published OnlineFirst September 17, 2015.

Objective 3-b

Paper 4: Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid

Bojan Jelaković, Xavier Castells, Karla Tomić, **Maude Ardin**, Sandra Karanović and Jiri Zavadil

International Journal of Cancer, 2014

Aim Investigate the contribution of aristolochic acid (AA) to the aetiology of new tumour types from the endemic nephropathy region.

Approach The mutational spectra of clear cell renal cell carcinomas (ccRCC) and kidney chromophobe RCC were analysed using WES of DNA from formalin-fixed paraffin-embedded (FFPE) tumour tissues.

Graphical summary

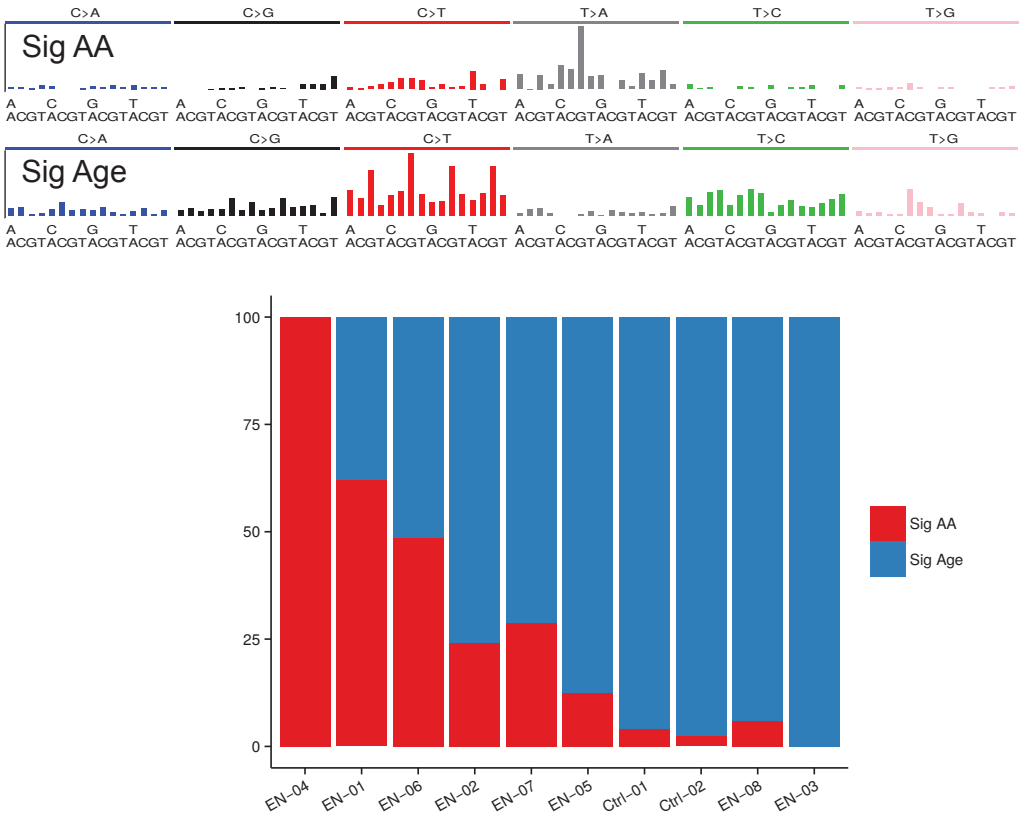


Figure 18: Mutational signatures found in kidney tumours from BEN region. On top mutational signature of AA (Sig AA) found in a subset of 5 patients and the age signature (Sig Age) found in almost all the patients. On bottom contribution of each signature to the mutation load in individual tumour samples.

Novelty and highlights

- AA mutational signature found in two tumour RCC subtypes diagnosed in AAN patients.
- Evidence provided for the potential role of AA in the development of RCC.

IJC
International Journal of Cancer

Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid

Bojan Jelaković[1], Xavier Castells[2], Karla Tomić[3], Maude Ardin[2], Sandra Karanović[1] and Jiri Zavadil[2]

[1] Department for Nephrology, Arterial Hypertension, Dialysis and Transplantation, University Hospital Center, School of Medicine, University of Zagreb, Zagreb, Croatia
[2] Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, Lyon, France
[3] Department of Pathology, Dr Josip Benčević General Hospital, Slavonski Brod, Croatia

Aristolochic acid (AA) is a potent dietary cytotoxin and carcinogen, and an established etiological agent underlying severe human nephropathies and associated upper urinary tract urothelial cancers, collectively designated aristolochic acid nephropathy (AAN). Its genome-wide mutational signature, marked by predominant A:T > T:A transversions occurring in the 5′-CpApG-3′ trinucleotide context and enriched on the nontranscribed gene strand, has been identified in human upper urinary tract urothelial carcinomas from East Asian patients and in experimental systems. Here we report a whole-exome sequencing screen performed on DNA from formalin-fixed, paraffin-embedded renal cell carcinomas (RCC) arising in chronic renal disease patients from a Balkan endemic nephropathy (EN) region. In the EN regions, the disease results from the consumption of bread made from wheat contaminated by seeds of *Aristolochia clematitis,* an AA-containing plant. In five of eight (62.5%) tested RCC tumor specimens, we observed the characteristic global mutational signature consistent with the mutagenic effects of AA. This signature was absent in the control RCC samples obtained from patients from a nonendemic, metropolitan region. By identifying a new tumor type associated with the AA-driven genome-wide mutagenic process in the context of renal disease, our results suggest new epidemiological and public health implications for the RCC incidence worldwide, particularly for the high-risk regions with unregulated use of AA-containing traditional herbal medicines.

Key words: aristolochic acid, endemic nephropathy, renal cell carcinoma, whole-exome sequencing, mutational signature

Abbreviations: AA: aristolochic acid; AAN: aristolochic acid nephropathy; bp: base-pair; CKD: chronic kidney disease; CTN: chronic tubulointerstitial nephropathy; eGFR: estimated glomerular filtration rate; EN: endemic nephropathy; FFPE: formalin-fixed, paraffin-embedded; HCC: hepatocellular carcinoma; IARC: International Agency for Research on Cancer; IDMS: isotope dilution mass spectrometry; Mbp: megabase-pairs; MDRD: Modification of Diet in Renal Disease; NMF: Non-negative Matrix Factorization; PCR: polymerase chain reaction; RCC: renal cell carcinoma; SBS: single base substitution; TCGA: The Cancer Genome Atlas; UTUC: upper urinary tract urothelial cell carcinomas; WES: whole-exome sequencing. ; WHO: World Health Organization

Grant sponsor: Croatian Science Foundation; **Grant number:** 04-38; **Grant sponsor:** NIH/NCI, partial support to the Genome Technology Center at the Laura and Isaac Perlmutter Cancer Center, New York University; **Grant number:** NIH/NCI P30CA016087; **Other Support:** IARC Regular Budget
DOI: 10.1002/ijc.29338
History: Received 5 Sep 2014; Accepted 3 Nov 2014; Online 17 Nov 2014

Correspondence to: Bojan Jelaković, School of Medicine, University of Zagreb, Kišpatićeva 12, 10000 Zagreb, Croatia, Tel.: +385-95-9030-751, E-mail: jelakovicbojan@gmail.com and Jiri Zavadil, International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon cedex 08, France, Tel.: +33-4-72-73-83-62, FAX: +33-4-72-73-83-22, E-mail: zavadilj@iarc.fr

Ingestion of nephrotoxic and carcinogenic aristolochic acid (AA) can lead to aristolochic acid nephropathy (AAN) marked by chronic tubulointerstitial nephropathy (CTN) and recurrent upper urinary tract urothelial cell carcinomas (UTUC), involving the renal pelvis and upper ureter. AAN has been designated a significant public health problem with millions of people world-wide being at risk.[1–3] Endemic nephropathy (EN) is an environmental form of AAN affecting particular regions of several Balkan countries, and manifesting by increased rates of chronic kidney disease (CKD) and upper urinary tract urothelial carcinogenesis that have been causally linked to the intake of AA through consumption of home-made bread prepared from wheat grains contaminated with seeds of *Aristolochia clematitis*.[4,5] Aristolactam-DNA adducts detected in renal cortex and/or A:T > T:A mutations in the 5′-CpApG-3′ context accumulating on the nontranscribed strand of the *TP53* gene in CTN and UTUC were reported as biomarkers of AA exposure in this geographical region.[4–6] Recent studies performed in Taiwanese patients with documented history of use of *Aristolochia*-containing traditional herbal medicine demonstrated that the A:T > T:A transversion, originally observed in the *TP53* gene,[7,8] is the predominant genome-wide mutation type in the UTUC.[9,10] The detailed characteristics of this somatic alteration such as its predominance among other mutation types, gene strand orientation bias and sequence context are highly specific to the genotoxic effects of AA. AA is classified as Group 1 carcinogen by the World Health Organization-

Short Report

What's new?
Ingestion of aristolochic acid (AA) causes severe nephropathies and carcinomas of the upper urinary tract, and represents a significant public health problem with millions of people at risk worldwide. In this study of renal disease patients in an endemic region, the authors identified a previously unrecognized type of renal cell carcinoma that harbors the mutational signature of this potent carcinogen. Their findings suggest that the putative causal role of AA in renal cortex carcinogenesis should be broadly addressed in high-risk regions marked by inadvertent exposure to AA or widespread use of AA-containing herbal remedies.

International Agency for Research on Cancer (WHO IARC) and its broader carcinogenic effects were demonstrated in animal models, by the induction of precancerous lesions and tumors in the forestomach, urinary tract and of fibrohistiocytic sarcomas at the AA injection site.[11–13] A limited number of hepatocellular carcinoma (HCC) cases of East Asian origin studied for the etiological effects of hepatitis B virus manifested with the AA signature.[10,14,15] The presence of the aristolactam-DNA adducts in the renal cortex has been reported previously in Taiwanese renal cancer patients[7] and observed in rats in other target tissues including forestomach, liver, kidney, urinary tract,[16] suggesting a wider tissue spectrum targeted by this highly potent mutagen. However, the association of AA with human malignancies other than UTUC and HCC remains largely unexplored.

In the last decade, a higher frequency of renal cell carcinomas (RCC) with distinct epidemiological and clinical features has been registered in the Croatian Centre for Endemic Nephropathy.[17] We thus aimed to investigate a possible role of AA in the etiology of RCC among CKD patients from the EN regions and close vicinity, by analyzing the genome-wide mutation spectra in the tumor DNA.

Materials and Methods

Patient samples

Eight RCC patients from the farming villages were analyzed: five from an EN area previously associated with exposure to AA due to consumption of contaminated bread[5,18] and three from villages close to the EN region with no EN cases reported in the past. In addition, two RCC cases from the metropolitan area of Croatia were analyzed as controls unlikely to have been exposed to AA. The clinical features of the patients are listed in Table 1. The study protocol included the patients' informed consent and ethical approvals were obtained from the Ethical Boards of the School of Medicine, University of Zagreb, of the General Hospital in Slavonski Brod and from the IARC Ethics Committee. Of the eight EN RCC patients, we identified four (EN-01, EN-02, EN-04 and EN-05) who had been baking own bread, three of whom were farmers harvesting grain from locally grown wheat; one patient presented with CTN (EN-01), one with concurrent UTUC (EN-02) and one (EN-06) had been diagnosed with UTUC five years prior to the diagnosis of RCC (see Table 1).

DNA isolation

Hematoxylin-eosin preparations from the formalin-fixed paraffin-embedded (FFPE) tumor blocks were used to identify tumor tissue free of necrotic areas by digital scanning at 20x magnification (Leica SCN400 Scanner, Leica Biosystems). The tumor areas to be macro-dissected were measured using the ImageJ free software or SlidePath Gateway Client, Leica Biosystems. Ten μm sections prepared by Leica RM 2145 microtome (Leica Microsystems) were used to isolate genomic DNA (2–3 μg, yield 5–10 ng/mm^2). Prior to genomic DNA isolation, slides were de-paraffinized for 5 min in 100% xylene, followed by 5 min in absolute ethanol, 5 min in 85% ethanol, 5 min in 75% ethanol and kept in milliQ water. DNA isolation was carried out using the QIAamp DNA FFPE Tissue kit (Qiagen). DNA yields and concentrations were measured using the Picogreen assay (Life Technologies) and Fluoroskan Ascent FL microplate fluorometer (Thermo Fisher Scientific). The purity was evaluated by the NanoDrop 8000 spectrophotometer (Thermo Fisher Scientific). The integrity of genomic DNA was assessed by 0.8% agarose gel electrophoresis.

Library preparation and whole-exome sequencing (WES)

Two hundred and fifty ng of genomic DNA were sheared using the adaptive focused acoustics^{TM} method (Covaris) to fragments of ~300 base-pair size on average, with water temperature of 4°C, one cycle at 175 Watt peak power, 10 duty factor and 200 cycles per burst. Resulting fragment size was assessed using the 2100 Bioanalyzer and the High Sensitivity DNA kit (Agilent Technologies). The sheared DNA went into library preparation using the KAPA LTP Library Preparation Kit (Kapa Biosystems). Briefly, the fragmented DNA was first subjected to end repair reaction followed by poly-A-tailing and adapter ligation. Excess adapters were removed by double solid-phase reversible immobilization clean-up using Agencourt AMPure XP beads (Beckman Coulter). Eight cycles of PCR were performed to amplify the libraries with correct adapter sequences on both ends. Next, exome capture was performed with pools of five libraries per hybridization (200 ng of each sample) using Nimblegen SeqCap EZ Exome v3.0 reagent. The exome-enriched populations were further amplified in a ten-cycle PCR amplification step. The post-enrichment libraries were pooled together to a final concentration of 6 pM in 420 μl. This volume was loaded on one lane of the rapid run mode flow cell for cluster generation on the HiSeq2500 (Illumina) and the samples were sequenced in paired-end 50 bp cycle run.

Sequencing data processing and analysis

Sequencing reads were aligned by Burrows-Wheeler Aligner, variants called by the Genome Analysis Toolkit (GATK,

Table 1. Demographic and clinicopathological features of the studied RCC cases

Case ID	Domicile	Gender	Age at surgery (years)	RCC subtype	ICD-O	Tumor grade	Tumor stage	Preoperative s-creatinine (μmol/L)	CKD stage	eGFR	Other kidney disease
EN-01	EN	M	66	Clear cell	8310/3	G4	pT3NxMx	710	5	7.0	Yes[1]
EN-02	EN	M	65	Clear cell	8310/3	G2	pT1N0Mx	162	3	39.6	UTUC[2]
EN-03	Near EN	M	42	Clear cell	8310/3	G2	pT1N0Mx	88	1	93.0	No
EN-04	EN	M	78	Clear cell	8310/3	G2	pT1NxMx	757	5	6.4	No
EN-05	Near EN	M	55	Clear cell	8310/3	G1	pT1NxMx	100	2	72.0	Yes[3]
EN-06	EN	M	76	Unclassified	8312/3	G4	pT3N2Mx	364	4	15.1	UTUC[4]
EN-07	Near EN	F	64	Chromophobe	8317/3	G3	pT2NxMx	75	2	73.0	No
EN-08	EN	F	47	Clear cell	8310/3	G2	pT2N0Mx	95	3	58.1	No
Ctrl-01	Non-EN	F	56	Clear cell	8310/3	G3	pT2N0Mx	62	1	91.8	No
Ctrl-02	Non-EN	M	74	Clear cell	8310/3	G2	pT3aN0Mx	193	3	31.5	No

Estimated glomerular filtration rate based on the Modification of Diet in Renal Disease (MDRD) equation and not measured by isotope dilution mass spectrometry (IDMS). The last column indicates the patient's kidney disease other than RCC (UTUC: upper urinary tract urothelial carcinoma).
[1]Chronic tubulointerstitial nephropathy (CTN).
[2]Concurrent tumor.
[3]Nephrolithiasis and RCC in the contralateral kidney.
[4]UTUC diagnosed 5 years prior to RCC.
Abbreviations: EN: endemic region; near EN: close vicinity; F: female; M: male; ICD-O: the International Classification of Diseases for Oncology Code; CKD: chronic kidney disease; eGFR: estimated glomerular filtration rate.

Broad Institute), annotated by ANNOVAR and filtered stringently to remove genetic variants observed in general population, using data from the public projects 1,000 genomes (1,000*g*, http://www.1000genomes.org/), Exome Sequencing Project (ESP, http://exome.gs.washington.edu/) and SNP database build 137 (dbSNP, http://www.ncbi.nlm.nih.gov/SNP/). Variants exhibiting a frequency higher than 0.1% in either 1,000*g* or ESP databases, annotated in dbSNP database and an in-house panel of germ line variants generated using data obtained from 560 cases from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) were removed, as were the variants in fragments mapping to repetitive sequences of the genome with a homology higher than 90%. The complete final list of 4,031 SBS variants in the tested as well as control RCC samples will be provided upon request. The raw sequencing data (fastQ and alignment (.bam) files) have been deposited in the Sequence Read Archive (SRA, access ID SRP049084) of the National Center for Biotechnology Information (NCBI) and can be obtained through the NCBI's database for Genotypes and Phenotypes (dbGaP) authorized access system.

Determination of the AA mutational signature

The genome-wide AA signature in UTUC tumors had been defined previously based on varying criteria, *e.g.*, ≥40 SBS per sample, increased A:T > T:A presence (20–80% of all SBS types) in the predominant 5′-CpApG-3′ context, and 1.25- to 2-fold bias for A:T > T:A accumulation on the nontranscribed strand.[9,10] In the sequencing data presented here we first followed analogous criteria, considering ≥15% of A:T > T:A per tumor (10.5% is the maximum frequency seen in the COSMIC database upon excluding the HCC class containing a small number of potentially AA signature-positive tumors found by previous studies[14,19]) and the concurrent predominance of the 5′-CpApG-3′ sequence context. For strand bias analysis, we applied additional stringency by calculating its statistical significance using Pearson χ^2 test (prop.test function available in the stats R package). For each sample the test calculated a *p*-value as the probability that the proportion of SBS in the nontranscribed strand is equal to 0.5, as expected by chance. As multiple conditions were assessed in parallel, a false discovery rate (FDR) correction was applied using the p.adjust function from the stats R package. The identified signature based on these criteria was visualized by customized R functions. To extract comprehensive gene signatures from all ten RCC samples and to validate the presence of the AA signature by an independent method, non-negative matrix factorization (NMF) was performed under optimized conditions using an R package.[20] The input matrix contained one column per sample and the rows contained the frequency of the six possible mutation types within a two-flank sequence context (*e.g.*, A > T in the C_G context *et cetera*). The context distribution was normalized to reflect the trinucleotide frequencies occurring in the portion of the human genome corresponding to the exome capture reagent coverage. The similarity between the NMF RCC signatures and published signatures from cancer and experimental settings was further evaluated by each signature represented as a vector in a 96-dimensional space. The tangent of the angle between each pair of vectors was taken as the distance metric. This distance was used to compute the grid of distance from each of the two RCC signatures to each of the 24 reference signatures and converted for graphical presentation to a similarity matrix by taking the negative logarithm of the distance.

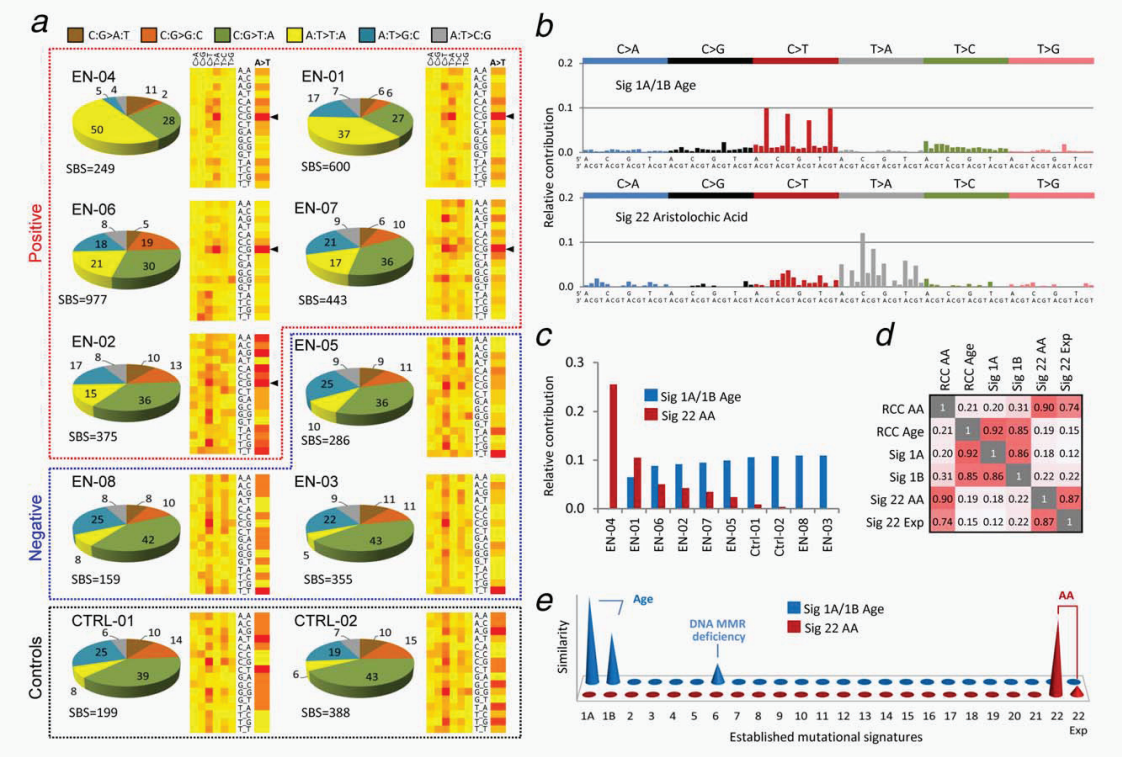


Figure 1. Summary of WES results for the AA signature-positive RCC cases and RCC controls. (*a*) Pie charts show distribution of mutation types (color-code shown on top), as percentage of all single base substitution (SBS) counts in each sample. Groups of AA signature-positive, negative samples and controls are demarcated by dotted lines. Two-dimensional heat-maps indicate the relative distribution of 96 possibilities of six mutation type categories in a nonstranded view (columns, the labels C > A, C > G, *etc.* represent C:G > A:T, C:G > G:C, *etc.*) and their trinucleotide context (rows). A detailed view of the relative distribution of the A:T > T:A context only is shown as a single column to the right under the **A>T** header. The most frequent 5′-CpApG-3′ context (labeled C_G) for the A:T > T:A transversion in the positive samples is marked by black arrowheads. Heat-map color code: Red-to-yellow = high-to-low context frequency. (*b*) Comprehensive identification of the mutational signatures in all 10 studied samples, using NMF. *X*-axis: 96 trinucleotide contexts; *Y*-axis: normalized values of each column represent the contribution of each mutation type in a particular sequence context to each signature, expressed as the value of a given factor obtained from the NMF decomposition process. (*c*) Relative contribution of the mutational load identified in individual samples to the NMF signatures shown in (*b*), expressed as the given factor value obtained from the decomposition process. See Table 2 for contributions expressed as mutation counts. (*d*) Correlation matrix (1 equals 100% similarity) of RCC NMF signatures (RCC AA and RCC Age) with previously validated signatures 1A and 1B (age),[22] with NMF-processed AA signatures in UTUC (Sig 22 AA)[9] and in AA-exposed clonally immortalized mouse embryonic fibroblasts (Sig 22 Exp).[21] (*e*) Graphical representation of the similarity distance of the RCC signatures (front-back axis) to 23 human cancer signatures (horizontal axis, 1A through 22[9,22]) and to an AA signature established experimentally in cultured cells (22 Exp).[21] *Y*-axis: similarity of signatures between the two systems, expressed as negative log(tan(angle)), see Materials and Methods. Negative values below the *x*–*z* plane correspond to angles >45° representing dissimilarity, and are thus suppressed.

Results

The WES analysis of DNA macrodissected from FFPE tumor specimens revealed the AA signature in five of eight (62.5%) RCC samples of patients from the EN regions. These findings were based on increased A:T > T:A transversion rates (1.2–3.5 SBS/Mbp, A:T > T:A frequencies of 17–50% and strand orientation bias ratios of 1.6–2.7) in four tested EN RCC samples (EN-01, EN-04, EN-06 and EN-07; Fig. 1*a* and Table 2). Interestingly, the EN-07 case was a resident of a village not considered an EN village. This observation is in line with recent evidence suggesting that ingestion of AA via contaminated bread was more widespread than previously thought, and cases of EN could be identified in villages and regions outside the established EN areas.[5] One EN RCC case (EN-02) showed lower presence of A:T > T:A (15.2%, 0.9 per Mbp) but a significant strand orientation bias of 2.8, and also enrichment for the 5′-CpApG-3′ sequence context for the A:T > T:A transversions (Table 2). It was thus considered borderline positive considering the AA signature features previously described in Taiwanese UTUC patients[9,10] and in experimental model systems.[10,21] In contrast, the two non-EN RCC controls exhibited low A:T > T:A frequencies (Ctrl-01: 7.5% and Ctrl-02: 6.4%), as did the remaining two RCC cases from villages near to the EN area (EN-03 4.8%, EN-05 9.8%)

Short Report

Table 2. Summary of the SBS mutation data and results of the AA signature analysis

Case ID	Total SBS	SBS per Mbp	A>T SBS	A>T per Mbp	% A>T	% CAG context	SB A>T	SB *p* value	SB FDR *q* value	Contribution to Sig 1A/1B (Age)	Contribution to Sig 22 (AA)	AA signature
EN-01	600	9.4	224	3.5	37.3	23.0	1.6 (128/78)	0.00064	0.010	228 (38%)	372 (62%)	Yes
EN-02	375	5.9	57	0.9	15.2	11.4	2.8 (39/14)	0.00098	0.013	256 (68%)	119 (32%)	Yes[1]
EN-03	355	5.5	17	0.3	4.8	7.0	2.5 (10/4)	0.181	0.746	355 (100%)	0 (0%)	–
EN-04	249	3.9	124	1.9	49.8	22.3	2.7 (82/30)	0	0	0 (0%)	249 (100%)	Yes
EN-05	286	4.5	28	0.4	9.8	8.2	1.3 (14/11)	0.689	1	231 (81%)	55 (19%)	–
EN-06	977	15.3	201	3.1	20.6	35.0	2.6 (140/53)	0	0	623 (64%)	354 (36%)	Yes
EN-07	443	6.9	76	1.2	17.2	23.7	1.8 (46/26)	0.025	0.231	325 (73%)	118 (27%)	Yes
EN-08	159	2.5	13	0.2	8.2	0.0	1.8 (7/4)	0.546	1	159 (100%)	0 (0%)	–
Ctrl-01	199	3.1	15	0.2	7.5	0.0	0.6 (5/8)	0.579	1	185 (93%)	14 (7%)	–
Ctrl-02	388	6.1	25	0.4	6.4	7.5	0.9 (12/13)	1	1	376 (97%)	12 (3%)	–

A>T: A:T>T:A transversion; SB A>T: strand bias expressed as the ratio of A:T>T:A transversions on the nontranscribed *versus* transcribed strand (the number of respective transversions is shown in brackets). % CAG context: percentage of A:T>T:A transversions in the most frequent context reported for the AA signature. SB *p* value and false discovery rate (FDR) *q* value = measures of significance of the strand bias, see Materials and Methods. Contribution to Sig (Sig = non-negative matrix factorization-determined mutational signature) is shown as the number of SBS and the corresponding percentage (in parentheses) of the total SBS number per sample. AA signature—sample positivity for AA signature considering the simultaneous occurrence of ≥50 SBS and ≥15% A>T, ≥20% CAG context and its significance (SB *p* and/or *q* value) and mutation load contribution to Sig 22 (AA).

[1]Lower percentage of A>T and CAG context but supported by NMF.

Abbreviation: SBS: single base substitutions.

and also one negative EN RCC case (EN-08, 8.2%) (see Table 2). In order to position these findings in a broader context, we analyzed 518 RCC samples unlikely to be exposed to AA, available from The Cancer Genome Atlas (TCGA). The 62.5% AA signature positivity rate in the samples of the EN region provenance appeared significant ($p < 2.2e-16$, using two-sample $\chi 2$ distribution test for equality of proportions with continuity correction) as the TCGA set contained only one sample meeting all the parameters of the AA signature (*i.e.*, simultaneous occurrence of ≥50 SBS and ≥15% A>T, ≥20% 5′-CpApG-3′ context and >1.5 strand bias ratios with significant *p* values and FDR corrected *q* values, see Table 2), and it contained another six samples with elevated A:T>T:A and strand bias (with nonsignificant *q* values) but lacking the expected predominant 5′-CpApG-3′ sequence context for the A:T>T:A transition.

To further strengthen the AA signature determination criteria and to identify additional mutational patterns reflecting effects of other etiological agents, we applied NMF, a mixed-pattern decomposition method used previously to successfully extract mutational signatures from human cancers and to pinpoint etiological factors.[22] NMF established that samples EN-01, EN-02, EN-04, EN-06 and EN-07 were positive for the AA signature (NMF Signature 22 described previously in human urothelial and liver tumors[23,24] and in immortalized clones arising from cultured, AA-treated primary embryonic fibroblasts[21]), and that all tumors except for EN-04 harbored high contents of C:G>T:A transitions within the 5′-XpCpG-3′ context (Figs. 1*b* and 1*c*). This observation is consistent with the age-specific signature 1A/1B[22] reflecting the generally advanced patients' age at the time of surgery (see Table 1). The NMF approach thus unequivocally grouped the borderline RCC sample EN-02 with the four AA signature-specific samples confirming the presence of the canonical AA signature in this tumor and providing additional information on sample-specific mutational load toward each signature (Fig. 1*c* and Table 2). The samples EN-03, EN-05, EN-08 and the two non-EN controls (Ctrl-01 and Ctrl-02) were negative for the AA signature but exhibited the signature of age (Figs. 1*b* and 1*c*). No additional markedly distinct signatures apart from the two shown in Figure 1*b* were found by the NMF approach when the number of expected signatures was increased. Next, we performed similarity analysis matching the two NMF signatures found in the tested RCC set (Fig. 1*b*) against the previously published data on known mutational signatures in human cancers[22,23] including AA-associated UTUC tumors,[9] as well as against the AA signature induced experimentally in cultured cells.[21] While the EN RCC age signature matched the published the Signatures 1A, 1B, and to a lesser extent the Signature 6 (attributed to DNA mismatch repair deficiency),[22] the RCC NMF AA signature matched the AA signatures previously identified both in the UTUC samples from Taiwanese patients and in cultured cells harboring an experimentally generated AA signature. The results are summarized graphically in Figures 1*d* and 1*e*.

Discussion

Our study identifies the specific AA mutational signature in a set of five RCC tumors from patients from a Balkan EN region and near vicinity, providing molecular evidence for AA-driven mutagenic process associated with carcinogenesis in the renal cortex and implicating AA in the etiology of this

cancer type in the context of CKD. While the mutation load in the AA signature-positive EN RCC confirms the highly mutagenic effects of AA,[9,10] it appears lower (4–15 SBS/Mbp) in our RCC set in comparison to the previously reported AA signature-positive UTUC (2–65[9] and 8–35[10] SBS/Mbp). These differences warrant further epidemiological and laboratory investigations using larger patient sets, in order to establish whether they reflect varying AA exposure modes in each geographical region, the metabolizing rates in the distinct target cell types, the sampling procedures affecting the heterogeneity of tested tissues or other factors.

Given the unregulated global market with AA-containing herbs that might be putting millions of people at risk, particularly in Eastern Asia,[3] our finding has potentially profound implications for increased awareness of additional cancer types associated with AAN. Robust designs of molecular epidemiology studies of AAN should now include RCC cases alongside patients with UTUC and HCC, to improve the worldwide surveillance of the disease. Furthermore, we document in a new region and a new cancer type that sensitive genome-wide screens coupled with sophisticated mutation pattern decomposition methods such as NMF, can identify AA as an environmental cancer risk factor in AAN and provide a robust evidence-base for diagnostic and preventive approaches toward eradication of this serious public health problem.[3,25]

Acknowledgements

We thank Drs. Želimir Stipančić and Tvrtko Hudolin for assistance with collection of clinical samples and data, Dr. Stephanie Villar and Ms. Christine Carreira for their assistance with archived tumor sample processing, and Dr. Magali Olivier for helpful comments on the sequencing data analysis. We thank the New York University Genome Technology Center at the New York University Laura and Isaac Perlmutter Cancer Center, for expert assistance with the Illumina HiSeq2500 sequencing experiments.

References

1. Debelle FD, Vanherweghem JL, Nortier JL. Aristolochic acid nephropathy: a worldwide problem. *Kidney Int* 2008;74:158–69.
2. Gokmen MR, Cosyns JP, Arlt VM, et al. The epidemiology, diagnosis, and management of aristolochic acid nephropathy: a narrative review. *Ann Intern Med* 2013;158:469–77.
3. Grollman AP. Aristolochic acid nephropathy: Harbinger of a global iatrogenic disease. *Environ Mol Mutagen* 2013;54:1–7.
4. Grollman AP, Shibutani S, Moriya M, et al. Aristolochic acid and the etiology of endemic (Balkan) nephropathy. *Proc Natl Acad Sci USA* 2007; 104:12129–34.
5. Jelakovic B, Karanovic S, Vukovic-Lela I, et al. Aristolactam-DNA adducts are a biomarker of environmental exposure to aristolochic acid. *Kidney Int* 2012;81:559–67.
6. Moriya M, Slade N, Brdar B, et al. TP53 Mutational signature for aristolochic acid: an environmental carcinogen. *Int J Cancer* 2011;129:1532–6.
7. Chen CH, Dickman KG, Moriya M, et al. Aristolochic acid-associated urothelial cancer in Taiwan. *Proc Natl Acad Sci USA* 2012;109:8241–6.
8. Hollstein M, Moriya M, Grollman AP, et al. Analysis of TP53 mutation spectra reveals the fingerprint of the potent environmental carcinogen, aristolochic acid. *Mutat Res* 2013;753:41–9.
9. Hoang ML, Chen CH, Sidorenko VS, et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci Transl Med* 2013;5:197ra02.
10. Poon SL, Pang ST, McPherson JR, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci Transl Med* 2013;5:197ra01.
11. Cosyns JP, Goebbels RM, Liberton V, et al. Chinese herbs nephropathy-associated slimming regimen induces tumours in the forestomach but no interstitial nephropathy in rats. *Arch Toxicol* 1998;72:738–43.
12. Cosyns JP, Dehoux JP, Guiot Y, et al. Chronic aristolochic acid toxicity in rabbits: a model of Chinese herbs nephropathy? *Kidney Int* 2001;59: 2164–73.
13. Debelle FD, Nortier JL, De Prez EG, et al. Aristolochic acids induce chronic renal failure with interstitial fibrosis in salt-depleted rats. *J Am Soc Nephrol: JASN* 2002;13:431–6.
14. Huang J, Deng Q, Wang Q, et al. Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. *Nat Genet* 2012;44: 1117–21.
15. Sung WK, Zheng H, Li S, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet* 2012;44:765–9.
16. Schmeiser HH, Schoepe KB, Wiessler M. DNA adduct formation of aristolochic acid I and II in vitro and in vivo. *Carcinogenesis* 1988;9: 297–303.
17. Belicza M, Demirovic A, Tomic K, et al. Comparison of occurrence of upper urinary tract carcinomas in the region with endemic villages and non-endemic nephropathy region in Croatia. *Coll Antropol* 2008;32:1203–7.
18. Hranjec T, Kovac A, Kos J, et al. Endemic nephropathy: the case for chronic poisoning by aristolochia. *Croat Med J* 2005;46:116–25.
19. Kan Z, Zheng H, Liu X, et al. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res* 2013;23: 1422–33.
20. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* 2010;11:367.
21. Olivier M, Weninger A, Ardin M, et al. Modelling mutational landscapes of human cancers in vitro. *Sci Rep* 2014;4:4482.
22. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature* 2013;500:415–21.
23. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* 2014;15:585–98.
24. Poon SL, McPherson JR, Tan P, et al. Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. *Genome Med* 2014;6:24.
25. Gokmen MR, Lord GM. Aristolochic acid nephropathy. *BMJ* 2012;344:e4000.

Short Report

Objective 3-c

Paper 5: Origins and impact of A:T>T:A mutations in human cancers

Maude Ardin, Liacine Bouaoun, Ludmil B. Alexandrov, Monica Hollstein, Graham Byrnes, Zdenko Herceg, Jiri Zavadil and Magali Olivier

In preparation

Aim Characterise the specificity of the aristolochic acid (AA) mutational signature and identify the variations of the rare A:T>T:A-based mutational signatures in human cancer.

Approach A comprehensive data mining of WES data from human public cancer repositories (COSMIC, TCGA and ICGC) and AA-studies was done in order to select T:A>A:T-enriched samples for extracting T:A>A:T-based mutational signatures.

Graphical summary

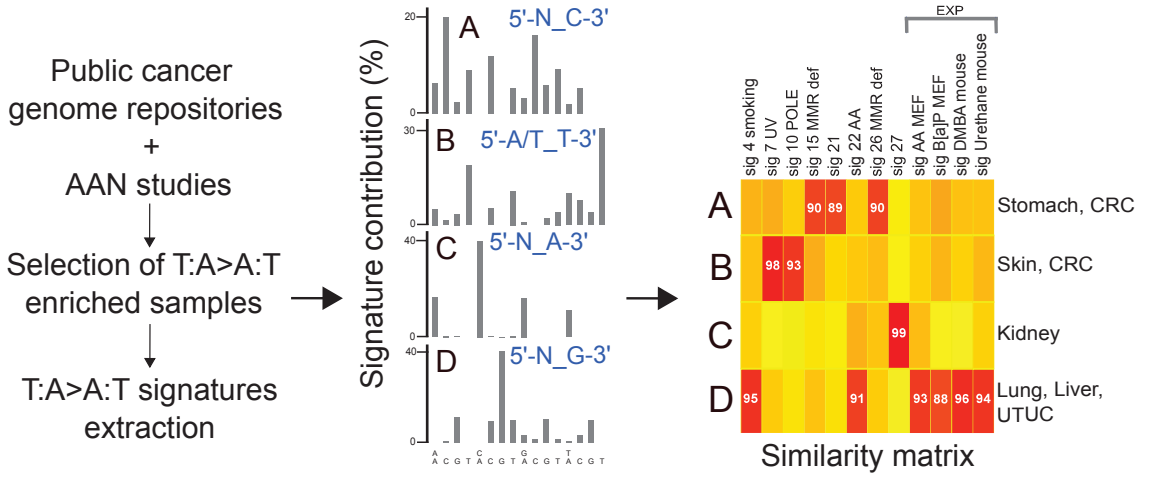


Figure 19: T:A>A:T-based mutational signatures and similarity with T:A>A:T component of published signatures. Samples enriched in T:A>A:T were selected after a comprehensive data mining of public cancer genome data repositories. Mutational signatures were extracted considering only T:A>A:T transversions and four signatures were obtained (Sig A,B,C and D). The predominant sequence context is indicated on top of each signature. The heatmap shows cosine similarity values, expressed in %, with the T:A>A:T-component of published signatures. The main tumour types in which the T:A>A:T signatures were found are listed on the right of the heatmap, CRC: colorectal cancer, UTUC: upper tract urothelial carcinoma cancer.

Novelty and highlights

- High contents of A>T transversions within 5'-Pyr-A-Pur-3' on the non-transcribed strand is specific of AA exposure.
- The sequence context, 5'-CpApG-3', is targeted by other carcinogenic compounds or processes but at a lower frequencies.
- AA mutational signature has been found in tumour samples (liver, kidney and bladder cancer) not previously reported in association with AA exposure.

Origins and impact of A:T>T:A mutations in human cancers

Maude Ardin[1], Liacine Bouaoun[2], Ludmil Alexandrov[3], Monica Hollstein[1], Graham Byrnes[2], Zdenko Herceg[4], Jiri Zavadil[1*], Magali Olivier[1*]

(1) Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, World Health Organization, 150 cours Albert Thomas, 69372 Lyon cedex 08, France

(2) Group of, International Agency for Research on Cancer, World Health Organization, 150 cours Albert Thomas, 69372 Lyon cedex 08, France

(3) Theoretical Biology and Biophysics (T-6) and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM, USA

(4) Epigenetics Group, International Agency for Research on Cancer, World Health Organization, 150 cours Albert Thomas, 69372 Lyon cedex 08, France

*corresponding authors: Jiri Zavadil, Email: zavadilj@iarc.fr or Magali Olivier, Email: olivierm@iarc.fr

Keywords: aristolochic acid, mutational signature, cancer genomes, mutation patterns, mutagenesis.

Databases: TCGA; ICGC; COSMIC

Abstract

Exposure to the genotoxic compound aristolochic acid (AA), an IARC Group 1 carcinogen, leads to renal disease and urological and hepatobilliary cancers. AA-related tumors carry specific types of mutations caused by AA mutagenesis which are characterized by a high prevalence of A:T>T:A transversions in a 5'-Py-**A**-Pu-3' (predominantly 5'-CpApG-3') sequence context, marked by transcriptional strand-bias. The presence of this mutational signature may thus indicate prior exposure to AA and mutagenesis via the aristolactam-DNA adduct formation pathway. However, the specificity of this signature towards AA has not been fully demonstrated. In order to address this question and to characterize the general landscape of A:T>T:A mutations in human cancers, we performed systematic and comprehensive mining of information on A:T>T:A transversions available from public cancer genome databases.

Here we analyzed somatic mutation data from TCGA, ICGC and COSMIC repositories to identify all possible A:T>T:A-based mutational signatures, using the non-negative matrix factorization (NMF) algorithm. Tumors from patients with documented AA exposure were included in the analysis as positive controls. The obtained signatures were compared to the A:T>T:A component of published signatures using the cosine similarity method to investigate their possible origin. From 19,466 samples available in public repositories, we identified 1,048 samples enriched for A:T>T:A mutations. In this sample set, we identified four A:T>T:A-based mutational signatures. Two were matching with the A:T>T:A compotent of signatures reported to be linked to endogenous mechanisms (e.g. repair deficiency, POLE) and one with the A:T>T:A component of a previously reported signature of unknown origin (COSMIC signature 27). Interestingly, one of the signature, Sig D, matched with the A:T>T:A component of several previously published signatures, including COSMIC signature 22 related to AA, COSMIC signature 4 linked to tobacco smoking, and experimental signatures caused by AA, urethane and dimethylbenzanthracene (DMBA). The contribution of these signatures to the overall mutation load in individual tumor samples varied greatly between sample types but the greatest contribution was observed for Sig D in UTUC samples known to be related to AA exposure. Our results show that the predominant sequence context of A:T>T:A mutations caused by AA is not uniquely characteristic of the effects of activated AA, as other carcinogens

may contribute to a similar A:T>T:A signature pattern. However, we demonstrate that A:T>T:A are rare somatic events in human cancers and that the presence of Sig D in a combination with a high A:T>T:A mutation load is correlative with AA exposure.

Introduction

The increasingly accumulating data from genome-scale sequencing studies of human cancers suggest that among the 12 possible types of single base substitutions (SBS), A:T>T:A transversions are less frequently represented [1–3]. However, in certain cancer types linked to particular exogenous mutagenic insults the mutation load can be highly enriched for this particular type of SBS mutation. A prominent example of this scenario is the predominant A:T>T:A pattern observed in urological cancers linked to ingestion of aristolochic acids [4–7]. Aristolochic acids AA-I and AA-II are classified by the International Agency for Research on Cancer (IARC) as Group 1 carcinogens. Exposure to AA present in the herbaceous plants of the *Aristolochia* genus, used as traditional medicines or occurring as a contaminating weed in wheat fields, can lead to progressive tubulo-interstitial nephropathy. This disease, named aristolochic acid nephropathy (AAN) is linked primarily to the cytotoxic effects of AA-I and is associated with elevated risk of developing cancer in the upper tract urothelium and urinary bladder [8,9]. A number of recent studies further characterized AA as a factor contributing to the development of renal cell and hepatobiliary cancers in patients from South Eastern Europe [6,10] and East Asia [11,12]. This rapidly emerging spectrum of AA-associated tumor types highlights the importance of reliable exposure history screening in the populations at risk.

Following ingestion, AA becomes metabolically activated into aristolactam nitrenium ions which in turn form covalent aristolactam (AL)-DNA adducts in the target tissues. AL-DNA adducts in renal cortex and other tissues are used as biomarkers of exposure, as they are very stable and in the cells of renal tubules they may persist for over two decades after the cessation of exposure. DNA sequencing established a characteristic AA mutational signature marked by accumulation of A:T>T:A transversions within the 5'-Pyr-A-Pur-3' sequence context enriched for 5'-CpApG-3', with slight variations reported for renal cancers of BEN patients (5'-CpAp(T/A)-3') and UTUC tumors from AA-exposed patients from Taiwan, China

(5'-(C/T)pApG-3'. The A>T transversions are preferentially located on the non-transcribed strand, i.e. manifesting a transcriptional strand bias. These characteristic mutations are considered highly specific to the effects of AA, and they arise due to the fact that AL-DNA adducts that block DNA replication, thereby exerting miscoding effects, remain resistant to the global genome-nucleotide excision repair but can be partially repaired by transcription-coupled nucleotide excision repair acting on the transcribed DNA strand [13].

Additional A:T>T:A-enriched patterns have been reported, for instance in primary human renal cancers sequenced as part of the international TCGA consortium project, providing the foundation for COSMIC mutational signature 27 based solely on T>A mutations in the 5'-NpTpA-3' context. Next, accumulation of A:T>T:A components of overall mutation patterns were reported in various experimental rodent models of controlled carcinogenesis. One example involves lung cancer development model upon exposure to ethyl carbamate (urethane, IARC Group 2A), primarily involving the formation of promutagenic 1,N^6-ethenodeoxyadenosine adducts consistent with the resulting A:T>T:A mutation component of the overall genomic mutation profile. Another experimental system for which genome-wide A:T>T:A pattern has been reported previously, involves induced carcinogenesis of the skin upon treatment with dimethylbenzanthracene (DMBA) [14,15]. The proposed mechanism for A:T>T:A mutations due to DMBA exposure is the loss of N3 and/or N7-dA adduct by depurination, followed by misreplication of unrepaired apurinic sites [16]. For both of these chemical compounds, the targeted trinucleotide sequence context of the A:T>T:A transversions is similar to that observed for AA as it exhibits the predominant 5'-Cp[A>T]pG-3' pattern. Similarly to urethane, ethenoadducts are also formed following the exposure to vinyl chloride (IARC Group 1 carcinogen) [17,18].

Therefore, an important question arises as to the specificity of AA mutational signature as it has been described to-date, particularly if the mutation pattern is to be used as exposure biomarker in molecular cancer epidemiology studies. It is thus important to determine additional human carcinogens distinct from AA that can generate A:T>T:A mutations in these particular sequence contexts.

In this study, we address these questions by identifying all possible variations of the relatively rare A:T>T:A-based mutational signatures in human cancer, using comprehensive data mining and analysis of whole-exome sequencing data from

human public cancer repositories (COSMIC, TCGA and ICGC) and cancer studies aimed at exposure to AA.

Results and discussion

Frequencies of A:T>T:A transversions in human cancers

Publicly available somatic mutation data from exome sequencing studies were retrieved and curated from the three main cancer genome repositories and from original publications describing tumor samples associated with exposure to AA (AA studies)(see Methods). A total of 19,466 samples were retrieved, including 8,000 from COSMIC, 9,000 from TCGA, 2,000 from ICGC and 53 from AA exposure studies. The tumor type distribution of these samples is heterogeneous. Three cancer types (blood, brain and head and neck) represent more than half of the samples, and the AA study dataset includes only two tumor types, UTUC and renal cell carcinomas from Taiwan, China and from the BEN region in Southern Europe (Figure 1A).

The total number of SBS per sample for all tumor types is shown in Figure 1B. As described before by Lawrence et al. [1], samples with the highest load of SBS are associated with exposure to known and potent carcinogens, such as UTUC, skin, urinary bladder and lung cancer (Figure 1B). Interestingly, samples from the AA exposure dataset have the highest mutation load with a median of 440 SBS (average of 759 SBS) per sample. A:T>T:A substitutions represent a small fractions of these mutations in most cancer types except in UTUC associated with AA exposure (Figure 1C). A:T>T:A substitutions are thus rare somatic events in cancer (median of 3 and average of 8 A:T>T:A transversions per sample) but are found at high levels in AA-related cancers (median of 224 and an average of 490 A:T>T:A transversions in AA exposed samples). Nevertheless, specific set of samples not known to be related to AA exposure exhibited a high proportion of A:T>T:A mutations (Figure 2). These samples correspond to various tumor types including carcinomas of the gall bladder, stomach, lung and liver. Plotting the sequence context of SBS in these A:T>T:A enriched samples shows that, in some samples, the A:T>T:A sequence context is close to the one described for AA and show a strand bias (Suppl fig 1).

A:T>T:A mutational signatures

To further characterize the patterns of A:T>T:A substitutions in human cancer and to better understand their possible origin, we extracted mutational signatures based on A:T>T:A mutations using the NMF approach described in Alexandrov et al. [2]. For this purpose, we parsed out all A:T>T:A substitutions within their trinucleotide sequence context (involving 5' and 3' 1-base flanks), for all samples carrying at least 32 A:T>T:A changes. We selected samples with at least 32 A:T>T:A mutations because A:T>T:A-based signatures are composed of 16 parameters, corresponding to 16 possible trinucleotide contexts, and we want to allow for at least two mutations per sequence context. Among the 19,466 samples, 1,048 samples from 22 tumor types passed this selection threshold (Suppl fig 2).

