Genome-wide modeling of mutation spectra of human cancer-risk agents using experimental systems

Maria Zhivagui

N°d'ordre NNT : 2017LYSE1278

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de

l'Université Claude Bernard Lyon 1

Ecole Doctorale N° accréditation

Biologie Moléculaire, Intégrative et Cellulaire de Lyon (BMIC)

Spécialité de doctorat : **Cancérologie**

Discipline : **Biologie moléculaire et cellulaire**

Soutenue publiquement/à huis clos le 30/11/2017, par :

Maria ZHIVAGUI

Genome-wide modeling of mutation spectra of human cancer-risk agents using experimental systems

Devant le jury composé de :

M. PHILLIPS David	Professeur, King's College London	Président
M. SICHEL Francois	Professeur, Université de Caen-Normandie	Rapporteur
M. BESARATINIA Ahmad	Professeur, Keck School of Medicine, University of Southern California	Rapporteur
M. BAROUKI Robert	Professeur, Inserm-Université Paris Descartes	Examinateur
M^{me}. FERVERS Beatrice	Professeur, Centre Léon Bérard	Examinatrice
M^{me}. BERNET Agnès	Professeur, Centre Léon Bérard	Examinatrice
M. ZAVADIL Jiri	Docteur, Centre International de Recherche sur le Cancer	Directeur de thèse
M. MCKAY James	Docteur, Centre International de Recherche sur le Cancer	Co-directeur de thèse
M. KORENJAK Michael	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 de Mécanismes Moléculaire et Biomarqueurs (MMB)

Centre international de Recherche sur le Cancer (CIRC)

150 Cours Albert Thomas

69372 Lyon cedex 08

France

Abstract

Résumé en français

Modélisation à l'échelle du génome des spectres de mutations des agents de risque de cancer humain en employant des systèmes expérimentaux

Les génomes du cancer présentent une mosaïque de types de mutations. Trente signatures mutationnelles ont été identifiées à partir d'un grand nombre de tumeurs humaines primaires. Déchiffrer l'origine de ces signatures mutationnelles pourrait aider à identifier les causes du cancer humain. Environ 40% des signatures décrites sont d'origine inconnue, soulignant la nécessité de modèles expérimentaux contrôlés pour étudier l'origine de ces signatures. Au cours de mon travail de doctorat, j'ai caractérisé et utilisé des modèles *in vitro* et *in vivo* d'exposition aux cancérogènes, en particulier, les cellules primaires Hupki MEF, les lignées cellulaires HepaRG et lymphoblastoïdes (LCL) ainsi que les tumeurs des rongeurs. Ensuite, que j'ai caractérisé les signatures mutationnelles au niveau de génome entier de plusieurs composés cancérogènes pour lesquels le spectre de mutations n'était pas connu ou controversé.

Tout d'abord, les conditions de cytotoxicités et genotoxicités pour chaque composé ont été établies et la formation d'adduits d'ADN a été évaluée. Suite au séquençage du gène *TP53*, un séquençage au niveau génomique a été effectué des clones de MEF immortalisés dérivés de l'exposition à l'acrylamide, au glycidamide et à l'ochratoxine A (OTA).

Le travail suggère une nouvelle signature mutationnelle unique pour l'acrylamide médiée par son métabolite actif, le glycidamide. En fait, le profil de la signature mutationnelle a récapitulé les types de mutations attendus en fonction de l'analyse des adduits d'ADN.

En outre, une analyse intégrée utilisant un modèles cellulaire, les Hupki MEF, et tumoral, les tumeurs rénales des rats exposés à l'OTA, suggère un manque de mutagénicité directe pour l'OTA avec une contribution potentielle d'un mode d'action lié à la production des radicaux libres observée dans la signature mutationnelle d'OTA dans les MEF.

Cette stratégie expérimentale simple et puissante peut faciliter l'interprétation des empreintes de mutations identifiées dans les tumeurs humaines, élucider l'étiologie du cancer et éventuellement soutenir la classification des cancérigènes par le CIRC en fournissant des preuves mécanistes.

Mots clés : Facteurs de risque de cancer, modèles d'expositions *in vitro*, tissues de tumeurs, séquençage du genome entier, spectres de mutations, signatures mutationnelles.

Résumé en anglais

Genome-wide modeling of mutation spectra of human cancer-risk agents using experimental systems

Cancer genomes harbour a mosaic of mutation patterns from which thirty mutational signatures have been identified, each attributable to a particular known or yet undetermined causal process. Deciphering the origins of these global mutational signatures in full could help identify the causes of human cancer, especially for about 40% of those signatures identified thus far that remain without a known etiological factor. Thus, well-controlled experimental exposure models can be used to assign particular mutational signatures to various mutagenic factors.

During the time frame of my PhD work, I characterized and employed innovative *in vitro* and *in vivo* models of carcinogen exposure, namely, primary Hupki MEF cells, HepaRG and lymphoblastoid cell lines as well as rodent tumors. The cytotoxic and genotoxic conditions for each tested exposure compound were established and DNA adduct formation was assessed in select cases. Following a pre-screen by *TP53* gene sequencing, genome-wide sequencing of immortalized Hupki MEF clones derived from exposure to acrylamide, glycidamide and ochratoxin A was performed, alongside whole genome sequencing of ochratoxin A induced rat renal tumors.

The results reveal a novel mutational signature of acrylamide mediated by its active metabolite, glycidamide, a pattern that can be explained by the parallel analysis of individual glycidamide-DNA adducts. In addition, an integrative mutation analysis using *in vitro* and *in vivo* models suggests a lack of direct mutagenicity for OTA and possible indirect effects ROS-mediated in MEF cells.

The presented robust experimental strategy can facilitate the interpretation of mutation fingerprints identified in human tumors, thereby elucidating cancer etiology, elucidating the relationship between mutagenesis and carcinogenesis and ultimately providing mechanistic evidence for IARC's carcinogen classification.

Key words: Cancer-risk factors, *in vitro* exposure models, FFPE tissues, genome-wide sequencing, mutation spectra, mutational signature.

ACKNOWLEDGMENTS

I would like to express my special appreciation and thanks to my thesis supervisor and advisor Dr. Jiri Zavadil. He has been a great mentor and teacher to me. I thank him for allowing me to grow as a research scientist. I would like to add to this special acknowledgement my co-supervisors Dr. Michael Korenjak and Dr. James McKay. I thank them all for their guidance at various levels throughout the different stages of my doctoral thesis, and for sustaining an environment where knowledge and expertise have been shared in frequently enriching and inspiring discourses.

I would also like to show gratitude to Dr. Zdenko Herceg for his wisdom, guidance and for lifting me up in difficult moments. I appreciate his support and thank him for his valuable encouragement.

My further thanks go to the key Molecular Mechanisms and Biomarker's collaborators who provided expertise and advice for parts of my project, namely to Dr. Dinesh Kumar Barupal for his help with the compounds prioritization method, Dr. Silvia Balbo and Dr. Andrea Carra for their eager contribution to screen for OTA-DNA adducts, Dr. Frederick Beland and Dr. Mona Churchwell for the establishment of GA-DNA adducts in the MEF cell lines, , Dr. Ron Herbert from the US NTP for providing access to NTP bioassays material, Dr. Steve Rozen and Dr. Arnoud Boot for assisting with the whole-genome data analysis and Dr. Ludmil B. Alexandrov for offering great explanations and advices at different stage of my PhD work.

I thank the members of my thesis committee for following on my PhD work throughout these three years as well as for their guidance, advices and discussions, namely, Prof. David Phillips and Dr. Virginie Petrilli.

I would like to thank Prof. Agnes Bernet, Prof. Béatrice Fervers, Prof. David Phillips, Prof. Francois Sichel and Prof. Robert Barouki for their willingness and graceful acceptance to serve as jury members.

I would like to thank again Prof. Francois Sichel as well as Prof. Ahmad Besaratinia for reviewing my thesis manuscript, for their time and for their comments.

I would like to also express my thanks to all the members of the IARC MCA section and other IARC colleagues, especially to Dr. Akram Ghantous for his professional guidance.

I thank my parents for their support, and a special thanks goes to my mom. She is a fighter and she has continuously supported us and encouraged us to always aim high and make our dreams come true. I owe her all the respect and the love. Thanks to my dad who worked

hard for our education. Thanks to my brothers and sisters for their support and for all the lovely memories we have shared during these distant years. Lastly, thanks to my friends who have been enormously supportive during my PhD work and for being there.

List of abbreviations

3-NBA	3-nitrobenzanthrone
AA	aristolochic acid
ACR	acrylamide
AFB1	aflatoxin B1
AHRR	Aryl-Hydrocarbon Receptor Repressor
APOBEC	Apolipoprotein B mRNA Editing Enzyme, Catalytic Polypeptide-Like
BaP	benzo[*a*]pyrene
BBCE	Barrier Bypass-Clonal Expansion
BBN	N-butyl-N(4-hydroxybutil)nitrosamine
BEN	Balkan Endemic Nephropathy
B-gal	β-Galactosidase
bp	base pair
BSA	Bovine Serum Albumin
C*	crisis
CB	crisis bypass
CE	clonal expansion
CK-19	cytokeratin-19
COSMIC	Catalogue of Somatic Mutation In Cancer
Cr(VI)	chromium (VI)
CYP40	cytochrome P450
DMBA	7,12-dimethylbenz[a]anthracene
DMSO	Dimethylsulfoxide
dR	deoxyribose
EBV	Epstein Barr Virus
EDTA	Ethylenediaminetetraacetic acid
ESLC	Evidence Suggesting Lack of Carcinogenicity

FACS	Fluorescence-Activated Cell Sorting
FASAY	Functional Analysis of Separated Allele In Yeast
FCS	Fetal Bovine Serum
FDR	False Discovery Rate
FF	Fresh-Frozen
FFPE	Formalin-Fixed Paraffin-Embedded
FSC	Forward-Scattered Light
fwd	forward
GA	glycidamide
GDC	Genomic Data Commons
γH2Ax	Phosphorylated H2Ax
GIV	Global Imbalance Variation
GLYPH	Glyphosate
H&E	Hematoxylin And Eosin
HBV	Hepatitis B Virus
HCC	Hepatocellular Carcinoma
HKG	Housekeeping Genes
HK-2	Human Kidney Cell Line
HMEC	Human Mammary Epithelial Cells
HPV	Human Papilloma Virus
IARC	International Agency of Research on Cancer
ICGC	International Cancer Genome Consortium
JIA	IARC Junior Investigator Award
IMO	IARC Monographs Section
indel	Insertion-Deletion
iPS	induced Pluripotent Stem cells
LC/MS	Liquid Chromatography/Mass Spectrometry
LC/MS-MS	Liquid Chromatography / Tandem Mass Spectrometry

LCL	Lymphoblastoid Cell Line
MCA	Mechanisms of Carcinogenesis
MEF	Mouse Embryonic Fibroblast
MMB	Molecular Mechanism and Biomarkers Group
MMR	DNA Mismatch Repair
MNNG	1-methyl-3-nitro-1-nitrosoguanidine
MNU	N-methyl-N-nitrosourea
MS	mass spectrometry
MTT	3-(4,5-Dimethylthiazol-2-Yl)-2,5-Diphenyltetrazolium Bromide
MutSpec	Mutation Spectra
N1-GA-Ade	N1-(2-carbamoyl-2-hydroxyethyl) Adenine
N3-GA-Ade	N3-(2-carbamoyl-2-hydroxyethyl) Adenine
N7-GA-Gua	N7-(2-carbamoyl-2-hydroxyethyl) Guanine
NCI	National Cancer Institute
N-GLYPH	N-nitroso-glyphosate
NGS	next-generation sequencing
NHGRI	National Human Genome Research Institute
NMF	Non-Negative Matrix Factorization
NL	neutral loss
NSLC	non-small-cell lung cancer
NTP	National Toxicology Program
OTA	ochratoxin A
PCA	Principal Component Analysis
PCAWG	Pan-Cancer Analysis of Whole Genomes
PCR	Polymerase Chain Reaction
PEG	Polyethylene Glycol
PhIP	2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine
PT	partial trypsinization

qRT-PCR	Quantitative Real Time-Polymerase Chain Reaction
rev	reverse
ROS	Reactive Oxygen Species
S*	senescence
SBI	Senescence Bypass/ Immortalization
SBS	single base substitution
SD	Standard Deviation
SNP	single nucleotide polymorphism
Spont	Spontaneous
SSC	Side-Scattered Light
TBP	TATA Box Binding Protein
TCGA	The Cancer Genome Atlas
TRC	Toronto Research Chemicals
TSG	tumor suppressor gene
UTUC	upper tract urothelial carcinoma
UV	Ultraviolet Light
WES	whole-exome sequencing
WGS	whole-genome sequencing
WHO	World Health Organization

Table of Contents

INTRODUCTION

1. Cancer prevalence, incidence and mortality:

Cancer is one of the leading causes of mortality worldwide, causing one of six deaths globally (Ferlay et al., 2015). As stated by the World Health Organization (WHO), 70% of new cancer cases will arise in the next two decades. Public health concerns have grown immensely trying to understand the biology and the burden of cancer on society.

According to the GLOBOCAN project (Global Cancer Reports 2014; Ferlay et al., 2015), prevalence estimates for 2012 indicate that for all cancers combined (excluding non-melanoma skin cancer) there were 32.6 million people (older than 15 years) alive who had been diagnosed with cancer in the previous five years. 48% of the 5-year prevalent cancer cases occurred in the less developed world, and 52% occurred in the high-income countries of North America and Western Europe, together with Japan, the Republic of Korea, Australia, and New Zealand. Figure 1 represents the 5-year prevalence of new cancer cases.

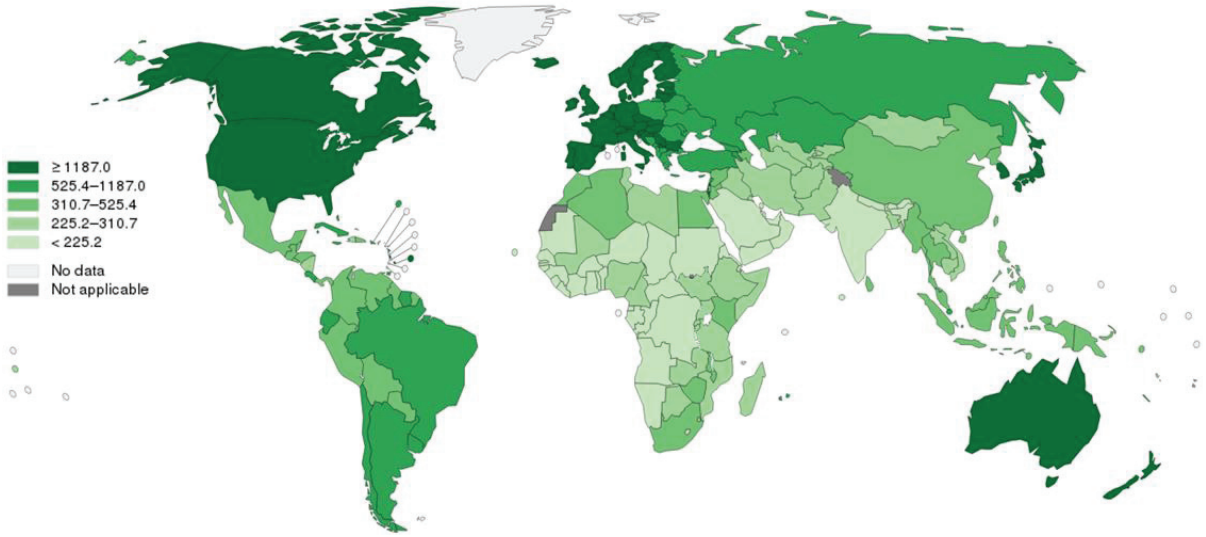


Figure 1: **Estimated numbers of prevalence cases (5-year)**, in both sexes, from all cancers excluding non-melanoma skin cancer, worlwide in 2012. Data source: GLOBOCAN 2012; Graph production: IARC, World Health Organization (http://gco.iarc.fr/today).

Despite the higher incidence rate of cancer in the developed world largely due to tobacco smoking, high overall calorie intake coupled with the sedentary lifestyle in the rich populations, the level of mortality in the less developed world is remarkably higher than in the rich countries, accounting for 64.9% and 35.1%, respectively. In fact, regions such as Africa,

Asia, and Central and South America represent about 70% of the cancer deaths worldwide (Ferlay et al., 2015). This increased death rate is caused by multiple challenges facing the less developed countries. Attempts to control cancer development are less effective in the less developed countries given the remarkable disparities in resources compared to the rich countries. Different factors contribute to a vicious cycle wherein the poor world is trapped including poverty and low education level, limited government funds for health care expenditure and lack of trained professionals and managing cancer (see Figure A.1, meaning Appendix A figure A.1). Escaping from this cycle would require improvements in health care as well as in the socioeconomic status of the countries (Internal Network for Cancer Treatment and Prevention, INCTR). The most common causes of cancer deaths in the world are lung, breast, colorectal, prostate, stomach, and liver cancers (Figure A.2). Worldwide distribution of particular cancer types indicates marked differences between populations, mostly attributed to discrepancies in risk-factors exposure. The substantial burden of cancer on societies in low- and high-income countries is a major driving force for continued research to better understand the causes of cancer, and hence the development of therapeutic and preventive measures (Ferlay et al., 2015).

2. Cancer biology:

Cancer is a generic term reflecting neoplasms that can affect different organs and tissues of the body. The complexity of this disease has been extensively studied in the past decades generating a rich knowledge on the dynamic changes that drive a normal cell to become malignant (Hanahan and Weinberg, 2000). One defining feature of cancer is the abnormal growth of cells beyond their boundaries, the ability to invade adjacent tissues and the blood circulation leading to the dispersal of the cells into different organs, a process termed metastasis. Tumorigenesis is defined by a number of molecular and cellular hallmarks driving cell transformation (Hanahan and Weinberg, 2000, 2011). This process follows the Darwinian evolution by which a cell is subjected to a succession of genetic or epigenetic changes that confer a growth advantage and lead to the progressive conversion of a normal cell into a malignant mass (Stratton et al., 2009).

2.1. Hallmarks of cancer:

Throughout cancer development, cells accumulate hallmark characteristics enabling their transformation into a malignant entity with the ability to proliferate indefinitely. The hallmarks of cancer development have been described and revised in (Hanahan and Weinberg, 2000, 2011), resulting in a total of ten biological capabilities, such as sustained proliferative signaling, resistance to cell death, replicative immortality, invasion and metastasis or genome instability. The ten hallmarks (summarized in Figure 2) are acquired differently and at various times across different cancer types and individuals. Among these hallmarks, the ability to invade the blood circulation and adjacent tissues, leading to the dispersal of cancer cells to different organs, a process termed metastasis, is the main cause of cancer deaths worldwide. Induction of genome instability, for example, is brought about by mutations affecting pathways that monitor genomic integrity, such as *TP53* (“the guardian of the genome”), which results in the accumulation of random mutations and structural rearrangements that can subsequently orchestrate other hallmark capabilities.

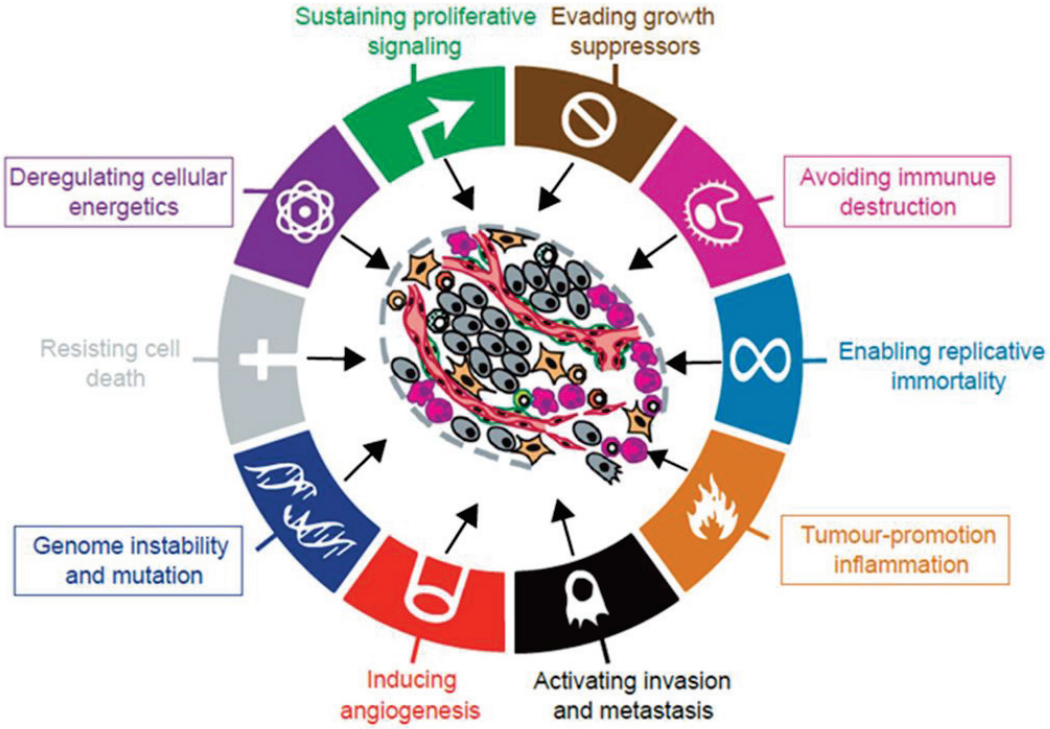


Figure 2: The hallmark of cancer: Hallmarks of cancer development. Taken from (Hanahan and Weinberg, 2000) and (Hanahan and Weinberg, 2011).

2.2. Cancer genome:

In spite of the more recent emergence of epigenetic changes during tumorigenesis, cancer is primarily considered a genetic disease, causing complex abnormalities in the genomes of cancer cells (Nowell, 2002; Vogelstein and Kinzler, 1993). In analogy to Darwinian evolution, cells continuously acquire stochastic, heritable genomic alterations, which through natural selection can give rise to the phenotypic diversity and heterogeneity of tumors. Some of the acquired mutations can be deleterious and others can provide a growth advantage to the cells, which ultimately allows cancer cells to survive, proliferate, invade and metastasize (Stratton et al., 2009).

2.2.1. Epigenetic changes in a cancer genome:

In recent years, evidence has emerged linking epigenetic changes to environmental factors and human malignancies (Feil and Fraga, 2012). Cancer genomes frequently undergo epigenetic changes, which follow the Darwinian natural selection process and favor the growth of cells with characteristically altered chromatin structure and deregulated gene expression (Stratton, 2013). These changes are brought about by epigenetic modifier genes, such as DNA methyltransferases/demethylases, histone modifiers or ATP-dependent

chromatin remodelers that are frequently mutated in human cancer (Feinberg et al., 2016). The results of epigenetic deregulation can range from the misexpression of individual oncogenes or tumor suppressor genes to large-scale chromatin structure alterations and genomic instability.

Well-established cancer-risk agents and lifestyle factors have been studied in terms of epigenome deregulation, improving the understanding of their long-lasting effects on cancer outcome. Tobacco smoking, diet, infections, inflammation and age are known to affect epigenetic states and can play a role in the early onset of cancers through different mechanisms. Smoking, which is the strongest exposure factor causing lung cancer, harbors an epigenetic signature characterized by consistent methylation changes in the Aryl-hydrocarbon receptor repressor (*AHRR*) gene. Age is the strongest demographic risk factor for cancer and, interestingly, DNA methylation profiles of chronological age established an "epigenetic clock" that can be affected by different external and endogenous factors (Horvath, 2013).

Progress in epigenetic research can open the door to a new era where epigenetic biomarkers can serve as surrogate for diagnostics and risk stratification of cancer in tissues and can provide evidence on the interactive role of epigenetic deregulation in the roadmap between environmental exposures and cancer (Herceg et al. 2017).

2.2.2. Somatic mutations in a cancer genome:

Throughout the lifetime of a cancer patient, mutations are accumulating in the genome. These acquired mutations are termed somatic mutations, differentiating them from germline mutations which are inherited changes linked to familial predisposition (Stratton et al., 2009).

Somatic mutations in cancer cells can encompass different structural classes of DNA sequence changes (Figure 3). They include:

1. Point mutations:

    a. Single base substitutions (SBS) of one base to another. Depending on the base change and position, these can have varying effects. Silent or synonymous mutations do not alter the protein sequence. Alternatively, they can lead to a truncated or inactive protein when the SBS introduces a stop codon (missense mutation) or induces an amino-acid change (non-synonymous mutation), respectively. Finally, mutations can fall in gene regulatory regions disrupting the transcriptional activity of the gene. For example, the first cancer-causing gene change was discovered in 1982 when

researchers identified a G>T substitution in codon 12 of the HRAS gene causing a glycine to valine substitution (Reddy et al., 1982; Tabin et al., 1982).

b. Small insertions and deletions (Indels) that result from loss or gain of nucleotide base pairs can produce abnormal protein sequences, thus affecting their function (Jego et al., 1993).

2. Chromosomal rearrangements, in which DNA segments break off and re-attach at a different genomic location, within the same chromosome or on a different chromosome, termed intra- and interchromosomal rearrangements, respectively (Figure 3). This can lead to gene disruption, the fusion of two genes or the translocation of a gene adjacent to regulatory elements, resulting in abnormal gene expression. Translocations are mostly operative in leukemias, lymphomas and sarcomas (Nowell et al., 1960; Rowley, 1973). More recently, rearranged cancer fusion genes were discovered in half of prostate cancer patients (Tomlins et al., 2005) as well as in non-small-cell lung cancer (NSCLC) cases (Soda et al., 2007).

3. Copy number variations:

   a. Copy number increases, from two copies in a diploid genome, to several hundreds of copies. These are referred to as gene amplifications, which are a common mechanism for the activation of oncogenes (Alitalo, 1984), by increasing mRNA levels and thus gene expression.

   b. Copy number reductions resulting from large deletions. This may induce the complete absence of a DNA segment, resulting in the loss of an associated gene, and most commonly observed as a mutational mechanism for TSG (Harris et al., 1991).

4. Insertion of new DNA sequence, originating from exogenous sources, notably viruses such as human papilloma virus (HPV), Epstein Barr virus (EBV), hepatitis B virus (HBV). These viruses have been unambiguously implicated in the development of different types of cancer (Talbot and Crawford, 2004).

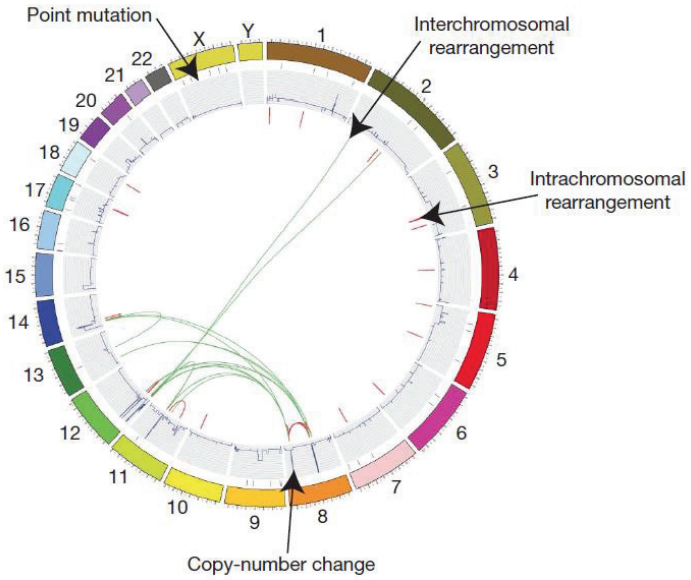


Figure 3: Visualisation of the different types of genomic alterations present in cancer genome. Circos plot are used to depict chromosomes, point mutations, copy number and rearrangements from the outer circle to the inner circle. Each alteration is represented relative to its position on chromosomes. Adopted from (Stratton et al., 2009).

2.2.3. Driver and passenger mutations

Mutations accumulate progressively during the lifespan of an individual. Nonetheless, not all of these mutations result in tumor development. To reflect this concept, mutations are classified according to their consequences on cancer development and referred to as driver or passenger mutations (Figure 4). A driver mutation is a mutation that is implicated in oncogenesis; it has been positively selected in the microenvironment of the tumor tissue by conferring a growth advantage to the tumor. Such mutations are carried along in the clonal growth of a cancer and can help maintain and promote its growth. A passenger mutation is a mutation that does not contribute to cancer development and that has not been selected for during the evolution of the cancer (Vogelstein et al., 2013). Passenger mutations do not have functional consequences on tumor growth. By exploiting the functional contribution of driver genes, such as oncogenes and TSG, to tumor development it is possible to define and distinguish the clustered nature of driver mutations, which occur in a small number of genes, from passenger mutations, which are randomly distributed throughout the genome. Nevertheless, this task remains challenging as some mutation processes target specific genomic regions, generating clusters of passenger mutations that can be mistaken for driver alterations (Stratton et al., 2009). In addition, identification of cancer driver genes hinges in part on mutation analysis of the most commonly mutated genes within a particular type of cancer (Cancer Genome Atlas Research Network, 2008). This suggests that there are more

driver genes still to be identified, including drivers infrequently mutated across cancers or driver gene rearrangements that demand advanced genomic annotations for identification (Greenman et al., 2007). In order to identify a driver gene that is mutated in more than 5% of tumors of the same type with sufficient confidence, sequencing of hundreds of cases will be required.

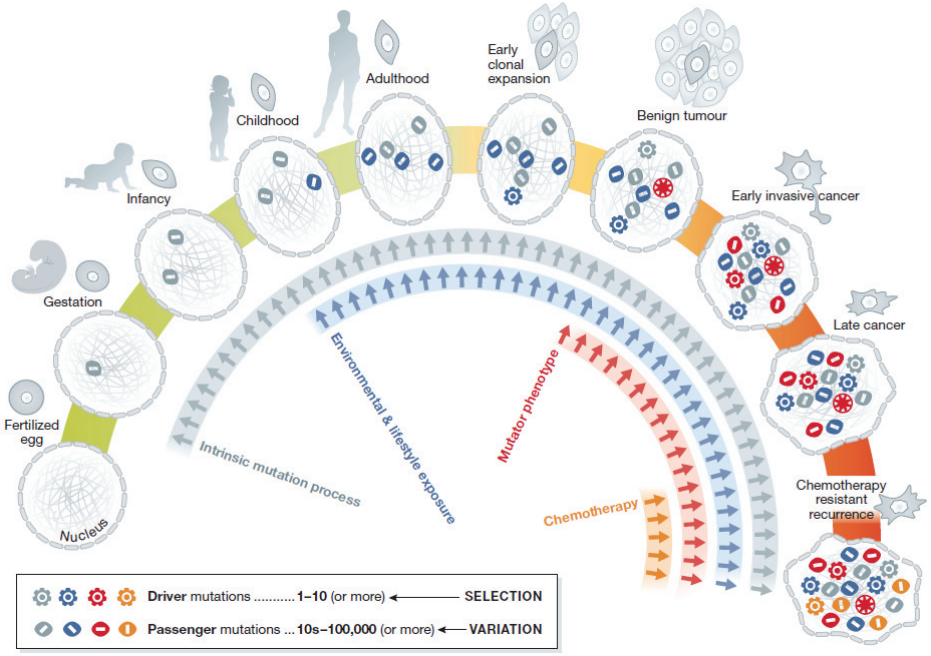


Figure 4: Cellular lineage of cancer cell. Coloured symbols represent the progressive accumulation of somatic mutations between the fertilised egg and a fully malignant cancer cell. Embryogenesis represents a sensitive stage where embryos are prone to intrinsic mutation processes. After birth and during childhood, more mutations start accumulating due to environmental or lifestyle exposure. Chronic exposure to some cancer-risk agents can promote a mutator phenotype leading to an over-proliferation of cells and finally provoking cancer. Moreover, therapeutic approaches to eradicate cancer can cause the development of mosaic cells resistant to chemotherapy and thus the recurrence of the disease. Adopted from (Stratton, 2013).

3. Causes of mutations in human cancer:

Regarded mostly as a genetic disease, research has been heavily focused on the identification of cancer driver genes, as they represent attractive targets for therapeutic drugs. With the announcement of the reference human genome sequence in 2002 and the development of NGS, human cancer genomes have been sequenced at an unprecedented rate, uncovering the identity of genes operative in cancer development and accounting for more than 1% of all human genes (Futreal et al., 2004). Interestingly, while examining the cancer data for cancer genes in a myriad of sequence changes, researchers noticed that the mutational patterns differed by frequency and mutations type across the cancer types. This was suggestive that each cancer type can be a consequence of distinct mutagenic processes.

3.1. Intrinsic versus extrinsic exposures leading to cancer mutations

Cancer genomes represent a historical archive of the different mutagens that acted on the organism and were ultimately responsible for the development of cancer. Mutations can be prompted by endogenous factors (intrinsic mutagens), such as the inherent genetic instability and defects in the DNA repair machinery, or exogenous factors (extrinsic mutagens), such as dietary compounds, smoking, viruses, occupational and environmental carcinogens (Nowell, 2002; Vogelstein and Kinzler, 1993).

Lately, the identification of the source of mutations found in human cancer genomes has become a subject of controversy.

It is widely accepted that mutations can be introduced upon exposure to carcinogens or via inherited predisposition to cancer. However, differences in incidence rate across cancer types have stimulated a discussion on whether stochastic DNA changes, due to variations in stem cell divisions in different organs, could explain this observation (Tomasetti and Vogelstein, 2015; Tomasetti et al., 2017a, 2017b). Quantitative correlation analysis of the lifetime risk of developing a certain cancer with the number of stem cells divisions within the same organ implied that replication-related mutations are required to drive neoplastic development (Tomasetti and Vogelstein, 2015).

An extended analysis based on a novel mathematical model and DNA sequencing and epidemiological data from 69 countries, representing 4.8 billion people, replicated the previous findings (Tomasetti et al., 2017b). This model computed 29% of driver gene mutations to be linked to environmental factors, whereas 66% were attributed to replicative errors.

However, it is difficult to completely separate replication-related mutations from exogenous factors. Consumption of very hot beverages for example, which has been linked to esophageal carcinogenesis, induces severe damage to the cellular lining of the esophagus, which in turn will trigger stem cells located in the deep layers to divide in order to replace the damaged cells. Therefore, stem cells divisions, caused by an exogenous factor, can introduce replication-related mutations (López-Lázaro). Moreover, the endogenous production of reactive oxygen species (ROS), which can trigger replicative mutations, has been considerably associated with several environmental compounds that act indirectly on the DNA, such as pesticides, mycotoxins and heavy metals (Frenkel, 1992).

Numerous epidemiological studies emphasize the contribution of the environment to the cancer burden observed in particular populations (Wild et al., 2015). For instance, hepatocellular carcinoma shows a high incidence rate in regions with a high risk of exposure to aflatoxin B_1 (AFB1) compared to other regions where AFB1 exposure is minimal (Wild et al., 1990). In addition, comparison of cancer incidence of Japanese subjects residing in Hawaii versus those living in Japan, especially from Okinawa, revealed a dramatic decrease in cancers of the mouth, pharynx and esophagus in all Japanese migrants, suggesting that they have escaped exposure to an environmental cancer risk factor peculiar to Okinawa region (Stemmermann et al., 1991). Furthermore, Wu et al. (Wu et al., 2016) provided extensive discussion on the causes of cancer mutations being attributed more significantly to environmental factors by employing different approaches. This resulted in an estimated 70-90% contribution of extrinsic factors to cancer development, with the rest being due to intrinsic factors.

Uncovering the causes of cancer allows cancer prevention measures, seeking to reduce or remove the exposure factors (Brennan and Wild, 2015; Colditz et al., 2012). Current estimates indicate that the majority of the global cancer risk could be preventable (Ferlay et al., 2015). Taking the ongoing discussion regarding the contribution of replication errors into account, a smaller proportion of cancers would be amenable to a reduction of environmental exposures, while the majority would require cancer prevention measures based on early detection and intervention.

3.2. Approaches to identify the sources of the somatic mutations:

Human cancer genomes harbour complex mutation patterns reflecting tumor heterogeneity that can stem from exposures to multiple carcinogenic agents (Greenman et al., 2007). Some mutagenic carcinogens leave specific SBS mutation imprints on the DNA, exemplified by tobacco smoke carcinogens, ultraviolet light (UV) and AFB1, causing characteristic

mutation patterns as seen in lung (G>T), skin (C>T) and liver (G>T) cancers, respectively (Hollstein et al., 1991; Pleasance et al., 2010a, 2010b). Further refined classification of these SBS mutations can be applied by taking the nucleotide bases flanking the mutated base on 5' and 3' into consideration. Thus, it became possible to discriminate between the mutation patterns of G>T transitions observed in lung and liver cancers, and the analysis of human cancer mutation spectra now offers the possibility to study cancer etiology (Hollstein et al., 2017).

3.2.1. Single-gene approaches

Previously, mutagenicity and genotoxicity evaluation of compounds relied on simple assays employing prokaryotic systems, such as the Ames test, and assays that are laborious, such as the comet and micronucleus assays. However, these assays do not provide insights regarding the specific base changes and the sequence context (Zhivagui et al., 2016).

Using single-gene sequencing experimental models as well as primary human tumors provides an alternative to study the mutagenic processes associated with specific carcinogenic exposures. The experimental systems used for this purpose depend either on a phenotypic selection method (e.g. bacterial reporter genes) or on genes that are frequently mutated in human cancers.

3.2.1.1. Reporter gene assays

Commonly utilized *in vitro* reporter genes rely on endogenous genes, e.g. HPRT, DHFR and TK, to convert certain media supplements to toxic metabolites implying the occurrence of genetic changes in the encoding genes. In contrast, animal *in vivo* model systems include the genomic integration of a transgene consisting of a reporter gene (such as *lacI*, *lacZ*, *gpt*, *gpa*, *hprt*, *aprt*, *supF* and *cII* genes) and a viral shuttle vector. After exposure, the transgene is packaged into phage particles ensuring the efficient delivery of the target gene into a bacterial host. Mutation detection is examined using chromogenic or viability selection (Boverhof et al., 2011). Reporter gene assay allowed the assessment of the mutagenicity of a number of carcinogens, for instance, the heterocyclic amine 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP), a common dietary carcinogen in cooked meat, which was associated with an increased rate of G>T transversions. Other examples include the dietary carcinogen acrylamide (A>T and G>C), AFB1 (G>T) and the chemotherapeutic agent 8-methoxypsoralen (T>A) (Zhivagui et al., 2016).

3.2.1.2. Single-gene mutation profiles

Sequencing of cancer genes facilitates the identification of driver mutations in a cancer type. This approach provided the first evidence regarding the molecular mechanisms by which environmental carcinogens leave characteristic imprints on the DNA. Examples of the most frequently mutated genes in human tumors include *TP53*, *KRAS* and *BRAF* genes. Single cancer gene sequencing in skin and lung tumors identified mutation patterns characteristic of exposures to UV-light and benzo[*a*]pyrene (BaP), respectively (Brash, 2015; Brash et al., 1991; Pfeifer et al., 2002). UV-light induces C>T transitions at dipyrimidines in skin cancers, and tobacco smoke prompts G>T transversions in lung tumors (Figure 5) (Olivier et al., 2010). These tumor-associated mutation patterns are in agreement with results from controlled experimental exposure studies of UV-light and BaP (Denissenko et al., 1996; Miller, 1982). Indeed, sequencing of the *TP53* gene in cancer patients with different exposures history (exposed vs. non-exposed) strengthened the link between environmental factors and cancer (Hollstein et al., 1991). Lung tumors from smokers display a predominant G>T mutation pattern which is not evident in lung tumors from non-smokers, and the G>T imprint correlates with the level of tobacco consumption (Pfeifer et al., 2002).

Human exposure to the plant carcinogen aristolochic acid (AA) has been linked to the endemic Balkan nephropathy and upper tract urothelial carcinoma (UTUC) (Grollman et al., 2007). Indeed, *TP53* mutation screening of these cancers identified a pronounced A>T mutation fingerprint associated with the unique mutation pattern of AA observed in the laboratory (Hollstein et al., 2013; Nedelko et al., 2009).

Finally, the *TP53* gene exhibits a unique mutation profile characterized by predominant G>T transversions in hepatocellular carcinoma (HCC) cases from regions where AFB1 exposure is prevalent (Bressac et al., 1991; Montesano et al., 1997; Wogan, 1992). Liver cancers from other populations where AFB1 exposure is minimal and other risk factors prevail exhibit distinct *TP53* mutation fingerprints (Montesano et al., 1997).

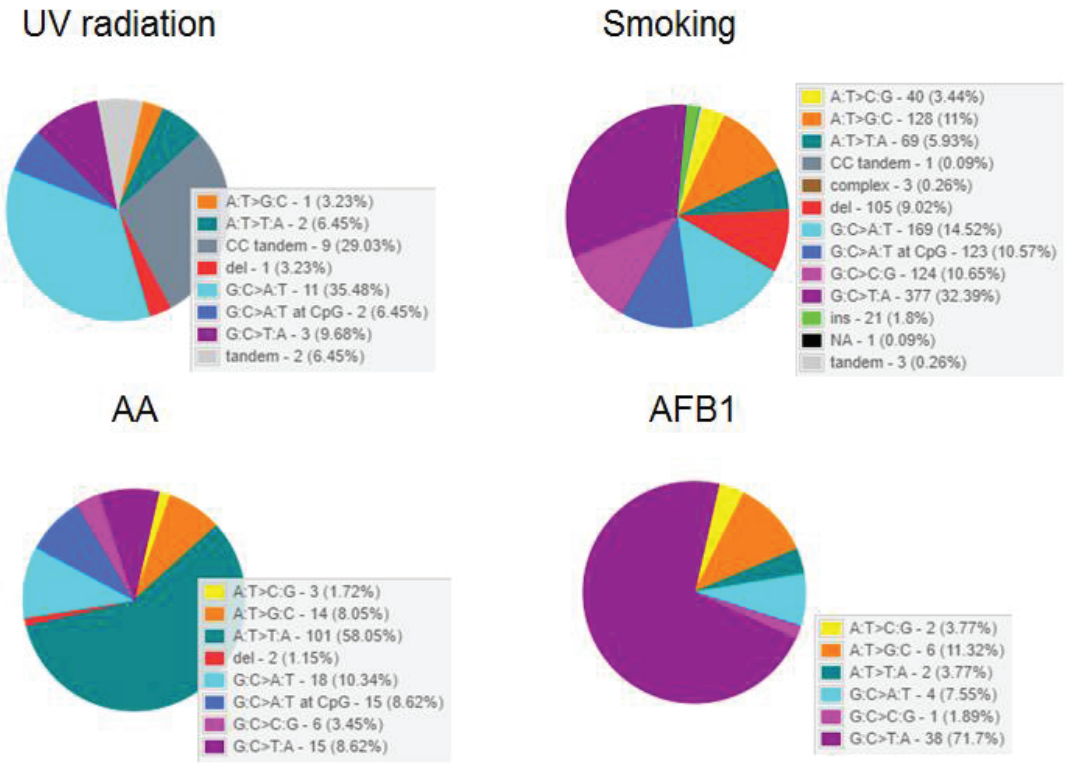


Figure 5: **Data extracted from IARC *TP53* database** (http://p53.iarc.fr/TP53SomaticMutations.aspx). Pie charts representation of the proportion of *TP53* SBS changes observed in human skin, lung, kidney and liver cancers linked to the external factors, UV, smoking, AA and AFB1 respectively.

Different *in vivo* and *in* vitro experimental systems contributed to the extraction of *TP53* mutation patterns, representing “rudimentary signatures” of mutagens and their association to specific cancer types. These models include a genetically engineered mouse system, harboring the human *TP53* gene, from which embryonic fibroblasts were derived. Hupki (Human *TP53* knock-in) mice were exposed to UVB inducing characteristic *TP53* gene mutations similar to those predominantly observed in human skin cancer (C>T) (Luo et al., 2001a, 2001b). Moreover, Hupki mouse embryonic fibroblasts (Hupki MEFs) were exposed to a number of carcinogens elucidating analogy between the Sanger sequenced *TP53* genes from the *in vitro* assay and human tumors associated with the same exposure (Zhivagui et al., 2016).

Yeast systems were also exploited for *TP53* mutagenesis using a strain transfected with an expression vector harboring human *TP53* cDNA that had been UV-irradiated *in vitro*. The results revealed CC>TT transversions which is in line with the observations from human skin cancer data (Inga et al., 1998).

Finally, normal human fibroblasts were treated with known carcinogens, such as BaP, AFB1 and acetaldehyde and mutation patterns of *TP53* were evaluated by functional analysis of separated allele in yeast (FASAY) (Paget et al., 2012).

Notably, experimental identification of carcinogen-specific mutation patterns demonstrates effectiveness in convergence of the mutation data with the epidemiological studies for the establishment of causal associations between environmental exposures and human cancers (Hollstein et al., 2013; Zhivagui et al., 2016).

Despite their significant contribution to understanding the sources of somatic mutations in cancer, single gene sequencing studies harbor major limitations: first, *TP53* mutation, which confers a selective growth advantage, may not always occur or be selected for during cell transformation; second, many samples from a specific cancer type are needed to accumulate enough alterations to extract a specific mutation profile (Hollstein et al., 2017; Zhivagui et al., 2016). Fortunately, advances in NGS and bioinformatics analysis can help address these challenges and allow efficient testing of hypotheses regarding putative cancer-risk factors.

3.2.2. Massively parallel sequencing and computational analysis

Massively parallel sequencing has revolutionized many aspects of biology, including mutation research, due to high speed sequencing capacities and the reduction in the overall sequencing cost. NGS enables the extraction of mutation patterns from individual tumor samples, overcoming the need to pool many individuals for single-gene mutation profiling.

The mutation spectra observed in cancer genomes are the consequence of exposure to multiple risk factors during the individuals' lifetime. In order to establish the contribution of individual exposures to the final mutation spectra, a simple mathematical model based on non-negative matrix factorization (NMF) can be used to deconvolute the spectra into mutational signatures characteristic for cancer-risk factors (Figure 6). The NMF algorithm was first used by Alexandrov and colleagues to achieve an elegant reconstruction of the original sources of mutations, using mutation data from 7042 cancer patients in 30 different cancer types to extract 21 distinct mutational signatures (Alexandrov et al., 2013a) that were later expanded to 30 and are available on the COSMIC website (source: http://cancer.sanger.ac.uk/cosmic/signatures) (Alexandrov et al., 2013b).

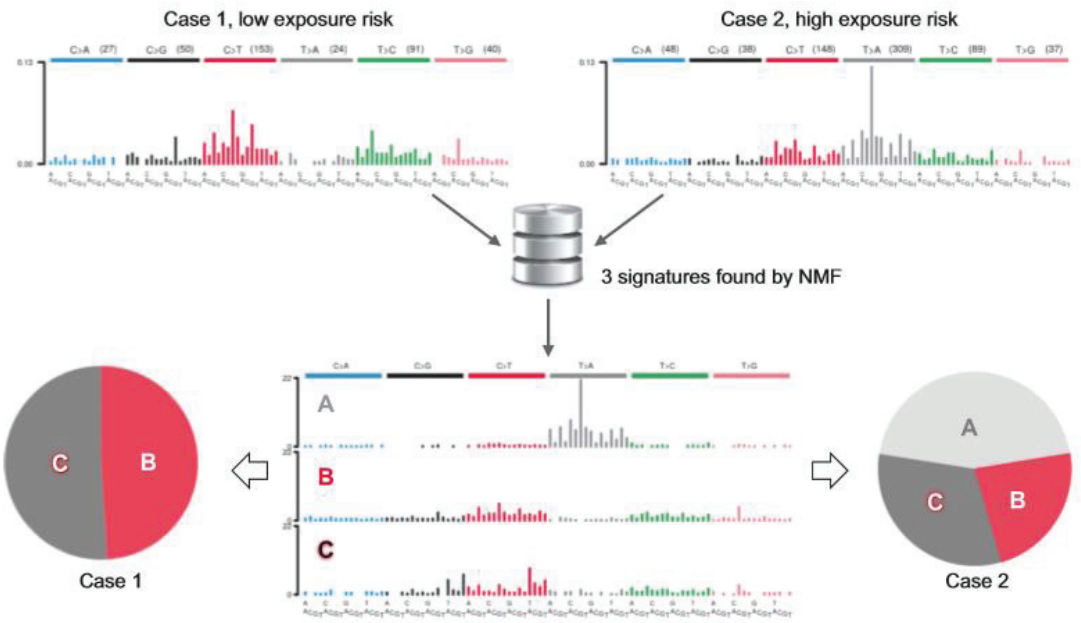


Figure 6: An example of using the NMF tool to tease apart different mutational signatures from tumor sequencing data. In the first case with low exposure risk, the only observed mutational signatures are C and B, whereas in case 2 with high exposure risk a new mutational signature A is extracted, and by the mean of epidemiological studies and patient exposure history the mutagenic factor can be identified. The pie charts represent the relative contribution of each signature to the overall mutation load in each tumor. These two cases illustrate the difference in cancer etiology (Hollstein et al., 2017).

Currently, modeling of mutation spectra and mutational signatures in human cancers addresses the SBS data only. 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. SBS are conventionally reported as the pyrimidine of the mutated Watson-Crick base pair. Thereby, a C:G>A:T substitution will refer to both C>A and G>T substitutions. The profile includes the trinucleotide sequence context of each mutated base (with the mutated base in the centre) on the x-axis, generating 96 possible SBS mutation types. The y-axis represents the proportion of each mutation with respect to the overall SBS counts (see Figure 6). Moreover, an additional feature can be attributed to a mutational signature taking in consideration the proportion of the mutations generated on the transcribed and the non-transcribed DNA strand, referred to as transcription strand bias. A ratio between the mutation counts on the non-transcribed (numerator) and the transcribed strand (denominator) greater than 1 denotes the activation of transcription-related DNA repair machinery, reducing the number of mutations in the denominator. Reversely, a replication-related strand bias can occur when the mutation counts are relatively increased on the transcribed strand compared to the non-transcribed strand.

Differences in mutation patterns in different cancer types are immediately evident. Small cell lung carcinoma, for example, displays a predominant C:G>A:T pattern with transcription strand bias (signature 4), related to tobacco smoke carcinogens. Upper tract urothelial carcinoma shows a unique mutational signature characterized by T:A>A:T in 5'-CAG-3' sequence context (signature 22), ascribed to AA exposure, whereas melanoma harbors a mutational signature characterized by predominant C:G>T:A at dipyrimidine nucleotide (signature 7) attributed to UV-light exposure (Alexandrov et al., 2013b). Importantly, mutational signatures with a strong link to a specific cancer type and its main etiological factor replicate observations from the single-gene *TP53* sequencing approach.

As tumor mutation spectra are a composite of superimposed mutational signatures left by various mutagenic insults, seven of the thirty known mutational signatures, for example, were identified in liver cancers. The known risk factors attributed to liver cancer occurrence are HBV and HCV, alcohol consumption and AFB1 exposure. Among the identified signatures, signature 16, characterized by T:A>G:C transitions at 5'-NAT-3' sites is observed exclusively in 90% of the liver tumors. Its etiology is nevertheless still unknown. Signature 24 was also uncovered in AFB1-exposed liver cancer spectra, characterized by a transcription asymmetry of C:G>A:T transversions (Schulze et al., 2015). This was elegantly confirmed using an integrated experimental analysis across human cell lines, animals and primary HCC tumors (Huang et al., 2017).

Surprisingly, endogenous mutagenic processes constitute almost half of the identified mutational signatures. Signature 1, attributed to the spontaneous deamination of 5'-methylcytosine, is seen in almost all tumors and is pronounced in some, such as in acute myeloid leukaemia. Together with signature 5, it has been linked to the clock-like cellular processes reflecting the chronological age of patients at diagnosis (Alexandrov et al., 2015). Furthermore, a number of signatures have been attributed to the disruption of processes regulating DNA homeostasis (Helleday et al., 2014), such as malfunction of DNA repair polymerases *eta* and *epsilon* (signatures 9 and 10, respectively), defective DNA mismatch repair (MMR) (signatures 6, 15 and 20), and BRCA1/2 mutations indicating a failure of DNA double strand repair by homologous recombination (signature 3). Notably, in more than half of the cancer types analysed to date, APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) deaminase mutational signatures (signatures 2 and 13) have been identified. APOBEC signatures are supported by extensive work in experimental model systems (Burns et al., 2013; Chan et al., 2015; Harris et al., 2002; Kazanov et al., 2015). These enzymes are implicated in virus restriction and suppression of retrotransposition (Smith et al., 2012). Signature 2 was found in cervical and in head and neck cancers, both of

which are related to HPV infection, implying the recurrent activation of APOBEC upon viral infection (Rebhandl et al., 2015; Warren et al., 2015).

Among the 30 distinct mutational signatures thus far identified, 40% remain with unknown etiology reflecting the need for controlled experimental mutation studies (Zhivagui et al., 2016).

4. Cancer genomics repositories:

Hand-in-hand with the remarkable technological advances in sequencing, a complete catalogue of all somatically acquired variants in cancer genomes was established in coordination with different data repositories, which keep the mutation data freely accessible and mineable. In addition to the somatic mutations, these repositories frequently contain additional omics data, such as gene expression data from RNA-seq, proteomics and epigenetic profiles, of the same cancer cases, and have been correlated with basic clinical features.

There are three large-scale data repositories that exist today:

- The Catalogue Of Somatic Mutations In Cancer (COSMIC) which is maintained by the Sanger Institute UK, using manually curated data from the scientific literature.
- The Cancer Genome Atlas (TCGA) which collaborates with the Pan-Cancer analysis project and is funded by the NIH.
- The International Cancer Genome Consortium (ICGC) which coordinates the generation of comprehensive catalogues of genomic abnormalities internationally.

4.1. The Catalogue Of Somatic Mutations in Cancer (COSMIC)

The Catalogue Of Somatic Mutations in Cancer (COSMIC) is easily accessible through its website (http://www.sanger.ac.uk/cosmic) (Forbes et al., 2011, 2017). COSMIC is the broadest database initiative of cancer mutation recurrence exploring targets and trends in the genome of human cancers worldwide. Release v78 (September 2016) includes 1,235,846 tumors samples, and 28,366 tumors genome-wide sequenced. Using 23,489 scientific publications for manual curation (accounting for 60% of COSMIC content), this high-resolution resource focuses on 186 key genes across all cancers. Molecular profiling of this large number of tumors allowed annotation of over 4 million coding mutations and one million copy number variants (Forbes et al., 2017). Around 30% of COSMIC content has been selected from consortia sources such as TCGA and ICGC.

4.2. The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) is accessible on http://cancergenome.nih.gov/. It is managed by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The primary aim of TCGA is to assimilate and interpret molecular profiles from DNA, RNA and protein sequencing as well as epigenetic patterns from clinical cases of different types of cancer. Data annotation is not confined to point mutations only, but also to

the characterization of copy number variations, DNA methylation, mRNA and miRNA expression and sequence and transcript splice variations (Weinstein et al., 2013). Cancer samples are chosen on the basis of poor prognosis and public health impact and on the availability of human tumor-matched normal tissue samples. In 2017, a launch of a new data portal, the Genomic Data Commons (GDC), will take place, which includes 29 different human tumor sites.

4.3. The International Cancer Genome Consortium (ICGC)

The International Cancer Genome Consortium (ICGC) is publically accessible on http://icgc.org/. The aim of the ICGC platform is to attain a comprehensive elucidation of cancer genome abnormalities by harnessing genomic, transcriptomic and epigenomic changes in 50 different tumor types denoting clinical and public health importance throughout the world (2010). To date, ICGC analyzed over 25,000 cancer genomes using different omics approaches and identified around 46 million somatic mutations in 21 different tumor types.

The Pan-Cancer Analysis of Whole Genomes (PCAWG) collaboration was established, covering about 700 researchers from around the world. PCAWG encompasses whole genome data of 2,834 donors with matched tumor/normal samples using ICGC data repositories. It aims for meaningful cross-tumor comparisons and standardized bioinformatic analyses using gold-standard, benchmarked, version-controlled algorithms (Campbell et al., 2017).

