

Univerzita Karlova v Praze

1. lékařská fakulta

Charles University in Prague

First Faculty of Medicine

Studijní program: Biomedicína

Studijní obor: Fyziologie a patofyziologie člověka



UNIVERZITA KARLOVA
1. lékařská fakulta

Mgr. Hana Hušková

**Výzkum klíčových mechanismů onkogeneze s použitím
modelových buněčných systémů**

**Investigating critical mechanisms of oncogenesis using cell
model systems**

Disertační práce / Doctoral Thesis

Supervisors: Prof. MUDr. Tomáš Stopka Ph.D., Dr. Ing. Jiří Zavadil

Prague 2017

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem řádně uvedla a citovala všechny použité prameny a literaturu. Současně prohlašuji, že práce nebyla využita k získání jiného nebo stejného titulu. Souhlasím s trvalým uložením elektronické verze mé práce v databázi systému meziuniverzitního projektu Theses.cz za účelem soustavné kontroly podobnosti kvalifikačních prací.

Declaration:

I hereby declare that I have written this Ph.D. Thesis using the cited work. The text of this Thesis was not used to apply for other degree than this. I hereby agree with electronic deposition of this Thesis.

V Praze/In Prague, 15.6.2017

Mgr. Hana Hušková
Podpis/Signature

Identifikační záznam

HUŠKOVÁ, Hana. *Výzkum klíčových mechanismů onkogeneze s použitím modelových buněčných systémů. [Investigating critical mechanisms of oncogenesis using cell model systems.]* Praha, 2017. 106 s, 1 příloha. Doktorská práce (Ph.D.). Univerzita Karlova, 1. lékařská fakulta, Ústav Biocev, Praha. Školitelé: Zavadil Jiří (International Agency for Research on Cancer), Stopka Tomáš (First Faculty of Medicine, Charles University in Prague).

Acknowledgments

I would like to thank my supervisors, Dr. Jiří Zavadil and Prof. Tomáš Stopka, for their mentorship, which helped me to grow both scientifically and personally. I would also like to thank the members of their research groups – namely Karin Vargová, Stephanie Villar, Maude Ardin, Xavier Castells Domingo, Tomáš Zikmund, Helena Paszeková, Michael Korenjak, and others – for all the scientific, technical and personal support I received during the work on my Thesis project, and to Liacine Bouaoun for his advice on statistical analyses.

Next, I would like to thank Prof. Terry Dwyer of Oxford Martin School for his friendship and career advice, and the Epiladies group – Athena Sklias, Andrea Halaburkova, Szilvi Ecsedi, Nora Fernandez Jimenez, and others – for the friendship without boundaries.

Last but not least, I want to thank my family – my parents Blažena and Jiří, my brothers Jan and Jiří and their families, as well as my grandparents Jana, Čestmír, Hana, Julian and Antonín, and other members of my rather large family. They are the source of my inspiration, and the immense support I have been receiving from them not only during these last years, but for my whole life, has been essential to me. They are the giants on whose shoulders I stand. I would like to dedicate this Thesis to them and to my beloved Jan, who always lifts my spirits and makes me see the bright side of life.

TABLE OF CONTENTS

ABSTRAKT (CZ)	7
ABSTRACT (EN)	8
LIST OF ABBREVIATIONS	9
GLOSSARY	11
1. INTRODUCTION	13
1.1. The origins of cancer	13
1.2. Alterations in DNA as a record of mutagenic processes	14
1.2.1. Mutation spectra of the <i>TP53</i> gene reflect cancer aetiology	14
1.2.2. Genome-wide signatures of mutational processes operative in cancer ...	16
1.2.3. Modelling mutational spectra and signatures using experimental systems	18
1.3. Alterations in DNA as the causes and effectors of tumour physiology	21
1.3.1. Drivers and passengers	21
1.3.2. Identification of cancer driver events using sequencing data	22
1.3.3. Cancer driver pathways and complexes	25
2. INTRODUCTION TO THE THESIS	27
2.1. Hypotheses	29
2.2. Aims of the Thesis	29
2.3. Specific aims of the Thesis	29
3. MATERIAL AND METHODS	31
3.1. Material	31
3.1.1. Cell lines	31
3.1.2. Chemicals	32

3.1.3. Enzymes, antibodies, DNA and protein ladders	33
3.1.4. Commercial kits.....	34
3.1.5. Custom buffers.....	35
3.1.6. Custom oligonucleotides	36
3.1.7. Equipment.....	36
3.1.8. Software	38
3.2. Methods.....	39
3.2.1. Generation of immortalized cell lines	39
3.2.2. Sequencing library preparation and whole exome sequencing	39
3.2.3. WES data processing.....	40
3.2.4. Mutation spectra and mutational signatures.....	40
3.2.5. Pathway analysis	41
3.2.6. Identification of candidate cancer driver mutations	41
3.2.7. Single cell subcloning.....	42
3.2.8. DNA extraction and Sanger sequencing	42
3.2.9. Inhibitor treatment	43
3.2.10. MTS proliferation assay	43
3.2.11. Colony formation assay.....	43
3.2.12. Protein extraction	43
3.2.13. Polyacrylamide gel electrophoresis, immunoblotting and antibodies	44
3.2.14. Mutation analysis in human tumour data.....	44
4. RESULTS	46
4.1. Global mutation analysis	46
4.1.1. Mutation burden in MEF cell lines	46
4.1.2. Estimation of clonality of MEF cell lines	47

4.1.3. Mutation spectra analysis.....	47
4.1.4. Analysis of mutational signatures.....	51
4.1.5. Functional annotation of mutations in immortalized MEF cell lines	56
4.1.6 Mutations in cancer genes	60
4.1.7. Mutations in genes and complexes involved in regulation of epigenome. 63	
4.2. Identification and functional testing of putative cancer driver events.....	67
4.2.1. A systematic prioritization scheme for high-confidence candidate driver events	67
4.2.2. Ras ^{Q61} mutation supports cell proliferation in nutrient-poor conditions.....	68
4.2.3. Inhibition of Ezh2 activity leads to cell death in BAF-mutant cell lines in an oncogenic Ras-dependent manner	69
5. DISCUSSION	75
5.1. MEF immortalization assay to decipher mutational processes operative in human cancer.....	75
5.2. MEF immortalization assay to identify and test putative cancer driver events	80
6. CONCLUSIONS	85
7. FUTURE DIRECTIONS AND CONTEXT	87
8. REFERENCES	89
Articles	89
URLs	102
SUPPLEMENTARY INFORMATION.....	103
List of publications.....	103
Articles related to the Thesis.....	103
Articles not related to the Thesis.....	103
Supplementary Data and Figure legends	103

ABSTRAKT (CZ)

Lidé jsou v průběhu života vystaveni různým faktorům způsobujícím poškození DNA, vedoucí ke změnám v buněčné fyziologii a potenciálně k expanzi imortalizovaného buněčného klonu a vzniku nádoru. Mutace v DNA jsou jak záznamem o působení mutagenních procesů, tak klíčem k biologii a patofyziologii nádorů. Masivně paralelní sekvenování umožňuje sekvenování všech kódujících sekvencí či dokonce celých genomů lidských nádorů. Z těchto dat je možné získat vzorce mutací typické pro jednotlivé mutagenní procesy, stejně jako poukázat na mutace a geny hrající roli při vzniku a vývoji nádoru. Řada popsaných vzorců mutací však nemá známou příčinu a řada známých karcinogenů nemá dosud přiřazen mutační vzorec. Stejně tak se předpokládá, že dosud není známa řada mutací a genů s vlivem na vznik nádoru. Tato disertační práce charakterizuje experimentální systém založený na imortalizaci myších embryonálních fibroblastů (MEF) za působení mutagenu, umožňující určení vzorců mutací daných mutagenů a určení mutovaných genů důležitých pro vznik nádoru. Kultivace buněk MEF vede k jejich senescenci, která může být překonána mutacemi ve funkčně důležitých genech, analogicky ke stádiím vzniku lidských nádorů. Sekvenování kódujících sekvencí 25 imortalizovaných buněčných linií, které vznikly za působení rozličných mutagenů ukázalo, že tento systém dokáže rekapitulovat vzorce mutací nalezené v lidských nádorech. Tyto buněčné linie také vykazovaly mutace v řadě genů důležitých pro vznik rakoviny u člověka a genů účastnících se epigenetické regulace. Skórovací systém, vyvinutý v rámci této práce, určil jako možné geny podporující vznik nádorů geny známé (např. Tp53 a Hras), ale i geny, jejichž vliv na vznik nádorů u člověka dosud nebyl zkoumán, jako je Smarcd2, kódující podjednotku komplexu BAF regulujícího chromatin. Použití molekulárního inhibitoru ukázalo, že MEF buněčná linie s mutací Smarcd2 je závislá na aktivitě komplexu PRC2, což koresponduje s výsledky získanými z lidských buněčných linií s mutacemi dalších podjednotek komplexu BAF. Předložená disertační práce ukazuje, že imortalizované linie z MEF buněk mohou být využity jako účinné modely pro studium důležitých aspektů vzniku nádorů.

Klíčová slova: mutace, vzorce mutací, mutagen, onkogen, tumor supresor, Ras, BAF

ABSTRACT (EN)

Humans and cells in their bodies are exposed to various mutagens in their lifetime that cause DNA damage and mutations, which affect the biology and physiology of the target cell, and can lead to the expansion of an immortalized cell clone. Genome-wide massively parallel sequencing allows the identification of DNA mutations in the coding sequences (whole exome sequencing, WES), or even the entire genome of a tumour. Mutational signatures of individual mutagenic processes can be extracted from these data, as well as mutations in genes potentially important for cancer development ('cancer drivers', as opposed to 'passengers', which do not confer a comparative growth advantage to a cell clone). Many known mutational signatures do not yet have an attributed cause; and many known mutagens do not have an attributed signature. Similarly, it is estimated that many cancer driver genes remain to be identified. This Thesis proposes a system based on immortalization of mouse embryonic fibroblasts (MEF) upon mutagen treatment for modelling of mutational signatures and identification and testing of cancer driver genes and mutations. The signatures extracted from WES data of 25 immortalized MEF cell lines, which arose upon treatment with a variety of mutagens, showed that the assay recapitulates the signatures of these compounds found in human tumours. The cell lines also harboured numerous mutations in genes known to act as cancer drivers in certain contexts, as well as mutations in a list of genes implicated in regulation of the epigenome. A scoring system devised for this study identified multiple putative drivers of the cancer-like phenotype of the cell lines, both well-known drivers (Tp53, Hras) as well as yet unrecognized putative ones (Smarcc1, Smarcd2 subunits of the BAF chromatin remodeling complex). Experiments using a small molecule inhibitor showed that the Smarcd2 mutation is likely to create a dependency of the affected cells on the PRC2 complex, as was previously demonstrated for other mutations in the BAF complex subunits in human cancer cell lines. In summary, the data presented in this Thesis show that the MEF cell lines are an invaluable resource for studies of certain aspects of human cancer development.

Keywords: mutations, mutational signature, mutagen, cancer driver, Ras, BAF

LIST OF ABBREVIATIONS

AA	Aristolochic acid
AF	Allelic fraction, allelic frequency
AFB1	Aflatoxin B1
AID	Activation-induced cytidine deaminase
B[a]P	Benzo[a]pyrene
BWA	Burrows-Wheeler Aligner
COSMIC	Catalogue of Somatic Mutations in Cancer
DAVID	The Database for Annotation, Visualization and Integrated Discovery
DKFZ	German Cancer Research Center
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DTT	DL-Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
GO	Gene Ontology
HCC	Hepatocellular carcinoma
HCl	Hydrochloric acid
HMEC	Human mammary epithelial cells
Hupki	Human p53 knock-in
IARC	International Agency for Research on Cancer
ICGC	International Cancer Genome Consortium
IPA	Ingenuity Pathway Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
MEF	Mouse embryonic fibroblast
MNNG	N-methyl-N'-nitro-N-nitrosoguanidine

NaCl	Sodium chloride
NaOH	Sodium hydroxide
NMF	Non-negative matrix factorization
NP-40	Nonidet P-40
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
qPCR	Quantitative polymerase chain reaction
P/S	Penicillin/streptomycin
Ras-wt	Ras wild type
ROS	Reactive oxygen species
SBS	Single base substitution
SDS	Sodium dodecyl sulfate
SNP	Single nucleotide polymorphism
TBS	Tris buffered saline
TCGA	The Cancer Genome Atlas
UTUC	Upper tract urothelial carcinoma
UV	Ultraviolet light
UVC	Ultraviolet light class C
WES	Whole exome sequencing
WGS	Whole genome sequencing

GLOSSARY

Cancer driver

event – any change in the biology of a cell clone which confers a comparative growth advantage to the clone.

gene – a gene, the function of which, when altered (e.g. by a mutation or an epigenetic modification), contributes to malignant development of a cell clone.

mutation – a mutation which confers a comparative growth advantage to a cell clone.

Genome-wide sequencing – sequencing of the whole genome or whole exome, as opposed to targeted sequencing, which only focuses on sequencing of specific amplicons.

Massively parallel sequencing – any of the techniques which allow high-throughput DNA sequencing (also called next generation sequencing). The most commonly used techniques are based on the sequencing-by-synthesis approach, where the added base is detected either by emission of fluorescent signal (Illumina), or emission of a proton and subsequent change of pH (Thermo Fisher Scientific – previously Life Technologies, this principle is also known as ion semiconductor sequencing).

Mutation

germline – variation in the DNA of the cells of the germ line. Germline mutations are transmitted to all cells of the offspring.

somatic – variation in the DNA which takes place in a somatic cell. Somatic mutation is present only in a specific cell clone.

Mutation spectrum – frequency of individual mutation types in a certain sample (cell line, tumour, collection of mutations in the *TP53* gene, etc.). It is frequently displayed either simply as the proportion of the individual mutation types (Fig. 7), or including the information on 3-nucleotide sequence context (Fig. 12). Importantly, the mutations are not subjected to any dimension reduction method before plotting. Compare with Mutational signature.

Mutational signature – a profile of mutations introduced by a specific mutagenic process. It is displayed as the frequency of 6 mutation types in 16 3-nucleotide sequence contexts (Fig. 9). Even in well-controlled laboratory conditions, it is difficult to ensure that only a single mutagenic process operates in the cells of interest. Therefore, the mutational signatures are extracted by a mathematical approach. Compare with Mutation spectrum.

Oncogene – a gene, the activity of which supports the oncogenic transformation of a cell clone. Oncogenes are usually overexpressed (*MIR155HG*, *MYC*), or bear mutations which confer a new function to the resulting protein, or lock it in an active state (*BRAF*, *RAS* genes).

Passenger mutation – a mutation present in a cell clone, but not conferring a growth advantage.

Sequence context – bases on 5' and 3' of the base of interest. Most frequently used is the 3-nucleotide sequence context, which constitutes of 1 base on 5', the base of interest, and 1 base on 3'. For example: 5'-GCT-3' is a 3-nucleotide sequence context for the underlined base C.

Senescence – biological aging. On cellular level, it is marked by the loss of the ability to divide due to various stress (DNA damage due to reactive oxygen species, shortening of telomeres, or other causes). Senescent cells often acquire a specific secretory phenotype and chromatin changes.

Transcriptional strand bias – proportion of a certain mutation type on a coding vs. non-coding DNA strand.

Tumour suppressor gene – a gene which, when inactivated, permits malignant development of a cell clone. The inactivation is often done on genetic (missense, nonsense, splice-site mutations) or epigenetic (transcriptional silencing) level.

1. INTRODUCTION

1.1. The origins of cancer

Cancer is a leading cause of mortality, accountable for 15% of deaths worldwide (Torre et al., 2015). It is characterized by uncontrolled proliferation of cells which do not respect normal tissue organization and can invade distant sites in the body. Cancer is in fact a group of more than 100 diseases which can originate from various cell types, have diverse risk factors as well as epidemiological and clinical characteristics. However varied the cancer types are, they all originate from a cell, the genetic information of which has been damaged due to innate processes or environmental mutagens, leading to expansion of an immortal cellular clone with specific pathological phenotype.

The hypothesis that mutations in genetic information cause cancer, or the somatic mutation theory, has been formulated in the early 20th century. It was proposed:

- a) that more than one mutation is needed for a cell to become malignant, which is consistent with the notion that cancer incidence increases in higher age groups, and
- b) that not any mutation is malignant, but only that which contributes to more efficient cell propagation (NORDLING, 1953 and references therein).

It was already recognized by then that the malignant process can be induced, or accelerated, by mutagenic agents. Pott's observation that soot is a causing agent of cancer was followed by experimental induction of carcinomas in rabbits treated with coal tar, by Yamagiwa and Ichikiwa at the beginning of the 20th century (Fujiki, 2014). The role of environmental agents such as smoking, asbestos, or aflatoxins for cancer development has been then demonstrated by epidemiological studies (DOLL and HILL, 1950, DOLL, 1955, Sporn et al., 1966, BARNES and BUTLER, 1964, Alpert et al., 1968), inspiring the formation of the program of evaluation of carcinogenic risks to humans (also known as the Monographs program) at the World Health Organization's International Agency for Research on Cancer (IARC). Another line of research supported the role of heritability for tumour formation (Knudson, 1971, Nielsen et al., 2016). Recently, it was proposed that replication errors are

primarily critical for a cell to become malignant (Tomasetti and Vogelstein, 2015). The discussion about the proportion of cancer cases attributable to the role of replication errors vs. environmental risk factors and heritability is ongoing (Tomasetti et al., 2017, Wild et al., 2015, Wu et al., 2015). The various reports estimate that environmental risk factors are responsible for 60-90 % of cancer cases.

Identification of deoxyribonucleic acid (DNA) as the molecular substance of genetic information (Avery et al., 1944) and determination of its structure (WATSON and CRICK, 1953) allowed investigation of molecular mechanisms of carcinogenesis. It was shown that chemical carcinogens form adducts on DNA (Harris et al., 1974), and that innate metabolic and repair pathways also play role in oncogenic transformation (Lindahl and Nyberg, 1972, Lindahl and Andersson, 1972, Loeb et al., 1974, Cleaver, 1968). The discovery of cell-transforming mutation of *Hras* proto-oncogene in chemically-induced mouse skin carcinomas (Balmain and Pragnell, 1983) ultimately placed the alterations in DNA on the interface between mutagenic processes on one side and cancer biology on the other, and stimulated multifaceted research of the links between mutagenesis and cancer pathophysiology.

1.2. Alterations in DNA as a record of mutagenic processes

1.2.1. Mutation spectra of the *TP53* gene reflect cancer aetiology

Mutations in tumour DNA are a result of mutagenic processes operative during the tumour's lifetime (Fig. 1). Most mutations are so called somatic mutations, originating and present in a specific somatic cell clone and are not inherited, as opposed to germline mutations which are present in the cells of the germ line and are thus present in all cells of the offspring. Different mutagens leave distinct mutation patterns on DNA, which depends of the mechanism of their action. The most common and easily detected mutations are single base substitutions (SBS). In a classic work, Hollstein *et al.* (Hollstein et al., 1991) compiled mutation spectra of SBS found in *TP53* gene in human cancers. *TP53* gene produces a transcription factor which plays a crucial role in regulating cell survival, senescence and apoptosis (Biegging et al., 2014). Mutations which impair *TP53* function are thus

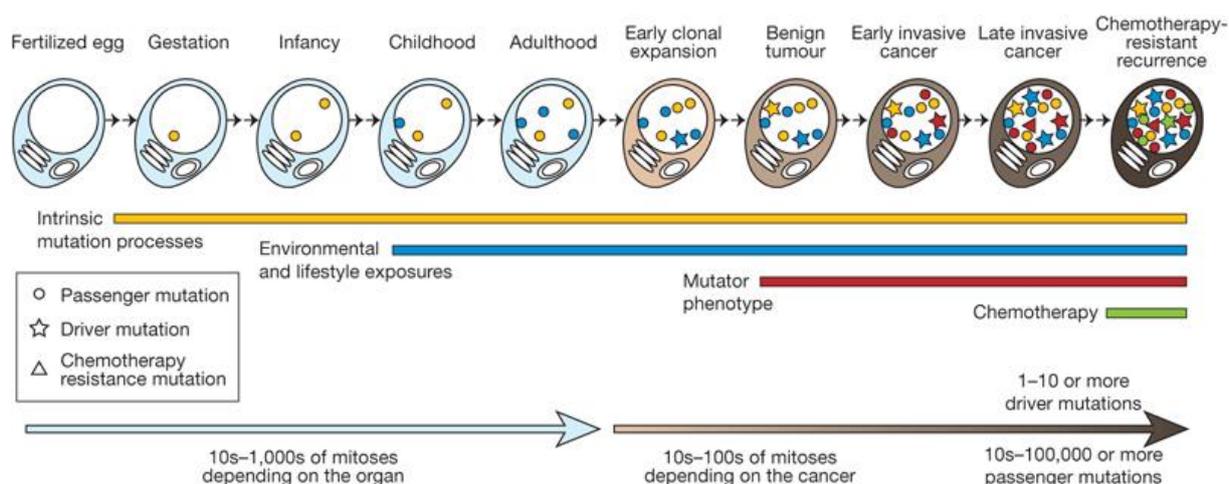


Figure 1: Intrinsic and environmental processes operate in cells and cause mutations in genes driving cancer development. During their lifetime, humans and cells in their bodies are exposed to a number of influences that cause mutations in the DNA. These influences can be intrinsic, such as replication and transcription, or environmental, such as smoking, diet or some traditional and modern medicines. Most of the mutations that cells acquire do not increase a cell's fitness, they are called 'passenger mutations'. However, some of the mutations confer a comparative growth advantage that lead to clonal expansion of a cell and formation of a tumour. These mutations are called 'cancer driver mutations' and genes affected by the mutations are called 'cancer driver genes' or just 'drivers'. It is proposed that different drivers operate in different stages of tumour development. Reproduced figure (Stratton et al., 2009).

common and widespread alterations in cancer (Hollstein et al., 1994, Bouaoun et al., 2016). Importantly, types of SBS in the *TP53* gene differed between tumour types with distinct aetiology. For example, lung tumours from smokers had high prevalence of G>T mutations, which was not the case for tumours from non-smokers. Similarly, hepatocellular carcinomas (HCC) from Chinese patients from Qidong region, had typically G>T mutations in the *TP53* gene, as opposed to the HCCs from Japanese patients (Hollstein et al., 1991). These findings were in line with the mechanisms of action of the principal mutagenic exposures in the specific cancer types – polycyclic aromatic hydrocarbons such as benzo[a]pyrene (B[a]P) for smoking-related lung cancer, and aflatoxin B1 (AFB1) for liver cancer in Qidong region (Sun et al., 1985). More recently, mutation spectra of *TP53* gene revealed high proportion of A>T mutations in upper tract urothelial carcinomas (UTUC) from patients with exposure to aristolochic acid, as validated by aristolactam adduct analysis in the renal cortex (Grollman et al., 2007, Jelaković et al., 2012, Schmeiser et al., 2012, Olivier et al., 2012). These results supported the conclusion that aristolochic acid (AA), the nephrotoxic compound found in the plants of the genus

Aristolochia, contribute to UTUC development. *TP53* variants in cancer are catalogued in the IARC p53 database (Bouaoun et al., 2016, Hollstein et al., 1994) which constitutes a useful resource for gaining insights into the origins of cancer via *TP53* biology.

1.2.2. Genome-wide signatures of mutational processes operative in cancer

The advances of massively parallel sequencing technology inspired efforts in sequencing of human tumours. Large amounts of data on somatic mutations in cancer are now accessible in dedicated repositories such as the Catalogue of Somatic Mutations in Cancer (COSMIC), The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). Recently, Alexandrov et al. (Alexandrov et al., 2013b) developed an algorithm to extract the combination of mutations typical for a mutational process from the catalogues of somatic mutations in tumours, using non-negative matrix factorization method (NMF). The algorithm reflected the notions that a specific mutagen or innate mutagenic process usually acts on a specific base and may prefer a certain sequence context. The method therefore divides substitutions to 6 different classes based on the pyrimidine of the mutated Watson-Crick pair (C>A, C>G, C>T, T>A, T>C, T>G) and includes information on the bases right next to the 5' and 3' end of the mutated base (5'-NNN-3', the mutated base is underlined). The combination of 6 mutation classes with 16 possible 3-nucleotide sequence contexts generates 96 categories of mutations, providing good amount of detail to distinguish mutagenic processes which generate mutations of the same type but in different sequence context. Finally, the algorithm also included the information on so called strand bias, ie. the difference in proportion of mutations in transcribed vs. non-transcribed strand. Strand bias is typically generated when there is a bulky adduct on the DNA. The adduct obstructs the transcription of a gene and is removed by transcription-coupled repair machinery on the transcribed strand. However, it is not removed as efficiently on the non-transcribed strand and leads to accumulation of more mutations on that strand.

The method was applied on somatic mutation data from more than 7,000 tumours from 30 cancer types and identified 21 patterns of mutations which were termed

'mutational signatures' (Alexandrov et al., 2013a). The analysis was then expanded, using data from more than 12,000 tumours of 40 cancer types, and identified 30 mutational signatures which are recorded as a reference on a web site within the COSMIC project (URL1).

A considerable number of the 30 signatures display high fraction of C>T (9 signatures), C>A (5 signatures), or both (2 signatures) mutation types. This probably reflects high prevalence of mutations due to 5'-methylcytosine deamination (C>T) and oxidation of guanine (C>A, or G>T on the other strand). Other signatures either display high proportion of another mutation type (e.g. signature 12: T>C, signature 22: T>A), or display a characteristic combination of multiple mutation types in specific sequence contexts (e.g. signature 5, signature 3).

The authors also attributed signatures to possible aetiological agents, based on known mechanisms of action of the agents and based on cancer type from which the signatures were typically extracted. Six mutational signatures were attributed to environmental mutagens: tobacco smoking and chewing (signatures 4 and 29), ultraviolet (UV) light exposure (signature 7), exposure to alkylating agent temozolomide (signature 11), aristolochic acid exposure (signature 22) and aflatoxin exposure (signature 24). Ten signatures were attributed to innate mutagenic processes, such as spontaneous deamination of 5-methylcytosine (signature 1, so called the 'age signature') or altered activity of error-prone polymerase POLE (signature 10). Interestingly, five signatures were attributed to defective repair machinery (signatures 3, 6, 15, 20, 26) and three were associated with the activity of APOBEC family enzymes (signatures 2, 9, 13). Importantly, the aetiology of 14 signatures was not attributed to any mutagenic process.

The work of Alexandrov et al. marked a new era, in showing how the vast and ever growing amount of somatic mutation data can be utilized to learn about the origins of cancer. Indeed, further development of mutational signatures analysis already provided results. For example, finding aristolochic acid signature in renal cell carcinomas from the Balkan countries and bladder and hepatobiliary carcinomas from China redefined the problem of aristolochic acid exposure with regards to new

cancer types (Scelo et al., 2014, Jelaković et al., 2015, Poon et al., 2013, Poon et al., 2015, Zou et al., 2014). Thorough analysis of mutational signatures in various cancer types associated with smoking indicated that in organs which are not directly affected by combustion products, such as bladder, cervix, kidney and pancreas, smoking seems to induce more general mutational processes, like the activity of APOBEC enzymes (Alexandrov et al., 2016), thus revealing another link between smoking and cancer. Lately, signature 18 was attributed to impaired base excision repair in colorectal carcinoma (Pilati et al., 2017, Viel et al., 2017) as well as prolonged culture times when studied in human organoids (Blokzijl et al., 2016). However, many carcinogens with epidemiological and molecular evidence do not yet have attributed signatures, and many signatures do not have a proposed aetiology. Some of the aetiologies which were attributed to certain signatures have been experimentally validated (Zámborszky et al., 2017, Chan et al., 2015, Segovia et al., 2015). Systematic experimental approach is needed to provide the explicit link between a mutagenic process and a mutational signature.

1.2.3. Modelling mutational spectra and signatures using experimental systems

1.2.3.1. Single and reporter gene approaches

Before the invention and spreading of massively parallel sequencing, the efforts to define mutation spectra of carcinogens were directed towards a single gene or reporter gene approaches.

The reporter gene techniques typically depended on the use of a shuttle vector and a gene, such as lacZ or gpt or cplII, which enabled the selection of colonies based on colour or viability. The reporter gene DNA was either treated *in vitro*, or integrated in multiple copies to genomes of transgenic animals, mice or rats, which were then treated with a mutagen *in vivo*. The DNA was then extracted and packed to phage vectors which were transduced to a bacterial host with the adequate genotype for selection. Colonies with mutated reporter gene were then selected and the amplicon corresponding to the gene of interest was sequenced. This approach generated experimental mutation spectra for example for aflatoxin B1, acrylamide, second hand smoke, or 8-methoxypsoralen (Besaratina et al., 2012, Manjanatha et

al., 2015, Levy et al., 1992, Trottier et al., 1992, Wattanawaraporn et al., 2012, Sage et al., 1993).