Four A:T>T:A signatures were found from this sample set (Figure 3, left panel). This signature analysis was confirmed by another method based on probabilistic models [19]. This method, *pmsignature*, identified the same four signatures in this sample set (Suppl fig 3). Furthermore, we also accounted for over-fitting by randomly generating 200 permutation-based NMF factorizations (see Methods). None of the randomized data had a cophenetic coefficient value above the one obtained from the original data, ensuring that the 4 signatures obtained are not due to random event (Suppl fig 4).

The obtained signatures were compared to the A:T>T:A components of a reference set of published signatures that include signatures extracted from human cancers and from experimental mouse models (see Methods) (Figure 3, middle panel). Finally, the number of samples contributing to each signature was computed (Figure 3, right panel).

The first signature, Sig A, is characterized by T>A in the 5'-NpTp(C/T)-3' sequence context and was found mainly in colorectal and stomach cancers. The majority of the colorectal samples are from TCGA and COSMIC United States-based studies and only a subset come from a Chinese study of the ICGC. In regards to stomach tumor samples, only a subset are from COSMIC studies from the U.S. and South Korea and all the other samples are from a TCGA study that collected the samples from Russia. Sig A matches COSMIC signatures 15 (cosine similarity cos=0.9) and 26 (cos=0.9) reported to be linked to DNA repair deficiency.

The second signature, Sig B, exhibited T>A in the 5'-(A/T)pTpT-3' sequence context and was found in skin cancers and a subset of colorectal tumors. Skin tumor samples are from TCGA and COSMIC datasets that reported data for U.S. patients. The subset of colorectal samples is from a Chinese study of ICGC. Sig B matches COSMIC signature 7 corresponding to the effects of UV (cos=0.98) and signature 10 (cos=0.93) corresponding to the altered activity of the error-prone polymerase POLE.

A:T>T:A mutations contributing to Sig A and Sig B represent only 3% and 2% respectively of the overall mutation load in their samples of origin (Suppl table 1). Therefore, these two signatures are not characteristic of a marked enrichment for A:T>T:A and are thus likely to be the result of a weaker mutagenic stimulus in these samples.

The third signature, Sig C, represented by T>A within the 5'-NpTpA-3' context was found exclusively in a subset of 20 kidney samples from TCGA and 9 liver samples from TCGA and 1 Chinese sample from COSMIC. Sig C exhibits a transcriptional strand bias for the sequence context 5'-CpTpA-3', and provides a perfect match (cos=1) with COSMIC signature 27 of unknown origin. Indeed, it is based on the same subset of kidney tumor samples from the U.S. patients, from which the signature 27 originates. In the liver cancer samples, A:T>T:A represented only 5% of mutations, whereas in the kidney cancer samples A:T>T:A represented 30%, suggesting the involvement of at least one potent mutagenic factors causing these alterations.

The fourth signature, Sig D, characterized by T>A in the 5'-NpTpG-3' context was found in carcinomas of the lung (the samples are from TCGA and COSMIC studies), liver (all the samples are from Asian patients from TCGA, COSMIC and ICGC studies), head and neck (all the samples are from American patients), bladder (all the samples are from a Chinese study of ICGC), esophagus (the samples are from Chinese patients from ICGC and COSMIC studies) and UTUC (the samples correspond to the AA-exposure positive UTUC dataset). This signature exhibits a transcriptional strand bias for all 5'-NpTpG-3' sequence contexts. Sig D matches several signatures, including COSMIC signature 22 (cos=0.91) corresponding to the mutagenic effects of AA, COSMIC signature 4 (cos=0.95) associated with smoking, and experimental signatures of carcinogens compounds such as AA (cos=0.91),

urethane (cos=0.94) and DMBA (cos=0.96) and to a lesser extent BaP (cos=0.88). These multiples matches are consistent with the contribution of several tumor types found in Sig D and also indicate that multiple carcinogens may preferentially target the 5'-CpTpG-3' sequence context. This could reflect the fact that CAG (CTG) is the most prominent trinucleotide in both human and mouse exomes and is thus more likely to be targeted by reactive metabolites of carcinogens.

In terms of the relative mutation contribution, UTUC samples had the highest number of the A:T>T:A with an overall 72% contribution to the entire mutation load, followed by kidney samples (30% contribution) and liver samples (20% contribution). These UTUC and most of the kidney samples are from studies addressing exposure to aristolochic acid, which is the likely mutagenic factor underlying Sig D in these tumors. The proportion of A:T>T:A among all substitutions in the lung, bladder and esophageal tumors is around 10%.

In regards to the two signatures (Sig C and D) marked with an enrichment of A:T>T:A transversions, we further investigated the possible origins of such rare mutations. For the signature D found in various tumor types, the enrichment of A:T>T:A in liver (28 samples), kidney (2 samples) and bladder (5 samples) tumor samples may be due to exposure and mutagenic effects of AA present in certain preparations of the Traditional Chinese Medicine (TCM), as recently reported [6,7,10,11,20,21] and as demographical information about those patients revealed that they were all of Chinese origin (a region of high AA exposure due to wide spread use of TCM). Regarding samples of lung and esophageal tumors (147 and 7 samples respectively), the AA exposure is very unlikely as these samples come from regions where AA has been banned. In these cases, A:T>T:A transversions may be caused by other exposures such as carcinogens present in the tobacco smoke. Indeed, these tumor types are known tobacco-related cancers and a number of the hundreds of compounds present in tobacco smoke may cause A:T>T:A transversions. Urethane is one candidate, as it has been shown to induce lung tumors carrying C:G>T:A and A:T>T:A mutations in the *cII* reporter gene [22], and, more recently, a genome-scale analysis [23] in mouse models of carcinogenesis showed that urethane-exposed mice developed lung adenocarcinomas which carry a high proportion of A:T>T:A mutations in the predominant sequence context 5'-CpApG-3'. Another candidate source of A:T>T:A transversions in the tobacco smoke is BaP.

Although BaP is predominantly inducing C:G>A:T mutations, it also causes A:T>T:A mutations in a similar sequence context in *in vitro* mouse-models, although A:T>T:A represent less than 9% of all mutations [24,25]. Sig D has the highest similarity (cos=0.96) with the A:T>T:A component of the spectrum generated by the research compound DMBA, although genome-wide studies showed that *in vivo* exposure to DMBA induced concomitant C:G>A:T mutations [14,15]. DMBA is used to induce skin cancers in experimental models and general human exposure is unlikely.

The signature Sig C found enriched in a subset of kidney tumors from the U.S, is marked by A:T>T:A transversions in the 5'-NpTpA-3' context. Previous cohort of patients and experimental animal studies have shown that psoralens, a natural compounds used in combination with UV radiation (PUVA therapy) to treat skin conditions such as vitiligo or psoriasis (e.g. methoxsalen), can exert similar mutations [26,27] originating from promutagenic covalent mono-adducts, di-adduct and DNA intercalation/interstrand crosslinks on thymidine [28–30]. Interestingly, increased incidence of cancer of the kidney and other organs has been reported in rats exposed to 8-methoxypsoralen [31] and a case report described an occurrence of renal carcinoma in a patient with a PUVA therapy history [32]. In humans, squamous cell carcinoma of the skin has also been reported, due to the combined effects of methoxsalen and topical UVA radiation [33–35]. However, the causal links between exposure to psoralens, human kidney cancer and COSMIC signature 27 and Sig C identified here remain to be investigated.

Other compounds may induce A:T>T:A transversions. For example, occupational exposure to vinyl chloride, an IARC group 1 carcinogen [36] has been linked to the development of angiosarcomas of the liver (ASL) and hepatocellular carcinomas (HCC) in factory workers. Vinyl chloride has been shown to induce an A:T>T:A mutational pattern in the tumor suppressor gene TP53 in human and rat [37,38]. However, the genome-wide mutational pattern of vinyl chloride and other potential A:T>T:A inducing carcinogens remain to be better characterized in order to determine their similarity with the other A:T>T:A-based signatures.

Conclusions

Here we identified four distinct A:T>T:A-based signatures and linked them to different cancer types and candidate causal processes. One of the signatures, Sig D, previously related exlusively to the exposure to AA, overlaps in its sequence context pattern with several additional A:T>T:A signatures and may thus reflect the effects of distinct carcinogens. Signature Sig D may thus result from mutagenic effects of compounds that act through promutagenic adduct formation. We propose that A:T>T:A mutation profile alone is thus generally not specific to AA exposure. In contrast, this profile combined with a high load of A:T>T:A mutations, due to high mutagenicity of AA, may be a more reliable biomarker of AA exposure, although further validations and analyses are warranted that would include more samples with documented exposure to AA. Identification of AL-DNA adducts by liquid chromatography-mass spectrometry provides the most specific and reliable biomarker of exposure to AA and should be performed alongside the mutational signature analyses.

Methods

Data sources and processing

Somatic mutation data obtained from whole exome sequencing studies were retrieved from public cancer genomics repositories and original studies. The Cancer Genome Atlas (TCGA) consensus MAF files from November 2016 were used. For International Cancer Genome Consortium (ICGC) data, lists of variants from version 20 were downloaded. For the Catalogue Of Somatic Mutations In Cancer (COSMIC), the list of coding mutations from release 76 (file CosmicMutantExport.tsv) was used. Samples from aristolochic acid exposure studies from Taiwan, China [4,5] and from the BEN region [10,39] were included in the analysis as positive control for aristolochic acid mutational signature. As COSMIC and ICGC datasets include TCGA data, we excluded TCGA samples using TCGA sample identifiers to avoid redundancy with TCGA dataset. Because COSMIC and ICGC datasets also include whole-genome data, samples with more than 10,000 mutations, likely to be derived from whole-genome, were removed from these datasets. To further eliminate duplicate samples, samples sharing more than 10% of mutations were considered as redundant and removed from the analysis. All datasets were annotated with the strand orientation and the gene name recovered from RefGene database and with the trinucleotide sequence context (immediate 5' and 3' sequence context extracted from GRCh37/hg19 reference genome) using the MutSpec-Annot tool [40].

Mutation signature analysis

We used the R package NMF [41] with the algorithm of Brunet and the Kullback-Leibler divergence penalty to extract mutational signatures from A:T>T:A mutations (only 16 parameters were considered, reflecting 16 possible trinucleotide contexts). A number of iterations of 200 was used in order to achieve stability for extracting A:T>T:A mutational signatures. NMF decomposes the mixture patterns of the input matrix (n x m) into two matrices W (n x r) and H (r x m) under the constraint of non-negativity [42]. The parameter r corresponds to the factorisation rank (i.e. number of signatures) to be extracted and it is thus necessary to first estimate it. For this we ran NMF using different rank values (between 2 and 8) and 50 iterations and computed

two statistics for evaluating the quality measures of the result: the cophenetic coefficient and the rss curve. The rank is chosen by taking the first value for which the cophenetic coefficient starts to decrease as proposed by Brunet et al. [43] and where the rss curve shows an inflection point as suggested by Hutchins et al. [44]. In order to further evaluate the stability of the A:T>T:A-based mutational signatures obtained and to account for over-fitting, we conducted 200 permuted NMF factorizations. This involves independent permutations of the rows (i.e. somatic mutations) for each column (i.e. cancer samples) of the original data. For each of 200 such random samples we re-ran the NMF factorization as it has been done for the original data. The quality measures (the cophenetic coefficient, the residuals and the explained variance) were estimated for the randomized data and compared to the ones obtained using the original data.

Comparison with published signatures

We used the cosine similarity method implemented in the LSAfun R package [45] for comparing the A:T>T:A-based mutational signatures extracted and the A:T>T:A-component of the reference signatures. Each signature was represented as a vector and the similarity between the two vectors was measured by calculating the cosine of the angle between them. The cosine values obtained ranged between 0 and 1, corresponding to an absence of similarity to a perfect match. We considered cosine values above 0.88 as a good match. The reference signatures included the 30 signatures extracted from human cancers and publicly available in COSMIC [46], 4 experimental signatures from in vitro MEF models exposed to four carcinogens (aristolochic acid (AA), activation-induced cytidine deaminase (AID), benzo[a]pyrene (B[a]P) and N-methyl-N'-nitro-N-nitrosoguanidine (MNNG)) [24] and 2 signatures from in vivo mouse models exposed to urethane [23] or 9,10-dimethyl-1,2-benzanthracene (DMBA) [14].

Acknowledgments This work was supported by the International Agency for Research on Cancer regular budget.

Figure legends

Figure 1: Characteristics of the sample sets analyzed.

(A) Distribution of samples across tumors types and data sources. (B) Overall mutation rates across tumor types. (C) A:T>T:A mutation rates across tumor types. The red horizontal bars indicate the median number of mutations in each tumor type. The y-axis is log2 scaled.

Figure 2: A:T>T:A enrichment compared to total SBS counts

Count of A:T>T:A transversions are plotted against the total number of mutations in each sample. Only tumor types with a calculated enrichment of A:T>T:A are represented in this plot and all other tumor types are combined in the "Other" category.

Samples with more than 300 A:T>T:A are labeled, except for the UTUC samples.

Figure 3: A:T>T:A signatures.

Left panel: A:T>T:A-based mutational signatures determined by NMF for A:T>T:A enriched samples. The x-axis represents the trinucleotide sequence context, with the 5' flanking base in the first row and the 3' flanking base in the second row and the y-axis represents the contribution of the signature expressed as a percentages.

Middle panel: Cosine similarity matrix between NMF signatures and previously described signatures (see Methods). Cosine values are expressed as percentages and only the ones with a match above 88% are shown.

Right panel: Contribution by tumor types to the individual signatures extracted with NMF. The x-axis represents the contribution expressed as a percentage of the median of A:T>T:A mutations contributing to each signature for each tumor type. The y-axis in the left list the tumor types and the y-axis on the right list the number of samples considered for each tumor types.

Supplementary Figure 1: Mutation spectrum and transcriptional strand bias for samples enriched in A:T>T:A

Bar graph in the left represent mutation types distribution within their trinucleotide sequence context.

Bar graph on the right represents the strand bias ratios (N: non-transcribed strand , T: transcribed strand).

Supplementary Figure 2: Overview of the samples considered for signature analysis

Distribution of samples with at least 32 A:T>T:A across tumors types and data sources.

Supplementary Figure 3: Comparison of A:T>A:T signatures

A:T>T:A signatures were extracted using the pmsignature package and the result is compared with the signatures obtained using the NMF method.

Supplementary Figure 4: Statistics evaluating the stability of the signatures

Supplementary Table 1: Samples contribution to each signature obtained

The contribution was calculated as a percentage of A:T>T:A mutations contributing to each signatures.

References

1. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8.

2. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.

3. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502:333–9.

4. Hoang ML, Chen C-H, Sidorenko VS, He J, Dickman KG, Yun BH, et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. Sci. Transl. Med. 2013;5:197ra102.

5. Poon SL, Pang S-T, McPherson JR, Yu W, Huang KK, Guan P, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. Sci. Transl. Med. 2013;5:197ra101.

6. Scelo G, Riazalhosseini Y, Greger L, Letourneau L, Gonzàlez-Porta M, Wozniak MB, et al. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. Nat. Commun. [Internet]. 2014 [cited 2014 Nov 14];5. Available from: http://www.nature.com/ncomms/2014/141029/ncomms6135/full/ncomms6135.html

7. Poon SL, Huang MN, Choo Y, McPherson JR, Yu W, Heng HL, et al. Mutation signatures implicate aristolochic acid in bladder cancer development. Genome Med. 2015;7:38.

8. Nortier JL, Martinez MC, Schmeiser HH, Arlt VM, Bieler CA, Petein M, et al. Urothelial carcinoma associated with the use of a Chinese herb (Aristolochia fangchi). N. Engl. J. Med. 2000;342:1686–92.

9. Lemy A, Wissing KM, Rorive S, Zlotta A, Roumeguere T, Muniz Martinez M-C, et al. Late onset of bladder urothelial carcinoma after kidney transplantation for end-stage aristolochic acid nephropathy: a case series with 15-year follow-up. Am. J. Kidney Dis. Off. J. Natl. Kidney Found. 2008;51:471–7.

10. Jelaković B, Castells X, Tomić K, Ardin M, Karanović S, Zavadil J. Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid. Int. J. Cancer J. Int. Cancer. 2015;136:2967–72.

11. Zou S, Li J, Zhou H, Frech C, Jiang X, Chu JSC, et al. Mutational landscape of intrahepatic cholangiocarcinoma. Nat. Commun. 2014;5:5696.

12. Sung W-K, Zheng H, Li S, Chen R, Liu X, Li Y, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat. Genet. 2012;44:765–9.

13. Sidorenko VS, Yeo J-E, Bonala RR, Johnson F, Schärer OD, Grollman AP. Lack of recognition by global-genome nucleotide excision repair accounts for the high mutagenicity and persistence of aristolactam-DNA adducts. Nucleic Acids Res. 2012;40:2494–505.

14. Nassar D, Latil M, Boeckx B, Lambrechts D, Blanpain C. Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. Nat. Med. 2015;21:946–54.

15. McCreery MQ, Halliwill KD, Chin D, Delrosario R, Hirst G, Vuong P, et al. Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. Nat. Med. 2015;21:1514–20.

16. Chakravarti D, Pelling JC, Cavalieri EL, Rogan EG. Relating aromatic hydrocarbon-induced DNA adducts and c-H-ras mutations in mouse skin papillomas: the role of apurinic sites. Proc. Natl. Acad. Sci. U. S. A. 1995;92:10422–6.

17. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans Volume 100F. International Agency for Research on Cancer; 2012.

18. Swenberg JA, Fedtke N, Ciroussel F, Barbin A, Bartsch H. Etheno adducts formed in DNA of vinyl chloride-exposed rats are highly persistent in liver. Carcinogenesis. 1992;13:727–9.

19. Shiraishi Y, Tremmel G, Miyano S, Stephens M. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. PLoS Genet. 2015;11:e1005657.

20. Totoki Y, Tatsuno K, Covington KR, Ueda H, Creighton CJ, Kato M, et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nat. Genet. [Internet]. 2014 [cited 2014 Nov 6];advance online publication. Available from: http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3126.html

21. Hoang ML, Chen C-H, Chen P-C, Roberts NJ, Dickman KG, Yun BH, et al. Aristolochic acid in the etiology of renal cell carcinoma. Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. 2016;

22. Hernandez LG, Forkert P-G. In vivo mutagenicity of vinyl carbamate and ethyl carbamate in lung and small intestine of F1 (Big Blue x A/J) transgenic mice. Int. J. Cancer J. Int. Cancer. 2007;120:1426–33.

23. Westcott PMK, Halliwill KD, To MD, Rashid M, Rust AG, Keane TM, et al. The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. Nature. 2015;517:489–92.

24. Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallée MP, et al. Modelling mutational landscapes of human cancers in vitro. Sci. Rep. 2014;4:4482.

25. Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM, et al. The genome as a record of environmental exposure. Mutagenesis. 2015;gev073.

26. Besaratinia A, Pfeifer GP. Biological consequences of 8-methoxypsoralen-photoinduced lesions: sequence-specificity of mutations and preponderance of T to C and T to a mutations. J. Invest. Dermatol. 2004;123:1140–6.

27. Lambertini L, Surin K, Ton T-VT, Clayton N, Dunnick JK, Kim Y, et al. Analysis of p53 tumor suppressor gene, H-ras protooncogene and proliferating cell nuclear antigen (PCNA) in squamous cell carcinomas of HRA/Skh mice following exposure to 8-methoxypsoralen (8-MOP) and UVA radiation (PUVA therapy). Toxicol. Pathol. 2005;33:292–9.

28. Johnston BH, Hearst JE. Low-level psoralen--deoxyribonucleic acid cross-links induced by single laser pulses. Biochemistry (Mosc.). 1981;20:739–45.

29. Tessman JW, Isaacs ST, Hearst JE. Photochemistry of the furan-side 8-methoxypsoralen-thymidine monoadduct inside the DNA helix. Conversion to diadduct and to pyrone-side monoadduct. Biochemistry (Mosc.). 1985;24:1669–76.

30. Cadet J, Voituriez L, Nardin R, Viari A, Vigny P. A new class of psoralen photoadducts to DNA components: isolation and characterization of 8-MOP adducts to the osidic moiety of 2'-deoxyadenosine. J. Photochem. Photobiol. B. 1988;2:321–39.

31. National Toxicology Program. Toxicology and Carcinogenesis Studies of 8-Methoxypsoralen (CAS No. 298-81-7) in F344/N Rats (Gavage Studies). Natl. Toxicol. Program Tech. Rep. Ser. 1989;359:1–130.

32. Zarur FP, d'Almeida LFV, Mafort MSP, Gusmão PR de, Avelleira JCR. Two cases of renal cell cancer during immunobiologic therapy for psoriasis. An. Bras. Dermatol. 2014;89:1017–8.

33. Forman AB, Roenigk HH, Caro WA, Magid ML. Long-term follow-up of skin cancer in the PUVA-48 cooperative study. Arch. Dermatol. 1989;125:515–9.

34. Chuang TY, Heinrich LA, Schultz MD, Reizner GT, Kumm RC, Cripps DJ. PUVA and skin cancer. A historical cohort study on 492 patients. J. Am. Acad. Dermatol. 1992;26:173–7.

35. Lindelöf B, Sigurgeirsson B, Tegner E, Larkö O, Johannesson A, Berne B, et al. PUVA and cancer risk: the Swedish follow-up study. Br. J. Dermatol. 1999;141:108–12.

36. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans Volume 100A [Internet]. Lyon: International Agency for Research on Cancer; 2012. Available from: http://monographs.iarc.fr/ENG/Monographs/vol100A/

37. Hollstein M, Marion MJ, Lehman T, Welsh J, Harris CC, Martel-Planche G, et al. p53 mutations at A:T base pairs in angiosarcomas of vinyl chloride-exposed factory workers. Carcinogenesis. 1994;15:1–3.

38. Barbin A, Froment O, Boivin S, Marion MJ, Belpoggi F, Maltoni C, et al. p53 gene mutation pattern in rat liver tumors induced by vinyl chloride. Cancer Res. 1997;57:1695–8.

39. Castells X, Karanović S, Ardin M, Tomić K, Xylinas E, Durand G, et al. Low-coverage exome sequencing screen in formalin-fixed paraffin-embedded tumors reveals evidence of exposure to carcinogenic aristolochic acid. Cancer Epidemiol. Biomarkers Prev. 2015;cebp.0553.2015.

40. 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.

41. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010;11:367.

42. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401:788–91.

43. 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;101:4164–9.

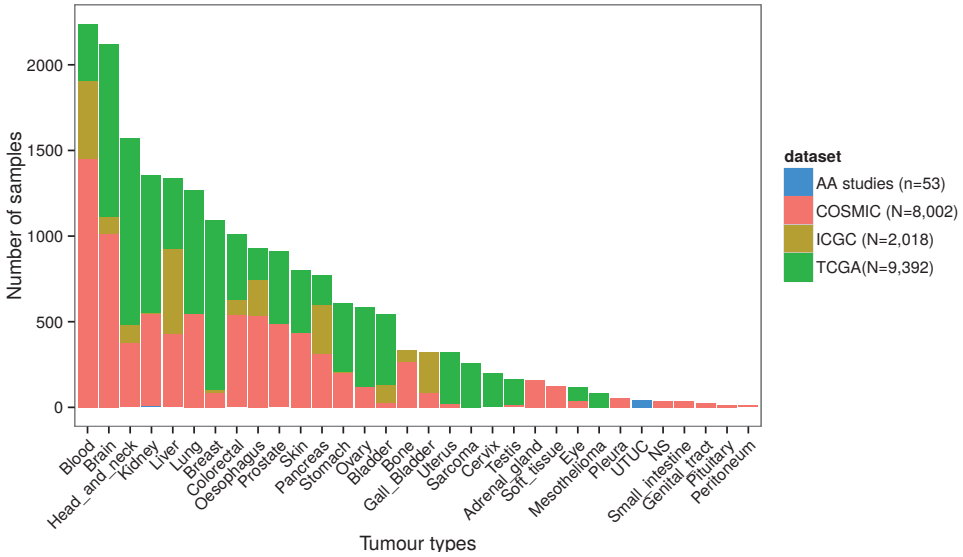
44. Hutchins LN, Murphy SM, Singh P, Graber JH. Position-dependent motif characterization using non-negative matrix factorization. Bioinforma. Oxf. Engl. 2008;24:2684–90.

45. Günther F, Dudschig C, Kaup B. LSAfun--An R package for computations based on Latent Semantic Analysis. Behav. Res. Methods. 2015;47:930–44.

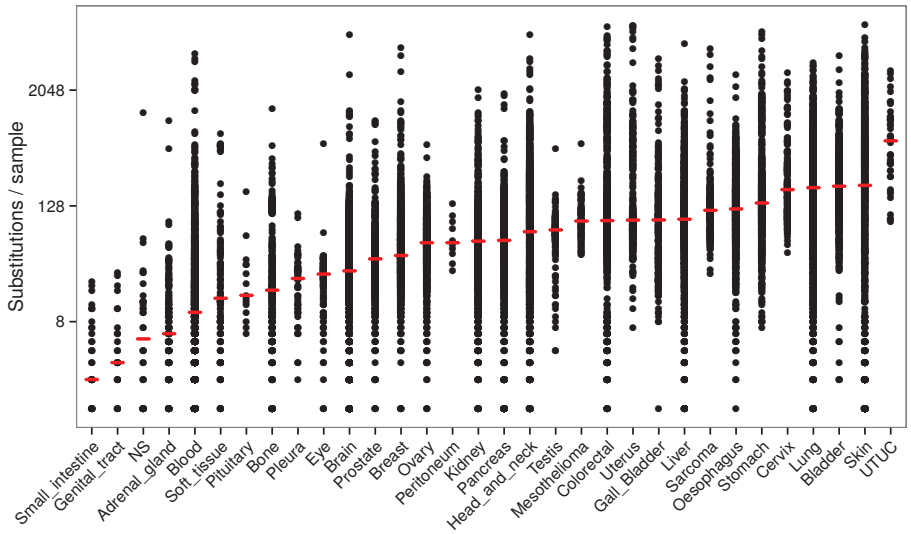
46. Petljak M, Alexandrov LB. Understanding mutagenesis through delineation of mutational signatures in human cancer. Carcinogenesis. 2016;bgw055.

Figure 1

A)


Number of samples
2000
1500
1000
500
0
dataset
AA studies (n=53)
COSMIC (N=8,002)
ICGC (N=2,018)
TCGA(N=9,392)
Blood
Brain
Head_and_neck
Kidney
Liver
Lung
Breast
Colorectal
Oesophagus
Prostate
Skin
Pancreas
Stomach
Ovary
Bladder
Bone
Gall_Bladder
Uterus
Sarcoma
Cervix
Testis
Adrenal_gland
Soft_tissue
Eye
Mesothelioma
Pleura
UTUC
NS
Small_intestine
Genital_tract
Pituitary
Peritoneum
Tumour types

B)


Substitutions / sample
2048
128
8
Small_intestine
Genital_tract
NS
Adrenal_gland
Blood
Soft_tissue
Pituitary
Bone
Pleura
Eye
Brain
Prostate
Breast
Ovary
Peritoneum
Kidney
Pancreas
Head_and_neck
Testis
Mesothelioma
Colorectal
Uterus
Gall_Bladder
Liver
Sarcoma
Oesophagus
Stomach
Cervix
Lung
Bladder
Skin
UTUC

C)

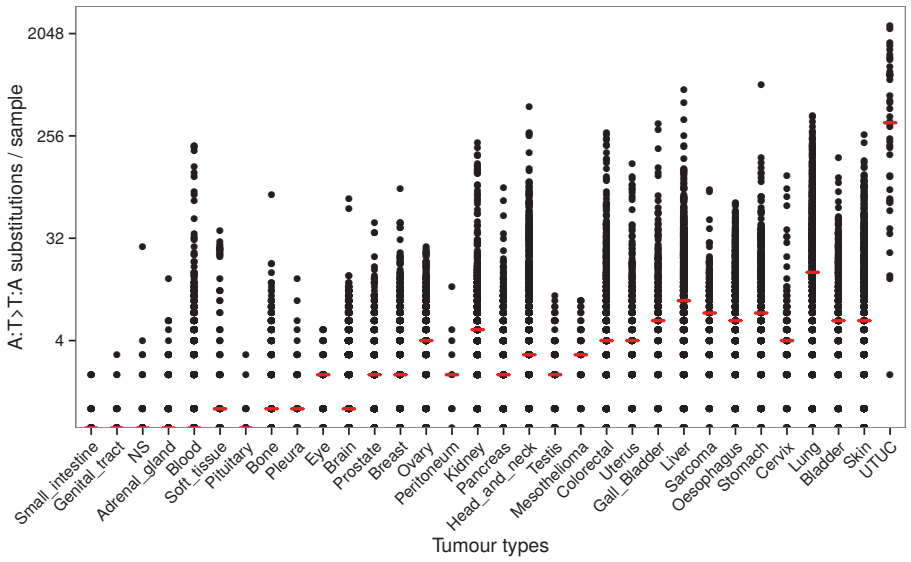

A:T>T:A substitutions / sample
2048
256
32
4
Small_intestine
Genital_tract
NS
Adrenal_gland
Blood
Soft_tissue
Pituitary
Bone
Pleura
Eye
Brain
Prostate
Breast
Ovary
Peritoneum
Kidney
Pancreas
Head_and_neck
Testis
Mesothelioma
Colorectal
Uterus
Gall_Bladder
Liver
Sarcoma
Oesophagus
Stomach
Cervix
Lung
Bladder
Skin
UTUC
Tumour types

Figure 2

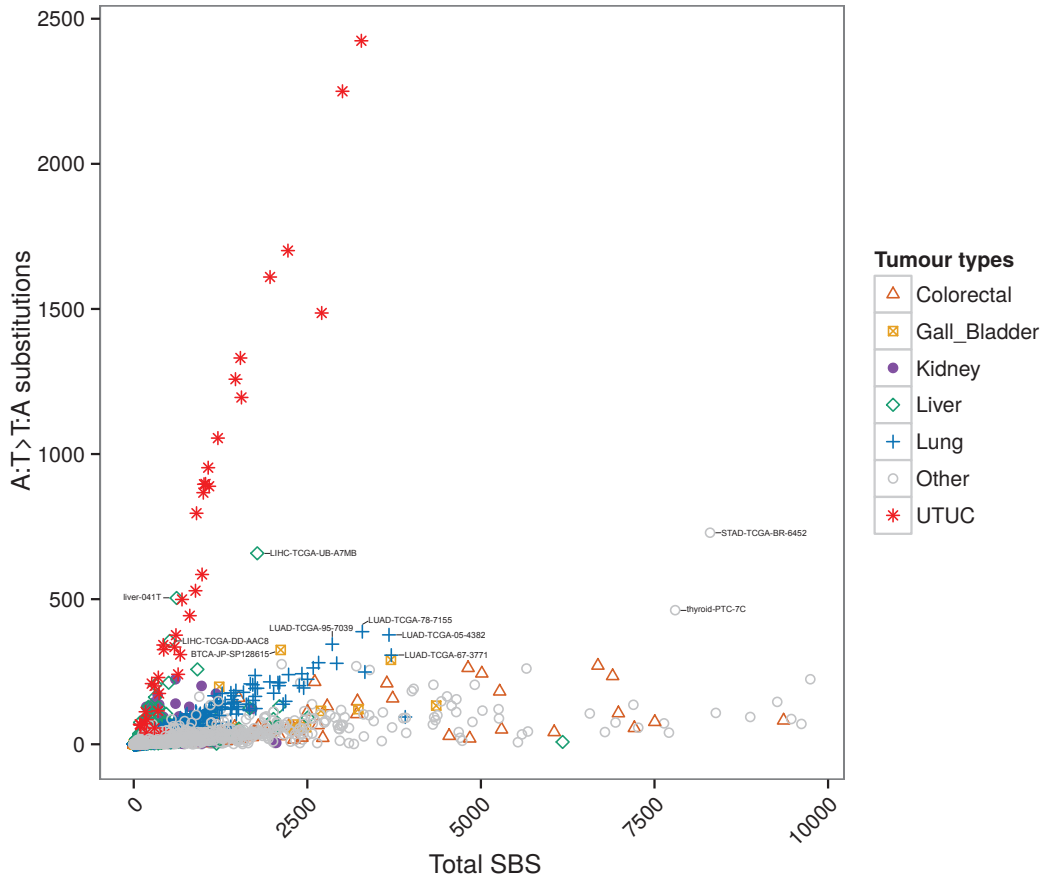
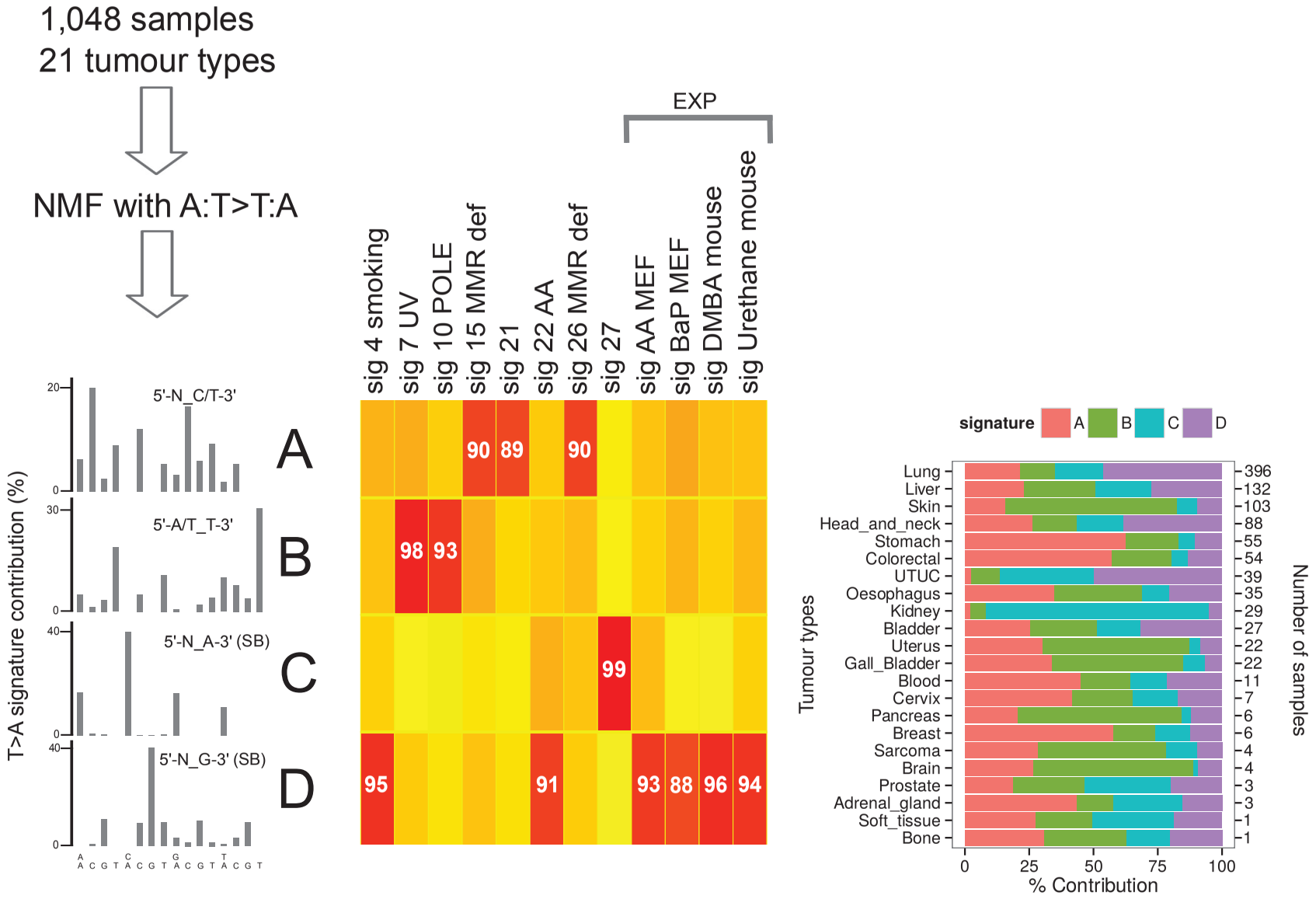
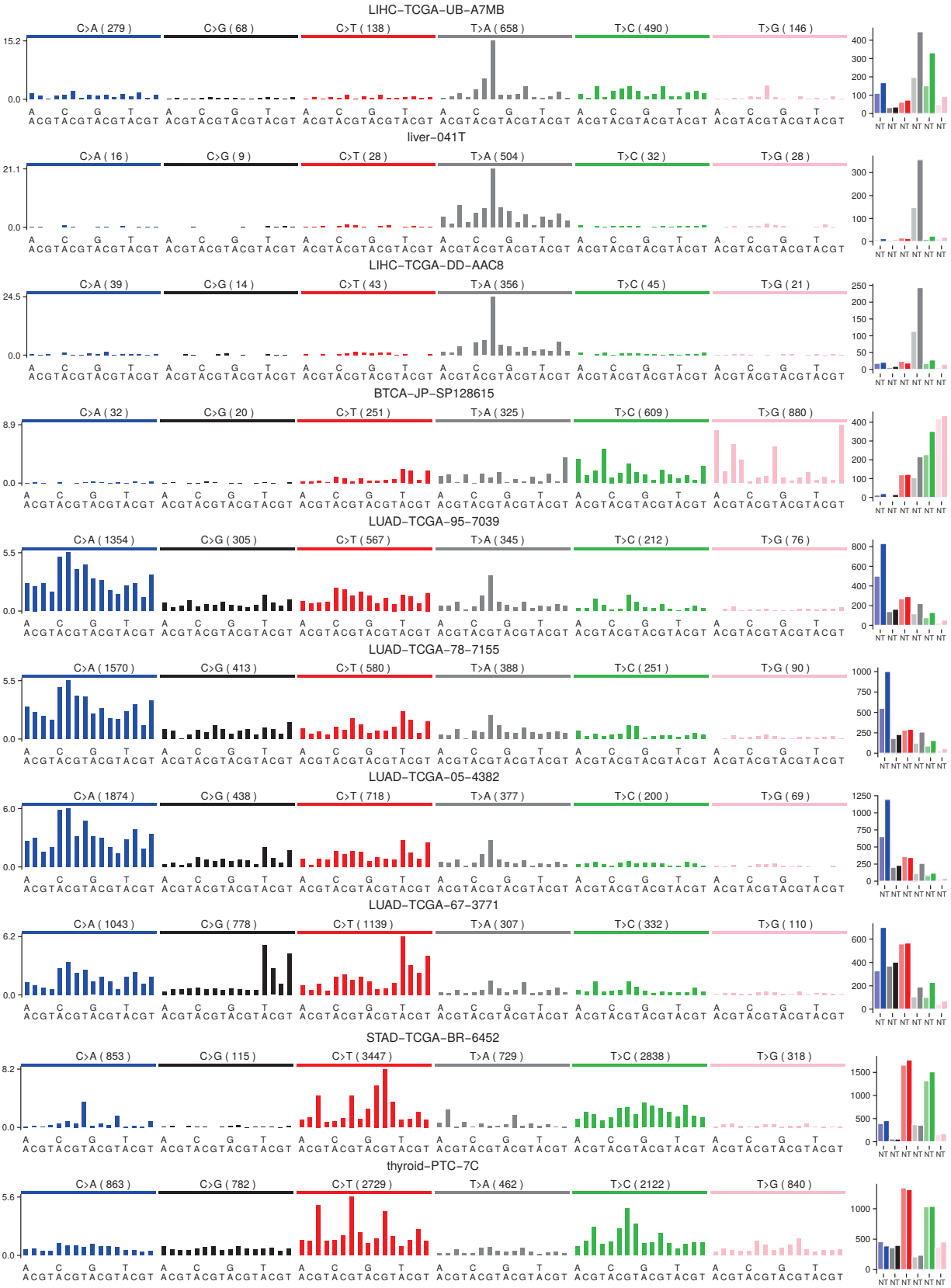

2500
2000
1500
1000
500
0
A:T>T:A substitutions
0
2500
5000
7500
10000
Total SBS
LIHC-TCGA-UB-A7MB
liver-041T
STAD-TCGA-BR-6452
thyroid-PTC-7C
LUAD-TCGA-78-7155
LUAD-TCGA-95-7039
LUAD-TCGA-05-4382
LUAD-TCGA-67-3771
LIHC-TCGA-DD-AAC8
BTCA-JP-SP128615
Tumour types
Colorectal
Gall_Bladder
Kidney
Liver
Lung
Other
UTUC

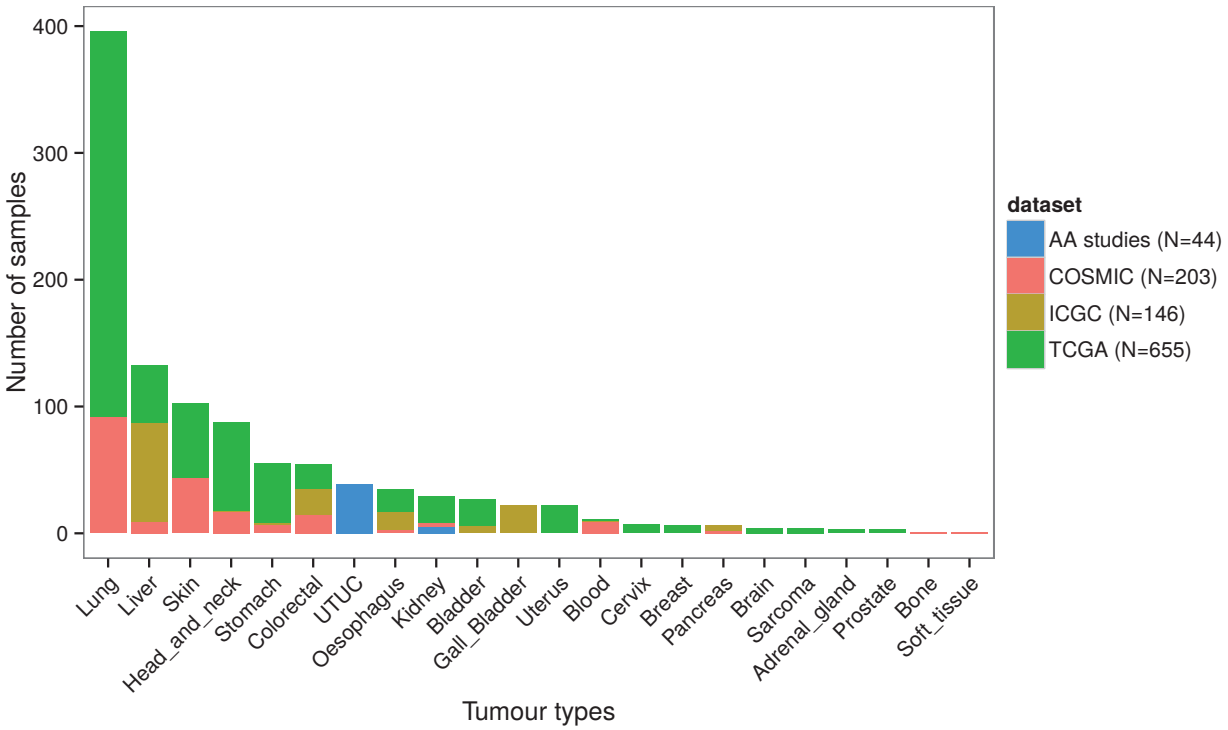
Figure 3


1,048 samples
21 tumour types
NMF with A:T>T:A
T>A signature contribution (%)
5'-N_C/T-3'
5'-A/T_T-3'
5'-N_A-3' (SB)
5'-N_G-3' (SB)
A
B
C
D
EXP
sig 4 smoking
sig 7 UV
sig 10 POLE
sig 15 MMR def
sig 21
sig 22 AA
sig 26 MMR def
sig 27
sig AA MEF
sig BaP MEF
sig DMBA mouse
sig Urethane mouse
90 89 90
98 93
99
95 91 93 88 96 94
signature A B C D
Tumour types
Lung 396
Liver 132
Skin 103
Head_and_neck 88
Stomach 55
Colorectal 54
UTUC 39
Oesophagus 35
Kidney 29
Bladder 27
Uterus 22
Gall_Bladder 22
Blood 11
Cervix 7
Pancreas 6
Breast 6
Sarcoma 4
Brain 4
Prostate 3
Adrenal_gland 3
Soft_tissue 1
Bone 1
Number of samples
0 25 50 75 100
% Contribution

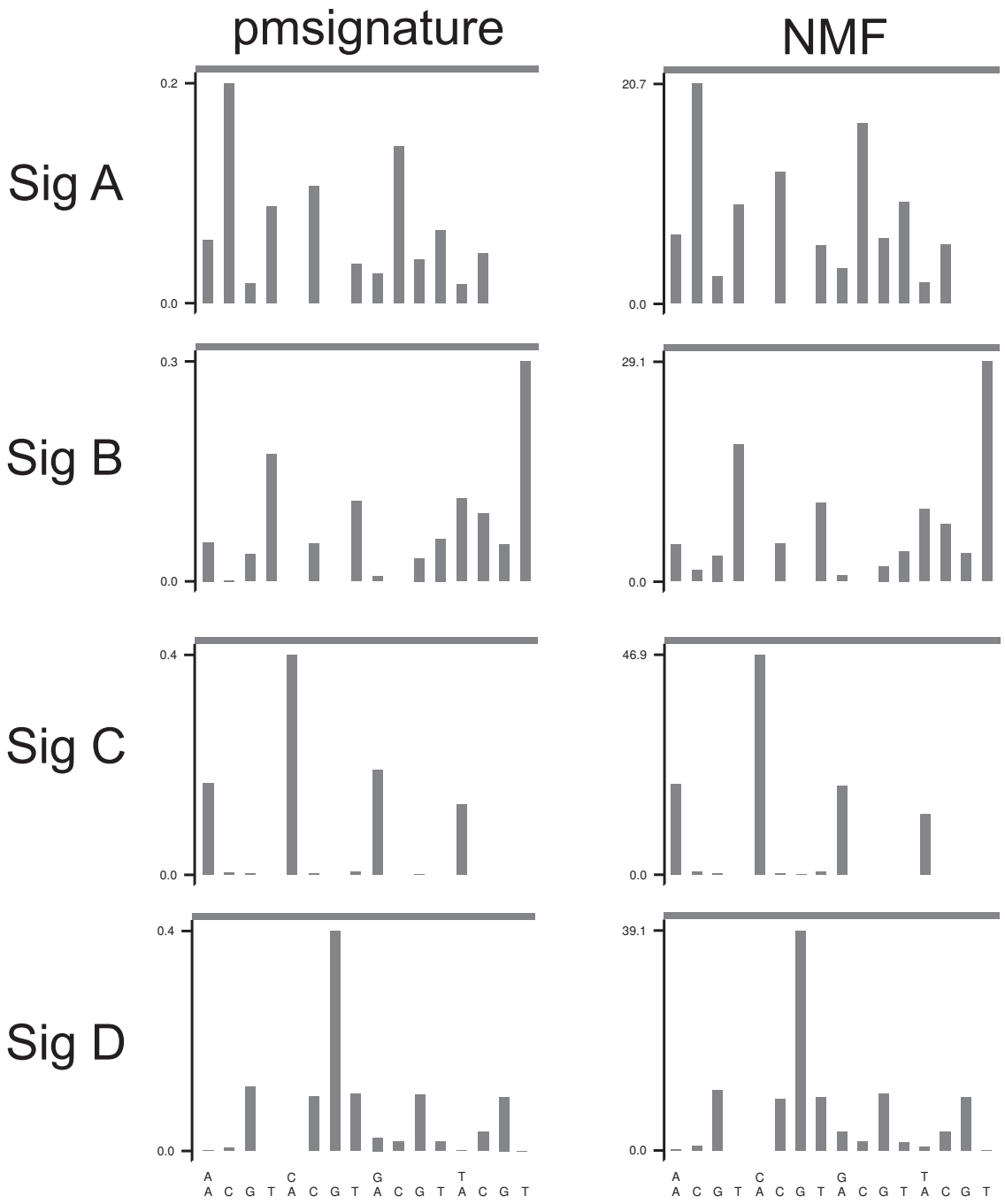
Supplementary Figure 1



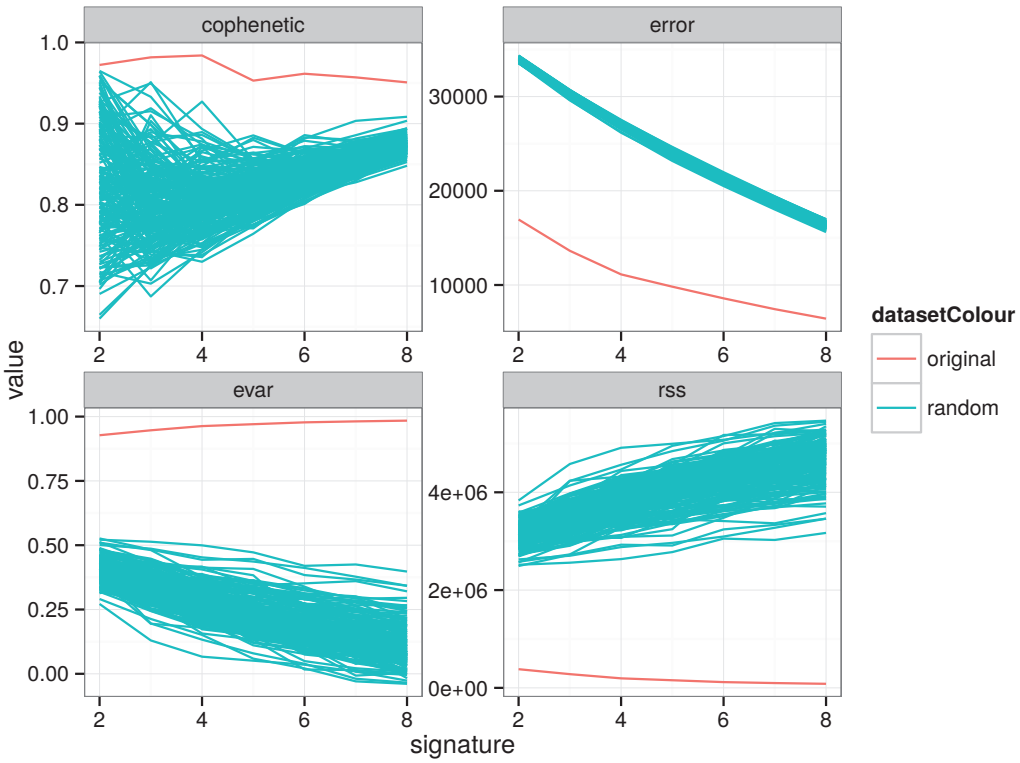
Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 4



DISCUSSION

&

CONCLUSION

DISCUSSION AND CONCLUSION

The analysis of cancer genomes at a genome-wide scale to identify mutational signatures in human cancers provides a unique opportunity to link mechanisms of carcinogenesis with the underlying cancer aetiology. Identifying the environmental or lifestyle factors behind the signatures patterns provides evidence-base rational for relevant preventive measures in populations at risk. While the interest in such analyses is growing in the scientific community, the complex expert skills required to perform such analyses poses limits to a broader access by the scientific community.

