The relationship between DNA modifications and mutations in cancer

Markéta Tomková

St Catherine's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2017

The relationship between DNA modifications and mutations in cancer

Markéta Tomková
St Catherine's College
Doctor of Philosophy
Trinity 2017

Abstract

Somatic mutations are the main triggers that initiate the formation of cancer. Large sequencing data sets in recent years revealed a substantial number of mutational processes, many of which are poorly understood or of completely unknown aetiology. These mutational processes leave characteristic sequence patterns, often called "signatures", in the DNA. Characterisation of the mutational patterns observed in cancer patients with respect to different genomic features and processes can help to unravel the aetiology and mechanisms of mutagenesis. Here, we explored the effects of DNA modifications and DNA replication on mutagenesis.

The most common mutation type, C>T mutations in a CpG context, is thought to result from spontaneous deamination of 5-methylcytosine (5mC), the major DNA modification. Much less is known about the mutational properties of the second most frequent modification, 5-hydroxymethylcytosine (5hmC). Integrating multiple genomic data sets, we demonstrate a twofold lower mutagenicity of 5hmC compared to 5mC, present across multiple tissues.

Subsequently, we show how DNA modifications may modulate various mutational processes. In addition to spontaneous deamination of 5mC, our analysis suggests a key role of replication in CpG>TpG mutagenesis in patients deficient in post-replicative proofreading or repair, and possibly also in other cancer patients. Together with an analysis of mutation patterns observed in cancers exposed to UV light, tobacco smoke, or editing by APOBEC enzymes, the results show that the role of DNA modifications goes beyond the well-known spontaneous deamination of 5mC.

Finally, we explored which of the known mutational processes might be modulated by DNA replication. We developed a novel method to quantify the magnitude of strand asymmetry of different mutational signatures in individual patients followed by evaluation of these exposures in early and late replicating regions. More than 75 % of mutational signatures exhibited a significant replication strand asymmetry or correlation with replication timing. The analysis gives new insights into mechanisms of mutagenicity in multiple signatures, particularly the so far enigmatic signature 17, where we suggest an involvement of oxidative damage in its aetiology. In conclusion, our results suggest that DNA replication or replication-associated DNA repair interacts with most mutagenic processes.

This thesis is submitted to the Nuffield Department of Medicine, University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise stated, describes my own research.

Word count: 49 721

Markéta Tomková, St Catherine's College

Those who fought for something better
Those who taught by how they lived
Loved ones taken long before their work was done

— Tripod and Austin Wintory; performed by
Peter Hollens *Underground*

This thesis is dedicated to

my dad Honza,

my grandma Miluška,

and my dance teacher Hanička.

Even the little children that no one cares for
He called up everyone
And he gave us fruit
And treated everyone

— Culture *Why Am I A Rastaman?*

Acknowledgements

I would like to thank my two amazing supervisors Dr Benjamin Schuster-Böckler and Professor Skirmantas Kriaučionis for introducing me into this beautiful world of science, for guiding me throughout the more and the less successful parts of the doctorate, for being frank in their comments, for creating such a friendly atmosphere in their labs, and for being enthusiastic about the actual science (as opposed to politics of academia). I am especially grateful for the combination of freedom in explorations and —at the same time— detailed and insightful guidance and feedback that I was given. I also would also like to thank Professor Xin Lu for her high-levels guidance and co-supervision in the first years of my doctorate.

I thank Dr Michael McClellan for his experimental results, which complement my bioinformatics observations, and I am very grateful for all his work on upgrading some of the bioinformatics results to the next level in the lab. I also thank him for all the discussions, for his scientific geekiness, and his enthusiasm. I thank all other members of Skirmantas' lab: Anandhakumar, Hiromi, Martin, Paolo, Sophie, Wenjun, and Ying, for the scientific and other discussions and for being so friendly. I thank David for ideas and discussions early in the project, Andy for his help with Bayesian inference and other bioinformatics methods, Gergana and other Ludwig students, who made my doctorate in Ludwig very enjoyable. I thank the researchers in whose projects I could be involved, in particular Professor Ian Tomlinson, Dr Chiara Bardella, Richard Owen, Michael White, Dr Benjamin Fairfax, and Dr Annabelle Lewis, with whom I have learned much new. I thank Jakub, Gergana, and Michael for proofreading of the text. I thank my Oxford friends, DTCers, neighbours, dancers, and musicians for all the fun, energy and for connecting me with the non-scientific reality. I thank my Prague dancers for their remote support and friendship. I thank the Bakala foundation, EPSRC, my college, and Ludwig Cancer Research for funding my studies. I thank the consortia and authors, who made their data available for me to use them in my thesis; and I thank all the enlightened people who initiated the data sharing in the cancer genomics field.

I thank my wonderful and loving family, who have always been supporting me in all my life decisions. I especially thank my mum that she endures listening to my passionate and chaotic explanations of my projects during our night calls, for the enthusiasm we can share in our scientific collaborations, and that she always charges me with a positive energy. I thank my dear brother for inspiring me and encouraging

me into computer science and research. I also thank my admirable grandparents, who initiated our Oxford adventure, and with whom we had so many beautiful discussions.

Finally, and most importantly, my greatest thanks belong to Jakub, my husband. It is thanks to him that I am in Oxford, that I could do and finish my doctorate, that we could both immerse into the scientific adventures, that he makes sure that we do other activities than work, that I enjoy every day in my life, and that we could spend together so many beautiful evenings and nights working on the projects, drinking tea, and listening to music. Some of the music pieces which contributed to this work (by creating a productive working environment), or which have links to the work, are shared with you in the headers where normal people put quotes of wisdom.

Contents

Peace of the unseen
Peace of the spirit
Peace of Iona

— The Waterboys *Peace Of Iona*

Ahrk fin norok paal graan
fod nust hon zindro zaan
Dovahkiin, fah hin kogaan mu draal!

— Jeremy Soule *Skyrim*

1 Introduction

In the last century, several theories have been formed about the cause and origin of cancer (e.g., reviewed in Vineis et al., 2010). The two major components of these theories are mutagenesis and epigenetics. Mutations are permanent changes to the DNA. Epigenetics can be described as the study of heritable information carried by other means than the sequence of the DNA bases. For example, this can be carried by small chemical modifications of the DNA bases (DNA modifications). Although DNA modifications are an indispensable component of living cells, they are also the cause of the most common type of mutations. This mutagenic process has been known for more than 40 years (Lindahl and Nyberg, 1974). However, the role of other, recently discovered, DNA modifications in mutagenesis is almost unexplored. Moreover, DNA modifications can interact also with other known mutational processes, but the effects of these interactions on mutations observed in whole-genomes of cancer patients are largely unknown. The interplay between DNA modifications and mutagenesis is therefore one of the two main topics of this thesis.

Another important source of mutations is replication of DNA performed before each cell division. Mutations can be introduced during replication due to random errors made by DNA polymerases. The knowledge of such source of mutations is very old, but also recently actively discussed in the cancer scientific community (Tomasetti and Vogelstein, 2015). Several mechanisms exist in cells to repair errors introduced during replication.

When all these mechanisms are intact, the entire process is remarkably accurate, making only one error in every 10^{9-10} bases (Rayner et al., 2016). The estimated error-rate was thought to be too low to account for the mutation load observed in cancer genomes, diminishing the role of replication in cancer mutagenesis (Loeb, 1991). However, other mutational processes were also shown to be linked to replication. Nevertheless, it is currently unknown which of the mutational processes interact with replication and how. The interplay between DNA replication and mutagenesis is therefore the second main topic of this thesis.

This chapter first provides a brief introduction into the necessary basics of genomics (section 1.1) and DNA replication (section 1.1.3), followed by an overview of epigenomics with a main focus on DNA modifications (section 1.2), leading into a section about DNA mutagenesis and repair (1.3). Sections 1.4 and 1.5 summarise previous research about the influence of DNA modifications and DNA replication on DNA mutagenesis. The chapter is concluded with the aims of the thesis (section 1.6).

1.1 Genomics: DNA, transcription, and replication

1.1.1 DNA

The hereditary information of cells is stored in molecules of deoxyribonucleic acid (DNA), composed of monomers called nucleotides (Fig. 1.1). Each DNA nucleotide consists of a sugar 2'-deoxyribose, phospate group, and a base. The sugars are joined by the phospate groups to form a polymer chain called DNA backbone. The hereditary information is encoded into the sequence of bases, which are connected to the backbone by the sugars. Four types of bases exist in the DNA; two purines: adenine (A) and guanine (G), and two pyrimidines: cytosine (C) and thymine (T). The sequence of bases is connected with the phosphate-deoxyribose backbone into a directional DNA strand. The terminology for the strand direction is based on the two ends of the molecule: *5' end* (which contains a phospate group attached to the 5' carbon of the sugar ring) and *3' end* (with a hydroxyl group attached to the 3' carbon of the sugar ring) (Fig. 1.1). DNA stays most of the time in the form of two antiparallel strands. The bases of the two strands are non-covalently connected by hydrogen bonds, pairing complementary

bases together, such that A is always opposite T (A:T pair) and C is opposite G (C:G pair). The two strands are coiled around a common axis, forming a double helix. The structure of DNA was discovered in 1953 (Watson and Crick, 1974, 1953; Franklin and Gosling, 1953) and names of Watson and Crick are also used to distinguish the two strands: the 5' to 3' "top" strand is sometimes called Watson strand and is used as a reference, whereas Crick strand refers to the opposite 5' to 3' bottom strand.

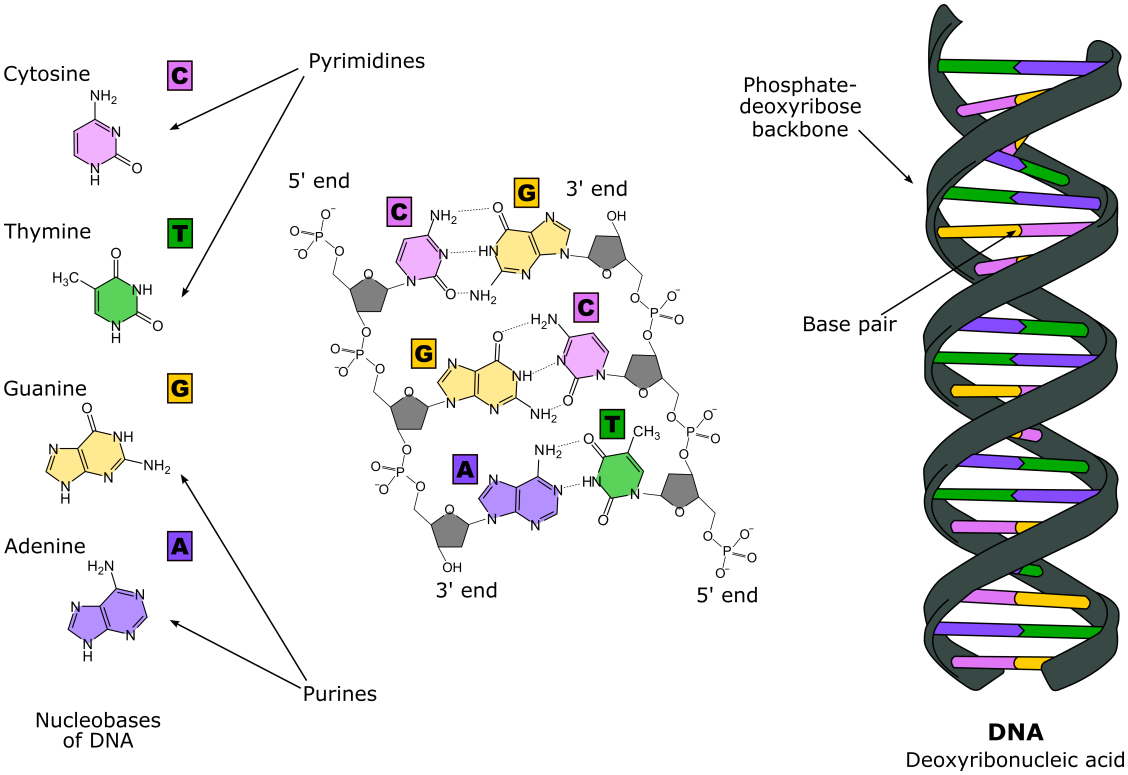

Figure 1.1. DNA Four bases in the DNA (left) follow given base-pairing rules (C:G and A:T) and are connected to the phospate-deoxyribose backbone (2'-deoxyribose is shown in grey) (middle), to form a double-helix structure (right). This figure was derived from Difference_DNA_RNA-EN.svg by Roland1952, licensed under CC-BY-SA.

In order to package the 2 metres of human nuclear DNA into a nucleus of a cell with an average diameter of less than 10 µm, it needs to be compacted on several levels (McGinty and Tan, 2015). On the lowest level, the DNA double helix is coiled around eight histone protein cores called nucleosomes, first observed in 1974 (Olins and Olins, 1974), and crystallised in 1997 (Luger et al., 1997) (Fig. 1.2).

Approximately 147 base-pairs of DNA (ca. 1.65 turns) are wrapped around the histone octamer, which consists of two copies of each histone: H2A, H2B, H3, and H4.

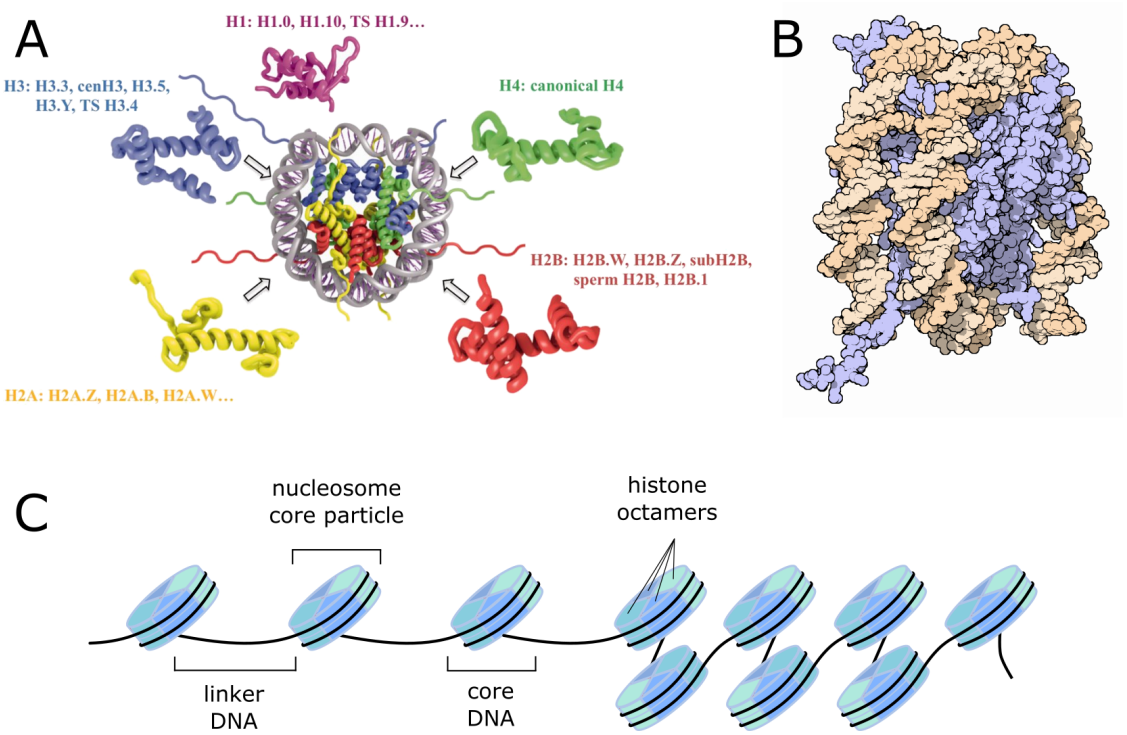

Figure 1.2. Nucleosome structure. A: Schematic representation of nucleosome structure and its composition of octamer histones (H3, H4, H2A and H2B) and H1 linker histone. Selected histone variant names for each histone type are shown. Reprinted from (Draizen et al., 2016), with permission from the publisher. **B:** Nucleosome structure, with DNA in orange and histone proteins in blue. Reprinted from Molecule of the Month by David S. Goodsell and the RCSB PDB, licensed under CC-BY-4.0. **C:** The nucleosome core is formed by histones and 147 bp of core DNA, while adjacent nucleosomes are separated by stretches of linker DNA of varying length up to about 100 bp.

The higher-order structure is stabilised by linker histone H1. The DNA between the nucleosome cores is called *linker DNA* and can be of variable length (typically between 10 and 80 bp). The centre of the nucleosome/DNA wrapped around the nucleosome is called *nucleosome dyad*. The locations of nucleosomes on the human DNA are to some extent shared among the cells and two terms are used to describe this similarity (Struhl and Segal, 2013). *Nucleosome occupancy* is defined as the fraction of cells from the population in which the base pair is occupied by any histone octamer. *Nucleosome positioning* in a base pair in the genome is defined as the fraction of cells from the population in which that base pair is at the nucleosome dyad.

The DNA wrapped around nucleosomes is further compacted to form higher-order chromatin structures, which are tightly coiled into the chromatids of chromosomes.

The complex of DNA and histone proteins is called chromatin and depending on the degree of condensation it can be found in two forms: euchromatin and heterochromatin. Euchromatin (open chromatin; gene rich and associated with active transcription) is less condensed, potentially allowing better access of proteins related to transcription or DNA repair, but also more exposed to DNA damage, whereas heterochromatin (closed chromatin; associated with inactive genes) stays highly condensed throughout the cell cycle (Margueron and Reinberg, 2010).

1.1.2 Transcription

The hereditary information for creating proteins is encoded in genes in the DNA. In eukaryotes, the sequence of bases in the gene body (between transcription start site (TSS) and transcription end site (TES)) is transcribed to a single-stranded molecule of ribonucleic acid (RNA) of type precursor messenger RNA (pre-mRNA). In contrast to DNA, RNA contains sugar ribose, the base uracil is used instead of thymine, and is mostly single-stranded, albeit folded in a three-dimensional structure with stretches of nucleotides paired with complementary sequences from different parts of the same molecule.

The principle of complementarity is used when RNA polymerase (e.g., the human RNA Pol II) copies bases from the DNA to the RNA. Transcription is initiated when transcription activators bind *promoter* region near TSS to attract RNA Pol II. The transcription initiation can be enabled by regulatory regions called *enhancers*, which can be located thousands of nucleotides away from TSS and provide binding sites for gene regulatory proteins. While the pre-mRNA molecule is being produced, it is concurrently processed by having a modified guanine nucleotide cap added to the 5' end, by removal of non-coding regions called introns through a process of pre-mRNA splicing, and by polyadenylation of the 3' end of pre-mRNA. The processed molecule is termed mRNA and is transported from the nucleus to the cytosol, where it is translated to a protein.

1.1.3 DNA replication

Before a cell divides, its DNA needs to be replicated. The replication of DNA is again based on the complementarity of bases in the DNA. The two strands are separated and serve as a template for synthesis of a new strand, creating two double helix copies of the original molecule. In eukaryotes, DNA replication is initiated in replication origins (ORI), simultaneously in many places in the genome. The recognition of ORI by origin recognition complex (ORC) in eukaryotes is still not fully understood but is thought to be defined by a wide variety of features and goes beyond a simple DNA sequence motif (Aladjem and Redon, 2016; Snedeker et al., 2017). In prokaryotes, ORI share similar sequence motifs (Fuller et al., 1984; Leonard and Méchali, 2013). In eukaryotes, regions with ORI were found enriched with CpG islands[1], GC-rich regions, G-rich repeats and motifs associated with G quadruplexes, four stranded helical structures of DNA (Leonard and Méchali, 2013; Cayrou et al., 2011; Besnard et al., 2012). However, some of the enrichments might be an artefact of the techniques used for ORI detection. For instance, techniques based on lambda exonuclease λ-exo digestion show a technical bias for GC-rich DNA and G4 motifs; and after a control for these biases using nonreplicating genomic DNA, the association between ORI and G quadruplex motifs and G+C content is markedly diminished (Foulk et al., 2015).

The actively used ("activated") ORI differ between cell types, age of the cell and other factors (Fragkos et al., 2015). Only a small proportion (e.g., 10–20%) of potential ORI are activated in a cell and their numbers are estimated to be in the order of tens of thousands (Huberman and Riggs, 1968; Méchali, 2010; Besnard et al., 2012; Leonard and Méchali, 2013; Langley et al., 2016). The distribution of ORI is not uniform, with clusters of early-firing ORI separated from late-firing ORI by ORI-poor temporal transition regions (TTR) (Desprat et al., 2009; Cayrou et al., 2011; Leonard and Méchali, 2013).

The replication proceeds in both directions from ORI, unwinding the parental strands with Cdc45, MCM2–7, and GINS (CMG) complex (Burgers and Kunkel, 2017) (Fig. 1.3A). The generated strands are coated with replication protein A (RPA) to keep them

[1]CpG refers to a cytosine-phoshate-guanine dinucleotide (in the 5'-to-3' direction). CpG islands are regions with a relatively high frequency of CpG dinucleotides, often found in promoter regions close to TSS, important for the regulation of gene transcription. See section 1.2.1 for more details.

stabilised and single-stranded (Burgers and Kunkel, 2017). A term *replication fork* is used for the actively replicated region, due to its Y-shape structure. The synthesis of daughter strands is performed by replicative DNA polymerases. Crucially, the known eukaryotic replicative DNA polymerases work directionally, synthesising the daughter strand from the 5' end to the 3' end (i.e., reading the template in the 3' to 5' direction). Only one of the strands can therefore be synthesised in a continuous fashion. This is called the *leading strand*, its template is the top (Watson) strand to the left from the ORI and the bottom (Crick) strand to the right from the ORI (Fig. 1.3B). On the other hand, the *lagging strand* is synthesised discontinuously in ca. 100–200 nucleotides long pieces of DNA called *Okazaki fragments*, which are then joined together into a continuous strand. The synthesis of each leading strand and each Okazaki fragment is initiated by Pol α-primase complex, which creates a 5–10 nucleotides long RNA primer extended with ca. 30 nucleotides of DNA (Stillman, 2008; Pellegrini, 2012; Burgers and Kunkel, 2017).

In the most accepted model of the eukaryotic replication fork, the bulk of the leading strand is synthesised by replicative polymerase ε (Pol ε) and the lagging strand by replicative polymerase δ (Pol δ) (Stillman, 2008; Mertz et al., 2017b; Burgers and Kunkel, 2017; Snedeker et al., 2017) (Fig. 1.3). The processivity[2] of Pol δ and to a lesser extent also Pol ε is enhanced by proliferating cell nuclear antigen (PCNA), a ring-shaped clamp, which encircles the DNA, tethers the replicative polymerases to the DNA, and by interacting with a number of other proteins coordinates different sub-processes involved in the replication (Moldovan et al., 2007; Burgers and Kunkel, 2017). On the lagging strand, the synthesis of each Okazaki fragment is initiated by RPA-recruited Pol α-primase complex and followed by elongation by Pol δ. At the end of the Okazaki fragment (called *Okazaki junction*), Pol δ carries out strand displacement of the RNA primers and Pol α-synthesised DNA, generating a nascent flap, which is cut by FEN1, allowing ligation of the Okazaki junction (Stith et al., 2008; Burgers and Kunkel, 2017) (Fig. 1.3C). Recent evidence from yeast shows that not all Pol α-synthesised DNA is removed, leading to approximately 1.5 % of the mature genome resulting from Pol α synthesis, possibly due to DNA-binding proteins blocking the

[2]Processivity of an enzyme is its ability to catalyse consecutive reactions without releasing its substrate.

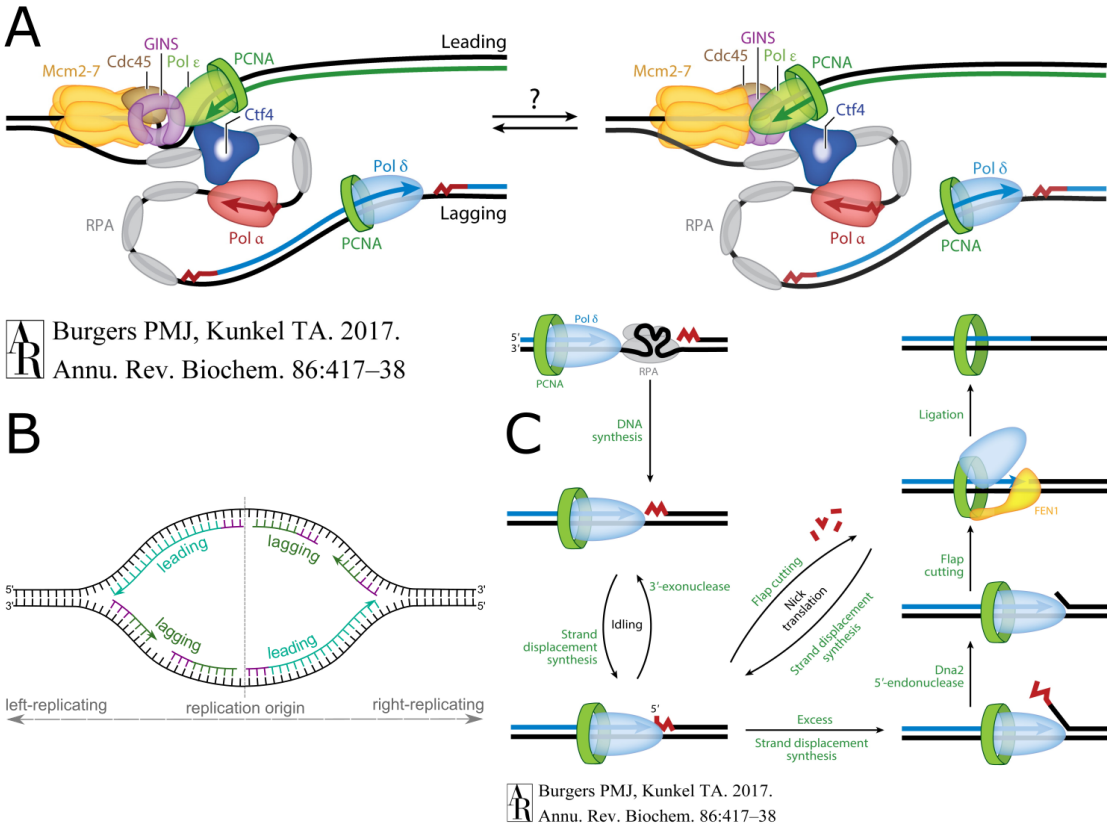

Figure 1.3. A model of eukaryotic replication. A: Replisome structure and interactions. Parental strands are separated by CMG complex (consisting of Cdc45, MCM2–7, and GINS) and the single-stranded DNA is coated with RPA. The leading strand is primed by Pol α/Primase complex and is synthesised by Pol ε, with only ~40-nt (left) or ~20-nt (right) lengths of single-stranded DNA in between CMG and the polymerase. The lagging strand is shown looped such that both Pol α and Pol ε move in the same direction while held in a complex by Ctf4. **B:** A schematic representation of the replication origin and DNA synthesis proceeding in both directions from the origin. **C:** Okazaki fragment maturation. Primers on the lagging strand are elongated by Pol δ until the downstream Okazaki fragment is reached. Subsequent strand displacement synthesis by Pol δ is counteracted by its 3'-exonuclease activity. In the presence of FEN1, the nascent flap is cut and strand displacement synthesis restarts. This iterative process predominantly releases mononucleotides. Occasional excess strand displacement synthesis yields very long 5'-flaps that are processed to short flaps by the nuclease activity of Dna2. After degradation of all primer RNA, ligation of the DNA–DNA nick is performed by DNA ligase 1. Figures (A) and (C) are reprinted from Burgers and Kunkel (2017), with permission from the publisher.

displacement by Pol δ (Reijns et al., 2015). The length of the Okazaki fragments is determined by nucleosome periodicity and Okazaki junctions preferentially occur near nucleosome midpoints (dyads), rather than in internucleosomal linker regions (Smith and Whitehouse, 2012; Williams et al., 2016).

The model of division of work between Pol ε and Pol δ on the leading and lagging

strands, respectively, has been recently challenged by Johnson et al. (2015), who proposed Pol δ to be the major polymerase for both the leading and the lagging strands in *Schizosaccharomyces pombe*, due to observed increase of mutation rate on both strands in strains with mutated Pol δ and defective in MMR (*pol3-L612M msh2* Δ). In this model, Pol ε is not involved in the synthesis of the leading strand, but might be involved in proofreading and correcting errors made by Pol δ. This alternative model has been actively discussed in the community; however it is disfavoured in normal undamaged DNA (Burgers et al., 2016; Lujan et al., 2016) due to multiple reasons: contradictory observations of mutation spectra in yeast strains with mutated Pol δ and Pol ε (summarised in Burgers et al., 2016), asymmetric ribonucleotide incorporation by Pol δ and Pol ε (summarised in Burgers et al., 2016), inability of Pol ε to proofread mistakes made by Pol δ (Flood et al., 2015), and enrichment of mutations on the leading strand in Pol ε exonuclease-defective human tumours (Shinbrot et al., 2014). Nevertheless, involvement of Pol δ on the leading strand is possible under stress, in such cases as re-priming following a DNA damage avoidance or replication fork restart (Lujan et al., 2016; Miyabe et al., 2015).

Polymerase	Subunit	Function	Gene	Protein
Pol ε	A	Polymerase, 3' to 5' exonuclease	POLE (POLE1)	p261
	B	Regulatory	POLE2	p59
	C	Double-stranded DNA binding	POLE3	p17
	D	Double-stranded DNA binding	POLE4	p12
Pol δ	A	Polymerase and 3' to 5' exonuclease	POLD1	p125
	B	Regulatory	POLD2	p50
	C	Regulatory	POLD3	p68 (p66)
	D	Regulatory	POLD4	p12
Pol α	A	Polymerase	POLA1	p180
	B	Regulatory	POLA2	p70
Primase	A	Catalytic	PRIM1	p49
	B	Regulatory	PRIM2	p58

Table 1.1. Human replicative polymerases.

Both Pol ε and Pol δ consist of one large catalytic subunit and three smaller subunits in humans (Table 1.1). The catalytic subunits (encoded in humans by *POLE* and *POLD1*, respectively) contain a polymerase domain and a 3' $\rightarrow$ 5' exonuclease domain (Ream et al., 2014, chapter 2). The exonuclease domain provides Pol ε and Pol δ with an important ability to proofread the newly synthesised DNA strand. Germline mutations

in the proofreading domains of Pol ε and Pol δ predispose to cancer and proofreading-null mice develop cancers (Rayner et al., 2016; Albertson et al., 2009). Somatic mutations in the proofreading domains are found in ultramutated cancer samples, with often more than 10^5 mutations per Gbp (Shinbrot et al., 2014; Shlien et al., 2015). Both the exonuclease activity and the high fidelity of the polymerase domain are essential for the accuracy of eukaryotic replication. As Pol α lacks the exonuclease domain, it is only moderately accurate, which is the reason why most of the DNA synthesised by Pol α is removed by the strand displacement activity of Pol δ (Williams et al., 2016).

The replication substantially differs when the template DNA contains a lesion. A summary of pathways involved in replication of damaged DNA is provided in section 1.3.2.

1.2 Epigenomics

A substantial part of information needed for correct functioning of a cell is encoded in epigenetic modifications (see Appendix 8.1 for more detailed definition and a historical context of the term epigenomics). Epigenomic information can be carried in a number of features: DNA modifications, histone modifications, nucleosome positioning, chromatin interactions and domains. This section introduces DNA methylation 1.2.1 and hydroxymethylation 1.2.2, and other DNA modifications 1.2.3, and their role in cancer 1.2.4. A brief summary of other types of epigenetic modifications can be found in Appendix 8.2.

