MEDICAL UNIVERSITY OF VIENNA

DNA damage and DNA repair in cancer genomes

Doctoral thesis at the Medical University of Vienna
for obtaining the academic degree

Doctor of Philosophy

Submitted by

Michel B.-B. Owusu, MSc

Supervisor:
Dr. Joanna Loizou
CeMM Research Center for Molecular Medicine
of the Austrian Academy of Sciences
Lazarettgasse 14, AKH BT25.3
1090 Vienna, Austria

Vienna, 08/2018

Declaration

The experimental part of the work that I present here was carried out in the laboratory of Dr. Joanna Loizou, DNA damage signaling group, CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria. The computational data analysis part was carried out in the laboratory of Dr. Serena Nik-Zainal, Mutation Signatures group, Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK and Department of Medical Genetics, The Clinical School, University of Cambridge, Cambridge, UK.

The project was conceived by Dr. Joanna Loizou, Prof. Stephen P. Jackson and Dr. Serena Nik-Zainal. The experimental part of the project was designed by Dr. Joanna Loizou and myself, with input from Mag. Marc Wiedner and Dr. Jana Stranska.

I performed the experimental work and was assisted by Mag. Marc Wiedner and Dr. Jana Stranska with the cell culture work. Dr. Xueqing Zou performed the computational data analysis. Dr. Xueqing Zou and I contributed equally to this work. Project Coordination assistance was received from Dr. Rebecca Harris at the lab of Prof. Stephen P. Jackson, The Gurdon Institute and Department of Biochemistry, University of Cambridge, Cambridge, UK.
Dr. Serena Nik-Zainal, Dr. Xueqing Zou, Dr. Joanna Loizou and myself wrote the manuscript, with input from all authors.

The resulting manuscript is my first-author publication that constitutes this thesis and has been reprinted according to the reprint and permission policies of the Nature Publishing Group. The article can be found in the Nature Communications journal:

Validating the concept of mutational signatures with isogenic cell models. Xueqing Zou#, Michel Owusu#, Rebecca Harris, Stephen P. Jackson, Joanna I. Loizou* & Serena Nik-Zainal*. Nature Communications volume 9, Article number: 1744 (2018).
https://doi.org/10.1038/s41467-018-04052-8
(# these authors contributed equally, * these authors are both corresponding authors)

I, the author, wrote all the chapters of this thesis.

Table of contents

List of figures

List of tables

Abstract

Cancer arises due to mutations in the genome that transform an otherwise healthy cell towards malignancy. During the evolutionary process that starts from the single cell, the rising cancer needs to break through existing physiological barriers, in place to prevent its growth. The most fundamental barriers of cancer are existing mechanisms that prevent mutations and the subsequent malignant transformation of cells. Among them, the most important ones are DNA synthesis and DNA repair mechanisms.

In this thesis, I give a broad description of the different DNA repair pathways, their key players and how changes in their activity can lead to mutagenesis and malignant transformation of cells. I also present the results from one of my PhD projects, where we successfully modeled different mutagenesis patterns that occur in cancer genomes in a controlled laboratory environment. This is important, because cancer genomes accumulate a complex mixture of mutation patterns in their genome, which have been well studied. It has however been very challenging for researchers to associate the mutation patterns to the biological processes from which they originate.

Our work is one of the first two to show that endogenous mutation patterns, that arise through DNA repair defects, can be reproduced and studied in a controlled *in vitro* experiment. This opens the window for a variety of applications in the field, including the study of timing and sequence of mutation events, which has hitherto been difficult to assess from observational *in vivo* data. We identify *in vitro* signatures that confirm the known mismatch repair signature, but also novel ones that could be used as biomarkers for future characterization and therapeutic interventions of rare cancer subtypes with mutations in mismatch repair, Fanconi Anemia and BRCA repair pathways.

Zusammenfassung

Krebs wird durch Mutationen im Genom verursacht, die eine ansonsten gesunde Zelle in eine Kranke verwandeln. Während des evolutionären Prozesses, der von der Einzelzelle ausgeht, muss der entstehende Krebs vorhandene physiologische Barrieren überwinden, die existieren um sein Wachstum zu verhindern. Die grundlegendste Krebsbarriere sind existierende Mechanismen, die Mutationen und die anschließende krankhafte Transformation von Zellen verhindern. Zu den wichtigsten zählen DNA-Synthese und DNA-Reparatur Mechanismen.
In dieser Arbeit beschreibe ich detailliert die verschiedenen DNA-Reparaturwege, ihre Schlüsselspieler und wie Veränderungen in ihrer Aktivität zur Mutagenese und krankhafter Transformation von Zellen führen können. Ich präsentiere auch die Ergebnisse eines meiner Promotionsprojekte, in denen wir erfolgreich verschiedene Mutagenese-Muster modellierten, die in echten Krebsgenomen in einer Kontrolllaborumgebung vorkommen. Dies ist wichtig, weil Krebsgenome eine komplexe Mischung von Mutationsmustern in ihrem Genom akkumulieren, die schon gut beschrieben worden waren. Es war bis jetzt jedoch eine große Herausforderung für Forscher, Mutationsmuster mit den biologischen Prozessen zu verknüpfen, aus denen sie stammen. Unsere Arbeit ist eine der ersten beiden in der Welt, die zeigen, dass endogene Mutationsmuster, die durch DNA-Reparaturdefekte verursacht werden, reproduziert und in einem kontrollierten *In vitro*-Experiment untersucht werden können. Dies öffnet die Türe für eine Vielzahl von Anwendungen auf dem Gebiet, einschließlich der Untersuchung des Zeitpunkts und der Sequenz von Mutationsereignissen, die zuvor schwierig aus der Beobachtung von *in vivo* Daten zu bewerten war. Wir identifizieren *In vitro*-Signaturen, die die bekannte Fehlpaarungsreparatursignatur bestätigen, sowie neue, die als Biomarker für die zukünftige Charakterisierung und therapeutische Intervention von seltenen Krebs-Subtypen mit Mutationen in Fehlpaarungsreparatur, Fanconi-Anämie und BRCA-Reparaturwegen verwendet werden können.

Publications arising from this thesis

Title:

Validating the concept of mutational signatures with isogenic cell models.

Authors:

Xueqing Zou#, Michel Owusu#, Rebecca Harris, Stephen P. Jackson, Joanna I. Loizou* & Serena Nik-Zainal*.

(# these authors contributed equally, * these authors are both corresponding authors)

Journal:

Nature Communications volume 9, Article number: 1744 (2018).

https://doi.org/10.1038/s41467-018-04052-8

List of Abbreviations

ABL1	Abl Proto-Oncogene 1, Non-Receptor Tyrosine Kinase
ADP	Adenosine Diphosphate
AGT	O6-Alkylguanine-Dna Alkyltransferase
AID	Activation Induced Cytidine Deaminase
ALK	Anaplastic Lymphoma Receptor Tyrosine Kinase
AP	Apurinic Or Apyrimidinic
APOBEC	Apolipoprotein B Mrna Editing Enzyme Catalytic Subunit
ATM	Ataxia Telangiectasia Mutated,.Atm Serine/Threonine Kinase
ATP	Adenosine Triphosphate
ATR	Ataxia Telangiectasia And Rad3 Related, Atr Serine/Threonine Kinase
BC	Before Christ
BCR	Bcr, Rhogef And Gtpase Activating Protein
BER	Base Excision Repair
BLM	Bloom Syndrome Recq Like Helicase
BRAF	B-Raf Proto-Oncogene, Serine/Threonine Kinase
BRCA	Brca1, Dna Repair Associated
BSS	Blind Signal Separation
CIN	Chromosomal Instability
CML	Chronic Myelogenous Leukemia
CNV	Copy Number Variation
CPD	Cyclobutane Pyrimidine Dimer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CS	Cockayne Syndrome
CSA	Ercc Excision Repair 8, Csa Ubiquitin Ligase Complex Subunit
CSB	Ercc Excision Repair 6, Chromatin Remodeling Factor
CSR	Class Switch Recombination
DDR	Dna Damage Response
DNA	Deoxyribonucleid Acid
DR	Direct Repair
DSB	Double-Strand Break
DSBR	Double-Strand Break Repair
EGFR	Epidermal Growth Factor Receptor
FA	Fanconi Anemia
FANC	Fanconi Anemia Complementation Group
FANCC	Fanconi Anemia Complementation Group C
FANCI	Fanconi Anemia Complementation Group I
FANCJ	Fanconi Anemia Complementation Group J
FANCL	Fanconi Anemia Complementation Group L
FANCM	Fanconi Anemia Complementation Group M
FANCP	Fanconi Anemia Complementation Group P
FANCS	Fanconi Anemia Complementation Group S
GGR	Global-Genome Repair
HDR	Homology Directed Repair
HGP	Human Genome Project
HR	Homogous Recombination

HSPC	Haematopoietic Stem And Progenitor Cell
ICL	Interstrand-Crosslink
IR	Ionizing Radiation
KRAS	Kras Proto-Oncogene, Gtpase
MGMT	O-6-Methylguanine-Dna Methyltransferase
MMR	Mismatch Repair
MRN	Mre11-Rad50-Nbs1-Complext
MSI	Microsatellite Instability
MUTYH	Muty Dna Glycosylase
NER	Nucleotide Excision Repair
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End Joining
NMF	Non-Negative Matrix Factorization
NRAS	Nras Proto-Oncogene, Gtpase
NSCLC	Non-Small Cell Lung Cancer
PARP	Poly (Adp-Ribose) Polymerase
PCA	Principal Component Analysis
PCNA	Proliferating Cell Nuclear Antigen
PCR	Polymerase Chain Reaction
PNKP	Polynucleotide Kinase 3'-Phosphatase
POLA	Polymerase Alpha
POLB	Polymerase Beta
POLD	Polymerase Delta
POLE	Polymerase Epsilon
POLL	Polymerase Lamda
POLM	Polymerase Mu
POLQ	Polymerase Theta
PTIP	Pax Interacting Protein 1
RAS	Proto-Oncogene, Gtpase
RNA	Ribonucleic Acid
RNAPII	Rna Polymerase Ii
ROS	Reactive Oxygen Species
RPA	Replication Protein A
SAM	S-Adenosylmethionine
SHM	Somatic Hyper Mutation
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
TCR	Transcription-Coupled Repair
TF	Transcription Factor
TFIIH	Ercc Excision Repair 2, Tfiih Core Complex Helicase Subunit
TLS	Translesion Synthesis
TTD	Trichothiodystrophy
UNG	Uracil Dna Glycosylase
UV	Ultra Violet Light
UVSSA	UV Stimulated Scaffold Protein A
VDJ	Variable, Joining, Diversity Gene Segments
VUS	Variance Of Uncertain Significance
WGS	Whole Genome Sequencing
XP	Xeroderma Pigmentosum

XPA	Xeroderma Pigmentosum Group A-Complementing Protein
XPB	Xeroderma Pigmentosum Group B-Complementing Protein
XPC	Xeroderma Pigmentosum Group C-Complementing Protein
XPD	Xeroderma Pigmentosum Group D-Complementing Protein
XPE	Xeroderma Pigmentosum Group E-Complementing Protein
XPF	Xeroderma Pigmentosum Group F-Complementing Protein
XPG	Xeroderma Pigmentosum Group G-Complementing Protein

Acknowledgements

First and foremost, I would like to thank my PhD supervisor Dr. Joanna Loizou, who has made all of this work possible by creating a scientifically stimulating atmosphere in her group and enabling her students to follow-up on their own ideas and intuitions, no matter how crazy those had been at times.

I would also like to thank our collaborators in Cambridge Dr. Nik-Zainal and Dr. Xueqing Zou for their contribution to our joint manuscript.

Next, I would like to thank my senior colleagues: Abdelghani Mazouzi, you inspired all of us with your never-ending enthusiasm for science. Georgia Velimezi, with your example, you taught us how to be diligent scientists. Jana Prochazkova, with your example, you taught me to work efficiently in the cell culture. Mark Wiedner, I do not know how our lab would have survived without your management and assistance. Thank you for everything you did.

I would also like to thank my junior colleagues, Joana Ferreira da Silva and Lydia Robinson, who have been the main driving force of the lab in recent times. I am very proud to be leaving the group in such capable hands. I would also like to express my appreciation for the work on my project that was done by my students Peter Bannauer and Claudia Doppler.

I am grateful to have worked at CeMM. I have met many inspiring people in the 5 years of my PhD, and have learned a great deal from them. Mentioning every single one would exceed the standard length of a PhD thesis, therefore I will resort to mentioning a few. Prof Giulio Superti-Furga, one of the most dynamic and inspiring leaders I have ever met in my life. Drs Stefan Kubicek and Sebastian Nijman who really inspired me to do screens, particularly Stefan who has worked closely with me on other manuscripts. Their group members, in particular Drs Markus Muellner, Claudia Kerzendorfer, Charles Lardeau and Ms Anna Ringler who have given support and assistance to me with regards to my projects. I would also like to thank Drs Joerg Menche, Andreas Bergthaler and Georg Winter for their support and mentoring, particularly Joerg who has been a collaborator on several projects in our lab and has given me bioinformatics support through his superb PhD student Michael Caldera.

I would also like to express appreciation for the people that kept me partying during my PhD, the hangover boys and girls: Ferran Fece de la Cruz, Mate Kiss, Adrian Cesar Razquin, Lindsay Kosack, Dijana Vitko and Aniko Fejes.

I am particularly grateful for Aniko Fejes, who over the past three years has been an essential support in all my professional and private endeavors. Finally, I am grateful to my family, particularly my mother, who inspired me to become the person that I am today.

CHAPTER ONE: INTRODUCTION

Our fight against cancer

Historic perspective

A description of breast cancer, from 3000 BC, found on Egyptian papyrus (Edwin Smith Papyrus) is currently the oldest recording of cancer that we know of (Hajdu, 2010). Other records dating back to 1500 BC indicate that ancient Egyptians knew of various types of cancer and treated them by cautery, with knives, or various sorts of chemicals, including salt or arsenic paste (Hajdu, 2010). This knowledge of cancer and its treatment was also present in Greek texts: Around 400 BC, Hippocrates, described a cancerous growth, which reminded him of a growing crab, henceforth the name cancer (Hajdu, 2010). Superficial cancer (e.g. skin) was treated with creams and ointments whereas deep rooted cancer was dissected by knife or classified as incurable. Scientists and philosophers hypothesized that there was a natural cause of cancer, rejecting many of the superstitious believes and stigmata around cancer that was prevalent in their time (Hajdu, 2004).

Interestingly, since the first recordings of cancer, up to the late 20th century, although much had been done, overall, not very much had changed about our knowledge of the fundamental causes of cancer, or our therapies for treating cancer. Most cancers are still diagnosed according to the organ or tissue of origin and treatment "with knives" (surgery), "chemicals" (chemotherapy) are still the most prevalent. Surgical removal of cancer is at times inefficient, since the cancer may be "deeply rooted" (metastatic) and continue growing after treatment. Chemotherapy does not only target cancerous cells but also healthy cells, leaving patients with horrible side effects and, in some cases, secondary therapy induced tumors. However, cancer therapy is experiencing innovative changes in the 21st century, with a focus on targeted, personalized therapy and the use of endogenous cancer killers (immunotherapy), thus focusing on causing minimal side effects and maximum depletion of the tumor. Much of the innovation has to do with our increased knowledge of what cancer actually is, which came with technological breakthroughs, such as cloning or sequencing but also cell and mouse models. Though much remains to be learned, we have moved from believing in superstitious causes of cancer to understanding the various levels of cancer disease, from the influences of environmental carcinogens, to human organs, tissues, cells and ultimately the deoxyribonucleic acid (DNA).

Mutations in deoxyribonucleic acid (DNA)

The information encoded in the sequence of DNA entails genes, the blueprint for the manufacture of proteins, one of the main molecules that operate and control almost all processes in a living cell. Changes in the sequence of DNA, pose a potential threat to the integrity of proteins and hence the life of a cell and a whole organism, such as a human being. Those changes arise in the form of damage inflicted to the DNA or mistakes that occur during the copy process of DNA. If those changes in the sequence of DNA are not recognized and repaired in time by DNA repair proteins, then those changes can become permanently integrated into the DNA. Mutations are permanent changes to DNA from its original sequence. Mutations in humans can either occur in germ cells (cells that develop to become sperm and ovum) or in somatic cells (any non-germ cells). A somatic mutation only occurs in a single cell but can be carried on to all of its potential daughter cells, if that cell has the ability to replicate. A germ cell of one sex, carrying a mutation, has the potential to fuse with a gamete from the opposite sex, forming a zygote which eventually gives rise to all the cells of a human offspring. Thus, mutations in germ cells (germline mutations) are potentially carried on to all the cells (germ and somatic cells alike) of an offspring. A person's germline genome is the individual's genome as inherited from the parents. Two randomly chosen individuals may have about twenty thousand genetic (germ line) variations distributed throughout their expressed genome (exome) (Vogelstein *et al*, 2013).

What is cancer

Cancer is a genetic disease (or at least strongly influenced by genetic factors). In general, it arises from mutations in a single cell, which then proliferates and evolves towards malignancy, with the cancer taking any possible path in order to grow and survive. The overgrowth of cancerous cells affects tissues and organs often leading to clinical symptoms.

Mutations have a strong malignant potential if they occur in cancer related genes (oncogenes or tumor suppressors). Those are genes, which upon gain of function or loss of function mutations respectively, can transform an otherwise normal cell towards malignancy (Pleasance *et al*, 2010a). The reasoning behind a mutated gene as causal in cancer stems from the observation that the number and type of mutations in affected genes were not probable to be caused by chance (Futreal *et al*, 2004).

Some of the mutations in the development of cancer are responsible for its initiation, progression or metastasis (driver mutations). Most of the mutations in a cancer genome however are not directly responsible for its development (passenger mutations) (Pleasance *et al*, 2010b). Depending on the conditions, passenger mutations can turn into driver mutations.

This has for instance been reported in the response of tumors to chemotherapy, where existing mutations may be responsible for "de-novo" driver mutations that confer resistance to treatment (Roche-Lestienne *et al*, 2002). Although driver mutations are not easily identified by DNA sequencing alone, drivers are typically the most commonly shared mutations between tumors, cluster around known cancer associated genes and tend to be non-silent mutations (Stratton *et al*, 2009). Passenger mutations on the other and are randomly distributed throughout the genome. Therefore, it is assumed that clones bearing driver mutations are positively selected in the evolution of cancer (Pon & Marra, 2015).

According to Bernd Vogelstein (Vogelstein *et al*, 2013), there are about 140 genes that can drive cancer when mutated and a typical cancer has 33 - 66 genes with mutations that are expected to affect protein integrity (non-synonymous mutations). On average there are only 2 – 8 driver gene mutations per tumor, the remaining are passenger gene mutations. Vogelstein also inferred that overall, all driver mutations fall into 12 signaling pathways that regulate three core processes: cell fate, cell survival and genome maintenance. Frequently mutated cancer genes include Tumor Suppressor p53 (TP53), RAS oncogenes, various DNA repair genes (Knijnenburg *et al*, 2018) and kinases (Greenman *et al*, 2007).

Epigenetic alterations

Epigenetic changes can alter the expression of a gene in an otherwise unperturbed genomic sequence (Plass *et al*, 2013). Such changes can be passed on from mother to daughter cells, functioning as genomic changes, constituting a stable inheritable trait for a cell. These type of inherited changes in gene expression are at times referred to as epimutations (Oey & Whitelaw, 2014), since they function as mutations. They can either be the cause for mutations in genes (primary epimutations) or the consequence of mutations in gene regulating factors (cis or trans regulators). Reports of primary epimutations in cancer are very rare. One prominent example is the methylation of MLH1 which is often found in cancer (Hitchins *et al*, 2007; Ward *et al*, 2013). An epimutation in a gene can function as a driver gene mutation (mut-driver genes). In that case it can be called an epi-driver gene (Vogelstein *et al*, 2013). The expression of epi-driver genes are frequently deregulated in cancer without actual mutations (changes in the sequence) of the gene.

Sequencing of genomes

Advances in the cloning and sequencing of genes allowed researchers to study the genome and changes therein. Initially the sequence of single bases or genes were studied by different methods including early sequencing technologies (Sanger & Coulson, 1975; Sanger *et al*, 1977; Capella *et al*, 1991). Automated sequencing technologies (Gocayne *et al*, 1987) and

computer algorithms for sequence annotations paved the way for the study of exomes and whole genomes (Fleischmann *et al*, 1995; Fraser *et al*, 1995; Bult *et al*, 1996; C. elegans Sequencing Consortium, 1998; Adams *et al*, 2000; Arabidopsis Genome Initiative, 2000), including the human genome.

Lessons from the human genome

The sequencing of the human genome, human genome project (HGP), was pursued in parallel by an international governmental organization (Lander *et al*, 2001) and a private company (Venter *et al*, 2001) with their initial results published in 2001. Since then, additional work has been done (Schmutz *et al*, 2004; International Human Genome Sequencing Consortium, 2004) with the final studies (Kidd *et al*, 2008; Boyer *et al*, 2001) revealing structural variations, such as insertions and deletions, between the genomes of eight individual people. Moreover, data scientists, who had hoped to identify novel cancer genes after the HGP, were met with disappointment because they did not find any (Boyer *et al*, 2001). Overall, revelations from the HGP implied that the goal of attaining the sequence of "the" human genome, required rethinking. In order to understand disease associated changes in the genome of humans, researchers would have to focus on studying many individual human genomes. This was made possible with the advancements in automated sequencing, Next Generation Sequencing (NGS), when sequencing became more affordable and available to many researchers. Thousands of individual genomes could be sequenced, including cancer genomes. The main goal of sequencing cancer genomes has been to look for driver mutations that increase the mutation rate in a cell, contributing to a more rapid evolution of the tumor and metastases formation (Wong *et al*, 2011). The thus gained genomic information could in theory be used to guide targeted therapies that may result in more effective treatment and reduced toxicity (Garraway, 2013).

Current state in our fight against cancer

The work on cancer over the past centuries has overall resulted in an improved outcome for patients (Figure 1). The work has rested on 4 main pillars: prevention, detection, diagnosis and treatment. According to Bert Vogelstein and colleagues (Tomasetti & Vogelstein, 2015), 5% of cancer is inherited, 29% is acquired due to environmental factors and 66% is due to random mutations in genomes. Thus, almost 30% of cancer incidences could be prevented by adequate measures in our behavior and society. The remaining 66% of cases that are acquired through random mutations in the genome require early detection, proper diagnosis and adequate treatment (including targeted therapy), similar so for the 5% of inherited cases.

Therefore, there is still much to be done to improve any of the four pillars of our fight against cancer.

Targeted approaches in cancer therapy

There have been some highlights in targeted therapeutic approaches, such as the BCR-Abl inhibitor imatinib for chronic myelogenous leukemia (CML) (Shaw *et al*, 2013), the ALK inhibitor crizotinib for ALK positive lung cancer, EGFR inhibitor erlotinib for non-small cell lung cancer (NSCLC) (Arteaga, 2003; Lynch *et al*, 2004) or Vemurafenib, a BRAF inhibitor for late stage melanoma. Though some of these targeted therapies have proven extremely effective in some cases (Kalia, 2015), cancer often finds ways to acquire resistance to treatment. Therefore, understanding the molecular mechanisms that underlie the evolution of each cancer type and subtypes will be essential for the next generation of cancer treatment.

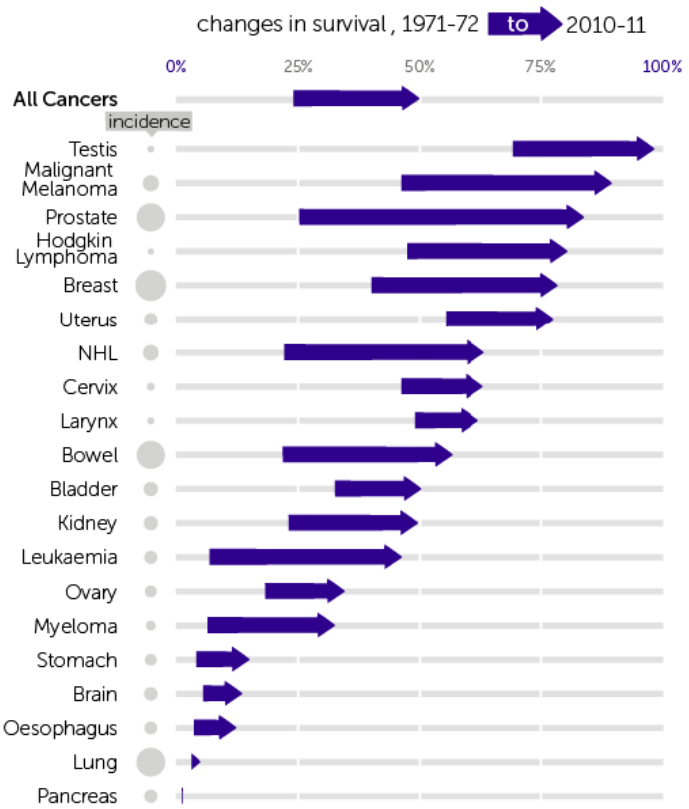


Figure 1: Advances in cancer therapy

Age-Standardized ten-year net survival of selected cancers. Patients were adults (Aged 15-99) from England and Wales in the years 2010-2011. Ten-year survival for 2005-2006 and 2010-2011 is predicted using a statistical model. Breast cancer is for female only. Laryngeal cancer is for male only. Figure is taken from Cancer Research UK (Cancer Research UK website).

Barriers against cancer

Tissue and cellular barriers

In many ways, cancer is age related (possibly part of the aging process) (Aunan *et al*, 2017). Some barriers against cancer diminish with age (e.g. genomic barriers), therefore, fighting cancer is possibly related to a fight against a natural evolutionary process in the life of human beings.

Humans are exposed to specific carcinogenic agents in occupational settings (e.g. asbestos), or due to medicinal or life-style choices (e.g. tobacco smoking, alcohol) (Carbone *et al*, 2004; Luch, 2005; Pfeifer & Besaratinia, 2009). However, some carcinogens are currently unavoidable, (e.g. in food, water, air or as byproducts of endogenous metabolic processes). There are many layers and mechanisms of defense in the human body that are in place to protect us from cancer (or in other words, a malignant overgrowth of cells). For instance, the skin and the gut represent highly specialized environments with distinct structures, cell types, and innate defense mechanisms adapted to support their individual challenges. Ultra violet (UV) light is a powerful mutagen that can cause cancer. The cells of the human body that are most exposed to UV light are on the skin surface. Interestingly, as skin cells migrate from the innermost layers (where proliferation takes place) towards the outermost layers, they lose their proliferation potential and eventually even their nucleus. Thus, the cells exposed to the highest degree of UV mutagen (and other environmental mutagens) are deprived of their cancer potential, since they neither acquire mutations in their genome nor grow or proliferate. Something very similar is true for cells in the gut: The dividing cells are hidden (and protected) in the depth of the crypts, while apical cells on the surface, exposed to highest degree of mutagens in the gut lumen, are non-proliferating or dead. Another example, are haematopoietic stem and progenitor cells (HSPCs) in the bone marrow. HSPCs belong to the most prolific cells in the human body. Interestingly enough, they can only reside and grow in a specific niche with a specific microenvironment. In adult humans that niche is in the bone marrow. Researchers recently made an intriguing discovery, which implies that HSPCs evolved to exclusively reside in the bone marrow as a protection from mutagenic / carcinogenic UV light from the sun (Kapp *et al*, 2018). A final example is the immune system, which plays a vital role in protecting the organism from the outgrowth of renegade cells. Cancer cells therefore develop strategies to escape the immune system (Houghton, 1994; Koster *et al*, 2015).

The cells of the human body are extremely specialized, this may (to some degree) explain the heterogeneity of cancer. Tissues / organs have specialized functions in specialized tissue environments (e.g. breast vs colon). Genes of cells of a tissue could have evolved to adapt to

a function and environment (Nunney & Muir, 2015). Cancer appears to evolve by any possible means (intrinsic and extrinsic factors). That may be why we, for instance, find BRCA mutations predominately in breast cancer, or MMR mutations in colon cancer.
In addition to tissue and cellular barriers, there are intracellular barriers at place to protect the human body from cancer, including programmed cell death of dysfunctional or malignant cells.

Genomic barriers

There are features in the structure, architecture and function of the genome (Figure 1) (Rübben & Nordhoff, 2013) that act as barriers against cancer. The DNA sequence consists of coding and non-coding units (or regions). The coding units (exons) are subsets of genes and contain the coding sequences for proteins. Non-coding units can be subsets of genes (introns) that do not code for proteins, or form sequences between genes (intergenic regions). Intergenic regions may contain sequences involved in the regulation of various diverse processes, such as transcription, translation, DNA replication, as well as in the stability of structures such as centromeres and telomeres. Coding regions make up about 1% of the genome (Elgar & Vavouri, 2008). The remaining ~99% is noncoding DNA, made up of ~75% intergenic regions (Shabalina *et al*, 2001) and ~24% intronic regions. Though some non-coding regions may indeed represent regulatory units of genes, given just the sequence of DNA, most random mutations in the genome would fall into non-coding regions, with no or minimal functional implications for a cell.
DNA is wrapped around histones (nucleosomes) with actively transcribed / expressed regions (euchromatin) and less actively transcribed regions (heterochromatin) which also have implications for mutations. Euchromatic regions tend to accumulate fewer mutations that heterochromatic regions, which means that actively expressed genes, which could induce cancer, are more protected from mutations than less expressed ones.
Damage of DNA, which may lead to mutations, may induce cellular death. Indeed, as proposed for some cases of spontaneous abortion in embryos with chromosomal instability (CIN) (Adjiri, 2017), keeping the integrity of DNA as a barrier against cancer is important to the degree that whole organisms may perish. Consequently, in addition to the packaging, architecture and structure of DNA, one of the most important genomic barriers against cancer is the response to DNA damage.
Though there are many barriers that prevent cancer (Hanahan & Weinberg, 2011) from environmental, to cellular, to genomic factors, this work focuses on the most fundamental origin of cancer: DNA, DNA damage, DNA damage response (including DNA damage repair) and mutations.

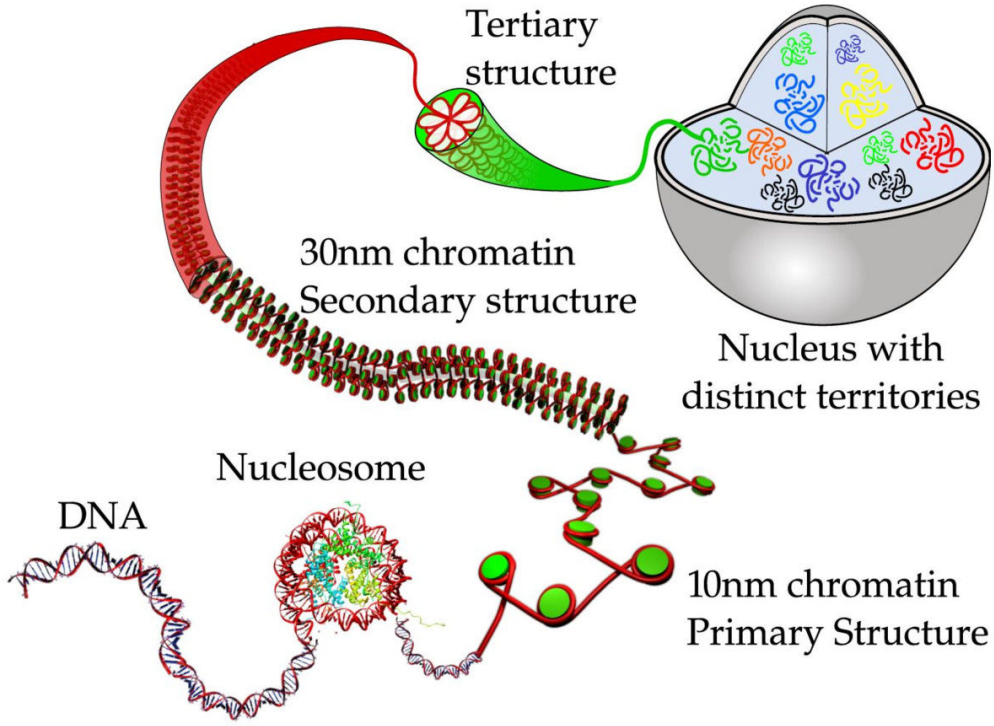


Figure 2: Genome structure

A cartoon of the 3D structure of the genome in a nucleus. Figure was taken from (Iyer *et al*, 2011).

DNA synthesis & replication

DNA synthesis

In human cells, DNA, is a double stranded helical molecule. Each strand is a chain of single nucleotides that form phosphodiester bonds, and the two strands bind together through base pairing of the nucleotides. A nucleotide in DNA consists of a deoxyribose (sugar), a phosphate group, and a base. The deoxyribose and phosphate residues form the backbone of the strand (phosphodiester bonds) and are called sugar phosphate backbone. There are four bases: adenine, cytosine, guanine and thymine. Adenine and guanine belong to the purine group and cytosine and thymine belong to the pyrimidine group. Normal base pairing in DNA is between adenine and thymidine or guanine and cytosine. Other base pairs may lead to mutations if unrepaired.

In order to give rise to functional daughter cells, a parent cell must produce two DNA molecules for them to inherit. DNA is copied in a semi conservative manner, which means, the two parent strands serve as templates for two newly synthesized daughter strands. Deoxyribonuclotide triphosphate (dNTP) are incorporated into DNA by polymerases in an enzymatic reaction: DNA_n + dNTP = DNA_{n+1} + diphosphate, where DNA_n is the DNA molecule with n nucleotides.

Nucleotides can only be added to the 3'-hydroxyl end of DNA, defining the direction of replication from the 5'-phosphate to 3'-hydroxyl end (5' to 3'). DNA damage can already occur by incorporation of damaged nucleotides into DNA. Oxidized nucleotides are one of the most common types of damaged nucleotides. There are different specialized DNA repair factors that deal with damaged nucleotides. One of them is the protein nudix hydrolase 1 (NUDT1, also known as MTH1), which can hydrolyze oxidized purinic dNTPs to dNMPs, thereby depleting cells from oxidized nucleotides that could otherwise be incorporated into DNA causing DNA damage. Besides damaged nucleotides, which are a source of mutations, nucleotide variants, such as the RNA nucleotide Uracil, can become integrated into DNA and require specialized DNA repair enzymes, Uracil-DNA glycosylase (UNG) in the case of Uracil, for their removal.

DNA replication

DNA replication belongs to one of the most conserved processes in a cell. The ability to pass information encoded in the DNA from a mother to daughter cell must have existed since the very early evolutionary stages of life (Taylor & Lehmann, 1998). DNA polymerases are specialized proteins, dedicated to the synthesis and repair of DNA. The human replication polymerases are DNA polymerase alpha (POLA), epsilon (POLE) and delta (POLD). A primase, associated with polymerase alpha, synthesizes an RNA primer to initiate replication, subsequently, polymerase alpha performs elongation for a few nucleotides, after which, the other two replication polymerases take over the remaining strand synthesis (O'Donnell *et al*, 2013). Polymerase epsilon and delta are different in their structure, subunit composition, processivity and fidelity. Polymerase epsilon is primarily responsible for replication of the forward strand, whereas polymerase delta is responsible for replication of the reverse strand (Lujan *et al*, 2016) with a possible involvement in replication of the forward strand (Johnson *et al*, 2015). Mistakes are rare and estimated to occur every 10^4 - 10^5 bases (Kunkel, 2004), with high variations depending on polymerase, nucleotides and sequence context (e.g. tandem repeats). On average, the rate is 100-fold lower than what is expected stochastically (10^2 - 10^3 *in vitro*) (Loeb & Kunkel, 1982), indicating that replication polymerases have a high selectivity to the polymerization reaction. The occasional mistakes however, can be a source for mutations if not repaired. The main elongation polymerases, POLE and POLD, have a proofreading function (exonuclease activity), which additionally increases their fidelity: upon detection of an adjacent misincorporated nucleotide, the polymerase can move one (or a few) nucleotides backwards (3' to 5') and excise the mismatched nucleotide via a 3'-to-5' exonuclease activity. Since DNA synthesis can only occur from the 5' to 3' direction, the forward strand (by definition) is synthesized in a continuous manner and the reverse strand is

synthesized in a discontinuous manner, fragment after fragment (Okazaki fragments). The Replication via Okazaki fragments require an extra step of DNA ligation, performed by DNA ligase 1 (Lig1), in order to link two adjacent fragments. Replication polymerases stall upon encounter of DNA damage (replication fork stalling) and either require lesion bypass or repair for fork progression. Okazaki fragments, that are not ligated, remain as nicks in the DNA strand, encountered as DNA damage, which can induce replication fork stalling. Stalled replication forks may collapse, leading to toxic double-strand breaks that can trigger apoptosis.

DNA damage

DNA damage to a cell may come from an internal (endogenous) or external (exogenous) source (Table 1). Endogenous sources can be mistakes during replication, which mostly result in base substitutions or indels, besides replication however, there are many other possible endogenous threats to the integrity of DNA. Most commonly, endogenous damage is inflicted by reactive oxygen species (Jena, 2012) that are produced from physiological activities in a cell, particularly metabolic processes. Other possible endogenous mutagens include S-adenosylmethionine (SAM) (De Bont, 2004), acetaldehydes (Matsuda *et al*, 1998) and the enzymatic activity of proteins (Supek & Lehner, 2017). Exogenous sources for DNA damage come from the environment of the cell, such as the tissue environment (e.g. metabolic processes of the tissue or organ) or the environment of the organism. Typical exogenous sources for DNA damage in human cells are UV-light, cigarette smoking, alcohol consumption, exposure to radioactive material or therapeutic induced DNA damage (chemotherapy, radiotherapy).

DNA damaging agents rarely produce one type of damage, but every damaging agent may have a predominant feature. Ionizing radiation (IR), used in radiotherapy, induces single-strand breaks, double-strand breaks, base damage and more, but its most distinguished feature (that induces cell death), are double-strand breaks. Furthermore, IR induced breaks are predominantly caused by free radicals, particularly reactive oxygen species (Sonntag, 2006). Since a single agent may produce different types of damage, different types of repair factors are often involved in damage repair.