The single gene approach was again centred on the *Tp53* gene. Experimental *Tp53* mutation spectra were generated using mouse embryonic fibroblasts (MEFs) from Hupki (human p53 knock-in) animals designed to study the alterations occurring in the human *Tp53* orthologue. Unlike human cells, MEF possess long telomeres and express telomerase (Ohtani et al., 2012). Though MEFs cultivated in atmospheric oxygen levels enter oxidative stress-induced senescence, characterized by accumulation of *Tp53*, *p21* and other components of the *Tp53* axis (Odell et al., 2010), they immortalize easily due to the absence of replicative senescence barrier. Molecular mechanisms of senescence bypass in MEFs commonly involve alterations in *p53* pathway, notably mutations in the *Tp53* gene (Odell et al., 2010). Hupki mice were genetically engineered to replace a portion of the murine *Tp53* gene between exons 4 and 9, which is the most frequently mutated region in cancer (URL2), with corresponding human *TP53* sequence (Luo et al., 2001b). This genetic modification did not affect the development of Hupki mice, the liver tumour response or *Tp53* response to DNA damaging agents (Whibley et al., 2010, Jaworski et al., 2005, Luo et al., 2001b) while providing an excellent model to study mutation spectra of carcinogens with direct applicability to the human *TP53* biology. Hupki MEF assay was useful in supporting the knowledge of mutation spectra of both well-known and less studied compounds (Besaratina and Pfeifer, 2010, vom Brocke et al., 2006) and also allowed to study the effect of mutations in the *Tp53* gene on its function (Odell et al., 2013). Single/reporter gene assays were helpful in gaining insight into the actions of mutagens on the molecular level. However, their main disadvantage lies in relatively small scale and thus low resolution and relatively high cost, also connected to the need of animal facilities.

1.2.3.2. Genome-wide model systems

More recently, the efforts were directed to generate experimental genome-wide mutation spectra and mutational signatures. Lower model organisms – mostly yeast and worm – have been successfully utilized. It is important to note that sequencing at standard 30× coverage requires a relatively clonal sample, as otherwise mutations

are not reliably detected (Fig. 2). Yeast and worm-based assays involve bottlenecks – plating a yeast cell or a single worm on a new plate. These assays have been helpful particularly in investigating the effects of various mutants in DNA repair pathways and other innate mutagenic processes. For example Chan et al. showed in their elegant work that experimental mutagenesis in yeast discriminates between APOBEC3A and APOBEC3B due to different preference for base on the position -2 from the mutated base (Chan et al., 2015). Application of this knowledge to sequencing data from human tumours led to the finding that APOBEC3A is probably the main APOBEC enzyme operating in human cancer.

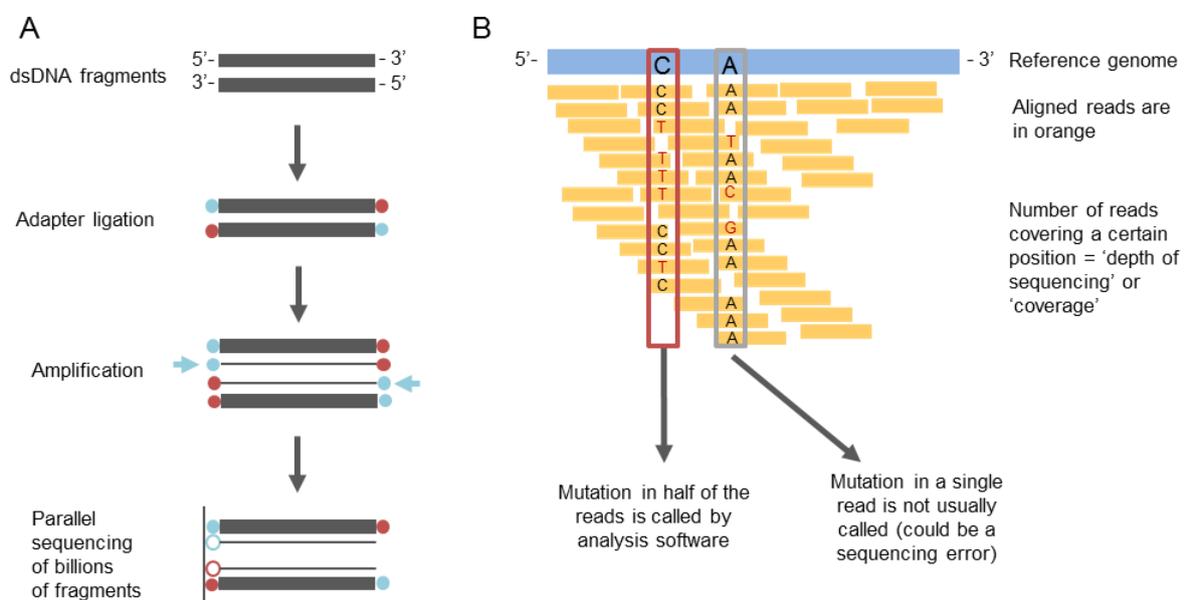


Figure 2: Principles of massively parallel sequencing. A – preparation of sequencing library and production of sequencing reads. DNA is fragmented and adaptor and other sequences are ligated. Library is then amplified and applied on a sequencing chip. There, the sequences are immobilized and subjected to a polymerase chain reaction (PCR). Extension of the fragment is recorded, either by the emission of a specific fluorescent signal by an added nucleotide, or due to the release of a proton during the addition of a nucleotide to the sequence (semiconductor sequencing). Technology of semiconductor sequencing currently does not allow sequencing from both ends of the DNA fragment (so called paired-end sequencing, which allows more downstream analyses). B – read alignment and variant calling. Sequences that are recorded during the sequencing are called reads. They have to be firstly aligned to a reference genome build, then variants can be called. It is important to note that variants in tumours are usually not called against the reference genome, but against a normal sample from the same individual ('tumour-normal pair') to avoid calling single nucleotide polymorphisms and germline mutations in somatic variant calling analysis.

The activities of various genes and enzymes, as well as sequence representations, may be different between yeast, worm and human. Indeed, there have been attempts to use human cells to model mutational signatures. Poon et al. treated human proximal tubule cell line HK-2 with sublethal dose of 10 μ M AA for 6 months (Poon et al., 2013). Two clones that emerged from the treated culture were sequenced and displayed a T>A-rich mutational signature, identical to that extracted from UTUC with AA aetiology. The results, though very informative in terms of mutational signatures, were based on a lengthy and laborious experimental effort. Also, the model was an already transformed human cell line, so the results have a limited applicability for biological effects of the mutations.

1.3. Alterations in DNA as the causes and effectors of tumour physiology

1.3.1. Drivers and passengers

Evolution theory teaches us that entities with higher fitness are selected over other entities in the population, and that mutagenesis is central for generating these differences. From a gene-centred view, the genes, or their variants, which are apt in reproducing themselves, are selected for during the evolution process.

Populations of cells in our body are subjected to intrinsic and environmental mutagenic processes. Some mutations can confer a selective growth advantage to a cell clone, which can lead to its expansion and result in development of a tumour (Fig. 1). The mutations that confer a selective growth advantage are called cancer driver mutations and genes bearing such mutations are referred to as cancer driver genes. In fact, majority of mutations, both somatic and germline, do not have an effect on cell fitness. These are called passenger mutations. Some driver genes are typically affected by gain-of-function mutations, and an acquired activity of these genes is important for cancer development. These driver genes are called oncogenes. Other genes, on the contrary, tend to accumulate inactivating mutations, and impaired function of these genes makes it easier for a cell clone to form a tumour. These driver genes are called tumour suppressor genes. Over 600 genes were found to be implicated in cancer so far and are now included in the Cancer

Gene Census, a manually-curated database of cancer driver genes (Futreal et al., 2004).

Cancer develops in a multistep process and a cell typically needs to acquire several driver mutations in order to give rise to a tumour. Different mutations may be needed for the early clonal outgrowth, for specific changes in metabolism, vascularization, invasion and, ultimately, resistance to therapy (Fig. 1). Along the way, the cell clone also acquires passenger mutations which are usually much more abundant than the driver mutations. Discriminating drivers from passengers is one of the main interests of cancer research.

Many key drivers that are frequently mutated in various cancer types, such as the *RAS* genes, *BRAF* or *TP53*, were identified based on experimental approaches such as cloning and cell transformation assays. However, genes with lower mutation frequencies can also shift normal cells towards a cancer phenotype. One challenge that remains is the identification of driver genes that are mutated with low frequency (Lawrence et al., 2014).

1.3.2. Identification of cancer driver events using sequencing data

Massively parallel sequencing allows identifying the totality of mutations in a tumour. Cancer genomes typically contain tens to thousands of mutations in protein-coding genes, but only a fraction of them drives the tumour development. Natural selection favours cells that carry functional mutations in cancer driver genes. Therefore, these genes are expected to be found mutated in cancer with certain level of recurrence. In fact, such genes should be more frequently mutated than expected by the background mutation rate, adjusted for many variables (gene size, sequence context, replication timing, gene expression, etc.). There are also other signs of positive selection that can point out possible cancer driver genes. For example, if a gene is an oncogene, activating mutations tend to be clustered in 'hotspot' regions of the gene and its corresponding protein product, while in tumour suppressor genes, inactivating mutations tend to be distributed along the entire length of the gene (Fig. 3). The standard '20/20 rule' requires, for a gene to be classified as an oncogene, that at least 20 % of the recorded mutations in the gene

are at recurrent positions and are missense. For a tumour suppressor gene, at least 20 % of the recorded mutations must be inactivating (Vogelstein et al., 2013). Cancer driver mutations can also be located outside the protein-coding sequence, in regulatory regions like promoters and enhancers. Databases exist that collect information on transcription factor binding sites, and these are used to annotate single nucleotide variants found in tumours (Bryne et al., 2008, Boyle et al., 2012) and to evaluate their effect (Hoffman and Birney, 2010, Pleasance et al., 2010). However, identification of mutations in regulatory regions requires sequencing of the whole tumour genome (WGS, whole genome sequencing), while identification of mutations in protein-coding regions only requires sequencing of the collection of exons (WES, whole-exome sequencing). Coding sequences cover ~2 % of human genome; WES is thus faster and cheaper, and less demanding in terms of data processing power and storage space. Due to these advantages, large amounts of tumours have been sequenced on the exome level and ample data exist on driver mutations in protein-coding genes. Coding sequences are also rather well annotated, compared to the so called ‘non-coding genome’. For these reasons, I will focus on mutations in protein-coding genes in this overview.

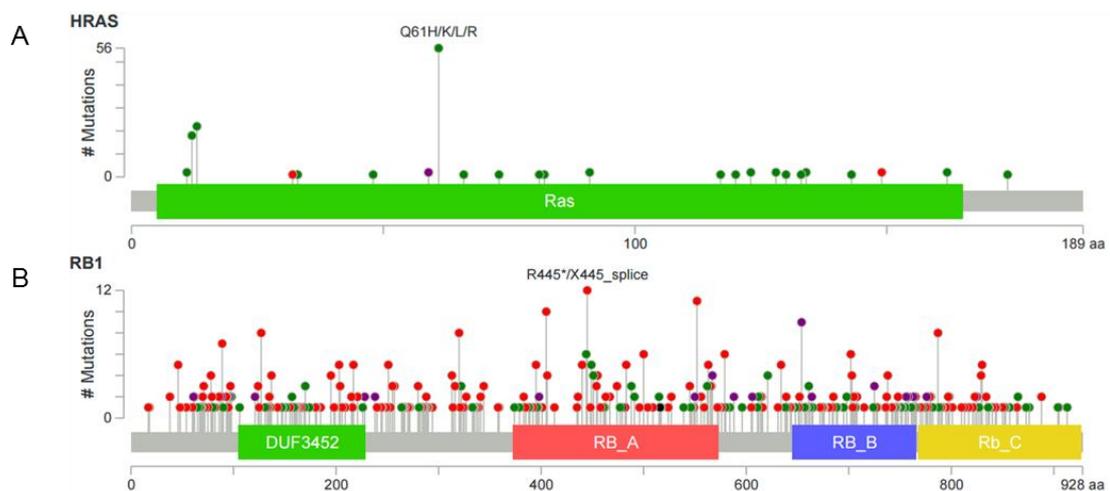


Figure 3: Distribution of mutations in cancer driver genes. Mutations in oncogenes are activating and show clustering in ‘hotspots’, while mutations in tumour suppressor genes are inactivating and distributed alongside their sequence. A - HRAS as an example of an oncogene with “hotspot” at the positions 12, 13 and 61 of the protein sequence. B - RB1 as an example of a tumour suppressor. Most of the mutations are nonsense, i.e. generating a premature stop codon, or mutations of splice sites. Green dots – missense mutations, red dots – nonsense and splicing mutations, purple dots – missense and nonsense mutations, black dots - deletions. Images generated from the TCGA data on cBioportal (Cerami et al., 2012, Gao et al., 2013).

1.3.2.1. Bioinformatic approaches to driver identification

Numerous bioinformatic methods were developed to take advantage of the large amounts of data which are readily available in public repositories, for driver gene identification (Lawrence et al., 2013, Tamborero et al., 2013a, Dees et al., 2012, Davoli et al., 2013, Reimand and Bader, 2013, Reimand et al., 2013). The algorithms were based on various concepts for identification of driver genes, as described above.

Applying the methods on human tumour sequencing data identified many new putative cancer driver genes; however, the problem was the reproducibility. For example, Lawrence *et al.* (Lawrence et al., 2014) analysed data from 4,742 tumours of 21 cancer types using MutSig methods based on the background mutation rate. They identified 219 cancer driver genes, using stringent criteria. Subset of the sample set, specifically 3,205 tumours of 12 cancer types, was earlier analysed by Tamborero *et al.* (Tamborero et al., 2013b) using MuSic (based on the background mutation rate) and Oncodrive (based on mutation clustering and functional bias) methods. They identified 291 high-confident drivers. One would expect that the drivers identified by both approaches would overlap to a large extent, since the datasets which were analysed were also highly overlapping. That was not the case, though. Only 109 genes were predicted as drivers in both analyses, and 80 of these genes overlapped with the Cancer Gene Census. Thus, the methods seemed good in identifying known, frequently-mutated drivers, but in addition discovered many separate candidate drivers which may, or may not, be true positives.

Recently, Tokheim et al. developed a framework for evaluation of the driver prediction methods (Tokheim et al., 2016). They tested 8 different prediction programs and compared those using metrics such as the number of significant genes, overlap of the results with each other and with the Cancer Gene Census, observed vs. expected p-value distribution, consistency of the methods, and others. They found that ratiometric methods, i.e. those based on evaluation of functional impact of a mutation, performed overall better than methods based on the background mutation rate. The reason is that features like mutation clustering

or functional impact of mutations on a gene are less variable than background mutation rates among cancer types, or even individual tumours (Lawrence et al., 2013).

1.3.3. Cancer driver pathways and complexes

Mutations drive cancer development because they affect gene products that are part of biological pathways, which, when destabilized, promote cancer development. Deregulation of biological processes and pathways can be thus considered the ultimate cancer driver event. Well-established pathways implicated in cancer development include (Vogelstein et al., 2013, Hanahan and Weinberg, 2011):

Signalling pathways involved in cell growth and proliferation (MAPK, PI3K, ErbB, TGF- β , etc.). Kinases in these pathways frequently harbour activating mutations that cause constitutive signal transduction, while phosphatases tend to accumulate inactivating mutations that abolish their inhibitory functions.

Signalling pathways involved in cell fate and differentiation (NOTCH, Hedgehog, etc.). Cancer-related mutations in these pathways shift the balance between differentiation and cell division favouring an undifferentiated, proliferative phenotype.

Cell cycle, senescence and apoptosis. Regulators of these pathways are frequently altered, allowing uncontrolled cell proliferation.

DNA repair. Defects in DNA repair pathways contribute to aggressive phenotype, due to increased mutation rates, while residual DNA repair activity is necessary to prevent elimination of cancerous cells. Therefore, distinct DNA repair pathways are frequently found inactivated in cancer cells (Helleday et al., 2008).

Some of the best-characterized cancer driver genes – *TP53*, *BRCA1*, *KRAS*, *RB1*, and others – function within these pathways, and their contribution to tumour development is rather well studied. In contrast, other pathways may not contain chief cancer driver genes while still widely affected in cancer. Such emerging cancer-related pathways are that of RNA processing (Sveen et al., 2015) and that of epigenetic modifications and chromatin organization (Feinberg et al., 2016).

Regulators of the epigenome are of particular interest. Comprehensive analysis of data from 4,623 tumours revealed frequent mutations of genes within BAF complex, PRC1 complex, and other complexes involved in regulation of chromatin, with the tendency to mutual exclusivity in the mutations of individual subunits (Gonzalez-Perez et al., 2013). This was later confirmed by an independent analysis with HotNet2 algorithm (Leiserson et al., 2015). HotNet2 evaluates both the mutations in single genes as well as the topology of interactions among the encoded proteins. Indeed, the BAF network was found significantly mutated, and mutations in the BAF subunits had a tendency towards mutual exclusivity indicating that a mutation of a single subunit is sufficient for destabilization of the complex. The algorithm identified also other new networks which had not yet been linked to cancer – for example the cohesin and condensin complexes, which are implicated in chromatin cohesion and condensation during mitosis, but also more broadly in gene regulation (Peters et al., 2008, Hirano, 2012). These networks are universally mutated across cancer types with frequency ~4 %, but the individual genes within the networks are mutated at much lower frequency and were mostly neglected by gene-centred approaches. The results prove the utility of pathway-based approaches for discovery of putative cancer driver events.

2. INTRODUCTION TO THE THESIS

Research to-date supports the view of mutations as central events in cancer development, which can provide information about environmental exposures and innate processes which act as mutagens, and also permit to identify and study the alterations in cellular physiology which lead to formation of a tumour with its specific characteristics.

Humans are exposed to various influences during their lifetime. Hence, the combinations of mutations in tumours are results of all these effects. Also, tumours are often removed and sequenced relatively late into their development, making it more complicated to identify drivers of distinct stages of the tumour evolution. Therefore, improving models of cancer development is of particular importance.

Human primary cells should ideally be the model system for recapitulating tumour evolution *in vitro*. It is essential to understand that tumour cells have to immortalize, because formation of a tumour requires many more cell divisions than the Hayflick limit permits (Armstrong and Tomita, 2017). However, generation of immortal human cell lines from primary cells (without targeted genetic manipulation techniques) is extremely difficult, particularly due to the presence of the replicative senescence barrier, and has rarely been reported (Stampfer and Bartley, 1985).

Here we propose primary MEF cells as a useful model for cancer development (Fig. 4). As mentioned earlier, primary MEFs cultivated in standard conditions undergo senescence, which is frequently bypassed, resulting in generation of an immortal cell line. Immortalization is not obstructed by Hayflick limit, because MEF cells possess long telomeres and express telomerase, and is driven by mutations in specific genes. The Tp53 pathway is often affected, with the *Tp53* gene itself being mutated in ~25 % of cases (Whibley et al., 2010), similar to the *TP53* mutation rate in human tumours (Olivier et al., 2010). Importantly, it was demonstrated that MEFs exposed to various compounds accumulate mutations in *Tp53*, which are specific to that compound (see chapter 1.2.3.1 for more details). However, not only the *Tp53*, but other genes and pathways – which are also

observed to act in human tumours – were suggested to play a role in MEF immortalization (vom Brocke et al., 2006).

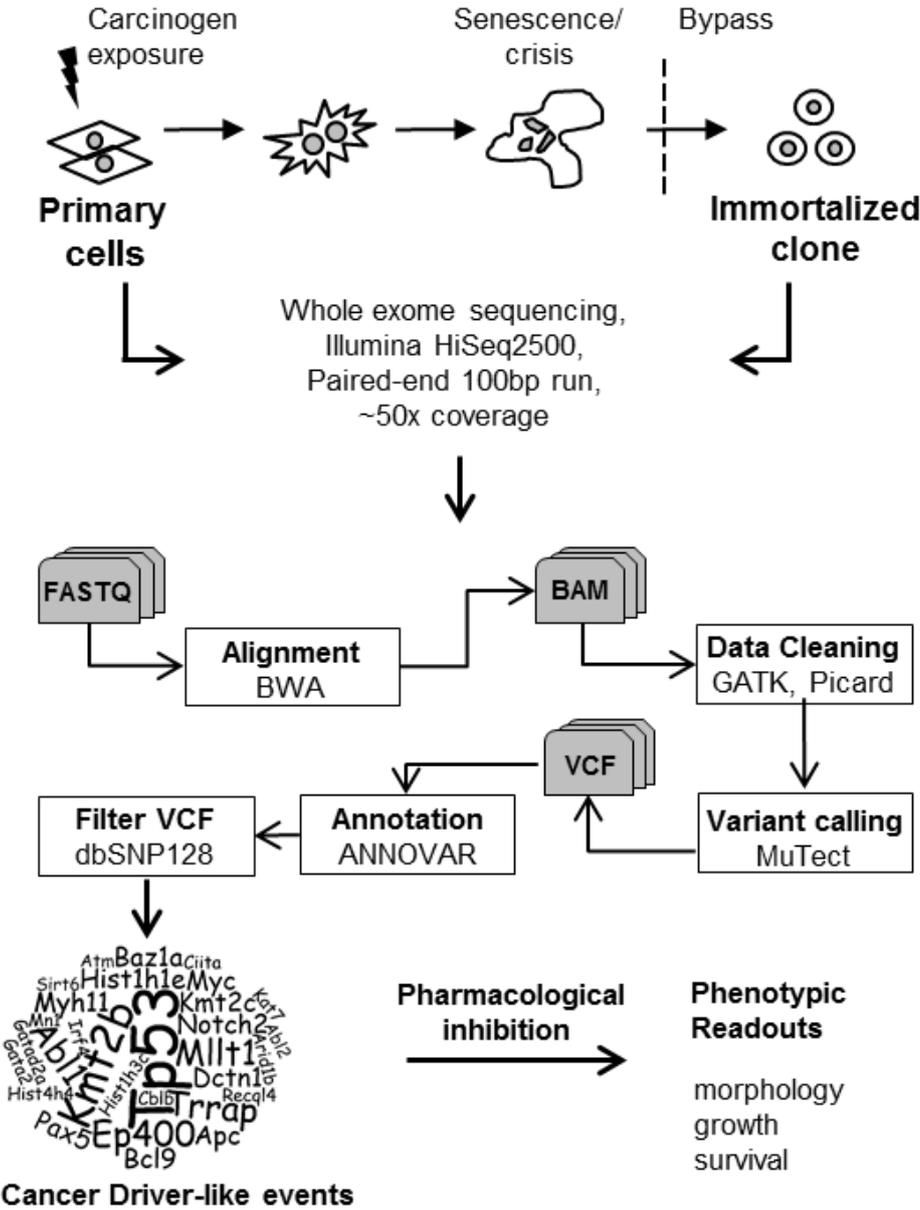


Figure 4: Study design. Mouse embryonic fibroblasts are exposed to a mutagen in an early passage and cultivated until senescence and immortalization. Exome of the resulting cultures is sequenced at ~50× coverage. Data are analysed by the indicated pipeline and the results are used to extract mutational signatures and mine putative driver mutations. Impact of selected mutations is tested using small molecule inhibitors.

2.1. Hypotheses

I hypothesized that the carcinogen exposure and immortalization MEF assay recapitulates features relevant to the activity of used mutagens and selects for mutations that contribute to the cancer-like phenotype of the immortalized cells. Mutations act in a combinatorial manner and each clonal cell line results from the selection of a specific combination of growth-promoting driver mutations and driver genes. These can involve alterations in known, frequently-mutated genes as well as yet uncharacterized events. I further hypothesized that the driver mutations are introduced early in the assay due to the carcinogen treatment and are likely to become components of the carcinogen-specific mutational signature. Driver mutations can thus be identified from the pool of non-synonymous exposure-specific mutations with predicted functional impact, and tested in downstream validation experiments.

2.2. Aims of the Thesis

- a) To generate **mutational signatures** of carcinogens using MEF immortalization assay, in order to recapitulate signatures observed in human tumour sequencing data.
- b) To identify acquired mutations acting as **potential drivers** during immortalization of MEF cells.
- c) To **functionally test** the impact and roles of select candidate driver mutations, both individually and in combination.

2.3. Specific aims of the Thesis

- a) *To generate mutational signatures of carcinogens using MEF immortalization assay, in order to recapitulate signatures observed in human tumour sequencing data.*

Experimental models of mutagenesis are essential to attribute mutational signatures to their causative processes. The specific approach is to sequence exomes of MEF

cells, which were treated in an early passage with a carcinogen with attributed or presumed signature, and were cultured until immortalization (Fig. 4). Sequencing of cells which immortalized spontaneously will reveal the mutation profile linked to the culture conditions, for comparison. The mutational signatures are extracted with the NMF algorithm and compared with COSMIC signatures and other relevant data.

b) To identify acquired mutations acting as potential drivers during immortalization of MEF cells.

Tp53 is a cancer driver gene; it also acquires mutations during the MEF immortalization assay that are characteristic of the compound with which the MEFs were treated (Reinbold et al., 2008, Liu et al., 2004). The aim therefore is to identify – going beyond *Tp53* – potential driver genes from the pool of mutations of the exposure-specific type which have predicted functional impact (non-synonymous, predicted *in silico* as deleterious, affecting a functional domain of a protein, etc.).

c) To functionally test the impact and roles of select candidate driver mutations, both individually and in combination.

Each tumour and each cell line have a unique combination of (driver) mutations which affect specific pathways. Therefore, it is desirable to study functions of mutations on their original mutation background, rather than in completely different settings, as it is often done, to assess their effects. The specific aim is, after identification of mutations of interest, to test their effects individually and in combination, using small molecule inhibitors.

3. MATERIAL AND METHODS

3.1. Material

3.1.1. Cell lines

Twenty-six cell lines were used in the study. Twenty-two cell lines were generated from primary Hupki MEF cells in the laboratory of Dr. Monica Hollstein in the German Cancer Research Center (DKFZ), Heidelberg (Liu et al., 2004, Liu et al., 2005, Nedelko et al., 2009, Feldmeyer et al., 2006). Two cell lines were generated from primary MEFs which were obtained from Hupki mice crossed with transgenic mice expressing activation-induced cytidine deaminase (AID) (Okazaki et al., 2003). The cross was generated in the laboratory of Dr. Hiroyuki Marusawa at the Kyoto University, Japan, and their MEFs were sent to DKFZ, where the actual cell lines were produced. One cell line was generated from primary Hupki MEFs by Hana Huskova in the MMB Group at IARC. Table 1 lists the cell lines and the conditions under which they were generated.

Table 1: Cell lines

Cell line ID	Exposure type	Exposure dose	Exposure duration	Origin
AA_1	AA	50 µM	4 days	DKFZ
AA_2	AA	50 µM	4 days	DKFZ
AA_3	AA	50 µM	4 days	DKFZ
AA_4	AA	50 µM	4 days	DKFZ
AA_5	AA	50 µM	4 days	DKFZ
AA_6	AA	50 µM	12 days	DKFZ
AA_7	AA	50 µM	8 days	DKFZ
AFB1_1	AFB1	2 µM	8 days	DKFZ
AFB1_2	AFB1	2 µM	8 days	DKFZ
AFB1_3	AFB1	2 µM	8 days	DKFZ
AID_1	none	n. a.	n. a.	DKFZ
AID_2	none	n. a.	n. a.	DKFZ
B[a]P_1	B[a]P	1 µM	6 days	DKFZ
B[a]P_2	B[a]P	1 µM	6 days	DKFZ
B[a]P_3	B[a]P	5 µM	2 days	DKFZ
MNNG_1	MNNG	20 µM	2 hours	DKFZ
MNNG_2	MNNG	20 µM	2 hours	DKFZ
MNNG_3	MNNG	20 µM	2 hours	DKFZ
MNNG_4	MNNG	20 µM	2 hours	DKFZ
Spont_1	none	n. a.	n. a.	DKFZ
Spont_2	none	n. a.	n. a.	DKFZ
Spont_3	none	n. a.	n. a.	DKFZ
Spont_4	none	n. a.	n. a.	DKFZ
Spont_5	none	n. a.	n. a.	IARC MMB
UVC_1	UVC	20 J/m ²	n. a.	DKFZ
UVC_2	UVC	20 J/m ²	n. a.	DKFZ

AA - aristolochic acid, AFB1 - aflatoxin B1, AID - activation-induced cytidine deaminase, B[a]P - benzo[a]pyrene, DKFZ - German Cancer Research Center, IARC MMG – Molecular Mechanisms and Biomarkers group, International Agency for Research on Cancer, MNNG - N-methyl-N'-nitro-N-nitrosoguanidine, n. a. – not applicable Spont - spontaneous immortalization, UVC - UV light class C.

3.1.2. Chemicals

This section lists the chemicals used in the study, with their catalogue number and the manufacturer.