1.2.1 DNA methylation in normal cells

The most extensively studied epigenetic modification is cytosine with a covalently attached methyl group, forming a molecule 5-methylcytosine (5mC). Methylation of cytosine is found in bacteria (such as *Escherichia coli*), plants (such as *Arabidopsis thaliana*), fungi (such as *Neurospora crassa*), insects and other invertebrates, and the genomes of all examined vertebrate species (Suzuki and Bird, 2008; Su et al., 2011; Capuano et al., 2014), but is absent from *Caenorhabditis elegans* (Simpson et al., 1986) and all examined yeast strains (Capuano et al., 2014). The amount, sequence context, genomic context and proposed functions of 5mC markedly differ between species

(Suzuki and Bird, 2008; Du et al., 2015). In the remainder of the introduction, we will focus on vertebrates and especially humans.

In human DNA, around 4 % of cytosines are methylated (which corresponds to ca. 1 % of all DNA bases) and most of 5mC resides in CpG dinucleotides (Breiling and Lyko, 2015; Bird, 2002; Bird and Taggart, 1980). Methylation outside CpG context (i.e., in a CpH sequence context, where H refers to A, C, or T) occurs predominantly in CpA dinucleotides and is most abundant in neurons (>2 % of CpA positions), relatively abundant (1–2 %) in other adult brain cells, H1 embryonic stem cells (ESC), oocytes, etc., while it is low (<1 %) in heart, aorta, stomach, etc., and undetectable in sigmoid colon, small bowel, sperm, or fibroblasts (He and Ecker, 2015). CpH methylation has started to attract more attention in recent years, especially in the context methyl-CpG binding protein 2 (MeCP2), which binds also 5mCpApC and is a critical protein in the neurological disorder Rett syndrome (He and Ecker, 2015; Luo and Ecker, 2015; Kinde et al., 2015). However, in this thesis we will focus mostly on the far more abundant CpG methylation.

Around 70–80 % of CpGs are methylated (Bird, 2002). As CpG is a palindromic sequence[3], the CpG methylation can be efficiently maintained after DNA replication before cell division simply by copying the methylation status from the template strand. This mechanism of methylation maintenance has been proposed already in the early papers suggesting 5mC to be a heritable epigenetic mark in vertebrates (Holliday and Pugh, 1975; Riggs, 1975). The copying of methylation groups during replication is performed mainly by the maintenance DNA methyltransferase DNMT1 guided by UHRF1 (Cheng, 2014; Du et al., 2015). Methylation can be deposited also *de novo*, by DNMT3A and DNMT3B enzymes[4] All these three methyltransferases are required for normal embryonic and neonatal development (Li et al., 1992; Okano et al., 1999). Moreover, DNMT3L, a catalytically inactive DNMT3 homologue, is required for *de novo* methylation primordial germ cells and Dnmt3L$^{-/-}$ male mice are sterile (Rose and

[3]Genomic palindromic sequence is a sequence, which is the same as its reverse complement, i.e., 5'-to-3' sequence on the complementary strand.

[4]However, this division of work is not absolute, as DNMT3A and DNMT3B enzymes are thought to be also required for methylation maintenance, as well as DNMT1 can exhibit *de novo* methylation activity at certain repetitive elements (Jones and Liang, 2009; Arand et al., 2012).

Klose, 2014). A complete lack of methylation is incompatible with viability of normal somatic cells and cancer cells, but not mouse ESC (Jones, 2012). DNA methylation has multiple functions: X-chromosome inactivation, imprinting, promoting genome and chromosomal stability, regulating transcription, preventing spurious transcription initiation, and it has been also proposed to play a role in regulating alternative splicing (for more details, see Appendix 8.3).

1.2.2 DNA hydroxymethylation in normal cells

The second most common DNA modification in human DNA is 5-hydroxymethylcytosine (5hmC), which was indisputably shown to exist in brain and other tissues only recently in 2009 (Kriaucionis and Heintz, 2009). It was concurrently shown that ten-eleven translocation (TET) enzymes are able to convert 5mC into 5hmC (Tahiliani et al., 2009). Unlike 5mC, which is observed at similar levels in many cell types (showing only 1–2.5-fold difference), the abundance of 5hmC varies widely (up to 22-fold difference), but is detectable in ESC and all examined tissues (Li and Liu, 2011; Tomkova et al., 2016; Tahiliani et al., 2009; Globisch et al., 2010; Szwagierczak et al., 2010; Wu and Zhang, 2011; Nestor et al., 2012; Liu et al., 2013) (Fig. 1.4).

The discoveries about 5hmC and TET enzymes gave a new direction in the long-term search for the mechanisms of DNA demethylation (transition from 5mC to C). Global DNA demethylation occurs in development and other contexts, including cancer. Although the mechanisms of *de novo* and maintenance methylation are well understood, the opposite process is still under debate. The simplest form is *passive demethylation*, in which replication dilutes 5mC due to missing, down-regulated, or inefficient DNA methylation maintenance. However, it has been shown that demethylation in the zygotic paternal genome occurs rapidly after fertilisation to such an extent that cannot be explained by the replication-dependent passive dilution (Mayer et al., 2000; Oswald et al., 2000). Several mechanisms of *active demethylation* have been proposed. The older and less supported ones (reviewed in Wu and Zhang, 2010) include:

- Enzymatic removal of the methyl group of 5mC by MBD2; however, Mbd2-null mice are viable and exhibit normal demethylation.

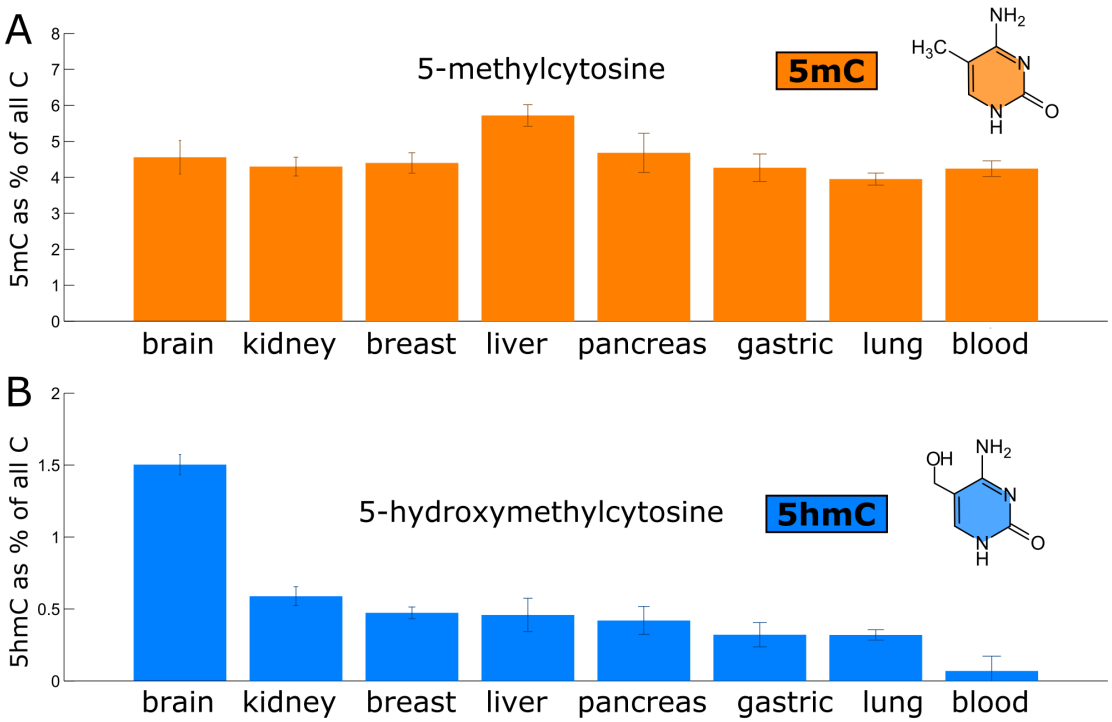

Figure 1.4. HPLC measurements of total 5hmC and 5mC in eight tissues: average values with standard deviation of 5mC and 5hmC (as a percentage of total cytosine). Measured by Michael McClellan, methods described in Tomkova et al. (2016).

- Direct excision of 5mC by BER (which is used in plants); however in mammals suitable glycosylases have not been found. Both of the proposed TDG and MBD4 have 30–40-fold lower activity against 5mC:G than T:G and Mbd4-null zygotes exhibit normal demethylation.

- Enzymatic deamination of 5mC by activation-induced deaminase (AID) or apolipoprotein B mRNA editing enzyme, catalytic polypeptide (APOBEC) family of proteins, followed by BER of the produced T:G mismatch; however, AID and all the examined APOBEC enzymes show higher efficiency for C than 5mC (with further decrease for 5hmC and the higher oxidative states), as summarised in section 1.4.4.

- Nucleotide excision repair; however, biochemical evidence is missing and some of the supporting results were irreproducible.

The most accepted models include involvement of TET enzymes (Fig. 1.5). They can not only mediate oxidation from 5mC to 5hmC, but also further to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Ito et al., 2011). Both 5fC and 5caC can be excised by TDG, creating an abasic site, which is then repaired by BER and unmodified C is restored (reviewed in Wu and Zhang, 2017). Alternatively, when DNA replication occurs after the oxidation, the created 5hmC, 5fC, and 5fC are paired with unmodified cytosine, which is referred to as *TET-assisted passive demethylation* (Hill et al., 2014) or *active modification–passive dilution* (Wu and Zhang, 2017).

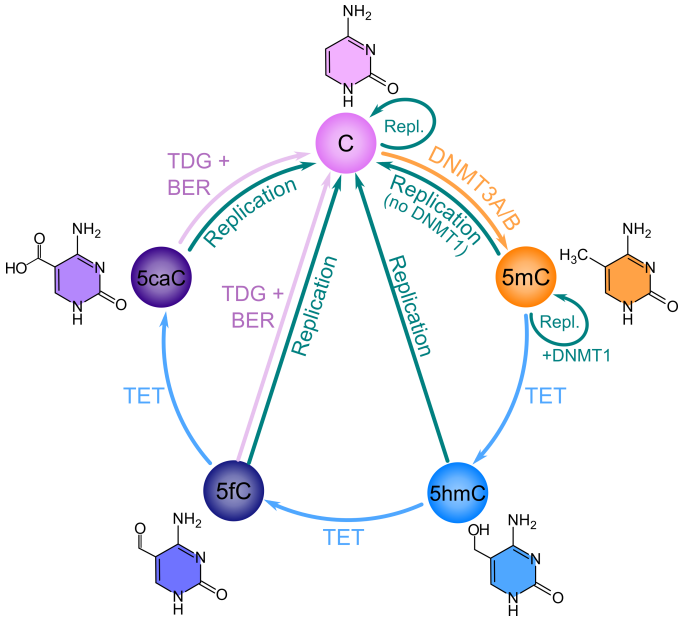

Figure 1.5. Passive and active demethylation. Passive demethylation happens during replication of methylated CpGs when DNMT1 does not copy the methylation mark to the other strand. TET-assisted passive demethylation involves oxidation of 5mC by TET enzymes into 5hmC, 5fC, or 5caC, followed by replication and pairing with unmodified C. In TET-assisted active demethylation, 5fC or 5caC are excised by TDG and restored to C by BER.

Initially, 5hmC was studied mainly for its roles in active demethylation. This was changed by the discovery that majority of 5hmC is found stable in the genomic DNA, as opposed to transient intermediate between methylated and unmodified cytosine (Bachman et al., 2014). A number of DNA binding proteins recognising 5hmC have been identified (Mellén et al., 2012; Spruijt et al., 2013; Takai et al., 2014). Moreover, 5hmC is particularly enriched in promoters and gene bodies of actively transcribed genes (Wu and Zhang, 2011; Williams et al., 2011) and the 5hmC levels in gene bodies

correlate with gene expression (Mellén et al., 2012). 5hmC has been therefore implicated in regulation of transcription and splicing (for more details, see Appendix 8.3).

1.2.3 Other DNA modifications in normal cells

Compared to 5mC and 5hmC, their oxidative products 5fC and 5caC are much less abundant, with levels in different mouse tissues ranging between 0.2–15 $\times 10^{-4}$ % of C and up to 2×10^{-4} % of C, respectively (Bachman et al., 2015), in line with measurements in other studies (Ito et al., 2011; Iurlaro et al., 2016). Both 5fC and 5caC are substrates for TDG (Maiti and Drohat, 2011) and *Tdg* deficiency in E11.5 mouse embryos causes approximately 7-fold increase of 5fC, indicating active TET-TDG-mediated demethylation (Iurlaro et al., 2016). This activity is enriched at exon-intron boundaries and CGI shores and 5fC is increased in active enhancers in both WT and *Tdg*-null animals, as shown by single-base resolution sequencing of 5fC (Iurlaro et al., 2016). Not only 5fC and 5caC are substrates for TDG, but they also recruit a number of other DNA-repair-associated proteins from BER and MMR pathways (Spruijt et al., 2013). This might be enhanced by the altered structure of DNA double helix in the presence of 5fC (Raiber et al., 2015). Moreover, 5caC:G pair stimulates Pol δ exonuclease activity and is recognised as a mismatch by MMR as strongly as T:G pair (Shibutani et al., 2014). In spite of this lesion-like treatment of 5fC and 5caC, it was shown that not only 5hmC, but also 5fC can be a stable DNA modification (Bachman et al., 2015). This supports the role of 5fC in other biological functions than only DNA demethylation intermediate.

It was thought that cytosine is the only base that carries modifications with epigenetic functions in the mammalian DNA. Recent discovery of N^6-methyladenine (N6mA) in the DNA of mouse ESC changed this view (Wu et al., 2016). This modification is rare ($6 - 7 \times 10^{-4}$ % of A) and largely unexplored, but has been suggested to control evolutionarily young LINE-1 retrotransposons (Koziol et al., 2015; Luo et al., 2015; Pfeifer, 2016).

1.2.4 DNA modifications in cancer cells

DNA modifications affect cancer in two major ways. First, they have an important effect on mutagenesis, already in normal cells, and thus alter the risk of cancer. The best known effect is spontaneous deamination of 5mC into T. As most 5mCs occur in a CpG context, this leads to a high number of CpG>TpG mutations, which represent the most common mutation type in cancer, normal somatic cells and germ line, as detailed in section 1.4. The role of 5mC and other DNA modifications on the origin of cancer mutations (via the best known spontaneous deamination, but also via other processes) is one of the main topics of this thesis.

Second, DNA modifications exhibit substantial changes during tumorigenesis. Historically, three types of changes have been identified. Hypermethylation of CpG islands of gene promoters was associated with silencing of these —often tumour-suppressor— genes (e.g., MLH1) (Sakai et al., 1991; Gonzalez-Zulueta et al., 1995; Herman et al., 1994; Hiltunen et al., 1997; Jones and Baylin, 2002). Genome-wide hypomethylation was associated with genomic instability (Esteller and Herman, 2002; Eden et al., 2003; Jones and Baylin, 2002). And hypomethylation of gene-specific promoters was linked to activation of oncogenes (Nishigaki et al., 2005; Oshimo et al., 2003; Akiyama et al., 2003; Cho et al., 2000; Sato et al., 2003).

However, more recent results suggest that the impact of DNA modification alterations in cancer is more complex. For instance, the CpG island hypermethylation happens after the genes are silenced by other means, such as Polycomb complexes (Keshet et al., 2006). Therefore methylation serves not as the primary silencing mechanism, but might prevent the gene to be activated (Klutstein et al., 2016). Nevertheless, specific hypermethylation found in CGIs, termed *CGI methylation phenotype* (CIMP), is a frequent feature of some types of cancers (Hughes et al., 2013). It was first identified in colorectal cancer (C-CIMP) (Toyota et al., 1999), where it was associated with *BRAF* mutations, microsatellite instability, and predictive of shorter survival (Zong et al., 2016). Similar phenotypes were subsequently shown to be predictive of better outcome in a distinct group of *IDH1*-mutated gliomas (G-CIMP) (Noushmehr et al., 2010), predictive of response to epigenetic treatment in infant ependymomas (Mack et al., 2014), mildly

predictive of worse survival in a group of oesophageal adenocarcinomas (Krause et al., 2016), and observed in many other cancer types (Hughes et al., 2013).

Similarly, the role of cancer hypomethylation might be more complex, such as contributing to transposable elements activation (Burns, 2017) or inducing spurious transcription and production of aberrant transcripts (Neri et al., 2017). The latter is supported also by the fact that loss/mutation of SETD2 and loss of H3K36me3 mark (two important players in prevention of spurious transcription; see Appendix section 8.3) are key events promoting cancer growth (Duns et al., 2010; Fontebasso et al., 2013; Kanu et al., 2015).

Most examined cancer types exhibit also a significant hypo-hydroxymethylation (Li and Liu, 2011; Jin et al., 2011; Haffner et al., 2011; Kraus et al., 2015). The reasons or impacts of this depletion of 5hmC are not well understood. It might be a simple consequence of higher proliferation rate of cancer cells and slow establishment of 5hmC on the nascent strand after DNA replication (Bachman et al., 2014). On the other hand, loss of 5hmC predicts poor prognosis in a number of cancer types (Lian et al., 2012; Chen et al., 2015; Yang et al., 2013b; Shi et al., 2016b; Zhang et al., 2016)[5].

Moreover, genes involved in TET-mediated demethylation are often mutated or downregulated in cancer. *TET2* gene is mutationally inactivated in about 15 % of myeloid cancers, including 22 % of acute myeloid leukemia (AML) (Delhommeau et al., 2009; Langemeijer et al., 2009) and TETs are often downregulated in human cancers (Yang et al., 2013a; Kohli and Zhang, 2013). *IDH1* and *IDH2* are mutated in more than 70 % of lower-grade gliomas (grades II and III) (Turcan et al., 2012), in some glioblastomas (Yan et al., 2009; Parsons et al., 2008), AML (Mardis et al., 2009), thyroid carcinomas (Hemerly et al., 2010; Murugan et al., 2010) and several other cancers (Sjöblom et al., 2006; Mardis et al., 2009; Pansuriya et al., 2011; Amary et al., 2011a,b). The wild-type IDH1 and IDH2 catalyse the conversion of isocitrate to α-ketoglutarate, which is a cofactor for many dioxygenases including TET enzymes. The most common mutations in these genes (R132 in *IDH1* and R140 and R172 in *IDH2*) lead to production of α-hydroxyglutarate, an oncometabolite that can competitively inhibit these α-ketoglutarate-dependent

[5] However, this also does not prove an active role of 5hmC loss in the carcinogenesis, as the predictivity might be also just a consequence of another common cause, such as increased proliferation rate.

TET enzymes, and thus oxidation of 5mC to 5hmC (Xu et al., 2011). These *IDH1* mutations do indeed lead to global hypermethylation and hypohydroxymethylation (Figueroa et al., 2010; Turcan et al., 2012; Bardella et al., 2016). However, how this global hypermethylation is restricted only to CGIs, as observed in the *IDH1*-mutated G-CIMP cancers, is currently unknown. Finally, also *DNMT3A* mutations are highly recurrent in AML patients (Ley et al., 2010).

Notwithstanding the unanswered questions about the mechanisms of DNA modifications in tumorigenesis, they have proved promising both as clinical markers (Heyn and Esteller, 2012; Bock et al., 2016; Moran et al., 2017; Guo et al., 2017) and in the design of cancer treatment (Yang et al., 2010, 2014; Zauri et al., 2015; Gustafson et al., 2015; Wongtrakoongate, 2015).

1.3 DNA mutagenesis and associated repair

DNA mutations are permanent changes to the DNA of different sizes: *single nucleotide variants* (SNVs; one base substitution), small-scale insertions and deletions (indels; up to 10 kbp), and large-scale chromosomal changes, such as *copy number variations* (CNVs; large-scale amplifications, deletions, and translocations). Mutations in the coding regions of DNA that lead to a change of amino acid (*missense mutations*) or truncation (*nonsense mutations*) of the translated protein are called *non-synonymous mutations*. This change can cause inactivation of the protein (*loss-of-function mutation*) or can cause increased or even novel activity of the protein (*gain-of-function mutation*), e.g., through altered structure of the folded protein. While loss-of-function mutations are often broadly distributed over a gene body, gain-of-function mutations usually happen in only very specific recurrently mutated positions (Baeissa et al., 2017). Finally, *synonymous mutations* are changes to the sequence of a gene that do not directly change the sequence of the encoded protein. Although these mutations are also called “silent”, they have been shown to have the potential to contribute to human cancer, such as through regulation of alternative splicing (Supek et al., 2014b).

DNA mutations in germ line[6] are one of the key components of evolution, providing variation in the genomes of individuals in the population. On the other hand, mutations that happen in *somatic cells* (other than germ line) are not passed to progeny and do not therefore influence evolution. Most of the mutations in somatic cells are harmless, but some can give rise to various diseases, including cancer.

In the current[7] model of tumorigenesis, cancer is caused by mutations in specific genes, which —when mutated— can lead to changes in the phenotype, such as increased growth, proliferation, or DNA repair deficiency, giving a selective advantages to the cells, and by that give rise to cancer (Vogelstein et al., 2013; Martincorena and Campbell, 2015). These mutations that causally drive the disease are called *cancer driver mutations*, while majority of somatic mutations are harmless (*passenger mutations*) (Stratton et al., 2009).

The view of mutations being the primary cause of cancer has been challenged several times. For instance, whole genome sequencing of ependymoma, a rare brain tumour type, found no significant recurrent mutations in the cohort of 47 patients, and several samples without any mutations in the entire genome, but with a potential epigenetic origin (Mack et al., 2014; Parker et al., 2014; Versteeg, 2014). Nevertheless, apart from few outlier cases, the current evidence supports the importance of mutations in tumorigenesis, albeit with important involvement of epigenetics (You and Jones, 2012; Klutstein et al., 2017), nutrition (Campbell, 2017), metabolism (Cao et al., 2015) and other factors (Moore and Chang, 2010; Elinav et al., 2013).

Given the importance of DNA mutations in tumorigenesis, the next important question is what causes the mutations. Although the process of acquiring mutations is to some extent stochastic, large-scale sequencing studies in the recent decade have revealed that the distribution of somatic mutations across the genome is not uniform (Lawrence et al., 2013). Apart from positive and negative selective pressure, a number of factors can influence distribution of mutation frequencies, such as chromatin organisation (Schuster-Böckler and Lehner, 2012), replication timing (Koren et al., 2012), metabolic load (Ames et al., 1993), transcription (Lawrence et al., 2013), sequence

[6] Germ line is a population of cells in sexually reproducing organisms that are/give rise to the gametes. The DNA of germ line cells is passed to the progeny during reproduction.

[7] A very brief historical perspective is in Appendix 8.4.

context (Alexandrov et al., 2013a; Nik-Zainal et al., 2012a), and exposure to different mutagens (Poon et al., 2013).

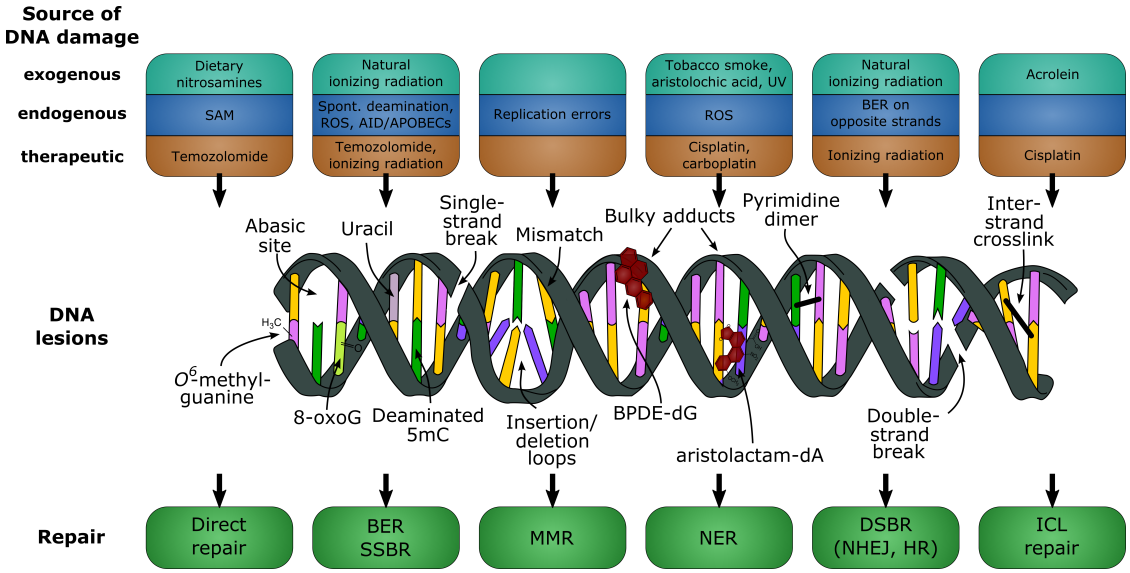

Figure 1.6. DNA damage and repair DNA damage can be caused by exogenous or endogenous mutagens or by therapeutic DNA-damaging agents. Schematic examples of the corresponding DNA lesions are shown in the middle, and the repair pathways are shown below the lesions that they repair. SAM: S-adenosyl methionine, ROS: reactive oxygen species, UV: ultraviolet, BER: base excision repair, SSBR: single-strand break repair, MMR: mismatch repair, NER: nucleotide excision repair, DSBR: double-strand break repair, NHEJ: non-homologous end joining, HR: homology repair, ICL: iterstrand crosslink.

The non-uniform distribution of mutations in cancer genomes is therefore likely a result of a number of non-uniform mutation-causing processes, DNA repair, fixation of mismatches/DNA damage into mutations, and selection (Fig. 1.6). *DNA damage* or *DNA lesion* is a chemical change of the DNA base (such as deamination, oxidation, attachment of bulky adducts, removal of the base, etc.) or DNA structure (single and double strand breaks, cross links between adjacent bases, etc.). In contrast, DNA mutations are changes in the DNA sequence (while the DNA structure, DNA bases and other elements defining the DNA remain unchanged) and arise from mis-incorporation of a wrong base during replication, or from DNA damage that was incorrectly repaired (Martincorena and Campbell, 2015).

Genomic sequencing of cancer mutations in the last decade has helped to identify a number of mutational processes. These processes often show characteristic mutational patterns, which can be described by the type of mutation (such as C>T) and

their trinucleotide sequence context (such as TCG>TTG). A mathematical method for separation of signals with different sources was recently applied to identify *mutational signatures* of the main mutational processes operating in cancer patients (Alexandrov et al., 2013a). The concept of mutational signatures and their detection is described in the General methods 2.1.1.

The next sections summarise the main pathways for DNA damage repair (1.3.1), how replication deals with unrepaired damage (1.3.2), and a brief summary of the main known endogenous and exogenous sources of DNA damage (1.3.3). The mutational processes influenced by DNA modifications (1.4) and replication (1.5) are described in detail.

1.3.1 DNA repair

1.3.1.1 Direct reversal repair

In some cases, a DNA lesion can be repaired by a simple direct reversal. The best known example of such repair is demethylation of O^6-methylguanine lesion by O^6-methylguanine DNA methyltransferase (MGMT) (Curtin, 2012). Direct reversal of DNA alkylation can be performed by AlkB family of DNA repair dioxygenases, such as ALKBH2/3, which exhibit demethylation activity against the cytotoxic N1-methyladenine and N3-methylcytosine DNA adducts (Duncan et al., 2002; Sedgwick et al., 2007; Yi et al., 2012).

1.3.1.2 Base excision repair (BER)

BER repairs most of the non-bulky DNA base lesions (oxidised, deaminated, and alkylated bases) that cannot be corrected by direct reversal (Bauer et al., 2015). First, a glycosylase recognises the DNA lesion (or mismatch, such as T:G mismatch recognition by TDG), hydrolyses the β-N-glycosidic bond between the base and the sugar, and removes the base, leaving an apurinic site or apyrimidinic site, jointly called *abasic site* (AP site). The glycosylases are usually lesion-specific (Iyama and Wilson, 2013). The created AP site is hydrolysed by an AP endonuclease, such as APE1, creating a nick (a single-strand break). The following steps differ between *short path BER* and *long path BER*, two modes of this pathway. In short patch BER, polymerase Pol β replaces the

missing nucleotide and the nick is sealed by XRCC1–LIG3α complex, in assistance of Poly(ADP-ribose) polymerase 1 (PARP1). In long patch BER, several nucleotides are replaced by strand-displacement synthesis, followed by flap removal and nick sealing (Curtin, 2012; Iyama and Wilson, 2013; Bauer et al., 2015). BER is important also for repair of AP sites and single-strand breaks created by other means.

1.3.1.3 Nucleotide excision repair (NER)

NER corrects helix-distorting bulky base adducts and intrastrand crosslinks (Curtin, 2012; Bauer et al., 2015). NER is initiated by recognition of the damage, followed by incision of an approximately 24–32 nucleotide long single-strand oligonucleotide fragment around the damage. The gap is filled by a DNA polymerase and sealed with a ligase (Marteijn et al., 2014). Two modes of NER operate on the DNA: global genome NER (GG-NER) and transcription-coupled NER (TC-NER). The two modes differ in the first recognition step. In GG-NER, damage sensor XPC complex constantly probes the DNA for helix-distorting lesions and the damage is recognised by XPC accompanied with UV-DBB complex (the damage is often "flipped out" to allow direct binding to XPC) (Marteijn et al., 2014). In TC-NER, the damage is recognised during transcription elongation by RNA Pol II, which stalls at the lesion, a complex of CSA-CSB proteins is formed, and RNA Pol II backtracks to leave the DNA lesion accessible for repair (Marteijn et al., 2014).

Deficiency in GG-NER leads to increased mutagenesis, photosensitivity and cancer, such as in a syndrome called Xeroderma pigmentosum (XP), caused by mutations in XPC and other GG-NER genes. On the other hand, deficiency in TC-NER results in premature ageing and neurological disorders, such as in a syndrome called Cockayne syndrome (CS), caused by mutations in CSA, CSB, and other TC-NER gene (Menck and Munford, 2014; Reid-Bayliss et al., 2016).

1.3.1.4 DNA double strand break repair (DSBR)

Double-strand breaks (DSBs) belong to the most deleterious DNA lesions, leading to genomic translocations and activating cell death if unrepaired (Iyama and Wilson, 2013). DSBs in dividing cells are repaired by homologous recombination (HR), whereas all cells

can be repaired by non-homologous end joining (NHEJ). During HR, the sister chromatid is used as a template for synthesis of the missing parts on both strands (BRCA1 and BRCA2 are used in the first part of HR before the sister chromatid double helix is opened and used as a template), whereas in NHEJ the ends of the DSB are re-ligated, possibly leaving insertions or deletions at the breakpoint (Tubbs and Nussenzweig, 2017). More details and different subtypes of NHEJ are reviewed, e.g., in (Chang et al., 2017). Germline (or somatic) mutations in *BRCA1/2* significantly increase the risk of cancer, lead to a characteristic mutational signature including deletions flanked by short repeats (possibly due to enhanced use of NHEJ instead of the deficient HR), but also enabled design of the first targeted treatment for inherited cancer disorder (Lord and Ashworth, 2016).