DNA damage can trigger cell death, but not all types of DNA damage are equally dangerous. Potent apoptosis inducing damages include N-alkylations, bulky DNA adducts, DNA crosslinks and DNA double-strand breaks. Another example is O6-methylguanine, although this damage requires the activity of DNA mismatch repair to trigger apoptosis (Roos & Kaina, 2006).

Table 1 Frequency of DNA damage in cells

	DNA damage	DNA lesions	Number of lesions per cell	Comment
endogenous	Depurination	AP site	12000 / day	substitutions
	Depyrimidination	AP site	600 / day	substitutions
	Cytosine deamination	base transition	192 / day	C:G > T:A
	SAM-induced methylation	3meA	600 / day	-
	SAM-induced methylation	7meG	4000/ day	-
	SAM-induced methylation	O6meG	10 - 30 / day	G:C > A:T
	Oxidation	8oxoG	400 - 1500 / day	G:C > T:A
	SSB	nick	55000 / day	substitutions
	DSB	break	25 / day	rarrangements
exogenous	Sunlight (at peak hour)	dimers+	100000 / day	C:G > T:A
	Smoking	DNA adducts+	45 - 1029	1 -2 packs / day for 40 years
	Chest X-rays	DSBs+	0.0008	0.02 mSv
	Dental X-rays	DSBs+	0.0002	0.005 mSv
	Mammography	DSBs+	0.016	0.4 mSv
	Body CT	DSBs+	0.28	7 mSv
	Head CT	DSBs+	0.08	2 mSv
	Tumor PET scan (F-18)	DSBs+	0.4	10 mSv
	Iodine-131 treatment	DSBs+	4.4	110 mSv
	External beam therapy (IR)	DSBs+	76	1900 mSv
	Airline travel	DSBs+	0.0002 / h	0.005 mSv / h
	Space mission (60 days)	DSBs+	2	50 mSv
	Chernobyl acccident	DSBs+	12	300 mSv
	atomic bombs (Hiroshima and Nagasaki)	DSBs+	0.2 - 160	5 - 4000 mSv

DSBs+, double-strands breaks and other lesions. Table adapted from (Tubbs & Nussenzweig, 2017) and (Roos & Kaina, 2013).

DNA damage response

The DNA damage response (DDR) consists of a complex network of proteins that act as sensors, transducers, mediators or effectors of DNA damage. Sensors are proteins that are able to detect DNA lesions in the genome. For instance, the MRN complex is a sensor for double-strand breaks. Sensors may recruit other factors to the site of damage. Some of those factors are transducers; they create or amplify a signal for DNA damage in the nucleus. An important example for that is the kinase ATM, which is recruited by MRN to sites of DNA damage. ATM phosphorylates a large number of proteins (Matsuoka *et al*, 2007), thereby recruiting them, or modulating their activity after DNA damage. Other recruited proteins act as mediators, they mediate interactions between proteins, an important example is the phosphorylated histone variant H2AX, called γH2AX, which is also commonly used as a marker for DNA damage (Sharma *et al*, 2012). The last group of proteins, effector proteins, regulate the outcome of damage-sensing and signal transduction. There is a variety of effector proteins. Some effectors function in the repair of DNA, this is the case for ligases such as LIG4, which can anneal the two DNA ends after double-strand breaks. Another group of effectors is involved in the regulation of the cell cycle, the checkpoint kinase CHK2 is an example for that. It is important for cells with DNA damage, to arrest in their cell cycle, in order

to provide time for DNA repair. In the case the DNA damage cannot be repaired, cells need to undergo apoptosis (programmed cell death), in order to maintain cellular homeostasis. The tumor suppressor TP53 is an important effector protein involved in the regulation of apoptosis (Fridman & Lowe, 2003). Other effectors may be involved in the regulation of other essential cellular pathways, such as gene expression or metabolism. We are still constantly discovering novel factors involved in the DNA damage response (Gupta *et al*, 2018). Perturbations in the DNA damage response network may lead to failure in the protection or repair of DNA, which can cause diseases associated with malignant growth, dysfunction or depletion of cells (Jackson & Bartek, 2009b).

DNA repair

In order to deal with a variety of lesions, cells have evolved specialized proteins, dedicated to the repair of DNA. Those proteins can be grouped into DNA repair pathways, depending on the repair mechanism. However, a strict separation into pathways is not always possible, due to the cooperation of repair factors from different pathways. Moreover, DNA lesions are often of a heterogeneous nature, for instance ionizing radiation induces a whole range of DNA lesions (Leadon, 1996), including base damages, single-strand as well as double-strand breaks, that require repair by proteins from different pathways. In general, DNA repair typically happens in five steps: sensing of the damage (sensing), nicking or accessing of DNA (incision), removal of damaged nucleotides or adducts (resectioning), replacement of nucleotides (synthesis), and ligation of the DNA backbone (ligation). DNA repair is conditional. For example, mismatch repair mainly functions in association with replication induced mistakes or damages, nucleotide excision repair, has a subpathway, specialized on the repair of transcription associated mistakes or damages, and homologous recombination is only active in S or G2 phase. In addition, there are conditions where closely related or non-related repair factors can repair DNA damage in the absence or failure of the specialized repair factors (Moder *et al*, 2017; Puddu *et al*, 2015).

There are two (non-mutually exclusive) ways of how DNA damage or repair may cause a disease. Unrepaired DNA damage triggers apoptosis or senescence, which if exacerbated leads to diseases associated with cellular depletion, such as anemia, progeria or mental retardation. On the other hand, erroneous repair of DNA damage results in mutations that may cause diseases associated with malignant cellular dysfunction or growth, such as cancer. The rates of different mutations is therefore increased in several rare inherited diseases, including Fanconi anemia, ataxia telangiectasia, and xeroderma pigmentosum, which are also associated with increased risks of cancer (Jackson & Bartek, 2009b; Stratton *et al*, 2009).

Polymerases in DNA repair

The synthesis of DNA during DNA repair requires polymerases. Some repair pathways utilize replication polymerases (POLE, POLD) for DNA synthesis, but there are many more polymerases specialized in the repair of lesions. Polymerase beta (POLB) for instance is the repair polymerase in base excision repair. Translesion synthesis (TLS) polymerases can assist DNA replication polymerases during replication or during DNA repair by their unique ability to synthesize DNA across sites of lesions (lesion bypass), which are inaccessible to replication polymerases. Some translesion polymerases are the main DNA synthesis polymerases for repair. An example for that is the double-strand break repair factor polymerase theta (POLQ), which functions in one of the DSB repair pathways (Mateos-Gomez *et al*, 2015). TLS polymerases most commonly use damaged DNA as a template and have no proofreading function, making repair by TLS highly error prone (McCulloch & Kunkel, 2008). Therefore, TLS can contribute greatly to mutations in the genome (Supek & Lehner, 2017).

Mismatch repair (MMR)

DNA mismatch repair is primarily responsible for the repair of single or short (few nucleotides) base substitutions that occur during DNA replication (Iyer *et al*, 2006). DNA replication polymerases can misincorporate single nucleotides, which can lead to base substitutions if not recognized and repaired by the polymerase or by mismatch repair. During replication, MMR, is mostly strand specific, and very likely mediated by transiently under-methylated GATC sequences that direct MMR to the daughter strand (evidence from experiments in bacteria) (Pukkila *et al*, 1983). Replication polymerases are innately more error prone at sites of DNA repeats in the genome. They can skip single (or short sequences of) nucleotides at sites of repeats by slipping forward or backwards during replication. This polymerase slippage can lead to an insertion or deletion of one nucleotide (or a few nucleotides) at sites of repeat sequences in the genome. The proofreading function of polymerases does not work for such single base insertions or deletions (Kroutil *et al*, 1996), they require MMR for efficient repair.
In general, there is complementarity between replication, proofreading and MMR (Lujan *et al*, 2014). For instance, the lagging strand accumulates about twice as much damage as the leading strand, but the lagging strand is also about twice as much efficiently repaired by MMR than the leading strand, indicating complementarity between replication and MMR. Another interesting example is polymerase proofreading and MMR. Polymerase proofreading does not work for damaged mismatches. For instance, polymerases preferentially incorporate adenosine nucleotides opposite of 8-oxo-guanine (8oxoG). This mistake, if not repaired, leads to a G:C > T:A substitution mutation in subsequent replication rounds. Polymerase proofreading is unable to detect the mistake because the mismatch forms a Hoogsteen base

pair, with a geometry that is alike to a correct base pair. Such lesions however, are efficiently corrected by MMR (Ni *et al*, 1999; Russo *et al*, 2003). On the other hand, MMR is less efficient in repairing the infrequent mismatches C:T or T:T, produced by replication polymerases (Lujan *et al*, 2014). Complementarity between replication, proofreading and MMR implicates that the three systems coevolved to ensure fidelity of DNA replication.

Single base-base and indel mismatches are primarily recognized by MutS alpha, consisting of the MSH2/MSH6 heterodimer. The complex scans DNA and stops at mismatches. The mismatch detection is associated with an ATP dependent conformational change of MutS alpha, such that it forms a moving clamp at the same time allowing recruitment and binding of MutL alpha. PCNA activates MutL alpha (containing an endonuclease domain), which (by mobility of MutS alpha) nicks the newly synthesized DNA strand upstream and downstream of the mismatch. From that point on, there are two to three models for DNA repair (Kunkel & Erie, 2015): One possibility involves MutS alpha promoted strand resectioning by EXO1. This results in a fragile single stranded DNA molecule that becomes coated and protected by RPA. Replication polymerases POLE or POLD can then fill in the missing nucleotides. A second possibility is strand displacement synthesis, wherein POLE or POLD directly invade the nicked, single-stranded DNA, and synthesize a new strand by displacing the old strand containing the mismatch. A third possibility, which still requires more evidence, could use the 3'- 5' exonuclease activity of POLE or POLD for resection, followed by strand extension. After resynthesis of missing nucleotides, all two or three subpathways require ligation of the remaining nick in DNA by ligase 1. This results in an error free repair of single base-base and indel mismatches that are frequently generated at sites of repeats. Therefore, in the absence of MMR, genomes of cells exhibit a large amount of repeat mediated indels, which has become a marker for MMR deficient tumors, called microsatellite instability (MSI) (Boland & Goel, 2010).

There are very few studies on other known MMR factors, most importantly, MSH2 and MSH3 (which form the heterodimer MutS beta), MLH1 and PMS1 (MutL beta), and MLH1 and MLH3 (MutL gamma). Notably, Mut Sbeta, is involved in the repair of large as well as one- and two-base indel mismatches (Harfe & Jinks-Robertson, 2000). One of the primary reason why MLH1, MSH2 and MSH6 are studied in much more detail than other MMR factors, is because mutations in those mismatch repair genes account for almost all tumors with MMR deficiencies (Korhonen *et al*, 2008).

Direct reversal / repair (DR)

The least complex form of DNA repair is the simple reversal the damage without excisions or insertions of nucleotides or bases. This type of repair often only requires the activity of a single

enzyme and is error free. A trivial example is a special case of single-strand break repair. In its most basic form, single strand breaks (nicks in DNA) are repaired by ligation of the broken backbone. The two adjacent nucleotides can be sealed together, provided that the 3'-hydroxyl and 5'-phosphate are intact and no base damages have occurred. In general, any one-step reversal of a damage can be considered part of the direct reversal repair pathway. Conventionally however, there are 3 most commonly known and studied DR pathways (Yi & He, 2013): DR of UV adducts, O-alkylation adducts or N-alkylation adducts.

UV light may induce the formation of dimers- cyclobutane pyrimidine dimers (CPDs)- and photoproducts- 6,4-photoproduct (6-4PPs)- at adjacent pyrimidine bases. CPDs are more abundant (3:1 ratio), while 6-4PPs are more toxic. From a structural point of view, both adducts are formed as two separate adjacent pyrimidine molecules are chemically bound to become a single dipyrimidine molecule. If not repaired, these dimers interfere with essential cellular processes such as transcription or replication. In humans, UV lesions are repaired by the nucleotide excision repair pathway, which uses a complex mechanism and has a wide substrate range in addition to UV lesions. Many other organisms use nucleotide excision repair or photolyases. Photolyases (not present in placental mammals, which are humans and mice) employ a very simple but efficient mechanism for the repair of UV lesions (Essen & Klar, 2006). In the case of CPDs, the CPD photolyase contains two chromophores. After binding to DNA, one of the chromophores absorbs light in the blue / near-UV spectrum and transfers the energy to the second chromophore. The second chromophore uses the energy to split the CPD dimer, thus restoring the original pyrimidine bases. 6-4PP is repaired in an analogous manner by 6-4PP photolyase (Essen & Klar, 2006).

Alkylation of DNA is the abnormal addition of alkyl (including methyl) groups, to DNA. Little is known about the endogenous sources of alkylation induced DNA damage. One reported example is S-adenosylmethionine (SAM), a reactive methyl group donor, which plays a role in physiological regulation of gene expression (De Bont, 2004). The repair of alkylation adducts on DNA is mediated by different proteins, dependent on the position of the methyl group.

AlkB functions in the reversal of N-alkylated bases and is a member of alpha-keto-glutarate-dependent and iron-dependent oxygenases (Fedeles *et al*, 2015). It is a bacterial protein, but there are nine AlkB homologs in humans. It uses an iron and oxygen intermediate to oxidize methylated bases. This results in the conversion of alpha-ketoglutarate to succinate and CO_2, and is coupled to the hydroxylation of the methyl group. The hydroxyl-methyl group spontaneously decomposes to formaldehyde, thereby restoring the original, unmethylated base. AlkB is primarily involved in the reversal of 1-methyl adenine and, the structurally similar, 3-methyl cytosine.

An important alkylation induced lesion is methylation of guanine at the oxygen in position 6, 6-O-methylguanine (O6meG). O6meG is extremely mutagenic, due to its ability to base pair with thymine instead of cytosine, leading to a G:C > A:T base substitution. The only known repair factor for O6meG is O-6-methylguanine-DNA methyltransferase (MGMT), also known as O6-alkylguanine-DNA alkyltransferase (AGT), since it can repair a larger range of alkyl adducts besides methylation (Kaina *et al*, 2007). Due to its unique role, loss of MGMT is associated with high incidences of mutagenesis and cancer. MGMT is a suicide enzyme, meaning that the protein becomes permanently inactive after the enzymatic repair of a lesion. A cysteine residue in the catalytic site of the enzyme forms a strong bond with alkyl groups. Continuous expression of MGMT is therefore required for continuous repair of O6meG in cells. In order to increase mutagenesis and promote carcinogenesis, cancer cells tend to turn off the expression of MGMT by methylation (or loss of function mutations), making them highly vulnerable to alkylation based chemotherapeutic agents, such as temozolomide or carmustine (Shiraishi, 2000).

Base excision repair (BER)

Base excision repair is the primary repair pathway for single base damages (short patch repair) or a few (2 – 10) base damages (long patch repair) that occur outside of DNA replication and hardly distort the helical duplex structure of DNA. Base damages are induced by different endogenous or exogenous sources. Reactive oxygen species (ROS), metabolites, enzymatic activities, UV, chemo- or radiotherapy, can all lead to single base damages. One of the most notorious examples of base damage is the ROS induced oxidation of guanine at position 8, known as 8-oxo-guanine, which may lead to G:C > T:A mutations if not repaired (Grollman & Moriya, 1993). Even spontaneous events in a cell may result in base damage. Such is the case for spontaneous deamination of cytosine to uracil or 5-methyl-cytosin (5meC) to thymine, resulting in C:G > T:A base substitutions, if unrepaired. Single base damages that lead to base substitution mutations are the most abundant type of mutations in the genome of cells, and are main contributors to many genetic diseases including cancer (Pleasance *et al*, 2010a).

Single-strand break repair is often regarded as a special case of base excision repair in scientific literature (Giglia-Mari *et al*, 2011; Curtin, 2012). The repair steps between long and short patch base excision repair and single-strand break repair are similar but require different proteins for some steps of the repair process (Krokan & Bjørås, 2013; Giglia-Mari *et al*, 2011; Curtin, 2012). After recognition of the damage, the specific base is removed by excision. Excision is performed by specialized glycosylases. In the case of 8oxoG, the glycosylase OGG1 is primarily responsible for removal of the base. Although each glycosylase is

specialized in a different manner for the repair of damaged bases, they also act in a redundant manner. Other important DNA glycosylases of oxidized lesions include MUTYH and NEIL1. Removal of the damaged base results in an apurinic or apyrimidinic (AP) site, typically referred to as abasic site (the absence of a purine, pyrimidine base) (Lindahl *et al*, 2004). APE1 and functionally related enzymes can cleave the DNA backbone, resulting in a lesion that resembles and is similarly repaired as a single-strand break. The AP site is subsequently removed and leaves either a 3' hydroxyl or phosphate group. In the latter case dual kinase and 3'-phosphotase PNKP can catalyze conversion of 3' phosphate to 3' hydroxyl, required for the addition of nucleotides (Jilani *et al*, 1999). Polymerase B (POLB) a specialized BER polymerase fills in the missing nucleotide and the opened strand is sealed by the BER ligase, ligase 3 (LIG3) in association with its cofactor XRCC1. Due to the functional redundancy of some base excision repair factors, especially DNA glycosylases, there are few examples of major diseases associated with mutations of single proteins in the pathway. Combined mutations of multiple BER proteins however, or BER proteins with other DNA repair factors are reported to be very toxic and mutagenic (Xie *et al*, 2004; Chan *et al*, 2009; Kemmerich *et al*, 2012; Krokan & Bjørås, 2013).

Nucleotide excision repair (NER)

There are two different types of adducts that threaten the genome integrity of cells. Some adducts, such as oxidation or methylation adducts, barely distorts the helical DNA duplex and are primarily repaired by BER. Other adducts distort the DNA duplex and interfere with replication, transcription or epigenetic regulation (Giglia-Mari *et al*, 2011; Kuper & Kisker, 2012). Helix distorting adducts most commonly come from environmental mutagens: UV exposure creates CPDs and 6-4PPs, alcohol consumption or tobacco smoke exposes cells to acetaldehydes or benzo[a]pyrene (BaP). In many organisms, including humans, NER is the main pathway for repair of bulky adducts and thus one of the main pathways to guard the genome against some of the most common environmental mutagens that humans are exposed to. Failure in the repair of CPDs predominantly results in C > T (including CC > TT) mutations, while failure in the repair of BaPs predominantly results in C > A (including CC > AA) mutations.

There are two known subpathways of NER, the transcription-coupled repair (TCR) and global-genome repair (GGR). Though there are more than 25 proteins known to be involved in NER, the mechanism of repair in the two subpathways only differs in the initial recognition of the lesion and then converges to one pathway (Hanawalt, 1994). Furthermore, TCR only functions in the repair of adducts associated with transcription, whereas GGR may function everywhere in the genome (Bukowska & Karwowski, 2018). DNA damage in TCR is recognized upon

stalling of RNA polymerase II (RNAPII) at sites of lesions. The recruitment and interactions of UVSSA, USP7, CSA / CSB, XAB2 results in backtracking of RNAPII making the lesion accessible for repair. In GGR recognition happens via XPC-RAD23B (damage sensor) and XPE. The damage sensor binds to the bulky adduct, initiating DNA repair. TFIIH, at the converging point of TCR and GGR, unwinds the DNA with its helicase subunits XPB (3'-5') and XPD (5'-3'). The pre-incision complex, consisting of XPA, RPA and XPG, stabilizes the lesion whiles the ERCC1-XPF heterodimer performs the 5'-end incision and XPG performs the 3'-end incision. PCNA coordinates DNA synthesis by POLD, E or K, and finally, depending on the cell cycle stage, the nick is either sealed by LIG3/XRCC1 or LIG1.

Besides their dedicated roles in NER, some of the repair proteins have additional roles in other cellular activities, including other DNA repair pathways (Kamileri *et al*, 2012). The versatility of NER and its factors may explain the versatility in the NER associated disease phenotype. NER deficiency results in xeroderma pigmentosum (XP), Cockayne syndrome (CS) and trichothiodystrophy (TTD). All NER disorders are associated with photosensitivity and neurological abnormalities, XP is distinguished by elevated skin cancer risk, CS by progeria syndrome and TTD by cutaneous abnormalities (Bukowska & Karwowski, 2018). There are no cures for these diseases, current medical treatment include strategies to avoid exposure to sun and other environmental mutagens, dietary restrictions and treatment of symptoms (Bukowska & Karwowski, 2018).

Double-strand break repair (DSBR)

Double-strand breaks are among the most toxic DNA lesions that a cell can possibly encounter. Double-strand breaks (DSBs) in the genome, if unrepaired, can trigger apoptosis of a cell (Lips & Kaina, 2001). Errors in the repair of double-strand breaks can lead to large (up to several kilo bases) deletions, insertions, and rearrangements, which are some of the most dangerous mutations and can cause gross genomic instabilities, one of the greatest hallmarks of cancer.

There are nonetheless instances where double-strand breaks are required in normal cell physiological processes. 1.) Physiological double-strand breaks in DNA replication: One of the initial steps of DNA replication is the unwinding of the DNA helix, which allows polymerases to bind and replicate DNA. Chromatin DNA exists in a supercoiled state. The unwinding of the two DNA strands creates torsional stress on the DNA molecule, which can lead to DNA damage. In order to release torsional stress as well as disentangle DNA, the protein topoisomerase 2 (TOP2) creates transient double-strand breaks by cleaving DNA on both strands (Nitiss, 2009). These breaks are efficiently ligated after entanglement. 2.) Physiological double-strand breaks in adaptive immune cell maturation: The adaptive immune

cells, T- and B-cells, undergo programmed and coordinated double-strand break inductions in the processes of VDJ recombination and somatic hypermutation (SHM) (Malu *et al*, 2012), or, but only in the case of B-cells, class switch recombination (CSR) (Xu *et al*, 2012). These forms of genetic recombination allow immune cell receptors to adapt to a large repertoire of antigens. 3.) Physiological double-strand breaks in meiosis: Cells can either divide through mitosis or meiosis (Kohl & Sekelsky, 2013). Mitosis results in two identical daughter cells from one mother cell, whereas meiosis results in four non-identical daughter cells. In the first step of meiosis, a diploid mother cell, containing one copy of the paternal and one copy of the maternal chromosome, duplicates both copies of DNA, resulting in a tetraploid cell (four copies of DNA). Genetic recombination takes place between homologous paternal and maternal chromosomes (crossing over). This is mediated by programmed DNA double-strand breaks and their repair (Andersen & Sekelsky, 2010). The resulting chromosomes are each a novel mixture of paternal and maternal chromosomes, which become equally segregated into two daughter cells in the first cell division. The two daughter cells divide as well, segregating their DNA equally, resulting in a sum of 4 daughter cells with one copy of mixed paternal / maternal chromosomes. 4.) Physiological double-strand breaks of telomere ends: The ends of telomeres, if unshielded by proteins (shelterin proteins) mimic a double-strand break and induce recruitment of DNA repair factors (Sfeir & de Lange, 2012). This can result in the fusion of chromosomes, constituting a gross genomic instability. Telomere ends therefore require constant protection from double-strand break induced repair.

Cells have evolved two main mechanisms to deal with double-strand breaks: homologous recombination (HR), also called homology directed repair (HDR), and repair by non-homologous end joining (NHEJ). Both pathways require the same damage recognition steps and initial signal transduction (Goodarzi & Jeggo, 2013). Double-strand breaks are sensed by the MRN complex, consisting of the proteins MRE11, RAD50 and NBS1. NBS1 binds DNA and recruits its cofactor ATM. MRE11 is a 3'- 5' exonuclease, which can perform initial resectioning of damaged DNA ends. RAD50 is thought to have a tethering function between the two ends of broken DNA (de Jager *et al*, 2001). Recruited ATM, dimerizes and auto phosphorylates for its activation. Activated ATM functions as a master kinase in the repair of double-strand breaks, recruiting and phosphorylating hundreds of substrates (Matsuoka *et al*, 2007), including H2AX, MDC1, 53BP1 and BRCA1. Phosphorylated H2AX (known as γH2AX), marks the site of DNA damage, and collocalizes with many DNA repair factors (at the site of DNA damage), including the scaffolding protein MDC1. Through mediation of MDC1, the ubiquitin ligases RNF8 and RNF168 recruit 53BP1 which opposes BRCA1 for repair pathway choice between HR and NHEJ (Chapman *et al*, 2012b).

Homology directed repair is the error-free repair choice for cells, but requires a long 3'-single-stranded DNA overhang (or tail) by resectioning and a homologous DNA template for repair.

53BP1 prevents 5'-3' resectioning by the nucleases EXO1, DNA2 and MRE11, via its cofactors RIF1 and PTIP and thereby prevents HR from taking place (Daley & Sung, 2014). At the same time, BRCA1 competes with 53BP1 to prevent 53BP1 activity. NHEJ repair is active throughout the entire cell cycle, but most dominantly in G1, while HR is only active in S and G2 phases of the cell cycle (Chiruvella *et al*, 2013a; Chapman *et al*, 2012a). Therefore, NHEJ, promoted by 53BP1, acts as the default repair mechanism for double-strand breaks, while HR is the favored repair mechanism in the presence of sister chromosomes. Although there have been reports of HR with homologous chromosomes as templates in human cells, such events are very rare compared to repair by sister chromosomes, and are less well studied (Rong & Golic, 2003). This may be due to a closer local proximity or higher sequence similarity (no allelic variances) of sister chromosomes compared to homologous chromosomes.

Homologous recombination (HR) / Homology directed repair (HDR)

5' to 3' resectioning needs to occur on either ends of the broken DNA for HR to take place (creating 3'-overhangs). This is mediated by the nucleases MRE11, EXO1, DNA2. The processing of DNA results in single stranded DNA (ssDNA) on both DNA strands. BRCA1 mediates the recruitment and loading of RPA proteins, which coat the otherwise fragile single-stranded DNA (Daley & Sung, 2014). RPA is replaced by RAD51 filament, via an interaction with BRCA2. The RAD51 filament is able to find, recognize and bind the homologous template sequence on the sister chromosome (strand invasion), forming a crossing over structure (Holliday junction). After strand invasion, DNA polymerases can perform synthesis of the damaged DNA region by using the homologous template. The crossing over structure is resolved by enzymatic activity (involving several possible pathways), thereby restoring the damaged DNA back to its original state without mutations (Matos & West, 2014).

Non-homologous end joining (NHEJ)

53BP1 prevents resectioning of broken DNA ends and promotes NHEJ, also called classical NHEJ (c-NHEJ). DNA-PKcs and its cofactors KU70/KU80, are recruited to the site of damage (Hiom, 2005). KU70/KU80 bind and tether the broken DNA ends. The endonuclease, Artemis performs minimal resectioning, if required, to clean DNA ends
(Povirk *et al*, 2007; Woodbine *et al*) before LIG4 (with its cofactor XRCC4) ligates the two DNA ends. Unless the double-strand break was a clean cut without loss of nucleotides, deleterious mutations would occur after end processing and ligation.
An alternative form of NHEJ (alt-NHEJ) exists in human cells, and requires the availability of short homologies (microhomologies) for repair. DNA is resected until microhomologies between the two strands are encountered and the two strands can bind. The remaining

overhanging DNA is also removed, leaving cells with the loss of DNA sequences at sites of microhomology (microhomology mediated deletions). This subpathway of NHEJ, primarily involves PARP1, POLQ and LIG3 (Chiruvella *et al*, 2013b).

In addition to the core components of NHEJ, other enzymes participate in DSB repair mainly through DNA end processing prior to ligation. These include the translesion synthesis polymerases POLL and POLM (Covo *et al*, 2009; Lee *et al*, 2004).

Fanconi Anemia (FA) pathway of repair

Some endogenous (e.g. acetaldehyde) or exogenous (e.g. cisplatin) chemicals are able to crosslink DNA. They can crosslink bases on the same strand, called intra-strand crosslink, or adjacent bases on opposite strands, called inter-strand crosslink (ICL). Intra-strand crosslinks pose a low threat to normal cells, they form bulky adducts and are readily repaired by the NER pathway (O'Donovan *et al*, 1994). ICLs however, pose a formidable threat to cells because they can block the unwinding of DNA during replication or transcription, and can produce toxic double-strand breaks. In the case of replication, the DNA polymerases stall at replication folks with ICLs. If not repaired, stalled replication forks can lead to replication fork collapse and double-strand breaks, which may trigger apoptosis. The error free repair of ICLs requires Fanconi Anemia (FA) proteins. Fanconi Anemia is a rare disorder, named after Guido Fanconi, a Swiss physician who first described the disorder. There are more than 15 known FA or FA-like proteins, and loss of each one is to some degree associated with the disorder (Yao *et al*, 2013).

Fanconi Anemia Complementation Group (FANC) M, FANCM, recognizes and binds ICLs. This leads to the recruitment of the FA-core complex, consisting of several FA proteins, including FANCC and FANCL. FANCL ubiquitinates and thereby activates FANCI and FANCD2 (FANCI-D2). Activated FANCI-D2 induces recruitment of other repair factors, including BRCA1 and BRCA2 from the homologous recombination pathway, NER factors, TLS polymerases and endonucleases (Nojima *et al*, 2005). The lesion is resolved via endonuclease cleavage of nucleotides flanking the crosslink, followed by NER mediated detachment of the crosslink from one of the two DNA strands (unhooking) and translesion synthesis over the unhooked crosslink (Klein Douwel *et al*, 2014). Finally, the homologous recombination pathway uses the sister chromatid as a template for error free repair of the crosslink. This type of repair is cell cycle dependent since it requires sister chromatids. The exact mechanism and proteins involved in FA repair are still being studied. Interestingly, there is little known about the endogenous sources of crosslinks. Acetaldehydes have only recently been established as a possible source of endogenous crosslinks (Stone *et al*, 2008), particularly as a byproduct of alcohol metabolism (Garaycoechea *et al*, 2018). There is

currently no cure for FA disorders. Understanding of the endogenous sources of crosslinks is very important, such a knowledge could potentially be used in diet-based symptom amelioration of FA disorders. Importantly, FA proteins may have other, hitherto uncharacterized functions, associated with the disease phenotype (Sumpter *et al*, 2016).

DNA damage & DNA damage response in cancer therapy

DNA damage and cancer therapy

Cancer cells are a malignant transformation of normal cells. As such there are characteristics that differentiate cancer from normal cells. One of the goals in cancer therapy, is to find specific vulnerabilities of the cancer, which would allow the targeting of malignant cells without harming healthy ones. The most commonly exploited characteristic of cancer cells is their heightened proliferation rate as compared to the average normal cell. Cells with a high proliferation rate spend more time dividing and replicating their DNA, making them (among others) exquisitely vulnerable to DNA damage by genotoxic agents (i.e. cheomo- or radiotherapy). This has been the main strategy for cancer treatment over the past decades (Jackson & Bartek, 2009a). Ionizing radiation, as well as compounds such as cisplatin or doxorubicin, target DNA or DNA repair factors. Cancer cells however are not the only fast dividing cells in the body, some healthy ones are also highly affected by treatment with genotoxic agents, including cells of the hair follicles and stem cells from different organs. This is the cause of some of the common side effects of cancer therapy (e.g. hair loss).

Exploiting DNA repair in cancer therapy

For a long time, researchers have been trying to find different targets for cancer with little success for all cancers but reasonable success for some cancers. For instance, chronic myelogenous leukemia (CML) is most commonly caused by a fusion of two proteins, BCR and ABL1 (BCR-ABL), due to a chromosomal aberration. This protein does not exist in healthy cells, such that inhibitors against BCR-ABL have been extremely successful in the treatment of CML (The Philadelphia chromosome: a mutant gene and the quest to cure cancer at the genetic level, 2013). Genomic instability is a hallmark of cancer (Hanahan & Weinberg, 2011) and associated with perturbations in DNA replication, DNA repair or DNA damage response pathways. As such, the DNA repair and associated proteins are promising targets for cancer therapy. A well-known example is the use of PARP inhibitors, such as the clinically approved drug, olaparib: BRCA mutations are one of the most common features of breast cancer, and result in ineffective DNA repair by the homologous recombination pathway. Cells with BRCA associated repair deficiency are often rewired to rely on a repair pathway that depends on

PARP. Thus, cancer with loss of function mutations in BRCA are specifically vulnerable to PARP inhibition (Fong *et al*, 2009). Finding novel cancer vulnerabilities is one of the main aims of studying cancer genomes (Futreal *et al*, 2004; Jackson & Helleday, 2016).

Cancer genomes

Mutations in cancer genes

Since the first discovery of oncogenes and tumor suppressors, a quest for finding novel driver genes (cancer genes) has dominated the field of cancer research (Lawrence *et al*, 2013). Modern sequencing technologies have permitted researchers to look for those driver genes in an unbiased manner, across the whole genome. Through sequencing efforts over the past 15 years, some experts in the field even propose that almost all of the most common driver genes have been discovered by now (Vogelstein *et al*, 2013), although others disagree (Martincorena *et al*, 2017). They include many genes involved in the DNA damage response, such as TP53, ATM, BRCA1/2 and MSH6 (Cancer Genome Atlas Network, 2012; Knijnenburg *et al*, 2018). However, some tumors do not have mutations in any of the known driver genes or the regulatory elements of those genes (Vogelstein *et al*, 2013). There are several possibilities for that: 1. There are probably more driver genes to be discovered (Martincorena *et al*, 2017), particularly for specific types of cancer (i.e. not commonly present in all cancers), and though many important driver genes have been discovered, there were also many non-validated reports (Lawrence *et al*, 2013). Finding tools that will accurately predict driver genes from cancer genome sequencing remains one of the most difficult challenges in the field (Lawrence *et al*, 2013). 2. There are also driver mutations, outside of cancer genes and associated regulators. Therefore, it is important to study also non-driver mutations and general patterns of mutations in the genome. 3. There are other driver events not detectable as mutations (i.e. no changes in the sequence of DNA, such as epigenetic changes or posttranslational modifications).

Why we look for mutation patterns

Since the human genome project, advances in sequencing technologies have allowed researchers to study the global structure of the genome, including their distribution across the genome. This has revealed many fascinating insights about the role of non-coding regions of DNA. For instance, at least 97% of DNA in the genome was previously presumed to have no function at all ("junk" DNA), because their sequence did not code for proteins (EHRET & DE HALLER, 1963; Ohno, 1972). It is now acknowledged that many non-coding regions of the DNA form important regulatory elements for genes, such as promoters, enhancers or suppressors (Pennisi, 2012). Mutations in regulatory regions of genes can alter the gene

expression and drive cancer (Shar *et al*, 2016). And even synonymous mutations, mutations in genes that do not alter the coding sequence of genes, can drive cancer (e.g. differential transcription factor binding due to enhancer elements in protein-coding regions) (Supek *et al*, 2014; Li *et al*, 1995). On the other hand, some mutations in cancer genes considered to be causative to cancer, may not actually contribute to cancer development (variants of uncertain significance, VUS (Findlay *et al*, 2018)). Thus, it is important to study mutations that are not in genes or cancer driver genes.

Proteins most often interact in a complex, or in a pathway, with other proteins. This principle is so important that we have terms to describe the collective of dysfunctions that result in the same or similar phenotype as mutations in a main driver gene. BRCAness (Lord & Ashworth, 2016) describes all molecular dysfunctions that mimic the molecular features of BRCA1/2 dysfunctions (e.g. EDC4 phenocopies BRCA (Hernández *et al*, 2018). Furthermore, a protein's localization, modifications and interaction with other proteins are important for its function, and perturbations on that level can influence cancer development but may not be detected by studying driver genes or driver mutations. Therefore, besides testing for the dysfunction of known driver mutations in cancer, it is also important to have methods that report the dysfunction of entire pathways and complexes or molecular modifications that are not detected by only looking at driver mutations. Testing for mutation patterns in genomes gives information about biological processes, and is thus a method that can actually report dysfunctions in pathways, protein complexes and protein modifications.

Well established driver mutations do not only associate with malignant tumors (cancer) but also with benign ones (Pollock *et al*, 2003; Bauer *et al*, 2007). The finding of mutations in cancer drivers does therefore not always imply a causative role of those mutations in malignant tumor development. Testing for mutation patterns can reveal mutator phenotypes that can be associated with malignant diseases as oppose to benign ones.

Furthermore, studying mutation patterns can also lead to identification of novel driver events in cancer. For instance, based on an identified mutation pattern, Alexandrov et al proposed a strong prevalence of APOBEC / AID family members' involvement in the development of some tumors (Alexandrov *et al*, 2013a).

Taken together, this means that once we have successfully catalogued all mutations in all cancer genes, some of these mutations may not have an influence on cancer evolution, or only have an influence in a context dependent manner (e.g. tissue). Furthermore, we would still find a substantial amount of mutations in the genome that promote cancer without being in known genes. It is therefore important to find other markers, besides cancer genes, that are associated with cancer development. Mutation patterns are promising markers (another targetable molecular feature) that can complement driver mutation screens (Garraway, 2013) and improve personalized cancer treatment (Davies *et al*, 2017).

Mutation patterns in cancer genomes

Mutations and mutation types (e.g. substitutions, indels, rearrangements, and copy number variations) are differentially distributed across the genome due to factors such as differences in the sequence, structure, localization and function of different regions in the genome as well as their exposure to endogenous and exogenous mutagens. This heterogeneous distribution of mutations defines specific mutational patterns in the genomes of cells that can serve as a proxy for biological processes (Figure 3).