5. IARC Monographs on the evaluation and classification of carcinogenic risks to humans

The International Agency of research on Cancer (IARC) is the specialized cancer agency of the World Health Organization (WHO). It seeks to prompt broad collaboration in cancer research to uncover the causes of cancer so that prevention measures can be adopted in order to reduce the burden of this disease and related suffering. The IARC Monographs are expert evaluations of a compendium of carcinogenic chemicals and their causal effect on the human population.

5.1. Objective and scope

The objective of the IARC Monographs program is to review public scientific reports for evidence linking a wide range of agents to human cancer occurrence. The Monograph represents the first step in carcinogen risk assessment through examination of published data, positive or negative, in order to evaluate whether or not a compound could induce cancer in humans. The term ‘agent’, frequently used, refers to a broad range of agent categories including chemicals, complex mixtures, occupational and environmental exposures, as well as biological organisms.

5.2. Evaluation and rationale

Evaluation of cancer risk agents depends on the strength of evidence available in the literature related to carcinogenesis in human and animals as well as to mechanistic data. The classification of a compound doesn’t revolve around the carcinogenic potency but around the strength of evidence (sufficient or insufficient). This classification can be changed when new evidence becomes available.

Establishing a causal association between exposure to a studied agent and human cancer relies on the available epidemiological studies that allow the categorization of the compound into one of the following: a) sufficient evidence of carcinogenicity; b) limited evidence of carcinogenicity; c) inadequate evidence of carcinogenicity; and d) evidence suggesting lack of carcinogenicity.

Similarly, the carcinogenicity of a compound in experimental animals is classified based on conventional animal bioassays (mostly rodents), including those that employ genetically-modified animals.

Mechanistic and other relevant data have the power to affect the evaluation and the classification of an agent by weighting data on preneoplastic lesions, tumour pathology,

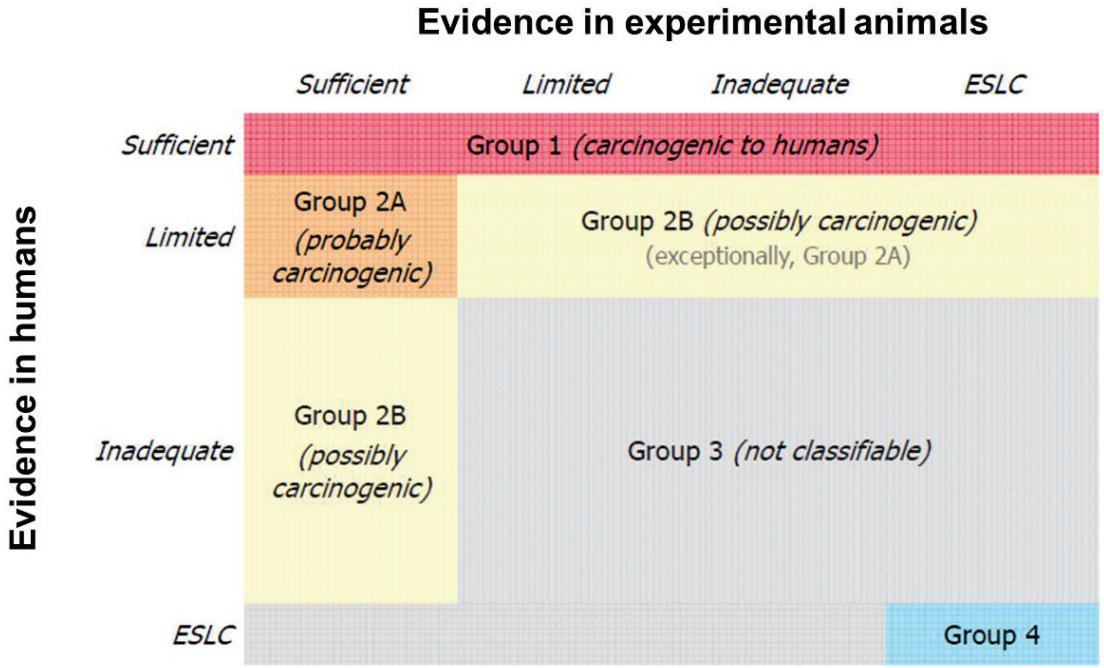
genetic and related effects, structure-activity relationships, metabolism and toxicokinetics, physicochemical parameters and analogous biological agents. The agent is then categorized based on the strength of evidence as “weak”, “moderate” or “strong”. The Working Group can identify mechanistic data that are likely to operate in humans and consider whether multiple mechanisms contribute to carcinogenesis, whether different mechanisms function at different dose ranges, whether distinct mechanisms drive tumorigenesis in humans compared to animals and whether a unique mechanism is activated only in a susceptible group. The evidence is strengthened when experiments are performed in different models with consistency in the results and biological plausibility.

5.3. Overall evaluation

Based on the strength of evidence inferred from the human epidemiology data, experimental animal studies and mechanistic and other relevant data, the agent is subsequently classified into one of the following categories (see also Table 1):

- Group 1: The agent is carcinogenic to humans. This group comprises 120 agents.
- Group 2: The agent is probably (2A) or possibly (2B) carcinogenic to humans. This group contains 81 and 299 agents, respectively.
- Group 3: The agent is not classifiable as to its carcinogenicity to humans. Group 3 embraces 502 agents.
- Group 4: The agent is probably not carcinogenic to humans. This group consists of 1 agent, caprolactam.

Table 1: **A summary of evaluation instructions** to classify an agent in one of the groups assigned by IARC monographs based on the strength of evidence in humans and experimental models. ESLC: evidence suggesting lack of carcinogenicity. Table source: IARC monographs.



Mechanistic data can be pivotal when the human data are not conclusive and can, therefore, result in the change of classification of a compound. Table C.1 is complementary to Table 1, depicting the impact of mechanistic data on cancer-risk agent classification.

6. The MutSpec project: Molecular Mechanisms and Biomarkers group, IARC

With respect to IARC's core activities, elucidating the mechanisms of environmental exposures through genetic and epigenetic alterations can provide evidence base for the etiology of cancer, strengthen the data on carcinogen evaluation and classification and may ultimately influence prevention measures.

As part of the Mechanisms of Carcinogenesis (MCA) section, a main focus of the Molecular Mechanisms and Biomarkers (MMB) group at IARC, led by Dr. Jiri Zavadil, is to decipher the origins of the molecular changes that shape human cancer genomes. Such changes can arise from environmental exposures or endogenous processes that leave fingerprints on the DNA. In 2014, the "MutSpec" project, short for Mutation Spectra, was launched in coordination with the IARC Monographs section (IMO) and other IARC groups, in order to experimentally generate mutational signatures specific to cancer-risk agents and to elucidate the enigmatic signatures observed in human tumors. For this purpose, a list of high priority compounds has been generated, reflecting MMB group interests as well as recommendations of the Advisory Group regarding compounds of interest for carcinogen classification by the IARC Monographs section (Straif et al., 2014). The "MutSpec" project seeks to identify carcinogen mutation spectra and signatures in well-controlled experimental settings, using robust mammalian *in vitro* exposure assays and tumor tissue from animal bioassays.

6.1. The experimental model systems

In vivo exposure bioassays as well as *in vitro* exposure assays are two roads that can lead to a controlled assessment of the genotoxicity, mutagenicity and carcinogenicity of a compound. Ideally, such exposure studies would use model systems that enable the testing of a large number of compounds within a reasonable timeframe. Cellular models suitable for mutation spectra analysis should include a bottleneck step followed by clonal expansion and mimic key steps of carcinogenesis (initiation via exposures, promotion and progression). There are two approaches to be considered for *in vitro* systems: 1) Bypass of a biological barrier, like crisis or senescence, and emergence of an immortalized clonal population, referred to as Barrier-Bypass Clonal Expansion (BBCE); 2) Cells to which a selective biological bypass step is not applicable require single-cell subcloning after exposure, referred to as Clonal Expansion (CE). Moreover, these models should be able to recapitulate key aspects of human biology (e.g. metabolism, DNA repair pathways) (Zhivagui et al., 2016).

6.1.1. Mouse embryonic fibroblast: Hupki MEF cells

Several model systems used for the inquiry of mutational signatures by the means of massively parallel sequencing meet some but not all of the above mentioned criteria (Zhivagui et al., 2016).

Hupki MEFs were first established for single-gene studies using Hupki mice (Liu et al., 2004). Using this cell system, exposures to UV light, AA, benzo[*a*]pyrene (B[a]P) and 3-nitrobenzanthrone (3-NBA) were carried out (vom Brocke et al., 2008; Feldmeyer et al., 2006; Liu et al., 2004, 2005). It is characterized by a biological barrier (senescence), which cells can bypass in a clonal manner (see Figure 7). Sanger sequencing of the *TP53* gene recapitulated human cancer *TP53* mutation profiles associated with the same exposures (Besaratinia and Pfeifer, 2010; Brocke et al., 2006; Kucab et al., 2010), namely in skin, kidney and lung tumors.

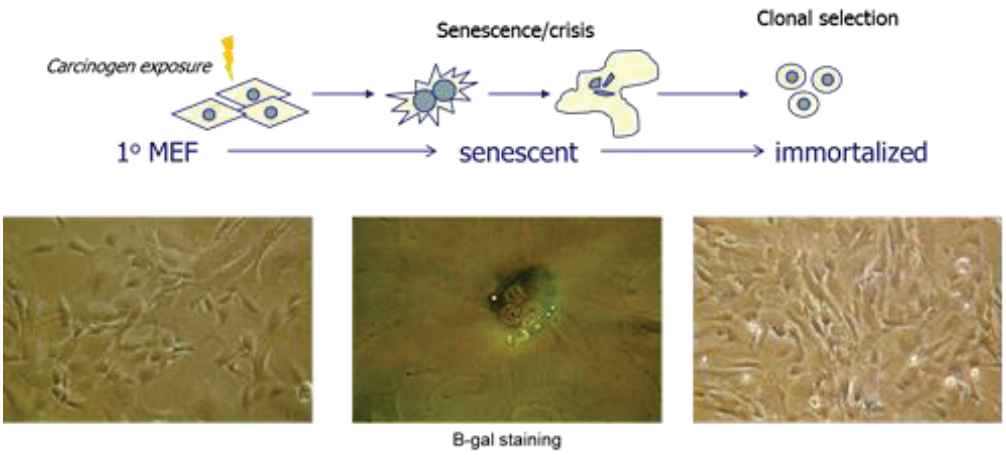


Figure 7: **Hupki MEF exposure**. MEFs are exposed as primary cells to carcinogens. The cells are propagated in culture until they reach senescence, manifested in modified cellular morphology (e.g. increase in cytoplasmic size) due to the inability of the cells to undergo a full cell cycle, hence, the formation of multi-nucleated cells. Senescence can be detected biochemically using beta-galactosidase staining. Mouse cells have the ability the bypass senescence generating immortalized cell lines representing a number of clones or subclones.

More recently, Hupki MEF cell lines derived from exposure to UV-light class C, AA, B[a]P and methylnitronitrosoguanidine (MMNG) were subjected to whole-exome sequencing. In agreement with the *TP53* sequencing studies, extracted SBS-mutational signatures recapitulated the mutation profiles observed in human cancer linked to same exposures, (melanoma, UTUC, lung and brain cancer, respectively) (Olivier et al., 2014) (Figure 8). The immortalized cell lines represent relatively homogenous populations of one predominant

clone and less represented subclones, which allows reliable identification of enriched SBS patterns upon sequencing at reasonable coverage (Zhivagui et al., 2016). These findings were validated at the whole-genome scale allowing investigations beyond SBS mutations and towards structural variations, large insertions and deletions and copy number alterations (Nik-Zainal et al., 2015).

Nevertheless, using mouse cell lines has caveats to recapitulate exposures in human beings due to limitations in the differences in genetic background, species-specific repair machineries and metabolic restrictions (Zhivagui et al., 2016). The addition of human S9 fraction, comprising active metabolic enzymes such as CYP450 and transferases, can boost metabolism of pro-carcinogens and thus circumvent the latter limitation. Interestingly, immortalization of primary mouse cells requires only one barrier bypass event such as disruption of the *p19/ARF/p53* axis, making it an easier and faster system compared with the human cells necessitating disruptions of several critical genetic pathways (Hahn and Weinberg, 2002).

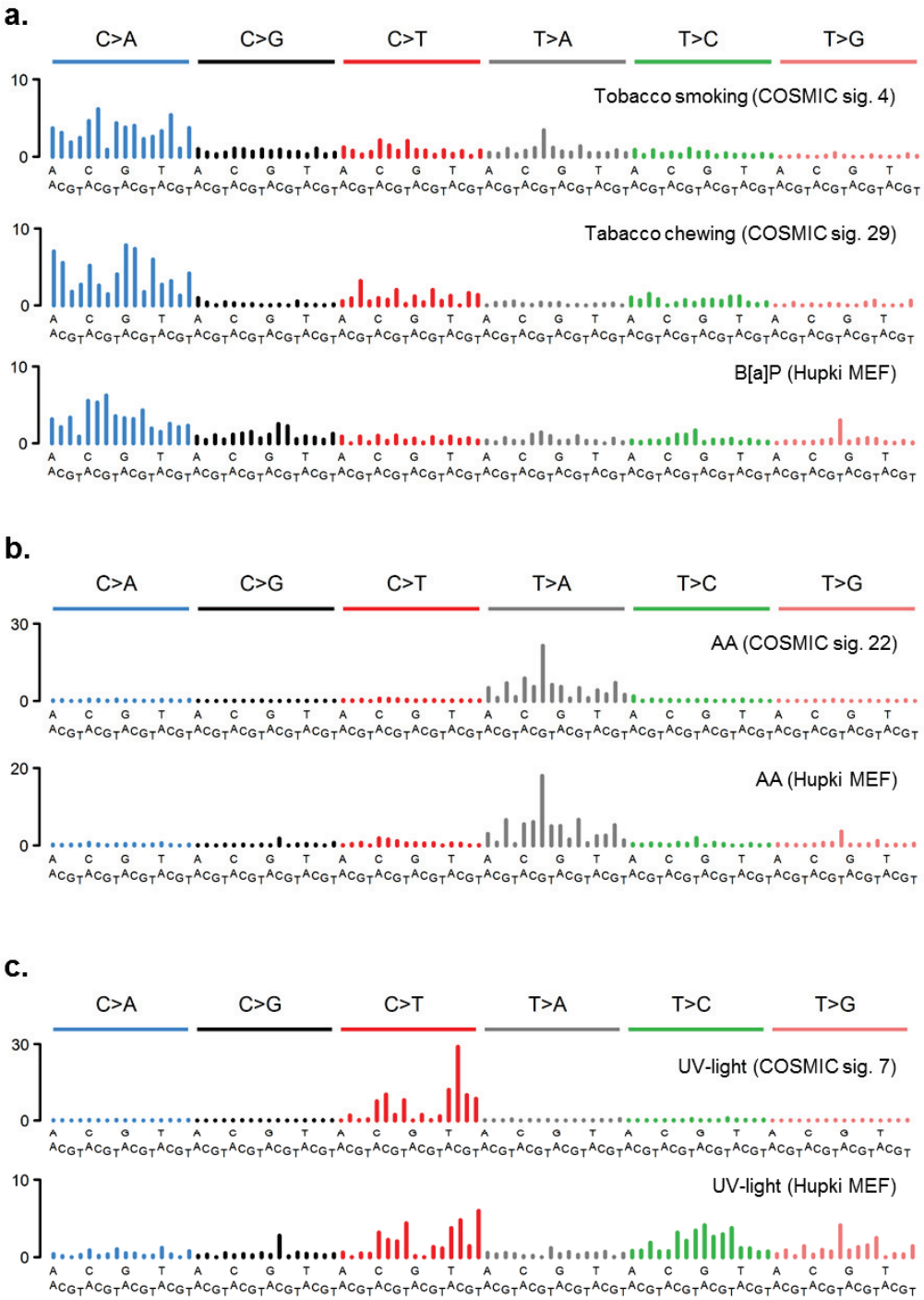


Figure 8: **Carcinogens' mutational fingerprints** in human primary tumors recapitulated in the Hupki MEF experimental system. (a) The upper panels show the mutational signature identified in smoking-related cancer patients (COSMIC signature 4 and 29). Lower panel: Hupki MEF cells treated with B[a]P under well controlled settings. (b) The upper panel represents the mutational signature identified in UTUC patients (COSMIC signature 22) correlating with AA exposure in Hupki MEFs (lower panel). (c) Mutational signature from skin cancer patients, attributed to UV-light (upper panel) (COSMIC signature 7) recapitulated by Hupki MEFs exposed to UV-light (lower panel).

6.1.2. Human cell models

6.1.2.1. HepaRG cells: human hepatic bipotent progenitor cells

HepaRG cells are hepatic progenitor cells isolated from a donor afflicted with hepatocarcinoma (Gripon et al., 2002; Guillouzo et al., 2007). HepaRG is a well-established hepatic cell line with the ability to grow as early hepatic progenitor cells, which express properties of stem cells, and can be differentiated to a dual population of hepatocyte-like and biliary-like cells. In addition, they have the capacity to completely transdifferentiate from mature cells back to progenitor cells (Cerec et al., 2007). HepaRG is an immortal cell line, with a highly stable karyotype, infinitely proliferative, and it does not give rise to tumors after transplantation into nude mice (Andersson et al., 2012). The cell line is particularly useful to evaluate drugs and perform drug metabolism studies as it expresses a full array of functions, responses, and regulatory pathways of primary human hepatocytes, including Phase I and II enzymes (Aninat, 2005). In addition, it represents an interesting tool to study aspects of progenitor biology (e.g., differentiation process), carcinogenesis, and pathogenic infections.

In culture, HepaRG cells can grow as progenitor cells, which proliferate until late passages, after which the cells seem to enter into a crisis-like event and lose some of their capacities, characterized by a slightly reduced ability to undergo differentiation towards active and mature hepatocytes. Alternatively, once the progenitor HepaRG cells reach confluency, a differentiation process is triggered and the cells start shaping their morphology towards the dual population of hepatocytes and biliary cells. The cells reach full maturity within 2 weeks after confluency. Differentiated cells have a short lifetime in culture and reach senescence within a couple of weeks. At low cell density, these cells can transdifferentiate at any stage back to progenitor cells (Cerec et al., 2007; Fukujin et al., 2000; Savary et al., 2015) (Figure 9).

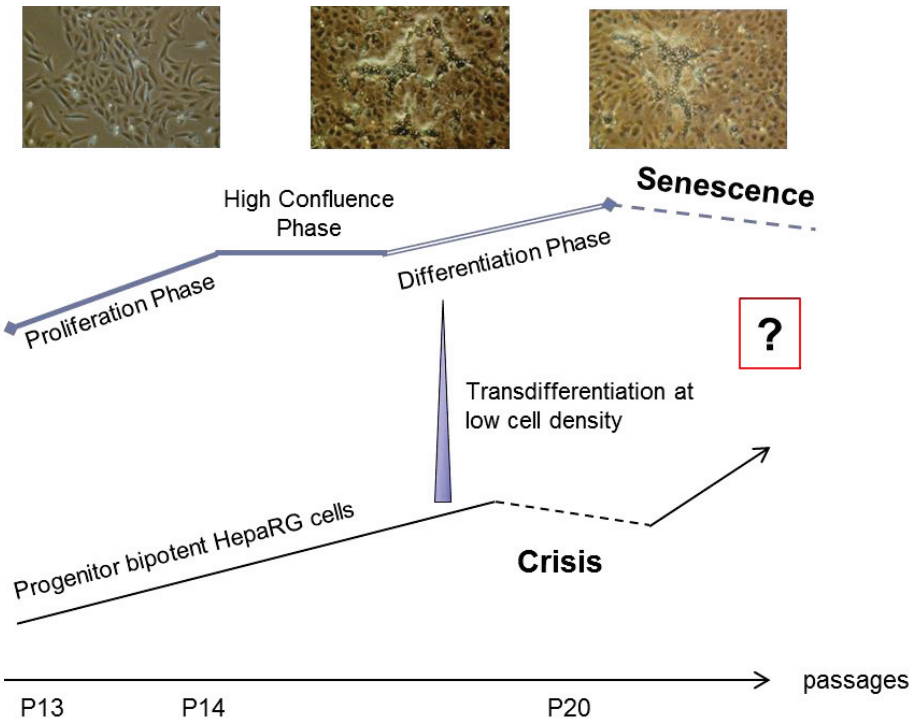


Figure 9: **Schematic representation of HepaRG cell culture** showing the two scenarios corresponding to progenitor cell proliferation (lower growth curve) as well as the differentiated dual population (upper growth curve). Images depict the cells at different stages: progenitor state (left), differentiation (center), senescence (right).

The versatility of this cellular model may offer different approaches, in order to address the effect of carcinogen exposure on the senescence and crisis-like states of HepaRG cells as well as on their potential to clonally expand.

6.1.2.2. Human lymphoblastoid cell lines: LCL

Lymphoblastoid cell lines (LCL) represent a surrogate for human isolated and cryopreserved peripheral blood lymphocytes, which are seldom available. They are established by *in vitro* infection of B-cells from human peripheral blood with the EBV virus resulting in unlimited replication of the B-cells (Hussain and Mulherkar, 2012). EBV imparts the least genetic changes on the B cells compared to other viruses, as EBV remains in the episome form in the host genome (Neitzel, 1986). Therefore, EBV-immortalized B-cells bear negligible genetic and phenotypic alterations and maintain similarities with their parental lymphocytes at the molecular and functional levels.

LCL have proven to be powerful tool for mutation analysis studies as well as for transcriptional and proteomic studies (Hussain and Mulherkar, 2012). They are used to study DNA damage, DNA repair and cytotoxicity responses to drugs, radiation and chemical compounds (Hussain and Mulherkar, 2012; Jagger et al., 2009). Importantly, genome-wide

sequencing of LCL has been documented (Schafer et al., 2013), rendering them potentially suitable for NGS-based mutation analysis of chemical compounds that effect lymphocytes.

6.1.3. Rodent bioassays: powerful in vivo exposure study systems

Animal bioassays represent another experimental system frequently used for mutation analysis. The United States National Toxicology Program (US NTP), established in 1978, is an interagency program that aims to evaluate agents of public health concern such as industrial chemicals, dietary compounds, pesticides and drugs for cancer risk assessment. To date, the program has tested the short- and long-term effects of around 600 agents in bioassays, using mice and rats. Short-term and 2-year, long-term studies are complemented with genotoxicity examination. Control and exposed animal groups are sacrificed at the end of the study, all organs undergo pathology review for cancer classification and tissues are then archived and stored (Table 2). This material can be exploited as a source of DNA for genome-wide mutational signature analyses.

Table 2: **The US National Toxicology Program.**

Source of bioassay	US NTP/NIEHS
Location	Research Triangle Park, NC, USA
Type of agent tested	Industrial chemicals, chemicals in industrial and consumer products, pesticides, water disinfection byproducts, hormones, drugs, fuels, food additives and contaminants, metals and metal compounds, particles and fibers, and non-ionizing radiation.
Number of agents tested	ca. 600
Animal models	B6C3F1 mice (also SKH-1 and Swiss mice) F344/N rats (also NBR, Sprague Dawley and Wistar rats)
Gender	Males and females
Usual number of experimental groups (Usual number of animals/group/sex)	1 control + 3-4 dose groups (50)
Routes of administration	Inhalation, feed, gavage, drinking-water (also irradiation, dermal and trans-placental)
Usual duration of studies	2-years
Histopathology	On all organs and tissues

6.2. High priority compounds, background and relative interests

A list of high priority compounds was established using a semi-automated approach focusing on evidence of human exposure and epidemiological data, evidence or suspicion of carcinogenicity and genotoxicity, and whether the additional mechanistic data would improve the classification by IARC monograph (see details of the prioritization scheme in the Materials and Methods section). It encompasses compounds of Group 1, Group 2A and 2B and Group 3 (Table 3).

Table 3: **List of high priority compounds after a multi-step prioritization process** (see details in Materials and Methods). Classification of the compounds follows IARC classification. Group 1: carcinogenic to human; Group 2A: probably carcinogenic to human; Group 2B: possibly carcinogenic to human.

Compounds	IARC Classification	Year of classification report
Acrylamide	2A	1994
Glycidamide	NA	NA
Ochratoxin A	2B	1993
Hexavalent chromium	1	1978
N-methyl-N-nitrosourea	2A	1978
N′-nitrosonornicotine	1	2007
Nicotine-derived nitrosamine ketone	1	2007
Methyleugenol	2B	2013
Glyphosate	2A	2016
N-nitroso-glyphosate	NA	NA

Among the high priority compounds, five agents were selected for testing during the framework of my PhD, namely, acrylamide, glycidamide, ochratoxin A, hexavalent chromium and N-nitroso-N-methylurea.

6.2.1. Acrylamide and glycidamide

Acrylamide is a vinyl monomer, widely used in the industries, such as water treatment and sugar production, as well as in laboratories for gel electrophoresis (IARC monographs, volume 60, 1994). In 2002, Tareke and colleagues, discovered acrylamide in food products processed at high temperatures. Acrylamide is formed in carbohydrate-rich foods upon

Maillard reactions involving heat, reducing sugars, such as glucose, and the amino acid asparagine, present in potatoes and cereals for example (Tareke et al., 2002). Other sources of acrylamide include coffee and cigarette smoke (Mojska et al., 2016; Takatsuki et al., 2003). Acrylamide is easily absorbed by an organism upon ingestion. It undergoes oxidation by cytochrome P450 in the liver, producing the epoxide metabolite glycidamide (Ghanayem et al., 2005; Sumner et al., 1999) (Figure 10). In contrast to acrylamide, glycidamide is highly reactive and can bind relatively faster to DNA (Segerbäck et al., 1995). Several DNA adducts of glycidamide have been described, namely, N7-(2-carbamoy-2-hydroxyethyl) guanine (N7-GA-Gua), N3-(2-carbamoy-2-hydroxyethyl) adenine (N3-GA-Ade) and N1-(2-carbamoy-2-hydroxyethyl) adenine (N1-GA-Ade) (Gamboa da Costa et al., 2003; Segerbäck et al., 1995) (Figure 10).

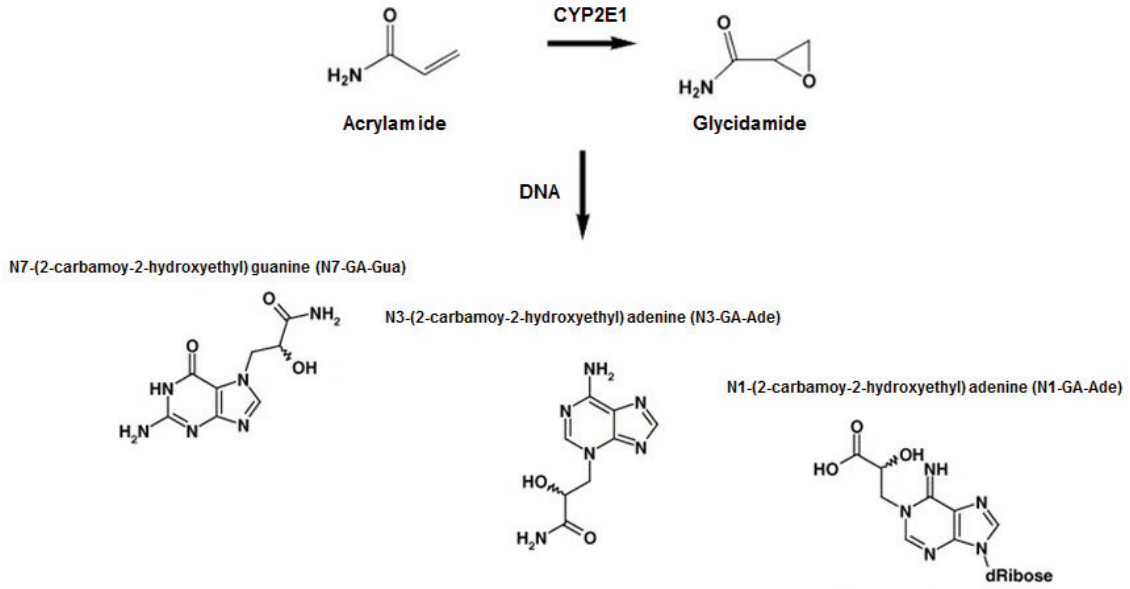


Figure 10: **Chemical structure of acrylamide and glycidamide**. Acrylamide is a vinyl that is metabolized via CYP2E1 to glycidamide. Because of its electrophilic structure, glycidamide can readily covalently bind to DNA, forming the three mostly observed DNA adducts, N7-GA-Gua, N3-GA-Ade and N1-GA-Ade. Assembled from (Doerge et al., 2005; Krishnapura et al., 2016).

Interestingly, animal bioassays show an increase in cancer development upon exposure to acrylamide and glycidamide at different sites, namely the Harderian gland, lung, forestomach, skin and mammary gland (Beland et al., 2013, 2015). In addition, epidemiological studies assessed the link between dietary acrylamide intake and renal, ovarian and endometrial cancers (Hogervorst et al., 2008; Virk-Baker et al., 2014). The results were not able to establish a clear association between acrylamide and cancer

development in humans. A number of mutagenesis assays *in vivo* and *in vitro,* based on alterations in driver genes and reporter genes showed an increased association of acrylamide and glycidamide exposure with T:A>C:G transitions, as well as T:A>A:T and C:G>G:C transversion mutations (Besaratinia and Pfeifer, 2003, 2004; Ishii et al., 2015; Manjanatha et al., 2015a; Von Tungeln et al., 2009, 2012), whereas glycidamide exposure was characterized by C:G>A:T transversions (Besaratinia and Pfeifer, 2004).

The IARC Monographs classified acrylamide as probably carcinogenic to humans (Group 2A) in 1994, based on sufficient evidence of carcinogenicity in experimental animals. Nonetheless, this classification precedes the discovery of acrylamide in food and re-evaluation of acrylamide as well as glycidamide may be warranted, considering new (molecular) epidemiological and mechanistic data.

6.2.2. Ochratoxin A

Ochratoxin A (OTA) is a toxin metabolite produced by various types of fungi (Aspergillus and Penicillium) (Figure 11). It is a widespread contaminant of animal feed and many food commodities such as cereals, coffee, cacao, grapes, wine, soy and beer (Bellver Soto et al., 2014; Kuiper-Goodman and Scott, National Toxicology Program, C56586, 1989). In Belgium, the estimated daily intake of OTA suggests that about 1% of the population exceed the tolerable daily intake level (Heyndrickx et al., 2015). OTA is a nephrotoxin (Gekle and Silbernagl, 1993; Schwerdt et al., 1999), it has been suggested to be an etiological factor for the development of the Balkan endemic nephropathy (BEN) and UTUC due to high levels of OTA detected in the blood, urine as well as breast milk of BEN patients (Clark and Snedeker, 2006; Krogh et al., 1977; Pfohl-Leszkowicz and Manderville, 2007; Radić et al., 1997). Nevertheless, AA showed stronger causal associations in BEN and UTUC patients given the levels of aristolactam-DNA adducts found in renal tissues together with the AA-specific mutation profile characterised by T:A>A:T transversions observed in the *TP53* gene from tumor tissues and supported experimentally (Arlt et al., 2007; Cosyns et al., 1994; Grollman et al., 2007; Nedelko et al., 2009). These findings were later accentuated by genome-wide sequencing of urological cancers arising in chronic renal disease patients from BEN regions (Castells et al., 2015; Jelaković et al., 2015).

Animal bioassays show a clear evidence of carcinogenicity of OTA in the kidney of F344/N rats (National Toxicology Program, 1989). The IARC Monographs classified OTA as possibly carcinogenic to humans (Group 2B) based on sufficient evidence of carcinogenicity in experimental animals (IARC monograph, volume 56, 1993).

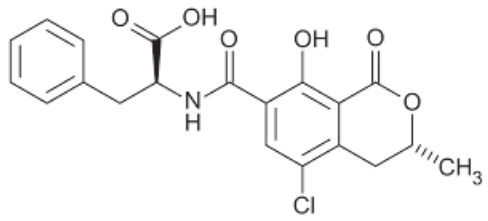


Figure 11: **Ochratoxin A chemical structure**. Taken from (Lee et al., 2012).

OTA's mode of action has been a matter of debate. Whether OTA can covalently bind to DNA to form DNA adducts or induces the production of reactive oxygen species (ROS) is still a matter of discussion. On one hand, a number of studies suggest an indirect mechanism of OTA mediated by oxidative DNA damage, as no evidence for the presence of OTA DNA adducts was found in these studies, using ^{32}P-postlabelling analyses (Jia et al., 2016; Mally et al., 2005; Turesky, 2005). On the other hand, a second set of studies have consistently detected adduct spots on chromatograms from DNA of mice, rats and pigs treated with OTA (Faucet et al., 2004; Manderville, 2005; Pfohl-Leszkowicz et al., 1991). In addition, a liquid chromatography/mass spectrometry (LC/MS) based approach suggested the presence of an OTA DNA adduct (C-C8 OTA 3'dGMP) *in vitro* (Mantle et al., 2010). Alternative mechanisms, such as effects on DNA ploidy and mitotic disruption have been suggested for OTA-mediated carcinogenicity (Brown et al., 2007; Mally, 2012).

Elucidation of potential OTA genotoxicity and mode of action using state-of-the-art technologies, such as whole-genome sequencing and adductomics analysis may provide adequate human risk assessment and carcinogen classification.

6.2.3. Hexavalent chromium

Hexavalent chromium and other chromium compounds are well established environmental carcinogens and human occupational respiratory carcinogens. Humans can be exposed to chromium through inhalation, burning cigarettes, ingestion, and water contamination due to chromium-containing wastes, dermal contact, or pressure treated woods. In addition, workers in industries that generate or use chromium (VI) are at high risk of exposure through burning of fossil fuels, waste incinerators, leather tanning and paint pigments (Nickens et al., 2010). Moreover, epidemiological studies in the UK, Europe, Japan and the U.S. have consistently shown an elevated risk of respiratory diseases in workers exposed to chromium (VI), namely, fibrosis, nasal perforation and ulceration, development of nasal polyps and lung cancer (Ishikawa et al., 1994; Nickens et al., 2010). Taken altogether, amassing human data on

industrial and environmental exposure to chromium (VI) and risk of lung cancer classified chromium (VI) as carcinogenic to human (Group 1) by IARC.

Chromium (VI) induces lung cancer in experimental animals (National Toxicology Program, 2008). Albeit the body of information supporting the genotoxicity and mutagenicity of chromium (VI) *in vivo* as well as *in vitro*, the specific mechanism of carcinogenicity for chromium (VI) remains unclear and debated (Figure 12). Hexavalent chromium can result in ROS production in response to cytotoxic effects and oxidative stress, following the reduction reaction forming chromium (III) (Bagchi et al., 1997; Nigam et al., 2014; Patlolla et al., 2009; Pratheeshkumar et al., 2016). Chromium (VI) carcinogenicity has been suggested to result in genomic instability and structural genetic lesions including DNA adducts, DNA strand breaks, DNA-protein crosslinks, oxidized bases, abasic sites, and DNA inter- and intrastrand crosslinks (O'Brien et al., 2003; Salnikow and Zhitkovich, 2008). Additionally, several lines of evidence suggest a major role of SBS in chromium (VI)-mediated mutagenicity *in vivo* and *in vitro,* mainly by targeting C:G nucleotides (Cheng et al., 2000; Holmes et al., 2008). Lastly, changes in epigenetic modifications have been observed upon exposure to chromium (VI), including aberrant methylation and gene silencing (Klein et al., 2002; Sun et al., 2009), as well as altered histone modifications, such as acetylation of histones H3 and H4 (Schnekenburger et al., 2007).

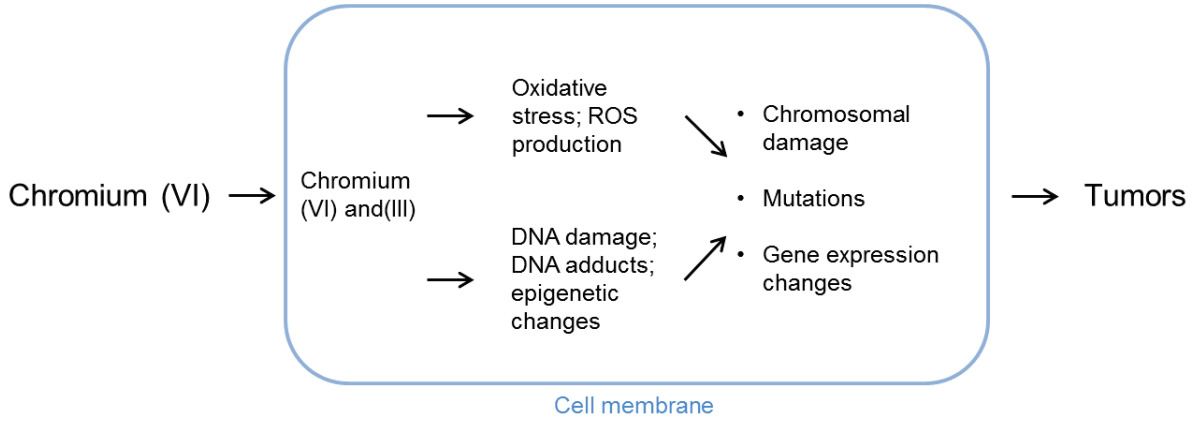


Figure 12: **Potential mechanisms of carcinogenesis upon exposure to hexavalent chromium**. After cellular uptake, chromium (VI) undergoes metabolic reduction to chromium (III) causing ROS generation and/or DNA damage through DNA adduct formation or altered gene expression. These damages can lead to genomic instabilities and acquisition of mutations (mainly SBS). Adopted from (Nigam et al., 2014).

While all these data are important and shed light on potential mechanisms of carcinogenesis mediated by chromium (VI), large-scale DNA damage studies can help solve the debate, and

with the advent of NGS a potential mutational signature of hexavalent chromium may be identified under well-controlled settings in experimental models, which may in turn be linked to exposed cancer patients.

6.2.4. N-Nitroso-N-methylurea

N-methyl-N-nitrosourea (MNU) is a nitrosoamine compound and an alkylating agent. It can be naturally formed in preserved food from nitrites, by high temperature or putrefaction. MNU has been first used in chemotherapy in combination with cyclophosphamide in solid tumors (Kolarić, 1977). Health professionals such as pharmacists, physicians, and nurses could have been exposed during clinical testing for its use as a chemotherapeutic agent, e.g. through preparation and administration of the drug or during clean-up (IARC Monographs, volume 17, 1978). MNU is carcinogenic in all animal species tested, ranging from mice and rats to dogs and monkeys (IARC Monographs, volume 17, 1978). Following administration by different routes, MNU induces tumor development at multiple sites, including the nervous tissue, stomach, esophagus, respiratory tract and kidney. Animals treated with *H. pylori* in combination with MNU showed an increase in gastric adenocarcinoma incidence, but not when they were treated with either agent alone (Sugiyama et al., 1998). Taken altogether, MNU is reasonably anticipated to be a human carcinogen based on sufficient evidence of carcinogenicity in experimental animals and classified by IARC as a probable human carcinogen (Group 2A).

MNU alkylates nucleic acids both *in vivo* and *in vitro* allowing the transfer of its methyl group onto 7-guanine producing mostly 7-methylguanine (Lijinsky et al., 1972). The genotoxicity and mutagenicity of MNU has been validated in different experimental models using assays detecting sister chromatid exchange, unscheduled DNA synthesis and bacterial phages for mutagenicity (IARC Monographs, volume 17, 1978). Recently, the first whole exome sequencing data on MNU exposure has been generated from mouse lung cancer. It defined the mutation spectra of MNU characterized by predominant C:G>T:A transitions, resembling the profile of other alkylating agents, such as MMNG and temozolomide (Alexandrov et al., 2013b; Olivier et al., 2014; Westcott et al., 2015) (Figure 13).

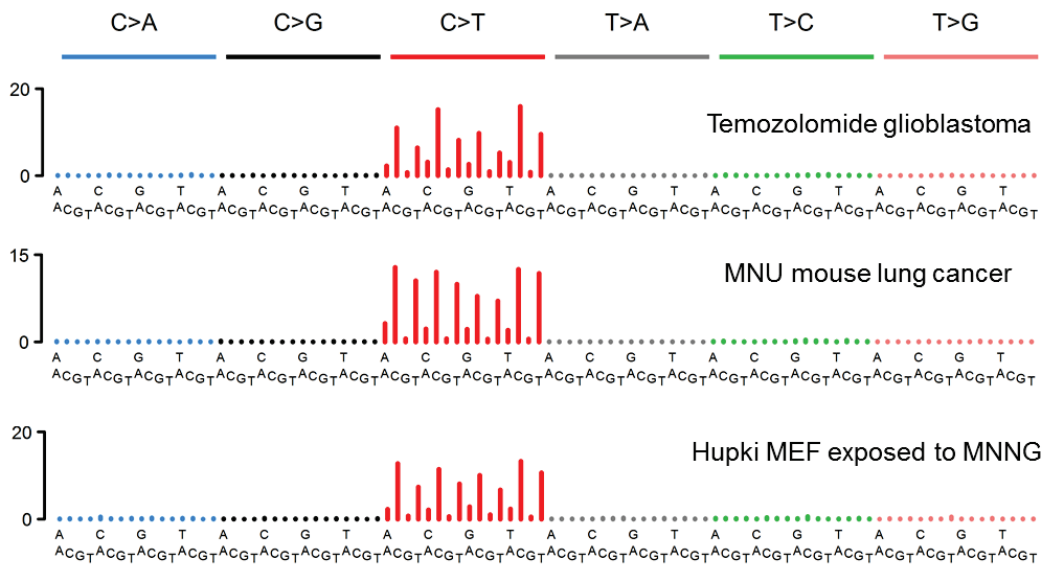


Figure 13: **Similarities between alkylating agents' mutational fingerprint**. Mutational signatures of temozolomide from human glioblastoma developed upon treatment to the chemotherapeutic agent (Data taken from COSMIC), MNU exposure-derived mouse lung cancer (Westcott et al., 2015), and MNNG-treated Hupki MEFs (Olivier et al., 2014).

In order to establish the Hupki MEF immortalization model MNU was used as a proof-of-principle compound for cytotoxicity and genotoxicity assays as well as for single-gene mutation screening after clonal expansion. MNU remains a compound of interest in the MutSpec project due to its potential implication in nasopharyngeal carcinoma, as well as an interesting alkylating compound to be evaluated for adductomics analyses.

OBJECTIVES

Mutagenic compounds can alter the DNA in characteristic ways, leaving imprints termed mutational signatures. The identification of carcinogen-specific mutational signatures can, therefore, help unravel cancer etiology. Among the 30 mutational signatures observed in human primary tumors, 40% remain of unknown origin and only 23% were attributed to specific external exposures, such as UV light, alkylating agents, dietary contaminants and tobacco smoke (Alexandrov et al., 2013b). Thus, it becomes urgent deciphering the mutational signatures of the hundreds of agents that are known to be carcinogenic to humans (IARC Group 1) as well as probably/possibly carcinogenic to humans (IARC Group 2A and B). In order to accomplish this task, well-controlled experimental systems are required to identify the causes of such orphan mutational signatures (Zhivagui et al., 2016).

The MutSpec project is an IARC cross-cutting program aiming at an integrative analysis across different experimental and primary systems (Figure 14), which include *in vitro* exposure of mammalian cells to a number of cancer-risk agents, cross validation of the resulting mutational signatures with those generated by concurrent sequencing of rodent tumors from the chemical bioassay collection of the US NTP and matching them to primary human tumor sequencing or public-domain human cancer data.

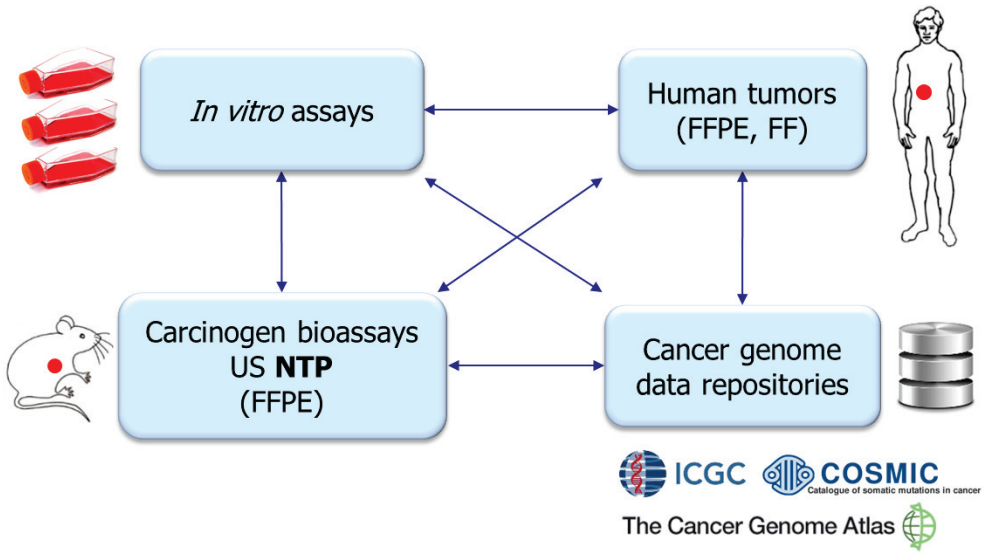


Figure 14: The MutSpec project: a multi-system analysis of experimentally identified mutational signatures with primary tumors (rodent and human tumors, public human cancer data repositories).

Within the framework of my PhD, my objectives were:

Aim 1: Development of mammalian experimental cellular models for the MutSpec project (Review 1, Zhivagui M et al., 2016, BCPT journal).

Aim 2: Identification of the cytotoxic and genotoxic potential of high priority compounds.

Aim 3: Characterization of the mutational signature of high priority compounds using suitable experimental models (Results summarized in Paper 1, Zhivagui M. et al., under review, Carcinogenesis journal; Paper 2, Zhivagui M et al., in preparation).

First, we adopted the Hupki mouse embryonic fibroblasts cell system for its ability to emulate critical steps of cell transformation and carcinogenesis: selective barrier bypass and clonal expansion of the resulting immortalized cells. Previous reports have shown that exposure in Hupki MEFs reproduced observations from genomic data derived from human cancers linked to identical exposures (Nik-Zainal et al., 2015; Olivier et al., 2014; Zhivagui et al., 2016). In addition, HepaRG cells and LCL were assessed for their application in the MutSpec project.
Second, we focused on the list of high-priority cancer-risk agents epidemiologically linked to human cancer and for which additional mechanistic data can help delineate cancer etiology and speed up carcinogen classification.
Third, agent-specific mutational signature was extracted using in vitro experimental models. The resulting *in vitro* mutational signatures were matched with those generated by concurrent sequencing of rodent tumors from the chemical bioassay collection of the US NTP and compared to own or public-domain human cancer data.

This integrative analysis can ensure the identification of high-confidence mutational signatures and can thus help interpret the mechanistic impact of the tested agents on human cancer burden.

MATERIALS AND METHODS

1. Prioritization of compounds for testing

Starting from a list consisting of compounds that have never been classified (NA), compounds with emerging toxicological data of concern (Group 3), and classified compounds for which the toxicological or epidemiological data have changed (Group 2A, 2B), a semi-automated data-mining approach using different databases (such as PubMed, ToxRef, NCBI) was used to rank compounds (Figure 15). The databases queries included epidemiological evidence, mutagenicity, genotoxicity, DNA adduct formation, chemical structure similarity and availability of animal bioassays, and the integration of the results was visualized by Cytoscape or MetaMapp (visualization platforms). The prioritization was, thus, based on evidence of human exposure and epidemiological data, evidence or suspicion of carcinogenicity and genotoxicity, and whether the additional mechanistic data would improve the classification by IARC.

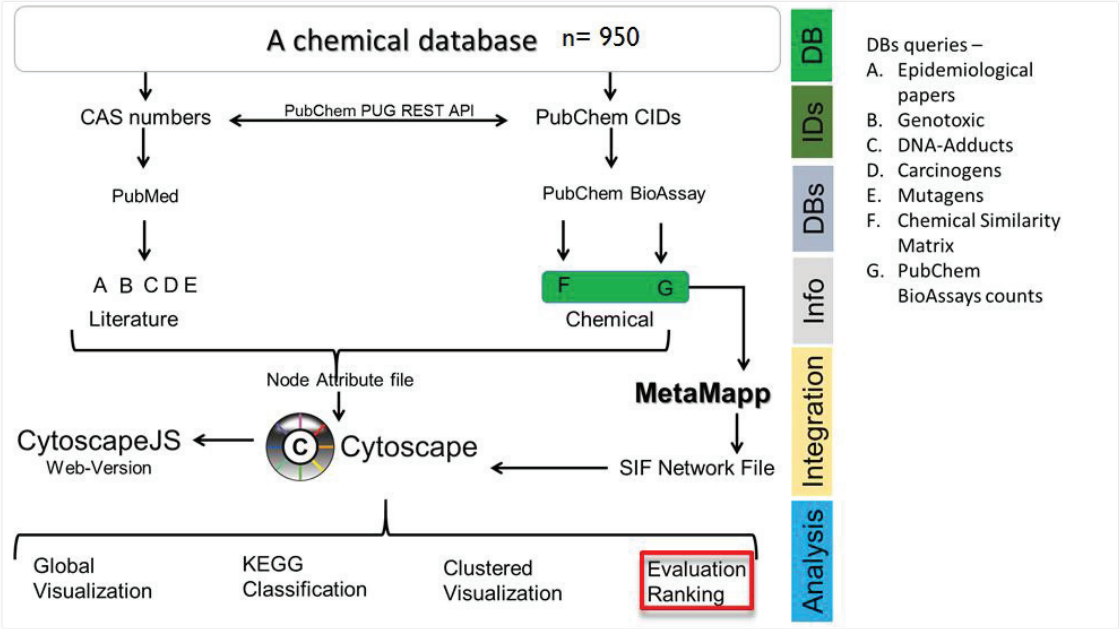


Figure 15: IARC's semi-automated data mining system used for initial prioritization of chemical compounds from various databases (such as PubMed) according to their associated information (database queries). The integration of the output results is visualized by MetaMapp and Cytoscape (visualization platforms). Courtesy: Dr. Dinesh Kumar Barupal.

Next, based on the evaluation ranking, in-depth literature searches of top-ranked compounds allowed the selection of 50 high priority compounds for testing.

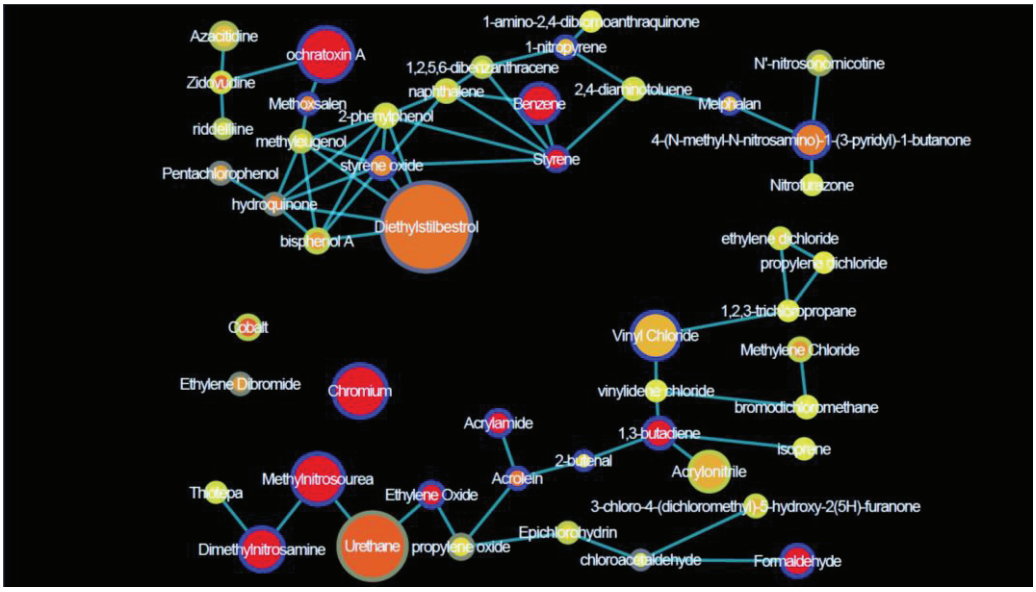


Figure 16: Compounds clustering based on chemical structures (distance between compounds); genotoxicity counts (node color: yellow – low, red – high); carcinogenicity counts (node size); and DNA adduct formation (node border color: yellow –low, blue – high).

These compounds were subsequently clustered based on structural similarities, data records on genotoxicity, carcinogenicity and DNA adducts formation (Figure 16). Finally, 10 compounds out of 50 were selected for the availability of sufficient information regarding their chemical nature and biological activity as well as their prevalence in the human environment, such as acrylamide and some pesticides. The final list encompasses compounds of Group 1, Group 2A and 2B and Group 3 (Table 4).

Table 4: **List of high priority compounds** after a multi-step prioritizing process. Numbers represent the output counts from PubMed search.

Compounds	IARC Classification	Mutagenicity	DNA adducts formation	Epidemiological studies
Acrylamide	2A	226	66	104
Glycidamide	NA	50	36	17
Ochratoxin A	2B	134	68	16
Hexavalent chromium	1	392	67	159
N-methyl-N-nitrosourea	2A	740	95	112
N′-nitrosonornicotine (NNN)	1	31	23	14
4'-(N'-nitrosomethylamino)-1-(3-pyridyl)-1-butanone (NNK)	1	114	97	81
Methyleugenol	2B	163	13	2
Glyphosate	2A	74	2	29
N-nitroso-glyphosate	NA	0	0	0

2. Compounds preparation

Most of the compounds were purchased from Sigma-Aldrich with the exception of the N-methyl-N-nitrosourea that was kindly shared by Prof. David Phillips, and the N-nitroso-glyphosate which was purchased from Toronto Research Chemicals (TRC). The compounds were diluted to a stock solution of 1 M or 500 mM and stored at -20°C (Table 5).

Table 5: Solvents used to dissolve each compound. PBS: Phosphate-Buffered Saline; DMSO: dimethylsulphoxide.

Compounds	**IARC Classification**	**Solvent**
Acrylamide	2A	PBS
Glycidamide	NA	PBS
Ochratoxin A	2B	DMSO
Hexavalent chromium	1	PBS
N-methyl-N-nitrosourea	2A	DMSO
N′-nitrosonornicotine (NNN)	1	DMSO
4'-(N'-nitrosomethylamino)-1-(3-pyridyl)-1-butanone (NNK)	1	DMSO
Methyleugenol	2B	DMSO
Glyphosate	2A	PBS
N-nitroso-glyphosate	NA	PBS

3. Hupki MEFs cell culture, exposure and immortalization

Human p53 knock-in mouse embryonic fibroblasts (Hupki MEFs), isolated from 13.5-day old *Trp53*$^{tm/HolI}$ mouse embryos harboring a humanized *Trp53* gene (Whibley et al., 2010), were cultured in Advanced DMEM supplemented with 15% fetal calf serum, 1% penicillin/streptomycin, 1% pyruvate, 1% glutamine and 0.1% β-mercapto-ethanol. Hupki MEFs from two different embryos were used for each exposure experiments. Primary MEFs were seeded in six-well plates and, at passage 2, exposed for 24 hours to the compound of interest or to the vehicle. Acrylamide and ochratoxin A exposures were carried out in the absence or presence of 2% human S9 fraction (Life Technologies) complemented with NADPH (Sigma). Exposed and control primary cells were cultivated until they bypassed senescence and immortalized so that clonal cell populations could be isolated (Figure 7) (Chen et al., 2013; Todaro and Green, 1963). A number of Hupki MEF immortalized clones were generated from exposure to acrylamide in the presence and absence of human S9 fraction, glycidamide, ochratoxin A in the presence and absence of human S9 fraction, chromium (VI) and MNU.