<i>Name</i>	<i>Catalogue no.</i>	<i>Company</i>
2-mercaptoethanol	31350010	Life Technologies
Acetic acid	100056	Merck
Advanced DMEM	12491023	Life Technologies
Agarose LE Ultrapure	GEPAGA0765	AbCys Eurobio
Bromphenol blue	805732	ICN Biomedical
cOmplete™Mini Protease Inhibitor Cocktail	11836153001	Sigma Aldrich
Crystal violet	34024	BDH Stains
Dimethyl sulfoxide (DMSO)	D2650-100ML	Sigma Aldrich
DL-Dithiothreitol (DTT)	D5545	Sigma Aldrich
DNA Gel Loading Dye 6x	R0611	Thermo Fisher Scientific
Ethylenediaminetetraacetic acid (EDTA) disodium salt dihydrate	E5134	Sigma Aldrich
Ethanol, absolute	107017	Merck
Fetal bovine serum	10270106	Life Technologies
Fetal calf serum	CVFSVF00-01	Eurobio Abcys
Glycerol	50405	Euromedex
GSK126	M60071-2S	Xcess Biosciences
Gel Red™	41003	Biotium via VWR
Halt phosphatase inhibitor Cocktail	78420	Life Technologies
Hydrochloric acid (HCl)	109063	Merck
Isopropanol	100272	Merck

L-Glutamine	25030024	Life Technologies
Nonidet p40 substitute (NP-40)	11332473001	Roche Life Science
Formaldehyde 37%	104002	Merck
Penicillin Streptomycin (P/S)	15140122	Life Technologies
Phosphate-buffered saline (PBS)	14190169	Life Technologies
Ponceau S	33429	BDH Stains
Protein Assay Dye Reagent Concentrate	5000006	Bio-Rad
Sodium dodecyl sulphate (SDS)	L5750	Sigma Aldrich
Sodium pyruvate	11360039	Life Technologies
Sodium chloride (NaCl)	S7653	Sigma Aldrich
Sodium hydroxide (NaOH)	106462	Merck
Tris-Borate-EDTA 10x	ET020-C	Euromedex
Tris-Glycine-SDS 10x	eu0510	Euromedex
Trizma base	T1503-1KG	Sigma Aldrich
Triton X-100	1610407	Bio-Rad
Trypan blue	15250061	Life Technologies
Trypsin EDTA 10x	15400054	Life Technologies
Tween-20	P7949-100ML	Sigma Aldrich
U0126-monoethanolate	U120-1MBG	Sigma Aldrich

3.1.3. Enzymes, antibodies, DNA and protein ladders

<i>Name</i>	<i>Catalogue no.</i>	<i>Company</i>
Antibody to Ccnd1	NCL-L-CYCLIN D1-GM	Novocastra*
Anti-Histone H3 (tri methyl K27) antibody [mAbcam 6002] - CHIP Grade	ab6002	Abcam

Anti-Histone H3 antibody – Nuclear Loading Control and CHIP Grade	ab1791	Abcam
Anti-rabbit IgG, HRP-linked Antibody	7074	Cell Signaling Technology
Benzonase	E1014-5KU	Sigma Aldrich
Gene Ruler 100 bp DNA Ladder	SM0242	Thermo Fisher Scientific
Gene Ruler 1kb DNA Ladder	SM0314	Thermo Fisher Scientific
Goat Anti-Mouse Immunoglobulins, Polyclonal, HRP	P044701	Agilent Technologies
Mouse anti-actin, monoclonal (clone: C4)	0869100	MP Biomedicals
p44/42 MAPK (Erk1/2) Antibody	9102	Cell Signaling Technology
Phospho-p44/42 MAPK (Erk1/2) (Thr202/Tyr204) Antibody	9101	Cell Signaling Technology
ProSieve™ QuadColor™ Protein Marker	LON00193837	Lonza via Ozyme
RNAse A	EN0531	Life Technologies
Proteinase K solution	part of 55114	Qiagen

* Kindly provided by Christine Carreira, IARC, Lyon, France

3.1.4. Commercial kits

<i>Name</i>	<i>Catalogue no.</i>	<i>Company</i>
CellTiter 96® AQueous One Solution Cell Proliferation Assay (MTS)	G3582	Promega
Clarity Western ECL Substrate	1705060	Bio-Rad
Nucleospin Tissue	74090150	Macherey-Nagel
MycoAlert™ Mycoplasma Detection Kit	LT07418	Lonza
Mini-PROTEAN® TGX™ Precast Protein Gel	4561096 and 4561084	Bio-Rad

HotStarTaq DNA Polymerase	203205	Qiagen
QiAmp DNA Mini Kit	51304	Qiagen
Qubit dsDNA HS Assay Kit	Q32854	Life Technologies
Trans-Blot® Turbo™ Midi PVDF	1704157	Bio-Rad
Trans-Blot® Turbo™ Mini PVDF	1704156	Bio-Rad
DNTP-SET, 4x 250 µl 100mM	033785	Dutscher

3.1.5. Custom buffers

6x Laemmli buffer

0.375M Tris pH 6.8, 12% SDS, 60% glycerol, 0.6M DTT, 0.06% bromophenol blue

Blocking buffer

5% or 3% non-fat dry milk in tris buffered saline (TBS) supplied with 0.1% Tween-20 ('TBS-Tween')

DNA lysis buffer

50 mM Tris-Cl pH8, 100 mM EDTA pH8, 100 mM NaCl, 1% SDS

Fix and stain solution (crystal violet)

4% paraformaldehyde in PBS, 0.005% crystal violet

Lysis buffer for whole cell protein extraction

50 mM Tris pH 7.4, 250 mM NaCl, 0.1% SDS, 0.5% NP-40, 2 mM DTT, protease and phosphatase inhibitors

Lysis buffer for histone extraction

0.5% Triton X-100 in PBS, protease inhibitors

TER buffer

10 mM TrisHCl pH 8 (10 µL of the 1M stock pH 8), 1mM EDTA (2µL of the 0.5M stock pH 8), 0.02 mg/mL RNase A (2 µL), add water to 1 mL, store at +4°C

TBS 10×

0.242 % Trizma base and 0.8% NaCl in water, adjust pH to 7.6 with HCl, use at 1 × diluted with water

3.1.6. Custom oligonucleotides

The oligonucleotides used to generate data for this Thesis were used as sequencing primers for Sanger sequencing. They are listed in Table 2.

3.1.7. Equipment

List of equipment with its manufacturer.

▪ 5424 Microcentrifuge	Eppendorf
▪ BioPhotometer	Eppendorf
▪ ChemiDoc XRS+ System	Bio-Rad
▪ D40 Camera	Nikon
▪ Forma™ Series II 3111 CO ₂ Incubator	Thermo Fisher Scientific
▪ GeneAmp® PCR System 2700	Applied Biosystems
▪ Gibco BRL Gel Electrophoresis System	Life Technologies
▪ HiSeq 2500 System	Illumina
▪ LB913 Apollo Absorbance Reader	Berthold Technologies
▪ NanoDrop™ 8000 Spectrophotometer	Thermo Fisher Scientific
▪ Polystat5 Heated Bath Circulator	BioBlock
▪ Power Supply Model 20013.0	Bio-Rad
▪ Qubit 3.0 Fluorometer	Life Technologies
▪ Regulated DC Power Supply	Kikusui Electronics Corp.
▪ TC20™ Automated Cell Counter	Bio-Rad
▪ Telaval 31 Inverted Microscope	Zeiss
▪ Thermomixer® R	Eppendorf
▪ Trans-Blot® Turbo™ Transfer System	Bio-Rad

Table 2: Sequencing primers

Primer	Sequence (5' → 3')
APC_C8278T_F	AAGACACCCATGGGAAACAG
APC_C8278T_R	CTTCTTCGTGTTGGTGCTCA
ATM_C3092T_F	GCACTGGCATTTCACATA
ATM_C3092T_R	TTCGGAATATGGATCAGCCTA
BAZ1A_G392A_F	GGAAGCTCTTGAATCCGAAA
BAZ1A_G392A_R	GGCCCAGGCTAACCTAGAAG
CDKN1A_F	CGGTGACTCCTACTTCTGTGG
CDKN1A_R	TCTCCGTGACGAAGTCAAAG
EP400_A970T_F	CCCCAGATCAGCAGCATTAT
EP400_A970T_R	TCCTTGAGTGCCTCCATTTC
EXT1_A1036T_F	GCTCTGCTCTGAACCTCCAT
EXT1_A1036T_R	CCCAATTCTGGCTCTTCAA
HRAS1_A182T_F	ATGGGGTATGATCCATCAGG
HRAS1_A182T_R	CTCACGGGCTAGCCATAGGT
JAK1_2MUTS_F	TCTCGAGAGGAAGCCTGTGTC
JAK1_2MUTS_R	CTAAGAGCCATGGCAGGAAA
KMT2A_C9755T_F	AGTGCCCCTTCAAATATTGC
KMT2A_C9755T_R	TAGGGGCTGCTGTAGTTTGC
SETD1A_G1499A_F	CCAGCCCTGAGAGAGAAGAA
SETD1A_G1499A_R	AATTAGCTGGTGCAGGAGGA
SETD1A_G5095A_F	GCTTACACTCCCACCTCCTG
SETD1A_G5095A_R	AAGGACTGAGGCTCCCTTGT
SIN3B_C2441T_F	CAGGGATAGGGCCTCCTTAG
SIN3B_C2441T_R	ACTTGCTGTGTGGACCCTGT
SMARCC1_A365T_F	GGCATCTGGACACCAGACTT
SMARCC1_A365T_R	AAAGGCCTTACCTTGCCATT
SMARCD2_G497A_F	ACTCACACAGGGAGCTGTCC
SMARCD2_G497A_R	GGACTTCTGAAAGAACGCTCA
TP53_EXON4_F	TGCTCTTTTCACCCATCTAC
TP53_EXON4_R	ATACGGCCAGGCATTGAAGT
TP53_EXON5_F	TGAGGTGTAGACGCCAACTCT
TP53_EXON5_R	AACCAGCCCTGTCGTCTCT
TP53_EXON6_F	GCCTCTGATTCCTCACTGAT
TP53_EXON6_R	CGAAAAGTGTCTGTGCATCC
TP53_EXON7_F	CTGCTTGCCACAGGTCTCCCC
TP53_EXON7_R	TGTGCAGGGTGGCAAGTG
TP53_EXON8_F	TCCTTACTGCCTCTTGCTTCTCT
TP53_EXON8_R	AGGCATAACTGCACCCTTGG
TRRAP_G6952A_F	TCTTTGCCAGGAGCCACTAT
TRRAP_G6952A_R	GCCATCGGGGAATTATTCTT
SMARCB1_G158A_F	TGAACCCTTGATGTCAGCAG
SMARCB1_G158A_R	GCTGCAATGAAGACACTGGA
SMARCA2_C535A_F	ATATCTGGAGGAGGCCCAAC
SMARCA2_C535A_R	TGCAGAGTTTCAGGGAGAGG

F - forward, R - reverse

3.1.8. Software

The following programs were used:

- *ANNOVAR* for variant annotation (Wang et al., 2010),
- *Burrows-Wheeler Aligner (BWA) v0.7.5a* for alignment of sequencing reads to the reference genome,
- *Circos* for graphics,
- *Chromas* for visualisation of Sanger sequencing results,
- *The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7* for pathway analysis (Huang et al., 2009),
- *FastQC* for quality control of fastq files,
- *Genome Analysis Toolkit (GATK) v3.6-0* and *Picard tools v2.4.1* for duplicate read marking, realignment around indels, and base recalibration,
- *Galaxy server* and *MutSpec suite in Galaxy* (Ardin et al., 2016) for mutation spectra and mutational signatures analysis,
- *Ingenuity Pathway Analysis (IPA)* for pathway analysis,
- *MuTect v1.1.4* for somatic variant calling (Cibulskis et al., 2013),
- *Oncoprinter v1.1* for visualisation of mutations in samples included in cBioportal (Gao et al., 2013, Cerami et al., 2012),
- *Qualimap2* for quality control of bam files (García-Alcalde et al., 2012),
- *R Studio, R program and R packages ggplot2, reshape, grid, scales, gridExtra, and basic R statistical packages* for graphics and statistical analyses,
- *Variant Effect Predictor* for prediction of the mutation impact (McLaren et al., 2016).

The list includes only programs which were not specific to individual equipment. All programs listed above are freely available except for IPA, which is the product of Ingenuity Systems and is available under license. The license of Jiri Zavadil, one of the Thesis supervisors, was used.

3.2. Methods

3.2.1. Generation of immortalized cell lines

Hupki MEF cell lines were generated using the 3T3 protocol with minor adaptations (Liu et al., 2007). Fibroblasts were harvested from 13.5-day old embryos harbouring humanized *Tp53* knock-in cassette. Primary cells were treated with a carcinogen or a carrier in an early passage and cultivated until senescence bypass. This was established by the ability of cells to populate a flask after high dilution (1:10). Immortalized cell lines with *Tp53* mutations were preferentially chosen for whole exome sequencing.

Cells were grown in Advanced DMEM high glucose medium (Life Technologies) supplemented with 15 % fetal calf serum, 1 % L-Glutamine, 1% P/S, 2% pyruvate and 50 μ M 2-mercaptoethanol at 37°C and 5% CO₂ and atmospheric oxygen level. All used cell lines were tested negative for mycoplasma contamination.

3.2.2. Sequencing library preparation and whole exome sequencing

DNA extraction was done in Dr. Monica Hollstein's laboratory at DKFZ and sequencing library preparation and WES were done in several batches and outsourced to Otogenetics, Oxford Gene Technology and the Genome Technology Center of the New York University (each sequenced a different batch of samples applying comparable protocols). The chemicals, kits, machines and protocols used to generate the libraries are specified below.

DNA was extracted with DNeasy Blood & Tissue Kit (Qiagen) according to manufacturer's instructions. DNA was fragmented by ultrasound using Covaris machine (Covaris) or Bioruptor (Diagenode) and purified using Agencourt AMPure XP Beads (Beckman Coulter). DNA samples were then tested for size distribution and concentration using an Agilent Bioanalyzer 2100 or Tapestation 2200. Illumina libraries were generated with NEBNext reagents (New England Biolabs). Exon capture was done using SureSelect XT Mouse All Exon Kit (Agilent Technologies). Exon enrichment was verified by quantitative polymerase chain reaction (qPCR) and the quality, quantity and fragment size distribution of DNA was determined by Agilent Bioanalyzer or Tapestation. The libraries were sequenced in paired-end 100

nucleotide reads on Illumina HiSeq 2500 platform according to manufacturer's protocols.

3.2.3. WES data processing

WES data processing was done by Dr. Maude Ardin, using the computational resources at IARC.

3.2.3.1. Alignment

The quality of fastq files was determined by FastQC. The reads were then aligned to the mm9 mouse genome build, using the BWA-MEM type of Burrows-Wheeler Aligner. Duplicate marking, realignment around indels and base recalibration was done using Picard and GATK tools, respectively. An average of 51.44 million reads was sequenced per sample, of which 98% were mapped, 75% on target, with a mean depth-of-coverage of 54. Bam files were uploaded to the National Center for Biotechnology Information BioProjects web site, accession number PRJNA238303.

3.2.3.2. Variant calling, annotation, filtering

Variants were called with MuTect software using default parameters. Each immortalized cell line was compared to multiple primary cultures and only overlapping calls were considered, to ensure robust variant calling and exclude potential polymorphisms. Variants were annotated with ANNOVAR and single nucleotide polymorphisms (SNPs) according to the dbSNP database were filtered out. The complete list of variants is provided as Supplementary Data 1.

3.2.4. Mutation spectra and mutational signatures

Mutational spectra and signature analysis was performed using the MutSpec toolbox in Galaxy (Ardin et al., 2016). Annotated and filtered variant calling files were uploaded to the IARC Galaxy server and various metrics (distribution of single base substitutions, strand bias, spectrum of mutations in 96 sequence contexts, estimated number of mutational signatures) were calculated using MutSpec Stat tool.

MutSpec NMF was used to extract mutational signatures from the data, and MutSpec Compare to analyse, how similar the identified signatures were to the signatures from the COSMIC database.

3.2.5. Pathway analysis

Variants were filtered for exonic non-synonymous single base substitutions and splice site mutations. RefSeq-annotated genes affected by these variants were analyzed using DAVID and IPA. If RefSeq gene names were not recognized, aliases were used. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were interrogated by DAVID using relaxed criteria, as deregulation of biological processes in transformed cells can occur in the absence of multiple hits. IPA was run with default settings and canonical pathways were extracted using either standard ($p < 0.05$) or relaxed criteria ($p < 0.175$). The identified biological processes and pathways were prioritized based on recurrence among cell lines and cancer relevance.

3.2.6. Identification of candidate cancer driver mutations

Variants were filtered for exonic non-synonymous and splicing mutations and these were inspected for mutations in cancer-related genes and regulators of the epigenome (Vogelstein et al., 2013, Gonzalez-Perez et al., 2013, Futreal et al., 2004). Mutations were prioritized using a simple scoring system. A point was added if the mutation was exposure-specific (for example A:T>T:A in a cell line immortalized upon exposure to aristolochic acid) and therefore likely to be introduced early in the essay. Another point was added if the mutation was in a known hotspot or if it was truncating or if it was a mutation of a splice site. For other mutations, $\frac{1}{2}$ point was added if the mutation was located in a functional domain and $\frac{1}{2}$ point if the mutation is predicted deleterious in protein by SIFT via Variant Effect Predictor. A $\frac{1}{2}$ point was also added if the allelic frequency of the mutation was higher than 25 %. The highest-scoring mutations are those that are likely to be functionally-important clonal mutations introduced early in the essay, and therefore likely drivers of the immortalization process.

3.2.7. Single cell subcloning

Clonal populations were generated from the cell lines by dilution cloning. Briefly, ~30 cells were diluted in 10 mL of medium and distributed to 96-well plate. Individual emerging colonies were expanded and used for subsequent experiments.

3.2.8. DNA extraction and Sanger sequencing

DNA was extracted using QIAmp DNA Mini Kit, or NucleospinTissue kit according to manufacturer's protocol. For some samples, NaCl-ethanol extraction was performed. In this protocol, 1 million of cells is resuspended in 500 μ L of lysis buffer with 12.5 μ L of proteinase K and incubated for 2 hours in 56°C. The sample is mixed for 5 minutes, then 200 μ L of saturated NaCl solution are added and the sample is mixed again for 5 minutes. The sample is centrifuged for 10 minutes at 21,000 \times g and supernatant is transferred to a new tube. Isopropanol (400 μ L) is added and the sample is precipitated for 1 h in -20°C. The sample is then centrifuged as previously, supernatant is discarded and the pellet is washed with 500 μ L of 75% ethanol. The sample is centrifuged again, supernatant is discarded, the pellet is let air-dry and resuspended in 50-200 μ L of water or TER buffer.

Purity, quality and quantity of the DNA were assessed using Nanodrop, 0.8% agarose gel electrophoresis and Qubit, respectively.

Regions of interest were amplified using HotStartTaq DNA polymerase kit with dNTPs. The composition of the reaction was following:

<i>Reagent</i>	<i>Volume [μL]</i>
10 \times buffer	2.5
Q solution	5
dNTP mix (5 mM each dA, dT, dG, dC)	1
25 mM MgCl ₂	1
Forward primer (10 μ M)	0.5
Reverse primer (10 μ M)	0.5
HotStart Taq	0.125
Double distilled H ₂ O	12.375
DNA (50 ng/ μ L)	2
Total volume of the reaction	25

The PCR program was: 95°C/15 min - (94°C/30s - 60°C/30s - 72°C/60s) × 35 cycles - 72°C/10min - 10°C/∞. Quality of the PCR was assessed by agarose gel electrophoresis (2% gel in TBE buffer).

Sanger sequencing was done by Biofidal (Vaulx-en-Velin, France).

3.2.9. Inhibitor treatment

Cells were treated with 20 µM MEK inhibitor U0126, or 4, 8, and 16 µM of EZH2 inhibitor GSK126, or DMSO carrier as indicated in the Results section.

3.2.10. MTS proliferation assay

Cells were plated in 96-well plate and kept under indicated conditions. Cell viability was measured using CellTiter 96® AQueous One Solution Cell Proliferation Assay. Plates were incubated for 2 hours at 37°C and absorbance was measured in a microplate reader at 492 nm.

3.2.11. Colony formation assay

Basic colony formation assay was performed in standard 6-well plates. Cells were seeded so there were about 10 thousand cells at the time of treatment. Then they were kept under indicated conditions. Colonies were visualized after 7 days using crystal violet staining.

3.2.12. Protein extraction

For whole-cell protein extraction, cell pellets were resuspended in lysis buffer with protease inhibitors and phosphatase inhibitors and incubated 30 minutes on ice. The lysates were centrifuged (15 minutes at 18,400×g) and protein concentration in supernatant was measured.

For histone extraction, cell pellets were resuspended in lysis buffer and incubated for 10 minutes on ice (shaking). Nuclei were pelleted (6,500×g for 10 minutes) and histones were acid-extracted with 0.2N HCl at 4°C overnight (shaking). The extracts were neutralized with 20% of 1M NaOH and treated with benzonase (100U/mL, 10 minutes at 4°C, shaking). Debris was pelleted and protein concentration in supernatant was measured.

3.2.13. Polyacrylamide gel electrophoresis, immunoblotting and antibodies

Polyacrylamide gel electrophoresis and immunoblotting were performed using Trans-Blot® Turbo™ Transfer System. Firstly, 20-30 µg of whole cell protein or 3-5 µg of histones were boiled for 5 minutes with Laemmli buffer. The reaction was then loaded to 4–15% or 4–20% Mini-PROTEAN® TGX™ Precast Protein Gel which was run in Tris-glycine-SDS buffer at 90-120 V for 30-60 minutes. The blotting was done using Trans-Blot® Turbo™ Mini or Midi PVDF Transfer Packs, which include membrane, paper and transfer buffer, with a custom program on the Transblot Turbo machine (25V, 1.0 A, 10 minutes). Success of the transfer was verified using Ponceau S staining.

The general protocol for incubation with antibodies and signal development was following: firstly, the membranes were blocked with non-fat dry milk solution for 1 h. Then they were incubated with a primary antibody, diluted in blocking buffer, for 1 h in room temperature (antibody for actin, and histone antibodies) or in 4°C overnight (the remaining antibodies). The membranes were then washed 3 times for 10 minutes with TBS-0.1% Tween-20 and incubated with a secondary antibody, diluted in blocking buffer, for 1 h in room temperature and washed as before. All the aforementioned steps were done on a shaker. The development of signal was done using Clarity Western ECL kit according to manufacturer's protocol, and ChemiDoc Imaging System.

Antibodies were used in the following dilutions:

- phospho-Erk1/2 – 1,000×, Erk1/2 – 1,000×, actin – 25,000×, Ccnd1 – 100×, H3 – 20,000×, anti-mouse – 2,000×, anti-rabbit – 2,000× in 5% non-fat dry milk in TBS-Tween 0.1%.
- H3K27me3 – 4,000× in 3% non-fat dry milk in TBS-Tween 0.1%.

3.2.14. Mutation analysis in human tumour data

Published studies included in cBioPortal for Cancer Genomics (Cerami et al., 2012, Gao et al., 2013) were mined for samples with non-synonymous mutations and small indels in the Ep400 and Ttrap subunits of the TIP60 complex and in BAF complex subunits. Duplicates and samples missing mutation annotation were excluded. Data

were visualized using OncoPrinter version 1.0.1. Mutual exclusivity was tested using the χ^2 -test.

4. RESULTS

4.1. Global mutation analysis

4.1.1. Mutation burden in MEF cell lines

Twenty-six immortalized cell lines derived from primary Hupki mouse embryonic fibroblasts using modified 3T3 protocol were selected. Nineteen of the cell lines emerged after treatment with five different carcinogens: aristolochic acid (AA, N=7), aflatoxin B1 (AFB1, N=3), benzo[a]pyrene (B[a]P, N=3), N-methyl-N'-nitro-N-nitrosoguanidine (MNNG, N=4) and ultraviolet light class C (UVC, N=2). Two cell lines developed from primary Hupki MEFs which were engineered to overexpress activation-induced cytidine deaminase (AID), a DNA-mutating enzyme of the APOBEC family. Five cell lines immortalized spontaneously (Spont). Twenty-five cell lines were assigned to a test set for analyses, whereas one spontaneously immortalized cell line (Spont_5) was used as a control for subsequent experiments.

Exomes of the selected cell lines were sequenced at ~50× coverage. Three primary Hupki MEF cultures were used as controls for single nucleotide variant calling. Sequencing analysis of the 25 cell lines included in the test set yielded total of 16,061 single-base substitutions (Supplementary Data 1), The mutation load in the cell lines varied from ~100 to ~1,500 variants per cell line (Fig. 5).

	1	2	3	4	5	6	7
AA	679	613	444	356	265	280	583
MNNG	1231	1132	1258	1539			
Spont	279	355	115	334			
AFB1	236	317	359				
B[a]P	1260	715	1395				
AID	362	625					
UVC	364	965					

Figure 5: Burden of mutations in Hupki MEF cell lines. Number of single base substitutions identified in WES data for each of the 25 MEF cell lines is indicated. Colour scale indicates cell lines with higher (red) and lower (white) mutation burden.

4.1.2. Estimation of clonality of MEF cell lines

MEF immortalization protocol used to produce cell lines for this study does not guarantee that only one cell clone passes the senescence barrier. More clones can be present in the resulting immortalized culture. We used the WES data to estimate clonality of the sequenced MEF cell lines.

Single base substitutions are usually random and therefore most probably heterozygous. Hence, if a cell line is a result of a clonal expansion, majority of SBS should have the allelic frequency (or allelic fraction, AF) around 50 %. We determined the proportion of mutations which have AF between 25 % and 75 % in each cell line (Fig. 6). Eighteen cell lines had more than 50 % of mutations in the indicated interval, which suggested the presence of one predominant clone in the cell lines. Seven cell lines had less than 50 % mutations in the indicated interval, two of them (B[a]P_1 and AA_7) only 33.3 % and 11.3 %, respectively. This suggests presence of more than one predominant clone in the cell lines. In summary, majority of the cell lines probably contain only one predominant clone, but in some cases, more than one clone arose.

4.1.3. Mutation spectra analysis

Six types of SBS are recognized in the COSMIC nomenclature, based on the pyrimidine of the Watson-Crick pair: C>A, C>T, C>G, T>A, T>C, T>G. This nomenclature will be used throughout the Results section, to comply with the common usage, although it does not always correspond to the mutated base.

Proportions of the distinct SBS types were assessed in the WES data for each cell line of the test set. Distinct mutation spectra were observed in cell lines exposed to specific carcinogenic insults (Fig. 7):

Cell lines derived from cells **treated with AA** (AA_1-AA_7) displayed high (~50-60%) proportion of T>A transversions. This uncommon mutation type is a typical result of aristolactam adducts on adenine. This leads to strand bias towards T>A mutations on the transcribed strand (which are, in fact A>T mutations on the non-transcribed strand, which were not repaired with transcription-coupled repair machinery)

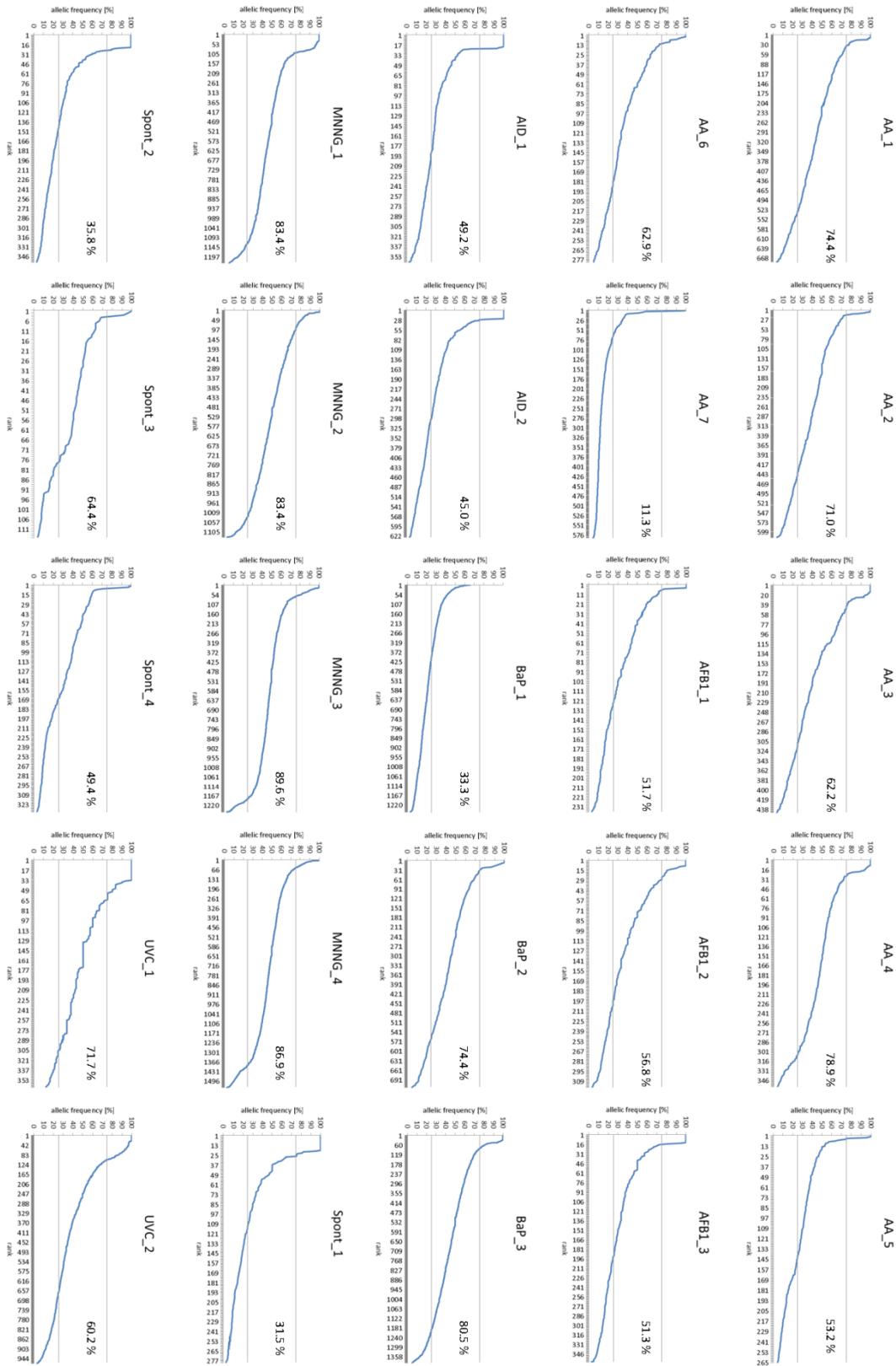


Figure 6: Distribution of allelic frequencies of mutations in Hupki MEF cell lines. Single base substitutions in the individual cell lines are ranked based on their allelic frequency. The proportion of mutations falling in the interval between 25% and 75% of allelic frequency is indicated.