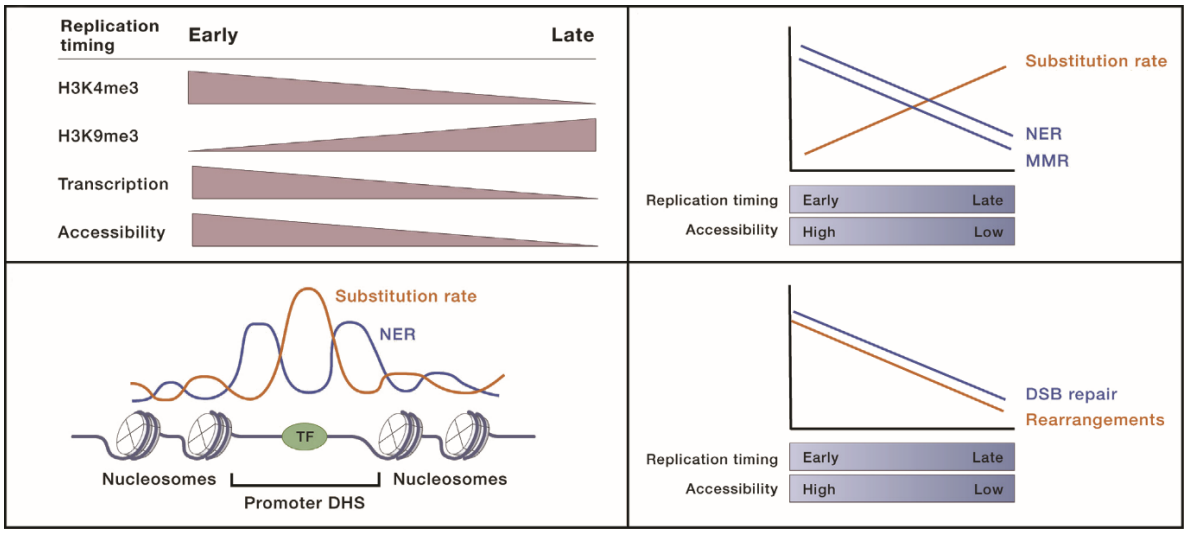


Figure 3: Differential distribution of mutations across the genome

Some of the factors that influence the differential distribution of mutations in the genome. Left: replication timing and chromatin accessibility. Right: DNA repair pathways. Figure was taken from (Tubbs & Nussenzweig, 2017).

TP53 is the most commonly mutated cancer gene. This is in part due to its function as a tumor suppressor, and to another part due to its sequence and sequence structure. Most of its loss of function mutations cluster between exons 5 to 8. These exons encompass the DNA binding domain, crucial for its functional activity as a transcription factor. Exons 5 to 8 together, span a region of 540 nucleotides (180 codons), including sequence structures, such as CpG dinucleotides and pyrimidine dimers. The large number of nucleotides and diverse sequence structures provide a wide window for possible mutations of different types that may result in a loss of function of the protein. Other cancer genes may require specific mutations at a shorter nucleotide sequence range or have a less dominant function compared to TP53 (Pfeifer & Besaratinia, 2009). BRCA1 for instance requires point mutations or insertion / deletions that result in protein truncation or mRNA decay. This narrows down the amount of random mutations and the chance of them resulting in a loss of function of the protein and is thus less

favorable for cancer evolution, compared to the TP53 gene. Another example is the oncogene KRAS, which is very frequently mutated in cancer. The known carcinogenic mutations are all in codon 12 or 13 of exon 2, which span a region of 6 nucleotides and therefore offer a very narrow nucleotide range for mutations. Natural selection in the tumor microenvironment would therefore favor mutations in TP53 compared to KRAS. The sequence of DNA is thus a factor for the heterogeneous distribution of mutations in cancer genomes.

DNA replication and repair influence the distribution of mutations in the genome. Watanabe et al observed a differential distribution of mutation frequency in early versus late replicating sites, looking at chromosomes 11q and 21q (Watanabe *et al*, 2002). Early replicating regions have a lower mutation burden than late replicating ones. Stamatoyannopoulos et al observed the same phenomena, looking at 1% of the human genome (Stamatoyannopoulos *et al*, 2009). The authors speculated that the bias was likely due to accumulation of ssDNA. However, Supek et al showed that this phenomena was actually due to MMR, which appears to be more active at early replication (and euchromatic) regions (Supek & Lehner, 2015), but the reason for this bias in MMR still remains to be revealed.

The occupancy of DNA by DNA-binding proteins influences the frequency of mutations. Sabarinathan et al observed that active transcription factor binding and nucleosome core DNA (bound by histones) sites, have a higher mutation burden than their flanking regions (Sabarinathan *et al*, 2016). The authors show that the higher mutation burden was due to a decreased activity of NER at those sites. The DNA-binding proteins could be physically preventing repair proteins from accessing sites of damage, leading to less efficient repair and a higher mutation rate.

The two strands of DNA are also differentially exposed to damage and differentially accessible for repair. The non-coding strand is less enriched for mutations compared to the coding strand. It has been suggested that this was due to TC-NER, which is active on the non-coding strand (Haradhvala *et al*, 2016), but in addition to that, transcription factors binding to the DNA could protect the non-coding strand from mutagenic attacks, whiles the coding strand would be both unprotected by TF and not repaired by TC-NER. Similarly, the leading and lagging strand show differential mutation burdens, with the lagging strand generally having a higher mutation burden (Pleasance *et al*, 2010a). Due to discontinuous replication of the lagging strand, it is plausible that the lagging strand is less shielded by replication proteins and thus more susceptible to mutagenic attacks than the leading strand (Haradhvala *et al*, 2016).

Mutations induced by extrinsic or intrinsic mutagens are more enriched in the non-coding and the lagging strand. The most likely reason for this is that parts of DNA are easier accessible to damage during transcription or replication (Haradhvala *et al*, 2016). Such mutations, at times, appear in the form of clusters (clustered mutations), regions in DNA with unusually high mutation burdens. The APOBEC signature for instance, which is a form of clustered mutations,

is highly associated with the non-coding or the lagging strand (Haradhvala *et al*, 2016). Erroneous repair by TLS polymerases may also cause clustered mutations by directing carcinogen induced mutations to particular sites in the genome (Supek & Lehner, 2017).
Moreover, some mutations are enriched in specific organs or tissues (Blokzijl *et al*, 2016). Other mutations, particularly spontaneous ones, increase with age (Blokzijl *et al*, 2016; Alexandrov *et al*, 2015).

Mutation signatures: a brief history and explanation

The first mutation patterns were identified by applying exogenous DNA damaging agents and analyzing nucleotide changes in the sequence of DNA (Tessman *et al*, 1964). Typical models for such studies were single stranded DNA viruses, due to their small and simple genome. Forward and reverse mutations at single nucleotide sites in the viral genome were coupled to a phenotypic readout, such as plaque formation capability or temperature sensitivity, in order to assess changes in the sequence. Studies with UV were the most elucidative once (Howard & Tessman, 1964; Setlow & Carrier, 1966; Witkin, 1969). It was discovered that, among other lesions, UV exposure predominantly caused a unique type of mutation pattern, C > T or CC > TT, resulting from pyrimidine dimers at dipyrimidine sites in DNA. As cloning and sequencing technologies improved and became more widely available (Sanger & Coulson, 1975; Mullis *et al*, 1986), so did the ability of researchers to synthesize and analyze specific genomic regions for mutational changes. Genomic loci, such as the locus for the TP53 or KRAS gene, which are frequently mutated in cancer, were analyzed extensively for endogenously or exogenously induced mutational changes, including specific patterns (Capella *et al*, 1991; Brash *et al*, 1991; Ozturk, 1991; Sidransky *et al*, 1991; Hollstein *et al*, 1991). It was not long until whole genome sequencing technologies were developed and a decade later, at the beginning of the 21st century, entire human genomes, including human cancer genomes were being sequenced. For the first time in history, researchers could look at variations between different types of cancer at an unprecedented scope and resolution.
Researchers immediately used the opportunity to study mutations in cancer genes on a genome scale. After a while, instead of looking at mutations in cancer genes, a few researchers, particularly a team lead by Sir Michael Stratton, had started developing algorithms for studying patterns of mutations in cancer genomes (Futreal *et al*, 2001; Greenman *et al*, 2007; Stratton *et al*, 2009). The goal was to identify disease relevant patterns and use them as biomarkers for inferring the etiology of the underlying mutagenic processes, as well as the etiology of the respective cancer. These specific patterns are also referred to as mutation signatures.

The meanings of the two words, mutation "pattern" and mutation "signature", are still evolving in the scientific community and are either used interchangeably, as in one of the first mentions of mutation signatures (Greenman *et al*, 2007), or such that mutation signature is a special case of mutation patterns (Pleasance *et al*, 2010a). In the first mentions of mutation signatures, it seemed to be restricted to disease associated mutational patterns with respect to base substitutions, due to technical limitations of the time. As the algorithms for data assembly and analysis improved, additional known disease associated patterns, especially rearrangements (Campbell *et al*, 2008) and indels (Ley *et al*, 2008; Mardis *et al*, 2009), could be analyzed. In addition, novel mutation patterns were discovered, such as patterns associated to: replication timing and chromatin accessibility (Supek & Lehner, 2015), DNA binding factors (Sabarinathan *et al*, 2016), replicated or transcribed strands (Haradhvala *et al*, 2016). Nowadays, with many researchers focused on the discovery of novel patterns, we expect the discovery of more mutagenesis-associated patterns in the years to come.

The first demonstration of the possibly causal nature of a specific mutation signature and cancer, was published in 2010 (Pleasance *et al*, 2010a). The authors sequenced one metastatic malignant melanoma and compared it to one lymphoblastoid cell line derived from the same patient. The most prominent signature in their analysis, was the C > T base substitution at dipyrimidine sites, which is uniquely associated with UV induced mutations and had previously been shown to cause skin cancer (Rusch & Baumann, 1939; Brash *et al*, 1991). Their analysis was however limited to base substitutions and chromosomal rearrangements due to lack of algorithms for the analysis of other mutation types. In the same year the authors published another article, where they confirmed a previously reported mutation signature, associated with tobacco smoke (Pfeifer *et al*, 2002), in the genome of a small cell lung cancer (Pleasance *et al*, 2010b). They mainly focused on base substitutions and rearrangements. Interestingly, similar to their previous paper (Pleasance *et al*, 2010a), they reported differences in mutation rates and types between the transcribed and non-transcribed DNA strand (and gene expression). They suggested differential DNA repair, instead of differential mutagenesis to be responsible for the observed transcription strand bias.

Shortly after Alexandrov and Nik-Zainal published several papers on mutation signatures, which laid the path for many researchers to follow (Nik-Zainal *et al*, 2012; Alexandrov *et al*, 2013b; 2013a).

Algorithms for analyzing mutation signatures

Since the beginning of the 21st century, sequencing technologies have been improving, making the sequencing of whole genomes more affordable. We moved from studying single cancer genomes with their matched controls (Pleasance *et al*, 2010a; 2010b), to studying

dozens (Nik-Zainal *et al*, 2012), hundreds (Alexandrov *et al*, 2013a) and now even thousands of cancer genomes with their matched controls (Alexandrov *et al*, 2018).

The availability of large data sets requires algorithms to extract meaningful information. Although dominant signatures can be observed by eye (Nik-Zainal *et al*, 2012), most signatures are subtle and require mathematical rigor for unbiased calls. The choice for such algorithms can be based on prior knowledge or reasonable assumptions about the data and the information that is to be gained from it. In the case of cancer genomes, we know that different genomes have different types of mutations that are prevalent to different extends (Pfeifer & Besaratinia, 2009). Different types of mutations are manifested as different patterns in cancer genomes (Figure 4). We can assume that the number of those patterns is finite, even if it was extremely large. Moreover, there could exist a small subset of all the numbers of mutation patterns, which in combination is sufficient to represent almost all of the mutational processes that are active in almost all cancer genomes.

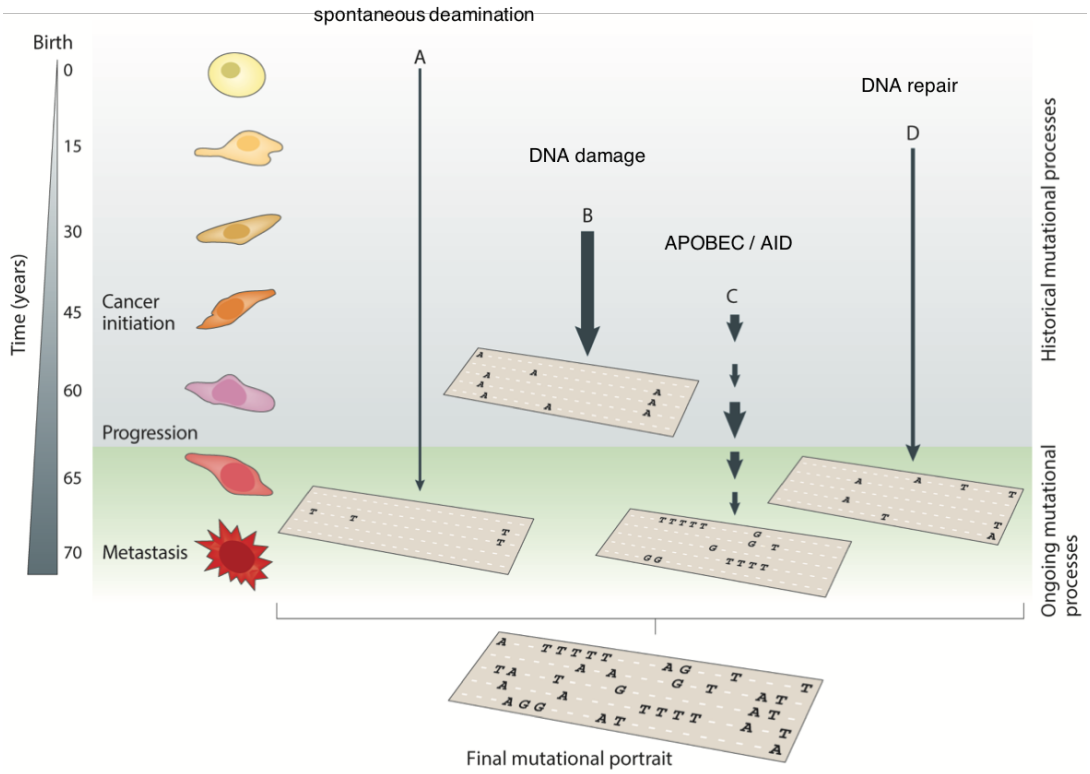


Figure 4: Clonal evolution of cancer and mutation signatures

Depiction of cancer development from normal to malignant state. Different mutagenic processes are active during the lifetime of a cell. These processes leave behind specific patterns in the genome. Figure was adapted from (Helleday *et al*, 2014).

Mathematically the set of processes could be called P, with elements p_n and n = 1, ..., N. According to our definition, the set of patterns or mutation types could be represented as Ξ with K letters where each k element of Ξ represents a specific mutational pattern and each process p_n would be a combination of patterns. We can therefore describe p_n as a probability, the probability that a pattern k contributes to a process p_n. Thus, a process p_n has a probabilistic distribution of patterns with the sum 1, this distribution defines a signature for each process. Mathematically, this means (Alexandrov *et al*, 2013b):

$$\sum_{k=1}^{K} p_n^k = 1 \; and \; p_n^k > 0, k = 1, \ldots, K.$$

In addition to patterns that define signatures for mutational processes, we can expect the impact of a signature to be different in different cancer genomes. For instance, UV acts as a mutational process with a distinct signature, C > T or CC > TT transitions. The UV signature contributes to a much larger amount of mutations in melanoma (skin cancer) than in other cancer types, due to the greater exposure of the skin to UV (Brash *et al*, 1991). In other words, different cancer genomes are differentially exposed to different mutational processes (Figure 4). We can introduce the term e to represent exposure, e, would be dependent on a cancer genome g and would be specific for the mutational process p_n: $e_g^n \geq 0, with \; g = 1, \ldots, G \; and \; n = 1, \ldots, N$.

An individual cancer genome could thus be described as a mutational catalogue m_g, the sum of all the signatures of mutational processes times their exposure. The mutation type k, of all operational processes and their exposures in a cancer genome g would be expressed in this manner (Alexandrov *et al*, 2013b):

$$m_g^k \approx \sum_{n=1}^{N} p_n^k e_g^n.$$

This term can be generalized for all K mutation patterns and G genomes by expressing exposures to mutational processes and mutational catalogs as matrices (Alexandrov *et al*, 2013b):

$$\begin{pmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \ldots & m_{G-1}^K & m_G^K \end{pmatrix} \approx \begin{pmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \ldots & p_{N-1}^K & p_N^K \end{pmatrix} \times \begin{pmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \ldots & e_{G-1}^N & e_G^N \end{pmatrix}$$

A simplified depiction of the matrix factorization is $M \approx P \times E$. For better understanding of the theoretical nature of our problem, we started our argument with mutational processes (P), exposures of genomes (E) and ended with mutational catalogues of cancer genomes (M). In a practical setting, our problem however starts with having the genome sequence of several cancer samples, which are analogous to the mutational catalogues of cancer genomes (M), and we are trying to decipher the mutational signatures by finding the factors (P) and (E). This was a strange problem in the field of molecular medicine but a well-known one in the field of mathematics, called blind signal separation or blind source (BSS) problem(Handbook of Blind Source Separation, 2010). The challenge is to find the source of signals from a mixture of signals, with little or no knowledge of the sources or how they were mixed. Originally applied for temporal separation of audio signals, it has found general application to multidimensional data, such as images or in our case mutation patterns in genomes. There are several approaches for solving such a problem, including principal component analysis (PCA) and non-negative matrix factorization (NMF). NMF is particularly convenient for our type of problem, since it requires the data to be non-negative, as is the case for genome data, and has previously produced meaningful information from biological data (Lee & Seung, 1999; Berry *et al*, 2007). Since its first applications in the analysis of genome data (Nik-Zainal *et al*, 2012; Alexandrov *et al*, 2013b; 2013a), NMF based methods have become the primary methods for extracting mutation signatures in cancer genomes (Alexandrov *et al*, 2018).

The accuracy of deciphering mutation signatures from cancer genomes depends on different factors, e.g. the number of available cancer genomes, number of mutations and number of signatures to be extracted (Alexandrov *et al*, 2013b). In general, the more genomes and the higher the exposures (more mutations), the more signatures can be deciphered with greater accuracy, but having more mutations facilitates the deciphering process more than having more genomes. Conversely, the higher the number of operational mutational processes or mutation patterns in the cancer genomes, the lower the accuracy of NMF algorithms to detect signatures. Therefore, combining different cancer types could result in more exposure for a given signature (which is the rational for studying mutation signatures in the combined genomes of different cancer types). On the other hand, combining different cancer types could potentially result in more signatures to be analyzed (which would require even more genomes to increase accuracy). Assuming that the number of main mutational processes in cancer genomes is limited, the potential risk of combining different types of cancer genomes would be lower than the potential benefit.

Types of patterns / mutation signatures in the genome

Base substitutions

There are four bases in DNA: C, T, A, G. In its most basic sense, a single base substitution is the change of a base (or nucleotide) in DNA to another base (or nucleotide). They are by far the most abundant type of mutations in genomes. For instance, Pleasance et al found 33345 base substitutions, 66 indels, and 37 rearrangements in the genome of a melanoma cancer (Pleasance *et al*, 2010a). There are 6 possibilities for base substitutions, which define 6 mutation types. Those are C:G > A:T, C:G > G:C, C:G > T:A, T:A > A:T, T:A > C:G, T:A > G:C. They are commonly abbreviated to C>A, C>G, C>T, T>A, T>C, T>G, in which the mutated base is represented by the pyrimidine of the Watson-Crick base pair. In that sense a C > A and G > T are both represented by C > A. Studying base substitutions alone, can already teach us a lot about mutational processes (Table 2).

Table 2 Example of base substitution mutations and associated processes

Repair	Source of damage	Type of lesion	Substitution	Cancer incidence	Comment
TC-NER	fungi toxin	aflatoxin	C > A	liver cancer	contaminated crops
BER	reactive oxygen species	not known	C > A	many cancer types	associated with loss of MUTYH
BER	anti viral defence	APOBEC / AID family	C > T	many cancer types	associated with kataegis
DR	alkylating agent	O6meG	C > T	glioblastoma	after TMZ treatment
MMR	replication	replication mistakes	C > T	colorectal cancer	repeat mediated deletions
NER	ultrviolet light	CPD dimer	C > T	skin cancer	transcripton strand bias
NER	benzo[a]pyrene (smoking)	bulky adducts (guanine)	G > T	lung cancer	transcription strand bias
TC-NER	herbal medicine	aristolochic acid	T > A	urothelial cancer	strand bias

Early reports of base substitutions in cancer genomes described base substitutions in the form of single nucleotide polymorphisms (SNP) or single nucleotide variants (SNV), without a particular focus on reporting the actual base changes or a pattern therein (Ley *et al*, 2008; Mardis *et al*, 2009). A single nucleotide variant is any change in a base at a specific locus, whereas a single nucleotide polymorphism is a SNV that is common in a population. Later publications focused on reporting base changes in the form of the 6 mutation types (Pleasance *et al*, 2010a), which could already be used as signatures for mutational processes. There is however some ambiguity in only looking at single base substitutions. As soon we are looking for signatures for more than 6 processes, at least 2 processes would have the same signature (Figure 5). Researchers found ways to increase the specificity of associating signatures with their underlying processes, by increasing the mutation types and subtypes.

One possibility is to take the sequence context into account, by for instance searching for dinucleotide changes instead of single nucleotide changes (Nik-Zainal *et al*, 2012). This may determine specific patterns, like the UV signature CC > TT. Another possibility is to look at the

adjacent 5' and 3' base, next to the mutated base (trinucleotides), e.g.: TpCpG > TpTpG, with the mutated base underlined (Figure 5). This results in 16 subtypes for each one of the 6 mutation types, and 96 mutation subtypes overall (Nik-Zainal *et al*, 2012). This can naturally be extended to many more nucleotides, and the challenge is to find a balance between an increase in mutation subtypes and the loss of information about specific mutations. The last base substitution subtypes, which are noteworthy, are the two adjacent 5' and two adjacent 3' bases next to the mutation (pentanucleotides), e.g.: ApTpCpGpT > ApTpTpGpT, with the mutated base underlined. This results in a total of 1536 mutation subtypes (Alexandrov *et al*, 2013b). Alexandrov et al tested NMF based algorithms on 21 breast cancers and found that using the 1536 pentanucleotide signatures resulted in fewer but more specific signatures than using the 96 trinucleotide signatures (Alexandrov *et al*, 2013b). Therefore dividing base substitutions into more mutation types leads to fewer mutations per mutation type, decreasing the amount of signatures or processes, but produces some highly specific patterns.

Figure 5: Rational for 96 base substitution subtypes

Figure was inspired and adapted from Lupski Lecture 2015 by Prof. Mike Stratton, Wellcome Trust Sanger Institute, UK, at Genomics of Rare Disease: Beyond the Exome (29 April - 1 May 2015).

On the next level of complexity, base substitutions can be combined with other mutation types, such as indels and rearrangements, or other factors that affect the distribution of mutations, such as transcription strand bias, replication strand bias, and clustered mutations. Different combinations may yield different results, and increasing the complexity may not always improve the accurate identification of specific signatures. Alexandrov et al found that either a combination of 96 trinucleotides with dinucleotides, indels and kataegis or a combination of 96 trinucleotides and strand bias (192 trinucleotide substitutions) resulted in largely unchanged signatures compared to the 96 trinucleotides alone, except for the emergence of an additional signature which was characterized by kataegis (Alexandrov *et al*, 2013b).

Overall the 96-base substitution subtypes along with indels, rearrangements and strand biases are the most commonly reported signatures and result in a proper representation of the mutational landscapes of cancer genomes.

Insertions / deletions (indels)

First algorithms for studying indels (along with substitutions) in genomes were published in 2008/2009 (Ley *et al*, 2008; Mardis *et al*, 2009). Indels are distinguished by their type (insertion or deletion), size (commonly 1 – 3 bp) and their sequence context (sites of repeat or microhomology). Repeat mediated indels are associated with failure in the MMR pathway (Figure 6). Microhomology mediated indels are associated with error prone NHEJ repair pathways (Figure 6). Indels are the second most common types of mutations identified in genomes (Pleasance *et al*, 2010a).

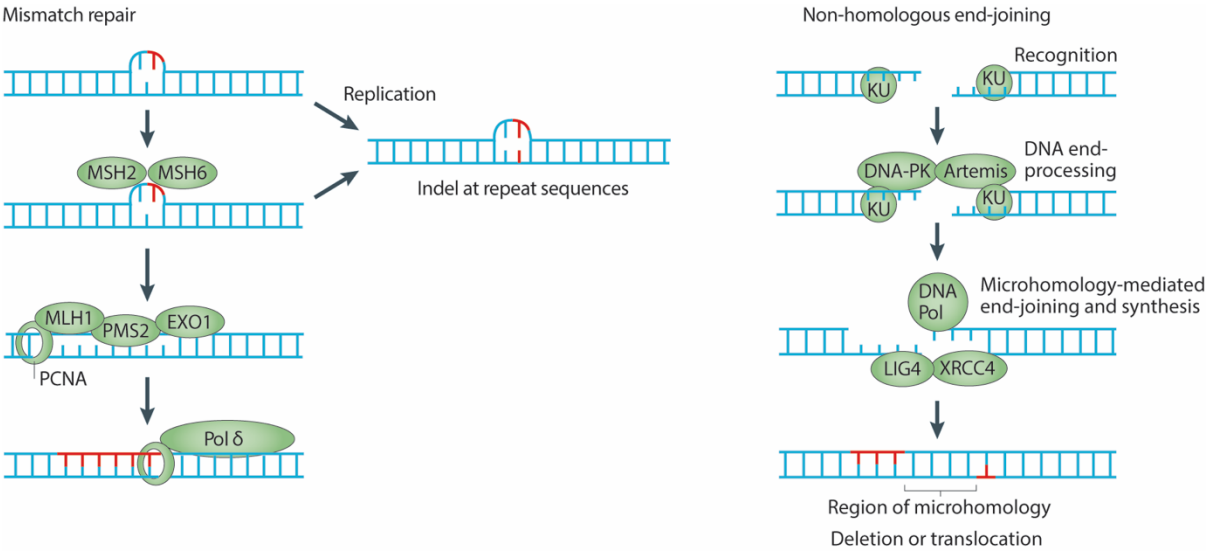


Figure 6: Repeat mediated versus micro-homology mediated insertions or deletions

Left: Mismatch repair (MMR) is responsible for the repair of single base insertions and deletions at repeat sequences. Timely recruitment of MMR factors leads to error-free repair. Deficiency in MMR results in insertions/deletions (indels) at repeat sequences. Right: Homologous Recombination (HR) repair is an error-free double-strand break repair (DSBR) pathway. In the absence of HR, lesions are repaired by the error-prone DSBR pathway, NHEJ. The repair process is associated with deletions and translocations at regions of microhomology. Figure was adapted from (Helleday *et al*, 2008).

Rearrangements

Rearrangements are mutations that juxtapose nucleotides that are normally separated in the genome (e.g. on two different chromosomes). They do not simply change the structure of the DNA sequence but also affect a higher order of DNA structure, the chromosomes.

Rearrangements can be separated into types: Insertions, Deletions, Duplications, Inversions and Translocations (Figure 7). Furthermore, the mutation types can be separated into sizes (e.g. 1 – 10 kb, 1 mb). Computationally detecting rearrangements had been quite challenging on the data analysis level. Some of the first algorithms were published in 2008/2009 (Campbell *et al*, 2008; Stephens *et al*, 2009) but algorithms have improved since then (Alexandrov *et al*, 2013b). Interestingly Alexandrov et al noticed a strong association in occurrence of rearrangements and kataegis (Alexandrov *et al*, 2013b), possibly suggesting an unknown link between the two events.

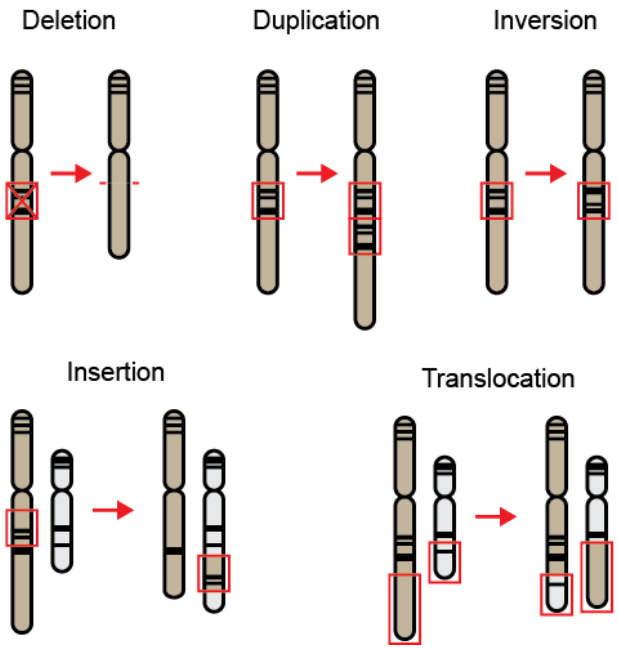


Figure 7: Types of rearrangement mutations

A depiction of the different types of rearrangement mutations: deletions, duplications, inversions, insertions and translocations.

Other structural or topographical changes

There are more mutation types or structural variations that are at times included in the study of mutation signatures, such as copy number variations (CNV) or topographical changes in the genome. One of the most mysterious phenomena is chromothripsis (Stephens *et al*, 2011) (and similar events such as chromoplexies (Baca *et al*, 2013)), where some parts of the genome are reshuffled in a single mutagenic event. The exact mechanisms for such genomic catastrophes are not resolved, but current models involve DNA damage, DNA repair and DNA replication processes. The study of mutation signatures is not yet at a fully mature stage, and thus the meaning is still evolving and adapting to novel observations. Perhaps in the near

future, mutation signatures may evolved to include epigenetic marks, gene expression profiles and more.

Mutation signatures & molecular markers in cancer

Mutation signatures as markers of biological processes

The study of mutation signatures is still at an early stage, but even now, applications are being developed for clinical purposes (Davies *et al*, 2017). Mutation signatures can serve as molecular markers for cancer (Alexandrov *et al*, 2013a) (Figure 8). They can be used to distinguish diseased and healthy state from each other, but more importantly, they can be used to distinguish different types of diseased states from each other. Mutation signatures are defined by DNA damage or DNA repair processes and have been predominantly studied in cancer, where they are used to profile different types of cancer (Alexandrov *et al*, 2013a) (Figure 8). However, the study of mutation signatures is already being extended to other biological problems, such as rare diseases (Garaycoechea *et al*, 2018) and neuronal development (Lodato *et al*, 2017).

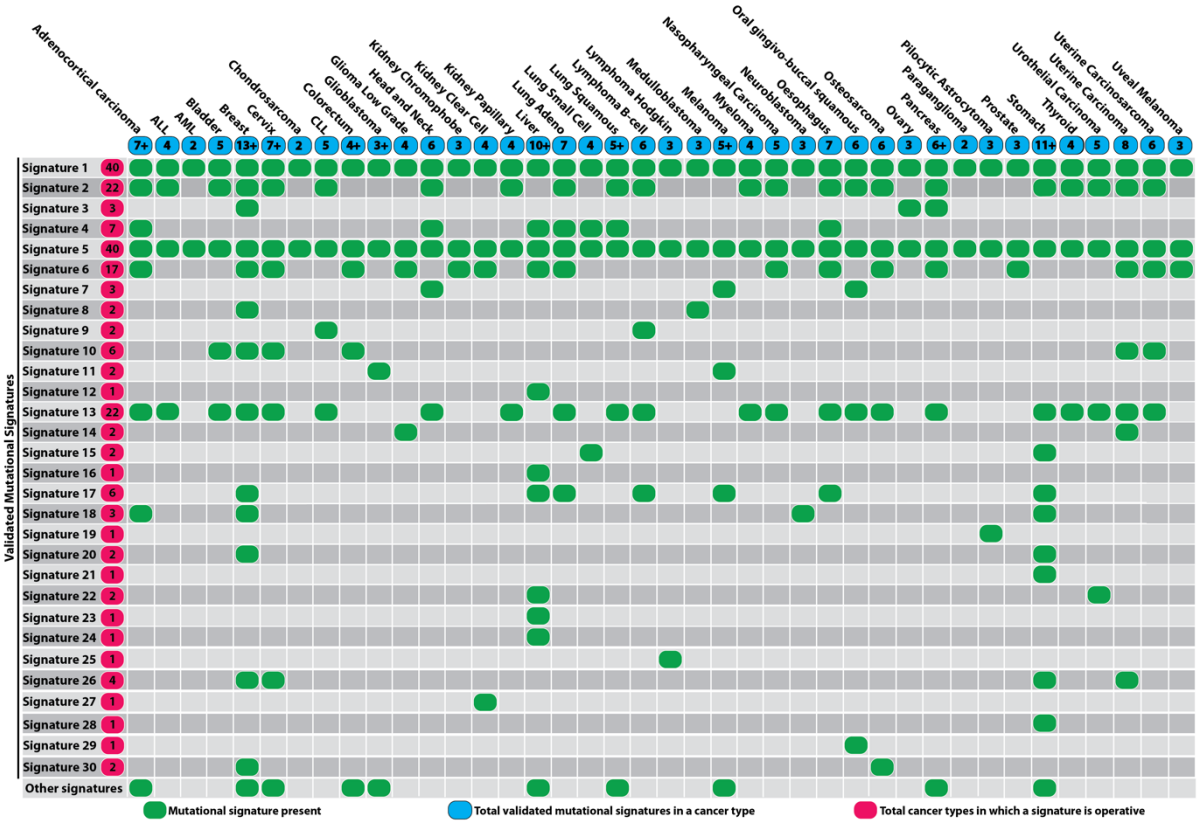


Figure 8: Mutation signatures across different cancer genomes

Thirty mutation signatures and their presence in different cancer types. Figure taken from Cosmic Catalogue of Somatic Mutations in Cancer (Wellcome Trust Sanger Institute).

Through the studies of mutation signatures, we gain an understanding of important mutagenic processes that result in cancer development. This knowledge can already prevent some cancer incidences through public education. On the other hand, diagnostic tools are being developed to detect cancer at an early stage (Lawrence *et al*, 2013; Davies *et al*, 2017). Finally, mutation signatures can be used to stratify cancer patients into treatment groups that will respond better to therapy, e.g. PARP inhibitors for cancers with HR deficiency signatures or immune checkpoint blockades for cancers with MMR deficiency signatures.

Re-categorizing cancer with mutation signature analysis

The current standard markers for breast cancer include the mutation status of BRCA1/2, TP53 or HER2 and the expression status of hormone receptors estrogen and progesterone. Nik-Zainal et al showed that using mutation signatures to classify breast cancers resulted in different categorizations than what could be obtained by using some of the standard molecular markers (Nik-Zainal *et al*, 2012; Morganella *et al*, 2016), indicating that there is space for improvement in categorizing breast cancer, and mutation signatures could complement current standard molecular markers (Davies *et al*, 2017). For instance, mutation signature analysis, identified a minority of MMR deficient breast cancers that could show better response to an alternative treatment rather than standard breast cancer therapies (Davies *et al*, 2017).
Microsatellites are sequences of repeats (tandem repeats). Repeats between 10 to 60 nucleotides are called minisatellite, whereas repeats with fewer nucleotides are known as short tandem repeats or microsatellites. Microsatellite instability (MSI) is a term for high mutation rates in microsatellites, and used as a molecular biomarker for some cancer types. For instance, MSI contributes to 15% colon (Boland & Goel, 2010), 22% gastric (Cancer Genome Atlas Research Network, 2014), 20 - 30% endometrial (Kunitomi *et al*, 2017) and 12% ovarian (Pal *et al*, 2008) cancer. MSI is commonly associated with a deficiency in MMR. However, there are MSI cancer without known mutations in MMR and MMR deficient cancers without MSI (Cortes-Ciriano *et al*, 2017). Recently, an immunotherapy, which works surprisingly well for MSI / MMR deficient cancers, has been developed. Immune checkpoint blockades allow immune cells to better recognize and eliminate cancer cells. Due to their high mutation rate, MSI / MMR deficient cancer, produce many tumor antigens that further facilitate their recognition by immune cells. Immune checkpoint blockade is widely assumed to be among the most promising emerging approaches in cancer treatment (Shen *et al*, 2018). However, it currently benefits only a limited subpopulation of patients. Therefore, there is an urgent clinical need to identify molecular tumor subtypes that are likely to benefit from specific immunotherapies (Shen *et al*, 2018). The application of mutation signatures has already shown the capacity to do so by recognizing MMR deficient tumors in different cancer types,

even if there is no detectable underlying mutation in a MMR gene (Davies *et al*, 2017). Furthermore, mutation signatures may be used to classify MMR deficiency into more meaningful categories based on their mutation patterns (Morganella *et al*, 2016) which could improve patient stratification into better responding treatment groups.

Though, up to date there are more than 30 signatures and different people using different algorithms to study mutation signatures, the results are predominantly similar (Alexandrov *et al*, 2018).

Modelling mutation signatures *in vitro*

Motivation

After a decade of studying mutation signatures, the mechanistic basis of some signatures is either partially or well understood. For many signatures however, the etiologies are unknown or remain speculative (Alexandrov *et al*, 2018). At the time of detection, the average cancer has undergone decades of evolutionary selective growth in a human body (Figure 9). During that time, the parental tumor will have given rise to millions of sub-clones that each have small to large variations to each other's genome. The most adapted to the tumor microenvironment will have been preserved, while many others will have perished, along with their unique genetic makeup. Such loss of information and other confinements make it virtually impossible to establish a causality between mutation signatures and their underlying mutational processes (Figure 9). Nor are we able to establish the timing or sequence of events that would allow us to better understand and possibly prevent the evolution of malignant tumors. Besides the biology, technical difficulties such as sample availability and tumor heterogeneity from patients also pose formidable challenges in the study of cancer genomes. Herein lies the strength of *in vitro* generated mutation signatures. We have established a protocol, wherein, we artificially control the microenvironment as well as the genetic starting material of isogenic cell lines. This allows us to induced individual mutagenic perturbations and follow the resulting mutation signatures over time, making it possible to determine a causal link between signatures and mutagenic processes (Figure 9). Moreover, *in vitro* studies may result in the revelation of novel mutation signatures, which have hitherto not been discovered *in vivo*, due to low prevalence, or other impairments. Thus, the *in vitro* studies may inform the *in vivo* studies and lead to better therapies for subgroups of patients.

In vitro mutation studies have already been used to confirm the mutation signatures of chemical or environmental mutagens, such as UV (Saini *et al*, 2016), aristolochic acid (Poon *et al*, 2013) and aflatoxin (Huang *et al*, 2017). Others have used genetic perturbations to study mutation signatures in C. elegance (Meier *et al*, 2014; 2017) and human organoids (Drost *et*

al, 2017; Blokzijl *et al*, 2016). To our knowledge, we are the first to obtain mutation signatures from human isogenic cell lines with loss of specific DNA repair genes.