During my thesis, I addressed this limitation by developing tools and methods for facilitating the identification of carcinogen-specific mutational signatures from next-generation sequencing (NGS) data that can be used by non-bioinformatician scientists at IARC and externally. Mutational signatures of human carcinogens were analysed using data obtained from mutagenesis experiments and human tumours with known history of carcinogenic exposure. I also focused on the mutational signature of aristolochic acid (AA) in exposed patients from the BEN regions and characterised the specificity of this unique signature.

Development of bioinformatics tools

Extracting mutational signatures from NGS data includes several steps and implies the use of bioinformatics software not easily accessible to biologists. Generally, extracting mutational signatures requires annotating the variants for recovering the immediate flanking bases (the trinucleotide sequence context) surrounding the variants and the transcriptional strand orientation, which enables to estimate and extract signatures and to compare them with the published signatures. Recently, several tools were developed using the R programming language for extracting mutational signatures from NGS data: somaticSignatures [113], pmsignature [114], signeR [115] and BayesNMF developed by the Broad Institute (`http://archive.broadinstitute.org/cancer/cga/msp`). Other packages were also developed using the C programming language such as EMu [116],

or using the proprietary programming language MATLAB such as the WTSI Mutational Signature Framework [42].

These tools propose to use VCF files, to retrieve the trinucleotide sequence context of the list of variants and to extract mutational signatures. However, the input format required by these tools is not easily obtained from the original VCFs generated by the various variant callers. The input format should be a single text tabular files combining all the informations (chromosomal location, reference and alternate changes) on the variants for the different samples. The order of these informations differ between the tools, complicating the conversion of VCF into the correct input format. Furthermore, none of these tools propose a comparison with existing signatures and their availability as R, C or MATLAB packages limits the use to scientists with programming skills.

In order enable easy analyses of mutational signatures to all scientists, we have developed a suite of tools named MutSpec, integrated within the web-based platform Galaxy. MutSpec tools are compatible with the VCFs files generated by most commonly-used variant callers in the cancer research field and combine annotation of the variants, the estimation and extraction of signatures and the comparison with a predefined list of published signatures or a custom user-defined list.

MutSpec tools are now freely available on the Galaxy toolshed, at `https://toolshed.g2.bx.psu.edu/`, and were successfully downloaded more than 30 times. The tools are routinely used by MMB group members as well as others scientists at IARC and have received positive feedbacks from users for the identification and interpretation of carcinogen-specific mutational signatures from NGS experiments.

Currently, only SBS are considered in mutational signatures but many other alterations occur in cancer genomes including indels, copy number alterations and structural variations, some of which form new types of non-SBS signatures [117]. These additional features could be incorporated into mutational signatures definition for providing a broader characterisation of specific carcinogens fingerprints on the genome. Furthermore, only the immediate flanking bases surrounding the mutations are currently considered in mutational signatures analysis. Incorporating additional bases could potentially help to differentiate compounds acting via similar mechanisms, even though the statistical power may decrease as the number of parameters will increase.

MutSpec tools could be adapted for inclusion of additional features/modules that will reflect the expanded definition of mutational signatures.

Characterisation of carcinogens mutational signatures

Experimental, preferably non-animal models are needed to characterise carcinogen-specific mutational signatures observed in human genomes as they provide mechanistic clues on cancer aetiology. Defining precisely carcinogen mutational signatures may also support the classification of carcinogens and the establishment of cancer prevention strategies.

One of these experimental models used in this work is based on Hupki mouse carrying a partial (exons 4-9) human *TP53* replacing the corresponding *Trp53* mouse exons. This model demonstrated its ability to recapitulate *TP53* mutation patterns observed in human tumours [118]. However, single-gene analysis requires extensive pooling of data from multiple experiments, as typically only one mutation is present in each sample, and they are not informative enough for interpreting the complex mutation patterns resulting from multiple exposures. To overcome this limitation, we demonstrated for the first time that exome-wide mutational signatures specific of known human carcinogens could be recapitulated in the MEF assay, expanding the information contents gained from *TP53* analysis. Furthermore, a whole-genome sequencing (WGS) analysis confirmed our results obtained with WES and provided the characterisation of additional alterations (indels and structural rearrangements) following carcinogens exposures [50].

The Hupki-MEF assay is thus a simple and powerful system for modelling cancer mutation landscapes of human tumours and will continue to be used by the MMB group to identify new carcinogen-specific signatures and formulate new hypotheses on the aetiology of human cancers.

To bypass the limitation of the differences between human and mouse organisms in metabolism and DNA repair mechanisms, the MMB group started exposures of carcinogens of interest in human liver progenitor cells (HepaRG). HepaRG cell lines have a greater capacity to metabolise chemical compounds making the model suitable for testing the genotoxicity of chemicals in human liver [119, 120]. Furthermore, to overcome the limitation of *in vitro* assays (metabolism, culture artefacts), the MMB group established

a collaboration with the U.S. NTP programme, that tested more than 600 compounds potentially carcinogenic in humans. This collaboration will allow us to collect tumours from *in vivo* bioassays of rodent (mice and rats) with clear evidence of carcinogenic effects of the same chemical exposures than the ones that will be tested *in vitro*.

Comparing carcinogen-specific mutational signatures from new *in vitro* and retrospective *in vivo* assays with human tumours data from cancer genomics repositories is expected to help defining the causal factors contributing to human tumours development. This unique multi-system approach should thus facilitate the understanding of the origin of cancer development and the establishment of public health policies.

Investigation of the aristolochic acid exposure

The investigation of exposure to the environmental and iatrogenic carcinogen aristolochic acid (AA) is important because potentially large populations are likely to be exposed through the consumption of traditional Chinese medecines, particularly in Asian countries [54] or through the inadvertent dietary contamination such as in the BEN region.

The detection of AA-derived aristolactam-DNA adducts accumulated in the renal cortex by mass spectrometry is routinely used as a biomarker of exposure. However, the detection is challenging in samples with low amount of DNA due to the lack of accessible protocols. In order to overcome this limitation we took advantage of the unique features of AA mutational signature for developing a low-coverage whole-exome sequencing (LC-WES) method to detect AA exposure in challenging archived samples with DNA of low amount and of poor quality. LC-WES approach demonstrated its ability to detect AA mutational signature in archived tumour samples from the BEN region, providing a new tool for molecular epidemiology for investigating AA exposure in human cancers in population at risk. This cost-effective method can be used in systematic screening studies of archived FFPE samples of AA-associated tumours or in new tumour types for revealing AA exposure. LC-WES proved to be an efficient method for detecting the AA exposure that can provide evidence for the establishment of prevention programs.

The investigation of AA in the aetiology of new tumour types in the BEN region using WES revealed the presence of AA mutational signature in two RCC subtypes. This result

highlights the importance to include additional tumour types in systematic screening programmes in order to improve the surveillance of AA related cancers.

For facilitating AA exposure detection in new UTUC cases, we aim to establish a targeted sequencing of 32 cancer driver genes selected after analysing recurrently mutated genes in three different studies of AA-associated tumours [68]. We started the targeted sequencing of these 32 genes in 15 UTUC samples from Korea in which the exposure was confirmed by DNA adducts analysis for some of the samples and in 45 UTUC samples from Japan were AA exposure is suspected, all this using the Illumina MiSeq platform available within MMB group. This approach enabled us to identify one sample with a characteristic AA signature. Further analysis is being conducted to determine if it is enough for capturing AA exposure and mutational signature.

Finally, in order to better characterise the AA mutational signature, we performed a systematic screening of A:T>T:A transversions in public repositories. Previously undescribed A:T>T:A mutational signatures were identified, as well as, new cases and target organs potentially associated with the AA exposure. This analysis revealed that a marked enrichment of A>T transversions preferentially located on the non-transcribed strand, appears to be specific of AA exposure. However, the predominant sequence context, 5'-CpApG-3', is also targeted by other carcinogens and is thus not specific to AA.

Currently, only AA has been found to induce this specific mutational signature characterised by a predominance of A>T preferentially located on the non-transcribed strand, but it is possible that other compounds acting through adduct targeting adenine bases could leave a similar pattern.

For example, additional compounds can target the 5'-CpApG-3' in a genome-wide manner, such as urethane [121], to which human exposure is likely, or the research compound dimethylbenzanthracene (DMBA) [122, 123]. However, unlike in the case of AA, in neither of these other scenarios the A:T>T:A transversions is the sole SBS component of the respective signatures.

In order to identify carcinogens potentially inducing A:T > T:A transversions, the MMB multi-system approach will be used. Data from *in vitro* and *in vivo* experimental mutagenesis in which the suspected carcinogens would have been tested, will be analysed and the resulting mutational signatures will be compared to AA.

To conclude, the MutSpec platform offers an easy-to-use framework for the analysis of mutational signatures from genome-wide data accessible through the user-friendly web-based platform Galaxy. Its free access by the scientific community is expected to facilitate the identification of new human carcinogens and the interpretation on how these mutations contribute to human carcinogenesis by a large community of researchers. Considering the worldwide use of traditional Chinese medicines, the detection of AA mutational signature provides an important molecular evidence for past AA exposure in human cancer genomes.

Furthermore, the establishment of a new method for detecting AA exposure in challenging archived tumour samples provides a cost-effective tool for screening programme of new tumour types.

To conclude this thesis work, analysing mutational signatures is important as it can provide invaluable insights into cancer aetiology, even though identifying the exact origins of the exposure remains a challenging task. Mutational signatures analysis continue to pose a great challenge for bioinformaticians in terms of identifying the mutational signatures of hundreds to thousands of potentially mutagenic factors to which we are exposed through our lifetime.

LIST OF FIGURES

&

LIST OF TABLES

List of Figures

List of Tables

BIBLIOGRAPHY

Bibliography

[1] Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. **2000**;100(1):57–70.

[2] Blasco MA. Telomeres and human disease: ageing, cancer and beyond. Nat Rev Genet. **2005**;6(8):611–622.

[3] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. **2011**;144(5):646–674.

[4] Vogelstein B, Kinzler KW. The multistep nature of cancer. Trends Genet. **1993**;9(4):138–141.

[5] Nowell PC. Tumor progression: a brief historical perspective. Semin Cancer Biol. **2002**;12(4):261–266.

[6] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. **2009**;458(7239):719–724.

[7] Stratton MR. Journeys into the genome of cancer cells. EMBO Mol Med. **2013**;5(2):169–172.

[8] Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. J Genet Genomics. **2011**;38(3):95–109.

[9] Marx V. Biology: The big challenges of big data. Nature. **2013**;498(7453):255–260.

[10] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. **2016**;17(6):333–351.

[11] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. **2015**;13(7).

[12] Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. **2016**;17:125.

[13] Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. **2010**;11(1):31–46.

[14] Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. **2010**;11(10):685–696.

[15] Yang Y, Dong X, Xie B, Ding N, Chen J, Li Y, et al. Databases and Web Tools for Cancer Genomics Study. Genomics Proteomics Bioinformatics. **2015**;13(1):46–50.

[16] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. **2011**;39:D945–950.

[17] Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. **2004**;91(2):355–358.

[18] Weinstein JN, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. **2013**;45(10):1113–1120.

[19] Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature. **2010**;464(7291):993–998.

[20] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery. **2012**;2(5):401–404.

[21] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. **2013**;6(269):pl1.

[22] Wogan GN, Hecht SS, Felton JS, Conney AH, Loeb LA. Environmental and chemical carcinogenesis. Semin Cancer Biol. **2004**;14(6):473–486.

[23] Han J, Haiman C, Niu T, Guo Q, Cox DG, Willett WC, et al. Genetic variation in DNA repair pathway genes and premenopausal breast cancer risk. Breast Cancer Res Treat. **2009**;115(3):613–622.

[24] Wu S, Powers S, Zhu W, Hannun YA. Substantial contribution of extrinsic risk factors to cancer development. Nature. **2016**;529(7584):43–47.

[25] Hyndman IJ. Review: the Contribution of both Nature and Nurture to Carcinogenesis and Progression in Solid Tumours. Cancer Microenviron. **2016**;9(1):63–69.

[26] Perera FP. Molecular Epidemiology: Insights Into Cancer Susceptibility, Risk Assessment, and Prevention. JNCI J Natl Cancer Inst. **1996**;88(8):496–509.

[27] Kemp CJ. Animal Models of Chemical Carcinogenesis: Driving Breakthroughs in Cancer Research for 100 Years. Cold Spring Harb Protoc. **2015**;2015(10):865–874.

[28] Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity/carcinogenicity experimental study designs and criteria used by the National Toxicology Program. Environ Health Perspect. **1990**;86:313–321.

[29] Monographs, editor. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. International Agency for Research on Cancer; **1988**.

[30] Tomatis L. The IARC monographs program: changing attitudes towards public health. Int J Occup Environ Health. **2002**;8(2):144–152.

[31] for Research on Cancer IA. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Preamble; **2006**.

[32] Pfeifer GP. How the environment shapes cancer genomes. Curr Opin Oncol. **2015**;27(1):71–77.

[33] Greenblatt MS, Bennett WP, Hollstein M, Harris CC. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer Res. **1994**;54(18):4855–4878.

[34] Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. Science. **1991**;253(5015):49–53.

[35] Hollstein M, Hergenhahn M, Yang Q, Bartsch H, Wang ZQ, Hainaut P. New approaches to understanding p53 gene tumor mutation spectra. Mutat Res. **1999**;431(2):199–209.

[36] Bouaoun L, Sonkin D, Ardin M, Hollstein M, Byrnes G, Zavadil J, et al. TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. Hum Mutat. **2016**;.

[37] Brash DE, Rudolph JA, Simon JA, Lin A, McKenna GJ, Baden HP, et al. A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. Proc Natl Acad Sci USA. **1991**;88(22):10124–10128.

[38] Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene. **2002**;21(48):7435–7451.

[39] Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb Perspect Biol. **2010**;2(1):a001008.

[40] Vogelstein B, Kinzler KW. Carcinogens leave fingerprints. Nature. **1992**;355(6357):209–210.

[41] Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. **2014**;24(100):52–60.

[42] Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. **2013**;500(7463):415–421.

[43] Petljak M, Alexandrov LB. Understanding mutagenesis through delineation of mutational signatures in human cancer. Carcinogenesis. **2016**;p. bgw055.

[44] Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. **2014**;15(9):585–598.

[45] Hsu IC, Metcalf RA, Sun T, Welsh JA, Wang NJ, Harris CC. Mutational hotspot in the p53 gene in human hepatocellular carcinomas. Nature. **1991**;350(6317):427–428.

[46] Bressac B, Kew M, Wands J, Ozturk M. Selective G to T mutations of p53 gene in hepatocellular carcinoma from southern Africa. Nature. **1991**;350(6317):429–431.

[47] Schulze K, Imbeaud S, Letouzé E, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet. **2015**;47(5):505–511.

[48] Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell. **2012**;150(6):1121–1134.

[49] Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallée MP, et al. Modelling mutational landscapes of human cancers in vitro. Scientific Reports. **2014**;4:4482.

[50] Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM, et al. The genome as a record of environmental exposure. Mutagenesis. **2015**;30(6):763–770.

[51] Kwak DH, Lee JH, Kim T, Ahn HS, Cho WK, Ha H, et al. Aristolochia manshuriensis Kom inhibits adipocyte differentiation by regulation of ERK1/2 and Akt pathway. PLoS ONE. **2012**;7(11):e49530.

[52] Nortier JL, Martinez MC, Schmeiser HH, Arlt VM, Bieler CA, Petein M, et al. Urothelial carcinoma associated with the use of a Chinese herb (Aristolochia fangchi). N Engl J Med. **2000**;342(23):1686–1692.

[53] Cosyns JP, Jadoul M, Squifflet JP, De Plaen JF, Ferluga D, van Ypersele de Strihou C. Chinese herbs nephropathy: a clue to Balkan endemic nephropathy? Kidney Int. **1994**;45(6):1680–1688.

[54] Grollman AP. Aristolochic acid nephropathy: Harbinger of a global iatrogenic disease. Environ Mol Mutagen. **2013**;54(1):1–7.

[55] Hranjec T, Kovac A, Kos J, Mao W, Chen JJ, Grollman AP, et al. Endemic nephropathy: the case for chronic poisoning by aristolochia. Croat Med J. **2005**;46(1):116–125.

[56] Shibutani S, Dong H, Suzuki N, Ueda S, Miller F, Grollman AP. Selective toxicity of aristolochic acids I and II. Drug Metab Dispos. **2007**;35(7):1217–1222.

[57] Schmeiser HH, Schoepe KB, Wiessler M. DNA adduct formation of aristolochic acid I and II in vitro and in vivo. Carcinogenesis. **1988**;9(2):297–303.

[58] Schmeiser HH, Bieler CA, Wiessler M, van Ypersele de Strihou C, Cosyns JP. Detection of DNA adducts formed by aristolochic acid in renal tissue from patients with Chinese herbs nephropathy. Cancer Res. **1996**;56(9):2025–2028.

[59] Stiborová M, Arlt VM, Schmeiser HH. Balkan endemic nephropathy: an update on its aetiology. Arch Toxicol. **2016**;.

[60] Jelaković B, Karanović S, Vuković-Lela I, Miller F, Edwards KL, Nikolić J, et al. Aristolactam-DNA adducts are a biomarker of environmental exposure to aristolochic acid. Kidney Int. **2012**;81(6):559–567.

[61] De Broe ME. Chinese herbs nephropathy and Balkan endemic nephropathy: toward a single entity, aristolochic acid nephropathy. Kidney Int. **2012**;81(6):513–515.

[62] Debelle FD, Vanherweghem JL, Nortier JL. Aristolochic acid nephropathy: a worldwide problem. Kidney Int. **2008**;74(2):158–169.

[63] Monographs, editor. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans Volume 100A. International Agency for Research on Cancer; **2012**.

[64] Nedelko T, Arlt VM, Phillips DH, Hollstein M. TP53 mutation signature supports involvement of aristolochic acid in the aetiology of endemic nephropathy-associated tumours. Int J Cancer. **2009**;124(4):987–990.

[65] Hollstein M, Moriya M, Grollman AP, Olivier M. Analysis of TP53 mutation spectra reveals the fingerprint of the potent environmental carcinogen, aristolochic acid. Mutat Res. **2013**;753(1):41–49.

[66] Hoang ML, Chen CH, Sidorenko VS, He J, Dickman KG, Yun BH, et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. Sci Transl Med. **2013**;5(197):197ra102.

[67] Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. Sci Transl Med. **2013**;5(197):197ra101.

[68] Castells X, Karanović S, Ardin M, Tomić K, Xylinas E, Durand G, et al. Low-coverage exome sequencing screen in formalin-fixed paraffin-embedded tumors reveals evidence of exposure to carcinogenic aristolochic acid. Cancer Epidemiol Biomarkers Prev. **2015**;p. cebp.0553.2015.

[69] Scelo G, Riazalhosseini Y, Greger L, Letourneau L, Gonzàlez-Porta M, Wozniak MB, et al. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. Nat Commun. **2014**;5.

[70] Turesky RJ, Yun BH, Brennan P, Mates D, Jinga V, Harnden P, et al. Aristolochic acid exposure in Romania and implications for renal cell carcinoma. British Journal of Cancer. **2015**;.

[71] Jelaković B, Castells X, Tomić K, Ardin M, Karanović S, Zavadil J. Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid. Int J Cancer. **2015**;136(12):2967–2972.

[72] Hoang ML, Chen CH, Chen PC, Roberts NJ, Dickman KG, Yun BH, et al. Aristolochic acid in the etiology of renal cell carcinoma. Cancer Epidemiol Biomarkers Prev. **2016**;.

[73] Zou S, Li J, Zhou H, Frech C, Jiang X, Chu JSC, et al. Mutational landscape of intrahepatic cholangiocarcinoma. Nat Commun. **2014**;5:5696.

[74] Totoki Y, Tatsuno K, Covington KR, Ueda H, Creighton CJ, Kato M, et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nat Genet. **2014**;advance online publication.

[75] Poon SL, Huang MN, Choo Y, McPherson JR, Yu W, Heng HL, et al. Mutation signatures implicate aristolochic acid in bladder cancer development. Genome Medicine. **2015**;7(1):38.

[76] Grollman AP, Marcus DM. Global hazards of herbal remedies: lessons from Aristolochia. EMBO rep. **2016**;17(5):619–625.

[77] Hesper B, Hogeweg P. Bioinformatica: een werkconcept. Kameleon. **1970**;.

[78] Vincent AT, Charette SJ. Who qualifies to be a bioinformatician? Front Genet. **2015**;6.

[79] Fejes AP. What is a bioinformatician | blog.fejes.ca;. http://blog.fejes.ca/?p=2418.

[80] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. **2010**;20(9):1297–1303.

[81] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. **2011**;17(1):pp. 10–12.

[82] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. **2014**;30(15):2114–2120.

[83] Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. **2014**;15(2):256–278.

[84] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth. **2012**;9(4):357–359.

[85] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. **2009**;25(14):1754–1760.

[86] Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. **2016**;32(2):292–294.

[87] Ding L, Wendl MC, McMichael JF, Raphael BJ. Expanding the computational toolbox for mining cancer genomes. Nat Rev Genet. **2014**;15(8):556–570.

[88] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. **2009**;6(9):677–681.

[89] McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MGF, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol. **2011**;7(5):e1001138.

[90] Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, Binder S, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. **2011**;27(19):2648–2654.

[91] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotech. **2013**;31(3):213–219.

[92] Wang C, Evans JM, Bhagwate AV, Prodduturi N, Sarangi V, Middha M, et al. PatternCNV: a versatile tool for detecting copy number changes from exome sequencing data. Bioinformatics. **2014**;30(18):2678–2680.

[93] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. **2009**;25(21):2865–2871.

[94] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. **2009**;25(16):2078–2079.

[95] Yang R, Nelson AC, Henzler C, Thyagarajan B, Silverstein KA. ScanIndel: a hybrid framework for indel detection via gapped alignment, split reads and de novo assembly. Genome Medicine. **2015**;7(1):127.

[96] Saunders CT, Wong WSW, 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.

[97] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. **2012**;22(3):568–576.

[98] Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Medicine. **2013**;5:91.

[99] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. **2011**;27(15):2156–2158.

[100] Boutros PC, Ewing AD, Ellrott K, Norman TC, Dang KK, Hu Y, et al. Global optimization of somatic variant identification in cancer genomes with a global community challenge. Nat Genet. **2014**;46(4):318–319.

[101] Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods. **2015**;12(7):623–630.

[102] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucl Acids Res. **2010**;38(16):e164–e164.

[103] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). **2012**;6(2):80–92.

[104] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. **2016**;17.

[105] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. **2003**;31(13):3812–3814.

[106] 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.

[107] Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. **2010**;7(8):575–576.

[108] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative Genomics Viewer. Nat Biotechnol. **2011**;29(1):24–26.

[109] Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y. GenomeView: a next-generation genome browser. Nucleic Acids Res. **2012**;40(2):e12.

[110] Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. Curr Protoc Bioinformatics. **2009**;CHAPTER:Unit1.4.

[111] Krzywinski M, Schein J, Birol n, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. **2009**;19(9):1639–1645.

[112] Goecks J, Nekrutenko A, Taylor J, Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. **2010**;11(8):R86.

[113] Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics. **2015**;31(22):3673–3675.

[114] Shiraishi Y, Tremmel G, Miyano S, Stephens M. A simple model-based approach to inferring and visualizing cancer mutation signatures. bioRxiv. **2015**;p. 019901.

[115] Rosales RA, Drummond RD, Valieris R, Dias-Neto E, Silva ITd. signeR: An empirical Bayesian approach to mutational signature discovery. Bioinformatics. **2016**;p. btw572.

[116] Fischer A, Illingworth CJR, Campbell PJ, Mustonen V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biol. **2013**;14(4):R39.

[117] Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. **2016**;534(7605):47–54.

[118] Liu Z, Hergenhahn M, Schmeiser HH, Wogan GN, Hong A, Hollstein M. Human tumor p53 mutations are selected for in mouse embryonic fibroblasts harboring a humanized p53 gene. Proc Natl Acad Sci USA. **2004**;101(9):2963–2968.

[119] Jossé R, Aninat C, Glaise D, Dumont J, Fessard V, Morel F, et al. Long-term functional stability of human HepaRG hepatocytes and use for chronic toxicity and genotoxicity studies. Drug Metab Dispos. **2008**;36(6):1111–1118.

[120] Doktorova TY, Yildirimman R, Ceelen L, Vilardell M, Vanhaecke T, Vinken M, et al. Testing chemical carcinogenicity by using a transcriptomics HepaRG-based model? EXCLI J. **2014**;13:623–637.

[121] Westcott PMK, Halliwill KD, To MD, Rashid M, Rust AG, Keane TM, et al. The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. Nature. **2015**;517(7535):489–492.

[122] Nassar D, Latil M, Boeckx B, Lambrechts D, Blanpain C. Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. Nat Med. **2015**;21(8):946–954.

[123] McCreery MQ, Halliwill KD, Chin D, Delrosario R, Hirst G, Vuong P, et al. Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. Nat Med. **2015**;21(12):1514–1520.

APPENDICES

APPENDICES

Appendix I: Revealing the molecular portrait of triple negative breast tumors in an understudied population through Omics analysis of formalin-fixed and paraffin-embedded tissues

Felipe Vaca-Paniagua, Rosa María Alvarez-Gomez, Hector Aquiles Maldonado-Martínez, Carlos Pérez-Plasencia, Veronica Fragoso-Ontiveros, Federico Lasa-Gonsebatt, Luis Alonso Herrera, David Cantú, Enrique Bargallo-Rocha, Alejandro Mohar, Geoffroy Durand, Nathalie Forey, Catherine Voegele, Maxime Vallée, Florence Le Calvez-Kelm, James McKay, **Maude Ardin**, Stéphanie Villar, Jiri Zavadil and Magali Olivier

PLoS ONE, 2015

PLOS | ONE

RESEARCH ARTICLE

Revealing the Molecular Portrait of Triple Negative Breast Tumors in an Understudied Population through Omics Analysis of Formalin-Fixed and Paraffin-Embedded Tissues

Felipe Vaca-Paniagua[1,3,4]*, Rosa María Alvarez-Gomez[5], Hector Aquiles Maldonado-Martínez[6], Carlos Pérez-Plasencia[3,4,5], Veronica Fragoso-Ontiveros[3,5], Federico Lasa-Gonsebatt[7], Luis Alonso Herrera[8], David Cantú[8], Enrique Bargallo-Rocha[9], Alejandro Mohar[7], Geoffroy Durand[2], Nathalie Forey[2], Catherine Voegele[2], Maxime Vallée[2], Florence Le Calvez-Kelm[2], James McKay[2], Maude Ardin[1], Stéphanie Villar[1], Jiri Zavadil[1], Magali Olivier[1]*

1 Group of Molecular Mechanisms and Biomarkers, International Agency for Research on Cancer, Lyon, France, **2** Group of Genetic Cancer Susceptibility, International Agency for Research on Cancer, Lyon, France, **3** Subdirección de Investigación Básica, Instituto Nacional de Cancerología, México D.F., México, **4** Unidad de Biomedicina, FES-Iztacala, Universidad Nacional Autónoma de México (UNAM), México D.F., México, **5** Unidad de Genómica y Secuenciación Masiva (UGESEM), Instituto Nacional de Cancerología, México D.F., México, **6** Departamento de Patología Molecular, Instituto Nacional de Cancerología, México D.F., México, **7** Departamento de Epidemiología, Instituto Nacional de Cancerología, México D.F., México, **8** Unidad de Investigaciones Biomédicas en Cáncer, Instituto Nacional de Cancerología, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México (UNAM), México D.F., México, **9** Departamento de Tumores Mamarios, Instituto Nacional de Cancerología, México D.F., México

* felipe.vaca@gmail.com (FVP); OlivierM@iarc.fr (MO)

Citation: Vaca-Paniagua F, Alvarez-Gomez RM, Maldonado-Martínez HA, Pérez-Plasencia C, Fragoso-Ontiveros V, Lasa-Gonsebatt F, et al. (2015) 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 10(5): e0126762. doi:10.1371/journal.pone.0126762

Academic Editor: Khalid Sossey-Alaoui, Cleveland Clinic Lerner Research Institute, UNITED STATES

Received: January 12, 2015

Accepted: April 7, 2015

Published: May 11, 2015

Data Availability Statement: The exome sequencing data are available as aligned BAM files at the Sequence Read Archive [SRA: PRJNA261515]. The expression data have been deposited in NCBI's Gene Expression Omnibus [GEO:GSE62502]. All processed data are available in supplementary tables.

Funding: This work was undertaken during the tenure of a Postdoctoral Fellowship from the International Agency for Research on Cancer,

Abstract

Triple negative breast cancer (TNBC), defined by the lack of expression of the estrogen receptor, progesterone receptor and human epidermal receptor 2, is an aggressive form of breast cancer that is more prevalent in certain populations, in particular in low- and middle-income regions. The detailed molecular features of TNBC in these regions remain unexplored as samples are mostly accessible as formalin-fixed paraffin embedded (FFPE) archived tissues, a challenging material for advanced genomic and transcriptomic studies. Using dedicated reagents and analysis pipelines, we performed whole exome sequencing and miRNA and mRNA profiling of 12 FFPE tumor tissues collected from pathological archives in Mexico. Sequencing analyses of the tumor tissues and their blood pairs identified *TP53* and *RB1* genes as the most frequently mutated genes, with a somatic mutation load of 1.7 mutations/exome Mb on average. Transcriptional analyses revealed an overexpression of growth-promoting signals (EGFR, PDGFR, VEGF, PIK3CA, FOXM1), a repression of cell cycle control pathways (TP53, RB1), a deregulation of DNA-repair pathways, and alterations in epigenetic modifiers through miRNA:mRNA network de-regulation. The molecular programs identified were typical of those described in basal-like tumors in other populations. This work demonstrates the feasibility of using archived clinical samples for

partially supported by the European Commission FP7 Marie Curie Actions—People—Co-funding of regional, national and international programs (COFUND). FV was supported by CONACyT 169082.

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

advanced integrated genomics analyses. It thus opens up opportunities for investigating molecular features of tumors from regions where only FFPE tissues are available, allowing retrospective studies on the search for treatment strategies or on the exploration of the geographic diversity of breast cancer.

Introduction

Triple negative breast cancer (TNBC), defined by the lack of expression of the estrogen receptor, progesterone receptor and human epidermal receptor 2 (HER2), is characterized by an aggressive clinical course, an earlier age of diagnosis, and lacks efficient treatment [1]. TNBC is a heterogeneous disease, both at the histopathological and molecular levels [2]. Histopathological subtypes range from ductal carcinoma, the most frequent phenotype, to rare phenotypes such as metaplastic, adenoid or medullary [3]. Transcriptomics analysis and microRNA (miRNA) profiling have identified up to four TNBC subtypes, the most frequent (80–90%) corresponding to the basal-like subtype first described in previous studies [2,4–6]. Whole genome sequencing efforts have shown that TNBC is defined by a predominance of *TP53* alterations that can be present in up to 80% of the cases, and by a large set of mutated genes occurring with minor mutation frequencies [7]. These studies have also revealed that TNBC can carry between just a handful to hundreds of somatic mutations [7, 8].

TNBC is associated with *BRCA1* germline mutations and has been reported to be more frequent in certain populations [9–11]. Indeed, studies conducted in the United States of America showed that women with TNBC are more likely to be of African and Hispanic descent and to live in socioeconomically deprived areas [12]. Studies in Mexican and African women showed prevalence from 25 to 55% [13,14] compared to 15% in Caucasian women [15]. In these populations, TNBC is also more prevalent in premenopausal compared to postmenopausal women [16]. A better understanding of the molecular heterogeneity and underlying biology of TNBC in these populations is thus essential to develop prevention and treatment strategies.

In countries where TNBCs are more prevalent, tumor samples are mainly accessible as formalin-fixed paraffin embedded (FFPE) archived tissues. The possibility to use these types of samples for comprehensive genomic analyses would allow a better characterization of the full spectrum of TNBC molecular features in these populations. Although recent developments have been made to adapt protocols and reagents to FFPE samples, performing advanced, multi-faceted molecular analyses such as genomics studies on this type of sample remains a technical challenge.

In this study, we assessed the feasibility of advanced molecular profiling of archived clinical samples of TNBC collected in Mexico. We performed transcriptomic (mRNA and miRNA) profiling and exome sequencing of 12 TNBC from Mexican women and showed that these integrated analyses can be achieved in archival FFPE samples. These molecular features were consistent with the basal-like subtype described in other populations.

Materials and Methods

Patients and samples

A retrospective series of 12 Mexican female patients diagnosed with primary TNBC (stages IIA-IIIB) at the National Cancer Institute of Mexico (INCAN) was selected based on the availability of tumor material in FFPE blocks (with tumor cellularity above 70%) and paired

peripheral blood DNA. Patients were treatment naïve at the time samples were collected. They provided written informed consent for participation in the study. Mean age of patients was 48 years (range 30–64). Five patients were premenopausal and seven were postmenopausal. Disease stages were from IIA or IIIB. Average tumor size was 3.8 cm (range 1–15) with an average tumor cellularity of 76% (range 50–90) (S1 Table). Punch biopsies from the blocks were done to enrich for regions containing the highest proportion of tumor cells for DNA extraction. Samples of punched tumor tissue, peripheral blood DNA and tumor sections on slides were anonymized and shipped to the International Agency for Research on Cancer (IARC) for molecular analyses.

Ethical approvals

This study was approved by the Ethics and Scientific committees of the National Cancer Institute of Mexico and by the IARC Ethics Committee.

Nucleic acids extraction

Genomic DNA was extracted from FFPE tumor tissues using the QIAamp DNA FFPE Tissue Kit (Qiagen) following manufacturer's instructions. Genomic DNA was isolated from peripheral blood with the Magna Pure System (Roche) following manufacturer instructions. The integrity of the material was verified by Bioanalyzer profiling (Agilent). Sample quantification was done with the Qubit dsDNA HS Assay Kit (Invitrogen). Total RNA (mRNA and miRNA) was extracted from FFPE tumor tissues with the RNeasy FFPE Kit (Qiagen) following manufacturer's instructions.

Whole gene expression profiling and differential gene expression analysis

Transcriptomic analysis of the 12 tumor samples was done with the FFPE-designed WG-DASL (Illumina) assay (assess 29,285 annotated transcripts) according to manufacturer's instructions. Briefly, 200 ng of total RNA were reverse-transcribed into biotinylated cDNA, which was then primer-extended with the Assay Specific Oligos. The cDNA was then amplified with universal primers and hybridized to Illumina Human WG DASL HT Expression BeadChip arrays. The Illumina Genome Studio V2010.2 was used to obtain the signal values (AVG-Signal), with no normalization and no background subtraction. A technical duplicate was performed for all samples. The performance of hybridizations was evaluated by assessing the presence of outliers and the noise-to-signal ratios by calculating the ratio of centiles P95/P05 prior to normalisation for each sample. We defined outliers as samples with P95/P05 ratio <9.5. All samples were found to show a correct noise-to-signal ratio (P95/P05>9.6). The expression data have been deposited in NCBI's Gene Expression Omnibus [GEO:GSE62502] [17].

For differential gene expression analysis, two public datasets generated on the Illumina HumanHT-12 v3.0 beadChip, which contains 99.98% of the 29,285 probes of the Human WG DASL HT BeadChip were used as normal breast tissue controls: NCBI GEO GSE17072 and GSE32124 including five and 33 fresh frozen tissue samples, respectively. Identity and concordance of the array probes between these public datasets and the probes used in this study were verified by direct comparison using R. Genespring software GX 11.2 (Agilent, Santa Clara, USA) was used to perform quantile normalization on all datasets and differential expression analysis using the paired t-test method. Significance levels (p-values) were corrected using the Benjamini-Hochberg false discovery rate (FDR) method to correct for multiple hypothesis testing. Probes with a FDR-adjusted p-value of <0.01 and probes with a minimum of 2-fold and 1.5-fold change were considered significantly differentially expressed for the GSE32124 and

GSE17072 datasets respectively. Genes for which probes showed contradictory direction of regulation were excluded from further analyses. The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 was used for classification of the differentially expressed genes according to biological and molecular processes. Gene Set Enrichment Analysis (GSEA) (Broad Institute) was used to define biologically relevant sets of genes based on experimentally validated gene sets deposited in the Molecular Signature Database (MSigDB). Only gene sets represented with a p-value of <0.01 were considered.

miRNA profiling and differential expression analysis

Human TaqMan Low Density Arrays were used for profiling 754 mature miRNAs (TLDA A v2, TLDA B v3; Applied Biosystems) following manufacturer's instructions. Briefly, 500 ng of total RNA (same aliquots as those used for whole gene expression) were reverse transcribed with the Megaplex RT Primers. cDNA was quantified using the 7900 HT Real-Time PCR system (Applied Biosystems). The signals obtained from the CT values were used for percentile shift normalization and correction using The Benjamini-Hochberg method in the Genespring software GX 11.2 (Agilent). For differential expression analysis, six normal fresh frozen breast tissues profiled with the same TLDAs (NCBI GEO accession number GSE35412) were used as control samples. MicroRNAs with a FDR- adjusted p-value of <0.01 and with a minimum of 2-fold change were considered as differentially expressed and were ranked by percentile values.

PAM50 tumor subtype classification

From the 50 genes included in the PAM50 classifier, 49 were present on the DASL array. Expression values from these genes obtained in the microarray experiments were used to classify samples in the four subtypes defined with PAM50. Correlation of PAM50 centroids was computed on summarized values. Classification was considered as the centroid with the biggest correlation. Expression values of probes corresponding to the same gene were averaged and the correlation was computed on these values. P-value was calculated as the proportion of times that the maximum correlation to each centroid is greater than or equal to the maximum observed correlation in 100,000 resamplings of the unsummarized data.

Integrated analysis of miRNA-mRNA relationships

The significantly modulated (up- or down-regulated) miRNAs and mRNA were used as input to the Ingenuity Pathways Analysis (IPA) suite (Ingenuity Systems, Redwood City, CA) for identifying potential miRNA-mRNA regulatory relationships. Only relationships defined as experimentally observed or high confidence predictions were considered. Predictions were based on data deposited in various databases (TargetScan, TarBase, miRecords, miRanda, DIANA-microT, miTarget, PicTar and PITA) taking into account stringent criteria such as strong evolutionary conservation, seed site number in the mRNA, site sequence context, thermodynamics parameters and pairing stability. From the list of interactions and networks found by the IPA suite (S10 Table), a reduced complexity network was built for each list (miRNAup/mRNAdn and miRNAdn/mRNAup), taking into account only interactions involving mRNA of genes involved in cancer (based on IPA disease annotations) and a miRNA connectivity score above 4. Networks were visualized using the Cytoscape 3.1.0 software. For the miRNAdn/mRNAup network, clusters were selected using the MCL algorithm of the ClusterMaker plugin using the default settings as reported elsewhere [18–20].

Exome sequencing and somatic mutation analysis

Genomic DNAs from tumor tissue and from peripheral blood were used for library preparation with the 5500XL SOLiD Fragment Library Core Kit (Life Technologies) and exome capture with the TargetSeq Exome Kit (Life Technologies) with minor modifications. Single fragment sequencing was then performed with the SOLiD 5500XL platform (Life Technologies). Briefly 1.5–3 ug of DNA were sonicated with a Covaris S220 to 160 bp mean size (120–200 bp range) with the program: 10% duty cycle; intensity 5; 100 cycles per burst; 6 cycles of 60 seconds for blood DNA and 6 cycles of 100s for FFPE DNA; in frequency sweeping mode. Sheared fragments were end-repaired, size-selected (100–250 bp) using the Agencourt AMPure XP magnetic beads (Beckman Coulter), dA-tailed and ligated to P1-T and Barcode-T adaptors and amplified in 6–8 PCR cycles. Quality controls included quantitation of DNA after size-selection, ligation of adaptors and amplification of DNAs using the Qubit dsDNA HS Assay Kit and the Qubit 2.0 fluorometer, and DNA profiling of at the same steps using the High Sensitivity DNA Kit and the 2100 Bioanalyzer (Agilent). 62.5 ng of amplified material were pooled in batches of four and used for the exome capture with the TargetSeq Exome Enrichment kit which is designed to target 37.3 Mb of exomic sequence representing >98% VEGA, CCDS coding regions, and RefSeq exons. The pooled barcoded libraries were purified with Dynabeads M-270 Streptavidin (Invitrogen) and amplified by 6–8 PCR cycles and purified with Agencourt AMPure XP (Beckman Coulter). Emulsion PCRs were done with the SOLiD EZ Bead system (Lifetechnologies). Captured and emulsion amplified libraries were sequenced with the SOLiD 5500XL (1x75bp) using the Exact Call Chemistry (ECC) module to a calculated average coverage of 100X for tumors and 50X for blood DNA across targeted regions.

Exome data analyses (mapping and variant calling) were performed with LifeScope suite software (Lifetechnologies), using a combination of LifeScope targeted resequencing (TR) and low frequency variant detection (LFVD) workflows with the default parameters: (i) the color space reads generated by the SOLiD5500XL were mapped against the human genome reference version hg19 generating bam files which were then enriched for targets and good quality reads; (ii) single-nucleotide variants (SNVs) were called on enriched bam in both TR and LFVD workflows and small insertions and deletions (indels) were called on enriched bam in TR workflow. The LFVD aligned BAM files were submitted to the Sequence Read Archive [SRA: PRJNA261515]. The Variant Call Format (VCF) output files were annotated using the Annovar software including functional annotations as well as frequencies in known databases and in our custom SOLID catalogue of variants. Finally the VCF files were annotated using a custom "back-to-bam" script that added the number of reads found at each position with each allele and on each strand. Only variants present in both strands and in uniquely mapped reads were considered. Annotation of variants was done with ANNOVAR. Tumor-specific somatic mutations were obtained by subtracting tumor mutations found in the paired DNA blood sample. Somatic mutations were further filtered as follows. Exclusion criteria: mutations where the variant allele was found in less than two sequence read start sites; mutation in the context of an homopolymer tract of more than six repeated nt upstream or downstream; synonymous mutation; mutation with depth >800X (to eliminate false-positive variants over-represented by misalignment); allelic frequency >0.1% in either the 1000 Genomes project or the Whole Exome Sequencing Project (Washington University, which includes 6500 sequenced samples); allelic frequency >5% in the IARC generic catalogues of sequenced samples and a TCGA dataset of 650 samples (to eliminate systematic sequencing errors). Inclusion criteria: mutation found in an exonic or splice site; mutation not present in a duplicated genomic region; mutation predicted as deleterious by SIFT or PolyPhen-2.

To further select potentially pathogenic alterations, mutations were annotated as driver events if present in genes (1) defined as tumor suppressors or oncogenes in GSEA (Broad

Institute) database, (2) considered as drivers by Vogelstein et al [21], (3) identified as significantly mutated genes in the five studies of breast cancer available in the TCGA database. Collectively, these databases contain more than 6,000 sequenced tumors. All filtered variants were verified by manual inspection of the BAM files using the Integrated Genome Viewer (Broad Institute). Recurrent somatic mutations were confirmed experimentally by sequencing both the tumor and blood DNA samples with the Ion Torrent PGM platform using targeted amplicon sequencing at 500X minimum coverage, following manufacturer's instructions.

Results

Deregulated transcriptional programs

To characterize the transcriptional programs expressed in TNBC tumors from Mexican women, archived tumor samples were analyzed by whole gene expression profiling using FFPE specific reagents for RNA extraction and analysis (see Methods). The FFPE-designed WG DASL HT (Illumina) expression assay covers 20,727 unique genes. We first classified samples using the PAM50 classifier and found that 75% (9/12) of the tumors belonged to basal-like molecular subtype while 25% corresponded to the HER2 subtype (Fig 1, S2 Table). Differential gene expression analysis (DGEA) was then performed using two independent public control datasets (obtained on the same type of analysis platform) of normal breast tissues as paired normal tissues were not available from the tissue archive (see Materials and Methods). Using the first control set (GSE32124, 33 samples of normal tissues), 7,459 of 20,727 (36.0%) genes were differentially expressed in the tumor tissues, with 18.0% and 17.9% of genes up- and down-regulated respectively (S3 and S4 Tables). Using the second control set (GSE17072, five samples of normal tissues), 1,590 of 20,727 (7.7%) genes were found differentially expressed with 4.4% and 3.3% of genes up- and down-regulated respectively (S5 Table). There was a near 90% overlap between these two independent analyses, with only one gene related to cancer that had contradictory direction of regulation (S1 Fig). Therefore, considering the stringency of the statistics used (T-test p = 0.01, FDR = 0.01), the substantial overlap between both analyses, and the fact that the larger number of controls may resemble more accurately the transcriptional nature of the normal breast tissue, we used results obtained with the GSE32124 dataset (33 controls). This DGEA analysis showed that, among genes expected to be overexpressed in the TNBC phenotype, *MKI67*, *TOP2A*, *CCNE1*, *CCNE2*, *EGFR*, *FGFR1*, *FGFR2*, *VEGFA*, *HIF1A*, *ARNT*, *FOXM1* and the BRCA1-repressor *ID4* were found up-regulated (S3 Table). Interestingly, *MYC* mRNA was not up-regulated, but there was a significant enrichment of 37 overexpressed genes in the gene set corresponding to cytogenetic band 8q24 that includes *MYC* gene (S6 Table). Significant enrichment was also found for other cytogenetic bands associated with breast cancer, including 1q21, 1q22, 1q32, 3q23 and 3q28 (S6 Table). Among other up-regulated genes, those that contain transcription-binding sites for *MYC*, *MAX*, *MYB*, *ELK1*, *ETS1*, *ETS2*, *ETV7*, *MAF* and *E2F* were significantly enriched (S7 Table). Significant enrichment was also found for growth promoting and tumor progression pathways such as *EGFR*, *MYC*, *VEGFR* and *E2F*; for biological processes linked to proliferation such as telomere maintenance, cell cycle, DNA repair, DNA replication chromosome organization; and for gene sets associated with ductal invasive breast cancer and exposure to *EGF* or *MYC* activity (Fig 2A). Interestingly, the isoforms of the apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like (*APOBEC3A*, *APOBEC3B* and *APOBEC3H*), implicated in breast cancer carcinogenesis, were also up-regulated (S3 Table). In addition, 10 of the 14 genes associated with the Basal-like Immune Activated (BLIA) subtype of TNBC recently reported by Burstein et al., where up-regulated, including *CXCL11*, *RARRES1*, *GBP5*, *CXCL10*, *CXCL13*, *LAMP3*, *STAT1*, *CTLA4*, *TOP2A*, *LCK* [22].