4. HepaRG cell culture, exposure and clonal expansion

Progenitor HepaRG cells were seeded at low density until they reached confluency within 2 weeks. Progenitor cells were cultivated in DMEM supplemented with 10% Fetal Bovine Serum (FCS), 1% penicillin/streptamycin, 1% glutamine, 1% sodium pyruvate and 0.001% dexamethasone. Once the cells reached confluency, the medium was complemented with the growth factor EGF to boost the differentiation into hepatocyte-like cells. 2% DMSO was then added leading to full differentiation of cells towards hepatocyte-like and biliary-like cells (Figure 9). Hepatocytes are smaller in size compared to biliary cells and they are less adherent to the collagen matrix. In order to isolate hepatocytes from the dual population, we performed FACS sorting based on the size of the cells and seeded the cells at high density to preserve the differentiation state of the cells. In addition, we tried another technique to separate the hepatocyte cells by partial trypsinization, i.e. incubating the cells with 0.025% of trypsin for less than 5 minutes and collecting the detached cells. These cells were then seeded at high density. In order to assess the metabolic functionality of isolated hepatocyte-like cells by partial trypsinization (PT), we collected samples at different time-points: immediately after PT (PT_0), 24 hours after seeding the cells at high density (PT_{24}), 4 days after PT (PT_4), 7 days after PT (PT_7), 10 days after PT (PT_{10}) and 14 days after PT ($PT_{2\ weeks}$). RT-qPCR was carried out using different hepatic markers including CYP3A4, CYP2E1, albumin and aldolase. Progenitor cell markers were also used, such as CK-19.

5. Cytotoxicity assessment upon compound exposure

In order to define the cell viability upon treatment to cancer-risk agents, cells were seeded in 96-well plates and treated with a range of concentrations of the compound in test. Cell viability was measured 48 hours after treatment cessation using CellTiter 96® Aqueous One solution Cell Proliferation Assay (Promega). After exposure, cells were washed with PBS and fresh medium containing 10% MTT reagent was added, in which the cells were incubated for 4 hours at 37°C. The absorbance was measured at 492 nm using the APOLLO 11 LB913 plate reader. The MTT assay was performed in triplicates for each experimental condition. LCL cytotoxicity evaluation was performed using the Trypan Blue exclusion test. As a result of cytotoxicity testing, exposure conditions for Hupki MEFs were established: 10 mM of ACR, 5 mM of ACR+S9, 3 mM of GA, 0.8 mM of OTA +/- S9, 25 mM of Cr(VI) and 10 mM of MNU for 24 hours. HepaRG cells were treated with 200uM of AA for 24 hours. Finally, LCL cells were chronically treated with 1.25 mM (non-cytotoxic dose) and 10 mM (cytotoxic dose) of glyphosate and N-nitrosoglyphosate.

6. Genotoxicity assessment upon compound exposure

Immunofluorescence staining was carried out using an antibody specific for Ser139-phosphorylated H2Ax (γH2Ax) (9718, Cell Signaling Technology). Briefly, primary MEFs were seeded on coverslips in 12 well-plates and, the following day, treated with the compound in duplicates for 24 hours. Four hours after treatment cessation, the cells were fixed with 4% formaldehyde at room temperature for 15 minutes. Following blocking in 5% normal goat serum (31872, Life Technologies) for 60 minutes, they were incubated with γH2Ax-antibody (1:500 in 1% BSA) at 4°C overnight. Subsequent incubation with a fluorochrome-conjugated secondary antibody (4412, Cell Signaling Technology) was carried out for 60 minutes at room temperature. Coverslips were mounted in Vectashield mounting medium with DAPI (Eurobio). Immunofluorescence images were captured using a Nikon Eclipse Ti microscope.

LCL cells were exposed in 6-well plates for 24 hours. After washing with PBS and centrifugation, the cells were transferred onto a glass slide and left to air dry for 15 minutes. Blocking solution comprised of PBS and 5% BSA. After 1 hour of blocking at room temperature, the cells were incubated for 1 hour with the primary antibody in PBS and 1% BSA. Finally, the samples were incubated in a fluorochrome-conjugated secondary antibody for 45 minutes at room temperature and the slides were mounted in Vectashield mounting medium with DAPI.

7. DNA adduct analysis

Liquid chromatography, tandem mass spectrometry (LC/MS/MS) is a highly sensitive and specific analytical technique that can accurately determine the identities and concentrations of DNA adducts within samples. Glycidamide-DNA adducts [N7-(2-carbamoy-2-hydroxyethyl)-guanine (N7-GA-Gua) and N3-(2-carbamoy-2-hydroxyethyl)-adenine (N3-GA-Ade)] were quantified at the National Center for Toxicological Research (NCTR) using LC/MS/MS with stable isotope dilution as previously described (Gamboa da Costa et al., 2003). The DNA was isolated from the cells using standard digestion with proteinase K, followed by phenol-chloroform extraction and ethanol precipitation. The DNA was subsequently treated with RNase A and T1, extracted with phenol-chloroform, and reprecipitated with ethanol. N7 GA-Gua and N3 GA-Ade were released by neutral thermal hydrolysis for 15 min, using Eppendorf Thermomixer R (Eppendorf North America, Hauppage, NY) set at 99°C. The samples were filtered through Amicon 3K molecular weight cutoff filters (Merck Millipore, Tullagreen, IRL) to separate the adducts from the intact DNA. The LC/MS/MS used for quantification consisted of an Acquity UPLC system (Waters, Milford, MA) and a Xevo TQ-S triple quadrupole mass spectrometer (Waters, Milford, MA). The same MRM transitions as previously described were monitored with a cone voltage of 50V and collision energy of 20eV for each adduct transition and its corresponding labeled isotope transition.

OTA-induced DNA adduct formation was measured by the LC/MS/MS permitting a screening of all the DNA adducts found in a sample at the University of Minnesota. This analysis is based on three consecutive detection events: Full scan (MS^1), data-dependent MS^2-acquisition (dd- MS^2) and Neutral Loss-acquisition (NL-MS^3). The full scan ionization measures the accurate mass of each individual ion making it possible to assign a molecular formula to each analyte. Every 100ms, the detector isolates the 5 most abundant ions (based on the intensity of the peaks) and fragments them looking for a specific signal corresponding to the neutral loss of the deoxyribose (dR) group (MS^2). A further fragmentation event (MS^3) is triggered upon the observation of the neutral loss allowing the release of the nucleobase and the adduct (for detailed protocol, see Appendix C). Different conditions were included, namely, treated and untreated Hupki MEFs, in the presence and the absence of the human S9 fraction. Samples were collected 4 hours after treatment cessation. Taking in consideration the previously suggested chemical structure of an OTA-DNA adduct on guanine, a fractioning collection was carried out in order to focus on the structure and the mass of this specific DNA adduct as well as any other chemical structure that would resemble OTA. Furthermore, we also assessed indirect mechanisms of OTA, mediated through ROS production and ROS DNA-adducts.

8. RNA extraction

A strong lysis and a good homogenization of the samples are a key for RNA isolation as they ensure quick breakdown of the cells to inactivate RNases in the lysis buffer. Briefly, 200uL of trizol was added to one million cells and left at room temperature for 5 minutes after vigorous vortexing. Chloroform (20% the volume of trizol) was then added followed by centrifugation at 11000 rpm for 10 minutes. The organic phase contains proteins and lipids; the interphase holds the DNA whereas the aqueous phase contains the RNA. Hence, the aqueous phase was aspirated and complemented with 200uL of chloroform. After centrifugation for 10 minutes, the upper (aqueous) phase was again extracted and maintained in 1uL of glycogen, isopropanol (50% of the aqueous phase) and ammonium acetate (10% of the final volume). The samples were incubated at -20°C for 3 hours or overnight then centrifuged for 30 minutes at 14000 rpm at 4°C. The RNA pellet was washed in 70% ethanol prepared with RNase-free water and spun at 4°C for 2 minutes at 14000 rpm. The remaining ethanol was left to evaporate at room temperature for 5 minutes. The pellet was dissolved in RNase-free water and the quality as well as the quantity of the RNA assessed using a NanoDrop. The 260/280 ratio gives a loose indication of the purity of the RNA in the sample. The absorbance at 260 nm measures the RNA concentration and at 280 nm it measures the protein concentration in the sample. The 260/280 ratio should range between 1.8 and 2.2. Samples with a low 260/230 ratio (below about 1.8) have a significant presence of organic contaminants that may interfere with other downstream processes like RT-PCR, lowering the efficiency of the enzymes.

9. Quantitative Real-Time Polymerase Chain Reaction (qRT-PCR)

The input RNA used was 500ng per sample (treated with DNase). To reverse transcribe the RNA and generate cDNA we used the Reverse Transcription kit from Life-technology and followed the manufacturer's instructions. A control sample (without reverse transcriptase enzyme) was also included. Following cDNA production, targeted primers for the genes of interest were used, such as CYP3A4, CYP2E1, albumin and aldolase. Polymerase Chain Reaction (PCR) was performed in duplicate using SYBR green master mix. Gene expression was normalized to the average of three housekeeping genes: *GAPDH*, *B2M* and TATA Box binding protein (*TBP*).

10. *TP53* genotyping of primary and immortalized cells

Exons 4 to 8 of the Hupki MEF-knocked-in human *TP53* gene (NC_000017.11) were sequenced using standard protocols. Sanger sequencing of PCR products was performed at Biofidal (Lyon, France), using the following primers (all in 5' to 3' orientation): Exon 4: fwd – TGCTCTTTTCACCCATCTAC, rev – ATACGGCCAGGCATTGAAGT; Exons 5-6: fwd –

TGTTCACTTGTGCCCTGACT, rev – TTAACCCCTCCTCCCAGAGA; Exon 7: fwd – CTTGCCACAGGTCTCCCC, rev – CACTTGCCACCCTGCACA; Exon 8: fwd – TCCTTACTGCCTCTTGCTTCTCTT; rev – CCAAGGGTGCAGTTATGCCT. Sequences were analyzed using the CodonCode Aligner software.

11. DNA extraction from cultured cells

DNA extraction followed the manufacturer's instructions using the nucleospin tissue DNA extraction kit (from Macherey Nagel). Briefly, cell pellets were resuspended in Buffer T1, proteinase K and Buffer B3 then incubated at 70°C for 15 minutes. In order to adjust DNA binding conditions, absolute ethanol was added to the mix and the samples vortexed. The samples were loaded onto provide columns and centrifuged for 1 minute at 11000 rpm. The flow-through was discarded and the column washed twice with BW buffer and B5 buffer, respectively. Finally, pure DNA was eluted in AE buffer (from Qiagen containing EDTA). Quantity and purity of the DNA were assessed using Qubit and NanoDrop as well as by agarose gel electrophoresis.

12. Animal bioassay FFPE sample processing

Tissue selection and retrieval were carried out in collaboration with the US NTP (Figure 17).

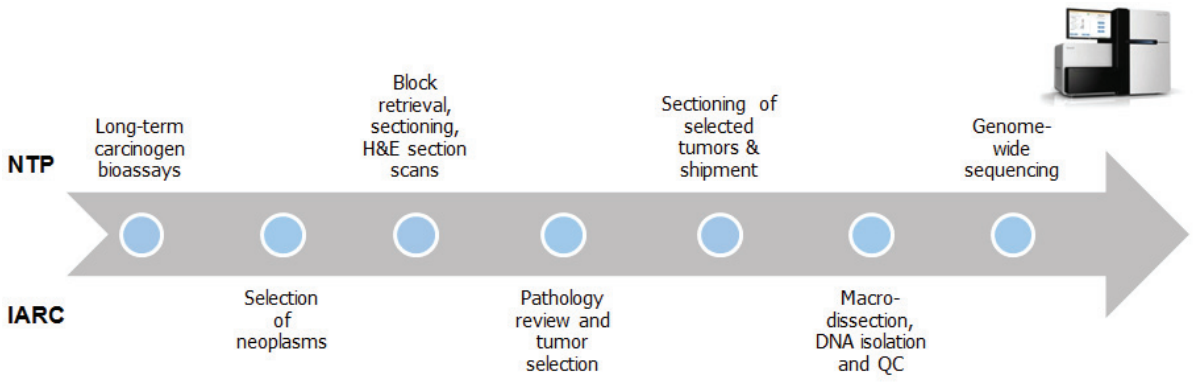


Figure 17: Tissue collection pipeline in collaboration with the US NTP.

Sample selection was based on long-term 2-year bioassays, for which statistical analysis on the carcinogenicity of the test compound had been assessed in comparison to spontaneous tumor occurrence in vehicle-exposed rodents by the US NTP. For the OTA study, male rats showed a clear evidence of carcinogenicity in the kidney. No spontaneous kidney tumors were present in the control groups. Following the request of haematoxylin and eosin (H&E)

stained scans, these were reviewed and the tumor area annotated by a pathologist at IARC (Figure 18). Final tumor selection was based on tumor size and the selected tissue sections (10 x 10um unstained sections with an H&E-stained 5um section at the beginning and end of the series) were provided by the US NTP. Tumor tissues were complemented with normal tissue originating from brain or liver of the same animal. Tumor tissue was macro-dissected and DNA extraction, including quality control to assess the purity and the quantity of the DNA, was performed (see sections below).

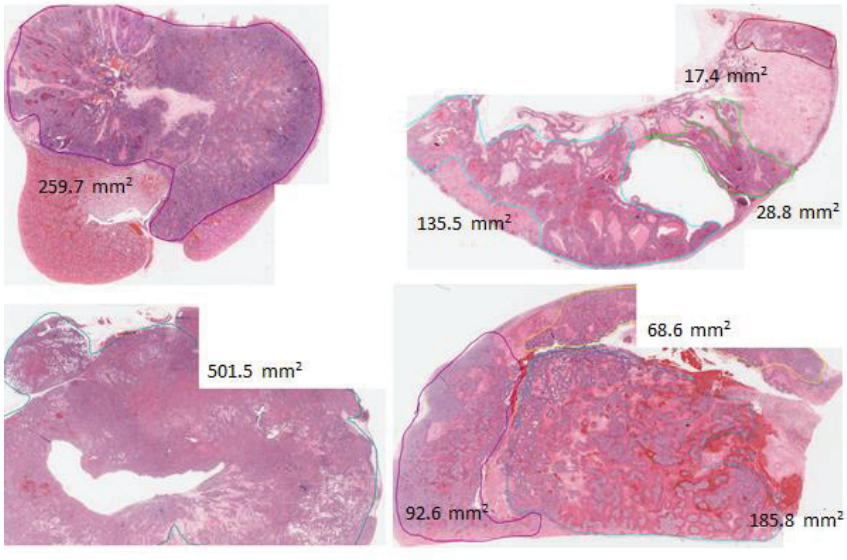


Figure 18: Tumor tissue sections stained with H&E. Pathological review allowed classification of the tumors and annotation of the tumor size within each sample.

13. DNA extraction from animal tissues

Genomic DNA was extracted from normal and tumor tissues (6 slides each) after deparaffinization. In order to enrich for tumor tissue and reduce wild-type background, we performed a macro-dissection of the slide by referring to the annotated H&E stained scans. The nucleospin tissue DNA extraction kit was used for DNA isolation (from Macherey Nagel) with some modifications. Briefly, tissues were resuspended in 180uL of Buffer T1 and 25uL of proteinase K and kept overnight at 56°C. If the lysis was not yet complete, 25uL of proteinase K were added for 1 hour at 56°C. 2uL of RNase A (100mg/mL) solution was added to the samples and they were incubated for 5 minutes at room temperature. After addition of 200uL of Buffer B3, and vigorous vortexing the samples were incubated at 70°C for 10 minutes. The nature of the samples (FFPE tissue) required a critical DNA de-crosslinking step. To do so, samples were incubated in 400uL of de-crosslink buffer for 3

hours at 65°C. In order to adjust DNA binding conditions, 800uL of absolute ethanol were added to the mix and vortexed. For each sample, one NucleoSpin®Tissue Column was placed into a Collection Tube, 700µl of the sample applied to the column and incubated for 15 minutes at room temperature. The samples were centrifuged for 1 minute at 5000 rpm, the flow-through discarded and the silica membrane washed twice: first, with 500uL of Buffer BW (3 minutes incubation at room temperature followed by centrifugation for 1 minute at 5000 rpm) and second, with 600uL of Buffer B5. Finally, the DNA was eluted in 100uL ultrapure water by centrifugation for 1 minute at 11000rpm. The DNA was quantified by Qubit. DNA quality control was performed to insure the suitability of the DNA for library preparation and next-generation sequencing, namely through PCR and qPCR.

14. Library preparation for WGS

WGS library preparation was carried out using the Kapa High-Throughput kit. Genomic DNA is fragmented mechanically by sonication using Covaris shearing instrument. The input DNA is 330ng diluted in 55uL of AE buffer from Qiagen containing 10 mM Tris-HCL (pH 9) and 0.5 mM EDTA. The DNA is then transferred into a snap-cap microtube and fragmented for an average size range between 350 and 550 bp. The cell line DNA was fragmented for 60 seconds whereas the FFPE DNA was fragmented for 130 seconds to get to the desired size range. The tubes were spun every 30 seconds as well as at the end of the shearing. 1uL of the fragmented DNA is then assessed on Bioanalyzer DNA high-sensitivity chip in duplicate, while the rest of the DNA is transferred into wells of a PCR plate. As the tissue fixation and storage significantly damage and compromise the quality of DNA from FFPE samples, an additional step is used to repair deaminated cytosine to uracil, nicks and gaps, oxidized bases and blocked 3’ ends by employing the NEBNext FFPE repair kit which contains a cocktail of repair enzymes permitting a better ligation of the adaptors and increasing the yield of the DNA library. After this step, the DNA is cleaned-up using 3x of AMPure XP beads and eluted in ultrapure water. We then proceed with the end repair reaction and A-tailing allowing the incorporation of non-template dAMP on the 3’ end of the DNA fragments. The DNA is re-attached to the beads using 3x of PEG/NaCl SPRI solution allowing purification of the DNA for the next step. For adapter ligation, we prepare a mix consisting of water, ligation buffer and T4 DNA ligase. For each sample we added different indexed adapter to allow distinction of the samples when pooled for sequencing. After incubation for 15 minutes at 20°C in a thermocycler, the DNA is re-attached to the beads using 1x of PEG/NaCl SPRI solution and we eluted in 52uL of EB buffer from Qiagen. In order to remove the adapter dimers as well as the big DNA fragments we performed a dual size selection before DNA amplification; first, using a 0.6x ratio of AMPure beads we removed fragments bigger than 550 bp by keeping the supernatant (for examples, for 50uL of DNA we added 30uL of beads). Second, using a

0.9x ratio of AMPure beads we allowed for DNA fragments larger than 250 bp to bind to the beads (Note: as PEG solution is left from the previous ratio the amount of beads to add is less than 0.9x. We added for 78uL of DNA 9.75uL of beads). We finally eluted the DNA in 22uL of EB buffer from Qiagen not containing EDTA as it can bother the PCR reaction. The PCR reaction is carried out using 7 cycles and the amplified library is finally purified with AMPure beads (1.8x) and eluted in ultrapure water. Lastly, library profiles are examined on Bioanalyzer assuring the size of fragments and the good ligation of the adaptors. The samples are then pooled, shipped and sequenced at GENEWIZ, New Jersey, USA, with 150 bp paired-end sequencing at 50X coverage.

15. Library preparation for WES

Exome sequencing is a capture based method developed to identify variants in the protein-coding region of the genome. The typical workflow for WES follows the same steps as the genome sequencing with an additional capture phase: Nucleic acid isolation, DNA fragmentation (350-450 bp), End-repair and A-tailing, target and capture exons using biotinylated probes from the SureSelect XT Mouse All Exon Kit (Agilent Technologies), and amplification of targets. The Kapa Hyper Plus kit was used for WES library preparation. It relies on enzymatic digestion using double strand DNA Fragmentase. While there are approximately 180,000 exons in the human genome, constituting less than 2% of total sequence, the exome contains ~80-90% of known variants causing disease making it a cost-effective alternative to whole genome sequencing.

16. Bioinformatics pipeline and processing of NGS data

For NGS data analysis, a bioinformatics pipeline was developed in the MMB group (Ardin et al., 2016) and implemented in a Galaxy web-based platform (Figure 19).

Fastq files were analyzed for data amount and quality using FastQC (0.11.3) and were processed with an in-house pipeline for adapter trimming and alignment to the mm10 genome (release GRCm38). These components of the pipeline are publicly available at https://github.com/IARCbioinfo/alignment-nf. Two somatic variant callers were employed with default parameters in order to detect single base substitutions (SBS) and small insertions/deletions (indels) (MuTect 1.1.6-4 and Strelka 1.015) in exposed clones, using primary cells and normal tissues as reference samples.

Mutation data obtained from the MuTect variant caller were processed with the MutSpec suite ((Ardin et al., 2016); https://github.com/IARCbioinfo/mutspec for annotation with Annovar and variant filtering to remove single-nucleotide polymorphism (SNP) contents (dbSNP142 and dbSNP146 for mouse and rat samples, respectively), segmental duplicates, repeats and tandem repeat regions. To maximize the chance of robust variant calls and to

exclude potential single nucleotide polymorphisms (SNP), we considered only variants unique to each sample. Principal Component Analysis (PCA) was used to establish similarities between MuTect and Strelka calls with respect to the six SBS types and their 96 possible trinucleotide contexts.

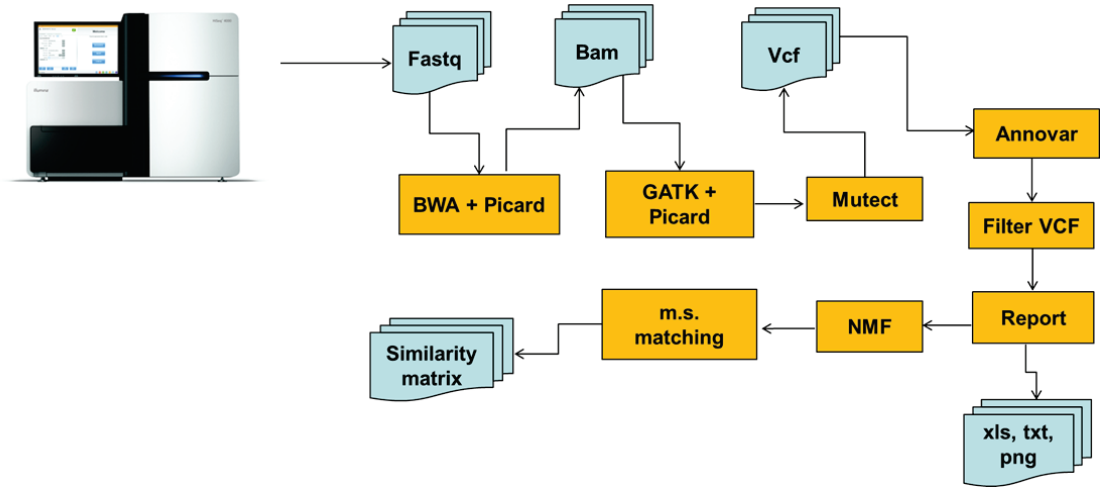


Figure 19: MutSpec Bioinformatics pipeline implemented in Galaxy interface.

17. Statistical analysis

The FactoMiner R package (R package version 3.3.2; https://cran.r-project.org/web/packages/FactoMineR) was used to perform PCA. Rainfall plots were generated using the Karyoplot R package (https://cran.r-project.org/web/packages/Karyoplot) used in (Nik-Zainal et al., 2015).

In order to perform the transcriptional strand bias (SB) analyses, *p*-values were calculated using Pearson's $\chi 2$ test. As multiple comparisons were assessed, the *p*-value was adjusted by applying a false discovery rate (FDR). Statistical analyses were carried out using the stats R package. The SB was considered statistically significant at $p\text{-value} \leq 0.05$.

To analyze sample mutation spectra and treatment-specific mutational signatures, filtered mutations were classified into 96 types corresponding to the six possible base substitutions (C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G, T:A>G:C) and the 16 combinations of flanking nucleotides immediately 5' and 3' of the mutated base. Mutation patterns were then deconvoluted into mutational signatures using the non-negative matrix factorization (NMF) algorithm of Brunet with the Kullback-Leibler divergence penalty (Alexandrov et al., 2013a; Brunet et al., 2004). We used the DNA damage estimator tool (as per (Chen et al., 2017); (https://github.com/Ettwiller/Damage-estimator)) to measure the Global Imbalance Value

(GIV) score and to exclude sequencing-related DNA damage and artefacts that can confound the determination of treatment-specific variants.

RESULTS

Objective 1: Development of mammalian cell models for exposure assays

During the course of conducting my doctoral thesis work, we focused on developing cellular exposure systems able to clonally expand and replicate observations from human primary cancers, namely, Hupki MEFs, HepaRG cells and LCL (Zhivagui et al., 2016. Publication attached as Appendix D).

1. Hupki MEF cells

Hupki MEF immortalization protocol was established allowing the cells to enter a biological barrier, the senescence, within 2 weeks of cell culture. Senescent cells have a decelerated cell division, visible in the growth curves (Figure 20A). In addition, we remark that the cells showed dramatic morphological changes compared to the primary cells manifested by an increased nuclear-cytoplasm ratio as well as the formation of multinuclear cells (Figure 20B). Histochemical staining of β-galactosidase was used as a marker for senescent cells and depicted in the middle image in Figure 20B. Furthermore, the cells were able to bypass this bottleneck step and clonally immortalize within 2-3 months of cell culture (Figure 20A). This was characterized by an accelerated cell division pace and changes in the cell morphology compared to the senescent cells (Figure 20B).

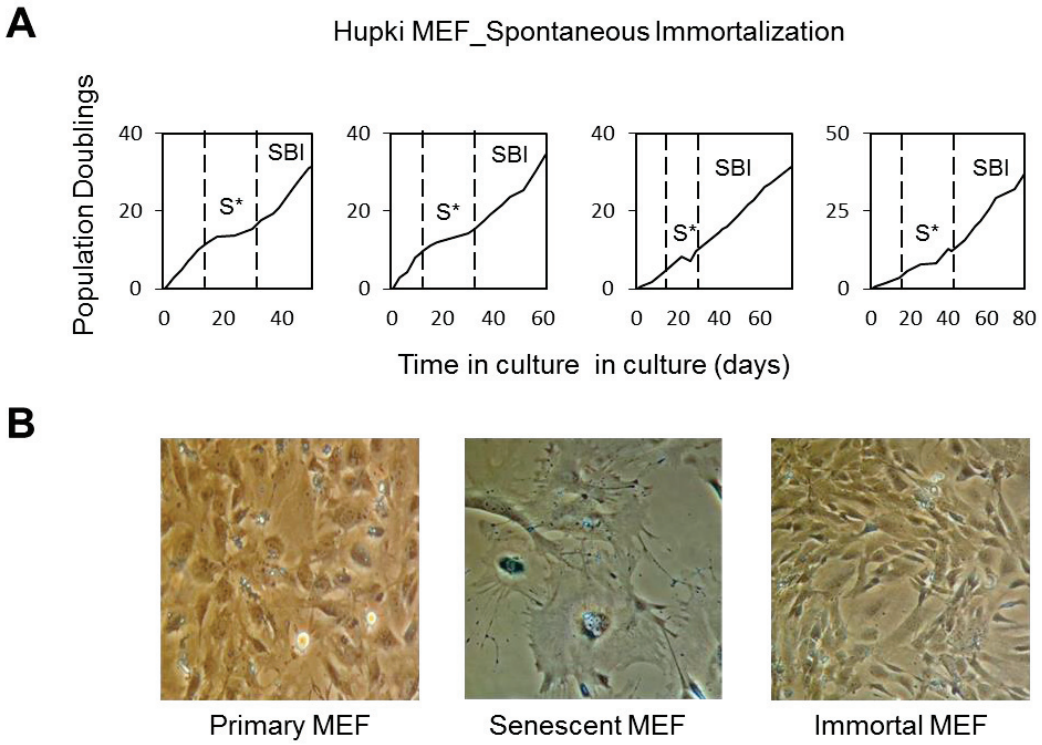


Figure 20: Hupki MEF immortalization. Growth curves representing the doubling population of control Hupki MEFs in prolonged cell culture (A). MEF cells underwent senescence (S*) reflected by a slower doubling population and cell morphology changes compared to the primary cells manifested by an increased nuclear-cytoplasm ratio as well as the formation of multinuclei (B). Histochemical staining of β-galactosidase can mark senescent cells (depicted in the middle image). Subsequently, Hupki MEFs bypassed senescence and propagated as immortal cell lines (SBI) characterized by an increased doubling population.

2. HepaRG cell model

Characterized by its versatility in culture, we aimed at establishing the best-possible protocol regarding simplicity and duration for exposure experiments using the HepaRG cell model. We designed a panel of strategies that take advantage of the unique biological properties of the HepaRG cells, thereby addressing their potential applicability to the MutSpec project (Figure 21). Two strategies were considered: single-cell subcloning (Figure 21A) and clonal outgrowth through potential crisis bypass of cells (Figure 21B).

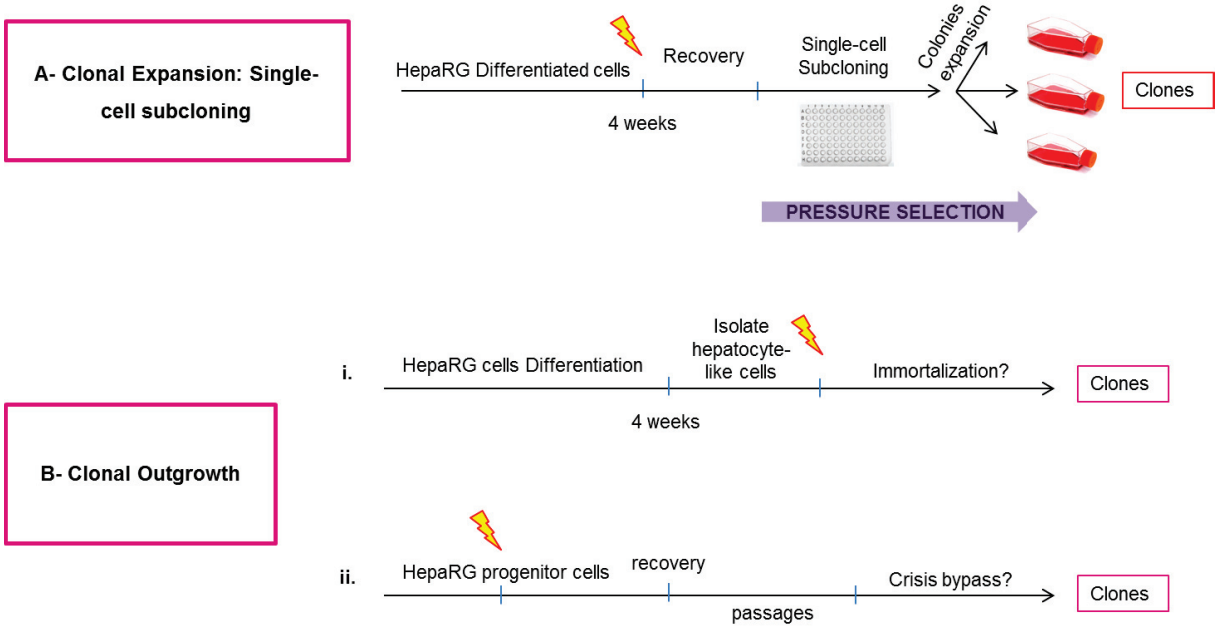


Figure 21: **Establishing the HepaRG cell system for exposure and clonal expansion assays**. (A) Single-cell subcloning: Differentiated HepaRG cells would be exposed to AA and, following dedifferentiation, maintained in culture as progenitor cells and single-cell subcloned. (B) Clonal outgrowth: test the ability of isolated hepatocyte cells to recover from carcinogen treatment and to clonally expand and immortalize (i). Progenitor HepaRG cells can be also exposed to a carcinogen and maintained in culture for recovery, crisis and crisis bypass leading to clonal outgrowth (ii).

2.1. Clonal Expansion assay: Single-cell subcloning

In respect to the **CE assay** (Figure 21A); we sought to test the ability of the cells to generate single-cell subclones at different stages of cell culture. Fully differentiated HepaRG cells were exposed to a carcinogen (AA) after which the cells transdifferentiated back to their progenitor-like origin. After recovery, treated and untreated cells were propagated in culture up to a point when we noticed a slow population doubling of the cells (Figure 22A) concomitant with dramatic changes in cellular morphology compared to normal progenitor cells manifested by an elongated cytoplasm (Figure 22B); this was suggestive of cells entering a crisis-like state (C*). Following continuous cell passaging, population doubling analysis showed an increase in growth rate (Figure 22A), suggesting that the cells may have bypassed crisis (CB). These cells were able to reach confluency within a week after serial dilutions.

Interestingly, once the cells overcame crisis, and only then, progenitor HepaRG cells were amenable to single-cell subcloning generating clones that were able to proliferate and expand.

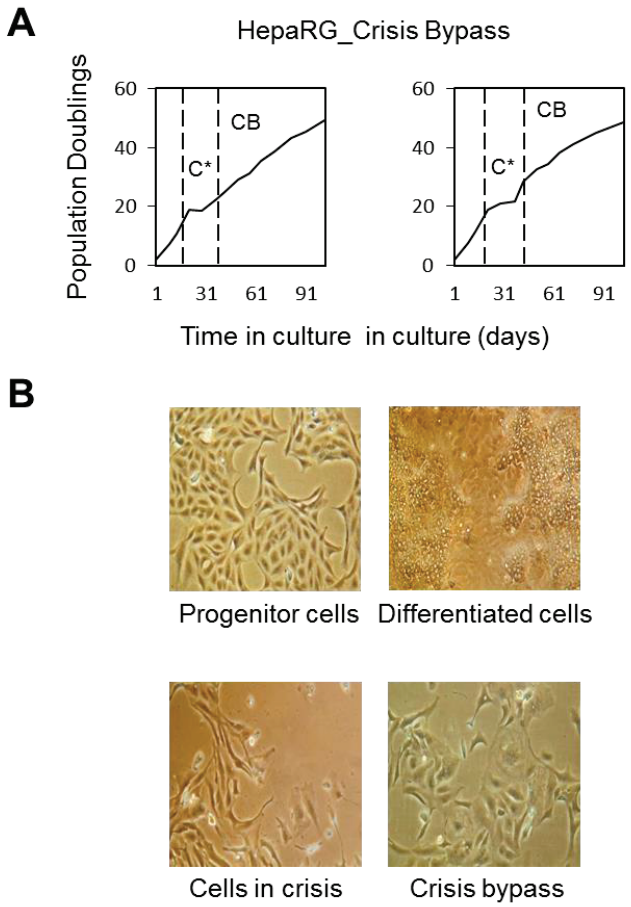


Figure 22: HepaRG cells in prolonged cell culture. The growth curves represent the population doubling of HepaRG cells in prolonged cell culture (A). Progenitor HepaRG cells (after transdifferentiation) underwent crisis (C*) reflected by the slower doubling population of the cells after one month of cell culture. Cells in crisis showed an altered cellular morphology compared to normal progenitor cells (B). Subsequently, the cells seemed to be able to circumvent the potential crisis barrier and restart their cell division, leading to an increase in their growth rate (CB). Following crisis bypass, the cells were able to generate single-cell subclones (shown in the last image).

2.2. Barrier-Bypass Clonal Expansion assay: Clonal outgrowth

With regard to the **BBCE assay** (Figure 21B), we tested two scenarios through which differentiated hepatocyte cells and progenitor cells might be able to bypass a bottleneck step (crisis) and lead to clonal cell populations.

2.2.1. Hepatocyte-like cells isolation

First, we developed a technique permitting efficient separation of the hepatocyte-like cells from the dual population of differentiated HepaRG cells. Given that the hepatocytes are smaller in size compared to the biliary cells, we tried to FACS sort the cells based on their size (Figure E.1). The isolated hepatocytes were able to re-attach to the wells, however, the number of hepatocytes thus isolated was not sufficient to maintain the cells at high density and preserve their differentiated state.

Second, as an alternative strategy, we based the hepatocyte isolation on their low cell anchorage and weak cell-matrix contacts (Cerec et al., 2007). Partial trypsinization was tested to allow the hepatocytes to detach first from the cell culture vessel, leaving mostly biliary cells behind (Figure 23). Visual inspection during the partial trypsinization step

suggested separation of the culture into hepatocytes and biliary cells (Figure 23 d-f). The isolated cells were seeded at high density to counteract transdifferentiation and were collected at different time-points to investigate metabolic functionality and activity of the isolated hepatocytes.

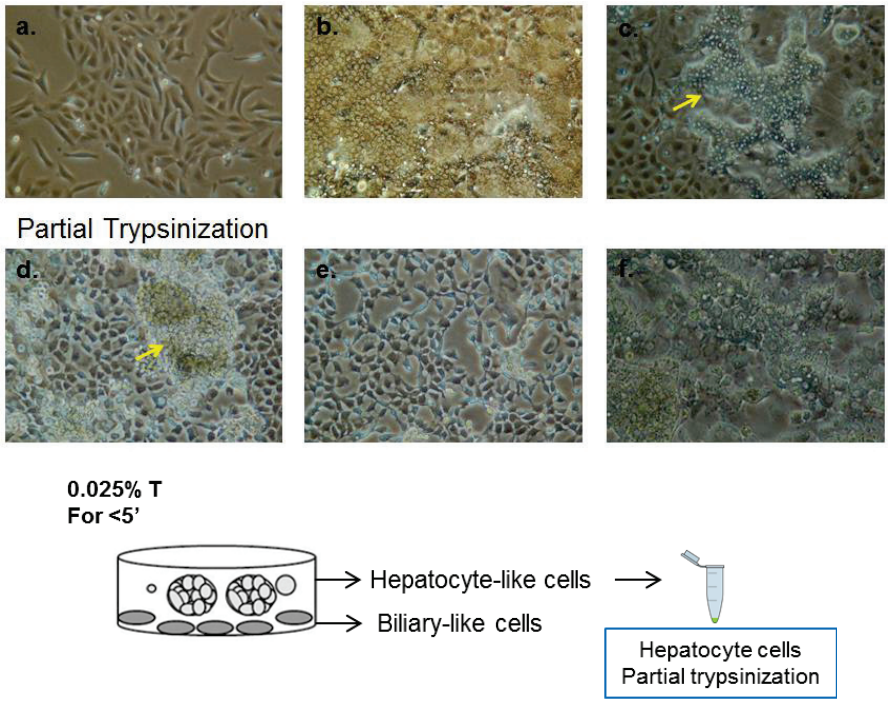


Figure 23: Real-time culture images and graphical representation of the partial trypsinization technique allowing efficient isolation of hepatocyte-like cells from the dual population in culture. HepaRG cells at different culture stages: (a) proliferating stage (3 days); (b) confluent stage (15 days); (c) differentiated stage (30 days) treated with 2% DMSO for 15 days, with the arrow pointing to hepatocyte islands; (d) Hepatocyte colonies under partial trypsinization treatment (<5 min), marked by the arrow; (e) the remaining attached biliary cells after hepatocyte collection; (f) pure hepatocytes suspension after 4 days of seeding in a 24 well-plate. The cells were either seeded at high confluency or collected as pellets to investigate the expression of a number of hepatic markers giving insights into the functionality of the hepatocytes.

This was achieved by measuring the level of expression of various metabolic enzymes, such as phase I (CYP3A4 and CYP2E1) and phase II enzymes (UGT1A1 and GSTA2), hepatocyte markers (albumin and aldolase) and a progenitor cell marker (CK-19) by qRT-PCR. Expression of phase I and phase II enzymes as well as the hepatic markers were highest after 4 to 7 days of seeding, suggesting that the hepatocyte population requires at least 4 days to re-arrange in a monolayer on the plate and re-gain their metabolic activity. However isolated hepatocytes left in culture for more than a week tend to lose their activity and convert into progenitor cells as evidenced by the decreased level of expression of hepatocyte markers and the increased expression of the progenitor marker (Figure 24).

Isolated hepatocyte-like cells can be exposed to a pro-carcinogen, allowing its metabolic activation, and upon recovery, differentiated cells might undergo senescence (Figure 9) after which some may acquire the ability to surpass this barrier step leading to the immortalization of hepatocytes (Figure 21B - i).

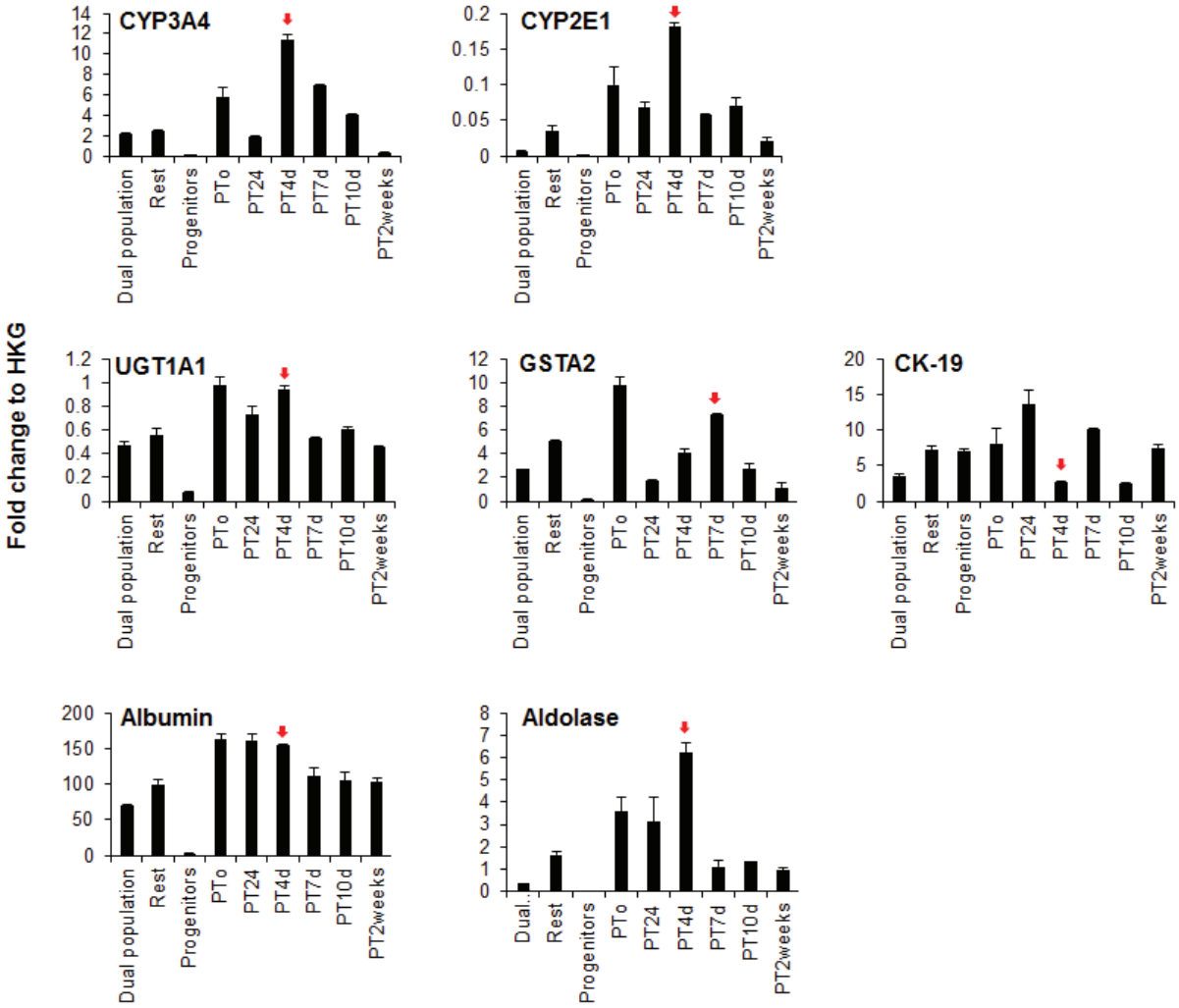


Figure 24: Evaluation of hepatocyte activity upon partial trypsinization (PT) by qRT-PCR for genes encoding for Phase I enzymes (*CYP3A4* and *CYP2E1*), Phase II enzymes (*UGT1A1* and *GSTA2*), hepatocyte markers (*Albumin* and *Aldolase*) and a progenitor cell marker (*CK-19*). The y-axis represents the fold-change normalized to the housekeeping genes *GAPDH, B2M, SFRS4* and *TBP*.

2.2.2. Exposure of progenitor bipotent cells

The ability of the progenitor cells to bypass crisis and clonally outgrow may follow the same trajectory as suggested by the growth curve results in Figure 22. In fact, progenitor cells were shown to bypass a potential crisis-related decline in growth after two months of cell culture

(Figure 22A). However, more work is needed to understand the state of the cells after bypassing crisis whether they became cancerous cells or remained progenitor cells. Interestingly, as illustrated in Figure 24, we discerned that the progenitor bipotent cells have significantly less to no metabolic activity compared to the fully differentiated, dual population of HepaRG cells. Therefore, the addition of human S9 fraction may be a critical strategy to boost compound metabolism in these cells (Figure 21B - ii).

Objective 2: Identification of cytotoxic and genotoxic effects of high priority compounds

1. Identification of the cytotoxic effect of high priority chemical agents

Upon exposure of primary Hupki MEF and HepaRG cells to a range of concentrations of high priority compounds, we observed dose-dependent cytotoxic effects of the various compounds (Figure 25). Hupki MEFs were exposed to acrylamide (ACR) (in the absence or presence of the S9 fraction) and its metabolite, glycidamide (GA), OTA (in the absence or presence of the S9 fraction), Cr(VI) and MNU. HepaRG cells were exposed to AA as different cell populations, namely progenitor cells, fully differentiated dual population of hepatocyte-like and biliary-like cells as well as isolated hepatocyte-like cells from partial trypsinization. The analysis informed the selection of the exposure conditions for the subsequent exposure/immortalization experiments, which was based on a 50% (range 30-70%) decrease in cell viability (Figure 25).

Exposure of primary Hupki MEFs showed cytotoxic effects across the employed compounds.

HepaRG cells were exposed to AA at different stages. We noticed that the progenitor cells were the least affected by AA treatment, except at high concentrations. Isolated hepatocyte-like cells from partial trypsinization were clearly affected by AA exposure starting at a concentration of 100uM. Fully differentiated HepaRG cells, consisting of the dual population of hepatocytes and biliary cells, showed the highest increase in cell death with a cytotoxic effect starting at 50uM.

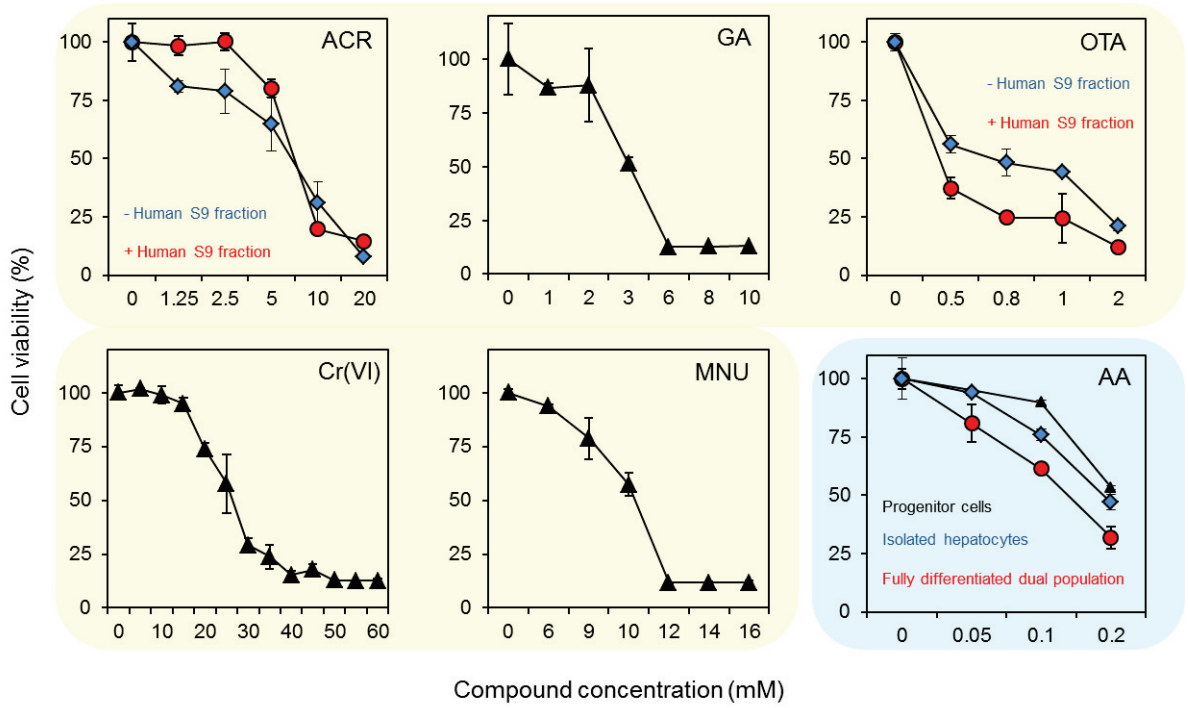


Figure 25: Compound-induced cytotoxicity *in vitro*. Panels in this figure represent the relative absorbance of formazan, indicative of cell viability determined by MTT assay, following 24-hour treatment of primary Hupki MEFs (in yellow) and HepaRG cells (in blue) with the indicated concentrations of chemical agents. The absorbance was measured 48 hours after treatment cessation and was normalized to the untreated cells. The results are expressed as mean percent ± standard deviation (SD) from three replicates.

2. DNA damage-dependent γH2Ax response to exposure to high priority compounds

In order to assess the genotoxicity of the tested compounds, γH2Ax immunofluorescence was carried out. Exposure to all compounds resulted in a marked increase in γH2Ax staining in the exposed Hupki MEFs, in comparison to the mock-treated control cells (Figure 26), suggesting a clear increase of DNA damage following exposure.

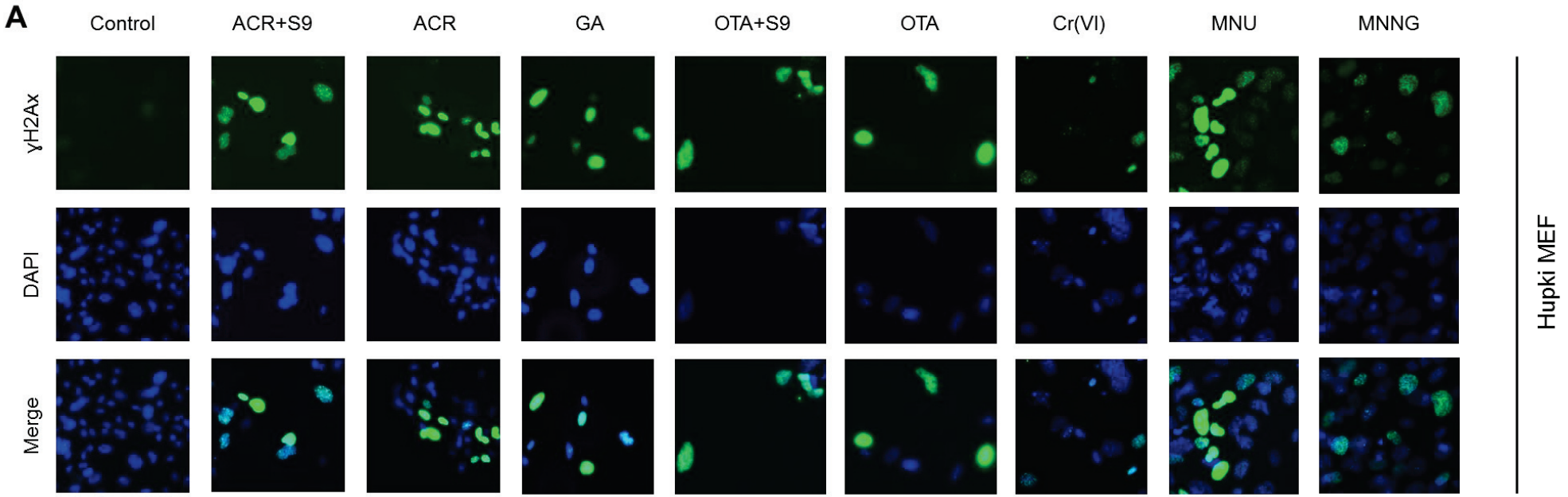


Figure 26: DNA damage assessment by immunofluorescence using a monoclonal antibody specific for Ser139-phosphorylated histone H2Ax (γH2Ax). Primary Hupki MEFs were treated with cytotoxicity-optimized concentrations for each compound for 24 hours prior to immunofluorescence. A negative-control of untreated cells was included as well as a positive-control of MNNG-treated primary MEFs.

Objective 3: Characterization of the mutational signatures specific to mutagens

Following compounds prioritization, we employed the well-established Hupki MEF system for exposure and clonal expansion assay in order to define the genome-wide mutational signatures of acrylamide and its metabolite, glycidamide (summarized in the attached manuscript: Paper 1), and OTA (described in the manuscript in preparation: Paper 2).

Paper 1: Summary of findings regarding the dietary compounds acrylamide and glycidamide

Title

Experimental analysis of exome-scale mutational signature of acrylamide and its metabolite glycidamide

Authors: Maria Zhivagui, Maude Ardin, Stephanie Villar, Mona I. Churchwell, Vincent Cahais, Alexis Robitaille, Liacine Bouaoun, Adriana Heguy, Kathryn Guyton, James McKay, Monica Hollstein, Magali Olivier, Frederick A. Beland, Michael Korenjak and Jiri Zavadil

Under review, Carcinogenesis, 2017

Aim: Identify the genome-wide mutational signatures of acrylamide and its metabolite glycidamide

Approach: Primary Hupki MEF exposed to acrylamide and glycidamide were used to generate immortalized clones harboring specific types of somatic mutations characteristic of the tested compounds. Protein-coding DNA sequencing coupled with sophisticated mathematical algorithms was used to define the putative mutational signature of acrylamide and glycidamide (Figure 27).

Graphical summary

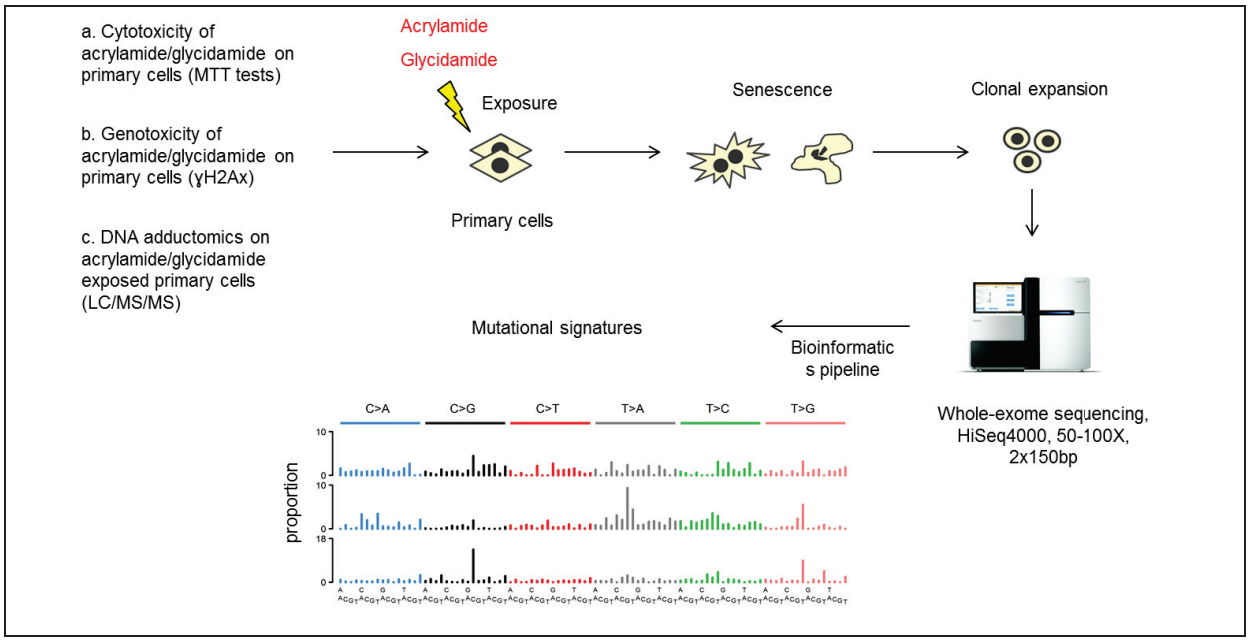


Figure 27: **Study design**. Compound-induced cytotoxicity and genotoxicity was assessed in primary cells. Hupki MEFs were exposed to acrylamide and glycidamide and maintained through senescence bypass and clonal expansion. Whole-exome sequencing followed by data analysis allowed extraction of putative mutational signatures specific to the tested compounds.