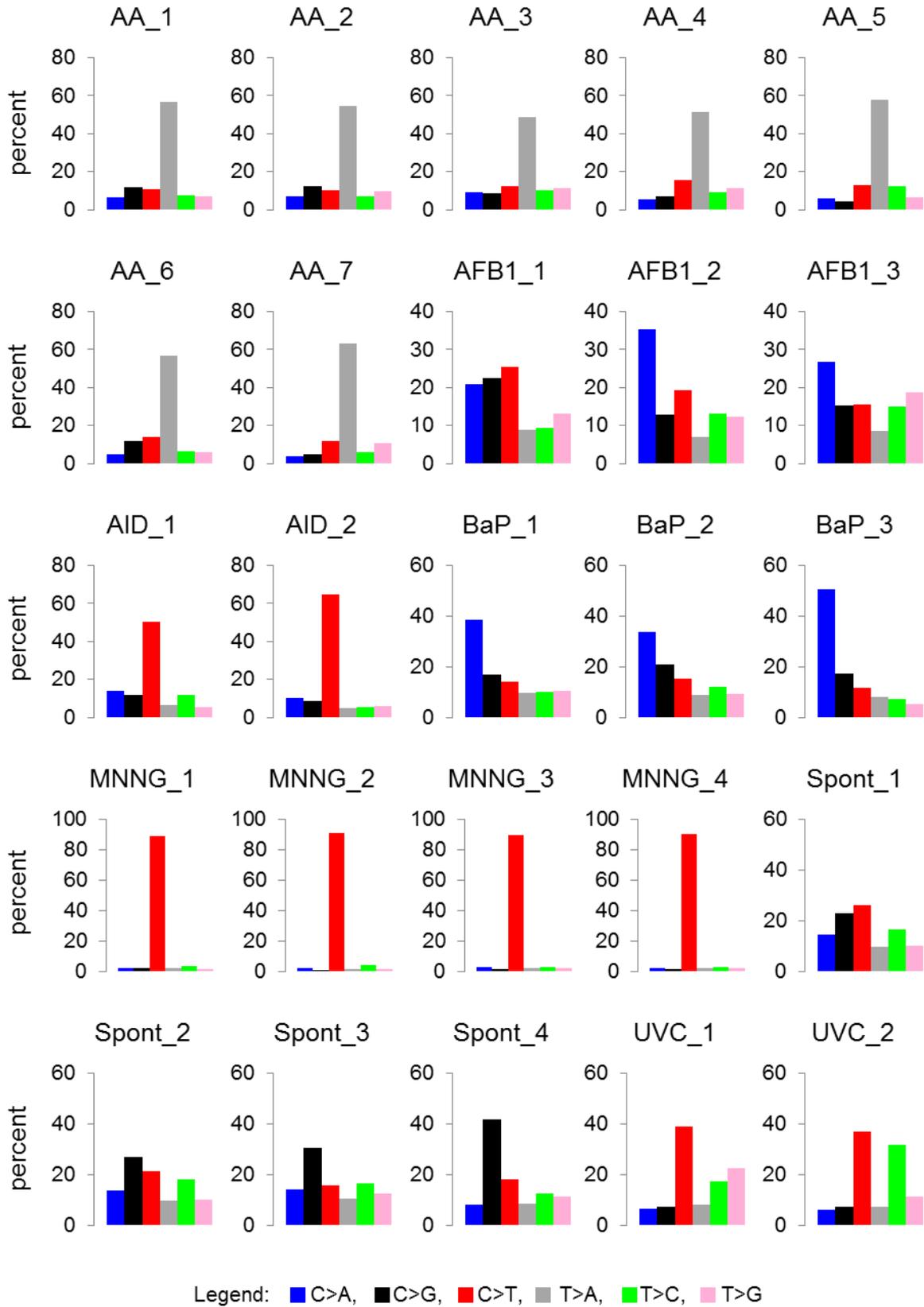


Figure 7: Mutation spectra of Hupki MEF cell lines. Proportion of each of the 6 mutation types as identified in WES data of 25 immortalized MEF cell lines.

(Sidorenko et al., 2012) (Fig. 8). This is also true in the case of aristolochic acid-treated MEF cell lines.

Cell lines derived from cells **treated with MNNG** (MNNG_1-MNNG_4) displayed high proportion of C>T transitions as expected from an alkylating agent. MNNG methylates guanine which causes mG-T mispairing during DNA replication, and ultimately creation of an A-T pair from the original G-C pair (Green et al., 1984, Loechler et al., 1984, Drabløs et al., 2004)

C>T was also a major mutation type in cell lines which developed from cells **overexpressing AID** (AID_1, AID_2), an enzyme which deaminates cytosine, causing its transition to thymine (Maul and Gearhart, 2010). On the contrary, C>T proportion in cell lines derived from the **UVC-treated** cells (UVC_1, UVC_2) is not that pronounced, even though it is the main mutation type introduced by the UV light (Pfeifer et al., 2005).

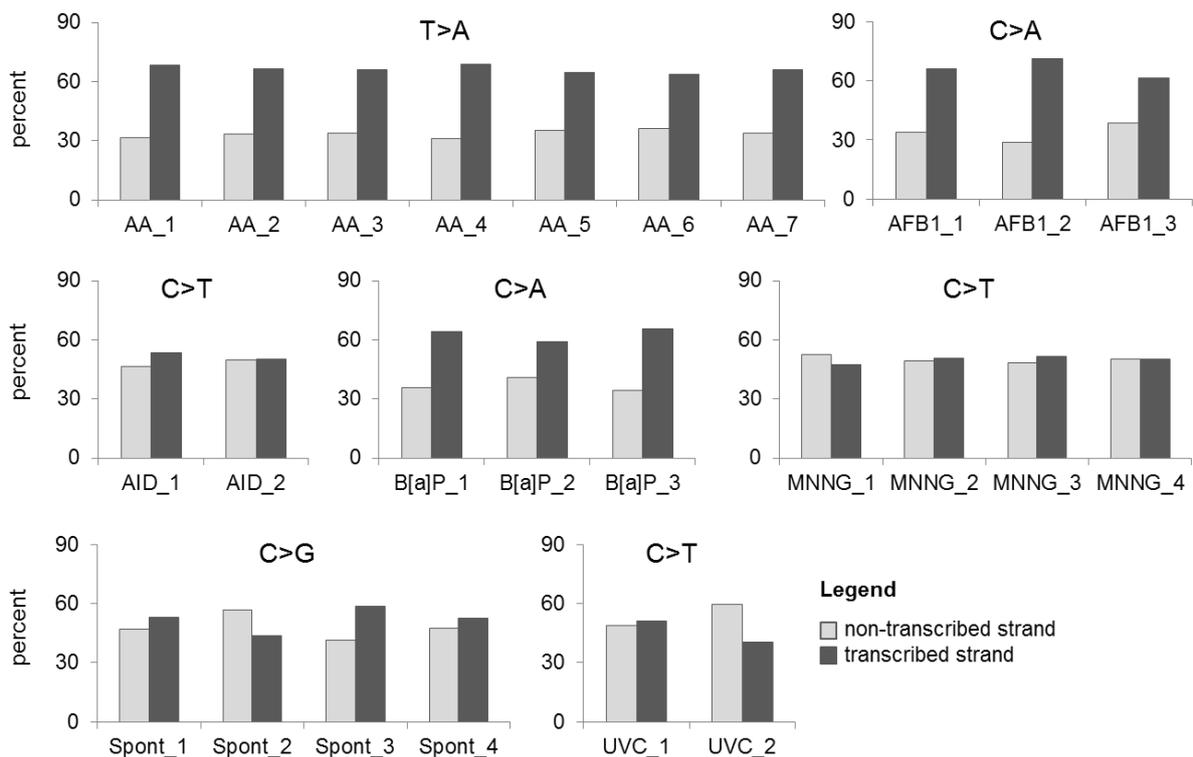


Figure 8: Mutation strand bias in Hupki MEF cell lines. Mutations belonging to the mutation type typical for each exposure were extracted. Proportion of mutations on transcribed and non-transcribed strand was plotted for each cell line.

Both **B[a]P** and **AFB1** create bulky adducts on guanine, which lead to introduction of C>A mutations preferentially on the transcribed strand (which are in fact G>T mutations on the non-transcribed strand) . This is also the case in the cell lines derived from cells treated with these agents (AFB1_1-AFB1_3, B[a]P_1-B[a]P_3) with the exception of the cell line AFB1_1, where C>T and C>G alterations are slightly more numerous. However, strand bias in C>A mutations, was present in all B[a]P and AFB1 cell lines (Fig. 8).

The most common mutation type in **spontaneously immortalized** cell lines (Spont_1-Spont_4) is C>G followed by C>T. The main stress of the culture conditions is supposed to be the reactive oxygen species (ROS). However, they typically produce C>A mutations; C>G and C>T are less commonly introduced by ROS (Yasui et al., 2014).

Together, the aforementioned results show that mutation spectra in cell lines derived from MEFs exposed to various mutagens correspond to the results expected from the mechanism of action of these mutagens. As a matter of fact, typical mutation types were experimentally identified earlier by sequencing of reference genes such as *Tp53* (Hollstein et al., 2013, Reinbold et al., 2008). However, genome-wide sequencing provides more data and higher resolution with fewer investments. Therefore, genome-wide sequencing of MEF cell lines is a valuable resource for assessment of mutation spectra of various agents.

4.1.4. Analysis of mutational signatures

We set to examine mutational signatures in 25 cell lines of the test set, to see if they can recapitulate signatures found in human cancers. We used the residual sum of square method to estimate the number of signatures present in the test sample set (Hutchins et al., 2008) and cosine similarity method to compare them to the 30 signatures extracted from more than 12,000 human tumours, listed in the COSMIC database.

Six mutational signatures were extracted and named A-F (Fig. 9).

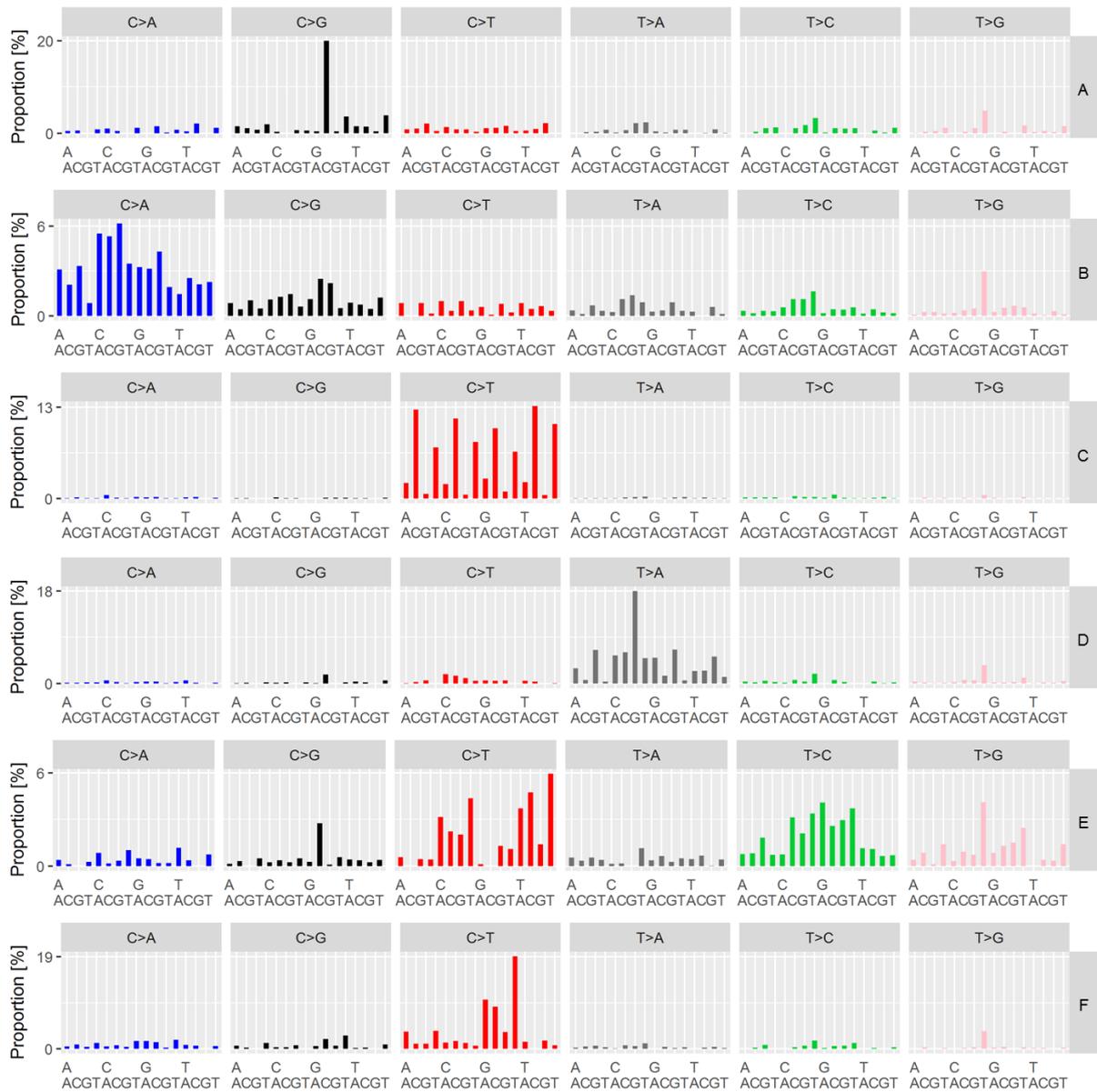


Figure 9: Mutational signatures identified in 25 immortalized MEF cell lines. Six mutational signatures were extracted using non-negative matrix factorization algorithm, based on the frequencies of 6 mutation types in 16 different sequence contexts (5' basis is in the first line below the graph, 3' basis is in the second line). The signatures were named A-F.

The main feature of the **signature A** was a pronounced peak of C>G mutations in 5'-G_C-3' context. This peak can be also spotted in signatures B, D, E and F, although it does not account for such a proportion of mutations as in the signature A. Signature A did not bear resemblance to any of the 30 COSMIC signatures (Fig. 10), and it consisted mostly of mutations found in spontaneously immortalized cell lines, although many other cell lines contributed to it, too (Fig. 11). We assume that this mutational signature reflects the effect of culture conditions on the cells.

Signature B displays high frequency of C>A mutations in various sequence contexts. The cell lines that mostly contributed to this signature were those derived from the exposures to B[a]P and AFB1. Signatures which were mostly similar to signature B were signature 4 (tobacco smoking, similarity 0.82), signature 24 (aflatoxin, similarity 0.76) and signature 29 (tobacco chewing, similarity 0.72). Importantly, performing NMF to extract more than 6 signatures did not separate AFB1 and B[a]P cell lines. Furthermore, performing NMF analysis to extract 6 signatures with all cell lines excluding AFB1-, or B[a]P-exposed cell lines did not improve correlation with the reference signatures for the respective exposures. In fact, C>A-rich mutational signature generated only with AFB1-treated cells had higher similarity to signature 29 (tobacco chewing) than to signature 24 (aflatoxin) (cosine 0.81 and 0.63, respectively). Indeed, there is not a given threshold for cosine value which would mean that two signatures are highly similar. As a rule of thumb, cosine > 0.9 is considered as highly relevant, similarity between 0.7 and 0.9 is considered as suggestive, similarity < 0.7 is considered as weak. The fact that the two signatures could not be separated and that the correlation with known signatures was not ideal, indicate limitations of MEF immortalization assay for some carcinogens. Treatment with AFB1 and B[a]P produced expected mutation spectra in the treated MEF cells (Fig. 7), but frequency of these mutations in specific sequence contexts was different than in human cancers, which could be a reflection of a different metabolism of the chemical agents and/or DNA adducts produced by these agents, or differences of the mixed real-life exposure and the effect of the specific compounds used to produce the experimental signatures.

Signature C consists of C>T mutations in 5'-N_R-3' context (N – any base, R – pyrimidine). It is specific to cell lines derived from cells treated with the alkylating agent MNNG. Signature C is identical to signature 11 (cosine 0.98), which has been attributed to exposure to the alkylating drug temozolomide. Similarly, **signature D** is identical to signature 22 (cosine 0.96), which has been linked to exposure to the aristolochic acid. The signature is rich in T>A mutations with a peak in 5'-C_G-3' context. Signature 22 is specific to cell lines derived from cells exposed to the AA. Alkylating agent MNNG and aristolochic acid are thus examples of carcinogens

where MEF immortalization assay gives the exact same results as the real-life exposure of human cells.

Signature E displays high proportion of C>T and T>C mutations, and, to a lesser extent, T>G mutations. This signature is mostly composed of mutations detected in the UVC-exposed cell lines, but many other cell lines also contribute to this signature. Signature E does not show a considerable similarity to any of the COSMIC

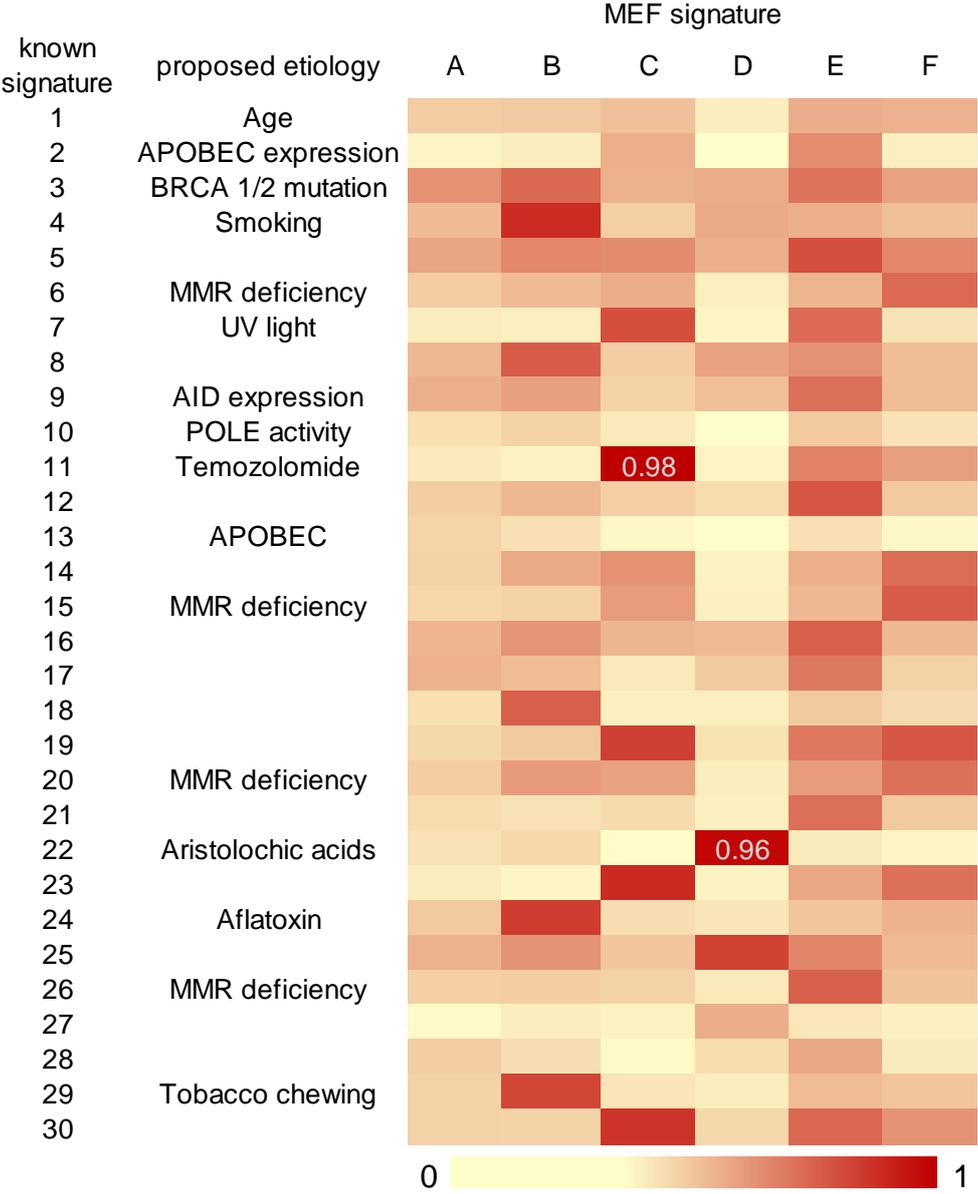


Figure 10: Similarity of signatures A-F to Cosmic mutational signatures. The similarity was established using cosine similarity method. Cosine values higher than 0.9 are displayed. AID – activation-induced cytidine deaminase, MMR – mismatch repair, UV – ultraviolet.

signatures. It is likely that this signature is a mixed one, scavenging mutations from many cell lines, but with high contribution of UVC-derived mutations. If we were to establish the UVC signature, we would ideally need to compare signature E to data from human tumours with known UVC exposure. However, UVC is almost entirely absorbed by the Earth's atmosphere and not used in indoor tanning devices. Therefore it is not a common exposure and it is not likely to find tumour samples with a signature specific to UVC for validation of the signature proposed from the *in vitro* experiments.

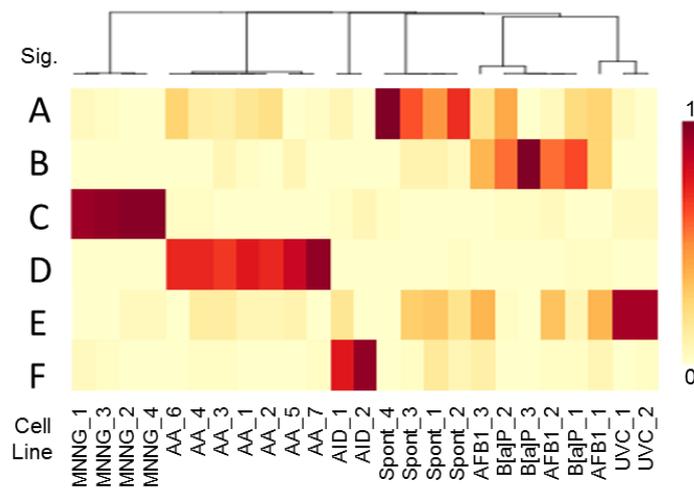


Figure 11: Relative contribution of mutations from individual MEF cell lines to mutational signatures A-F, identified in the pooled data set. Darker colour indicates higher proportion of mutations of an individual cell line contributing to the specific signature.

Signature F is defined by a high proportion of C>T mutations with a noticeable peak in the 5'-G_T-3' context. This signature is specific to cell lines developed from cells overexpressing the AID transgene. It is not similar to any of the 30 reference signatures. However, C>T mutations in 5'-G_T-3', 5'-G_A-3' and 5'-G_C-3' contexts are typical for AID activity in the immunoglobulin gene (Rogozin and Kolchanov, 1992, Puente et al., 2015). An analogous signature was identified in whole genome sequencing data from 30 samples of chronic lymphocytic leukaemia and attributed to ectopic activity of AID (Kasar et al., 2015). Signature F from Hupki MEF cell lines recapitulates signature of AID activity in human cancer.

Interestingly, all the six signatures identified in MEF cell lines contained a T>G fraction, which resembled the T>G fraction which is the main feature of COSMIC

signature 17. The origin of the signature 17 is unknown. A very recent study found, using data from mouse tumours, that T>G mutations in the contexts typical for the signature 17 were found in lower AF (Huang et al., 2017). Similar picture arises from mutation spectra (plotted as variant frequency by trinucleotide sequence contexts) of variants, from pooled WES MEF cell line data, segregated to bins according to their AF (Fig. 12). The AF of signature 17-predominant T>G mutations was less than 40 %. Remarkably, C>G mutations in 5'-G_C-3' context, the main feature of MEF signature A, had an analogous character. They were also more abundant in the bins containing alterations with the AF smaller than 40 %. This suggests that signature 17 and MEF signature A are produced by processes, (either innate, or environmentally induced), which operate in long-term during the development of the cell lines.

In summary, MEF immortalization assay allows to recapitulate mutational signatures found in human cancers and can be, with some limitations, used to assess mutation spectra and signatures for mutagenic agents. MEF immortalization assay may be equally useful for studying origins of signatures with yet unknown aetiology.

4.1.5. Functional annotation of mutations in immortalized MEF cell lines

The total of 16,061 single-base substitutions were identified in 25 immortalized MEF cell lines (Supplementary Data 1). 10,687 of them were annotated as exonic or splicing mutations using the RefGene database via ANNOVAR. Figure 13 shows the breakdown of these mutations to finer categories. Most were nonsynonymous missense mutations (7,081), other types of nonsynonymous alterations were less numerous (stopgain – 344, stoploss – 19). Furthermore, 171 mutations were modifying the splice sites. Lastly, 3,009 alterations were identified as synonymous and 63 were annotated as 'unknown' (meaning that a transcript maps to multiple locations in the reference genome build, none having a complete open reading frame).

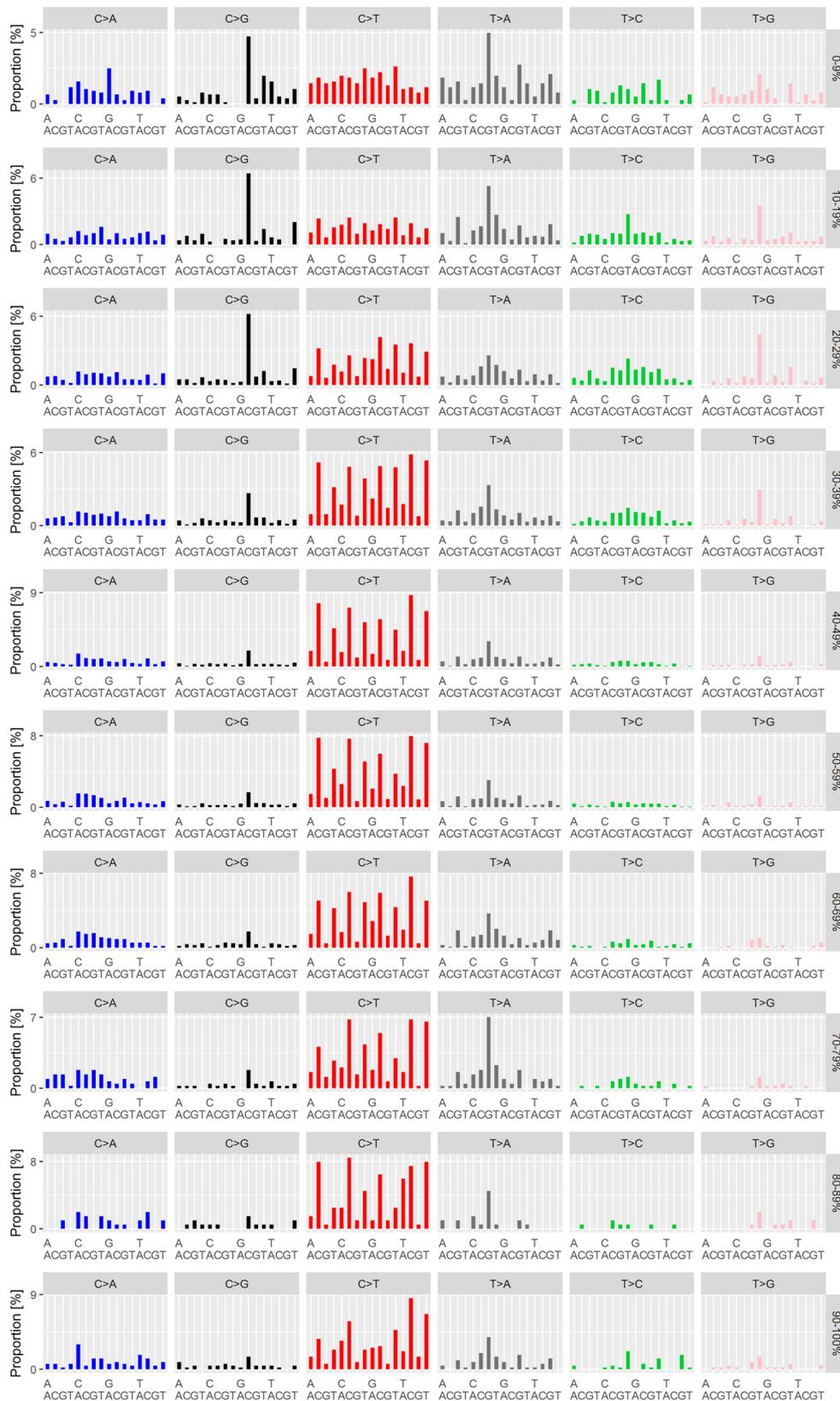


Figure 12: Distribution of mutation spectra based on allelic frequency. Mutations were divided to bins based on allelic frequency. Graphs show proportion of alterations per context (5' basis is in the first line below the graph, 3' basis is in the second line) and mutation type in each bin.