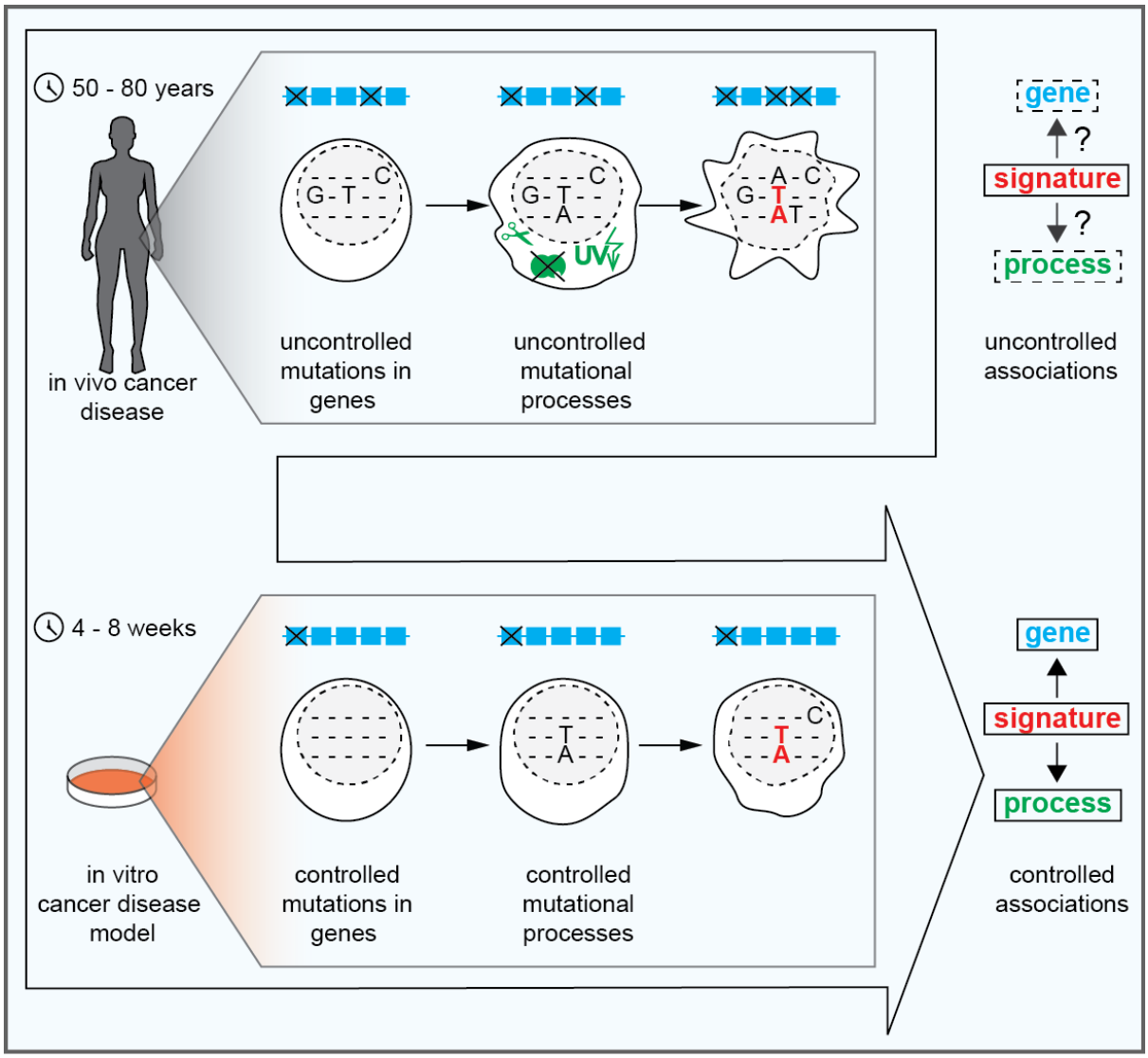


Figure 9: *In vitro* cancer mutation signatures

The strategy and goal of *in vitro* mutagenesis: starting from a signature with no validated association and resulting in causal relationships between signatures and mutagenic processes.

Strategy

The human HAP1 cell line is a derivative of the human KBM7 cell line, which was isolated from a male CML patient (Andersson *et al*, 1995; Kotecki *et al*, 1999; Carette *et al*, 2011). HAP1 cells are almost entirely haploid (except for a portion of chromosome 15), and in contrast to KBM7 cells, are not dependent on the BCR-ABL1 fusion protein. The advantage of working with haploid cells is that genome editing on one copy of a gene is more efficient than on two or more. In addition, twice the amount of experimental conditions can be tested on a haploid

genome compared to a diploid genome when it comes to sequencing costs. We expanded a single HAP1 cell to obtain a pool of isogenic cells, and used CRISPR-Cas9 gene editing technology to delete individual DNA repair genes in those cells. Since the generated knockout cell lines were all derived from the same pool of mother cells, mutations that accumulated over time in the knockouts could be traced back to specific mutagenic processes (i.e. the missing DNA repair genes).

We selected a panel of versatile DNA repair genes that cover different pathways and stages of DNA repair. NUDT1 removes damaged nucleotides from the nucleotide pool, preventing mutations during DNA synthesis. DNA replication polymerase, POLE, is responsible for DNA synthesis, detection and repair of single base substitutions. MMR genes MSH6 and EXO1 are involved in the repair of mistakes that escape the proofreading function of POLE (and POLD). If required, TLS polymerase, POLM, assists replication polymerases with its translesion bypass function, at times creating de-novo mutations in that process. Base excision repair genes NEIL1 and POLB take part in the repair of single mutated bases, while FANCC and EXO1 are involved in the repair of toxic DSBs, especially after replicative stress or the encounter of endogenous ICLs (Lopez-Martinez *et al*, 2016). The checkpoint kinase CHK2 plays an important role in cell cycle arrest after DNA damage, protecting cells from mutations by providing them with sufficient time for DNA repair.

Aims of this thesis

The aims of this thesis were to: 1.) Establish experimental conditions that would allow for *in vitro* mutagenesis by deletion of endogenous genes in human isogenic cell lines. 2.) Develop algorithms for the analysis of *in vitro* mutational patterns in the in vitro mutagenized cell lines. 3.) Compare mutational patterns identified *in vitro* with existing *in vivo* mutation patterns. 4.) Suggest mutational patterns that could serve as novel molecular markers in cancer diagnosis and cancer therapy.

CHAPTER TWO: RESULTS

Prologue

Here I present the results of one of my PhD projects that was published in Nature Communications, entitled “Validating the concept of mutational signatures with isogenic cell models”. In this publication, we demonstrate for the first time that mutation patterns of cancer genomes, originating from defects in DNA repair, can be replicated in human isogenic cell lines. The manuscript starts with a brief introduction to the field. After that, we present mutation patterns of human isogenic cell lines, defective in the following DNA repair genes (by CRISPR-Cas9): CHK2, EXO1, FANCC, MSH6, NEIL1, POLB, POLE, POLM. Those mutation patterns include base substitutions, indels and rearrangements, as well as replication timing associated mutation patterns. We also show validation of the deficient cell lines and their proliferation rate, which is used to determine mutation rates. Finally, we provide a concise discussion about the interpretation of our data.

I carried out the experimental work with the help of Marc Wiedner and Jana Stranska. The data was analyzed by our colleagues in Cambridge, UK, who have been pioneers and are experts of the mutation signature field. The manuscript was written by Serena Nik-Zainal, Xueqing Zou, Joanna Loizou and myself with input from all authors.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04052-8 OPEN

Validating the concept of mutational signatures with isogenic cell models

Xueqing Zou[1], Michel Owusu[2], Rebecca Harris[1], Stephen P. Jackson[3], Joanna I. Loizou[2] & Serena Nik-Zainal[1,4]

The diversity of somatic mutations in human cancers can be decomposed into individual mutational signatures, patterns of mutagenesis that arise because of DNA damage and DNA repair processes that have occurred in cells as they evolved towards malignancy. Correlations between mutational signatures and environmental exposures, enzymatic activities and genetic defects have been described, but human cancers are not ideal experimental systems —the exposures to different mutational processes in a patient's lifetime are uncontrolled and any relationships observed can only be described as an association. Here, we demonstrate the proof-of-principle that it is possible to recreate cancer mutational signatures in vitro using CRISPR-Cas9-based gene-editing experiments in an isogenic human-cell system. We provide experimental and algorithmic methods to discover mutational signatures generated under highly experimentally-controlled conditions. Our in vitro findings strikingly recapitulate in vivo observations of cancer data, fundamentally validating the concept of (particularly) endogenously-arising mutational signatures.

[1] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK. [2] CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria. [3] The Gurdon Institute and Department of Biochemistry, University of Cambridge, Cambridge CB2 1QN, UK. [4] Department of Medical Genetics, The Clinical School, University of Cambridge, Cambridge CB2 0QQ, UK. These authors contributed equally: Xueqing Zou, Michel Owusu. Correspondence and requests for materials should be addressed to J.I.L. (email: jloizou@cemm.oeaw.ac.at) or to S.N.-Z. (email: snz@sanger.ac.uk)

The concept of mutational signatures was postulated in 2012: The catalogue of somatic mutations uncovered through tumour sequencing is the outcome of one or more mutational processes that have been operative through the lifetime of a cancer patient[1,2]. Each mutational process, defined by DNA damage and DNA repair components, leaves a characteristic pattern or *mutational signature* on the tumour genome[1–4]. The final mutational portrait of each patient's cancer is determined by the intensity and duration of exposure to each mutational process[4,5].

As an analytical principle, mutational signatures have gained considerable traction, and are regularly featured in cancer genomics literature[6–8]. Already, there are multiple algorithms to extract mutational signatures[5,9–12], though each has its own mathematical idiosyncrasies leading to results that are broadly similar, but never identical. This has caused some to question the robustness of the concept. Nevertheless, as a field, mutational signature research has progressed remarkably. Mutational signatures have been sought across tens of thousands of cancers, revealing over 40 different base substitution signatures (paper in preparation), further supplemented by assessments of how these signatures are distributed across various genomic architectures including replication-timing domains, replication strands, nucleosome occupancy and transcription factor binding sites[13,14]. More recently, genome rearrangement signatures have been unveiled, assisting in the categorization of breast cancer subtypes[13,15,16] and clinical applications based on mutational signatures are currently being developed[17].

No matter how sophisticated the analyses of in vivo mutagenesis of cancers, there are limitations to studying tumours—it is an uncontrolled and noisy system[18–21], and even the best clinical metadata collections will at most, provide associations. Critics of the concept have highlighted that this purely mathematically-based idea, although compelling, lacks definitive validation through in vitro methods.

Historic *TP53* and *HPRT* reporter assays and experiments exposing mouse embryonic fibroblasts (MEFs) to various exogenous agents have already provided convincing evidence that mutation patterns can be generated, particularly for environmental agents such as ultraviolet light and tobacco carcinogens[22,23]. Yet, there have been limited efforts to demonstrate similarly clear relationships for endogenous mutational processes. Few would dispute that substitution Signature 1 composed primarily of C>T transitions at an NpCpG sequence context is linked with deamination of methyl-cytosines, and substitution Signatures 2 and 13 characterised by the distinctive C>T transitions and C > G transversions at a TpCpN trinucleotide context are initiated by the activity of the APOBEC family of enzymes[3,4]. However, many of the mutational signatures that are likely to be endogenous in origin have not been verified. Associations of specific substitution and insertion/deletion (indel) signatures with mismatch repair (MMR) deficiency[24–26], as well as substitution, indel and rearrangement signatures with homologous recombinational (HR) repair deficiency[27–30] though conspicuous, have not been confirmed. Many other genes are also involved in the myriad DNA repair pathways in our cells, and it is not clear whether genetic defects in alternative, related genes could produce mutational signatures as well. Even if mutational signatures could be reproduced using in vitro techniques, it is not known whether these signatures would mimic what is observed in vivo.

Here, we explore whether targeted CRISPR-Cas9-based[31–33] knockouts of selected DNA repair genes can recreate mutational signatures. We describe the experimental cell-based system and develop the computational methodologies to confirm or refute whether each gene knockout generates mutation patterns, thus, providing a general approach for exploring mutational signatures. We further seek whether experimentally-generated mutation patterns bear similar appearances and/or behaviours to mutational signatures seen in primary cancers. If so, this would serve to endorse that mutational signatures are not simply mathematical extractions, but are the consequences of true biological processes.

Results

Generation of DNA repair gene knockouts. We used the immortalised human near-haploid cell line HAP1 to generate isogenic CRISPR-Cas9-mediated knockouts[34]. The advantage of using a haploid cell line is that CRISPR-Cas9-mediated editing is simplified because only one genetic allele needs to be altered to generate a null phenotype. Moreover, because only half the genomic DNA is present, next generation sequencing (NGS) needs are substantially reduced making the experiment more affordable. To determine whether we could detect mutational signatures that result from defects in DNA repair pathways we chose to target genes that play diverse and independent roles in the detection, signalling or repair of DNA damage (Table 1).

Aliquots of the HAP1 cell line were exposed to constructs that express the endonuclease Cas9 and guide RNAs (gRNAs) that were designed to target individual genes of interest. Single clones were selected and those carrying a frame-shift mutation in the given gene were designated as the parental cell line (Fig. 1a), which were amplified and analysed by high-depth whole genome

Table 1 List of DNA repair genes targeted and their functions

Gene symbol	Gene name	Function	Repair pathway	Position
CHK2	Checkpoint kinase 2	Serine threonine kinase	Cell cycle and apoptotic regulation in response to DNA damage	22q12.1
EXO1	Exonuclease 1	5′ to 3′ exonuclease; RNase H activity	Homologous recombination; mismatch repair	1q43
FANCC	Fanconi anemia, Complementation group C	Component of Fanconi repair system core complex	DNA cross-link repair	9q22.32
MSH6	MutS homolog 6	Mismatch recognition	Mismatch repair	2p16.3
NEIL1	Endonuclease VIII-like 1	DNA glycosylase and apurinic/apyrimidinic lyase	Base excision repair	15q24.2
NUDT1	Nudix hydrolase 1	Hydrolyzes oxidized purine nucleoside triphosphates	Modulation of nucleotide pools	7p22.3
POLB	DNA polymerase beta	DNA polymerase (catalytic subunit)	Base excision repair	8p11.21
POLE	DNA polymerase epsilon	DNA polymerase (catalytic subunit)	Nucleotide excision repair and mismatch repair	12q24.33
POLM	DNA polymerase mu	DNA polymerase (catalytic subunit)	Gap filling during non-homologous end-joining	7p13

sequencing (WGS). The parental cell lines (labelled as 'parental clone' in Fig. 1a) were subsequently cultured for one month, from which seven 'subclones' were derived, amplified and analysed by WGS. This workflow served to allow for the identification of mutations that occurred over approximately 36 cellular divisions, considering that the doubling time is approximately 20 h.

Each parental clone and subclone was successfully sequenced to ~15-fold depth. Short read sequences were aligned to the human reference genome assembly GRCh37/hg19 and all classes of somatic mutations were called in the parental clones (subtracting from the primary bulk HAP1 population) and in subclones (subtracting from the parental clones). Targeting of the genes of interest was confirmed by identifying frameshift indels in the relevant gene in short-read data (see Supplementary Fig. 1a and Supplementary Data 1), and loss of protein expression was confirmed through immunoblotting (Supplementary Fig. 1b). Potential off-target edits were also systematically sought in an agnostic manner, whether generating small or large (multi-kb) insertion or deletions, and none were identified. Proliferation rates were also determined for each knockout cell line (Supplementary Fig. 1c). Moreover, potential off-target sites were also searched using COSMID (http://crispr.bme.gatech.edu), a web-based tool to identify and validate CRISPR/CAS9 off-target sites[35] (see Supplementary Data 2 for a ranked-list of potential off-target sites of the relevant guide RNA sequences generated by COSMID). Furthermore, we also confirmed in all subclones, that no additional mutations were acquired in other DNA repair genes during the early clonal expansion phase (see Supplementary Data 3 for a list of DNA repair genes) that could affect the final mutational signature obtained in each subclone.

Knockouts of DNA repair genes instigates mutagenesis. A level of background mutagenesis was observed in parental clones (average ~1200 substitutions, ~60 indels, ~6 rearrangements) and in all subclones (Fig. 1b–d and Supplementary Figs. 2–4). Above the background mutations, subclones associated with particular gene knockouts also had greater numbers of specific classes of mutations, although effect sizes were notably variable. For example, the knockout of *MSH6* was associated with a surge of substitutions and indels. By contrast, the *FANCC* knockout was associated with a possibly small increase in indels but a large increase in rearrangements. Knockout of *EXO1* appeared to cause modest elevations of all classes of mutation (Fig. 1b–d). For each gene knockout, a high level of consistency was observed between all seven subclones in terms of total counts (Fig. 1b–d) and overall patterns of mutations (Supplementary Figs. 2–4). Thus, at first pass, it is possible to crudely discriminate between the effects of gene knockouts through these experiments, suggesting that this is a rational experimental system for exploring the mutational effects conferred by defects in specific genes.

Understanding the signal-to-noise issue. There are however a number of issues to acknowledge and resolve which are universal to all human cell-based systems used for exploring mutagenesis. First, the background mutagenesis was easily detectable: for example, for base substitutions approximately 700–2000 mutations were detected per colony and this comprises a distinctive C>A/G>T substitution pattern with tallest peaks at TCT, GCA, GCT and ACA (in decreasing order; Supplementary Fig. 2). This ubiquitous signature shares considerable similarity with previously reported Signature 18, first observed in primary neuroblastoma[3]. Subsequently, this mutational signature was described in breast and adrenocortical cancers. A very similar signature (cosine similarity of 0.94 to Signature 18) has been associated with mutations in the *MUTYH* gene, hinting that it is a final outcome of a primary mutational process that could involve oxidative damage[8]. Regardless, this mutational process was effectively noise in our system, and was pervasive in parental clones and subclones in our experiments, supporting the possibility of it being due to DNA damage incurred during the experimental process. Background mutagenesis was also detectable in indels (Supplementary Fig. 3) and rearrangements (Supplementary Fig. 4).

Second, this inescapable and abundant mutational process contributed a very large volume of background mutagenesis, which could complicate the detection of true mutational signatures for each target knockout gene. The mutation signals of various gene knockouts were highly different—some were strong in nature while others may be considerably weaker, and could be obscured by the overwhelming background signature. These two issues of high noise and potentially low signal are generic and arise in other cell-based models including induced pluripotent stem cells (iPSCs)[36], embryonic stem cells (ESCs) (manuscript in preparation) and organoids[36–38]. As described below, we thus developed methods to quantitatively and reliably discern whether mutational signatures are present in cell-based experimental systems in order that they may be applied to similar approaches in the future.

Detecting mutational signatures in experimental systems. The pervasive background signature was present in all parental clones and subclones regardless of gene knockout. By contrast, if a gene knockout produced a mutational signature, then the signature should be observed in all relevant subclones and would not be detectable (or be present at a greatly reduced level) in the parental clone. We do however, expect some variation between subclones and must therefore take this into consideration in the modelling. Our aim therefore was to determine whether there is robust and consistent divergence of subclones from parental clones, both qualitatively (mutation spectrum) and quantitatively (mutation count), indicative that targeting particular genes does indeed produce mutational signatures.

To account for the limited number of samples and mutations per sample, and the potentially limited signal-to-noise ratio, we used a bootstrap resampling method of the 96-channel mutation profile for all parental clones and subclones (Fig. 2 and Online Methods for details). This provided us with distributions of subclones and of parental clones from which reliable estimates of the qualitative differences in mutation spectra could be calculated (Fig. 2a; see Online Methods for details). An additional tier to discriminate whether a gene knockout is associated with mutagenesis came from taking mutation count into consideration: an "expected" mutation density was used to deduce a p value to detect an alteration in mutation burden for subclones of a given gene knockout (Fig. 2b, see Online Methods for details). Once a gene knockout was confirmed to be associated with generating a mutation pattern, the final mutational profile (which is a linear combination of background mutagenesis and the gene knockout) was obtained by subtracting the background mutagenesis from the mutational profile of the subclones (see Online Methods).

This principle of signature discrimination (Fig. 2c) was applied to indel and rearrangement patterns as well, although different classifications were used. For indels, a vector of eight features was used comprising the following categories: 1 bp insertion, >=2 bp insertion, 2 bp microhomology-mediated deletion, >= 3 bp microhomology-mediated deletion, 1 bp repeat-mediated deletion, >=2 bp repeat-mediated deletion, other deletion (where there are no specific junctional features associated) and complex indels. For rearrangements, a vector of ten features was applied

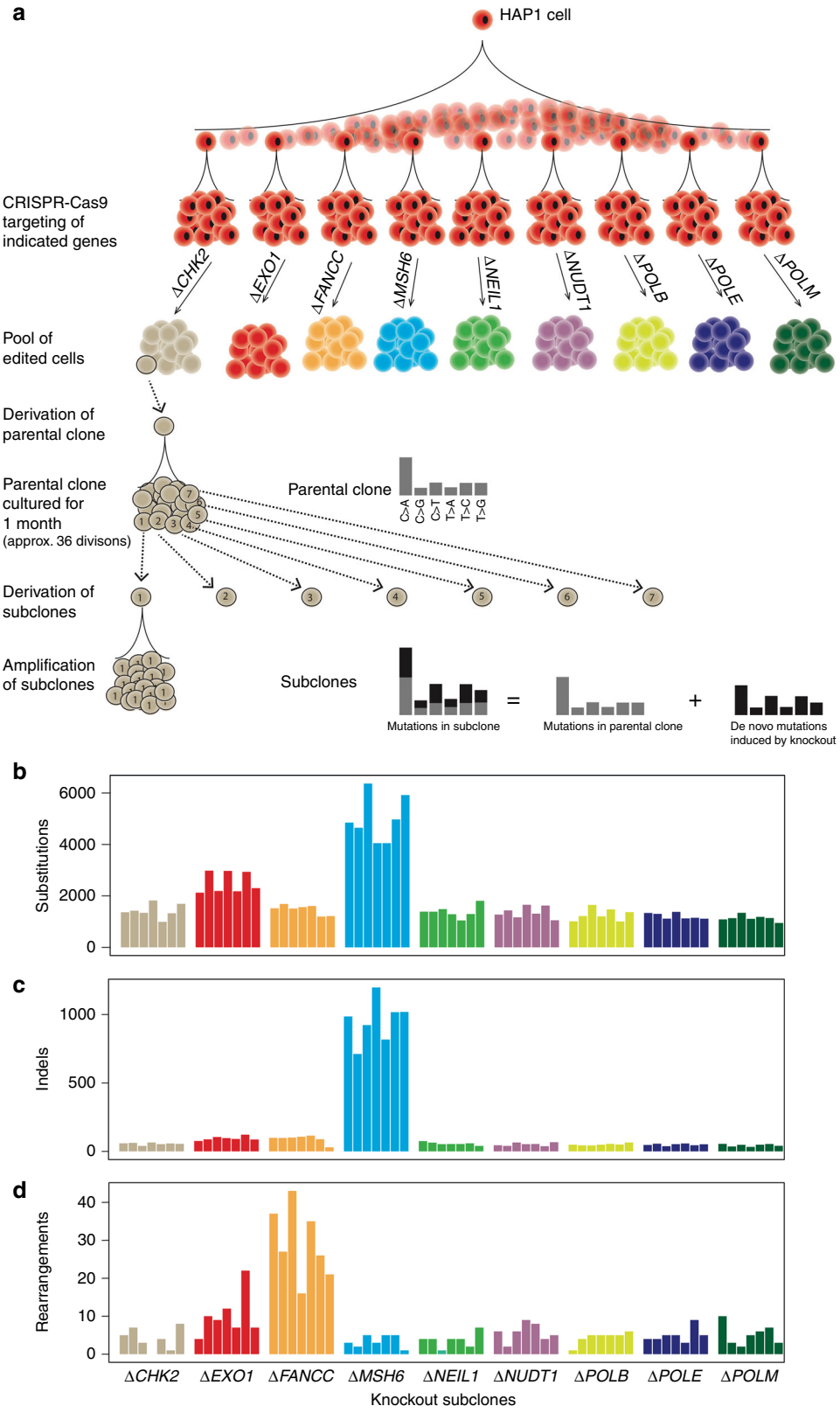
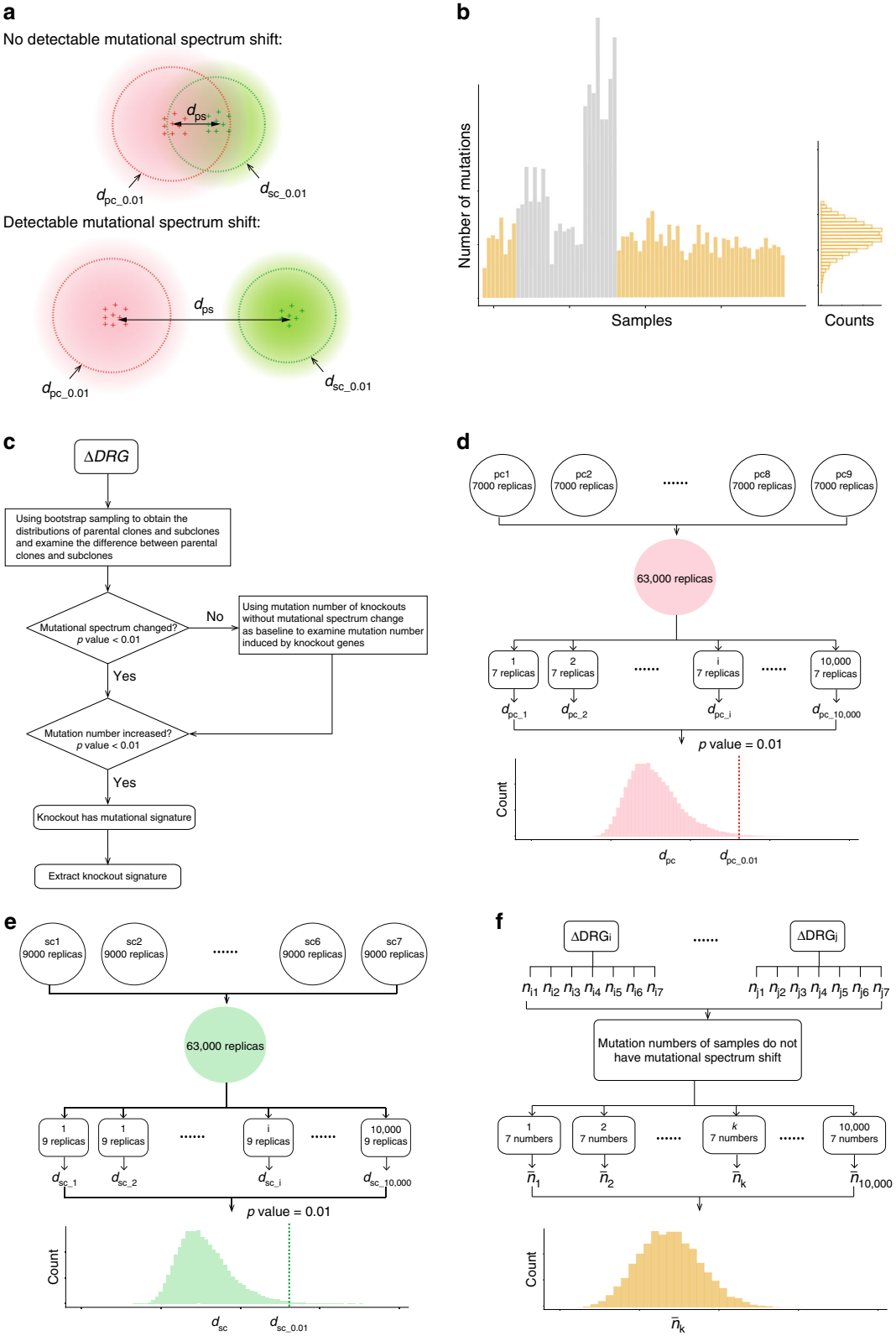


Fig. 1 Knockouts of DNA repair genes instigate mutagenesis. **a** The experimental strategy for investigating whether DNA repair gene knockouts produced mutagenic effects. Parental HAP1 cells are split into multiple aliquots and used for CRISPR-Cas9-mediated gene-editing of the indicated genes. Resulting clones carrying frame-shift mutations are identified by Sanger sequencing and immunoblotting, amplified, cultured for one month (approximately 36 divisions), and seven subclones are derived through a single-cell bottleneck. DNA is extracted and whole genome sequenced for the seven subclones, and the parental clone. De novo mutations in a subclone that is subject to a particular knockout can be obtained by removing mutations in the parental clone from mutations in the subclone. De novo mutations are identified for all classes of mutation: **b** substitutions, **c** indels and **d** rearrangements, of the seven subclones for each knockout gene



comprising: 1–10 kb, 10 kb–1 Mb and >1 Mb size groups of the three classes of deletions, inversions and tandem-duplications, and the last category was translocations.

By using these methods, we conclusively identified seven mutational signatures from nine gene knockouts in this HAP1-based experimental system: two substitution signatures were induced by knockouts of *EXO1* and *MSH6* (Fig. 3); three indel signatures produced by knockouts of *EXO1*, *FANCC* and *MSH6* (Fig. 4); and two rearrangement signatures associated with knockouts of *EXO1* and *FANCC* (Fig. 5), as described in detail below.

Experimentally-generated gene knockout mutational signatures. MSH6 is a protein involved in DNA MMR. MSH6 forms a heterodimer with MSH2 and helps to maintain a low error rate during replication[39]. Inherited mutations in this gene are associated with elevated risks particularly of colorectal and endometrial cancer[40,41]. Inherited and somatic mutations with loss of the wild-type allele are associated with elevated mutation rates in primary human cancers, particularly at polynucleotide repeat tracts conferring a diagnostic phenotype called microsatellite instability (MSI)[26]. In-keeping with previous observations, the *MSH6* knockout was associated with considerably elevated substitution density (~4 fold) over background and had a characteristic pattern dominated by C>T and T>C mutations (Fig. 3a, d). This mutational signature bears a resemblance to the multiple substitution signatures (extracted from many different tumour-types) that have been associated with MMR deficiency in cancers (Signatures 6, 12, 14, 15, 20 and 26), but was not perfectly identical to any one of them. Interestingly, when mutational signatures are extracted from breast cancers alone and all analyses restricted to just this tissue-type, we find that the in vitro signature is strikingly similar to the MMR deficiency signature in breast cancers. This is also the case for tumour-specific signature extractions of 52 colorectal and 44 endometrial cancers, both being cancer-types that are associated with *MSH6* mutations. Furthermore, the *MSH6* knockout had a very high level of 1 bp deletions occurring at polynucleotide repeat tracts, with ~7 fold more deletions than insertions overall, in-keeping with MSI (Fig. 4a, d). Intriguingly, an *MSH6* knockout in an alternative iPSC model generated an identical signature (cosine similarity is 0.94) suggesting that in different cell lines, the signature associated with *MSH6* knockout is very stable (unpublished data).

EXO1 encodes an enzyme that functions as a 5′-3′ DNA exonuclease as well as an endonuclease cleaving RNA on DNA/RNA hybrids (RNase H activity)[42–44]. It plays a role in, and interacts with, components of both the DNA double-strand break repair (DSBR) and MMR[45] pathways. The *EXO1* knockout resulted in a substitution signature with predominantly C>A/G>T transversions with peaks at GCT, GCC and TCT (Fig. 3d) and smaller contributions from C>G/G>C, C>T/G>A and T>C/A>G. The *EXO1* knockout also had an indel pattern that featured a high percentage of 1 bp repeat-mediated deletions and a smaller proportion of long (>=3 bp) microhomology-mediated deletions (mm-del) (Fig. 4d). This is an example where the indel knockout signature and background signature are qualitatively similar (cosine similarity is 0.97, Fig. 4b) but quantitatively distinct (Fig. 4c). Additionally, the *EXO1* knockout produced a rearrangement signature characterised predominantly by a high percentage (60%) of medium-to-large (10 Kb–1 Mb) tandem duplications (Fig. 5d). Knockout of *EXO1* thus created multiple signatures of all mutation classes, probably as a consequence of *EXO1* operating at the junction of several DNA repair pathways.

FANCC is a component of the Fanconi anemia (FA) DNA repair system that functions in the processing of DNA crosslinks that are encountered in S phase via a mechanism that ultimately employs homologous recombination (HR)[28,46,47]. In-keeping with this role, the *FANCC* knockout created a number of mutational signatures that are predicted to be initiated by a DNA double-strand break. These included a characteristic indel pattern of long deletions (⩾3 bp in length) with microhomology observed at the indel junction (Fig. 4d). Furthermore, the *FANCC* knockout produced a rearrangement pattern characterised by chromosomal deletions of between 1–10 Kb in size, inversions in all size ranges, as well as short (=<10 Kb) and long (>1 Mb) tandem duplications (Fig. 5d). This combination of indel and rearrangement patterns showed a high degree of similarity to those seen in primary tumours with defects of other well-known HR components such as *BRCA1* and *BRCA2*[15,17].

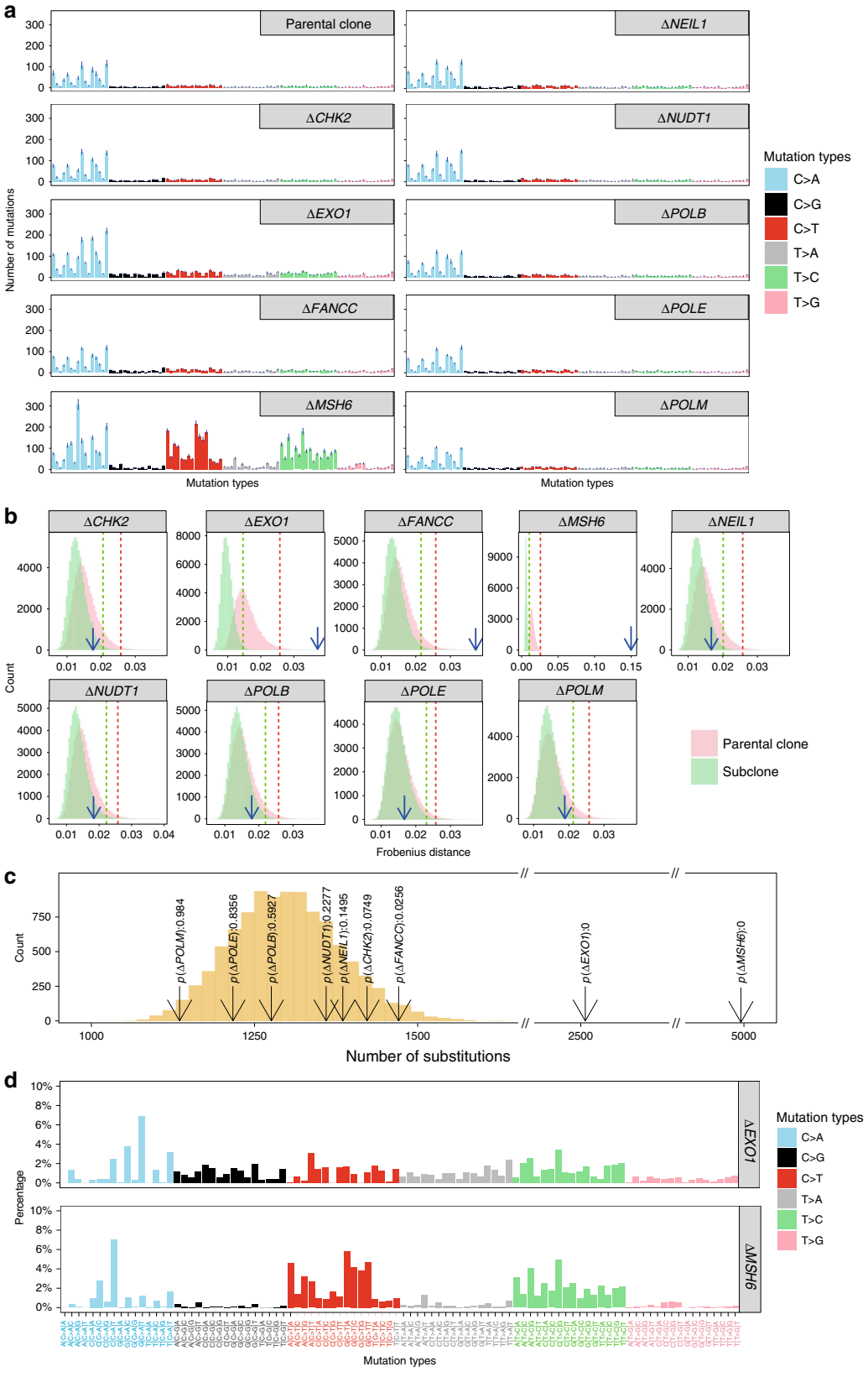
To understand whether the targeting of these DNA repair genes could affect proliferation, we measured the proliferation rates of the given cell lines over a period of ten days (Supplementary Fig. 1c). The *MSH6*, *EXO1* and *FANCC* knockouts had the slowest proliferation rate, indicating that loss of these genes is not associated with an increased proliferative rate. Hence, the elevated numbers of mutations in *MSH6, EXO1* and *FANCC* knockouts were not simply due to an increase in the rate of cell division. Based on these assays, the mutation rates of the seven mutational signatures can be calculated: *MSH6* knockout signatures produced ~148 substitutions and ~36 indels per cell division; *EXO1* knockout signatures produced ~16 substitutions, ~0.58 indels and ~0.19 rearrangements per cell division; *FANCC* knockout signatures produced ~0.58 indels and ~0.68 rearrangements per cell division (Supplementary Data 4).

The knockouts of *CHK2*[48–50], *NEIL1*[51], *NUDT1*[52], *POLB*[53], *POLE*[54] and *POLM*[55] did not appear to produce detectable mutational signatures under these experimental conditions. Additionally, apart from the gene-edits themselves, there were no additional recurrent activating mutations or loss-of-function mutations identified in subclones after culture, suggesting that the enrichment of "driver" events was not a feature in these experiments.

Somatic mutations in DNA polymerase epsilon (*POLE*) have been reported to be associated with a characteristic mutational process in Signature 10[56,57]. We found however that knockout of *POLE*, did not appear to be associated with a striking signature in our study. This is not surprising, given that the identified mutational signature is associated with mutations in the proofreading domain of *POLE* (dominant negative effect), which is not mimicked by the knockout.