Novelty and highlights: We identified a novel mutational signature of acrylamide exerted by its metabolite glycidamide that is unique and characterized by an increased frequency of T:A>A:T, T:A>C:G and C:G>A:T mutations. Given the ubiquitous presence of acrylamide in our environment, reaching beyond occupational settings into daily life through long-term exposure to dietary acrylamide, the newly identified mutational signature can serve as a powerful tool in epidemiological studies for cancer risk assessment of acrylamide and possibly establish a link between exposure to acrylamide and cancer development.

Paper 1: Full manuscript, under review in Carcinogenesis journal

Title

Experimental analysis of exome-scale mutational signature of glycidamide, the reactive metabolite of acrylamide

Authors

Maria Zhivagui[1], Maude Ardin[1], Alvin W. T. Ng[2,3,4], Mona I. Churchwell[5], Manuraj Pandey[1], Stephanie Villar[1], Vincent Cahais[6], Alexis Robitaille[7], Liacine Bouaoun[8], Adriana Heguy[9], Kathryn Guyton[10], Martha R. Stampfer[11], James McKay[12], Monica Hollstein[1,13,14], Magali Olivier[1], Steven G. Rozen[2,3], Frederick A. Beland[5], Michael Korenjak[1] and Jiri Zavadil[1]

Affiliations

[1] Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer, Lyon 69008, France

[2] Centre for Computational Biology, Duke-NUS Medical School, Singapore 169857, Singapore

[3] Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, 169857, Singapore

[4] NUS Graduate School for Integrative Sciences and Engineering, 117456, Singapore

[5] Division of Biochemical Toxicology, National Center for Toxicological Research, Jefferson, AR 72079, USA

[6] Epigenetics Group, International Agency for Research on Cancer, Lyon 69008, France

[7] Infections and Cancer Biology Group, International Agency for Research on Cancer, Lyon 69008, France

[8] Environment and Radiation Section, International Agency for Research on Cancer, Lyon 69008, France

[9] Department of Pathology and Genome Technology Center, New York University, Langone Medical Center, New York, NY 10016, USA

[10] IARC Monographs Section, International Agency for Research on Cancer, Lyon 69008, France

[11] Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

[12] Genetic Cancer Susceptibility Group, International Agency for Research on Cancer, Lyon 69008, France

[13] Deutsches Krebsforschungszentrum, 69120 Heidelberg, Germany

[14] Faculty of Medicine and Health, University of Leeds, LIGHT Laboratories, Leeds LS2 9JT, United Kingdom

Keywords: Acrylamide, glycidamide, DNA adducts, massively parallel sequencing, mutational signatures

Correspondence:

ZavadilJ@iarc.fr and/or KorenjakM@iarc.fr

Abstract

Acrylamide, a probable human carcinogen, is ubiquitously present in the human environment, with sources including heated starchy foods, coffee and cigarette smoke. Humans are also exposed to acrylamide occupationally. Acrylamide is genotoxic, inducing gene mutations and chromosomal aberrations in various experimental settings. Covalent haemoglobin adducts were reported in acrylamide-exposed humans and DNA adducts in experimental systems. The carcinogenicity of acrylamide has been attributed to the effects of glycidamide, its reactive and mutagenic metabolite capable of inducing rodent tumors at various anatomical sites. In order to characterize the pre-mutagenic DNA lesions and global mutation spectra induced by acrylamide and glycidamide, we combined DNA-adduct and whole-exome sequencing analyses in an established exposure-clonal immortalization system based on mouse embryonic fibroblasts. Sequencing and computational analysis revealed a unique mutational signature of glycidamide, characterized by predominant T:A>A:T transversions, followed by T:A>C:G and C:G>A:T mutations exhibiting specific trinucleotide contexts and significant transcription strand bias. Computational interrogation of human cancer genome sequencing data indicated that a combination of the glycidamide signature and an experimental benzo[*a*]pyrene signature are nearly equivalent to the COSMIC tobacco-smoking related signature 4 in lung adenocarcinomas and squamous cell carcinomas. We found a more variable relationship between the glycidamide- and benzo[*a*]pyrene-signatures and COSMIC signature 4 in liver cancer, indicating more complex exposures in the liver. Our study demonstrates that the controlled experimental characterization of specific genetic damage associated with glycidamide exposure facilitates identifying corresponding patterns in cancer genome data, thereby underscoring how mutation signature laboratory experimentation contributes to the elucidation of cancer causation.

A 40-word summary

Innovative experimental approaches identify a novel mutational signature of glycidamide, a metabolite of the probable human carcinogen acrylamide. The results may elucidate the cancer risks associated with exposure to acrylamide, commonly found in tobacco smoke, thermally processed foods and beverages.

Introduction

Cancer can be caused by chemicals, complex mixtures, occupational exposures, physical agents, and biological agents, as well as lifestyle factors. Many human carcinogens show a number of characteristics that are shared among carcinogenic agents (1). Different human carcinogens may exhibit a spectrum of these key characteristics, and operate through separate mechanisms to generate patterns of genetic alterations. Recognizable patterns of genetic alterations or mutational signatures characterize carcinogens that are genotoxic. Recent work shows that these DNA sequence changes can be expressed in simple mathematical terms that enable mutational signatures to be extracted from thousands of cancer genome sequencing data sets (2). Several of the over 30 identified mutational signatures have been attributed to specific external exposures or endogenous factors through epidemiological and experimental studies (2). However, about 40% of the current signatures remain of unknown origin, and additional, thus far unrecognized, signatures are likely to be defined in rapidly accumulating cancer genome data. Well-controlled experimental exposure systems can thus help identify the underlying causes of known orphan mutational signatures as well as define new patterns generated by candidate carcinogens (reviewed in (3,4)).

Various diet-related exposures contribute to the human cancer burden. Examples include contaminants in food or alternative medicines, such as aflatoxin B1 (AFB1) or aristolochic acid (AA). The mutagenicity of these compounds is well-documented; AFB1 induces predominantly C:G>A:T base substitutions and AA causes T:A>A:T transversions. The characteristic mutations coupled with information on the preferred sequence contexts in which they are likely to arise allowed unequivocal association of exposure to AFB1 or AA with specific subtypes of hepatobiliary or urological cancers, respectively (5-13).

Among dietary compounds with carcinogenic potential, acrylamide is of special interest due to extensive human exposure. Important sources of exposure to acrylamide include tobacco smoke (14), coffee (15), and a broad spectrum of occupational settings (16). Dietary sources of acrylamide comprise carbohydrate-rich food products that have been subject to heating at high temperatures. This is due to Maillard reactions, which involve reducing sugars and the amino acid asparagine, present in potatoes and cereals (17). There is sufficient evidence that acrylamide is carcinogenic in experimental animals (18,19) and it has been classified as a probable carcinogen (Group 2A) by the International Agency for Research on Cancer in 1994 (16). The association of dietary acrylamide exposure with renal, endometrial and ovarian cancers has been explored in recent epidemiological studies (20,21). However, accurate acrylamide exposure assessment in epidemiological studies based on questionnaires has been difficult, and more direct measures of molecular markers,

such as hemoglobin adduct levels, may not yield conclusive findings on past exposures (22-27). An improved understanding of its mechanism of action using well-controlled experimental systems is critical for understanding the potential carcinogenic risk associated with exposure.

Acrylamide undergoes oxidation by cytochrome P450, producing the reactive metabolite glycidamide that is highly efficient in DNA binding due to its electrophilic epoxide structure (28-30). The *Hras* mutation load in neoplasms of mice exposed to acrylamide or glycidamide was found to be considerably higher in mice treated with glycidamide (31). This finding is corroborated by a considerably higher mutation frequency in the *cII* reporter gene of Big Blue mouse embryonic fibroblasts treated with glycidamide in comparison to acrylamide (32,33). Mutation analysis in different experimental *in vivo* and *in vitro* models using reporter genes showed an increased association of acrylamide and glycidamide exposure with T:A>C:G transitions, as well as T:A>A:T and C:G>G:C transversion mutations (31-36), whereas glycidamide exposure was also characterized by C:G>A:T transversions (33). However, these proposed acrylamide- and glycidamide-specific mutation patterns were based on limited mutation counts in reporter genes and thus do not reflect the complexity of genome-wide distributions and profiles. Based on the limited data available thus far, it is not possible to translate adequately the reported mutation types (T:A>C:G, T:A>A:T, C:G>G:C, C:G>A:T) to global alteration patterns.

The advent of massively parallel sequencing has created the opportunity to study a large number of mutations in a single sample, thus significantly enhancing the power of mutation analysis in experimental models and enabling reliable identification of specific sequence contexts for the induced alterations. Analogously to human cancer genome projects, genome-scale mutational signatures can be extracted from highly controlled carcinogen exposure experiments using mammalian cell and animal models coupled with advanced mathematical approaches (2,3,37,38).

Here we report the systematic assessment of acrylamide and glycidamide mutagenicity based on DNA adduct formation and mutation profile analysis using massively parallel sequencing in a cell model amenable to the analysis of carcinogen-induced mutation patterns and their impact on the resulting cell phenotype (3,37-39). We identify a specific and robust mutational signature attributable to glycidamide, and by computationally interrogating human cancer genome-wide mutation data, we characterize glycidamide signature-positive tumors, thereby highlighting a potential contribution of acrylamide/glycidamide exposure to carcinogenesis in humans.

Materials and methods

Source and authentication of primary cells

Primary Human-p53 knock-in mouse embryonic fibroblasts (Hupki MEFs) were isolated from 13.5-day old *Trp53*$^{tm/HolI}$ mouse embryos from the Central Animal Laboratory of the Deutsches Krebsforschungszentrum, Heidelberg, as described previously (40). The mice had been tested for Specific Pathogen-Free (SPF) status. The derived primary cells were genotyped for the human *TP53* codon 72 polymorphism (Table 1) to authenticate the embryo of origin. Cells from three different embryos (E210, E213 and E214) were used for the exposure experiments (Table 1). All subsequent cell cultures were routinely tested at all stages for the absence of mycoplasma.

Cell culture, exposure and immortalization

The primary MEF cells were expanded in Advanced DMEM supplemented with 15% fetal calf serum, 1% penicillin/streptomycin, 1% pyruvate, 1% glutamine, and 0.1% β-mercapto-ethanol. The cells were then seeded in six-well plates and, at passage 2, exposed for 24 hours to acrylamide (A4058, Sigma), glycidamide (04704, Sigma), or vehicle (PBS). Acrylamide exposure was carried out in the absence or presence of 2% human S9 fraction (Life Technologies) complemented with NADPH (Sigma). Exposed and control primary cells were cultivated until they bypassed senescence and immortalized clonal cell populations could be isolated (41). The human mammary epithelial cell (HMEC) cultures utilized in this study for whole-genome sequencing (WGS) were generated from benzo[*a*]pyrene (B[a]P) exposed HMEC described previously (42,43).

MTT assay for cell metabolic activity and viability

Cells were seeded in 96-well plates and treated as indicated. Cell viability was measured 48 hours after treatment cessation using CellTiter 96® Aqueous One solution Cell Proliferation Assay (Promega). Plates were incubated for 4 hours at 37°C and absorbance was measured at 492 nm using the APOLLO 11 LB913 plate reader. The MTT assay was performed in triplicates for each experimental condition.

γH2Ax Immunofluorescence

Immunofluorescence staining was carried out using an antibody specific for Ser139-phosphorylated H2Ax (γH2Ax) (9718, Cell Signaling Technology). Primary MEFs were seeded on coverslips in 12 well-plates. The cells were incubated in with γH2Ax-antibody (1:500 in 1% BSA) at 4°C overnight. Subsequent incubation with a fluorochrome-conjugated secondary antibody (4412, Cell Signaling Technology) was carried out for 60 minutes at

room temperature. Coverslips were mounted in Vectashield mounting medium with DAPI (Eurobio). Immunofluorescence images were captured using a Nikon Eclipse Ti.

DNA adduct analysis

Glycidamide-DNA adducts (N7-(2-carbamoy-2-hydroxyethyl)-guanine (N7-GA-Gua) and N3-(2-carbamoy-2-hydroxyethyl)-adenine (N3-GA-Ade)) were quantified by liquid chromatography-mass spectrometry (LC-MS/MS) with stable isotope dilution as previously described (44) (see Supplementary Materials and Methods for details). The LC-MS/MS used for quantification consisted of an Acquity UPLC system (Waters) and a Xevo TQ-S triple quadrupole mass spectrometer (Waters). The same MRM transitions as previously described (44) were monitored with a cone voltage of 50V and collision energy of 20eV for each adduct transition and its corresponding labeled isotope transition.

***TP53* genotyping**

Exons 4 to 8 of the knocked-in human *TP53* gene (NC_000017.11) were sequenced using standard protocols. Sanger sequencing of PCR products was performed at Biofidal (Lyon, France). *TP53* primer sequences are listed in Supplementary Materials and Methods. Resulting sequences were analyzed using the CodonCode Aligner software.

Library preparation and whole-exome sequencing (WES)

Library preparation was carried out using the Kapa Hyper Plus library preparation kit (Kapa Biosystems) according the manufacturer's instructions. Exome capture was performed using the SureSelect XT Mouse All Exon Kit (Agilent Technologies). Eighteen exome-captured libraries were sequenced in the paired-end 150 base-pair run mode using the Illumina HiSeq4000 sequencer.

Processing of WES data

Fastq files were analyzed for data amount and quality using FastQC (0.11.3) and were processed with an in-house pipeline for adapter trimming and alignment to the mm10 genome (release GRCm38). These components of the pipeline are publicly available at https://github.com/IARCbioinfo/alignment-nf. The resulting alignment files had a mean depth-of-coverage of 135 and 175 for acrylamide and glycidamide samples, respectively. All alignment files can be accessed from the NCBI Sequence Read Archive (SRA) data portal under the BioProject accession number PRJNA238303. Two somatic variant callers were employed with default parameters in order to detect single base substitutions (SBS) and small insertions/deletions (indels) (MuTect 1.1.6-4 and Strelka 1.015) in exposed clones, using primary cells as normal samples. Each immortalized clone was compared to primary

MEFs from three different embryos (conditions Prim_1, Prim_2, and Prim_3). The overlap of the variant calling outcome with respect to the different primary MEFs showed concordance close to 80% (Suppl. Fig. S1) with MuTect exhibiting more stringent calling performance. Thus, mutation data obtained from the MuTect variant caller were further processed with the MutSpec suite ((45); https://github.com/IARCbioinfo/mutspec). For more details, see Supplementary Materials and Methods and the summary of sequencing metrics (Suppl. Table S1), the list of identified MuTect SBS variants (Suppl. Table S2) and indels (Suppl. Table S3).

Bioinformatics and statistical analyses

The FactoMiner R package (R package version 3.3.2; https://cran.r-project.org/web/packages/FactoMineR) was used to perform the principal component analysis (PCA). To perform the transcription strand bias (SB) analyses, *p*-values were calculated using Pearson's $\chi 2$ test. As multiple comparisons were assessed, the *p*-value was adjusted by applying a false discovery rate (FDR). Statistical analyses were carried out using the stats R package. The SB was considered statistically significant at *p*-value ≤ 0.05. To analyze samples mutation spectra and treatment-specific mutational signatures, filtered mutations were classified into 96 types corresponding to the six possible base substitutions (C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G, T:A>G:C) and the 16 combinations of flanking nucleotides immediately 5' and 3' of the mutated base. Mutation patterns were then deconvoluted into mutational signatures using the non-negative matrix factorization (NMF) algorithm (46,47). The reconstruction error calculation evaluated the accuracy with which the deciphered mutational signatures describe the original mutation spectra of each sample by applying Pearson correlation and cosine similarity.

In order to clean up the profile of the glycidamide mutational signature from the residual signature 17 signal and to increase the stability of NMF decomposition, we supplied the NMF input by adding samples with a high level of signature 17 (over 65% contribution as determined by independent NMF analysis, see Supplementary Materials and Methods).

Cosine similarity analysis was used to evaluate the concordance of the newly identified T:A>A:T-rich mutational signature of glycidamide with the previously reported mutational signatures characterized by a predominant T:A>A:T content. These comprised COSMIC signatures 22 (AA), 25 and 27 (both of unknown etiology(2)), the experimentally derived mutational signature of AA (37,45), 7,12-dimethylbenz[*a*]anthracene (DMBA) (48,49), and urethane (50).

We employed the mutational signature activity (mSigAct) software's sparse signature assignment function (sparse.assign.activity) (13) to assess the presence of the experimental

mutational signatures of glycidamide and benzo[*a*]pyrene in whole-genome somatic mutation data from 38 lung adenocarcinomas, 48 lung squamous carcinomas, and 320 liver cancers from the ICGC Pan-Cancer Analysis of Whole Genomes (PCAWG) study. We excluded 244 hyper-mutated microsatellite unstable and aristolochic acid signature-containing liver tumors as the presence of high numbers of T>A mutations adversely prevented assessment of the possible presence of the glycidamide signature. A set of 11 active COSMIC mutational signatures were identified in the remaining tumor samples (excluding COSMIC signature 4).

We defined a 'pure' experimental C>N benzo[*a*]pyrene signature by WGS (using Illumina HiSeq4000 by Genewiz, NJ, USA) of finite lifespan post-stasis clones derived from primary human mammary epithelial cells (HMEC) treated with B[a]P as previously described (42,43,51). The read alignment to NCBI GRCh38 genome build, variant calling, filtering and annotation were consistent with the MutSpec pipeline described above (45). Proportion matrices of the experimental GA-signature, the GA-signature normalized to the human genome trinucleotide frequency to allow for human PCAWG data screening, and the whole-genome B[a]P signature are available in Suppl. Table S4.

Results

Acrylamide and glycidamide induce cytotoxic and genotoxic responses in Hupki MEFs

Upon exposure of primary Hupki MEFs to a range of concentrations of acrylamide (ACR) (in the absence or presence of the S9 fraction) and its metabolite, glycidamide (GA), we observed a dose-dependent cytotoxic effect on the cells for either compound (Fig. 1A). This analysis informed the selection of two conditions for the ACR exposure to be used in the subsequent exposure/immortalization experiments, 10 mM ACR for 24 hours in the absence of human S9 fraction, and 5 mM ACR for 24 hours in the presence of S9 fraction, which elicited 50% (range 30-70%) decrease in cell viability. The IC50 condition for GA was used for subsequent mutagenesis analysis, corresponding to a 24-hour treatment with 3 mM of the compound. The genotoxic effects of either ACR or GA manifested by a marked increase in γH2Ax staining in the exposed cell populations, in comparison to the mock-treated control cells (Fig. 1B).

Immortalized MEF cells accumulate *TP53* mutations following acrylamide or glycidamide treatment

Primary MEF cultures from three different embryos (Prim_1, Prim_2, and Prim_3) were exposed to ACR or GA using the established conditions and multiple immortalized clones were derived. MEF senescence and immortalization phases were evident from the growth curves generated for each culture (Suppl. Fig. S2). Subsequently, the clones derived from ACR exposure (ACR clones) and GA exposure (GA clones) and spontaneous

immortalization (Spont), were pre-screened for *TP53* mutations by Sanger sequencing, to assess the mutagenic process prior to exome-scale analysis. In the context of ACR treatment, clones obtained from the Prim_2 MEFs that were heterozygous for the polymorphic site in codon 72 showed a loss of heterozygosity involving a loss of the proline allele in the ACR_1 clone whereas the arginine allele was lost in ACR_2, giving rise to a hemizygous clone (Table 1). No *TP53* mutations were observed in any of the three Spont clones, whereas 3 out of 7 ACR clones and 1 of 5 GA clones carried non-synonymous *TP53* mutations (Table 1). The detected mutations indicated specific selection for mutations in the *TP53* gene during cell immortalization and confirmed the clonal nature of MEF immortalization.

Analysis of mutation spectra

Whole-exome sequencing of all spontaneously immortalized and exposed clones and subsequent extraction of acquired variants revealed that the total number of acquired SBS did not differ markedly between the ACR and Spont clones. The Spont clones harbored on average 190 (median = 151, range = 141-277) SBS, whereas the ACR clones had on average 208 (median = 173, range = 151-262) SBS. In contrast, the total number of SBS was considerably increased in the GA clones, with an average of 485 SBS (median = 448, range = 370-592) (Suppl. Table S1 and S2). This finding suggests markedly stronger mutagenic properties of GA in the MEFs. To estimate the extent of sequencing-related damage in our samples, we determined the GIV score of each sample as described in Materials and Methods and in (52). No detectable damage for any of the mutation types was observed in our dataset (data not shown). The ACR exposed samples exhibited an overall diffuse pattern across the six different SBS types (Suppl. Fig. S3). The Spont clones showed an enrichment of C:G>G:C SBS in the 5'-GCC-3' context, which was also present at varying levels in the exposed cultures. This particular mutation type appears to be related to the culture conditions used for the immortalization assay, as its presence has previously been noted upon spontaneous as well as exposure-driven MEF immortalization (37). No significant transcription strand bias was observed for any of the mutation classes in the Spont or ACR clones (Suppl. Fig. S4). In the five clones derived from the GA-treated primary MEF cultures, we observed an enrichment of acquired T:A>A:T and C:G>A:T transversions and T:A>C:G transitions (Suppl. Fig. S3B), marked by significant transcription strand bias (Suppl. Fig. S4).

PCA performed on the resulting 6-class SBS spectra unambiguously separated the GA clones from the remaining experimental conditions (Fig. 2A). The analysis of indels (listed in Suppl. Table S3) showed lower numbers of these alterations in the GA-associated clones compared to the ACR or Spont clones (Fig. 2B). This suggests that a higher

accumulation of SBS may selectively promote the senescence bypass and selection of the GA clones, with a decreased functional contribution of indels, while an inverse scenario is plausible in case of the Spont and ACR clones, reminiscent of a previous report based on the Big Blue mouse embryonic fibroblasts and c*II* transgene (53).

Variant allele frequency analysis

Variant allele frequency (VAF) analysis was carried out for GA clones. Overall, a significant proportion of acquired mutations was present at allelic frequencies between 25-75% (Suppl. Fig. S5). Upon grouping of substitutions into bins of high (67-100%), medium (34-66%) and low (0-33%) VAF, the predominant GA-specific mutation types (T:A>A:T, T:A>C:G and C:G>A:T) started manifesting at high VAF, whereas the 5'-NTT-3' alterations, corresponding to the COSMIC signature 17 previously reported to arise in cultured mouse cells including MEFs (38,54,55) showed lower VAF, therefore a later appearance in the cultures (Suppl. Fig. S6). This observation suggests the early effects of the GA exposure and the reproducible contribution of the induced mutations to the senescence bypass and their clonal propagation during the immortalization stage.

Mutational signature analysis

Using NMF, we extracted the mutational signatures from all the MEF clones. Using computed statistics for estimating the number of signatures, three signatures were identified as an optimal number, with signatures A and C enriched in the Spont and ACR clones, and signature B selectively enriched in the GA clones (Fig. 2C,D). Reconstruction of the observed mutation spectra supports the robustness of the signature analysis with strong Pearson's correlation and cosine similarity in GA-derived clones (Fig. 2D). In signature C and also to a lesser extent in signatures A and B, we observed an admixture of a pattern identical to the orphan COSMIC signature 17 (T:A>G:C in a 5'-NTT-3' trinucleotide context), described in various human cancers (most notably esophageal adenocarcinoma), but also seen in aflatoxin B1-driven mouse liver cancers (11), as well as primary MEF-derived clones (37,38). In *in vitro* contexts, this signature has been linked to cell culture conditions and associated oxidative stress (54,55). To refine further the obtained experimental signatures, we developed a signature 'baiting' approach that combined the MEF clones data with signature 17-rich data from esophageal adenocarcinomas from the ICGC ESAD-UK study for new NMF analysis (56). This resulted in considerable reduction (average = 47%, median = 48%) of the signature 17-specific most prominent T>G peaks and a more refined pattern for signature B, associated primarily with GA treatment (Fig. 3A and Suppl. Fig. S7). This putative GA signature retains the predominant enrichment for the T:A>A:T transversions and

T:A>C:G transitions in the 5'-CTG-3' and 5'-CTT-3' trinucleotide contexts, and the C:G>A:T component. Moreover, these mutation types were marked by significant transcription strand bias (Fig. 3B and Suppl. Fig. S4), exhibiting higher accumulation of mutations on the non-transcribed strand consistent with the decreased efficiency of the transcription-coupled nucleotide excision repair due to adduct formation.

DNA adduct analysis

Following metabolic activation, acrylamide induces well-characterized glycidamide DNA adducts at the N7- and N3-positions of guanine and adenine, respectively. LC-MS/MS-based adduct quantification revealed the absence of these adducts in the spontaneously immortalized control samples as well as in MEFs exposed to acrylamide in the absence of S9 fraction (levels below the limit of detection). This suggests the lack of CYP2E1 activity, which is required for the metabolism of acrylamide to glycidamide, in the MEFs. Upon addition of human S9 fraction, N7-GA-Gua levels increased to 11adducts/10^8 nucleotides, suggesting limited metabolic activation of acrylamide due to the presence of enzymatic activity in the S9 fraction (Fig. 3C and Suppl. Fig. S8). Glycidamide-exposed cells exhibited significantly increased DNA adduct levels, with both N7-GA-Gua and N3-GA-Ade observed at very high average levels, 49 000 adducts/10^8 nucleotides and 350 adducts/10^8 nucleotides, respectively, after subtracting the trace amount of contamination from the internal standard (Fig. 3C and Suppl. Fig. S8).

Comparison of the glycidamide signature to known signatures characterized by prominent T:A>A:T profiles

We next performed cosine similarity analysis of the putative GA signature and all known T:A>A:T-rich signatures extracted from primary cancers as well as experimental systems (Fig. 3D and Suppl. Fig. S9). The best match was 84% pattern similarity with COSMIC signature 25 (derived from four Hodgkin lymphoma cell lines) (Fig. 3D). However, unlike the GA signature, COSMIC signature 25 exhibits strand bias for only T:A>A:T mutations and no transcription strand bias for the T:A>C:G mutations. Thus, the mutation patterns and strand bias on all three main mutation types generated by GA treatment (Fig. 3A,B) appear specific and novel.

Glycidamide signature screening in human tumor data from the ICGC PCAWG

The initial mSigAct test performed on PCAWG data from lung and liver tumors indicated a marked presence of the GA signature. This observation was in keeping with the presence of acrylamide in tobacco smoke and was further corroborated by a cosine similarity of 94% between the adenine (T>N) components of COSMIC signature 4 (tobacco smoking) and the

GA signature (Fig. 4A). We thus hypothesized that COSMIC signature 4 reflects co-exposure to B[a]P (generating C>N/guanine mutations with transcription strand bias) and to GA (generating T>N/adenine mutations with transcription strand bias) (Fig. 4A,B). To provide further experimental evidence, we generated a 'pure' B[a]P mutational signature by whole-genome sequencing of cell clones derived from B[a]P-exposed normal human mammary epithelial cells (HMEC). This yielded a robust signature characterized by predominant strand biased guanine (mainly C>A) mutation levels and negligibly mutated adenines (T>N) (Fig. 4A,B). Next, we used mSigAct to interrogate the PCAWG tumor samples for the level of exposure to the experimentally defined GA and B[a]P signatures (alongside other COSMIC mutational signatures) in 48 lung squamous carcinomas, 38 lung adenocarcinomas, and 320 liver cancers. We compared these to estimated levels of exposure to COSMIC signature 4, and found that in the lung cancers, a combination of the GA and B[a]P signatures accounted for very similar numbers of mutations as COSMIC signature 4, thus further supporting the hypothesis that COSMIC signature 4 represents combined and highly correlated exposure to GA and B[a]P (Fig. 4C). Compared to lung cancers, we found more variability in the assignment of mutation numbers to GA and B[a]P versus COSMIC signature 4 in liver cancers (Fig. 4C), which may reflect a decreased relationship between GA and B[a]P exposure due to generally more complex exposure history in the liver. The successful reconstruction of COSMIC signature 4 by the experimental GA- and B[a]P- signatures in the lung and liver human tumors enabled correct assignment of the GA-signature in a subset of 29 lung adenocarcinomas, 46 lung SCC and 26 liver tumors (Fig. 4D). The SBS counts corresponding to GA-mutational signature ranged between 300 up to 43,000 mutations/per sample in lung tumors, and between 190 to 23,000 mutations/per sample in liver tumors (Fig. 4D and Suppl. Table S5). These findings indicate exposure to glycidamide linked to tobacco smoking – when concomitant with B[a]P-signature, or through diet or occupation – in the absence of B[a]P signature (samples Liver-HCC::SP112224; Liver-HCC::SP49551; Liver-HCC::SP50105; Liver-HCC::SP98861; Liver-HCC::SP50183, see Suppl. Fig. S10 and Suppl. Table S5).

Discussion

In this study we report the identification of an exome-wide mutational signature for glycidamide, a metabolite of the probable human carcinogen acrylamide. The newly identified signature is based on massively parallel sequencing performed in a well-controlled experimental carcinogen exposure-clonal immortalization model, revealing characteristic mutagenic effects of glycidamide. The glycidamide mutational signature presented here and the results of statistical assessment of its presence in multiple human tumor types may help clarify the thus-far tenuous association of acrylamide with human cancer.

In concordance with its *in vivo* carcinogenicity in rodents (16,19,31,57), our findings in the established MEF carcinogen exposure and immortalization system suggest that characteristic mutagenic effects may play a role during acrylamide/glycidamide-driven tumor development. In contrast to glycidamide, acrylamide exposure led neither to an increased number of SBS nor did it induce characteristic mutation types in the MEF exposure system. Despite the absence of a mutagenic effect of acrylamide in our experiments, acrylamide and glycidamide exposures induce an almost identical set of tumors in both mice and rats, providing a substantial argument for a glycidamide-mediated tumorigenic effect of acrylamide (19). This is further supported by mechanistic studies showing that lung tissue from mice exposed to acrylamide and glycidamide displays comparable DNA adduct patterns as well as similar mutation frequencies in the *cII* transgene (36). Similar observations had been made in the context of *in vitro* mutagenicity of acrylamide in human and mouse cells, suggesting the key role for epoxide metabolite glycidamide to form pre-mutagenic DNA adducts (33).

As shown by our adduct analysis, acrylamide is not efficiently metabolized by MEFs. This finding is in keeping with the results from previous animal carcinogenicity studies. In fact, glycidamide induces hepatocellular carcinomas in neonatal B6C3F1 mice, whereas administration of acrylamide does not increase the tumor incidence. This has been attributed to the inability of neonatal mice to efficiently metabolize acrylamide (31). Moreover, in contrast to acrylamide treatment, glycidamide induces tumors of the small intestine in a dose-dependent manner upon perinatal exposure (57) and similar observations were made for glycidamide mutagenicity *in vitro* (33). We compensated for the lack of proper acrylamide metabolic activation by the addition of human S9 fraction, and the assessment of DNA adducts indeed suggests acrylamide metabolic activation upon addition of S9. However, the adduct levels are substantially lower compared to glycidamide exposure, which may account for the observed differences in mutagenicity. Interestingly, a consistent minor contribution of the glycidamide mutational signature was detected in the majority of ACR clones, whereas it was absent in the Spont clones. This raises the possibility that partial metabolic activation of acrylamide in the MEF system resulted in low levels of glycidamide. However, a clear mutational signature in the employed experimental setting was achieved only by exposing the cells directly to glycidamide.

Single reporter gene studies had previously linked acrylamide and glycidamide exposure to multiple different mutation types. Thanks to the larger number of mutations captured by exome sequencing, we were able to attribute to the glycidamide exposure a particular mutational signature characterized by strand-biased C:G>A:T and T:A>A:T transversions, and T:A>C:A transitions towards the non-transcribed strand suggesting a formation of DNA-adducts. The presence of N7-GA-Gua and N3-GA-Ade, two well-

characterized glycidamide DNA adducts originating from the metabolic conversion of acrylamide (30,44,53), shows a remarkable relationship between DNA adduct profiles and the putative mutational signature of glycidamide. N3-GA-Ade and N7-GA-Gua are depurinating adducts. They can result in apurinic/apyrimidinic sites, which, during replication, induce the mis-incorporation of deoxyadenine, leading to the observed T:A>A:T and C:G>A:T transversions of the glycidamide signature, respectively. The third mutation type specifically enriched in the glycidamide signature, T:A>C:G transitions, has been ascribed to the N1-GA-Ade adduct, a miscoding adduct and the most commonly identified adenine adduct *in vitro* (35,44,53,58). Levels of the guanine adduct were especially high in the exposed MEF cells, whereas the associated C:G>A:T transversions in the resulting post-senescence clones were less represented. This could reflect differences in DNA repair efficiency concerning individual GA-DNA adduct species, or the fact that the resulting clones are derived from single cells whereas the GA-DNA adducts were measured on average in the bulk primary cell population. A mechanism of negative selection of cells with high N7-GA-Gua adduct burden is also plausible.

We observed consistent presence of COSMIC signature 17 in the data generated from the untreated and treated MEF clones. The etiology of signature 17 remains unknown. While some candidate causal factors have been proposed in esophageal adenocarcinoma and gastric cancers (e.g., inflammatory conditions due to acid reflux, *H. pylori*) (56) and in cultured mouse cell systems (54,55), further studies are required to establish why signature 17 tends to arise *in vitro* in immortalized clones derived from mouse embryonic fibroblasts as observed in our study and also previous work (38).

Genome-scale sequencing of tumor tissues will be needed to verify, *in vivo*, the glycidamide mutational signature identified in this study. The established animal models (18,19) of acrylamide- and glycidamide-mediated tumorigenesis provide a suitable starting point, and it would be interesting to compare mutational signatures derived from these models with the *in vitro* results. The identified glycidamide signature with its extended features of transcription strand bias for the major mutation types differs from the currently known COSMIC signatures (Fig. 3D). In addition, we show that in the cancer genome sequencing data sets from the ICGC PCAWG effort, the putative glycidamide-mutational signature can be identified in a subset of tumors of the lung and liver (sites of possible acrylamide exposure due to tobacco smoking), based on combining experimentally derived signatures with sophisticated computational signature reconstruction approaches (Fig. 4).

The continued interest in understanding the contribution of acrylamide and its electrophilic metabolite glycidamide to cancer development reflects recent accumulation of new mechanistic data on the animal carcinogenicity of the compounds. The possible

carcinogenic effects in humans have been recommended for re-evaluation by the Advisory Group to the Monographs Program of the International Agency for Research on Cancer (59). Our findings related to the reconstruction of COSMIC signature 4 using the experimental GA-signature and B[a]P signature, together with the presence of the GA signature in the lung and liver cancer data are relevant given the established high contents of acrylamide in tobacco smoke. Despite the absence of prominent T>N (adenine) mutations in the experimental B[a]P exposure setting, we cannot exclude a possibility that in the human lung cells the adenine residues can be additionally targeted by other tobacco carcinogens such as benzo[*a*]pyrene derivatives or nitrosamines. Importantly, five liver tumor samples identified in this study harbored the GA signature but the major features of signature 4 as represented by the experimental B[a]P signature were absent (Suppl. Fig. S10, Suppl. Table S5). These tumors are thus of particular interest as they could reflect dietary or occupational exposure to acrylamide.

The presented mutational signature of glycidamide and its potential use for screening of cancer genome sequencing data may provide a basis for relevant assessment of cancer risk through new carefully designed molecular cancer epidemiology studies. Future validation analyses involving e.g. GA-DNA adduct monitoring in non-tumor tissue of cancer patients or in animal exposure models are warranted to provide additional evidence that the predominant T>N mutations in the cancers identified in this study indeed originate from exposure to acrylamide and its reactive metabolite glycidamide.

Acknowledgments

The views expressed in this manuscript do not necessarily represent those of the U.S. Food and Drug Administration. The study was supported by funding obtained from INCa-INSERM (Plan Cancer 2015 grant to J.Z.), NIH/NIEHS (1R03ES025023-01A1 grant to M.O.), and the Singapore National Medical Research Council (NMRC/CIRG/1422/2015 grant to S.G.R.) and the Singapore Ministry of Health via the Duke-NUS Signature Research Programmes to S.G.R.. M.R.S. was supported by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank the NYU Genome Technology Center, funded in part by the NIH/NCI Cancer Center Support Grant P30CA016087, and GENEWIZ, South Plainfield, NJ, USA, for expert assistance with Illumina sequencing.

References

1. Smith, M.T.*, et al.* (2016) Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. *Environ Health Perspect*, **124**, 713-21.

2. Alexandrov, L.B.*, et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415-421.
3. Zhivagui, M.*, et al.* (2017) Modelling Mutation Spectra of Human Carcinogens Using Experimental Systems. *Basic Clin Pharmacol Toxicol*, **121 Suppl 3**, 16-22.
4. Hollstein, M.*, et al.* (2017) Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. *Oncogene*, **36**, 158-167.
5. Poon, S.L.*, et al.* (2013) Genome-Wide Mutational Signatures of Aristolochic Acid and Its Application as a Screening Tool. *Science Translational Medicine*, **5**, 197ra101-197ra101.
6. Meier, B.*, et al.* (2014) C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res*, **24**, 1624-36.
7. Scelo, G.*, et al.* (2014) Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun*, **5**, 5135.
8. Jelakovic, B.*, et al.* (2015) Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid. *Int J Cancer*, **136**, 2967-72.
9. Hoang, M.L.*, et al.* (2016) Aristolochic Acid in the Etiology of Renal Cell Carcinoma. *Cancer Epidemiology, Biomarkers & Prevention*, **25**, 1600-1608.
10. Chawanthayatham, S.*, et al.* (2017) Mutational spectra of aflatoxin B1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc Natl Acad Sci U S A*, **114**, E3101-E3109.
11. Huang, M.N.*, et al.* (2017) Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res*, **27**, 1475-1486.
12. Zhang, W.*, et al.* (2017) Genetic Features of Aflatoxin-Associated Hepatocellular Carcinoma. *Gastroenterology*, **153**, 249-262 e2.
13. Ng, A.W.T.*, et al.* (2017) Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. *Sci Transl Med*, **9**.
14. Mojska, H.*, et al.* (2016) Acrylamide content in cigarette mainstream smoke and estimation of exposure to acrylamide from tobacco smoke in Poland. *Annals of agricultural and environmental medicine: AAEM*, **23**, 456-461.
15. Takatsuki, S.*, et al.* (2003) Determination of acrylamide in processed foods by LC/MS using column switching. *Shokuhin Eiseigaku Zasshi. Journal of the Food Hygienic Society of Japan*, **44**, 89-95.
16. IARC Monograph vol. 60 (1994) *Some industrial chemicals. Lyon, 15 - 22 February 1994*, Lyon.
17. Tareke, E.*, et al.* (2002) Analysis of Acrylamide, a Carcinogen Formed in Heated Foodstuffs. *Journal of Agricultural and Food Chemistry*, **50**, 4998-5006.
18. Beland, F.A.*, et al.* (2013) Carcinogenicity of acrylamide in B6C3F(1) mice and F344/N rats from a 2-year drinking water exposure. *Food and Chemical Toxicology,* **51**, 149-159.
19. Beland, F.A.*, et al.* (2015) Carcinogenicity of glycidamide in B6C3F1 mice and F344/N rats from a two-year drinking water exposure. *Food and Chemical Toxicology*, **86**, 104-115.
20. Hogervorst, J.G.*, et al.* (2008) Dietary acrylamide intake and the risk of renal cell, bladder, and prostate cancer. *The American Journal of Clinical Nutrition*, **87**, 1428-1438.
21. Virk-Baker, M.K.*, et al.* (2014) Dietary Acrylamide and Human Cancer: A Systematic Review of Literature. *Nutrition and Cancer*, **66**, 774-790.
22. Olesen, P.T.*, et al.* (2008) Acrylamide exposure and incidence of breast cancer among postmenopausal women in the Danish Diet, Cancer and Health Study. *International Journal of Cancer*, **122**, 2094-2100.

23. Wilson, K.M.*, et al.* (2009) Acrylamide exposure measured by food frequency questionnaire and hemoglobin adduct levels and prostate cancer risk in the Cancer of the Prostate in Sweden Study. *International Journal of Cancer*, **124**, 2384-2390.
24. Xie, J.*, et al.* (2013) Acrylamide Hemoglobin Adduct Levels and Ovarian Cancer Risk: A Nested Case-Control Study. *Cancer Epidemiology Biomarkers & Prevention*, **22**, 653-660.
25. Obón-Santacana, M.*, et al.* (2016) Acrylamide and glycidamide hemoglobin adduct levels and endometrial cancer risk: A nested case-control study in nonsmoking postmenopausal women from the EPIC cohort. *International Journal of Cancer*, **138**, 1129-1138.
26. Obón-Santacana, M.*, et al.* (2016) Acrylamide and Glycidamide Hemoglobin Adducts and Epithelial Ovarian Cancer: A Nested Case-Control Study in Nonsmoking Postmenopausal Women from the EPIC Cohort. *Cancer Epidemiology, Biomarkers & Prevention*, **25**, 127-134.
27. Obón-Santacana, M.*, et al.* (2016) Dietary and lifestyle determinants of acrylamide and glycidamide hemoglobin adducts in non-smoking postmenopausal women from the EPIC cohort. *European Journal of Nutrition*.
28. Sumner, S.C.*, et al.* (1999) Role of cytochrome P450 2E1 in the metabolism of acrylamide and acrylonitrile in mice. *Chemical Research in Toxicology*, **12**, 1110-1116.
29. Ghanayem, B.I.*, et al.* (2005) Role of CYP2E1 in the epoxidation of acrylamide to glycidamide and formation of DNA and hemoglobin adducts. *Toxicological Sciences*, **88**, 311-318.
30. Segerbäck, D.*, et al.* (1995) Formation of N-7-(2-carbamoyl-2-hydroxyethyl) guanine in DNA of the mouse and the rat following intraperitoneal administration of [14C] acrylamide. *Carcinogenesis*, **16**, 1161-1165.
31. Von Tungeln, L.S.*, et al.* (2012) Tumorigenicity of acrylamide and its metabolite glycidamide in the neonatal mouse bioassay. *International Journal of Cancer*, **131**, 2008-2015.
32. Besaratinia, A.*, et al.* (2003) Weak yet distinct mutagenicity of acrylamide in mammalian cells. *Journal of the National Cancer Institute*, **95**, 889-896.
33. Besaratinia, A.*, et al.* (2004) Genotoxicity of acrylamide and glycidamide. *Journal of the National Cancer Institute*, **96**, 1023-1029.
34. Von Tungeln, L.S.*, et al.* (2009) DNA adduct formation and induction of micronuclei and mutations in B6C3F1/Tk mice treated neonatally with acrylamide or glycidamide. *International Journal of Cancer*, **124**, 2006-2015.
35. Ishii, Y.*, et al.* (2015) Acrylamide induces specific DNA adduct formation and gene mutations in a carcinogenic target site, the mouse lung. *Mutagenesis*, **30**, 227-235.
36. Manjanatha, M.G.*, et al.* (2015) Acrylamide-induced carcinogenicity in mouse lung involves mutagenicity: c*II* gene mutations in the lung of big blue mice exposed to acrylamide and glycidamide for up to 4 weeks. *Environ Mol Mutagen*, **56**, 446-56.
37. Olivier, M.*, et al.* (2014) Modelling mutational landscapes of human cancers in vitro. *Scientific Reports*, **4**.
38. Nik-Zainal, S.*, et al.* (2015) The genome as a record of environmental exposure. *Mutagenesis*, **30**, 763-70.
39. Huskova, H.*, et al.* (2017) Modeling cancer driver events in vitro using barrier bypass-clonal expansion assays and massively parallel sequencing. *Oncogene*, **36**, 6041-6048.
40. Liu, Z.*, et al.* (2004) Human tumor p53 mutations are selected for in mouse embryonic fibroblasts harboring a humanized p53 gene. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 2963-2968.
41. Todaro, G.J.*, et al.* (1963) Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. *The Journal of Cell Biology*, **17**, 299-313.

42. Severson, P.L.*, et al.* (2014) Exome-wide mutation profile in benzo[a]pyrene-derived post-stasis and immortal human mammary epithelial cells. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, **775-776**, 48-54.
43. Stampfer, M.R.*, et al.* (1985) Induction of transformation and continuous cell lines from normal human mammary epithelial cells after exposure to benzo[a]pyrene. *Proc Natl Acad Sci U S A*, **82**, 2394-8.
44. Gamboa da Costa, G.*, et al.* (2003) DNA adduct formation from acrylamide via conversion to glycidamide in adult and neonatal mice. *Chemical Research in Toxicology*, **16**, 1328-1337.
45. Ardin, M.*, et al.* (2016) MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics*, **17**, 170.
46. Brunet, J.-P.*, et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4164-4169.
47. Alexandrov, Ludmil B.*, et al.* (2013) Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, **3**, 246-259.
48. McCreery, M.Q.*, et al.* (2015) Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. *Nature Medicine*, **21**, 1514-1520.
49. Nassar, D.*, et al.* (2015) Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nature Medicine*, **21**, 946-954.
50. Westcott, P.M.K.*, et al.* (2014) The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. *Nature*, **517**, 489-492.
51. Stampfer, M.R.*, et al.* (1988) Human mammary epithelial cells in culture: differentiation and transformation. *Cancer Treat Res*, **40**, 1-24.
52. Chen, L.*, et al.* (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, **355**, 752-756.
53. Besaratinia, A.*, et al.* (2005) DNA adduction and mutagenic properties of acrylamide. *Mutation Research*, **580**, 31-40.
54. Behjati, S.*, et al.* (2014) Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, **513**, 422-425.
55. Milholland, B.*, et al.* (2017) Differences between germline and somatic mutation rates in humans and mice. *Nat Commun*, **8**, 15183.
56. Secrier, M.*, et al.* (2016) Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nature Genetics*, **48**, 1131-1141.
57. Olstørn, H.B.A.*, et al.* (2007) Effects of perinatal exposure to acrylamide and glycidamide on intestinal tumorigenesis in Min/+ mice and their wild-type litter mates. *Anticancer Research*, **27**, 3855-3864.
58. Randall, S.K.*, et al.* (1987) Nucleotide insertion kinetics opposite abasic lesions in DNA. *Journal of Biological Chemistry*, **262**, 6864-6870.
59. Straif, K.*, et al.* (2014) Future priorities for the IARC Monographs. *The Lancet Oncology*, **15**, 683-684.

Figure legends

Figure 1: Acrylamide- and glycidamide-induced cytotoxicity and genotoxicity *in vitro*. (**A**) Cell viability, following 24-hour treatment of primary MEFs with the indicated concentrations of acrylamide (top panel), in the absence (diamonds) and presence (circles) of human S9 fraction, and glycidamide (bottom panel), as determined by MTT assay. Absorbance was measured 48 hours after treatment cessation and was normalized to untreated cells. The

results are expressed as mean percent ±SD of three replicates. (**B**) DNA damage assessment by immunofluorescence with an antibody specific for Ser139-phosphorylated histone H2Ax (ɣH2Ax). Primary MEFs were treated with acrylamide or glycidamide for 24 hours prior to immunofluorescence. Compound concentrations used were based on 20-70% viability reduction in the MTT assay: 10 mM acrylamide, 5 mM acrylamide in the presence of S9 fraction and 3 mM glycidamide. ACR: acrylamide; GA: glycidamide.

Figure 2: Analysis of the mutation patterns derived from exome sequencing data from immortalized Hupki MEF clones. (**A**) Principle component analysis (PCA) of WES data. PCA was computed using as input the mutation count matrix of the clones that immortalized spontaneously (Spont) or were derived from exposure to acrylamide (ACR) or glycidamide (GA). Each sample is plotted considering the value of the first and second principal components (Dim1 and Dim2). The percentage of variance explained by each component is indicated within brackets on each axis. Spont, ACR- and GA-exposed samples are represented by differently colored symbols. (**B**) Representation of small insertions and deletions (indels) counts within the immortalized clones as determined by the Strelka variant caller. (**C**) Mutational signatures identified by non-negative matrix factorization (NMF) in the 15 Hupki MEF-derived clones (sig A, sig B, and sig C). X-axis represents the trinucleotide sequence context. Y-axis represents the frequency distribution of the mutations. The predominant trinucleotide context for T:A > A:T mutations is indicated in sig B (5'-C<u>T</u>G-3'). The trinucleotide contexts for C:G > G:C (5'-G<u>C</u>C-3') and T:A > G:C mutations (5'-N<u>T</u>T-3') are highlighted in sig C. (**D**) Contribution of the identified signatures to each sample (X-axis), assigned either by absolute SBS counts or by proportion (bar graphs). The reconstruction accuracy of the identified mutational signatures in individual samples is shown in the bottom scatter plot (Y-axis value of 1 = 100% accuracy).

Figure 3: (**A**) Refinement of GA signature. The contribution of signature 17 (T:A>G:C in 5'-N<u>T</u>T-3' context), present in all clones, was decreased by performing NMF on Hupki samples pooled with primary tumor samples with high levels of signature 17 (see Methods). (**B**) Transcription strand bias analysis for the six mutation types in GA-exposed clones. For each mutation type, the number of mutations occurring on the transcribed (T) and non-transcribed (N) strand is shown on the Y-axis. *** $p < 10^{-8}$; * $p < 10^{-2}$. (**C**) DNA adducts analysis as determined by LC-MS/MS. Levels of N7-GA-Gua adduct in ACR+S9 and GA treated MEFs and N3-GA-Ade DNA adduct level in GA treated MEFs. The data are presented as the number of adducts in 10^8 nucleotides. $n \geq 2$. (**D**) Cosine similarity matrix comparing the putative glycidamide mutational signature with other A>T rich mutational signatures from

COSMIC (signatures 22, 25, and 27) and from experimental exposure assays using specific carcinogens (7,12-dimethylbenz[a]anthracene (DMBA), urethane, and aristolochic acid (AA)).

Figure 4: GA signature in human primary cancer genome PCAWG data. (**A**) Comparison of COSMIC signature 4 with two experimentally derived signatures (B[a]P_Exp = signature in clones from benzo[*a*]pyrene treated HMEC cells; GA_Exp = signature in clones from glycidamide-treated MEF cells). Cosine similarity between the T>N (adenine) components of signature 4 and GA signature is shown to the right. (**B**) Transcription strand bias analysis for the six mutation types underlying the signatures in panel A). For each mutation type, the number of mutations occurring on the transcribed (T) and non-transcribed (N) strand is shown on the left Y-axis. The significance is expressed as –log10(p-value) indicated on the right Y-axis. *** $p < 10^{-8}$; ** $p < 10^{-4}$; * $p < 10^{-2}$. (**C**) Scatter plots show reconstruction of COSMIC signature 4 using B[a]P- and glycidamide- experimental mutational signatures in lung adenocarcinoma, lung squamous cell carcinoma and hepatocellular carcinoma from the PCAWG data set. (**D**) mSigAct analysis identifies the assignment and the contributions of mutational signatures (including the experimental signature_GA_Exp (red) and signature_B[a]P_Exp (blue)) to the mutation burden of a total of 101 PCAWG lung and liver tumors identified as positive for the GA signature signal.

Table 1: Summary of cell lines, treatment conditions and *TP53* mutation status.

Sample ID	Embryo	Exposure	Conc. (mM)	Exposure duration (hrs)	cDNA change	gDNA change*	AA change	Codon 72 (rs1042522)**
Prim_1	E210	-	-	-				Pro/Pro
Prim_2	E213	-	-	-				Arg/Pro
Prim_3	E214	-	-	-				Pro/Pro
Spont_1	E213	-	-	-				Arg/Pro
Spont_2	E214	-	-	-				Pro/Pro
Spont_3	E214	-	-	-				Pro/Pro
ACR_S9_1	E213	ACR	5	24				Arg/Pro
ACR_S9_2	E213	ACR	5	24				Arg/Pro
ACR_1	E213	ACR	10	24	c.881delA	g.7577057delA	p.E294fs	Arg/-
ACR_2	E213	ACR	10	24	c.818G>T	g.7577120C>A	p.R273L	Pro/-
ACR_3	E214	ACR	10	24	c.740A>T; c.839G>C	g.7577541T>A; g.7577099C>G	p.N247I; p.R280T	Pro/Pro
ACR_4	E214	ACR	10	24				Pro/Pro
ACR_5	E214	ACR	10	24				Pro/Pro
GA_1	E210	GA	3	24				Pro/Pro
GA_2	E210	GA	3	24				Pro/Pro
GA_3	E210	GA	3	24	c.309-310CC>TA	g.7579377-7579378GG>TA	[p.Y103Y; p.Q104K]	Pro/Pro
GA_4	E214	GA	3	24				Pro/Pro
GA_5	E214	GA	3	24				Pro/Pro

Prim = Primary cells; Spont = spontaneously immortalized clones; ACR = acrylamide-exposure derived clones; GA = glycidamide-exposure derived clones. Each exposure condition was carried out in two biological replicates (embryos). S9 = human S9 fraction; Pro = proline; Arg = arginine; Arg/- or Pro/- = loss of allele; fs = frameshift; c = codon; g = genomic position; p = protein. * = hg19 coordinates; ** = human polymorphic site (rs1042522).