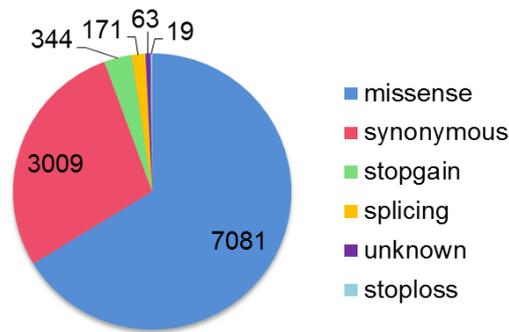


Figure 13: Annotation of exonic and splicing variants found in 25 MEF immortalized cell lines.

Pathway analysis was performed to reveal the processes affected by nonsynonymous mutations. DAVID was used to query the GO and KEGG databases. Ingenuity Knowledge Base was queried using the IPA tool. The pathway analysis was performed with relaxed criteria, and number of genes mutated in the pathway/process, rather than p-values, were taken into account when prioritizing the hits. Among the frequently affected processes were those involved in structural integrity (adhesion, extracellular matrix, cytoskeleton) and signalling pathways connected to these entities. Furthermore, pathways involved in regulation of cell cycle, proliferation, apoptosis and differentiation were affected (MAPK signalling, Wnt signalling, Tp53 pathway, Notch and Hedgehog signalling, etc.), as well as chromatin modification and transcription regulation (Table 3). The processes, which were found frequently affected in the set of 25 immortalized MEF cell lines, are also frequently deregulated in human tumours, and many can be classified under the hallmarks of cancer (Hanahan and Weinberg, 2011). These results indicate that processes affected by mutations in human cancers are also affected in immortalized MEF cell lines.

Table 3: Selection of biological processes and pathways recurrently affected in immortalized MEFs.

Class	ID	Term	# cell lines affected
GOTERM_BP_FAT	go:0007155	cell adhesion	20
	go:0051726	regulation of cell cycle	12
	go:0051056	regulation of small gtpase mediated signal transduction	12
	go:0006915	apoptosis	11
	go:0007267	cell-cell signaling	11
	go:0007186	g-protein coupled receptor protein signaling pathway	8
	go:0006974	response to dna damage stimulus	7
	go:0006281	dna repair	7
	go:0006511	ubiquitin-dependent protein catabolic process	7
	go:0016568	chromatin modification	7
go:0045449	regulation of transcription	6	
GOTERM_CC_FAT	go:0005856	cytoskeleton	23
	go:0031012	extracellular matrix	22
	go:0005886	plasma membrane	22
	go:0030054	cell junction	19
	go:0044427	chromosomal part	13
	go:0005783	endoplasmic reticulum	11
	go:0005813	centrosome	7
	go:0005819	spindle	4
GOTERM_MF_FAT	go:0000166	nucleotide binding	24
	go:0032553	ribonucleotide binding	24
	go:0005524	atp binding	24
	go:0008092	cytoskeletal protein binding	22
	go:0004672	protein kinase activity	22
	go:0016887	atpase activity	18
	go:0003677	dna binding	16
	go:0005085	guanyl-nucleotide exchange factor activity	16
	go:0003700	transcription factor activity	12
	go:0005089	rho guanyl-nucleotide exchange factor activity	11
	go:0005088	ras guanyl-nucleotide exchange factor activity	10
	go:0008233	peptidase activity	10
	go:0008528	peptide receptor activity, g-protein coupled	6
KEGG_PATHWAY	mmu04010	mapk signaling pathway	21
	mmu04510	focal adhesion	20
	mmu05200	pathways in cancer	19
	mmu04310	wnt signaling pathway	18
	mmu04512	ecm-receptor interaction	16
	mmu04115	p53 signaling pathway	12
	mmu04012	erbb signaling pathway	10
	mmu00983;mmu00982	drug metabolism	9
	mmu04120	ubiquitin mediated proteolysis	8
	mmu04350	tgf-beta signaling pathway	7
	mmu04910	insulin signaling pathway	6
	mmu04330	notch signaling pathway	6
	mmu04340	hedgheg signaling pathway	4
	IPA_PATHWAY		actin cytoskeleton signaling
		xenobiotic metabolism signaling	12
		apoptosis signaling	11
		atm signaling	8
		mouse embryonic stem cell pluripotency	8
		wnt/ β -catenin signaling	8
		epithelial adherens junction signaling	7
		telomerase signaling	7
		molecular mechanisms of cancer	6
		sapk/jnk signaling	6
		p53 signaling	6
		pi3k/akt signaling	5
		stat3 pathway	5
		integrin signaling	5
		hif1 α signaling	5
	notch signaling	3	

4.1.6 Mutations in cancer genes

It was published earlier – using the evidence from several of the cell lines included in our study – that Hupki MEF immortalization assay selects for *Tp53* mutations typical for human cancer (Liu et al., 2004), and that these mutations reflect the mutation spectra of carcinogenic exposures used to generate the cells (Liu et al., 2004, Besaratinia and Pfeifer, 2010, vom Brocke et al., 2006, Luo et al., 2001a).

In the laboratory of Dr. Monica Hollstein, where most of the cell lines included in our study was generated, as well as in our laboratory, mutations in the humanized cassette of the *Tp53* gene are routinely tested in immortalized cell lines. Cell lines bearing a non-synonymous mutation in the *Tp53* gene were preferentially selected for WES. Table 4 lists non-synonymous mutations in the *Tp53* gene identified in the set of 25 Hupki MEF cell lines. Twenty different mutations were identified in the cell lines. As published previously, mutations in the *Tp53* gene reflected mutations typical for carcinogenic compounds in human tumours. For example, the cell line AA_2 has two *Tp53* mutations, both of the T>A mutation type. One of them, N131Y, was found in urothelial tumours linked to the exposure to aristolochic acid (Odell et al., 2013). The set of cell lines also contained typical *Tp53* mutations in codons 245, 248, 249 and 273 (URL3).

Table 4: Nonsynonymous *Tp53* mutations in 25 immortalized MEF cell lines.

MUT_ID	cDNA sequence	protein sequence	Effect	Domain Function	SIFT	Cell line
1002	c.C296T	p.S99F	missense	NA	deleterious	MNNG_1
1066	c.G314C	p.G105A	missense	DNA binding	deleterious	Spont_3
1341	c.C380T	p.S127F	missense	DNA binding	deleterious	B[a]P_1
1392	c.A391T	p.N131Y	missense	DNA binding	deleterious	AA_2
1567	c.C423G	p.C141W	missense	DNA binding	deleterious	B[a]P_2
1736	c.C454T	p.P152S	missense	DNA binding	deleterious	MNNG_4
1881	c.C476T	p.A159V	missense	DNA binding	deleterious	MNNG_4
2220	c.C535T	p.H179Y	missense	DNA binding	deleterious	UVC_2
3244	c.G734A	p.G245D	missense	DNA binding	deleterious	MNNG_2
3297	c.G743A	p.R248Q	missense	DNA binding	deleterious	UVC_1
3299	c.G743T	p.R248L	missense	DNA binding	deleterious	AFB1_2, AFB1_3
3314	c.A745T	p.R249W	missense	DNA binding	deleterious	AA_1
3347	c.C749A	p.P250H	missense	DNA binding	deleterious	UVC_1
3494	c.A774C	p.E258D	missense	DNA binding	deleterious	AA_4
3730	c.C817T	p.R273C	missense	DNA binding	deleterious	AA_3
3739	c.G818T	p.R273L	missense	DNA binding	deleterious	AFB1_1
3866	c.C843G	p.D281E	missense	DNA binding	deleterious	Spont_4
3884	c.G845C	p.R282P	missense	DNA binding	deleterious	B[a]P_1
4021	c.A871T	p.K291*	nonsense	DNA binding	NA	AA_2
4407	c.A961T	p.K321*	nonsense	NLS	NA	MNNG_3

MUT_ID - mutation identifier from the IARC p53 database, NA - not available, NLS - nuclear localization signal.

Next it was examined, whether more known cancer driver genes, in addition to *Tp53*, are mutated in the set of cell lines. To that end, a catalogue of cancer driver genes was built by merging entries listed in the Cancer Gene Census (Futreal et al., 2004), a manually curated database of genes implicated in oncogenesis, and a list of cancer driver genes from an authoritative review (Vogelstein et al., 2013). More than 300 hits were found by filtering nonsynonymous exonic and splicing mutations against the catalogue. Among these alterations were Hras^{Q61L} and Kras^{Q61R} (Fig. 14), well-known driver mutations in human tumours, which lead to constitutive growth signalling (Prior et al., 2012). Murine Hras and Kras proteins are almost identical to their human orthologues (Fig. 14), and the applicability of human data to mouse proteins was established for the two specific alterations (Westcott et al., 2015).

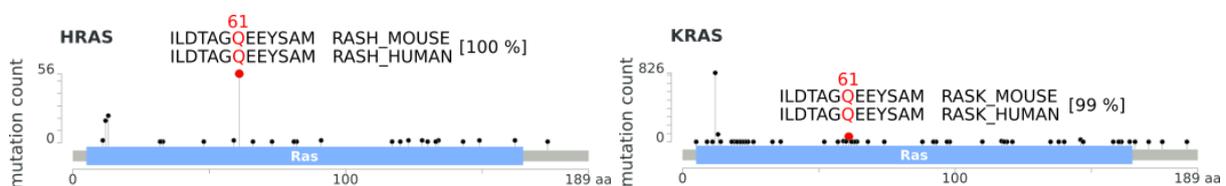


Figure 14: Mutations, found in MEF BBCE cell lines, which were previously identified as tumour hotspots. Plots, showing mutations in human HRAS and KRAS proteins based on the TCGA data, were generated using cBioPortal (Cerami et al., 2012, Gao et al., 2013). The mutated residue in MEFs is highlighted by a red circle. Alignment of human and mouse protein sequence around the mutated residue is shown in the insert, the mutated codon is indicated above the alignment. The overall similarity of human and mouse protein sequence is indicated in square brackets.

Since driver genes confer a selective growth advantage to a cell clone, they tend to be found frequently mutated in tumours. Thus, a common approach to identification of putative cancer driver genes in tumour sequencing data is the analysis of recurrence. We performed a simple recurrence analysis for the genes found in our catalogue of cancer driver genes. Fifty-one elements from the catalogue, including tumour suppressors *Atm*, *Apc* and *Arid1b*, were mutated in more than one cell line (Fig. 15). Interestingly, four of the 51 recurrently mutated genes were bearing the exact same mutations at all instances. These were: *Cdh11* (5×), *Hist1h1e* (2×), *Pax5* (3×), and *Rbm15* (2×). Observing frequent hotspots in a relatively small set of samples is not expected; such hotspots could be unfiltered germline variants or sequencing errors. Mining bam files of additional 73 Hupki cell lines from other

Hist1h1e, *Rbm15*) did not show strong features of artefacts and were thus kept on the list.

Altogether, immortalized MEF cell lines harbour mutations in genes known to be important for human cancer development.

4.1.7. Mutations in genes and complexes involved in regulation of epigenome

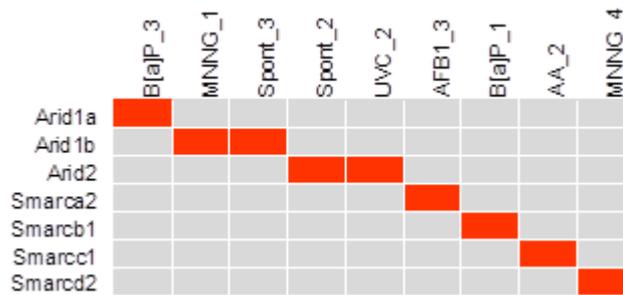
Epigenetic modifications (DNA methylation, histone modifications, chromatin remodelling) have a crucial role in regulation of gene expression. Epigenetic modifiers (genes modifying DNA or chromatin) have been recognized as important players in neoplastic development and were found frequently mutated in human tumours (Feinberg et al., 2016, Gonzalez-Perez et al., 2013, Leiserson et al., 2015).

To identify mutations in genes involved in epigenetic regulation, nonsynonymous exonic and splicing mutations were filtered against the list of bona-fide epigenetic modifiers (Gonzalez-Perez et al., 2013) and histone genes. 105 nonsynonymous and splicing mutations were found in 66 epigenetic modifiers, the most frequently affected epigenetic processes were histone methylation and demethylation and ATP-dependent chromatin remodelling (Fig. 16). Also, there were 34 mutations in 23 histone genes. Most of the histone mutations, 24, were found in two cell lines with the AID transgene. Twenty epigenetic modifiers and four histone genes were mutated recurrently (Fig. 15).

Proteins involved in regulation of the epigenome usually act in complexes. Protein complexes are the true effectors, therefore recurrent mutations of a chromatin modifying complex, rather than its individual components, could be the ultimate driver event. If a mutation in one subunit of a complex impairs the function of the whole complex, then one would expect to see mutual exclusivity in mutations of the complex subunits. Such a pattern was observed for the subunits of the BAF complex, an ATP-dependent chromatin remodelling complex, in the Hupki cell line dataset, as well as in an analysis of human tumour sequencing data from the cBioportal database (Cerami et al., 2012, Gao et al., 2013) (Fig. 17). This is in line with previously published results showing mutual exclusivity of BAF complex

Group	Gene	AA_1	AA_2	AA_3	AA_4	AA_5	AA_6	AA_7	AFB1_1	AFB1_2	AFB1_3	AID_1	AID_2	BaP_1	BaP_2	BaP_3	MNING_1	MNING_2	MNING_3	MNING_4	Spont_1	Spont_2	Spont_3	Spont_4	UVC_1	UVC_2	SUM	
ATP-dependent chromatin remodeling	Gatad2a							1												1					1	0		
	Rbbp7															1											1	
	Arid1a															1								1			1	
	Arid1b																2							1			3	
	Arid2																						1				2	
	Baz1a	1																							1		2	
	Baz1b																									1	3	
	Baz2a															1										1		2
	Chd1																									1		1
	Chd3																									1		1
	Chd4																									1		1
	Smarca2										1					1										1		2
	Smarca2																									1		1
	Smarca1														1											1		1
	Smarca1																									1		1
Smarca2																									1		1	
DNA methylation	Aicda											1	1													1	2	
	Dnmt1													1												1	2	
	Dnmt3a														1											1	1	
	Tet1																									1	1	
Tet2																									1		1	
Histone Acetylation	Ep300																									1	1	
	Ep400	1	1																							1	4	
	Kat6a																									1	1	
Kat7																									1	3		
Histone Deacetylation	Hdac10															1										1	1	
	Hdac2																									1	1	
	Hdac6						1											1								2	2	
	Hdac9									1																1	1	
	Ncor1																									1	1	
	Sirt5																			1						1	1	
Sirt6																									2	2		
Tb11xr1				1			1																		1	1		
Histone Demethylation	Kdm1b	1																									1	
	Kdm2a														1												1	
	Kdm3a																										2	
	Kdm3b		1														1	1									1	
	Kdm4a				1						1																2	
	Kdm4d															1											1	
	Kdm6a																									1	1	
	Kdm6b	1																								1	3	
	PH2																									1	1	
	Rbp2															1											1	1
Histone Methylation	Ash1							1																			2	
	Ash2																										1	
	Ehmt2			1																							1	
	Kmt2a																										1	
	Kmt2b															1	1	2			1						6	
	Kmt2c					1																					4	
	Kmt2d	1			1										2											1	8	
	Nsd1																										2	
	Prdm9																										1	
	Rtf1																										1	
	Setd1a															1											2	
	Setd1b																1										1	
	Setd7																										1	
	Smyd1																										1	
Other	Bag6																										1	
	Lmna																										1	
Mum1																										1		
PRC1	Cbx2																									1	1	
	Cbx7																										1	
	Ezh1																										1	
	L3mbtl1																										1	
	Phc2																									1	1	
Ring1																									1	1		
PRC2	Asx1																									1	1	
Histone genes	Hist1h1b																										2	
	Hist1h1d																										3	
	Hist1h1e																										5	
	Hist1h1t																										1	
	Hist1h2ab																										1	
	Hist1h2ag																										1	
	Hist1h2ai																										1	
	Hist1h2ak																										1	
	Hist1h2bb																										1	
	Hist1h2bc																										2	
	Hist1h2be																										1	
	Hist1h2bf																										1	
	Hist1h2bg																											

A



B

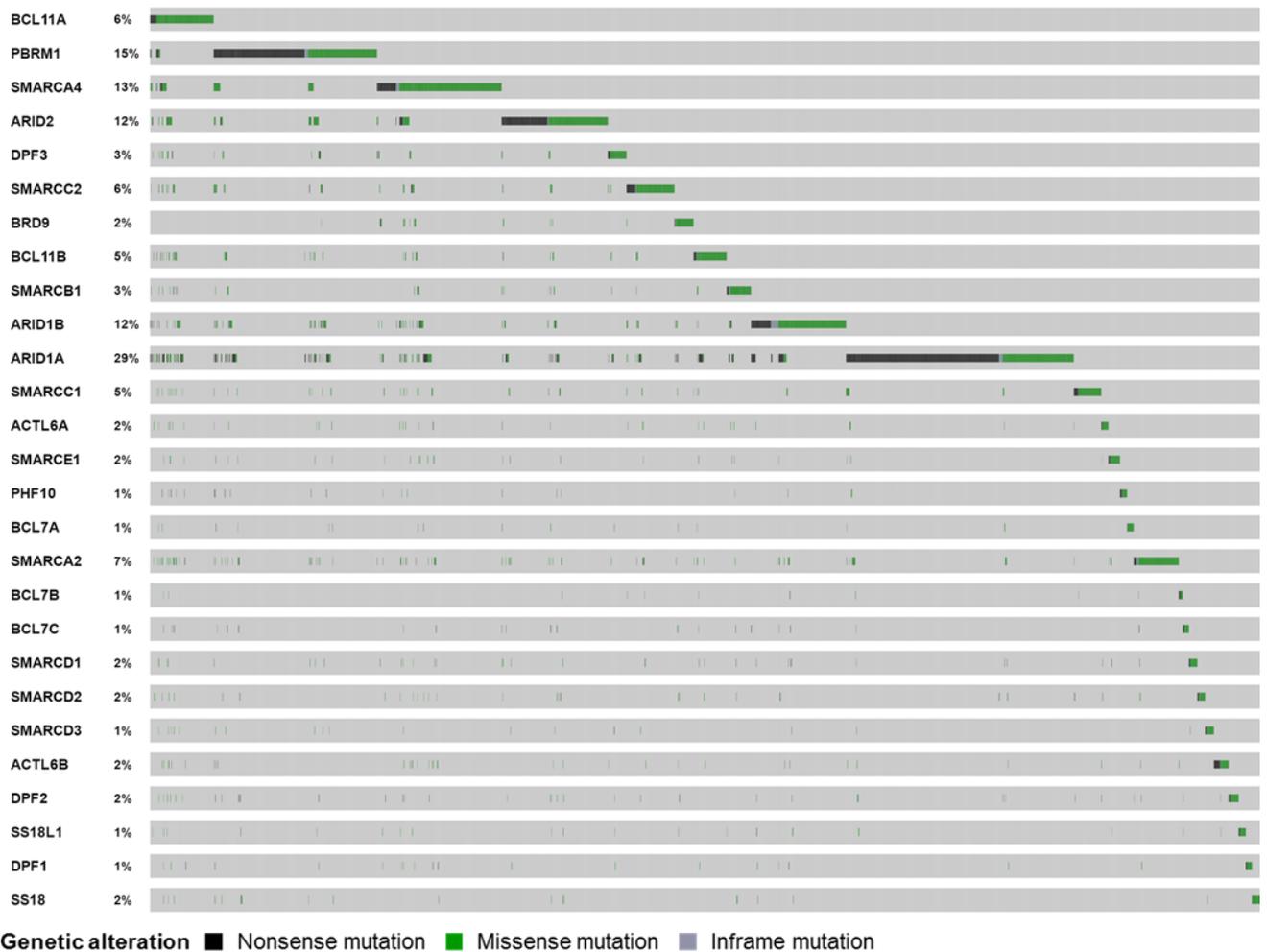


Figure 17: Analysis of BAF complex mutations in mouse and human samples. A – BAF complex subunits mutated in MEF BBCE cell lines (in red). B – Mutational analysis of BAF complex subunits in human tumours. Data from whole exome and whole genome sequencing of human tumours were downloaded from cBioPortal (published studies only) (Cerami et al., 2012, Gao et al., 2013). 2,860 samples with at least one mutation in a BAF complex subunit were found. The results were plotted using Oncoprinter v1.0.1.

subunits in human tumours (Leiserson et al., 2015, Gonzalez-Perez et al., 2013) building on the studies describing loss-of-function mutations in specific BAF subunits (Helming et al., 2014).

Besides BAF complex subunits, Ep400 and Trrap subunits of the TIP60 histone acetyltransferase complex, too, displayed a tendency towards mutual exclusivity in the Hupki cell line dataset. Each was mutated in four different cell lines, but only one cell line had mutations in both Ep400 and Trrap (Fig. 18). Analysis of data obtained by sequencing of human tumours showed statistically significant mutual exclusivity of EP400 and TRRAP mutations (Fig.18). These results indicate that TIP60 complex is another chromatin modifying protein complex whose impaired activity could be important for neoplastic development.

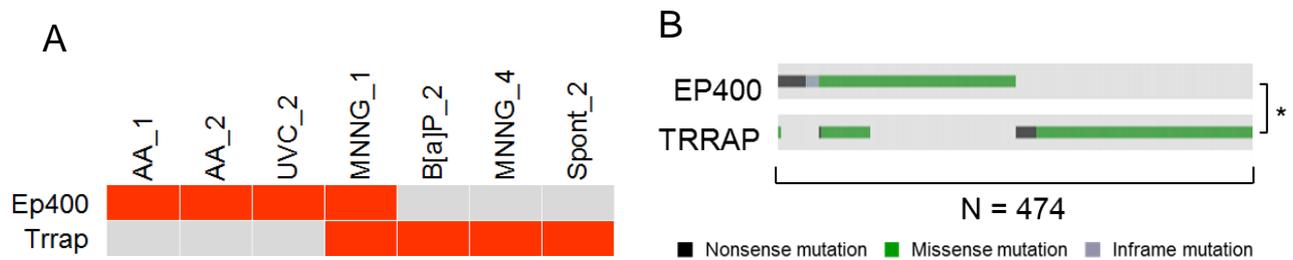


Figure 18: Analysis of Trrap and Ep400 mutations in mouse and human samples. A - TIP60 complex subunits Ep400 and Trrap mutated in MEF BBCE cell lines. B - TIP60 complex subunits EP400 and TRRAP mutated in human sequencing studies included in cBioPortal. Result of X²-test indicated, *p<0.001.

Notably, Trrap was not a part of the list of *bona-fide* epigenetic modifiers, although it is a well-known part of histone acetyltransferase complexes (Murr et al., 2007). This prompted us to search for other epigenome regulators, not included in the original list (Gonzalez-Perez et al., 2013), but mutated in the set of Hupki MEF cell lines. The resulting list included 123 mutations of 78 epigenome regulators, belonging mostly to the ATP-dependent chromatin remodelling class and histone modification classes, and is attached as Supplementary Figure 1 (Stanley et al., 2013, Jahan and Davie, 2015, Eom et al., 2011, Grozeva et al., 2014, Cohn et al., 2016).

4.2. Identification and functional testing of putative cancer driver events

4.2.1. A systematic prioritization scheme for high-confidence candidate driver events

The finding that MEF immortalization assay selects for mutations in known cancer genes led us to the idea that it could be possible to use the cell lines to identify and investigate potential cancer driver events.

The challenge was to distinguish the potential drivers from passengers. We reasoned that a driver mutation would:

- be of the mutation type typically introduced by the carcinogen used to generate the cell line. The carcinogen treatment was administered in the early passages. Therefore mutations of the exposure-specific type are likely introduced early and contribute to the immortalized, cancer-like phenotype of the cell lines.
- have the allelic frequency of around 50%. Single base substitutions are most likely heterozygous. Mutations which facilitate clonal expansion would therefore have the allelic frequency around 50%.
- affect the function of a gene. Mutations affecting splice sites, introducing or eliminating a stop codon, mutations in DNA sequence encoding a functional domain of a protein, and recurrent mutations ('hotspots') are more likely to have a deleterious effect on a gene. Specialized algorithms, such as SIFT (Ng and Henikoff, 2001), can also be used to assess the effect of an SBS on a gene's function.

These criteria were used to score mutations in immortalized MEF cell lines (see Material and Methods for details). In this proof-of-principle study, we assessed the score for genes included in the Cancer Gene Census and in genes involved in regulation of the epigenome. Table 5 shows high-scoring mutations in two cell lines: AA_2 and MNNG_4. These cell lines were derived from cells treated with aristolochic acid and methyl-nitro-nitrosoguanidine, respectively. These compounds

produced the clearest mutational signatures from the chemicals tested (Fig. 9, Fig. 10), and AA- and MNNG-treated cell lines were therefore chosen for the discovery phase. The cell lines were subcloned to produce truly clonal cultures. Notably, all but three high-scoring mutations which were validated by Sanger sequencing were present in all subclones (Supplementary Figures 2 & 3). This finding supported the overall clonal nature of the two cell lines and clonal character of the putative driver mutations.

Table 5: Candidate driver mutations in two immortalized MEF cell lines.

Cell line	Gene symbol	Function	transcript ID	cDNA change	AA change	Mutated in human tumors (Cosmic) [%]	known to be involved in senescence
AA_2	Cbx7	PRC1 complex	NM_144811	c.T32A	p.F11Y	0.2	
	Cdkn1a*	Cell cycle	NM_007669.5	c.94-2A>T		0.3	YES
	Ep400*	TIP60 complex	NM_029337	c.A970T	p.R324X	2.2	YES
	Ext1	Glycosaminoglycan metabolism	NM_010162	c.A103T	p.S35C	0.5	
	Ext1*			c.A1036T	p.R346X		
	Hras*	MAPK signaling	NM_001130443	c.A182T	p.Q61L	2.7	
	Jak2	JAK-STAT signaling	NM_001048177	c.A2479T	p.I827L	30.0	
	Smarcc1*	BAF complex	NM_009211	c.A356T	p.H119L	0.7	YES
	Smyd1	H3K4 methylation	NM_009762	c.T1072A	p.S358T	0.7	
	Tp53*	Transcription, DNA repair	NM_000546.4	c.A391T	p.N131Y	26.8	YES
Tp53*		c.A871T		p.K291X			
MNNG_4	Apc*	Wnt signaling	NM_007462	c.C8278T	p.P2760S	10.9	YES
	Atm*	DNA repair	NM_007499	c.C3092T	p.T1031I	4.2	YES
	Baz1a*	ACF complex	NM_013815	c.G392A	p.R131K	0.8	
	Brca1	DNA repair	NM_009764	c.C4322T	p.P1441L	1.3	YES
	Gatad2a	Histone deacetylation	NM_001113345	c.G243A	p.M81I	0.3	
	Jak1*	JAK-STAT signaling	NM_146145	c.C1327T	p.P443S	1.0	
	Jak1*			c.C1286T	p.P429L		
	Kmt2a*	H3K4 methylation	NM_001081049	c.C9755T	p.P3252L	1.6	
	Prdm1	Transcription	NM_007548	c.C2420T	p.P807L	1.1	
	Setd1a*	H3K4 methylation	NM_178029	c.G1499A	p.S500N	1.2	
	Setd1a*			c.G5095A	p.D1699N		
	Sin3b*	HDAC	NM_009188	c.C2441T	p.T814I	0.7	YES
	Smarcd2*	BAF complex	NM_031878	c.G497A	p.G166E	0.3	
	Trrap*	TIP60 complex	NM_001081362	c.G6952A	p.V2318M	2.6	
	Tp53*	Transcription, DNA repair	NM_000546.4	c.C454T	p.P152S	26.8	YES
Tp53*	c.C476T			p.A159V			

* Validated by Sanger sequencing (Supplementary Figure 2 and Supplementary Figure 3).