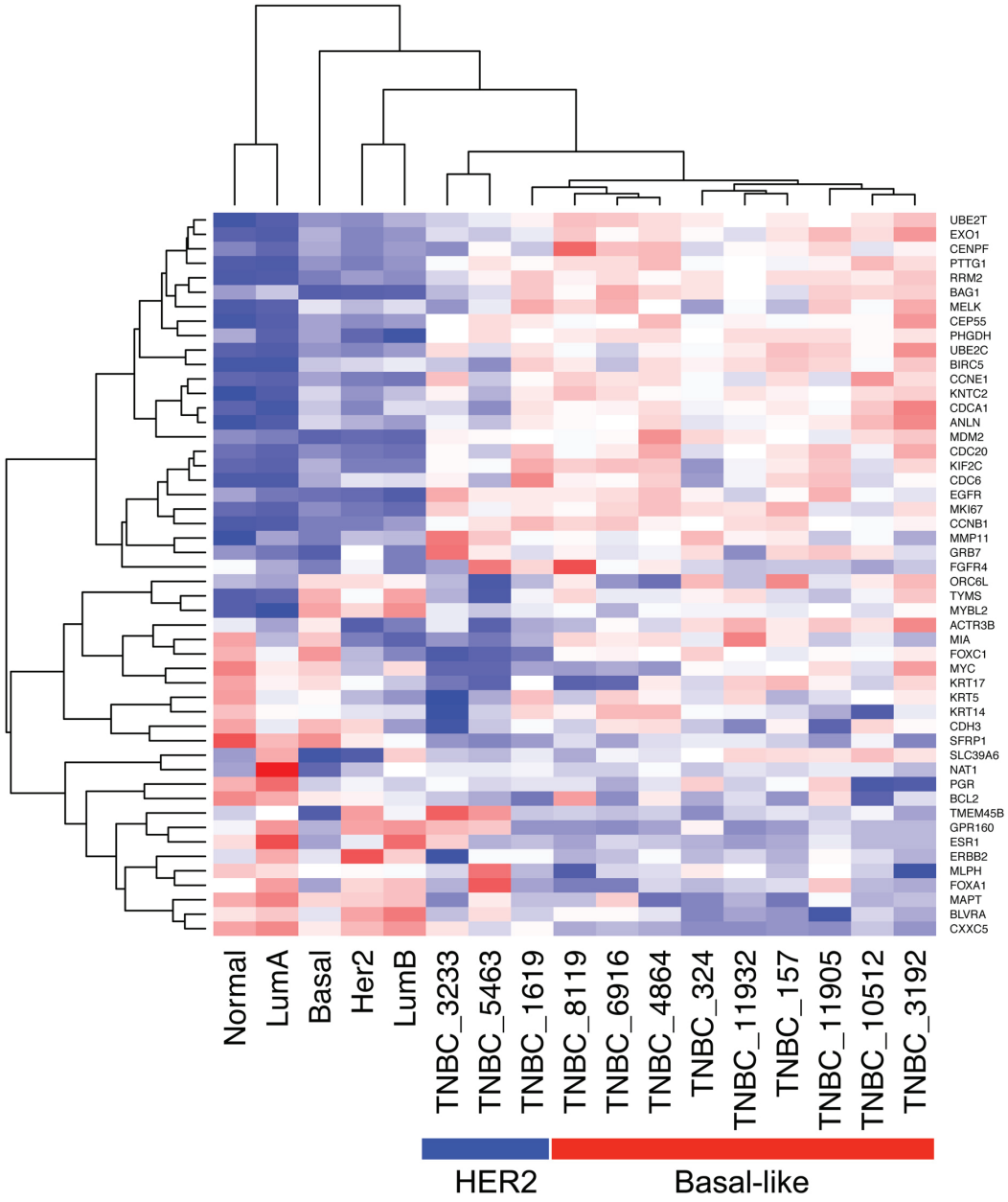


Fig 1. PAM50 classification of TNBC samples. The hierarchically-clustered normalized expression values of the PAM50 classifier genes is shown for the 12 triple negative breast cancers (TNBCs) analyzed and the five centroids. The samples were classified according to their correlation with the PAM50 centroids. Red and blue boxes represent overexpressed and down-regulated genes, respectively.

doi:10.1371/journal.pone.0126762.g001

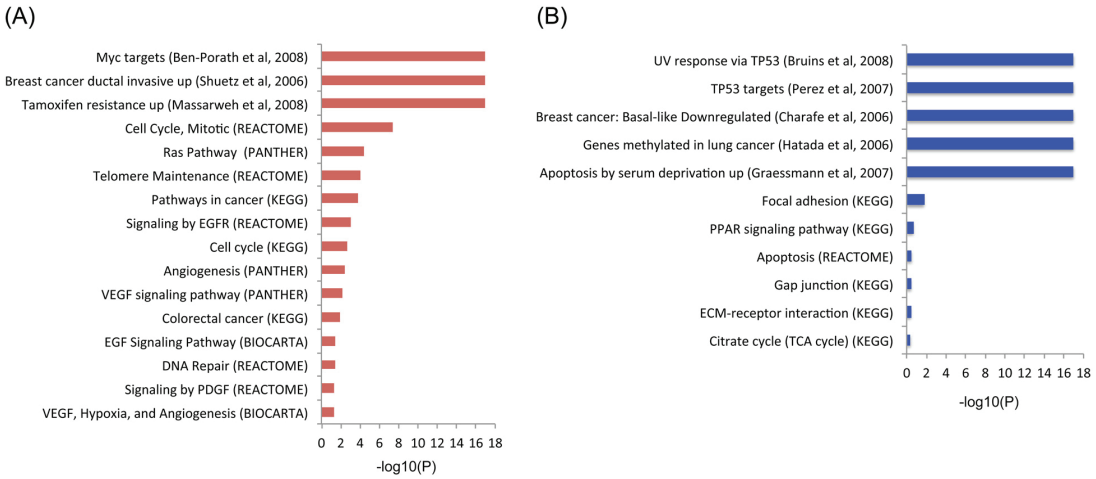


Fig 2. Enriched pathways from the differential gene expression analysis. (A) Up-regulated pathways from KEGG, PANTHER, REACTOME and Chemical and Genetic perturbations (GSEA). (B) Down-regulated pathways from the same sources. Enrichment is shown as-log10(P) values.

doi:10.1371/journal.pone.0126762.g002

Consistent with the triple negative phenotype, *ESR1* and *ERBB2* were down-regulated (although *PGR* levels did not show a statistically significant down-regulation) (S4 Table). Within down-regulated genes, there was significant enrichment for *TP53* targets, genes down-regulated in basal-like breast cancer, respiratory electron transport and cell adhesion pathways, *MYC* repressors, *BRCA*, *CHEK2* and *ATM* networks (Fig 2B and S4 Table).

miRNA expression analysis and networks of miRNA:mRNA connectivity

The expression profile of 754 miRNAs was determined in the 12 tumor samples using TaqMan arrays. To identify deregulated miRNAs, a control dataset of six normal tissues (deposited in GSE35412) obtained with the same analysis platform was used for differential expression analysis (see Materials and Methods). We found a pattern characterized by high expression of oncogenic miRNAs and repression of tumor suppressor miRNAs (S8 and S9 Tables). The clusters of highest enrichment in the up-regulated miRNAs were the miR-1283 (7 miRNAs), miR-1185 (6 miRNAs) and miR-17 cluster (5 miRNAs). Consistent with its polycistronic nature, all members of the miR-17-92 oncogenic cluster showed co-regulation, with the exception of miR-19a. Furthermore, the miR-17 and miR-548 families were the most enriched ones with five and seven members involved, respectively. Overall, the top five up-regulated miRNAs were miR-624, miR-339-5p, miR-191 and miR-651. The top five up-regulated oncogenic miRNAs were miR-93, miR-20a, miR-214, miR-146b and miR-92a (percentile rank of overexpressed miRNAs 92.9, 91.4, 84.3, 75.8 and 68, respectively). Strongly down-regulated tumor suppressing miRNAs included miR-1, miR-34c, let-7a, let-7b, miR-127, with a percentile rank of inhibition of 96.2, 84.2, 69.1, 61.6, 59.3, respectively. Other down-regulated tumor suppressors were let-7c, miR-101, let-7e, miR-125b, miR-141, miR-126, miR-34a, miR-34c and miR-200a.

In order to identify potential networks of post-transcriptional regulation of gene expression, we analyzed inverse relationships of miRNAs and their target mRNAs in the differentially expressed paired profiles focusing on genes involved in cancer (see Materials and Methods). We found a repressive network composed of 53 overexpressed miRNAs and 68 down-regulated genes (Fig 3A, S10 Table). The repressed genes were significantly enriched for the ontology terms “regulation of apoptosis” and “tight junction”. Tumor associated miRNAs present in this

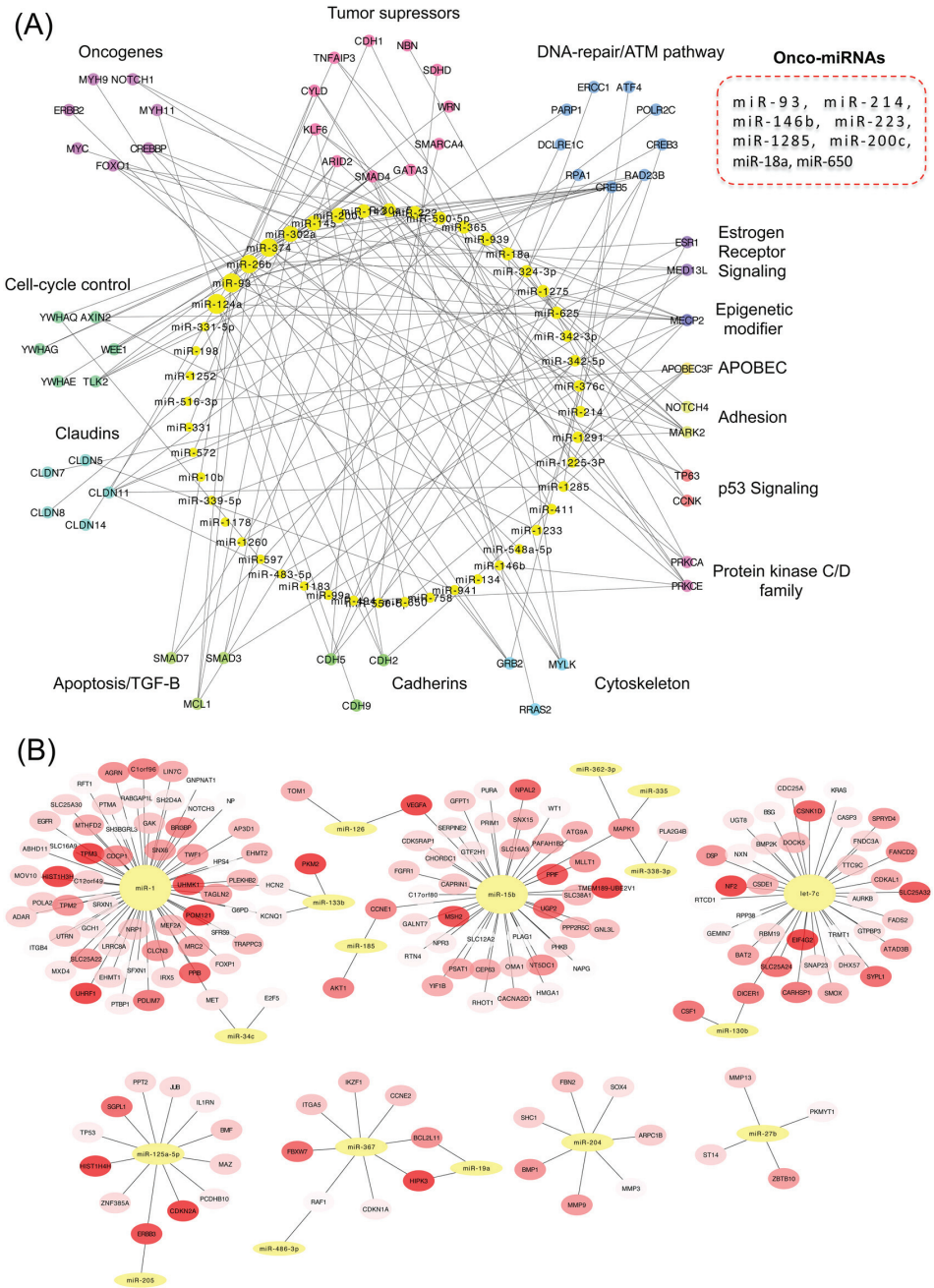


Fig 3. Networks of connectivity between miRNA and mRNA. (A) Networks of (A) up-regulated miRNAs and down-regulated mRNAs, (B) down-regulated miRNAs and up-regulated mRNAs, identified in the TNBC samples (only high confident interactions and genes involved in cancer were selected).

doi:10.1371/journal.pone.0126762.g003

network included miR-18a, miR-93, miR-146b, miR-200c, miR-214, miR-223, miR-650 and miR-1285, collectively repressing 9 tumor suppressors genes (including *ARID2*, *ATM*, *CYLD*, *KLF6*, *NBN*, *SDHD*, *SMAD4*, *SMARCA4*, *TNFAIP3*), nine oncogenes (including *BCL2*, *CREB1*, *CREBBP*, *FOXO1*, *MYC*, *MYH11*, *MYH9*, *NOTCH1*), growth inhibitors (*BMPR2*, *BTG2*, *BTG3*, *LITAF* and *PA2G4*), chromatin remodelers, six claudin genes (*CLDN11*, *CLDN14*, *CLDN5*, *CLDN7*,*CLDN8*, *CLDND1*) 11 DNA-repair/ATM pathway genes (*ATF4*, *CREB3*, *CREB5*, *DCLRE1C*, *ERCC1*, *PARP1*, *POLR2C*, *RAD23B*, *RAD51*, *RPA1*, *XRCC5*) and six proapoptotic genes (*BAK1*, *BCL2L1*, *MCL1*, *SMAD3*, *SMAD5*, *SMAD7*). Interestingly, the transcripts of the estrogen receptor (*ESR1*), *HER2* (*ERBB2*) and E-cadherin (*CDH1*) were part of this suppressive network (Fig 3A). In the reverse association, there were 32 connected clusters of down-regulated miRNAs and up-regulated target mRNAs. Seven of these clusters harbored four or more genes (Fig 3B) and showed a significant functional enrichment for kinase and cell cycle activity, and for focal adhesion pathways. The clusters with the highest number of mRNAs included tumor suppressing miRNAs such as miR-1 (connected to 63 mRNAs), miR-15b (linked to 43 mRNAs), let-7b (connected to 37 mRNAs), miR-125a (connected to 13 mRNAs) and miR-204 (linked to 7 mRNAs) (Fig 3B). Oncogenes present in these clusters included *EGFR*, *FGFR1*, *HRAS*, *KRAS*, *FOXP1*, *HMGA1*, *MET*, *MLLT1*, *NOTCH2*, *PAFAH1B2*, *PICALM*, *PLAG1*, *RUNX1*, *TPM3*, and five growth factors (*BMP1*, *CSF1*, *IL1RN*, *TGFB3* and *VEGFA*), some of which are typically overexpressed in TNBC (Fig 3B).

Landscape of somatic mutations

The 12 tumors and their paired blood DNA were sequenced by whole exome sequencing at a mean coverage of 123x (range, 54–269) and 65x (range, 56–82) respectively (Table 1). Using a dedicated bioinformatics pipeline combining high mapping and calling stringency, and exclusion of platform-specific sequencing errors (see Materials and Methods), we found a mean frequency of 1.7 mutations/exome Mb (range, 0.46–5.56), which is concordant with previous studies in TNBC [7,23]. To define potential driving mutations, we applied a filtering approach based on mutation position (splicing sites, exonic), nature of the mutation (non-synonymous with predicted deleterious impact on protein), its genomic context, absence in human populations without cancer, and the biological relevance of the affected gene (see Materials and Methods). With this strategy, we found samples with 1 to 20 driving mutations with a mean of 5.25 driving mutations per sample (Fig 4; S11 Table). Two tumors with only one driving mutation had *TP53* affected. Recurrent driving genes with pathogenic mutations were *TP53* (54%), *RB1* (27%), *ARID1A* (18%), *BRCA1* (18%), *KDM6A* (18%), *PTEN* (18%) and *SETD2* (18%). Other relevant cancer genes that were mutated in only one sample included *ATR*, *ATX*, *BRCA2*, *CDH1*, *GATA3*, *PAX7*, *PIK3CA* and *ESR1*. Thirty percent of the mutated driving genes were druggable with 91 anti-neo-neoplastic agents, of which 18 are drugs approved by the Food and Drug Administration (FDA) (S12 Table). Interestingly, pathogenic mutations were found in 19 DNA repair genes and tumors with higher numbers of mutations in DNA repair genes showed

Table 1. Whole exome sequencing summary statistics.

Metric	Value (range)
Tumors-normal pairs sequenced	12
Total sequenced (GB)	198.43
Mean fold tumor target coverage (range)	123x (54–269)
Mean fold normal target coverage (range)	65x (56–82)
Mean somatic mutation rate per megabase (range)	1.7 (0.46–5.56)

doi:10.1371/journal.pone.0126762.t001

higher mutation load (Fig 4; S2 Fig). Given that we found three isoforms of the *APOBEC3* cytidine deaminase overexpressed in these tumors, we looked for the expected signature of the APOBEC3 family of enzymes. The APOBEC signature is characterized by C to T and C to G mutations in a T**C**W sequence context [24–26]. In our set of tumors the main mutation type was C to T with a majority occurring in the context of 5'-T**C**T-3', resembling the APOBEC signature. Mutations with this signature were found in cancer driving genes such as *ATRX*, *FGFR2*, *SETBP1* and *TP53*. However, there was no correlation between APOBEC mRNA expression and the total number of C to T transitions in the context of T**C**W at the sample level.

Summary integrated analysis

TP53 pathway. *TP53* was found mutated in 50% of the samples with a high proportion of truncating mutations (50%; Fig 4). Evidence for *TP53* inactivation by other mechanisms was also found. Indeed, the p53 inhibitors *MDM2* and *MDM4* were up-regulated at the mRNA level, and several miRNA acting downstream of p53 or impacting p53 were deregulated. A set of onco-suppressive microRNAs with p53 binding sites in their promoters were down-regulated. These microRNAs, that included miR-15b, miR-205, miR-34a, miR-34c, miR-192, miR-200a and miR-107 target p53 inhibitors (such as *MDM2*), cell cycle genes (*CDK4/6*, *E2F*, *CCNE1/2*, *CDC7*), oncogenes (*MET*), growth factors (*IGF1*), antiapoptotic genes (*BCL2*), metabolic genes (*LDHA*), stemness genes (*NANOG*, *SOX2*, *KLF4*, *OCT4*, *CD44*), proangiogenic genes (*ANRT*) [27–29] or genes involved in proliferation and metastasis (*LAMC1*). miR-449b

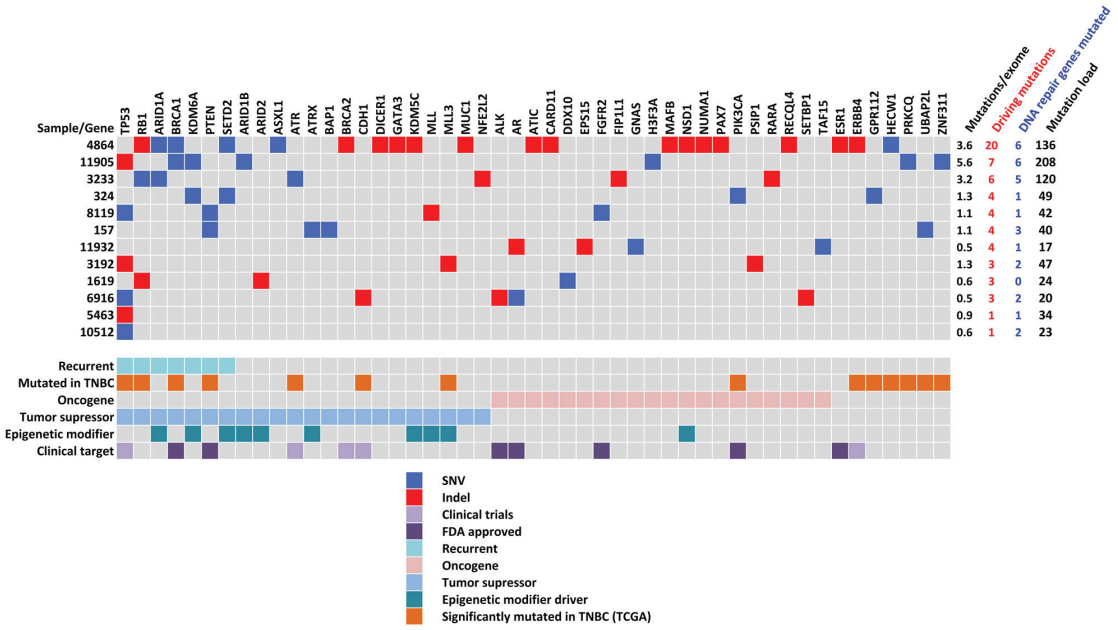


Fig 4. Summary results of whole exome sequencing for all matched samples (n = 12). Top panel, details of mutations found in tumors in each specific samples (blue squares are single-nucleotide variants (SNVs) and red squares are insertions and deletions (Indels). Right panel shows the mutations per exome, the amount of mutations in driving and DNA repair genes and total number of mutations per tumor. Lower panel shows gene annotations derived from present results and other studies: Recurrent gene mutated in more than one sample in present series; Mutated in triple negative breast cancer (TNBC) [7], significantly mutated in TNBC from TCGA [49,50]; Oncogenes and tumor suppressors class derived from references [21,51]; Epigenetic modifiers class derived from references [52,53]; Clinical target, genes that are targets of drugs in clinical trials or approved by FDA. (http://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/default.htm).

doi:10.1371/journal.pone.0126762.g004

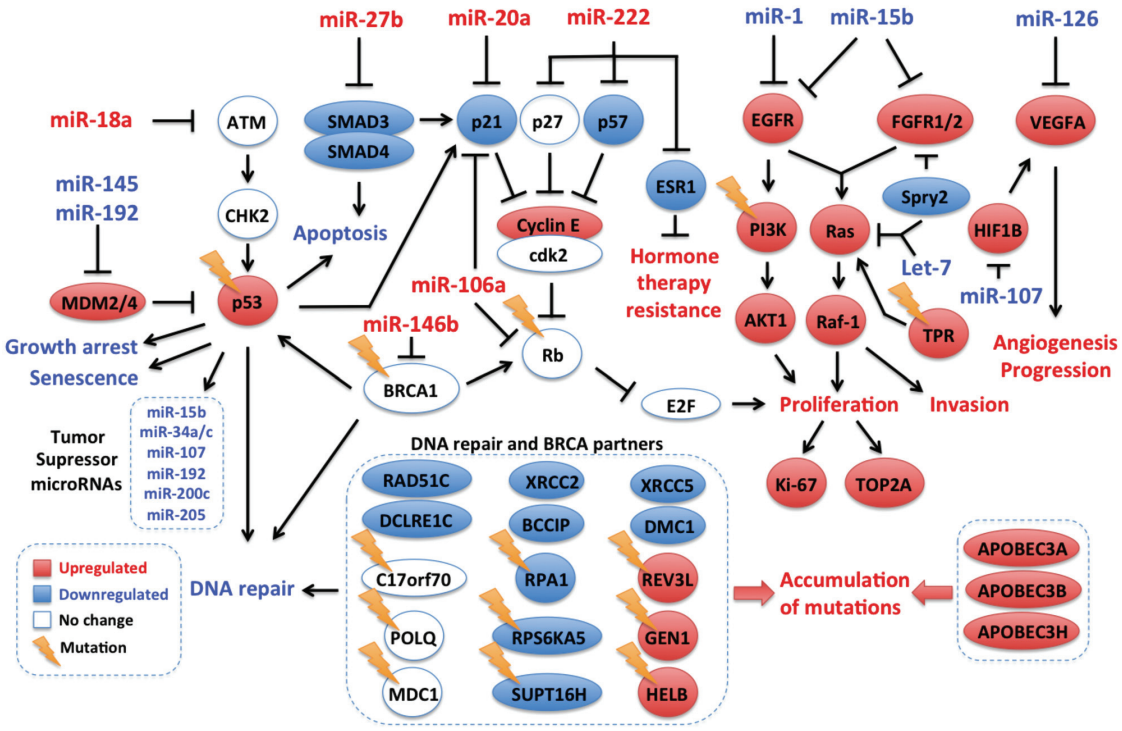


Fig 5. Integrated molecular portrait of triple negative breast tumors from Mexican women using archived clinical samples. Data from whole exome sequencing and miRNA and mRNA expression experiments were integrated. Tumor suppressor pathways and DNA repair genes are repressed through different mechanisms, while tumor growth and progression pathways and genes related to endogenous mutation promotion exhibit up-regulation. Gene, miRNA and Pathway expression levels are depicted in red for up-regulation and blue for down-regulation; genes without differential expression are marked in white. Mutations are shown as lightning icons.

doi:10.1371/journal.pone.0126762.g005

and miR34 which block the p53 repressors *SIRT1* and *HDAC1* were down-regulated. miR-1285, which is a direct repressor of TP53 was overexpressed [30]. Finally, The MDM-blocking miRNA miR-192 was down-regulated (Fig 5).

PIK3CA/PTEN pathway. Only three samples showed direct activation of the *PIK3CA* pathway, with one sample (8%) carrying an activating mutation in *PIK3CA* (p.H1047L) and two samples carrying *PTEN* inactivating mutations. However, there was evidence for global activation of this pathway in the miRNA/mRNA expression analyses (Fig 5). Indeed, key *PIK3CA* effectors were overexpressed (*AKT1*, *PIK3CB*, *PIK3CD*, *PTK2*, *KRAS*, *CD19*, *HSP90AA1* and the integrins *ITGA1*, *ITGA5*, *ITGA8*, *ITGB4*, *ITGB6* and *ITGB8*), and the PTEN-targeting miRNAs miR-21, miR-222 and the AKT downstream targets *MDM2*, *NOS3*, *YWHAZ* were up-regulated. In addition the *FOXO1*, *BAD*, *CDKN1A*, *CDKN1B* genes, which encode proteins repressed by *AKT* were found down-regulated, as well as the *AKT* inhibitors miR-125b, miR-149, miR-184, miR-708, *PPP2CA*, *PPP2R1A*, *PPP2R1B* and *PPP2R5E* (S3 Fig).

RB1 pathway. An overall down-regulation of the Rb pathway was apparent, with mutations in *RB1* in three samples and global up-regulation of *MDM2* (RB1 repressor) and of RB1-targeting miRNAs (miR-215, miR-106a, miR-17, miR-20a, miR-93, miR-215, miR-21).

DNA repair pathway. Mutations in DNA repair genes were found in all samples but one, including four deleterious somatic mutations in *BRCA1/2* genes. Although no change was observed at mRNA level in BRCA genes, some of their important molecular partners were down-

regulated including *BCCIP*, *RAD51C*, *DMC1*, *DCLRE1C* (Artemis), *XRCC2*, *XRCC5* and *RPA1* (Fig 5). Furthermore, the BRCA1-targeting miR-146b was among the top up-regulated miRNAs, which may result in BRCA1 repression through translation inhibition [31]. These results suggest a low capacity for genome maintenance and may thus reflect a genomic instability state in these tumors [32–34].

FOXM1 pathway. Consistent with other TNBC studies, *FOXM1* mRNA was overexpressed in these tumors and several *FOXM1* transcriptional targets were found up-regulated (*AURKB*, *BIRC5*, *CCNB1*, *CCNB2*, *CENPA*, *CENPF*, *NEK2*, *SKP2*, *XRCC1* and *MMP9)*. Furthermore, *FOXM1*, the inducers *RAF1*, *HIF1A*, *CCNA*, *CCNB* and *CCNE* were overexpressed, while the repressors miR-149 and miR-186 were down-regulated (S4 Fig). There was thus evidence for *FOXM1* pathway activation in these tumors.

Epigenetic modifiers. Interestingly, we found a total of 14 mutations in chromatin remodeling genes, including recurrent mutations in *ARID1A*, *KDM6A* and *SETD2* (Fig 4). Additionally, the expression levels of several genes with epigenome regulatory activity were altered, including 51 up-regulated and 23 down-regulated (S13 Table).

Invasiveness. Invasive/EMT-like phenotype has been reported in TNBC and could reflect the metastatic nature of these tumors. Although regulators of invasiveness such as Snail, Slug and Twist were not found deregulated, molecular markers like *MMP9* and Fibronectin and miRNA related to invasion (miR-222) were up-regulated, while *CDH1* was repressed. These results may reflect an invasive-like signature.

Discussion

This study represents the most comprehensive genomic analysis performed on clinical FFPE samples of TNBC from an ethnically homogenous cohort of Latin American women. Combining whole exome sequencing, miRNA and mRNA transcriptomic profiling, we found a gene expression program and a profile of somatic mutations consistent with findings reported in recent integrated genomic analyses of large series of high quality breast cancer samples, including two studies focused on TNBC [7,35]. Although we analyzed a small set of samples (12 tumors), these samples have been carefully selected as molecularly homogeneous TNBC based on IHC analyses with more than 70% tumor cellularity. Our transcriptional analysis found that 9 out of 12 samples (75%) had an expression pattern of basal-like breast cancer (based on the PAM50 gene signature). These findings are thus consistent with the fact that basal-like subtype of breast cancer is found in 70–80% of TNBC [8]. Moreover, our samples may be more similar to the immune-activated basal-like subtype of TNBC that has recently identified in several sample sets [22].

The tumors analyzed harbored one to 20 driving mutations per tumor. *TP53* was the most frequently mutated gene, with a high proportion of truncating mutations, a feature previously described in basal-like breast cancer [36]. *RB1* was the second most frequent hit as reported before [7,35]. *PIK3CA* gene, which is frequently mutated in breast cancer, except in the TNBC/basal-like subtype, was also found rarely mutated (in 1/12, 8.3% of samples). Overall, half of the samples (6/12) carried genetic alterations potentially actionable by currently available drugs approved by the FDA. Deregulated transcriptional programs were characterized by growth and tumor promoting signals, with in particular the overexpression of EGFR, PDGFR and VEGF pathways, recognized as some of the main pathways driving TNBC [7,37]. The combined analysis of mutations and transcriptional programs showed that well-recognized pathways deregulated in basal-like tumors were found deregulated in these set of TNBC tumors (see Summary integrated analysis section and Fig 5). Thus, no specific molecular characteristics were found in these series of tumors, suggesting that TNBC in Mexican women develops through similar mechanisms to TNBC in other populations. It is of note that our analysis was focused on genes involved in cancer pathways

and on well-defined molecular interactions, as our primary aim was to assess whether the major molecular alterations detected in this series of tumors matched the one described in other TNBC cases. However, this study shows that TNBC oncogenic molecular programs converge in key deregulated pathways that may be targeted, not only by chemotherapeutic agents and protein-targeting drugs, but also oncogenic miRNA blockers.

One of the challenging limitations of our study was the lack of normal tissue for the analysis of mRNA and miRNA differential expression. We circumvented this issue by employing public expression data of normal breast tissue obtained on the same or similar arrays as the one used here and by applying high stringency statistics, an approach that has been successfully applied in other breast cancer studies [38,39]. For the DGEA with mRNA we compared the performance of two datasets and found a high degree of concordance and no notable difference in the sense of regulation of the genes. For miRNA analysis, we could identify only one control set. Despite this limitation, the overall pathway and miRNA analyses confirmed expected molecular alterations found in other breast cancer and TNBC genomic studies, and corroborated the histological classification, which collectively reinforced the validity of our analysis.

While other studies showed the possibility of using FFPE samples for Next-Generation Sequencing (NGS) analysis [40,41], this remains challenging. For mRNA analysis we used reagents specially dedicated to FFPE samples that allow the detection of molecules in partially degraded samples [42,43]. In contrast, the stability of miRNA in FFPE samples has been described to be high in different studies [44,45] and we used a classical TaqMan platform that has been validated for FFPE samples [46,47]. Tissue fixation with formalin is known to increase C to T substitution artifacts during the PCR steps. In breast cancer, the up-regulation of the APOBEC3 enzymes has been identified as a major source of C to T mutations within the TCW sequence context [48]. In our analysis, although we found an overall up-regulation of *APOBEC3B*, *APOBEC3A* and *APOBEC3H* mRNA and a mutation signature enriched for C to T transitions, the level of mRNA expression of APOBEC members at the sample level did not correlate with the overall mutation load or the abundance of C to T mutations. Whether potential C>T artifacts due to the FFPE origin of the samples may have masked APOBEC mutation signature thus remains unclear. It is of note that, although we could not test all mutations in validation experiments due to lack of sufficient material, we could validate 20/22 (91%) mutations in recurrently mutated genes using targeted resequencing with the Ion Torrent (data not shown). Thus, although mutation confirmation is important for interpreting sequencing data from FFPE material, our work confirms the feasibility of using these biospecimens for NGS analysis. It is expected that technology developments will soon improve the quality of sequencing obtained with these samples.

Conclusions

Our results show that an integrated molecular analysis including exome and miRNA:mRNA transcriptome is feasible on archival FFPE samples. The possibility to use FFPE material, which represents the most frequent sample type in biorepositories, is of broad interest since it opens the opportunity to perform large retrospective studies on treatment outcomes, to investigate rare tumor types and to explore the geographic diversity of breast cancer.

Supporting Information

S1 Fig. Differentially expressed genes against different control datasets. Differential expression analysis was done using the GSE32124 and GSE17072 GEO public datasets. The statistical conditions used, the number of differentially expressed genes and genes with contradictory direction of regulation are shown.
(PDF)

S2 Fig. DNA repair genes mutated and sample mutation load. For each sample the burden of somatic mutations was plotted against the total number of DNA repair genes with pathogenic mutations. Pearson correlation is shown (n = 12).
(PDF)

S3 Fig. Molecular alterations in PI3K pathway. Genes with pathogenic somatic mutations and differentially expressed genes and miRNAs that regulate PI3K pathway are shown. Gene, miRNA and Pathway expression levels are depicted in red for up-regulation and blue for down-regulation; genes without differential expression are marked in white. Mutations are shown as lightning icons.
(PDF)

S4 Fig. Molecular alterations in FOXM1 pathway. Differentially expressed genes and miRNAs that regulate FOXM1 pathway are shown. Gene, miRNA and Pathway expression levels are depicted in red for up-regulation and blue for down-regulation; genes without differential expression are marked in white.
(PDF)

S1 Table. Patients and sample characteristics.
(XLSX)

S2 Table. PAM50 tumor subtype classification.
(XLSX)

S3 Table. Probes of genes up-regulated in DGEA with GSE32124 (33 controls).
(XLSX)

S4 Table. Probes of genes down-regulated in DGEA with GSE32124 (33 controls).
(XLSX)

S5 Table. Genes up- and down-regulated in DGEA with GSE17072 (5 controls).
(XLSX)

S6 Table. Up-regulated genes enriched in cytogenetic bands.
(XLSX)

S7 Table. Transcription factor binding sites enriched in up-regulated genes.
(XLSX)

S8 Table. MicroRNAs up-regulated.
(XLSX)

S9 Table. MicroRNAs down-regulated.
(XLSX)

S10 Table. Integrated analysis of miRNA-mRNA relationships.
(XLSX)

S11 Table. Somatic mutations identified by whole exome sequencing.
(XLSX)

S12 Table. Genes with somatic mutations targeted with FDA-approved drugs (data base in: http://www.fda.gov).
(XLSX)

S13 Table. Molecular alterations in epigenetic modifier genes.
(XLSX)

Acknowledgments

We are thankful to Dr Behnoush Abedi-Ardekani for her kind assistance on the pathology confirmation analysis of samples.

Author Contributions

Conceived and designed the experiments: FVP MO JZ. Performed the experiments: FVP SV JZ GD RMAG NF VFO HAMM. Analyzed the data: FVP MO JZ MA CV MV FLG. Contributed reagents/materials/analysis tools: RMAG VFO. Wrote the paper: FVP MO. Performed the pathology work and analysis: HAMM. Obtained the clinical information: RMAG VFO. Blood sampling and blood DNA isolation: RMAG VFO. Contributed to critical review of the manuscript: JZ JM FLK LAH DC HAMM AM EBR VFO CPP RMAG. Prepared the final manuscript: FVP MO. Read and approved the final manuscript: FVP RMAG HAMM CPP VFO FLG LAH DC EBR AM GD NF CV MV FLK JM MA SV JZ MO.

References

1. Podo F, Buydens LM, Degani H, Hilhorst R, Klipp E, Gribbestad IS, et al. Triple-negative breast cancer: present challenges and new perspectives. Mol Oncol. 2010; 4: 209–229. doi: 10.1016/j.molonc.2010.04.006 PMID: 20537966
2. Prat A, Adamo B, Cheang MC, Anders CK, Carey LA, Perou CM. Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. Oncologist. 2013; 18: 123–133. doi: 10.1634/theoncologist.2012-0397 PMID: 23404817
3. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. N Engl J Med. 2010; 363: 1938–1948. doi: 10.1056/NEJMra1001389 PMID: 21067385
4. Cascione L, Gasparini P, Lovat F, Carasi S, Pulvirenti A, Ferro A, et al. Integrated microRNA and mRNA signatures associated with survival in triple negative breast cancer. PLoS One. 2013; 8: e55910. doi: 10.1371/journal.pone.0055910 PMID: 23405235
5. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011; 121: 2750–2767. doi: 10.1172/JCI45014 PMID: 21633166
6. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406: 747–752. PMID: 10963602
7. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012; 486: 395–399. doi: 10.1038/nature10933 PMID: 22495314
8. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490: 61–70. doi: 10.1038/nature11412 PMID: 23000897
9. Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, et al. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA. 2006; 295: 2492–2502. PMID: 16757721
10. Lund MJ, Trivers KF, Porter PL, Coates RJ, Leyland-Jones B, Brawley OW, et al. Race and triple negative threats to breast cancer survival: a population-based study in Atlanta, GA. Breast Cancer Res Treat. 2009; 113: 357–370. doi: 10.1007/s10549-008-9926-3 PMID: 18324472
11. Stevens KN, Vachon CM, Couch FJ. Genetic susceptibility to triple-negative breast cancer. Cancer Res. 2013; 73: 2025–2030. doi: 10.1158/0008-5472.CAN-12-1699 PMID: 23536562
12. Amirikia KC, Mills P, Bush J, Newman LA. Higher population-based incidence rates of triple-negative breast cancer among young African-American women: Implications for breast cancer screening recommendations. Cancer. 2011; 117: 2747–2753. doi: 10.1002/cncr.25862 PMID: 21656753
13. Lara-Medina F, Perez-Sanchez V, Saavedra-Perez D, Blake-Cerda M, Arce C, Motola-Kuba D, 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: 3658–3669. doi: 10.1002/cncr.25961 PMID: 21387260
14. Ly M, Antoine M, Andre F, Callard P, Bernaudin JF, Diallo DA. [Breast cancer in Sub-Saharan African women: review]. Bull Cancer. 2011; 98: 797–806. doi: 10.1684/bdc.2011.1392 PMID: 21700549
15. Stead LA, Lash TL, Sobieraj JE, Chi DD, Westrup JL, Charlot M, et al. Triple-negative breast cancers are increased in black women regardless of age or body mass index. Breast Cancer Res. 2009; 11: R18. doi: 10.1186/bcr2242 PMID: 19320967

16. Huo D, Ikpatt F, Khramtsov A, Dangou JM, Nanda R, Dignam J, et al. Population differences in breast cancer: survey in indigenous African women reveals over-representation of triple-negative breast cancer. J Clin Oncol. 2009; 27: 4515–4521. doi: 10.1200/JCO.2008.19.6873 PMID: 19704069
17. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30: 207–210. PMID: 11752295
18. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, et al. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc. 2007; 2: 2366–2382. PMID: 17947979
19. Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, et al. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics. 2011; 12: 436. doi: 10.1186/1471-2105-12-436 PMID: 22070249
20. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011; 27: 431–432. doi: 10.1093/bioinformatics/btq675 PMID: 21149340
21. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LAJ, Kinzler KW. Cancer genome landscapes. Science. 2013; 339: 1546–1558. doi: 10.1126/science.1235122 PMID: 23539594
22. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua SA, et al. Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-Negative Breast Cancer. Clin Cancer Res. 2014
23. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499: 214–218. doi: 10.1038/nature12213 PMID: 23770567
24. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, et al. APOBEC3B is an enzymatic source of mutation in breast cancer. Nature. 2013; 494: 366–370. doi: 10.1038/nature11881 PMID: 23389445
25. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat Genet. 2013; 45: 977–983. doi: 10.1038/ng.2701 PMID: 23852168
26. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013; 45: 970–976. doi: 10.1038/ng.2702 PMID: 23852170
27. Corney DC, Flesken-Nikitin A, Godwin AK, Wang W, Nikitin AY. MicroRNA-34b and MicroRNA-34c are targets of p53 and cooperate in control of cell proliferation and adhesion-independent growth. Cancer Res. 2007; 67: 8433–8438. PMID: 17823410
28. Hermeking H. p53 enters the microRNA world. Cancer Cell. 2007; 12: 414–418. PMID: 17996645
29. Yamakuchi M, Lotterman CD, Bao C, Hruban RH, Karim B, Mendell JT, et al. P53-induced microRNA-107 inhibits HIF-1 and tumor angiogenesis. Proc Natl Acad Sci U S A. 2010; 107: 6334–6339. doi: 10.1073/pnas.0911082107 PMID: 20308559
30. Tian S, Huang S, Wu S, Guo W, Li J, He X. MicroRNA-1285 inhibits the expression of p53 by directly targeting its 3' untranslated region. Biochem Biophys Res Commun. 2010; 396: 435–439. doi: 10.1016/j.bbrc.2010.04.112 PMID: 20417621
31. Garcia AI, Buisson M, Bertrand P, Rimokh R, Rouleau E, Lopez BS, et al. Down-regulation of BRCA1 expression by miR-146a and miR-146b-5p in triple negative sporadic breast cancers. EMBO Mol Med. 2011; 3: 279–290. doi: 10.1002/emmm.201100136 PMID: 21472990
32. Lu H, Yue J, Meng X, Nickoloff JA, Shen Z. BCCIP regulates homologous recombination by distinct domains and suppresses spontaneous DNA damage. Nucleic Acids Res. 2007; 35: 7160–7170. PMID: 17947333
33. Meng X, Yue J, Liu Z, Shen Z. Abrogation of the transactivation activity of p53 by BCCIP down-regulation. J Biol Chem. 2007; 282: 1570–1576. PMID: 17135243
34. Powell SN, Kachnic LA. Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation. Oncogene. 2003; 22: 5784–5791. PMID: 12947386
35. Craig DW, O'Shaughnessy JA, Kiefer JA, Aldrich J, Sinari S, Moses TM, et al. Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. Mol Cancer Ther. 2013; 12: 104–116. doi: 10.1158/1535-7163.MCT-12-0781 PMID: 23171949
36. Holstege H, Horlings HM, Velds A, Langerod A, Borresen-Dale AL, van de Vijver MJ, et al. 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. doi: 10.1186/1471-2407-10-654 PMID: 21118481
37. Xu H, Eirew P, Mullaly SC, Aparicio S. The omics of triple-negative breast cancers. Clin Chem. 2014; 60: 122–133. doi: 10.1373/clinchem.2013.207167 PMID: 24298072

38. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat Med. 2009; 15: 907–913. doi: 10.1038/nm.2000 PMID: 19648928

39. Maia AT, Spiteri I, Lee AJ, O'Reilly M, Jones L, Caldas C, et al. Extent of differential allelic expression of candidate breast cancer genes is similar in blood and breast. Breast Cancer Res. 2009; 11: R88. doi: 10.1186/bcr2458 PMID: 20003265

40. Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, Parkhomchuk D, et al. Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis. PLoS One. 2009; 4: e5548. doi: 10.1371/journal.pone.0005548 PMID: 19440246

41. Wagle N, Berger MF, Davis MJ, Blumenstiel B, Defelice M, Pochanard P, et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov. 2012; 2: 82–93. doi: 10.1158/2159-8290.CD-11-0184 PMID: 22585170

42. April C, Klotzle B, Royce T, Wickham-Garcia E, Boyaniwsky T, Izzo J, et al. Whole-genome gene expression profiling of formalin-fixed, paraffin-embedded tissue samples. PLoS One. 2009; 4: e8162. doi: 10.1371/journal.pone.0008162 PMID: 19997620

43. Sfakianos GP, Iversen ES, Whitaker R, Akushevich L, Schildkraut JM, Murphy SK, et al. Validation of ovarian cancer gene expression signatures for survival and subtype in formalin fixed paraffin embedded tissues. Gynecol Oncol. 2013; 129: 159–164. doi: 10.1016/j.ygyno.2012.12.030 PMID: 23274563

44. Hall JS, Taylor J, Valentine HR, Irlam JJ, Eustace A, Hoskin PJ, et al. Enhanced stability of microRNA expression facilitates classification of FFPE tumour samples exhibiting near total mRNA degradation. Br J Cancer. 2012; 107: 684–694. doi: 10.1038/bjc.2012.294 PMID: 22805332

45. Xi Y, Nakajima G, Gavin E, Morris CG, Kudo K, Hayashi K, et al. Systematic analysis of microRNA expression of RNA extracted from fresh frozen and formalin-fixed paraffin-embedded samples. RNA. 2007; 13: 1668–1674. PMID: 17698639

46. Hui AB, Shi W, Boutros PC, Miller N, Pintilie M, Fyles T, et al. Robust global micro-RNA profiling with formalin-fixed paraffin-embedded breast cancer tissues. Lab Invest. 2009; 89: 597–606. doi: 10.1038/labinvest.2009.12 PMID: 19290006

47. Goswami RS, Waldron L, Machado J, Cervigne NK, Xu W, Reis PP, et al. Optimization and analysis of a quantitative real-time PCR-based technique to determine microRNA expression in formalin-fixed paraffin-embedded samples. BMC Biotechnol. 2010; 10: 47. doi: 10.1186/1472-6750-10-47 PMID: 20573258

48. 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: 415–421. doi: 10.1038/nature12477 PMID: 23945592

49. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012; 486: 405–409. doi: 10.1038/nature11154 PMID: 22722202

50. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012; 486: 400–404. doi: 10.1038/nature11017 PMID: 22722201

51. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102: 15545–15550. PMID: 16199517

52. Shen H, Laird PW. Interplay between the cancer genome and epigenome. Cell. 2013; 153: 38–55. doi: 10.1016/j.cell.2013.03.008 PMID: 23540689

53. Timp W, Feinberg AP. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nat Rev Cancer. 2013; 13: 497–510. doi: 10.1038/nrc3486 PMID: 23760024

Appendix II: Base changes in tumour DNA have the power to reveal the causes and evolution of cancer

Monica Hollstein, Ludmil B. Alexandrov, Christopher P. Wild, **Maude Ardin** and Jiri Zavadil

Oncogene, 2016

OPEN

Oncogene (2016), 1–10

www.nature.com/onc

REVIEW

Base changes in tumour DNA have the power to reveal the causes and evolution of cancer

M Hollstein[1,2], LB Alexandrov[3,4], CP Wild[5], M Ardin[1] and J Zavadil[1]

Next-generation sequencing (NGS) technology has demonstrated that the cancer genomes are peppered with mutations. Although most somatic tumour mutations are unlikely to have any role in the cancer process *per se*, the spectra of DNA sequence changes in tumour mutation catalogues have the potential to identify the mutagens, and to reveal the mutagenic processes responsible for human cancer. Very recently, a novel approach for data mining of the vast compilations of tumour NGS data succeeded in separating and precisely defining at least 30 distinct patterns of sequence change hidden in mutation databases. At least half of these mutational signatures can be readily assigned to known human carcinogenic exposures or endogenous mechanisms of mutagenesis. A quantum leap in our knowledge of mutagenesis in human cancers has resulted, stimulating a flurry of research activity. We trace here the major findings leading first to the hypothesis that carcinogenic insults leave characteristic imprints on the DNA sequence of tumours, and culminating in empirical evidence from NGS data that well-defined carcinogen mutational signatures are indeed present in tumour genomic DNA from a variety of cancer types. The notion that tumour DNAs can divulge environmental sources of mutation is now a well-accepted fact. This approach to cancer aetiology has also incriminated various endogenous, enzyme-driven processes that increase the somatic mutation load in sporadic cancers. The tasks now confronting the field of molecular epidemiology are to assign mutagenic processes to orphan and newly discovered tumour mutation patterns, and to determine whether avoidable cancer risk factors influence signatures produced by endogenous enzymatic mechanisms. Innovative research with experimental models and exploitation of the geographical heterogeneity in cancer incidence can address these challenges.

Oncogene advance online publication, 6 June 2016; doi:10.1038/onc.2016.192

INTRODUCTION

Cancer is a genetic disease, and mutations in genes that drive cancer constitute the overriding molecular events leading to malignant growth. During the first decade of the next-generation sequencing (NGS) revolution, the focus of whole-genome and whole-exome cancer sequencing projects was to describe the genome landscape of major human cancers, that is, to identify groups of genes (driver genes) that contribute to the growth of different types of tumours when mutated.[1,2] This massive effort has made it clear that the set of mutated driver genes in cancer genomes typically consists of fewer than 10 in any given tumour. Driver genes provide a blueprint of the malignant process, and offer targets for specific therapies.[3–5] In less than a decade, NGS identified most genes in the genome that can provide a growth advantage to a cell if mutated. Fewer than 1% of all human genes appear to have this potential to drive neoplastic development. A characteristic subset of driver genes harbouring deleterious mutations has been identified for each major cancer type, corroborating the notion that cancer is many diseases, each type following an underlying developmental path. Although there is considerable heterogeneity in the genome landscapes of different cancers, it appears that all driver gene products affect a common set of biological pathways.[5,6]

Although the first goal of tumour NGS data analysis was to identify driver gene mutations buried amongst a plethora of accumulated sequence changes, it became apparent that the frequency and types of common base substitutions differed substantially across cancer types. Furthermore, mutation pattern heterogeneity could arise in distinct sets of tumours of the same type.[3,7] Although it is generally accepted that some of this diversity stems from differences in patient exposure history, cursory perusal of mutation profiles did not lead significantly further in identifying the sources of mutations beyond what had been achieved previously through mutation spectra analysis of single cancer genes. Once methods were applied to parse enigmatic mutation catalogues into specific mutational signatures, however, the picture changed entirely. Computational mining of information previously locked in mutation databases allowed tight associations to be made between specific cancer risk factors and unique patterns of sequence changes in tumours.