Figure 1

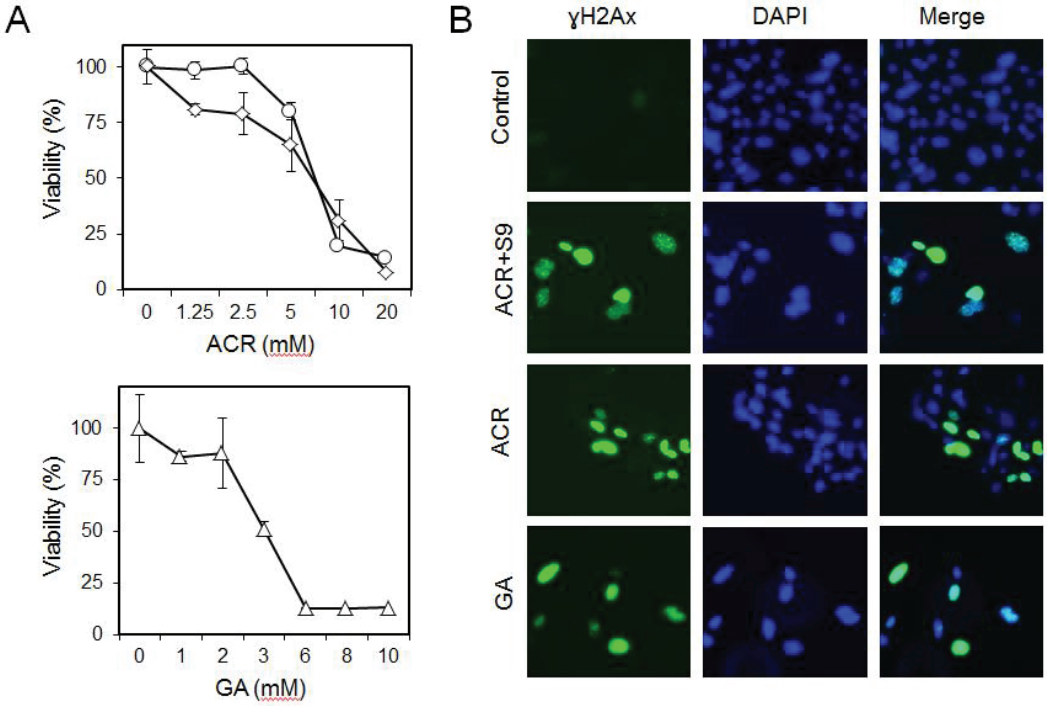


Figure 2

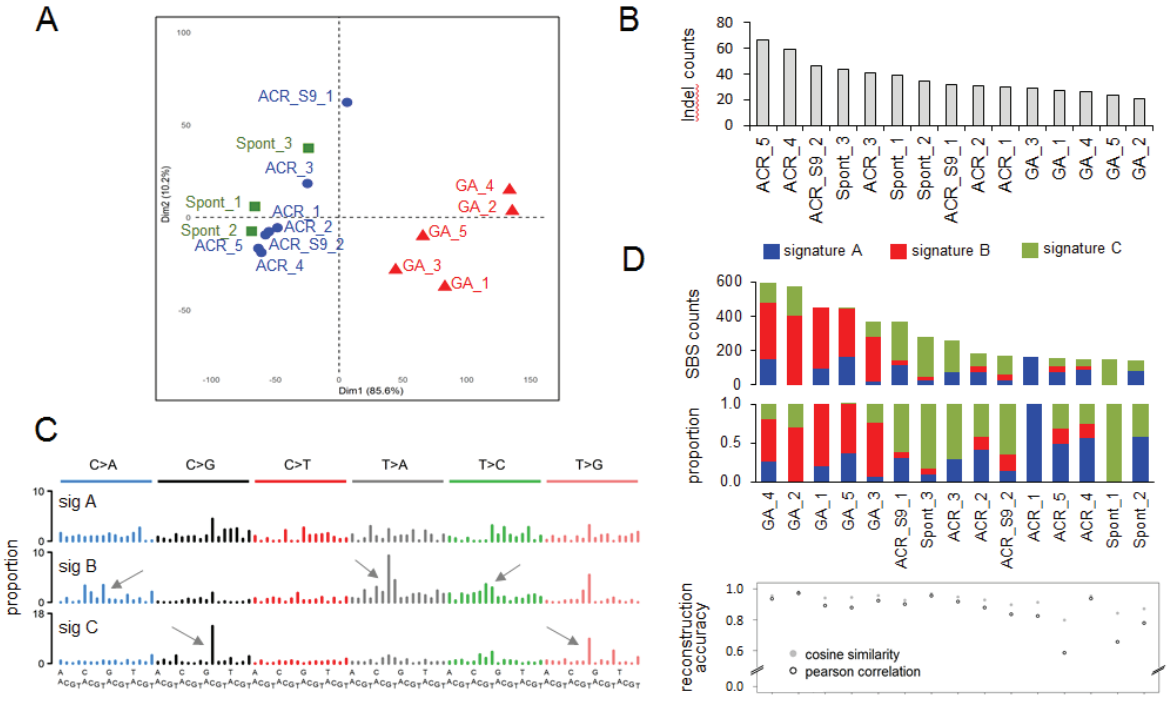


Figure 3

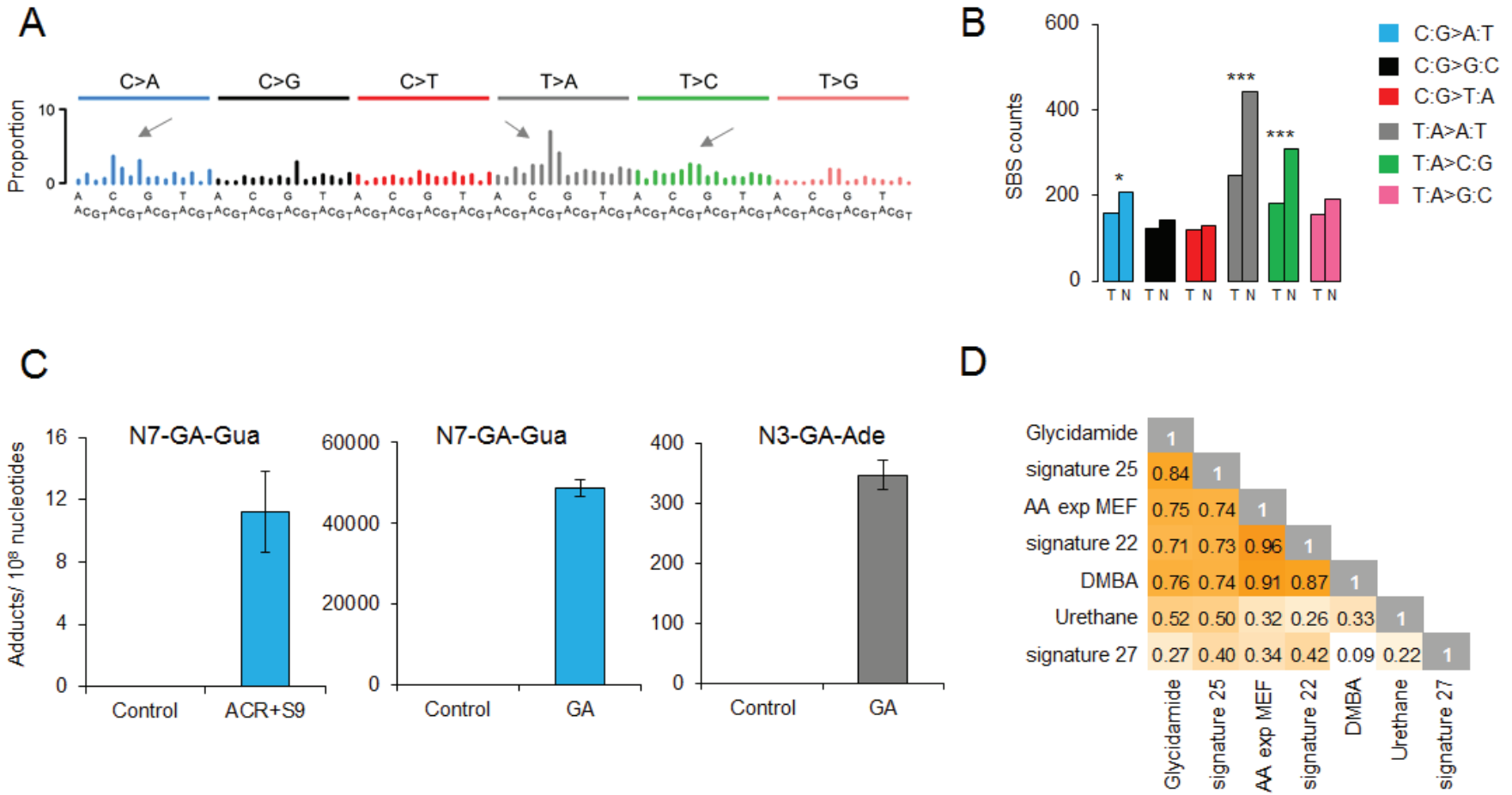
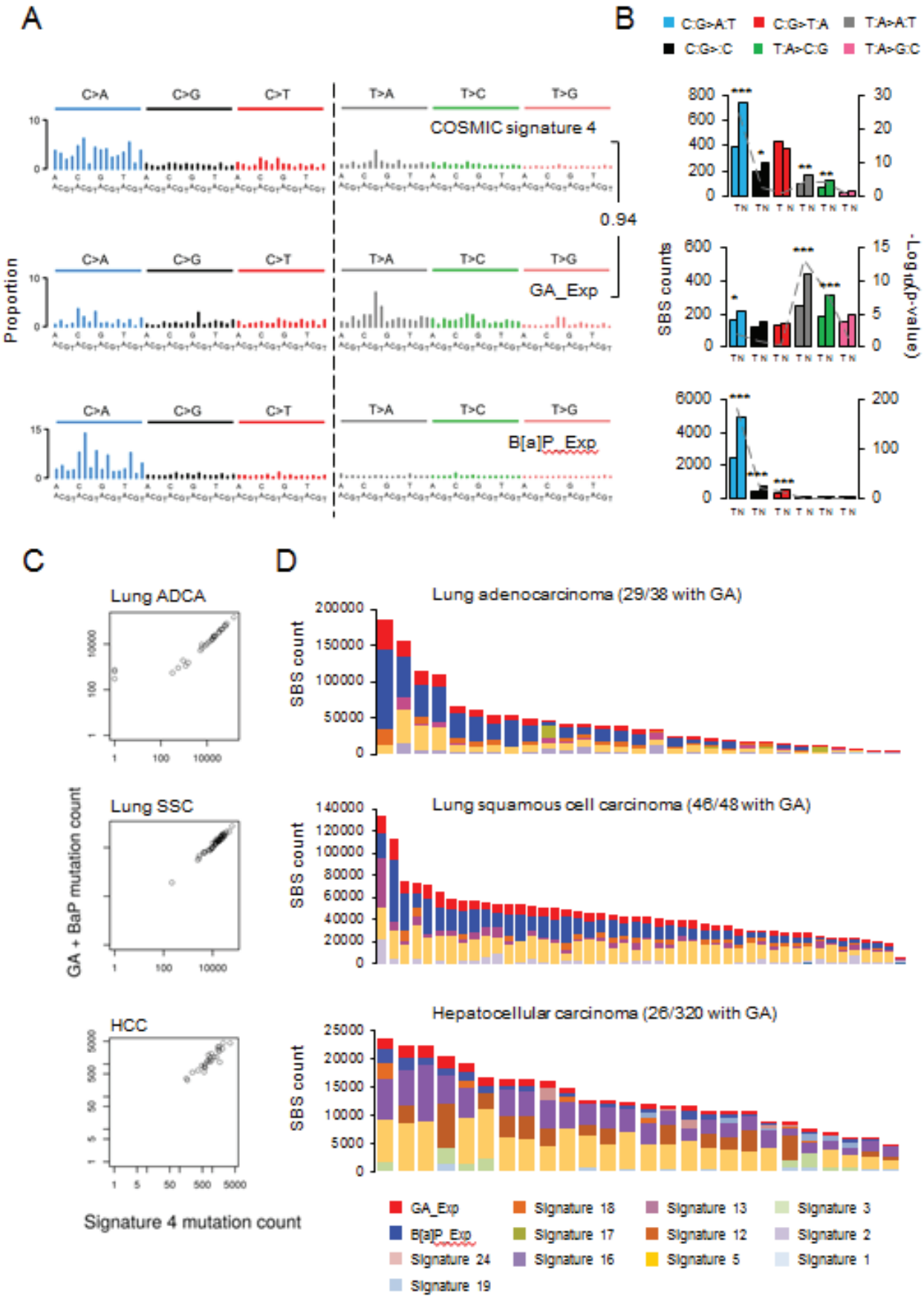

A
C>A
C>G
C>T
T>A
T>C
T>G
Proportion
10
0
B
SBS counts
600
400
200
0
C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C
*
***
***
T N
C
N7-GA-Gua
N7-GA-Gua
N3-GA-Ade
Adducts/ 10⁸ nucleotides
16
12
8
4
0
60000
40000
20000
400
300
200
100
Control
ACR+S9
Control
GA
Control
GA
D
Glycidamide
signature 25
AA exp MEF
signature 22
DMBA
Urethane
signature 27
1
0.84 1
0.75 0.74 1
0.71 0.73 0.96 1
0.76 0.74 0.91 0.87 1
0.52 0.50 0.32 0.26 0.33 1
0.27 0.40 0.34 0.42 0.09 0.22 1

Figure 4


A
B
C
D
C>A
C>G
C>T
T>A
T>C
T>G
COSMIC signature 4
GA_Exp
B[a]P_Exp
0.94
Proportion
C:G>A:T
C:G>T:A
T:A>A:T
C:G>:C
T:A>C:G
T:A>G:C
SBS counts
-Log10(p-Value)
Lung ADCA
Lung SSC
HCC
GA + BaP mutation count
Signature 4 mutation count
Lung adenocarcinoma (29/38 with GA)
Lung squamous cell carcinoma (46/48 with GA)
Hepatocellular carcinoma (26/320 with GA)
SBS count
GA_Exp
B[a]P_Exp
Signature 24
Signature 19
Signature 18
Signature 17
Signature 16
Signature 13
Signature 12
Signature 5
Signature 3
Signature 2
Signature 1

Supplementary Materials and Methods

DNA adduct analysis

The DNA was isolated from the cells using standard digestion with proteinase K, followed by phenol-chloroform extraction and ethanol precipitation. The DNA was subsequently treated with RNase A and T1, extracted with phenol-chloroform, and reprecipitated with ethanol. N7 GA-Gua and N3 GA-Ade were released by neutral thermal hydrolysis for 15 minutes, using Eppendorf Thermomixer R (Eppendorf North America) set to 99 °C. The samples were filtered through Amicon 3K molecular weight cutoff filters (Merck Millipore) to separate the adducts from the intact DNA.

TP53 genotyping

The following are the *TP53* primers used for amplicon sequencing of mutations accumulated in human TP53 of the Hupki MEFs. The sequences are presented in 5' to 3' orientation: Exon 4: fwd – TGCTCTTTTCACCCATCTAC, rev – ATACGGCCAGGCATTGAAGT; Exons 5-6: fwd – TGTTCACTTGTGCCCTGACT, rev – TTAACCCCTCCTCCCAGAGA; Exon 7: fwd – CTTGCCACAGGTCTCCCC, rev – CACTTGCCACCCTGCACA; Exon 8: fwd – TCCTTACTGCCTCTTGCTTCTCTT; rev – CCAAGGGTGCAGTTATGCCT. Sequences and their alterations were analyzed using the CodonCode Aligner software.

Processing of WES data

Prior to variant calling, recalibrated .bam files were interrogated for imbalanced base mismatch distribution between Read 1 and Read 2 sequences. We used the DNA damage estimator tool (as per (1); (https://github.com/Ettwiller/Damage-estimator)) to measure the Global Imbalance Value (GIV) score and to exclude sequencing-related DNA damage and artefacts due to oxidative damage that can confound the determination of treatment-specific variants. The MutSpec suite included tools for annotation of the vcf files with Annovar and variant filtering to remove dbSNP142 contents, segmental duplicates, repeats, and tandem repeat regions. Finally, to maximize the chance of robust variant calls and to exclude potential unfiltered single nucleotide polymorphisms (SNP), we considered only variants unique to each sample.

Bioinformatics and statistical analyses

The following are the International Cancer Genome Consortium (ICGC) esophageal carcinoma patient data (2,3) that were used in the step of cleaning the experimental signature from the COSMIC signature 17 signal: ESAD-UK-SP119768.hg19; ESAD-UK-SP191660.hg19; ESAD-UK-SP111113.hg19; ESAD-UK-SP111173.hg19; ESAD-UK-

SP192267.hg19; ESAD-UK-SP111026.hg19; ESAD-UK-SP192494.hg19; ESAD-UK-SP111019.hg19; ESAD-UK-SP111058.hg19.

References

1. Chen, L.*, et al.* (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, **355**, 752-756.
2. Secrier, M.*, et al.* (2016) Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nature Genetics*, **48**, 1131-1141.
3. Cancer Genome Atlas Research Network (2017) Integrated genomic characterization of oesophageal carcinoma. *Nature*, **541**, 169-175.

Supplementary Figure Legends

Supplementary Fig. S1: Comparison of different normalization and single-nucleotide variant calling strategies. Variant calling with respect to primary cell normalization. Venn diagrams show the overlap of variants called in glycidamide (GA)-derived clones after normalization to three different batches of primary cells (Prim_1, Prim_2, and Prim_3).

Supplementary Fig. S2: Growth curves of Hupki MEFs. Primary cells were either left untreated (Spont) or were exposed to acrylamide (ACR±S9) or glycidamide (GA). X-axis represents days in culture. Y-axis represents the cumulative doubling populations. The dashed vertical line represents the threshold of p-value < 0.05. Arrow: compound exposure; S*: senescence; SBI: senescence bypass/immortalization.

Supplementary Fig. S3: Mutation spectra derived from exome sequencing data from immortalized Hupki MEF clones derived from exposure to (**A**) acrylamide (ACR) or (**B**) glycidamide (GA), or (**C**) by spontaneous immortalization (Spont). X-axis represents the trinucleotide sequence context. Y-axis represents the frequency distribution of the mutations in each context.

Supplementary Fig. S4: Illustration of the transcription strand bias derived from the analysis of exome sequencing data from immortalized Hupki MEF cell lines. GA: glycidamide-derived clones; ACR: acrylamide-derived clones; Spont: spontaneously immortalized clones. The six mutation types are represented by different colors. For each mutation type, the number of mutations occurring on the transcribed (T) and non-transcribed (N) strand, as well as the p-

values for strand bias is shown on the y-axes. The dashed grey line in each graph indicates the p-values for strand bias for each mutation type. The horizontal, dashed black line represents a significance threshold of $p < 0.05$.

Supplementary Fig. S5: Distribution of mutations based on their allelic frequencies in the five glycidamide (GA)-derived clones (left). 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. Percentages of the six mutation types, color-coded, among all mutations identified in GA clones (right). The overall mutation number for each sample is indicated in the centre of the pie chart.

Supplementary Fig. S6: Mutation type and mutation spectra analysis with respect to variant allele frequency (VAF). The analysis was carried out using exome sequencing data from immortalized Hupki MEF clones derived from exposure to glycidamide. Top left: Mutation counts were stratified into three VAF bins ([0-33% = low VAF]; [34-66% = medium VAF]; [67-100% = high VAF]). Top right: The relative contribution of the six mutation types to the overall number of mutations in each VAF bin is shown on the y-axis. Bottom panel: Mutation spectra (left) and strand bias (right) analysis for the different VAF bins. Mutation spectra analysis: X-axis represents the trinucleotide sequence context. Y-axis represents the frequency distribution of the mutations. The counts for each mutation type are indicated in parentheses. Strand bias analysis: For each mutation type, the number of mutations occurring on the transcribed and non-transcribed strand is shown on the y-axis. T: transcribed strand; N: non-transcribed strand.

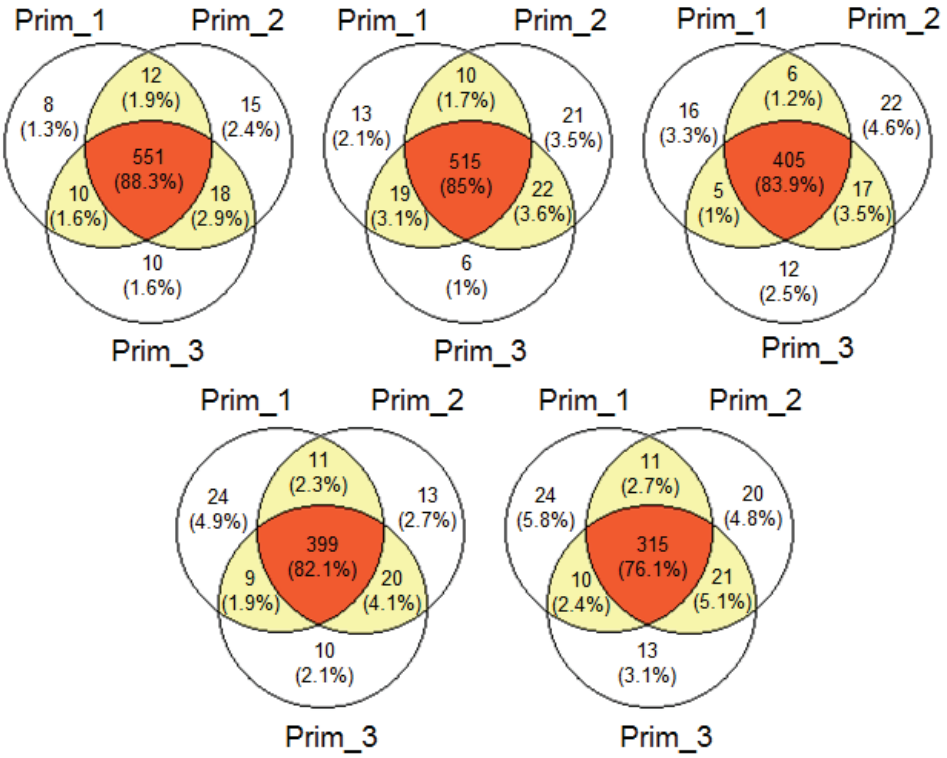
Supplementary Fig. S7: The 'baiting' clean-up of background signature 17 and the quantification of its efficiency. COSMIC signature 17 (top track) marked by the arrows observed in GA mutation spectra as well as in GA-mutational signature before and after baiting (clean). The heat-map table on the right indicates the final proportionate reduction of signature 17-specific peaks after re-running the NMF with signature 17-rich ICGC ESAD data sets listed in the Supplementary Materials and Methods section.

Supplementary Fig. S8: (**A**) The structures of N7-GA-Gua and N3-GA-Ade adducts analyzed by LC-MS/MS. (**B**) Representative multiple-reaction monitoring chromatograms (relative signal intensity vs time) for N7-GA-Gua and N3-GA-Ade adducts in DNA from ACR treatment in the presence of S9 fraction (ACR+S9) and GA-treated (GA) primary Hupki MEF. Internal standards (IS) were added in amounts of 1000 fmol for N7-GA-Gua and 200 fmol for N3-GA-Ade.

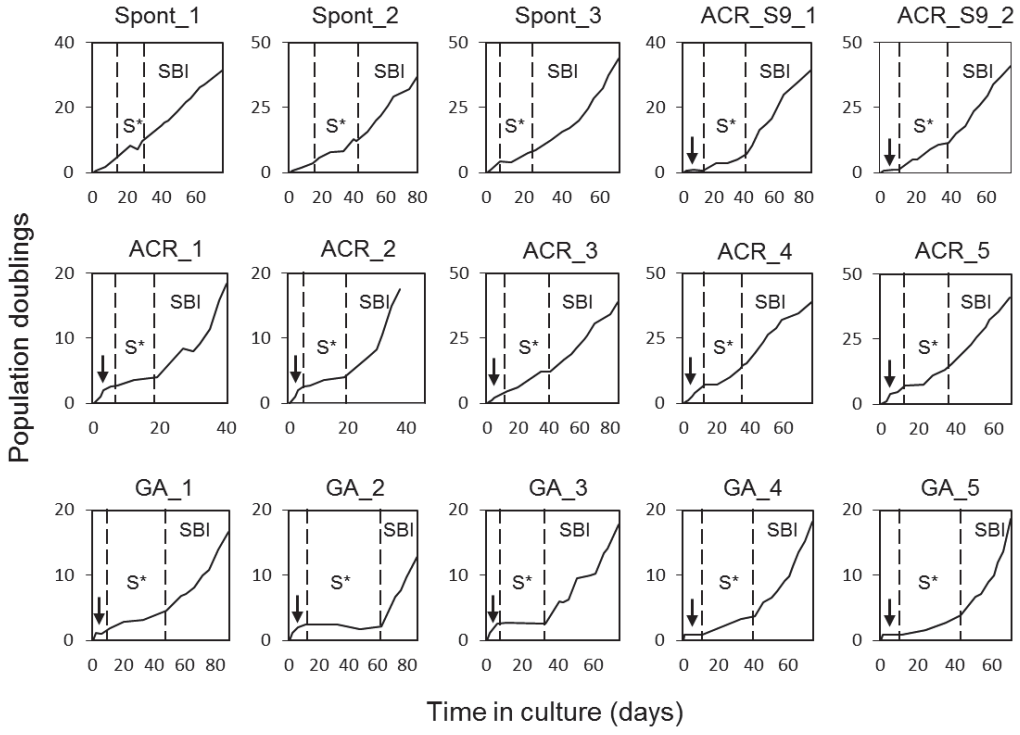
Supplementary Fig. S9: T:A>A:T enriched mutational signatures used for cosine similarity analysis (see Fig. 3D). The individual signatures were originally derived from human cancer sequencing data or experimental models (animal bioassays, cell lines) of carcinogen exposure. X-axis represents the trinucleotide sequence context. Y-axis represents the frequency distribution of the mutations. The predominant trinucleotide context for T:A>A:T mutations is indicated by an arrow in the signature landscape. AA: aristolochic acid; DMBA: 7,12-dimethylbenz[a]anthracene.

Supplementary Fig. S10: (**A**) Scatter plots show the measure of correlation of the GA-signature versus B[a]P-signature (used to reconstruct COSMIC signature 4) in PCAWG lung adenocarcinomas (ADCA), lung squamous cell carcinomas (SCC) and hepatocellular carcinomas (HCC). (**B**) Bar-plots representing the proportion of the assignment of the experimental GA_Exp and B[a]P_Exp signatures in lung adenocarcinomas, lung squamous cell carcinomas and hepatocellular carcinomas from the PCAWG data set. The asterisk denotes liver HCC samples harboring GA-signature only (no B[a]P-signature detected), indicating possible dietary or occupational exposure. Full list of these samples is accessible from Suppl. Table S5.

Suppl. Fig. S1

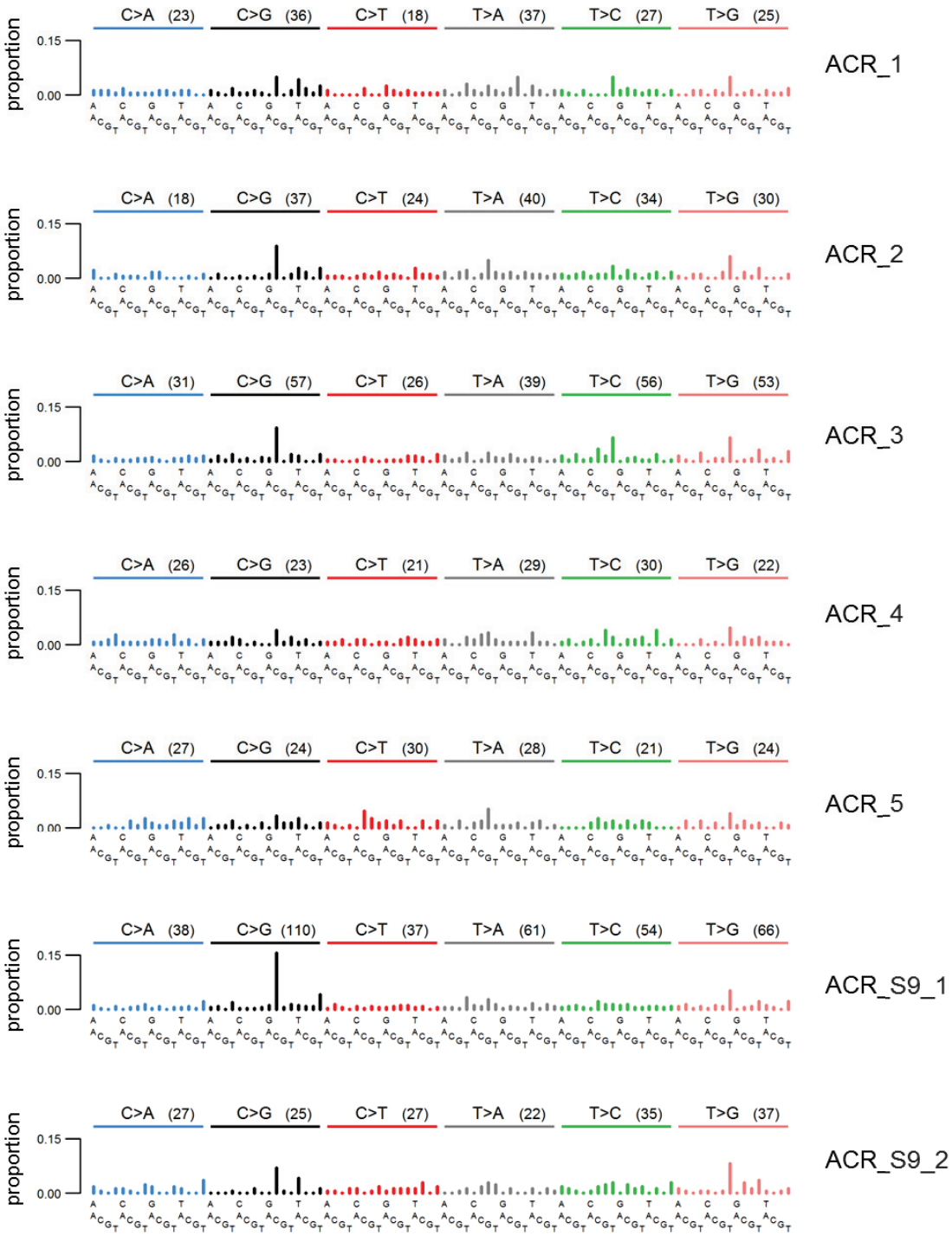


Suppl. Fig. S2

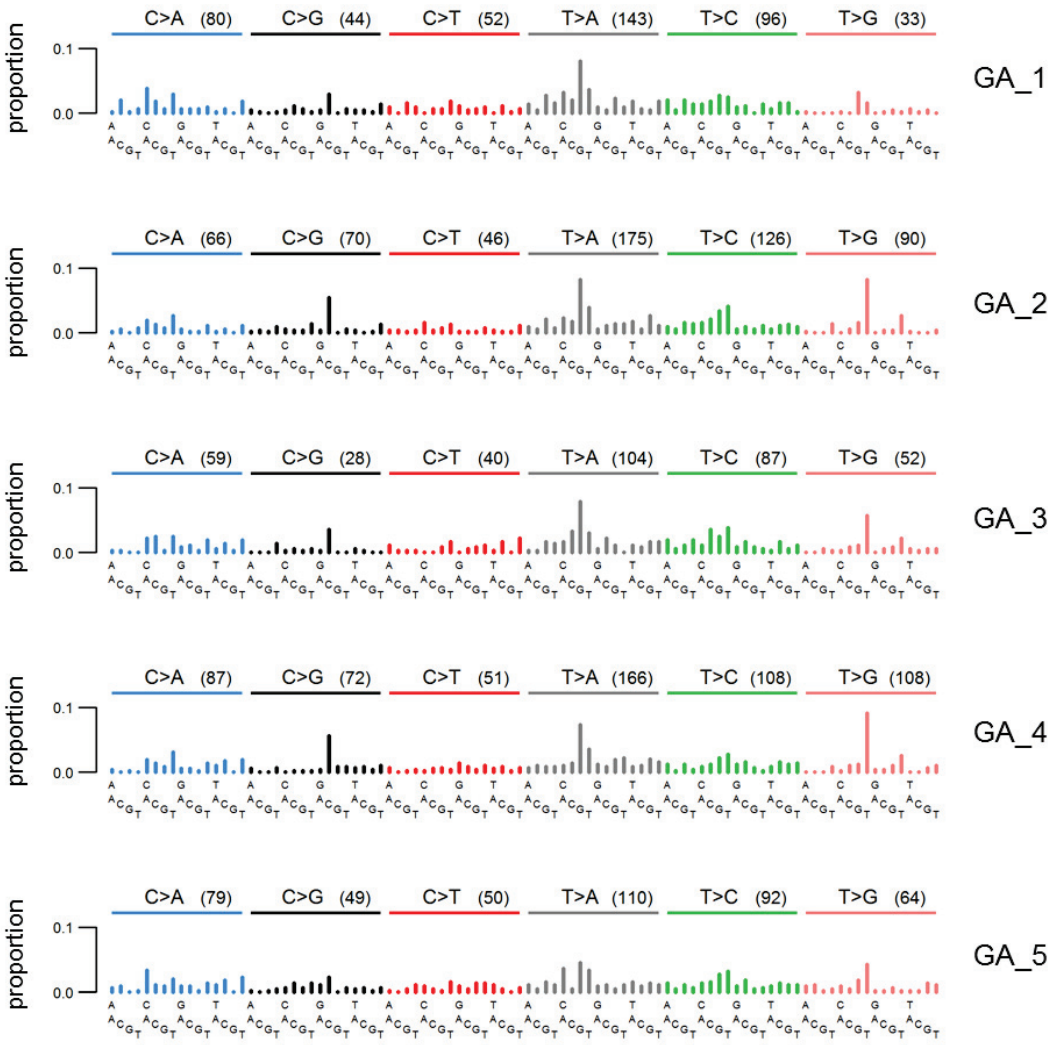


Suppl. Fig. S3

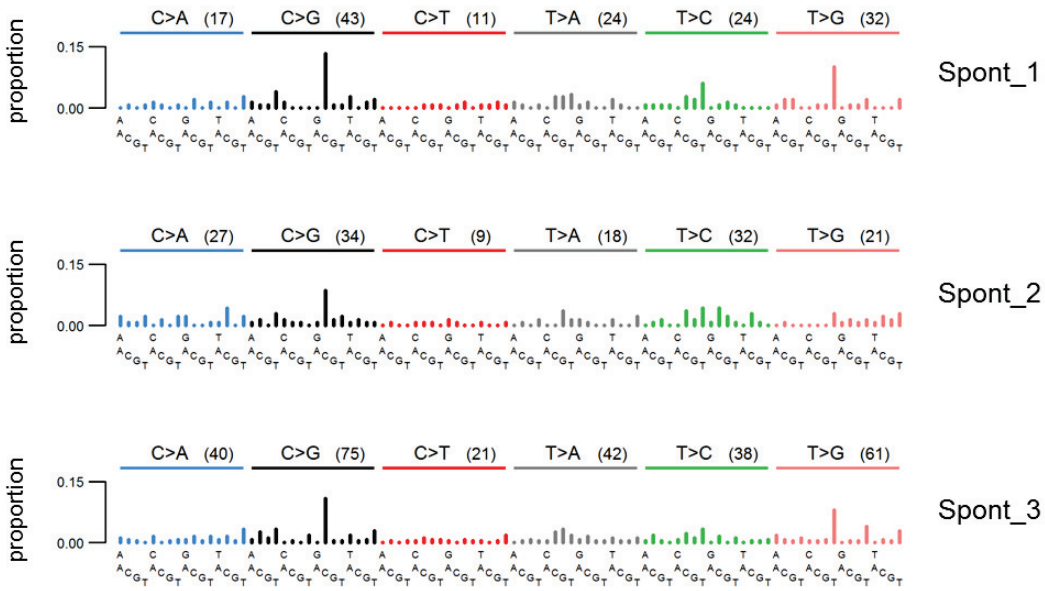
A



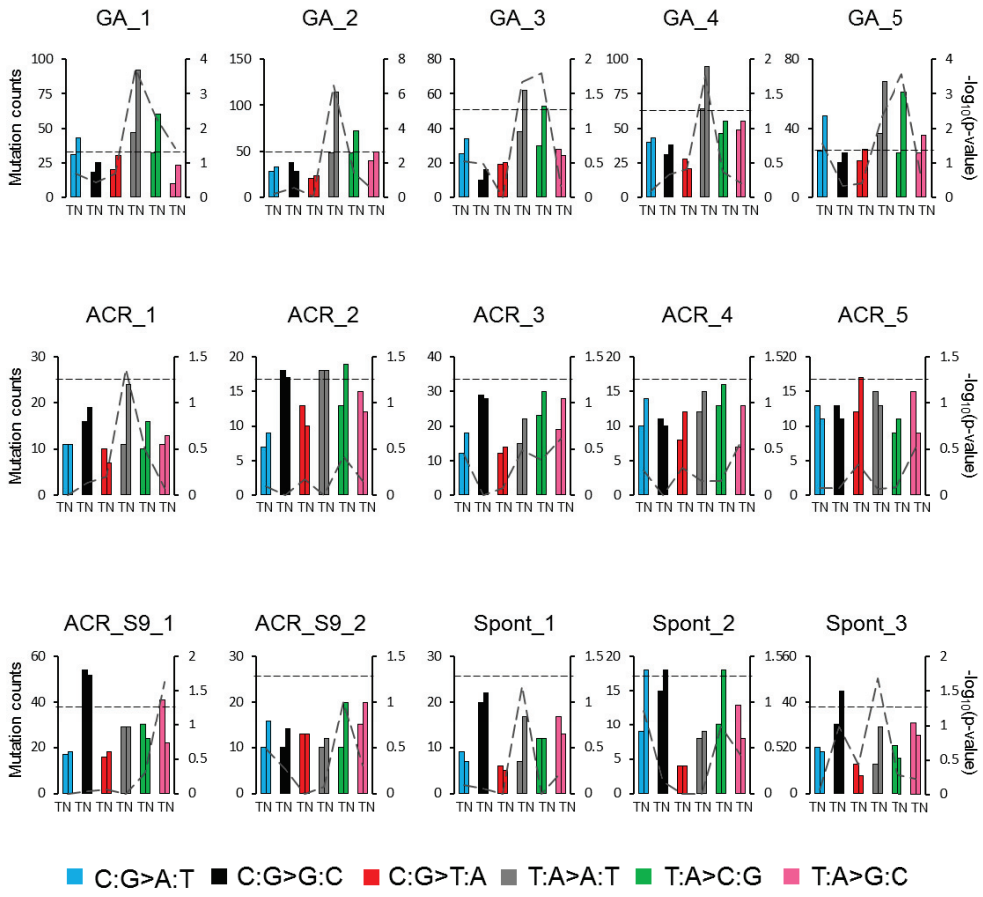
B


C>A (80)
C>G (44)
C>T (52)
T>A (143)
T>C (96)
T>G (33)
GA_1
C>A (66)
C>G (70)
C>T (46)
T>A (175)
T>C (126)
T>G (90)
GA_2
C>A (59)
C>G (28)
C>T (40)
T>A (104)
T>C (87)
T>G (52)
GA_3
C>A (87)
C>G (72)
C>T (51)
T>A (166)
T>C (108)
T>G (108)
GA_4
C>A (79)
C>G (49)
C>T (50)
T>A (110)
T>C (92)
T>G (64)
GA_5
proportion
0.1
0.0

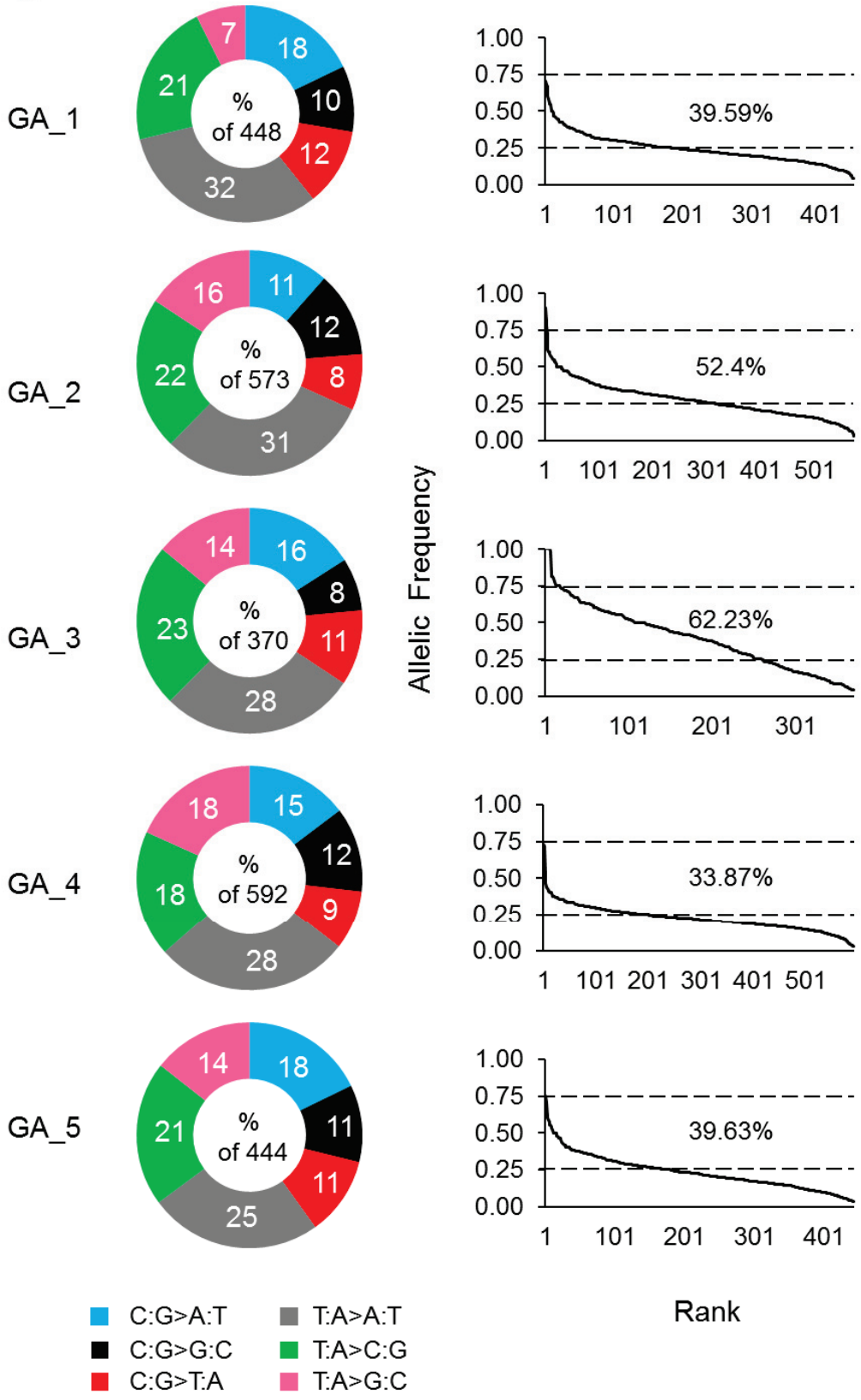
C


C>A (17)
C>G (43)
C>T (11)
T>A (24)
T>C (24)
T>G (32)
proportion
0.15
0.00
Spont_1
C>A (27)
C>G (34)
C>T (9)
T>A (18)
T>C (32)
T>G (21)
proportion
0.15
0.00
Spont_2
C>A (40)
C>G (75)
C>T (21)
T>A (42)
T>C (38)
T>G (61)
proportion
0.15
0.00
Spont_3

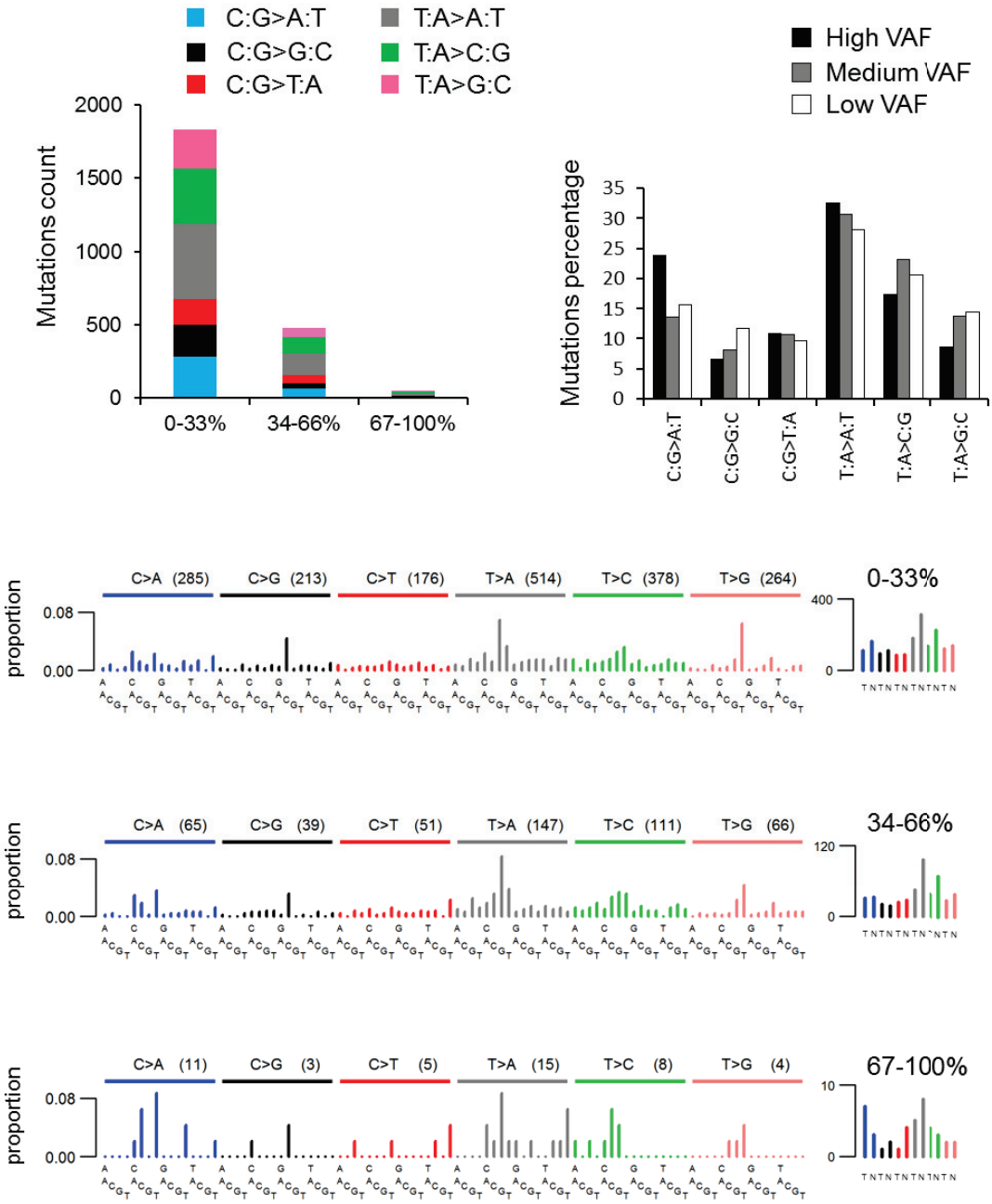
Suppl. Fig. S4


GA_1
GA_2
GA_3
GA_4
GA_5
ACR_1
ACR_2
ACR_3
ACR_4
ACR_5
ACR_S9_1
ACR_S9_2
Spont_1
Spont_2
Spont_3
Mutation counts
-log10(p-value)
TN TN TN TN TN TN
C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C

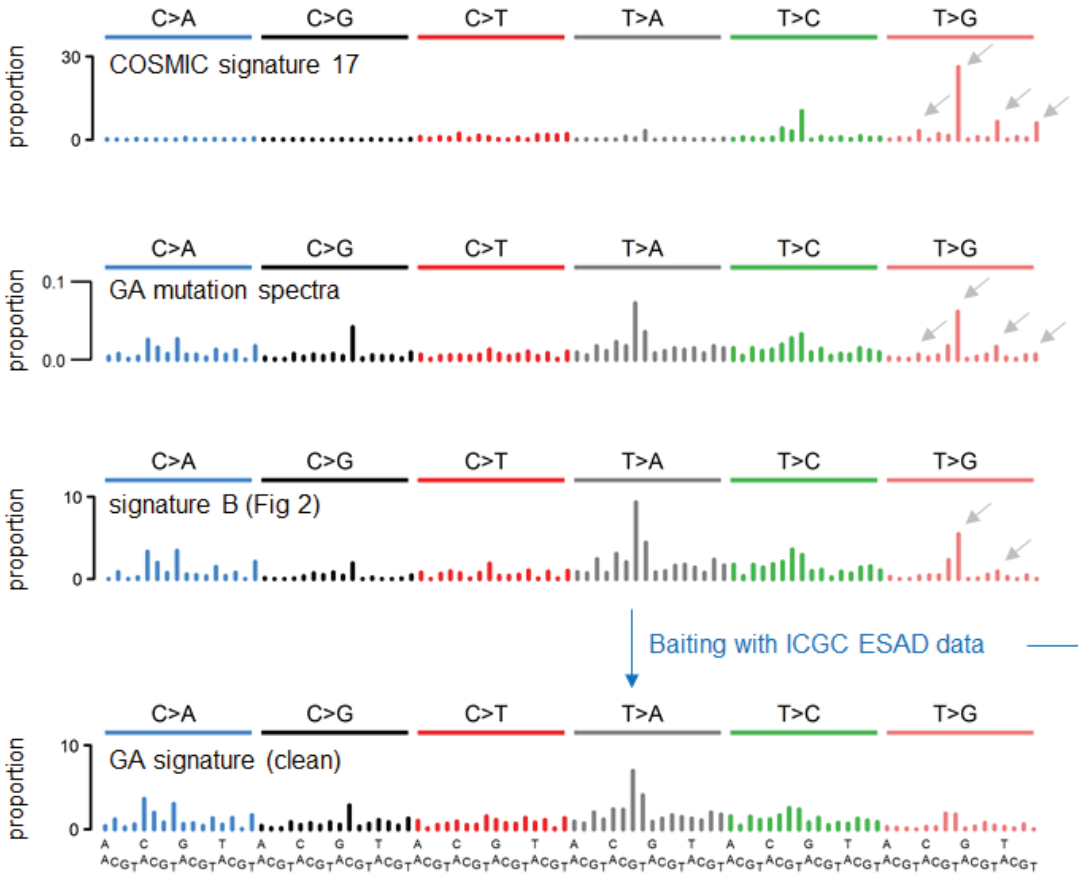
Suppl. Fig. S5


GA_1
18
10
12
32
21
7
% of 448
39.59%
GA_2
11
12
8
31
22
16
% of 573
52.4%
GA_3
16
8
11
28
23
14
% of 370
62.23%
GA_4
15
12
9
28
18
18
% of 592
33.87%
GA_5
18
11
11
25
21
14
% of 444
39.63%
1.00
0.75
0.50
0.25
0.00
1 101 201 301 401
1 101 201 301 401 501
1 101 201 301
Allelic Frequency
Rank
C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C

Suppl. Fig. S6

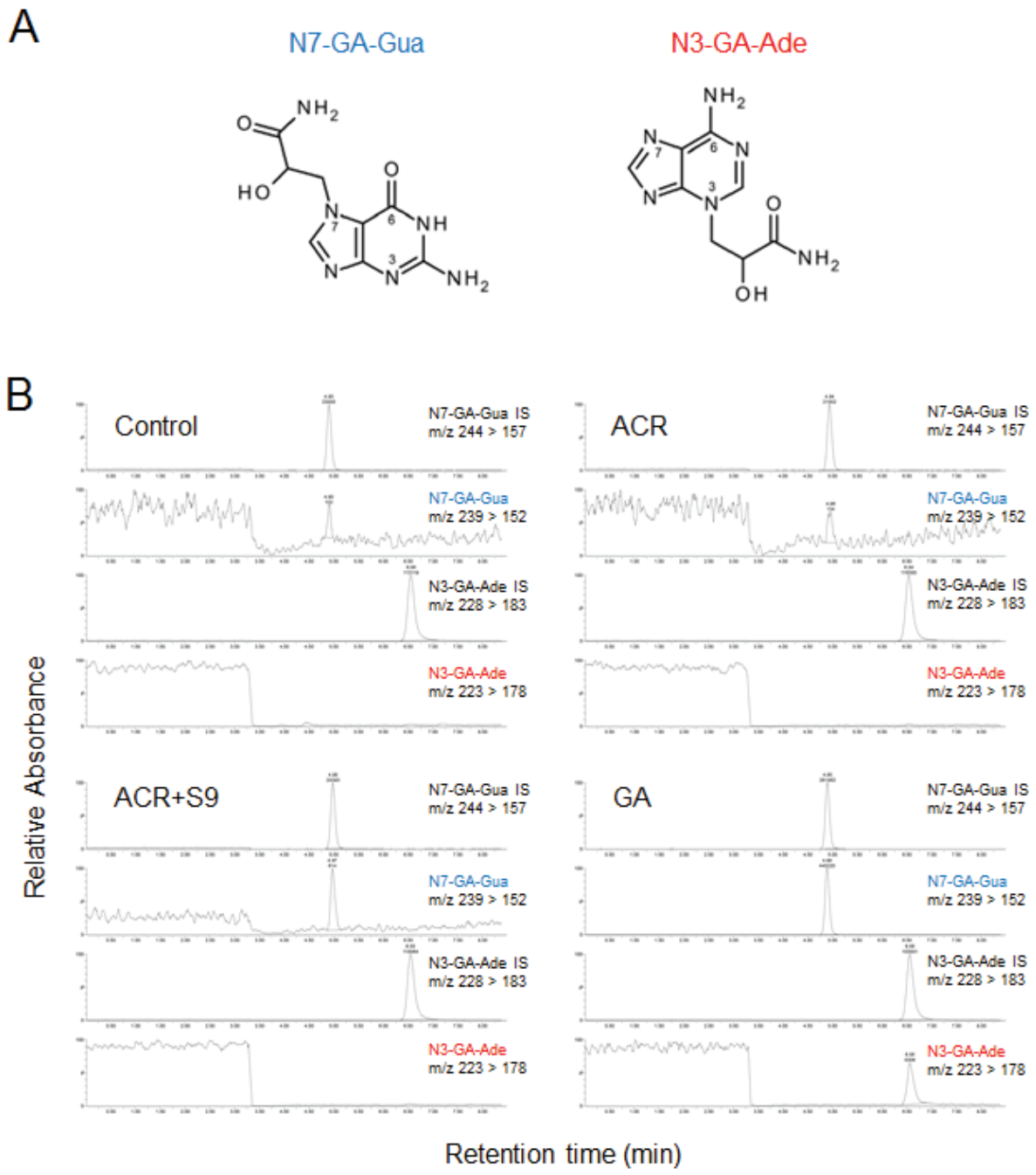

C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C
Mutations count
2000
1500
1000
500
0
0-33%
34-66%
67-100%
High VAF
Medium VAF
Low VAF
Mutations percentage
35
30
25
20
15
10
5
0
C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C
proportion
0.08
0.00
C>A (285)
C>G (213)
C>T (176)
T>A (514)
T>C (378)
T>G (264)
0-33%
400
0
C>A (65)
C>G (39)
C>T (51)
T>A (147)
T>C (111)
T>G (66)
34-66%
120
0
C>A (11)
C>G (3)
C>T (5)
T>A (15)
T>C (8)
T>G (4)
67-100%
10
0

Suppl. Fig. S7

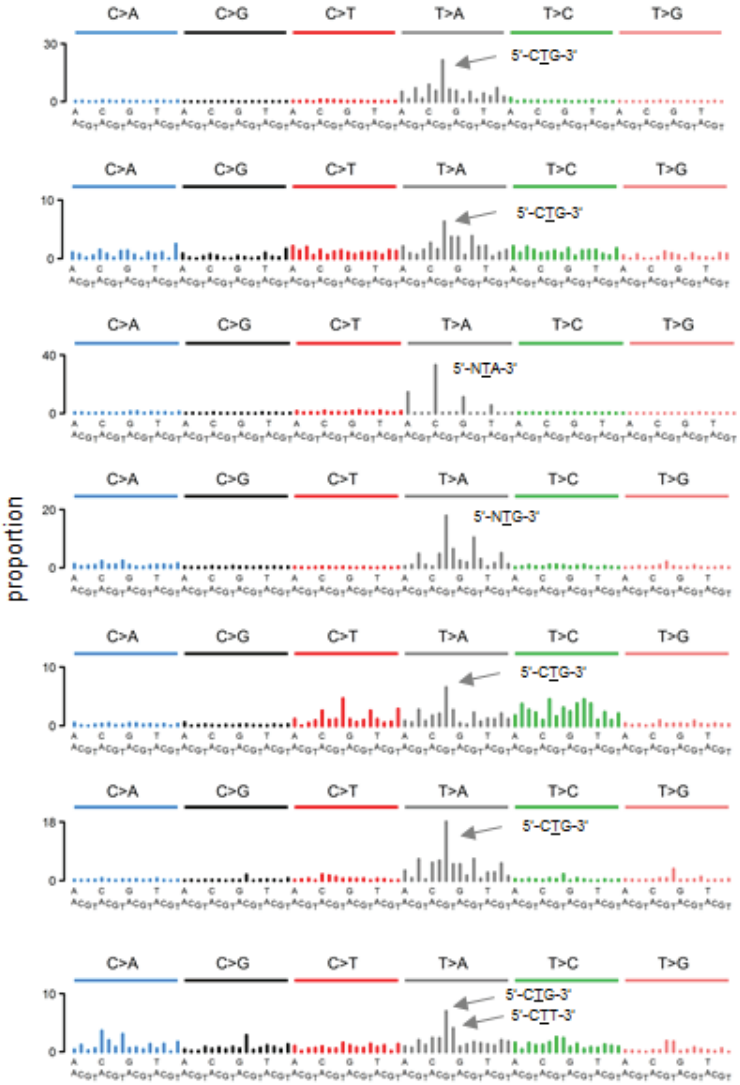


Sig. 17 peaks	Decrease (%)
C[T>A]C	0.0
C[T>A]G	25.7
C[T>A]T	8.0
C[T>C]C	22.4
C[T>C]G	29.7
C[T>C]T	20.0
A[T>G]T	100.0
C[T>G]C	42.9
C[T>G]G	19.1
C[T>G]T	68.2
G[T>G]T	52.7
T[T>G]T	0.0

Suppl. Fig. S8

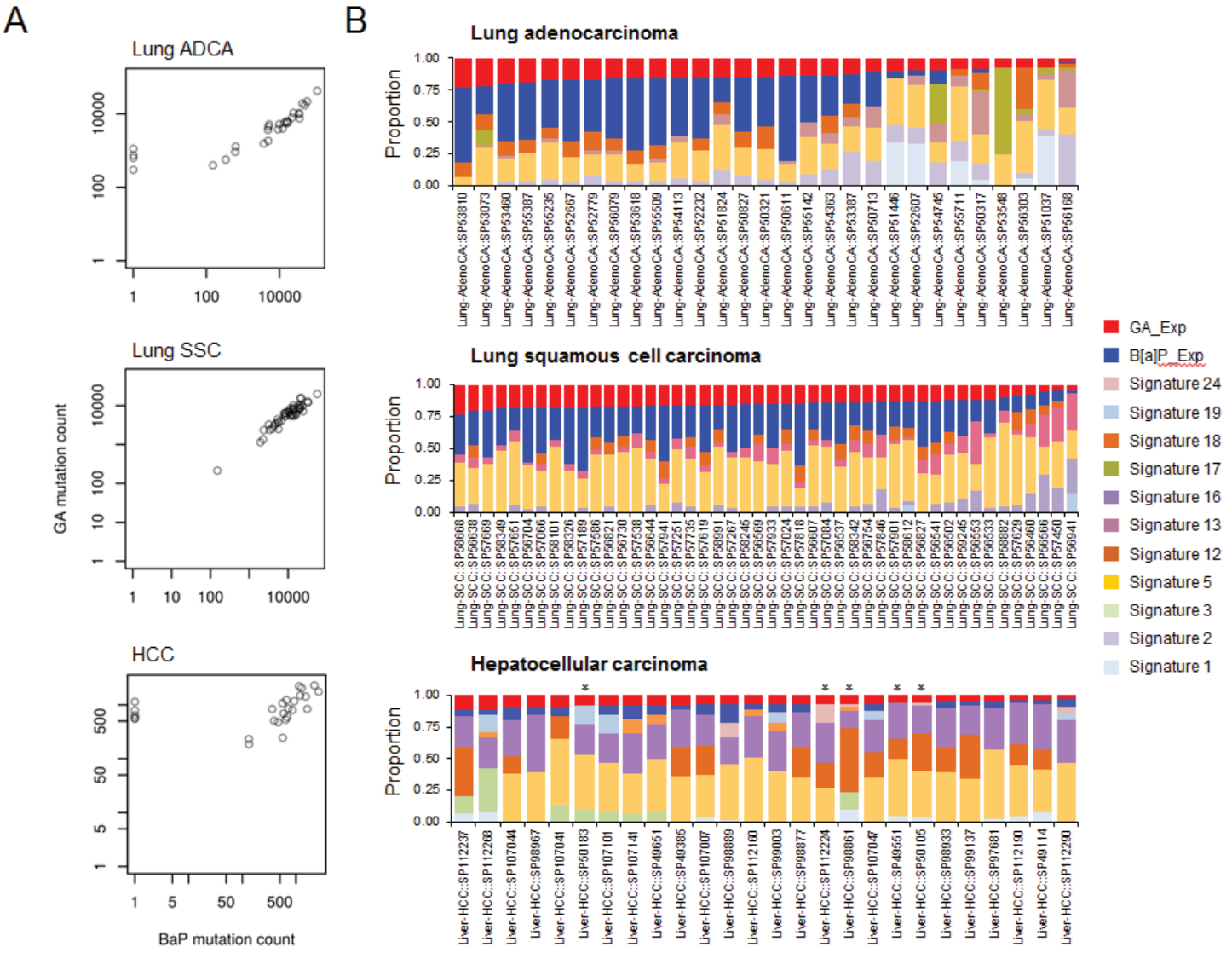

A
N7-GA-Gua
N3-GA-Ade
B
Control
ACR
ACR+S9
GA
N7-GA-Gua IS
m/z 244 > 157
N7-GA-Gua
m/z 239 > 152
N3-GA-Ade IS
m/z 228 > 183
N3-GA-Ade
m/z 223 > 178
Relative Absorbance
Retention time (min)

Suppl. Fig. S9



Exposure	Signature	Sample origin
AA [ref. 2]	sig 22	Urothelial carcinoma
Unknown [ref. 2]	sig 25	Hodgkin's lymphomas
Unknown [ref. 2]	sig 27	Kidney clear cell carcinomas
DMBA [ref. 48,49]	-	Skin cancer (mouse)
Urethane [ref. 50]	-	Lung cancer (mouse)
AA [ref. 37]	-	Hupki MEF
Glycidamide	-	Hupki MEF

Supplementary Figure S10



Paper 2: Summary of findings regarding the mycotoxin compound ochratoxin A

Title

Integrative analysis of whole-genome mutational signature of ochratoxin A in cells and rodent kidney tumors

Authors: Maria Zhivagui, Arnoud Boot, Andrea Carra', Vincent Cahais, Stephanie Villar, Steve G Rozen, Silvia Balbo, Michael Korenjak and Jiri Zavadil

Introduction:

OTA is a mycotoxin widely spread in the human diet. Exposure of rodents to OTA shows a clear evidence of carcinogenicity in the kidney of F344/N rats (National Toxicology Program, 1989). The IARC Monographs classified OTA as possibly carcinogenic to humans (Group 2B). Yet, the mode of action of OTA remains a matter of debate since the 90's. We speculate that using a validated exposure-clonal immortalization cell model may provide evidence on the mechanism of OTA on the DNA using DNA analysis and whole genome sequencing. These findings can be further corroborated by complementing the *in vitro* system with rat kidney tumors from the US NTP.