4.2.2. Ras^{Q61} mutation supports cell proliferation in nutrient-poor conditions

Among the high-scoring alterations in the AA_2 cell line was Hras^{Q61L}, a well-characterized mutation which causes enduring oncogenic signalling. Another activating Ras mutation, Kras^{Q61R}, was found in the UVC_2 cell line. AA_2-1

and UVC_2-3 subclones were compared with two immortal cultures lacking Ras mutations (AA_3-3, Spont_5). Both Ras-mutated clones were growing in multilayers, appeared to be less tightly attached to the surface of culture vessel, and had the population doubling time around 12 hours. On the contrary, Ras wild-type (Ras-wt) cultures grew in single layers well attached to the culture vessel and had population doubling time of around 24 hours, which was a standard doubling time for most cell lines in the collection.

These phenotypic differences could be, at least in part, caused by constitutive activation of the Ras pathway in the mutant cell lines. Therefore, the cultures were treated with the U0126 agent, which inhibits Ras/Raf/Mek/Erk signaling on the level of Mek kinase. Indeed, U0126 treatment led to elimination of Erk1/2 phosphorylation and a slightly delayed downregulation of its target gene *Ccnd1* (Fig. 19B,C). 24-hour treatment of AA_2-1 and UVC_2-3 cultures with 20 μ M of U0126 induced pronounced change of the cells' morphology, while no effect was observed in the case of AA_3-3 and Spont_5 (Fig. 19A).

Next, cell viability was determined after 24 hours of 20 μ M U0126 treatment. Since the physiological role of Ras is to transduce exogenous mitogenic signals, we hypothesised that the effect of the Mek inhibitor would be increased in environment deficient for such signals. To mimic such conditions, the experiment was performed in medium with normal and limited amount of serum. Upon serum starvation, Ras-mutant, but not Ras-wt cultures showed significant decrease in amount of metabolically active cells as determined by MTS assay (Fig. 19D). Together these results confirm the impact of activating Ras mutations in the MEF system and suggest that the Ras-mutant, but not Ras-wt cultures develop dependency on activated Ras signaling.

4.2.3. Inhibition of Ezh2 activity leads to cell death in BAF-mutant cell lines in an oncogenic Ras-dependent manner

AA_2 and MNNG_4 cell lines each contained one nonsynonymous mutation affecting a BAF complex subunit: *Smarcc1*^{H119L} and *Smarcd2*^{G166E}, respectively. These genes have not yet been recognized as putative cancer driver genes in human tumour

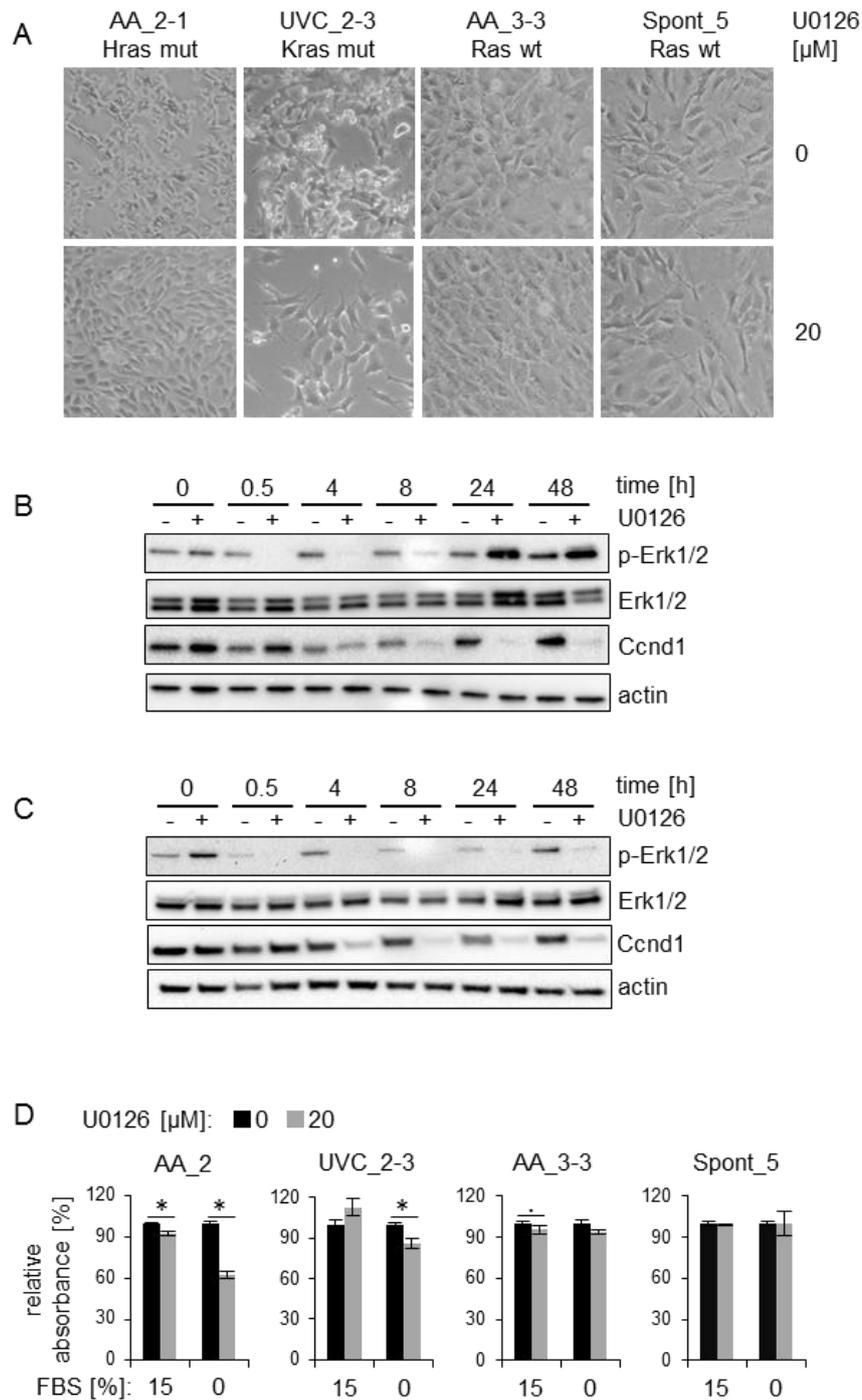


Figure 19: Effect of Mek inhibitor treatment. A - Morphology of cell cultures with activating Ras mutations (AA_2-1, UVC_2-3) and cell cultures with wild type Ras genes (AA_3-3, Spont_5) after 24-hour treatment with 20 μ M of Mek inhibitor U0126, or carrier (DMSO). Magnification 100 \times . B and C – phosphorylation of Erk1/2 kinases during Mek inhibitor treatment, and Erk target Ccnd1, in AA_2-1 (B) and UVC_2-3 (C) cell clones. Actin is used as loading control. D – number of cells, measured by relative absorbance in MTS assay, after 24-hour treatment with 20 μ M Mek inhibitor U0126, or carrier (DMSO) in medium with 15% serum and in serum-deprived medium. Results of at least three independent experiments are plotted as average and standard error of mean. Result of Wilcoxon two-sample test is indicated: \cdot $p < 0.05$, * $p < 0.01$.

sequencing projects. However, they fitted in the mutually-exclusive mutational pattern of BAF complex subunits (Fig. 17).

The result of the activity of the BAF complex – or complexes, since there is variability in the composition of the complex in different cell types – seems to be opening of chromatin and enabling gene expression (Kadoch and Crabtree, 2015). On the contrary, PRC2, a Polycomb group complex, introduces the H3K27me3 mark, leading to tighter packing of chromatin and gene silencing (Golbabapour et al., 2013). The balance between BAF and PRC2 complex was found to be important in development and cell fate decisions (Kadoch et al., 2016). Previous elegant work showed that mutations in BAF complex subunits constitute dependence on PRC2 complex, specifically its catalytic subunit, EZH2 (Kim et al., 2015). Treatment with EZH2 inhibitor and/or EZH2 siRNA resulted in more cell death and less colony formation ability in human BAF-mutant cancer cell lines. However, data from Project Achilles, which uses genome-scale RNA interference and CRISPR/Cas9 reagents to find vulnerabilities in cancer cell lines, indicated that the BAF-PRC2 dependency does not operate in cell lines with a concurrent RAS mutation (Kim et al., 2015). We set out to test whether the functional relationship between BAF and PRC2 could be recapitulated using MEF cell clones with previously untested BAF mutations, either alone (Smarcd2 – MNNG_4-2 clone) or in combination with activating Ras mutation (Hras^{Q61L}, Smarcc1 – AA2_1 clone). Cell line Spont_5, wild type for mutations in Ras genes and the BAF complex, was chosen as a negative control.

Cultures were treated with the Ezh2 inhibitor GSK126; cell viability was assessed using MTS assay and colony formation assay (Fig. 20A,B). Cell line Spont_5 displayed a decrease in the proportion of viable cells and in the number of colonies after 4 and 7 days, respectively, of treatment with the Ezh2 inhibitor. However, a fraction of cells survived in both assays. This was in sharp contrast with MNNG_4-2 clone, which was highly sensitive to GSK126 and did not show any remaining viability after 3 days (MTS assay) and 7 days (colony formation assay), respectively, of treatment. Finally, the clone AA_2-1, which harboured Hras and Smarcc1 mutations, was the most resistant to the Ezh2 inhibitor treatment. The same order of sensitivity was observed in an independent experiment with another set of MEF

immortalized cell clones: UVC_2-3 (Arid2, Kras^{Q61R}), MNNG_1-1 (Arid1b), AFB1_3-2 (Smarca2), BaP_1-2 (Smarcb1) (Fig. 21). In this case, the threshold of sensitivity was higher than in the previous set of experiments. The change could be attributed to the change of serum provider, and the batch of inhibitor, given that the positive control MNNG_4-2 clone was also less sensitive to the treatment (12 μ M inhibitor was needed to kill the cells in three days in the MTS assay, versus 8 μ M in the first set of experiments). Importantly, immunoblotting shows a decrease in H3K27me3 mark upon inhibitor treatment in all cell lines (Fig. 20C, Fig. 21C).

In summary, these data show that mutations in BAF complex subunits consistently sensitize the cells to Ezh2 inhibitor treatment in an (oncogenic) Ras dependent fashion.

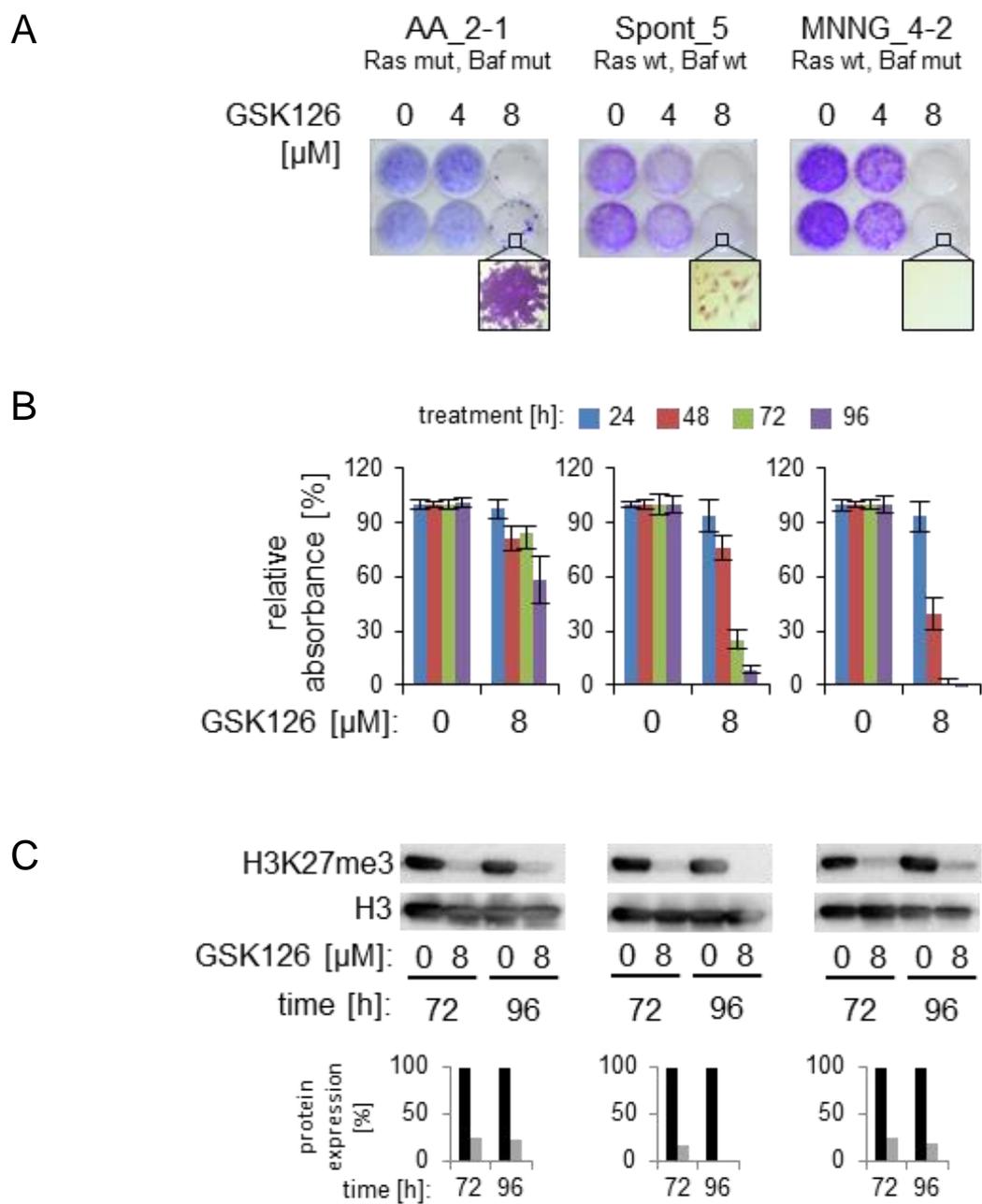


Figure 20: Effect of Ezh2 inhibitor treatment – first set of experiments. A – results of colony formation assay. Cells were seeded in a low density and treated with Ezh2 inhibitor GSK126, or with carrier (DMSO). Colonies were visualized after 7 days using crystal violet staining. Window shows 100 \times magnification. B – results of MTS assay. Cells were treated with Ezh2 inhibitor or carrier (DMSO) and absorbance was measured at indicated time points. Results of three independent experiments are plotted as mean and standard error of mean. C – Immunoblot for H3K27me3 mark in cells treated with Ezh2 inhibitor and a carrier. H3 was used as loading control. Abundance of H3K27me3 is plotted (treated relative to untreated cells).

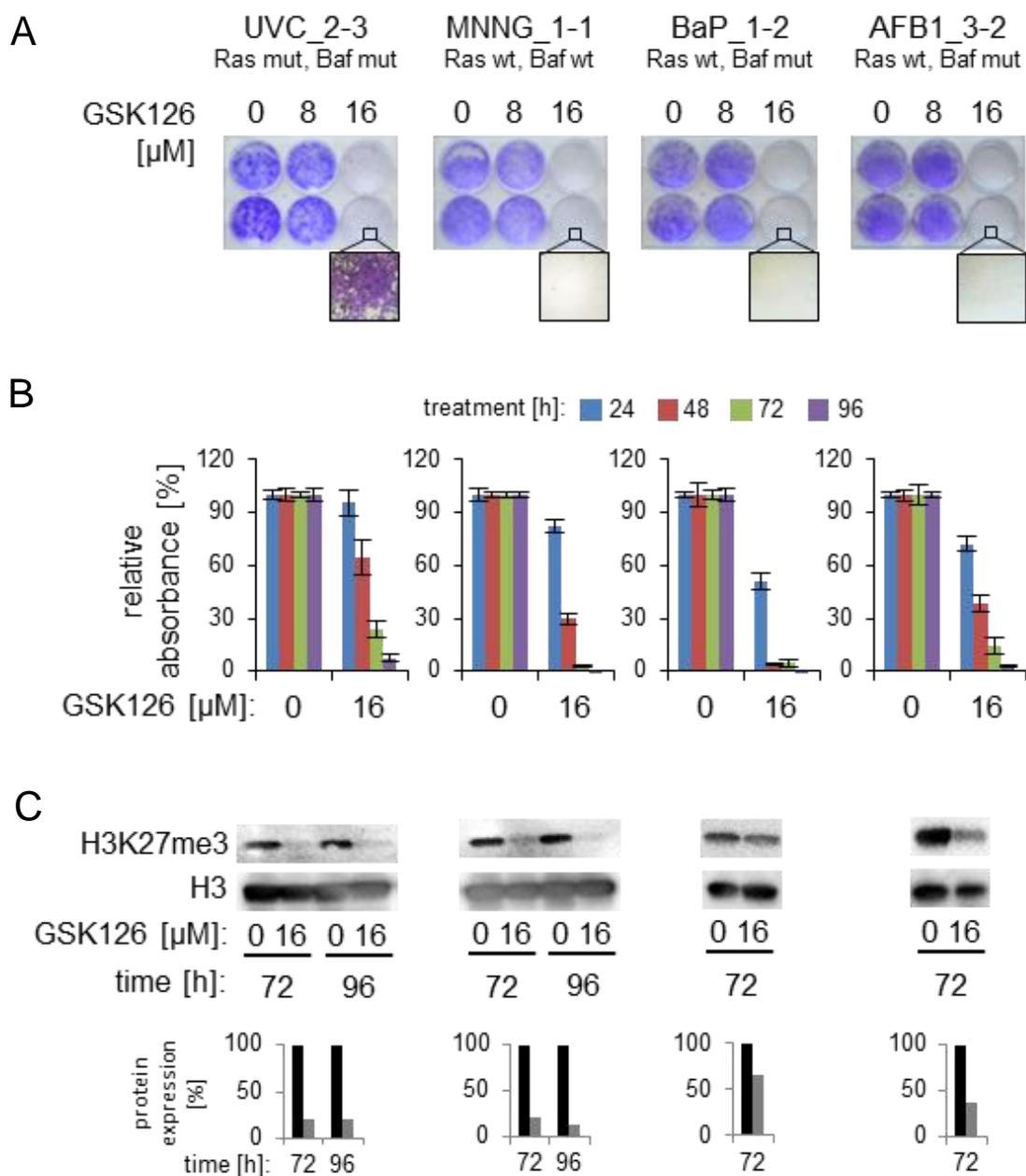


Figure 21: Effect of Ezh2 inhibitor treatment – first set of experiments. A – results of colony formation assay. Cells were seeded in a low density and treated with Ezh2 inhibitor GSK126, or with carrier (DMSO). Colonies were visualized after 7 days using crystal violet staining. Window shows 100× magnification. B – results of MTS assay. Cells were treated with Ezh2 inhibitor or carrier (DMSO) and absorbance was measured at indicated time points. Results of three independent experiments are plotted as mean and standard error of mean. C – Immunoblot for H3K27me3 mark in cells treated with Ezh2 inhibitor and a carrier. H3 was used as loading control. Abundance of H3K27me3 is plotted (treated relative to untreated cells).

5. DISCUSSION

In this Thesis, I demonstrated the use of a simple cell-based *in vitro* carcinogen exposure system coupled with massively parallel sequencing as a model approximately mimicking steps of cancer initiation, promotion and progression. As in many experimental systems in science, it allows to study certain features of a real-life phenomenon, while not considering others. In other words, it cannot recapitulate tumour development in its complexity; however, it can be useful in gaining insight in some of its key aspects.

The system depends on immortalization of mouse embryonic fibroblasts upon treatment with a (potentially) carcinogenic compound. This assay has been previously used to generate mutation spectra of carcinogenic compounds with mutations in the *Tp53* gene (Liu et al., 2004, Liu et al., 2005, Liu et al., 2007, Nedelko et al., 2009, Odell et al., 2013, Reinbold et al., 2008). The results presented in this Thesis demonstrate that genome-wide sequencing of the MEF cell lines provides sufficient detail to a) extract mutational signatures of mutagens with which the cells were treated, and b) to identify drivers of the cancer-like phenotype of the cell lines distinct from the well-known *Tp53* gene.

5.1. MEF immortalization assay to decipher mutational processes operative in human cancer

Though *Tp53* mutations are relatively common in immortalized MEF cells and the incidence can be increased by certain approaches (Kucab et al., 2017), generating experimental mutation spectra only with *Tp53* gene requires production and selection of many cell lines, and is therefore laborious and expensive. However, the same assay generates enough detail with fewer cell lines used when genome-wide sequencing is employed.

The results of this proof-of-principle study show that the MEF immortalization assay coupled with genome-wide sequencing recapitulates mutation spectra of environmental mutagenic compounds (AA, AFB1, B[a]P, MNNG) found in human

cancer and by other experimental approaches (Severson et al., 2014, Poon et al., 2013, Alexandrov et al., 2013a).

The assay generated enough data, so that the mutational signatures could be extracted by the NMF method. Mutational signatures of compounds used in the study were also very similar to those identified in human cancers and recorded in the COSMIC database of mutational signatures. This is especially true for AA and MNNG. AFB1 and B[a]P produced the expected mutation spectra: majority of mutations were C>A (G>T) with transcriptional strand bias. However, in the MEF assay, both compounds participated on a single signature which bore similarities to both tobacco-related signatures (COSMIC signatures 4 and 29) and the aflatoxin-related signature 24. When the analysis was done separately, using data from either AFB1-treated cell lines, or B[a]P-treated cell lines, together with the data from the rest of the cell lines (to have enough data for running the NMF algorithm), the B[a]P signature was very similar to signature 29 (tobacco chewing, similarity 0.9) and signature 4 (tobacco smoking, similarity 0.83), while the AFB1 signature was more similar to signature 29 (similarity 0.81) than to signature 24 (aflatoxin, similarity 0.63). These data suggest that the mutational signature of aflatoxin is different in human and mouse.

Potential mouse-human differences (in metabolism, gene expression, activity of DNA repair) can create a mutational signature specific to the mouse, which is usually of lesser relevance to human cancer studies. On the other hand, it must be stressed that COSMIC mutational signatures are not a final, permanent reference as they are bound to change with the expansion of cancer studies such as the Pancancer Analysis of Whole Genomes of the ICGC. Mutational signatures extracted from sequencing data of human tumours, and the causes, to which these signatures are attributed, depend on the number of analysed variants, current state of knowledge on the action of various compounds and innate cellular and physiological processes, and the quality of medical records connected to the sequencing data. Also, humans are exposed to various agents in their lives, which are then demonstrated in the mutational signatures present in human tumours. For instance, there are many different aflatoxins, and food is usually contaminated

by several of them. Aflatoxin B1 is a well-recognized carcinogen, and widely studied, but other aflatoxins also have mutagenic and genotoxic activities (Kumar et al., 2016). The COSMIC signature 24 might be caused by a mixture of aflatoxins. Similarly, tobacco consumption generates many chemicals – B[a]P being of the best known carcinogenic one, but the smoking and chewing signatures might reflect the presence of other chemicals than B[a]P. Thus, signatures generated under well-controlled, experimental conditions are needed to decipher the origins of human cancer.

Hupki MEF immortalization coupled with massively parallel sequencing was one of the first approaches allowing modelling mutational signatures of human cancers using mammalian cells (Olivier et al., 2014). Other systems like human renal tubule HK-2 cell line (Poon et al., 2013), human mammary epithelial cells HMEC (Severson et al., 2014) (Severson 2014) provide useful data from human systems. The real asset is of these systems is that they produce mutational signatures specific for the cells which are the targets of the compounds used in the assays (AA in the case of HK-2, B[a]P in the case of HMEC). However, MEFs produce correct mutational signatures – at least for the compounds described in this Thesis –, are much easier to handle compared to the aforementioned human cells, and the assay is relatively short (2 months vs. 6 months for human-cell systems). The current state of the field was recently summarized elsewhere (Zhivagui et al., 2016, Hollstein et al., 2017).

Many chemical, environmental carcinogens do not act as the original molecules, but must be metabolically activated. In the present study, it is the case of AA, AFB1 and B[a]P which are metabolized with the assistance of cytochrome P450 enzymes, which are expressed in Hupki MEFs (Liu et al., 2004). However, other compounds might need enzymes which are not expressed in the Hupki MEFs. In such cases, use of the human liver S9 fraction, which contains liver metabolic enzymes, should solve the problem. The same goes for the human-cell assays.

Besides the well-known carcinogens, AA, AFB1, B[a]P and MNNG, this Thesis also shows data on mutational signature of UVC radiation. A study by Liu *et al.* (Liu et al., 2004) showed that Hupki MEF cell lines, which developed after treatment

with UVC, exhibit *Tp53* mutations in dipyrimidine sites. Such mutations are typically seen in human melanoma and are the result of UV light exposure (Pfeifer et al., 2005). It would seem that MEF assay recapitulates the effects of the UV light, observable in human cancers. This may be true; however, the genome-wide UVC signature obtained from MEF assay only partially resembles the UV signature observed in human melanoma (COSMIC signature 7). The reason for this may be that UVC is not a common exposure in human, because it is almost entirely absorbed by the atmosphere. Neither are there any other genome-wide studies of UVC experimental exposures, so the MEF UVC signature cannot be compared to other data. Arguably, it does not matter, since there is no clinical significance for UVC exposure. It would be interesting to have a UVC signature, however, because in our case the signature which is mostly present in the UVC lines also contains a considerable proportion of mutations from other cell lines (Fig. 11).

This can be due to the mathematical strategy for extraction of the signatures. NMF has become a popular decomposition for dimensional reduction and is widely used in various domains; in particular in genomics research to extract mutational signatures in tumour samples. Several implementations or types of NMF techniques have been proposed (Alexandrov et al., 2013b, Rosales et al., 2017, Fischer et al., 2013, Gaujoux and Seoighe, 2010). Determining the number of signatures to extract is challenging for NMF, as is the case for other reduction methods. Many methods depend on indices (such as residual sum of squares) that improve monotonically with the number of signatures and hence encourage over-fitting: this is a major problem in statistical modelling (Hastie et al., 2009). Conversely, the number of signatures extracted must be smaller than the number of samples, so in many circumstances we can at best hope to obtain combinations of the signatures really acting. An alternative is to use a regression-based method, which decomposes the mutational spectrum of each sample separately according to the pre-determined signatures provided, such as those catalogued in COSMIC. These methods, such as the non-negative least square method in R or *deconstructSigs* (Rosenthal et al., 2016) are highly sensitive to the presence of known signatures, but have the disadvantage that they cannot extract novel signatures.

More than half of the signatures in the COSMIC database have been attributed to endogenous mutational processes – defects in DNA repair and the activity of enzymes from the APOBEC family. Two cell lines from the cell line collection used in this study were generated from Hupki MEF cells overexpressing AID, which is an APOBEC family enzyme, critical for somatic hypermutation and class-switch recombination of immunoglobulin genes (Maul and Gearhart, 2010). It was proposed that ectopic activity of this enzyme contributes to the development of B-cell cancer (diffuse large B-cell lymphoma) and cancers of the digestive system (gastric, gallbladder, colorectal) (Komori et al., 2008, Endo et al., 2008, Khodabakhshi et al., 2012, Pasqualucci et al., 2001, Matsumoto et al., 2007). The major peaks of MEF AID signature are identical to human AID signatures recently constructed from mutations in immunoglobulin genes in CLL (Puente et al., 2015), and extracted from WGS of 30 CLL cases (Kasar et al., 2015). These signatures are not present in the COSMIC database. The COSMIC signatures 2 and 13 were attributed to APOBEC activity, but were later found to be produced by another enzyme of the family, APOBEC3A.

The last signature extracted from Hupki MEF cell lines had a prominent peak of C>G mutations in 5'-G_C-3' context. To our knowledge, this signature hasn't been described before. It was the main signature in the spontaneously immortalized cell lines, but it was also present in most of the other cell lines (Fig. 11). Oxidative stress is a major cause of senescence in MEF cells cultivated in atmospheric oxygen levels; cells grown in 3-5% oxygen can be cultivated for a very long time without showing marks of senescence (Parrinello et al., 2003). Oxidative stress usually leads to C>A transversions, and, to a lesser extent, C>G transversions (Yasui et al., 2014). Yeast cells with the deletion of peroxiredoxin Tsa1 displayed similar proportions of C>A and C>G mutations, and about three times more C>T mutations (Serero et al., 2014). We propose that MEF signature with C>G predominance is caused by oxidative damage during the standard culture conditions, potentially together with a sensitising genetic cause.