These results highlight successful, methodically-generated genome-wide mutation patterns of all classes, in a human cell-based system, demonstrating that biological abrogation of some DNA repair genes not only initiates mutagenesis, but

Fig. 2 Schematic illustration of algorithm developed in the present study. **a** Schematic illustration distinguishing the mutational spectrum of parental clones and subclones. Each red "+" represents a parent clone and green "+" represents a subclone. Red and green clouds represent bootstrapped samples for parental clones and subclones respectively. d_{ps} is the distance between the centroid of parental clones and that of subclones. Red dashed circle shows the boundary of distance d_{pc} with p value = 0.01 and green dashed circle shows the boundary of distance d_{sc} with p value = 0.01 (see online Methods). The mutational spectra of parental clones and subclones are considered to be different only when $d_{ps} > d_{pc\_0.01}$ and $d_{ps} > d_{sc\_0.01}$. **b** Distribution of background mutation number in subclones. Left: The number of mutations in each sample. Cyber yellow and grey highlight the samples that do not have or do have mutational spectrum shifts from parental clones, respectively. Right: Mutation numbers of the samples that do not have mutational spectrum shifts (cyber yellow samples) are used to construct a distribution indicating expected numbers of mutations in cells where the gene knockout does not have an effect. **c** Workflow of characterisation on knockout signatures. **d**, **e** Detailed workflow of quantitative estimation of the difference between the mutation spectrum of parental clones and that of subclones by bootstrapping parental clones (**d**) and subclones (**e**) (see Online Methods). **f** Detailed work flow of the construction of the distribution of mutation numbers generated in cells where the gene knockout does not have an effect, using bootstrap sampling methods (see Online Methods)



creates distinctive mutation patterns, or mutational signatures, conclusively validating the abstract concept of mutational signatures in human cancers. Furthermore, single gene targeting in vitro in some cases generated not just one but multiple mutational signatures, buttressing previous reports that multiple in vivo cancer-derived signatures could arise from single gene defects such as in *BRCA1*/*BRCA2*[17]. This is likely to be due to the multitude of compensatory DSB repair pathways that are brought into play in the absence of conservative, error-free HR and due to some activity of translesion synthesis. Whatever are the

mechanisms that underpin these observations, this is important authentication—because multiple mutational signatures are now starting to be exploited as a principle for designing clinical biomarker assays[17]. This notion of using multiple signals as a biomarker would predict more sensitive and more specific tumour stratification—critical for clinical trials that are currently still largely based on single-channel assays with all their attendant limitations.

Similarities between experimental and cancer signatures. When mutational signatures were first mathematically extracted from cancers, several mutational signatures were found to be associated with inactivation of DNA repair genes. To investigate how in vitro experimentally-generated mutational signatures of gene knockouts compared with in in vivo cancer-derived signatures, we calculated cosine similarities between the in vivo and in vitro mutational signatures for substitutions (Fig. 6a) and rearrangements (Fig. 6b) (cancer-derived indel signatures are not available). Then, we compared overall mutational profiles of knockouts with those of patient cancers.

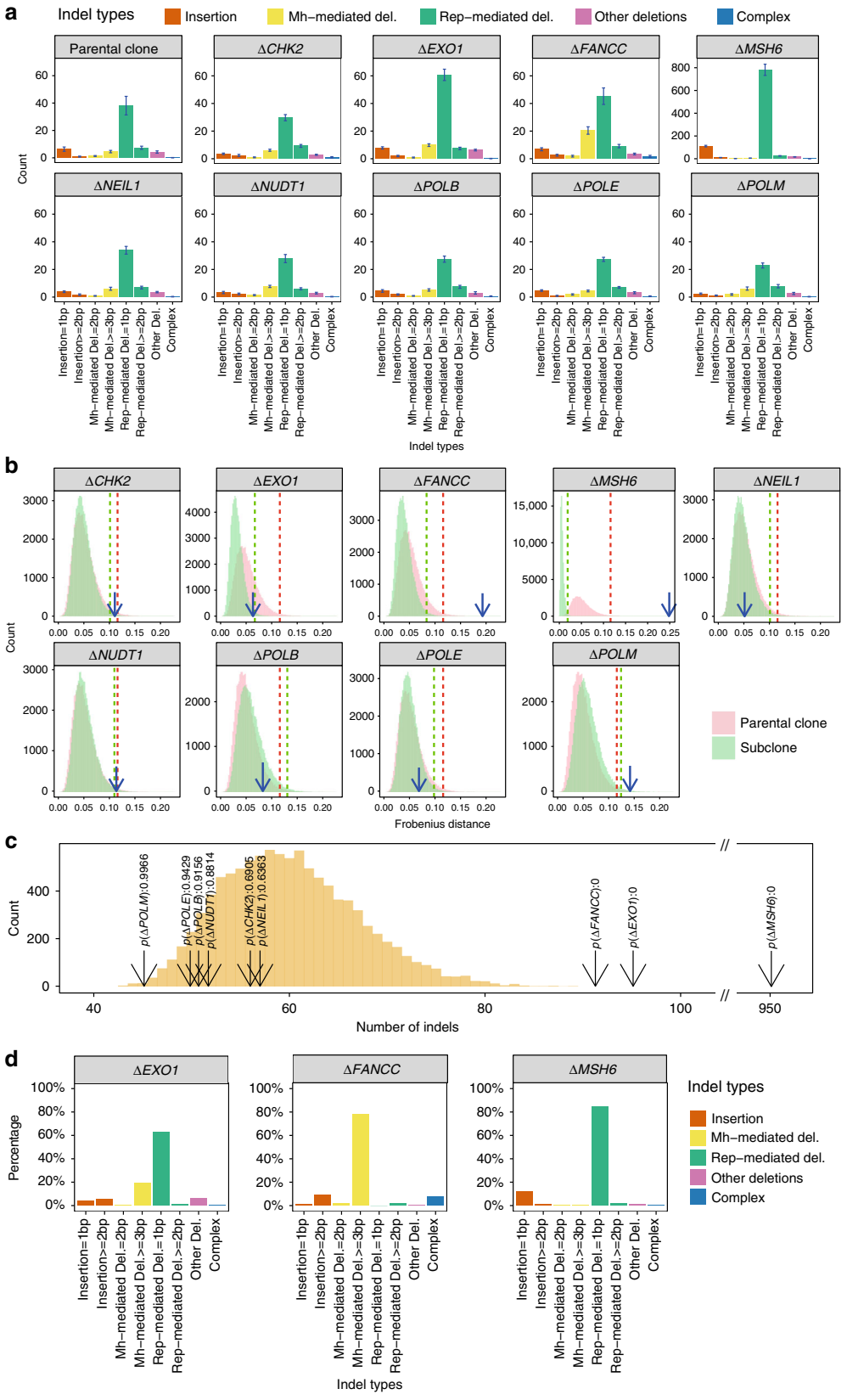
The substitution signatures of *MSH6* and *EXO1* knockouts were compared with cancer-derived 30 COSMIC signatures (http://cancer.sanger.ac.uk/cosmic/signatures). The *MSH6* knockout signature is most similar to COSMIC signature 20 with cosine similarity of 0.91 (Fig. 6a), although there are relatively high cosine similarities when compared to other cancer-derived signatures associated with MMR-deficiency (all $\geqslant$0.6). The *EXO1* knockout substitution signature is most similar to COSMIC Signatures 3 (0.71) and 5 (0.71). Whole genome profiles of experimentally-generated gene knockouts bear uncanny resemblances to whole genome profiles of primarily repair-deficient tumours (Fig. 6c). The *MSH6* knockout, for example, bears striking similarity to those in MMR-deficient tumours—characterised by C>T and T<C substitution signatures and high burden of indels at polynucleotide repeat tracts (Supplementary Fig. 5). By contrast, the *FANCC* and *EXO1* knockouts are more similar to HR-deficient cancers; defined by general genomic instability and an excess of deletions with microhomology at the breakpoint junction (Fig. 6c, Supplementary Figs. 6 and 7). This is an interesting observation because although both of these proteins are not typical HR genes, they do play a role in promoting HR repair of DNA double-strand breaks. These data also provide additional experimental evidence to support how cancers that are deemed to be "HR-deficient", can be sub-classified further genetically.

In a previous analytical exercise exploring structural variation in breast cancer, six classes of rearrangement signatures were identified[15], including two types of tandem duplication signatures —Rearrangement Signature 3 (RS3) comprising short (<10 Kb) tandem duplications and enriched in *BRCA1*-null tumours and Rearrangement Signature 1 (RS1) comprising long (>100 Kb) tandem duplications, not associated with *BRCA1* mutations although a genetic cause has not been identified. The rearrangement signatures of *EXO1* and *FANCC* knockouts were compared with cancer-derived rearrangement signatures (RS1-RS6). The *EXO1*-knockout rearrangement signature is strikingly similar (0.93) to RS1 which is defined by long tandem duplications (Fig. 6b). By contrast, the *FANCC*-knockout rearrangement signature shows little similarity (0.09) to RS1, and instead shows elements of RS3 (0.43) and RS5 (0.59), which have short tandem duplications and deletions. Hence, we show that these rearrangement signatures are not just mathematical abstractions but indeed separate biological entities—that is, the two tandem duplication patterns, namely RS1 and RS3, are able to be recreated by knocking out disparate genes. The *FANCC* knockout rearrangement pattern comprised mainly short tandem duplications and short deletions (<10 Kb) and also had other rearrangement classes but essentially echoed those of *BRCA1*-null cancers (Fig. 6c and Supplementary Fig. 6). This is consistent with the role played by *BRCA1* in HR, downstream of the FA pathway[46,58]. By contrast, the *EXO1* knockout rearrangement signature was dominated by medium-to-long tandem duplications emulating the alternative cohort of genomically unstable (but *BRCA1*-intact) tumours (Fig. 6c and Supplementary Fig. 7).

Genomic architecture of experimentally-generated signatures. Previous analyses of breast-cancer-derived mutational signatures revealed diverse relationships with replicative strand and replicative time domains, as well as transcriptional strands. We thus explored whether experimentally-generated mutational signatures mirrored are thereby validated these mathematically-derived observations.

Of the experimentally-generated mutational signatures, first, we did not find evidence of transcriptional strand bias (Fig. 7a and Supplementary Fig. 8). Second, replication strand asymmetry was not observed for the signatures caused by knockouts of *EXO1*, though it was observed for the C>T/G>A (1.27 fold, p value = 0.021, t test) and T>C/A>G (1.38, p value = 0.018, t test) components of the *MSH6* knockout (Fig. 7b). This interesting bias was consistent with the observation that MMR deficiency associated mutational signatures 6, 20 and 26 have either an excess of damage to G and T on the lagging replicative strand or C and A on the leading replicative strand (Fig. 7c). This implied that MSH6 must have a particular role in directing the repair of damage of these nucleotides during replication. Third, while *EXO1* knockout mutational signatures were consistently increased in regions of the genome associated with late replication, the mutational signature of *MSH6* demonstrated a

Fig. 3 Determination of substitution mutational signatures in gene knockouts. **a** Profile of 96 mutation types (6 types of substitution * 4 types of 5′ base * 4 types of 3′ base) of parental clones and DNA repair gene knockouts. A strong background signature is observed in all samples. The substitution spectrum of each sample is shown in Supplementary Fig. 2. Error bars were referred to as standard error of means ($n = 7$). **b** Discrimination of mutation spectrum of parental clone and subclones. Bootstrap sampling method was used to construct a population of parental clones. The distribution of distance of parental clone replicates to the centroid of parental clones is shown as the pink histogram. The red dashed line indicates a cutoff ($d_{pc\_0.01}$) where 99% replicates are within this distance to the centroid of parental clones. The distribution of subclone replicates is shown as the light green histogram. The green dashed line indicates a cutoff ($d_{sc\_0.01}$) where 99% subclones are within this distance to the centroid of all subclones. The blue arrow indicates the distance (d_{ps}) between centroid of subclones to the centroid of parental clones. A knockout is considered to have an effect on the substitution spectrum, when $d_{ps} > d_{pc\_0.01}$ and $d_{ps} > d_{sc\_0.01}$ are observed, e.g., *EXO1*, *FANCC*, *MSH6*. **c** Identification of mutation number increase in subclones due to gene knockout. From (**b**), one can discriminate the knockouts that do not generate mutational signatures. The number of mutations in these knockout backgrounds can be used as a baseline; through bootstrap sampling method, we obtained the distribution of the number of mutations in subclones in a wildtype background and, therefore calculated the p value of mutation number of each knockout. *EXO1* and *MSH6* show significantly elevated mutation numbers as well as mutational profile change. **d** Substitution signatures of *EXO1* and *MSH6* knockouts. The mutational signatures associated with gene knockouts are obtained by removing the substitution profile of parental clones from the mean of the substitution spectrum of the seven subclones



notably flatter slope, with more mutations early in replication compared to the other knockout signatures (Fig. 7d). This strikingly echoed in vivo observations—a base substitution signature associated with tumour MMR deficiency also exhibits a flattened profile across replication timing domains, unlike most other substitution signatures in breast cancer[13]. Crucially, this result from an experimentally-generated knockout of *MSH6* provided support for a previous hypothesis that MMR activity is essential for reducing mutagenesis in gene-rich, early replicative domains. When abolished, the protective role usually played by

MMR on mutagenesis in these regions, is lost, thus resulting in the excess of mutations in early domains and a flattened replication timing profile[13]. In conclusion, our findings collectively show that mutational signature behaviours across genomic architecture are corroborated by in vitro studies.

Discussion

The gene-edited human cell-based model system used here has permitted validation of the mutational signatures concept across all classes of mutations. This system, however, is not without issues. A challenge posed by the considerable cell culture-related signature resulted in an encumbered signal-to-noise ratio. Here, we combine the experimental set-up with algorithmic developments in order to successfully view mutational patterns generated by knockout of DNA repair genes. These principles lend themselves to a thorough, systematic screen of all genes involved in maintaining genome integrity and of all potential genotoxic agents in order to comprehensively understand the repertoire of mutational signatures in human cells.

We found that in our experimental setup, not all knockouts of genes associated with DNA repair produced detectable mutational signatures. While this could reflect lack of a mutational signature, it is also possible that some gene knockouts produce signals that are too weak to be detected under these experimental conditions. They require intensification through elevating mutation rates. One way this could be achieved is by increasing cumulative time in culture—but the data here already suggest that mutation accumulation rates are variable between genes and a one-size-fits-all approach will therefore always have its limitations. Alternatively, increasing DNA damage experimentally (using acute or chronic regimes) could help to amplify mutagenesis. However, mutational signatures spawned through assisted methods have arisen under subtly different conditions and should be interpreted with this in mind. Using alternative isogenic models that are more permissive for mutagenesis (e.g., MEFs) could also help to increase mutation rates. However, using different cell-based systems with different genetic backgrounds could result in diverse mutational signatures, if similar studies are performed. Lastly, because of the nature of growing cells in culture, it is possible that this is associated with some loss of insights. Copy number changes are often poorly tolerated in cell-based systems and copy number patterns may perhaps be underrepresented using these approaches.

Nevertheless, we present a proof-of-principle, demonstrating how experimentally-generated mutation patterns recapitulate those seen through analysis of primary tumours, thus authenticating the abstract concept of mutational signatures. Our findings also validate previously observed mutational signature relationships with replication, both spatially and temporally. We also note that our findings have also highlighted how a single gene defect is not restricted to creating one mutational signature—it can engender multiple mutational signatures of different classes. The converse is also true: a mutational signature may not necessarily reflect a defect in a single gene, as it could arise through dysregulation of a number of related genes in a pathway. Herein, we have conclusively demonstrated in vitro that endogenous mutational signatures are a direct, mechanistic read-out of pathway dysfunction and could thus be used as biomarkers of pathway dysregulation even in the absence of knowing the precise gene defect or even which gene is compromised.

Methods

Culture conditions. HAP1 cells were grown in Iscove's Modified Dulbecco's Medium (IMDM; GIBCO), containing L-Glutamine and 25 mM HEPES and supplemented with 10% fetal bovine serum (FBS) and 1% Penicillin/Streptomycin (P/S). Cells were grown at 37 °C, with 20% oxygen and 5% carbon dioxide. HAP1 cells were passaged every 3 days and maintained sub-confluent for 1 month. The cell lines were tested negative for mycoplasma contamination using MycoAlert Mycoplasma Detection Kit. HAP1 is not listed in the database of commonly misidentified cell lines by ICLAC. The parental HAP1 cell line has been characterized and authenticated by our collaborators at Horizon Genomics.

Gene editing by CRISPR-Cas9. CRISPR-Cas9 knockouts were generated in collaboration with Horizon Genomics. HAP1 cells were transfected with a Cas9 expressing plasmid, a guide RNA (gRNA) expressing plasmid and a plasmid conferring Blasticidin resistance, using Xfect (Clontech). Guide RNA sequences were 5′-AGGTAAAGCTGGCTTTCGAG-3′ (*CHK2*), 5′-ATCCATCAAATACG AGAAT-3′ (*EXO1*), 5′-GCCAACAGTTGACCAATTGT-3′ (*FANCC*), 5′-CCAAG ATGGAGGGTTACCCC-3′ (*MSH6*), 5′-TGCCCACCTGCGCTTTTACA-3′ (*NEIL1*), 5′-TTCGGGGCCGGCCGGTGGAA-3′ (*NUDT1*), 5′- GAGCAAACGGA AGGCGCCGC-3′ (*POLB*), 5′-AGTTTCGGCACTCAAGCGCC-3′ (*POLE*), and 5′-ACAGGCCTGGCGCGCTCCAA-3′ (*POLM*).

Subsequently, the cells were treated with 20 µg/ml Blasticidin for 24 h in order to eliminate untransfected cells. After 5–7 days of recovery from Blasticidin selection, clonal cell lines were isolated by limiting dilutions.

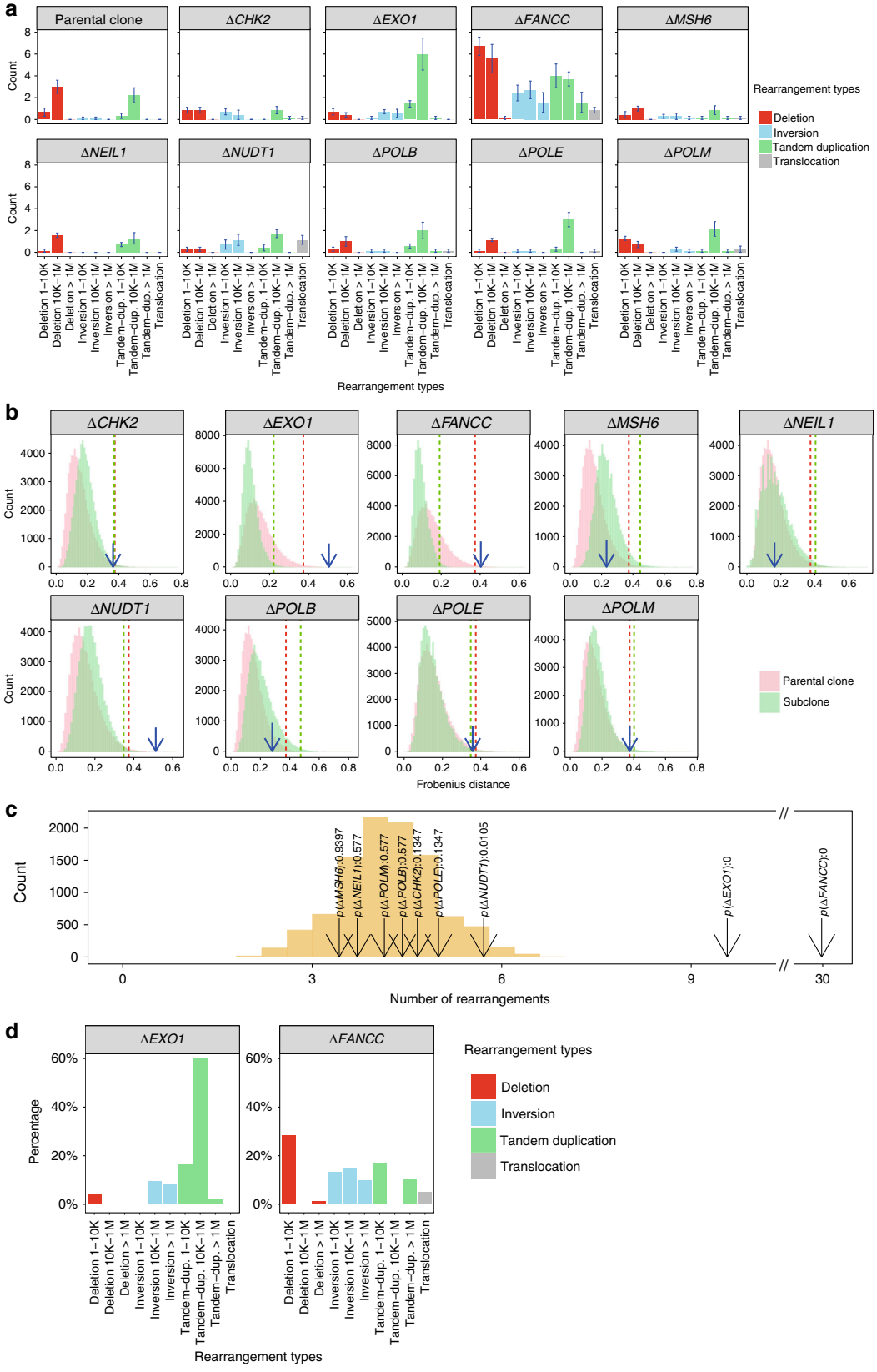
Sanger sequencing. Genomic DNA was extracted using Viagen Bitoech DirectPRC Lysis Reagent (Cell) adhering to the protocol provided by the manufacturer. The genomic region targeted by the gRNA was amplified using the primers and PCR amplification conditions provided below. Frameshift mutations were identified using Nucleotide BLAST against the reference genome GCF_000001405.33. Clones with frameshift mutations were selected as parental cell lines.

Forward primers (For) were 5′-TCAAAGATGCCCCAAAATTTTCCAT-3′ (*CHK2*-For), 5′-CTCGTAAGTATCCAAGGCAGGATTT-3′ (*EXO1*-For), 5′-CA AACCTACACACACATACATGGAC-3′ (*FANCC*-For), 5′-TGGCAGTAGTGAC TCTTACCTGTAT-3′ (*MSH6*-For), 5′-TGGCCAGCCAGTTTGTGAAT-3′ (*NEIL1*-For), 5′-GCTGGGGAGTTACAGCATACC-3′ (*NUDT1*-For), 5′-ACTTG TGAATAATTTTGTGTGGGTCA-3′ (*POLB*-For), 5′-CACTCTTTAGATAA GGACCACGCTA-3′ (*POLE*-For) and 5′-TCGCCCTAATTAATAGCACCCTT TA-3′ (*POLM*-For).

Reverse primers (Rev) were 5′-CTTTGTTTTTCCCTCTAGTGGTGC -3′ (*CHK2*-Rev), 5′-ATCATAGGGTACTAAGGTGCTGAAC-3′ (*EXO1*-Rev), 5′-ACTAAACAAGAAGCATTCACGTTCC-3′ (*FANCC*-Rev), 5′-AATGCCA GAAGACTTGGAATTGTTT-3′ (*MSH6*-Rev), 5′-TGGTACTCCTGCAAGA CACA-3′ (*NEIL1*-Rev), 5′-GAAACCAAGGGTGTGGCCCTA-3′ (*NUDT1*-Rev), 5′-CAGATCATAAGCTATGGAAGGGTGA-3′ (*POLB*-Rev), 5′-AGAGCAAGA CTCCGTCTCAAAAA-3′ (*POLE*-Rev) and 5′-CGGAGTTTCCCTCTGCGTT-3′ (*POLM*-Rev).

PCR amplification: heat lid to 110 °C; start reaction with 94 °C for 2 min; loop 35 × (94 °C for 30 s; 55 °C for 30 s; 68 °C for 1 min), then finish with 68 °C for 7 min.

Fig. 4 Determination of indel signatures in gene knockouts. **a** Indel spectra of parental clones and DNA repair gene knockouts are represented by a 8-channel indel profile which takes the type, length of indel motif and the characteristics at the indel junction into account: 1 bp insertion, >=2 bp insertion, 2 bp microhomology-mediated deletion, >=3 bp microhomology-mediated deletion, 1 bp repeat-mediated deletion, >=2 bp repeat-mediated deletion, other deletions and complex indels. Error bars were referred to as standard error of means ($n = 7$). The indel spectrum of each sample is shown in Supplementary Fig. 3. **b** Distribution of bootstrapped indel spectra of parental clone (pink) and subclones (green). *FANCC*, *MSH6* and *POLM* show significant changes in indel spectrums. **c** Comparison of indels numbers among subclones. The cyber yellow distribution is generated by bootstrapping the indel number of knockout subclones without significant changes in indel profiles. *EXO1*, *FANCC* and *MSH6* show significant increases in indel numbers, indicating the effect of gene knockout on indels. In contrast, although *POLM* shows a detectable indel spectrum shift, it did not show a clear increase in indel number (p value = 0.9966). Hence, *POLM* cannot be determined to have an indel signature. **d** Indel signature of *EXO1*, *FANCC* and *MSH6*. Indel signature of *EXO1* is similar to the culture indels signature. Indel signature of *FANCC* is dominated by microhomology-mediated deletions of 3 bp or more. Indel signature of *MSH6* is dominated by 1 bp deletions at poly-nucleotide repeat tracts



Proliferation assay. Knockout cell lines were plated in triplicates at a density of 0.32×10^6 cells ml^{-1} and allowed to proliferate. Every second day, cells were dissociated with Trypsin-EDTA (Gibco), living cells were counted using CASY Cell Counter and Analyzer system (Innovatis), and replated at 1:2, 1:3 or 1:4 dilutions, depending on the growth rate of the cell line. The experiment was carried out for 10 days. Proliferation was plotted for each time point considering the dilution rates. The average growth rate is a mean over 10 days.

Protein extracts and immunoblotting. Cell extracts were prepared using RIPA lysis buffer (NEB) with protease (Sigma) and phosphatase (Sigma) inhibitors. Immunoblots were performed using standard procedures. Samples containing proteins were separated using SDS PAGE 4–12% gradient gels (Invitrogen) and transferred onto nitrocellulose membranes. The membranes were incubated with primary and secondary antibodies. The primary antibodies were NUDT1 (NB100-109, Novus Biologicals), CHK2 (05–649, Millipore), POLM (C1, Santa Cruz),

EXO1 (A302-639A, Bethyl Laboratories), FANCC (MABC524- clone 8F3, Millipore), POLE (GTX132100, GeneTex), Actin (A5060, SIGMA), NEIL1 (12145-1-AP, Proteintech), POLB (ab26343, Abcam), and MSH6 (D60G2, Cell Signalling). Catalogue numbers and working dilutions for antibodies are provided in Supplementary Table 1. Uncropped immunoblot images are shown in Supplementary Fig. 9.

DNA library preparation and sequencing. Five hundred nanogram of genomic DNA was fragmented (average size distribution ~500 bp, LE220, Covaris Inc), purified, libraries prepared (Agilent SureSelect XT custom kits, Agilent Technologies), and index tags applied (Sanger 168 tag set). Index tagged samples were amplified (6 cycles of PCR, KAPA HiFi kit, KAPA Biosystems), quantified (dsDNA BR assay, HS assay, *Thermo Fisher Scientific*), normalized (~0.85 ng/μl), then pooled together in an equivolume fashion. Pooled samples were submitted to cluster formation for HiSeq ×10 sequencing (32 lanes, 150 bp PE read length, Illumina Inc). The average sequencing coverage is 15-fold for all samples given that HAP1 is a haploid cell line. The details of sequence coverage for all clones and subclones are provided in Supplementary Data 5.

Alignment and somatic variant-calling. Short reads were aligned to human reference genome GRCh37/hg19. Somatic substitutions, indels and rearrangements in clones and subclones were called by CaVEMan[59] (http://cancerit.github.io/CaVEMan/), Pindel[60,61] (http://cancerit.github.io/cgpPindel) and BRASS[15] (https://github.com/cancerit/BRASS), respectively.

De novo somatic mutations of substitutions, indels and rearrangements in subclones were obtained by removing all mutations seen in parental clones. The summary of de novo somatic mutations for each gene knockout is provided in Supplementary Data 6.

Determination of mutational signatures for gene knockouts. The mutational landscape of a cell over a certain period of time reflects a balance point between DNA damage and repair processes in the cell. Exposure to exogenous mutagenic agents or abrogation of DNA repair activity could affect this balance, thereby inducing changes in the mutational landscape. Based on this principle, if the knockout of a gene effectively generates a mutation pattern, then one could observe two changes: First, a shift in the mutational spectrum of cells between subclones and parental clones (shown schematically in Fig. 2a); Second, a change in numbers of mutations in subclones when compared to background (Fig. 2b).

To conclusively identify an effect of a gene knockout, three steps are required: (1) Detecting a qualitative difference between mutational spectra of knockout subclones and that of parental clones; (2) Detecting a quantitative difference in numbers of mutations. (3) Extracting knockout signature. Figure 2c demonstrates the workflow. A more detailed method is described below.

In step 1, we applied a bootstrap resampling method on parental clones and subclones, and calculated the Frobenius distance between parental clones and subclones to quantify the difference between the mutational spectrum of parental clone (without gene knockout effects) and that of subclones (with gene knockout effects).

First, mutation profiles for parental clones (M_p) and subclones (M_s) for each gene KO were defined as:

$$M_p = \begin{bmatrix} m_p^1 \\ \vdots \\ m_p^K \end{bmatrix} \text{ and } M_s = \begin{bmatrix} m_{s1}^1 & \cdots & m_{s7}^1 \\ \vdots & \ddots & \vdots \\ m_{s1}^K & \cdots & m_{s7}^K \end{bmatrix},$$

where m is the mutation number of each mutation feature in each sample, p and s refer to the parental and subclones of different gene knockouts respectively.

The substitution spectrum is made up of a 96-channel vector ($K = 96$), where for each of the six classes of C>A, C>G, C>T, T>A, T>C and T>G, the flanking 5′ and 3′ sequence context for each of the mutated bases is also taken into account (6 types of substitution * 4 types of 5′ base * 4 types of 3′ base = 96 channels). For indels, the profiles are made up of eight features ($K = 8$), including 1 bp insertion, >= 2 bp insertion, 2 bp microhomology-mediated deletion, >= 3 bp microhomology-mediated deletion, 1 bp repeat-mediated deletion, >= 2 bp repeat-mediated deletion, other type of deletion and complex indels, are used. For rearrangements, ten mutation features ($K = 10$) are employed: 1–10 Kb, 10 Kb–1 Mb, and >1 Mb sized deletions, inversions and tandem-duplications respectively and translocations. The profile of substitutions, indels and rearrangements for all samples are shown in Supplementary Figs. 2–4, respectively.

Second, a bootstrap distribution for parental clones was generated. Bootstrap resampling was applied to each parental clone to generate 7000 replicates where the frequency of each mutation type corresponded to its probability in the clone multiplied by the total counts. In total, for nine parental clones, 63,000 replicates are generated. From 63,000 replicates, seven samples are randomly selected and the normalized distance between the centroid of the seven chosen replicates and the centroid of original parental clones, is calculated as d_{pc}. By repeating this step 10,000 times, we obtain a distribution of d_{pc} (shown in Fig. 2d), and the distance associated with p value = 0.01, $d_{pc\_0.01}$, is identified.

Third, bootstrap distributions for subclones of knockouts were generated. The application of bootstrapping on subclones is similar to that of parental clones, see Fig. 2e. For each knockout, 63,000 replicates of subclones are generated (9000 replicates * 7 subclones). Nine replicates are randomly chosen from 63,000 replicates and are used to calculate the normalized distance between the centroid of replicates and the centroid of original subclones, d_{sc}. The distribution of d_{sc} is therefore obtained by repeating the previous step for 10,000 times and the threshold distance with p value = 0.01, $d_{sc\_0.01}$, can be calculated.

Finally, changes in mutational spectrum between parental clones and subclones were determined. For each of the gene knockouts, the distance between centroid of parental clones and centroid of subclones (d_{ps}) is compared with $d_{pc\_0.01}$ and $d_{sc\_0.01}$. The criterion to determine whether the mutational spectrum associated with a given gene knockout is significantly different to the parental clone is $d_{ps} > d_{pc\_0.01}$ and $d_{ps} > d_{sc\_0.01}$, see Fig. 2a.

Step 2 involves determination of increase of mutation number associated with a gene knockout. Aggregated mutation numbers of gene knockouts that do not have a change in mutation spectrum (results from step 1) are used to construct a distribution of baseline mutation counts (i.e., no effect of gene knockout), as shown in Fig. 2f. According to this distribution, a p value of aggregated mutation number of each gene knockout can be calculated. Gene knockouts with p value < 0.01 are considered to have a significantly elevated mutation count, indicative of mutational signatures associated with abrogation of these genes.

In step 3, we extracted knockout signatures based on quantile analysis. The mutational spectrum of subclones can be seen as a linear combination of the mutational spectrum present in parental clones (background mutagenesis) and the mutational spectrum associated with the specific gene knockout:

$$\overline{M}_s \approx e_p \times \overline{P}_p + e_{ko} \times P_{ko}$$

where $\overline{P}_p = \sum_p M_p / \sum_p \sum_k m_p^k$ and $\overline{M}_s$ is the centroid of seven subclones of each knockout gene. ko refers to different gene knockouts. e_p and e_{ko} are the number of mutations caused by parental clone signature and knockout gene signature respectively.

Hence, once a knockout gene is considered to have a mutational signature, its signature (P_{ko}) can be obtained by removing mutations associated with parental clones from the mutation profile of the subclone:

$$P_{ko} \approx (\overline{M}_s - e_p \times \overline{P}_p)/e_{ko}$$

The detailed steps are as described below:

First, we generated bootstrap distributions of subclones. For each knockout gene, 10,000 replicates of subclones are generated to construct a distribution of mutation number in k^{th} of K features of each of the subclones. According to that distribution, the upper and lower boundaries (99% CI) for each k^{th} feature are identified.

Second, the initial status is assumed that there is no knockout signature, i.e., background exposure, e_p, is the total mutation number of subclones. Thus, the background signature profile, $e_p \times \overline{P}_p$, can be calculated. Each number in k^{th} of K features of background signature profile was compared with the upper and lower boundaries of each k^{th} feature of subclones calculated from step 1. For each step,

Fig. 5 Determination of rearrangement signatures in gene knockouts. **a** The rearrangement spectra of parental clones and DNA repair gene knockouts are represented by a 10-channel profile that takes the type and length of rearrangements into account. The rearrangement spectrum of each sample is shown in Supplementary Fig. 4. Error bars were referred to as standard error of means ($n = 7$). **b** Distribution of bootstrapped rearrangement spectra of parental clone (pink) and subclones (green) of the knockouts. *EXO1*, *FANCC* and *NUDT1* knockouts show significant changes in their rearrangement profiles. **c** Identification of elevated rearrangement numbers in knockouts. *EXO1* and *FANCC* knockouts show high number of rearrangements (p value <= 0.01), while *NUDT1* has a p value of 0.0105, which is at the border of our threshold. To be conservative, *NUDT1* is not determined to have a rearrangement signature. **d** Rearrangement signature of *EXO1* and *FANCC*. The rearrangement signature associated with knockout of *EXO1* is characterised by median tandem duplications (10 kb–1 Mb). The rearrangement signature associated with knockout of *FANCC* is characterised by short deletions (1–10 kb), deletions and tandem duplications of 1–10 kb and 10 kb–1 Mb

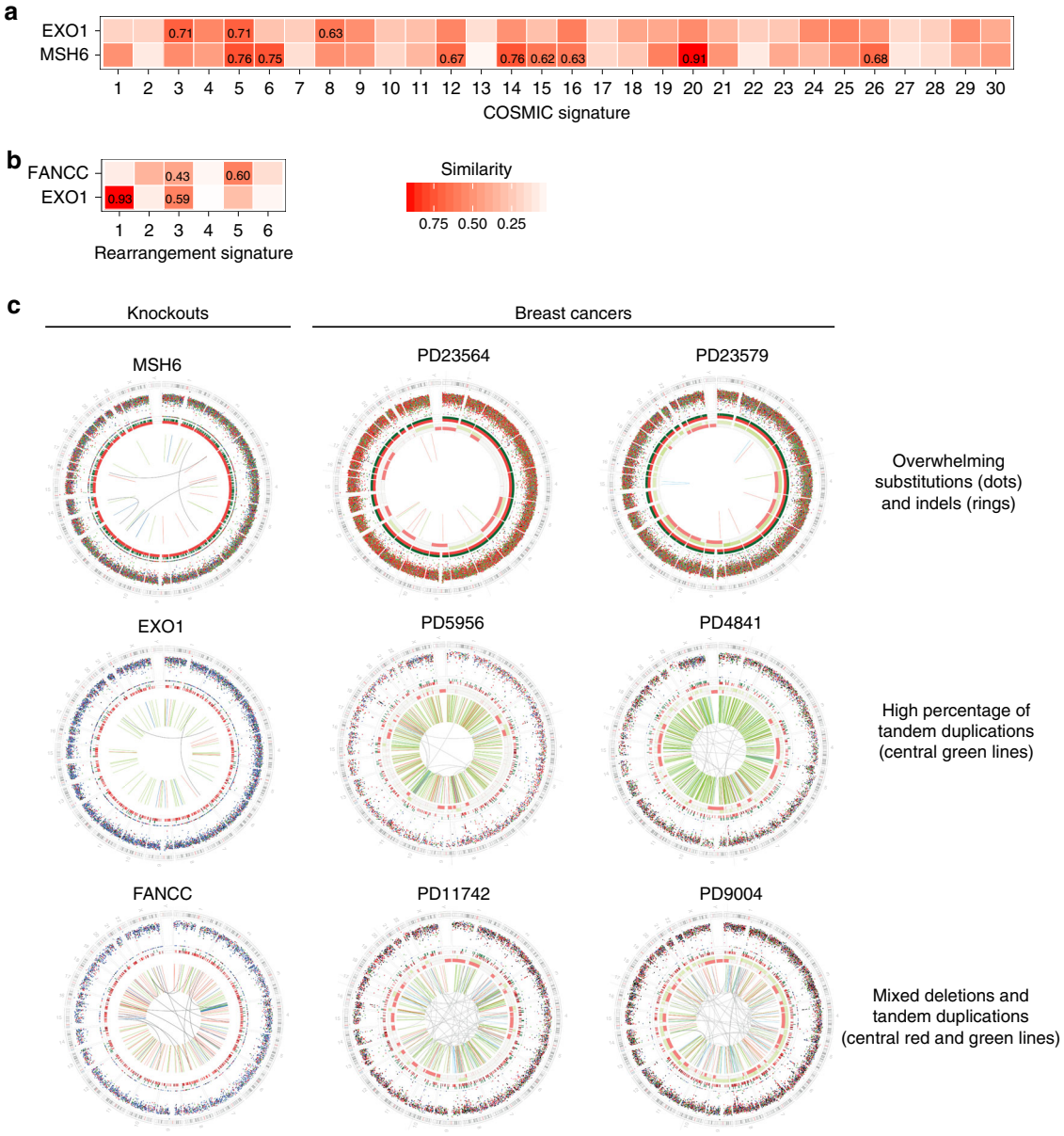


Fig. 6 Comparison of mutational signatures between cancer (in vivo) and knockouts (in vitro). **a** Cosine similarity between 30 COSMIC substitution signatures (http://cancer.sanger.ac.uk/cosmic/signatures) and *EXO1/MSH6* knockout substitution signatures. **b** Cosine similarity between six cancer-derived rearrangement signatures and *EXO1/FANCC* knockout rearrangement signatures. **c** Genome plots of *MSH6*, *EXO1* and *FANCC* knockouts and of cancer samples. Genome plots show somatic mutations including substitutions (outermost, dots represent six mutation types: C>A, blue; C>G, black; C>T, red; T>A, grey; T>C, green; T>G, pink), indels (the second outer circle, colour bars represent five types of indels: complex, grey; insertion, green; deletion other, red; repeat-mediated deletion, light red; microhomology-mediated deletion, dark red) and rearrangements (innermost, lines representing different types of rearrangements: tandem duplications, green; deletions, orange; inversions, blue; translocations, grey). Genome plot of *MSH6/EXO1/FANCC* HAP1 knockouts are aggregations of seven subclones. PD23564 and PD23579 are breast cancers with microsatellite instability which is resulted from impaired mismatch repair. PD5956 and PD4841 are two breast cancers that would historically have been termed as having HR deficiency but are enriched for rearrangement signature 1 and distinct from *BRCA1/BRCA2*-mutated cancers. PD11742 and PD9004 are two breast cancers with *BRCA1/BRCA2*-null HR deficiency

100 bootstrapping background exposure profiles are generated, and if there are at least five parental signature profiles fall within the boundary of subclones, the current background exposure is determined as the final background exposure, and iteration stops. Otherwise, e_p will reduce by 1 in the next step and the newly constructed status will be compared with mutational profiles of subclones.