1.3.1.5 Mismatch repair (MMR)

MMR recognises and repairs errors on the nascent strand of DNA replication. Despite the name, single-nucleotide mismatches are only one (and perhaps the least important one) of the types of errors recognised by MMR (Crouse, 2016). It suppresses insertion/deletion loops that resulted from slipped mispairing (illustrated in Fig. 1.6), it can recognise ribonucleotides, and it was suggested to play an important role in preventing mutations due to damaged bases (Crouse, 2016).

The errors are recognised by MutS complex. Mismatches and short insertion/deletion loops are recognised by MutSα complex (MSH2–MSH6 dimer), whereas recognition of longer insertions and deletions is performed by MutSβ complex (MSH2–MSH3 dimer). The bound MutS recruits MutL complex (comprising of MLH1 and PMS2), which coordinates the recruitment of additional proteins for excision of the damaged strand, filling the gap, and ligation of the nick (Curtin, 2012; Hewish et al., 2010).

Defects in the MMR pathway lead to *microsatellite instability* (MSI)[8], 100–1 000-fold increase of mutations, and association with cancer (Hewish et al., 2010; Curtin, 2012; Iyama and Wilson, 2013; Helleday et al., 2014). Defects in MMR are also observed in Lynch syndrome, a hereditary dominant condition, predisposing for cancer (also known

[8] MSI is defined as *variability in the length of base pair repeated sequences (< 5 bp) that is caused by replication slippage and that is normally kept stable by mismatch repair* (Helleday et al., 2014).

as hereditary non-polyposis colon cancer, HNPCC) and accounting for approximately 3 % of all colorectal cancer patients (Hewish et al., 2010). MMR deficiency in Lynch syndrome is either due to germline loss-of-function mutation in one of the MMR genes (*MLH1*, *MSH2*, *PMS2*, or *MSH6*), or due to hemiallelic methylation of *MLH1* or *MSH2* (Hewish et al., 2010).

In its canonical, replicative function, MMR is strand-specific, correcting the daughter strand, but cannot repair damage to the template replicating strand (Curtin, 2012; Crouse, 2016). MMR has been observed to act sometimes also outside the context of replication (named non-canonical function of MMR), but due to lost discrimination of the correct and erroneous strand, it often acts mutagenically. Such behaviour can be intentional and physiological, such as in the case of somatic hypermutation at the immunoglobulin locus (Crouse, 2016).

1.3.2 Replication of damaged DNA

If a mismatch is not repaired before replication, after the template strands are separated, the mismatch is fixated into a mutation. Even more dangerous is that unrepaired DNA damage (different than mismatch) can cause replication fork collapse, leading to genome rearrangements, cell death, and disease (Cortez, 2015). Multiple DNA damage tolerance (DDT) pathways therefore exist to allow bypass/avoidance of the DNA damage and normal continuation of replication. The three main DDT pathways are: translesion synthesis (TLS), template switching (TS), and homologous recombination (HR, salvage pathway) (Branzei and Szakal, 2016b) (Fig. 1.7).

1.3.2.1 Translesion synthesis (TLS)

TLS has the ability to replicate across the DNA lesion without need of a different template. This is enabled by a class of TLS polymerases, which lack proofreading activity, but can recognise modified nucleotides or other DNA lesions and are able to insert nucleotides opposite them (Bi, 2015). Unsurprisingly, the incorporated nucleotide may be wrong[9], making TLS a mutagenic pathway.

[9]Sometimes, it is not even clear what is the right nucleotide, as it depends on how the DNA lesion originated.

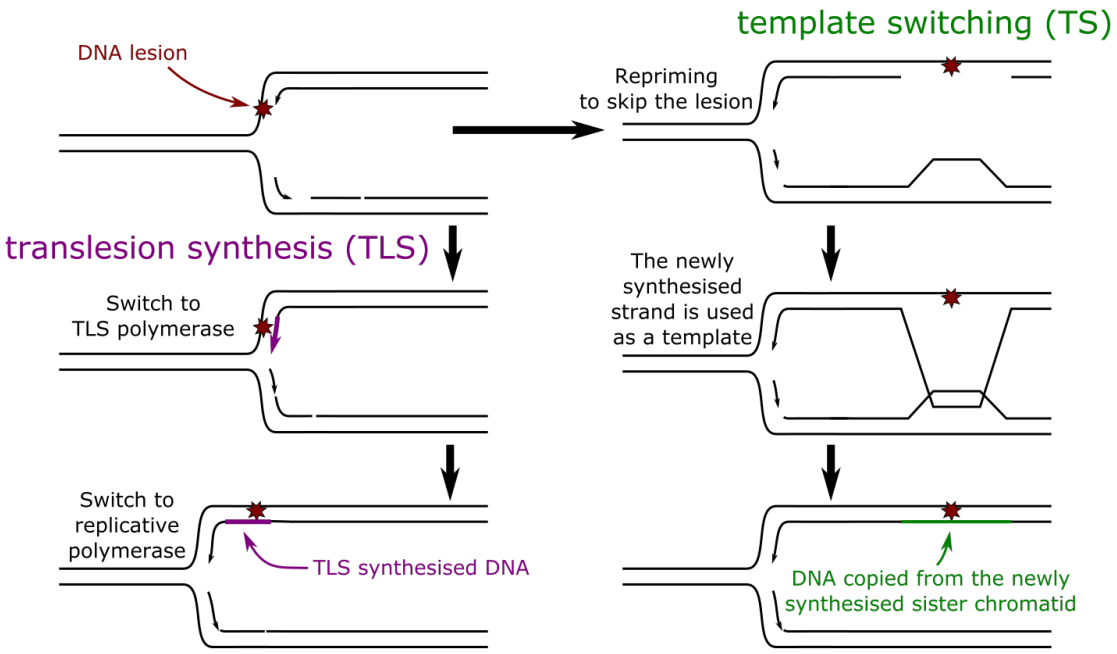

Figure 1.7. DNA damage tolerance during replication. In translesion synthesis (left), DNA polymerases that can synthesize DNA past DNA lesions are used. In template switching (right), the sister chromatid is used as a template instead of the damaged strand. Here, TLS is shown in a polymerase switching mode, in which the replication fork stalls until the lesion is bypassed. TS is shown in post-replicative gap-filling mode, in which the lesion is skipped by the replicative polymerase, a gap is created and filled afterwards. However, both TLS and TS might operate in both modes, depending on the lesion and other factors.

Seven TLS polymerases are known in human cells: Pol η (gene *POLH*), Pol ι (gene *POLI*), Pol κ (gene *POLK*), REV1 (gene *REV1*), Pol ζ (the catalytic subunit encoded by gene *REV3L*), Pol θ (gene *POLQ*), Pol ν (gene *POLN*), and the recently discovered PrimPol (gene *PRIMPOL*) (Lange et al., 2011; Rudd et al., 2014) (Fig. 1.8).

Two modes of TLS exist (Zhao and Todd Washington, 2017). In the first mode ("polymerase switching"), the stalled replicative polymerase is replaced with a TLS polymerase, which inserts a base opposite the lesion (Fig. 1.7 left). Extension from the base and synthesis of several more nucleotides can be performed by the same or different TLS polymerase (such as Pol ζ). The TLS polymerases are subsequently replaced back with the replicative polymerase (Sale et al., 2012; Mailand et al., 2013). The polymerase switch is promoted by PCNA monoubiquitination by RAD18–RAD6 complex (Mailand et al., 2013).

In the second mode ("post-replicative gap filling"), a gap is left around the lesion,

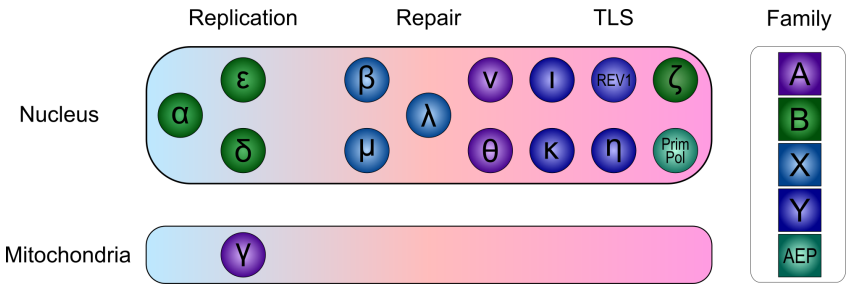

Figure 1.8. Overview of DNA polymerases DNA polymerases can be grouped based on amino acid sequence relationships into five families: A, B, X, Y, and archaeo-eukaryotic primase (AEP) superfamily. All of them operate in nucleus, only Pol γ is responsible for replication of the mitochondrial DNA. The polymerases can also be grouped by their main function: replicative DNA polymerases synthesise bulk of the DNA during replication, TLS polymerases replicate damaged DNA, and repair polymerases synthesis DNA during damage repair; however, the grouping is not absolute, as many polymerases have more than one of these three functions.

while the replication complex reprimes downstream of the lesion. The gap is filled by TLS polymerases later, in a similar two-step process as described for the first mode (Mailand et al., 2013). The gap can be filled shortly after it was generated, or as late as in the G2 phase (Mailand et al., 2013; Branzei and Szakal, 2016a).

Recent studies suggest that both modes might be in use, depending on the type of the lesion and used polymerases. For instance, Pol η and Rev1 bypass UV-induced CPD and 6-4PP at replication forks, whereas only 6-4PP are also tolerated by a Pol ζ-dependent gap-filling mechanism, independent of S phase (Quinet et al., 2016). While repriming on the lagging strand is natural, as it happens for each Okazaki fragment, skipping the lesion on a leading strand seemed more complex. However, it was shown that repriming happens also on the leading strand and the recently discovered PrimPol has been suggested to enable it, due to its ability to act as a RNA/DNA primase (next to being a TLS polymerase) (Guilliam and Doherty, 2017).

1.3.2.2 Template switching (TS) and homologous recombination (HR)

In TS, the newly synthesised sister chromatid is used as a template to synthesise DNA past the lesion. The current model of TS prefers post-replicative TS on gaps behind replication fork proximal to an origin of replication. The gap region anneals to the homologous duplex, while the newly synthesised strand is used as a template for the other strand (Branzei and Szakal, 2016b) (Fig. 1.7 right).

HR (also called salvage pathway and described already in the section about DSBR) is very similar to TS, but differs by the involved proteins and regulation: while TS is dependent on RAD5, RAD6 and RAD18 and is promoted by PCNA poly-ubiquitination, HR is mediated by RAD51 and RAD52 and is prone to crossover. (Bi, 2015; Branzei and Szakal, 2016b).

1.3.2.3 Regulation of DDT pathways

The usage of the three DDT pathways is regulated by modifications of PCNA. Increasing evidence suggests that TS is favoured at early times during replication, while TLS and HR act in late S or G2/M phase (Bi, 2015; Branzei and Szakal, 2016b). This is supported by lower mutation rates in early replicating regions, as TS is generally error-free, while TLS is more mutagenic (Lang and Murray, 2011). Moreover, disruption of TLS in yeast leads to decreased mutation frequency in late-replicating regions and therefore a more even distribution of mutation frequency between early and late-replicating regions (Lang and Murray, 2011). The mechanism for this temporal division of work is unknown, but it has been suggested that this might be due to open chromatin in the early-replicated regions, which better enables DNA bending needed for TS (Branzei and Szakal, 2016a).

1.3.3 DNA damage

Individual proteins involved in the described pathways of DNA damage replication and repair often depend on the precise nature of the DNA damage. This section gives a general introduction into the types of DNA damage, while detailed mechanisms of types most relevant for this thesis are described in sections 1.4 and 1.5.

1.3.3.1 Hydrolytic deamination

Cytosine, adenine, and guanine, the three bases with an amino group, can undergo spontaneous or enzymatic deamination (Fig. 1.9). Physiologically most relevant is deamination of cytosine, which produces 70–200 uracil bases in human DNA per day (Visnes et al., 2009; Lindahl, 1993). On the contrary, deamination rates of adenine and guanine in DNA (to hypoxanthine and xanthine, respectively) are at 2–3 % of the rate of cytosine deamination (Lindahl, 1993; Karran and Lindahl, 1980). To prevent mutations,

uracil, hypoxanthine, and xanthine are repaired by base excision repair enzymes (see section 1.3.1): uracil DNA glycosylase (UDG) and xanthine DNA glycosylases (XDGs) including SMUG1 (Visnes et al., 2009; Mi et al., 2009; Saparbaev and Laval, 1994; Lee et al., 2015a). If unrepaired, two rounds of replication can fixate the damaged base into a mutation, as uracil pairs with adenine, creating a C:G>T:A mutation, hypoxanthine pairs with cytosine, inducing a A:T>G:C mutation, and depurination of xanthine can give rise to G:C>A:T mutations by insertion of adenine opposite the resulting abasic site (Lindahl, 1993; Vongchampa et al., 2003; Terato et al., 2002).

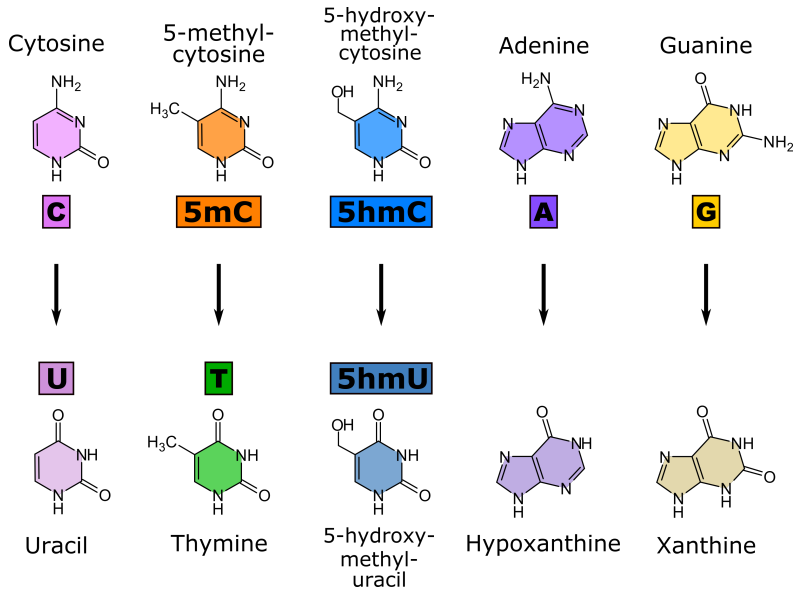

Figure 1.9. Deamination of DNA bases. The deamination products of three main bases (C, A, G) and two major DNA modifications (5mC, 5hmC) are shown below the deaminated bases.

Importantly, 5mC also deaminates and the deamination rate is two to four fold higher than that of C (Lindahl and Nyberg, 1974; Shen et al., 1994). Moreover, the deamination product of 5mC is thymine (Fig. 1.9). The resulting T:G mismatch is thought to be less efficiently repaired than the U:G pair, due to the risk of excising the wrong base (Lindahl, 1993). This is in turn thought to be the cause of the most common type of mutations: C>T transitions in a CpG context (Lindahl and Nyberg, 1974; Ehrlich et al., 1986; Lindahl, 1993). Finally, 5hmC deaminates into 5-hydroxymethyluracil (5hmC), but the rate and biological importance of this reaction are currently unknown.

Cytosine deamination can happen spontaneously, or be induced by an enzymatic activity, such as by activation induced deaminase (AID) or members of the apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) enzyme family. AID is a key enzyme in adaptive immunity, initiating antigen-dependent antibody diversification through somatic hypermutation (SHM) and class switch recombination (CSR) (Di Noia and Neuberger, 2007; Rebhandl, 2015).

APOBEC enzymes function in RNA editing and in innate immunity (Smith et al., 2012; Rebhandl, 2015). The seven known APOBEC3 proteins (A-D, F-H) use cytosine deamination in defence against viruses and retroviruses, such as HIV, human T-cell lymphotropic virus, hepatitis B virus, hepatitis C virus, human papillomavirus and human herpesviruses (Vieira and Soares, 2013) and prevent movement of retrotransposable elements, such as LINEs, SINEs, and LTRs (Chiu and Greene, 2008).

Due to their ability to convert single-stranded DNA cytosines to uracils, AID/APOBEC3 enzymes have been implicated also in genomic DNA mutagenesis (Nik-Zainal et al., 2012b; Roberts et al., 2012, 2013; Burns et al., 2013b,a; Starrett et al., 2016). In particular, AID, APOBEC3A, APOBEC3AB, APOBEC3AC and APOBEC3AH have access to the nucleus (Li and Emerman, 2011; Zhen et al., 2012; Lackey et al., 2013; Rebhandl, 2015). They deaminate cytosine in a TCN sequence context in single-stranded DNA (Smith et al., 2012; Rebhandl, 2015; Shi et al., 2016a). Clusters of C>T and C>G mutations in this sequence context have been observed in cancer, often in regions that are known to spend some time as single-stranded DNA: near double-strand breaks (Roberts et al., 2012, 2013), on the non-template strand of transcribed genes (Nordentoft et al., 2014), and on the lagging strand template of DNA replication (Haradhvala et al., 2016; Hoopes et al., 2016; Seplyarskiy et al., 2016b). Two mutational signatures have been attributed to the activity of the APOBEC enzymes: signature 2 (higher in C>T mutations) and signature 13 (higher in C>G mutations) (Alexandrov et al., 2013a). Three of the APOBEC enzymes show most convincing evidence for being involved in cancer mutagenesis:

- APOBEC3B mRNA levels are upregulated in most primary breast tumours and breast cancer cell lines, expression of APOBEC3B correlates with increased levels of genomic uracil, increased mutation frequencies, and C>T transitions, and

induced APOBEC3B overexpression causes cell cycle deviations, cell death, DNA fragmentation, c-H2AX accumulation and C>T mutations (Burns et al., 2013a). APOBEC3B is upregulated in cancer types with strong APOBEC mutational signature and the expression levels correlate with the enrichment of this signature in individual samples (Burns et al., 2013b; Seplyarskiy et al., 2016b). Moreover, high APOBEC3B levels are predictive of poor clinical outcome in a number of cancer types (Starrett et al., 2016).

- APOBEC3A overexpression causes DNA damage and cell death and the damage is especially strong during replication when the DNA is single-stranded (Landry et al., 2011; Green et al., 2016), and its broader binding sequence context in yeast is more prevalent in the observed cancer mutations than the sequence context of APOBEC3B (Chan et al., 2015). However, the specific expression of APOBEC3A in myeloid linage as opposed broad expression of APOBEC3B in a number of tissues suggests importance of the latter one.

- APOBEC3H has been linked to cancer mutations, such as in cases with polymorphism causing deletion of A3B (Starrett et al., 2016).

Due to the importance of deamination of C, 5mC, and 5hmC for this thesis, both spontaneous and enzymatic deamination types are described with further details in section 1.4.

1.3.3.2 Depurination and depyrimidination

The DNA base is covalently joined to the sugar 2'-deoxyribose by β-N-glycosidic bond. This bond can be hydrolytically cleaved, creating an AP site (Fig. 1.10). AP site can be produced by spontaneous hydrolysis, alkylation-induced hydrolysis, or glycosylase-catalysed base-excision repair (Marnett and Plastaras, 2001). Spontaneous depurination occurs at 20-fold higher rate than depyrimidination (Lindahl, 1993). AP sites are thought to be among the most common DNA lesions, with the estimated frequency of 2 000–10 000 depurination events (followed by a repair) per day in each human cell (Lindahl, 1993). When in double-strand DNA, AP sites are repaired by AP-endonuclease, DNA

polymerase, and DNA ligase (Jacobs and Schär, 2012). Trying to repair an AP site in single-stranded DNA would risk strand breakage. Instead, either A or C tend to be incorporated opposite AP sites, both at similar frequencies (Chan et al., 2013).

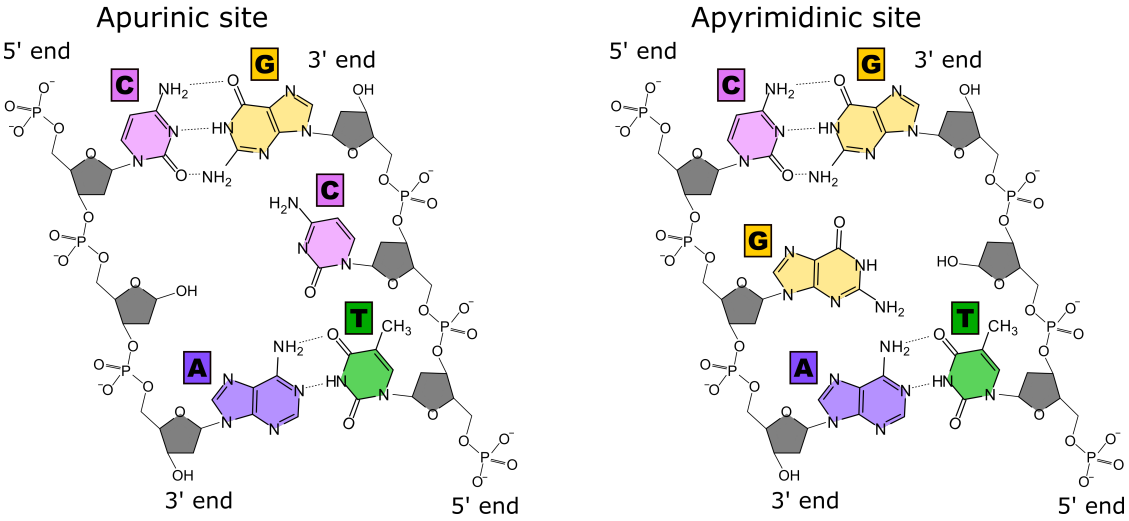

Figure 1.10. Abasic site (AP site). An example of apurinic site (left) and apyrimidinic site (right).

1.3.3.3 Oxidation

Reactive oxygen species (ROS), such as hydrogen peroxide (H_2O_2), superoxide ($O_2^{-\bullet}$) and hydroxyl radicals ($OH^{\bullet}$) are formed in all living cells as a consequence of metabolism (mainly cellular respiration), inflammation, other biochemical reactions, and external factors (De Bont and van Larebeke, 2004; David et al., 2007). Oxidative damage of DNA by ROS is highly abundant in cancer (Klaunig and Kamendulis, 2004; De Bont and van Larebeke, 2004). DNA bases —and guanine especially— are particularly susceptible to ROS-mediated oxidation (Neeley and Essigmann, 2006). The most common oxidised guanine product is 7,8-dihydro-8-oxoguanine (8-oxoguanine, 8-oxoG, 8-hydroxyguanine, 8-OH-G) (De Bont and van Larebeke, 2004), with the estimated frequency of 2 800 lesions per cell per day (Tubbs and Nussenzweig, 2017).

Mutagenic potential of 8-oxoG results from its ability to form a stable pair with C, as well as A (McAuley-Hecht et al., 1994; Le Page et al., 1998) (Fig. 1.11). This allows a relatively efficient bypass of 8-oxoG during replication, but at the cost of potentially generating mutations (Shibutani et al., 1991; Maga et al., 2007; Rodriguez et al., 2013).

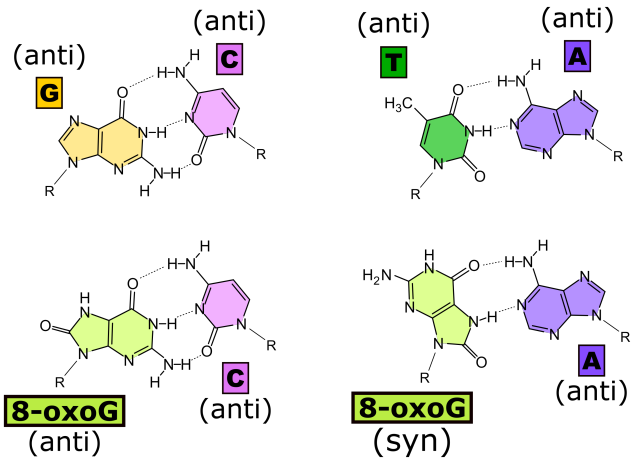

Figure 1.11. 8-oxoG can base-pair with C or A. The differences in structure of 8-oxoG compared to G (i.e., oxo group at C8 and NH at N7) allow 8-oxoG to form 8-oxoG(anti):C(anti) as well as 8-oxoG(syn):A(anti) base pairs.

Mutations associated with oxidised guanine in the DNA are G:C>T:A, resulting from the 8-oxoG paired with A (Neeley and Essigmann, 2006).

Repair of 8-oxoG is executed by proteins of the BER pathway (Fig. 1.12). OGG1 excises 8-oxoG from the 8-oxoG:C pair, allowing APE1-mediated restoration of the G:C pair (David et al., 2007). If the 8-oxoG:C pair is not repaired before replication, an 8-oxoG:A pair is often created. This is repaired by MUTYH which excises A from the mutagenic 8-oxoG:A pair, allowing incorporation of C opposite 8-oxoG by APE1, Pol λ, and FEN1, creating a substrate for OGG1 (David et al., 2007; Markkanen et al., 2013). The role of OGG1 and MUTYH in prevention of 8-oxoG-induced G:C>T:A mutations can be seen also in cells lacking these proteins. For example, disruption of MUTYH, such as due to germ line mutations causing MUTYH-associated polyposis (MAP), increases the risk of cancer and leads to a high number of G:C>T:A mutations in cancer patients with MAP (Rashid et al., 2016). The G:C>T:A mutations are likely to be prevented also by the activity of NTH1 and NEIL1 glycosylases, other two members of BER pathway (Suzuki and Kamiya, 2017).

1.3.3.4 Incorporation of damaged or incorrect nucleotides

DNA damage can happen not only to the nucleotides in the DNA, but also to the deoxyribonucleoside triphosphates (dNTPs) of the cellular DNA precursor pool. In fact,

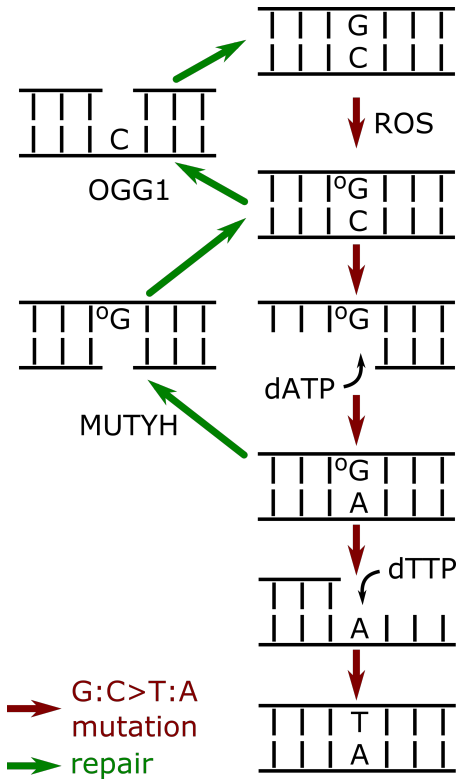

Figure 1.12. Oxidation of guanine in the DNA can cause G:C>T:A mutations. Reactive oxygen species (ROS) can oxidise guanine in G:C pair, creating 8-oxoG, which can be paired with A during replication, leading to a G:C>T:A mutation. This is prevented by MUTYH, which excises A from the 8-oxoG:A pair, creating an 8-oxoG:C pair, which is a substrate for OGG1, which restores the G:C.

the cellular DNA precursor pool is orders of magnitude more susceptible to modification than the DNA molecule itself (Topal and Baker, 1982; Rudd et al., 2016). Such damaged dNTPs can then be incorporated into DNA by replicative or TLS polymerases. Moreover, mutations can happen also due to an imbalance of dNTPs in the pool (Mathews, 2015; Mertz et al., 2015; Williams et al., 2015) and incorporation of normal but incorrect dNTPs by replicative or more frequently the less selective TLS polymerases (Lange et al., 2011), and incorporation (without a subsequent removal) of ribonucleotides (Williams et al., 2016).

The most studied damaged dNTP is 7,8-dihydro-8-oxo-2'-deoxyguanosine-5'-triphosphate (8-oxo-dGTP), the oxidation product of dGTP. DNA polymerases can incorporate 8-oxo-dGTP opposite dC or dA, but the template dA is usually favoured (dA:dC preference is higher than 100:1 in Pol ι, Pol η; dA:dC preference is higher than 10:1 in Pol λ, Pol β, Pol

γ, Pol κ; while only Pol α prefers dC) (Katafuchi and Nohmi, 2010; Patro et al., 2009). Incorporation of 8-oxo-dGTP opposite template dA can lead to T:A>G:C mutations, as observed in *in vitro* gap-filling assay by Pol η, where presence of 8-oxo-dGTP (at an equimolar concentration to the normal dNTPs) increased T:A>G:C mutation frequency 17-fold (Hidaka et al., 2008). Similarly, 8-oxo-dGTP-induced T:A>G:C mutations were also observed *in vivo* in E.coli (Inoue et al., 1998), simian cells (Satou et al., 2007), and human cells (Kamiya, 2007; Satou et al., 2009). This mutagenesis was mediated by Pol η, Rev1, and Pol ζ, but not Pol ι, as shown by siRNA knock-downs (Satou et al., 2009).

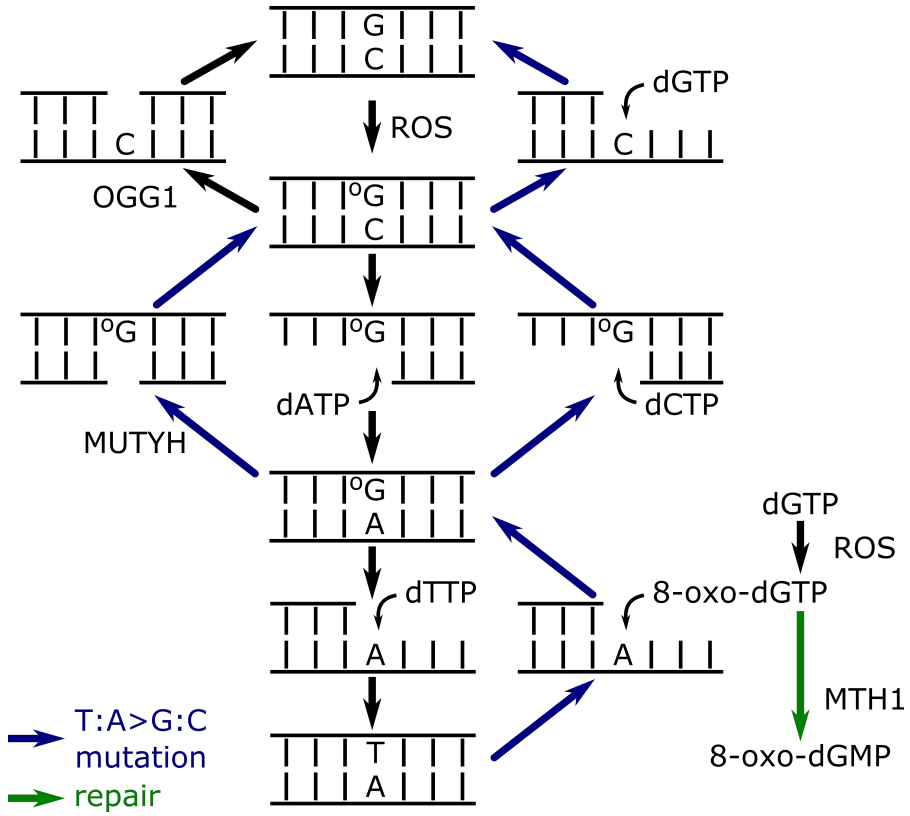

Figure 1.13. Oxidation of guanine precursor in the dNTP pool can cause T:A>G:C mutations. Reactive oxygen species (ROS) can oxidise dGTP in the nucleotide pool, creating 8-oxo-dGTP, which can be incorporated into the DNA, most commonly opposite A. During replication, the 8-oxoG can be paired with C, leading to a T:A>G:C mutation. The excision of A from the 8-oxoG:A pair by MUTYH can also contribute to the T:A>G:C mutagenesis. The main prevention this mutagenesis is MTH1, a "sanitising enzyme" in the nucleotide pool, which converts 8-oxo-dGTP into 8-oxo-dGMP.