Oxidative stress is present during the whole time of experiment, thus, it should continually produce mutations which would, as a result, be present in lower allelic

fractions. This was the case for C>G mutations, as shown in Figure 12. Interestingly, T>G mutations in sequence contexts typical for the signature 17 also were more present in lower allelic fractions (<40 %). It could be, thus, that the signature 17 is also caused by a process steadily operating in the cells.

MEF immortalization assay has proven useful in deciphering mutational signatures of known, as well as suspected human carcinogenic processes. The main limitations of the assay could be potentially insufficient metabolic activity for activation of some chemical compounds, and differences in DNA repair between mouse and human. The assay should be very useful in an integrated approach where mutational signatures are also extracted from human tumour sequencing data, *in vivo* experimental exposures (as in the US National Toxicology Program) and human cells exposed to the mutagen of interest. Comparison of *in vitro* vs *in vivo* and rodent vs. human should gain robust mutational signatures, necessary to decipher mutagenic processes operative in human cancers.

5.2. MEF immortalization assay to identify and test putative cancer driver events

Human cells acquire genetic alterations during their lifetime. It is widely acknowledged that successive genetic changes, providing comparative growth advantage to a cell clone ('drivers'), are the causes of tumour development and progression. It is relatively difficult to identify driver mutations among the numerous alterations found in tumours. Most of the mutations found in tumours are believed to be 'passenger' (not contributing to the fitness of a cell clone). MEF cells grow in culture until they reach senescence barrier and crisis. Some cells from the culture are able to bypass senescence, often due to mutations in the *Tp53* gene, or due to other genetic alterations. This process is analogous to human cancer development and is much more easily achieved in mouse than in human, due to the lack of replicative senescence barrier in mouse cells.

Data from genome-wide sequencing of thousands of human tumours are readily accessible from public repositories and can be used to identify driver alterations. Multiple programs were developed, based on various assumptions, to identify

putative cancer driver genes in these data. The programs perform well at identifying well-known frequently-mutated driver genes included in the Cancer Gene Census, but give very different predictions for drivers mutated with low frequency. Studies on mutation landscapes of various cancer types are published repeatedly. But the prediction of driver genes usually lack experimental validation, thus providing little mechanistic insight.

WES of MEF cell lines identified nonsynonymous (i.e. potentially functional) mutations in genes listed in the Cancer Gene Census, including well-known driver mutations – Hras^{Q61L} and Kras^{Q61R}. These alterations lock Ras genes in active state, causing constant pro-survival and pro-proliferation signalling, and are particularly important in cells in environment with low level of mitogenic signals. Inhibition of Ras signalling in Ras-mutated MEF cell clones led to lower level of cell survival in serum-deprived medium, experimentally validating the driver properties of activating Ras mutations in Hupki MEF cell lines.

Mutations in known cancer driver genes are not automatically functionally relevant for the pathophysiology of the malignant process. This Thesis presents a scoring system to narrow down mutations which could potentially contribute to the cancer-like phenotype of the immortalized MEF cells. The system is based on rational assumptions about the nature of coding driver mutations (exposure-predominant mutation type, predicted functional impact on the protein, present in an adequate AF). Actually, in a recent study, ratiometric methods have been found to be more reliable in identifying driver genes than methods based on mutation rate (Tokheim et al., 2016). When applied to mutations from two cell lines, the scoring system identified the well-known cancer driver mutations in *Hras* and *Tp53*, but also in other genes, among them the *Smarcc1* and *Smarcd2*, which produce the subunits of the BAF chromatin remodelling complex. The subsequent experiments, described in this Thesis, demonstrated that these mutations sensitise cells to the Ezh2 inhibitor in an (activated) Ras-dependent fashion (Kim et al., 2015). The cell line AA_2-1, bearing *Smarcc1* and activating *Hras* mutation, was much more resistant than *Smarcd2* mutant and Ras-wt cell clone MNNG_4-2. The same order of sensitivity was observed in validation set of 4 cell lines (one Ras and BAF mutant, three BAF

mutants), though the concentrations of the treatment were different, potentially due to the change of serum provider and inhibitor lot between the two experiments.

It has not been resolved, in MEF and human cell lines, if the resistance to Ezh2 inhibition in Ras mutants is due to the fact that the BAF mutations are functional, but the effect is masked by the activated Ras, or if the mutations are not drivers in any context (i.e. they are passenger), or if the mutations are not functional in the mutant Ras context, but would be driver if the cell clones developed in a genetic context without activated Ras. This can be tested by combination of genetic manipulations by CRISPR/Cas9 system and treatment with small molecule inhibitors. In general, putative drivers identified in MEF cell line data can be readily experimentally examined in the context in which they developed. This is a big advantage in contrast with putative driver events identified in human tumour sequencing studies. If there had been any experimental validation, it was done in human cell lines which are usually established from aggressive tumours and could yield a quite different result than if the mutation was tested in the original network (Puente et al., 2015, Guichard et al., 2012).

The variability in cancer alterations can be reduced by analysing it from a higher perspective: pathways, processes, protein complexes. Cancer variants usually alternate the physiological function of the higher-level process, providing a specific advantage to the tumour. The data presented in this Thesis show that pathways affected in MEF cell lines resemble to those altered in human cancer. As in human cancer, the subunits of BAF complex were mutated in ~30% of MEF cell lines in a mutually-exclusive pattern. The analysis of mutations in MEF cell lines also led to the discovery of mutually-exclusive mutations in TIP60 complex subunits Ep400 and Trrap. Thus, MEF immortalization assay can be useful in identification (and testing) new mutational patterns with potential role in human cancer.

This Thesis presents a genetic view on cancer development, placing in the centre the nonsynonymous mutations in protein-coding genes. We presented a proof-of-principle prioritization analysis of genes included in the Cancer Gene Census, and of epigenetic modifiers, which were of our interest. The scoring was done manually, and therefore it was not possible to evaluate all nonsynonymous

mutations in the 25 cell lines of the test set. Developing a program based on the scoring system would be very helpful. Moreover, previous research suggested that synonymous mutations, which are generally flatly considered as passengers, also could have some function in tumour development (Supek et al., 2014). Here are, thus, the limitations of our scoring system, which complies with the general paradigm.

Moreover, protein-coding genes constitute only a small fraction of the genome, though an important one. The rest of the genome is not 'junk', as it was once thought, but contains regulatory sequences, genes encoding regulatory ribonucleic acids ('non-coding genes') and many other potentially important elements. WES is sufficient to generate data to extract mutational signatures (Nik-Zainal et al., 2015, Olivier et al., 2014). However, WGS also provides data about mutations in non-coding sequences which could be potentially important for the cancer-like phenotype. Furthermore, tumours display deregulation on multiple layers: genetic, epigenetic, transcriptional, proteomic. The Thesis only provides data on a part of the genetic variation. It was already demonstrated that spontaneously immortalized Hupki cell lines resemble human cancers in terms of methylation and transcription (Tommasi et al., 2013). It would be interesting to examine these and other levels of complexity in cell lines present in this Thesis. Importantly, the cell lines have been made available to the research community as a resource for investigation of cancer development.

Surely, there are limitations in terms of the type of research to which the MEF cell lines can serve. Tumours are kind of organs of themselves. They start from a cell clone with specific properties, but as they develop, they have to organize and adapt (eg. induce vascularization to get enough nutrition, manipulate immune cells to evade immune response). In the end, a tumour is composed not only from the tumorigenic cell clone, but also from other cell types which serve to the tumour development (pericytes, cancer-associated fibroblasts, immune cells). MEF cell line assay in its current state cannot be used to study the interplay between different tumour cell types. However, it can provide an insight into the tumour founding-clone biology: alterations in cell cycle and apoptosis, epigenetics, metabolism, migration properties. Other tools, such as tissue and cancer organoids, are helpful in studying characteristics of a tumour on the organ level. The technology is very well

established for organs and tumours of the gastrointestinal system, but not so for other organs and tissues of the body. According to our data, MEF cell lines come from fibroblasts, but, from genetic point of view, do not resemble any specific cancer type.

Integration of research from MEF immortalization assay, organoids and human tumour cell lines is needed to provide a plastic picture of processes important for cancer development.

6. CONCLUSIONS

- a. MEF immortalization under mutagen treatment, coupled with massively parallel sequencing, produces mutation spectra expected from the mutagens used in the assay. This is true for the tested mutagens: AA, AID, AFB1, B[a]P, MNNG, and , to some extent, for the UVC. Mutation spectra of spontaneously-immortalized MEF cell lines had high proportion of C>G mutations which was not expected.
- b. WES of immortalized MEF cell lines produced sufficient data to extract mutational signatures. MEF mutational signatures of AA and MNNG closely resembled their corresponding signatures extracted from human tumours, as did the MEF mutational signatures of AID and B[a]P. The MEF mutational signature of AFB1 did not closely resemble the one extracted from human tumours presumably exposed to aflatoxins, which could be due to differences between mouse and human metabolism or DNA repair and/or more complex aflatoxin composition in the real-life human exposure. MEF UVC and spontaneous signatures did not resemble any COSMIC signature.
- c. MEF cell lines bore *Tp53* and *Ras* mutations typical for human cancers. Numerous genes included in the Cancer Gene Census were mutated by nonsynonymous SBS, as well as many genes involved in regulation of the epigenome. Furthermore, MEF cell lines displayed cancer-like mutation profile in terms of affected pathways, as well as alterations in the BAF and TIP60 chromatin-modifying complexes. Subunits of BAF complex were mutated in 9 out of 25 MEF cell lines in a mutual exclusive manner, as described in human cancers earlier (Gonzalez-Perez et al., 2013, Leiserson et al., 2015). Similarly, Ep400 and Trrap subunits of TIP60 histone acetyltransferase complex were nearly mutually exclusively mutated in 7 out of 25 MEF cell lines. The pattern was confirmed in sequencing data from human tumours, describing for the first time this feature of human cancers.

- d. Driver mutations can be identified in MEF cell lines. We devised a scoring system to identify mutations potentially driving the cancer-like phenotype of MEF cell lines among the wealth of mutations identified by WES. The scoring system is based on mutation type, AF and prediction of functional effect of the mutation. When applied on the Cancer Gene Census and epimodifier genes from two cell lines, the algorithm identified known Tp53 and Ras driver mutations, but also mutations in Smarcc1 and Smarcd2 BAF complex subunits, which were not previously identified as driver genes in human tumour sequencing studies.
- e. MEF cell lines, in contrary to human tumours, permit *in vitro* manipulations, and thus functional testing of putative driver mutations. Inhibition of Ras signalling by Mek inhibitor in Ras-mutant cell lines led to decreased cell viability in serum-deprived medium, while this effect was not observed for Ras-wt cell lines. Inhibition of Ezh2 activity in BAF-mutant cell lines led to elimination of the cells, in contrary to Ras/BAF double mutants. This was shown earlier in human cancer cell lines (Kim et al., 2015). This Thesis extends the list of BAF complex mutations conferring a vulnerability to Ezh2 inhibition for Smarcd2^{G166E}.

7. FUTURE DIRECTIONS AND CONTEXT

The proof-of-principle experiments described in this Thesis demonstrate that MEF immortalization assay recapitulates mutational signatures of human cancers and can be used for identification of mutational signatures of putative or known carcinogens.

Various reports indicate that proportion of cancer cases attributable to environmental risk factors is 60-90 %. These cancers can be prevented by avoiding the exposure. Firstly, however, the populations at risk have to be identified. Analysis of mutational signatures has already proven helpful in redefining the problem of the AA exposure, providing support for the notion that the problem is not specific to the Danube river basin, but is widespread in South-East Asia as well. Thus, mutational signature analysis has become a useful tool for molecular epidemiology. Also, a number of mutational signatures have been linked to DNA repair deficiencies. Understanding the causes behind these signatures and their effects can inform therapy decisions (Torgovnick and Schumacher, 2015, Pilati et al., 2017). MEF immortalization assay coupled with massively parallel sequencing has already been incorporated into collaborative projects on mutational signatures of various compounds, and the data are expected to support the evaluation process in the IARC Monographs program.

Identification of putative cancer driver mutations is a common type of analysis included in genome-wide sequencing studies of tumours; however, experimental validation is rather rare, and usually performed in tumour cell lines, which can be rather different from the tumours which served for the driver discovery. MEF cell lines are sequenced in a relatively early stage of cancer-like development which facilitates driver identification, and the testing of putative drivers can be done directly in the cell line, in the original mutation landscape. The original combination of destabilized networks can make a difference, as was presented on the example of Ras-BAF interplay.

I propose that the results presented in this Thesis directly inspire further research. One possible direction consists of better characterization of the cell lines in terms of other 'omics' approaches (genomics, epigenomics, transcriptomics, proteomics), potentially identifying new multi-level patterns of deregulation that could be then

tested in human data. Second line of investigations can be based on detailed testing of selected putative driver events, for example the effects of Ep400 and Trrap mutations on physiology of the affected cell clone. Gene silencing, CRISPR/Cas9 gene editing and small molecule inhibitors are well-developed tools which can be used to answer research questions of this kind.

Together, MEF cell lines presented in this Thesis are a unique resource, freely available to other researches, with the potential to facilitate new discoveries in the biology and physiology of cancer.

8. REFERENCES

Articles

- Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P. J., Vineis, P., Phillips, D. H. and Stratton, M. R. (2016) 'Mutational signatures associated with tobacco smoking in human cancer', *Science*, 354(6312), pp. 618-622.
- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A. L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Imielinski, M., Jäger, N., Jones, D. T., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichten, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., Stratton, M. R., Initiative, A. P. C. G., Consortium, I. B. C., Consortium, I. M.-S. and PedBrain, I. (2013a) 'Signatures of mutational processes in human cancer', *Nature*, 500(7463), pp. 415-21.
- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. and Stratton, M. R. (2013b) 'Deciphering signatures of mutational processes operative in human cancer', *Cell Rep*, 3(1), pp. 246-59.
- Alpert, M. E., Hutt, M. S. and Davidson, C. S. (1968) 'Hepatoma in Uganda. A study in geographic pathology', *Lancet*, 1(7555), pp. 1265-7.
- 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, pp. 170.
- Armstrong, C. A. and Tomita, K. (2017) 'Fundamental mechanisms of telomerase action in yeasts and mammals: understanding telomeres and telomerase in cancer cells', *Open Biol*, 7(3).
- Avery, O. T., Macleod, C. M. and McCarty, M. (1944) 'STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III', *J Exp Med*, 79(2), pp. 137-58.
- Balmain, A. and Pragnell, I. B. (1983) 'Mouse skin carcinomas induced in vivo by chemical carcinogens have a transforming Harvey-ras oncogene', *Nature*, 303(5912), pp. 72-4.

- BARNES, J. and BUTLER, W. H. (1964) 'CARCINOGENIC ACTIVITY OF AFLATOXIN TO RATS', *Nature*, 202, pp. 1016.
- Besaratinia, A., Li, H., Yoon, J. I., Zheng, A., Gao, H. and Tommasi, S. (2012) 'A high-throughput next-generation sequencing-based method for detecting the mutational fingerprint of carcinogens', *Nucleic Acids Res*, 40(15), pp. e116.
- Besaratinia, A. and Pfeifer, G. P. (2010) 'Applications of the human p53 knock-in (Hupki) mouse model for human carcinogen testing', *FASEB J*, 24(8), pp. 2612-9.
- Bieging, K. T., Mello, S. S. and Attardi, L. D. (2014) 'Unravelling mechanisms of p53-mediated tumour suppression', *Nat Rev Cancer*, 14(5), pp. 359-70.
- Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., Nijman, I. J., Martincorena, I., Mokry, M., Wiegerinck, C. L., Middendorp, S., Sato, T., Schwank, G., Nieuwenhuis, E. E., Verstegen, M. M., van der Laan, L. J., de Jonge, J., IJzermans, J. N., Vries, R. G., van de Wetering, M., Stratton, M. R., Clevers, H., Cuppen, E. and van Boxtel, R. (2016) 'Tissue-specific mutation accumulation in human adult stem cells during life', *Nature*, 538(7624), pp. 260-264.
- Bouaoun, L., Sonkin, D., Ardin, M., Hollstein, M., Byrnes, G., Zavadil, J. and Olivier, M. (2016) 'TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data', *Hum Mutat*, 37(9), pp. 865-76.
- Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M. and Snyder, M. (2012) 'Annotation of functional variation in personal genomes using RegulomeDB', *Genome Res*, 22(9), pp. 1790-7.
- Bryne, J. C., Valen, E., Tang, M. H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B. and Sandelin, A. (2008) 'JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update', *Nucleic Acids Res*, 36(Database issue), pp. D102-6.
- Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C. and Schultz, N. (2012) 'The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data', *Cancer Discov*, 2(5), pp. 401-4.
- 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., Getz, G. and Gordenin, D. A. (2015) 'An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers', *Nat Genet*, 47(9), pp. 1067-72.
- 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(6326), pp. 752-756.
- Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. and Getz, G. (2013) 'Sensitive

- detection of somatic point mutations in impure and heterogeneous cancer samples', *Nat Biotechnol*, 31(3), pp. 213-9.
- Cleaver, J. E. (1968) 'Defective repair replication of DNA in xeroderma pigmentosum', *Nature*, 218(5142), pp. 652-6.
- Cohn, O., Chen, A., Feldman, M. and Levy, D. (2016) 'Proteomic analysis of SETD6 interacting proteins', *Data Brief*, 6, pp. 799-802.
- Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J. and Elledge, S. J. (2013) 'Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome', *Cell*, 155(4), pp. 948-62.
- Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., Wilson, R. K. and Ding, L. (2012) 'MuSiC: identifying mutational significance in cancer genomes', *Genome Res*, 22(8), pp. 1589-98.
- DOLL, R. (1955) 'Mortality from lung cancer in asbestos workers', *Br J Ind Med*, 12(2), pp. 81-6.
- DOLL, R. and HILL, A. B. (1950) 'Smoking and carcinoma of the lung; preliminary report', *Br Med J*, 2(4682), pp. 739-48.
- Drabløs, F., Feyzi, E., Aas, P. A., Vaagbø, C. B., Kavli, B., Bratlie, M. S., Peña-Díaz, J., Otterlei, M., Slupphaug, G. and Krokan, H. E. (2004) 'Alkylation damage in DNA and RNA--repair mechanisms and medical significance', *DNA Repair (Amst)*, 3(11), pp. 1389-407.
- Endo, Y., Marusawa, H., Kou, T., Nakase, H., Fujii, S., Fujimori, T., Kinoshita, K., Honjo, T. and Chiba, T. (2008) 'Activation-induced cytidine deaminase links between inflammation and the development of colitis-associated colorectal cancers', *Gastroenterology*, 135(3), pp. 889-98, 898.e1-3.
- Eom, G. H., Kim, K. B., Kim, J. H., Kim, J. Y., Kim, J. R., Kee, H. J., Kim, D. W., Choe, N., Park, H. J., Son, H. J., Choi, S. Y., Kook, H. and Seo, S. B. (2011) 'Histone methyltransferase SETD3 regulates muscle differentiation', *J Biol Chem*, 286(40), pp. 34733-42.
- Feinberg, A. P., Koldobskiy, M. A. and Göndör, A. (2016) 'Epigenetic modulators, modifiers and mediators in cancer aetiology and progression', *Nat Rev Genet*, 17(5), pp. 284-99.
- 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(2), pp. 163-8.
- Fischer, A., Illingworth, C. J., Campbell, P. J. and Mustonen, V. (2013) 'EMu: probabilistic inference of mutational processes and their localization in the cancer genome', *Genome Biol*, 14(4), pp. R39.
- Fujiki, H. (2014) 'Gist of Dr. Katsusaburo Yamagiwa's papers entitled "Experimental study on the pathogenesis of epithelial tumors" (I to VI reports)', *Cancer Sci*, 105(2), pp. 143-9.

- 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(3), pp. 177-83.
- Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C. and Schultz, N. (2013) 'Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal', *Sci Signal*, 6(269), pp. p11.
- García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F. and Conesa, A. (2012) 'Qualimap: evaluating next-generation sequencing alignment data', *Bioinformatics*, 28(20), pp. 2678-9.
- Gaujoux, R. and Seoighe, C. (2010) 'A flexible R package for nonnegative matrix factorization', *BMC Bioinformatics*, 11, pp. 367.
- Golbabapour, S., Majid, N. A., Hassandarvish, P., Hajrezaie, M., Abdulla, M. A. and Hadi, A. H. (2013) 'Gene silencing and Polycomb group proteins: an overview of their structure, mechanisms and phylogenetics', *OMICS*, 17(6), pp. 283-96.
- Gonzalez-Perez, A., Jene-Sanz, A. and Lopez-Bigas, N. (2013) 'The mutational landscape of chromatin regulatory factors across 4,623 tumor samples', *Genome Biol*, 14(9), pp. r106.
- Green, C. L., Loechler, E. L., Fowler, K. W. and Essigmann, J. M. (1984) 'Construction and characterization of extrachromosomal probes for mutagenesis by carcinogens: site-specific incorporation of O6-methylguanine into viral and plasmid genomes', *Proc Natl Acad Sci U S A*, 81(1), pp. 13-7.
- Grollman, A. P., Shibutani, S., Moriya, M., Miller, F., Wu, L., Moll, U., Suzuki, N., Fernandes, A., Rosenquist, T., Medverec, Z., Jakovina, K., Brdar, B., Slade, N., Turesky, R. J., Goodenough, A. K., Rieger, R., Vukelić, M. and Jelaković, B. (2007) 'Aristolochic acid and the etiology of endemic (Balkan) nephropathy', *Proc Natl Acad Sci U S A*, 104(29), pp. 12129-34.
- Grozeva, D., Carss, K., Spasic-Boskovic, O., Parker, M. J., Archer, H., Firth, H. V., Park, S. M., Canham, N., Holder, S. E., Wilson, M., Hackett, A., Field, M., Floyd, J. A., Hurles, M., Raymond, F. L. and Consortium, U. K. (2014) 'De novo loss-of-function mutations in SETD5, encoding a methyltransferase in a 3p25 microdeletion syndrome critical region, cause intellectual disability', *Am J Hum Genet*, 94(4), pp. 618-24.
- Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I. B., Calderaro, J., Bioulac-Sage, P., Letexier, M., Degos, F., Clément, B., Balabaud, C., Chevet, E., Laurent, A., Couchy, G., Letouzé, E., Calvo, F. and Zucman-Rossi, J. (2012) 'Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma', *Nat Genet*, 44(6), pp. 694-8.
- Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of cancer: the next generation', *Cell*, 144(5), pp. 646-74.

- Harris, C. C., Genta, V. M., Frank, A. L., Kaufman, D. G., Barrett, L. A., McDowell, E. M. and Trump, B. F. (1974) 'Carcinogenic polynuclear hydrocarbons bind to macromolecules in cultured human bronchi', *Nature*, 252(5478), pp. 68-9.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics 2 edn.: Springer.
- Helleday, T., Petermann, E., Lundin, C., Hodgson, B. and Sharma, R. A. (2008) 'DNA repair pathways as targets for cancer therapy', *Nat Rev Cancer*, 8(3), pp. 193-204.
- Helming, K. C., Wang, X. and Roberts, C. W. (2014) 'Vulnerabilities of mutant SWI/SNF complexes in cancer', *Cancer Cell*, 26(3), pp. 309-17.
- Hirano, T. (2012) 'Condensins: universal organizers of chromosomes with diverse functions', *Genes Dev*, 26(15), pp. 1659-78.
- Hoffman, M. M. and Birney, E. (2010) 'An effective model for natural selection in promoters', *Genome Res*, 20(5), pp. 685-92.
- 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(2), pp. 158-167.
- 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*, 753(1), pp. 41-9.
- Hollstein, M., Rice, K., Greenblatt, M. S., Soussi, T., Fuchs, R., Sørli, T., Hovig, E., Smith-Sørensen, B., Montesano, R. and Harris, C. C. (1994) 'Database of p53 gene somatic mutations in human tumors and cell lines', *Nucleic Acids Res*, 22(17), pp. 3551-5.
- Hollstein, M., Sidransky, D., Vogelstein, B. and Harris, C. C. (1991) 'p53 mutations in human cancers', *Science*, 253(5015), pp. 49-53.
- Huang, d. W., Sherman, B. T. and Lempicki, R. A. (2009) 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nat Protoc*, 4(1), pp. 44-57.
- Huang, M. N., Yu, W., Teoh, W. W., Ardin, M., Jusakul, A., Ng, A., Boot, A., Abedi-Ardekani, B., Villar, S., Myint, S. S., Othman, R., Poon, S. L., Heguy, A., Olivier, M., Hollstein, M., Tan, P., Teh, B. T., Sabapathy, K., Zavadil, J. and Rozen, S. (2017) 'Genome-Scale Mutational Signatures Of Aflatoxin In Cells, Mice And Human Tumors', *bioRxiv*, Accepted for publication in Genome Research.
- Hutchins, L. N., Murphy, S. M., Singh, P. and Graber, J. H. (2008) 'Position-dependent motif characterization using non-negative matrix factorization', *Bioinformatics*, 24(23), pp. 2684-90.
- Jahan, S. and Davie, J. R. (2015) 'Protein arginine methyltransferases (PRMTs): role in chromatin organization', *Adv Biol Regul*, 57, pp. 173-84.

- Jaworski, M., Hailfinger, S., Buchmann, A., Hergenbahn, M., Hollstein, M., Ittrich, C. and Schwarz, M. (2005) 'Human p53 knock-in (hupki) mice do not differ in liver tumor response from their counterparts with murine p53', *Carcinogenesis*, 26(10), pp. 1829-34.
- 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', *Int J Cancer*, 136(12), pp. 2967-72.
- Jelaković, B., Karanović, S., Vuković-Lela, I., Miller, F., Edwards, K. L., Nikolić, J., Tomić, K., Slade, N., Brdar, B., Turesky, R. J., Stipančić, Ž., Dittrich, D., Grollman, A. P. and Dickman, K. G. (2012) 'Aristolactam-DNA adducts are a biomarker of environmental exposure to aristolochic acid', *Kidney Int*, 81(6), pp. 559-67.
- Kadoch, C., Copeland, R. A. and Keilhack, H. (2016) 'PRC2 and SWI/SNF Chromatin Remodeling Complexes in Health and Disease', *Biochemistry*, 55(11), pp. 1600-14.
- Kadoch, C. and Crabtree, G. R. (2015) 'Mammalian SWI/SNF chromatin remodeling complexes and cancer: Mechanistic insights gained from human genomics', *Sci Adv*, 1(5), pp. e1500447.
- Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., Haradhvala, N., Lawrence, M. S., Kiezun, A., Fernandes, S. M., Bahl, S., Sougnez, C., Gabriel, S., Lander, E. S., Kim, H. T., Getz, G. and Brown, J. R. (2015) 'Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution', *Nat Commun*, 6, pp. 8866.
- Khodabakhshi, A. H., Morin, R. D., Fejes, A. P., Mungall, A. J., Mungall, K. L., Bolger-Munro, M., Johnson, N. A., Connors, J. M., Gascoyne, R. D., Marra, M. A., Birol, I. and Jones, S. J. (2012) 'Recurrent targets of aberrant somatic hypermutation in lymphoma', *Oncotarget*, 3(11), pp. 1308-19.
- Kim, K. H., Kim, W., Howard, T. P., Vazquez, F., Tsherniak, A., Wu, J. N., Wang, W., Haswell, J. R., Walensky, L. D., Hahn, W. C., Orkin, S. H. and Roberts, C. W. (2015) 'SWI/SNF-mutant cancers depend on catalytic and non-catalytic activity of EZH2', *Nat Med*, 21(12), pp. 1491-6.
- Knudson, A. G. (1971) 'Mutation and cancer: statistical study of retinoblastoma', *Proc Natl Acad Sci U S A*, 68(4), pp. 820-3.
- Komori, J., Marusawa, H., Machimoto, T., Endo, Y., Kinoshita, K., Kou, T., Haga, H., Ikai, I., Uemoto, S. and Chiba, T. (2008) 'Activation-induced cytidine deaminase links bile duct inflammation to human cholangiocarcinoma', *Hepatology*, 47(3), pp. 888-96.
- Kucab, J. E., Hollstein, M., Arlt, V. M. and Phillips, D. H. (2017) 'Nutlin-3a selects for cells harbouring TP53 mutations', *Int J Cancer*, 140(4), pp. 877-887.
- Kumar, P., Mahato, D. K., Kamle, M., Mohanta, T. K. and Kang, S. G. (2016) 'Aflatoxins: A Global Concern for Food Safety, Human Health and Their Management', *Front Microbiol*, 7, pp. 2170.

- Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S. and Getz, G. (2014) 'Discovery and saturation analysis of cancer genes across 21 tumour types', *Nature*, 505(7484), pp. 495-501.
- Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D. A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S. and Getz, G. (2013) 'Mutational heterogeneity in cancer and the search for new cancer-associated genes', *Nature*, 499(7457), pp. 214-8.
- Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., Lawrence, M. S., Gonzalez-Perez, A., Tamborero, D., Cheng, Y., Ryslik, G. A., Lopez-Bigas, N., Getz, G., Ding, L. and Raphael, B. J. (2015) 'Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes', *Nat Genet*, 47(2), pp. 106-14.
- Levy, D. D., Groopman, J. D., Lim, S. E., Seidman, M. M. and Kraemer, K. H. (1992) 'Sequence specificity of aflatoxin B1-induced mutations in a plasmid replicated in xeroderma pigmentosum and DNA repair proficient human cells', *Cancer Res*, 52(20), pp. 5668-73.
- Lindahl, T. and Andersson, A. (1972) 'Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid', *Biochemistry*, 11(19), pp. 3618-23.
- Lindahl, T. and Nyberg, B. (1972) 'Rate of depurination of native deoxyribonucleic acid', *Biochemistry*, 11(19), pp. 3610-8.
- Liu, Z., Belharazem, D., Muehlbauer, K. R., Nedelko, T., Knyazev, Y. and Hollstein, M. (2007) 'Mutagenesis of human p53 tumor suppressor gene sequences in embryonic fibroblasts of genetically-engineered mice', *Genet Eng (N Y)*, 28, pp. 45-54.
- 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(9), pp. 2963-8.
- 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(7), pp. 2583-7.
- Loeb, L. A., Springgate, C. F. and Battula, N. (1974) 'Errors in DNA replication as a basis of malignant changes', *Cancer Res*, 34(9), pp. 2311-21.

- Loechler, E. L., Green, C. L. and Essigmann, J. M. (1984) 'In vivo mutagenesis by O6-methylguanine built into a unique site in a viral genome', *Proc Natl Acad Sci U S A*, 81(20), pp. 6271-5.
- Luo, J. L., Tong, W. M., Yoon, J. H., Hergenbahn, M., Koomagi, R., Yang, Q., Galendo, D., Pfeifer, G. P., Wang, Z. Q. and Hollstein, M. (2001a) '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(22), pp. 8158-63.
- Luo, J. L., Yang, Q., Tong, W. M., Hergenbahn, M., Wang, Z. Q. and Hollstein, M. (2001b) '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(3), pp. 320-8.
- Manjanatha, M. G., Guo, L. W., Shelton, S. D. and Doerge, D. R. (2015) 'Acrylamide-induced carcinogenicity in mouse lung involves mutagenicity: cll gene mutations in the lung of big blue mice exposed to acrylamide and glycidamide for up to 4 weeks', *Environ Mol Mutagen*, 56(5), pp. 446-56.
- Matsumoto, Y., Marusawa, H., Kinoshita, K., Endo, Y., Kou, T., Morisawa, T., Azuma, T., Okazaki, I. M., Honjo, T. and Chiba, T. (2007) 'Helicobacter pylori infection triggers aberrant expression of activation-induced cytidine deaminase in gastric epithelium', *Nat Med*, 13(4), pp. 470-6.
- Maul, R. W. and Gearhart, P. J. (2010) 'AID and somatic hypermutation', *Adv Immunol*, 105, pp. 159-91.
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P. and Cunningham, F. (2016) 'The Ensembl Variant Effect Predictor', *Genome Biol*, 17(1), pp. 122.
- Murr, R., Vaissière, T., Sawan, C., Shukla, V. and Herceg, Z. (2007) 'Orchestration of chromatin-based processes: mind the TRRAP', *Oncogene*, 26(37), pp. 5358-72.
- 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(4), pp. 987-90.
- Ng, P. C. and Henikoff, S. (2001) 'Predicting deleterious amino acid substitutions', *Genome Res*, 11(5), pp. 863-74.
- Nielsen, F. C., van Overeem Hansen, T. and Sørensen, C. S. (2016) 'Hereditary breast and ovarian cancer: new genes in confined pathways', *Nat Rev Cancer*, 16(9), pp. 599-612.
- Nik-Zainal, S., Kucab, J. E., Morganella, S., Glodzik, D., Alexandrov, L. B., Arlt, V. M., Wenginger, A., Hollstein, M., Stratton, M. R. and Phillips, D. H. (2015) 'The genome as a record of environmental exposure', *Mutagenesis*, 30(6), pp. 763-70.
- NORDLING, C. O. (1953) 'A new theory on cancer-inducing mechanism', *Br J Cancer*, 7(1), pp. 68-72.

- Odell, A., Askham, J., Whibley, C. and Hollstein, M. (2010) 'How to become immortal: let MEFs count the ways', *Aging (Albany NY)*, 2(3), pp. 160-5.
- Odell, A. F., Odell, L. R., Askham, J. M., Alogheli, H., Ponnambalam, S. and Hollstein, M. (2013) 'A novel p53 mutant found in iatrogenic urothelial cancers is dysfunctional and can be rescued by a second-site global suppressor mutation', *J Biol Chem*, 288(23), pp. 16704-14.
- Ohtani, N., Takahashi, A., Mann, D. J. and Hara, E. (2012) 'Cellular senescence: a double-edged sword in the fight against cancer', *Exp Dermatol*, 21 Suppl 1, pp. 1-4.
- Okazaki, I. M., Hiai, H., Kakazu, N., Yamada, S., Muramatsu, M., Kinoshita, K. and Honjo, T. (2003) 'Constitutive expression of AID leads to tumorigenesis', *J Exp Med*, 197(9), pp. 1173-81.
- Olivier, M., Hollstein, M. and Hainaut, P. (2010) 'TP53 mutations in human cancers: origins, consequences, and clinical use', *Cold Spring Harb Perspect Biol*, 2(1), pp. a001008.
- Olivier, M., Hollstein, M., Schmeiser, H. H., Straif, K. and Wild, C. P. (2012) 'Upper urinary tract urothelial cancers: where it is A:T', *Nat Rev Cancer*, 12(8), pp. 503-4.
- Olivier, M., Weninger, A., Ardin, M., Huskova, H., Castells, X., Vallée, M. P., McKay, J., Nedelko, T., Muehlbauer, K. R., Marusawa, H., Alexander, J., Hazelwood, L., Byrnes, G., Hollstein, M. and Zavadil, J. (2014) 'Modelling mutational landscapes of human cancers in vitro', *Sci Rep*, 4, pp. 4482.
- 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(8), pp. 741-7.
- Pasqualucci, L., Neumeister, P., Goossens, T., Nanjangud, G., Chaganti, R. S., Küppers, R. and Dalla-Favera, R. (2001) 'Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas', *Nature*, 412(6844), pp. 341-6.
- Peters, J. M., Tedeschi, A. and Schmitz, J. (2008) 'The cohesin complex and its roles in chromosome biology', *Genes Dev*, 22(22), pp. 3089-114.
- Pfeifer, G. P., You, Y. H. and Besaratinia, A. (2005) 'Mutations induced by ultraviolet light', *Mutat Res*, 571(1-2), pp. 19-31.
- 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., Bertherat, J., Letouzé, E. and Laurent-Puig, P. (2017) 'Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas', *J Pathol*, 242(1), pp. 10-15.
- 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., Varela, I., Nik-Zainal, S., Davies, H. R., Ordoñez, G. R., Mudie, L. J., Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J. W., Mangion, J., Sun, Y. A., McLaughlin, S. F., Peckham, H. E., Tsung, E. F.,

- Costa, G. L., Lee, C. C., Minna, J. D., Gazdar, A., Birney, E., Rhodes, M. D., McKernan, K. J., Stratton, M. R., Futreal, P. A. and Campbell, P. J. (2010) 'A small-cell lung cancer genome with complex signatures of tobacco exposure', *Nature*, 463(7278), pp. 184-90.
- 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., Ng, L. G., Weng, W. H., Chuang, C. K., Yuen, J. S., Pang, S. T., Tan, P., Teh, B. T. and Rozen, S. G. (2015) 'Mutation signatures implicate aristolochic acid in bladder cancer development', *Genome Med*, 7(1), pp. 38.
- 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., Chong, S. C., Gan, A., Tay, S. T., Lim, W. K., Cutcutache, I., Huang, D., Ler, L. D., Nairismägi, M. L., Lee, M. H., Chang, Y. H., Yu, K. J., Chan-On, W., Li, B. K., Yuan, Y. F., Qian, C. N., Ng, K. F., Wu, C. F., Hsu, C. L., Bunte, R. M., Stratton, M. R., Futreal, P. A., Sung, W. K., Chuang, C. K., Ong, C. K., Rozen, S. G., Tan, P. and Teh, B. T. (2013) 'Genome-wide mutational signatures of aristolochic acid and its application as a screening tool', *Sci Transl Med*, 5(197), pp. 197ra101.
- Prior, I. A., Lewis, P. D. and Mattos, C. (2012) 'A comprehensive survey of Ras mutations in cancer', *Cancer Res*, 72(10), pp. 2457-67.
- Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., Munar, M., Rubio-Pérez, C., Jares, P., Aymerich, M., Baumann, T., Beekman, R., Belver, L., Carrio, A., Castellano, G., Clot, G., Colado, E., Colomer, D., Costa, D., Delgado, J., Enjuanes, A., Estivill, X., Ferrando, A. A., Gelpí, J. L., González, B., González, S., González, M., Gut, M., Hernández-Rivas, J. M., López-Guerra, M., Martín-García, D., Navarro, A., Nicolás, P., Orozco, M., Payer, Á., Pinyol, M., Pisano, D. G., Puente, D. A., Queirós, A. C., Quesada, V., Romeo-Casabona, C. M., Royo, C., Royo, R., Rozman, M., Russiñol, N., Salaverría, I., Stamatopoulos, K., Stunnenberg, H. G., Tamborero, D., Terol, M. J., Valencia, A., López-Bigas, N., Torrents, D., Gut, I., López-Guillermo, A., López-Otín, C. and Campo, E. (2015) 'Non-coding recurrent mutations in chronic lymphocytic leukaemia', *Nature*, 526(7574), pp. 519-24.
- Reimand, J. and Bader, G. D. (2013) 'Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers', *Mol Syst Biol*, 9, pp. 637.
- Reimand, J., Wagih, O. and Bader, G. D. (2013) 'The mutational landscape of phosphorylation signaling in cancer', *Sci Rep*, 3, pp. 2651.
- 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(19), pp. 2788-94.
- Rogozin, I. B. and Kolchanov, N. A. (1992) 'Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis', *Biochim Biophys Acta*, 1171(1), pp. 11-8.

- Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. and da Silva, I. T. (2017) 'signeR: an empirical Bayesian approach to mutational signature discovery', *Bioinformatics*, 33(1), pp. 8-16.
- Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. and Swanton, C. (2016) 'DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution', *Genome Biol*, 17, pp. 31.
- Sage, E., Drobetsky, E. A. and Moustacchi, E. (1993) '8-Methoxypsoralen induced mutations are highly targeted at crosslinkable sites of photoaddition on the non-transcribed strand of a mammalian chromosomal gene', *EMBO J*, 12(2), pp. 397-402.
- Scelo, G., Riazalhosseini, Y., Greger, L., Letourneau, L., Gonzàlez-Porta, M., Wozniak, M. B., Bourgey, M., Harnden, P., Egevad, L., Jackson, S. M., Karimzadeh, M., Arseneault, M., Lepage, P., How-Kit, A., Daunay, A., Renault, V., Blanché, H., Tubacher, E., Sehmoun, J., Viksna, J., Celms, E., Opmanis, M., Zarins, A., Vasudev, N. S., Seywright, M., Abedi-Ardekani, B., Carreira, C., Selby, P. J., Cartledge, J. J., Byrnes, G., Zavadil, J., Su, J., Holcatova, I., Brisuda, A., Zaridze, D., Moukeria, A., Foretova, L., Navratilova, M., Mates, D., Jinga, V., Artemov, A., Nedoluzhko, A., Mazur, A., Rastorguev, S., Boulygina, E., Heath, S., Gut, M., Bihoreau, M. T., Lechner, D., Foglio, M., Gut, I. G., Skryabin, K., Prokhortchouk, E., Cambon-Thomsen, A., Rung, J., Bourque, G., Brennan, P., Tost, J., Banks, R. E., Brazma, A. and Lathrop, G. M. (2014) 'Variation in genomic landscape of clear cell renal cell carcinoma across Europe', *Nat Commun*, 5, pp. 5135.
- Schmeiser, H. H., Kucab, J. E., Arlt, V. M., Phillips, D. H., Hollstein, M., Gluhovschi, G., Gluhovschi, C., Modilca, M., Daminescu, L., Petrica, L. and Velcirov, S. (2012) 'Evidence of exposure to aristolochic acid in patients with urothelial cancer from a Balkan endemic nephropathy region of Romania', *Environ Mol Mutagen*, 53(8), pp. 636-41.
- Segovia, R., Tam, A. S. and Stirling, P. C. (2015) 'Dissecting genetic and environmental mutation signatures with model organisms', *Trends Genet*, 31(8), pp. 465-74.
- Serero, A., Jubin, C., Loeillet, S., Legoix-Né, P. and Nicolas, A. G. (2014) 'Mutational landscape of yeast mutator strains', *Proc Natl Acad Sci U S A*, 111(5), pp. 1897-902.
- 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 Genet Toxicol Environ Mutagen*, 775-776, pp. 48-54.
- Sidorenko, V. S., Yeo, J. E., Bonala, R. R., Johnson, F., Schäerer, O. D. and Grollman, A. P. (2012) 'Lack of recognition by global-genome nucleotide excision repair accounts for the high mutagenicity and persistence of aristolactam-DNA adducts', *Nucleic Acids Res*, 40(6), pp. 2494-505.

- Sporn, M. B., Dingman, C. W., Phelps, H. L. and Wogan, G. N. (1966) 'Aflatoxin B1: binding to DNA in vitro and alteration of RNA metabolism in vivo', *Science*, 151(3717), pp. 1539-41.
- Stampfer, M. R. and Bartley, J. C. (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(8), pp. 2394-8.
- Stanley, F. K., Moore, S. and Goodarzi, A. A. (2013) 'CHD chromatin remodelling enzymes and the DNA damage response', *Mutat Res*, 750(1-2), pp. 31-44.
- Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009) 'The cancer genome', *Nature*, 458(7239), pp. 719-24.
- Sun, T. T., Wu, S. M., Wu, Y. Y. and Chu, Y. R. (1985) 'Measurement of individual aflatoxin exposure among people having different risk to primary hepatocellular carcinoma', *Princess Takamatsu Symp*, 16, pp. 225-35.
- Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. and Lehner, B. (2014) 'Synonymous mutations frequently act as driver mutations in human cancers', *Cell*, 156(6), pp. 1324-35.
- Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A. and Skotheim, R. I. (2015) 'Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes', *Oncogene*.
- Tamborero, D., Gonzalez-Perez, A. and Lopez-Bigas, N. (2013a) 'OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes', *Bioinformatics*, 29(18), pp. 2238-44.
- Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandath, C., Reimand, J., Lawrence, M. S., Getz, G., Bader, G. D., Ding, L. and Lopez-Bigas, N. (2013b) 'Comprehensive identification of mutational cancer driver genes across 12 tumor types', *Sci Rep*, 3, pp. 2650.
- Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. and Karchin, R. (2016) 'Evaluating the evaluation of cancer driver genes', *Proc Natl Acad Sci U S A*, 113(50), pp. 14330-14335.
- Tomasetti, C., Li, L. and Vogelstein, B. (2017) 'Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention', *Science*, 355(6331), pp. 1330-1334.
- Tomasetti, C. and Vogelstein, B. (2015) 'Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions', *Science*, 347(6217), pp. 78-81.
- Tommasi, S., Zheng, A., Weninger, A., Bates, S. E., Li, X. A., Wu, X., Hollstein, M. and Besaratinia, A. (2013) 'Mammalian cells acquire epigenetic hallmarks of human cancer during immortalization', *Nucleic Acids Res*, 41(1), pp. 182-95.
- Torgovnick, A. and Schumacher, B. (2015) 'DNA repair mechanisms in cancer development and therapy', *Front Genet*, 6, pp. 157.
- Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J. and Jemal, A. (2015) 'Global cancer statistics, 2012', *CA Cancer J Clin*, 65(2), pp. 87-108.

- Trottier, Y., Waithe, W. I. and Anderson, A. (1992) 'Kinds of mutations induced by aflatoxin B1 in a shuttle vector replicating in human cells transiently expressing cytochrome P4501A2 cDNA', *Mol Carcinog*, 6(2), pp. 140-7.
- Viel, A., Bruselles, A., Meccia, E., Fornasarig, M., Quaia, M., Canzonieri, V., Policicchio, E., Urso, E. D., Agostini, M., Genuardi, M., Lucci-Cordisco, E., Venesio, T., Martayan, A., Diodoro, M. G., Sanchez-Mete, L., Stigliano, V., Mazzei, F., Grasso, F., Giuliani, A., Baiocchi, M., Maestro, R., Giannini, G., Tartaglia, M., Alexandrov, L. B. and Bignami, M. (2017) 'A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer', *EBioMedicine*.
- Vogelstein, B., Papadopoulos, N., Velculescu, V., Zhou, S., Diaz, L. and Kinzler, K. (2013) 'Cancer Genome Landscapes', *Science*, 339(6127), pp. 1546-1558.
- vom Brocke, J., Schmeiser, H. H., Reinbold, M. and Hollstein, M. (2006) 'MEF immortalization to investigate the ins and outs of mutagenesis', *Carcinogenesis*, 27(11), pp. 2141-7.
- Wang, K., Li, M. and Hakonarson, H. (2010) 'ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Res*, 38(16), pp. e164.
- WATSON, J. D. and CRICK, F. H. (1953) 'The structure of DNA', *Cold Spring Harb Symp Quant Biol*, 18, pp. 123-31.
- Wattanawaraporn, R., Woo, L. L., Belanger, C., Chang, S. C., Adams, J. E., Trudel, L. J., Bouhenguel, J. T., Egner, P. A., Groopman, J. D., Croy, R. G., Essigmann, J. M. and Wogan, G. N. (2012) 'A single neonatal exposure to aflatoxin b1 induces prolonged genetic damage in two loci of mouse liver', *Toxicol Sci*, 128(2), pp. 326-33.
- Westcott, P. M., 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., Fredlund, E., Quigley, D. A., Adams, D. J. and Balmain, A. (2015) 'The mutational landscapes of genetic and chemical models of Kras-driven lung cancer', *Nature*, 517(7535), pp. 489-92.
- 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(15), pp. 11326-35.
- Wild, C., Brennan, P., Plummer, M., Bray, F., Straif, K. and Zavadil, J. (2015) 'Cancer risk: role of chance overstated', *Science*, 347(6223), pp. 728.
- Wu, S., Powers, S., Zhu, W. and Hannun, Y. A. (2015) 'Substantial contribution of extrinsic risk factors to cancer development', *Nature*.
- Yasui, M., Kanemaru, Y., Kamoshita, N., Suzuki, T., Arakawa, T. and Honma, M. (2014) 'Tracing the fates of site-specifically introduced DNA adducts in the human genome', *DNA Repair (Amst)*, 15, pp. 11-20.

- Zhivagui, M., Korenjak, M. and Zavadil, J. (2016) 'Modelling Mutation Spectra of Human Carcinogens Using Experimental Systems', *Basic Clin Pharmacol Toxicol*.
- Zou, S., Li, J., Zhou, H., Frech, C., Jiang, X., Chu, J. S., Zhao, X., Li, Y., Li, Q., Wang, H., Hu, J., Kong, G., Wu, M., Ding, C., Chen, N. and Hu, H. (2014) 'Mutational landscape of intrahepatic cholangiocarcinoma', *Nat Commun*, 5, pp. 5696.
- Zámborszky, J., Szikriszt, B., Gervai, J. Z., Pipek, O., Póti, Á., Krzystanek, M., Ribli, D., Szalai-Gindl, J. M., Csabai, I., Szallasi, Z., Swanton, C., Richardson, A. L. and Szüts, D. (2017) 'Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions', *Oncogene*, 36(6), pp. 746-755.

URLs

- URL1: <http://cancer.sanger.ac.uk/cosmic/signatures>
URL2: <http://p53.iarc.fr/TP53SomaticMutations.aspx>
URL3: <http://p53.iarc.fr/TP53SomaticMutations>

SUPPLEMENTARY INFORMATION

List of publications

Articles related to the Thesis

Huskova H, Ardin M, Weninger A, Vargova K, Barrin S, Villar S, Olivier M, Stopka T, Herceg Z, Hollstein M, Zavadil J, Korenjak M. 2017. Modeling cancer driver events in vitro using barrier bypass-clonal expansion assays and massively parallel sequencing. *Oncogene* (accepted for publication). (**IF=7.932**)

Olivier M, Weninger A, Ardin M, **Huskova H**, Castells X, Vallée MP, McKay J, Nedelko T, Muehlbauer KR, Marusawa H, Alexander J, Hazelwood L, Byrnes G, Hollstein M and Zavadil J. 2014. Modelling mutational landscapes of human cancers in vitro. *Sci Rep* 4: 4482. (**IF=5.578**)

Articles not related to the Thesis

Huskova H, Korecka K, Karban J, Vargova J, Vargova K, Dusilkova N, Trneny M, Stopka T. 2015. Oncogenic microRNA-155 and its target PU.1: an integrative gene expression study in six of the most prevalent lymphomas. *Int J Hematol* 102: 441-450. (**IF=1.846**)

Supplementary Data and Figure legends

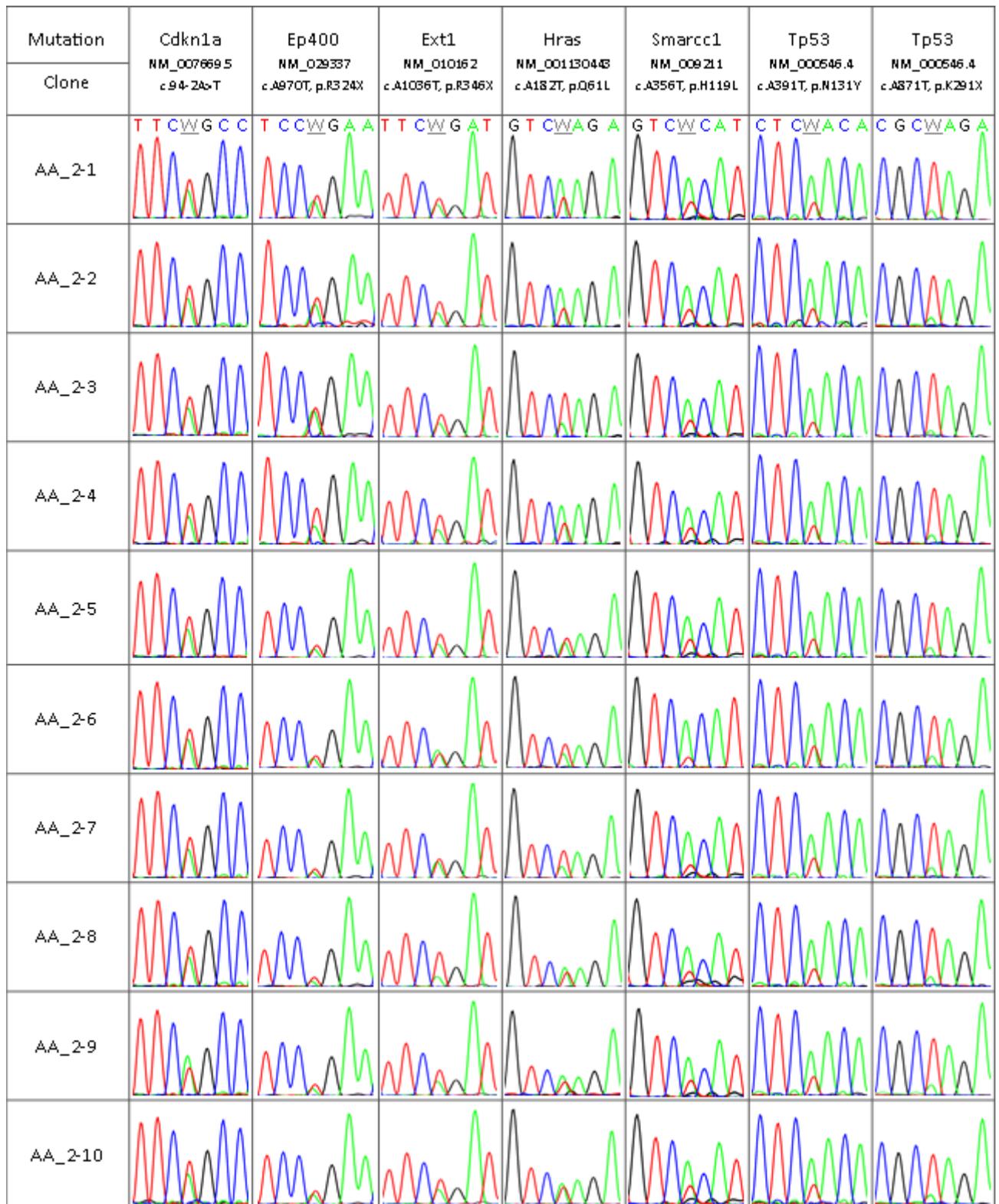
Supplementary Data 1: List of all mutations identified by WES in 26 Hupki MEF cell lines (25 of the test set and 1 control). As a .txt file on an electronic medium attached to this Thesis.

Supplementary Figure 1: Extended list of epigenetic modifiers mutated by nonsynonymous mutations in the 25 Hupki MEF cell lines. Yellow – exposure-specific mutation type, blue – other than exposure-specific mutation type, numbers in yellow and blue fields – mutation count. To be found below.

Supplementary Figure 2: Sanger sequencing of candidate driver mutations in clones derived from AA_2 cell line. DNA sequence is displayed above the chromatogram of the first clone using IUPAC code. To be found below.

Supplementary Figure 3: Sanger sequencing of candidate driver mutations in clones derived from MNNG_4 cell line. DNA sequence is displayed above the chromatogram of the first clone using IUPAC code. If a base is not identical in all clones, it is marked with the respective IUPAC letter in each respective chromatogram. To be found below.

Group	Gene	AA_1	AA_2	AA_3	AA_4	AA_5	AA_6	AA_7	AFB1_1	AFB1_2	AFB1_3	AID_1	AID_2	BaP_1	BaP_2	BaP_3	MNNG_1	MNNG_2	MNNG_3	MNNG_4	Spont_1	Spont_2	Spont_3	Spont_4	UVC_1	UVC_2	SUM	
ATP-dependent chromatin remodeling	Arid1a															1											1	
	Arid1b																2											3
	Arid2																					1	1					2
	Smarca2									1																		1
	Smarca1														1													1
	Smarcc1		1																									1
	Smarcd2																								1			1
	Rbbp7							1																				1
	Baz1a	1																										1
	Baz1b																1											1
	Baz2a																											1
	Baz2b																											1
	Chd1																											1
	Chd3																											1
	Chd4															1												1
	Chd5																											1
	Chd7																											1
Chd8															1												1	
Chd9	1																										1	
Gatad2a																											1	
DNA methylation	Aicda											1	1														2	
	Dnmt1														1													2
	Dnmt3a																											1
	Tet1																											1
	Tet2																											1
Histone Acetylation	Ep300																										1	
	Gtf3c4																										1	
	Kat6a																										1	
	Ka7																										1	
Histone Acetylation	Ep400	1	1																								4	
	Trrap														2												4	
Histone Deacetylation	Hdac10																										1	
	Hdac2																										1	
	Hdac6						1																				2	
	Hdac9												1														1	
	Ncor1																										1	
	Sirt5																										1	
	Sirt6																										1	
	Tbl1xr1																										2	
Histone Demethylation	Kdm1b	1																									1	
	Kdm2a																										1	
	Kdm3a																										1	
	Kdm3b		1																								1	
	Kdm4a																										1	
	Kdm4d																										1	
	Kdm6a																										1	
	Kdm6b																										1	
	Phf2																										1	
	Rbp2																										1	
Histone Methylation	Ash1l																										2	
	Ash2l																										1	
	Ehmt2																										1	
	Kmt2a																										1	
	Kmt2b																										1	
	Kmt2c																										6	
	Kmt2d																										4	
	Setd1a																										1	
	Setd1b																										2	
	Nsd1																											1
	Prdm9																										1	
	Prmt1																										1	
	Prmt7																										1	
Prmt8																										1		
Rtf1																										1		
Setd3																										2		
Setd5																										1		
Setd6																										1		
Setd7																										1		
Smyd1																										1		
Other	Bag6																										1	
	Lmna																										1	
	Mum1																										1	
PRC1	Cbx2																										1	
	Cbx7																										1	
	Ezh1																										1	
	L3mbtl1																										1	
	Phc2																										1	
PRC2	Ring1																										1	
	Asxl1																										1	
SUM		6	6	3	2	2	1	4	1	3	3																	



Supplementary Figure 2