Third, once the background exposure, e_p, is identified from step 2, the exposure associated with a knockout is thus obtained by subtracting parental exposure from centroid of subclones.

Topography of mutations associated with knockout genes. We explored the relationships between genomic features, e.g., DNA replication and transcription, and mutations associated with knockout genes. Reference information of replicative strands and replication timing regions were obtained from the ENCODE project Repli-seq data (https://www.encodeproject.org/)[62]. Regions of protein coding gene in the genome were used to assign transcriptional strand coordinates. Here, all substitutions are represented in pyrimidine context and the coordinates of transcriptional and replicative strands are given on the +strand of the reference genome, therefore the transcriptional/replicative strand information associated with each substitution is adjusted to the pyrimidine-based mutation, e.g., a G>C mutation on the transcribed strand is described as a C>G mutation on the non-transcribed strand.

Code availability. The code for determination and extraction of knockout signatures associated with this study is available from corresponding author (S.N.-Z.) upon request.

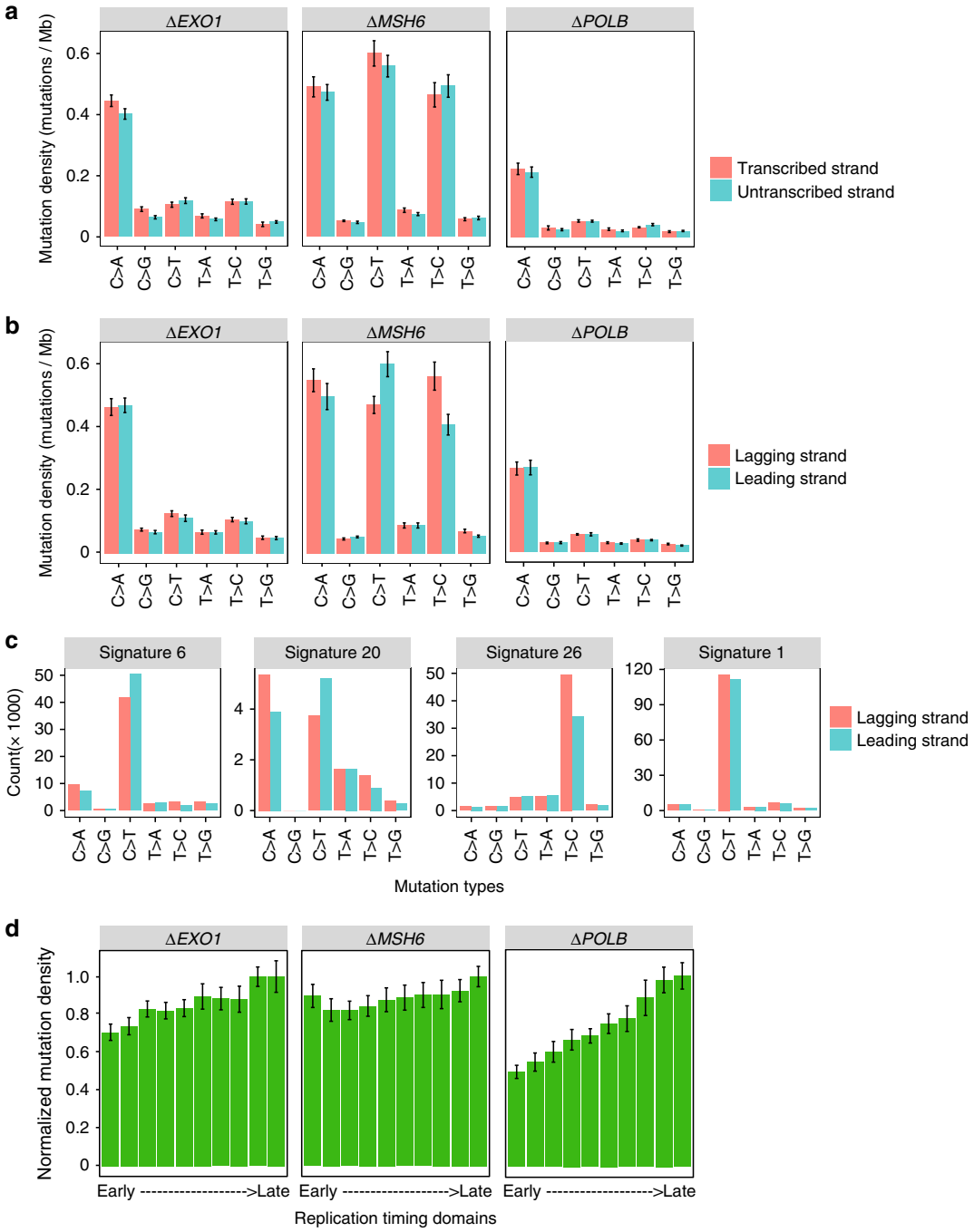


Fig. 7 The topography of experimentally-generated mutations of *EXO1*, *MSH6* and *POLB* knockouts. *POLB* does not show a mutational signature in substitutions. It is shown here as a contrast against *EXO1* and *MSH6* signatures. The topography of mutational signatures associated with the remaining six knockout genes is shown in Supplementary Fig. 8. **a** Histograms exploring transcriptional strand asymmetry. **b** Histograms exploring replication strand asymmetry. **c** Histograms showing replicative strand asymmetry of mutational signatures in breast cancers. Twelve mutational signatures were identified from 560 breast cancers[15]. Here only four signatures are shown: Signatures 6, 20 and 26 are associated with mismatch repair (MMR) deficiency; Signature 1 is associated with hydrolytic deamination of methylated CpG is shown as a contrast. **d** Distribution of normalized mutation density across the replication timing domains. The G2/S phase was separated into ten replication timing domains[13]. Mutation densities in replication timing domains were corrected for genomic size of each domain

Data availability. All mutation data can be obtained from: ftp://ftp.sanger.ac.uk/pub/cancer/Zou_et_al_2017

All other remaining data are available within the Article and Supplementary Files, or available from the authors upon request.

Received: 21 August 2017 Accepted: 29 March 2018
Published online: 01 May 2018

References

1. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
2. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
3. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
4. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).
5. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
6. Alexandrov, L. B. et al. Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618 (2016).
7. Secrier, M. et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat. Genet.* **48**, 1131–1141 (2016).
8. Pilati, C. et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J. Pathol.* **242**, 10–15 (2017).
9. Dees, N. D. et al. MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
10. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).
11. Shiraishi, Y., Tremmel, G., Miyano, S. & Stephens, M. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet.* **11**, e1005657 (2015).
12. Fischer, A., Illingworth, C. J. R., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.* **14**, R39 (2013).
13. Morganella, S. et al. The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 EP (2016).
14. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
15. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
16. Glodzik, D. et al. A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat. Genet.* **49**, 341–348 (2017).
17. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
18. Stephens, P. et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
19. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
20. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
21. Pon, J. R. & Marra, M. A. Driver and passenger mutations in cancer. *Annu. Rev. Pathol.: Mech. Dis.* **10**, 25–50 (2015).
22. Besaratinia, A. & Pfeifer, G. P. Applications of the human p53 knock-in (Hupki) mouse model for human carcinogen testing. *FASEB J.* **24**, 2612–2619 (2010).
23. Liu, Z. et al. Human tumor p53 mutations are selected for in mouse embryonic fibroblasts harboring a humanized p53 gene. *Proc. Natl Acad. Sci. USA* **101**, 2963–2968 (2004).
24. Li, G.-M. DNA mismatch repair and cancer. *Front. Biosci.* **8**, d997–d1017 (2003).
25. Hsieh, P. & Yamane, K. DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mech. Ageing Dev.* **129**, 391–407 (2008).
26. Xiao, X., Melton, D. W. & Gourley, C. Mismatch repair deficiency in ovarian cancer—Molecular characteristics and clinical implications. *Gynecol. Oncol.* **132**, 506–512 (2014).
27. Venkitaraman, A. R. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* **108**, 171–182 (2002).
28. Niedzwiedz, W. et al. The Fanconi anaemia gene FANCC promotes homologous recombination and error-prone DNA repair. *Mol. Cell* **15**, 607–620 (2004).
29. Jasin, M. & Rothstein, R. Repair of strand breaks by homologous recombination. *Cold Spring Harb. Perspect. Biol.* **5**, a012740 (2013).
30. Spies, M. & Fishel, R. Mismatch repair during homologous and homeologous recombination. *Cold Spring Harb. Perspect. Biol.* **7**, a022657 (2015).
31. Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
32. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotech.* **32**, 347–355 (2014).
33. Zhang, F., Wen, Y. & Guo, X. CRISPR/Cas9 for genome editing: progress, implications and challenges. *Hum. Mol. Genet.* **23**, R40–R46 (2014).
34. Carette, J. E. et al. Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. *Nature* **477**, 340–343 (2011).
35. Cradick, T. J., Qiu, P., Lee, C. M., Fine, E. J. & Bao, G. COSMID: a web-based tool for identifying and validating CRISPR/Cas off-target sites. *Mol. Ther. Nucleic Acids* **3**, e214 (2014).
36. Rouhani, F. J. et al. Mutational history of a human cell lineage from somatic to induced pluripotent stem cells. *PLoS Genet.* **12**, e1005932 (2016).
37. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
38. Drost, J. et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234 (2017).
39. Li, G.-M. Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18**, 85–98 (2008).
40. Poulogiannis, G., Frayling, I. M. & Arends, M. J. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. *Histopathology* **56**, 167–179 (2010).
41. Li, Z., Pearlman, A. H. & Hsieh, P. DNA mismatch repair and the DNA damage response. *DNA Repair* **38**, 94–101 (2016).
42. Genschel, J., Bazemore, L. R. & Modrich, P. Human Exonuclease I is required for 5′ and 3′ mismatch repair. *J. Biol. Chem.* **277**, 13302–13311 (2002).
43. Wei, K. et al. Inactivation of Exonuclease 1 in mice results in DNA mismatch repair defects, increased cancer susceptibility, and male and female sterility. *Genes Dev.* **17**, 603–614 (2003).
44. Liberti, S. E. & Rasmussen, L. J. Is hEXO1 a cancer predisposing gene? *Mol. Cancer Res.* **2**, 427 (2004).
45. Branzei, D. & Foiani, M. The DNA damage response during DNA replication. *Curr. Opin. Cell Biol.* **17**, 568–575 (2005).
46. Garcia-Higuera, I. et al. Interaction of the Fanconi anemia proteins and BRCA1 in a common pathway. *Mol. Cell* **7**, 249–262 (2001).
47. Kottemann, M. C. & Smogorzewska, A. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. *Nature* **493**, 356–363 (2013).
48. Hirao, A. et al. Chk2 is a tumor suppressor that regulates apoptosis in both an Ataxia Telangiectasia mutated (ATM)-dependent and an ATM-independent manner. *Mol. Cell Biol.* **22**, 6521–6532 (2002).
49. Bartek, J. & Lukas, J. Chk1 and Chk2 kinases in checkpoint control and cancer. *Cancer Cell* **3**, 421–429 (2003).
50. Zannini, L., Delia, D. & Buscemi, G. CHK2 kinase in the DNA damage response and beyond. *J. Mol. Cell Biol.* **6**, 442–457 (2014).
51. Krishnamurthy, N., Zhao, X., Burrows, C. J. & David, S. S. Superior removal of Hydantoin lesions relative to other oxidized bases by the human DNA glycosylase hNEIL1. *Biochemistry* **47**, 7137–7146 (2008).
52. Gad, H. et al. MTH1 inhibition eradicates cancer by preventing sanitation of the dNTP pool. *Nature* **508**, 215–221 (2014).
53. Ray, S., Menezes, M. R., Senejani, A. & Sweasy, J. B. Cellular roles of DNA polymerase beta. *Yale J. Biol. Med.* **86**, 463–469 (2013).
54. Mozzherin, D. J. & Fisher, P. A. Human DNA polymerase ε: enzymologic mechanism and gap-filling synthesis. *Biochemistry* **35**, 3572–3577 (1996).
55. Martin, M. J. & Blanco, L. Decision-making during NHEJ: a network of interactions in human Polμ implicated in substrate recognition and end-bridging. *Nucleic Acids Res.* **42**, 7923–7934 (2014).
56. The Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
57. The Cancer Genome Atlas Research, N. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
58. Long, D. T. & Walter, J. C. A novel function for BRCA1 in crosslink repair. *Mol. Cell* **46**, 111–112 (2012).
59. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinforma.* **56**, 15.10.1–15.10.18 (2016).
60. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
61. Raine Keiran, M. et al. cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr. Protoc. Bioinforma.* **52**, 15.7.1–15.7.12 (2015).
62. The, E.P.C. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).

Acknowledgements

Funding: Wet-lab consumables were provided by the J.I.L. lab, that is funded by the Austrian Academy of Sciences and was additionally funded by a Marie-Curie Career Integration Grant at the start of the project (Project number: 321602-NonCanATM). Sequencing experiments were funded by a Wellcome Trust Intermediate Clinical Fellowship (WT100183MA).

Personal funding: X.Z. is funded by a Wellcome Trust Strategic Award (WT101126/B/13/Z). M.O. is supported by a project grant from the FWF to J.I.L. (Project number P 29763-B27). S.N.-Z. was funded by a Wellcome Trust Intermediate Clinical Fellowship

(WT100183MA) at the start of the project and is now funded by a CRUK Advanced Clinician Scientist Award (C60100/A23916). S.P.J. laboratory is funded by Cancer Research UK (programme grant C6/A18796) and a Wellcome Trust Investigator Award (206388/Z/17/Z). Institute core infrastructure funding is provided by Cancer Research UK (C6946/A24843) and the Wellcome Trust (WT203144).
Project Coordination assistance was received from Dr Rebecca Harris. Mr Marc Wiedner and Dr Jana Stranka assisted with cell culture.

Author contributions

S.N.-Z., J.I.L. and S.P.J. designed the project. M.O. designed, optimized, coordinated and performed all wet-lab experiments. X.Z. performed all steps of data curation, post-hoc processing and performed downstream bioinformatic analysis. R.H. assisted in coordinating the project. S.N.-Z., X.Z., J.I.L and M.O. wrote the manuscript with comments from all authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04052-8.

Competing interests: S.N.-Z. has patents filed with the UK IPO and is a scientific consultant for Artios Pharma Ltd. J.I.L. has filed a patent with the European Patent Office (EPO). All other authors declare no competing interests.

Supplementary Information

Validating the concept of mutational signatures with isogenic cell models

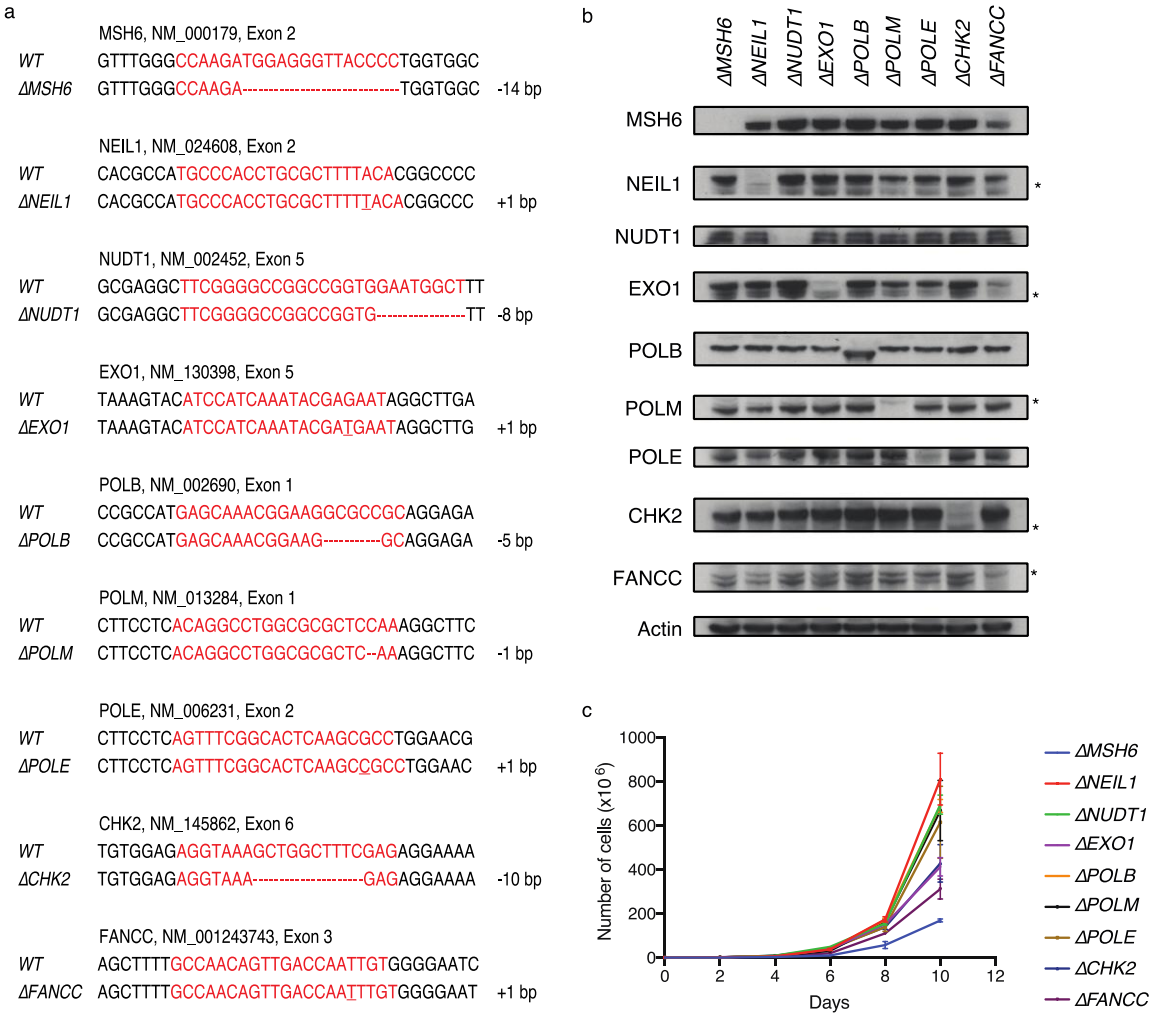
Xueqing Zou, Michel Owusu, Rebecca Harris, Stephen P. Jackson, Joanna I. Loizou, Serena Nik-Zainal

Supplementary Table

Name	Catalogue Number	Company	Dilution
NUDT1	NB100-109	Novus Biologicals	1:1000
NEIL1	12145-1-AP	Proteintech	1:500
CHK2	05-649	Millipore	1:1000
EXO1	A302-639A	Bethyl Laboratories	1:1000
POLE	GTX132100	GeneTex	1:1000
POLB	ab26343	Abcam	1:1000
POLM	C-1	Santa Cruz	1:500
FANCC	MABC524	Millipore	1:500
MSH6	D60G2	Cell Signaling	1:1000
ACTIN	A5060	SIGMA	1:5000

Supplementary Table 1. Information of catalogue numbers and working dilutions for antibodies.

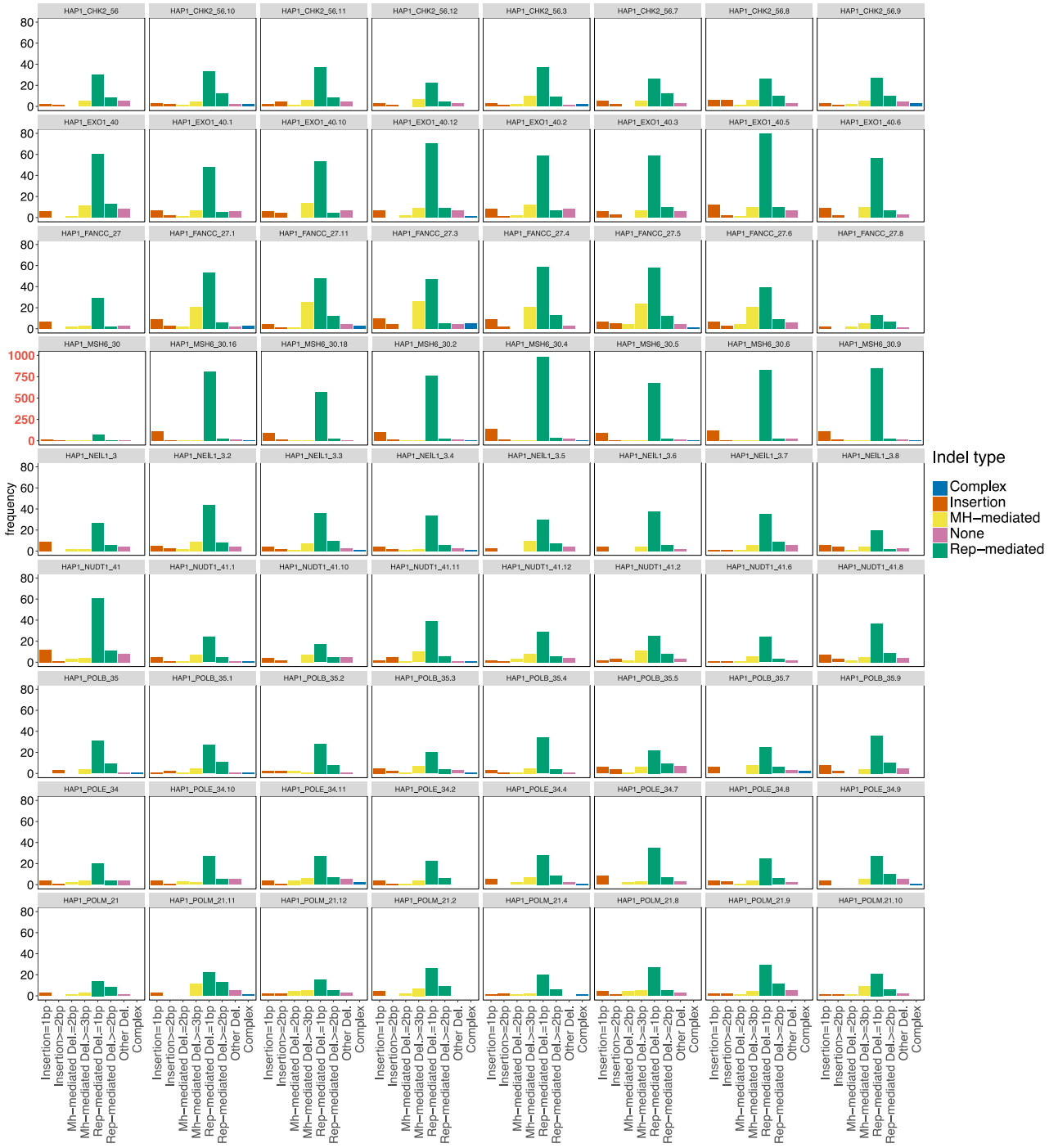
Supplementary Figures



Supplementary Figure 1. Generation and characterisation of CRISPR-Cas9 edited human HAP1 cells. (a) Sanger sequencing confirming CRISPR-Cas9-induced frame-shift mutations in illustrated genes. The red sequence in the wild type (WT) gene corresponds to the guide RNA (gRNA) sequence used. Insertions are marked by an underlined character and deletions by missing sequences (dashes). (b) Immunoblots for expression of MSH6, NEIL1, NUDT1, EXO1, POLB, POLM, POLE, CHK2 and FANCC in CRISPR-Cas9 edited human HAP1 clones. Actin serves as a loading control. "*" denotes a non-specific band. Immunoblot images are shown in Supplementary Figure 9. (c) Proliferation of indicated knockout cell lines over a period of ten days. Living cells were counted every second day using CASY Cell Counter and Analyzer system starting with 0.32 million cells. The plot shows a mean and standard deviation of 3 replicates for each time point. Error bars are defined as standard error of the mean.

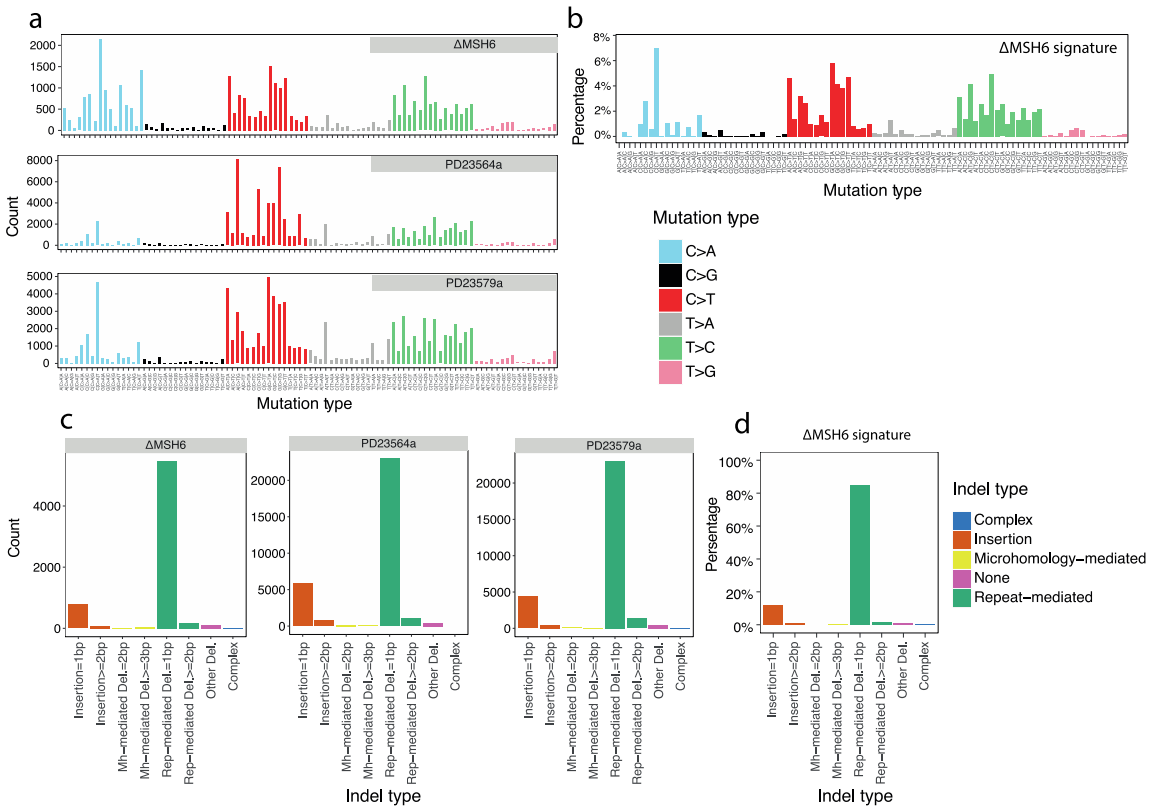

HAP1_CHK2_56
HAP1_CHK2_56−10
HAP1_CHK2_56−11
HAP1_CHK2_56−12
HAP1_CHK2_56−3
HAP1_CHK2_56−7
HAP1_CHK2_56−8
HAP1_CHK2_56−9
HAP1_EXO1_40
HAP1_EXO1_40−1
HAP1_EXO1_40−10
HAP1_EXO1_40−12
HAP1_EXO1_40−2
HAP1_EXO1_40−3
HAP1_EXO1_40−5
HAP1_EXO1_40−6
HAP1_FANCC_27
HAP1_FANCC_27−1
HAP1_FANCC_27−11
HAP1_FANCC_27−3
HAP1_FANCC_27−4
HAP1_FANCC_27−5
HAP1_FANCC_27−6
HAP1_FANCC_27−8
HAP1_MSH6_30
HAP1_MSH6_30−16
HAP1_MSH6_30−18
HAP1_MSH6_30−2
HAP1_MSH6_30−4
HAP1_MSH6_30−5
HAP1_MSH6_30−6
HAP1_MSH6_30−9
HAP1_NEIL1_3
HAP1_NEIL1_3−2
HAP1_NEIL1_3−3
HAP1_NEIL1_3−4
HAP1_NEIL1_3−5
HAP1_NEIL1_3−6
HAP1_NEIL1_3−7
HAP1_NEIL1_3−8
HAP1_NUDT1_41
HAP1_NUDT1_41−1
HAP1_NUDT1_41−10
HAP1_NUDT1_41−11
HAP1_NUDT1_41−12
HAP1_NUDT1_41−2
HAP1_NUDT1_41−6
HAP1_NUDT1_41−8
HAP1_POLB_35
HAP1_POLB_35−1
HAP1_POLB_35−2
HAP1_POLB_35−3
HAP1_POLB_35−4
HAP1_POLB_35−5
HAP1_POLB_35−7
HAP1_POLB_35−9
HAP1_POLE_34
HAP1_POLE_34−10
HAP1_POLE_34−11
HAP1_POLE_34−2
HAP1_POLE_34−4
HAP1_POLE_34−7
HAP1_POLE_34−8
HAP1_POLE_34−9
HAP1_POLM_21
HAP1_POLM_21−11
HAP1_POLM_21−12
HAP1_POLM_21−2
HAP1_POLM_21−4
HAP1_POLM_21−8
HAP1_POLM_21−9
HAP1_POLM−21−10
relative frequency
mutation type
mutation
C>A
C>G
C>T
T>A
T>C
T>G

Supplementary Figure 2. 96-element spectra of substitutions of all parental clones (first column) and subclones. A strong and consistent background signature comprising C>A mutations can be seen for all samples.

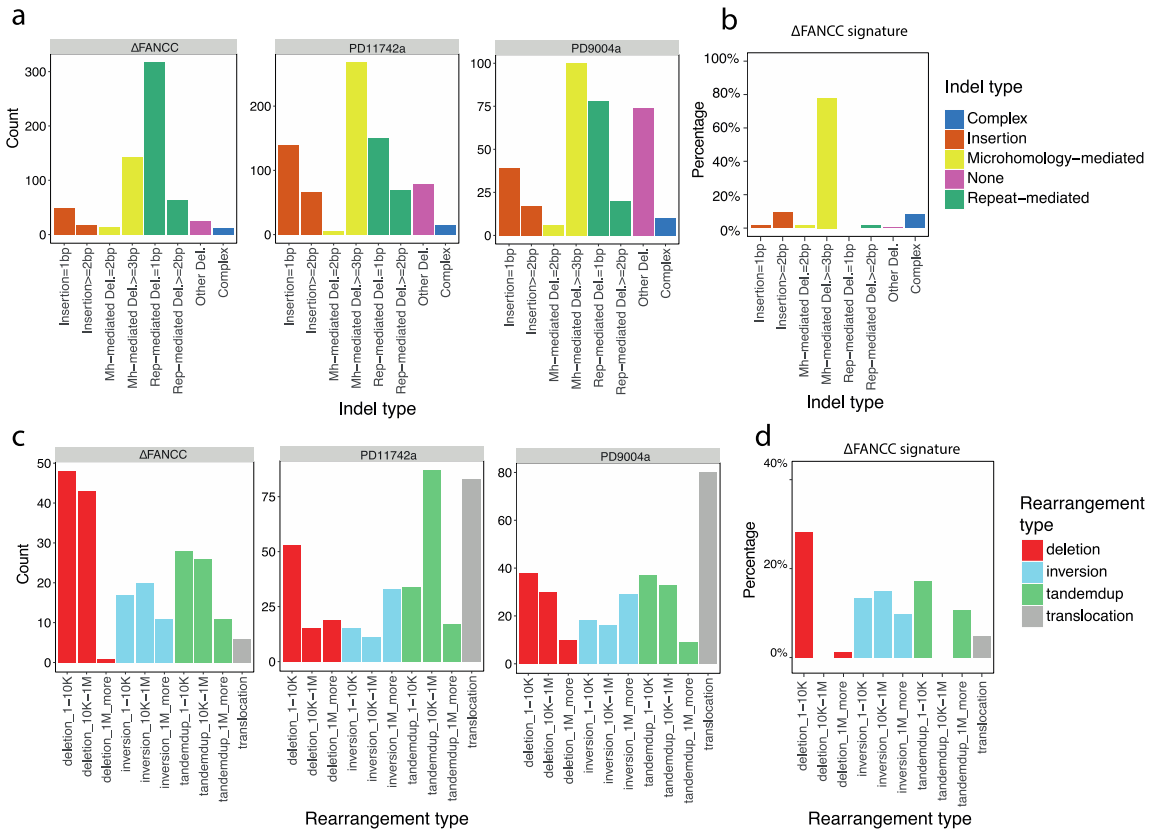


Supplementary Figure 3. 8-element spectra of indels of all parental clones (first column) and subclones. Similar to substitutions, a strong and consistent indel background is observed in all samples.

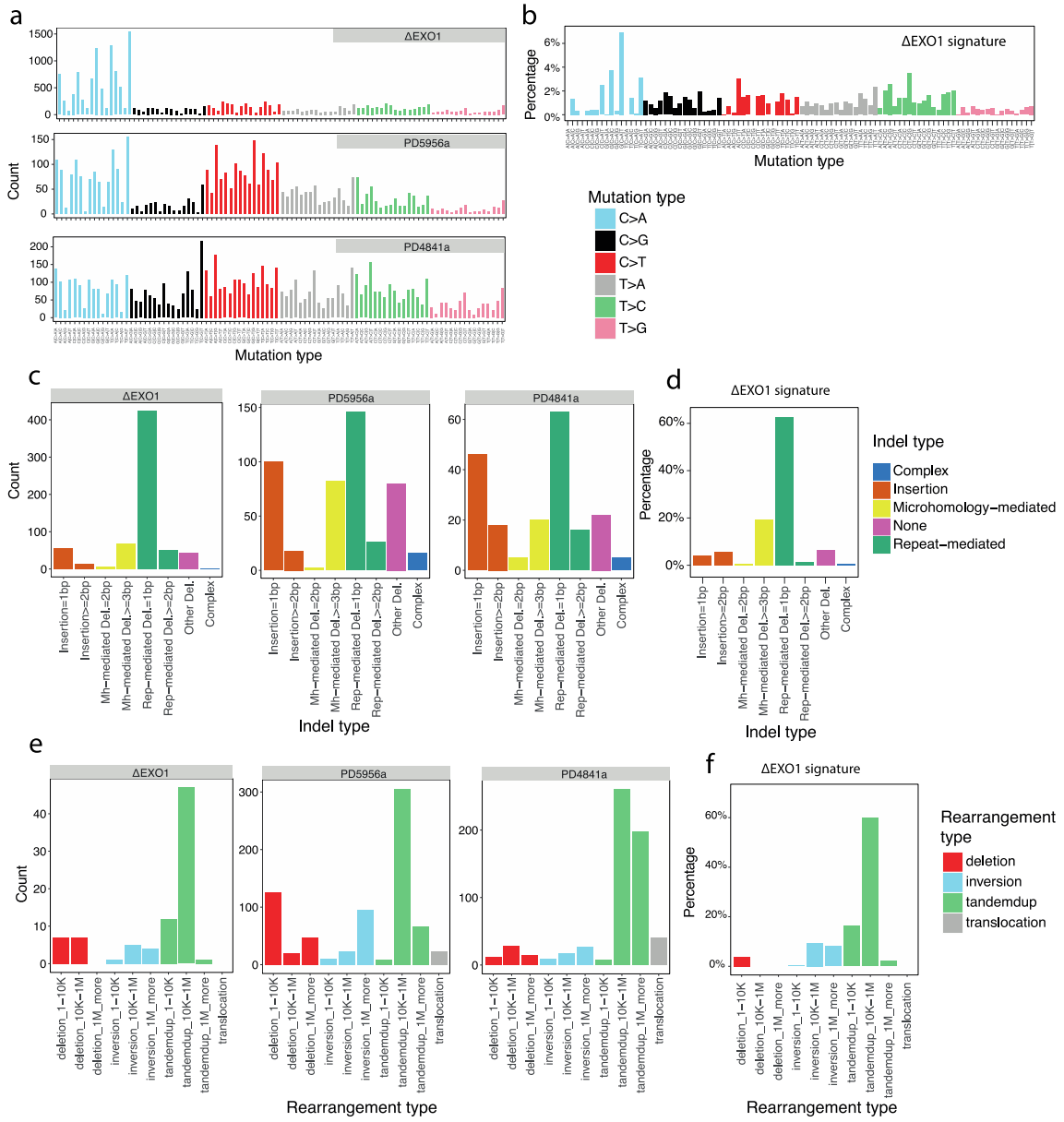
Supplementary Figure 4. 10-element spectra of rearrangements of all parental clones (first column) and subclones. Due to a low number of rearrangements, the profile of rearrangements of each sample is sparser than the profile of indels/substitutions of each sample.