ESSENTIAL OBSERVATIONS FROM SEQUENCING SINGLE CANCER GENES IN TUMOURS: IMPLICATIONS FOR CANCER AETIOLOGY

Mutation patterns amongst different cancer types are different

Mutation analysis of individual cancer genes, which preceded scrutiny of NGS mutational catalogues, provided the first evidence that carcinogenic insults leave mutational 'fingerprints' on tumour DNA. In the decades leading up to tumour NGS studies, catalogues

[1]Molecular Mechanisms and Biomarkers, International Agency for Research on Cancer, World Health Organization, Lyon, France; [2]Faculty of Medicine and Health, University of Leeds, Leeds, UK; [3]Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, Los Alamos, NM, USA; [4]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM, USA and [5]International Agency for Research on Cancer, World Health Organization, Lyon, France. Correspondence: Dr M Hollstein or Dr J Zavadil, Molecular Mechanisms and Biomarkers, International Agency for Research on Cancer, World Health Organization, 150 cours Albert Thomas, Lyon 69008, France or Dr LB Alexandrov, Theoretical Biology and Biophysics Group (T-6), Los Alamos National laboratory, P.O. Box 1663, Los Alamos, NM 87545, USA.
E-mail: M.Hollstein@leeds.ac.uk or zavadilj@iarc.fr or lba@lanl.gov
Received 13 February 2016; revised 31 March 2016; accepted 31 March 2016

of DNA sequence changes in frequently mutated genes such as the *TP53* tumour suppressor gene or the *K-ras*, and *B-raf* oncogenes offered a first glimpse of the mutational pathways operating in human cancers. Analysis of skin and lung tumours provided convincing demonstration of an environmental impact on tumour mutation patterns.[8–10] Numerous reports contributed to the understanding that exposure to ultraviolet (UV) light, the primary cause of skin cancer, is responsible for the uniquely characteristic C to T transitions at dipyrimidines in skin tumours, and that tobacco smoking causes G to T transversions, the predominant sequence changes present in lung tumours.[11] (Note: When possible, we describe a mutation by naming the base proposed to carry the pre-mutagenic lesion rather than by using the COSMIC system (Catalogue of Somatic Mutations in Cancer; cancer.sanger.ac.uk), which uniformly names the pyrimidine of the Watson-Crick base pair. When the pre-mutagenic lesion is currently unknown, we employ the COSMIC system.) Despite the limited scope of single-gene sequencing, these and other valuable insights from such projects continue to emerge, particularly from the analysis of *TP53*. One reason why sequencing *TP53* is particularly informative in revealing sources of mutagenic insult is that any one of numerous single base changes along the coding sequence is sufficient to disrupt its proper function.[12,13] Such diversity of potential mutations and sequence contexts can reveal discrete mutation profiles. Each *TP53* mutation in a set of tumours of a specific type is classified according to the type of base change, strand orientation, and sequence location, and the frequencies of specific alterations are then analysed. The tumour-specific patterns that emerge (such as the *TP53* G to T transversions on the non-transcribed strand in smokers' lung tumours clustering at hotspots in codons 157, 158 and 273) represent rudimentary 'signatures' produced by the action of mutagenic processes.[11,14] As fundamental DNA-damaging properties of human carcinogenic agents such as UV light and tobacco carcinogens had been well-characterized in the laboratory,[15–17] the effects of these agents on DNA were promptly recognized in skin and lung tumour *TP53* mutation spectra, a major step forward at the time.

Within this single-gene framework, however, the mutation spectrum is small in scale and the approach is fraught with limitations. First, as each patient analysis typically contributes just one mutation, fingerprints only begin to emerge as data from many individuals are pooled. Second, as driver gene mutations are selected during cancer development, the types of tumour mutations likely to be detected are generally limited to the specific changes and gene locations capable of unleashing oncogenic potential are not necessarily characteristic of the genome's mutation load as a whole. The *B-raf* mutation spectrum in melanomas illustrates the limitations in single-gene analysis in revealing sources of a somatic mutation burden. In the *B-raf* driver gene, the mutagenic risk factor fails to leave its identifying fingerprint.[8] Almost all *B-raf* mutations in melanoma are T to A transversions, yet the primary risk factor is UV light, powerful mutagen that produces C to T and CC to TT base changes at pyrimidine dinucleotides. An explanation for this anomaly is that most oncogenic *B-raf* mutations occur at a hotspot, the second nucleotide of *B-raf* codon 600. The sequence context (ACA GTG AAA) cannot capture the hallmark dinucleotide target of UV radiation. In contrast, melanoma mutations in *TP53* are dispersed across the locus and do indeed display the UV-characteristic C to T transitions at dipyrimidines and CC to TT tandem mutations. (Of general note, not all mutations that a carcinogen induces will be typical of its action on DNA. Thus, T to A mutations in the *B-raf* gene of melanoma, although uncharacteristic of UV exposure, may well have arisen from exposure to UV, even though a T to A substitution is not the most likely molecular change that sunlight generates.)

Within a cancer type, mutation patterns in a single gene can diverge widely when groups of patients with different exposure histories are examined

Whilst it was highly plausible that risk factors are responsible for some of the mutation pattern diversity amongst different types of cancer, demonstration of a specific risk factor mutation pattern present in tumours from exposed patients, but absent in non-exposed patients with the same type of cancer, strengthens the argument considerably. Extensive supporting evidence has come from *TP53* analysis of lung, urothelial and liver cancers. The G to T mutation fingerprint discovered in lung cancers and linked to tobacco smoking is not evident in lung cancer patients who are never-smokers, and the greater the tobacco smoke exposure, the more pronounced is the G to T mutation load in sentinel driver genes such as *TP53*.[10] In urothelial cancers from patients exposed to the plant carcinogen aristolochic acid (AA), there is a striking preponderance of *TP53* A to T mutations on the non-transcribed strand of DNA, the primary type of mutation induced in laboratory mutagenesis experiments with AA.[18,19] The signature does not appear in patients with no history of AA exposure. Finally, a unique liver cancer *TP53* mutation pattern, characterized by strand-biased G to T substitutions predominantly at codon 249, is present in hepatocellular carcinomas (HCC) from geographical regions (for example, parts of China and sub-Saharan Africa) where there is chronic, high-level exposure to aflatoxin B1 (AFB1), and hepatitis B virus infection is prevalent.[17,20,21] In populations where other risk factors prevail and exposure to AFB1 is minimal or absent, *TP53* mutations in HCC are diverse in type and location.[22] A variety of laboratory test systems demonstrated that AFB1 induces primarily G to T mutations. The codon 249 G to T hotspot mutation has shown its use as a powerful molecular biomarker of HCC risk and disease burden in regions where exposure to AFB1 is high, but it would be of little value as a biomarker in cohorts with no exposure to this carcinogen.

Overall, DNA sequencing of *TP53* continues to generate evidence supporting the prediction that two cohorts with the same cancer type but exposed to different environmental risk factors can have different characteristic mutations in their tumours. Mutation spectra in oncogenes and tumour suppressor genes have also indicated that the multiplicity of distinct risk-associated *TP53* mutation patterns in human tumours presaged the diversity in mutation patterns now emerging from tumour NGS data.

THE GAME CHANGER: GENOME-WIDE SEQUENCING DATA AND COMPUTATIONAL ANALYSIS

Mutation research has been witness to three seminal advances, each of which prompted a flurry of activity in laboratories around the world. First, in the 1970s, development of the rapid *Salmonella*/microsome assay for testing mutagenicity of chemicals, and subsequently the report on test results with 300 chemicals, established the fact that the majority of known and suspected human carcinogens are mutagenic.[23] More than a decade later, Vogelstein and colleagues discovered that colorectal cancers harbour a variety of inactivating point mutations in *TP53*.[24] This finding prompted a deluge of reports describing *TP53* mutations in a variety of human tumours. The fact that the mutations were found in target sequences large and complex enough to reveal different mutation patterns in various tumour types was a turning point because tumour *TP53* mutations provided the first comprehensive evidence in clinical samples that exposure to mutagenic carcinogens leave fingerprints on tumour DNA.[25] With the advent of NGS technologies, the third quantum leap in mutation research on cancer aetiology is now upon us. NGS-derived mutation data constitutes a blurred mixture of fingerprints from different mutagenic processes, however,

necessitating de-confounding computational procedures to identify discrete mutational signatures in simple mathematical terms.[26] The somatic mutations found in cancer genomes are approximated as a linear mixture of multiple mutational signatures, each contributing a different number of mutations to different genomes:

$$\text{Mutations} = \text{Signatures} \times \text{Exposures}$$

In principle, the known set of mutations in cancer genomes is used to find the optimal set of signatures and respective exposures that best describe the original catalogues of somatic mutations. This problem can be considered as a specific case of a blind source separation problem, and the challenge is to unscramble not-observed latent variables (that is, mutational signatures and their exposures) from a set of mixtures (that is, somatic mutations in cancer genomes). To 'unmix' and reconstruct the original sources from the records, a blind source separation algorithm is needed for best possible extraction of original signals from mixtures. The unmixing and reconstruction of the original signals is based on constrained and/or regularized optimization procedure minimizing an objective cost function together with a few imposed constraints, such as maximum variability, statistical independence, non-negativity, smoothness, sparsity, simplicity, and so on. The choice of optimization constraints is based on prior knowledge about the processed data, and hence the constraints could be different for every particular case. The non-negative nature of somatic mutations requires at the very least applying a non-negative constraint for solving the cancer genomics blind source separation problem. Alexandrov *et al.*[26] used a widely applied approach designated non-negative matrix factorization (NMF; Figure 1) to provide an effective solution.[27] NMF does not seek statistical independence or constrain any other statistical property of the mixed signals, and thus allows the estimated sources to be partially or entirely correlated. When tumour mutational catalogues are analysed with mathematical procedures such as NMF, numerous carcinogenic fingerprints hidden in a vast set of human NGS-analysed tumours can be separated and identified with unprecedented clarity, fast-forwarding our understanding of mutation origins during the evolution of cancer.[26,28]

Despite the apparent neutrality of bystander mutations in the cancer process, their sheer numbers promise to provide a far more powerful way than individual onco-mutation analysis to observe signatures of mutagenic activity. Understanding the mutagenic processes corresponding to NGS mutational signatures, however, continues to rely on finding matches with experimentally induced signatures or other laboratory data.

Diverse mutational processes are responsible for the heterogeneity in tumour NGS mutation spectra amongst different cancer types

The first NMF-based pan-analysis of NGS data from a broad assortment of different cancers demonstrated unequivocally that tumour types differ in their genome-wide mutation profiles, and presented compelling argument that distinct risk factors associated with each cancer type are likely to explain much of the heterogeneity in mutation spectra across tumour types.[28] Twenty-one distinct mutational signatures were extracted from mutation data on 30 types of cancer from 7042 patients in this unprecedented study, and a known cancer risk factor or endogenous molecular process was putatively assigned to many of the signatures. The number of distinct mutational signatures is now at 30 (source: COSMIC) and may soon approach 50 as the results of pan-cancer analyses become validated, and as patients

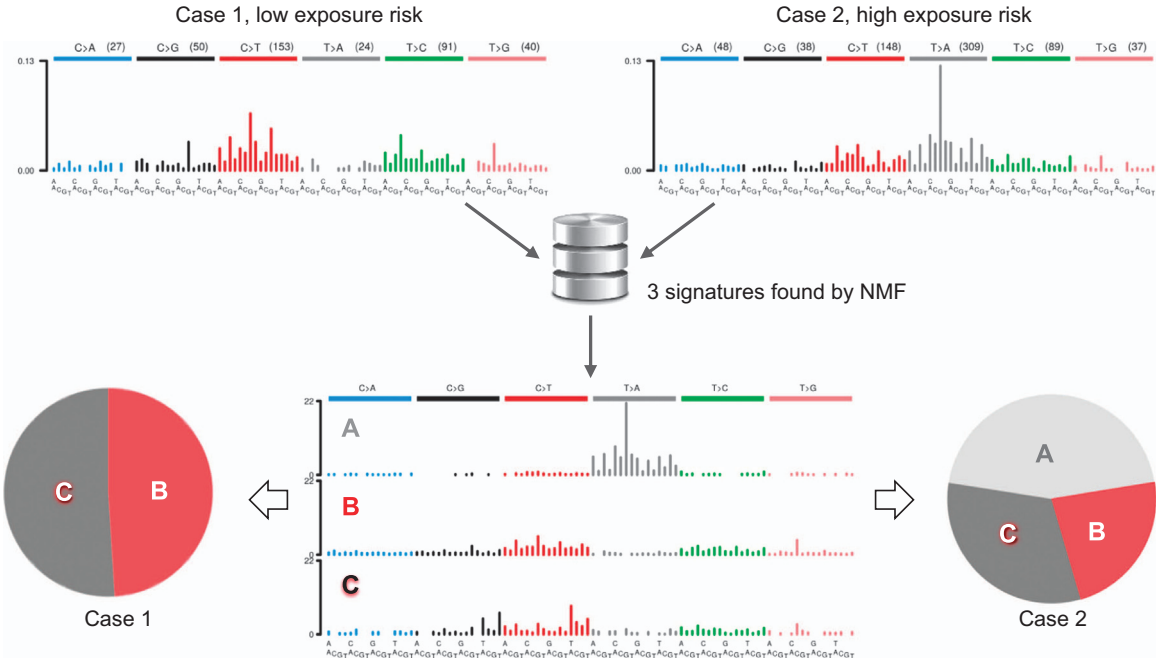


Figure 1. When patients with the same cancer type have different exposure histories, the mutation patterns in their tumours can be strikingly different. Two representative cases of upper urinary tract urothelial tumours from regions of either low or high risk of exposure to the carcinogen aristolochic acid[97] were analysed using whole-exome sequencing. The single-base substitution distribution spectra are shown on top. Performing NMF on the studied case series identified three distinct mutational signatures (**A**, **B** and **C**; middle panel). The pie charts show the proportionate contribution of individual signatures to the mutational load in each tumour. The absence of signature A in case 1 argues that the two tumours have distinct aetiologies.

Table 1. Mutational signatures assigned to IARC Group 1 carcinogen exposures

Name	*Exposure*	*Group 1 carcinogen*	*Chemical class*	*Characteristic pre-mutagenic DNA lesion*	*Signature hallmarks*	*Prominent trinucleotide target in the signature*
Signature 4	Tobacco smoke	Benzo[a]pyrene	PAH	(+)benzo[*a*]pyrene-7,8-dihydrodiol-9, 10-epoxide-dG adduct	G to T GG to TT tandem mutations Transcriptional strand bias	GGG GGA
Signature 7	Sunlight	Ultraviolet light	NA	Py-Py photodimers	C to T at dipyrimidines CC to TT tandem mutations Transcriptional strand bias	TCC
Signature 11	Chemotherapy	Temozolomide	Alkylating agent	O^6-methylguanine	G to A Transcriptional strand bias	GGG GGA
Signature 22	Dietary contaminant (grain); herbal medicine	Aristolochic acid	Plant alkaloid	7-(deoxyadenosin-N(6)-yl) aristolactam I adduct	A to T Transcriptional strand bias	CAG
Signature 24	Dietary contaminant (groundnuts)	Aflatoxins	Fungal toxin	8,9-dihydro-8-(N7-guanyl)-9-hydroxyaflatoxin B_1 adduct	G to T Transcriptional strand bias	GGC
Signature 29	Tobacco chewing	Unspecified	Unspecified	Unspecified	G to T Transcriptional strand bias	TGC GGC TGT

Abbreviations: IARC, International Agency for Research on Cancer; NA, not applicable; PAH, polycyclic aromatic hydrocarbon; Py, pyrimidine. Group 1 refers to the IARC classification of substances for which there is sufficient evidence of carcinogenicity to humans. The third column lists sources of exposure with documented links to cancer risk. Bases and context sequence are shown to reflect the base targeted by the mutagen (for example, as 5′-GGA-3′ for B[a]P, but as 5′-TCC-3′ for UV). The targeted base in its preferred sequence context is underlined in the last column. The targeted trinucleotides in the last column are extracted from signature analysis of human tumours; analysis of experimental models with single, controlled exposures recapitulates major features from the human data (Figure 2).

from geographic areas not previously tested become examined. Table 1 describes five signatures assigned to specific human carcinogenic exposures. (Note: In the following discussion, different signatures are referred to according to their unique identifying number. See http://cancer.sanger.ac.uk/cosmic/signatures)

The diversity in mutation patterns amongst cancer types can be illustrated by a comparison of signatures in small cell lung cancer, acute myeloid leukaemia and cutaneous melanoma.[28] In each of these cancer types one signature (but not the same one) contributed >85% of the total mutational burden. The tobacco smoking-associated signature 4, characterized by G to T transversions with transcriptional strand bias, dominated in small cell lung cancers, whereas acute myeloid leukaemia mutations were overwhelmingly C to T transitions at CpG dinucleotides (signature 1), presumably attributable to spontaneous deamination of 5-methylcytosine, and clearly distinguishable from the UV signature C to T transitions at dipyrimidines (signature 7) in melanoma.

The mutation spectrum derived from NGS of a tumour is composed of superimposed signatures left by various mutagenic insults

In most cancer types, parsing of NGS mutational catalogues demonstrated the presence of several distinct mutational signatures, in keeping with cancer aetiologies where multiple exposures are thought to significantly contribute to risk. The fact that in NGS analysis, each tumour provides an entire spectrum of mutations (rather than a set of tumours required for single gene-based analysis) has offered unprecedented opportunity to explore the multi-factor aspect of human cancer. Despite caveats mentioned below regarding signatures in branch mutations accumulating during clonal evolution, genome-wide mutations in a tumour can be displayed as a weighted composite of distinct mutational signatures, allowing a first approximation of the relative contribution of each risk-associated signature to the total mutation burden in the tumour. With NMF or similar mathematical approaches,[27–30] a rough estimate of the relative impact of multiple risk factors on the total mutation load can be obtained, a goal that was out of reach in the single-gene mutational analysis era. In the initial study applying NMF to NGS data from 30 different tumour types, liver cancer displayed the greatest number of distinct mutational signatures, presumably reflecting the multi-factorial aetiology of cancer at this site discernible from the data archives used. Seven signatures were identified, amongst them signature 16, apparently unique to liver cancers, which was detected in 90% of the tumours sequenced, and contributed anywhere from a few percentages to over half of all the somatic mutations recorded in a given sample. The cause of signature 16 mutations, characterized by strand-biased A to G transitions at NpApT sites, is unclear. This observation is intriguing because HCC is one of the few cancers with several known major risk factors, notably infection by hepatitis B or C viruses, alcohol consumption and exposure to AFB1. A recent study uncovered signature 24, one of the signatures characterized by frequent strand-biased G to T transversions, in six hepatitis B virus-infected HCC patients originating from subtropical Africa.[31] Extended cohort-specific as well as experimental studies are warranted to strengthen the proposed link between this signature and aflatoxin B1 exposure.

At present, of the first 30 distinct signatures defined, 60% have been provisionally assigned to known carcinogens or mutational processes. The remaining orphan signatures highlight the dearth of experimental mutation research, sending out a priority research call.

Specific endogenous mutational processes have a major impact on the mutation burden in human populations

The risk of sporadic adult cancer increases exponentially with age.[32] Deamination of 5-methylcytosine, a well-studied endogenous spontaneous mutagenic process known to erode DNA sequence integrity, presents as C to T transitions at CpG dinucleotides, the ubiquitous age-associated signature labelled signature 1.[33] Tumour mutation catalogues from almost all 30 types of cancers in the seminal study of Alexandrov *et al.*[28] had at least some trace of this signature, and in some cancers signature 1 predominated.

The accumulation of this and other classes of mutational events, such as those stemming from spontaneous base hydrolysis or the inherent infidelity of DNA replication and repair,[34] is to a certain extent essentially inevitable, as are some cancers. A recent study suggested that 10–30% of cancers can be primarily attributed to intrinsic factors,[35] although some argument persists regarding the proportion of human cancers that presumably cannot be avoided by changes in lifestyle or environment.[36] However, much remains to be understood with regard to the effect of external exposures on endogenous pro-mutagenic processes mentioned. On the basis of geographical disparities in cancer incidence within cancer types,[37,38] current estimates suggest ~90% of the global cancer burden could in principle be avoided, a large fraction of which may harbour mutational signatures that could be linked to patient exposure history. In contrast, two signatures of endogenous mutational processes discernible in practically all cancer types, signature 1 (C to T at CpG) mentioned above, and signature 5 (a diffuse pattern produced by unknown underlying molecular mechanism(s)), have been linked to age, the most inevitable and ubiquitous cancer risk factor. These two mutation patterns, attributed to 'clock-like' cellular processes, are the only signatures described thus far for which a correlation was found between the number of such mutations and the chronological age of patients at diagnosis.[33] Although it is unclear to what extent genetic background or external factors can accelerate this internal clock in normal cells, the tumours in which these signatures predominate are more likely to be those that contribute to the baseline incidence of cancer in humans.[35]

Surprisingly, of the first 30 signatures revealed by NMF, almost half correspond to patterns generated by enzymatic processes affecting DNA homoeostasis.[39] For example, signatures 9 and 10 are similar to mutation patterns left in the wake of DNA repair polymerases *eta* and *epsilon*, respectively, and signatures 6, 15 and 20 imply defective DNA mismatch repair. Further, signature 3 has been found in the majority of samples harbouring pathogenic BRCA1/2 mutations indicating that this signature reflects failure of DNA double strand repair by homologous recombination.[40] It has been long recognized that cancer patients with inherited deleterious mutations in DNA repair enzymes have tumours with a hypermutator phenotype.[41] However, inherited cancer syndromes of this class are relatively rare, so the demonstration that enzymatic DNA maintenance mechanisms appear to contribute to diverse types of sporadic cancers raises the question as to whether avoidable, known cancer risk factors can influence the impact from these pathways on the human mutation burden. In particular, the extent to which cancer risk factors that do not act through a direct mutational mechanism exert an influence on genome-altering cellular processes is one of the most enticing areas of cancer research, offering rich opportunities for laboratory science and epidemiology.

It is worth remembering that the human tumours subjected to NGS in the first phase of studies were not selected to address hypotheses about aetiology. Patients were not necessarily representative of the patient population for a given type of cancer, being typically recruited from a small number of high-income countries, and little epidemiological data were available or collected on the exposure history of the subjects. It is thus premature to draw conclusions about the number or prevalence of distinct mutational signatures occurring for a given cancer worldwide.

Modulation of the activity of the APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) family of deaminases

Remarkably, the first signature analysis of NGS data[28] revealed that 16 of the 30 cancer types displayed signatures that matched the mutator activities of APOBEC deaminases (signatures 2 and 13). In connection with its eponymous function, this large family of enzymes has several biological tasks, including viral restriction and suppression of retrotransposition.[42] The collateral damage these enzymes inflict on single-stranded genomic DNA has been characterized extensively in experimental model systems, facilitating recognition of their mutational impact on human tumour DNA.[43–46] On the basis of this characterization, a role for APOBEC3A and/or APOBEC3B in human cancer is more likely than for other members of the family. The putative contribution of the APOBEC3 enzyme activity to the total tumour mutation load reported in several independent NGS studies of breast tumours[47,48] is an important clue in elucidating the incompletely understood aetiology of sporadic breast cancer. With respect to APOBEC3 dysregulation in this cancer type, alterations at the gene locus itself (coding sequence or promoter mutation, gene copy-number polymorphism) and induction of enzymatic activity by factors in the cellular environment may be responsible. [49–51]

In-depth exploration of APOBEC expression modulation by cancer risk factors is needed in the wake of these recent surprising discoveries on the putative impact of APOBEC on the human mutation burden. Interestingly, significant numbers of signature 2 mutations are present in cervical cancer and in head and neck tumours,[52] two types of cancer with human papillomavirus (HPV) involvement.[53] Elevated APOBEC3 activity in HPV-infected cells would be a further manifestation of the APOBEC gene family responses to viral infection.[54,55] In a recent study, mutations related to the APOBEC signatures 2 and 13 found in HPV-positive head and neck cancers were reported enriched relative to the HPV-negative counterparts.[56] In most cancer types exhibiting APOBEC dysregulation, however, the underlying causes remain enigmatic, with the exception of the small numbers of tumours found to harbour gene copy polymorphisms or deleterious mutations involving the APOBEC locus.

Physicochemical mutational processes, 'amorphous' risk factors and co-mutagenic agents: SEVERAL elephants in the room?

In general, the mutagenic impact of reactive chemicals in the internal environment of the cell, and external influences on the mutagenic potential of endogenous enzymes are difficult to assess. Furthermore, it is not known how or to what extent established but 'amorphous' risk factors with no assigned genome-wide mutational signature, such as obesity, chronic inflammation, physical inactivity, and reproductive history, modulate mutation patterns. The chemical properties of reactive oxygen species, nitrogen radicals and lipid peroxidation products associated with oxidative stress and chronic inflammation link them directly to DNA damage and these molecules are considered an important source of tumour mutations.[57,58] Nevertheless, information on the relative contribution from such sources to tumour mutation load is imprecise, and the specific patterns in base substitution distribution they might produce are ill-defined. Attack on DNA by endogenous cellular chemicals has been shown in numerous studies to elicit specific classes of base substitution, however. A recent study reported that DNA exposed to hydrochlorous acid, a chemical secreted by neutrophils in inflamed tissues, acquired 5-chlorocytosine residues, a modification that caused transitions to T, a common mutation type overall in human cancers even when the particular subclass CpG to TpG, attributable to deamination of 5-methylcytosine (signature 1), is not considered.[59]

Mathematical analysis of data from fit-for-purpose NGS studies, for example by comparing mutations in distinct risk cohorts, should bring more clarity to this prickly topic.[60] 'Amorphous' risk factors present no small challenge; whereas many chemical carcinogens produce unique DNA adducts that serve as traces of exposure, episodic exposures from endogenous chemical flux, or exposure to a non-mutagenic agent acting on endogenous mutational processes from a distance, are difficult to pinpoint.

Finally, some risk factors may impact risk primarily by modulating a different trajectory of cancer development such as immune surveillance of cancerous cells, and not by increasing the mutation load.

Episodic exposures in cancer evolution

Two recent reports on the multi-clonal evolution of lung cancer and the role of APOBEC3B activity, in which truncal (early) mutations were compared against branch (more recent) mutations illustrate how temporal shifts in mutation patterns feed into the mutational landscape of a full-blown cancer.[61,62] The two studies, which traced lung cancer development by sampling tumours at multiple locations, concluded that APOBEC3B dysfunction typically exerts effects later in the evolution of the primary clone. The enzyme's signature was evident amongst branch mutations but not in truncal mutations. Thus, whilst parsing of a mutational catalogue can estimate the relative importance of multiple signatures, and hence exposures in the natural history of the cancer, the percentage of the total mutation burden in the late stages of cancer that are attributable to a given signature/exposure may not necessarily indicate the relative importance of multiple environmental exposures in initiating a cancer. Obtaining multiple biopsies of an exposed organ or cancer to assess tissue burden of mutant cells, or to retrace the evolution of the mutational load and the timing of distinct mutational insults, is a strategy that gains power from NGS and mutational signature analysis.[61,63–68] In principle, one could revisit the migrant studies or time-trend studies of descriptive epidemiology but with genome-wide mutational analysis. For example, changes in mutation pattern following changes in risk factor exposure over a lifetime could be tracked in cancers from migrant populations, particularly when the difference in exposure patterns and cancer incidence between the patients' country of origin and the subsequent place of residence is extreme. An example would be migrants from Africa to Europe where exposure to the dietary carcinogen AFB1 is markedly different. The International Agency for Research on Cancer World Cancer Report 2014 contains numerous examples of widely differing exposure patterns and more than 10-fold geographical discrepancy in incidence for a number of common cancers.[38] Alternatively, one could examine cancer types that have seen rapid changes in incidence over time, an example being the increases in countries undergoing rapid development, offering opportunities to compare spectra for the same tumour in the same population but in the face of different environmental and lifestyle exposures.

EXAMPLES OF MUTATION SPECTRA HETEROGENEITY WITHIN A CANCER TYPE, ATTRIBUTABLE TO DIFFERENCES IN RISK FACTOR EXPOSURES

Evidence for a direct role of external risk factors in shaping human tumour mutation spectra is now accumulating from NGS projects specifically designed to capture this information by comparison of groups of patients with the same type of cancer, but differing in exposure to a known cancer-causing agent. Investigations along these lines show that mutation patterns can indeed be heterogeneous within a cancer type and that differences in risk factor exposures explain this variation. Three prominent examples are discussed here that parallel observations from earlier single-gene studies.

Mutation patterns attributable to tobacco smoking

The great majority of lung cancers worldwide arise in patients who smoke or have smoked tobacco. The outstanding features of the lung cancer NGS mutation spectrum, corroborated by several projects involving hundreds of lung cancer patients, are (i) the presence of a distinct strand-biased G to T transversion signature in smokers but, crucially, absent in never-smokers, and (ii) the high numbers of somatic mutations per tumour.[7,28,69,70] Computational methods have defined signatures provisionally attributable to tobacco smoke exposure although preferred sequence contexts where presumptive tobacco-associated transversions accumulate in respiratory tract cancers are not fully established, perhaps reflecting the chemical complexity of tobacco smoke. It is unlikely that NMF of mutational catalogues will differentiate between fingerprints of two distinct carcinogens that both induce strand-biased G to T substitutions should the preferred sequence contexts of the two chemicals overlap significantly. Tobacco smoke (along with alcohol consumption and HPV infection) is a principal risk factor for head and neck cancers as well as lung cancer. As expected, NGS analysis of 74 head and neck cancers, 89% of which were from patients with a history of tobacco use, identified a prominent strand-biased G to T mutation pattern similar to findings in smokers' lung tumours.[71] The highest prevalence of the transversions occurred in tumours with the highest mutation burden overall, suggesting that G to T mutations could serve as a readout of tobacco smoke exposure. The mutagenic impact of tobacco carcinogens across various tissues is not uniform, however; bladder cancers of smokers do not have the same mutation profile as smokers' lung tumours.[72] Differences in tissue distribution and metabolism of carcinogens in tobacco smoke are two of the many factors potentially responsible for multiple tumour type-specific mutation patterns produced by a given exposure. With respect to head and neck cancers, the tissue-specific effect of tobacco smoke is a particularly complex issue when tumours of many different cell types and subsites are grouped together for analysis. A recent NGS study that addressed this problem revealed that mutations in tongue squamous cell carcinomas do not exhibit a pattern corresponding to the spectrum found in smokers' lung cancers, whereas mutations in tumours of the larynx do.[73]

The AA fingerprint

Epidemiological and experimental evidence have long conspired to incriminate AA in the aetiology of upper urinary tract urothelial carcinoma (UTUC).[74] AA is a potent plant mutagen that contaminates grain in some regions, such as rural areas along the lower Danube River, and is present in *Aristolochia*-containing herbal medicines popular in a number of countries. In two recent groundbreaking NGS studies in which the specified objective was to examine genome-wide mutation patterns in AA-associated UTUC,[75,76] the causal link between AA exposure and cancer could be established beyond reasonable doubt because of the convergence of several findings. First, the AA mutational signature was confined primarily to patients with documented exposure to AA (measurements of AA-derived adducts on adenine and/or patient exposure history). Second, the AA signature was reproduced in cells experimentally, and third, AA signature mutations (A to T transversions on the non-transcribed DNA strand at CpApG trinucleotides) were detected in somatically mutated driver genes. Clear cell and chromophobe renal cancers of patients from some regions of Eastern Europe also display this remarkably distinctive mutation pattern.[77,78] From mutational signature analysis it is now suspected that AA exposure may also be a contributing factor in causing hepatobiliary and bladder cancers.[76,79,80]

The mutational signature of a chemotherapeutic alkylating agent

Temozolomide (TMZ) is a human carcinogen, and a strong DNA alkylating agent used in the treatment of brain cancer and melanoma. Given its mutagenic properties, it came as no surprise to find that recurrent tumours of glioma patients treated with the compound displayed a heavy burden of G to A transitions, a base substitution induced by this class of chemicals when the alkylated deoxyguanine (O^6-methyldeoxyguanine) mispairs with thymine.

The naturally occurring 'control group', patients not offered TMZ therapy, also have C:G to T:A transitions in their recurrent tumours, but these occur primarily at the CpG sequence contexts (attributable to spontaneous deamination of methylated cytosine), unlike the TMZ-associated transitions clustering at CpC and CpT dinucleotides.[28] In a study of 23 patients, the mutation burden in recurrent tumours of patients treated with TMZ was up to 10-fold higher than in cancers of individuals not exposed to TMZ, and 98% of this mutation load were 'TMZ-type' transitions.[81] The study also revealed that TMZ exposure influenced not only the type of mutation but also the identity of the driver genes mutated in the tumours. In other words, this study suggests that a risk factor can participate in determining not only which types of mutations appear in the tumour, but also which genes become dysfunctional and drive the cancer process.

ORPHAN SIGNATURES AND THE CALL FOR MORE MUTAGENESIS STUDIES IN EXPERIMENTAL MODELS

Genome-wide mutation data have unveiled 'orphan' signatures undecipherable with the experimental and epidemiological data currently at hand, providing a major incentive for further targeted experimental work to decode enigmatic patterns and link them to causes of cancer. The oesophageal adenocarcinoma-linked mutation profile characterized by T to G substitutions at NpTpT is an example of a profile not readily linked to the major risk factors for

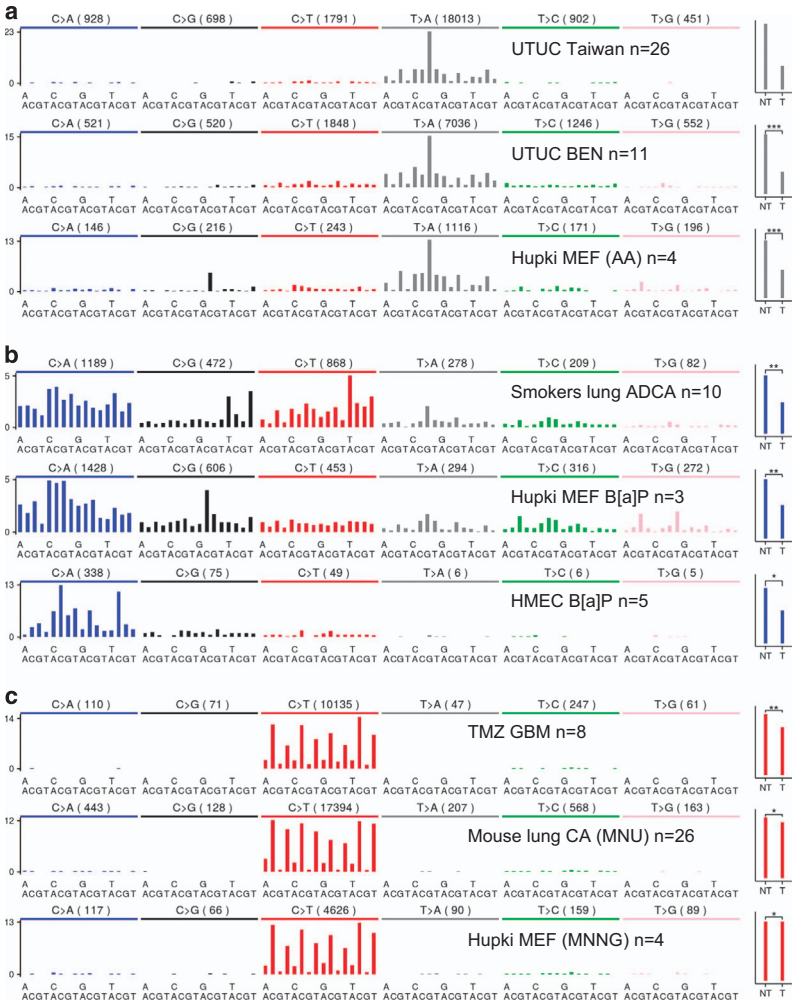


Figure 2. A carcinogen's fingerprint in human tumour DNA can be reproduced in experimental systems. Mutation distribution spectra (showing frequency of base substitution type and context) from exome sequencing of primary human tumours, cells exposed in culture, or tumours of exposed mice. (**a**) Upper panels: spectra in upper urinary tract urothelial carcinomas (UTUC) of patients from Taiwan, China and from Balkan Endemic nephropathy (BEN) regions of Europe, two populations known to be exposed to AA.[75,76,97] The lower panel shows that exposure of Hupki MEF to AA[92] induces a similar mutational profile. Pooled data from multiple samples are shown for each data set. (**b**) Mutational spectra observed in lung adenocarcinomas (ADCA) of heavy smokers (upper panel) have features in common with spectra in Hupki MEF[92] (middle panel) and human mammary epithelial cells (HMEC, lower panel) exposed to B[a]P,[83] a tobacco carcinogen. (**c**) Spectra attributable to alkylation agents; upper panel: temozolomide treatment-related glioblastoma (TMZ GBM);[81] middle panel: lung carcinoma of mice treated with methylnitrosourea (MNU);[90] lower panel: Hupki MEF cells treated with methylnitrosoguanidine (MNNG).[92] The bar graphs to the right show strand bias ratios. Strand bias reflects transcription-coupled repair of chemically damaged DNA bases (NT, non-transcribed strand; T, transcribed strand). Asterisks indicate X^2 test *P*-values for strand bias significance (*$P < 10E-5$; **$P < 10E-20$; ***$P < 10E-320$; $P = 0$ for UTUC Taiwan, in top panel of (**a**)). Note the less pronounced transcriptional strand bias ratios associated with the effects of alkylating agents.

the cohort in which the signature was observed, namely physical inactivity, obesity and gastro-intestinal reflux.[82] This illustrates how a mutational signature *per se* reveals little about its author. Without hypotheses on the nature of the cancer risk factor from epidemiological and patient exposure data, and without experimental information on the mutagenic and chemical properties of carcinogens or endogenous mutational processes, a signature is undecipherable. A key demonstration of the convergence of multiple lines of information to establish cause was provided by the example cited above linking AA exposure to the unusual tumour A to T mutational signature. The only clues the signature could have provided entirely on its own were that: (a) the transversions were probably induced by an external agent, as this base substitution is a universally rare type of sequence change, and (b) the inducing agent probably generated bulky adducts on DNA bases, because these lead to transcription-coupled repair and thus a strand bias in the mutations that persist unrepaired. It was the confluence of experimental studies, epidemiology and patient exposure information that provided the necessary basis upon which a plausible cause of this signature was derived. Information on pro-mutagenic DNA adducts and other DNA lesions as well as the mutation spectra they generate in experimental systems have been essential factors in the assignment of signatures to risk factors.

The genome-wide impact of carcinogens and endogenous enzymes on DNA sequences can be efficiently captured in animal models, lower organisms and in cell-based *in vitro* assays.[76,83–90] For example, exposure of normal murine embryonic fibroblasts (MEF) to known human carcinogens and sequencing of clones following immortalization is a rapid procedure that can generate mutational signatures corresponding to signatures in human tumours from patients exposed to the same agents (Figure 2).[91,92] This simple experimental procedure[93,94] is also suited to investigation of signatures linked to endogenous mutational processes. As proof of principle, we compared mutational signatures in immortalized MEF clones derived from MEFs isolated from mice harbouring an activation-induced cytidine deaminase (AID) transgene against signatures in non-transgenic mice, and demonstrated the expected excess of AID signature mutations in the clones derived from AID-expressing mice.[92] AID, a hypermutator enzyme that promotes antibody diversity, causes off-target mutations in B-cell lymphomas and possibly other cancer types when inappropriately expressed.[95,96] Another source of experimentally induced genome-wide mutation patterns is potentially available from past *in vivo* toxicology projects. There is an untapped reservoir of archived tumour samples from animal carcinogen tests that can be mined using robust protocols for extraction and NGS of DNA derived from formalin-fixed, paraffin-embedded tumours already developed for human studies,[77,97,98] allowing immediate access to information from this valuable source.

CONCLUDING REMARKS

Mutational signature analysis clearly incriminates environmental factors in shaping tumour mutation spectra. Risk factor-linked diversity in mutational signatures provides a framework for establishing which and to what extent certain factors do indeed contribute to the mutation burden of a tumour. The diversity is likely to be even more evident when well-designed international comparisons of mutation profiles are conducted, for example, with studies that take advantage of unusually high rates of incidence of specific tumour types in relatively restricted geographic areas (for example, gallbladder cancer in Chile). New tools for de-convoluting inherent genetic components and external factors in migrant studies are now at hand.[31,99] Heterogeneity of mutation signatures in a single cancer type implies that a one-size-fits-all approach to early detection biomarkers and molecular therapies requires refinement.

The resounding discovery from NMF-based analysis of NGS data that specific endogenous enzymatic processes appear responsible for prominent mutational signatures in a broad variety of cancers sends out a research call to identify environmental or lifestyle factors that could act by proxy, stimulating the endogenous mutators. It is important to know whether and which avoidable factors regulate these endogenous mutational processes in the natural history of cancer. An interdisciplinary approach that harnesses epidemiology, experimental models, NGS and mathematical analysis of mutations should meet these challenges.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We acknowledge these funding sources: INCa-INSERM 2015 Plan Cancer grant to JZ; LBA is supported through the J. Robert Oppenheimer Fellowship at Los Alamos National Laboratory.

REFERENCES

1 Stratton MR. Exploring the genomes of cancer cells: progress and promise. *Science* 2011; **331**: 1553–1558.
2 Garraway LA, Lander ES. Lessons from the cancer genome. *Cell* 2013; **153**: 17–37.
3 Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* 2007; **446**: 153–158.
4 Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009; **458**: 719–724.
5 Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. Cancer genome landscapes. *Science* 2013; **339**: 1546–1558.
6 Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011; **144**: 646–674.
7 Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* 2012; **150**: 1121–1134.
8 Brash DE. UV signature mutations. *Photochem Photobiol* 2015; **91**: 15–26.
9 Brash DE, Rudolph JA, Simon JA, Lin A, McKenna GJ, Baden HP *et al.* A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proc Natl Acad Sci USA* 1991; **88**: 10124–10128.
10 Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* 2002; **21**: 7435–7451.
11 Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb Perspect Biol* 2010; **2**: a001008.
12 Oren M, Rotter V. Mutant p53 gain-of-function in cancer. *Cold Spring Harb Perspect Biol* 2010; **2**: a001107.
13 Vousden KH, Prives C. Blinded by the light: the growing complexity of p53. *Cell* 2009; **137**: 413–431.
14 Hollstein M, Hergenhahn M, Yang Q, Bartsch H, Wang ZQ, Hainaut P. New approaches to understanding p53 gene tumor mutation spectra. *Mutat Res* 1999; **431**: 199–209.
15 Denissenko MF, Pao A, Tang M, Pfeifer GP. Preferential formation of benzo[a] pyrene adducts at lung cancer mutational hotspots in P53. *Science* 1996; **274**: 430–432.
16 Miller JH. Carcinogens induce targeted mutations in *Escherichia coli*. *Cell* 1982; **31**: 5–7.
17 Wogan GN. Aflatoxins as risk factors for hepatocellular carcinoma in humans. *Cancer Res* 1992; **52**: 2114s–2118s.
18 Grollman AP, Shibutani S, Moriya M, Miller F, Wu L, Moll U *et al.* Aristolochic acid and the etiology of endemic (Balkan) nephropathy. *Proc Natl Acad Sci USA* 2007; **104**: 12129–12134.
19 Hollstein M, Moriya M, Grollman AP, Olivier M. Analysis of TP53 mutation spectra reveals the fingerprint of the potent environmental carcinogen, aristolochic acid. *Mutat Res* 2013; **753**: 41–49.
20 Hsu IC, Metcalf RA, Sun T, Welsh JA, Wang NJ, Harris CC. Mutational hotspot in the p53 gene in human hepatocellular carcinomas. *Nature* 1991; **350**: 427–428.
21 Bressac B, Kew M, Wands J, Ozturk M. Selective G to T mutations of p53 gene in hepatocellular carcinoma from southern Africa. *Nature* 1991; **350**: 429–431.
22 Montesano R, Hainaut P, Wild CP. Hepatocellular carcinoma: from gene to public health. *J Natl Cancer Inst* 1997; **89**: 1844–1851.

23 McCann J, Choi E, Yamasaki E, Ames BN. Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals. *Proc Natl Acad Sci USA* 1975; **72**: 5135–5139.
24 Baker SJ, Fearon ER, Nigro JM, Hamilton SR, Preisinger AC, Jessup JM *et al.* Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science* 1989; **244**: 217–221.
25 Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. *Science* 1991; **253**: 49–53.
26 Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* 2013; **3**: 246–259.
27 Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999; **401**: 788–791.
28 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**: 415–421.
29 Fischer A, Illingworth CJ, Campbell PJ, Mustonen V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol* 2013; **14**: R39.
30 Shiraishi Y, Tremmel G, Miyano S, Stephens M. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet* 2015; **11**: e1005657.
31 Schulze K, Imbeaud S, Letouze E, Alexandrov LB, Calderaro J, Rebouissou S *et al.* Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet* 2015; **47**: 505–511.
32 Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer* 1954; **8**: 1–12.
33 Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* 2015; **47**: 1402–1407.
34 Lindahl T. Instability and decay of the primary structure of DNA. *Nature* 1993; **362**: 709–715.
35 Wu S, Powers S, Zhu W, Hannun YA. Substantial contribution of extrinsic risk factors to cancer development. *Nature* 2016; **529**: 43–47.
36 Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* 2015; **347**: 78–81.
37 Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015; **136**: E359–E386.
38 Stewart BW, Wild CP (eds). *World cancer report 2014*. International Agency for Research on Cancer: Lyon, France; Geneva, Switzerland, 2014.
39 Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* 2014; **15**: 585–598.
40 Alexandrov LB, Nik-Zainal S, Siu HC, Leung SY, Stratton MR. A mutational signature in gastric cancer suggests therapeutic strategies. *Nat Commun* 2015; **6**: 8683.
41 Shlien A, Campbell BB, de Borja R, Alexandrov LB, Merico D, Wedge D *et al.* Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nat Genet* 2015; **47**: 257–262.
42 Smith HC, Bennett RP, Kizilyer A, McDougall WM, Prohaska KM. Functions and regulation of the APOBEC family of proteins. *Semin Cell Dev Biol* 2012; **23**: 258–268.
43 Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat Genet* 2013; **45**: 977–983.
44 Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet* 2015; **47**: 1067–1072.
45 Harris RS, Petersen-Mahrt SK, Neuberger MS. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol Cell* 2002; **10**: 1247–1253.
46 Kazanov MD, Roberts SA, Polak P, Stamatoyannopoulos J, Klimczak LJ, Gordenin DA *et al.* APOBEC-induced cancer mutations are uniquely enriched in early-replicating, gene-dense, and active chromatin regions. *Cell Rep* 2015; **13**: 1103–1109.
47 Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* 2013; **494**: 366–370.
48 Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* 2012; **149**: 979–993.
49 Gohler S, Da Silva Filho MI, Johansson R, Enquist-Olsson K, Henriksson R, Hemminki K *et al.* Impact of functional germline variants and a deletion polymorphism in APOBEC3A and APOBEC3B on breast cancer risk and survival in a Swedish study population. *J Cancer Res Clin Oncol* 2015; **142**: 273–276.
50 Nik-Zainal S, Wedge DC, Alexandrov LB, Petljak M, Butler AP, Bolli N *et al.* Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat Genet* 2014; **46**: 487–491.
51 Roberts SA, Gordenin DA. Clustered and genome-wide transient mutagenesis in human cancers: hypermutation without permanent mutators or loss of fitness. *Bioessays* 2014; **36**: 382–393.
52 Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* 2014; **24**: 52–60.
53 zur Hausen H. Papillomaviruses in the causation of human cancers - a brief historical account. *Virology* 2009; **384**: 260–265.
54 Rebhandl S, Huemer M, Greil R, Geisberger R. AID/APOBEC deaminases and cancer. *Oncoscience* 2015; **2**: 320–333.
55 Warren CJ, Xu T, Guo K, Griffin LM, Westrich JA, Lee D *et al.* APOBEC3A functions as a restriction factor of human papillomavirus. *J Virol* 2015; **89**: 688–702.
56 Henderson S, Chakravarthy A, Su X, Boshoff C, Fenton TR. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep* 2014; **7**: 1833–1841.
57 Hussain SP, Hofseth LJ, Harris CC. Radical causes of cancer. *Nat Rev Cancer* 2003; **3**: 276–285.
58 Marnett LJ, Plastaras JP. Endogenous DNA damage and mutation. *Trends Genet* 2001; **17**: 214–221.
59 Fedeles BI, Freudenthal BD, Yau E, Singh V, Chang SC, Li D *et al.* Intrinsic mutagenic properties of 5-chlorocytosine: A mechanistic connection between chronic inflammation and cancer. *Proc Natl Acad Sci USA* 2015; **112**: E4571–E4580.
60 Brennan P, Wild CP. Genomics of Cancer and a New Era for Cancer Prevention. *PLoS Genet* 2015; **11**: e1005522.
61 de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* 2014; **346**: 251–256.
62 Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* 2014; **346**: 256–259.
63 Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* 2015; **22**: 105–113.
64 Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 2012; **366**: 883–892.
65 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**: 880–886.
66 Ross-Innes CS, Becq J, Warren A, Cheetham RK, Northen H, O'Donovan M *et al.* Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nat Genet* 2015; **47**: 1038–1046.
67 Weaver JM, Ross-Innes CS, Shannon N, Lynch AG, Forshew T, Barbera M *et al.* Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat Genet* 2014; **46**: 837–843.
68 Zhang L, Zhou Y, Cheng C, Cui H, Cheng L, Kong P *et al.* Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. *Am J Hum Genet* 2015; **96**: 597–611.
69 Krishnan VG, Ebert PJ, Ting JC, Lim E, Wong SS, Teo AS *et al.* Whole-genome sequencing of asian lung cancers: second-hand smoke unlikely to be responsible for higher incidence of lung cancer among Asian never-smokers. *Cancer Res* 2014; **74**: 6071–6081.
70 Pfeifer GP. How the environment shapes cancer genomes. *Curr Opin Oncol* 2015; **27**: 71–77.
71 Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* 2011; **333**: 1157–1160.
72 Cancer Genome Atlas Research N, Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 2014; **507**: 315–322.
73 Pickering CR, Zhang J, Neskey DM, Zhao M, Jasser SA, Wang J *et al.* Squamous cell carcinoma of the oral tongue in young non-smokers is genomically similar to tumors in older smokers. *Clin Cancer Res* 2014; **20**: 3842–3848.
74 Grollman AP. Aristolochic acid nephropathy: harbinger of a global iatrogenic disease. *Environ Mol Mutagen* 2013; **54**: 1–7.
75 Hoang ML, Chen CH, Sidorenko VS, He J, Dickman KG, Yun BH *et al.* Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci Transl Med* 2013; **5**: 197ra102.
76 Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P *et al.* Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci Transl Med* 2013; **5**: 197ra101.
77 Jelakovic B, Castells X, Tomic K, Ardin M, Karanovic S, Zavadil J. Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid. *Int J Cancer* 2015; **136**: 2967–2972.