Although 8-oxoG is efficiently repaired by BER (as described in the previous section), this repair can paradoxically in certain cases promote mutagenesis instead of preventing

it (Suzuki and Kamiya, 2017; Rudd et al., 2016) (Fig. 1.13). The 8-oxoG:A pair can either be a result of oxidised guanine on the DNA, paired with dATP during replication, or it can originate from 8-oxo-dGTP inserted opposite template dA during replication. While in the first case, it should be repaired into G:C pair, in the second case it should be repaired into T:A pair. As MUTYH excises A from the 8-oxoG:A pair (to be restored as a G:C pair), it helps with repair in the first case, but fixates a mutation in the second case (Suzuki and Kamiya, 2017). Indeed, knockdown of MUTYH reduces T:A>G:C mutations induced by the introduction of 8-oxo-dGTP into cells (Suzuki et al., 2010).

Involvement of MMR in repair of 8-oxo-dGTP incorporated into DNA during replication is under debate. MutSα binds poorly to 8-oxoG:A in both repetitive and non-repetitive sequences, but binds extensively substrates that contain an extra base in the 8-oxoG strand or in the complementary strand (Macpherson et al., 2005). This suggest a role of MMR in preventing insertions/deletions due to 8-oxo-dGTP-induced slipped/mis-paired repeat sequences. Such a role is in line with an increase of 8-oxoG levels and frameshift mutations in MMR-defective *msh2*$^{-/-}$ mouse embryonic fibroblasts and subsequent attenuation of the frameshift mutations and 8-oxoG levels by expression of the hMTH1 protein (Russo et al., 2004). Interestingly, the smallest reduction factor was observed in A:T>C:G mutations, confirming that MMR is not likely to prevent incorporation of 8-oxo-dGTP opposite template adenine.

Instead of repairing 8-oxoG after it is incorporated into DNA, cells seem to prefer to prevent the incorporation itself. The damaged nucleotide precursors are hydrolysed by sanitation enzymes, mainly from the nudix hydrolase family (Rudd et al., 2016). For instance MutT homologue 1 (MTH1; Nudix-type 1, NUDT1) hydrolyses oxidised nucleotides, including 8-oxo-dGTP, producing 8-oxo-dGMP. Inhibition of MTH1 leads to increased concentrations of 8-oxo-dGTP (Ganai and Johansson, 2016). Other sanitation enzymes (NUDT15 (MTH2), NUDT18 (MTH3), NUDT5, DCTPP1, etc.) prevent incorporation of 8-oxo-dGTP, 2-OH-dATP and other modified dNTPs, such as dUTP, 6-thio(d)GTP, 5-methyldCTP, (d)ITP, (d)XTP, and others (Rudd et al., 2016). Knockdowns of MTH1, MTH2, and NUDT5 increase the frequency of T:A>G:C induced by 8-oxo-dGTP and the mutagenesis is further enhanced in triple-knockdown cells, suggesting that the

three enzymes have mutually complementary roles in the elimination of 8-oxo-dGTP from the nucleotide pool (Hori et al., 2010).

MTH1 is often over-expressed in cancer cells and the over-expression is associated with poor prognosis in lung cancer (Fujishita et al., 2017; Nakabeppu et al., 2017). Targeting MTH1 was therefore suggested as a promising anti-cancer strategy (Gad et al., 2014; Huber et al., 2014). This possibility has attracted much attention in the last three years. It was speculated that such treatment could be especially potent in combination with radiotherapy or chemotherapy, which generate high ROS (Tu et al., 2016), or for cisplatin resistant tumours over-expressing Pol κ (Sanjiv et al., 2016). However, more research is needed in this area, as some of the promising anticancer results with MTH1 inhibitors were not reproduced with different MTH1 inhibitors (Kettle et al., 2016; Ellermann et al., 2017).

1.3.3.5 Bulky adducts

Many exogenous mutagens form bulky adducts by covalent binding to various sites on DNA bases. These are mutagens from tobacco smoke, aristolochic acid, aflatoxin B_1 produced by *Aspergillus flavus* mould, cisplatin treatment and others (Helleday et al., 2014; Hu et al., 2016).

Tobacco smoke contains thousands of chemicals and over 60 of them were classified as carcinogens (Hang, 2010). The most studied mutagenic tobacco carcinogens are polycyclic aromatic hydrocarbons (PAHs), mainly benzo[a]pyrene (B[a]P), and acrolein. These carcinogens react with DNA to form bulky DNA adducts, such as benzo[a]pyrene diol-epoxide adduct on dexoyguanosine (BPDE-dG adduct) and acrolein-induced Acr-dG adduct. Already two decades ago, PAH- and acrolein-DNA adducts were observed to be preferentially formed in the same positions in *P53* gene, as mutational hotspots in the lung cancers of smokers (Denissenko and Pao, 1996; Denissenko et al., 1997; Pfeifer, 2000; Pfeifer et al., 2002; Feng et al., 2006). The BPDE-dG adducts induce G:C>T:A mutations, the predominant mutation type in lung cancers with a smoking history (Alexandrov et al., 2013a, 2016). The types and sequence contexts of these cancer mutations were very similar to those induced *in vitro* by exposing cells to benzo[a]pyrene (Nik-Zainal

et al., 2015), suggesting that most lung cancer mutations are indeed from B[a]P as opposed to the plethora of other carcinogens present in tobacco smoke. All types of DNA damage tolerance pathways are used to deal with BPDE-dG adducts during DNA replication (Jha et al., 2016; Cohen et al., 2015). The bypass of BPDE-dG can be error-free, possibly by Pol κ or TS/HR (Avkin et al., 2004; Jha et al., 2016; Cohen et al., 2015), or error-prone, such as by Pol η or REV1-recruited Pol ζ (Zhao et al., 2006; Klarer et al., 2012; Hashimoto et al., 2012b).

Aristolochic acid is another example of a carcinogen causing a formation of DNA adducts. It is a natural compound found in *Aristolochia* plants, commonly used in traditional herbal medicines (Poon et al., 2013). Aristolochic acid contains metabolites that react with DNA to form covalent aristolactam-dA adducts (Hashimoto et al., 2016). These adducts have been detected in the stomach, kidney, urinary tract, bladder, and liver (Schmeiser et al., 1988). Exposure to aristolochic acid is associated with a high risk of urothelial carcinomas of the upper urinary tract and is thought to be the reason of much higher incidence of these cancers in Asia compared to the West (Poon et al., 2013). The aristolactam-dA adducts induce A:T>T:A mutations enriched in [C|T]AG context (Poon et al., 2013). A similar mutational signature as in cancer patients was also observed *in vitro* by exposing cells to Aristolochic acid I (Nik-Zainal et al., 2015). Error-prone bypass of aristolactam-dA adducts during replication can be performed by Pol ζ (Hashimoto et al., 2016).

The main pathway repairing bulky adducts is NER, including TC-NER (as reviewed in section 1.3.1). It is therefore not surprising that the mutations induced by the above described carcinogens exhibit transcriptional strand bias with a decrease of mutations on the strand that is used as a template for the transcription (Alexandrov et al., 2016).

1.3.3.6 Photoproducts and dimers induced by ultraviolet light

Exposure to ultraviolet (UV) light induces DNA damage, mainly in the form of photochemical reactions between adjacent pyrimidine bases. The two major resulting lesions are cis-syn cyclobutane pyrimidine dimer (CPD) and pyrimidine (6-4) pyrimidone photoproduct (6-4PP). 6-4PP is known to be more toxic to cells than CPD due to

pronounced distortion in the DNA molecule (Ikehata et al., 2015; Quinet et al., 2016). However, 6-4PP is up to eight times less frequently formed in the DNA than CPD (Bryan et al., 2014). Moreover, 6-4PP lesions are rapidly removed from the DNA, most of them being repaired between 5 min and 4 h after UV exposure (mostly by GG-NER), while repair of CPD is much slower, in certain regions persisting even 2 days after UV exposure, and both GG-NER and TC-NER types of repair are used (Adar et al., 2016; Hu et al., 2015). Therefore, most mutagenesis in skin cancer is due to CPD rather than 6-4PP. The most frequent form of CPDs is TT-CPD (Bryan et al., 2014). However, TT-CPD is very efficiently bypassed by Pol η in an error-free manner (Silverstein et al., 2010; Pfeifer and Besaratinia, 2012). Therefore most of the observed UV-induced mutations in human skin cells are C>T and CC>TT transitions due to CPDs with at least one cytosine (Brash, 2015). In line with the role of CPDs in skin mutagenesis and involvement of TC-NER in repair of CPDs, the mutations characteristic for skin cancers are decreased on the transcribed strand, especially in highly expressed genes (Alexandrov et al., 2013a; Haradhvala et al., 2016).

TLS polymerase Pol η (POLH) specialises in efficient and mostly error-free bypass of CPD (Ikehata et al., 2014). Mutations in POLH gene lead to Xeroderma pigmentosum variant (XP-V), a genetic disease associated with high sensitivity to sun and UV exposure and high incidence of cancer (Ikehata and Ono, 2011).

1.3.3.7 Other types of DNA damage

Other types of DNA damage due to endogenous or exogenous factors are described in a number of reviews (e.g., Marnett and Plastaras, 2001; De Bont and van Larebeke, 2004; Loeb and Harris, 2008; Benigni and Bossa, 2011; Tubbs and Nussenzweig, 2017). A special source of DNA damage are different anti-cancer chemotherapies and radiotherapies (Venkatesan et al., 2017). Examples of such treatment-induced DNA damage are:

- Chemotherapeutic drug Temozolomide (TMZ), an alkylating agent, which transfers an alkyl group to DNA purines, such as creating O^6-methylguanine (Zhang et al., 2012). A hypermutation phenotype (mostly C:G>T:A mutations) was

observed in patients treated with TMZ and the phenotype was linked to resistance to TMZ (Venkatesan et al., 2017).

- Platinum-based compounds, such as cisplatin, cause not only DNA adducts, but also interstrand cross-links, or intrastrand cross-links between adjacent guanines (Hu et al., 2016).

- Phototherapeutic agents, such as psoralen, which leads to mutations in TpA dinucleotides (Helleday et al., 2014).

- Ionizing radiation causes double-strand breaks that lead to a large number of deletions, uniformly distributed across the genome (Behjati et al., 2016).

Understanding the mechanisms of mutagenesis is therefore not only important in order to know how to improve cancer prevention, but also for design of anti-cancer therapies and understanding of resistance to the existing anti-cancer therapies.

1.4 Influence of DNA modifications on mutagenesis

The most important influence of DNA modifications on DNA mutagenesis is undoubtedly the spontaneous deamination of 5mC causing C to T transition, mostly in a CpG context (CpG>TpG). This is the most common type of mutations observed in cancer and genetic disorders (Cooper et al., 2010; Alexandrov et al., 2013a; Lawrence et al., 2013). CpG>TpG mutations form a basis of the most frequent mutational signature (signature 1), which is present in nearly all cancers (Alexandrov et al., 2013a; Wellcome Trust Sanger Institute, 2017). This signature is one of only two mutational signatures with clock-like properties, correlating with the age of patient, and therefore likely to be operating in normal somatic cells throughout life (Alexandrov et al., 2015). It is not only the major mutation type observed in cancer samples, but also in somatic mutations in healthy tissue (Blokzijl et al., 2016) and it is the most frequent type of germline variants (Kong et al., 2012; Rahbari et al., 2015). Moreover, spontaneous deamination of 5mC is thought to be the reason for depletion of CpG dinucleotides in the vertebrate

genomes, as the only regions not depleted of CpG sites, the CpG islands, are largely unmethylated (Jones and Baylin, 2002).

Mutations in CpG motifs are therefore often described as solely resulting from spontaneous deamination and DNA modifications are generally viewed as inducing mutations (Roberts and Gordenin, 2014). However, evidence in the literature shows that the role of cytosine modifications in mutagenicity goes beyond the well-described spontaneous deamination. Moreover, the effect of DNA modifications is not in all cases only pro-mutagenic. Nevertheless, many questions about the role of individual DNA modifications in different mutational processes remain unanswered. Here I review the current knowledge and highlight the unknown parts, which form the basis of motivation for chapters 3 and 4.

1.4.1 Spontaneous deamination

Both C and 5mC can undergo spontaneous hydrolytic deamination. The deamination rate of 5mC is two to four fold higher than that of C (Lindahl and Nyberg, 1974; Shen et al., 1994) and the deamination product, thymine, is a natural base in the DNA and therefore thought to be less efficiently repaired (Lindahl, 1993; Bellacosa and Drohat, 2015). BER is the main pathway responsible for repair of deamination-induced mismatches. Uracil is efficiently excised from U:G mismatches by UNG2, complemented by SMUG1, MBD4, and TDG (Jacobs and Schär, 2012; Krokan et al., 2014). Repair of T:G mismatch is less straightforward, as a canonical base (T) needs to be excised from the pair, to prevent a fixation into a C:G>T:A mutation. This specific excision is performed by two DNA glycosylases: TDG uses sequence specificity, as it removes T from a TpG dinucleotide, and MBD4 binds methylated CpG sites so that the glycosylase domain can more efficiently find deaminated 5mC bases (Bellacosa and Drohat, 2015). However, the efficiency of this repair is thought to be suboptimal (possibly in order to prevent incorrect excisions), and thus leaves relatively high numbers of 5mC:G>T:A mutations. Although 5hmC can also spontaneously deaminate, its effect on mutagenesis was largely unknown.

1.4.2 UV/sunlight mutagenesis

Most of the observed UV-induced mutations in human skin cells are C>T and CC>TT transitions, enriched in a TCG context, and are thought to result from UV-induced formation of CPDs (Brash, 2015). Methylation of cytosine increases the frequency of CPD formation by sunlight (Tommasi and Pfeifer, 1997) and UVB exposure (Mitchell, 2007; Rochette et al., 2009), but not UVC exposure (Rochette et al., 2009). It was proposed that this is due to ca. 5-fold higher molar absorption coefficient of methylated vs. unmethylated cytosine at 290 nm (UVB), but more similar values (1.3-fold lower for 5mC vs. C) at 254 nm (UVC) (Pfeifer, 2000; Schmidt et al., 2006; Rochette et al., 2009), but other conformational and electronic factors are likely to be involved (Martinez-Fernandez et al., 2017).

Bypass of T, C, and 5mC in CPDs by polymerase η is efficient and mostly error-free (Yu et al., 2001; McCulloch et al., 2004; Johnson et al., 2005; Vu et al., 2006; Song et al., 2012). However, C and 5mC in CPDs are unstable and spontaneously deaminate within hours to days due to loss of aromatic stabilization, compared to deamination half-life of thousands of years when in undamaged double-stranded DNA outside CPD (Cannistraro and Taylor, 2009). The deamination rate is strongly affected by the sequence context, with the highest rate for TCG and the lowest rate (*in vitro* ca. 50-fold lower) for CCG context (Cannistraro and Taylor, 2009). The increased deamination rate in TCG was observed after UVC exposure and further increased when exposed to longer wavelengths (UVB, UVA2, UVA1) (Ikehata et al., 2015). The TCG preference is likely dependent on Pol η, because sequencing of Xeroderma pigmentosum variant (XP-V, $Polh^{-/-}$) mouse model shows a relative decrease of TCG>T mutations and increase of C>T mutations in other sequence contexts compared to $Polh^{+/+}$ mice (Ikehata et al., 2014).

As most of the CpGs are methylated in human DNA (Bird and Taggart, 1980), also the deamination rate in TCG context has been mostly studied in methylated cytosine in the CPD. However, the influence of methylation on the deamination rate of cytosine in CPD is less clear. Cannistraro and Taylor (2009) measured in vitro deamination rates of C- and 5mC-containing CPDs in duplex DNA using site-specifically radiolabelled

nucleotides and showed that methylation slows deamination by a factor of 1.2–3.8, depending on the sequence context. Lee and Pfeifer (2003) measured deamination rate of CPD in methylated and unmethylated supF shuttle vector irradiated with UVB and then incubated at 37 °C to allow time for deamination before passage through a human cell line to establish mutations using the mismatch glycosylase activities of MBD4 protein in combination with ligation-mediated PCR. The methylated plasmids contained a relative increase of mutations in a CpG context (45/87) compared to unmethylated plasmids (18/86), transfected after 96h. However, there was no major difference in the mutant frequency between unmethylated (23.84×10^{-3}) and methylated DNA (22.30×10^{-3}).

Interestingly, the formation of CPD and the deamination rate of 5mC in a TCG context in CPD depend on the rotational positioning in the nucleosome: positioning away from the nucleosome surface has two-fold higher frequency of CPD formation and 42-fold higher deamination rate than positioning against the histone core surface (Song et al., 2011; Cannistraro et al., 2015). Similarly, genome-wide single-nucleotide resolution mapping of CPDs in yeast genome by CPD-seq revealed a strong inhibition of CPD formation in nucleosomal DNA with an inward rotation setting (Mao et al., 2016).

In conclusion, two mechanisms of CPD-induced mutagenesis have been proposed (Pfeifer et al., 2005) (Fig. 1.14). In the first mechanism, cytosine in the CPD deaminates into uracil (or 5mC deaminates into thymine), which is then "correctly" paired with adenine by Pol η during replication, causing a C>T mutation. In the second mechanism, cytosine in the CPD is directly paired with adenine during replication by an error-prone TLS polymerase (different than Pol η), such as Pol ι or Pol δ, with an extension by Pol ζ or Pol κ (Ikehata et al., 2014). The first mechanism is thought to prevail in Pol η-proficient cells, while the second is likely to act in Pol η-deficient cells (Ikehata et al., 2015).

The increased CPD formation in methylated cytosine would suggest that methylation increases the frequency of skin mutagenesis, but it has never been verified in the sequencing data sets of cancer somatic mutations[10]. Also the effect of nucleosome rotational positioning has not been explored in the actual cancer mutation data sets.

[10] One study looked at this relationship simultaneously with us; the results are compared and discussed in chapter 4.

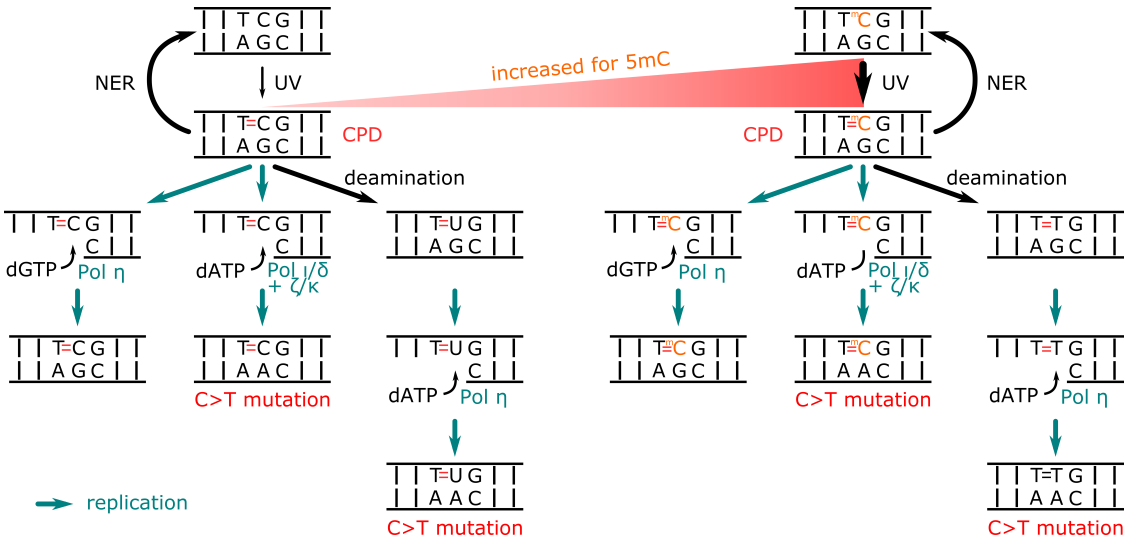

Figure 1.14. A model of the UV-induced mutagenesis and the role of 5mC in this process; based on known literature. Formation of cytosines-including cyclobutane pyrimidine dimers (CPDs) after UV light exposure is enhanced by methylation (left: unmethylated scenario, right: methylated scenario). If the CPD is not repaired by NER, it can: (a) be correctly replicated by Pol η, (b) be erroneously replicated by a different polymerase, (c) deaminate and be paired with A by Pol η, leading to a C:G>T:A mutation.

Finally, the combined effects of methylation and nucleosome positioning are yet to be determined.

1.4.3 Tobacco smoking mutagenesis

The most common mutations resulting from tobacco-induced damage are C:G>A:T transversions, often due to mispairing of bulky DNA adducts on guanine, such as BPDE-dG or Acr-dG, during replication. Mapping of BPDE adducts in the human P53 gene has shown that BPDE binds preferentially at guanines in a CpG context, previously observed as mutational hotspots in lung cancer, and that this preference is dependent on cytosine methylation (Denissenko et al., 1997). The preference is strongest for guanine which is directly paired with 5mC, compared to methylation in other neighbouring positions (Guza et al., 2011). This enhancement of adduct formation by 5mC is likely due to pre-covalent intercalative complexes with BPDE and effects of 5mC on altered diastereomeric composition of the resulting DNA adducts (Guza et al., 2011). The pre-covalent binding model is also in line with an observed increased BPDE binding

constant by conformational and hydrophobicity changes of 5mC in a CpG context in codon 248 of the *TP53* gene (Malla et al., 2017).

The preference for a CpG context in tobacco-smoking C>A mutagenesis is present in human lung cancer samples (Alexandrov et al., 2013a), in vitro exposure of cells to B[a]P (Nik-Zainal et al., 2015), BPDE-treated embryo fibroblasts derived from Xpa-knockout mice (deficient in both TC-NER and GG-NER) crossed with human TP53 knock-in mice, in sperm and bone marrow cells of B[a]P exposed mice (O'Brien et al., 2016) and in lung of B[a]P treated mice (Aoki et al., 2015). Therefore, although involvement of other tobacco carcinogens in the CpG>ApG cancer mutagenesis cannot be excluded (such as Acr-dG adducts (Feng et al., 2006; Wang et al., 2013) or oxidation of guanine in 5mCpG (Ming et al., 2014)), the similarity of mutational properties after B[a]P/BPDE treatment with mutations observed in lung cancer patients confirms the major role of B[a]P carcinogen in the mutagenesis in lung cancer.

The effect of tobacco smoking on C>G and C>T mutations in a CpG context is less clear. Increase of CpG>GpG mutations in bone marrow but not sperm after B[a]P exposure (O'Brien et al., 2016) suggests tissue specific effects. Moreover, some of the effects might be dependent on the time of exposure or other conditions, as B[a]P treatment led to ca. two-fold decrease of CpG>TpG mutations after 3 month, but then ca. two-fold increase after 24 month, compared to age-matched controls (Aoki et al., 2015).

In summary, the experimental evidence predicts increased tobacco-induced mutagenesis in 5mC:G pairs compared to C:G pairs. The BPDE-dG adduct can be then replicated in an error-free (such as by Pol κ (Avkin et al., 2004; Jha et al., 2016)), or error-prone (by Pol η (Zhao et al., 2006; Klarer et al., 2012)) manner, paired with adenine on the daughter strand, and by that creating an C>A mutation (Fig. 1.15). The C>A mutations should be therefore positively correlated with methylation levels. However, this has never been verified on a genome-wide scale in human cancer samples. The only indirect evidence is an observed decrease of C>A mutation frequency inside CpG islands in small-cell lung cancer cell line (Pleasance et al., 2010). However, this could be influenced by the regional differences of CpG islands (and associated repair), which occur near transcription start sites and can be affected by bound transcription factors. Moreover,

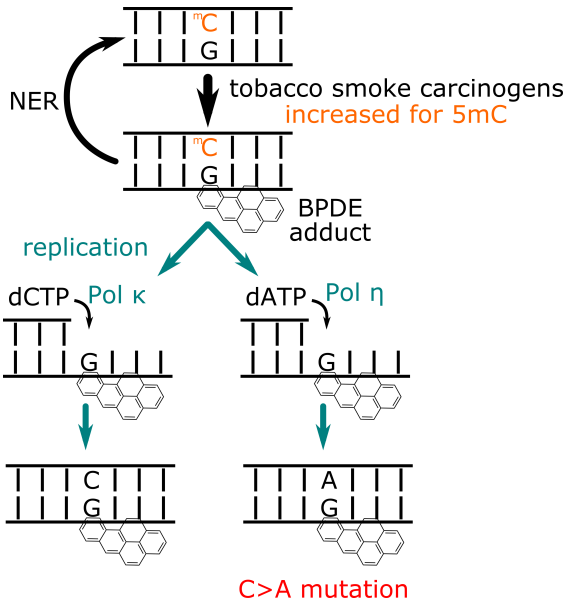

Figure 1.15. A model of the tobacco-induced mutagenesis and the role of 5mC in this process; based on known literature. Formation of BPDE adducts on guanine is enhanced by methylation of the opposite cytosine. If the BPDE-dG adduct is not repaired by NER, it can be erroneously paired with adenine during replication, creating a A:G-BPDE mismatch, which would be in the next replication fixated into a C:G>A:T mutation.

the effect of methylation on C>T and C>G tobacco-induced mutations is unknown, as well as the role of other DNA modifications in the tobacco-induced mutagenesis.

1.4.4 APOBEC/AID mutagenesis

Much research on AID/APOBEC activity has been fuelled by a question whether they could play a role in active demethylation. In the proposed model, AID/APOBEC enzymes deaminate 5mC, 5hmC, 5fC, or 5caC, and the deamination product is excised by BER, and unmodified C is restored (Teperek-Tkacz et al., 2011; Nabel et al., 2012). Therefore, the deamination activity of AID/APOBEC enzymes on different modifications of cytosines has been extensively researched:

- APOBEC3A efficiently deaminates both C and 5mC, but the efficiency is ca. 5–10-fold lower for 5mC (Carpenter et al., 2012; Nabel et al., 2012; Wijesinghe and Bhagwat, 2012; Siriwardena et al., 2015; Schutsky et al., 2017). Deamination of the higher oxidative states by APOBEC3A is markedly less efficient compared

to unmodified C, the proficiency decrease was estimated as 5600-fold for 5hmC, 3700-fold for 5fC, and more than 20,000-fold for 5caC (Schutsky et al., 2017).

- APOBEC3B showed even lower deamination activities on 5mC, estimated as 50-fold less than for C (Fu et al., 2015) and thousands-fold less than APOBEC3A.
- APOBEC3G almost exclusively prefers C (Carpenter et al., 2012; Wijesinghe and Bhagwat, 2012), estimated as 100-fold decreased activity on 5mC than C and no detected activity on 5hmC (Kamba et al., 2015).
- Most similar values of 5mC and C deamination activity were observed for APOBEC3H. The values of $100 \cdot 5mC/C$ preference were 85 for APOBEC3H haplotype II (hap II), 29 for APOBEC3H hap VII, 15 for APOBEC3H hap I, and 13 for APOBEC3H hap V, compared to 13 for APOBEC3A, and 2 for APOBEC3B and AID (Gu et al., 2016).
- Finally, AID showed a strong preference for deaminating unmodified cytosine, with a ca. 10-fold lower efficiency for 5mC (Nabel et al., 2012; Wijesinghe and Bhagwat, 2012; Siriwardena et al., 2015) and even larger decrease for 5hmC (Nabel et al., 2012; Rangam et al., 2012).

In summary, all the measurements show higher efficiency for unmodified than modified cytosine, making a role of AID/APOBEC in active demethylation unlikely.

Based on these experimental observations, we would expect the APOBEC-induced mutations to be happening mostly in unmodified cytosine. Since APOBEC3B shows most convincing evidence for causing cancer mutations, the expected difference would be 50-fold. The second most discussed mutagenic APOBEC enzyme is APOBEC3A, where the expected difference would be 5-10-fold.

However, this has never been verified in human cancer samples. The only current evidence from human cancer sequencing was done last year by comparing mutation frequencies of APOBEC-rich and poor samples, in mostly unmodified vs. modified positions (Seplyarskiy et al., 2016b). The authors observed ca. two-fold lower frequency of TCG mutations in modified cytosines compared to unmodified cytosines. Although

the data are in line with the expected direction of 5mC vs. C mutagenesis, the difference is much lower than expected. However, the authors used RRBS-seq derived modification maps, which cover only approximately 3.7 % of the genome (Stirzaker et al., 2014). Further research is therefore needed to determine the role of different DNA modifications in the APOBEC-induced mutagenesis.

1.4.5 The role of 5hmC, 5fC, and 5caC in mutagenesis

The role of 5hmC, 5fC, and 5caC in DNA mutagenesis is largely unexplored. Of these three DNA modifications, the effects of 5hmC would be most interesting, as it is the second most abundant modification and it is enriched in functionally important regions of the genome (exons, especially highly transcribed exons, and enhancers). Little is known about the frequency and mutagenicity of spontaneous deamination of 5hmC. Experimental data suggest that 5hmC should be highly protected from APOBEC-induced mutagenesis. 5hmC does not show increased formation of CPDs after UV exposure and in some sequence contexts, CPDs containing 5hmC are formed at very low levels (Kim et al., 2013), but more detailed data about the role of 5hmC in UV-induced or tobacco-induced mutagenesis are missing.

In contrast, 5fC and 5caC show several links to DNA damage/mutagenesis. *In vitro* mutagenic assay experiments of base-pairing stability and primer extension showed that 5fC and 5caC are only marginally mutagenic (5fC was paired with adenine by DNA polymerases Klenow exo$^-$, Pol η, and Pol κ in ca. 1–2 % of measurements) (Münzel et al., 2011). However, 5fC and 5caC have been suggested to cause a range of C>G, C>A, and C>T mutations *in vivo* (Kamiya et al., 2002; Xing et al., 2013). 5caC:G pairs can be recognised as a mismatch by proofreading of Pol δ and MutSα of MMR during replication (Shibutani et al., 2014), and both 5fC:G and 5caC:G pairs are recognised and excised by TDG (Maiti and Drohat, 2011). These observations have been speculated to underlie C>G (Supek et al., 2014a) or C>T (Mahfoudhi et al., 2016) mutagenesis. Exploring to what extent these modifications impact mutagenesis in cancer patients might be however challenging due to their low abundance in the genome: they are 2–4 orders of magnitude less frequent than 5hmC (Liu et al., 2013).

1.5 Influence of DNA replication on mutagenesis

The role of DNA replication in mutagenesis is often viewed solely as random misincorporation of wrong bases by the replicative polymerases. Such view was also used in a recent study published in Science (Tomasetti and Vogelstein, 2015), which showed that the lifetime risk of cancer correlates with the number of stem cell divisions in the lifetime of a given tissue. The authors interpreted this correlation as being due to random mutations introduced during replication and suggested that most of the variation in cancer risk is due to "bad luck" and therefore, in many cancer types, early cancer detection will be more effective than cancer prevention strategies (Tomasetti and Vogelstein, 2015). The study started a heated debate in the field, especially due to the suggested interpretation. One of the criticisms was that many different mutational signatures were found in cancers and associated with a number of internal and external mutagenic processes different from random mistakes during replication (Gao et al., 2016; Crossan et al., 2015). It is however unknown to what extent and in which ways replication influences all these mutational processes. This question was the main motivation for chapter 4. Here, I summarise the known replication-linked mutational processes (Fig. 1.16).