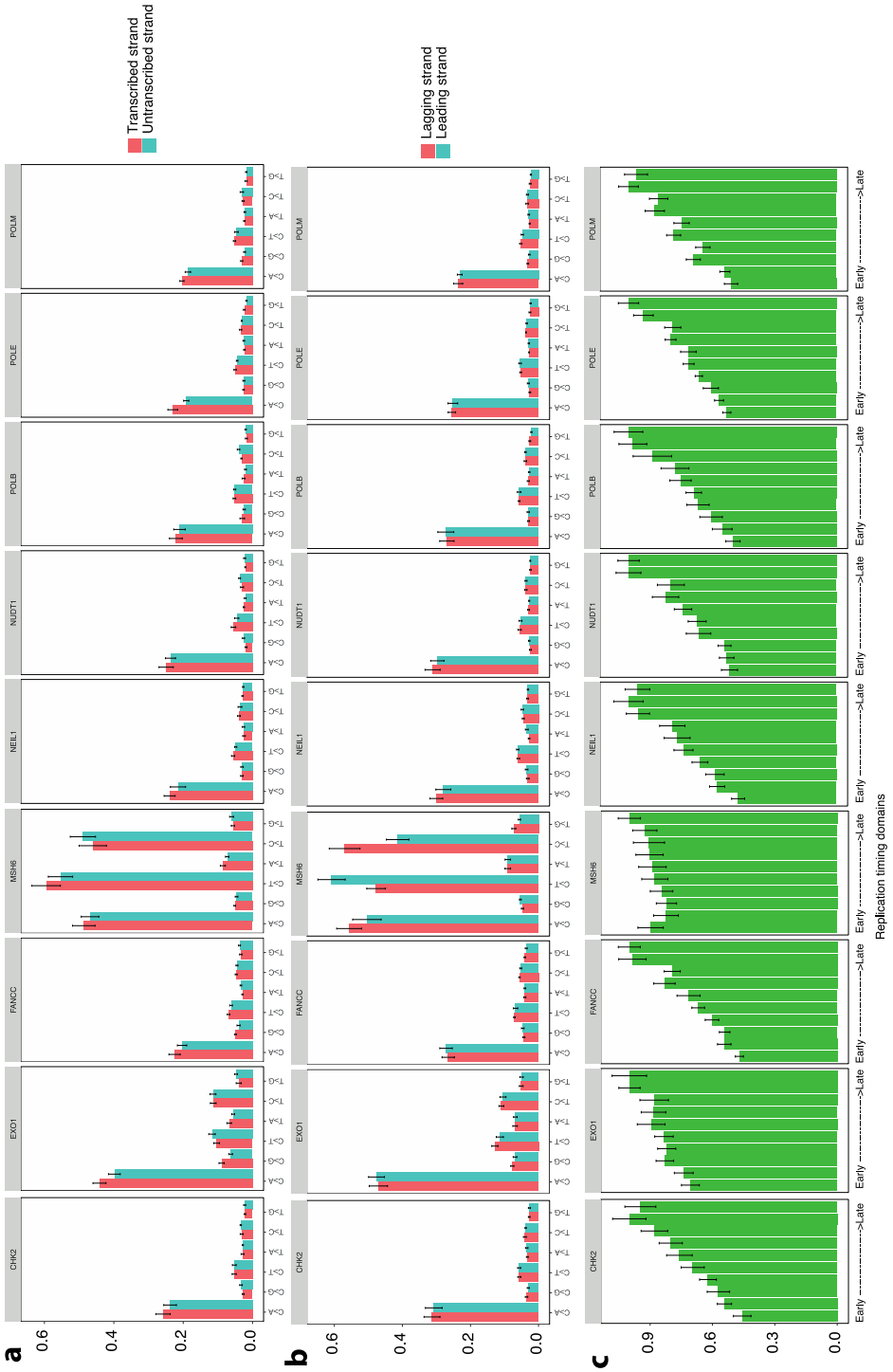
Supplementary Figure 5. Comparison of experimentally generated signatures in mismatch repair (MMR) gene *MSH6* knockouts and mutational signatures in MMR-deficient cancers. (a) 96-element substitution spectra of *MSH6* knockouts (seven subclones were aggregated) and two breast cancers with MMR deficiency, PD23564 and PD23579. (b) Substitution signatures associated with knockout of *MSH6*. (c) 8-element indel spectra of *MSH6* knockouts (seven subclones were aggregated), PD23564 and PD23579. (d) Indel signatures associated with knockout of *MSH6*. Overarching substitution and indel whole genome profiles of MMR-deficient samples are very similar to that of the *MSH6* knockout, particularly the individual substitution and indel profiles.



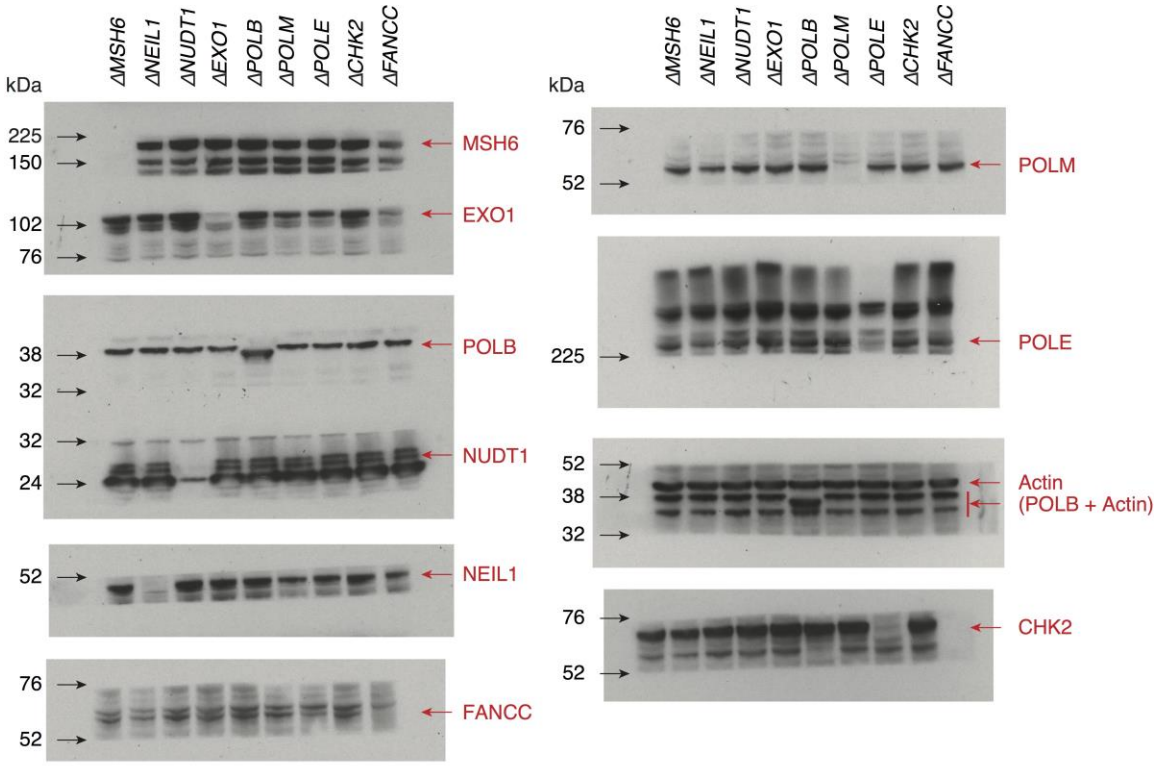
Supplementary Figure 6. Comparison of experimentally generated signatures in DNA cross-link repair via homologous recombination (HR) gene *FANCC* knockouts and cancer-derived mutational signatures in HR-deficient *BRCA1*-null samples. (a) 8-element indel spectra of *FANCC* knockouts (seven subclones were aggregated) and two *BRCA1*-null breast cancers associated with HR deficiency, PD11742 and PD9004. (b) Indel signatures associated with *FANCC* knockouts. (c) 10-element rearrangement spectra of *FANCC* knockouts (seven subclones were aggregated), PD11742 and PD9004. (d) Rearrangement signature associated with *FANCC* knockouts. Indel and rearrangement profiles of *BRCA1*-null samples are very similar to the *FANCC* knockout: microhomology-mediated deletions and a high number of rearrangements (1-10Kb tandem-duplications and 1-10Kb deletions).



Supplementary Figure 7. Comparison of experimentally-generated signatures in *EXO1* knockouts (*EXO1* is involved in both HR and MMR pathways) and cancer-derived mutational signatures in HR-deficient but *BRCA1*-intact samples. (a) 96-element substitution spectra of *EXO1* knockouts (seven subclones were aggregated) and two *BRCA1*-intact breast cancers with HR deficiency, PD5956 and PD4841. (b) Substitution signatures associated with *EXO1* knockouts. (c) 8-element indel spectra of *EXO1* knockouts (seven subclones were aggregated), PD5956 and PD4841. (d) Indel signature associated with *EXO1* knockouts. (e) 10-element rearrangement spectra of *EXO1* knockouts (seven subclones were aggregated), PD5956 and PD4841. (f) Rearrangement signature associated with *EXO1* knockouts. Indel and rearrangement profiles of *BRCA1*-intact samples are very similar to the *EXO1* knockout: featuring a high number of repeat-mediated deletions and long (>100kb) tandem duplication rearrangements.



Supplementary Figure 8. The topography of experimentally-generated mutations of nine knockout genes. (a) Histograms exploring transcriptional strand asymmetry. (b) Histograms exploring replication strand asymmetry. (c) Distribution of normalized mutation density across the replication timing domains. Error bars are defined as standard error of the mean.



Supplementary Figure 9. Uncropped Immunoblots from Supplementary Figure 1b.

CHAPTER THREE: DISCUSSION

Functional interpretation of *in vitro* mutation signatures

MSH6 associated signatures

Mismatch repair is essential in guarding the genome of replicating cells against single (or a few) base pair mutations. Mismatches are particularly high at microsatellites. The MMR protein MSH6 is part of the MUT Salpha complex which scans DNA for mismatched lesions. Given its upstream role in MMR, we expected loss of MSH6 to result in MMR associated signatures. Interestingly there are 4 - 5 mismatch repair associated mutation signatures in the COSMIC data base (Wellcome Trust Sanger Institute), Signature 6, 15, 20, 26 and 21. The latter one however, needs to be taken with caution, since it may not be MMR related. Signature 21's association to MMR is that, so far, it has only been found in samples that also had Signatures 15 and 20. Morganella et al studied the genomes of 560 breast cancers (WGS) and identified some unique features of the 4 main MMR signatures in the COSMIC data base.

COSMIC Signature 15 was entirely absent from 560 breast cancers. This could mean that the mutagenic process responsible for Signature 15 is not active or relevant in the development of breast cancer. This signature could be specific to some cancer types due to their specific tissue environment. So far (to my knowledge) it has only been detected in several stomach and a single lung cancer. It would be exciting to find stomach specific mutagens, which in combination with a MMR deficiency, could possibly explain the etiology of Signature 15. The lack of such mutagens in our study could also explain the low cosine similarity between our *in vitro* MMR deficiency model and Signature 15 (0.62).

COSMIC Signatures 6, 20 and 26 were all detected in 560 breast cancers, studied by Morganella et al. Signature 20 and 26 showed a replication time bias, which had previously been linked to MMR deficiency by Supek et al (Supek & Lehner, 2015), but intriguingly, Signature 6 did not. However, Signature 6 is still very likely to be MMR related since it is associated with defective MMR and predominantly found in microsatellite instable tumors (Wellcome Trust Sanger Institute). It may therefore be due to a non-canonical MMR pathway, which is not enriched in early versus late replicons. Our *in vitro* MMR deficiency model shows a very clear replication time bias and, only a modest cosine similarity of 0.76, when compared to Signature 6. Our model does therefore does not explain COSMIC Signature 6.

Morganella et al also showed that Signature 26 had a strong replication strand bias (but no transcription strand bias), which is the case for our MMR deficiency model as well. However, the cosine similarity between our model and Signature 26 is only 0.68. Furthermore, Signature 26 has a peculiar feature, which we did not find in our model: It is less abundant at nucleosome binding sites than their flanking regions. This is the opposite to what is found with other MMR

signatures (Signature 6 and 20) (Morganella *et al*, 2016), and what is known about NER and mutations at nucleosome core DNA (Sabarinathan *et al*, 2016). A possible explanation for this is that Signature 26 is a compound of MMR deficiency and another dominant biological process, which we have not identified yet.

Our MMR deficiency model shows the strongest similarity to COSMIC Signature 20 (0.91), but was not identical (0.95). We therefore suspect that canonical MMR, which is not intact in MSH6 knockout cell lines, could be the mutagenic process underlying Signature 20. Interestingly, a recent publication that appeared after ours (Haradhvala *et al*, 2018), demonstrated that Signature 20 was a compound signature of replication errors (by catalytic dead POLE) and MSI (highly associated with MMR deficiency). The cosine similarity was 0.93, only 0.02 points different from our reported signature. It would be interesting to investigate their model *in vitro* by using our approach with isogenic cell lines. Our data compared to their data, very strongly suggests that a deficiency in the canonical MMR pathway (i.e. MSH6) is largely responsible for COSMIC Signature 20.

In future it will be interesting to decipher the mutational processes that lead to COSMIC Signatures 6, 15, 26 and 21.

FANCC associated signatures

The Fanconi Anemia (FA) repair pathway is primarily known for its role in the repair of crosslinks. This pathway consists of more than 15 proteins (Yao *et al*, 2013) and cooperates closely with HR, NER and TLS. Indeed, the HR proteins BRCA1 and BRCA2 play such an important role in FA repair that they are also known by their FA repair names FANCS and FANCD1. FANCC (Fanconi Anemia Complementation Group C), is part of the FA-core complex, a complex of proteins which is recruited after damage recognition and required for downstream activation of FANCD2. Loss of FANCC disrupts the core complex leading to failure of FA repair (Muniandy *et al*, 2010). FANCC deleted cells are therefore a good model to study the mutagenic process induced by an absence of FA repair.

FA deficiency is associated with chromosomal aberrations, genomic instability and susceptibility to cancer (Crossan & Patel, 2011). This is in line with the observations of our *in vitro* generated mutation signatures for FANCC deleted cells. The dominant mutation types that we see are chromosomal rearrangements, by far more frequent than in any of our other knockout cell lines. Interestingly, the type of rearrangements (and indels) resembles the mutations in BRCA1/2 mutated breast cancer. As a control, we tested our cell line for mutations in other DNA repair genes, and except for FANCC, we found none. The mutator phenotype was not an obvious expectation but is in line with the close association between FA and HR repair. Indeed, FANCC and other FA proteins have previously been described to

be BRCAness genes. Here we show that FANCC also shows a very strong BRACAness mutator phenotype, to our knowledge, the first time this has been done looking at mutation signatures.

Though we can explain the similarity in mutation signatures between our *in vitro* cultivated FANCC and cancer patient derived BRCA1/2 deficient cells, it is not clear why our *in vitro* model would accumulate large numbers of rearrangements. Recently, FA proteins have been shown to be involved in replication folk stability, particularly FANCD2, BRCA2 and RAD51, which causes genomic instability. However, others have shown that the FA-core complex (which includes FANCC) was not involved in replication fork stability (Raghunandan *et al*, 2015).

FA deficient cells are exquisitely sensitive to crosslinking agents but we did not apply exogenous crosslinking agents in our study. The main known endogenous crosslinking agents are aldehydes (Stone *et al*, 2008). Our data suggests that if endogenous crosslinks were the actual source of the genomic instability of FANCC deleted cells in culture, then judging from the mutator phenotype acquired after one month, they could be of greater importance to endogenous DNA damage than hitherto thought.

FA patients are also sensitive to ionizing radiation (Pollard & Gatti, 2009), which is a potent double-strand break inducing agent. Though ionizing radiation also produces a variety of other DNA lesions, presumably including interstrand crosslinks (Dextraze *et al*, 2010), our data suggests that FA proteins could play an important role in other DNA repair processes besides crosslink repair.

EXO1 associated signatures

The nuclease EXO1 is involved in both mismatch repair and double-strand break repair. It removes nucleotides around damaged sites allowing for error free repair processes. Our EXO1 deficient cell line is thus a suitable model to study the crosstalk between mismatch repair and homologous recombination. In accordance with that, we found that EXO1 deleted cells showed a mutator phenotype resembling MMR deficiency to one and HR deficiency to another part. In contrast to other cell lines, invitro cultured EXO1 deleted cells showed elevated base substitution, high levels of 1 bp repeat mediated deletions and a loss of replication time bias, similar to MSH6 deleted cells, albeit milder. At the same time, microhomology mediated 3 bp deletions and rearrangements were elevated, similar to FANCC deleted cells and invivo BRCA1/2 deficient tumors. This is a consequence of EXO1's function but was not necessarily expected, since there are many factors that could have masked this mutation signature. For instance, other nucleases (especially DNA2) which is also recruited to sites of double-strand breaks could have taken over the role of EXO1, as has been observed

in yeast (Cejka, 2015). On the other hand, MMR does not rely on EXO1, therefore EXO1 independent pathways could have compensated for its loss in MMR. The non-essentiality of EXO1 in either MMR or HR may also explain the somewhat milder phenotype compared to MSH6 or FANCC deleted cells.

The elevated mutations in late versus early replicons, was a phenomena observed by multiple researchers between 2000 and 2015 (Watanabe *et al*, 2002; Stamatoyannopoulos *et al*, 2009). Unexpectedly, Supek et al proposed that MMR favored early replicons as opposed to late ones, leading to elevated mutations in the latter (Supek & Lehner, 2015). However, cancer genomes have undergone a long evolutionary process from initial occurrence until clinical detection and associations are not proofs beyond doubt. Here we show that by selectively depleting MMR factors, we also delete the mutation bias in late versus early replicons. Indeed to strengthen the argument, deficiency in core MMR, through loss of MSH6, entirely delete s the bias, while deficiency in alternative MMR repair, through loss of EXO1, delete s the bias to a moderate degree.

POLE associated signatures

POLE is one of the two main replication polymerases in human cells. It functions in DNA synthesis as well as in DNA repair through its proofreading domain. Mutations in POLE are known to induce very large numbers of mutations (Haradhvala *et al*, 2018) and have been proposed to underlie the etiology of COSMIC Signature 10 (Wellcome Trust Sanger Institute). Interestingly, our POLE deleted cells showed no significant mutation increase compared to our other knockout cell lines, such that no base substitution signature could be extracted for comparison to the COSMIC data base signatures. A possible explanation for this surprising result, is that specific mutations in POLE and not its entire depletion are required to produce its known mutator phenotype. This is in fact very likely the case, since the know hypermutation and COSMIC Signature 10 were associated with POLE somatic mutations in its catalytic exonuclease domain (Pro286Arg and Val411Leu 1 (Briggs & Tomlinson, 2013)). It is not unusual that total loss of an enzyme may not recapitulate a loss of function phenotype. For instance, ATM mouse models(Choi *et al*, 2010; Yamamoto *et al*, 2012) showed that ATM knockout mice were viable and mostly recapitulated the human AT phenotype, whereas catalytic inactive ATM (kinase-dead ATM) was embryonic lethal and its expression resulted in higher genomic instability than observed in knockout lymphocytes. Researchers commonly agree that this phenomena could be due to the activity of other related kinases (e.g. ATR, DNA-PKcs), which can compensate for loss of ATM. In the presence of kinase-dead ATM, the compensatory kinases may experience a physical hindrance, presented by the inactive enzyme, leading to a lethal DNA repair deficiency. This activity of catalytic dead enzymes is

termed a dominant negative effect. Likewise, exonuclease deficient POLE could be able to elongate DNA strands, without being able to repair mistakes, resulting in hypermutations. In the total absence of the protein however, other polymerases, presumably POLD with an intact exonuclease domain, could compensate for its loss, resulting in the absence of hypermutations. This is not implausible, since the main consensus in the literature on the function of replication polymerases is that POLE replicates the leading strand (Lujan *et al*, 2016), whilst POLD replicates the leading and lagging strand (Johnson *et al*, 2015). In this context, it would be interesting to introduce the Pro286Arg or Val411Leu mutation, instead of knocking out the gene, in order to test this hypothesis.

Other signatures

The knockout of CHEK2, NEIL1, NUDT1, POLB and POLM did not appear to produce detectable mutational signatures under our experimental conditions. Additionally, apart from the gene-edits themselves, there were no additional recurrent activating mutations or loss-of-function mutations identified in sub-clones after culture, suggesting that the enrichment of driver events was not a feature in the experiment with these cell lines.

DNA glycosylases of the base excision repair pathway are known to show functional redundancy (Krokan & Bjørås, 2013). Consequently, there are no known major diseases associated with a deficiency in single glycosylases (Xie *et al*, 2004; Chan *et al*, 2009; Kemmerich *et al*, 2012). NEIL1 is a DNA glycosylase involved in the removal of oxidized lesions, but MUTYH and OGG1 are also capable of removing oxidized lesions, perhaps compensating for loss of NEIL1 and suppressing mutations (Xie *et al*, 2004).

In the case of the other base excision repair factor in our experiment, the polymerase POLB, our gene editing resulted in a frame-shift mutation (confirmed by Sanger sequencing), which should have resulted in a knockout. However, a truncated version of the protein was still expressed (as evident on western blot), probably due to an alternative transcription start site downstream of the gRNA targeting the gene. Because of this, we hypothesize that the primary reason for lack of mutations in our *in vitro* model of POLB deficiency was an incomplete loss of the protein after gene editing.

Although Sanger sequencing as well as western blot confirmed the knockout for CHK2, we did not observe enrichment of mutations in this cell line. CHK2 and CHK1 share many substrates (Lazzaro *et al*, 2009). It is possible that a functional compensation between the two prevented enrichment of mutations in CHK2 deleted cells. We also assessed sensitivity of CHK2 to DNA damaging agents and did not observe a sensitivity to double-strand break inducing agents as we would have expected. This DNA damage tolerance is very peculiar, and unexpected, since CHK2 is canonically activated after double-strand breaks in an ATM

dependent manner (Manic *et al*, 2015). It is possibly that CHK1 is compensating for loss of CHK2. For a future experiments, it would be interesting to replicate the phenotype of the CHK2 knockout in at least one other cell line.

NUDT1 is an enzyme specialized on the removal of oxidized nucleotides, one of the most common damages, from the nucleotides pool. We do not know why our knockout cell line did not produce mutations, but it is possible that DNA repair pathways, such as MMR or BER compensate for loss of NUDT1 by removing inserted damaged nucleotides from the DNA.

POLM deficient cells showed the least substitutions and indels and average rearrangements. It is involved in NHEJ, which is more mutagenic than HR. Furthermore, POLM deficiency increases resistance to oxidative damage and reduces apoptosis (Martin & Blanco, 2014). The POLM knockout may therefore have a stronger shift towards HR due to inefficient NHEJ (Escudero *et al*, 2014), which may produce a subtle reduction in indel rates. This is however very speculative, and we remain to properly test POLM deficient cells for improved DNA damage repair. It has been reported that the catalytic mutant POLM was mutagenic, but not the knockout (Escudero *et al*, 2014). This argues for the possibility of a dominant negative effect in this model system.

Clinical significance of i*n vitro* mutation signatures

The signatures that we identified could be of clinical significance. The MSH6 mutant derived signature was in line with the known MMR associated patterns. This confirms that our *in vitro* generated signatures are indeed the same or similar to *in vivo* signatures. In light of this, mutant EXO1 or FANCC derived signatures may represent a subset of tumors, too small to have been identified in whole genome sequencing of thousands of tumors, where strong mutation signatures may cover subtle ones. These novel BRCA related signatures may thus represent markers for a hitherto neglected subset of tumors, which may be targeted through specific vulnerabilities conferred by EXO1 or FANCC deficiency. For instance, FANCC deficient cells are extremely vulnerable to crosslinking chemotherapeutics, such as cisplatin or oxaliplatin.

Optimization of experimental setup

Sequencing technologies have technical limitations, sample preparation as well as the sequencing process itself introduce mutational artefacts (Costello *et al*, 2013; Wong *et al*, 2014). Each protocol and instrument may have its own unique features. We controlled for such potential technical flaws by using isogenic cell lines including a pool of original parental cells and by sequencing all our samples on the same instrument using the same protocol. *In vivo*

signatures however cannot be controlled to such a high degree due to heterogeneity of tumor and control samples.

Known signatures could be a compound of two or more mutational processes which are strongly associated with each other, e.g. interaction of genetic mutagens with tissue specific mutagens. A recent publication by Haradhvala et al reported to have revealed the etiology of COSMIC Signature 20 (cosine similarity 0.93) by demonstrating that the combination of exonuclease mutant POLD and MSI did not merely produce an additive mutation signature of the two respective mutagenic processes but rather compounded into an unrelated, distinct one (Haradhvala *et al*, 2018). There are many other possible combinations that could arise *in vivo* or *in vitro*. Testing combinations of mutagens *in vivo* is not feasible, whereas with *in vitro* mutagens, we can control the mutagens and their combinations. The knowledge of the *in vitro* processes and their signatures could then be used to understand the *in vivo* ones.

One source of noise in our *in vitro* generated mutation signatures was the highly abundant C > A mutation in all knockout clones and parental cell lines. We are not sure of the source of this lesion, but it appeared to be associated with our model system or cell culture conditions. The signature is very similar to COSMIC Signature 18, first observed in breast and stomach carcinomas. Signature 18 is presumed to stem from loss of MUTYH (Viel *et al*, 2017; Pilati *et al*, 2017), which repairs lesions by reactive oxygen species. Sequencing all of our clones, we did not detect any deleterious mutations in MUTYH. Almost all cells of the human body live in tissue environments with low exposure to oxygen (Jagannathan *et al*, 2016). The majority of cell culture laboratories however, are set up to culture cells in an incubator environment of 20% oxygen (atmospheric oxygen level). It is feasible that the 20% oxygen level in our incubator is responsible for the enrichment of C > A mutations by reactive oxygen species. Furthermore, cell culture requires handling of cells outside of the incubator, in a lamina or in front of a microscope. In support of this argument, a signature similar to Signature 18 has also been reported in organoid based *in vitro* generated mutation signatures (Blokzijl *et al*, 2016; Drost *et al*, 2017). This is something that remains to be tested in the future. For instance, cells could be cultured in 20% oxygen and 3% oxygen (physiological level (Jagannathan *et al*, 2016)) in order asses the link between oxygen levels and C > A mutations. Such an experiment would confirm the association between C > A mutations and reactive oxygen radicals, and future *in vitro* mutation signature studies could be performed under low oxygen conditions.

In addition to the background signature which may have covered actual knockout signatures, we cannot exclude that the culture time was insufficient for the enrichment of mutation patterns in some knockouts. Moreover, the cell lines do not all have same proliferation rate. Application of an exogenous mutagen may also be required to amplify mutational patterns in some of our knockout cell lines. Some knockouts may only produce a signature in specific cell lines, due to tissue specific differences. Therefore, increasing the culture time and adding exogenous

mutagens as well as other cell line models are conditions that need to be tested in future experiments.

Future of mutation signatures

Though mutation signatures are by definition due to DNA damage and DNA repair processes, it is important to study other DNA repair associated proteins in the context of mutation signatures. Fascinating targets are epigenetic regulators. For instance, ARID1A is one of the most commonly mutated genes in cancer (Shen *et al*, 2018). Loss of function is associated with loss of chromatin accessibility. Researchers found that ARID1A is recruited by MSH2 to chromatin to promote MMR. Loss of ARID1A decreased the activity of MMR and resulted in increased mutation. The increased mutation load resulted in an increased susceptibility of ARID1A deficient cancer to immune checkpoint blockade therapy (Shen *et al*, 2018).
Complex signatures may result from a combination of mutational processes, therefore some of the current signatures may require further delineation. For instance, loss of POLE exonuclease (proofreading) activity is associated with a distinct base substitution signature (COSMIC signature 10). Loss of MMR factors (identified through MSI) is associated with at least 4 distinct signatures. However, loss of POLE exonuclease in combination with loss of MMR does not simply result in a combinational (additive) mutation pattern, but rather in one or two distinct de novo signatures (Haradhvala *et al*, 2018).

According to Alexandrov et (Alexandrov *et al*, 2018) elucidating the underlying mutational processes of mutation signatures will depend on two major streams of investigation: (i.) The generation of mutational signatures from model systems exposed to known mutagens or genetic perturbations and the comparison of those signatures with those found in human cancer genomes. (ii.) An overlay of mutation signatures with other characteristics of cancer through different approaches ranging from molecular profiling to epidemiology. Collectively, these studies will advance our understanding of cancer etiology with potential implications for prevention and treatment.

Conclusion

We have demonstrated for the first time in the known scientific literature, that mutations signatures associated to defective DNA repair mechanisms can be generated *in vitro* using isogenic cell lines. This novel approach for studying mutagenesis is complementary to the study of genomes from cancer patients, and may inform about their etiology. We therefore present a novel approach that adds to the repertoire of tools that researchers have developed to aid in our the fight against cancer.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, et al (2000) The genome sequence of Drosophila melanogaster. *Science* **287:** 2185–2195

Adjiri A (2017) DNA Mutations May Not Be the Cause of Cancer. *Oncol Ther* **5:** 85–101

Alexandrov L, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Boot A, Covington KR, Gordenin DA, Bergstrom E, López-Bigas N, Klimczak LJ, McPherson JR, Morganella S, Sabarinathan R, Wheeler DA, Mustonen V, Getz G, Rozen SG, Stratton MR, PCAWG Mutational Signatures Working Group, et al (2018) The Repertoire of Mutational Signatures in Human Cancer.

Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S & Stratton MR (2015) Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47:** 1402–1407

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg Å, Børresen-Dale A-L, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, et al (2013a) Signatures of mutational processes in human cancer. *Nature* **500:** 415–421

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ & Stratton MR (2013b) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3:** 246–259

Andersen SL & Sekelsky J (2010) Meiotic versus mitotic recombination: two different routes for double-strand break repair: the different functions of meiotic versus mitotic DSB repair are reflected in different pathway usage and different outcomes. *Bioessays* **32:** 1058–1066

Andersson BS, Collins VP, Kurzrock R, Larkin DW, Childs C, Ost A, Cork A, Trujillo JM, Freireich EJ & Siciliano MJ (1995) KBM-7, a human myeloid leukemia cell line with double Philadelphia chromosomes lacking normal c-ABL and BCR transcripts. *Leukemia* **9:** 2100–2108

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796–815

Arteaga C (2003) Targeting HER1/EGFR: a molecular approach to cancer therapy. *Semin. Oncol.* **30:** 3–14

Aunan JR, Cho WC & Søreide K (2017) The Biology of Aging and Cancer: A Brief Overview of Shared and Divergent Molecular Hallmarks. *Aging Dis* **8:** 628–642

Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, Van Allen E, Kryukov GV, Sboner A, Theurillat J-P, Soong TD, Nickerson E, Auclair D, Tewari A, Beltran H, Onofrio RC, et al (2013) Punctuated Evolution of Prostate Cancer Genomes. *Cell* **153:** 666–677

Bauer J, Curtin JA, Pinkel D & Bastian BC (2007) Congenital melanocytic nevi frequently harbor NRAS mutations but no BRAF mutations. *J. Invest. Dermatol.* **127:** 179–182

Berry MW, Browne M, Langville AN, Pauca VP & Plemmons RJ (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis* **52:** 155–173

Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, Huch M, Boymans S, Kuijk E, Prins P, Nijman IJ, Martincorena I, Mokry M, Wiegerinck CL, Middendorp S, Sato T, Schwank G, Nieuwenhuis EES, Verstegen MMA, van der Laan LJW, et al (2016) Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538:** 260–264

Boland CR & Goel A (2010) Microsatellite instability in colorectal cancer.

Boyer TG, Chen P-L & Lee W-H (2001) Genome mining for human cancer genes: whereforeartthou? *Trends Mol Med* **7:** 187–189

Brash DE, Rudolph JA, Simon JA, Lin A, McKenna GJ, Baden HP, Halperin AJ & Pontén J (1991) A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proceedings of the National Academy of Sciences* **88:** 10124–10128

Briggs S & Tomlinson I (2013) Germline and somatic polymerase ε and δ mutations define a new class of hypermutated colorectal and endometrial cancers. *The Journal of Pathology* **230:** 148–153

Bukowska B & Karwowski BT (2018) Actual state of knowledge in the field of diseases related with defective nucleotide excision repair. *Life Sci.* **195:** 6–18

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, et al (1996) Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science* **273:** 1058–1073

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282:** 2012–2018

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, et al (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40:** 722–729

Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487:** 330–337

Cancer Genome Atlas Research Network (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513:** 202–209

Cancer Research UK website. Available at: https://www.cancerresearchuk.org

Capella G, Cronauer-Mitra S, Pienado MA & Perucho M (1991) Frequency and spectrum of mutations at codons 12 and 13 of the c-K-ras gene in human tumors. *Environ. Health Perspect.* **93:** 125–131

Carbone M, Gruber J & Wong M (2004) Modern criteria to establish human cancer etiology. *Semin. Cancer Biol.* **14:** 397–398

Carette JE, Raaben M, Wong AC, Herbert AS, Obernosterer G, Mulherkar N, Kuehne AI, Kranzusch PJ, Griffin AM, Ruthel G, Dal Cin P, Dye JM, Whelan SP, Chandran K & Brummelkamp TR (2011) Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. *Nature* **477:** 340–343

Cejka P (2015) DNA End Resection: Nucleases Team Up with the Right Partners to Initiate Homologous Recombination. *J. Biol. Chem.* **290:** 22931–22938

Chan MK, Ocampo-Hafalla MT, Vartanian V, Jaruga P, Kirkali G, Koenig KL, Brown S, Lloyd RS, Dizdaroglu M & Teebor GW (2009) Targeted deletion of the genes encoding NTH1 and NEIL1 DNA N-glycosylases reveals the existence of novel carcinogenic oxidative damage to DNA. *DNA Repair (Amst.)* **8:** 786–794

Chapman JR, Sossick AJ, Boulton SJ & Jackson SP (2012a) BRCA1-associated exclusion of 53BP1 from DNA damage sites underlies temporal control of DNA repair. *J. Cell. Sci.* **125:** 3529–3534

Chapman JR, Taylor MRG & Boulton SJ (2012b) Playing the end game: DNA double-strand break repair pathway choice. *Mol. Cell* **47:** 497–510

Chiruvella KK, Liang Z & Wilson TE (2013a) Repair of double-strand breaks by end joining. *Cold Spring Harb Perspect Biol* **5:** a012757–a012757

Chiruvella KK, Liang Z & Wilson TE (2013b) Repair of Double-Strand Breaks by End Joining. *Cold Spring Harb Perspect Biol* **5:** a012757–a012757

Choi S, Gamper AM, White JS & Bakkenist CJ (2010) Inhibition of ATM kinase activity does not phenocopy ATM protein disruption: implications for the clinical utility of ATM kinase inhibitors. *Cell Cycle* **9:** 4052–4057

Cortes-Ciriano I, Lee S, Park W-Y, Kim TM & Park PJ (2017) A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun* **8:** 15180

Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, Fostel JL, Friedrich DC, Perrin D, Dionne D, Kim S, Gabriel SB, Lander ES, Fisher S & Getz G (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41:** e67–e67

Covo S, de Villartay J-P, Jeggo PA & Livneh Z (2009) Translesion DNA synthesis-assisted non-homologous end-joining of complex double-strand breaks prevents loss of DNA sequences in mammalian cells. *Nucleic Acids Res.* **37:** 6737–6745

Crossan GP & Patel KJ (2011) The Fanconi anaemia pathway orchestrates incisions at sites of crosslinked DNA. *The Journal of Pathology* **226:** 326–337

Curtin NJ (2012) DNA repair dysregulation from cancer driver to therapeutic target. *Nature Reviews Cancer* **12:** 801–817

Daley JM & Sung P (2014) 53BP1, BRCA1, and the choice between recombination and end joining at DNA double-strand breaks. *Molecular and Cellular Biology* **34:** 1380–1388

Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, Ramakrishna M, Martin S, Boyault S, Sieuwerts AM, Simpson PT, King TA, Raine K, Eyfjord JE, Kong G, Borg Å, Birney E, Stunnenberg HG, van de Vijver MJ, Børresen-Dale A-L, et al (2017) HRDetect

is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23:** 517–525

De Bont R (2004) Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19:** 169–185

de Jager M, van Noort J, van Gent DC, Dekker C, Kanaar R & Wyman C (2001) Human Rad50/Mre11 is a flexible complex that can tether DNA ends. *Mol. Cell* **8:** 1129–1135

Dextraze M-E, Gantchev T, Girouard S & Hunting D (2010) DNA interstrand cross-links induced by ionizing radiation: an unsung lesion. *Mutat. Res.* **704:** 101–107

Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, de Ligt J, Behjati S, Grolleman JE, van Wezel T, Nik-Zainal S, Kuiper RP, Cuppen E & Clevers H (2017) Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358:** 234–238

EHRET CF & DE HALLER G (1963) ORIGIN, DEVELOPMENT AND MATURATION OF ORGANELLES AND ORGANELLE SYSTEMS OF THE CELL SURFACE IN PARAMECIUM. *J. Ultrastruct. Res.* **23:** SUPPL6:1–42

Elgar G & Vavouri T (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet.* **24:** 344–352

Escudero B, Lucas D, Albo C, Dhup S, Bacher JW, Sánchez-Muñoz A, Fernández M, Rivera-Torres J, Carmona RM, Fuster E, Carreiro C, Bernad R, González MA, Andrés V, Blanco L, Roche E, Fabregat I, Samper E & Bernad A (2014) Polμ Deficiency Increases Resistance to Oxidative Damage and Delays Liver Aging. *PLoS ONE* **9:** e93074

Essen LO & Klar T (2006) Light-driven DNA repair by photolyases. *Cell. Mol. Life Sci.* **63:** 1266–1277

Fedeles BI, Singh V, Delaney JC, Li D & Essigmann JM (2015) The AlkB Family of Fe(II)/α-Ketoglutarate-dependent Dioxygenases: Repairing Nucleic Acid Alkylation Damage and Beyond. *J. Biol. Chem.* **290:** 20734–20742

Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, Janizek JD, Huang X, Starita LM & Shendure J (2018) Accurate functional classification of thousands of BRCA1variants with saturation genome editing. *bioRxiv***:** 294520

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA & Merrick JM (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269:** 496–512

Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, Mortimer P, Swaisland H, Lau A, O'Connor MJ, Ashworth A, Carmichael J, Kaye SB, Schellens JHM & de Bono JS (2009) Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N. Engl. J. Med.* **361:** 123–134

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, et al (1995) The minimal gene complement of Mycoplasma genitalium. *Science* **270:** 397–403

Fridman JS & Lowe SW (2003) Control of apoptosis by p53. *Oncogene* **22:** 9030–9040

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N & Stratton MR (2004) A census of human cancer genes. *Nature Reviews Cancer* **4:** 177–183

Futreal PA, Kasprzyk A, Birney E, Mullikin JC, Wooster R & Stratton MR (2001) Cancer and genomics. *Nature* **409:** 850–852

Garaycoechea JI, Crossan GP, Langevin F, Mulderrig L, Louzada S, Yang F, Guilbaud G, Park N, Roerink S, Nik-Zainal S, Stratton MR & Patel KJ (2018) Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* **553:** 171–177

Garraway LA (2013) Genomics-Driven Oncology: Framework for an Emerging Paradigm. *Journal of Clinical Oncology* **31:** 1806–1814

Giglia-Mari G, Zotter A & Vermeulen W (2011) DNA damage response. *Cold Spring Harb Perspect Biol* **3:** a000745–a000745

Gocayne J, Robinson DA, FitzGerald MG, Chung FZ, Kerlavage AR, Lentes KU, Lai J, Wang CD, Fraser CM & Venter JC (1987) Primary structure of rat cardiac beta-adrenergic and muscarinic cholinergic receptors obtained by automated DNA sequence analysis: further evidence for a multigene family. *Proceedings of the National Academy of Sciences* **84:** 8296–8300

Goodarzi AA & Jeggo PA (2013) The repair and signaling responses to DNA double-strand breaks. *Adv. Genet.* **82:** 1–45

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, et al (2007) Patterns of somatic mutation in human cancer genomes. *Nature* **446:** 153–158

Grollman AP & Moriya M (1993) Mutagenesis by 8-oxoguanine: an enemy within. *Trends in Genetics* **9:** 246–249

Gupta R, Somyajit K, Narita T, Maskey E, Stanlie A, Kremer M, Typas D, Lammers M, Mailand N, Nussenzweig A, Lukas J & Choudhary C (2018) DNA Repair Network Analysis Reveals Shieldin as a Key Regulator of NHEJ and PARP Inhibitor Sensitivity. *Cell*

Hajdu SI (2004) Greco-Roman thought about cancer. *Cancer* **100:** 2048–2051

Hajdu SI (2010) A note from history: Landmarks in history of cancer, part 1. *Cancer* **117:** 1097–1102

Hanahan D & Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* **144:** 646–674

Hanawalt PC (1994) Transcription-coupled repair and human disease. *Science* **266:** 1957–1958

Handbook of Blind Source Separation (2010) Handbook of Blind Source Separation.