78 Scelo G, Riazalhosseini Y, Greger L, Letourneau L, Gonzalez-Porta M, Wozniak MB *et al.* Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun* 2014; **5**: 5135.
79 Poon SL, Huang MN, Choo Y, McPherson JR, Yu W, Heng HL *et al.* Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med* 2015; **7**: 38.
80 Zou S, Li J, Zhou H, Frech C, Jiang X, Chu JS *et al.* Mutational landscape of intrahepatic cholangiocarcinoma. *Nat Commun* 2014; **5**: 5696.
81 Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY *et al.* Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* 2014; **343**: 189–193.
82 Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, Stewart C *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet* 2013; **45**: 478–486.
83 Severson PL, Vrba L, Stampfer MR, Futscher BW. Exome-wide mutation profile in benzo[a]pyrene-derived post-stasis and immortal human mammary epithelial cells. *Mutat Res Genet Toxicol Environ Mutagen* 2014; **775-776**: 48–54.
84 Flibotte S, Edgley ML, Chaudhry I, Taylor J, Neil SE, Rogula A *et al.* Whole-genome profiling of mutagenesis in *Caenorhabditis elegans*. *Genetics* 2010; **185**: 431–441.
85 Maslov AY, Quispe-Tintaya W, Gorbacheva T, White RR, Vijg J. High-throughput sequencing in mutation detection: a new generation of genotoxicity tests? *Mutat Res* 2015; **776**: 136–143.
86 Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J *et al.* C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res* 2014; **24**: 1624–1636.
87 Nassar D, Latil M, Boeckx B, Lambrechts D, Blanpain C. Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nat Med* 2015; **21**: 946–954.
88 Segovia R, Tam AS, Stirling PC. Dissecting genetic and environmental mutation signatures with model organisms. *Trends Genet* 2015; **31**: 465–474.
89 Tam AS, Chu JS, Rose AM. Genome-wide mutational signature of the chemotherapeutic agent mitomycin C in *Caenorhabditis elegans*. *G3 (Bethesda)* 2015; **6**: 133–140.
90 Westcott PM, Halliwill KD, To MD, Rashid M, Rust AG, Keane TM *et al.* The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. *Nature* 2015; **517**: 489–492.
91 Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM *et al.* The genome as a record of environmental exposure. *Mutagenesis* 2015; **30**: 763–770.
92 Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallee MP *et al.* Modelling mutational landscapes of human cancers in vitro. *Sci Rep* 2014; **4**: 4482.
93 Liu Z, Belharazem D, Muehlbauer KR, Nedelko T, Knyazev Y, Hollstein M. Mutagenesis of human p53 tumor suppressor gene sequences in embryonic fibroblasts of genetically-engineered mice. *Genet Eng (NY)* 2007; **28**: 45–54.
94 Liu Z, Hergenhahn M, Schmeiser HH, Wogan GN, Hong A, Hollstein M. `Human tumor p53 mutations are selected for in mouse embryonic fibroblasts harboring a humanized p53 gene. *Proc Natl Acad Sci USA* 2004; **101**: 2963–2968.
95 Gu X, Shivarov V, Strout MP. The role of activation-induced cytidine deaminase in lymphomagenesis. *Curr Opin Hematol* 2012; **19**: 292–298.
96 Pettersen HS, Galashevskaya A, Doseth B, Sousa MM, Sarno A, Visnes T *et al.* AID expression in B-cell lymphomas causes accumulation of genomic uracil and a distinct AID mutational signature. *DNA Repair (Amst)* 2015; **25**: 60–71.
97 Castells X, Karanovic S, Ardin M, Tomic K, Xylinas E, Durand G *et al.* Low-coverage exome sequencing screen in formalin-fixed paraffin-embedded tumors reveals evidence of exposure to carcinogenic aristolochic acid. *Cancer Epidemiol Biomarkers Prev* 2015; **24**: 1873–1881.
98 Hedegaard J, Thorsen K, Lund MK, Hein AM, Hamilton-Dutoit SJ, Vang S *et al.* Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. *PLoS One* 2014; **9**: e98187.
99 Totoki Y, Tatsuno K, Covington KR, Ueda H, Creighton CJ, Kato M *et al.* Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat Genet* 2014; **46**: 1267–1273.

Appendix III: TP53 variations in human cancers: new lessons from the IARC TP53 Database and genomics data

Liacine Bouaoun, Dmitriy Sonkin, **Maude Ardin**, Monica Hollstein, Graham Byrnes, Jiri Zavadil and Magali Olivier

Human mutation, 2016

DATABASES

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data

Liacine Bouaoun,[1] Dmitriy Sonkin,[2] Maude Ardin,[3] Monica Hollstein,[3,4] Graham Byrnes,[1] Jiri Zavadil,[3] and Magali Olivier[3]*

[1]*Group of Biostatistics, International Agency for Research on Cancer, Lyon Cedex 08 69372, France;* [2]*Division of Cancer Treatment and Diagnosis, Biometric Research Program, National Cancer Institute, Rockville 9609, Maryland;* [3]*Group of Molecular Mechanisms and Biomarkers, International Agency for Research on Cancer, Lyon Cedex 08 69372, France;* [4]*Faculty of Medicine and Health, University of Leeds, Leeds LS2 9JT, UK*

Communicated by Stephen J. Chanock

Received 2 February 2016; accepted revised manuscript 18 June 2016.
Published online 22 June 2016 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.23035

ABSTRACT: *TP53* gene mutations are one of the most frequent somatic events in cancer. The IARC TP53 Database (http://p53.iarc.fr) is a popular resource that compiles occurrence and phenotype data on *TP53* germline and somatic variations linked to human cancer. The deluge of data coming from cancer genomic studies generates new data on *TP53* variations and attracts a growing number of database users for the interpretation of *TP53* variants. Here, we present the current contents and functionalities of the IARC TP53 Database and perform a systematic analysis of *TP53* somatic mutation data extracted from this database and from genomic data repositories. This analysis showed that IARC has more *TP53* somatic mutation data than genomic repositories (29,000 vs. 4,000). However, the more complete screening achieved by genomic studies highlighted some overlooked facts about *TP53* mutations, such as the presence of a significant number of mutations occurring outside the DNA-binding domain in specific cancer types. We also provide an update on *TP53* inherited variants including the ones that should be considered as neutral frequent variations. We thus provide an update of current knowledge on *TP53* variations in human cancer as well as inform users on the efficient use of the IARC TP53 Database.**
Hum Mutat 00:1–12, 2016. © 2016 Wiley Periodicals, Inc.

KEY WORDS: annotations; cancer; germline variations; locus-specific database; mutation hotspots; somatic mutations; *TP53*

Introduction

For over 20 years, the tumor suppressor gene *TP53* (MIM# 191117) has been recognized as the most frequently somatically altered gene in human cancer, a fact confirmed by recent genome-wide analyses [Kandoth et al., 2013]. This master role in cancer has driven a large and diverse number of studies on the characteristics and role of *TP53* gene alterations in all types of human cancers as well as on the biology of the p53 protein and its mutated forms. The functional impact of all p53 mutant proteins has been extensively investigated in various types of experimental assays reflecting the versatile biological activities of p53 [Petitjean et al., 2007]. Hundreds of clinical studies on the prognostic or predictive role of *TP53* somatic mutations have also been published [Olivier et al., 2005]. *TP53* mutations may also be inherited causing the Li–Fraumeni syndrome (LFS) that predisposes to a wide spectrum of early-onset cancers [Malkin et al., 1990]. How different types of germline mutation influence the spectrum of tumors observed in LFS is an ongoing research topic. The data generated on all these aspects of *TP53* and published in an unstructured way in peer-reviewed articles have been curated and integrated in the IARC TP53 Database (http://p53.iarc.fr/), providing the scientific community with a unique resource to retrieve and mine this information [Petitjean et al., 2007].

The knowledge accumulated so far on *TP53* mutations is based on the Sanger sequencing technology and is biased by the fact that many studies did not sequence the entire gene but focused on hotspot regions. Moreover, many studies have used prescreening strategies, such as immunohistochemistry, thus may have missed mutations resulting in unchanged or loss of protein expression. Since the advent of next-generation sequencing (NGS) technologies, thousands of tumors have been sequenced genome- or exome-wide, generating new data on the prevalence and type of *TP53* mutations in various cancer types. These data, obtained using a technology that is more sensitive and less biased, constitute a new source of information to revisit our knowledge on *TP53* mutations in human cancers. At the same time, the information contained in the IARC TP53 Database can be of great help to interpret variants found in NGS studies, as the data coverage of this database remains broader than the one currently available in NGS data repositories.

In the last 5 years, several developments have been made to the IARC TP53 Database, adding new types of data and new annotations and upgrading the Web interface to expand its search and mining functionalities. Here, we present the current contents and functionalities of the database and how they can be used to interpret *TP53* variants. We also present a systematic analysis of *TP53* mutations accumulated in The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) data repositories and compare these new data with those reported in the IARC database to revisit our knowledge on the landscape of *TP53* somatic mutations in human cancers and cell-lines.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Magali Olivier, Group of Molecular Mechanisms and Biomarkers, International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon Cedex 08, France. E-mail: olivierm@iarc.fr

Contract Grant Sponsors: International Agency for Research on Cancer (IARC).

Methods

Data Sources

IARC data correspond to datasets extracted from the R18 release of the IARC TP53 Database (http://p53.iarc.fr/). For the dataset of somatic mutations, we excluded NGS studies, cell-lines, body fluids, xenographs, non-primary tissues, studies identified as poor-quality and considered only studies that used DNA as starting material and that screened at least exons 5 to 8.

NGS data were extracted from publicly available data from TCGA and ICGC data portals. TCGA data were downloaded on 26 March 2015 via a https protocol: https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/). Data available in ICGC data repository were downloaded via https://dcc.icgc.org/releases using the release R19 (June 2015, accessed on 10 August 2015) and excluding TCGA data to avoid duplicates. Only samples corresponding to primary tumors were extracted by selecting the following features in the annotations: "primary solid tumor," "primary blood derived cancer," "peripheral blood" for leukemias", "primary tumor—other" for both myeloid and pancreatic cancers, "primary tumor—lymph node" for lymphomas, except for melanoma were both primary tumors and metastasis were included.

The resulting NGS dataset included somatic mutation data for 32 cancer types corresponding to the major cancers (see Supp. Table S1 for the detailed list).

Cell-lines data were extracted from the Cancer Cell Line Encyclopedia (CCLE) resource. Affymetrix U133Plus2 mRNA expression, Affymetrix SNP 6.0 data, exome sequencing data [Hodges et al., 2007], and OncoMap data [MacConaill et al. 2009] have been obtained from the CCLE Website (http://www.broadinstitute.org/ccle/home). *TP53* mutation calls have been obtained from two files available from the CCLE Website: CCLE_hybrid_capture1650_hg19_NoCommonSNPs_NoNeutralVariants_CDS_2012.5.07.maf (22-May-2012) and 1650_HC_plus_RD_muts.maf.annotated (24-Nov-2014). Genomic characterization section in Supplementary Methods of Barretina et al. (2012) provides detailed description of sequencing data generation and variant calling pipeline. CCLE RNA-seq BAM files from CGHub and/or Hybrid Capture (HC) sequencing BAM files available from the Broad Institute CCLE Website were reviewed using the Integrative Genomics Viewer [Robinson et al., 2011] for 93 cell-lines scored as *TP53* wild-type by CCLE, but with documented mutation in other studies (COSMIC or IARC sources). Copy Number (CN) ratio is the ratio of signal intensity in a tumor sample versus normal reference samples normalized to total DNA quantity; thus a CN ratio of 1 corresponds to a diploid locus. CN ratio <= 0.6 indicates "allelic loss." CN ratio <= 0.25 indicates "bi-allelic loss."

All mRNA expression values are MAS5 normalized, with a 2% trimmed mean of 150 [Hubbell et al., 2002]. *TP53* Affymetrix (201746_at) mRNA MAS5-150 normalized expression values <=32 are considered to be indicative of "mRNA loss." Mutations were classified into functional types according to IARC annotations.

Reference Sequences

All *TP53* variants are described on the following reference sequences: hg19 genome, NM_000546.5 NCBI transcript, and P04637 UniProtKB protein.

Statistics

Codon distributions were calculated for all point mutations in exonic regions and for indel separately. Data management and statistical analyses were performed using the *R* and Bioconductor softwares [Gentleman et al. 2004; R Development Core Team, 2014]. Figures were drawn using the ggplot2 package in R.

IARC TP53 Database Structure and Contents

Datasets

The IARC TP53 Database integrates different datasets structured around a central module that describes *TP53* variations and their sequence context (Fig. 1). These datasets correspond to six main types of data: somatic variations, germline variations, variants found in cell-lines, functional assessment of variants, variants found in mutagenesis assays, and mouse-models of variants. Data on somatic variations include a compilation of somatic mutations found in tumor tissues, of the prevalence of somatic mutations by cancer type, and of studies investigating the prognostic and predictive value of somatic mutations. Detailed information on tumor histopathology, patient clinical, life-style, and demographic features are provided when available from original publications. Data on germline variations are a compilation of published and unpublished cases of *TP53* inherited variants, with detailed pedigree data. The prevalence of germline variations in various case series is also available in a recently implemented dataset. Germline variations frequent in healthy populations are also compiled. Data on the predicted and experimentally assessed functional impacts of missense substitutions are available in two datasets, one compiling data from diverse studies and one corresponding to a systematic screen of all missense mutants from one study using yeast assays. Data on *TP53* gene status in over 2,500 human cell-lines are annotated with Web links to the American Type Culture Collection (ATCC) cell-line repository and the COSMIC cell-line database. Data on mouse models carrying a mutant *Trp53* gene are listed with links to source databases. Finally, data on experimentally induced mutations are compiled in a new dataset. This dataset includes studies using the human *TP53* gene in eukaryotic cell assays designed to test the mutagenicity of carcinogenic compounds. The current dataset has data on over 900 experimentally induced mutations obtained after exposure with 18 different carcinogens in different types of assays, including the Hupki MEF assay [Luo et al., 2001], and a yeast reporter system [Yu et al., 2002]. These two assays involve a step of clonal expansion before mutation assessment, thus mimic the biological selection of functional mutations in processes relevant to cancer development.

All these datasets can be fully downloaded or queried through our Web search interface (see detailed list of datasets and statistics about their contents in Supp. Table S2).

Annotations

A wide range of annotations are provided for each dataset described above. They include: the description of gene variations and their functional impact; tumor histopathological classification; patient clinical characteristics and demographics; family structure and data sources. Gene variations are mapped to the genome build hg19 and to the canonical transcript and protein sequences (LRG_321_t1, NM_000546, Uniprot_P04637). Predicted nucleotide and amino-acid substitution rates were derived from the comparison of

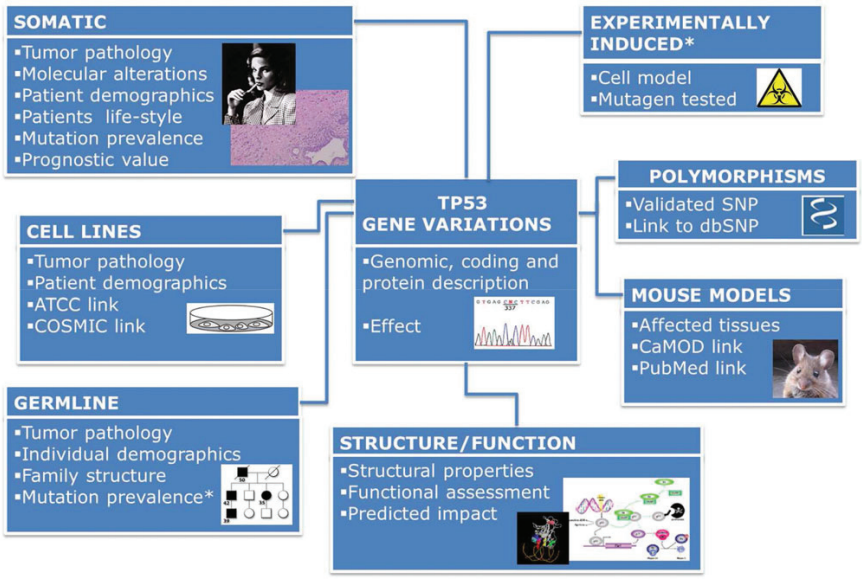


Figure 1. Overview of the structure and contents of the IARC TP53 Database.

human-mouse aligned sequence of chromosomes 21 and 10 and taking into account the trinucleotide sequence context [Lunter and Hein 2004; Mathe et al., 2006]. These rates reflect background mutation rates (due to replication errors and other endogenous processes) that can be used to weight mutation frequencies. Annotations on the functional consequences of mutations include experimentally assessed activities (such as promoter-specific transactivation, dominant-negative effect, gain of function, splicing, and so on) as well as predictions based on tools such as SIFT [Kumar et al., 2009], AGVGD [Mathe et al., 2006], Polyphen [Adzhubei et al., 2010], or HumanSpliceFinder [Desmet et al., 2009].

Annotations are based on international standards (ICD-O for tumor classifications, HUGO Mutation Database Initiative/Human Genome Variation Society [HGVS] nomenclature for mutation description) or internally developed systems (for describing functional impact in experimental assays). Detailed descriptions of all datasets and annotations are documented online (http://p53.iarc.fr/Manual.aspx).

Website Features

The Website has been redesigned to comply with recent Web standards and accommodate new search features and datasets. Advanced search options are available for all datasets allowing searches by various criteria, including mutation features, sample and patient characteristics, screening method, and publication citation. Selected subsets of data are summarized in graphical or tabular formats and can be downloaded. Entire datasets can also be fully downloaded. A number of resources such as protocols, guidelines, and recommendations on reporting mutation data, links to relevant publications and related Websites are also provided. Finally, a slide-show on *TP53* gene biology and the relevance of *TP53* mutations in cancer can be downloaded. Below we describe one example of a search option and its display outputs.

Searching for Occurrence and Functional Impacts of Specific Variants

The "gene variations" search option allows searching for a single variant or a list of variants described at the genomic, coding or protein level using the HGVS recommended nomenclature (http://www.hgvs.org/varnomen/). Groups of variants may also be searched for based on various criteria related to their transcript and protein positions, mutation types and functional impacts (not shown). In this example, we selected one variant described in the protein sequence (Fig. 2A). On the results page (Fig. 2B), the selected variant is displayed in a table that provides summary data on its location in the coding sequence, predicted and experimentally assessed functional classifications and frequency counts in the somatic, germline, and cell-line datasets. From this table, several options are available. First, mutation distributions by localization, type or functional impacts may be graphically displayed (Fig. 2C). Second, further annotations on the characteristics and functional impacts of a specific variant can be obtained (Supp. Fig. S1). This option requires the selection of a single variant and will return summary data from all datasets. This includes detailed sequence context of the variant, predicted effects on splicing and protein structure/function, experimental assessments of transactivation activity in yeast and human cell assays, mouse models carrying the variant, and exposure under which the variant was observed in experimental mutagenesis assays. Finally, all variants in the table or a selected subset may be used to query the "cell-lines," "somatic," and "germline" datasets. The cell-line query will return a table showing the characteristics of the cell-lines carrying the selected variants, including, when available, links to the ATCC catalog and to the COSMIC database (allowing to view other genomic alterations reported in the selected cell-line) (Fig. 2D). The tumor spectrum query will return two bar graphs showing the tumor distributions of the selected variants in the somatic and germline datasets (Fig. 2E).

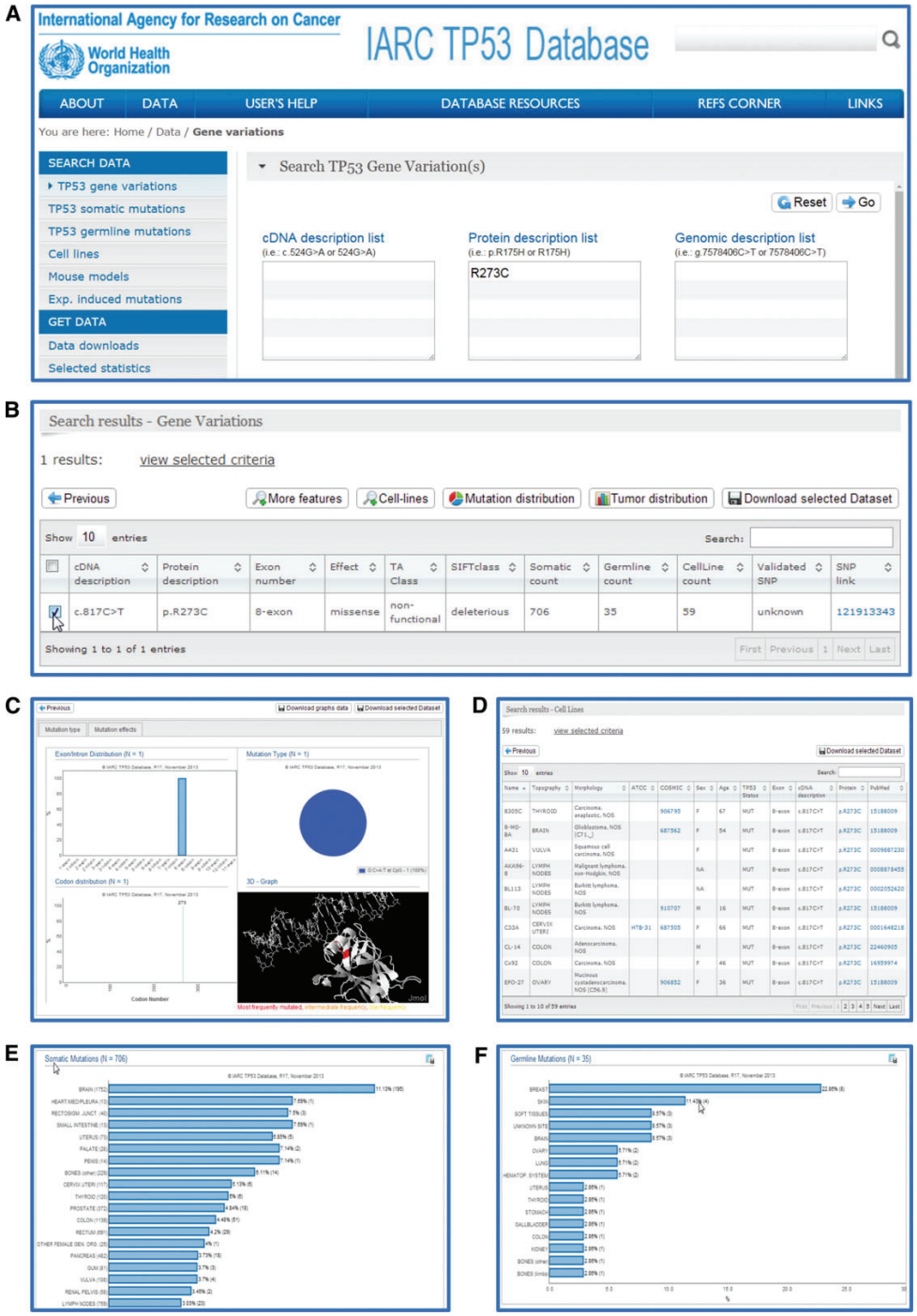


Figure 2. Example of screenshots of database query outputs using the Web search interface.

Thus, with this search option it is possible to retrieve all available structural and phenotypic data derived from bioinformatics predictions, experimental assays and human observations, for a specific variant or a group of variants. This unique knowledgebase is very useful for interpreting the functional relevance of variants.

TP53 Germline Variations: Identifying Disease-Causing Versus Neutral Variations

The dataset of germline variations provides a compilation of all *TP53* germline variants identified in cancer families and reported

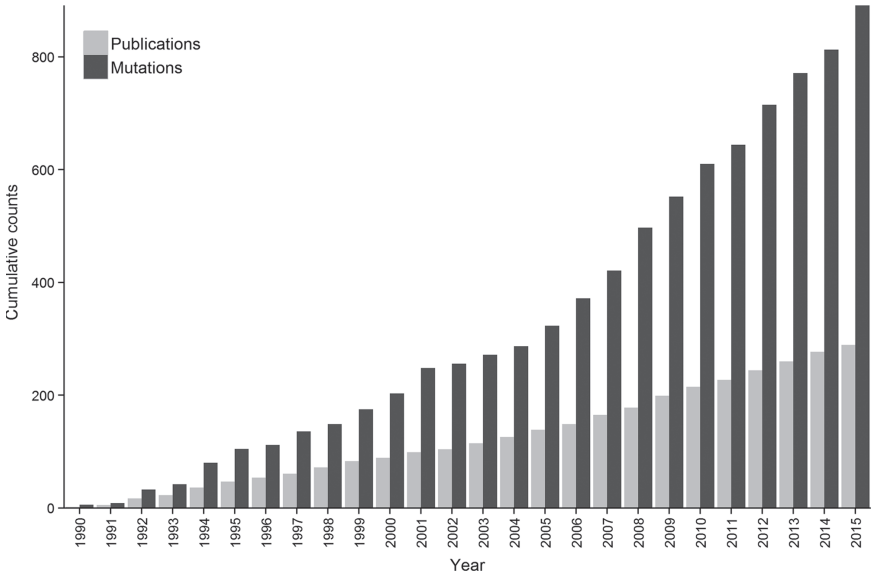


Figure 3. Time trend in the reporting of germline variations.

in the scientific literature since 1990. Variant annotations include the description of the pedigree structure, the characteristics of each individual in the pedigree (such as age, gender, variant carrier status, and relationship with proband) and the characteristics of all tumors identified in these individuals. The available data and annotations allow advanced genotype–phenotype relationship analyses. We and others have identified such genotype–phenotype relationships using limited datasets available at that time [Birch et al., 2001; Olivier et al., 2003]. Since then, the dataset has almost tripled (767 pedigrees in 2013 compared with 265 in 2003). In fact, the number of variants reported every year shows a marked increase after 2005 (Fig. 3). Although it coincides with the advent of cheaper sequencing technologies, all variants included in R18 have been screened by Sanger sequencing.

With the volume of data currently accumulating from NGS studies and clinical genetics laboratories, more *TP53* germline variants are expected to be found and interpreting their phenotype will be crucial to adapt clinical surveillance and management in carriers. In the current R18 release, there are 891 families carrying a *TP53* mutation, representing 210 unaffected carriers and 2,165 individuals affected with over 3,000 tumors. Missense variants are the most frequent (73%) as previously described [Olivier et al., 2003]. The distribution of single base substitution variants in these families and the spectrum of tumors observed in carriers are similar to what was derived from earlier datasets except for a higher prevalence of mutations at codon 337 and of adrenal cortical carcinoma (ADC) tumors (Fig. 4A and B). This is due to a founder variant, p.R337H (c.1010G>A), discovered in Brazil as linked to ADC [Ribeiro et al., 2001]. This variant is highly prevalent in Brazil and has been analyzed in many studies that have been included in the IARC TP53 Database. These studies have shown that the type of tumors arising in carrier is not restricted to ADC, as other LFS tumor types have been observed in carriers of this germline mutation [Achatz et al., 2007]. Other frequently mutated codons are classical hotspots common to both somatic and germline variants. The spectrum of tumors in *TP53* carriers (Fig. 4B) confirms previous observations such as the higher proportion of male among individuals affected with brain and stomach cancers, and females among those affected with ADC [Olivier et al., 2003]. The age distribution of the most frequent cancers is also similar to previous analyses. ADC, brain tumors, and soft tissues sarcomas affect mainly children, bone sarcomas are more prevalent in adolescents, and the incidence peak of breast cancer in women is under 30 years (Supp. Fig. S2). Brain tumors and soft tissues sarcomas show a biphasic distribution (Supp. Fig. S2), the later onset cases corresponding to different cancer subtypes compared with the early onset cases [Olivier et al., 2003]. A detailed genotype–phenotype analysis is beyond the scope of this article but the data accumulated is a unique resource to investigate the phenotype of *TP53* germline mutations and to interpret variants found in NGS-based studies.

Among inherited variants, approximately one hundred have been described as frequent (above 1%) in unaffected human populations (included in the "POLYMORPHISM" dataset). These variants are considered as neutral, although some of them have been shown to be associated with cancer susceptibility [Whibley et al., 2009, Sagne et al., 2013]. Of all *TP53* variants in dbSNP147, 21 are most probably deleterious variants (predicted deleterious by SIFT, exhibit loss of function in yeast assays and are frequently found as somatic mutations) but are not recorded as pathogenic based on ClinVar annotations (Supp. Table S3A). These variants have reported MAFs below 0.0001 or have never been reported in control populations. In fact, among the 100 most frequent somatic variants in the IARC database (all predicted as deleterious or causing loss of function in experimental assays), 65 are present in dbSNP147 (Supp. Table S3B) and only 34 are annotated as pathogenic. Using dbSNP to filter out germline neutral variants, even taking into account annotations on pathogenicity, should thus be done with caution as some true somatic and deleterious variants could be filtered out. We thus recommend considering only variants annotated as validated SNP in the IARC TP53 Database as likely *TP53* germline calls that should be filtered out.

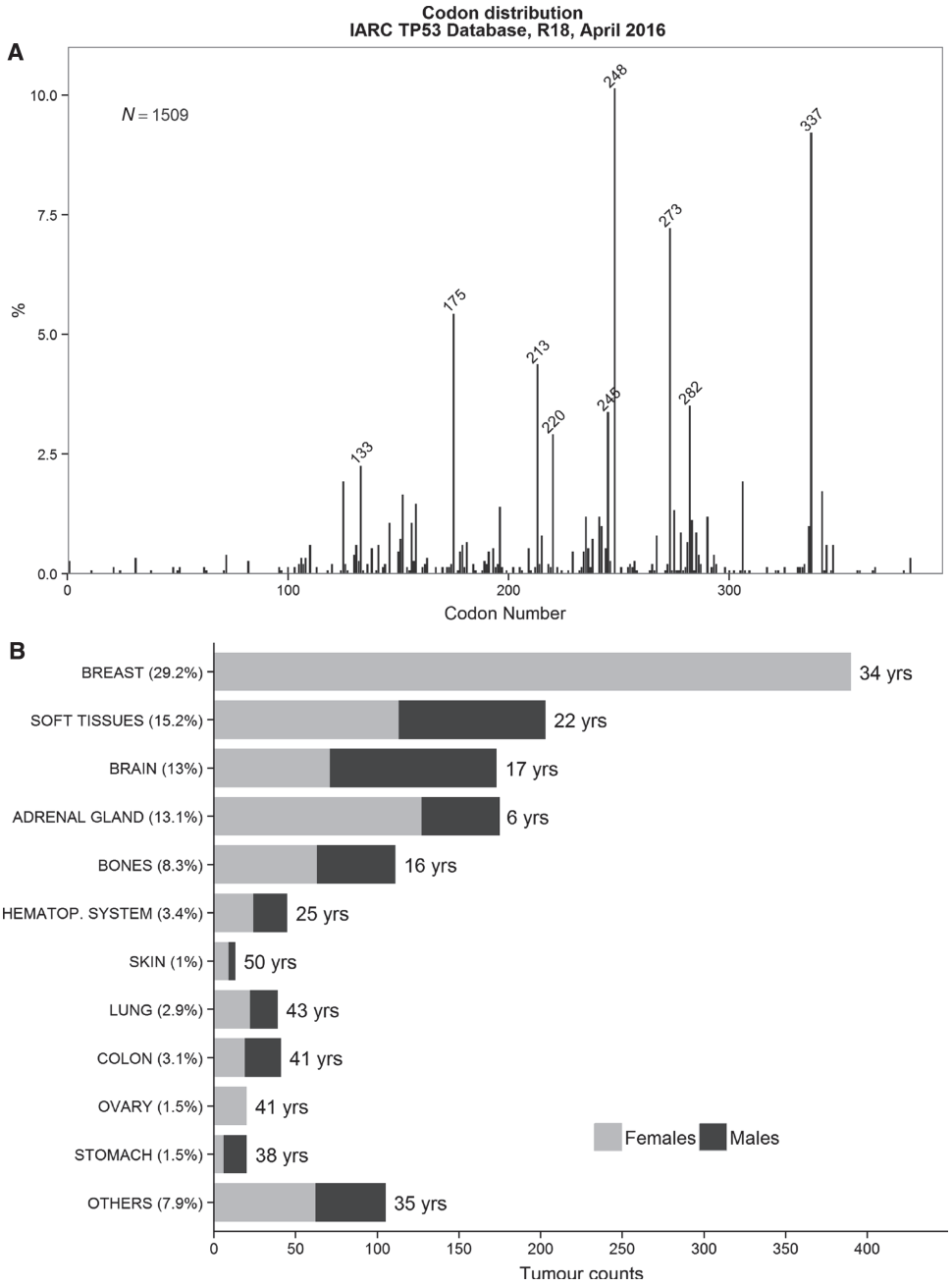


Figure 4. Mutation and tumor patterns in confirmed *TP53* germline mutation carriers. **A**: Codon distribution of *TP53* variations in confirmed mutation carriers (tumor counts). **B**: Distribution of tumor types and gender by cancer type in confirmed carriers.

TP53 Somatic Mutations: New Insights from NGS Data

TP53 is a Master Driver Gene in Most Cancers

Curated somatic mutation data from TCGA and ICGC exome sequencing projects (referred thereafter as NGS data) were analyzed to extract the prevalence of *TP53* mutations in 32 cancer types analyzed by NGS. Prevalence data for the same cancer types were then extracted from the IARC TP53 Database (referred to as IARC data) including only studies based on Sanger sequencing technology (R18 release, NGS studies excluded, see *Methods*). The overall mutation prevalence for all cancers combined was 29% in IARC data and 35% in NGS data. The range of cancer-specific prevalences was very similar between IARC and NGS data although prevalences in NGS

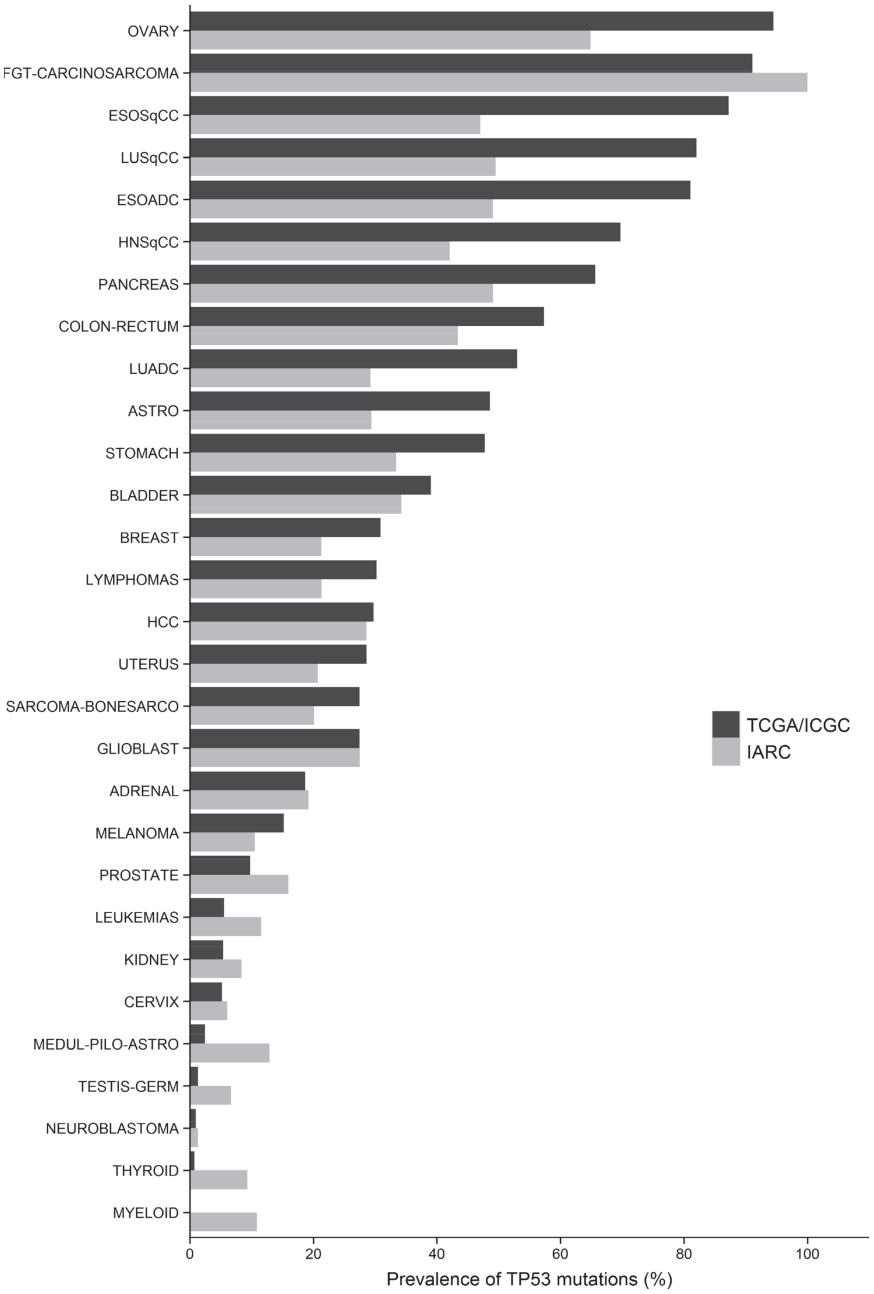


Figure 5. Prevalence of *TP53* somatic mutation in data extracted from IARC and NGS data for the 29 cancer types covered by both datasets.

studies tended to be higher than those derived from IARC (Fig. 5 and Supp. Table S4). Cancers with the highest mutation prevalence included ovarian cystadenocarcinomas, adeno, and squamous carcinomas of the esophagus and lung, head and neck squamous carcinomas, carcinomas of the pancreas, colorectal carcinomas, and female genital tract carcinosarcomas. This later cancer is interesting as only one small study, reporting nine out of nine mutated samples, was compiled in the IARC database. One NGS study reports 51 out of 56 mutated cases, thus confirming that this rare aggressive cancer is among the top *TP53* mutated cancers. The fact that higher prevalences are found in NGS studies may be due to several factors. First, the entire gene is screened in NGS studies, while in the IARC database half of the studies have only screened exons 5 to 8 where most mutations are located. Mutations outside this region may thus have been missed. Second, NGS is more sensitive than Sanger sequencing and may thus detect more mutations with low allelic fractions. It is interesting to note that cancer types that show the largest discrepancy are cancers associated with smoking

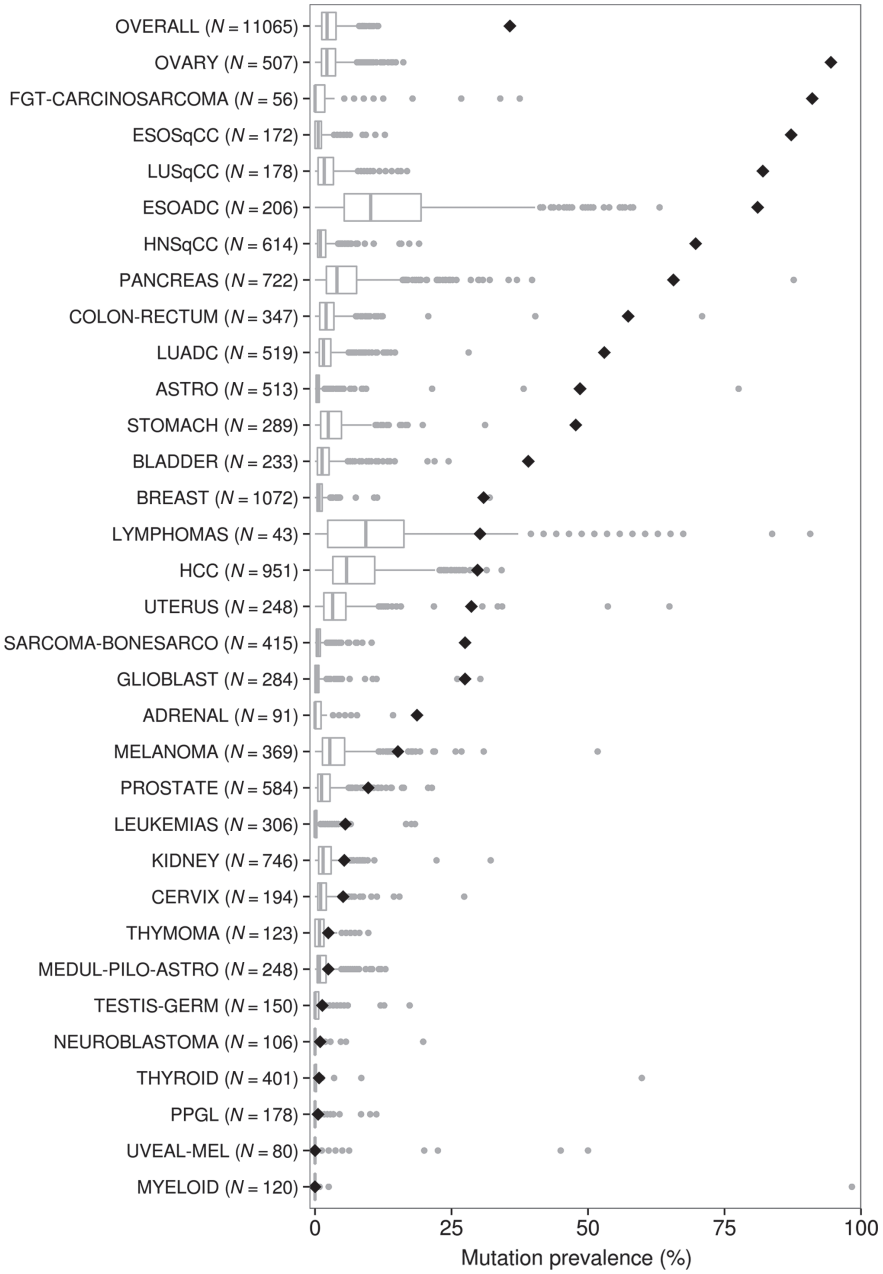


Figure 6. Prevalence of cancer gene somatic mutation for all genes reported in the cancer gene census. The overall (pooled samples) and cancer-specific prevalences from the NGS dataset are shown. Each dot corresponds to a cancer gene, with the diamond corresponding to the *TP53* gene. Numbers in parenthesis correspond to the total number of samples analyzed.

(carcinomas of the esophagus, lung, and head and neck). Because smoking is a strong mutagen, it would be interesting to investigate whether smoking-related cancers accumulate higher number of low allelic fraction mutations (potentially neutral passenger mutations) that would have been missed by Sanger sequencing in the IARC dataset. For cancers known to be rarely mutated, NGS studies confirmed the low prevalence (<15%) of *TP53* mutations in these cancer types. In myeloid, thyroid, testis, and pilo-astrocytic groups of tumors, NGS studies showed prevalences close to zero, whereas IARC data showed prevalences around 10%. This difference may be due to differences in tumor subtypes between the two datasets or to variability due to small numbers.

Using NGS data, we also calculated mutation frequencies for the 572 cancer genes listed in the cancer census (http://cancer.sanger.ac.uk/census/). Among these genes, *TP53* is the most frequently mutated overall (Fig. 6, top box plot), and it is

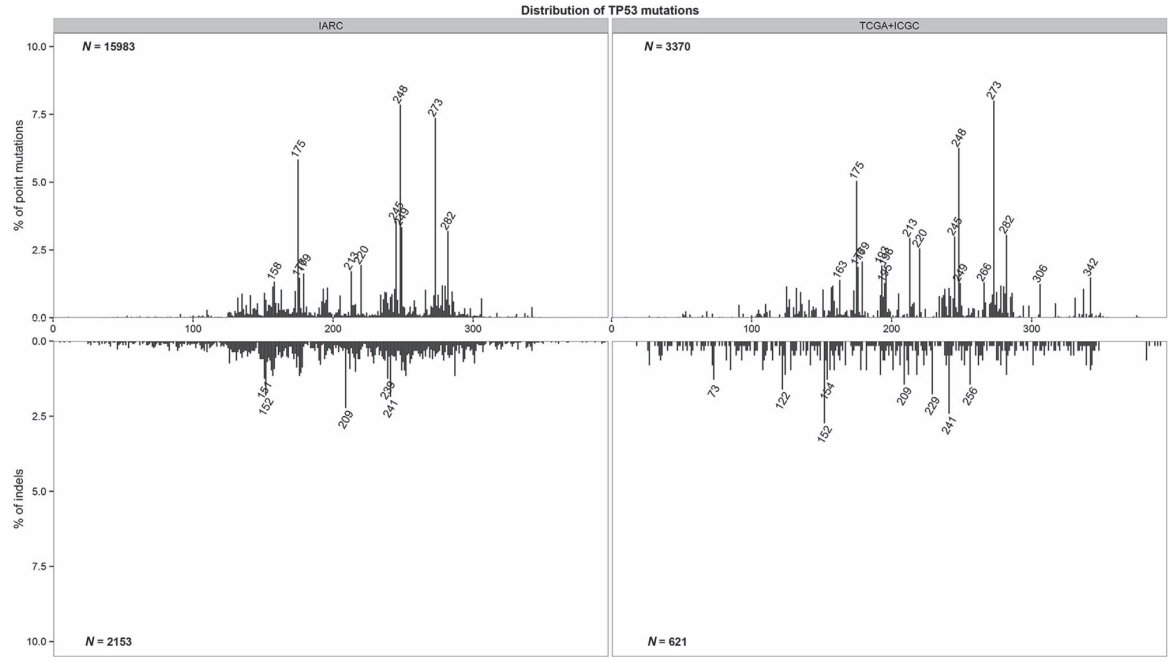


Figure 7. Codon distributions of *TP53* point mutations (top) and indels (bottom) in IARC and NGS datasets.

among the top three mutated genes in a majority (17 out of 32) of cancers (Fig. 6).

These data confirm that *TP53* is the most frequently mutated gene in human cancer overall, and highlight the fact that cancer-specific mutation prevalences are higher than previously determined by Sanger sequencing studies.

TP53 Hotspot Mutations

To further evaluate similarities and differences between IARC and NGS data at mutation position resolution, we plotted the overall codon distribution of point mutations and indels for IARC and NGS datasets. As shown in Figure 7, the distributions are very similar in the central portion of the protein corresponding to exons 5 to 8, the DNA-binding domain preferentially screened in studies compiled in the IARC TP53 Database. The well-known hotspots for single base substitutions include codons 175, 220, 245, 248, 273, and 282, where missense mutations affect either the 3D structure of the p53 protein or its contact with DNA and result in loss of function. Other hotspots, notably codons 196 and 213 are sites where nonsense mutations accumulate. Interestingly, hotspots for indels are also evident in both datasets, such as codons 152, 209, and 241. The codon 152 hotspot is located within a GC-rich region including a stretch of five cytosines, a sequence context known to favor strand slippage during replication. Mutations at this hotspot are single nucleotide insertions and deletions. At codon 209, a deletion of two nucleotides (TC) is observed in both datasets. For codon 241, single deletions of C or T are observed within a TTCCT sequence context. It remains to be determined whether sequence contexts at codons 209 and 241 are more prone to replication/repair errors that may explain the high frequency of indels at these two hotspots.

The unbiased screening with NGS shows that a significant number of mutations occur outside exons 5 to 8, in particular in the oligomerization domain located in the C-terminus of the protein (codons 323–356) where approximately 6% of point mutations and 7% of indels are found (Supp. Table S5). Three hotspots in the oligomerization domain have been underestimated in IARC data at codons 331, 337, and 342. Truncating mutations (nonsense, splice, and deletions) occur at codons 331 and 342, whereas missense mutations resulting in loss of function are more frequent at codon 337. The oligomerization domain is important for p53 activity because its transactivation activities require the formation of tetramers. Mutations in this domain are thus more prevalent than previously recognized. Overall, based on NGS data, about 15% of all mutations occur outside the DNA-binding domain, which may also explain the differences in mutation prevalence between IARC and NGS data. However, mutations outside the DNA-binding domain do not occur at the same frequency in all cancer types. Indeed, in glioblastomas, less than 4% of point mutations are located outside the DNA-binding domain, whereas in some cancers such as lung adenocarcinoma, esophagus squamous cell carcinoma, or breast cancer, 15%–20% of mutations may occur outside the DNA-binding domain (Supp. Table S5).