1. DNA adduct analysis

The comparison of all the data sets obtained by analysis of the samples revealed a higher number of DNA adducts detected in OTA-treated compared to the control samples. In particular, 3500 different putative DNA adducts were detected in the OTA-treated MEFs and only 1550 putative DNA adducts were detected in the untreated cells. Excluding the DNA adducts in common between the two data sets, as well as the analytes that are considered redundant, a group of 220 potential candidates was evaluated as a class of DNA adducts present in the treated cells (Figure 28).

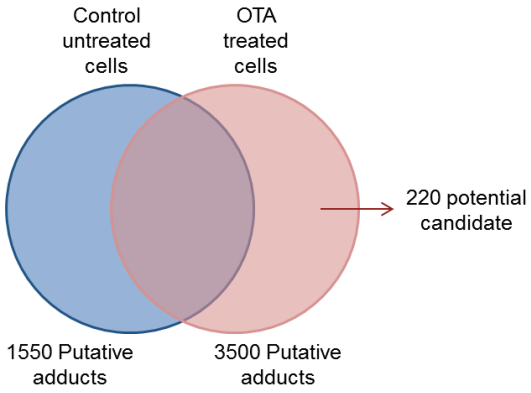


Figure 28: Putative OTA DNA adducts. Control untreated cells harbored 1550 putative DNA adducts (in blue). OTA treated primary MEFs comprised 3500 potential DNA adducts (in pink). After data mining, 220 DNA adducts were listed as potential candidates unique to OTA treatment.

The data analysis was then refined over by selecting among the 220 potential candidates those: 1) deriving from signals due to the neutral loss of the deoxyribose moiety (NL - 116.0479 m/z) and the ionization of the nucleobases (e.g. $[G+H]^+$ 152.0573 m/z) in the MS^2 and in the MS^3 detection events respectively; 2) corresponding to a clear chromatographic pattern; 3) showing a distinctive fragmentation signature (see Appendix C). These criteria allowed refining the resulting dataset. Table 6 summarizes the output from this analysis.

Full Scan	MS^2	NL^{1-2}	MS^3	NL^{2-3}	Sample name	RT (min)	Modification	Formula	Adduct Class
497.1900	381.1417	116.0483	127.0507	254.091	OTA+S9	24.79	$C_{11}H_{14}O_5N_2$	$C_{21}H_{29}O_{10}N_4$	dT
440.1944	324.1465	116.0479	152.0573	172.0892	OTA+S9	33.02	$C_{12}H_{12}O$	$C_{22}H_{26}O_5N_5$	dG
358.0999	242.0523	116.0476	152.0569	89.9954	OTA+S9	8.4	$C_2H_2O_4$	$C_{12}H_{16}O_8N_5$	dG
280.1414	164.0937	116.0477	136.0623	28.0314	OTA+S9	17.68	C_2H_4	$C_{12}H_{18}O_3N_5$	dA
332.1206	216.0731	116.0475	136.0619	80.0112	OTA+S9	9.79	C_4H_2ON	$C_{14}H_{16}O_4N_6$	dA
322.1512	206.1039	116.0473	136.062	70.0419	OTA+S9	7.35	C_4H_6O	$C_{14}H_{20}O_4N_5$	dA
456.1928	340.1444	116.0484	112.0508	228.0936	OTA+S9	6.69	$C_6H_{16}O_7N_2$	$C_{15}H_{30}O_{11}N_5$	dC
396.1893	280.1412	116.0481	152.0573	128.0839	OTA+S9	32.33	$C_7H_{12}O_2$	$C_{17}H_{26}O_6N_5$	dG
314.1101	198.0624	116.0477	136.0623	62.0001	OTA+S9	9.31	CH_2O_3	$C_{11}H_{16}O_6N_5$	dA
330.105	214.0574	116.0476	136.062	77.9954	OTA+S9	9.31	CH_2O_4	$C_{11}H_{16}O_7N_5$	dA

Table 5: **Data output**. The columns entitled Full Scan, MS^2 and MS^3 refer to lists of the ions detected in the course the three detection events. The columns NL^{1-2} and NL^{2-3} correspond to the neutral loss signals observed during the MS^2 and MS^3 detection events. The RT column reports the chromatographic retention time and finally the Modification Formula and Adduct Class columns refer to the structural information achieved calculating the chemical formulas which properly fit with the Full Scan, NL^{2-3} and MS^3 detected signals.

Moreover, these results demonstrate the induction of DNA damage that is unique and specific to OTA treatment. Yet, the assigned molecular formulas of potential DNA adducts cannot be linked to OTA structure before accounting for the variables occurring during samples treatment and LC/MS^3 chemical reactions. Further investigation and analytical approaches are needed in order to investigate potential OTA-derived DNA adducts.

2. Hupki MEFs immortalization and *TP53* mutations

Primary MEF cultures from three different embryos were exposed to OTA in the presence and absence of human S9 fraction. Applying the established cytotoxic conditions, multiple immortalized clones were derived. MEF senescence and immortalization were evident from the growth curves generated for each culture (Figure F.1a). Subsequently, the *TP53* gene was sequenced in order to assess mutagenicity of OTA and clonality of the lines derived from OTA exposure (OTA clones) and spontaneous immortalization (Spont). In the context of OTA treatment in the presence of S9 fraction, no *TP53* mutations were observed. One clone derived from OTA exposure in the absence of S9 and another derived from spontaneous immortalization carried non-synonymous mutations, T>A in codon 138 and G>T in codon 237, respectively (Figure F.1b). The chromatogram of the detected mutations suggests a high allelic frequency confirming the clonal nature of the MEF immortalization. However, due to the small number of mutations no conclusions could be drawn regarding the mutagenicity of OTA compared to untreated controls.

3. FFPE tissues processing

Working with old FFPE tissues from the US NTP, dating from 1984, was anticipated to be challenging. Therefore, we first assessed the quality of the isolated DNA to facilitate the generation of WGS libraries. The level of DNA degradation and the ability to amplify short fragments were used to guide the selection of better quality samples for library preparation. We examined the ability of tissues fixed in formalin for variable durations (between 3 – 72 days) to amplify short DNA segments using a standard PCR protocol (Figure F.2a). Tissue fixation for longer than 8 days in formalin caused the destruction of the DNA to an extent that hindered amplification of the test fragment, potentially due to cross-linking carry-over. Consequently, we prepared DNA libraries for a number of rat FFPE tissues fixed between 3 to 9 days, including normal and kidney tumor tissues (Figure F.2b). The resulting library profiles and concentrations, assessed using a Bioanalyzer, presented satisfying results with DNA fragments ranging between 200 and 350 bp (Figure F.2c).

4. Mutation spectra analysis

Whole-genome sequencing of OTA-derived MEF clones and rat kidney tumors together with the extraction of acquired somatic variants revealed that the total number of coding mutations accounted for an average of 2.3% and 2.6% of the total number of mutations in MEFs and rat tumors, respectively. To estimate the extent of sequencing-related damage in our samples, we determined the global imbalance variation (GIV) score of each sample as described in methods and in (Chen et al., 2017). No detectable sequencing artifacts for any of the mutation types were observed in our dataset (data not shown). The Hupki MEF clones derived from OTA exposure showed an enrichment of T:A>G:C transversions related to signature 17 from COSMIC (Alexandrov et al., 2013b). The kidney tumor samples exhibited an overall diffuse pattern across the six different SBS types, in addition to the more prominent spontaneous deamination of methylcytosine characterized by C:G>T:A transitions at CpG islands (Figure 29). No significant transcription strand bias was observed for any of the mutation classes in the kidney tumors or MEF clones. PCA analysis comparing the variants called by MuTect versus Strelka demonstrated similarities in profiles for both MEF clones derived from OTA exposure and rat kidney tumors developed upon OTA treatment after the elimination of low allelic frequency mutations (AF<20%) (Figure F.3 a-b).

5. Mutational signature analysis

Using NMF, we analyzed all the OTA-derived samples for the presence of mutational signatures. Two mutational signatures were identified (Figure 30 a-b), with signature A enriched in OTA clones, and signature B enriched in the rat kidney tumors. In signature A we observed a major enrichment of a pattern identical to COSMIC signature 17 (T:A>G:C in 5'-N<u>T</u>T-3' trinucleotide context). As signature 17 is ubiquitously found in Hupki MEF clones, we used a baiting approach to reduce its contribution to signature A (Figure F.4). Within signature A, we uncovered a mutation pattern potentially linked to treatment with OTA, characterized by C:G>A:T transversions (Figure 30c). Focusing on the C>N pattern only, we observed similarities between signature A and signature 18 and 36 from COSMIC (cosine similarity=0.86 and 0.897, respectively; COSMIC signature 36 is unpublished, data have been provided by Dr. L. Alexandrov). Signatures 18 and 36 have been related to an ongoing ROS production and ROS production against the background of mutated *MUTYH* gene, respectively. Furthermore, we separated COSMIC signature 1, related to age, from signature B. This yielded 0.91 cosine similarity of signature B with COSMIC signature 5 attributed to the clock-like mutational signature process (Figure 30d).

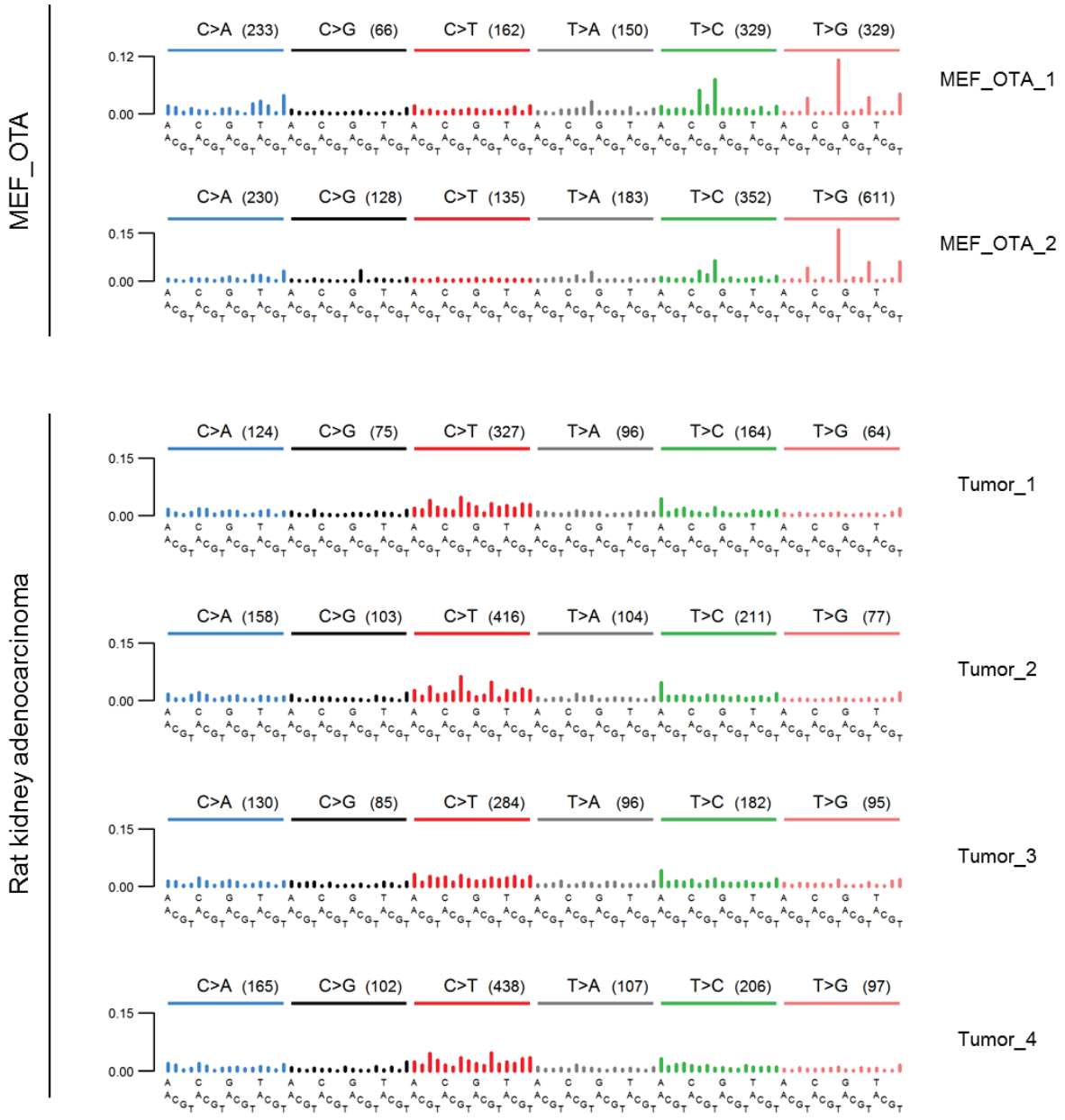


Figure 29: Genome-wide mutation spectra of OTA-derived MEF clones (upper panel) and OTA-induced kidney adenocarcinoma tumors (lower panel). X-axis represents the trinucleotide sequence context. Y-axis represents the frequency distribution of the mutations.

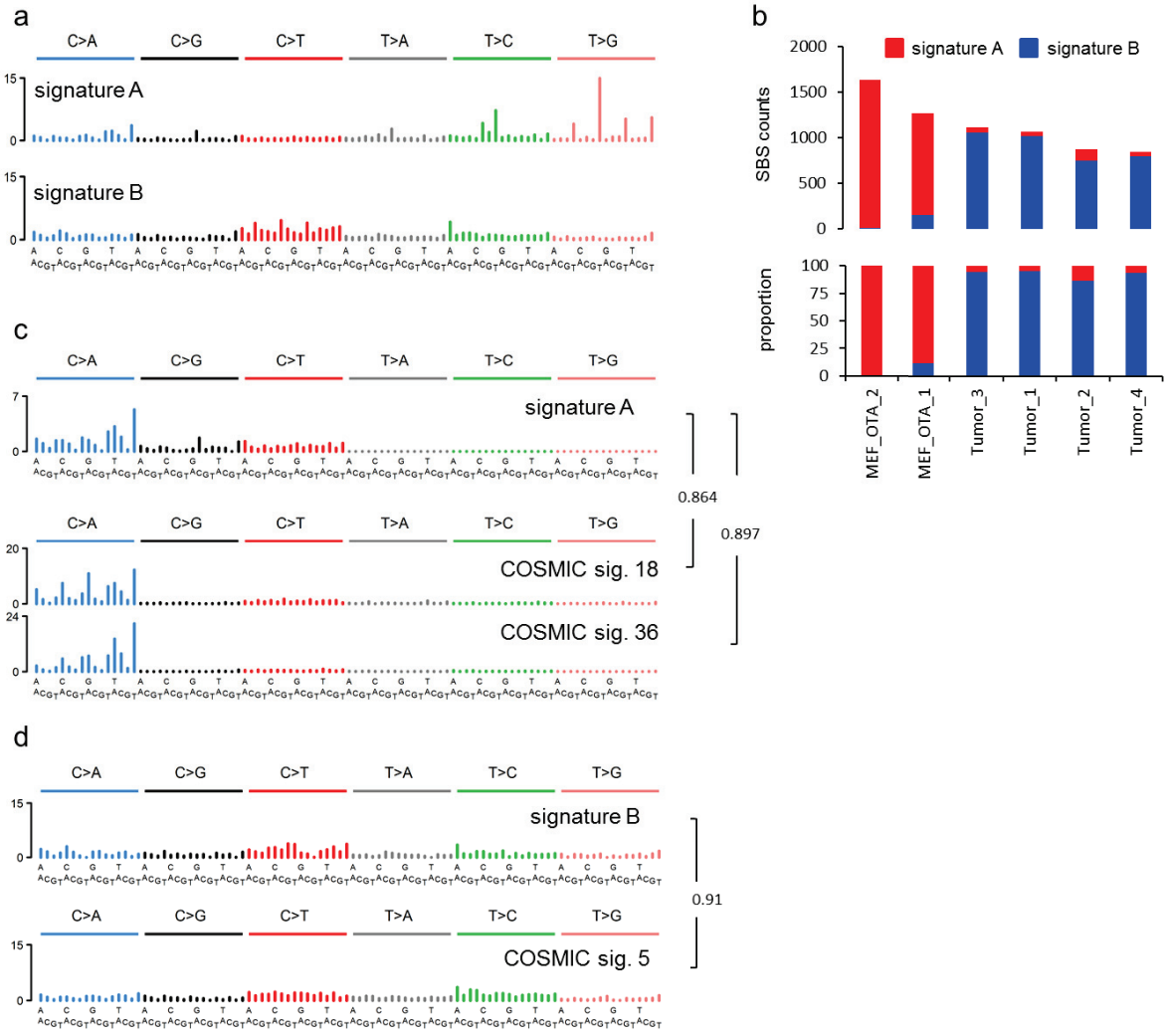
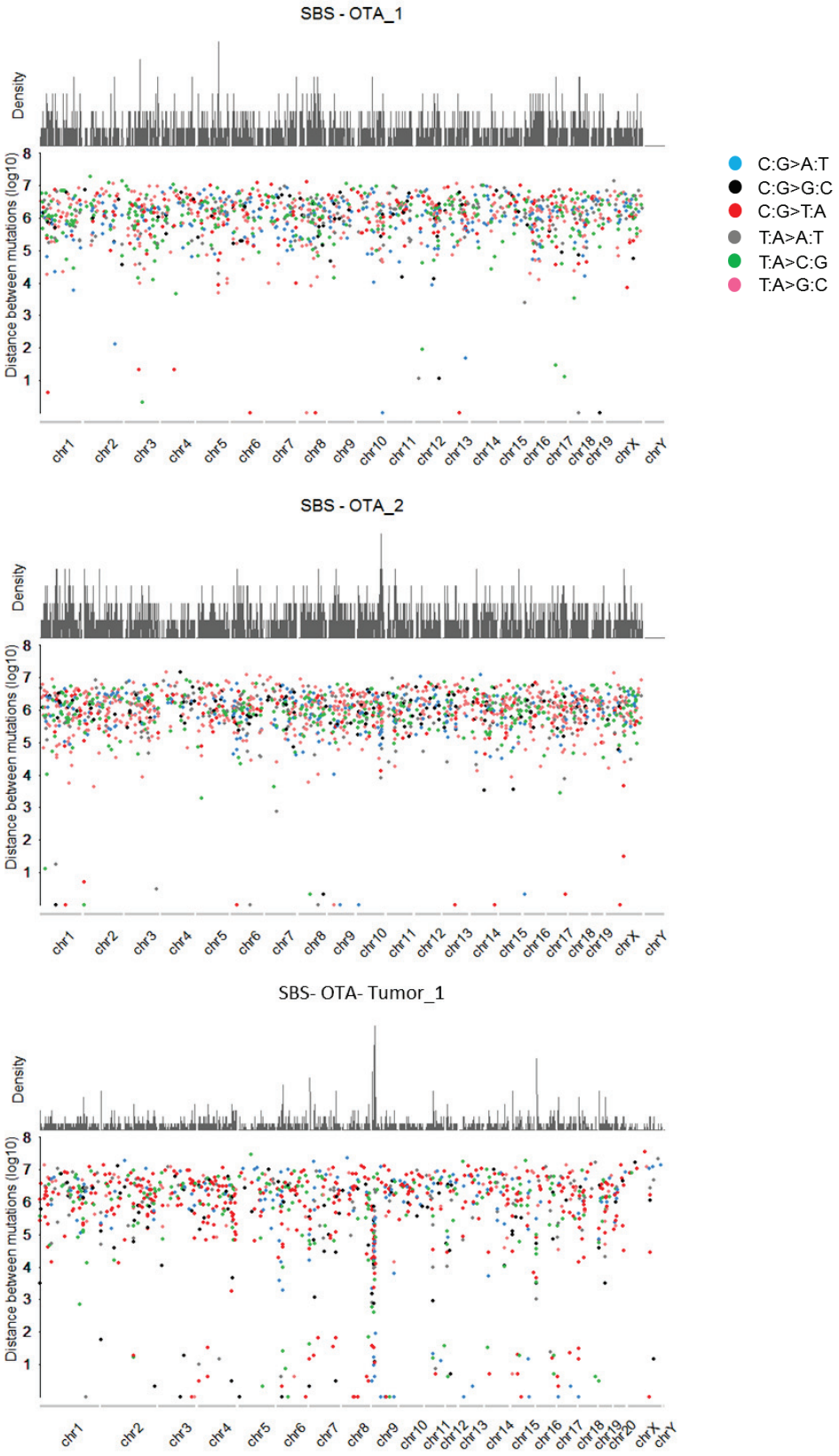


Figure 30: Mutational signature analysis. a) Mutational signatures identified by non-negative matrix factorization (NMF) in the 2 Hupki MEF cell lines and 4 rat kidney tumors from exposure to OTA (signature A and signature B). X-axis represents the trinucleotide sequence context. Y-axis represents the frequency distribution of the mutations. (b) Contribution of the identified signatures to each sample (X-axis), assigned either by absolute SBS counts (top track) or by proportion (bottom track). (c) C>N mutation profile of signature A compared to ROS-related mutational signatures from COSMIC, signature 18 and 36. (d) Mutational signature B resembling signature 5 after removal of the age signature. Numbers on the right represent cosine similarity values.

6. Distribution of somatic mutations on the genome

Rainfall plots show the distance between consecutive mutations in a logarithmic scale. Looking for kataegis-like events based on SBS variants, we noticed some regions in the tumor samples with increased load of mutations (Figure 31). OTA-MEF clones did not show obvious aberrant mutated regions, whereas OTA-tumor-1 manifested an increased mutation density on chromosome 9. Further analysis is ongoing including other genomic alteration features such as indels and CNV.


SBS - OTA_1
Density
Distance between mutations (log10)
C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chrX chrY
SBS - OTA_2
Density
Distance between mutations (log10)
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chrX chrY
SBS- OTA- Tumor_1
Density
Distance between mutations (log10)
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chrX chrY

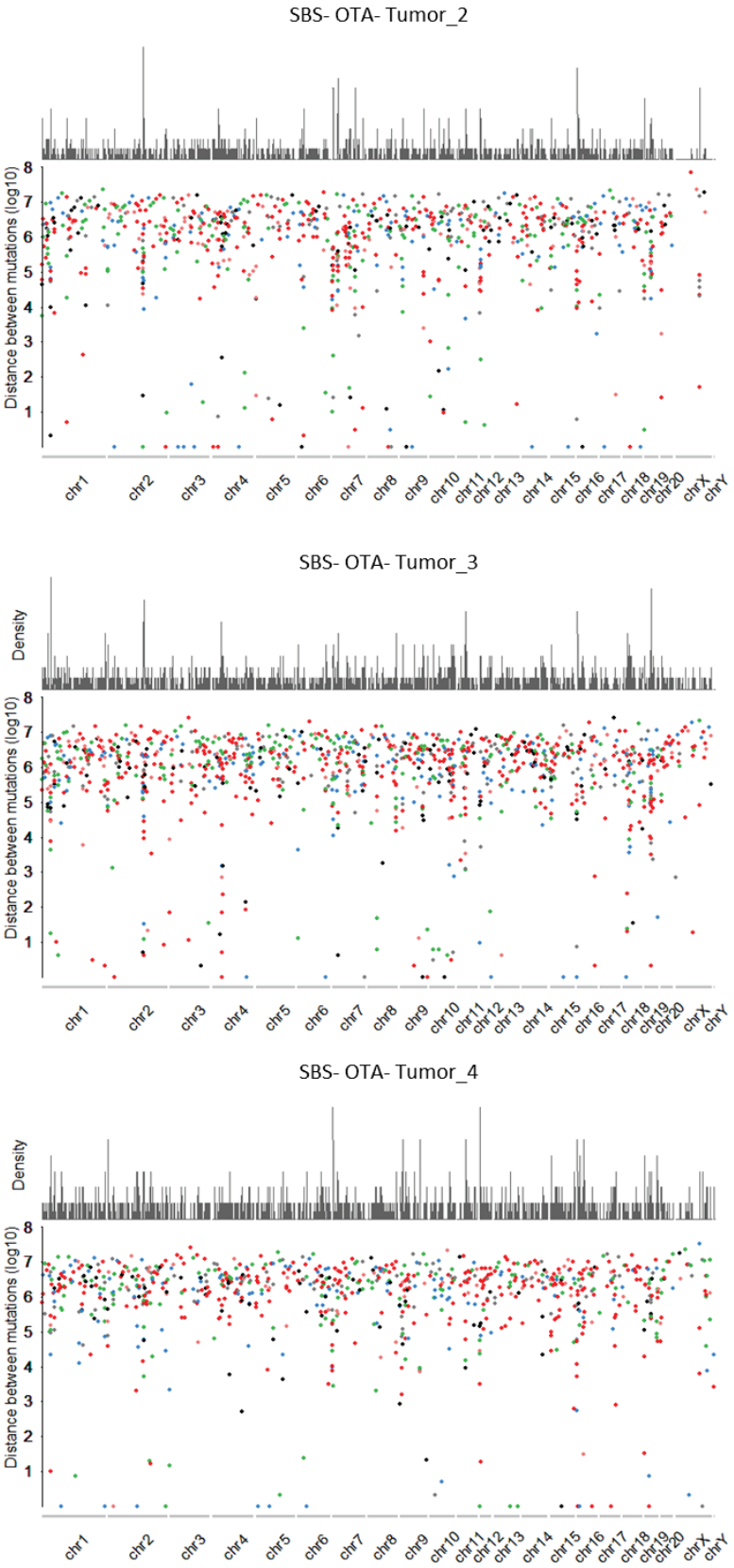


Figure 31: The distance between SBS mutations in both OTA clones and kidney tumors is plotted in the log scale (y-axis) versus the genomic position on the x-axis.

DISCUSSION

The cancer genome reflects various complex assaults accumulated throughout the life of the patient. Genome-wide mutation analysis of thousands of human cancers highlighted a number of mutational signatures attributed to endogenous and exogenous exposures. In order to reveal the causal factor underlying a mutational signature, the convergence of multiple lines of evidence from different areas of research is needed, including experimental studies, epidemiology and individual patient exposure history (Hollstein et al., 2017). In addition to well-understood mutational signatures, several orphan signatures were identified and due to the lack of relevant data and biological information, there is an opportunity to experimentally investigate the genome-wide mutagenic effects of candidate cancer-risk factors using relatively simple cellular models (Zhivagui et al., 2016).

During my PhD Thesis work, I explored the establishment of such new human cellular models for the IARC MutSpec project and I was able to generate NGS-based mechanistic evidence regarding the mutagenicity of high priority compounds found in the human diet and in the environment. Thus, the presented work and its results address the timely opportunity in applying experimental systems to the analysis of mutational signatures and revealing the potential associations with human cancers.

1. Establishing mammalian *in vitro* models for exposure assays

Normal primary cells, both human and rodent, have a limited lifespan when transferred from their *in vivo* environment to culture. They undergo stress-associated senescence, with permanent exit from the cell cycle and eventual death of the cell population. Occasionally, one cell may bypass this fate due to genetic changes and resume the cell cycle producing a clone of immortalized cells that replicate indefinitely *in vitro*. Clonal expansion is a prerequisite property of the system allowing to investigate the acquired somatic mutations in more or less homogeneous cell populations by deep sequencing analysis. Through extensive laboratory work using murine and human cells, we characterized various advantages and disadvantages for each cellular model.

Hupki MEF cells proved suitable in terms of cell culture protocol and reliable immortalization due to their sufficiently long telomeres and the telomerase enzymatic activity. The telomere shortening-mediated replicative senescence as observed in human primary cells (Espejel and Blasco, 2002) thus does not occur in Hupki MEFs. This cell model uses primary normal diploid cells to give rise to clones that have bypassed the selective biological barrier step within 2-3 months, which is usually achieved through disruption of the *p19/ARF/p53* pathway (Olivier et al., 2014; Zhivagui et al., 2016; Liu et al., 2004, 2005). Yet, potential limitations of

the MEF assay system include inadequate metabolic activation of certain carcinogens, the high rate of spontaneous immortalization, and potential species-specific differences in key biological pathways involved in cell transformation (Zhivagui et al., 2016). 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 metabolically activate pro-carcinogens. Importantly, however, the MEFs in culture have been shown to convert a variety of pro-carcinogens to their reactive intermediates (vom Brocke et al., 2006, 2008; Feldmeyer et al., 2006; Liu et al., 2004, 2005; Luo et al., 2001b; Reinbold et al., 2008). Frequent spontaneous immortalization of MEFs has been attributed to high mutation rates resulting from oxidative stress under standard culture conditions (20% oxygen) (Busuttil et al., 2003; Parrinello et al., 2003). Thus, growing MEFs under physiological oxygen levels (3-5%) can reduce background mutation and spontaneous immortalization rates, and should improve the stringency of the MEF system as well as the reliable identification of exposure-related mutation patterns.

In order to circumvent some of these constraints, we complemented the Hupki MEF system with a newly developed system that meets the key requirements. The HepaRG cell line is a human liver progenitor bipotent cell model. The cells can fully differentiate towards mature hepatocytes that can serve as surrogate to human primary hepatocytes (Lübberstedt et al., 2011). Testing the HepaRG model for both, BBCE and CE assays enabled us to address several questions that arose during manipulation of the HepaRG cells. We show that progenitor HepaRG cells experience a crisis-like state, characterized by reduced cell growth, followed by its bypass. Single-cell subcloning was only successful after the cells overcame this potential crisis, raising the possibility that a cell that acquired a proliferation advantage and a more transformed state was able to clonally outgrow. We also examined the ability of the cells to differentiate following bypass of the crisis-like state. As proposed on the supplier's web page (http://www.HepaRG.com), we found that the progenitor cells, at late passages, lost their bipotent potential and failed to differentiate into hepatocyte-like cells in culture producing a biliary-like cell population solely. This crisis bypass may be due to karyotype instability, genetic alterations, epigenetic modifications or chromatin structure changes.

Moreover, we implemented a technique to isolate hepatocyte-like cells for exposure using partial trypsinization (Cerec et al., 2007), and further gene expression assessment inferred that the hepatocytes necessitate 4 days of incubation in order to re-arrange in a monolayer and attain their metabolic activity. Such a pure culture of hepatocyte-like cells may be exposed to carcinogens under the BBCE scenario. Nonetheless, we show that isolated hepatocytes tend to progressively lose cell type-specific and metabolic markers and revert

back to the progenitor cell state, complicating thus the protocol relying on senescence, senescence-bypass and immortalization of hepatocytes upon carcinogen treatment. Moreover, in order to maintain the differentiated state of the cells by keeping them at high density, chronic exposures, using non-cytotoxic doses, should be of choice for this assay. The HepaRG model requires lengthy exposure experimentation (about 6 months) and further verification regarding its “immortalization” or “clonal expansion” states (Zhivagui et al., 2016). Recently, an elegant study to which our team contributed, conducted a chronic exposure assay using HepaRG cells after hepatocytes isolation in order to discern the genome-wide mutational signature of AFB1. The results underscored the relevance of the HepaRG system replicating findings from human HCC exposed to AFB1 (Huang et al., 2017). Another advantage of the HepaRG cell model is its applicability to investigate compounds known to target the liver, such as AA, AFB1 and methyleugenol.

Despite our prioritization of the rapid Hupki MEF system over the HepaRG cells (reflecting the 3-year timeframe for the PhD work), we made substantial progress and achievements in characterizing a potential HepaRG crisis-like event and in establishing protocols for crisis bypass, hepatocyte isolation and cytotoxicity evaluation. Based on this work, we conclude that more work is warranted to fully develop the HepaRG system into a model for assessing exposures related to human malignancies, namely those of the liver.

Moreover, the use of human cancer target cell models is an elegant strategy to identify the mutation pattern of a carcinogen that triggers cancer development in a specific site.

In addition to the applied model systems, other human cell line models have been explored for the analysis of mutation patterns, namely a proximal tubule human kidney cell line (HK-2), HepG2 cells derived from HCC and Human Mammary Epithelial Cells (HMEC) (Hoang et al., 2013; Huang et al., 2017; Poon et al., 2013; Severson et al., 2014; Zhivagui et al., 2016). As for most human model systems, their use requires long-term exposure and the number of compounds that can be tested is a limiting factor.

Emerging models such as induced Pluripotent Stem cells (iPSc) and organoids could be versatile systems that can be generated from different cell types and tissues and are capable of clonal expansion (Blokzijl et al., 2016; Zhivagui et al., 2016). However, these models lack the biological barrier step of HMEC and Hupki MEFs and it is not clear how this may influence mutational signature formation.

Lastly, taking in consideration *in vivo* carcinogen-animal models may provide new avenues for a better understanding of the molecular alterations observed in human diseases, and thus improve our knowledge on tumor initiation, progression, diagnosis and treatment.

Remarkably, an attractive paper published in 2018 used a chemically-induced mouse model to recapitulate human disease. The N-butyl-N(4-hydroxybutil)nitrosamine (BBN) mouse model developed muscle-invasive bladder cancer characterized by many analogies with primary human bladder cancer at various molecular levels, including gene expression, pathways, and mutation patterns. This model proved to be suitable for studying bladder carcinogenesis (Fantini et al., 2018).

Nevertheless, *in vivo* models are time consuming, labor extensive, costly and most importantly involve the direct use of animals. In addition, replication of real life human exposure to carcinogens is limited in experimental animals *in vivo* and more so, in *in vitro* cell culture models. Generally, humans are exposed to chronic doses of carcinogens over a span of several years to a few decades. The finite lifespan of (primary) cells in culture as compared to relatively longer lifetime of animals (i.e., days/weeks vs. a few years) makes modeling of human exposure to carcinogens much less realistic in the former models. Furthermore, both the *in vitro* and *in vivo* model systems can't fully recapitulate all aspects of human carcinogenesis due to differences in pharmacokinetic and pharmacodynamic properties of chemicals between the cultured cells *in vitro* or experimental animals *in vivo* and humans. This said, however, carcinogenicity studies in *in vitro* cell culture models can provide an initial indication of the cancer-causing potential of a given chemical/agent(s), and the results can be used as a guide to design 'refined' *in vivo* experiments with 'reduced' number of animals (to comply with the '3Rs' as guiding principles of the ethical use of animals for experimental research), followed by well-designed population-based/clinical studies.

2. Considerations for applying NGS to analyze FFPE tissues

FFPE samples represent an invaluable resource for retrospective and prospective molecular studies, especially when Fresh-Frozen (FF) tissues are not available. In addition, FFPE tissues from bio-archives offer indispensable materials that can help reduce new laboratory animal manipulation as well as generate substantial added value from these past studies. The US National Toxicology Program archives data and tissues from animal bioassays exposed to environmental agents. The design of NTP studies has historically been focused on (histo)pathological examination of the samples, however, technological advances have also resulted in a more recent increase in molecular studies. Exploiting FFPE tissues from the US NTP biobank permits data integration across studies and systems for novel meta-analysis. Nevertheless, applying NGS for the analysis of FFPE samples remains a challenging task. In order to extract nucleic acids from FFPE tissues, the paraffin needs to be removed and protein-DNA interactions resulting from fixation have to be reversed. Moreover, tissue preparation, the fixation process, fixation delay, paraffin embedding, archiving conditions and storage time are, in some cases, inevitable factors that can cause cross-linking reactions and chemical modifications of the DNA as well as DNA fragmentation (Einaga et al., 2017; Hedegaard et al., 2014). Therefore, optimized protocols for DNA extraction and library preparation using FFPE tissues are warranted in order to yield sufficient amount of DNA that is of good quality for NGS studies.

Failure of amplification of the DNA sequencing library can often be due to inefficient adapter ligation or DNA polymerization blockage caused by extended fixation times or DNA degradation caused by long storage times of the FFPE blocks. Hedegaard and colleagues highlighted the effects of tissue storage time on library preparation and sequencing quality by demonstrating that the concordance between FF and FFPE tissue in the context of DNA sequencing is affected by the storage time of the FFPE tissues. Samples stored for more than 3 years showed less reliable results, when compared to their FF counterparts. This manifested through higher duplication rates, smaller insert sizes, a lower fraction of mappable reads, a larger fraction of imperfectly mapped reads and reads mapping with unaligned ends. (Hedegaard et al., 2014).

Different DNA isolation strategies were investigated in our laboratory to establish a protocol that yielded optimal DNA amounts from FFPE tissues for NGS. However, DNA quality does not correlate with DNA quantity and further testing was necessary in order to select samples most suitable for molecular analyses from the available FFPE tissues. PCR amplification of two rat genes, *P53* and *Kras* (data not shown), showed that FFPE samples fixed for more than 8 days in formalin failed to efficiently amplify the test regions (Figure F.2). Additional

protocol adjustments, such as inclusion of the DNA damage repair kit, extended adaptor ligation and removal of small DNA fragments were applied in order to aid library preparation and sequencing. Using these adjustments, we were ultimately able to produce libraries that resulted in good quality NGS of very old rat FFPE tissues (Table F.1).

3. NGS and mutational signature identification using Hupki MEF system

The Hupki MEF immortalisation assay was shown to recapitulate *TP53* mutation patterns in the context of specific mutagenic carcinogen exposures (Liu et al., 2004, 2005). With the advent of massively parallel sequencing, it is now possible to extend the screen to genome-wide genetic alterations. Genome-wide sequencing of Hupki MEFs exposed to a number of carcinogens harboured a suite of base substitutions that recapitulate exome-wide mutation data derived from human cancers (Nik-Zainal et al., 2015; Olivier et al., 2014). The large number of mutations that can be derived from single samples using NGS-based approaches eliminates the need to interrogate a single gene, such as *TP53.* In practice*,* many cell populations are needed to accumulate the number of mutations for a single gene assay, whereas, NGS of one single immortalized cell can provide enough information to identify a mutational signature. A high coverage is required to perform single-cell sequencing in order to minimize the level of spurious variants due to sequencing errors. The clonal expansion step in the MEF immortalization protocol helps to enrich for a more or less homogeneous population of cells that have acquired cancer-like properties and characteristic mutation patterns, which can be identified using NGS. Using multiple cell line replicates per exposure or condition is warranted to generate highly robust mutation signatures. The scope of overlap between mutation patterns in human datasets and immortalised MEF cell lines includes: (a) the predominant mutation patterns, (b) the transcription strand bias of the specific mutation types, and (c) the sequence context of the dominant mutation type (Olivier et al., 2014). The Hupki MEF cell model, coupled with exposure to cancer-risk agents under well-controlled experimental settings, allowed the identification of a novel exome-scale mutational signature of glycidamide and suggested a lack of mutagenicity of OTA at the whole genome sequencing level.

3.1. Acrylamide and its metabolite glycidamide

Hupki MEF exposure to acrylamide and glycidamide provided high reproducibility of the exome-wide mutation profiles between the different cell line replicates (n≥5). In contrast to the diffuse mutation pattern induced by acrylamide, the glycidamide mutational signature was characterised by the predominance of T:A>A:T, T:A>C:G and C:G>A:T mutation patterns coupled with transcription strand bias towards the non-transcribed strand for the first two mutation types, implying the contribution of transcription-couple DNA repair to the signature. The main sequence contexts of the identified mutation types included 5'-C<u>T</u>G-3', 5'-C<u>T</u>T-3', 5'-C<u>C</u>A-3' and 5'-C<u>C</u>T-3'. This mutational signature was shown to be novel and unique when

compared to other established mutational signatures (Paper 1, Figure 3G). Interestingly, the predominant mutation types observed in the glycidamide-mutational signature corroborate the findings of the DNA adducts analysis. Increased levels of N3-GA-Ade and N7-GA-Gua DNA adducts have been linked to the formation of abasic site lesions that can bypass DNA repair and cause misincorporation of adenine during DNA replication, which leads to A>T and G>T substitutions (Besaratinia and Pfeifer, 2005; Ishii et al., 2015; Randall et al., 1987). Another prominent glycidamide DNA adduct, N1-GA-Ade, has been suggested to act as a miscoding DNA adduct generating A>G mutations, which could explain the high levels of T:A>C:G mutations in glycidamide-derived clones. Hence, this controlled study established a clear link between glycidamide DNA adducts and the resulting mutational signature.
Although WES data analysis provided satisfying results and proved to be a good and a cost-effective methodology to identify the mutational signature of glycidamide, we hypothesize that the significantly higher mutation numbers that can be derived from whole-genome sequencing would increase the reliability of the identified mutational signature, possibly including the significance of the trend-like strand bias for C:G>A:T mutations.

Using this unique mutational signature of glycidamide as a starting point, further analysis is required to screen for acrylamide/glycidamide signature in the large number of human tumor data. For this purpose, it will be necessary to first thoroughly define the mutational signature *in vitro* using human cell models, followed by *in vivo* studies exploiting animal tumors. Ideally, results from these model systems would be complemented by a well-controlled epidemiological study focusing on dietary as well as occupational settings with extensive and reliable exposure assessment (e.g. novel biomarkers).
We anticipate that such complementary studies can assist in the classification of glycidamide as well as potential reclassification of acrylamide by programs such as the IARC Monographs. The evaluation of acrylamide has not been updated since 1994, despite a considerable body of information that has emerged since then, especially with the discovery of acrylamide in the human diet, suggesting that its toxicity might reach beyond occupational settings into the daily and long term exposure of humans.

3.2. Ochratoxin A

Due to the findings that OTA induces kidney cancer in animals and that increased levels of OTA were found in human specimens (blood, urine and milk), indicating a potential chronic human exposure to OTA in a variety of settings (Clark and Snedeker, 2006; Krogh et al., 1977; Pfohl-Leszkowicz and Manderville, 2007; Radić et al., 1997), a better understanding of the mechanism of toxicity of OTA is warranted in order to provide adequate human risk assessment and carcinogen classification. We applied genome-wide mutation analysis using

in vitro model systems and rat kidney tumor samples, together with state-of-the-art DNA adduct analysis to help elucidate the mode of action by which OTA prompts carcinogenesis.

Given the lack of a specific mutational signature of OTA in either experimental model, despite using high concentrations of the mycotoxin that resulted in approximately 50% primary cell death, our findings are in agreement with a subset of previous reports that argue against a direct genotoxic effect of OTA.

We observed an enrichment of signature 17 from COSMIC in OTA-derived MEF clones as well as in spontaneous clones. Signature 17 has been lacking a known etiological factor. While some candidate causal factors have been proposed in esophageal adenocarcinoma and gastric cancers (acid reflux, *H. pylori*) (Secrier et al., 2016) further studies are required to establish the presence of this signature in *in vitro* immortalized clones derived from Hupki mouse embryonic fibroblasts. At the genome level, we were able to detect a well-defined mutation pattern manifested by C:G>A:T transversions with a lack of transcription strand bias, similar to signature 18 from COSMIC. This pattern showed a cosine similarity of 0.89 with signature 18, attributed to ROS damage as well as to MUTYH mutations (Pilati et al., 2017; Viel et al., 2017). ROS production is not uncommon in cell culture, however, as per the previously sequenced Hupki MEF clones at the exome level, signature 18 has not been previously observed for dozens of clones processed thus far. In fact, Hupki MEFs were grown in culture in medium supplemented with an antioxidant reagent, β-mercaptoethanol, likely to reduce oxidative stress. It is possible that WGS data and the more mutation counts detected in the non-coding regions of the genome, could allow the detection of ROS-mediated mutational signature. For this purpose, additional spontaneously immortalized clones are being whole-genome sequenced and analyzed. In contrast to the *in vitro* model, the mutational signatures extracted from rat kidney tumors were characterised by the presence of C:G>T:A mutation at CpG sites, corresponding to COSMIC signature 1, as well as by COSMIC signature 5. These signatures have been linked to the aging process. Signature 1 is ascribed to the spontaneous deamination of methylcytosine at CpG islands inducing the conversion of cytosine to thymidine. Signature 5 is a less defined mutational signature displaying a diffused pattern across mutation types. Alexandrov et al. revealed that patient age correlates with the contribution of mutational signatures 1 and 5 to the overall mutation pattern (Alexandrov et al., 2015). Hupki MEF clones do not show similar age-mutational signatures but rather manifest a prominent signature 17 upon immortalization. The lack of signatures 1 and 5 in the Hupki MEFs might be attributed to the *in vitro* culture settings, where murine cells are grown for a few passages until they reach immortalization, whereas *in vivo* in patients or animal models, tumor development is a much longer process.

OTA-induced ROS production has been well studied previously using various kidney cell models from human and rodent origins (Costa et al., 2016; Giromini et al., 2016; Jia et al., 2016; Mally et al., 2005; Sheu et al., 2017; Yang et al., 2014). We speculate that the observed 18-like signature in OTA-derived MEF clones may be specific to OTA treatment mediated through the production of ROS and oxidative stress. In contrast, we did not detect this mutational signature in the rat kidney tumors (cosine similarity of 0.46 – data not shown). In fact, it has been suggested that OTA exerts its carcinogenicity in rats through cell proliferation rather than oxidative stress (Qi et al., 2014). This is underpinned by the high levels of COSMIC signatures 1 and 5 observed in rat kidney tumors, reflecting that OTA might have triggered cell proliferation and cell division making the cells prone to replicative damage leading to cancer development.

Preliminary results on DNA adduct formation showed a possible future direction for the characterization of OTA-derived DNA damage. Indeed, further investigation and data analysis is warranted in order to draw conclusive results regarding the chemistry of OTA and its reaction with the DNA by including labeled OTA exposure, extensive scrutiny of the structures that have lost the deoxyribose and showed clear MS^3 fragmentation peak and investigation of possible indirect metabolite-mediated DNA adduct formation through ROS, for instance.

In addition to SBS-based genome alterations and mutational signatures, whole-genome sequencing enables the analysis of complex chromosomal aberrations, structural and copy number variations, clustered mutations, replication timing and mutations along the transcript length. It is conceivable that further bioinformatics analyses will provide additional insight into the potential mechanism of OTA during cell transformation.

3.3. Other compounds

The treatment by Cr(VI) did not induce any mutations in 8 different Hupki MEF clones, as assessed by *TP53* gene sequencing, whereas MNU caused a GCC>GTC non-synonymous mutation in *TP53* codon 138. The generated clones are expected to be analyzed by genome-wide sequencing for the discovery of unique genetic alterations and possible mutational signatures.

CONCLUSION

In conclusion, this PhD work provided insights into the applicability of different experimental models to the identification of exogenously induced mutational signatures. Characterization of novel mutational signatures specific to cancer-risk agents, such as the one identified for the probable dietary carcinogen acrylamide/glycidamide, may ultimately contribute to the overall, interdisciplinary mission of cancer research for cancer prevention.

BIBLIOGRAPHY

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013a). Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep. *3*, 246–259.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013b). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407.

Alitalo, K. (1984). Amplification of cellular oncogenes in cancer cells. Med. Biol. *62*, 304–317.

Andersson, T.B., Kanebratt, K.P., and Kenna, J.G. (2012). The HepaRG cell line: a unique *in vitro* tool for understanding drug metabolism and toxicology in human. Expert Opin. Drug Metab. Toxicol. *8*, 909–920.

Aninat, C. (2005). EXPRESSION OF CYTOCHROMES P450, CONJUGATING ENZYMES AND NUCLEAR RECEPTORS IN HUMAN HEPATOMA HepaRG CELLS. Drug Metab. Dispos. *34*, 75–83.

Ardin, M., Cahais, V., Castells, X., Bouaoun, L., Byrnes, G., Herceg, Z., Zavadil, J., and Olivier, M. (2016). MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. BMC Bioinformatics *17*, 170.

Arlt, V.M., Stiborova, M., vom Brocke, J., Simoes, M.L., Lord, G.M., Nortier, J.L., Hollstein, M., Phillips, D.H., and Schmeiser, H.H. (2007). Aristolochic acid mutagenesis: molecular clues to the aetiology of Balkan endemic nephropathy-associated urothelial cancer. Carcinogenesis *28*, 2253–2261.

Bagchi, D., Vuchetich, P.J., Bagchi, M., Hassoun, E.A., Tran, M.X., Tang, L., and Stohs, S.J. (1997). Induction of oxidative stress by chronic administration of sodium dichromate [chromium VI] and cadmium chloride [cadmium II] to rats. Free Radic. Biol. Med. *22*, 471–478.

Beland, F.A., Mellick, P.W., Olson, G.R., Mendoza, M.C.B., Marques, M.M., and Doerge, D.R. (2013). Carcinogenicity of acrylamide in B6C3F(1) mice and F344/N rats from a 2-year drinking water exposure. Food Chem. Toxicol. Int. J. Publ. Br. Ind. Biol. Res. Assoc. *51*, 149–159.

Beland, F.A., Olson, G.R., Mendoza, M.C.B., Marques, M.M., and Doerge, D.R. (2015). Carcinogenicity of glycidamide in B6C3F1 mice and F344/N rats from a two-year drinking water exposure. Food Chem. Toxicol. *86*, 104–115.

Bellver Soto, J., Fernández-Franzón, M., Ruiz, M.-J., and Juan-García, A. (2014). Presence of Ochratoxin A (OTA) Mycotoxin in Alcoholic Drinks from Southern European Countries: Wine and Beer. J. Agric. Food Chem. *62*, 7643–7651.

Besaratinia, A., and Pfeifer, G.P. (2003). Weak yet distinct mutagenicity of acrylamide in mammalian cells. J. Natl. Cancer Inst. *95*, 889–896.

Besaratinia, A., and Pfeifer, G.P. (2004). Genotoxicity of acrylamide and glycidamide. J. Natl. Cancer Inst. *96*, 1023–1029.

Besaratinia, A., and Pfeifer, G.P. (2005). DNA adduction and mutagenic properties of acrylamide. Mutat. Res. *580*, 31–40.

Besaratinia, A., and Pfeifer, G.P. (2010). Applications of the human p53 knock-in (Hupki) mouse model for human carcinogen testing. FASEB J. *24*, 2612–2619.

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., et al. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. Nature.

Boverhof, D.R., Chamberlain, M.P., Elcombe, C.R., Gonzalez, F.J., Heflich, R.H., Hernández, L.G., Jacobs, A.C., Jacobson-Kram, D., Luijten, M., Maggi, A., et al. (2011). Transgenic Animal Models in Toxicology: Historical Perspectives and Future Outlook. Toxicol. Sci. *121*, 207–233.

Brash, D.E. (2015). UV signature mutations. Photochem. Photobiol. *91*, 15–26.

Brash, D.E., Rudolph, J.A., Simon, J.A., Lin, A., McKenna, G.J., Baden, H.P., Halperin, A.J., and Pontén, J. (1991). A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. Proc. Natl. Acad. Sci. U. S. A. *88*, 10124–10128.

Brennan, P., and Wild, C.P. (2015). Genomics of Cancer and a New Era for Cancer Prevention. PLoS Genet. *11*.

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

Brocke, J. v., Schmeiser, H.H., Reinbold, M., and Hollstein, M. (2006). MEF immortalization to investigate the ins and outs of mutagenesis. Carcinogenesis *27*, 2141–2147.

vom Brocke, J., Schmeiser, H.H., Reinbold, M., and Hollstein, M. (2006). MEF immortalization to investigate the ins and outs of mutagenesis. Carcinogenesis *27*, 2141–2147.

vom Brocke, J., Krais, A., Whibley, C., Hollstein, M.C., and Schmeiser, H.H. (2008). The carcinogenic air pollutant 3-nitrobenzanthrone induces GC to TA transversion mutations in human p53 sequences. Mutagenesis *24*, 17–23.

Brown, A.L., Odell, E.W., and Mantle, P.G. (2007). DNA ploidy distribution in renal tumours induced in male rats by dietary ochratoxin A. Exp. Toxicol. Pathol. Off. J. Ges. Toxikol. Pathol. *59*, 85–95.