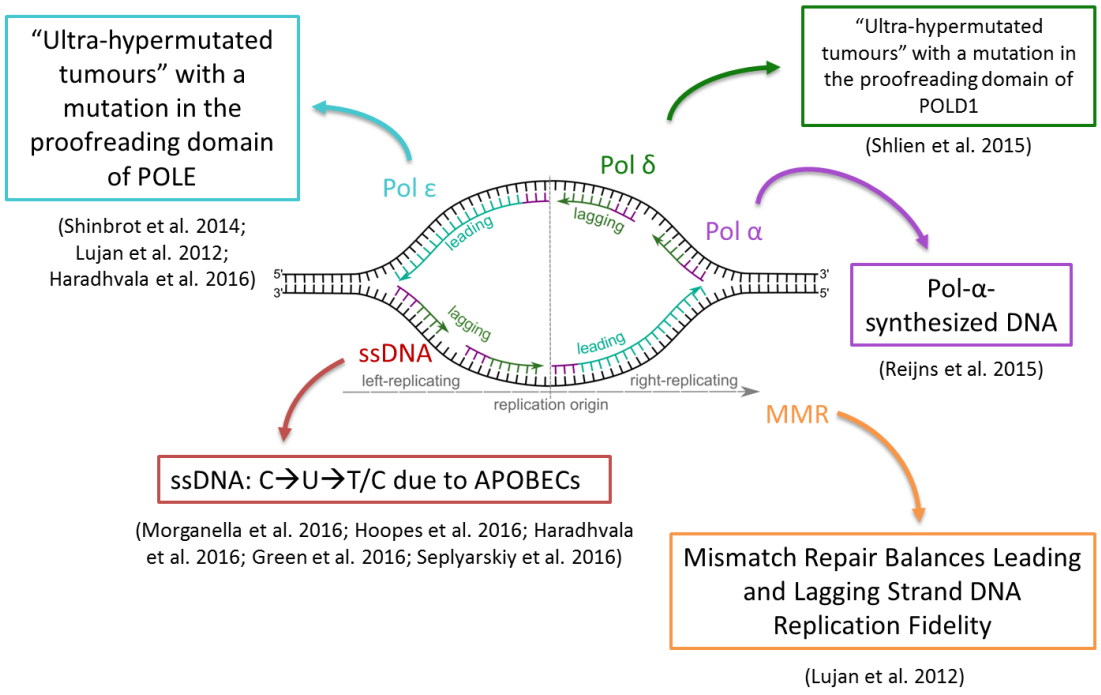

Figure 1.16. A schema of known mutagenic processes at the replication fork.

1.5.1 Errors made by replicative polymerases Pol ε and δ

As the vast majority of the human genome is synthesised by the replicative polymerases Pol ε and Pol δ, they need to be extremely accurate. Their fidelity results from high accuracy in base selection (with an error-rate of 10^{-5}) and proofreading by the exonuclease domain. This is completed by MMR proofreading, leading to a remarkable fidelity of replication: the total *in vivo* replication error-rate has been estimated to be $10^{-10} - 10^{-9}$ per base (Rayner et al., 2016; Lange et al., 2011; McCulloch and Kunkel, 2008; Loeb, 1991).

Disruptions in the two layers of proofreading dramatically increase the mutation rate, increasing the risk of cancer. Germline mutations in *POLE* and *POLD1* predispose individuals to intestinal and colonic polyposis (causing a syndrome named *Polymerase proofreading-associated polyposis* (PPAP)), to colorectal cancer, endometrial cancer, and other malignancies (Rayner et al., 2016). Somatic mutations in these genes are also found in 1–2 % of sporadic colorectal cancers, 7–12 % of sporadic endometrial cancers, and with lower frequencies in tumours of brain, pancreas, ovary, and other tissues (Rayner et al., 2016). Germline mutations in MMR cause Lynch syndrome, characterised by microsatellite instability and increased cancer risk in colon/rectum, endometrium, ovary, stomach and other tissues (Rayner et al., 2016). Lynch syndrome accounts for 3 % of colorectal patients and somatic MMR deficiency is found in 15–20 % sporadic colorectal cancer patients (Lynch et al., 2009; Poulogiannis et al., 2010). Most of the cancers due to mutations in *POLE*/*POLD1* are microsatellite stable and the combined MMR deficiency and DNA polymerase proofreading deficiency is synthetic lethal in *S. cerevisiae* and mice (Albertson et al., 2009; Herr et al., 2014), suggesting a possibility that the combined deficiency increases the mutation frequency above a threshold compatible with cancer-cell survival. Such a plateau of mutation burden was also observed in highly mutated patients with inherited biallelic mismatch repair deficiency and acquired DNA polymerase proofreading deficiency (Shlien et al., 2015).

The vast majority of cancer-associated *POLE*/*POLD1*-mutations are located in the proofreading domain (Mertz et al., 2017b; Rayner et al., 2016). This suggests that the high mutation burden in these cancers is caused by deficiency in the proofreading activity of

the polymerase and therefore reflects the natural errors made by the polymerase. The spectrum of mutations found in these cancers is very unique. They exhibit markedly elevated TCT>TAT, TCG>TTG, and TTT>TGT mutations (Shinbrot et al., 2014; Shlien et al., 2015), identified as mutational signature 10 (Alexandrov et al., 2013a). Some of the ultramutated *POLE*-mutated cancers also contain a strong component of mutational signature 14, characteristic by NCT>NAT (where N means any base) and C>T mutations (Wellcome Trust Sanger Institute, 2017). The C:G>A:T mutations in *POLE*-mutated cancers exhibit a replication strand bias with an enrichment of C>A mutations of the leading strand, as expected by the most accepted model of replication with Pol ε synthesising the leading strand (Shinbrot et al., 2014; Haradhvala et al., 2016). The spectrum of mutations in ultramutated *POLD1*-mutated cancers is very different, with an enrichment of CCN>CAN, C>T, T>A, and T>C mutations (Shlien et al., 2015). Enrichment of CCN>CAN mutations was also observed in *S. cerevisiae* with the same variant in the yeast gene encoding Pol δ (Lujan et al., 2014).

It is unknown why some of the sequence contexts are much more mutated than others. It has been suggested that the mechanisms of mutagenicity in these cancers might be linked to altered levels of dNTP pools (Williams et al., 2015; Mertz et al., 2015; Flood et al., 2015), or affected DNA binding of the polymerase domain (Church et al., 2013), potentially reducing efficiency of extrinsic proofreading (Barbari and Shcherbakova, 2017). More research is however needed to unravel the mechanisms of mutagenesis in this important group of highly mutated cancers.

1.5.2 Errors made by replicative polymerases Pol α

Polymerase Pol α is the least accurate of the three replicative human polymerases and lacks a proofreading domain (Walsh and Eckert, 2014). As Pol α initiates the synthesis on the leading strand and each Okazaki fragment by providing RNA primer and synthesising approximately 20–30 bases of DNA (Lang and Murray, 2011) (described in section 1.1.3), it could introduce dangerous mutations into the DNA.

Three mechanisms may be therefore involved in suppressing DNA mutations resulting from the errors due to Pol α (Zheng and Shen, 2011). First, the errors might be

proofread by the exonuclease domain of Pol δ which was suggested to form a complex with Pol α. Second, the Pol α-synthesised DNA can be removed together with the RNA primers in strand displacement activity of Pol δ. Third, the errors incorporated by Pol α are recognised by MMR. Nevertheless, it has been shown in yeast that a part of the Pol α-synthesised DNA is not removed, comprising approximately 1.5 % of the mature genome (Reijns et al., 2015). It was proposed that this is due to DNA-protein binding proteins that rapidly re-associate after replication and act as partial barriers to Pol-δ-mediated displacement of Pol-α-synthesized DNA (Reijns et al., 2015). The impact of these observations on human mutagenesis is currently unknown.

1.5.3 Errors made by TLS polymerases

TLS polymerases represent an important source of mutations, by incorporating incorrect bases, either by inserting damaged dNTPs or rNTPs, or by performing an error-prone bypass of DNA lesions, as was described in sections 1.3.2 and 1.3.3.4. It is still largely unexplored to what extent these different mutagenic activities of individual TLS polymerases taint the human genome. Pol η has been linked to mutational signature 9 comprising of a spectrum of T>G (enriched in ATN and TTN context), T>C, and C>A mutations, as this signatures was observed in cancers that have undergone somatic immunoglobulin gene hypermutation, in which Pol η is known to be involved (Alexandrov et al., 2013a).

1.5.4 Mutagenesis prevented by MMR

As summarised in section 1.3.1.5, the canonical role of MMR is to prevent mutations introduced during DNA replication. These include mostly insertion/deletion loops and to a lesser extent single nucleotide mismatches. Cancers deficient in MMR exhibit a hypermutation phenotype with high amounts of C>T, T>C, and C>A mutations (Zhao et al., 2014a). The C>T mutations of MMR-deficient tumours observed enriched on the lagging strand (Haradhvala et al., 2016). It was proposed that this is due to MMR's role in balancing mutational asymmetries generated during DNA replication, in particular by Pol δ and Pol α on the lagging strand (Lujan et al., 2012; Andrianova et al., 2017).

Next to mutation prevention, MMR is thought to have also non-canonical pro-mutagenic functions in certain contexts, such as in antibody maturation (Peña-Diaz and Jiricny, 2012; Crouse, 2016). The non-canonical MMR can have also detrimental effects, as it promotes repeat expansions associated with neuromuscular and neurodegenerative diseases and may contribute to cancer mutagenesis (Bak et al., 2014; Crouse, 2016; Peña-Diaz et al., 2012).

1.5.5 Damage to single-stranded DNA on the lagging strand

The discontinuous nature of DNA synthesis on the lagging strand leads to formation of single-stranded DNA. The single-stranded DNA is normally protected by RPA. The template of lagging strand is thought to be single-stranded for a longer period of time than the leading strand template (Okazaki et al., 1968; Seplyarskiy et al., 2016b; Hoopes et al., 2016). The exposed single-stranded DNA is prone to different types of damage. The best described type of such damage, which attracted much attention in the last four years, is cytosine deamination by AID/APOBEC enzymes (Hoopes et al., 2016; Seplyarskiy et al., 2016b; Morganella et al., 2016). This type of damage has been already described in sections 1.3.3.1 and 1.4.4. The C>T and C>G APOBEC-associated mutational signatures are found enriched on the leading strand in human cancers (Seplyarskiy et al., 2016b; Morganella et al., 2016; Haradhvala et al., 2016) and to some extent also in germline mutations (Seplyarskiy et al., 2016a). While mutations are often enriched in late-replicated regions, the APOBEC-associated mutagenesis (especially the C>G mutations and mutational signature 13) have similar frequency in early- and late-replicating regions (Haradhvala et al., 2016; Morganella et al., 2016) and in some cases seem to be even enriched in early-replicating regions (Kazanov et al., 2015). The reason for this observation (and what causes the different in signatures 2 and 13 in this respect) is not fully elucidated. The current model of APOBEC-induced mutagenesis is that APOBEC first deaminates cytosine on a single-stranded DNA into uracil. This can generate C>T mutations, as uracil is paired with adenine during replication. Alternatively, uracil can be excised by UNG and the resulting AP site can be paired with adenine (by the so called “A-rule”), also generating a C>T mutation, or C

is inserted opposite the AP site by REV1, creating a C>G mutation (Morganella et al., 2016). Based on this model, UNG and REV1 are needed to generate the C>G dominated mutational signature 13. It was therefore postulated that the UNG/REV1-dependent mechanism occurs earlier in replication, while later in replication the unrepaired uracils and AP sites are left unrepaired and lead to C>T dominated mutational signature 2 (Morganella et al., 2016).

While the involvement of AID/APOBEC enzymes in cancer mutagenesis is generally accepted, the initiating events leading the their dysregulation and the proportional involvement of the individual enzymes are still unknown (Mertz et al., 2017b).

1.6 Aims of the thesis

The recent advances in sequencing technologies have dramatically changed our means to research cancer mutagenesis. Compared to sequencing studies on single genes in a small number of replicates, we can now study mutagenesis in the entire genomes of thousands of cancer patients. Moreover, novel technologies allow measurements of DNA modifications, other epigenetic marks, DNA replication, DNA expression, and many other genomic features on a genome-wide scale. With the use of high throughput computing, bioinformatics, and mathematics, all these large scale genomic data sets can be integrated to investigate the mechanisms of mutagenesis. For instance, a mathematical method for separation of signals with different sources was recently used to identify signatures of the main mutational processes operating in cancer patients (Alexandrov et al., 2013a). For 16 of the 30 identified signatures, a known underlying mutational process could be identified (Alexandrov et al., 2013a, 2015; Helleday et al., 2014; Wellcome Trust Sanger Institute, 2017). This means that there might be at least 14 mutational processes contributing to mutagenesis in cancer which are currently unknown. Moreover, mechanisms of the 16 known mutational processes are also not always fully understood. However, the currently available large-scale genomic data are far from fully harvested, allowing investigation of the unexplained mechanisms causing mutations observed in cancer samples.

The main goal of this thesis was to utilise large-scale genomic data sets in order to examine how mutagenesis is affected by DNA modifications and DNA replication. The specific aims were:

1. The mutagenic potential of 5mC (by spontaneous deamination) is well documented. However, mutational properties of 5hmC are mostly unknown. The first aim was to investigate how frequently hydroxymethylated positions are mutated in cancer (Chapter 3).

2. The best known mutational process in CpG sites is a spontaneous deamination of 5mC. The second aim was to explore the role of DNA modifications in other processes than spontaneous deamination; in particular we focused on mutational processes associated with replication, UV exposure, tobacco exposure, and APOBEC enzymes (Chapter 4).
3. DNA replication was first thought to induce mutagenesis mainly by misincorporation of bases by the replicative polymerases. Other links between replication and mutagenesis are starting to appear, but it is currently unknown which of the individual mutagenic mechanisms are affected by replication and how. The third aim was therefore to assess the role of DNA replication in individual mutational processes by analysing mutational signatures with respect to replication strand asymmetry and replication timing (Chapter 5).

The publications associated with this thesis are listed in Appendix 7. A majority of Chapter 3 is published in Tomkova et al. (2016). The first half of Chapter 4 and most of the Chapter 5 form two manuscripts, currently in review. A manuscript about bsQC tool described in Chapter 2 is in preparation.

Kayinga voyo tethe
Teyinga yoh wodeke
Wumu bada pafisinga
Cungada bamise

— Alex Boyé *Peponi*

Teenage Mutant Ninja Turtles
Teenage Mutant Ninja Turtles
Teenage Mutant Ninja Turtles
Heroes in a half shell
Turtle power!

— *Teenage Mutant Ninja Turtles Theme Song*

2 General methods

This thesis makes use of a number of genomic data sets: measurements of mutation, DNA modifications, replication, and other genomic features. Although details of methods used and newly developed in this thesis are given in each chapter separately, here I summarise the main aspects of the sequencing techniques of the used data sets and their bioinformatics processing. Apart from one exception of a newly developed package (section 2.2.1), this chapter does not contain new methods development, but it is instead an overview of methods used in this thesis, which generally represent gold-standard approaches used in the cancer genomics field. In the entire thesis, the genome build GRCh37 (hg19) is used.

2.1 Mutations

Since the time when the first human genome was sequenced, the sequencing costs have undergone an incredible million-fold drop (Wetterstrand, 2016), mainly thanks to development of "second-generation" or "next-generation" sequencing (NGS), the first truly high-throughput sequencing platforms in the mid-2000s (Goodwin et al., 2016). These technologies are now routinely used in research and increasingly being applied in clinical diagnosis (Hardwick et al., 2017). They allow identification of inherited mutations in exomes (i.e., the part of the genome formed by exons) or even whole

genomes. Sequencing of somatic mutations is possible, as long as they are clonal (existing in the vast majority of cells in the sample) or the clonality is sufficiently high (Martincorena and Campbell, 2015). The reason for this limitation is the cost of sequencing and the relatively high sequencing error rate of the used sequencing methods (Wall et al., 2014). The classical identification of variants (*variant calling*) is therefore based on sequencing a number of cells at the same time. A position is then called to carry a variant only if the variant is present in a substantial proportion of cells. Therefore most of the current knowledge about somatic mutagenesis comes from cancer genomes, as tumours evolve through clonal expansion originated from a single cell (Greaves and Maley, 2012; Schwartz and Schäffer, 2017). Sequencing of non-clonal, rare, mutations is still challenging; nevertheless a number of promising methods have emerged in the last three years (Jee et al., 2016; Hoang et al., 2016).

The typical process to measure mutations using whole-genome sequencing (WGS) contains the following main steps (more detailed information can be found in (Nielsen et al., 2011; Goodwin et al., 2016)):

- DNA library preparation (extraction of DNA from cells, cutting the DNA into random fragments, e.g., around 300 bp long, ligating adapters to ends of the fragments),
- sequencing of the fragments from one end (*single-end sequencing*) or both ends (*paired-end sequencing*) and storing this information into *reads* (typically, each position should be covered by 30 reads on average),
- aligning the reads to a reference genome,
- various filtering steps, such as removal of duplicate reads (technical artefacts from polymerase chain reaction (PCR) used to amplify the library, or from optical sensors) and re-aligning long insertions/deletions for better detection of chromosomal rearrangements,
- calling variants in individual position.

The steps for whole-exome sequencing (WES, WXS) are similar, but the library preparation involves a step for capturing DNA fragments from exome only.

In this thesis, a number of publicly available WGS and WES datasets are used (Tables 9.3 and 9.4). In all but TCGA WGS data sets, variant calls are publicly available. For TCGA WGS, we downloaded the aligned reads for tumour and normal samples from the UCSC CGHub website under TCGA access request #10140 and called somatic variants using Strelka (Saunders et al., 2012) with default parameters. In most analyses, we focus on single-nucleotide variants (SNVs) only.

2.1.1 Mutational signatures

Individual mutational processes leave footprints on the genomes in the form of mutations of different types (SNVs, small indels, rearrangements, and copy number changes) and different genomic and sequence contexts. For example, as summarised in the introduction 1, UV light induces C>T transitions enriched in T<u>C</u>G trinucleotides and aristolochic acid causes an increased rate of A>T mutations in a [C|T]<u>A</u>G context. Identification of such characteristics of all main mutational processes in the sequenced cancer genomes was a motivation of a recent approach called mutational signatures (Nik-Zainal et al., 2012a; Alexandrov et al., 2013a,b). All base substitutions are first classified into six subtypes: C:G > A:T, C:G > G:C, C:G > T:A, T:A > A:T, T:A > C:G, and T:A > G:C (these subtypes are later called according to the mutated pyrimidine, i.e., C>N and T>N) and 16 possible sequence contexts of the mutated base (5' and 3'), giving a total of 96 possible mutation types.

The numbers of mutations of the 96 types in n samples can be written in a matrix $M \in (\mathbb{Z}_{\geq 0})^{96 \times n}$:

$$M = \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,n} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{96,1} & m_{96,2} & \cdots & m_{96,n} \end{pmatrix} \tag{2.1}$$

where columns correspond to samples, rows to the 96 mutation types, and each $m_{i,j}$ represents the number of mutations of type i in sample j, such as number of T<u>C</u>G>T<u>T</u>G mutations in sample j. The matrix $P \in (\mathbb{R}_{\geq 0})^{96 \times K}$ of mutational signatures is defined as:

$$S = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,K} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ s_{96,1} & s_{96,2} & \cdots & s_{96,K} \end{pmatrix} \tag{2.2}$$

where columns correspond to individual signatures, rows to the 96 mutation types, and each $s_{i,k}$ represents the component of the i-th mutation type in the k-th mutational signature. Each mutational signature is normalised to sum to one:

$$\forall k : \sum_{i=1}^{96} s_{i,k} = 1 \tag{2.3}$$

Finally, the goal is to find a matrix of exposures $E \in (\mathbb{R}_{\geq 0})^{K \times n}$:

$$E = \begin{pmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,n} \\ e_{2,1} & e_{2,2} & \cdots & e_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{K,1} & e_{K,2} & \cdots & e_{K,n} \end{pmatrix} \tag{2.4}$$

where columns correspond to individual signatures, rows to the samples, and each $e_{j,k}$ represents exposure to the k-th mutational signature in the j-th sample, such that:

$$M \approx S \times E \tag{2.5}$$

In other words:

$$\forall \text{ mutation type } i, \text{ sample } j : m_{i,j} \approx \sum_{k=1}^{K} s_{i,k} \cdot e_{k,j} \tag{2.6}$$

Such defined problem exactly corresponds to the so-called non-negative matrix factorisation (NMF), in which a known matrix M is factorised into unknown matrices S and E, with the property that all three matrices do not contain negative values. The exact definition of the problem is to find non-negative matrices S and E minimising the error function:

$$||M - S \times E|| \tag{2.7}$$

A number of algorithms can be used to solve the NMF problem, the most popular being *multiplicative update rule* (Lee and Seung, 1999), in which S and E are initialised randomly and then iteratively updated, until convergence or the maximum number of iterations are reached. This solution was also used by Alexandrov et al. (2013a,b).

Alexandrov et al. (2013a) applied this approach on 507 WGS and 6 535 WES samples of 30 different cancer types (Alexandrov et al., 2013a) and extracted more than 20 mutational signatures, which were later extended into 30 mutational signature (Alexandrov, 2015; Wellcome Trust Sanger Institute, 2017) and novel signatures are likely to be detected in the future. The current knowledge about the detected signatures is summarised in Fig. 2.1 and Table 2.1 and most up-to-date information is kept at http://cancer.sanger.ac.uk/cosmic/signatures. In the remainder of the thesis, these signatures are therefore referred to as "COSMIC signatures".

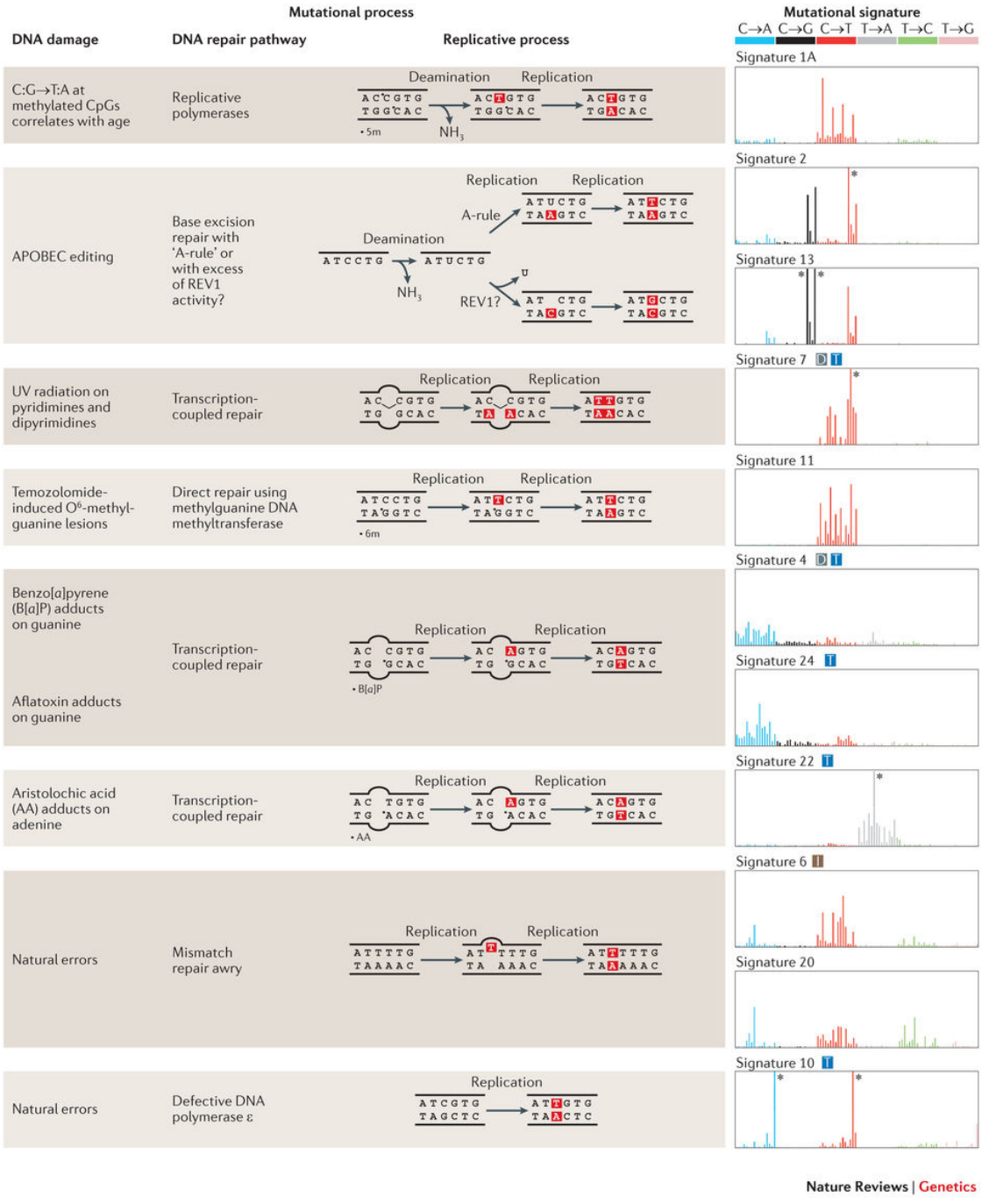

Figure 2.1. Summary of known mutational signatures, and the components of DNA damage and repair that constitute the mutational processes. Reprinted from Helleday et al. (2014), with permission from the publisher.

Sign.	Mutations	Cancer types/tissues	Proposed aetiology	Associations and comments
1	C>T at CpG	All cancer types	Spontaneous deamination of 5mC	Correlates with age of cancer diagnosis
2	C>T at TpC	Many cancer types	APOBEC enzymes	Replication strand bias in breast
3	Many types	Breast, ovary, and pancreas	Failure of DSBR by HR	Associated with mutations in BRCA1/2. Associated with larger indels
4	C>A	Lung, head and neck, liver, oesophagus	Tobacco smoking carcinogen benzo[a]pyrene	Tx strand bias
5	Many types	All cancer types	Unknown	Correlates with age of cancer diagnosis, Tx strand bias for T>C at ApT
6	C>A, C>T	Colorectal and uterus	Defective mismatch repair	Small indels at repeats
7	C>T	Skin cancers	Ultraviolet light exposure	Tx strand bias
8	C>A	Breast, medulloblastoma	Unknown	Weak tx strand bias
9	T>G	CLL and malignant B-cell lymphomas	Pol η and AID-mediated somatic hypermutation	Elevated in CLL samples with immunoglobulin gene hypermutation
10	TCT>TAT, TCG>CTG, TTT>TGT	POLE-mutated cancers	Errors in Pol ε	Tx strand bias for C>A and T>G mutations
11	C>T	Melanoma and GBM	Temozolomide treatment	Strong tx strand bias
12	T>C	Liver	Unknown	Strong tx strand bias
13	C>G at TpC	Many cancer types	APOBEC enzymes	Replication strand bias in breast
14	C>A, C>T at GpC	Uterus and low-grade glioma	Unknown	Hypermutation
15	C>T at GpC	Gastric, lung	Defective mismatch repair	Small indels at repeats
16	T>C	Liver	Unknown	Extremely strong tx strand bias for T>C at ATN
17	T>G at TpT and T>C at CpT	Oesophagus, gastric, breast, liver, lung, and other	Unknown	-
18	C>A	Neuroblastoma, breast, gastric	Unknown	-
19	C>T	Pilocytic astrocytoma	Unknown	-
20	C>A, C>T	Gastric and breast	Defective mismatch repair	Small indels at repeats
21	T>C	Gastric	Unknown	Defective mismatch repair
22	T>A	Urothelial and liver	Aristolochic acid exposure	Strong tx strand bias
23	C>T	Liver	Unknown	Strong tx strand bias
24	C>A	Liver	Aflatoxin exposure	Strong tx strand bias
25	T>A	Hodgkin lymphomas	Unknown	Tx strand bias
26	T>C	Breast, cervix, gastric, and uterus	Defective mismatch repair	Small indels at repeats
27	T>A	Kidney clear cell	Unknown	Very strong tx strand bias, associated with small indels at repeats
28	T>G	Gastric	Unknown	-
29	C>A	Gingivo-buccal oral sq. cell carcinoma	Tobacco chewing	Tx strand bias
30	C>T	Breast cancers	Unknown	-

Table 2.1. Summary of known mutational signatures. Compiled from (Wellcome Trust Sanger Institute, 2017; Alexandrov et al., 2013a; Alexandrov, 2015; Morganella et al., 2016).

2.2 DNA modifications

Since the turn of the twenty-first century, a number of methylome profiling techniques have been developed: gel-based, array-based, and sequencing-based methods (reviewed in Beck and Rakyan, 2008; Plongthongkum et al., 2014). Most of the existing data sets are from HumanMethylation450 arrays (HM450, 450K), enrichment-based methyl-DNA immunoprecipitation sequencing (MeDIP-seq) and MBD capture sequencing (MBD-seq), or reduced representative bisulfite sequencing (RRBS-seq). For instance in MeDIP-seq, monoclonal antibodies specific to 5mC are used to enrich for methylated DNA fragments before sequencing. Antibody specific for 5hmC can be used in a similar way to obtain regional measurements of hydroxymethylation, as in hydroxymethylated DNA immunoprecipitation sequencing (hMeDIP-seq).

Although these methods are very popular cost-effective approaches, they cover only a small proportion of the CpGs in the genome: 1.7 % with HM450, 3.7 % with RRBS-seq, and 17.8 % with MBDCap-seq (Stirzaker et al., 2014)). Moreover, the enrichment-based methods do not provide single-base resolution and are not fully quantitative. In order to study the effects of DNA modifications on mutagenesis, large statistical power is needed, and we therefore focus on truly whole-genome sequencing methods with single-base resolution mostly, namely whole-genome bisulfite sequencing (WGBS, BS-seq), and its derivatives: TET-assisted bisulfite sequencing (TAB-seq) (Yu et al., 2012) and oxidative bisulfite sequencing (oxBS-seq) (Booth et al., 2012).

BS-seq is based on bisulfite conversion, a chemical treatment of DNA, in which cytosines are deaminated into uracils. During a following amplification, the uracils are paired with adenines, and read as thymines after sequencing . Methylated cytosines are protected from this conversion, and therefore are paired with guanines, and ultimately read as cytosines (Fig. 2.2). Importantly, hydroxymethylated cytosines are also protected from bisulfite conversion. The sequenced reads need specific bioinformatics pipelines to be mapped to a reference genome. In each position, the number of reads with C (unconverted, i.e., with 5mC or 5hmC) and T (converted, i.e., with C, 5fC, or 5caC) are counted. The modification level is then defined as:

$$\text{mod level} = \frac{\text{unconverted reads}}{\text{unconverted reads} + \text{converted reads}} \tag{2.8}$$

It should be noted that terminology of the BS-seq measurements is not always unified. As the method was used before the discovery of 5hmC in human DNA, the BS-seq measurements were called methylation levels and this is still used in many publications. Since BS-seq measures the combined amount of 5mC + 5hmC, we term the BS-seq measurements "modification levels" (or "mod"). Strictly speaking, this is also not accurate, as both 5fC and 5caC get converted by the bisulfite treatment, and the BS-seq measurements therefore do not represent levels of all types of modifications, but only 5mC and 5hmC (Plongthongkum et al., 2014). However, we keep this terminology, because any better terminology has not been introduced and because the levels of 5fC and 5caC are 2–4 orders of magnitude less frequent than 5hmC (Liu et al., 2013).