To further compare codon distributions between the two datasets by cancer type, we analyzed cancer types with sufficient data in both datasets (at least 60 point mutations). In low grade astrocytomas, more than 20% of point mutations occur at codon 273 where a missense mutation, p.R273C, is predominant in both datasets (Fig. 8, ASTRO). Although codon 273 is a general hotspot, this specific substitution stands out from other hotspots in both datasets only in low grade astrocytomas. Its possible functional impact in astrocytoma development remains to be determined. In hepatocellular carcinoma (HCC), a well-known hotspot at codon 249 is the site of over 30% of point mutations in IARC data. The specific mutation

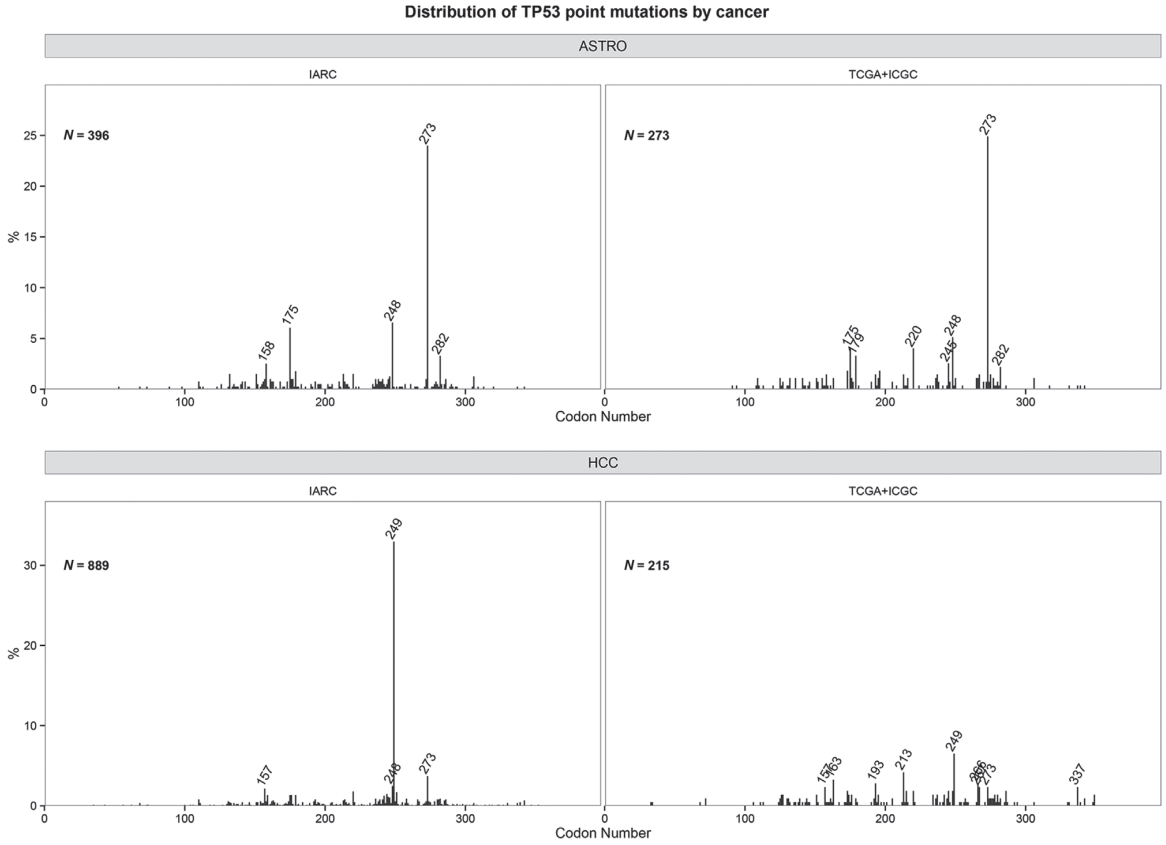


Figure 8. Codon distributions of *TP53* point mutations in low grade astrocytoma (ASTRO) and in hepatocellular carcinoma (HCC) in IARC and NGS datasets.

at this site, p.R249S, is linked to exposure to aflatoxin B1 (AFB1) in populations exposed to both AFB1 and hepatitis B virus in Africa and Asia [Bressac et al., 1991; Hsu et al., 1991]. Interestingly, it is also the most prevalent point mutation in NGS data, although with a frequency below 10% (Fig. 8, HCC). In this dataset, the seven cases with this mutation were from the TCGA data source, which has only records from Northern American populations. However, the seven cases were documented as African or Asian. In IARC data, 75% of HCC cases are from Asian or African populations. The difference in p.R249S mutation proportion in the two datasets may thus be explained by the difference in populations, and their underlying mutagenic exposures, covered in the two datasets. In bladder cancer, two hotspots at codons 280 and 285 are similarly represented in IARC and NGS data (Supp. Fig. S3). These hotspots have been linked to tobacco smoking and aromatic amines but their exact origin and functional role in bladder cancer remain unknown [Pfeifer et al., 2002]. Several tobacco-related hotspots (G>T mutations at codons 158, 249, 248, and 273) are also similarly represented in both datasets in lung cancer (Supp. Fig. S3). Overall, cancer-specific hotspots are very similar between IARC and NGS datasets (Supp. Fig. S3).

In conclusion, these data show that up to 20% of *TP53* somatic mutations may occur in the N- and C-terminus of p53 protein, a proportion underestimated in previous analyses. They also show that cancer-specific hotspots, several of which have been linked to specific mutagens or mutational processes, are well conserved across populations covered in IARC and NGS studies.

TP53 Status in Human Cell-Lines: Reassessment Using Omics Data from the CCLE Project

Because *TP53* has been shown to be mainly inactivated by gene mutations, the *TP53* status of cancer cell-lines reported in the literature has been determined by direct sequencing of the gene. The IARC dataset on cell-line status includes over 2,000 cell-lines for which *TP53* gene status has been assessed by Sanger sequencing. Other less frequent modes of inactivation have been described including inhibitory protein–protein interactions, copy number variations (CNV) and loss of mRNA expression. The extents to which these alternative mechanisms occur in human cancers and participate in *TP53* inactivation remain unclear. Recently, the CCLE (http://www.broadinstitute.org/ccle/home) project has performed omics analyses, including gene expression, chromosomal CN, and DNA and RNA sequencing data from over 1,000 cancer cell-lines [Barretina et al., 2012].

We have analyzed CCLE data and data from IARC and COMSIC to reassess *TP53* status in these cell-lines taking into account gene sequence together with CNV and mRNA expression. A total of 997 cell-lines were analyzed (Supp. Table S6). *TP53* mutations were not

detected by CCLE screen in 320 cell-lines and among them, 11 had bi-allelic loss of *TP53* accompanied by loss of mRNA expression (*TP53* CN ratio <= 0.25 and *TP53* MAS5-150 <= 32), four had loss of mRNA expression with mono-allelic loss of *TP53* (*TP53* CN ratio >0.25 and <=0.6, and *TP53* MAS5-150 <= 32), and 16 had loss of mRNA expression without CNV changes. These lines may thus be considered as defective for *TP53* activity. In three of these 31 cell-lines, a *TP53* mutation has been reported by others (data from COSMIC or IARC database, see Supp. Table S6) and may thus have been missed by CCLE screens, although we cannot exclude that the difference may be due to clonal divergence between the cell-lines analyzed by CCLE and the other studies. Interestingly, in some cell-lines with bi-allelic deletion such as cor-l23, hs766t and sk-n-mc, a robust *TP53* mRNA expression was observed. Deletion of coding exons 2,3,4 have been reported in these cell-lines [Lu et al., 2001; Slebos et al., 1998; Westwood et al., 2002]. Such cases may hint to the possibility that some of *TP53* isoforms may be due to deletions of subsets of *TP53* exons and may not be naturally occurring in the normal healthy tissues.

Among the 714 cell-lines carrying a *TP53* gene mutation, mRNA expression correlated with mutation type. Indeed, mRNA expression levels were lower in cell-lines with truncating mutations (frameshift, nonsense, and splice mutations) compared with cell-lines with missense mutations or inframe indels (median MAS5-150 normalized expression values: 39 vs. 331; average: 107 vs. 412, respectively). If only cell lines with bi-allelic truncating mutations (at least 95% of reads are mutated) are considered, decrease in *TP53* mRNA expression levels is even more pronounced (median MAS5-150 normalized expression values: 37 and average: 71). This loss of expression in samples with truncating mutations may be explained by the established nonsense-mediated decay mechanism.

Pharmacogenomics data, such as the one available from the Genomics of Drug Sensitivity in Cancer project [Yang et al., 2013], can provide additional insights into *TP53* and *TP53* pathway functionality in the cell-line in question. Indeed, from a pharmacogenomics perspective loss of *TP53* functionality by different mechanisms have very similar consequences [Sonkin et al., 2013] and studies have shown that it is critical to account for multiple mechanisms of *TP53* inactivation in order to correctly interpret pharmacological responses [Saiki et al., 2015, Sonkin, 2015]. The *TP53-MDM2* inhibitor, Nutlin-3a, has been shown to be effective only in cell-lines that have a functioning or at least partially functioning *TP53* pathway [Efeyan et al., 2007]. A total of 475 cell-lines included in the CCLE panels have been analyzed in other studies for their sensitivity to two inhibitors currently tested in a number of clinical trials [Jeay et al., 2015; Saiki et al., 2015]. Matching data from these studies to the updated *TP53* status (Supp. Table S6), revealed that 69% (79 out of 115) of wild-type (WT) cell-lines were sensitive and 96% (341 out of 355) of mutated cell-lines were insensitive to one of the *TP53-MDMD2* inhibitors. Five cell-lines with loss of mRNA expression, but no mutation detected, were all insensitive to *TP53-MDMD2* inhibitors. In 36 WT cell-lines that were insensitive to *TP53-MDM2* inhibitors, it is possible that a *TP53* alteration may have been missed in few cell-lines. However, insensitivity to *TP53-MDM2* inhibitors could be potentially due to other mechanisms or to an alternative mechanism of *TP53* inactivation. The melanoma cell-line C32 for example has an MDM4 amplification which is known to result in *TP53* inhibition via MDM4-TP53 protein–protein interaction [Wade et al., 2013] and may be the reason for C32 insensitivity to *TP53-MDM2* inhibitor(s).

The IARC TP53 Database contains mutation information for cell-lines from multiple sources, this helps to increase certainty of the *TP53* status and highlight potential discrepancies. In the analysis performed here, 93 cell-lines scored as *TP53* wild-type by CCLE, contained documented mutation in other studies (COSMIC or IARC sources). Careful manual review of CCLE RNA-seq BAM files and/or HC sequencing BAM files for these lines identified the expected mutation in 70 cases. This highlights the importance of considering multiple sources of *TP53* alterations, especially if *TP53* status is critical for the particular research. In 13 cases, CCLE BAM files contained no evidence of mutation. These discrepancies could be due to several reasons, including false positive calls, misalignment, cell-lines misidentification, or clonal divergence due to cell passaging.

Overall, these data show that *TP53* is inactivated in a majority of cell-lines (74%, 737 out of 997), with less than 3% (19 out of 737) of cell-lines that are altered by large bi-allelic deletions that were detected by CN analysis.

Conclusions

The IARC TP53 Database has been a reference resource for all data reported on *TP53* gene mutations and their phenotypic consequences over the last 20 years. The wide range of data types and annotations gathered from various resources and centralized in this unique resource provides a powerful framework for the analysis of the relevance of *TP53* variants in various aspects of cancer medicine (etiology, diagnosis, and prognosis). Massive genome-wide screens of human tumor samples have confirmed the master role of *TP53* in tumor suppression. It is indeed the most altered gene in human cancers. These screens are also generating new data on *TP53* somatic mutations at high speed. These NGS-derived data are not curated in the IARC TP53 Database as they are freely available from other public repositories (COSMIC, ICGC, TCGA, cBioportal). The analysis presented here shows that the number of *TP53* somatic mutations in the IARC database is larger than what is currently available in the genomic repositories (29,000 vs. 4,000 mutations). However, the higher sensitivity and more complete screening achieved by NGS highlighted some hitherto overlooked facts about *TP53* mutations, such as the presence of a significant number of mutations occurring outside the DNA-binding domain in certain cancer types or the loss of *TP53* expression through unknown mechanisms. Despite the accumulated data, there are still many unresolved questions concerning *TP53* functions, including its mode of inactivation in some cancer types and the phenotype of different mutants. It is expected that continuing integration of knowledge derived from data accumulating from various omics technologies and system biology approaches into resources such as the IARC TP53 Database will help resolve these questions.

Acknowledgments

We thank all authors of the articles compiled in the database in particular those who provided additional information or data in advance of print. We also thank Sylvie Nouveau and the LIB Group at IARC for their assistance in the collection of articles. The development of the IARC TP53 Database is funded by IARC.

Disclosure statement: The authors have no conflict of interest to declare.

References

Achatz MI, Olivier M, Le Calvez F, Martel-Planche G, Lopes A, Rossi BM, Ashton-Prolla P, Giugliani R, Palmero EI, Vargas FR, Da Rocha JC, Vettore AL, et al. 2007. The TP53 mutation, R337H, is associated with Li-Fraumeni and Li-Fraumeni-like syndromes in Brazilian families. Cancer Lett 245:96–102.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483:603–607.

Birch JM, Alston RD, McNally RJ, Evans DG, Kelsey AM, Harris M, Eden OB, Varley JM. 2001. Relative frequency and morphology of cancers in carriers of germline TP53 mutations. Oncogene 20:4621–4628.

Bressac B, Kew M, Wands J, Ozturk M. 1991. Selective G to T mutations of p53 gene in hepatocellular carcinoma from southern Africa. Nature 350:429–431.

Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37:e67.

Efeyan A, Ortega-Molina A, Velasco-Miguel S, Herranz D, Vassilev LT, Serrano M. 2007. Induction of p53-dependent senescence by the MDM2 antagonist nutlin-3a in mouse cells of fibroblast origin. Cancer Res 67:7350–7357.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR. 2007. Genome-wide in situ exon capture for selective resequencing. Nat Genet 39:1522–1527.

Hsu IC, Metcalf RA, Sun T, Welsh JA, Wang NJ, Harris CC. 1991. Mutational hotspot in the p53 gene in human hepatocellular carcinomas. Nature 350:427–428.

Hubbell E, Liu WM, Mei R. 2002. Robust estimators for expression analysis. Bioinformatics 18:1585–1592.

Jeay S, Gaulis S, Ferretti S, Bitter H, Ito M, Valat T, Murakami M, Ruetz S, Guthy DA, Rynn C, Jensen MR, Wiesmann M, et al. 2015. A distinct p53 target gene set predicts for response to the selective p53-HDM2 inhibitor NVP-CGM097. Elife 4.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MD, Miller CA, et al. 2013. Mutational landscape and significance across 12 major cancer types. Nature 502:333–339.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081.

Lu X, Errington J, Curtin NJ, Lunec J, Newell DR. 2001. The impact of p53 status on cellular sensitivity to antifolate drugs. Clin Cancer Res 7:2114–2123.

Lunter G, Hein J. 2004. A nucleotide substitution model with nearest-neighbour interactions. Bioinformatics 20 Suppl 1:I216–I223.

Luo JL, Yang Q, Tong WM, Hergenhahn M, Wang ZQ, Hollstein M. 2001. Knock-in mice with a chimeric human/murine p53 gene develop normally and show wild-type p53 responses to DNA damaging agents: a new biomedical research tool. Oncogene 20:320–328.

MacConaill LE, Campbell CD, Kehoe SM, Bass AJ, Hatton C, Niu L, Davis M, Yao K, Hanna M, Mondal C, Luongo L, Emery CM, et al. 2009. Profiling critical cancer gene mutations in clinical tumor samples. PLoS One 4:e7887.

Malkin D, Li FP, Strong LC, Fraumeni JF, Jr, Nelson CE, Kim DH, Kassel J, Gryka MA, Bischoff FZ, Tainsky MA. 1990. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 250:1233–1238.

Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV. 2006. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res 34:1317–1325.

Olivier M, Goldgar DE, Sodha N, Ohgaki H, Kleihues P, Hainaut P, Eeles RA. 2003. Li-Fraumeni and related syndromes: correlation between tumor type, family structure, and TP53 genotype. Cancer Res 63:6643–6650.

Olivier M, Hainaut P, Borresen-Dale AL. 2005. Prognostic and predictive value of TP53 mutations in human cancer. In:Hainaut P, Wiman K, editors. *25 Years of p53 Research*. Dordrecht, The Netherlands: Springer. p 320–328.

Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M. 2007. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum Mutat 28:622–629.

Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. 2002. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking- associated cancers. Oncogene 21:7435–7451.

R Development Core Team. 2014. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ribeiro RC, Sandrini F, Figueiredo B, Zambetti GP, Michalkiewicz E, Lafferty AR, DeLacerda L, Rabin M, Cadwell C, Sampaio G, Cat I, Stratakis CA, et al. 2001. An inherited p53 mutation that contributes in a tissue-specific manner to pediatric adrenal cortical carcinoma. Proc Natl Acad Sci USA 98:9330–9335.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26.

Sagne C, Marcel V, Amadou A, Hainaut P, Olivier M, Hall J. 2013. A meta-analysis of cancer risk associated with the TP53 intron 3 duplication polymorphism (rs17878362): geographic and tumor-specific effects. Cell Death Dis 4:e492.

Saiki AY, Caenepeel S, Cosgrove E, Su C, Boedigheimer M, Oliner JD. 2015. Identifying the determinants of response to MDM2 inhibition. Oncotarget 6:7701–7712.

Slebos RJ, Resnick MA, Taylor JA. 1998. Inactivation of the p53 tumor suppressor gene via a novel Alu rearrangement. Cancer Res 58:5333–5336.

Sonkin D. 2015. Expression signature based on TP53 target genes doesn't predict response to TP53-MDM2 inhibitor in wild type TP53 tumors. Elife 4.

Sonkin D, Hassan M, Murphy DJ, Tatarinova TV. 2013. Tumor suppressors status in cancer cell line Encyclopedia. Mol Oncol 7:791–798.

Wade M, Li YC, Wahl GM. 2013. MDM2, MDMX and p53 in oncogenesis and cancer therapy. Nat Rev Cancer 13:83–96.

Westwood G, Dibling BC, Cuthbert-Heavens D, Burchill SA. 2002. Basic fibroblast growth factor (bFGF)-induced cell death is mediated through a caspase-dependent and p53-independent cell death receptor pathway. Oncogene 21:809–824.

Whibley C, Pharoah PD, Hollstein M. 2009. p53 polymorphisms: cancer implications. Nat Rev Cancer 9:95–107.

Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, Ramaswamy S, Futreal PA, et al. 2013. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res 41:D955–961.

Yu D, Berlin JA, Penning TM, Field J. 2002. Reactive oxygen species generated by PAH o-quinones cause change-in-function mutations in p53. Chem Res Toxicol 15:832–842.

Appendix IV: Modeling cancer driver events in vitro using barrier bypass-clonal expansion assays and massively parallel sequencing

Hana Huskova, **Maude Ardin**, Annette Weninger, Karina Vargova, Sarah Barrind, Stephanie Villar, Magali Olivier, Tomas Stopka, Zdenko Herceg, Monica Hollstein, Jiri Zavadil and Michael Korenjak

Under first revision, Oncogene, 2017

Modeling cancer driver events *in vitro* using barrier bypass-clonal expansion assays and massively parallel sequencing

Hana Huskova[1,2], Maude Ardin[1], Annette Weninger[3], Karina Vargova[2], Sarah Barrin[4], Stephanie Villar[1], Magali Olivier[1], Tomas Stopka[2], Zdenko Herceg[5], Monica Hollstein[1,3,6,*], Jiri Zavadil[1,*] and Michael Korenjak[1,*]

[1] Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, 69008 Lyon, France.
[2] Biocev, First Faculty of Medicine, Charles University in Prague, 25250 Vestec, Czech Republic.
[3] Deutsches Krebsforschungszentrum, D-69120 Heidelberg, Germany.
[4] Dynamics of T cell interactions Team, Institut Cochin, Inserm U1016, 75014 Paris, France.
[5] Epigenetics Group, International Agency for Research on Cancer, 69008 Lyon, France.
[6] Faculty of Medicine and Health, University of Leeds, Leeds LS2 9JT, United Kingdom.

*To whom correspondence should be addressed. Michael Korenjak, Tel: +33 4 72 73 85 39, Fax: +33 4 72 73 83 22, Email: korenjakm@iarc.fr. Correspondence may be also addressed to: Jiri Zavadil, Tel: +33 4 72 73 85 39, Fax: +33 4 72 73 83 22, Email: zavadilj@iarc.fr, and Monica Hollstein, Tel: +44 19 43 60 55 48, Fax: +33 4 72 73 83 22, Email: M.Hollstein@leeds.ac.uk.

Author contributions:

MK, HH: Designed the study, acquired data and led the analysis and interpretation of results. Wrote the manuscript and approved the final version.

JZ, MH: Conceived and designed the study and interpreted the results. Revised the manuscript and approved the final version.

MO: Designed the study and analysed data. Revised the manuscript and approved the final version.

MA, AW, KV, SB, SV: Acquired and analysed data. Revised the manuscript and approved the final version.

TS, ZH: Interpreted the results, revised the manuscript and approved the final version.

Running Title: *In vitro*-modeling of cancer driver events

Conflict of interest: The authors declare no conflict of interest

ABSTRACT

The information on candidate cancer driver alterations available from public databases is often descriptive and of limited mechanistic insight, which poses difficulties for reliable distinction between true driver and passenger events. To address this challenge, we performed in-depth analysis of whole-exome sequencing data from cell lines generated by a barrier bypass-clonal expansion (BBCE) protocol. The employed strategy is based on carcinogen-driven immortalization of primary mouse embryonic fibroblasts and recapitulates early steps of cell transformation. Remarkably, almost 200 COSMIC Cancer Gene Census genes were found mutated, many of them recurrently, in a set of twenty-five immortalized cell lines. These recurrent alterations affected pathways regulating DNA damage response and repair, transcription and chromatin structure, cell cycle, and cell death, as well as developmental pathways. The functional impact of the mutations was strongly supported by the manifestation of several known cancer hotspot mutations among the identified alterations. When investigating the contribution of additional, putative cancer driver mutations to the immortalized phenotype, we identified a novel Ras-mediated dependence of coding *Smarcd2* and *Smarcc1* mutations on PRC2 histone methyltransferase activity. This interplay is in keeping with the roles of mutations in other BAF complex subunits in cancer cells. We propose that the information on the genetic contents generated by deep sequencing of the BBCE cell lines constitutes a unique resource that can yield mechanistic insights into driver events relevant to human cancer development.

Keywords: cancer driver, mutation, carcinogen, primary cell immortalization, massively parallel sequencing

INTRODUCTION

During the course of their lifetime, eukaryotic cells are exposed to various mutagenic processes that cause DNA damage and mutations. Mutation analysis can help uncover specific mutational signatures associated with active or past mutational processes (1-3), as well as shed light on biological mechanisms critical for tumor development. Most alterations found in tumors are passenger mutations that accumulate during tumorigenesis but do not critically affect cell fitness. However, a small subset of alterations, so-called cancer driver mutations, can confer a selective growth advantage to a cell, which, in turn, can lead to the expansion of a clonal cell population and tumor development (4). Discriminating driver from passenger events is one of the priorities in cancer research. In order to pinpoint candidate cancer driver alterations amongst the myriad of somatic mutations available from cancer genome sequencing studies, numerous computational approaches have been developed. These are either gene-centred methods that are based on the mutation frequency of individual genes compared to the background mutation rate (5-10), or network approaches that identify driver genes based on mutual exclusivity of genomic alterations (11-16). Application of these approaches to mutation data generated by large sequencing consortia led to the following important observations: First, hundreds of high-confidence candidate driver genes have been extracted using these methods, many of which are novel findings (17-19). Almost 600 genes have been implicated in cancer development to date and are included in the Cancer Gene Census (20). Second, even analyses that are based on highly overlapping mutation datasets vary considerably in the candidate drivers that they identify (17, 18), raising the possibility of a sizable number of false positives among the candidate driver events.

Despite the progress made in recent years, much of the knowledge regarding candidate cancer driver alterations remains descriptive and of limited mechanistic insight, emphasizing the need for rapid and robust systems that provide well-controlled experimental investigation of the functional impact of candidate driver events. The necessity of a cell to bypass senescence and become immortal in order for a tumor to develop is well established (21). Senescence bypass in rodent cells, which express telomerase and possess long telomeres, can be achieved by mutations in oncogenes and tumor suppressor genes, most importantly those belonging to the p53-p19^{ARF} tumor suppressor pathway (22). In contrast, human cells must also reactivate telomerase in order to bypass senescence, which likely explains why immortalization following exposure of primary diploid human cells to carcinogenic

insult is difficult to achieve and has rarely been reported (23-25). Therefore, rodent cell lines have been extensively studied for their potential to model the events associated with cell immortalization and transformation (22, 26). However, some of the major concerns regarding their applicability include the dependence of these assays on phenotypic readouts to assess transformation and an incomplete understanding of mechanisms. Recent advances in genome sequencing, together with the development of highly specific pharmacological inhibitors, have created exciting opportunities for mechanistic characterization of candidate cancer driver alterations using *in vitro* carcinogen exposure assays.

In the present study, we explored how whole-exome sequencing (WES) of carcinogen-immortalized primary mouse embryonic fibroblasts (MEFs) can be utilized to identify candidate cancer driver events. The MEF exposure system takes advantage of a biological barrier, which creates a selective pressure for clonal outgrowth of immortalized cells that have acquired a genetically-driven growth advantage. Importantly, it has already been shown that such barrier bypass-clonal expansion (BBCE) protocols select for immortalized cells with human *TP53* hotspot mutations (26-29). We present a comprehensive analysis of exome-wide sequencing of DNA from twenty-six immortalized MEF cell lines derived upon carcinogen exposure or other stimulus. Our results reveal frequent alterations in high-confidence cancer genes and genes affecting biological pathways implicated in critical steps of cell transformation, and *bona fide* cancer driver mutations. On the basis of these results and our functional studies of established and candidate cancer driver alterations using specific pharmacological inhibitors, we propose the MEF BBCE assay coupled with massively parallel sequencing as a tool for characterizing the functional impact of cancer driver-like events.

MATERIAL AND METHODS

Cell lines. Sixteen of the 26 Hupki (humanized p53 knock-in) MEF cell lines (Table S1) were generated from carcinogen-exposed and unexposed primary MEFs (27-32). The additional 10 cell lines were established for this study using the same procedure (33, 34). Briefly, carcinogen-exposed or untreated primary cells were cultivated until senescence bypass and immortalized cell lines with non-synonymous *TP53* mutations were preferentially chosen for WES (Table S1). Clonal populations were generated from cell lines AA_2, AA_3 and MNNG_4 by dilution cloning.

Whole exome sequencing and data processing. WES data used in this study were generated previously for 14 of the 26 cell lines (30); for this report, 12 additional cell lines were sequenced and the data processed as described in Olivier *et al.* (30). An average of 51.44 million reads (100 bp) were sequenced per sample, of which 98% were mapped and 75% on target (mm9 reference genome), with a mean depth-of-coverage of 54. Variants were called with MuTect software (version 1.1.4) using default parameters. Each immortalized cell line was compared to two or three primary MEF cultures and only overlapping calls were considered, to ensure robust variant calling and exclude potential polymorphisms.

Pathway analysis. RefSeq-annotated mouse genes containing non-synonymous single-base substitutions were analyzed using DAVID (35) and Ingenuity Pathway Analysis (IPA, Ingenuity Systems, Redwood City, CA). If RefSeq gene names were not recognized, aliases were used. Gene Ontology and KEGG pathways were interrogated by DAVID using relaxed criteria, as deregulation of biological processes in transformed cells can occur in the absence of multiple hits. IPA was run with default settings and canonical pathways were extracted using either standard ($p<0.05$) or relaxed criteria ($p<0.175$). The identified biological processes and pathways were prioritized based on recurrence among cell lines and cancer relevance.

Identification of candidate cancer driver mutations. Variants were filtered for exonic non-synonymous and splicing mutations and cross-referenced with cancer-related and chromatin associated genes (4, 20, 36, 37). Mutations were prioritized using a simple scoring system. A score of 1 was added if the mutation was of the exposure-predominant type and therefore likely to have been introduced early in the assay. A score of 1 was added if the mutation was in a known human hotspot, if it was truncating or if it was a splice site mutation. For other mutations, a score of 0.5 was added if

the mutation was located in a functional domain and if the mutation was predicted deleterious in the protein by SIFT via Variant Effect Predictor (38). A score of 0.5 was also added if the allelic frequency of the mutation was higher than 25 %.

Sanger sequencing. Primers for sequencing are listed in Table S3.

Inhibitor treatment. Cells were treated with 20 μM Mek inhibitor U0126 (Sigma Aldrich), 4 μM and 8 μM Ezh2 inhibitor GSK126 (Xcess Biosciences), and DMSO carrier (Sigma Aldrich).

MTT proliferation assay. Cells were plated in 96-well plates and treated as indicated. Cell viability was measured using CellTiter 96® AQueous One Solution Cell Proliferation Assay (Promega). Plates were incubated for 2 hours at 37°C and absorbance was measured in a microplate reader at 492 nm.

Colony formation assay. Cells were seeded in 6-well plates to ensure ~5 000 cells at treatment onset for each cell line and condition. Colonies were visualized after 7 days using crystal violet staining.

Immunoblotting and antibodies. Electrophoresis was performed using 4-20% Mini-Protean TGX Precast Protein Gels (BioRad). The following antibodies were used for immunoblotting: phospho-Erk1/2 (9101, 1:1 000, Cell Signaling), total Erk1/2 (9102, 1:1 000, Cell Signaling), histone H3K27me3 (ab6002, 1:4 000, Abcam), histone H3 (ab1791, 1:20 000, Abcam), Ccnd1 (NCL-L-CYCLIN D1-GM, 1:100, Novocastra), Actin (08691001, 1:25 000, MP Biomedicals),.

Analysis of Ep400 and Trrap mutations in human tumor data. Published studies included in cBioPortal for Cancer Genomics (39, 40) were mined for samples with non-synonymous mutations and small indels in *Ep400* and *Trrap*. The data were visualized using OncoPrinter version 1.0.1. Mutual exclusivity was tested using χ^2-test.

RESULTS

WES of MEF clones from BBCE assays as a means to identify potential cancer driver events

While analyzing WES data from 14 immortalized cell lines to identify mutational signatures introduced by specific carcinogens, we noticed frequent mutations in known and suspected cancer driver genes (30). Therefore, we expanded our effort and generated additional cell lines, in order to perform exome-wide, systematic sequencing analysis of cancer driver-like events followed by in-depth functional characterization of selected alterations (Fig. 1). Twenty-six cell lines were derived from primary MEFs (27-29, 31, 32), 19 of which were generated by treatment with a potent human carcinogen (aristolochic acid (AA), aflatoxin B1 (AFB1), benzo[a]pyrene (B[a]P), N-methyl-N'-nitro-N-nitrosoguanidine (MNNG) or ultraviolet light subclass C (UVC)). Five of the immortalized lines arose spontaneously from untreated cells and were included in the analysis. In addition, we examined two cell lines from MEFs genetically engineered to express activation-induced cytidine deaminase (AID) (Table S1). The immortalized cell lines differ in their morphologies (Data S1 and S2), suggesting that the cells can take distinct immortalizing paths, each driven by different sets of acquired mutations.

WES was carried out at a depth-of-coverage of 25-50x on primary MEFs and the 26 immortalized cell lines (this study; (30). Twenty-five of these lines were used as a test set for analyses, and one line was used as a control in subsequent functional validation experiments. Due to the clonal nature of the immortalized cell lines, sequencing at relatively low coverage was sufficient to identify high-confidence variant sequences (Fig. 1). In most cell lines, the majority of non-synonymous mutations were detected at an allelic frequency (AF) ranging from 25 to 75%, as might be expected for the accumulation of heterozygous mutations combined with clonal expansion in the MEF BBCE assay (Fig. S1). The number of mutations varied substantially depending on the compound or treatment protocol and was highest in MNNG- and B[a]P-treated cells (Table S2). In the analyzed set of cell lines, we identified 16 082 single base substitutions, about half of which were non-synonymous, with missense mutations accounting for the vast majority of non-synonymous alterations (Fig. 2A, Table S2, Data S3). Most carcinogen-induced mutations could be attributed to the predominant substitutions found in human tumors associated with the same exposure (Table S2) and amounted to about double the frequency of the two most common mutation types in spontaneously immortalized cells (69% vs 37%). Among other criteria, this exposure-specific enrichment of mutations introduced early and in a

controlled manner was subsequently exploited to identify and track alterations with a potential functional impact on immortalization (see below).

Hotspot mutations and recurrently mutated cancer genes

The clustering of mutations in certain regions of a gene (hotspots) is potentially indicative of cancer driver events. Interestingly, we identified well-known human cancer hotspot mutations in several clones: missense mutations in *Hras* (c.A182T/p.Q61L) and *Kras* (c.A182G/p.Q61R) and a hotspot mutation in the gene encoding the chromatin remodeling factor Smarcb1 (c.G158A/p.R53Q) (Fig. 2B). The applicability of the human mutation data to the mouse proteins is strongly supported by the sequence homology of the proteins between the two species (Fig. 2B) and is well-established for the Ras Q61L and Q61R mutations (41).

Natural selection favors cells that carry functional mutations in driver genes, which explains why these genes are recurrently mutated in cancer. In a proof-of-principle analysis, we compared recurrently mutated genes in the MEFs to human cancer genes (4, 20) and high-confidence epigenetic modifier genes (due to their recent emergence as critical cancer drivers) (36, 42). In total, 68 cancer and epigenetic modifier genes were found recurrently mutated in the MEF cell lines (Fig. 2C). Besides *TP53*, the mutated status of which was used in most cell lines as an additional indicator of clonality and thus preferentially chosen for WES, several other well-established tumor suppressors (*Apc, Atm, Brca2, Ptch1*) and oncogenes (*Hras, Abl1, Egfr, Myc*) were among the recurrently affected genes. Recurrent, non-synonymous mutations were also found in a number of genes encoding epigenetic modifiers, most frequently affecting histone H3K4-methylation, histone acetylation and ATP-dependent chromatin remodeling (*Kmt2b, Kmt2d, Ep400, Baz1a*) (Fig. 2C).

Cancer driver events often contribute to the deregulation of critical biological processes such as cell proliferation, apoptosis or DNA repair (4, 21). To assess whether similar processes are also affected in the MEF BBCE assay system, we analyzed the non-synonymous mutations from each one of the 25 cell lines for commonly targeted biological processes and pathways, integrating DAVID and Ingenuity Pathway Analysis (IPA) (see Material and Methods for details). In concordance with the importance of deregulation of cell proliferation, apoptosis and DNA repair during cellular transformation, we identified these processes to be among the most frequently and recurrently altered in the set of immortalized cell lines (Data S4). In addition, cell matrix organization,

transcription/chromatin structure, pluripotency, cancer-related signaling pathways and drug metabolism (probably due to selective pressure inflicted by carcinogen treatment) were also affected in the majority of cell lines. These findings suggest that MEF BBCE assays can contribute to the identification and characterization of cancer driver-like events.

A systematic prioritization scheme for high-confidence candidate driver events

In order to explore the functional impact of potential driver alterations on cell transformation and the propagation of immortalized cell lines, we first devised a ranking system to select alterations of high interest from the overall mutation pool. Starting from a list of mutations in established and suspected cancer genes, alterations were scored based on multiple criteria, including carcinogen-specificity, allelic frequency and potential functional impact (hotspot, truncation, functional domain, predicted deleterious *in silico*) (see Material and Methods for details). Using this strategy, we identified candidate driver events in several exposed cell lines, and we focused our subsequent analyses on two cell lines, derived from AA and MNNG treatment (AA_2; MNNG_4). As shown in Table 1, the highest-scoring mutations were distributed among biological processes closely linked to cell transformation and cancer development. Some of them affected well-known cancer genes, such as *TP53, Hras, Jak2, Apc, Atm* or *Brca1*, or were genes that have previously been implicated in the regulation of cellular senescence. Next, we derived multiple cultures by single-cell subcloning to confirm the overall clonal nature of the immortalized AA_2 and MNNG_4 cell lines by Sanger sequencing (Fig. S2 and S3). With the exception of three mutations, all tested candidate driver events were found consistently altered across the subclones. These results confirm the clonality of AA_2 and MNNG_4 regarding the highest-scoring mutations, permitting follow-up studies on the interplay of co-occurring alterations.

The oncogenic Ras Q61L hotspot mutation mediates increased proliferation in cells derived from BBCE assays

Hras (c.A182T/p.Q61L) is a well-characterized driver mutation, which results in constitutive activation of the Ras signaling pathway. Phenotypic comparison of AA_2-1 with two immortal clones lacking the activating *Hras* mutation (AA_3-3, Spont_5) revealed clear differences in cell morphology (Fig. 3A, top panel). AA_2-1 grew in multilayers and the cells appeared less tightly attached to the surface than

AA_3-3 and Spont_5, reminiscent of a partial transition towards anchorage-independent growth. Moreover, the doubling time of AA_2-1 was approximately 12 hours, whereas the other two cell lines had doubling times of around 24 hours, which reflects the standard generation time for most immortalized MEFs in this study. These differences could, at least in part, be due to constitutive activation of the Ras pathway in AA_2-1. Therefore, we treated all three cell lines with the Mek inhibitor U0126 to inhibit the Ras/Raf/Mek/Erk signaling pathway. Using immunoblotting, we observed the most pronounced decrease in phospho-Erk levels at early exposure times (up to 8 hours), and a delayed downregulation of the Ras pathway target Ccnd1 at 24 hours (Fig. 3B). Treatment of cells with 20μM U0126 for 24 hours almost completely reverted the AA_2-1 phenotype, but it had no effect on the morphology of the other cell lines (Fig. 3A). Next, we determined the proliferation rate of AA_2, AA_3-3 and Spont_5 in response to Mek inhibitor treatment (Fig. 3C). In contrast to the *Hras* wild-type cell lines, AA_2 showed a small (7%) but statistically significant decrease in proliferation upon treatment. Given the role of Hras in translating exogenous mitogenic signals, we hypothesized that the effect of the inhibitor on mutant Hras would be potentiated in conditions that limit such exogenous signals. Indeed, upon serum starvation of the cells, Mek inhibitor treatment resulted in a striking decrease in cell proliferation (Fig. 3C). As this effect was not observed in *Hras* wild-type cells, it suggests that only the mutant cell lines have developed a dependency on the Ras signaling pathway, and this might be partly responsible for their overall increased proliferation rate.

Novel BAF complex mutations confer sensitivity to Ezh2 inhibition in a Ras-dependent manner in carcinogen-immortalized MEFs

Recent studies have highlighted loss-of-function mutations in subunits of the SWI/SNF (BAF) chromatin remodeling complex in a large number of human cancers (43). Intriguingly, we observed an equally frequent rate of non-synonymous mutations in genes encoding BAF complex subunits in our set of MEF lines (9 out of 26 cell lines). Moreover, all identified BAF complex mutations were mutually exclusive across the BBCE cell line panel (Fig. 4A), in keeping with the mostly non-overlapping nature of mutations in BAF subunits in sequencing data derived from more than 3 000 TCGA samples (19). Previous elegant work revealed an increased dependence of BAF mutant animal tumors and human cancer cell lines on the PRC2 histone methyltransferase complex, and this dependency is alleviated in human cancer cells with oncogenic *RAS* mutations (44, 45). We set out to test whether we could

recapitulate this functional relationship in immortalized MEF cell lines harboring either a BAF complex mutation alone (*Smarcd2*; MNNG_4), or a combination of BAF and oncogenic Ras mutations (*Smarcc1*, *Hras* Q61L; AA_2). *SMARCD2* and *SMARCC1* mutations had not yet been characterized in the context of PRC2 inhibition. Spontaneously immortalized Spont_5 cells (BAF & Ras wild-type), MNNG_4-2 (BAF mutant, Ras wild-type) and AA_2-1 (BAF & Ras mutant) were treated with Ezh2 inhibitor (GSK126), and cell viability was assessed using MTT and colony formation assays (Fig. 4B, C). Spont_5 cells exhibited a significant decrease in cell viability following treatment, but both assays indicated that a fraction of cells survived. In contrast, the MNNG_4-2 cell line was highly sensitive to GSK126 treatment and showed no remaining viability under the same treatment conditions. The observed GSK126 sensitivity of both BAF wild-type and mutant cells, with a more pronounced effect in the mutants, recapitulates previous findings in MEFs isolated from wild-type and *Arid1a* mutant animals (45). Finally, AA_2-1, which harbors BAF and Ras alterations, was much more resistant to the Ezh2 inhibitor than the other cell lines. Importantly, the same order of sensitivity was observed in both the MTT and the colony formation assays. Immunoblotting analysis showed a striking decrease in H3K27me3 levels upon inhibitor treatment with no observable differences amongst the three cell lines (Fig. 4D). Taken together, we show that mutations in *Smarcd2* and *Smarcc1* confer a (oncogenic) Ras-dependent sensitivity to Ezh2 inhibition, consistent with previous findings for other BAF complex subunits in human cancer cell lines.

DISCUSSION

Building on knowledge gained from single-gene sequencing studies, large-scale tumor sequencing efforts have transformed the field of cancer genetics over the last years. This has led to an explosion in the number of genes that have been implicated in cancer development. In this study we report that a simple, cell-based *in vitro* carcinogen exposure assay, combined with massively parallel sequencing, can contribute to the identification of candidate cancer driver events from human tumor sequencing data. The MEF BBCE assay system provides functional information regarding the impact of particular alterations in the Ras pathway and BAF chromatin remodelling complex on the immortalized cell phenotype and, potentially, on subsequent steps leading to transformation. Given that driver mutations from spontaneous processes or from insult continue to accumulate in the immortalized MEF lines, it is likely that WES of well-established cell lines captures events that contribute to both immortalization and transformation.

Several key characteristics of the BBCE assay highlight its applicability as a promising *in vitro* screening strategy for the identification and investigation of driver events. First, senescence bypass and immortalization of MEFs, a well-known *in vitro* phenomenon with parallels to the conversion of normal cells to tumor cells *in vivo* (22, 25, 30), captures cancer driver events, as previously shown for human tumor p53 hotspot mutations (26-28). Second, MEF BBCE assays with individual genotoxic compounds can help to unravel the effects of complex combinations of factors (genotoxic and non-genotoxic) that contribute to human tumorigenesis, and the use of strong mutagens with characteristic single-base substitution patterns can facilitate the recognition of early driver mutations. Third, genome editing using the CRISPR/Cas9 system is used extensively in mouse cells and can be employed to study the functional impact of candidate driver events by correcting them to the wild-type sequences. Recent work has led to significant improvements in the efficiency of homology-directed repair-mediated genome editing (46, 47), allowing medium- to high-throughput functional screening of cancer driver-like events identified in cell lines generated from BBCE assays. Furthermore, the MEF BBCE assay is a simple and relatively fast *in vitro* procedure that uses primary normal diploid cells that can immortalize within a period of 6-8 weeks (27).

Potential limitations of the MEF BBCE assay system include inadequate metabolic activation of certain pro-carcinogens, the high rate of spontaneous immortalization, and potential species-specific differences in key biological pathways involved in cell transformation. Many of these concerns can be

addressed by simple adjustments to the assay protocol, such as the use of exogenous human liver S9 fraction to activate pro-carcinogens. It is of note, however, that MEFs in culture have been shown to convert a variety of pro-carcinogens to their reactive intermediates (26-28, 31, 32, 48, 49). Frequent spontaneous immortalization of MEFs has been attributed to high mutation rates resulting from oxidative stress under standard culture conditions (20% oxygen) (50, 51). Growth of MEFs under physiological oxygen levels (3%) can reduce background mutation and spontaneous immortalization rates, which should improve the stringency of the MEF system.

Tumor sequencing databases put the mutation frequency of all three human *RAS* genes combined between 9 and 30% (all cancer types), with 98% of mutations affecting amino acid residues G12, G13 or Q61 (52). The Q61L alteration, found in AA_2, is one of the most highly transforming mutations in this residue in NIH-3T3 cells (53). It locks Ras in a constitutively active state and is the most frequent *HRAS* mutation in human prostate adenocarcinoma (54). Positive selection for Ras pathway activation during the immortalization/transformation process is further highlighted by another oncogenic mutation, *Kras* (c.A182G/p.Q61R), in a cell line derived from UVC-exposure. The Q61L and Q61R Ras protein alterations found in our *in vitro* exposure system are the most predominant cancer-associated changes identified at this position (54).

Genome sequencing studies established the BAF ATP-dependent chromatin remodeling complex as one of the most commonly mutated human tumor suppressors (55, 56). Our finding that more than 30% of cell lines derived from MEF BBCE assays harbor non-synonymous mutations in BAF subunits resembles their mutation frequency in human cancers (>20%). The presence of a known cancer hotspot mutation in *Smarcb1* among the *in vitro*-induced mutations, and the non-overlapping nature of mutations in BAF complex components in human cancers (19) and exposed MEFs (Fig. 4A) support the selective deregulation of common pathways. We also observed near mutual exclusivity of mutations in genes encoding the Ep400 and Trrap subunits of the NuA4 histone acetyltransferase complex in our cell line panel as well as in a set of 474 TCGA samples (Fig. S4). Although based on a limited number of 25 BBCE cell lines, these findings are consistent with the notion from cancer genome sequencing that the selection of a mutation in a single component is sufficient to alter the activity of a pathway or protein complex while obviating the need for additional changes (19). Extended sequencing analysis of additional BBCE cell lines is warranted to further investigate broader commonalities between MEFs and human tumors.

Mutation in several BAF complex subunits (SMARCB1, SMARCA4, SMARCA2, ARID1A, PBRM1), sensitizes cancer cells to either inactivation of the EZH2 Polycomb protein, or treatment with an EZH2 inhibitor, and this effect is alleviated upon co-occurring *RAS* mutation (44, 45). In agreement with these findings, the MEF BBCE assay system revealed a high sensitivity of cells harboring a *Smarcd2* mutation to Ezh2 inhibitor treatment, whereas a cell line with simultaneous *Smarcc1* and *Ras* mutations was relatively unresponsive (Fig. 4B, C). The mechanism for this antagonism may be based on the defective removal of Polycomb complexes and their respective H3K27me3 histone modification mark in BAF mutant cells (44). Interestingly, Polycomb repressive complexes have also been postulated to cooperate with DNA methylation in the epigenetic reprogramming of pluripotency genes during MEF immortalization (57). Compared to other BAF complex subunits, among them several suspected cancer drivers, both SMARCD2 and SMARCC1 are infrequently mutated in human cancer. However, their significant positive selection in the *in vitro* assay, combined with the results from the Ezh2 inhibitor experiment, certainly warrants additional studies regarding the role of *SMARCD2* and *SMARCC1* in tumor development. In fact, saturation analysis suggests the existence of many more infrequently mutated cancer drivers (18).

It is important to keep in mind that, due to the criteria applied for choosing cell lines for exome sequencing, which usually included the presence of heterozygous or homo/hemizygous *TP53* mutations as an indicator of clonality, most candidate driver events we identified act as such in the context of *TP53* alterations. This situation resembles what is commonly found in human tumors. Some immortalized cell lines, however, retain wild-type p53. It will be interesting to see if in this context other key regulators of the p53 pathway are affected (58), and if CRISPR/Cas9-mediated correction of potential driver events in the same gene, protein complex or pathway, results in diverse functional outcomes depending on p53 status.

Assessment of tumorigenicity in nude mice of immortalized cell lines that have undergone massively parallel sequencing analysis, and the development of clonal expansion assays using normal diploid human cells, as previously reported for mammary epithelial cells (23), are two promising avenues for future research. Moreover, the ever-evolving arsenal of molecular readouts, and the possibility of studying candidate events in an isogenic background using CRISPR/Cas9-mediated mutation correction, reinforce the potential of BBCE assays for the detailed functional characterization of candidate cancer driver events.

411 **ACKNOWLEDGMENTS**

412 We thank the group of Hiroyuki Marusawa, Kyoto University, Japan, for providing cells from Hupki-
413 AID transgenic mice, Christine Carreira (IARC) for antibodies, the Centre Leon Berard in Lyon,
414 France, for providing computational capacity, and Leigh Ellis, Roswell Park Cancer Institute, and
415 members of the MMB and EGE groups for helpful discussions.

416 This work was supported by INCa – INSERM Plan Cancer 2015 ENV201507 Grant to J.Z; Ministry of
417 Education, Youth and sports of the Czech Republic [LQ1604] together with European Regional
418 Development Fund [CZ.1.05/1.1.00/02.0109] to T.S., Czech Science Foundation GAČR [GAČR 16-
419 27790A] to T.S., Ministry of Health/Grant Agency for Health Research of the Czech Republic [AZV
420 16-27790A] to T.S., Charles University in Prague institutional financing [LH15170, UNCE 204021],
421 and the Mobility fund, First Faculty of Medicine, Charles University in Prague, Czech Republic to K.V.

422

423

424

425

426

427

428

429

430

431

432

433

FIGURE LEGENDS

Figure 1. Study design. Mouse embryonic fibroblasts (MEF) were treated with a carcinogen in an early passage and cultivated until senescence bypass. The resulting cell line was subjected to whole exome sequencing on an Illumina HiSeq2500 sequencer and data were analyzed using the indicated pipeline. Sequence variants were systematically analyzed to identify cancer driver-like events, which were further investigated in functional assays with the help of small molecule inhibitors.

Figure 2. Global mutation analysis. (A) Overview of whole-exome sequencing results from 25 MEF BBCE cell lines. (B) Cancer hotspot mutations identified in MEF BBCE cell lines. Plots, showing mutations in *HRAS*, *KRAS* and *SMARCB1* based on TCGA data, were generated using cBioPortal (39, 40). The mutated residue in MEFs is highlighted by a red circle. Alignment of human and mouse protein sequence around the mutated residue is shown in the inset, the mutated codon is indicated above the alignment. The overall similarity of human and mouse protein sequence is indicated in square brackets. (C) Recurrently mutated cancer and epigenetic modifier genes in the MEF BBCE cell lines. Genes listed in the Cancer Gene Census (black, (20)), oncogenes (red) and tumor suppressor genes (blue) by Vogelstein *et al.* (4) and epigenetic modifiers (green, (36) modified) are indicated. Epigenetic modifiers that are also listed in the Cancer Gene Census are indicated in bold black. Epigenetic modifiers that are also listed as tumor suppressor genes by Vogelstein *et al.* are in bold blue. Epigenetic modifiers that are also listed as oncogenes by Vogelstein et al. are in bold red. Cell lines are arranged concentrically and grouped by carcinogen exposure. Red and black dots represent exposure-predominant and exposure non-predominant mutation types, respectively.

Figure 3. Morphology and survival of Ras-mutant and Ras-wild type cell lines upon Mek inhibitor treatment. (A) Morphology of Ras-mutant (AA_2-1) and Ras-wild type (AA_3-3, Spont_5) cells after 24-hour treatment with 20 μM of Mek inhibitor U0126. (B) Immunoblot showing levels of Erk1/2 phosphorylation, total Erk1/2 protein and a target of MAP kinase pathway (Ccnd1) in AA_2-1 cells upon Mek inhibitor U0126 treatment. The immunoblot was carried out using whole-cell protein extracts. Actin was used as loading control. For clarity reason the blot was cropped for display. (C) Proliferation of Ras-mutant and Ras-wild type cells after 24-hour treatment with 20 μM of Mek inhibitor U0126 in complete (15% FBS) and serum-free (0% FBS) growth medium. Relative absorbance (related to treatment with DMSO carrier), indicative of cell viability, was measured.