Brunet, J.-P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. U. S. A. *101*, 4164–4169.

Burns, M.B., Temiz, N.A., and Harris, R.S. (2013). Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat. Genet. *45*, 977–983.

Busuttil, R.A., Rubio, M., Dollé, M.E.T., Campisi, J., and Vijg, J. (2003). Oxygen accelerates the accumulation of mutations during the senescence and immortalization of murine cells in culture. Aging Cell *2*, 287–294.

Campbell, P.J., Getz, G., Stuart, J.M., Korbel, J.O., Stein, L.D., and Net, -ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (2017). Pan-cancer analysis of whole genomes. bioRxiv 162784.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068.

Cerec, V., Glaise, D., Garnier, D., Morosan, S., Turlin, B., Drenou, B., Gripon, P., Kremsdorf, D., Guguen-Guillouzo, C., and Corlu, A. (2007). Transdifferentiation of hepatocyte-like cells from the human hepatoma HepaRG cell line through bipotent progenitor. Hepatology *45*, 957–967.

Chan, K., Roberts, S.A., Klimczak, L.J., Sterling, J.F., Saini, N., Malc, E.P., Kim, J., Kwiatkowski, D.J., Fargo, D.C., Mieczkowski, P.A., et al. (2015). An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nat. Genet. *47*, 1067–1072.

Chen, H., Li, Y., and Tollefsbol, T.O. (2013). Cell Senescence Culturing Methods. In Biological Aging, T.O. Tollefsbol, ed. (Totowa, NJ: Humana Press), pp. 1–10.

Chen, L., Liu, P., Evans, T.C., and Ettwiller, L.M. (2017). DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science *355*, 752–756.

Cheng, L., Sonntag, D.M., de Boer, J., and Dixon, K. (2000). Chromium(VI)-induced mutagenesis in the lungs of big blue transgenic mice. J. Environ. Pathol. Toxicol. Oncol. Off. Organ Int. Soc. Environ. Toxicol. Cancer *19*, 239–249.

Clark, H.A., and Snedeker, S.M. (2006). Ochratoxin a: its cancer risk and potential for exposure. J. Toxicol. Environ. Health B Crit. Rev. *9*, 265–296.

Colditz, G.A., Wolin, K.Y., and Gehlert, S. (2012). Applying what we know to accelerate cancer prevention. Sci. Transl. Med. *4*, 127rv4.

Costa, J.G., Saraiva, N., Guerreiro, P.S., Louro, H., Silva, M.J., Miranda, J.P., Castro, M., Batinic-Haberle, I., Fernandes, A.S., and Oliveira, N.G. (2016). Ochratoxin A-induced cytotoxicity, genotoxicity and reactive oxygen species in kidney cells: An integrative approach of complementary endpoints. Food Chem. Toxicol. Int. J. Publ. Br. Ind. Biol. Res. Assoc. *87*, 65–76.

Cosyns, J.P., Jadoul, M., Squifflet, J.P., De Plaen, J.F., Ferluga, D., and van Ypersele de Strihou, C. (1994). Chinese herbs nephropathy: a clue to Balkan endemic nephropathy? Kidney Int. *45*, 1680–1688.

Denissenko, M.F., Pao, A., Tang, M., and Pfeifer, G.P. (1996). Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. Science *274*, 430–432.

Doerge, D.R., Gamboa da Costa, G., McDaniel, L.P., Churchwell, M.I., Twaddle, N.C., and Beland, F.A. (2005). DNA adducts derived from administration of acrylamide and glycidamide to mice and rats. Mutat. Res. Toxicol. Environ. Mutagen. *580*, 131–141.

Einaga, N., Yoshida, A., Noda, H., Suemitsu, M., Nakayama, Y., Sakurada, A., Kawaji, Y., Yamaguchi, H., Sasaki, Y., Tokino, T., et al. (2017). Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation. PloS One *12*, e0176280.

Espejel, S., and Blasco, M.A. (2002). Identification of telomere-dependent “senescence-like” arrest in mouse embryonic fibroblasts. Exp. Cell Res. *276*, 242–248.

Fantini, D., Glaser, A.P., Rimar, K.J., Wang, Y., Schipma, M., Varghese, N., Rademaker, A., Behdad, A., Yellapa, A., Yu, Y., et al. (2018). A Carcinogen-induced mouse model recapitulates the molecular alterations of human muscle invasive bladder cancer. Oncogene 1.

Faucet, V., Pfohl-Leszkowicz, A., Dai, J., Castegnaro, M., and Manderville, R.A. (2004). Evidence for covalent DNA adduction by ochratoxin A following chronic exposure to rat and subacute exposure to pig. Chem. Res. Toxicol. *17*, 1289–1296.

Feil, R., and Fraga, M.F. (2012). Epigenetics and the environment: emerging patterns and implications. Nat. Rev. Genet. *13*, 97–109.

Feinberg, A.P., Koldobskiy, M.A., and Göndör, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nat. Rev. Genet. *17*, 284–299.

Feldmeyer, N., Schmeiser, H.H., Muehlbauer, K.-R., Belharazem, D., Knyazev, Y., Nedelko, T., and Hollstein, M. (2006). Further studies with a cell immortalization assay to investigate the mutation signature of aristolochic acid in human p53 sequences. Mutat. Res. *608*, 163–168.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int. J. Cancer *136*, E359-386.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*, D945–D950.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45*, D777–D783.

Frenkel, K. (1992). Carcinogen-mediated oxidant formation and oxidative DNA damage. Pharmacol. Ther. *53*, 127–166.

Fukujin, H., Fujita, T., and Mine, T. (2000). Additivity of the proliferative effects of HGF/SF and EGF on hepatocytes. Biochem. Biophys. Res. Commun. *278*, 698–703.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177–183.

Gamboa da Costa, G., Churchwell, M.I., Hamilton, L.P., Von Tungeln, L.S., Beland, F.A., Marques, M.M., and Doerge, D.R. (2003). DNA adduct formation from acrylamide via conversion to glycidamide in adult and neonatal mice. Chem. Res. Toxicol. *16*, 1328–1337.

Gekle, M., and Silbernagl, S. (1993). Mechanism of ochratoxin A-induced reduction of glomerular filtration rate in rats. J. Pharmacol. Exp. Ther. *267*, 316–321.

Ghanayem, B.I., McDaniel, L.P., Churchwell, M.I., Twaddle, N.C., Snyder, R., Fennell, T.R., and Doerge, D.R. (2005). Role of CYP2E1 in the epoxidation of acrylamide to glycidamide and formation of DNA and hemoglobin adducts. Toxicol. Sci. Off. J. Soc. Toxicol. *88*, 311–318.

Giromini, C., Rebucci, R., Fusi, E., Rossi, L., Saccone, F., and Baldi, A. (2016). Cytotoxicity, apoptosis, DNA damage and methylation in mammary and kidney epithelial cell lines exposed to ochratoxin A. Cell Biol. Toxicol. *32*, 249–258.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158.

Gripon, P., Rumin, S., Urban, S., Le Seyec, J., Glaise, D., Cannie, I., Guyomard, C., Lucas, J., Trepo, C., and Guguen-Guillouzo, C. (2002). Infection of a human hepatoma cell line by hepatitis B virus. Proc. Natl. Acad. Sci. U. S. A. *99*, 15655–15660.

Grollman, A.P., Shibutani, S., Moriya, M., Miller, F., Wu, L., Moll, U., Suzuki, N., Fernandes, A., Rosenquist, T., Medverec, Z., et al. (2007). Aristolochic acid and the etiology of endemic (Balkan) nephropathy. Proc. Natl. Acad. Sci. U. S. A. *104*, 12129–12134.

Guillouzo, A., Corlu, A., Aninat, C., Glaise, D., Morel, F., and Guguen-Guillouzo, C. (2007). The human hepatoma HepaRG cells: A highly differentiated model for studies of liver metabolism and toxicity of xenobiotics. Chem. Biol. Interact. *168*, 66–73.

Hahn, W.C., and Weinberg, R.A. (2002). Modelling the molecular circuitry of cancer. Nat. Rev. Cancer *2*, 331–341.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Harris, C.C., Reddel, R., Pfeifer, A., Iman, D., McMenamin, M., Trump, B.F., and Weston, A. (1991). Role of oncogenes and tumour suppressor genes in human lung carcinogenesis. IARC Sci. Publ. 294–304.

Harris, R.S., Petersen-Mahrt, S.K., and Neuberger, M.S. (2002). RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. Mol. Cell *10*, 1247–1253.

He, L., Diedrich, J., Chu, Y.-Y., and Yates, J.R. (2015). Extracting Accurate Precursor Information for Tandem Mass Spectra by RawConverter. Anal. Chem. *87*, 11361–11367.

Hedegaard, J., Thorsen, K., Lund, M.K., Hein, A.-M.K., Hamilton-Dutoit, S.J., Vang, S., Nordentoft, I., Birkenkamp-Demtröder, K., Kruhøffer, M., Hager, H., et al. (2014). 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 *9*, e98187.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. Nat. Rev. Genet. *15*, 585–598.

Herceg, Z., Ghantous, A., Wild, C.P., Sklias, A., Casati, L., Duthie, S.J., Fry, R., Issa, J.-P., Kellermayer, R., Koturbash, I., et al. Roadmap for Investigating Epigenome Deregulation and Environmental Origins of Cancer. Int. J. Cancer n/a-n/a.

Heyndrickx, E., Sioen, I., Huybrechts, B., Callebaut, A., De Henauw, S., and De Saeger, S. (2015). Human biomonitoring of multiple mycotoxins in the Belgian population: Results of the BIOMYCO study. Environ. Int. *84*, 82–89.

Hoang, M.L., Chen, C.-H., Sidorenko, V.S., He, J., Dickman, K.G., Yun, B.H., Moriya, M., Niknafs, N., Douville, C., Karchin, R., et al. (2013). Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. Sci. Transl. Med. *5*, 197ra102.

Hoang, M.L., Chen, C.-H., Chen, P.-C., Roberts, N.J., Dickman, K.G., Yun, B.H., Turesky, R.J., Pu, Y.-S., Vogelstein, B., Papadopoulos, N., et al. (2016). Aristolochic Acid in the Etiology of Renal Cell Carcinoma. Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. *25*, 1600–1608.

Hogervorst, J.G., Schouten, L.J., Konings, E.J., Goldbohm, R.A., and van den Brandt, P.A. (2008). Dietary acrylamide intake and the risk of renal cell, bladder, and prostate cancer. Am. J. Clin. Nutr. *87*, 1428–1438.

Hollstein, M., Sidransky, D., Vogelstein, B., and Harris, C.C. (1991). p53 mutations in human cancers. Science *253*, 49–53.

Hollstein, M., Moriya, M., Grollman, A.P., and Olivier, M. (2013). Analysis of TP53 mutation spectra reveals the fingerprint of the potent environmental carcinogen, aristolochic acid. Mutat. Res. Mutat. Res. *753*, 41–49.

Hollstein, M., Alexandrov, L.B., Wild, C.P., Ardin, M., and Zavadil, J. (2017). Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. Oncogene *36*, 158–167.

Holmes, A.L., Wise, S.S., and Wise, J.P. (2008). Carcinogenicity of hexavalent chromium. Indian J. Med. Res. *128*, 353–372.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115.

Huang, M.N., Yu, W., Teoh, W.W., Ardin, M., Jusakul, A., Ng, A., Boot, A., Abedi-Ardekani, B., Villar, S., Myint, S.S., et al. (2017). Genome-Scale Mutational Signatures Of Aflatoxin In Cells, Mice And Human Tumors. bioRxiv 130179.

Hussain, T., and Mulherkar, R. (2012). Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair. Int. J. Mol. Cell. Med. *1*, 75–87.

Inga, A., Scott, G., Monti, P., Aprile, A., Abbondandolo, A., Burns, P.A., and Fronza, G. (1998). Ultraviolet-light induced p53 mutational spectrum in yeast is indistinguishable from p53 mutations in human skin cancer. Carcinogenesis *19*, 741–746.

Ishii, Y., Matsushita, K., Kuroda, K., Yokoo, Y., Kijima, A., Takasu, S., Kodama, Y., Nishikawa, A., and Umemura, T. (2015). Acrylamide induces specific DNA adduct formation and gene mutations in a carcinogenic target site, the mouse lung. Mutagenesis *30*, 227–235.

Ishikawa, Y., Nakagawa, K., Satoh, Y., Kitagawa, T., Sugano, H., Hirano, T., and Tsuchiya, E. (1994). "Hot spots" of chromium accumulation at bifurcations of chromate workers' bronchi. Cancer Res. *54*, 2342–2346.

Jagger, C., Tate, M., Cahill, P.A., Hughes, C., Knight, A.W., Billinton, N., and Walmsley, R.M. (2009). Assessment of the genotoxicity of S9-generated metabolites using the GreenScreen HC GADD45a-GFP assay. Mutagenesis *24*, 35–50.

Jego, N., Thomas, G., and Hamelin, R. (1993). Short direct repeats flanking deletions, and duplicating insertions in p53 gene in human cancers. Oncogene *8*, 209–213.

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

Jia, X., Cui, J., Meng, X., Xing, L., Shen, H., Wang, J., Liu, J., Wang, Y., Lian, W., and Zhang, X. (2016). Malignant transformation of human gastric epithelium cells via reactive oxygen species production and Wnt/β-catenin pathway activation following 40-week exposure to ochratoxin A. Cancer Lett. *372*, 36–47.

Kazanov, M.D., Roberts, S.A., Polak, P., Stamatoyannopoulos, J., Klimczak, L.J., Gordenin, D.A., and Sunyaev, S.R. (2015). APOBEC-induced cancer mutations are uniquely enriched in early replicating, gene dense, and active chromatin regions. Cell Rep. *13*, 1103–1109.

Klein, C.B., Su, L., Bowser, D., and Leszczynska, J. (2002). Chromate-induced epimutations in mammalian cells. Environ. Health Perspect. *110 Suppl 5*, 739–743.

Kolarić, K. (1977). Combination chemotherapy with 1-methyl-1-nitrosourea (MNU) and cyclophosphamide in solid tumors. Z. Für Krebsforsch. Klin. Onkol. *89*, 311–319.

Krishnapura, P.R., Belur, P.D., and Subramanya, S. (2016). A critical review on properties and applications of microbial l-asparaginases. Crit. Rev. Microbiol. *42*, 720–737.

Krogh, P., Hald, B., Plestina, R., and Ceović, S. (1977). Balkan (endemic) nephropathy and foodborn ochratoxin A: preliminary results of a survey of foodstuffs. Acta Pathol. Microbiol. Scand. [B] *85*, 238–240.

Kucab, J.E., Phillips, D.H., and Arlt, V.M. (2010). Linking environmental carcinogen exposure to TP53 mutations in human tumours using the human TP53 knock-in (Hupki) mouse model. FEBS J. *277*, 2567–2583.

Kuiper-Goodman, T., and Scott, P.M. (1989). Risk assessment of the mycotoxin ochratoxin A. Biomed. Environ. Sci. BES *2*, 179–248.

Lee, T.P., Saad, B., Ng, E.P., and Salleh, B. (2012). Zeolite Linde Type L as micro-solid phase extraction sorbent for the high performance liquid chromatography determination of ochratoxin A in coffee and cereal. J. Chromatogr. A *1237*, 46–54.

Lijinsky, W., Garcia, H., Keefer, L., Loo, J., and Ross, A.E. (1972). Carcinogenesis and alkylation of rat liver nucleic acids by nitrosomethylurea and nitrosoethylurea administered by intraportal injection. Cancer Res. *32*, 893–897.

Liu, Z., Hergenhahn, M., Schmeiser, H.H., Wogan, G.N., Hong, A., and Hollstein, M. (2004). 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.

Liu, Z., Muehlbauer, K.-R., Schmeiser, H.H., Hergenhahn, M., Belharazem, D., and Hollstein, M.C. (2005). 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.

López-Lázaro, M. Comment on'Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention'.

Lübberstedt, M., Müller-Vieira, U., Mayer, M., Biemel, K.M., Knöspel, F., Knobeloch, D., Nüssler, A.K., Gerlach, J.C., and Zeilinger, K. (2011). HepaRG human hepatic cell line utility as a surrogate for primary human hepatocytes in drug metabolism assessment in vitro. J. Pharmacol. Toxicol. Methods *63*, 59–68.

Luo, J.L., Yang, Q., Tong, W.M., Hergenhahn, M., Wang, Z.Q., and Hollstein, M. (2001a). 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.

Luo, J.-L., Tong, W.-M., Yoon, J.-H., Hergenhahn, M., Koomagi, R., Yang, Q., Galendo, D., Pfeifer, G.P., Wang, Z.-Q., and Hollstein, M. (2001b). UV-induced DNA damage and mutations in Hupki (human p53 knock-in) mice recapitulate p53 hotspot alterations in sun-exposed human skin. Cancer Res. *61*, 8158–8163.

Mally, A. (2012). Ochratoxin a and mitotic disruption: mode of action analysis of renal tumor formation by ochratoxin A. Toxicol. Sci. Off. J. Soc. Toxicol. *127*, 315–330.

Mally, A., Pepe, G., Ravoori, S., Fiore, M., Gupta, R.C., Dekant, W., and Mosesso, P. (2005). Ochratoxin a causes DNA damage and cytogenetic effects but no DNA adducts in rats. Chem. Res. Toxicol. *18*, 1253–1261.

Manderville, R.A. (2005). A case for the genotoxicity of ochratoxin A by bioactivation and covalent DNA adduction. Chem. Res. Toxicol. *18*, 1091–1097.

Manjanatha, M.G., Guo, L.-W., Shelton, S.D., and Doerge, D.R. (2015a). Acrylamide-induced carcinogenicity in mouse lung involves mutagenicity: *cII* gene mutations in the lung of big blue mice exposed to acrylamide and glycidamide for up to 4 weeks: AA-Induced Carcinogenicity in Mouse Involves Mutagenicity. Environ. Mol. Mutagen. *56*, 446–456.

Manjanatha, M.G., Guo, L.-W., Shelton, S.D., and Doerge, D.R. (2015b). Acrylamide-induced carcinogenicity in mouse lung involves mutagenicity: *cII* gene mutations in the lung of big blue mice exposed to acrylamide and glycidamide for up to 4 weeks: AA-Induced Carcinogenicity in Mouse Involves Mutagenicity. Environ. Mol. Mutagen. *56*, 446–456.

Mantle, P.G., Faucet-Marquis, V., Manderville, R.A., Squillaci, B., and Pfohl-Leszkowicz, A. (2010). Structures of covalent adducts between DNA and ochratoxin a: a new factor in debate about genotoxicity and human risk assessment. Chem. Res. Toxicol. *23*, 89–98.

Miller, J.H. (1982). Carcinogens induce targeted mutations in Escherichia coli. Cell *31*, 5–7.

Mojska, H., Gielecińska, I., and Cendrowski, A. (2016). Acrylamide content in cigarette mainstream smoke and estimation of exposure to acrylamide from tobacco smoke in Poland. Ann. Agric. Environ. Med. AAEM *23*, 456–461.

Montesano, R., Hainaut, P., and Wild, C.P. (1997). Hepatocellular Carcinoma: From Gene to Public Health. JNCI J. Natl. Cancer Inst. *89*, 1844–1851.

National Toxicology Program (1989). Toxicology and Carcinogenesis Studies of Ochratoxin A (CAS No. 303-47-9) in F344/N Rats (Gavage Studies). Natl. Toxicol. Program Tech. Rep. Ser. *358*, 1–142.

National Toxicology Program (2008). Toxicology and carcinogenesis studies of sodium dichromate dihydrate (Cas No. 7789-12-0) in F344/N rats and B6C3F1 mice (drinking water studies). Natl. Toxicol. Program Tech. Rep. Ser. 1–192.

Nedelko, T., Arlt, V.M., Phillips, D.H., and Hollstein, M. (2009). TP53 mutation signature supports involvement of aristolochic acid in the aetiology of endemic nephropathy-associated tumours. Int. J. Cancer *124*, 987–990.

Neitzel, H. (1986). A routine method for the establishment of permanent growing lymphoblastoid cell lines. Hum. Genet. *73*, 320–326.

Nickens, K.P., Patierno, S.R., and Ceryak, S. (2010). Chromium genotoxicity: A double-edged sword. Chem. Biol. Interact. *188*, 276–288.

Nigam, A., Priya, S., Bajpai, P., and Kumar, S. (2014). Cytogenomics of hexavalent chromium (Cr6+) exposed cells: A comprehensive review. Indian J. Med. Res. *139*, 349.

Nik-Zainal, S., Kucab, J.E., Morganella, S., Glodzik, D., Alexandrov, L.B., Arlt, V.M., Weninger, A., Hollstein, M., Stratton, M.R., and Phillips, D.H. (2015). The genome as a record of environmental exposure. Mutagenesis gev073.

Nowell, P.C. (2002). Tumor progression: a brief historical perspective. Semin. Cancer Biol. *12*, 261–266.

Nowell, P., Hungerford, D., and Nowell, P.C. (1960). A minute chromosome in human chronic granulocytic leukemia.

Obón-Santacana, M., Freisling, H., Peeters, P.H., Lujan-Barroso, L., Ferrari, P., Boutron-Ruault, M.-C., Mesrine, S., Baglietto, L., Turzanski-Fortner, R., Katzke, V.A., et al. (2016a). Acrylamide and glycidamide hemoglobin adduct levels and endometrial cancer risk: A nested case-control study in nonsmoking postmenopausal women from the EPIC cohort. Int. J. Cancer *138*, 1129–1138.

Obón-Santacana, M., Lujan-Barroso, L., Travis, R.C., Freisling, H., Ferrari, P., Severi, G., Baglietto, L., Boutron-Ruault, M.-C., Fortner, R.T., Ose, J., et al. (2016b). Acrylamide and Glycidamide Hemoglobin Adducts and Epithelial Ovarian Cancer: A Nested Case-Control Study in Nonsmoking Postmenopausal Women from the EPIC Cohort. Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. *25*, 127–134.

Obón-Santacana, M., Lujan-Barroso, L., Freisling, H., Cadeau, C., Fagherazzi, G., Boutron-Ruault, M.-C., Kaaks, R., Fortner, R.T., Boeing, H., Ramón Quirós, J., et al. (2016c). Dietary and lifestyle determinants of acrylamide and glycidamide hemoglobin adducts in non-smoking postmenopausal women from the EPIC cohort. Eur. J. Nutr.

O'Brien, T.J., Ceryak, S., and Patierno, S.R. (2003). Complexities of chromium carcinogenesis: role of cellular response, repair and recovery mechanisms. Mutat. Res. *533*, 3–36.

OCHRATOXIN, A. HHH 0.

Olesen, P.T., Olsen, A., Frandsen, H., Frederiksen, K., Overvad, K., and Tjønneland, A. (2008). Acrylamide exposure and incidence of breast cancer among postmenopausal women in the Danish Diet, Cancer and Health Study. Int. J. Cancer *122*, 2094–2100.

Olivier, M., Hollstein, M., and Hainaut, P. (2010). TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. Cold Spring Harb. Perspect. Biol. *2*.

Olivier, M., Weninger, A., Ardin, M., Huskova, H., Castells, X., Vallée, M.P., McKay, J., Nedelko, T., Muehlbauer, K.-R., Marusawa, H., et al. (2014). Modelling mutational landscapes of human cancers in vitro. Sci. Rep. *4*.

Paget, V., Lechevrel, M., André, V., Le Goff, J., Pottier, D., Billet, S., Garçon, G., Shirali, P., and Sichel, F. (2012). Benzo[a]pyrene, Aflatoxine B1 and Acetaldehyde Mutational Patterns in TP53 Gene Using a Functional Assay: Relevance to Human Cancer Aetiology. PLoS ONE *7*, e30921.

Parrinello, S., Samper, E., Krtolica, A., Goldstein, J., Melov, S., and Campisi, J. (2003). Oxygen sensitivity severely limits the replicative lifespan of murine fibroblasts. Nat. Cell Biol. *5*, 741–747.

Patlolla, A.K., Barnes, C., Hackett, D., and Tchounwou, P.B. (2009). Potassium dichromate induced cytotoxicity, genotoxicity and oxidative stress in human liver carcinoma (HepG2) cells. Int. J. Environ. Res. Public. Health *6*, 643–653.

Pfeifer, G.P., Denissenko, M.F., Olivier, M., Tretyakova, N., Hecht, S.S., and Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p 53 mutations in smoking-associated cancers. Oncogene *21*, 7435–7451.

Pfohl-Leszkowicz, A., and Manderville, R.A. (2007). Ochratoxin A: An overview on toxicity and carcinogenicity in animals and humans. Mol. Nutr. Food Res. *51*, 61–99.

Pfohl-Leszkowicz, A., Chakor, K., Creppy, E.E., and Dirheimer, G. (1991). DNA adduct formation in mice treated with ochratoxin A. IARC Sci. Publ. 245–253.

Pilati, C., Shinde, J., Alexandrov, L.B., Assié, G., André, T., Hélias-Rodzewicz, Z., Ducoudray, R., Le Corre, D., Zucman-Rossi, J., Emile, J.-F., et al. (2017). Mutational signature analysis identifies *MUTYH* deficiency in colorectal cancers and adrenocortical carcinomas: Mutational signature associated with *MUTYH* deficiency in cancers. J. Pathol. *242*, 10–15.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.-L., Ordóñez, G.R., Bignell, G.R., et al. (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191–196.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.-L., Beare, D., Lau, K.W., Greenman, C., et al. (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature *463*, 184–190.

Poon, S.L., Pang, S.-T., McPherson, J.R., Yu, W., Huang, K.K., Guan, P., Weng, W.-H., Siew, E.Y., Liu, Y., Heng, H.L., et al. (2013). Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. Sci. Transl. Med. *5*, 197ra101–197ra101.

Poon, S.L., Huang, M.N., Choo, Y., McPherson, J.R., Yu, W., Heng, H.L., Gan, A., Myint, S.S., Siew, E.Y., Ler, L.D., et al. (2015). Mutation signatures implicate aristolochic acid in bladder cancer development. Genome Med. *7*, 38.

Pratheeshkumar, P., Son, Y.-O., Divya, S.P., Turcios, L., Roy, R.V., Hitron, J.A., Wang, L., Kim, D., Dai, J., Asha, P., et al. (2016). Hexavalent chromium induces malignant transformation of human lung bronchial epithelial cells via ROS-dependent activation of miR-21-PDCD4 signaling. Oncotarget *7*, 51193–51210.

Qi, X., Yu, T., Zhu, L., Gao, J., He, X., Huang, K., Luo, Y., and Xu, W. (2014). Ochratoxin A induces rat renal carcinogenicity with limited induction of oxidative stress responses. Toxicol. Appl. Pharmacol. *280*, 543–549.

Radić, B., Fuchs, R., Peraica, M., and Lucić, A. (1997). Ochratoxin A in human sera in the area with endemic nephropathy in Croatia. Toxicol. Lett. *91*, 105–109.

Randall, S.K., Eritja, R., Kaplan, B.E., Petruska, J., and Goodman, M.F. (1987). Nucleotide insertion kinetics opposite abasic lesions in DNA. J. Biol. Chem. *262*, 6864–6870.

Rebhandl, S., Huemer, M., Greil, R., and Geisberger, R. (2015). AID/APOBEC deaminases and cancer. Oncoscience *2*, 320–333.

Reddy, E.P., Reynolds, R.K., Santos, E., and Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. Nature *300*, 149–152.

Reinbold, M., Luo, J.L., Nedelko, T., Jerchow, B., Murphy, M.E., Whibley, C., Wei, Q., and Hollstein, M. (2008). Common tumour p53 mutations in immortalized cells from Hupki mice heterozygous at codon 72. Oncogene *27*, 2788–2794.

Rowley, J.D. (1973). Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature *243*, 290–293.

Salnikow, K., and Zhitkovich, A. (2008). Genetic and Epigenetic Mechanisms in Metal Carcinogenesis and Cocarcinogenesis: Nickel, Arsenic and Chromium. Chem. Res. Toxicol. *21*, 28–44.

Savary, C.C., Jiang, X., Aubry, M., Jossé, R., Kopp-Schneider, A., Hewitt, P., and Guillouzo, A. (2015). Transcriptomic analysis of untreated and drug-treated differentiated HepaRG cells over a 2-week period. Toxicol. In Vitro.

Schafer, C.M., Campbell, N.G., Cai, G., Yu, F., Makarov, V., Yoon, S., Daly, M.J., Gibbs, R.A., Schellenberg, G.D., Devlin, B., et al. (2013). Whole exome sequencing reveals minimal differences between cell line and whole blood derived DNA. Genomics *102*, 270–277.

Schnekenburger, M., Talaska, G., and Puga, A. (2007). Chromium cross-links histone deacetylase 1-DNA methyltransferase 1 complexes to chromatin, inhibiting histone-remodeling marks critical for transcriptional activation. Mol. Cell. Biol. *27*, 7089–7101.

Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L.B., Calderaro, J., Rebouissou, S., Couchy, G., Meiller, C., Shinde, J., Soysouvanh, F., et al. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat. Genet. *47*, 505–511.

Schwerdt, G., Freudinger, R., Mildenberger, S., Silbernagl, S., and Gekle, M. (1999). The nephrotoxin ochratoxin A induces apoptosis in cultured human proximal tubule cells. Cell Biol. Toxicol. *15*, 405–415.

Secrier, M., Li, X., de Silva, N., Eldridge, M.D., Contino, G., Bornschein, J., MacRae, S., Grehan, N., O'Donovan, M., Miremadi, A., et al. (2016). Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. Nat. Genet. *48*, 1131–1141.

Segerbäck, D., Calleman, C.J., Schroeder, J.L., Costa, L.G., and Faustman, E.M. (1995). Formation of N-7-(2-carbamoyl-2-hydroxyethyl) guanine in DNA of the mouse and the rat following intraperitoneal administration of [14C] acrylamide. Carcinogenesis *16*, 1161–1165.

Severson, P.L., Vrba, L., Stampfer, M.R., and Futscher, B.W. (2014). Exome-wide mutation profile in benzo[a]pyrene-derived post-stasis and immortal human mammary epithelial cells. Mutat. Res. Toxicol. Environ. Mutagen. *775–776*, 48–54.

Sheu, M.-L., Shen, C.-C., Chen, Y.-S., and Chiang, C.-K. (2017). Ochratoxin A induces ER stress and apoptosis in mesangial cells via a NADPH oxidase-derived reactive oxygen species-mediated calpain activation pathway. Oncotarget *8*, 19376–19388.

Smith, H.C., Bennett, R.P., Kizilyer, A., McDougall, W.M., and Prohaska, K.M. (2012). Functions and regulation of the APOBEC family of proteins. Semin. Cell Dev. Biol. *23*, 258–268.

Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., et al. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature *448*, 561–566.

Stemmermann, G.N., Nomura, A.M., Chyou, P.H., Kato, I., and Kuroishi, T. (1991). Cancer incidence in Hawaiian Japanese: migrants from Okinawa compared with those from other prefectures. Jpn. J. Cancer Res. Gann *82*, 1366–1370.

Straif, K., Loomis, D., Guyton, K., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., and Mattock, H. (2014). Future priorities for the IARC Monographs. Lancet Oncol. *15*, 683–684.

Stratton, M.R. (2013). Journeys into the genome of cancer cells. EMBO Mol. Med. *5*, 169–172.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Sugiyama, A., Maruta, F., Ikeno, T., Ishida, K., Kawasaki, S., Katsuyama, T., Shimizu, N., and Tatematsu, M. (1998). Helicobacter pylori infection enhances N-methyl-N-nitrosourea-induced stomach carcinogenesis in the Mongolian gerbil. Cancer Res. *58*, 2067–2069.

Sumner, S.C., Fennell, T.R., Moore, T.A., Chanas, B., Gonzalez, F., and Ghanayem, B.I. (1999). Role of cytochrome P450 2E1 in the metabolism of acrylamide and acrylonitrile in mice. Chem. Res. Toxicol. *12*, 1110–1116.

Sun, H., Zhou, X., Chen, H., Li, Q., and Costa, M. (2009). Modulation of histone methylation and MLH1 gene silencing by hexavalent chromium. Toxicol. Appl. Pharmacol. *237*, 258–266.

Tabin, C.J., Bradley, S.M., Bargmann, C.I., Weinberg, R.A., Papageorge, A.G., Scolnick, E.M., Dhar, R., Lowy, D.R., and Chang, E.H. (1982). Mechanism of activation of a human oncogene. Nature *300*, 143–149.

Takatsuki, S., Nemoto, S., Sasaki, K., and Maitani, T. (2003). Determination of acrylamide in processed foods by LC/MS using column switching. Shokuhin Eiseigaku Zasshi J. Food Hyg. Soc. Jpn. *44*, 89–95.

Talbot, S.J., and Crawford, D.H. (2004). Viruses and tumours--an update. Eur. J. Cancer Oxf. Engl. 1990 *40*, 1998–2005.

Tareke, E., Rydberg, P., Karlsson, P., Eriksson, S., and Törnqvist, M. (2002). Analysis of Acrylamide, a Carcinogen Formed in Heated Foodstuffs. J. Agric. Food Chem. *50*, 4998–5006.

Todaro, G.J., and Green, H. (1963). QUANTITATIVE STUDIES OF THE GROWTH OF MOUSE EMBRYO CELLS IN CULTURE AND THEIR DEVELOPMENT INTO ESTABLISHED LINES. J. Cell Biol. *17*, 299–313.

Tomasetti, C., and Vogelstein, B. (2015). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science *347*, 78–81.

Tomasetti, C., Durrett, R., Kimmel, M., Lambert, A., Parmigiani, G., Zauber, A., and Vogelstein, B. (2017a). Role of stem-cell divisions in cancer risk. Nature *548*, E13–E14.

Tomasetti, C., Li, L., and Vogelstein, B. (2017b). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science *355*, 1330–1334.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science *310*, 644–648.

Turesky, R.J. (2005). Perspective: ochratoxin A is not a genotoxic carcinogen. Chem. Res. Toxicol. *18*, 1082–1090.

Viel, A., Bruselles, A., Meccia, E., Fornasarig, M., Quaia, M., Canzonieri, V., Policicchio, E., Urso, E.D., Agostini, M., Genuardi, M., et al. (2017). A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. EBioMedicine *20*, 39–49.

Virk-Baker, M.K., Nagy, T.R., Barnes, S., and Groopman, J. (2014). Dietary Acrylamide and Human Cancer: A Systematic Review of Literature. Nutr. Cancer *66*, 774–790.

Vogelstein, B., and Kinzler, K.W. (1993). The multistep nature of cancer. Trends Genet. *9*, 138–141.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Von Tungeln, L.S., Churchwell, M.I., Doerge, D.R., Shaddock, J.G., McGarrity, L.J., Heflich, R.H., Gamboa da Costa, G., Marques, M.M., and Beland, F.A. (2009). DNA adduct formation and induction of micronuclei and mutations in B6C3F1/Tk mice treated neonatally with acrylamide or glycidamide. Int. J. Cancer *124*, 2006–2015.

Von Tungeln, L.S., Doerge, D.R., Gamboa da Costa, G., Matilde Marques, M., Witt, W.M., Koturbash, I., Pogribny, I.P., and Beland, F.A. (2012). Tumorigenicity of acrylamide and its metabolite glycidamide in the neonatal mouse bioassay. Int. J. Cancer *131*, 2008–2015.

Warren, C.J., Xu, T., Guo, K., Griffin, L.M., Westrich, J.A., Lee, D., Lambert, P.F., Santiago, M.L., and Pyeon, D. (2015). APOBEC3A Functions as a Restriction Factor of Human Papillomavirus. J. Virol. *89*, 688–702.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer Analysis Project. Nat. Genet. *45*, 1113–1120.

Westcott, P.M.K., Halliwill, K.D., To, M.D., Rashid, M., Rust, A.G., Keane, T.M., Delrosario, R., Jen, K.-Y., Gurley, K.E., Kemp, C.J., et al. (2015). The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. Nature *517*, 489–492.

Whibley, C., Odell, A.F., Nedelko, T., Balaburski, G., Murphy, M., Liu, Z., Stevens, L., Walker, J.H., Routledge, M., and Hollstein, M. (2010). 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.

Wild, C., Brennan, P., Plummer, M., Bray, F., Straif, K., and Zavadil, J. (2015). Cancer risk: role of chance overstated. Science *347*, 728.

Wild, C.P., Jiang, Y.Z., Allen, S.J., Jansen, L.A., Hall, A.J., and Montesano, R. (1990). Aflatoxin-albumin adducts in human sera from different regions of the world. Carcinogenesis *11*, 2271–2274.

Wilson, K.M., Bälter, K., Adami, H.-O., Grönberg, H., Vikström, A.C., Paulsson, B., Törnqvist, M., and Mucci, L.A. (2009). Acrylamide exposure measured by food frequency questionnaire and hemoglobin adduct levels and prostate cancer risk in the Cancer of the Prostate in Sweden Study. Int. J. Cancer *124*, 2384–2390.

Wogan, G.N. (1992). Aflatoxins as risk factors for hepatocellular carcinoma in humans. Cancer Res. *52*, 2114s–2118s.

Wu, S., Powers, S., Zhu, W., and Hannun, Y.A. (2016). Substantial contribution of extrinsic risk factors to cancer development. Nature *529*, 43–47.

Xie, J., Terry, K.L., Poole, E.M., Wilson, K.M., Rosner, B.A., Willett, W.C., Vesper, H.W., and Tworoger, S.S. (2013). Acrylamide Hemoglobin Adduct Levels and Ovarian Cancer Risk: A Nested Case-Control Study. Cancer Epidemiol. Biomarkers Prev. *22*, 653–660.

Yang, Q., He, X., Li, X., Xu, W., Luo, Y., Yang, X., Wang, Y., Li, Y., and Huang, K. (2014). DNA damage and S phase arrest induced by Ochratoxin A in human embryonic kidney cells (HEK 293). Mutat. Res. *765*, 22–31.

Zhivagui, M., Korenjak, M., and Zavadil, J. (2016). Modelling Mutation Spectra of Human Carcinogens Using Experimental Systems. Basic Clin. Pharmacol. Toxicol. doi: 10.1111/bcpt.12690. Epub 2017 Feb 3

(1978). Some N-Nitroso compounds: this publication represents the views and expert opinions of an IARC working Group on the Evaluation of the Carcinogenetic Risk of chemicals to humans which met in Lyon, 10 - 15 October 1977 (Lyon: IARC).

(1994). Some industrial chemicals: ... views and expert opinions of an IARC Working Group on the Evaluation of Carcinogenesis Risks to Humans, which met in Lyon, 15 - 22 February 1994 (Lyon).

(2010). International network of cancer genome projects. Nature *464*, 993–998.

APPENDICES

Appendix A

Figure A.1: The vicious cycle that feed off cancer patients in the poor populations such as poverty, education, knowledge, evidence, access to care, prevention, early detection and treatment outcome. Adopted from the International Network for Cancer Treatment and Prevention (INCTR).

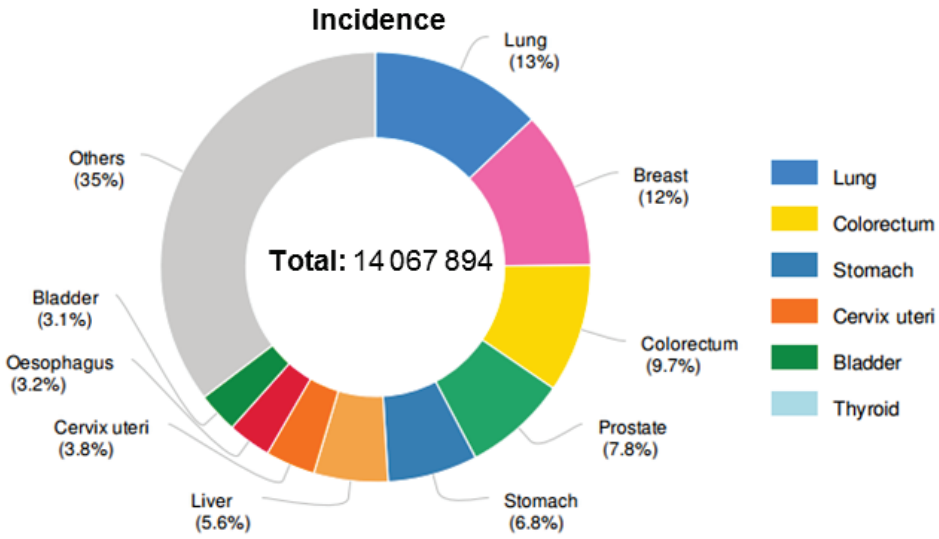
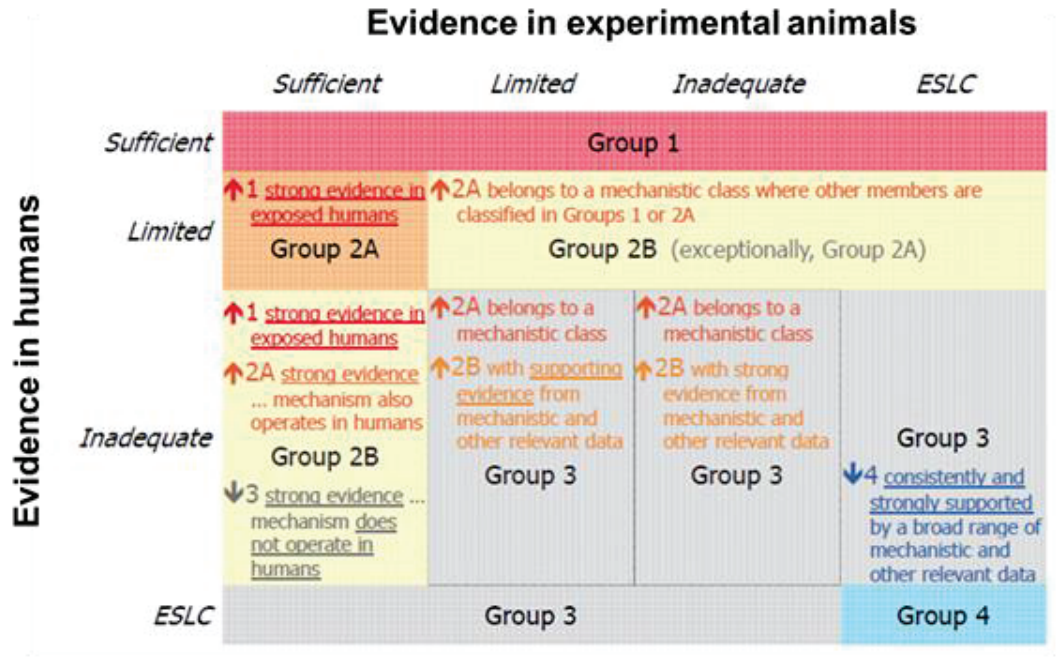


Figure A.2: **Cancer incidence worldwide in 2012**. Distribution of new cancer cases across the different target sites within 5 years. Taken from the Global Cancer Observatory.

Appendix B

Table B.1: Mechanistic data can be pivotal in classification of cancer-risk factors when the human data is inconclusive. Arrows represent the mechanistic data evaluation. Up-arrow means an upgrade of the classification. Down-arrow signifies a downgrade of the compound classification.



Appendix C: DNA adduct analysis protocol

1. DNA extraction for adductomics analysis

The cell cultures were centrifuged at 2500xg for 5 minutes. The supernatants were discarded and the cell pellets were resuspended in 3 mL of Cell Lysis Solution, purchased from Qiagen. The cell membranes were disrupted by keeping the tubes under shaking at room temperature for 24 hours. The RNA was digested by incubating the samples for 2 hours at room temperature with 40 µL of RNase-A (Qiagen). At the end of the digestion the proteins were precipitated by adding 1 mL of Protein Precipitation Solution (Qiagen). Protein pellets were obtained by spinning the tubes down at 4500xg for 3 min. The supernatants were saved and the pellets discarded. The DNA was precipitated from the supernatants by adding 4 mL of cold IPA (100 $\%_{v/v}$, 0 °C). The DNA was then pelleted by spinning down the tubes (14000xg, 4 °C) for 3 minutes. The supernatants were gently discarded. The DNA samples were washed by resuspending them in 1 mL of IPA 70 $\%_{v/v}$, pelleting at 14000xg (3 min at 4 °C) and isolating the DNA by discarding the supernatants. This was repeated, by resuspending the DNA in 1 mL of IPA 100 $\%_{v/v.}$. Once isolated and dried, the DNA was quantified by a UV/Vis spectrophotometer equipped with a NanoDrop cuvette. An example of the DNA spectrum is reported in Figure C.1. The DNA purity was evaluated normalizing the molar extinction measure at 280 and 260 nm wavelength (optimal $\lambda_{260/280}$ 1.8).

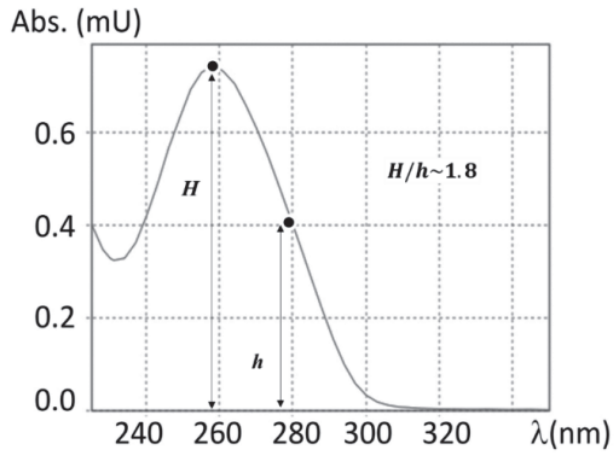


Figure C.1: UV spectrum of double strand DNA.

2. DNA enzymatic digestion

The absolute amount of DNA was preliminarily quantified by dissolving the samples in Tris buffer (Trizima/MgCl2 10 and 5 mM, pH7) and measuring its concentration by the UV/Vis spectrophotometer. The DNA digestion was carried out by using a cocktail of enzymes consisting of DNase (from E.coli, Aldrich), Phospodiesterase-1 (PDE-1) (from Crotalus

adamanteus, Aldrich) and Alkaline Phosphatase (ALP) (from from Pichia Pastoris, Aldrich). All the enzymes were purified by using a double filtration membrane Amico Ultra (0.5 mL, cutoff 10 k Da). The hydrolysis consisted of a two steps process during which a first aliquot of DNase was added to each sample prior to the treatment performed with the full enzyme mixture. Both treatments were carried out by incubating the samples for 24 hours at room temperature. The enzymes concentrations used in these treatments were optimized for digestion of 1 µg of DNA. The first digestion step used 0.5 Units of DNase. The second digestion step required 0.5 U, 0.2 U and 0.02 mU of DNase, ALP and PDE-1 respectively, to bring the DNA digestion to completion. To stop the hydrolysis, the enzymes were removed by using an Amicon Microcone single filtration membrane (0.5 mL, cutoff 10 kDa). The digestion yield was assessed by measuring the concentration of dG via an LC/UV measurement.

3. dG quantitation method

The chromatographic separation of the four 2'-deoxyribonucleosides was carried out using a HPLC Ultimate 3000 equipped with a reversed phase column, Luna C18 (250x0.5 mm, 5 µm, 100 Å). The LC system operated at 40 °C with a flow rate of 15 $\mu L \cdot min^{-1}$ and the separation was performed using a gradient. The A and B mobile phases consisted of H_2O and MeOH. The elution program started with an isocratic step at 5 % B (3 min), followed by a first linear gradient of 0.58 % $B \cdot min^{-1}$ (12 min), a second linear gradient of 27.67 % $B \cdot min^{-1}$ (3 min) and it concluded with a second isocratic step at 95 % B (3 min). Finally, the column was equilibrated (9 min) with a post-time isocratic step of 5 % B. The UV detector operated in absorbance optical mode, monitoring the 254 nm wavelength. The analyte of interest (dG) was quantified using a calibration curve consisting of eight different standard points (0.0625, 0.125, 0.25, 0.50, 1.00, 2.00, 4.0, 8.0 ng/µL pf dG).

Figure C.2 shows the chromatogram of a DNA sample. In this chromatogram dC has been eluted at 8 min, dG at 13 min, dT at 15 min and dA at 18 min.

Figure C.3 shows the calibration curve used to quantify dG in DNA-samples. An eight-point calibration curve was generated by injecting standard solutions with different concentration of dG. The limit of detection (LOD, 0.04 ng/µL) was assessed by spiking decreasing amounts of dG in water and calculating the concentration required to give a s/n-ratio equal to 3. The stability of the method was assessed by injecting the same calibration curve in three different days. The coefficient of variation was found lower than 5%.

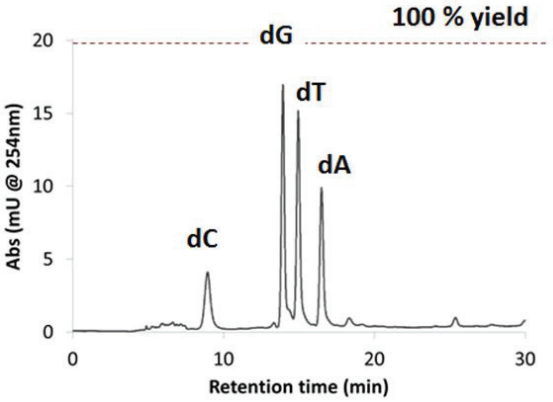


Figure C.2: Representative chromatogram of a DNA-sample enzymatically digested. The analysis was performed with LC/UV system, probing the absorbance of 254 nm wavelength.

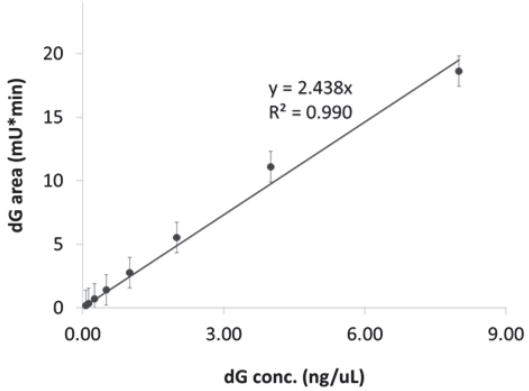


Figure C.3: Representative example of dG calibration curve. Measurements carried out via HPLC/UV.

4. Hydrophobic reversed phase fraction collection

In order to purify the raw reaction media at the end of the enzymatic digestion and enrich the sample with the analyte of interest a fraction collection methodology was used. The purification was carried out on an HPLC (Ultimate 3000, Thermo Scientific, Waltham, MA) equipped with a C18-Column (4.6 x 250 mm, 100Å, 5µm Luna-Phenomenex, Torrace, CA) operating at 25°C, with a flow rate of 1.0 mL·min^{-1}. The A and B mobile phases consisted of H_2O and MeOH. The elution program involved an isocratic step at 2% of B (5 min), followed by a linear gradient of 0.7 $\%_B$·min^{-1} (25 min) and a second isocratic step at 100% of B (15 min). At the end of the elution, the LC-system was equilibrated in isocratic condition (2% of B) for 20 min. The detector operated at 4Hz in absorbance-mode, probing two different wavelengths (λ^1 190 nm and λ^2 254 nm). The collection was optimized using the λ^1 and λ^2 to fractionate properly the gradient on column and to monitor the elution of the deoxyribonucleotides. A representative example of fraction collection is reported in Figure C.4.

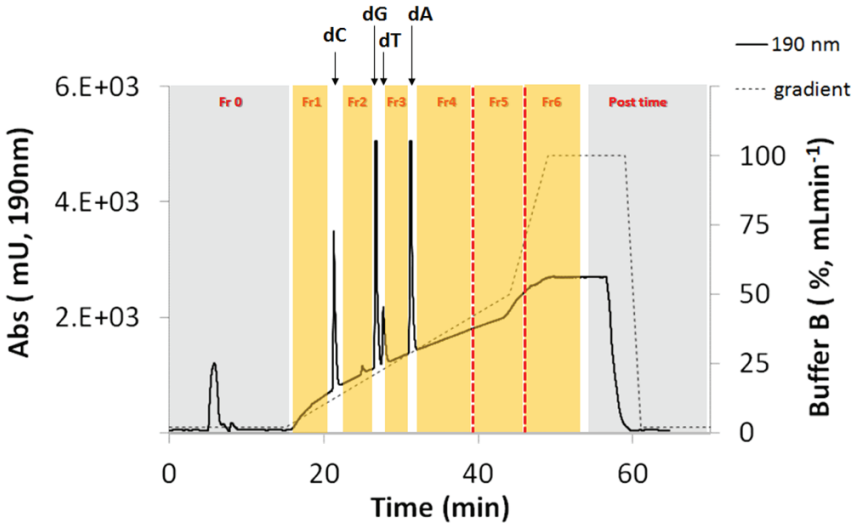


Figure C.4: Fraction collection protocols. Fraction-0 (Fr-0) (0-15 min), Fr-1 (15-19 min), Fr-dC (19-22 min), Fr-2 (22-27.0 min), Fr-dG (27.0-27.5 min), Fr-dT (27.5.-28.5 min), Fr-3 (28.5-32 min), Fr-dA (32-33.5min), Fr-4 (33.5-38.75 min), Fr-5 (38.75 – 45.75 min) and Fr-6 (45.75-53.00). The bolted line chromatogram is recorded at 190nm, the dotted line represents the elution program, and the orange boxes refer to the collected fraction while the gray boxes refer to the discarded fractions.

The sample enrichment and purification protocol was optimized by spiking a mix of DNA-adduct (100 fmol each) in 300uL of Tris-Buffer. The samples obtained from the fractionation were finally analyzed via LC-MS/MS. All the analytes of interested were eluted after 2′-deoxyadenosine.

The protocol resulted in the the isolation of ten different fractions. The collection program is here reported: fraction-0 (Fr-0, 0- 15 min), fraction-1 (Fr-1, 15-19 min), fraction-dC (Fr-dC, 19-22 min), fraction-2 (Fr-2, 22-27.0 min), fraction-dG (Fr-dG, 27.0-27.5 min), fraction-dT (Fr-dT, 27.5.-28.5 min), fraction-3 (Fr-3, 28.5-32 min), fraction-dA (Fr-dA, 32-33.5min), fraction-4 (Fr-4, 33.5-38.75 min), fraction-5 (Fr-5, 38.75 – 45.75 min) and fraction-6 (Fr-6, 45.75-53.00). All the fractions collected after the elution of dA were, unified and dried at reduced pressure. Once dried, the samples were stored at -20 °C.

5. LC/MS^3 Adductomic Analysis

The dried DNA samples were reconstituted in 20µL of LC-MS water (LCMS grade, Fluka) and then analyzed with a NanoUPLC system (Ultimate 3000, Thermo Scientific, Waltham, MA) coupled to an Orbitrap mass detector (Fusion-Thermo Scientific, Waltham, MA). The UPLC system operated with a 5µL loop. The chromatographic separation was performed with an a RP-column created by hand packing a commercially available fused-silica emitter (230x0.075 mm, 15 µm orifice, New Objective, Woburn MA) with C18 stationary phase (5 µm, 100Å, Luna-Phenomenex, Torrace, CA). The mobile phase consists of formic acid (0.05

$\%_{v/v}$ in H_2O, phase-A) and acetonitrile (100$\%_{v/v}$, phase-B). The elution program involved an isocratic step (2 % of B for 5 min at 1μL·min^{-1}), followed by a linear gradient of B (1.5 %·min^{-1} for 25 min at 0.3 μL·min^{-1}) and it concluded with a washing isocratic step, performed at 98% of B for 5 min at 0.3μLmin^{-1}. At the end of the elution program, the LC-system was equilibrated for 5 min in isocratic condition (2% of B, 1 μL·min^{-1}). In the course of the LC run, the injection valve switched at 6 min, excluding the sample loop from hydraulic path. This operation allowed performing several washes of the injection system, avoiding carryover and preventing memory effects. The LC system was interfaced to the MS-detector using a Nanoflex ESI ion source (Nanoflex Thermo Scientific, Waltham, MA). The source operated in positive ion mode at RT conditions. The electrospray voltage was set at 2.5 kV and the temperature of the ion tube was set up at 350 °C. The overall ion optics were optimized monitoring the background signal 371.1012 m/z (oligosilorxane, $[C_2H_6SiO]_5$).