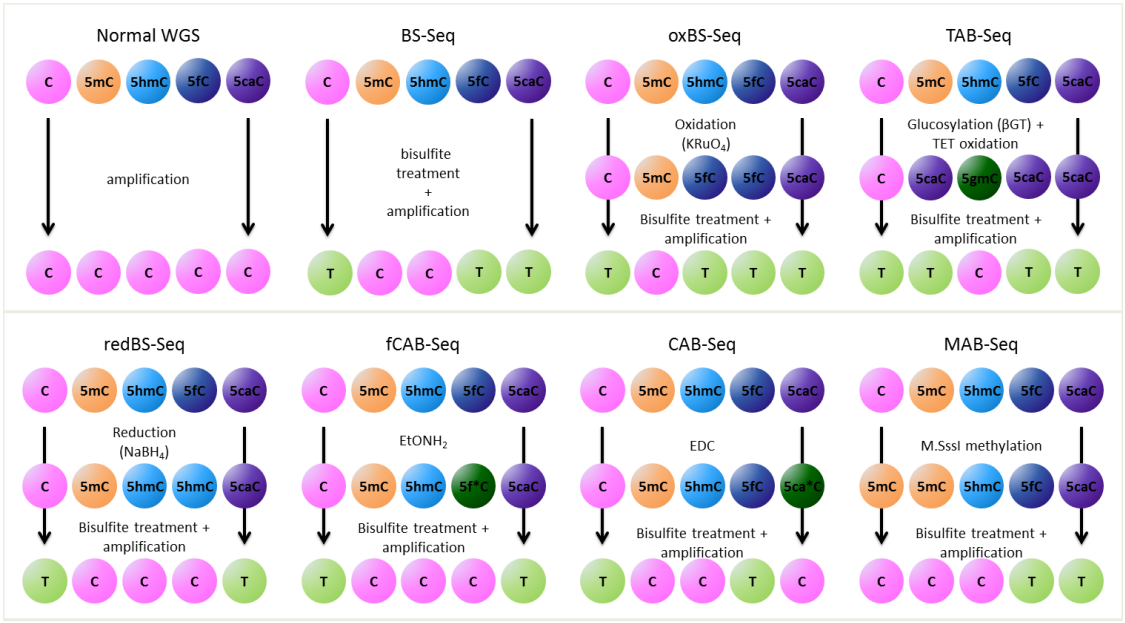

Figure 2.2. Overview of bisulfite-treatment-based sequencing techniques to measure DNA modifications with a single-base resolution. Data sets based on the methods in the top row are used in this thesis.

Increasing number of observations from different areas of epigenomics show that it is important to distinguish between 5mC and 5hmC (Li et al., 2016; Thomson and Meehan, 2017). This can be achieved with TAB-seq and oxBS-seq. In TAB-seq, the DNA is first

treated with b-glucosyltransferase (bGT) to introduce a glucose onto 5hmC, generating b-glucosyl-5-hydroxymethylcytosine (5gmC) (Yu et al., 2012). Next, TET oxidation is applied to convert 5mC and 5fC into 5caC, while C and 5gmC remain protected. Finally, a standard bisulfite treatment followed by sequencing are used, so that in the end 5hmC is read as C, but all the other four forms of cytosine are read as T (Fig. 2.2). TAB-seq therefore directly measures the 5hmC levels (as the percentage of unconverted reads).

In oxBS-seq, the DNA is first treated by potassium perruthenate ($KRuO_4$), which leads to specific oxidation of 5hmC into 5fC (Booth et al., 2012). This is followed by standard bisulfite treatment and sequencing. In the sequenced reads, only 5mC is read as C, while all the others are read as T (Fig. 2.2). This allows direct measurements of the 5mC levels (as the percentage of unconverted reads). Levels of 5hmC can be estimated from a combination of oxBS-seq and BS-seq measurements. 5hmC estimates can be in theory obtained as the difference of BS-seq and oxBS-seq measurements in individual positions. However, due to the noise in the measurements, variation among the cells, suboptimal conversion rates, and limited coverage due to high cost of sequencing, the resulting 5hmC estimates are only approximate. From our experience (on 18 whole-genome oxBS-seq samples) and experience of others, the BS-seq values are often smaller than oxBS-seq values, resulting in the histogram of 5hmC=BS-oxBS values nearly symmetrical around zero, especially in tissues with low 5hmC (Skvortsova et al., 2017). A number of strategies can be used to distinguish truly hydroxymethylated sites, such as multiple testing-corrected significantly higher proportion of unconverted reads in BS-seq than oxBS-seq, Bayesian approaches using probabilistic modelling (Äijö et al., 2016b,a; Johnson et al., 2016; Houseman et al., 2016; Xu et al., 2016), or grouping CpGs into regions to obtain higher coverage (Li et al., 2016).

Finally, several methods have been developed to measure 5fC and 5caC (Fig. 2.2). In redBS-seq (Booth et al., 2014) and fCAB-seq (Song et al., 2013), only C and 5caC are converted to T. In CAB-seq, only C and 5fC are converted to T (Lu et al., 2013). In MAB-seq, only 5fC and 5caC are converted to T (Wu et al., 2014).

The list of publicly available bisulfite-based sequencing datasets used in this thesis are listed in Table 9.1 and Table 9.2. Most of the data sets contained available modifica-

tion values and coverage per each covered CpG. Data sets by Chen et al. (2015) and Vandiver et al. (2015) were analysed using bsQC (described below).

2.2.1 bsQC: a newly developed package for analysis and quality control of bisulfite-sequencing data

Many tools have been developed to process sequencing data from BS-seq. However, as the first WGBS studies were performed on small numbers of samples, many of the tools were not designed for comfortable and efficient analysis of larger data sets. Moreover, the standards for quality control of different aspects of bisulfite-based sequencing data are still evolving. Most of the tools generate quality control outputs, but examination of these files generated for each sample separately is cumbersome and extremely time consuming. Finally, not all of them are tailored to bisulfite-based sequencing and therefore may generate unnecessary warnings (e.g. per base sequence content warning in FastQC).

Therefore, I developed bsQC, a command line pipeline for the processing of multiple bisulfite-based sequencing samples in parallel, with a special focus on quality control assessment. It combines several gold-standard tools for analysis of BS-seq data (detailed later), outputs methylation calls, and generates a single report with a comprehensive visual and tabular summary of quality control of all the involved steps side-by-side for all samples (Fig. 2.3). In addition, bsQC also contains tools for conversion rate estimation, batch effect detection, and basic exploration of the DNA modification levels across genomic features.

The steps performed in bsQC are summarised in (Fig. 2.3). The input of the pipeline are Fastq reads and a configuration file. The configuration file enables setting names of input files and other resources, parameters of the intermediate steps, running only a part of the pipeline, and setting CPU and memory, allowing for efficient usage of the available resources on a cluster.

1. The first step is estimation of conversion efficiencies. On one hand, the bisulfite conversion of unmodified cytosines can be inefficient, causing false positive modification calls. On the other hand, conversion of 5mC and 5hmC positions

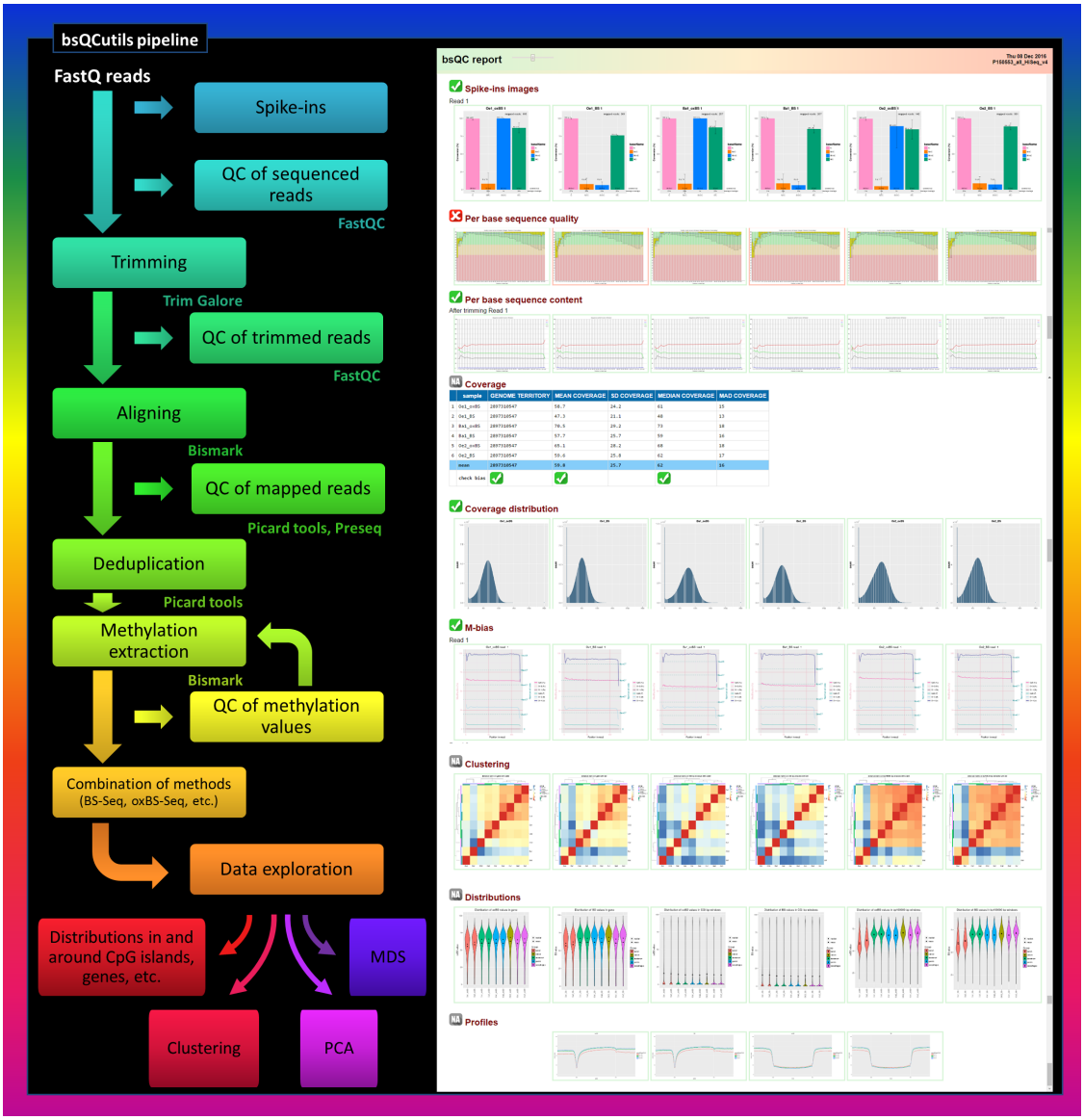

Figure 2.3. Overview of bsQC. Left: the bsQC pipeline and tools used for the individual steps. Right: selection of example sections of the output html bsQC report. Each section contains an icon indicating quality control result of all samples together (pass/warn/fail/NA). The three bottom rows in the example report (Clustering, Distributions, and Profiles) are exploratory visualisations of the entire dataset, shown separately for different genomic windows and features, such as genes, CpG islands, and exons.

would lead to false negative modification calls. Therefore, control DNA with known DNA modifications can be spiked-in into each sample. bsQC contains an option for detection of spike-ins and estimation of conversion efficiencies in C, 5mC, and 5hmC in each sample (standard processing as below is used, but with mapping the reads to a reference sequence of the used controls; the extracted

modification values are then compared with the ground truth). This allows simple detection of experiments with failed bisulfite conversion (or oxidation in TAB-seq/oxBS-seq) (Fig. 2.4A). Moreover, bsQC outputs the bisulfite conversion rates, which can be subsequently used for corrections for the differences between samples, e.g., using LUX (Äijö et al., 2016a,b).

2. Trimming of low quality ends of reads and adapters is performed using Trim Galore!, a wrapper tool around Cutadapt (Martin, 2011) and FastQC, which is used for quality control of the input reads (before and after trimming).

3. The trimmed reads are mapped to a reference genome using Bismark (Krueger and Andrews, 2011) and the quality of the mapped reads is evaluated using Picard tools, SAMtools (Li et al., 2009), and Preseq (Daley and Smith, 2013) to estimate library complexity, insert size distribution, duplication rate, prediction of library complexity with further sequencing, etc.

4. PCR and optical duplicates are removed using Picard tools.

5. Bismark is used to extract the coverage and the percentage of unconverted reads in individual positions, and modification levels along reads are visualised to allow for detection of M-bias (Fig. 2.4B). The M-bias shows average levels of modifications along the reads. The values should be constant along the read, however, beginnings or ends of reads sometimes have a bias for increased or decreased values. This can be easily assessed in the report, allowing to re-run this step with parameters adjusted to ignore the relevant parts of the reads.

6. Additional exploratory visualisations (described later, examples in Fig. 2.4D, E).

When all samples are processed, a single html report is generated for the entire dataset. The report style is inspired by FastQC; but in bsQC, the results are presented side-by-side for all samples in each quality control step. This enables a quick comparison of the samples and identification of outliers without having to tediously examine many separate documents.

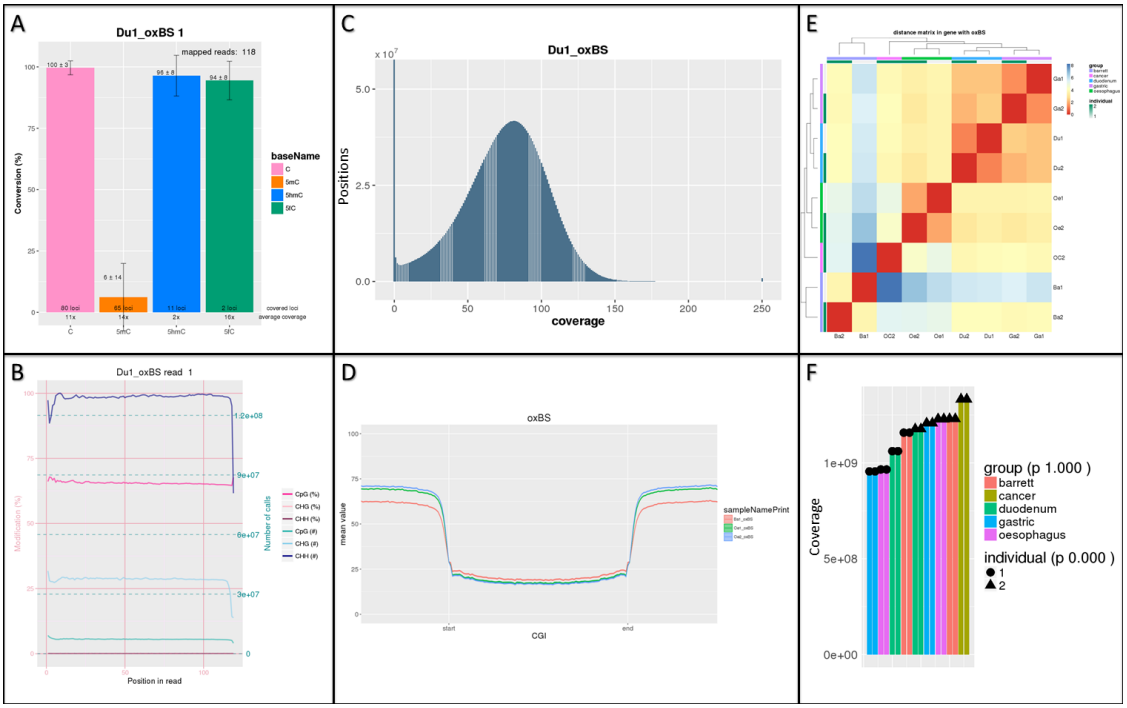

Figure 2.4. Example of figures in the bsQC report. A: Quality control of bisulfite conversion efficiency (and other sources of inaccuracies) on spiked-in controls. The bars represent mean ± standard deviation of converted reads in individual positions of the control reference. **B:** M-bias plot showing average modification levels and number of modification calls along the first reads. **C:** Distribution of coverage in individual positions of the reference genome. **D:** Meta-CGI plot of average modification levels inside and around CpG islands shown separately for individual samples. **E:** Hierarchical clustering on distance matrix between samples, using average oxBS values in genes. **F:** Detection of technical biases between pre-defined groups of samples. In this example, all samples from patient 1 have lower coverage than samples from the second patient, which could confound down-stream analyses, if left uncorrected.

Most of the visualisations and tables contain quality control icons (pass/warn/fail/NA). This allows quick identification of potential quality issues: bisulfite conversion rate (specific for BS-seq, oxBS-seq and TAB-seq), FastQC modules (adjusted for bisulfite-based sequencing), trimming, mapping efficiency, duplicate reads and evaluation of M-bias plots. Based on specific needs and expectations of each project, users can adjust parameters of the evaluations in a configuration files (what is considered to be a pass/warn/fail output) and combine this information with the visual-only quality controls (such as histogram of insert size).

The ultimate goal of many studies is to compare DNA modifications between several groups of samples: different tissues, treatment, genetic conditions, batch, etc. However, these biological differences might be confounded by technical differences, such as

different coverage or sequencing quality between the groups. Therefore, bsQC can detect some of these technical differences by allowing the user to annotate the samples (with group number, individual number, and used method: BS/oxBS/TAB) and then compare the groups in terms of several technical features, such as bisulfite conversion rate, insert size, duplication rate, and coverage (Fig. 2.4F). If there is such a technical difference between the groups, the user might use this knowledge to correct for the bias in a subsequent analysis (adjusting the methylation values based on the bisulfite conversion rates, sub-sampling reads in the case of differences in coverage, etc.), or perform additional sequencing, to ensure that the differences in DNA modifications are biological and not a mere technical artefact.

Finally, bsQC provides an optional basic exploration of the resulting modification levels in the data set. This might serve for the user as a first familiarisation with the data set and might also guide further explorations. In particular, the user can provide a set of bedfiles with genomic features, such as CpG islands, genes, exons, enhancers, etc., or genomic windows. These are then used to create meta-gene plots (profiles of DNA modifications along genes, or other provided features, for each sample) (Fig. 2.4D). Furthermore, distributions of average modification levels in the features are plotted (one violin plot per sample). Third, hierarchical clustering of the samples is computed based on Euclidean distances of the average values in the features between the samples (Fig. 2.4E). Finally, principal component analysis (PCA) and multidimensional scaling (MDS) of the sample values are shown. All these plots are displayed separately for each type of provided experiment (BS-seq/oxBS-seq/TAB-seq) and eventually also on the combination of these methods (e.g., 5hmC as max(0, BS-seq – oxBS-seq)). These basic estimates are also provided, or the user can use more sophisticated methods to get the estimates (such as using LUX (Äijö et al., 2016a,b)). We did not include any such method in the pipeline, because the selection of such method might be project specific.

Since the development of this tool, Bismark has been updated with an html report partially overlapping with bsQC; however, it does not show more samples together. Also, a recently published tool MultiQC (Ewels et al., 2016) has partially overlapping functionality (to combine quality logs from multiple tools and samples into one report),

but does not provide the additional bisulfite-based sequencing tailored quality controls, automatic detection of batch effects/technical artefacts, and exploratory visualisations of the modification values across samples and genomic features. The bsQC tool was originally developed for internal use on projects in this thesis (and used in chapters 3 and 4) and other projects, such as in (Bardella et al., 2016) and three ongoing collaborations. Several finishing technical code-edits are still needed for full deployment of the tool for public use.

2.3 DNA replication

2.3.1 Techniques to measure replication timing

Repli-Seq is a commonly used method to measure profiles of replication timing along the genome (Hansen et al., 2010; Ryba et al., 2011; Rivera-Mulia et al., 2015). In short, newly synthesised DNA (of asynchronously dividing cells) is *in vivo* labelled with the nucleotide analogue 5-bromo-2-deoxyuridine (BrdU), which is incorporated into the nascent strand instead of thymidine. The cells are then sorted into several S-phase fractions using flow cytometry. BrdU-labelled DNA from each fraction is then immunoprecipitated (i.e., isolated using an anti-BrdU monoclonal antibody), and a DNA library is prepared from each fraction for sequencing. The sequenced reads are mapped to genome reference and normalised density of individual S-phase fractions are computed in genomic windows (such as 50 kbp sliding windows with 1 kbp intervals) (Hansen et al., 2010; Ryba et al., 2011). Hybridisation microarrays (Repli-Chip) can be used instead of sequencing, producing comparable results (Ryba et al., 2011; Pope et al., 2014). Other methods with different protocols but also based on flow cytometry sorting of cells can be used to measure replication timing profiles (Koren et al., 2012).

2.3.2 Techniques to measure replication origins

An indirect approach to measure regions rich for replication origins (initiation zones, ORI clusters) is based on the replication timing profiles. The replication valleys (early replicating regions) correspond to the approximate locations of initiation zones, while the replication peaks (late replication regions) are regions of replication termini (Baker

et al., 2012; Hansen et al., 2010; Dellino et al., 2013; Haradhvala et al., 2016). Although this approach does not give precise locations of individual replication origins (as the replication valleys are the source of most tissue-specific variation in the profiles (Ryba et al., 2010)), the regions in between valleys and peaks represent conserved predominantly uni-directional timing transition regions, with the direction of replication given by the sign of the slope (negative slope for left-replicating regions, positive slope for right-replicating regions) (Haradhvala et al., 2016; Ryba et al., 2010).

Multiple techniques for a direct measurement of replication origins genome-wide have been developed. Short nascent strand sequencing (SNS-Seq; Lexo-enriched nascent strands sequencing, Lexo-NS-Seq) is one of the most popular techniques (Urban et al., 2015). DNA from asynchronous cells is made single-stranded, phosphorylated, and treated with lambda exonuclease enzyme (Lexo, λ-exo). Lexo is a 5'-to-3' DNA exonuclease, which digests the parental DNA strand, while the short nascent strands are protected by their 5' RNA primers. Fragments of size corresponding to the nascent leading strand (500–1 500 nt) are then purified, to exclude the very short (ca. 200 nt) Okazaki fragments of the nascent lagging strand. The resulting fragments are made double-stranded, sonicated, sequenced and mapped to a reference genome. Peaks of the mapped reads (determined using a peak calling such as from MACS (Zhang et al., 2008)) represent the replication origins. Microarrays (SNS-chip) can be used instead of sequencing (Urban et al., 2015).

SNS-Seq and other related techniques are dependent on the ability of Lexo to efficiently digest the parental DNA. However, Lexo digestion of nonreplicating genomic DNA showed that Lexo digests inefficiently DNA with G-quadruplex (G4) structures and GC-rich DNA, introducing a systematic bias into the resulting ORI measurements (Foulk et al., 2015). However, these biases can be controlled for by using Lexo digested nonreplicating genomic DNA as a control in the peak calling (Foulk et al., 2015).

Other methods to measure ORI genome-wide include: ORC-ChIP-Seq, BrIP-NS-Seq, and Bubble-trap (reviewed in Urban et al., 2015), and recently developed OK-Seq (Petryk et al., 2016) and ini-seq (Langley et al., 2016). The concordance between these methods is limited. For example OK-Seq very well corresponded to the regions with

OR predicted by replication timing and generally better aligned to Bubble-trap than SNS-Seq (Petryk et al., 2016). On the other hand, ini-seq showed highest concordance with SNS-Seq, followed by OK-Seq, and Bubble-trap (Langley et al., 2016). In summary, the current methods show promising results for detection of replication origins, but further research is needed to determine their accuracy and minimise the false positive and false negative calls.

Ich brauche Zeit, kein Heroin,
kein Alkohol, kein Nikotin
Brauch keine Hilfe, doch Koffein
und Hydroxymethylcytosin!

— Rammstein *Benzin (adapted)*

Calling the five kings of the genetic code
To inspire him with black pearls of their wisdom

— Pinar Ayata *Sounds of Science - Silver Halo*

3 Mutational properties of 5hmC compared to 5mC

3.1 Introduction

The mutagenic effects of cytosine methylation have been researched already for more than four decades, showing that spontaneous deamination of 5mC can cause CpG>TpG mutations, the most common mutation type in somatic and germline mutations (reviewed in Introduction 1). However, the mutational properties of cytosine hydroxymethylation are mostly unexplored. Although hydroxymethylation is not as frequent as methylation, 5hmC is the dominant modification in a considerable fraction of CpG dinucleotides (e.g., 13.4% in brain (Wen et al., 2014)) and the vast majority of 5hmC is found as a stable, long-lived modification in adult mouse tissue that undergoes little cell division (Bachman et al., 2014; Brazauskas and Kriaucionis, 2014). Moreover, 5hmC is elevated in highly expressed genes and enhancers (Stroud et al., 2011; Mellén et al., 2012; Yu et al., 2012; Lister et al., 2013; Wen et al., 2014; Chen et al., 2015). Therefore, the mutational properties of 5hmC are of great interest, as they could have a substantial influence on the mutability of important regions of DNA.

A large proportion of mutations observed in any cancer genome originate in its pre-cancerous cell of origin (Nik-Zainal et al., 2012a; Stephens et al., 2012; Tomasetti

et al., 2013; Wu et al., 2015) and will have been influenced by its epigenetic landscape. Moreover, signature 1 (consisting mainly of CpG>TpG mutations) is one of the two mutational signatures correlating with age at diagnosis (Alexandrov et al., 2015), supporting the fact that these mutations were gathered during the entire lives of the patients, not only after the origin of cancer. Therefore, reasonable estimates of mutational properties of 5mC and 5hmC can be obtained by combining information about positions of 5mC and 5hmC in normal (healthy) tissues with mutation frequency in cancers of the same tissues. This is now possible thanks to recent development of techniques that enable single-nucleotide resolution mapping of DNA modifications and distinguishing between 5mC and 5hmC, such as BS-seq combined with TAB-seq or oxBS-seq (Yu et al., 2012; Booth et al., 2012).

Recently, Supek et al. (2014a) reported elevated C>G transversion rates at 5hmC sites, using 5hmC maps from human and mouse embryonic stem cells. However, these findings are limited by the fact that embryonic stem cells differ substantially from the somatic tissues in which mutations were observed (Schultz et al., 2015). The publication of single-base resolution maps of 5mC and 5hmC occupancy in samples of human brain, kidney and blood (Wen et al., 2014; Chen et al., 2015; Pacis et al., 2015) now enables us to interrogate the tissue-specific effect of cytosine modifications on somatic mutation rates in unprecedented detail.

3.2 Materials and methods

3.2.1 Modification data

As 5hmC predominantly occurs in a CpG context (97.4% in adult brain (Wen et al., 2014)), we focussed the analysis on CpG sites. BS-seq and TAB-seq DNA modification measurements (Table 9.1) for brain were extracted from supplementary information provided by Wen et al. (2014). Only sites with more than 5 TAB-seq reads were taken into account. In blood, BS-seq and TAB-seq values in CpG sites were taken from supplementary files provided by (Pacis et al., 2015). For kidney and ESC maps, raw reads were processed with the bsQC pipeline (section 2.2.1). Multiple replicates were processed both independently and together (adding the reads from the replicates together).

Sequencing reads come from heterogeneous populations of cells. Hence, a single genomic position usually cannot be assigned a single state (C, 5mC or 5hmC). Instead, for each position we estimated:

- **mod level** as the number of unconverted BS-seq reads / number of all BS-seq reads,
- **5hmC level** as the number of unconverted TAB-seq reads / number of all TAB-seq reads,
- **5mC level** as mod level – 5hmC level,
- for positions with mod level > 10%, we define **$\mathbf{5hmC_{rel}}$ level** as 5hmC level / mod level; this represents how much the modified positions are methylated (low $5hmC_{rel}$) vs. hydroxymethylated (high $5hmC_{rel}$).

We next defined highly methylated and highly hydroxymethylated positions as:

- **$\mathbf{5mC_{high}}$**: mod level > 10% and $5hmC_{rel} \leq threshold_{5mC}$,
- **$\mathbf{5hmC_{high}}$**: mod level > 10% and $5hmC_{rel} \geq threshold_{5hmC}$.

The values of $threshold_{5mC}$ = 0.3 and $threshold_{5hmC}$ = 0.5 were used, as they represent a good combination of stringent criteria and sufficient statical power, but their robustness was validated by exploring a range of values for both thresholds for the main analysis (see section 3.3.1, Fig. 3.12, 3.13).

To compute the number of modified sites inside the exome, the reference Illumina Truseq definition of exon loci was downloaded from the Illumina website. Overlapping exons were merged using bedtools so that each genomic site is covered by at most one exon. Two-sided paired Wilcoxon signed-rank test was used for testing significance between mutation frequency of $5mC_{high}$ and $5hmC_{high}$ sites (i.e., two values in each patient). The same test was used for all the following statistics, if not stated otherwise.

3.2.2 Mutation data

Publicly available WGS and WES data sets used in this chapter are listed in Table 9.3. All single-nucleotide variants were classified by the pyrimidine of the mutated Watson-Crick base pair (C or T) and the immediate 5' and 3' sequence context into 96 possible mutation types as described by Alexandrov et al. (2013a).

3.2.3 Gene expression data

Gene expression (in FPKM) from RNAseq experiments on 630 brain tissue samples were downloaded from the GTEx human tissue expression project (`http://www.gtexportal.org/home/`).

3.2.4 Brain cancer driver genes

We classified genes into three classes:

- Brain cancer driver genes (19): BCOR, CDK4, CDKN2A, CSNK2B, CTNNB1, DDX3X, EGFR, ERBB2, IDH1, KDM6A, LDB1, NF1, PIK3CA, PIK3R1, PTCH1, PTEN, RB1, SMARCA4, TP53 (Parsons et al., 2008; TCGA, 2008; Pugh et al., 2012).

- Other cancer driver genes (108): table S2A in (Vogelstein et al., 2013).

- Non-driver genes (17 816): genes that are neither classified as brain cancer driver genes, nor as other cancer driver genes.

Non-driver genes similarly expressed as the brain driver genes were chosen using the following algorithm: for each driver gene, the N=8 most similarly expressed unique non-driver genes were selected, resulting in 152 expression-matched non-driver genes. The distributions of $5hmC_{rel}$ in all CpGs in brain driver genes (25 582 CpGs) and non-driver genes (120 918 CpGs) were compared using a two-sided Wilcoxon ranksum test.

3.3 Results

3.3.1 5hmC sites in brain exhibit lower frequency of CpG>TpG mutations than 5mC sites

Since brain exhibits particularly high levels of 5hmC (Fig. 3.1), we first investigated the relationship between the regional distribution of 5hmC, 5mC and mutagenesis in brain tumours. We reasoned that this approach would provide the highest sensitivity to detect any correlation between 5hmC and mutation frequency.

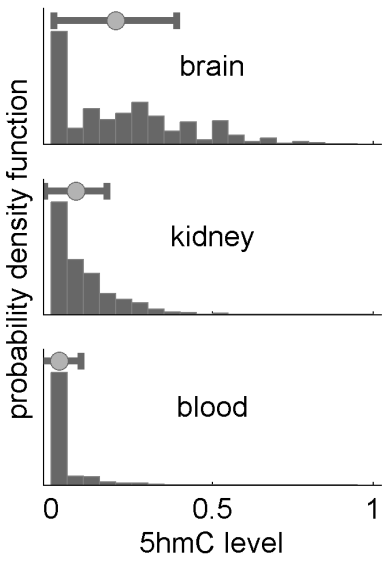

Figure 3.1. Distribution of 5hmC levels in a CpG context in brain, kidney and blood.