Columns represent the mean value and standard error of the mean derived from 3 independent experiments. Significance of two sample two-tailed Wilcoxon test is displayed, · p<0.05, * p<0.01.

Figure 4. Effect of Ezh2 inhibitor treatment in BAF-mutant, BAF-wild type and BAF/Ras double mutant MEF BBCE cell lines. (A) BAF complex members affected by non-synonymous mutations in MEF BBCE cell lines. (B) Cells mutant in Ras and BAF (AA_2-1), wild type in Ras and BAF (Spont_5) and cells with BAF mutation in a wild type-Ras background (MNNG_4-2) were seeded at low density in standard 6-well plates and treated with Ezh2 inhibitor GSK126 for 7 days. Cells were then visualized using crystal violet. The result is representative of 3 independent experiments. (C) Cells mutant in Ras and BAF (AA_2-1), wild type in Ras and BAF (Spont_5) and cells with BAF mutation in a wild type-Ras background (MNNG_4-2) were plated in 96-well plates and treated with GSK126 for up to 96 hours. Relative absorbance (related to treatment with DMSO carrier), indicative of cell viability, was measured at 24, 48, 72 and 96 hours. Columns represent mean value and standard error of the mean derived from 3 independent experiments. (D) Immunoblots for the H3K27me3 chromatin mark in all tested cell lines upon GSK126 treatment. Immunoblot was carried out using acid-extracted histones. Histone H3 immunoblot was performed to control for the baseline level of the protein. For clarity reason the blot was cropped for display.

Supplementary Information accompanies the paper on the Oncogene website (http://www.nature.com/onc).

492 **REFERENCES**

493 1. Hollstein M, Alexandrov LB, Wild CP, Ardin M, Zavadil J. Base changes in tumour DNA have
494 the power to reveal the causes and evolution of cancer. Oncogene. 2016.
495 2. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human
496 cancers. Nat Rev Genet. 2014;15(9):585-98.
497 3. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden
498 in cancer genomes. Curr Opin Genet Dev. 2014;24:52-60.
499 4. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer
500 genome landscapes. Science. 2013;339(6127):1546-58.
501 5. Youn A, Simon R. Identifying cancer driver genes in tumor genome sequencing studies.
502 Bioinformatics. 2011;27(2):175-81.
503 6. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC:
504 identifying mutational significance in cancer genomes. Genome Res. 2012;22(8):1589-98.
505 7. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic
506 Acids Res. 2012;40(21):e169.
507 8. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational
508 heterogeneity in cancer and the search for new cancer-associated genes. Nature.
509 2013;499(7457):214-8.
510 9. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional
511 clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29(18):2238-44.
512 10. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling
513 predicts novel cancer drivers. Mol Syst Biol. 2013;9:637.
514 11. Yeang CH, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in
515 cancer. FASEB J. 2008;22(8):2605-22.
516 12. Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A. Discovering functional modules
517 by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC Med Genomics.
518 2011;4:34.
519 13. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic
520 network modules. Genome Res. 2012;22(2):398-406.

14. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res. 2012;22(2):375-85.
15. Zhao J, Zhang S, Wu LY, Zhang XS. Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics. 2012;28(22):2940-7.
16. Leiserson MD, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput Biol. 2013;9(5):e1003054.
17. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci Rep. 2013;3:2650.
18. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505(7484):495-501.
19. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47(2):106-14.
20. 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.
21. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
22. Odell A, Askham J, Whibley C, Hollstein M. How to become immortal: let MEFs count the ways. Aging (Albany NY). 2010;2(3):160-5.
23. Stampfer MR, Bartley JC. 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. 1985;82(8):2394-8.
24. Severson PL, Vrba L, Stampfer MR, Futscher BW. Exome-wide mutation profile in benzo[a]pyrene-derived post-stasis and immortal human mammary epithelial cells. Mutat Res Genet Toxicol Environ Mutagen. 2014;775-776:48-54.
25. Hahn WC, Weinberg RA. Rules for making human tumor cells. N Engl J Med. 2002;347(20):1593-603.

550 26. vom Brocke J, Schmeiser HH, Reinbold M, Hollstein M. MEF immortalization to investigate
551 the ins and outs of mutagenesis. Carcinogenesis. 2006;27(11):2141-7.
552 27. Liu Z, Hergenhahn M, Schmeiser HH, Wogan GN, Hong A, Hollstein M. Human tumor p53
553 mutations are selected for in mouse embryonic fibroblasts harboring a humanized p53 gene. Proc
554 Natl Acad Sci U S A. 2004;101(9):2963-8.
555 28. Liu Z, Muehlbauer KR, Schmeiser HH, Hergenhahn M, Belharazem D, Hollstein MC. p53
556 mutations in benzo(a)pyrene-exposed human p53 knock-in murine fibroblasts correlate with p53
557 mutations in human lung tumors. Cancer Res. 2005;65(7):2583-7.
558 29. Nedelko T, Arlt VM, Phillips DH, Hollstein M. TP53 mutation signature supports involvement
559 of aristolochic acid in the aetiology of endemic nephropathy-associated tumours. Int J Cancer.
560 2009;124(4):987-90.
561 30. Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallee MP, et al. Modelling
562 mutational landscapes of human cancers in vitro. Sci Rep. 2014;4:4482.
563 31. Feldmeyer N, Schmeiser HH, Muehlbauer KR, Belharazem D, Knyazev Y, Nedelko T, et al.
564 Further studies with a cell immortalization assay to investigate the mutation signature of aristolochic
565 acid in human p53 sequences. Mutat Res. 2006;608(2):163-8.
566 32. Reinbold M, Luo JL, Nedelko T, Jerchow B, Murphy ME, Whibley C, et al. Common tumour
567 p53 mutations in immortalized cells from Hupki mice heterozygous at codon 72. Oncogene.
568 2008;27(19):2788-94.
569 33. Celis JE. Cell biology : a laboratory handbook. 3rd ed ed. Amsterdam: Elsevier Academic
570 Press; 2006. 4 vols. p.
571 34. Liu Z, Belharazem D, Muehlbauer KR, Nedelko T, Knyazev Y, Hollstein M. Mutagenesis of
572 human p53 tumor suppressor gene sequences in embryonic fibroblasts of genetically-engineered
573 mice. Genet Eng (N Y). 2007;28:45-54.
574 35. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database
575 for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4(5):P3.
576 36. Gonzalez-Perez A, Jene-Sanz A, Lopez-Bigas N. The mutational landscape of chromatin
577 regulatory factors across 4,623 tumor samples. Genome Biol. 2013;14(9):r106.
578 37. Cancer Gene Census online. [http://cancer.sanger.ac.uk/census].
579 38. Ensembl Variant Effect Predictor web interface. [http://www.ensembl.org/vep].

580 39. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer
581 genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer
582 Discov. 2012;2(5):401-4.
583 40. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of
584 complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1.
585 41. Westcott PM, Halliwill KD, To MD, Rashid M, Rust AG, Keane TM, et al. The mutational
586 landscapes of genetic and chemical models of Kras-driven lung cancer. Nature. 2015;517(7535):489-
587 92.
588 42. Plass C, Pfister SM, Lindroth AM, Bogatyrova O, Claus R, Lichter P. Mutations in regulators
589 of the epigenome and their connections to global chromatin patterns in cancer. Nat Rev Genet.
590 2013;14(11):765-80.
591 43. Helming KC, Wang X, Roberts CW. Vulnerabilities of mutant SWI/SNF complexes in cancer.
592 Cancer Cell. 2014;26(3):309-17.
593 44. Wilson BG, Wang X, Shen X, McKenna ES, Lemieux ME, Cho YJ, et al. Epigenetic
594 antagonism between polycomb and SWI/SNF complexes during oncogenic transformation. Cancer
595 Cell. 2010;18(4):316-28.
596 45. Kim KH, Kim W, Howard TP, Vazquez F, Tsherniak A, Wu JN, et al. SWI/SNF-mutant
597 cancers depend on catalytic and non-catalytic activity of EZH2. Nat Med. 2015;21(12):1491-6.
598 46. Chu VT, Weber T, Wefers B, Wurst W, Sander S, Rajewsky K, et al. Increasing the efficiency
599 of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat
600 Biotechnol. 2015;33(5):543-8.
601 47. Maruyama T, Dougan SK, Truttmann MC, Bilate AM, Ingram JR, Ploegh HL. Increasing the
602 efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining.
603 Nat Biotechnol. 2015;33(5):538-42.
604 48. Luo JL, Tong WM, Yoon JH, Hergenhahn M, Koomagi R, Yang Q, et al. UV-induced DNA
605 damage and mutations in Hupki (human p53 knock-in) mice recapitulate p53 hotspot alterations in
606 sun-exposed human skin. Cancer Res. 2001;61(22):8158-63.
607 49. vom Brocke J, Krais A, Whibley C, Hollstein MC, Schmeiser HH. The carcinogenic air
608 pollutant 3-nitrobenzanthrone induces GC to TA transversion mutations in human p53 sequences.
609 Mutagenesis. 2009;24(1):17-23.

610 50. Parrinello S, Samper E, Krtolica A, Goldstein J, Melov S, Campisi J. Oxygen sensitivity
611 severely limits the replicative lifespan of murine fibroblasts. Nat Cell Biol. 2003;5(8):741-7.
612 51. Busuttil RA, Rubio M, Dolle ME, Campisi J, Vijg J. Oxygen accelerates the accumulation of
613 mutations during the senescence and immortalization of murine cells in culture. Aging Cell.
614 2003;2(6):287-94.
615 52. Cox AD, Fesik SW, Kimmelman AC, Luo J, Der CJ. Drugging the undruggable RAS: Mission
616 possible? Nat Rev Drug Discov. 2014;13(11):828-51.
617 53. Der CJ, Finkel T, Cooper GM. Biological and biochemical properties of human rasH genes
618 mutated at codon 61. Cell. 1986;44(1):167-76.
619 54. Prior IA, Lewis PD, Mattos C. A comprehensive survey of Ras mutations in cancer. Cancer
620 Res. 2012;72(10):2457-67.
621 55. Kadoch C, Hargreaves DC, Hodges C, Elias L, Ho L, Ranish J, et al. Proteomic and
622 bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human
623 malignancy. Nat Genet. 2013;45(6):592-601.
624 56. Shain AH, Pollack JR. The spectrum of SWI/SNF mutations, ubiquitous in human cancers.
625 PLoS One. 2013;8(1):e55119.
626 57. Tommasi S, Zheng A, Weninger A, Bates SE, Li XA, Wu X, et al. Mammalian cells acquire
627 epigenetic hallmarks of human cancer during immortalization. Nucleic Acids Res. 2013;41(1):182-95.
628 58. Kim JE, Shin JS, Moon JH, Hong SW, Jung DJ, Kim JH, et al. Foxp3 is a key downstream
629 regulator of p53-mediated cellular senescence. Oncogene. 2016.

630

Figure 1

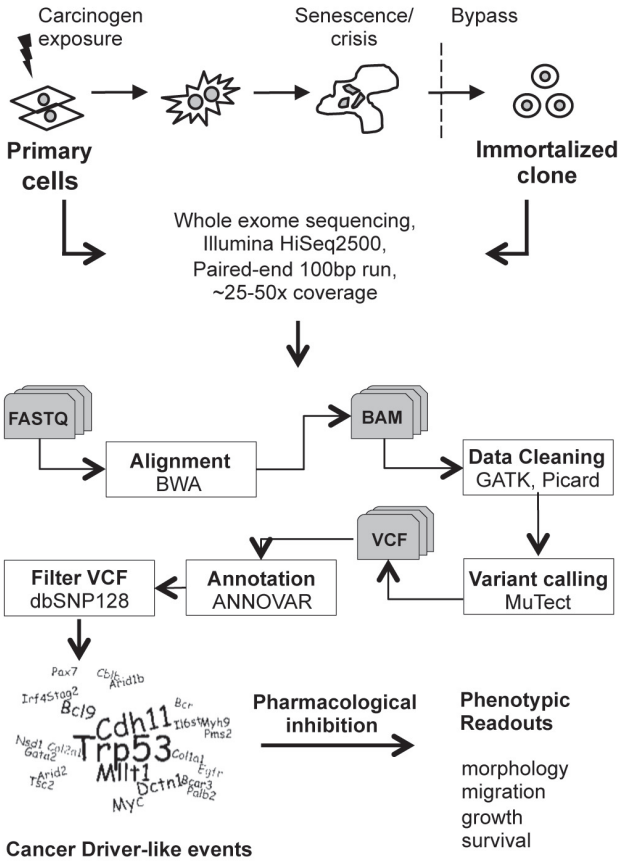

Carcinogen exposure
Senescence/ crisis
Bypass
Primary cells
Immortalized clone
Whole exome sequencing, Illumina HiSeq2500, Paired-end 100bp run, ~25-50x coverage
FASTQ
Alignment BWA
BAM
Data Cleaning GATK, Picard
VCF
Filter VCF dbSNP128
Annotation ANNOVAR
Variant calling MuTect
Pax7
Cblk
Arid1b
Irf4
Stag2
Bcl9
Cdh11
Bcr
Tl6s
Myh9
Pms2
Nsd1
Gata2
Gnl2nl
Trp53
Col1a1
Eif1r
Arid2
Tsc2
Mllt1
Dctn1
Bcar3
Palb2
Myc
Pharmacological inhibition
Phenotypic Readouts
morphology
migration
growth
survival
Cancer Driver-like events

Figure 2

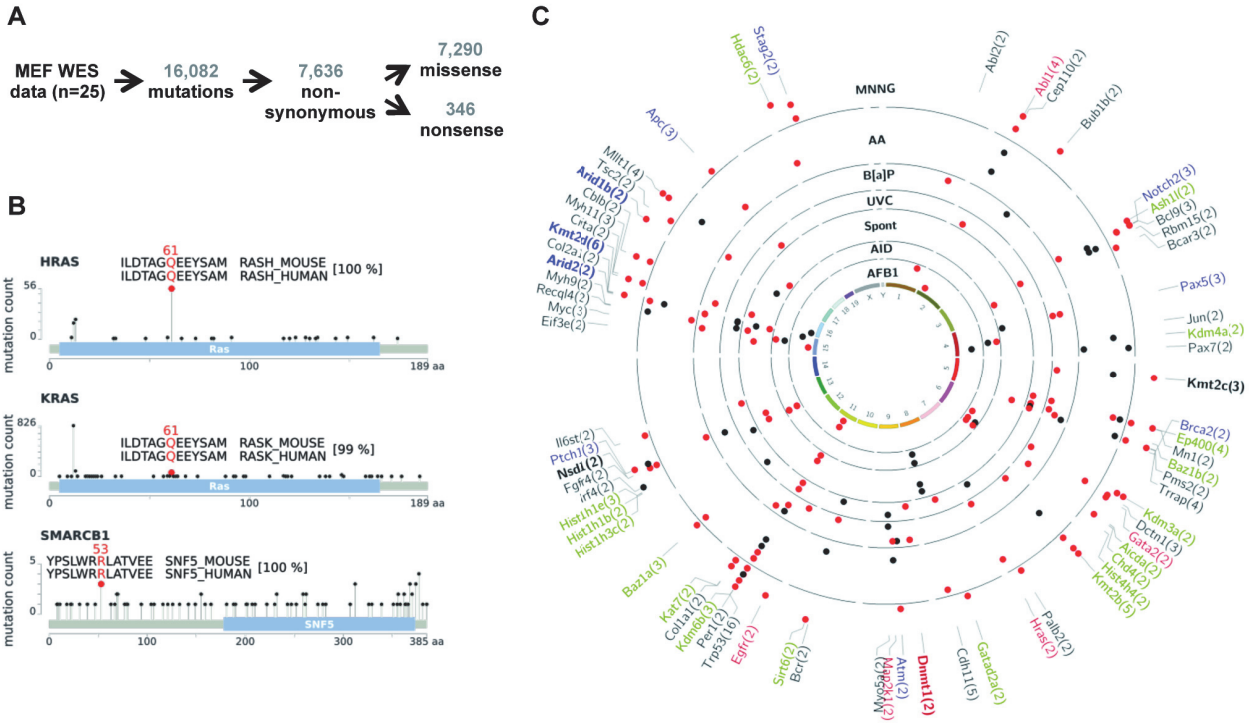

A
MEF WES data (n=25)
16,082 mutations
7,636 non-synonymous
7,290 missense
346 nonsense
B
HRAS
ILDTAGQEEYSAM RASH_MOUSE
ILDTAGQEEYSAM RASH_HUMAN [100 %]
KRAS
ILDTAGQEEYSAM RASK_MOUSE
ILDTAGQEEYSAM RASK_HUMAN [99 %]
SMARCB1
YPSLWRRLATVEE SNF5_MOUSE
YPSLWRRLATVEE SNF5_HUMAN [100 %]
mutation count
Ras
SNF5
aa
C
MNNG
AA
B[a]P
UVC
Spont
AID
AFB1
Apc(3)
Hdac6(2)
Stag2(2)
Abl2(2)
Abil1(4)
Cep110(2)
Bub1b(2)
Notch2(3)
Ash1l(2)
Bcl9(3)
Rbm15(2)
Bcar3(2)
Pax5(3)
Jun(2)
Kdm4a(2)
Pax7(2)
Kmt2c(3)
Brca2(2)
Ep400(4)
Mn1(2)
Baz1b(2)
Pms2(2)
Trrap(4)
Kdm3a(2)
Dctn1(3)
Gata2(2)
Aicda(2)
Chd4(2)
Hist4h4(2)
Kmt2b(5)
Palb2(2)
Hras(2)
Gata6a2a(2)
Cdh11(5)
Dnmt1(2)
Atm(2)
Map2k1(2)
Cyp26a1(2)
Sin6(2)
Bcr(2)
Egfr(2)
Trp53(16)
Per1(2)
Kdm6b(3)
Col1a1(2)
Kat7(2)
Baz1a(3)
Il6st(2)
Ptch1(3)
Nsd1(2)
Fgfr4(2)
Irf4(2)
Hist1h1e(3)
Hist1h1b(2)
Hist1h3c(2)
Mllt1(4)
Tsc2(2)
Arid1b(2)
Cblb(2)
Myh11(3)
Crta(2)
Kmt2d(6)
Col2a1(2)
Arid2(2)
Myh9(2)
Recql4(2)
Myc(3)
Eif3e(2)

Figure 3

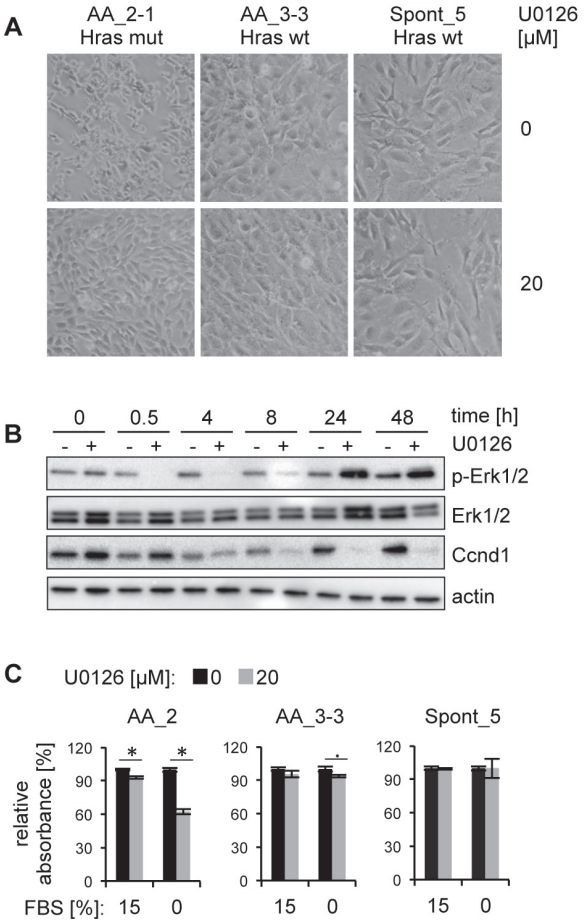

A
AA_2-1
Hras mut
AA_3-3
Hras wt
Spont_5
Hras wt
U0126
[μM]
0
20
B
0
0.5
4
8
24
48
time [h]
- + - + - + - + - + - +
U0126
p-Erk1/2
Erk1/2
Ccnd1
actin
C
U0126 [μM]: 0 20
AA_2
AA_3-3
Spont_5
relative absorbance [%]
120
90
60
30
0
*
*
FBS [%]: 15 0
15 0
15 0

Figure 4

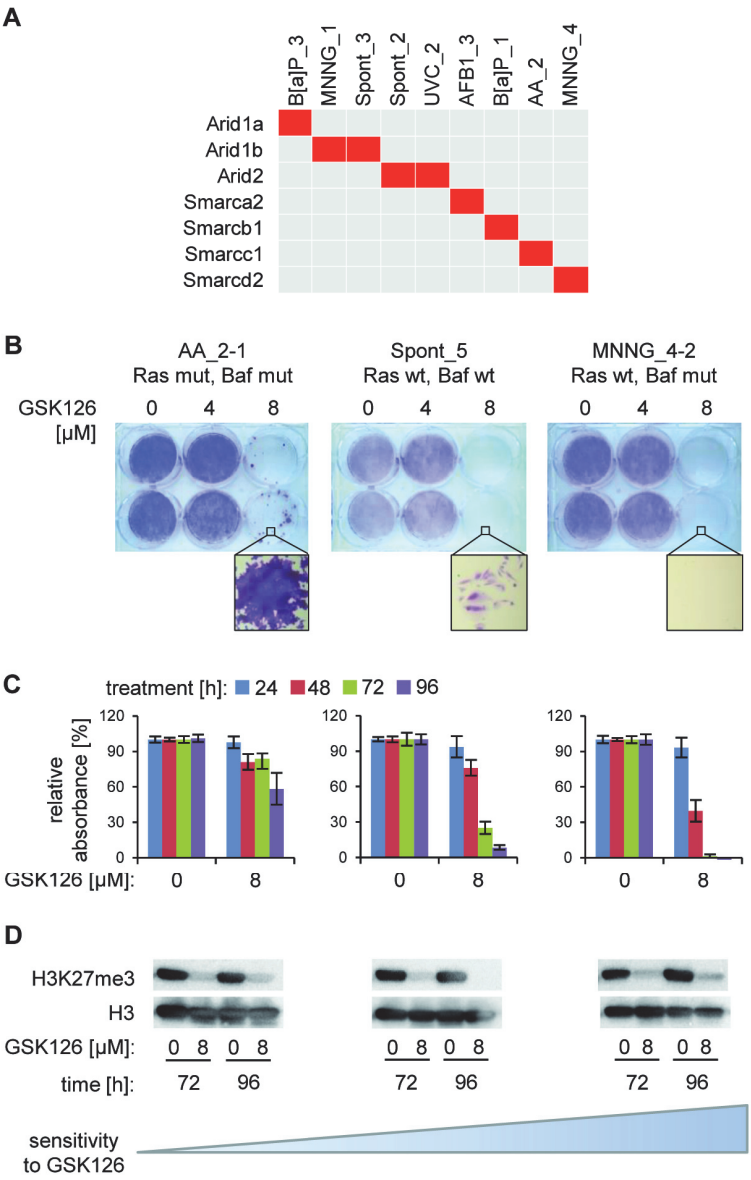

A
B[a]P_3
MNNG_1
Spont_3
Spont_2
UVC_2
AFB1_3
B[a]P_1
AA_2
MNNG_4
Arid1a
Arid1b
Arid2
Smarca2
Smarcb1
Smarcc1
Smarcd2
B
AA_2-1
Ras mut, Baf mut
Spont_5
Ras wt, Baf wt
MNNG_4-2
Ras wt, Baf mut
GSK126 [μM]
0 4 8
C
treatment [h]: 24 48 72 96
relative absorbance [%]
120
90
60
30
0
GSK126 [μM]: 0 8
D
H3K27me3
H3
GSK126 [μM]: 0 8 0 8
time [h]: 72 96
sensitivity to GSK126

Table 1: Candidate driver mutations in two immortalized MEF cell lines.

Cell line	Gene symbol	Function	transcript ID	cDNA change	AA change	Mutated in human tumors (Cosmic) [%]	known to be involved in senescence
AA_2	Cbx7	PRC1 complex	NM_144811	c.T32A	p.F11Y	0.2	
	Cdkn1a*	Cell cycle	NM_007669.5	c.94-2A>T		0.3	YES
	Ep400*	TIP60 complex	NM_029337	c.A970T	p.R324X	2.2	YES
	Ext1 Ext1*	Glycosaminoglycan metabolism	NM_010162	c.A103T c.A1036T	p.S35C p.R346X	0.5	
	Hras*	MAPK signaling	NM_001130443	c.A182T	p.Q61L	2.7	
	Jak2	JAK-STAT singaling	NM_001048177	c.A2479T	p.I827L	30.0	
	Smarcc1*	BAF complex	NM_009211	c.A356T	p.H119L	0.7	YES
	Smyd1	H3K4 methylation	NM_009762	c.T1072A	p.S358T	0.7	
	Tp53* Tp53*	Transcription, DNA repair	NM_000546.4	c.A391T c.A871T	p.N131Y p.K291X	26.8	YES
MNNG_4	Apc*	Wnt signaling	NM_007462	c.C8278T	p.P2760S	10.9	YES
	Atm*	DNA repair	NM_007499	c.C3092T	p.T1031I	4.2	YES
	Baz1a*	ACF complex	NM_013815	c.G392A	p.R131K	0.8	
	Brca1	DNA repair	NM_009764	c.C4322T	p.P1441L	1.3	YES
	Gatad2a	Histone deacetylation	NM_001113345	c.G243A	p.M81I	0.3	
	Jak1* Jak1*	JAK-STAT singaling	NM_146145	c.C1327T c.C1286T	p.P443S p.P429L	1.0	
	Kmt2a*	H3K4 methylation	NM_001081049	c.C9755T	p.P3252L	1.6	
	Prdm1	Transcription	NM_007548	c.C2420T	p.P807L	1.1	
	Setd1a* Setd1a*	H3K4 methylation	NM_178029	c.G1499A c.G5095A	p.S500N p.D1699N	1.2	
	Sin3b*	HDAC	NM_009188	c.C2441T	p.T814I	0.7	YES
	Smarcd2*	BAF complex	NM_031878	c.G497A	p.G166E	0.3	
	Trrap*	TIP60 complex	NM_001081362	c.G6952A	p.V2318M	2.6	
	Tp53* Tp53*	Transcription, DNA repair	NM_000546.4	c.C454T c.C476T	p.P152S p.A159V	26.8	YES

* Validated by Sanger sequencing (Figures S2 and S3).

Supplementary Information

SUPPLEMENTARY DATA, TABLES AND FIGURES LEGENDS

Data S1. A 24-hour time-lapse video of the AA_2 cell line immortalized upon exposure to aristolochic acid. Movie in mpeg format.

Data S2. A 48-hour time-lapse video of a spontaneously immortalized MEF cell line. Movie in mpeg format.

Data S3. Single base substitutions identified in immortalized MEF cell lines. Excel file (xlsx).

Data S4. Selection of biological processes and pathways recurrently affected in immortalized MEFs. Excel file (xlsx).

Table S1. Overview of *TP53* mutations in the cell line collection. In this pdf file.

Table S2. Summary of the distribution and types of mutations in immortalized MEFs. In this pdf file.

Table S3. List of primers used for Sanger sequencing. In this pdf file.

Figure S1. Distribution of mutations based on their allelic frequencies in 25 MEF BBCE cell lines. Mutations in individual cell lines were ranked and plotted based on decreasing allelic frequency. Percentage of mutations with allelic frequency between 25% and 75% is indicated. In this pdf file.

Figure S2. Sanger sequencing of candidate driver mutations in clones derived from AA_2 cell line. DNA sequence is displayed above the chromatogram of the first clone using IUPAC code. In this pdf file.

Figure S3. Sanger sequencing of candidate driver mutations in clones derived from MNNG_4 cell line. DNA sequence is displayed above the chromatogram of the first clone using IUPAC code. If a base is not identical in all clones, it is marked with the respective IUPAC letter in each respective chromatogram. In this pdf file.

Figure S4. Mutations in *Ep400* and *Trrap*. (A) *Ep400* and *Trrap*, core subunits of the NuA4 complex, are affected by non-synonymous mutations in seven MEF BBCE cell lines. (B) Detailed information regarding non-synonymous mutations in *Ep400* and *Trrap* in MEF BBCE cells. (C) Distribution of non-synonymous mutations, as well as small insertions and deletions affecting *EP400* and *TRRAP* in human tumor samples listed in the cBioPortal for Cancer Genomics. Starting from a total of 11966 samples, 474 harbored well-annotated mutations; 239 had mutations in Ep400 and 292 had mutations in Trrap. Fifty-seven samples had both Ep400 and Trrap mutations. Results of χ^2-test for mutual exclusivity of *EP400* and *TRRAP* mutations is indicated, * $p < 10^{-15}$. In this pdf file.

Table S1: Overview of non-synonymous TP53 mutations in the cell line collection.				
Cell line ID	**Exposure type**	**Exposure dose**	**Exposure duration**	**TP53 mutation (reference transcript NM_000546.4)**
AA_1	AA	50 µM	4 days	c.A745T:p.R249W
AA_2	AA	50 µM	4 days	c.A391T:p.N31Y, c.A871T:p.K291X
AA_3	AA	50 µM	4 days	c.C817T:p.R273C
AA_4	AA	50 µM	4 days	c.A774C:p.E258D
AA_5	AA	50 µM	4 days	n. s.
AA_6	AA	50 µM	12 days	n. s.
AA_7	AA	50 µM	8 days	n. s.
AFB1_1	AFB1	2 µM	8 days	c.G818T:p.R273L
AFB1_2	AFB1	2 µM	8 days	c.G743T:p.R248L
AFB1_3	AFB1	2 µM	8 days	c.G743T:p.R248L
AID_1	none	n. a.	n. a.	n. s.
AID_2	none	n. a.	n. a.	n. s.
B[a]P_1	BaP	1 µM	6 days	c.C380T:pS127F, c.G845C:p.R282P
B[a]P_2	BaP	1 µM	6 days	c.C423G:p.C141W
B[a]P_3	BaP	5 µM	2 days	n. s.
MNNG_1	MNNG	20 µM	2 hours	c.C296T:p.S99F
MNNG_2	MNNG	20 µM	2 hours	c.G734A:p.G245D
MNNG_3	MNNG	20 µM	2 hours	c.A961T:p.K321X
MNNG_4	MNNG	20 µM	2 hours	c.C454T:p.P152S, c.C476T:p.A159V
Spont_1	none	n. a.	n. a.	n. s.
Spont_2	none	n. a.	n. a.	n. s.
Spont_3	none	n. a.	n. a.	c.G314C:p.G105A
Spont_4	none	n. a.	n. a.	c.C843G:p.D281E
UVC_1	UVC	20 J/m^2	n. a.	c.G743A:p.R248Q, c.C749A:p.P250H
UVC_2	UVC	20 J/m^2	n. a.	c.C535T:p.H179Y
Spont_5	none	none	n. a.	wild type

AA - aristolochic acid, AFB1 - aflatoxin B1, AID - activation-induced cytidine deaminase, B[a]P - benzo[a]pyrene, MNNG - N-methyl-N'-nitro-N-nitrosoguanidine, Spont - spontaneous immortalization, UVC - UV light class C, n. a. - not applicable, n. s. - not sequenced.

Table S2: Summary of the distribution and types of mutations in immortalized MEFs.

Cell line	Mutations	Non-synonymous	Missense	Nonsense	Predominant Mutation Type
AA_1	680	347	319	28	T:A>A:T
AA_2	615	326	298	28	
AA_3	445	227	213	14	
AA_4	357	180	171	9	
AA_5	265	146	130	16	
AA_6	280	126	123	3	
AA_7	583	310	297	13	
MNNG_1	1232	556	539	17	C:G>T:A
MNNG_2	1133	459	436	23	
MNNG_3	1259	530	505	25	
MNNG_4	1541	663	637	26	
Spont_1	279	124	122	2	C:G>T:A, T:A>C:G
Spont_2	355	173	166	7	
Spont_3	116	69	68	1	
Spont_4	335	165	159	6	
AFB1_1	237	115	111	4	C:G>A:T, C:G>T:A
AFB1_2	318	154	145	9	
AFB1_3	360	205	195	10	
B[a]P_1	1262	675	644	31	C:G>A:T, C:G>G:C
B[a]P_2	716	356	342	14	
B[a]P_3	1395	696	657	39	
AID_1	362	158	155	3	C:G>T:A
AID_2	625	250	245	5	
UVC_1	366	171	167	4	C:G>T:A, T:A>C:G
UVC_2	966	455	446	9	

AA - aristolochic acid, AFB1 - aflatoxin B1, AID - activation-induced cytidine deaminase, B[a]P - benzo[a]pyrene, MNNG - N-methyl-N'-nitro-N-nitrosoguanidine, Spont - spontaneous immortalization, UVC - UV light class C.

Table S3: Sequencing primers.

Primer	Sequence (5' → 3')
APC_C8278T_F	AAGACACCCATGGGAAACAG
APC_C8278T_R	CTTCTTCGTGTTGGTGCTCA
ATM_C3092T_F	GCACTGGCATTTTCCACATA
ATM_C3092T_R	TTCGGAATATGGATCAGCCTA
BAZ1A_G392A_F	GGAAGCTCTTGAATCCGAAA
BAZ1A_G392A_R	GGCCCAGGCTAACCTAGAAG
CDKN1A_F	CGGTGACTCCTACTTCTGTGG
CDKN1A_R	TCTCCGTGACGAAGTCAAAG
EP400_A970T_F	CCCCAGATCAGCAGCATTAT
EP400_A970T_R	TCCTTGAGTGCCTCCATTTC
EXT1_A1036T_F	GCTCTGCTCTGAACCTCCAT
EXT1_A1036T_R	CCCAATTCTGGCTCTTCAAA
HRAS1_A182T_F	ATGGGGTATGATCCATCAGG
HRAS1_A182T_R	CTCACGGGCTAGCCATAGGT
JAK1_2MUTS_F	TCTCGAGAGGAAGCCTTGTC
JAK1_2MUTS_R	CTAAGAGCCATGGCAGGAAA
KMT2A_C9755T_F	AGTGCCCCTTCAAATATTGC
KMT2A_C9755T_R	TAGGGGCTGCTGTAGTTTGC
SETD1A_G1499A_F	CCAGCCCTGAGAGAGAAGAA
SETD1A_G1499A_R	AATTAGCTGGTGCAGGAGGA
SETD1A_G5095A_F	GCTTACACTCCCACCTCCTG
SETD1A_G5095A_R	AAGGACTGAGGCTCCCTTGT
SIN3B_C2441T_F	CAGGGATAGGGCCTCCTTAG
SIN3B_C2441T_R	ACTTGCTGTGTGGACCCTGT
SMARCC1_A365T_F	GGCATCTGGACACCAGACTT
SMARCC1_A365T_R	AAAGGCCTTACCTTGCCATT
SMARCD2_G497A_F	ACTCACACAGGGAGCTGTCC
SMARCD2_G497A_R	GGACTTCTGAAAGAACGCTCA
TP53_EXON4_F	TGCTCTTTTCACCCATCTAC
TP53_EXON4_R	ATACGGCCAGGCATTGAAGT
TP53_EXON5_F	TGAGGTGTAGACGCCAACTCT
TP53_EXON5_R	AACCAGCCCTGTCGTCTCT
TP53_EXON6_F	GCCTCTGATTCCTCACTGAT
TP53_EXON6_R	CGAAAAGTGTTTCTGTCATCC
TP53_EXON7_F	CTGCTTGCCACAGGTCTCCCC
TP53_EXON7_R	TGTGCAGGGTGGCAAGTG
TP53_EXON8_F	TCCTTACTGCCTCTTGCTTCTCT
TP53_EXON8_R	AGGCATAACTGCACCCTTGG
TRRAP_G6952A_F	TCTTTGCCAGGAGCCACTAT
TRRAP_G6952A_R	GCCATCGGGGAATTATTCTT

F - forward, R - reverse

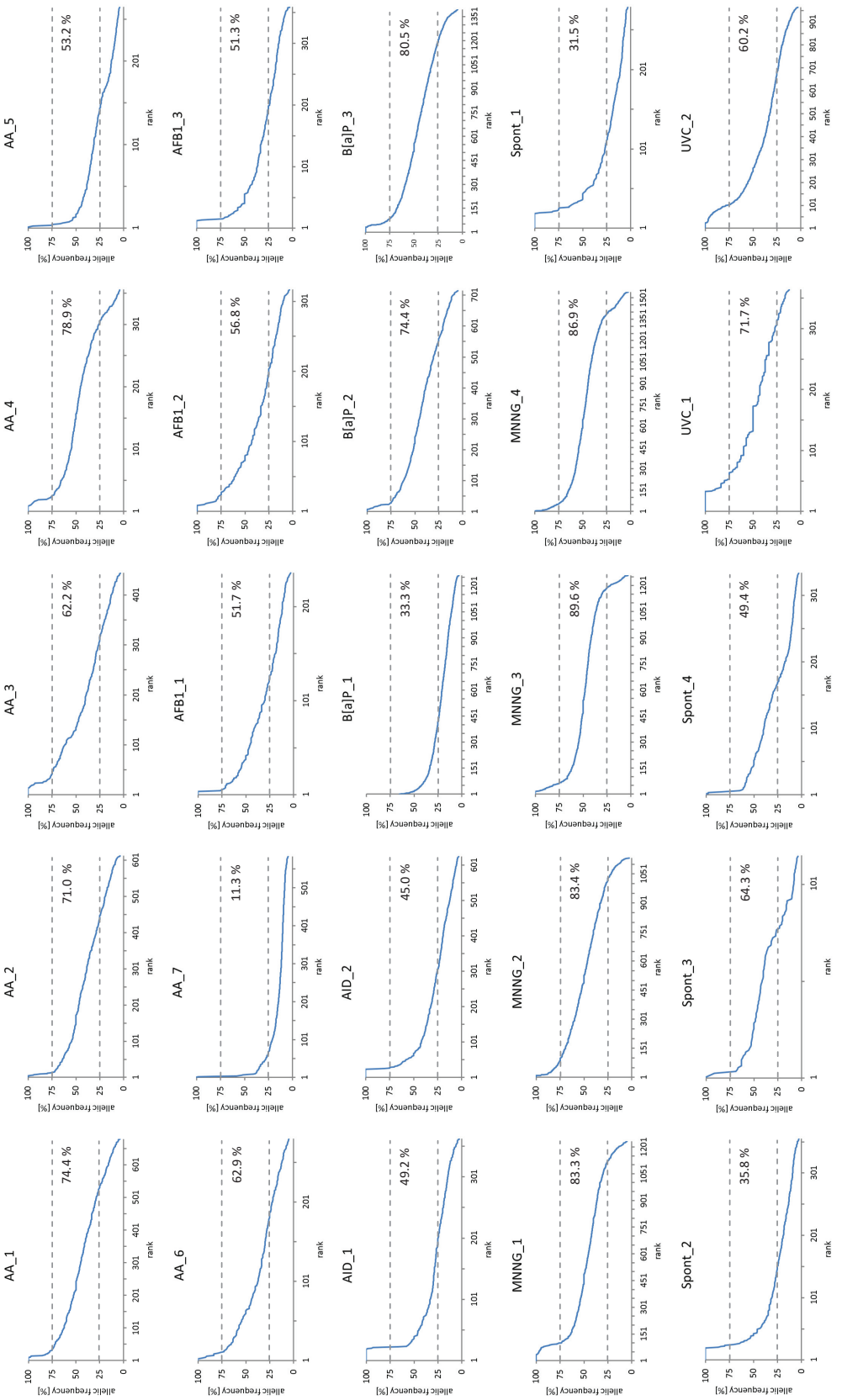

AA_1
74.4 %
AA_2
71.0 %
AA_3
62.2 %
AA_4
78.9 %
AA_5
53.2 %
AA_6
62.9 %
AA_7
11.3 %
AFB1_1
51.7 %
AFB1_2
56.8 %
AFB1_3
51.3 %
AID_1
49.2 %
AID_2
45.0 %
B[a]P_1
33.3 %
B[a]P_2
74.4 %
B[a]P_3
80.5 %
MNNG_1
83.3 %
MNNG_2
83.4 %
MNNG_3
89.6 %
MNNG_4
86.9 %
Spont_1
31.5 %
Spont_2
35.8 %
Spont_3
64.3 %
Spont_4
49.4 %
UVC_1
71.7 %
UVC_2
60.2 %
allelic frequency [%]
rank

Figure S1

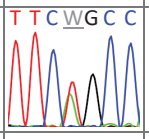
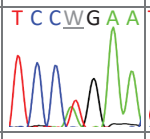
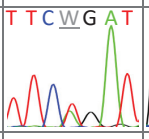
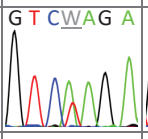
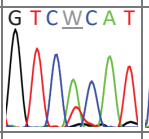
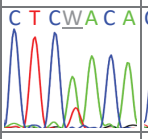
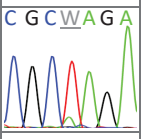
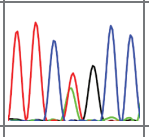
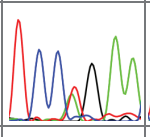
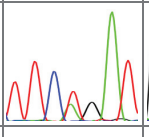
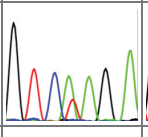
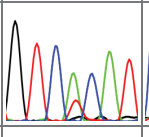
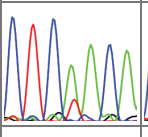
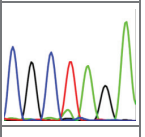
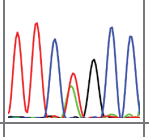
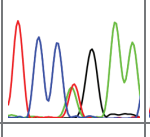
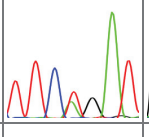
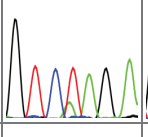
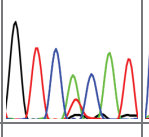
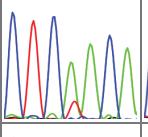
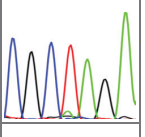
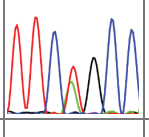
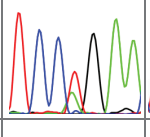
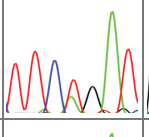
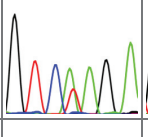
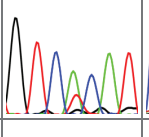
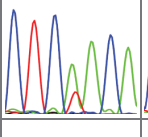
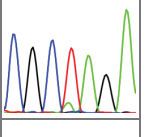
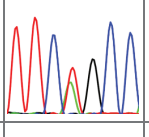
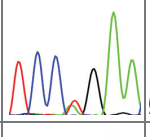
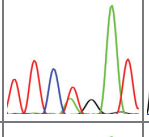
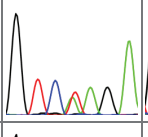
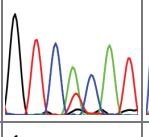
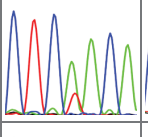
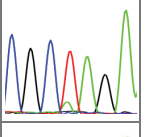
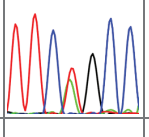
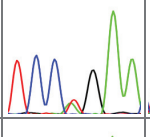
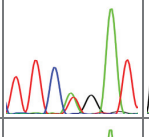
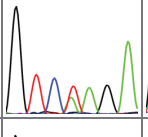
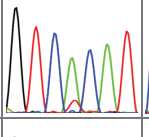
Mutation / Clone	Cdkn1a NM_007669.5 c.94-2A>T	Ep400 NM_029337 c.A970T, p.R324X	Ext1 NM_010162 c.A1036T, p.R346X	Hras NM_001130443 c.A182T, p.Q61L	Smarcc1 NM_009211 c.A356T, p.H119L	Tp53 NM_000546.4 c.A391T, p.N131Y	Tp53 NM_000546.4 c.A871T, p.K291X
AA_2-1							
AA_2-2							
AA_2-3							
AA_2-4							
AA_2-5							
AA_2-6							
AA_2-7							
AA_2-8							
AA_2-9							
AA_2-10							

Figure S2

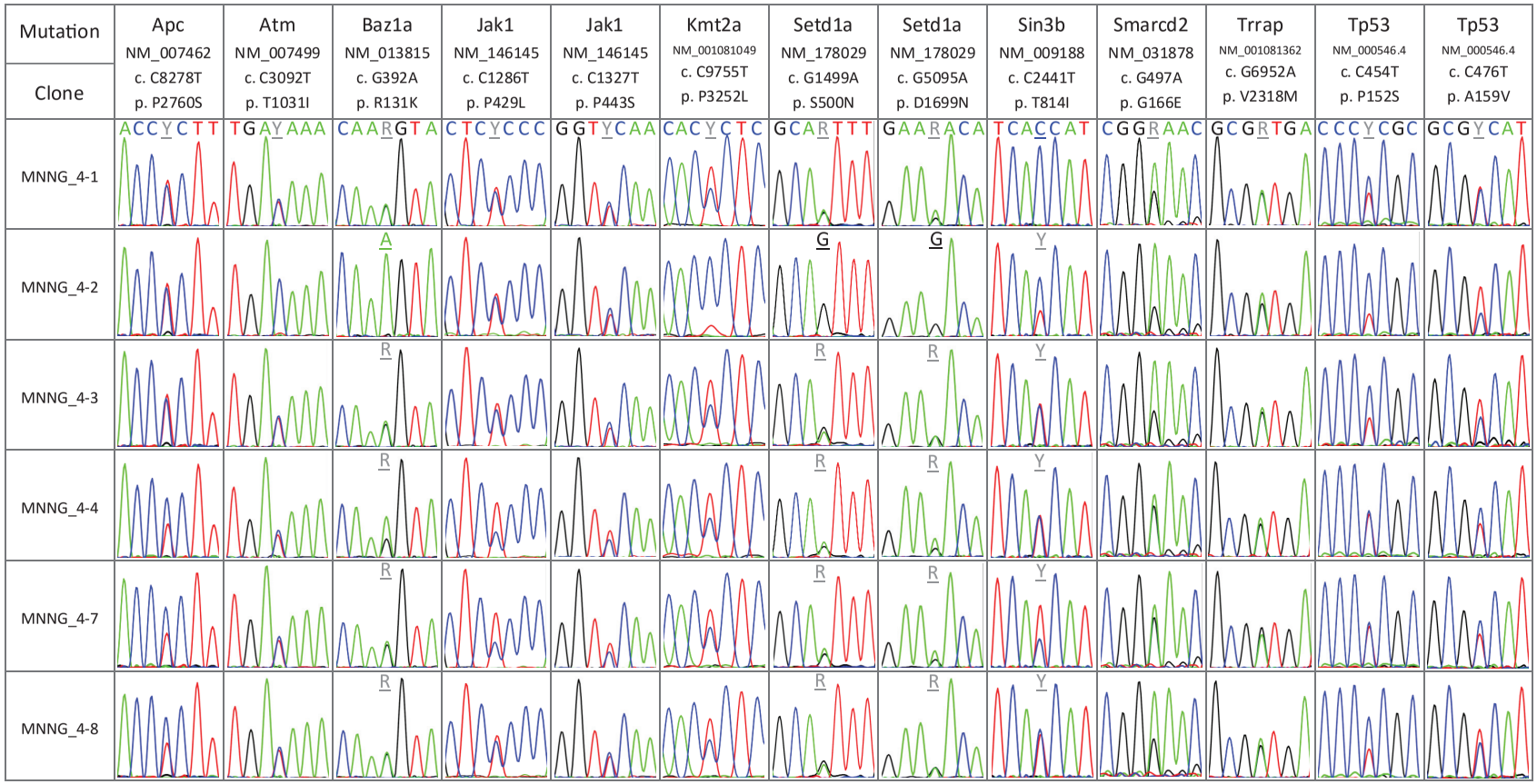
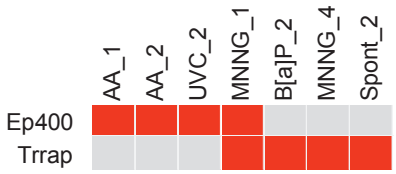

Mutation
Clone
Apc
NM_007462
c. C8278T
p. P2760S
Atm
NM_007499
c. C3092T
p. T1031I
Baz1a
NM_013815
c. G392A
p. R131K
Jak1
NM_146145
c. C1286T
p. P429L
Jak1
NM_146145
c. C1327T
p. P443S
Kmt2a
NM_001081049
c. C9755T
p. P3252L
Setd1a
NM_178029
c. G1499A
p. S500N
Setd1a
NM_178029
c. G5095A
p. D1699N
Sin3b
NM_009188
c. C2441T
p. T814I
Smarcd2
NM_031878
c. G497A
p. G166E
Trrap
NM_001081362
c. G6952A
p. V2318M
Tp53
NM_000546.4
c. C454T
p. P152S
Tp53
NM_000546.4
c. C476T
p. A159V
MNNG_4-1
MNNG_4-2
MNNG_4-3
MNNG_4-4
MNNG_4-7
MNNG_4-8

Figure S3

Figure S4

A



B

Gene symbol	Cell line	transcript ID	cDNA change	AA change
Ep400	AA_1	NM_029337	c.C602G	p.A201G
Ep400	AA_2	NM_029337	c.A970T	p.R324X
Ep400	UVC_2	NM_173066	c.T7649C	p.V2550A
Ep400	MNNG_1	NM_173066	c.C7315T	p.P2439S
Trrap	MNNG_1	NM_001081362	c.G8246A	p.R2749Q
Trrap	B[a]P_2	NM_001081362	c.G8293C	p.G2765R
Trrap	MNNG_4	NM_001081362	c.G6952A	p.V2318M
Trrap	Spont_2	NM_001081362	c.A8125G	p.T2709A

C