Haradhvala NJ, Kim J, Maruvka YE, Polak P, Rosebrock D, Livitz D, Hess JM, Leshchiner I, Kamburov A, Mouw KW, Lawrence MS & Getz G (2018) Distinct mutational signatures

characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat Commun* **9:** 1746

Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, Rheinbay E, Kim J, Maruvka YE, Braunstein LZ, Kamburov A, Hanawalt PC, Wheeler DA, Koren A, Lawrence MS & Getz G (2016) Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164:** 538–549

Harfe BD & Jinks-Robertson S (2000) DNA mismatch repair and genetic instability. *Annu. Rev. Genet.* **34:** 359–399

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

Helleday T, Petermann E, Lundin C, Hodgson B & Sharma RA (2008) DNA repair pathways as targets for cancer therapy. *Nature Reviews Cancer* **8:** 193–204

Hernández G, Ramírez MJ, Minguillón J, Quiles P, de Garibay GR, Aza-Carmona M, Bogliolo M, Pujol R, Prados-Carvajal R, Fernández J, García N, López A, Gutiérrez-Enríquez S, Diez O, Benítez J, Salinas M, Teulé A, Brunet J, Radice P, Peterlongo P, et al (2018) Decapping protein EDC4 regulates DNA repair and phenocopies BRCA1. *Nat Commun* **9:** 967

Hiom K (2005) DNA repair: how to PIKK a partner. *Curr. Biol.* **15:** R473–5

Hitchins MP, Wong JJL, Suthers G, Suter CM, Martin DIK, Hawkins NJ & Ward RL (2007) Inheritance of a cancer-associated MLH1 germ-line epimutation. *N. Engl. J. Med.* **356:** 697–705

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

Houghton AN (1994) Cancer antigens: immune recognition of self and altered self. *J. Exp. Med.* **180:** 1–4

Howard BD & Tessman I (1964) Identification of the altered bases in mutated single-stranded DNA: II. In vivo mutagenesis by 5-bromodeoxyuridine and 2-aminopurine. *J. Mol. Biol.* **9:** 364–371

Huang MN, Yu W, Teoh WW, Ardin M, Jusakul A, Ng AWT, Boot A, Abedi-Ardekani B, Villar S, Myint SS, Othman R, Poon SL, Heguy A, Olivier M, Hollstein M, Tan P, Teh BT, Sabapathy K, Zavadil J & Rozen SG (2017) Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res.* **27:** 1475–1486

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945

Iyer BV, Kenward M & Arya G (2011) Hierarchies in eukaryotic genome organization: Insights from polymer theory and simulations. *BMC Biophys* **4:** 8

Iyer RR, Pluciennik A, Burdett V & Modrich PL (2006) DNA mismatch repair: functions and mechanisms. *Chem. Rev.* **106:** 302–323

Jackson SP & Bartek J (2009a) The DNA-damage response in human biology and disease. *Nature* **461:** 1071–1078

Jackson SP & Bartek J (2009b) The DNA-damage response in human biology and disease. *Nature* **461:** 1071–1078

Jackson SP & Helleday T (2016) DNA REPAIR. Drugging DNA repair. *Science* **352:** 1178–1179

Jagannathan L, Cuddapah S & Costa M (2016) Oxidative stress under ambient and physiological oxygen tension in tissue culture. *Curr Pharmacol Rep* **2:** 64–72

Jena NR (2012) DNA damage by reactive species: Mechanisms, mutation and repair. *J. Biosci.* **37:** 503–517

Jilani A, Ramotar D, Slack C, Ong C, Yang XM, Scherer SW & Lasko DD (1999) Molecular Cloning of the Human Gene, PNKP, Encoding a Polynucleotide Kinase 3′-Phosphatase and Evidence for Its Role in Repair of DNA Strand Breaks Caused by Oxidative Damage. *J. Biol. Chem.* **274:** 24176–24186

Johnson RE, Klassen R, Prakash L & Prakash S (2015) A Major Role of DNA Polymerase δ in Replication of Both the Leading and Lagging DNA Strands. *Mol. Cell* **59:** 163–175

Kaina B, Christmann M, Naumann S & Roos WP (2007) MGMT: Key node in the battle against genotoxicity, carcinogenicity and apoptosis induced by alkylating agents. *DNA Repair (Amst.)* **6:** 1079–1099

Kalia M (2015) Biomarkers for personalized oncology: recent advances and future challenges. *Metab. Clin. Exp.* **64:** S16–21

Kamileri I, Karakasilioti I & Garinis GA (2012) Nucleotide excision repair: new tricks with old bricks. *Trends Genet.* **28:** 566–573

Kapp FG, Perlin JR, Hagedorn EJ, Gansner JM, Schwarz DE, O'Connell LA, Johnson NS, Amemiya C, Fisher DE, Wölfle U, Trompouki E, Niemeyer CM, Driever W & Zon LI (2018) Protection from UV light is an evolutionarily conserved feature of the haematopoietic niche. *Nature* **558:** 445–448

Kemmerich K, Dingler FA, Rada C & Neuberger MS (2012) Germline ablation of SMUG1 DNA glycosylase causes loss of 5-hydroxymethyluracil- and UNG-backup uracil-excision activities and increases cancer predisposition of Ung−/−Msh2−/− mice. *Nucleic Acids Res.* **40:** 6016–6025

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, et al (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453:** 56–64

Klein Douwel D, Boonen RACM, Long DT, Szypowska AA, Räschle M, Walter JC & Knipscheer P (2014) XPF-ERCC1 acts in Unhooking DNA interstrand crosslinks in cooperation with FANCD2 and FANCP/SLX4. *Mol. Cell* **54:** 460–471

Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, Fan H, Shen H, Way GP, Greene CS, Liu Y, Akbani R, Feng B, Donehower LA, Miller C, Shen Y, Karimi M, Chen H, Kim P, Jia P, et al (2018) Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep* **23:** 239–254.e6

Kohl KP & Sekelsky J (2013) Meiotic and mitotic recombination in meiosis. *Genetics* **194:** 327–334

Korhonen MK, Vuorenmaa E & Nyström M (2008) The first functional study of MLH3 mutations found in cancer patients. *Genes Chromosomes Cancer* **47:** 803–809

Koster BD, de Gruijl TD & van den Eertwegh AJM (2015) Recent developments and future challenges in immune checkpoint inhibitory cancer treatment. *Current Opinion in Oncology* **27:** 482–488

Kotecki M, Reddy PS & Cochran BH (1999) Isolation and characterization of a near-haploid human cell line. *Exp. Cell Res.* **252:** 273–280

Krokan HE & Bjørås M (2013) Base excision repair. *Cold Spring Harb Perspect Biol* **5:** a012583–a012583

Kroutil LC, Register K, Bebenek K & Kunkel TA (1996) Exonucleolytic proofreading during replication of repetitive DNA. *Biochemistry* **35:** 1046–1053

Kunitomi H, Banno K, Yanokura M, Takeda T, Iijima M, Nakamura K, Iida M, Adachi M, Watanabe K, Matoba Y, Kobayashi Y, Tominaga E & Aoki D (2017) New use of microsatellite instability analysis in endometrial cancer. *Oncol Lett* **14:** 3297–3301

Kunkel TA (2004) DNA replication fidelity. *J. Biol. Chem.* **279:** 16895–16898

Kunkel TA & Erie DA (2015) Eukaryotic Mismatch Repair in Relation to DNA Replication. *Annu. Rev. Genet.* **49:** 291–313

Kuper J & Kisker C (2012) Damage recognition in nucleotide excision DNA repair. *Curr. Opin. Struct. Biol.* **22:** 88–93

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, et al (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, et al (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499:** 214–218

Lazzaro F, Giannattasio M, Puddu F, Granata M, Pellicioli A, Plevani P & Muzi-Falconi M (2009) Checkpoint mechanisms at the intersection between DNA damage and repair. *DNA Repair (Amst.)* **8:** 1055–1067

Leadon S (1996) Repair of DNA Damage Produced by Ionizing Radiation: A Minireview. *Semin Radiat Oncol* **6:** 295–305

Lee DD & Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* **401:** 788–791

Lee JW, Blanco L, Zhou T, Garcia-Diaz M, Bebenek K, Kunkel TA, Wang Z & Povirk LF (2004) Implication of DNA polymerase lambda in alignment-based gap filling for nonhomologous DNA end joining in human nuclear extracts. *J. Biol. Chem.* **279:** 805–

811

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, et al (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456:** 66–72

Li X, Park W-J, Pyeritz RE & Jabs EW (1995) Effect on splicing of a silent FGFR2 mutation in Crouzon syndrome. *Nat. Genet.* **9:** 232–233

Lindahl T, Verly WG & Paquette Y (2004) Inroads into base excision repair I. The discovery of apurinic/apyrimidinic (AP) endonuclease. 'An endonuclease for depurinated DNA in Escherichia coli B,' Canadian Journal of Biochemistry, 1972.

Lips J & Kaina B (2001) DNA double-strand breaks trigger apoptosis in p53-deficient fibroblasts. *Carcinogenesis* **22:** 579–585

Lodato MA, Rodin RE, Bohrson CL, Coulter ME, Barton AR, Kwon M, Sherman MA, Vitzthum CM, Luquette LJ, Yandava C, Yang P, Chittenden TW, Hatem NE, Ryu SC, Woodworth MB, Park PJ & Walsh CA (2017) Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359:** 555–559

Loeb LA & Kunkel TA (1982) Fidelity of DNA synthesis. *Annu. Rev. Biochem.* **51:** 429–457

Lopez-Martinez D, Liang C-C & Cohn MA (2016) Cellular response to DNA interstrand crosslinks: the Fanconi anemia pathway. *Cell. Mol. Life Sci.* **73:** 3097–3114

Lord CJ & Ashworth A (2016) BRCAness revisited. *Nature Reviews Cancer* **16:** 110–120

Luch A (2005) Nature and nurture - lessons from chemical carcinogenesis. *Nature Reviews Cancer* **5:** 113–125

Lujan SA, Clausen AR, Clark AB, MacAlpine HK, MacAlpine DM, Malc EP, Mieczkowski PA, Burkholder AB, Fargo DC, Gordenin DA & Kunkel TA (2014) Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res.* **24:** 1751–1764

Lujan SA, Williams JS & Kunkel TA (2016) DNA Polymerases Divide the Labor of Genome Replication. *Trends Cell Biol.* **26:** 640–654

Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, Louis DN, Christiani DC, Settleman J & Haber DA (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **350:** 2129–2139

Malu S, Malshetty V, Francis D & Cortes P (2012) Role of non-homologous end joining in V(D)J recombination. *Immunol. Res.* **54:** 233–246

Manic G, Obrist F, Sistigu A & Vitale I (2015) Trial Watch: Targeting ATM-CHK2 and ATR-CHK1 pathways for anticancer therapy. *Mol Cell Oncol* **2:** e1012976

Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, Koboldt DC, Fulton RS, Delehaunty KD, McGrath SD, Fulton LA, Locke DP, Magrini VJ, Abbott RM, Vickery TL, Reed JS, Robinson JS, Wylie T, Smith SM, Carmichael L, et al (2009) Recurring

Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. *N. Engl. J. Med.* **361:** 1058–1066

Martin MJ & Blanco L (2014) Decision-making during NHEJ: a network of interactions in human Polμ implicated in substrate recognition and end-bridging. *Nucleic Acids Res.* **42:** 7923–7934

Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR & Campbell PJ (2017) Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171:** 1029–1041.e21

Mateos-Gomez PA, Gong F, Nair N, Miller KM, Lazzerini-Denchi E & Sfeir A (2015) Mammalian polymerase θ promotes alternative NHEJ and suppresses recombination. *Nature* **518:** 254–257

Matos J & West SC (2014) Holliday junction resolution: Regulation in space and time. *DNA Repair (Amst.)* **19:** 176–181

Matsuda T, Kawanishi M, Matsui S, Yagi T & Takebe H (1998) Specific tandem GG to TT base substitutions induced by acetaldehyde are due to intra-strand crosslinks between adjacent guanine bases. *Nucleic Acids Res.* **26:** 1769–1774

Matsuoka S, Ballif BA, Smogorzewska A, McDonald ER, Hurov KE, Luo J, Bakalarski CE, Zhao Z, Solimini N, Lerenthal Y, Shiloh Y, Gygi SP & Elledge SJ (2007) ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316:** 1160–1166

McCulloch SD & Kunkel TA (2008) The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res.* **18:** 148–161

Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, Raine K, Maddison M, Anderson E, Stratton MR, Gartner A & Campbell PJ (2014) C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res.* **24:** 1624–1636

Meier B, Volkova N, Hong Y, Schofield P, Campbell PJ, Gerstung M & Gartner A (2017) Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. *bioRxiv***:** 149153

Moder M, Velimezi G, Owusu M, Mazouzi A, Wiedner M, Ferreira da Silva J, Robinson-Garcia L, Schischlik F, Slavkovsky R, Kralovics R, Schuster M, Bock C, Ideker T, Jackson SP, Menche J & Loizou JI (2017) Parallel genome-wide screens identify synthetic viable interactions between the BLM helicase complex and Fanconi anemia. *Nat Commun* **8:** 1238

Morganella S, Alexandrov LB, Glodzik D, Zou X, Davies H, Staaf J, Sieuwerts AM, Brinkman AB, Martin S, Ramakrishna M, Butler A, Kim H-Y, Borg Å, Sotiriou C, Futreal PA, Campbell PJ, Span PN, Van Laere S, Lakhani SR, Eyfjord JE, et al (2016) The topography of mutational processes in breast cancer genomes. *Nat Commun* **7:** 11383

Mullis K, Faloona F, Scharf S, Saiki R, Horn G & Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51 Pt 1:** 263–273

Muniandy PA, Liu J, Majumdar A, Liu S-T & Seidman MM (2010) DNA interstrand crosslink

repair in mammalian cells: step by step. *Crit. Rev. Biochem. Mol. Biol.* **45:** 23–49

Ni TT, Marsischky GT & Kolodner RD (1999) MSH2 and MSH6 Are Required for Removal of Adenine Misincorporated Opposite 8-Oxo-Guanine in S. cerevisiae. *Mol. Cell* **4:** 439–444

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, et al (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* **149:** 979–993

Nitiss JL (2009) DNA topoisomerase II and its growing repertoire of biological functions. *Nature Reviews Cancer* **9:** 327–337

Nojima K, Hochegger H, Saberi A, Fukushima T, Kikuchi K, Yoshimura M, Orelli BJ, Bishop DK, Hirano S, Ohzeki M, Ishiai M, Yamamoto K, Takata M, Arakawa H, Buerstedde J-M, Yamazoe M, Kawamoto T, Araki K, Takahashi JA, Hashimoto N, et al (2005) Multiple Repair Pathways Mediate Tolerance to Chemotherapeutic Cross-linking Agents in Vertebrate Cells. *Cancer Res.* **65:** 11704–11711

Nunney L & Muir B (2015) Peto's paradox and the hallmarks of cancer: constructing an evolutionary framework for understanding the incidence of cancer. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **370:** 20150161

O'Donnell M, Langston L & Stillman B (2013) Principles and concepts of DNA replication in bacteria, archaea, and eukarya. *Cold Spring Harb Perspect Biol* **5:** a010108–a010108

O'Donovan A, Davies AA, Moggs JG, West SC & Wood RD (1994) XPG endonuclease makes the 3′ incision in human DNA nucleotide excision repair. *Nature* **371:** 432–435

Oey H & Whitelaw E (2014) On the meaning of the word 'epimutation'. *Trends Genet.* **30:** 519–520

Ohno S (1972) So much ‘junk’ DNA in our genome. *Brookhaven Symp. Biol.* **23:** 366–370

Ozturk M (1991) p53 mutation in hepatocellular carcinoma after aflatoxin exposure. *Lancet* **338:** 1356–1359

Pal T, Permuth-Wey J, Kumar A & Sellers TA (2008) Systematic review and meta-analysis of ovarian cancers: estimation of microsatellite-high frequency and characterization of mismatch repair deficient tumor histology. *Clin. Cancer Res.* **14:** 6847–6854

Pennisi E (2012) Genomics. ENCODE project writes eulogy for junk DNA. *Science* **337:** 1159–1161

Pfeifer GP & Besaratinia A (2009) Mutational spectra of human cancer. *Hum. Genet.* **125:** 493–506

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

Pilati C, Shinde J, Alexandrov LB, Assié G, André T, Hélias-Rodzewicz Z, Ducoudray R, Le Corre D, Zucman-Rossi J, Emile J-F, Bertherat J, Letouzé E & Laurent-Puig P (2017) Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and

adrenocortical carcinomas. *The Journal of Pathology* **242:** 10–15

Plass C, Pfister SM, Lindroth AM, Bogatyrova O, Claus R & Lichter P (2013) Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nat. Rev. Genet.* **14:** 765–780

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin M-L, Ordóñez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, et al (2010a) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463:** 191–196

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin M-L, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordóñez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, et al (2010b) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463:** 184–190

Pollard JM & Gatti RA (2009) Clinical radiation sensitivity with DNA repair disorders: an overview. *Int. J. Radiat. Oncol. Biol. Phys.* **74:** 1323–1331

Pollock PM, Harper UL, Hansen KS, Yudt LM, Stark M, Robbins CM, Moses TY, Hostetter G, Wagner U, Kakareka J, Salem G, Pohida T, Heenan P, Duray P, Kallioniemi O, Hayward NK, Trent JM & Meltzer PS (2003) High frequency of BRAF mutations in nevi. *Nat. Genet.* **33:** 19–20

Pon JR & Marra MA (2015) Driver and passenger mutations in cancer. *Annu Rev Pathol* **10:** 25–50

Poon SL, Pang S-T, McPherson JR, Yu W, Huang KK, Guan P, Weng W-H, Siew EY, Liu Y, Heng HL, Chong SC, Gan A, Tay ST, Lim WK, Cutcutache I, Huang D, Ler LD, Nairismägi M-L, Lee MH, Chang Y-H, et al (2013) Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci Transl Med* **5:** 197ra101–197ra101

Povirk LF, Zhou T, Zhou R, Cowan MJ & Yannone SM (2007) Processing of 3'-phosphoglycolate-terminated DNA double strand breaks by Artemis nuclease. *J. Biol. Chem.* **282:** 3547–3558

Puddu F, Oelschlaegel T, Guerini I, Geisler NJ, Niu H, Herzog M, Salguero I, Ochoa-Montaño B, Viré E, Sung P, Adams DJ, Keane TM & Jackson SP (2015) Synthetic viability genomic screening defines Sae2 function in DNA repair. *EMBO J.* **34:** 1509–1522

Pukkila PJ, Peterson J, Herman G, Modrich P & Meselson M (1983) Effects of high levels of DNA adenine methylation on methyl-directed mismatch repair in Escherichia coli. *Genetics* **104:** 571–582

Raghunandan M, Chaudhury I, Kelich SL, Hanenberg H & Sobeck A (2015) FANCD2, FANCJ and BRCA2 cooperate to promote replication fork recovery independently of the Fanconi Anemia core complex. *Cell Cycle* **14:** 342–353

Roche-Lestienne C, Soenen-Cornu V, Grardel-Duflos N, Laï J-L, Philippe N, Facon T, Fenaux P & Preudhomme C (2002) Several types of mutations of the Abl gene can be found in chronic myeloid leukemia patients resistant to STI571, and they can pre-exist to the onset of treatment. *Blood* **100:** 1014–1018

Rong YS & Golic KG (2003) The Homologous Chromosome Is an Effective Template for the Repair of Mitotic DNA Double-Strand Breaks in Drosophila. *Genetics* **165:** 1831–1842

Roos WP & Kaina B (2006) DNA damage-induced cell death by apoptosis. *Trends Mol Med* **12:** 440–450

Roos WP & Kaina B (2013) DNA damage-induced cell death: from specific DNA lesions to the DNA damage response and apoptosis. *Cancer Lett.* **332:** 237–248

Rusch HP & Baumann CA (1939) Tumor Production in Mice with Ultraviolet Irradiation. *The American Journal of Cancer* **35:** 55–62

Russo MT, Blasi MF, Chiera F, Fortini P, Degan P, Macpherson P, Furuichi M, Nakabeppu Y, Karran P, Aquilina G & Bignami M (2003) The Oxidized Deoxynucleoside Triphosphate Pool Is a Significant Contributor to Genetic Instability in Mismatch Repair-Deficient Cells. *Molecular and Cellular Biology* **24:** 465–474

Rübben A & Nordhoff O (2013) A systems approach defining constraints of the genome architecture on lineage selection and evolvability during somatic cancer evolution. *Biol Open* **2:** 49–62

Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A & López-Bigas N (2016) Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532:** 264–267

Saini N, Roberts SA, Klimczak LJ, Chan K, Grimm SA, Dai S, Fargo DC, Boyer JC, Kaufmann WK, Taylor JA, Lee E, Cortes-Ciriano I, Park PJ, Schurman SH, Malc EP, Mieczkowski PA & Gordenin DA (2016) The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts. *PLoS Genet.* **12:** e1006385

Sanger F & Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94:** 441–448

Sanger F, Nicklen S & Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74:** 5463–5467

Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, Bajorek E, Black S, Chan YM, Denys M, Escobar J, Flowers D, Fotopulos D, Garcia C, Gomez M, Gonzales E, Haydu L, Lopez F, Ramirez L, Retterer J, et al (2004) Quality assessment of the human genome sequence. *Nature* **429:** 365–368

Setlow RB & Carrier WL (1966) Pyrimidine dimers in ultraviolet-irradiated DNA's. *J. Mol. Biol.* **17:** 237–254

Sfeir A & de Lange T (2012) Removal of shelterin reveals the telomere end-protection problem. *Science* **336:** 593–597

Shabalina SA, Ogurtsov AY, Kondrashov VA & Kondrashov AS (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17:** 373–376

Shar NA, Vijayabaskar MS & Westhead DR (2016) Cancer somatic mutations cluster in a subset of regulatory sites predicted from the ENCODE data. *Mol. Cancer* **15:** 76

Sharma A, Singh K & Almasan A (2012) Histone H2AX phosphorylation: a marker for DNA

damage. *Methods Mol. Biol.* **920:** 613–626

Shaw AT, Kim D-W, Nakagawa K, Seto T, Crinó L, Ahn M-J, De Pas T, Besse B, Solomon BJ, Blackhall F, Wu Y-L, Thomas M, O'Byrne KJ, Moro-Sibilot D, Camidge DR, Mok T, Hirsh V, Riely GJ, Iyer S, Tassell V, et al (2013) Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N. Engl. J. Med.* **368:** 2385–2394

Shen J, Ju Z, Zhao W, Wang L, Peng Y, Ge Z, Nagel ZD, Zou J, Wang C, Kapoor P, Ma X, Ma D, Liang J, Song S, Liu J, Samson LD, Ajani JA, Li G-M, Liang H, Shen X, et al (2018) ARID1A deficiency promotes mutability and potentiates therapeutic antitumor immunity unleashed by immune checkpoint blockade. *Nat. Med.* **24:** 556–562

Shiraishi A (2000) Increased susceptibility to chemotherapeutic alkylating agents of mice deficient in DNA repair methyltransferase. *Carcinogenesis* **21:** 1879–1883

Sidransky D, Eschenbach Von A, Tsai YC, Jones P, Summerhayes I, Marshall F, Paul M, Green P, Hamilton SR & Frost P (1991) Identification of p53 gene mutations in bladder cancers and urine samples. *Science* **252:** 706–709

Sonntag von C (2006) Ionizing Radiation Damage to DNA Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM & Sunyaev SR (2009) Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41:** 393–395

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin M-L, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, et al (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144:** 27–40

Stephens PJ, McBride DJ, Lin M-L, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, Greenman CD, Jia M, Latimer C, Teague JW, Lau KW, Burton J, Quail MA, Swerdlow H, Churcher C, Natrajan R, et al (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462:** 1005–1010

Stone MP, Cho Y-J, Huang H, Kim H-Y, Kozekov ID, Kozekova A, Wang H, Minko IG, Lloyd RS, Harris TM & Rizzo CJ (2008) Interstrand DNA cross-links induced by alpha,beta-unsaturated aldehydes derived from lipid peroxidation and environmental sources. *Acc. Chem. Res.* **41:** 793–804

Stratton MR, Campbell PJ & Futreal PA (2009) The cancer genome. *Nature* **458:** 719–724

Sumpter R Jr., Sirasanagandla S, Fernández ÁF, Wei Y, Dong X, Franco L, Zou Z, Marchal C, Lee MY, Clapp DW, Hanenberg H & Levine B (2016) Fanconi Anemia Proteins Function in Mitophagy and Immunity. *Cell* **165:** 867–881

Supek F & Lehner B (2015) Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521:** 81–84

Supek F & Lehner B (2017) Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell* **170:** 534–547.e23

Supek F, Miñana B, Valcárcel J, Gabaldón T & Lehner B (2014) Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156:** 1324–1335

Taylor EM & Lehmann AR (1998) Conservation of eukaryotic DNA repair mechanisms. *Int. J. Radiat. Biol.* **74:** 277–286

Tessman I, Poddar RK & Kumar S (1964) Identification of the altered bases in mutated single-stranded DNA: I. In vitro mutagenesis by hydroxylamine, ethyl methanesulfonate and nitrous acid. *J. Mol. Biol.* **9:** 352–363

The Philadelphia chromosome: a mutant gene and the quest to cure cancer at the genetic level (2013) The Philadelphia chromosome: a mutant gene and the quest to cure cancer at the genetic level. *Choice Reviews Online* **51:** 51–0924–51–0924

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

Tubbs A & Nussenzweig A (2017) Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168:** 644–656

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, et al (2001) The sequence of the human genome. *Science* **291:** 1304–1351

Viel A, Bruselles A, Meccia E, Fornasarig M, Quaia M, Canzonieri V, Policicchio E, Urso ED, Agostini M, Genuardi M, Lucci-Cordisco E, Venesio T, Martayan A, Diodoro MG, Sanchez-Mete L, Stigliano V, Mazzei F, Grasso F, Giuliani A, Baiocchi M, et al (2017) A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* **20:** 39–49

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA & Kinzler KW (2013) Cancer genome landscapes. *Science* **339:** 1546–1558

Ward RL, Dobbins T, Lindor NM, Rapkins RW & Hitchins MP (2013) Identification of constitutional MLH1 epimutations and promoter variants in colorectal cancer patients from the Colon Cancer Family Registry. *Genetics in Medicine* **15:** 25–35

Watanabe Y, Fujiyama A, Ichiba Y, Hattori M, Yada T, Sakaki Y & Ikemura T (2002) Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Hum. Mol. Genet.* **11:** 13–21

Wellcome Trust Sanger Institute COSMIC, Catalog of Somatic Mutations in Cancer - Signatures of Mutational Processes in Human Cancer http://cancer.sanger.ac.uk/cosmic/signatures Available at: http://cancer.sanger.ac.uk/cosmic/signatures

Witkin EM (1969) Ultraviolet-induced mutation and DNA repair. *Annu. Rev. Microbiol.* **23:** 487–514

Wong KM, Hudson TJ & McPherson JD (2011) Unraveling the genetics of cancer: genome sequencing and beyond. *Annu Rev Genomics Hum Genet* **12:** 407–430

Wong SQ, Li J, Tan AY-C, Vedururu R, Pang J-MB, Do H, Ellul J, Doig K, Bell A, MacArthur GA, Fox SB, Thomas DM, Fellowes A, Parisot JP, Dobrovic ACANCER 2015 Cohort

(2014) Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. *BMC Med Genomics* **7:** 23

Woodbine L, Brunton H, acids AGN2011 Endogenously induced DNA double strand breaks arise in heterochromatic DNA regions and require ataxia telangiectasia mutated and Artemis for their repair. *academic.oup.com*

Xie Y, Yang H, Cunanan C, Okamoto K, Shibata D, Pan J, Barnes DE, Lindahl T, McIlhatton M, Fishel R & Miller JH (2004) Deficiencies in Mouse Myhand Ogg1Result in Tumor Predisposition and G to T Mutations in Codon 12 of the K-RasOncogene in Lung Tumors. *Cancer Res.* **64:** 3096–3102

Xu Z, Zan H, Pone EJ, Mai T & Casali P (2012) Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. *Nat. Rev. Immunol.* **12:** 517–531

Yamamoto K, Wang Y, Jiang W, Liu X, Dubois RL, Lin C-S, Ludwig T, Bakkenist CJ & Zha S (2012) Kinase-dead ATM protein causes genomic instability and early embryonic lethality in mice. *The Journal of Cell Biology* **198:** 305–313

Yao CJ, Du W, Zhang Q, Zhang F, Zeng F & Chen FP (2013) Fanconi anemia pathway--the way of DNA interstrand cross-link repair. *Pharmazie* **68:** 5–11

Yi C & He C (2013) DNA repair by reversal of DNA damage. *Cold Spring Harb Perspect Biol* **5:** a012575–a012575

Curriculum Vitae

Personal information

Family & first name: B.-B. Owusu, Michel
E-mail: mowusu@cemm.oeaw.ac.at
Nationality: Austrian

Academic career

2013 – present, PhD student at the Medical University of Vienna.
Oct. 2011 – Nov. 2013, MSc. in Molecular Medicine at the University of Vienna.
Sep. 2008 – Jun. 2011, BSc. in Molecular Biotechnology at the University of Applied Sciences FH-Campus.
Mar. 2008 – Jun. 2008, Mathematics at the University of Vienna (30 ECTS, average score: excellent).
1996 – 2005, High School diploma / Matura in Vienna.

Work experience in scientific research

Oct. 2013 – present, PhD at CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences. Supervisor: Dr. Joanna Loizou.
May. 2012 – Mar. 2013, Master's internship at CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences. Supervisor: Dr. Joanna Loizou.
Sep. 2009 – Aug. 2011 (6 months interruption due to Erasmus internship), Part time job as a technical assistant at the Institute for Molecular Biotechnology of the Austrian Academy of Sciences (IMBA). Supervisor: Dr. Leonie Ringrose.
Jan. – May. 2011 (Erasmus internship), Bachelor's internship at the Institut de Genetique Humaine / CNRS in Montpellier, France. Supervisor: Dr. Jean-Maurice Dura.
Sep. – Oct. 2010, Internship at IMBA, Supervisor: Dr.Ringrose.
Aug. – Sep. 2009, Internship at the Medical University of Vienna (MUW) - Vascular Biology and Thrombosis Research. Supervisor: Prof. Bernd Binder.
Feb. 2009, Internship at the Children's Cancer Research Institute in Vienna. Supervisor: Dr. Thomas Felzmann.

Publications

Mapping the non-essential kinome upon chemotherapy induced DNA lesions. **Michel Owusu**, Peter Bannauer, Joana Ferreira da Silva , Marc Wiedner, Charles H. Lardeau, Stefan Kubicek, Francesca Ciccarelli, Joanna I. Loizou. (under revision in Cell Reports, CELL-REPORTS-D-18-01087 and published on bioRviv: bioRxiv 385344; doi: https://doi.org/10.1101/385344)

Map of synthetic rescue interactions for the Fanconi anemia DNA repair pathway identifies USP48. Georgia Velimezi#, Lydia Robinson-Garcia#, Francisco Muñoz-Martinez, Wouter W. Wiegant, Joana Ferreira da Silva, **Michel Owusu**, Martin Moder, Marc Wiedner, Sara Brin Rosenthal, Kathleen M. Fisch, Jason Moffat, Jörg Menche, Haico Van Attikum, Stephen P. Jackson and Joanna I. Loizou. Nat Commun. 2018 Jun 11;9(1):2280. doi: 10.1038/s41467-018-04649-z. (#contributed equally)

Validating the concept of mutational signatures with isogenic cell models. Xueqing Zou#, **Michel Owusu#**, Rebecca Harris, Stephen P. Jackson, Joanna Loizou*, Serena Nik-Zainal* (#contributed equally, *corresponding authors. Nat Commun. 2018 May 1;9(1):1744. doi: 10.1038/s41467-018-04052-8.

Repair of UV-induced DNA Damage independent of nucleotide excision repair is masked by MUTYH. Abdelghani Mazouzi, Federica Battistini, Sarah C. Moser, Joana Ferreira da Silva, Marc Wiedner, **Michel Owusu**, Charles-Hugues Lardeau, Anna Ringler, Beatrix Weil, Jürgen Neesen, Modesto Orozco, Stefan Kubicek and Joanna I. Loizou. Mol Cell. 2017 Nov 16;68(4):797-807.e7. doi: 10.1016/j.molcel.2017.10.021

Parallel genome-wide screens identify synthetic viable interactions between the BLM helicase complex and Fanconi anemia: Martin Moder#, Georgia Velimezi#, **Michel Owusu**, Abdelghani Mazouzi, Marc Wiedner, Joana Ferreira da Silva, Lydia Robinson-Garcia, Fiorella Schischlik, Rastislav Slavkovsky, Robert Kralovics, Michael Schuster, Christoph Bock, Trey Ideker, Stephen P. Jackson, Jörg Menche, Joanna I. Loizou. Nat Commun. 2017 Nov 1;8(1):1238. doi: 10.1038/s41467-017-01439-x. (#contributed equally)

DNA Repair Cofactors ATMIN and NBS1 Are Required to Suppress T Cell Activation. Jana Prochazkova, Shinya Sakaguchi, **Michel Owusu**, Abdelghani Mazouzi, Marc Wiedner, Georgia Velimezi, Martin Moder, Gleb Turchinovich, Anastasiya Hladik, Elisabeth Gurnhofer, Adrian Hayday, Axel Behrens, Sylvia Knapp, Lukas Kenner, Wilfried Ellmeier, Joanna I. Loizou. PLoS Genet. 2015 Nov 6;11(11):e1005645. doi: 10.1371/journal.pgen.1005645. eCollection 2015 Nov.

ATMIN is required for the ATM-mediated signaling and recruitment of 53BP1 to DNA damage sites upon replication stress. Luisa Schmidt, Marc Wiedner, Georgia Velimezi, Jana Prochazkova, **Michel Owusu**, Sabine Bauer, Joanna I. Loizou.

DNA Repair (Amst). 2014 Dec;24:122-30. doi: 10.1016/j.dnarep.2014.09.001. Epub 2014 Sep 26.

Conference presentations

PhD-Symposium 2014 of the Young Scientists Association of the Medical University of Vienna, Vienna, Austria: Mapping the kinome in response to DNA damage. Participation and poster presentation.
30th Ernst Klenk Symposium in Molecular Medicine, DNA Damage Response and Repair Mechanisms in Aging and Disease, Cologne, Germany: Mapping the kinome in response to DNA damage. Participation and poster presentation.

Languages

Trilingual competence:

- Native / educational language: German
- Additional language: English (full professional proficiency)
- Mother language: Ghanaian (proficient)

Other languages: French (independent), Spanish (beginner)