We analysed 344 370 somatic single nucleotide variants (SNVs) from 665 samples derived from exome and whole genome sequencing of the following cancer types: Glioblastoma (GBM), Low grade glioma (LGG), Neuroblastoma (NRB), Medulloblastoma (MDB) and Pilocytic astrocytoma (PA) (Alexandrov et al., 2013a). Out of all SNVs, one quarter occurred in CpG dinucleotides and most of them were transitions from C to T (Fig. 3.2). Combined with the fact that CpG dinucleotides are the least frequent dinucleotides in the human genome, CpG>TpG mutations were clearly the most mutated type in brain tumours (Fig. 3.3).

Mutations and DNA modifications are not distributed uniformly along the chromosomes. First, we computed average 5hmC, 5mC, and $5hmC_{rel}$ in 100 kbp genomic

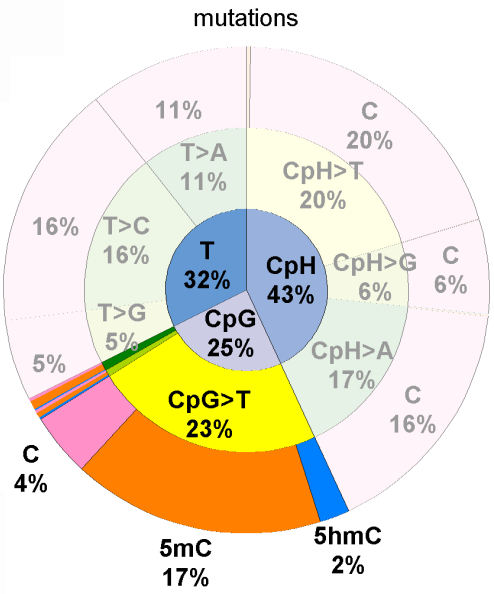

Figure 3.2. Distribution of SNVs in brain cancer whole genomes according to type, context and modification state.

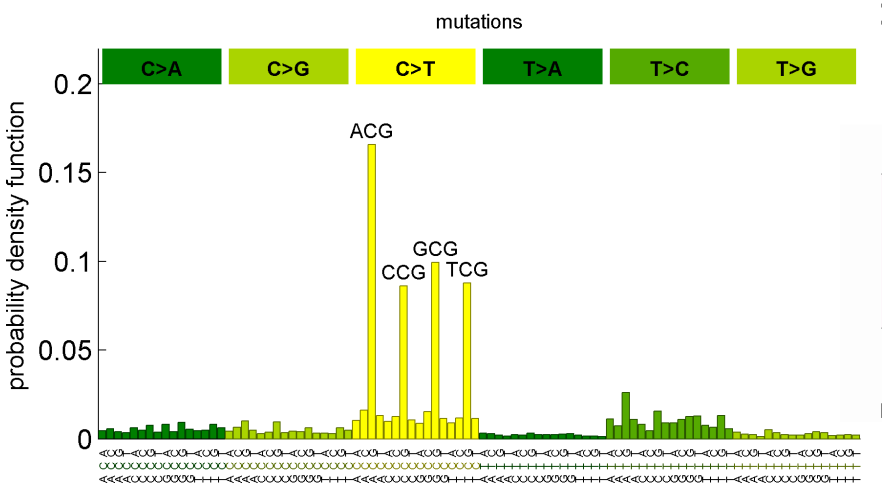

Figure 3.3. Frequency of SNVs in brain cancer exomes, stratified by sequence context, normalised by frequency of trinucleotides.

windows and compared them with the frequency of C>T mutations in CpG dinucleotides in the same windows (in each window computed as the number of CpG>TpG mutations divided by the number of Cs in CpG dinucleotides; only WGS samples were included). All traces were z-score normalised and plotted per chromosome after Gaussian smoothing with parameters n = 50, sigma = 2.5.

As expected, 5mC levels displayed a positive correlation with the frequency of CpG>TpG (r=0.66 for chr3 in Fig. 3.4; other chromosomes in Fig. 3.5, 3.6). In contrast, 5hmC levels were significantly anti-correlated with the frequency of CpG>TpG (r=-0.71 for chr3 in Fig. 3.4; other chromosomes in Fig. 3.5, 3.6). This correlation is not a simple consequence of the uneven distribution of genes, exons, CpG islands or levels of gene expression (Fig. 3.7), as these genomic features show weaker correlation with CpG>TpG mutation frequency.

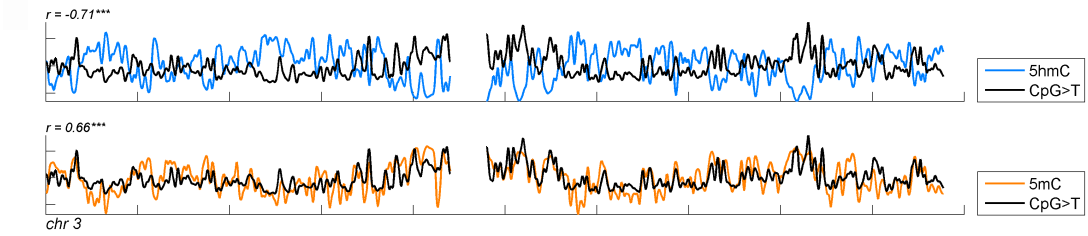

Figure 3.4. CpG>TpG mutations correlate positively with 5mC levels and negatively with 5hmC levels. CpG>TpG mutation frequency (black), 5hmC (blue) and 5mC (orange) density in 100 kbp windows of chromosome 3, smoothed with a Gaussian filter (n=50, sigma=2.5).

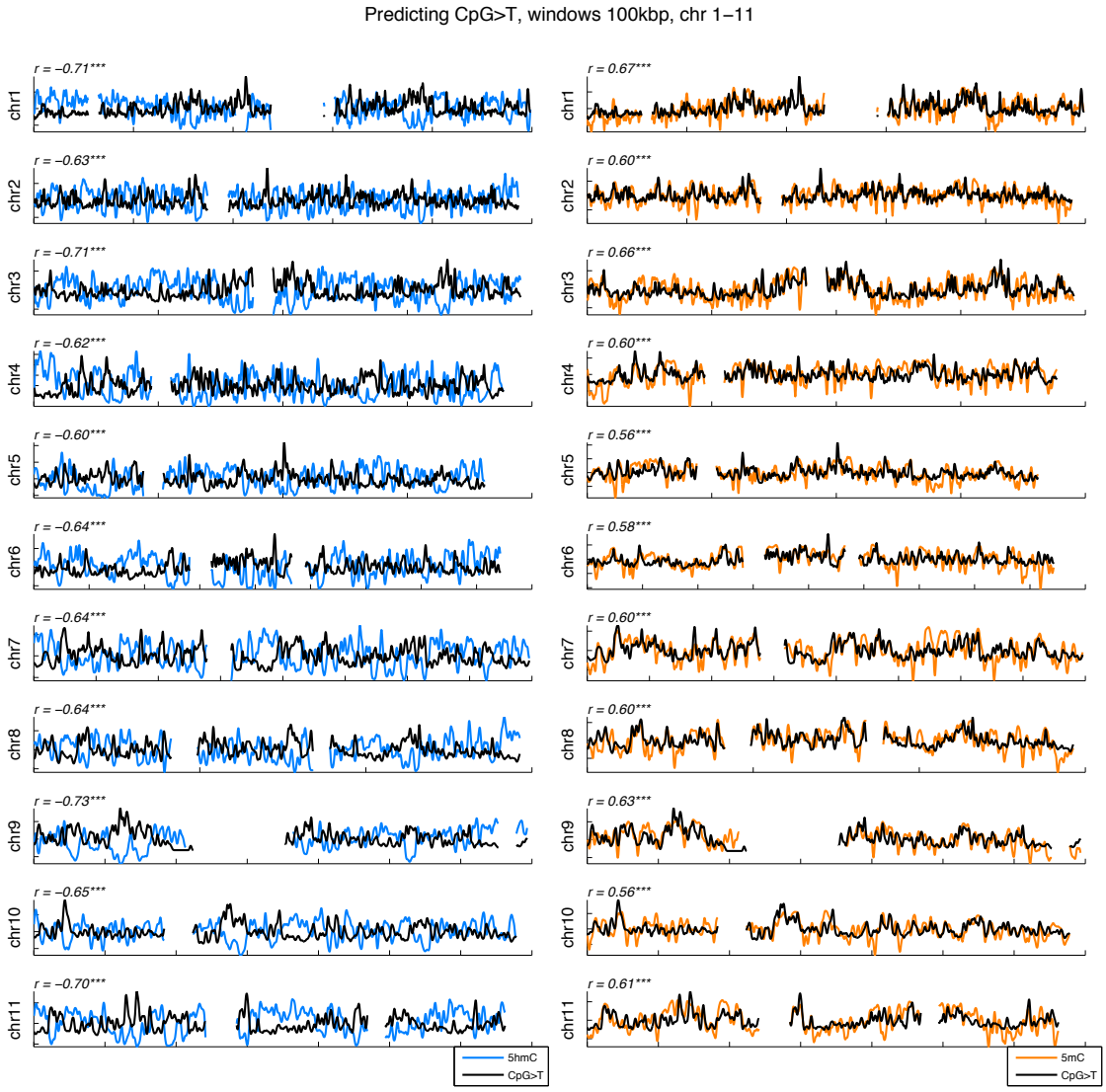

Figure 3.5. Distribution of CpG>TpG mutations in comparison with modifications across all chromosomes. CpG>TpG mutation frequency (black), 5hmC (blue) and 5mC (orange) density in 100 kbp windows, smoothed with a Gaussian filter (n=50, sigma=2.5). Chromosomes 1–11.

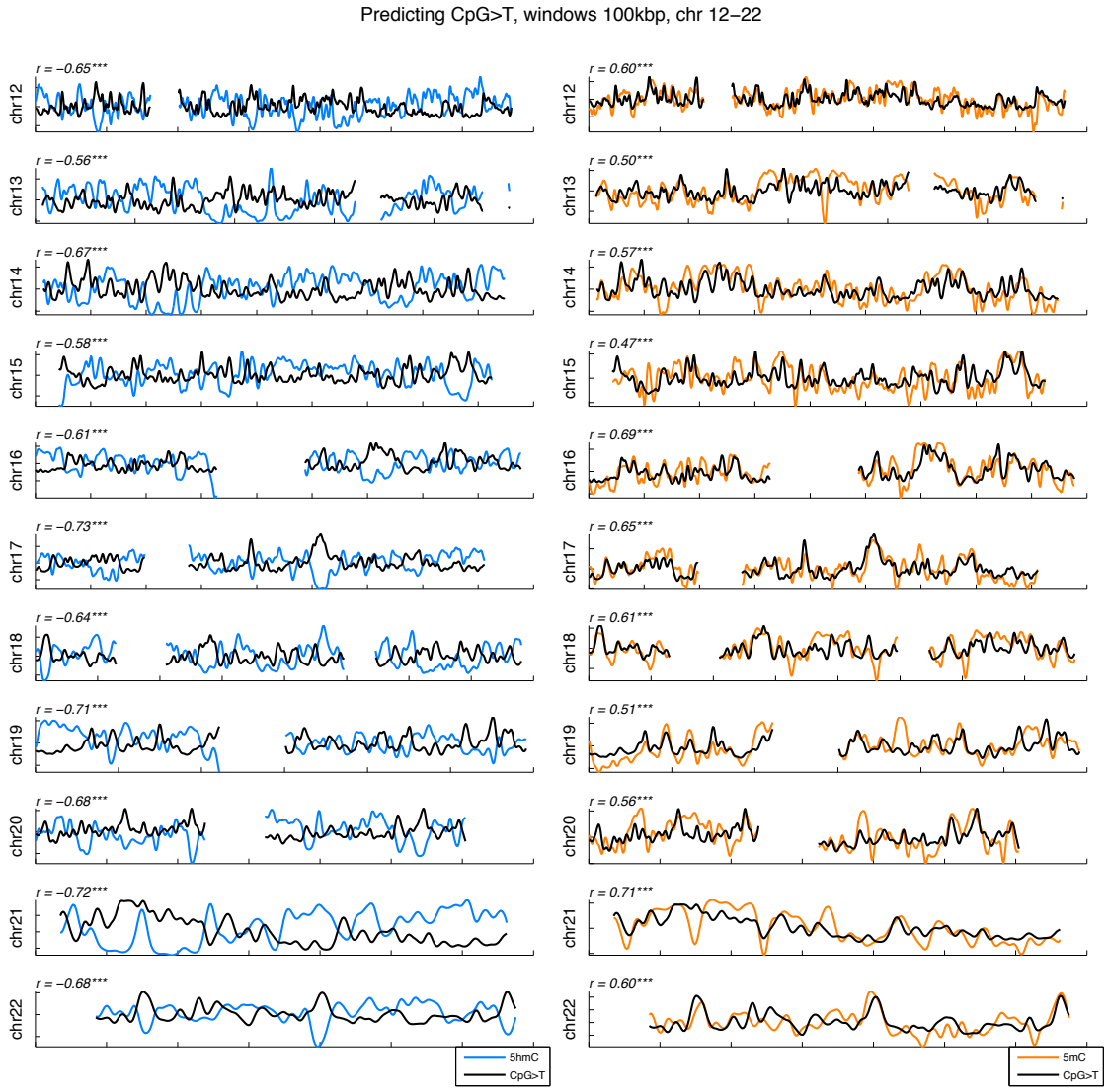

Figure 3.6. Distribution of CpG>TpG mutations in comparison with modifications across all chromosomes. CpG>TpG mutation frequency (black), 5hmC (blue) and 5mC (orange) density in 100 kbp windows, smoothed with a Gaussian filter (n=50, sigma=2.5). Chromosomes 12–22.

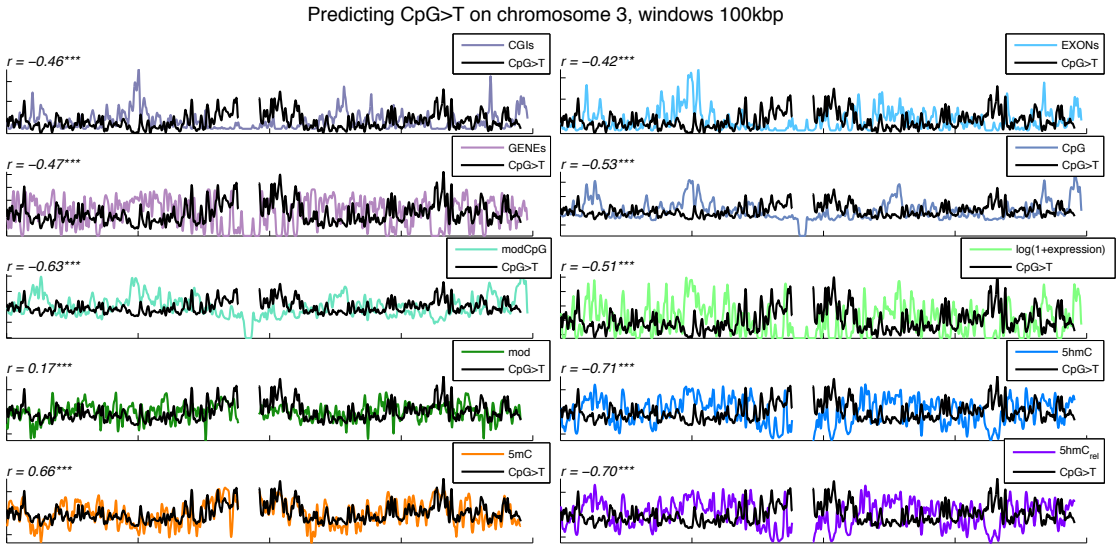

Figure 3.7. Distribution of CpG>TpG mutations in comparison with other genomic features. CpG>TpG mutation frequency (black) and several genomic features in 100 kbp windows on chromosome 3, smoothed with a Gaussian filter (n=50, sigma=2.5). CGIs: density of CpG islands, EXONs: density of exons, GENEs: density of genes, CpG: density of CpGs, modCpG: density of CpGs with mod level $\geq$ 10%; and average modification levels: mod, 5hmC, 5mC, and $5hmC_{rel}$.

Averaging over the entire genome, the frequency of C>T mutations differed substantially between $5mC_{high}$ and $5hmC_{high}$ sites. The fraction of mutated $5hmC_{high}$ sites was significantly lower than the fraction of mutated $5mC_{high}$ sites (Fig. 3.8). The lower mutation frequency was consistently observed in data derived from both exome and whole genome sequencing projects (P<0.001, Wilcoxon signed-rank test). Moreover, all brain cancer types individually displayed a significant (28–53%, P<0.05 in all types) reduction of C>T mutations in $5hmC_{high}$ sites (Fig. 3.10).

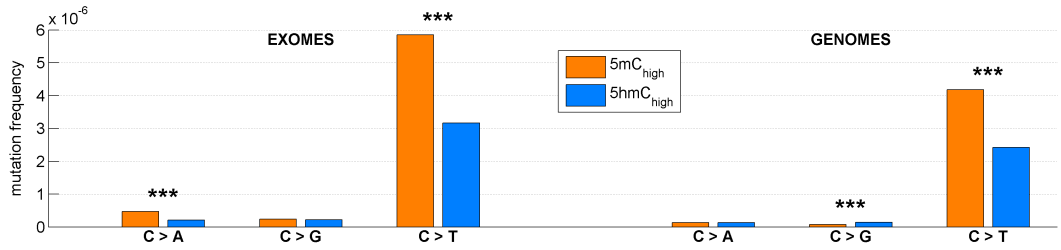

Figure 3.8. C>T mutations are common in the genome but depleted in 5hmC sites compared to 5mC sites. Average fraction of mutated sites for $5mC_{high}$ vs. $5hmC_{high}$ over all patient samples (CpG sites only; ***P < 0.001; **P < 0.01; *P < 0.05).

Since the minimal coverage per strand (5x) is lower than the commonly used standard of 15x per strand (i.e., 30x per both strands; http://ihec-epigenomes.org/research/reference-epigenome-standards/ and even higher thresholds recommended by Libertini et al. (2016)), we next explored the effects of coverage on the observed results. The ratio of average CpG>TpG mutation frequency in $5mC_{high}$ and in $5hmC_{high}$ sites remained similar for a range of minimal coverage thresholds (5–20x) per strand (Fig. 3.9; only WGS samples were used in this analysis and the given coverage was required in both BS-seq and TAB-seq). A slight increase of the ratio was present in sites with higher coverage. This could be due to mis-annotations of $5mC_{high}$ and $5hmC_{high}$ in positions with low coverage.

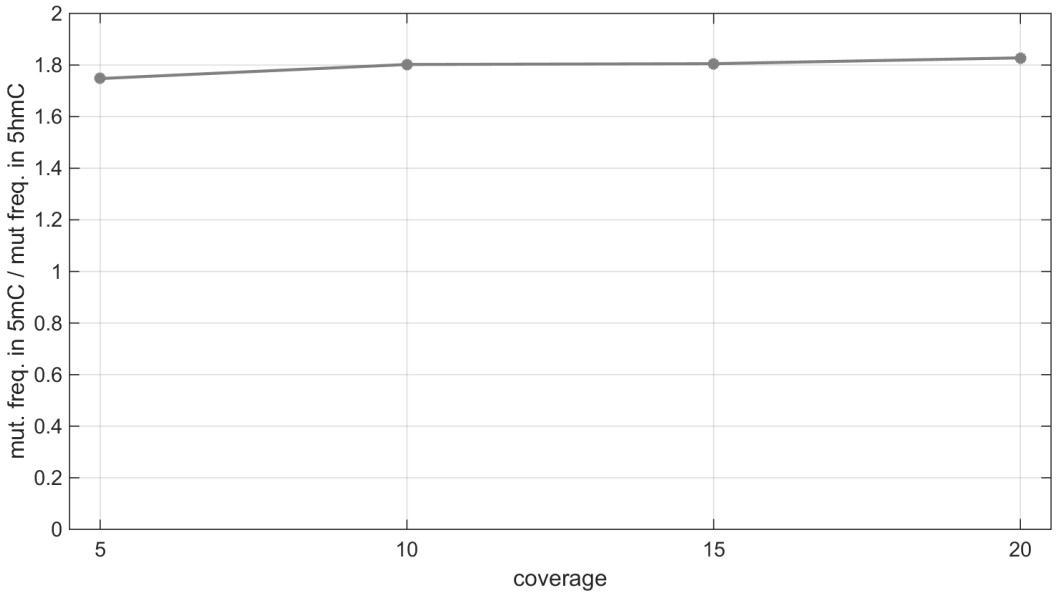

Figure 3.9. The ratio between mutation frequency in 5hmC and 5mC sites is similar for different coverage thresholds. Ratio of the average mutation frequency in $5mC_{high}$ sites and the average mutation frequency in $5hmC_{high}$ sites (y-axis) plotted against minimal coverage threshold of BS-seq and TAB-seq (x-axis). Only WGS samples and C>T mutations in a CpG context were included.

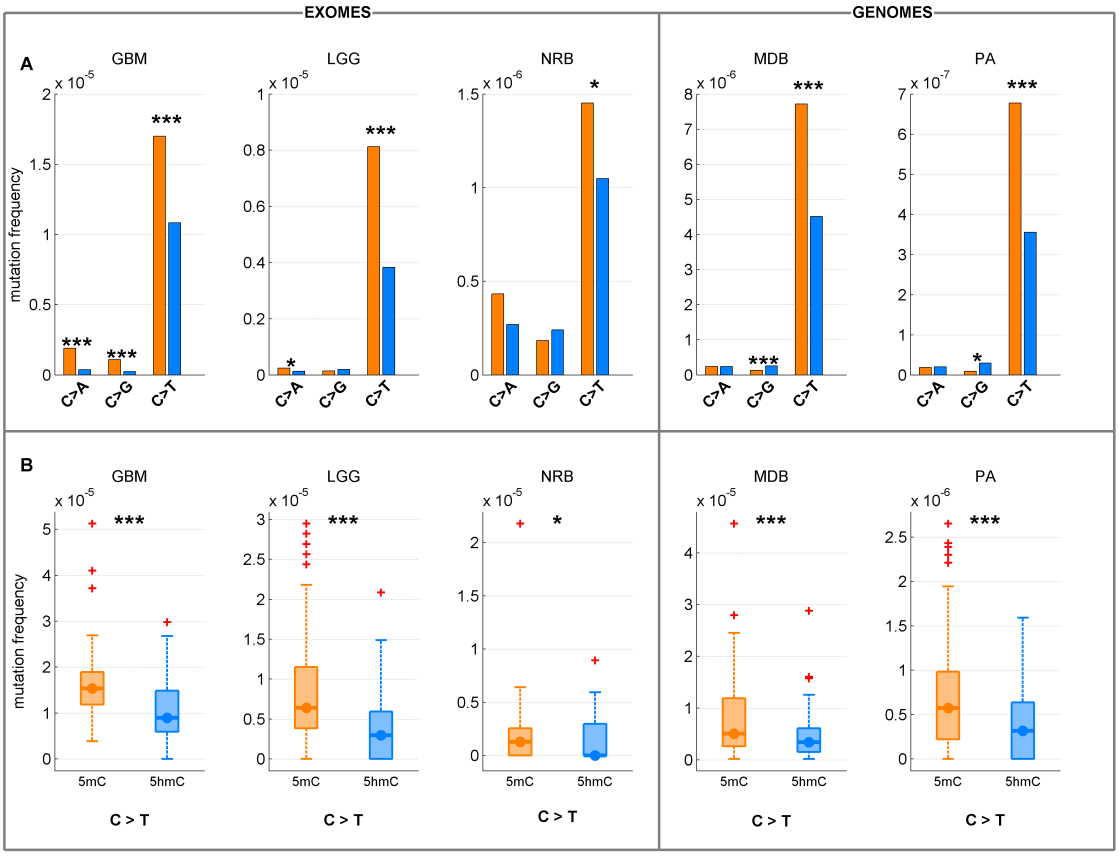

Figure 3.10. Differential mutation frequency between 5mC and 5hmC is present in all 5 brain cancer types. A: Average fraction of mutated CpG sites for $5mC_{high}$ vs. $5hmC_{high}$ computed separately for each cancer type. **B:** Box plot of C>T mutation frequency, as shown in A.

It is known that CpG>TpG mutations correlate with age at diagnosis, representing one of the only two known mutational signatures with "clock-like" properties (Alexandrov et al., 2015). Here we observed that this correlation is present in both methylated and hydroxymethylated sites (Fig. 3.11). Moreover, the slope for 5mC was steeper than for 5hmC, suggesting that even the mechanism causing the difference of CpG>TpG mutability between 5mC and 5hmC was present in the pre-cancerous cell of origin.

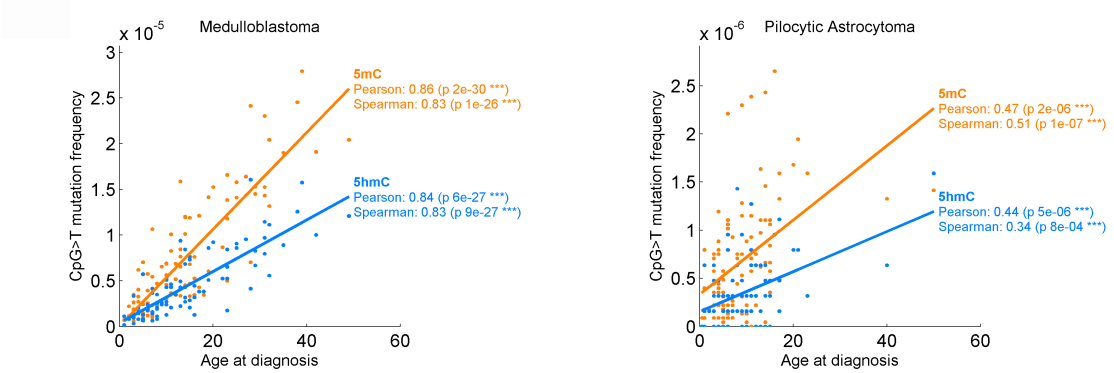

Figure 3.11. C>T mutations in both methylated and hydroxymethylated CpGs correlated with age at diagnosis. Correlation of whole genome CpG>TpG mutation frequency with age at the time of diagnosis in patients with Medulloblastoma and Pilocytic Astrocytoma.

We also compared the fraction of mutated $5mC_{high}$ and $5hmC_{high}$ sites for the other two possible types of mutations: C>A and C>G. As shown in Fig. 3.8, C>A or C>G transversions are an order of magnitude less frequent than C>T transitions in both 5mC and 5hmC sites. The relationship between C>A and C>G mutations and 5hmC varied between cancer types (Fig. 3.10). In GBM and LGG the frequency of C>A mutations was significantly higher in $5mC_{high}$ compared to $5hmC_{high}$ sites, but in NRB, MDB and PA we detected no significant difference. The frequency of C>G mutations in $5mC_{high}$ sites differed significantly from $5hmC_{high}$ sites only in MDB, PA and GBM. In MDB and PA, $5hmC_{high}$ sites were slightly enriched for C>G mutations, whereas in GBM an enrichment was observed at $5mC_{high}$ sites. Since C>T transitions are the most common somatic mutation type in brain and the difference in C>T mutations between $5mC_{high}$ and $5hmC_{high}$ sites is more consistent among cancer types than in the C>A and C>G transversions, we focus mainly on C>T mutations in the further analyses.

We confirmed that C>T mutations are significantly depleted at 5hmC sites across a wide range of thresholds in definitions of $5mC_{high}$ and $5hmC_{high}$ (Fig. 3.12A–F). While C>T mutations were significantly enriched in $5mC_{high}$ compared to $5hmC_{high}$ sites in all explored values of $threshold_{5mC}$ and $threshold_{5hmC}$, C>A and C>G mutations were markedly more sensitive to the choice of the threshold values. This lack of robustness in C>G and C>A mutations might be a result of relatively low numbers of these mutations in CpG sites, i.e., a lack of statistical power, and the results should be therefore interpreted with caution. On the other hand, the sensitivity analysis showed sufficient robustness of C>T mutations with regards to the choice of the threshold values. In fact, more stringent definitions of $5hmC_{high}$ (e.g., $5hmC_{rel} \geq 0.7$) result in even greater differences (42–59%) in C>T mutation frequencies between $5mC_{high}$ and $5hmC_{high}$ sites (Fig. 3.12G–I, Fig. 3.13).

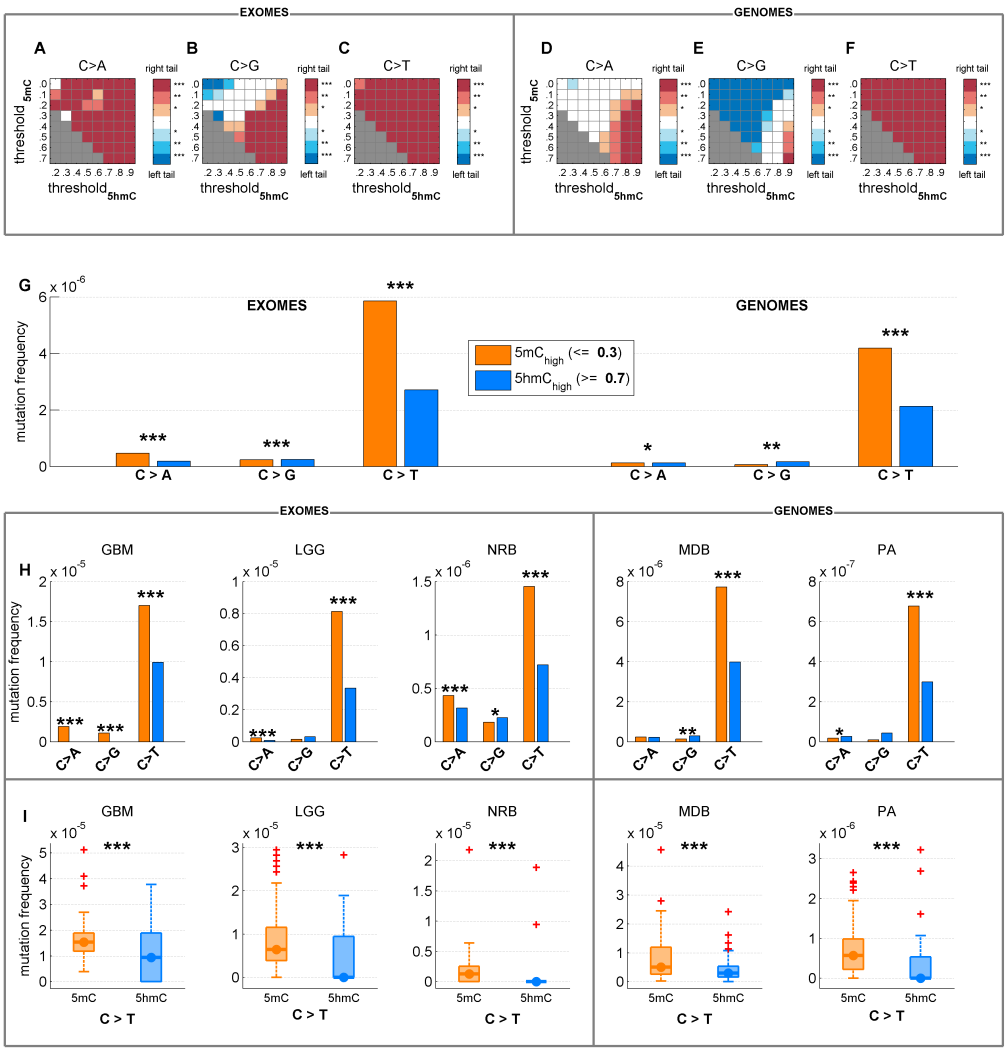

Figure 3.12. Depletion of C>T mutations in $5hmC_{high}$ is relatively insensitive to varying definitions of $5mC_{high}$ and $5hmC_{high}$. **A–F:** Significance of a difference in mutation frequency in $5mC_{high}$ and $5hmC_{high}$, for a range of values of $threshold_{5mC}$ and $threshold_{5hmC}$ ($5mC_{high}$ is defined as sufficiently modified sites with $5hmC_{rel} \leq threshold_{5mC}$; $5hmC_{high}$ is defined as sufficiently modified sites with $5hmC_{rel} \geq threshold_{5hmC}$). One-sided paired Wilcoxon sign-rank test was used. Red colour represents a significant increase of mutation frequency in $5mC_{high}$ (right tail test) whereas blue colour represents elevated mutations in $5hmC_{high}$ (left tail test). **G–I:** C>T mutation frequency for $5mC_{high}$ vs. $5hmC_{high}$ with $threshold_{5mC} = 0.3$ and $threshold_{5hmC} = 0.7$.