The MS-analyses consist of three detection events: full scan, untargeted data dependent MS^2-acquisition (dd-MS^2) and a neutral loss MS^3-data acquisition (NL-MS^3). The full scan (100-1000 m/z) was performed using the front quadrupole to fill up the C-Trap, which worked with a maximum injection time of 50 ms and automatic gain control (AGC) of $5 \cdot 10^4$. The MS-spectra were acquired by the Orbitrap at resolution of 60000 (ref. 400 m/z). The five most abundant ions detected during each full scan event were picked to trigger the dd-MS^2 fragmentation events. The mass tolerance required to trigger the MS^2 data acquisition was set at 5 ppm. A dynamic exclusion of 20s, and an intensity threshold of 10^4 counts were introduced to better manage the instrumental dwell time. In the course of the dd-MS^2 acquisitions, the front quadrupole was used to isolate each individual top 5 precursor ion (isolation width ± 1.5 m/z). The fragmentations were performed in the high pressure stage of the linear ion trap (LIT), which operated with a normalized collision energy of 30 % CID and an activation time of 10 ms. In order to measure the accurate mass of the fragment ions, the MS^2 spectra were recorded with the Orbitrap detector, which operated with a resolution of 15000 (ref. 400 m/z) and a max injection time of 200 ms. In the course of the NL-MS^3 data acquisitions, the ion trap was used to isolate the three most abundant MS^2-fragment ions (isolation width of ±3.0 m/z), which gave the neutral loss signal comparable to the release of the deoxyribose moiety (-dR; 116.0474 ± 0.0006 m/z, 5ppm). The MS^3-fragmentations were performed with the ion routing multipole, which operated with normalized collision energy of 50 % HCD. The MS^3-spectra were recorded with the Orbitrap, which performed a single microscan with a resolution of 15000 (ref. 400 m/z) and operated with injection time of 300 ms (Figure C.5).

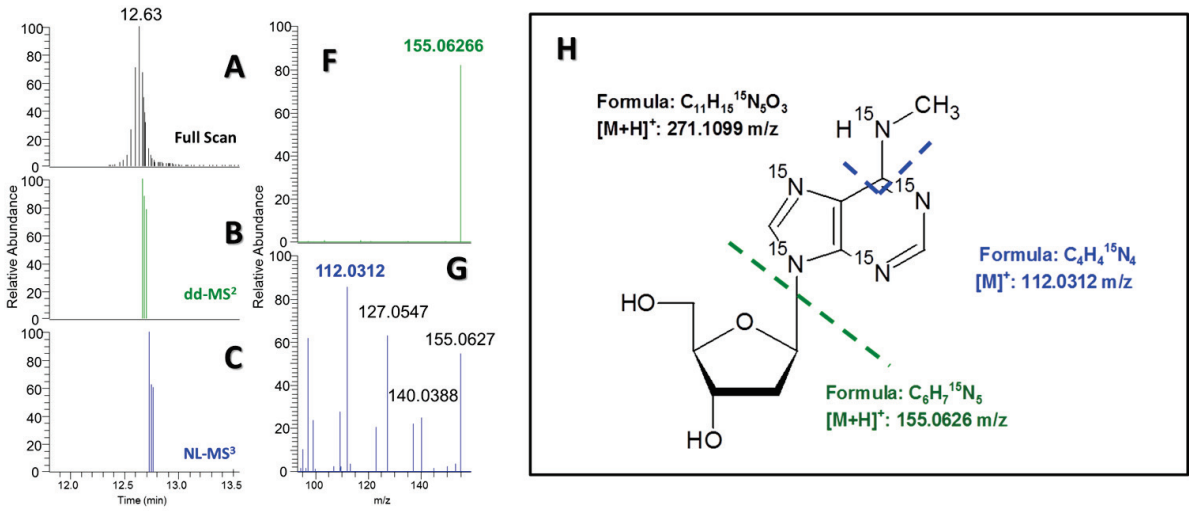


Figure C.5: Representative example of a ^{15}N isotopically labelled DNA adduct spiked into the sample ($^{15}N^6$-Me-dA). Box-A reports the chromatogram of the molecular ion $[M+H]^+$. Box-B reports the MS^2 chromatogram and Box-F the corresponding MS^2 fragmentation spectrum related to the neutral loss of the deoxyribose moiety. Box-C and Box-G report the MS^3 chromatogram and the MS^3 fragmentation spectrum related to the residual modified nucleobases. Box-H summarizes the fragmentation pathways observed in both the MS^2 and the MS^3 detection events.

6. Adductomic Data Analysis

The raw data files are extracted and converted into an ASCII format by a customized program, developed by Lin He at the Scripps Research Institute (He et al., 2015). Files are then analyzed with a homemade script operating in Excel® and MATLAB® environments. The script can load the ASCII data files and it automatically extracts all the MS^2 fragmentations which involved the neutral loss of the deoxyribose moiety (NL -116.0474 m/z). The software excludes all the redundant signals present in each data set using Boolean operators to filter out all the signals which have been simultaneously detected within a retention window of ± 1 min and any signal which has comparable molecular weights within a mass tolerance of ± 5 ppm. The filtered data sets from an exposed sample and a control can been merged together in a common data file, where a second subroutine excludes the signals common to both data files (time tolerance ±1 and mass tolerance ±5 ppm). In the end the script identifies in the MS^3 data set all the signals, which account for appearance of one of the nucleobases (guanosine, adenosine, cytosine or thymine). These signals are the diagnostic feature used for the identification of candidate DNA adducts. The description of the algorithms is summarized in the flow chart depicted in Figure C.6.

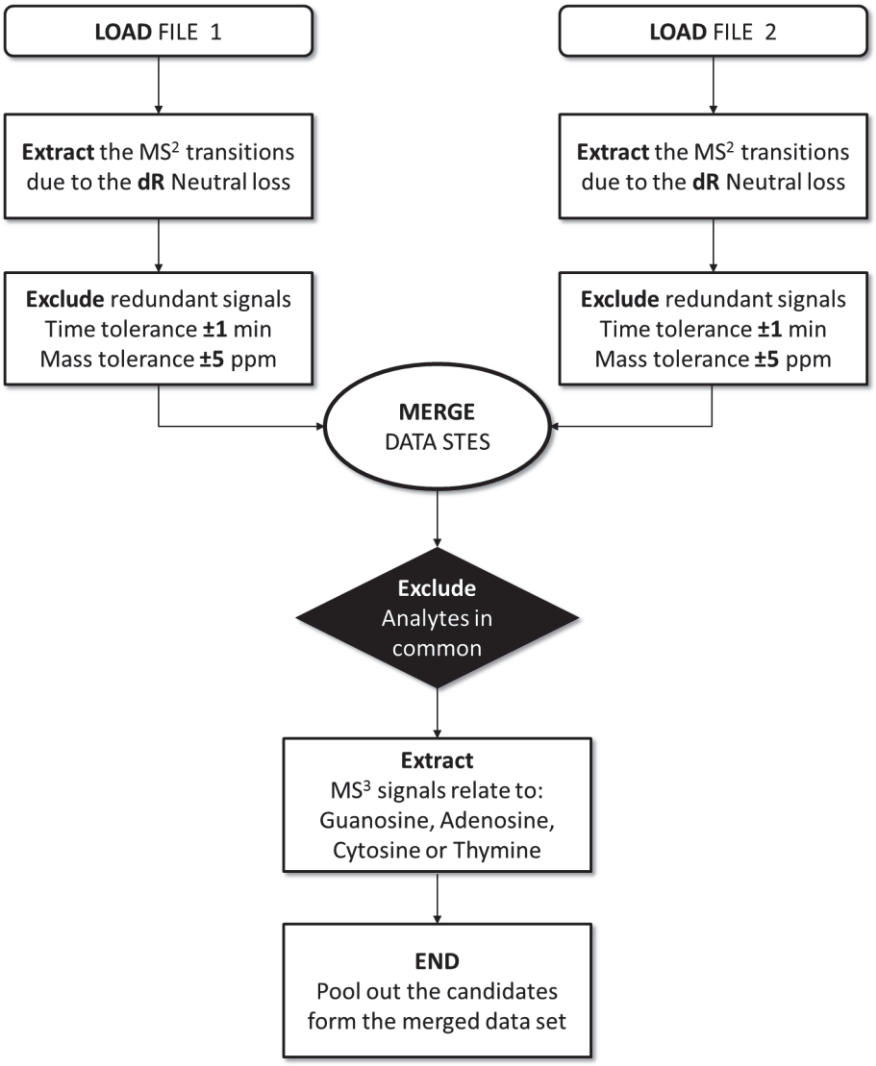


Figure C.6: Data analysis flow chart where File 1 is a data set deriving from the analysis of a control and file 2 from the analysis of an exposed sample.

Appendix D: Published review (Zhivagui et al., 2016)

BCPT Basic & Clinical Pharmacology & Toxicology

Basic & Clinical Pharmacology & Toxicology Doi: 10.1111/bcpt.12690

MiniReview

Modelling Mutation Spectra of Human Carcinogens Using Experimental Systems

Maria Zhivagui, Michael Korenjak and Jiri Zavadil

Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer (WHO), Lyon, France

(Received 22 August 2016; Accepted 13 October 2016)

Abstract: Mutation spectra in cancer genomes provide information on the disease aetiology and the causality underlying the evolution and progression of cancer. Genome-wide mutation patterns reflect the effects of mutagenic insults and can thus reveal past carcinogen-specific exposures and inform hypotheses on the causative factors for specific cancer types. To identify mutation profiles in human cancers, single-gene studies were first employed, focusing mainly on the tumour suppressor gene *TP53*. Furthermore, experimental studies had been developed in model organisms. They allowed the characterization of the mutation patterns specific to known human carcinogens, such as polycyclic aromatic hydrocarbons or ultraviolet light. With the advent of massively parallel sequencing, mutation landscapes become revealed on a large scale, in human primary tumours and in experimental models, enabling deeper investigations of the functional and structural impact of mutations on the genome, including exposure-specific base-change fingerprints known as mutational signatures. These studies can now accelerate the identification of aetiological factors, contribute to carcinogen evaluation and classification and ultimately inform cancer prevention measures.

Cancer in humans is characterised by a wide range of somatic mutations that confer a growth advantage on cells, leading to the development of a neoplasm [1,2]. These mutations often result from either endogenous defects in homeostatic biological pathways (*e.g.* DNA damage repair) or exogenous factors, such as exposures to chemical carcinogens. Various assays have been used to evaluate the genotoxic impact of the studied compounds. The Ames test employs the bacterium *Salmonella typhimurium* and is commonly used to investigate the mutagenic properties of chemicals. It allows the evaluation of a large number of compounds in a short time, and, depending on the bacterial strain, point or frameshift mutations can be investigated [3]. In eukaryotic cells, comet and micronucleus assays are frequently used to assess the potential of test chemicals to induce DNA breaks. These assays have been instrumental in assessing mechanistic information on compound genotoxicity by programs such as the IARC Monographs. However, these tests rely on prokaryotic systems (Ames), can be laborious (comet, micronucleus) and, most importantly, do not provide insight regarding the specific base changes and other features such as the sequence context.

Some mutagenic carcinogens leave specific mutation imprints on the DNA, as exemplified by tobacco smoke carcinogens and ultraviolet (UV) light, causing characteristic mutation patterns in cancers of the lung and skin, respectively [4,5]. Human tumours arise from various causes, and this is reflected in heterogeneous mutation patterns, often a composite result of the action of multiple mutagenic processes throughout the cell lineage life-time. With the advent of massively parallel sequencing, cancer genome studies have accumulated large amounts of mutation data accessible from dedicated data repositories (COSMIC, TCGA, ICGC data portals). Recently, Alexandrov *et al.* [6] analysed single-base substitutions of about 12,000 human tumours using advanced mathematical approaches and extracted mutational signatures of critical mutagenic processes operating in cancer cells. This approach revealed over 30 discrete mutational signatures in about 40 cancer types (see http://cancer.sanger.ac.uk/cosmic/signatures). Some (~37%) of the identified mutational signatures were attributed to endogenous mutagenic processes, for example spontaneous deamination of 5-methylcytosine, activity of APOBEC cytidine deaminases, deficiency of DNA repair mechanisms and polymerase η. Others (~23%) were linked to exogenous environmental exposures including tobacco smoking and chewing, aflatoxins, aristolochic acids (AA), alkylating therapeutic agents or UV light. However, about 40% of the identified signatures remain of unknown aetiology. This knowledge gap can be closed by extended molecular cancer epidemiology studies and concurrent development of new experimental models for systematic genome-wide mutagenicity testing of candidate carcinogenic exposures. Here, we present a brief overview of experimental models developed to date to investigate mutagenic processes associated with specific carcinogenic exposures. We also discuss additional, emerging model systems that can be explored for modelling of mutational signatures.

Author for correspondence: Jiri Zavadil, Molecular Mechanisms and Biomarkers Group, International Agency for Research on Cancer (WHO), 150 cours Albert Thomas, 69372 Lyon cedex 08, France (e-mail zavadilj@iarc.fr).

Experimental Mutation Spectra

Single-gene approaches.

Experimental systems that are based on single-gene screening approaches either rely on an efficient (phenotypic) selection method to detect enrichment of mutations (*e.g.* in bacterial reporter genes) or depend on genes that are frequently mutated in the context of a biological barrier bypass and clonal selection step, with the *TP53* tumour suppressor gene being the best example.

Reporter genes.

Reporter gene-based mammalian *in vitro* and *in vivo* model systems were designed for chemical mutagen or radiation exposure studies. Commonly applied *in vitro* models utilise the property of endogenous enzymes (*e.g.* HPRT, DHFR, TK) to convert certain media supplements to toxic metabolites, as a means of selecting for mutations in the encoding genes. In contrast, rodent *in vivo* model systems are characterised by the genomic integration of an engineered transgene that frequently consists of a reporter (such as *lacI, lacZ, gpt, gpa, hprt, aprt, supF* or *cII*) and a viral shuttle vector. After carcinogen exposure and genomic DNA isolation, the bacterial reporter gene is packaged into phage particles. These ensure efficient delivery of the target gene into a bacterial host, in which mutation screening can be carried out using chromogenic or viability selection [7]. Both *in vitro* and *in vivo* model systems have been extensively used to assess the mutagenicity of various carcinogens. Multiple assays linked the heterocyclic amine 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP), a common dietary carcinogen in cooked meat, to an increased rate of G>T transversions [8,9]. Other examples of carcinogens for which characteristic mutations in reporter genes have been observed include the dietary carcinogen acrylamide (A>T) [10], dietary and environmental carcinogens such as aflatoxin B1 (G>T) [11–13], and chemotherapeutic agents such as 8-methoxypsoralen (T>A) [14]. Hence, data generated using the reporter gene approach can be used to extract mutation profiles specific to the exposure. However, *in vivo* mutation analysis of reporter genes requires access to an animal facility and is tailored to the investigation at small-scale, single-locus level.

The TP53 *tumour suppressor gene.*

Due to the high frequency (~50%) of *TP53* mutations across many tumour types, comparison of its mutation spectra in distinct tumours of different proposed aetiologies became a focus of many studies. In a classic 1991 paper, Hollstein *et al.* [15] reviewed *TP53* mutation data in a variety of cancer types including colon, breast, lung and oesophageal cancer, hepatocellular carcinoma (HCC), lymphoma and leukaemia. *TP53* exhibited varying mutation spectra depending on the cancer type. In smokers' lung cancer, guanine on the non-transcribed strand was predominantly substituted with thymine (G>T), whereas this mutation was not seen to the same extent in non-smokers. Moreover, HCC was evaluated in patients from different geographical regions. A comparison with non-malignant cells from the same individuals revealed frequent G>T transversions in codon 249 of *TP53* in regions of high risk of liver cancer due to aflatoxin B1 exposure. In HCC cases from Japan, where aflatoxin B1 incidence is relatively low, the types and sequence of point mutations were markedly different.

Genetically engineered model systems. Genetically engineered model systems expressing human *TP53* gene have been devised to study exposure-specific mutation patterns.

In vivo animal studies. Hupki (human *p53* knock-in) mice, in which exons 4–9 of mouse *p53* were replaced by human *TP53* exons in the germ line [16], were exposed to class B UV light [17]. Tumours isolated from these mice exhibited characteristic *TP53* gene mutations similar to those predominantly observed in human skin cancer (C>T).

In vitro studies. Hupki mouse embryonic fibroblasts (Hupki MEFs), isolated from Hupki mice, were exposed as primary cells to UV light [18], AA [19], benzo[a]pyrene (B[a]P) [20] or 3-nitrobenzanthrone (3-NBA) [21] and passaged to senescence. Cells were propagated until the senescence bypass followed by clonal expansion, using the 3T3 protocol [22]. Senescence bypass in MEFs depends on the functional inactivation of the p53-p19^{ARF} tumour suppressor pathway, and resulting cell lines were screened for mutations in the *TP53* gene by Sanger sequencing. In combination with mining of *TP53* human cancer databases, this approach showed that the arising immortalised Hupki MEFs recapitulated the human cancer *TP53* mutation profiles associated with the same exposures [23–25].

Yeast systems were also exploited as *in vitro* models of *TP53* mutagenesis using a strain transfected with an expression vector harbouring human wild-type *TP53* cDNA that had been UV-irradiated *in vitro*. The results revealed CC>TT transitions, in keeping with the pattern observed in human skin cancer [26].

Human cell lines. Normal human fibroblasts were exposed to known carcinogens, including B[a]P, aflatoxin B1 and acetaldehyde, and mutation patterns of *TP53* were evaluated by functional analysis of separated alleles in yeast (FASAY). Similar to the experiments in genetically engineered *TP53* model systems, this assay replicated mutation profiles observed in human *TP53* mutant tumours linked to each particular exposure [27].

These approaches highlight the convergence of epidemiological and experimental data for the establishment of causal association between environmental exposures and human cancers [28]. However, there are limitations: *TP53* mutations conferring selective advantage may not always occur or become selected for during cell transformation. Additionally, these aforementioned assays tend to be laborious, time-consuming and resource-intensive. For instance, many cell lines must be generated to accumulate enough *TP53* mutations for extracting a specific mutation profile [29].

In sum, the reporter and single-gene sequencing studies yield mutation imprints that can be helpful in extracting rudimentary 'signatures' of mutagens, although with limitations. These low-complexity methods are becoming gradually replaced by increasingly affordable and more robust genome-scale approaches, as discussed below.

Massively parallel sequencing approaches.

Due to the wealth of data that can be derived from individual high-throughput sequencing experiments, carcinogen exposure combined with massively parallel sequencing has become a time- and labour-efficient alternative for the identification of mutation spectra. As a result of more recent advances in massively parallel sequencing technologies and related bioinformatics analyses, hypotheses regarding putative cancer-risk factors can now be tested by extracting genome-scale mutational signatures from genome-wide sequencing studies in primary as well as model systems [6,30,31]. In practice, replicates (biological and experimental) are a prerequisite to ensure the robustness of the assay and interpretation of results. The experimental approaches using massively parallel sequencing rely on clonal expansion of cells from a *bona fide* single-cell founder, to enrich for a homogeneous cell population and consequently a clonal enrichment of acquired variants.

Due to its small genome and the correspondingly lower sequencing cost, budding yeast (*Saccharomyces cerevisiae*) was the first model system applied to systematically study mutation introduction in combination with massively parallel sequencing. Work in yeast focused primarily on the effect of gene inactivation, either singly or in combination, on overall mutation spectra [32]. Most of the tested strains were deficient in well-conserved human gene orthologues with critical roles in genetic diseases and cancer and included mismatch repair [33,34] and DNA replication genes [35] as well as a large set of yeast mutator alleles [36,37]. In these assays, yeast strains were propagated for long-term mutation accumulation (MA) by performing up to 100 single-cell bottleneck passages, and the resulting MA strains were investigated for accumulation of single-nucleotide variants, small indels and large structural variants. This led, for example, to the identification of frequent C>T transitions and indels at homopolymeric repeats in mismatch repair mutants, features that were later attributed to a mismatch repair mutational signature in human cancers. The MA approach has proven to be an elegant way to functionally test the impact of mutations in specific genes on the mutation spectra of somatic cells and provides a valuable resource for comparison with human tumour-sequencing data. The integrated single-cell bottleneck steps ensure the identification of active mutagenic processes rather than selection-bias effects.

Analogous experiments were performed in another model organism, *Caenorhabditis elegans,* to investigate mutation landscapes in DNA repair-deficient worms upon exposure to carcinogenic agents [38]. One hundred and eighty-three worm populations were either followed for 20 generations, each of which represented a single-cell bottleneck at the zygote stage of the hermaphrodite, or were exposed to one of three mutagens (aflatoxin B1, cisplatin or mechlorethamine), and the offspring was collected after one single-cell bottleneck for genome-wide sequencing. The resulting mutation patterns based on single-base substitutions (SBS) were the predominant event in most of the backgrounds, with additional effects of the exposures manifesting by large structural variants. This approach allows investigating the interactions between endogenous factors and external exposures.

Yeast and *C. elegans* can be easily genetically manipulated and have well-annotated reference genomes. They are thus promising simple model systems for extracting mutation patterns of intrinsic mutagenic processes and exposure to carcinogenic compounds, as well as for functional studies. However, in comparison with mammalian systems, their small genome size limits the number of mutations that can be extracted. In the context of weak carcinogens and the less potent nature of several intrinsic mutagenic processes, this limitation can complicate the identification of high-confidence mutational signatures, particularly due to low-information contents on the trinucleotide context [6]. This limitation also applies to the single-gene studies discussed above, and it often needs to be overcome by increasing the number of samples sequenced. In addition, potential differences between simple and higher organisms in pro-carcinogen metabolization, DNA repair pathways or the effects of chromatin structure on mutation distribution are likely to play a role in the resulting patterns and should be carefully considered.

In mammalian cells, a pioneering carcinogen exposure study based on parallel sequencing of 150 copies of a reporter gene (*cII*) from a mouse *in vivo* model system was able to recapitulate results from conventional low-throughput analysis for multiple mutagens, with high sensitivity [39]. Subsequently, next-generation sequencing was carried out in Hupki MEFs that had been exposed, as primary cells, to several known mutagens, including AA, B[a]P, methylnitronitrosoguanidine (MNNG) and class C UV light (UVC) [30]. After senescence–barrier bypass and clonal expansion, the exome of immortalised cells was sequenced and mutational signatures based on SBS were extracted by the non-negative matrix factorization algorithm. Immortalised cell lines represent homogeneous populations of one predominant clone and less represented subclones, which allows reliable identification of enriched SBS and corresponding mutation profiles upon sequencing at reasonable coverage. Importantly, the exome-wide mutation patterns identified in Hupki MEF cells recapitulated those observed in human cancers with known links to the tested exposures [30].

Going beyond the exome level, a more recent study conducted genome-wide sequencing on Hupki MEFs exposed to AA, B[a]P and UVC and reproduced the previous exome-scale findings of signatures based on significantly higher mutation counts [31]. In addition, the whole-genome scale allows for comprehensive evaluation of structural variants, small insertions and deletions, copy number variants and mutations in non-coding regions.

Hupki MEFs have proven to be an important *in vitro* model system for mutagenicity testing of various chemical

carcinogens. Yet, the system has limited capacities in metabolic activation of certain pro-carcinogens. Moreover, using mouse cell lines may not be an optimal way to recapitulate exposures in human beings due to the overall differences in the genetic background and species-specific repair mechanisms. Therefore, there has been increased focus on devising assays using human cell lines to be used for experimental extraction of genome-wide mutational signatures.

HK-2 human renal proximal tubular cells were first used to identify the mutation fingerprint of AA [40]. In this approach, the cells were treated with sublethal doses of AA for 6 months until single clones could be recovered for sequencing, and the extracted mutation profile recapitulated the one identified in urothelial carcinomas of AA-exposed patients [40,41].

Next, Severson *et al.* [42] used primary human mammary epithelial cells (HMECs) to extract a B[a]P mutation pattern. On the path to immortalization, HMECs bypass two well-defined biological barriers. The first one is stress-induced senescence or stasis and is dependent on the p16–RB pathway. This hurdle was overcome by exposing the cells to mutagenic B[a]P. The second barrier is replicative senescence, which is due to critical telomere shortening and subsequent genome instability. In rare cases, exposed HMECs were able to overcome replicative senescence and become immortal. Importantly, clones harbouring sufficient numbers of mutations to extract carcinogen-specific mutation profiles can be generated after the stasis bypass with no need to rely on the immortalization step [42]. This makes HMECs a suitable model for the analysis of mutational signatures of carcinogens acting on epithelial cells.

Despite the obvious advantages that the human experimental systems offer, the actual experiments still tend to be lengthy and laborious as human cells do not immortalise as readily as Hupki MEFs.

Requirements, existing and emerging experimental models for compound testing

Experimental investigation of carcinogen fingerprints on the DNA sequence can ensure the elucidation of the enigmatic signatures observed in human tumours. Ideally, such studies would require model systems that enable analysis of large numbers of compounds within a reasonable time frame, include a single-cell bottleneck or barrier bypass–clonal expansion step and be able to recapitulate key aspects of human tumour biology (*e.g.* metabolism, DNA repair pathways). Most model systems that are applicable for the study of mutational signatures by the means of massively parallel sequencing meet some, but not all, of these criteria (Table 1). Yeast and *C. elegans* have short generation times, enabling time-efficient investigations of many compounds. Their large evolutionary distance from humans, however, can pose a limitation to their usability. Mouse embryonic fibroblasts are a proven model for the extraction of carcinogen-induced mutational signatures. Nonetheless, similar to yeast and *C. elegans*, critical interspecies differences in DNA repair mechanisms and the metabolic activation of pro-carcinogens have to be considered. The addition of the human S9 fraction, comprising active metabolic enzymes such as cytochrome P450 and transferases, can boost the metabolic pathways in MEFs and thus circumvent the latter limitation. Experimental models in human cell lines that meet important requirements for systematic analyses of cancer-risk agents include the HMEC and HK-2 cells. Whereas the HMEC system is based on primary cells and their abilities to bypass senescence and clonally expand, the HK-2 model utilises a clonal expansion step due to chronic carcinogen exposure in immortalised cells and can potentially be extrapolated to any other immortal cell line. However, the use of both these human systems is laborious and the number of compounds that can be tested is limited.

Additional emerging models could be explored for analysis of mutational signatures (Table 1). The generation of iPS cells offers a time-efficient strategy for clonal expansion, but lacks the barrier bypass step of MEF and HMEC and is usually based on metabolically less active cells originating from fibroblast cultures. iPS cells can, however, be generated from a number of different cell types, which could be exploited to match certain carcinogenic exposures to their target cell or tissue types. Higher-structure organoids would also offer a similar tissue-of-origin-specific approach to study carcinogen exposure. They can be generated from multiple tissues and have been used for clonal expansion from colon, small intestinal crypts and liver single cells followed by whole-genome sequencing [43]. It is conceivable that such a strategy could be combined with *in vitro* carcinogen exposure before the clonal expansion outgrowth, but the potential adaptation of such an experimental strategy warrants further investigations. Finally, animal carcinogen exposure bioassays are lengthy experiments, but over the last decades, several genetic toxicology programs have archived thousands of exposure-related tumour tissues in their repositories. These can be accessed for retrospective *in vivo* characterization of carcinogen-induced mutational signatures.

Application in cancer aetiology studies

The knowledge of distinct mutational signatures identified from cancer genome sequencing [6] or experimental model systems, combined with the availability of thousands of sequenced human cancer samples covering virtually all cancer types, allows screening of the available data for known exposure-related signatures, to infer aetiological factors. For instance, the AA signature identified in HK-2 cells and human urothelial tumours with known exposure was used to screen hepatocellular carcinomas, a new cancer type in which AA was identified as a strongly contributing mutagenic factor [40].

Furthermore, deciphering the mutagenic effects of novel carcinogens epidemiologically linked to human cancers may unravel new cancer causes [44]. To this end, systematic analyses in experimental models are required to validate such findings. Similarly, genome-wide effects of endogenous mutational processes (*e.g.* deregulated APOBEC and AID cytidine deaminase activities) can be investigated using the discussed *in vitro* models, to test hypotheses generated by analysis of human tumour-sequencing data [6,30,45–48] and to characterise novel mutational signatures that can be further used to

Table 1.
Examples of *in vitro* and *in vivo* experimental models for next-generation sequencing-based assessment of mutational signatures of carcinogens.

	Model system	Experimental time frame	Clonal expansion	Relevance to human tumour biology	Advantages	Limitations	Reference
Tested experimental models	*S. cerevisiae*	Mutation accumulation followed by 1 to 100 single colony bottlenecks (7 days to 10 months)	Clonally expand from single cells	Limited	• Well-annotated reference genome • Mutation accumulation approach ensures the identification of active mutagenic processes	• Interspecies differences to human models (*e.g.* metabolism, DNA repair) • Not tested for exogenous mutagens	[32–37]
	C. elegans	Mutation accumulation followed by one single-cell bottleneck (~1 month)	Each generation represents a single-cell bottleneck (zygote)	Limited	• Well-annotated reference genome • Mutation accumulation approach ensures the identification of active mutagenic processes	• Interspecies differences to human models (*e.g.* metabolism, DNA repair) • Requires specific equipment for manipulation	[38]
	Hupki MEF cells	Immortalization within 2 months	Combination of biological barrier bypass (senescence) and clonal expansion	Intermediate	• Mammalian model system • Extensively tested for extraction of mutation patterns and mutational signatures	• Interspecies differences to human models (*e.g.* metabolism, DNA repair)	[30,31]
	HMEC primary cells	Immortalization within 6 months	Combination of biological barrier bypass (stasis, senescence) and clonal expansion	High	• Human model system • Potential for organoid culture	• Difficult to manipulate and maintain	[42]
	HK-2 cell line	Clonal expansion within 6 months	Clonally expand (due to chronic exposure)	High	• Human model system • Potential to match carcinogens to target tissue (through adaptation to other cell lines)	• Immortalised cell line • Virally transformed	[40]
Emerging models	Animal bioassays	2-year carcinogenesis studies	Clonal expansion in the context of *in vivo* tumour formation	Intermediate	• Availability of archived tissue from toxicology programs • *In vivo* mutation screening of tumours	• Costly and lengthy process • High burden of animal use • Technical challenges (*e.g.* FFPE-derived DNA)	[50]
	Induced pluripotent stem cells (iPSC)	3–4 week iPS cell induction protocols, excluding a potential exposure step	Clonally derived	High	• No manual selection of individual iPSC colonies • Time-efficient human model system • Potential to match carcinogens to target tissue	• Metabolic limitations of the most commonly used cell type (fibroblasts) • Not tested for mutational signature extraction	[51]
	Organoids	Clonal organoid line within 1 to 15 weeks (tissue type-dependent) excluding a potential exposure step	Clonally expand	Very high	• Better resemblance of *in vivo* context than 2D cultures • Potential to match carcinogens to target tissue	• The establishment of organoid cultures is laborious • Technical challenges (*e.g.* exposure and clonal expansion) • Not tested for mutational signatures of chemical exposures	[43,52]

study the contribution of different processes to the mutation burden of human cancers [37,38].

Carcinogen-immortalised Hupki MEF and HMEC cells have been shown to acquire mutations in genes and pathways involved in human cancer development [30,42]. These findings are promising with respect to the potential use of the models not only to extract mutational signatures, but also to investigate the functional impact of the incurred mutations. *In vitro* models of carcinogen-induced cell immortalization could be used to investigate driver genes and events contributing to early steps of cell transformation. Some of these genes, such as the *RAS* oncogene or the *TP53* tumour suppressor gene, have been long studied functionally using *in vitro* systems. Other less frequently mutated potential drivers are now being identified thanks to computational analyses of tumour-sequencing data [49] and could be cross-referenced with sequencing results from well-controlled experimental exposure models.

Readily available resources from animal cancer bioassays, such as archived tissues, in combination with the increased use and the establishment of improved *in vitro* systems hold a great potential for reducing the use of animals in carcinogen testing. Based on compound selection guided by known mechanistic features (*e.g.* pro-mutagenic DNA-adduct formation) and single-gene approaches, the cross-comparison of the genome-wide patterns derived from *in vivo* and *in vitro* experimental exposure systems with data from human primary tumours and cancer genome repositories, such as ICGC, COSMIC and TCGA, can ensure the identification of high-confidence, carcinogen-specific mutational signatures. The availability of evidence-based mutagenicity data from experimental models will contribute significantly to the pace and accuracy of carcinogen evaluation and classification by programs such as the IARC Monographs, and such data could be used to verify exposure estimates in epidemiological studies. Finally, the knowledge regarding mutational signatures of putative aetiological agents could inform diagnostic procedures with respect to potential carcinogen exposures and has the potential to improve prevention as well as early detection of cancer.

Acknowledgements

M.Z. is a recipient of the European Environmental Mutagenesis and Genomics Society 2016 Young Scientist grant. The study was partially supported by the INCa-INSERM Plan Cancer 2015 grant to J. Z.

Conflict of interest

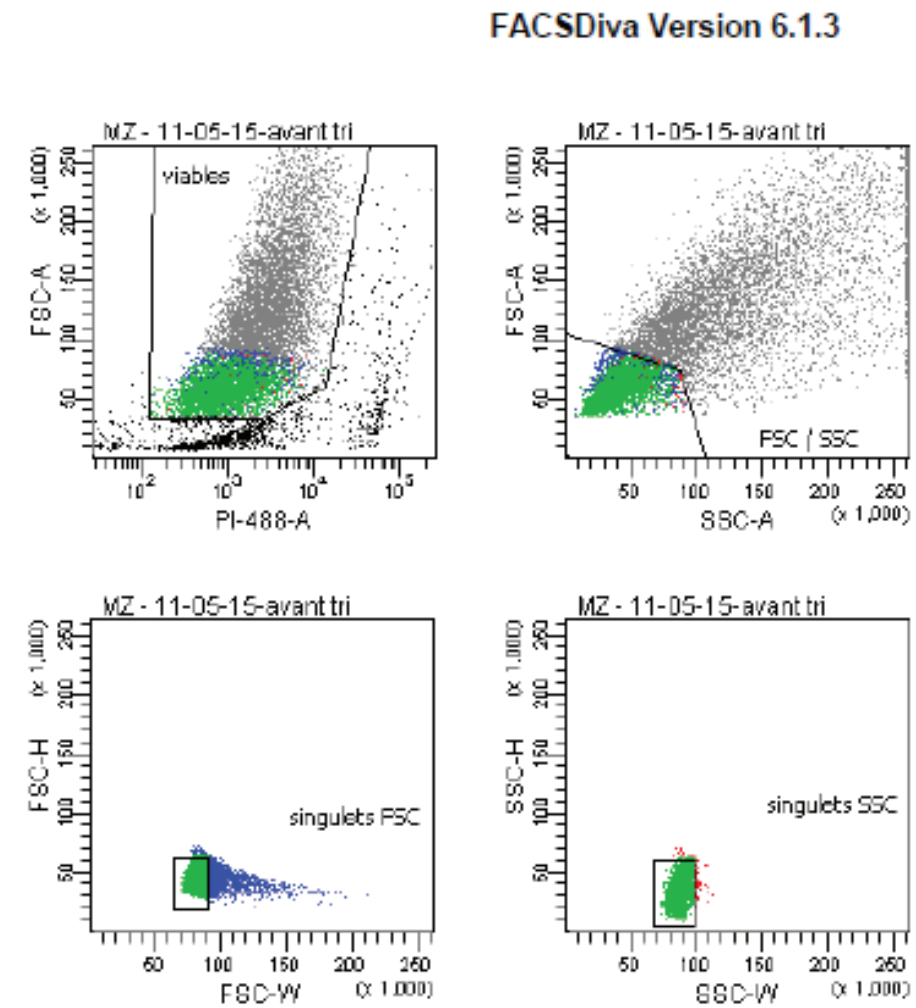
The authors have no conflict of interest to disclose.

References

1 Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;**144**:646–74.
2 Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science 2013;**339**:1546–58.
3 Ames BN. Identifying environmental chemicals causing mutations and cancer. Science 1979;**204**:587–93.
4 Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010a;**463**:191–6.
5 Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 2010b;**463**:184–90.
6 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.
7 Boverhof DR, Chamberlain MP, Elcombe CR, Gonzalez FJ, Heflich RH, Hernández LG *et al.* Transgenic animal models in toxicology: historical perspectives and future outlook. Toxicol Sci 2011;**121**:207–33.
8 Carothers AM, Yuan W, Hingerty BE, Broyde S, Grunberger D, Snyderwine EG. Mutation and repair induced by the carcinogen 2-(hydroxyamino)-1-methyl-6-phenylimidazo[4,5-b]pyridine (N-OH-PhIP) in the dihydrofolate reductase gene of Chinese hamster ovary cells and conformational modeling of the dG-C8-PhIP adduct in DNA. Chem Res Toxicol 1994;7:209–18.
9 Lynch AM, Gooderham NJ, Davies DS, Boobis AR. Genetic analysis of PHIP intestinal mutations in MutaMouse. Mutagenesis 1998 Nov;**13**:601–5.
10 Manjanatha MG, Guo L-W, Shelton SD, Doerge DR. Acrylamide-induced carcinogenicity in mouse lung involves mutagenicity: *cII* gene mutations in the lung of big blue mice exposed to acrylamide and glycidamide for up to 4 weeks: AA-Induced Carcinogenicity in Mouse Involves Mutagenicity. Environ Mol Mutagen 2015;**56**:446–56.
11 Levy DD, Groopman JD, Lim SE, Seidman MM, Kraemer KH. Sequence specificity of aflatoxin B1-induced mutations in a plasmid replicated in xeroderma pigmentosum and DNA repair proficient human cells. Cancer Res 1992;**52**:5668–73.
12 Trottier Y, Waithe WI, Anderson A. Kinds of mutations induced by aflatoxin B1 in a shuttle vector replicating in human cells transiently expressing cytochrome P4501A2 cDNA. Mol Carcinog 1992;**6**:140–7.
13 Wattanawaraporn R, Woo LL, Belanger C, Chang S-C, Adams JE, Trudel LJ *et al.* A single neonatal exposure to aflatoxin b1 induces prolonged genetic damage in two loci of mouse liver. Toxicol Sci 2012;**128**:326–33.
14 Sage E, Drobetsky EA, Moustacchi E. 8-Methoxypsoralen induced mutations are highly targeted at crosslinkable sites of photoaddition on the non-transcribed strand of a mammalian chromosomal gene. EMBO J 1993;**12**:397–402.
15 Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. Science 1991;**253**:49–53.
16 Luo JL, Yang Q, Tong WM, Hergenhahn M, Wang ZQ, Hollstein M. 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 2001a;**20**:320–8.
17 Luo J-L, Tong W-M, Yoon J-H, Hergenhahn M, Koomagi R, Yang Q *et al.* UV-induced DNA damage and mutations in Hupki (human p53 knock-in) mice recapitulate p53 hotspot alterations in sun-exposed human skin. Cancer Res 2001b;**61**:8158–63.
18 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 U S A 2004;**101**:2963–8.
19 Feldmeyer N, Schmeiser HH, Muehlbauer K-R, Belharazem D, Knyazev Y, Nedelko T *et al.* Further studies with a cell immortalization assay to investigate the mutation signature of aristolochic acid in human p53 sequences. Mutat Res Toxicol Environ Mutagen 2006;**608**:163–8.
20 Liu Z, Muehlbauer K-R, Schmeiser HH, Hergenhahn M, Belharazem D, Hollstein MC. p53 mutations in benzo (a) pyrene-exposed human p53 knock-in murine fibroblasts correlate with p53 mutations in human lung tumors. Cancer Res 2005;**65**:2583–7.

21 Vom Brocke J, Krais A, Whibley C, Hollstein MC, Schmeiser HH. The carcinogenic air pollutant 3-nitrobenzanthrone induces GC to TA transversion mutations in human p53 sequences. Mutagenesis 2008;**24**:17–23.
22 Sun H, Taneja R. Analysis of transformation and tumorigenicity using mouse embryonic fibroblast cells. Methods Mol Biol 2007;**383**:303–10.
23 Brocke J, Schmeiser HH, Reinbold M, Hollstein M. MEF immortalization to investigate the ins and outs of mutagenesis. Carcinogenesis 2006;**27**:2141–7.
24 Besaratinia A, Pfeifer GP. Applications of the human p53 knock-in (Hupki) mouse model for human carcinogen testing. FASEB J 2010;**24**:2612–9.
25 Kucab JE, Phillips DH, Arlt VM. Linking environmental carcinogen exposure to TP53 mutations in human tumours using the human TP53 knock-in (Hupki) mouse model. FEBS J 2010;**277**:2567–83.
26 Inga A, Scott G, Monti P, Aprile A, Abbondandolo A, Burns PA *et al.* Ultraviolet-light induced p53 mutational spectrum in yeast is indistinguishable from p53 mutations in human skin cancer. Carcinogenesis 1998;**19**:741–6.
27 Paget V, Lechevrel M, André V, Le Goff J, Pottier D, Billet S *et al.* Benzo[a]pyrene, aflatoxine B1 and acetaldehyde mutational patterns in TP53 gene using a functional assay: relevance to human cancer aetiology. PLoS ONE 2012;7:e30921.
28 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.
29 Hollstein M, Hergenhahn M, Yang Q, Bartsch H, Wang Z-Q, Hainaut P. New approaches to understanding p53 gene tumor mutation spectra. Mutat Res 1999;**431**:199–209.
30 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**. http://www.nature.com/doifinder/10.1038/srep04482.
31 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**:763–70.
32 Segovia R, Tam AS, Stirling PC. Dissecting genetic and environmental mutation signatures with model organisms. Trends Genet 2015;**31**:465–74.
33 Zanders S, Ma X, Roychoudhury A, Hernandez RD, Demogines A, Barker B *et al.* Detection of heterozygous mutations in the genome of mismatch repair defective diploid yeast using a Bayesian approach. Genetics 2010;**186**:493–503.
34 Lang GI, Parsons L, Gammie AE. Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast. G3 (Bethesda) 2013;**3**:1453–65.
35 Larrea AA, Lujan SA, Nick McElhinny SA, Mieczkowski PA, Resnick MA, Gordenin DA *et al.* Genome-wide model for the normal eukaryotic DNA replication fork. Proc Natl Acad Sci U S A 2010;**107**:17674–9.
36 Stirling PC, Shen Y, Corbett R, Jones SJM, Hieter P. Genome destabilizing mutator alleles drive specific mutational trajectories in Saccharomyces cerevisiae. Genetics 2014;**196**:403–12.
37 Serero A, Jubin C, Loeillet S, Legoix-Né P, Nicolas AG. Mutational landscape of yeast mutator strains. Proc Natl Acad Sci U S A 2014;**111**:1897–902.
38 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**:1624–36.
39 Besaratinia A, Li H, Yoon J-I, Zheng A, Gao H, Tommasi S. A high-throughput next-generation sequencing-based method for detecting the mutational fingerprint of carcinogens. Nucleic Acids Res 2012;**40**:e116.
40 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.
41 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.
42 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 Toxicol Environ Mutagen 2014;**775–776**:48–54.
43 Blokzijl F, deLigt J, Jager M, Sasselli V, Roerink S, Sasaki N *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. Nature 2016; http://www.nature.com/doifinder/10.1038/nature19768.
44 Hollstein M, Alexandrov LB, Wild CP, Ardin M, Zavadil J. Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. Oncogene 2016 Jun 6. [Epub ahead of print].
45 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**:1067–72.
46 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**:487–91.
47 Saraconi G, Severi F, Sala C, Mattiuz G, Conticello SG. The RNA editing enzyme APOBEC1 induces somatic mutations and a compatible mutational signature is present in esophageal adenocarcinomas. Genome Biol 2014;**15**:417.
48 Taylor BJ, Nik-Zainal S, Wu YL, Stebbings LA, Raine K, Campbell PJ *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. Elife 2013;**2**:e00534.
49 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** http://ww.ncbi.nlm.nih.gov/pmc/articles/PMC3788361/.
50 Hoenerhoff MJ, Hong HH, Ton T-V, Lahousse SA, Sills RC. A review of the molecular mechanisms of chemically-induced neoplasia in rat and mouse models in National Toxicology Program bioassays and their relevance to human cancer. Toxicol Pathol 2009;**37**:835–48.
51 Willmann CA, Hemeda H, Pieper LA, Lenz M, Qin J, Joussen S *et al.* To clone or not to clone? Induced pluripotent stem cells can be generated in bulk culture. PLoS ONE 2013;**8**:e65324.
52 Turner DA, Baillie-Johnson P, Martinez Arias A. Organoids and the genetically encoded self-assembly of embryonic stem cells. BioEssays 2016;**38**:181–91.

Appendix E



Tube: avant tri

Population	#Events	%Parent	%Total
All Events	11,296	####	100.0
viables	10,000	88.5	88.5
FSC / SSC	4,379	43.8	38.8
singulets FSC	3,174	72.5	28.1
singulets SSC	3,085	97.2	27.3

Figure E.1: FACS sorting of hepatocyte-like cells from the fully differentiated HepaRG dual cell population. The cells were sorted based on their viability status (propidium iodite), size (FSC) and viscosity (SSC). The number of cells sorted was not sufficient to maintain the cells at high density in culture.

Appendix F

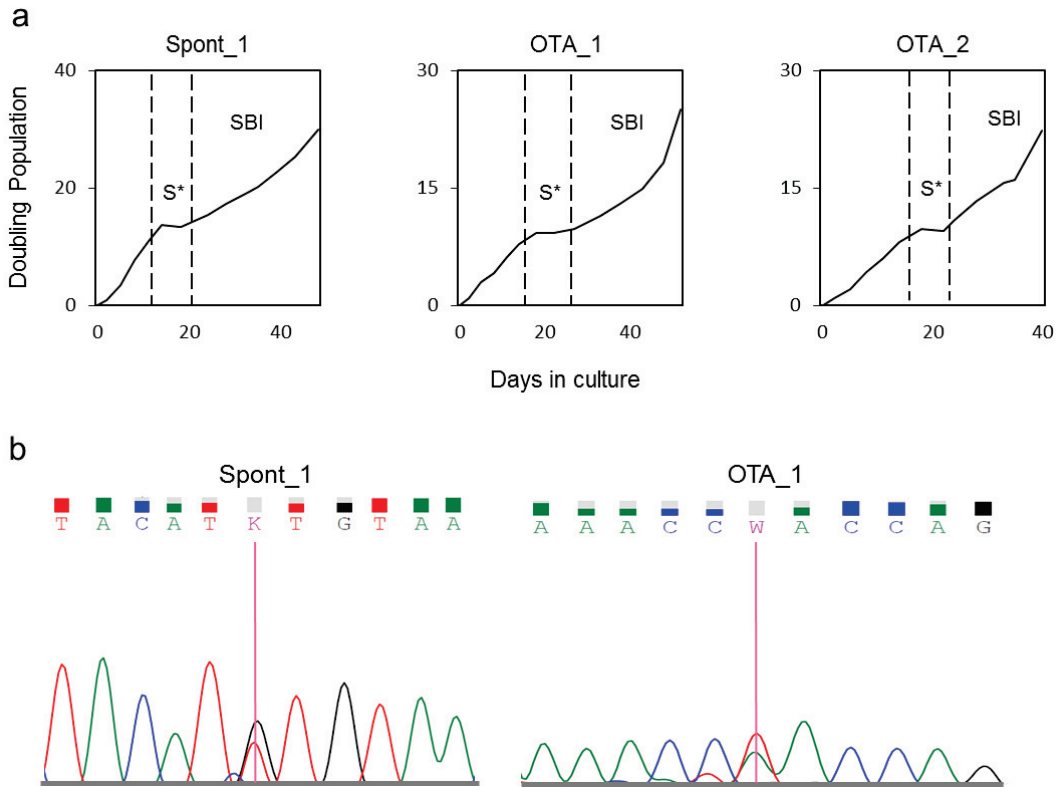


Figure F.1: Hupki immortalization and *TP53* mutation screening. a) Growth curves of Hupki MEFs. Primary cells were either left untreated (Spont) or were exposed to OTA (in the absence of human S9 fraction). X-axis represents days in culture. Y-axis represents the cumulative doubling populations. S*: senescence; SBI: senescence bypass/immortalization.

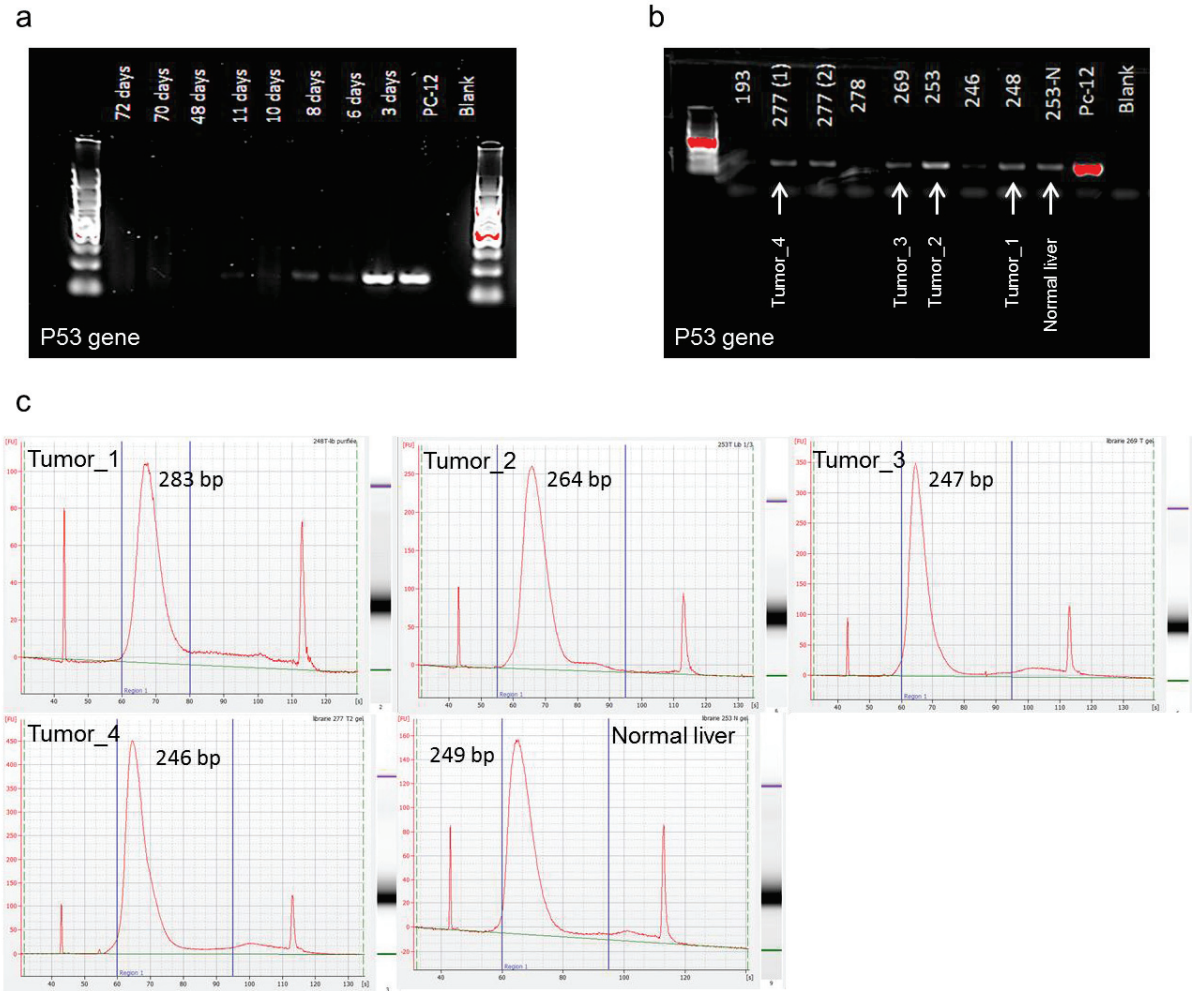


Figure F.2: **FFPE DNA quality control and library preparation for WGS**. a) PCR reaction of *P53* gene in rat FFPE tumor tissues with different fixation time, ranging from 3 to 72 days. PC-12 represents a positive control consisting of rat brain cells. b) A selection of a number of normal and tumor tissues for subsequent WGS, marked by the arrows. c) Bioanalyzer profile of 4 rat kidney tumors and 1 rat normal liver libraries. The peak size is mentioned in bp.

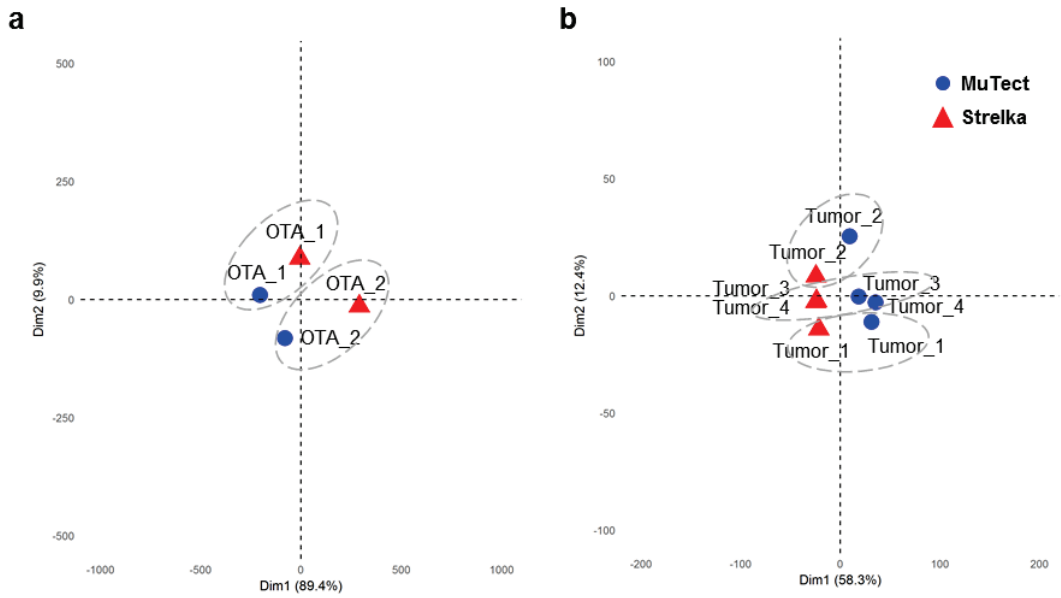


Figure F.3: PCA analysis comparing variants called by MuTect and Strelka using the 96 possible mutation types (a) for MEF clones derived from OTA exposure and (b) for rat kidney tumors developed upon treatment with OTA, after removal of low allelic frequency (<20%).

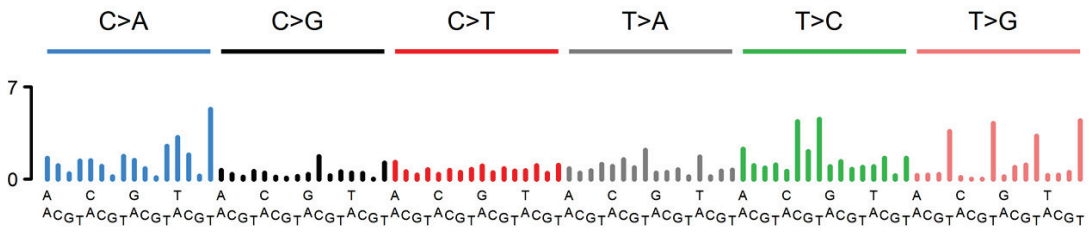


Figure F.4: Mutational signature B after baiting out signature 17 as much as possible unravelling other potential mutational patterns.

Table F.1: Quality control analysis of Hupki MEFs and FFPE rat tissues data.**Sample**	**Yield (Mbases)**	**% of >= Q30 Bases**	**Mean Quality Score**	**Mouse Genome Coverage**	**Rat Genome Coverage**
277-T2	126,941	85.88	36.57	-	45
269-T	129,156	85.75	36.53	-	46
248-T	129,376	80.08	35.04	-	46
253-T	126,969	85.39	36.42	-	45

253-N	120,532	83.85	36.01	-	43
E210	117,441	81.52	35.34	43	-
0.5 mM-OTA-S9-1	114,343	81.13	35.22	42	-
0.5 mM-OTA-S9-2	121,436	83.30	35.84	45	-