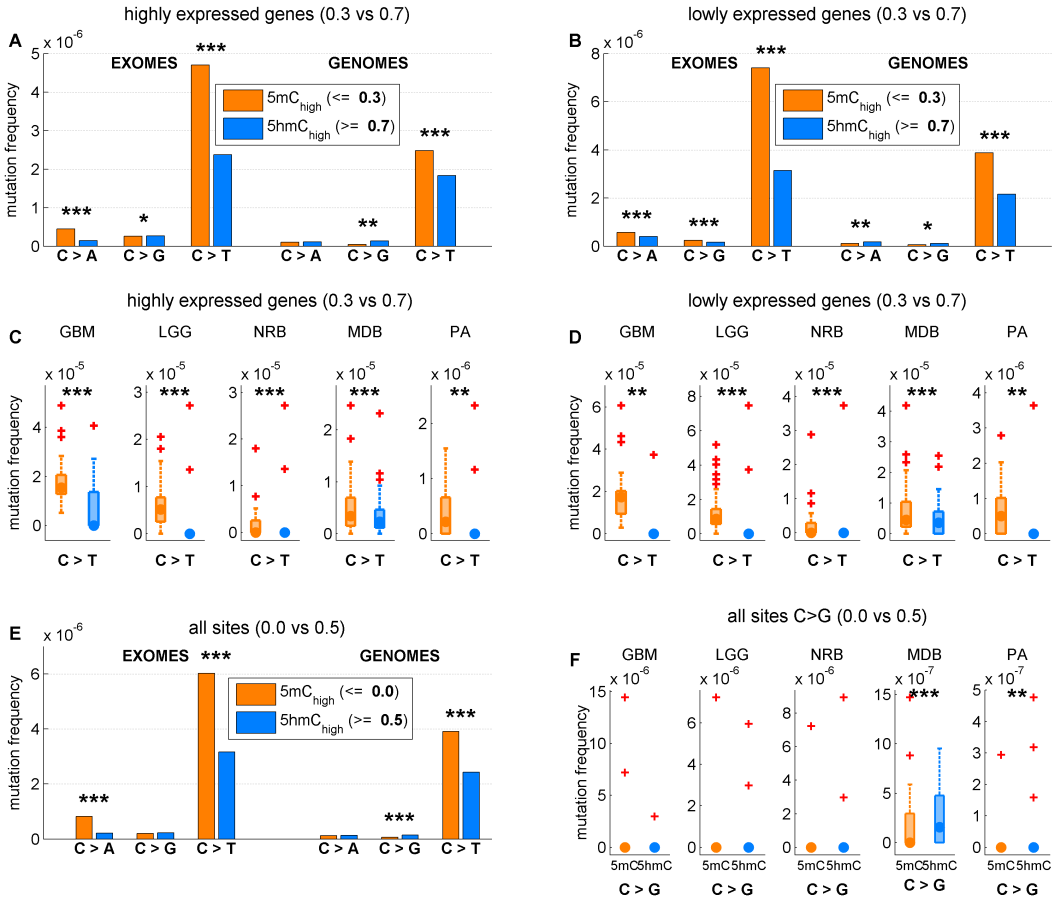

Figure 3.13. Depletion of C>T mutations in $5hmC_{high}$ is relatively insensitive to varying definitions of $5mC_{high}$ and $5hmC_{high}$. **A–D:** C>T mutation frequency for $5mC_{high}$ vs. $5hmC_{high}$ in highly vs. lowly expressed genes with $threshold_{5mC}$ = 0.3 and $threshold_{5hmC}$ = 0.7. **E–F:** C>G mutation frequency with $threshold_{5mC}$ = 0.0 and $threshold_{5hmC}$ = 0.5.

3.3.2 Reduced 5hmC mutability in brain is not accounted for by genomic regions or gene expression

We next examined whether the decreased frequency of C>T transitions at 5hmC vs. 5mC sites might be an indirect effect of 5hmC being associated with genomic regions of lower mutability. Levels of 5mC and 5hmC vary across genomic regions. For example, 5hmC density is elevated in highly expressed genes in brain (Mellén et al., 2012; Yu et al., 2012; Lister et al., 2013; Wen et al., 2014; Chen et al., 2015). The observed decrease in C>T mutation frequencies might therefore be attributable to higher gene expression, which would correlate with higher transcription coupled repair. We therefore performed the analysis described above separately for regions with high vs. low gene expression. Genes were sorted according to their median expression values in human brain (see section 3.2). The upper 50-percentile (9 701 genes) were classified as highly expressed, the rest as lowly expressed. Introns were included only for WGS samples.

There was a lower overall mutation frequency in highly expressed genes (Fig. 3.14A–B), but both highly and lowly expressed genes exhibited significantly lower C>T transition rates at 5hmC sites compared to 5mC sites (Fig. 3.14). This suggests that the observed difference between 5mC and 5hmC is not a result of transcription coupled repair.

Gene expression is only one of many possible region-related confounding factors. Hence, to correct for any regional variation, we performed the analysis on groups of sites generated by pairing the modified CpGs: each 5hmC site was paired with the nearest yet unpaired 5mC site from an equivalent genomic and sequence context (an approach adapted from (Supek et al., 2014a)). For each $5hmC_{high}$ site in random order, the nearest yet unselected $5mC_{high}$ site was selected such that the 5mC-5hmC pair fulfilled the following conditions: both $5hmC_{high}$ and $5mC_{high}$ sites are inside an exon or both are outside exons, and both share the same context (CG, CHG, and CHH, where H is T, A or C). This resulted in 6 801 374 pairs with a median distance of 1 and 25th and 75th quantiles of -177 and +177, respectively. Thereby we compared the mutation frequencies of two groups (one group comprising 5mC sites and one group comprising 5hmC sites) containing the same number of loci.

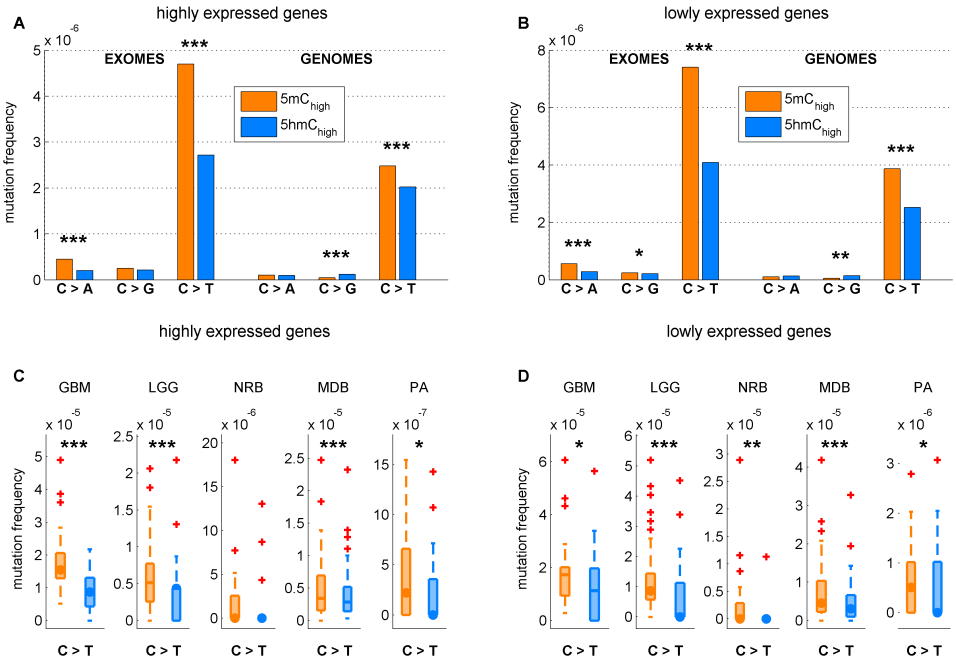

Figure 3.14. Depletion of C>T mutations in 5hmC sites is not explained by gene expression. **A–B:** Frequency of mutations in $5mC_{high}$ vs $5hmC_{high}$ sites within highly expressed (A) or lowly expressed (B) genes. **C–D:** Boxplot visualisation of C>T mutation frequency for each cancer type.

As a result of this experimental set-up, a substantial fraction of mutated 5mC sites were excluded, greatly reducing the statistical power of this "paired" analysis. Nevertheless, the frequency of C>T mutations in 5hmC remained significantly lower than in 5mC in both exomes and genomes (Fig. 3.15), supporting that the difference between 5mC and 5hmC mutation frequency is not caused by regional differences.

Finally, to complement the analysis of regional mutation rate variation with a third approach, we computed mutation frequencies around 5mC and 5hmC sites. First, modified sites with no other modifications in a 2 kbp radius were randomly selected (374 000 sites with 5mC and the same number of 5hmC sites). Next, the frequency of all mutation types in distance up 2 kbp upstream and downstream (in bins without other modifications) was computed. The mutation frequency differed substantially in the aligned positions of DNA modifications but was indistinguishable in the surrounding area (Fig. 3.16). In conclusion, regional mutation rate variability is unlikely to account for the difference in C>T mutational load in 5mC and 5hmC sites.

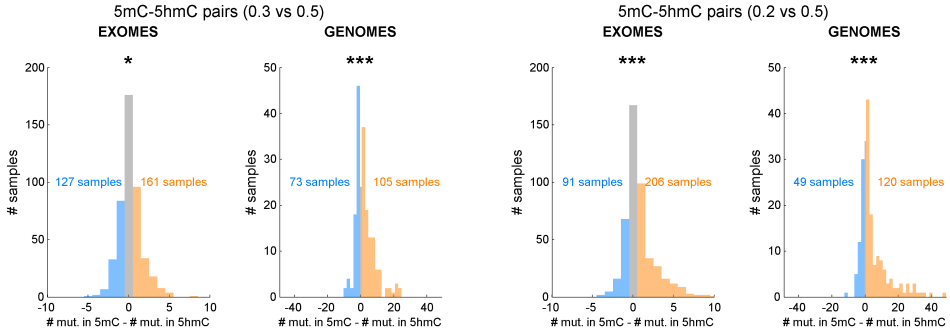

Figure 3.15. Depletion of C>T mutations in 5hmC sites is not explained by regional mutation rate variation. For each patient sample, the overall difference in mutations in paired sites was calculated and compared using a Wilcoxon signed-rank test. Shown here is a histogram of samples by the difference in mutations for paired 5mC and 5hmC sites (negative values shown blue, positive in orange). Mutations in 5mC sites exceed paired 5hmC sites, causing a shift to the right. Left: basic definition of $5mC_{high}$ and $5hmC_{high}$. Right: more stringent definition $5mC_{high}$ (threshold$_{5mC}$ = 0.2).

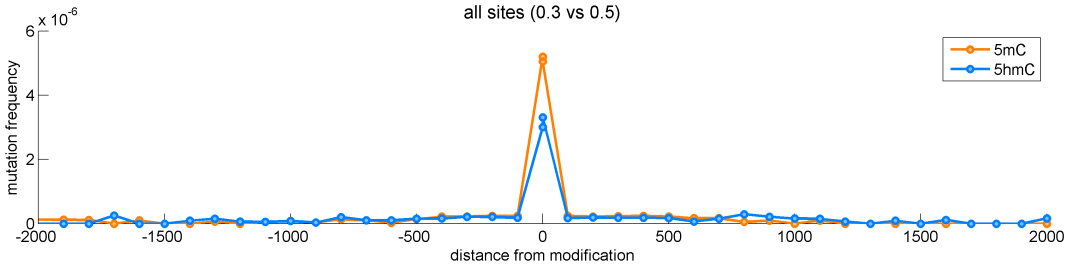

Figure 3.16. Depletion of C>T mutations in $5hmC_{high}$ is relatively insensitive to varying definitions of $5mC_{high}$ and $5hmC_{high}$. Mutation frequency around aligned 5mC and 5hmC sites.

3.3.3 Relative 5hmC correlates with CpG>TpG mutation frequency

The 5mC and 5hmC frequency at each base reflect the prevalence of each modification within the sequenced cell population. We hypothesised that if 5hmC had a direct effect on C>T mutation likelihood, we would observe an increase in mutation frequency with decreasing 5hmC occupancy. To test this, all modified cytosines (i.e., mod level > 10%) in the CpG context were divided into 9 right-open intervals according to their ratio of $5hmC_{rel}$ level. The leftmost bin contained cytosines where the major modification is 5mC, while the rightmost bin contained cytosines where the major modification is 5hmC. In each bin, the frequency of mutations was computed. A linear regression model was fitted to the data (function `fitlm` in Matlab) and the significance of the linear

coefficient was tested using F-test for the hypothesis that the regression coefficient is zero (function `coefTest` in Matlab).

We observed a clear linear relationship between $5hmC_{rel}$ values and C>T mutation frequencies (Fig. 3.17A). Notably, the correlation was present in all the tested brain cancer types in exome- and whole genome-sequenced samples. A regression slope test confirmed significance of this relationship in all the cancer types.

To confirm that the results are not influenced by an uneven distribution of information across bins, we performed also quantile binning so that each bin contains an approximately equal number of positions (apart from the first bin, which included all values with $5hmC_{rel}$=0). The results of quantile bins were equivalent to evenly spaced bins (Fig. 3.18H).

For comparison, we also evaluated the relationship between $5hmC_{rel}$ and the frequency of C>A and C>G mutations (Fig. 3.17A). Consistently with our previous results, an increase in $5hmC_{rel}$ is associated with an increase in C>G mutations in whole genomes (from MDB and PA samples), but the relationship in other cancer types shows no significant trend. C>A mutations decrease with $5hmC_{rel}$ levels in GBM but exhibit no significant signal in the remaining tumour types.

Next we compared mutation frequencies at 5mC and 5hmC sites to that of unmodified cytosines. We divided all the sequenced CpG sites into 9×9 bins according to their levels of 5mC and 5hmC. We observed that the mutation frequency of unmodified cytosine is similar to 5hmC, whereas 5mC exhibited much higher mutation frequency (Fig. 3.17B). Further, we calculated the mutation frequency distribution in sites that exhibited almost no methylation or almost no hydroxymethylation, respectively. When methylated sites are excluded, the mutation frequency does not show any significant trend with increasing levels of 5hmC (Fig. 3.17C). Conversely, excluding hydroxymethylated sites leads to a significant gradient in mutation frequency with increasing levels of 5mC (Fig. 3.17D). When only fully modified sites (mod level $\geq 90\%$) are taken into account, increasing levels of 5hmC (i.e., decreasing levels of 5mC) are associated with a significant decrease in C>T mutation frequency (Fig. 3.17E).

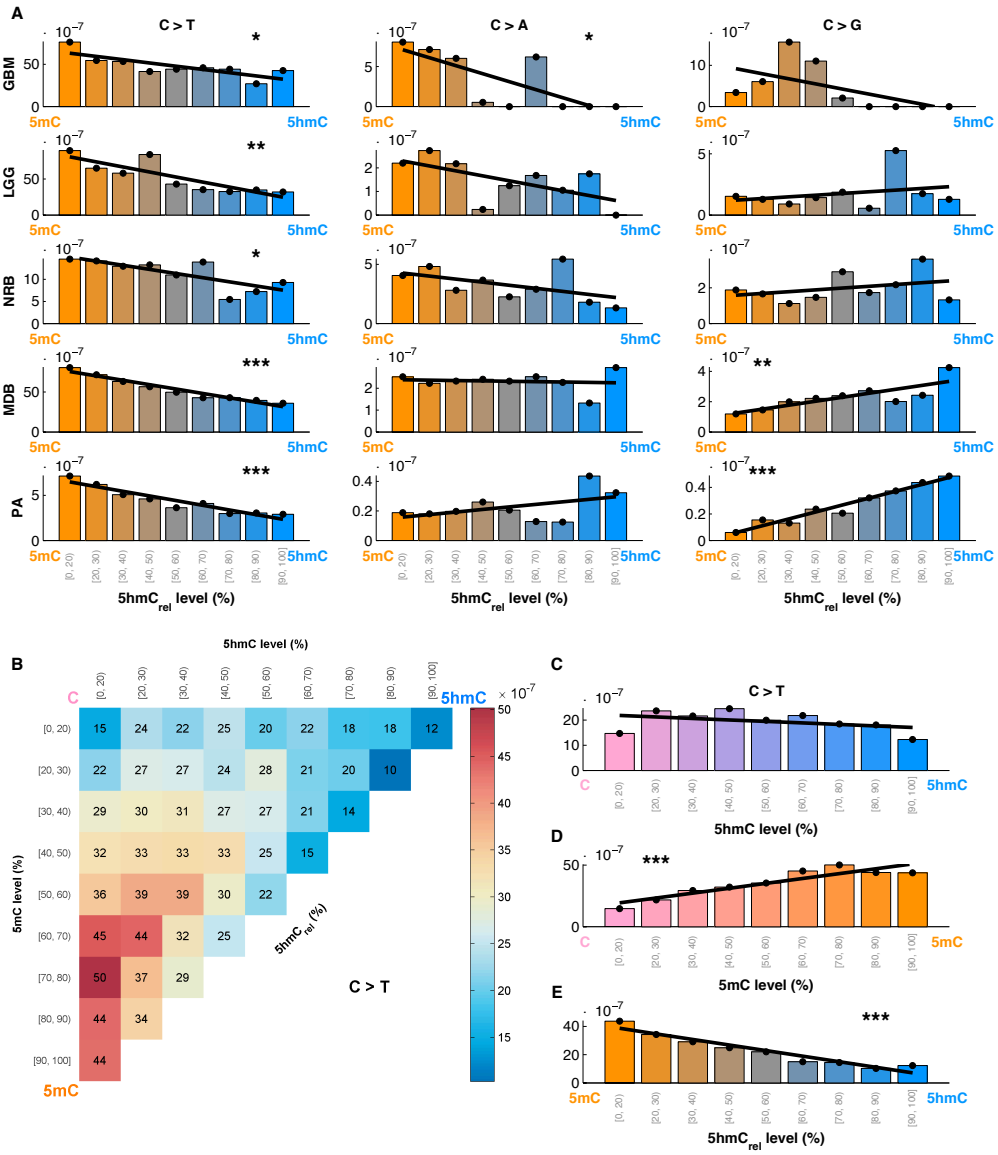

Figure 3.17. Mutation frequency negatively correlates with 5hmC$_{rel}$ level per base. **A:** Fraction of mutated CpG sites as a function of 5hmC$_{rel}$ levels by mutation and cancer type. Bins to the left represent sites predominantly methylated, while bins to the right contain increasingly hydroxymethylated sites. Black line denotes linear regression fit (F-test for coefficient deviation from 0). **B:** Distribution of CpG>TpG mutation frequency by modification type. The top left bin contains cytosines that are mostly unmodified, the bottom left bin contains exclusively methylated cytosines and the top right bin contains cytosines that are mostly hydroxymethylated. **C:** Top row of B, i.e. distribution of mutations in unmethylated sites. **D:** First column of B, i.e., distribution of mutations in sites without 5hmC. **E:** Diagonal of B, i.e., distribution of mutations in highly modified sites.

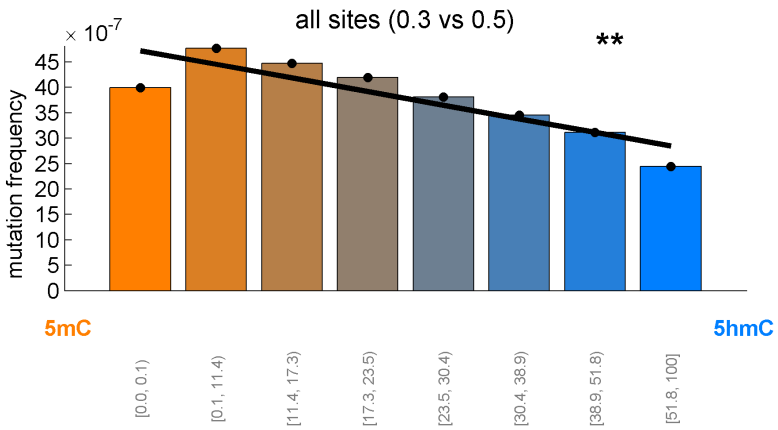

Figure 3.18. CpG>TpG mutation frequency as a function of $5hmC_{rel}$ levels with equal binning (each bin contains approximately the same number of sites).

In summary, the results in this section support the conclusion that the decrease in C>T mutation frequency at 5hmC sites is not an artefact of our chosen definition of 5mC or 5hmC. Even more importantly, it supports the notion that this decrease is directly caused by the properties of these DNA modifications.

3.3.4 5hmC is a predictor of CpG>TpG mutation frequency across the genome

To examine the exclusive impact of DNA modifications on regional frequencies of mutations, we compared DNA modifications with other genomic features in their ability to predict C>T mutations in CpG context. We used a generalised linear model (GLM) with CpG>TpG mutation frequency as a response variable and genomic features as individual predictors: average 5mC, average 5hmC, average $5hmC_{rel}$, average mod levels, gene density, exon density, CpG island density, density of modified CpGs, and gene expression (as log(1+expression)). Only whole genome sequencing data were used for this analysis. Values of the response variable and individual predictors were computed in genomic windows of sizes 3 kbp–3 Mbp. Then a generalised linear model (`fitglm`) assuming Poisson distribution of the response variable was fitted with a linear model specification (i.e., intercept + linear term for each predictor) and `DispersionFlag` set to true. To compare the resulting models, we calculated their respective "explained deviance" D^2 (`model.devianceTest`), a generalisation of

explained variance that is more appropriate for comparing generalised linear models, as recommended, e.g., in (Guisan and Zimmermann, 2000; Mittlböck and Heinzl, 2004). We compared the predictors in an iterative way, starting with a best individual predictor, and then in each step adding the predictor which leads to the best improvement of the explained deviance.

For genomic windows of 100 kbp, the best individual predictor of CpG>TpG mutation frequency was $5hmC_{rel}$ (D^2 = 0.11), outperforming all other features (Fig. 3.19A). When all features were combined into one model, the total explained deviance for 100 kbp windows was 16%. Both 5mC and 5hmC levels were amongst the top three predictors in the combined (step-wise) model, suggesting that they contain to some extent independent information predictive of the CpG>TpG mutations. On the other hand, average mod levels (the sum of 5mC and 5hmC levels) performed worst, possibly due to opposing effects of 5mC and 5hmC. This has an important consequence, suggesting that bisulfite sequencing measurements alone are a poor predictor of mutagenicity, at least in tissues with higher levels of 5hmC.

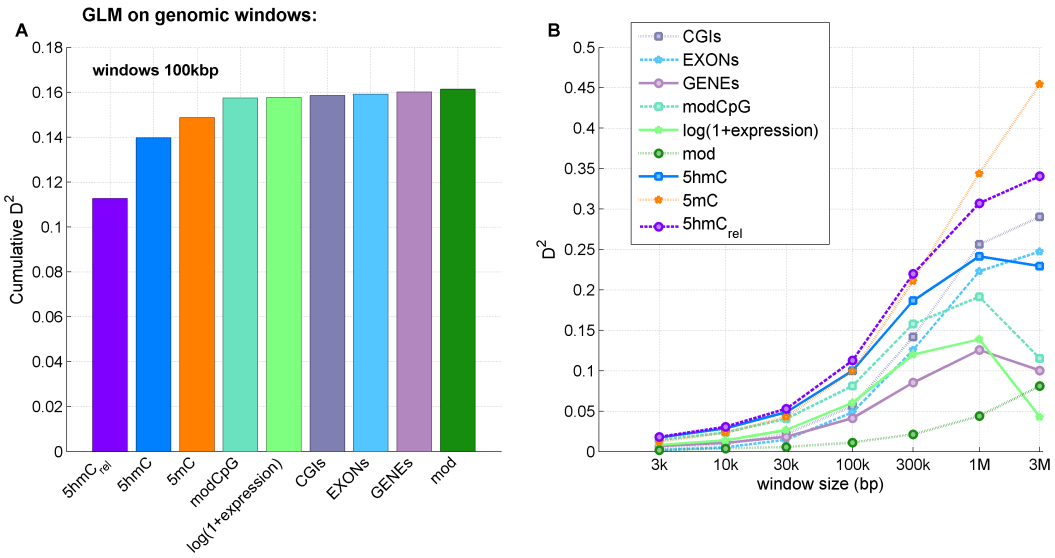

Figure 3.19. Predictors of CpG>TpG mutations: $5hmC_{rel}$ compared to other genomic features. **A:** Prediction of CpG>TpG mutation frequency (using whole genome sequencing only) in 100 kbp genomic windows. Predictors are sorted according to the D^2 in a univariate model. The height of the kth bar denotes the D^2 of a model with the first k predictors. **B:** Comparison of the nine predictors of CpG>TpG mutation features by D^2 in a univariate models, in a range of window sizes.

Varying the chosen window size (3 kbp–3 Mbp; Fig. 3.19B, Fig. 3.20A–C) did not substantially change the comparison of the predictive power of the respective features and similar order of the features was obtained also with p-value of the univariate models, and Spearman and Pearson correlation coefficients. In all metrics and window sizes, 5mC and $5hmC_{rel}$ were the two best predictors, with $5hmC_{rel}$ performing slightly better with smaller windows. However, the window sizes differed in the total explained deviance, which increased with window size, reaching values as high as 45% for univariate models and 60% for models with all predictors. This led us to question whether the increasing predictive power of larger windows has a biological reason, or whether it is a consequence of the lower data density in small windows.

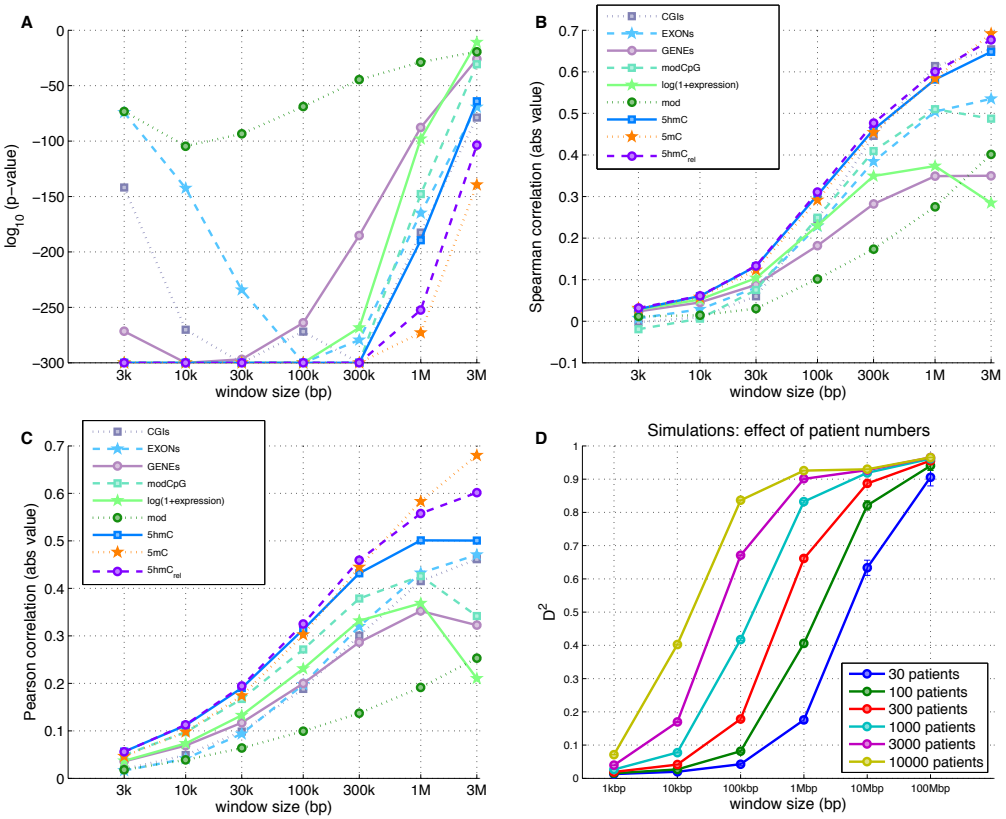

Figure 3.20. Genome-wide prediction of CpG>TpG mutation frequency: $5hmC_{rel}$ compared to other genomic features. A–C: Comparison of nine predictors of CpG>TpG mutation frequency in a range of window sizes by p-value of univariate generalised linear models (A), Spearman correlation (B), and Pearson correlation (C). **D:** Effects of window size and patient numbers on D^2 of GLM with one response variable (simulated mutation frequency) generated proportionally from a single ideal predictor.

Since many smaller windows contain no observed mutations, low D^2 values could

simply reflect a lack of data. To test this, we generated simulated mutations so that a "perfect" predictor was linearly related to the mutation likelihood per window. Each chromosome was split into windows of a given window size. For each window, all CpG sites were counted. A random `predictor` was generated in each window with a beta distribution Beta(3,4). For each patient, a random number of mutations was generated in each window w of size s_w as

$$\text{Binomial}\left(n = s_w, p = \frac{\texttt{predictor}(w)}{c}\right) \tag{3.1}$$

where:

$$c = \frac{\sum_{\text{window } w} s_w \cdot \texttt{predictor}(w)}{174} \tag{3.2}$$

The coefficient c was set so that the expected total number of mutations per patient summed to 174, the observed average number of CpG>TpG mutations in brain WGS data. The response variable was set as the average CpG>TpG mutation frequency over all patients. A GLM was fit on the given predictor and response variable and D^2 was measured. The process was repeated 10 times for each combination of window size and number of patients.

We then measured the effect of window and sample size (number of patients) on the observed D^2, repeating the simulations 10 times. The resulting curves of the explained deviance resemble those measured in the real data (Fig. 3.20D). Moreover, in the simulated data, higher numbers of patients lead to higher D^2 even for smaller window sizes, suggesting that lower D^2 values in smaller windows indeed are a consequence of lower data density.

3.3.5 Level of genic 5hmC correlates with decrease of CpG>TpG

It has been reported that 5hmC is enriched in gene bodies. We therefore tested whether the relationship between 5hmC and mutations, which we observed across the whole genome, is also detectable in the exome part of genes alone. We used again the same

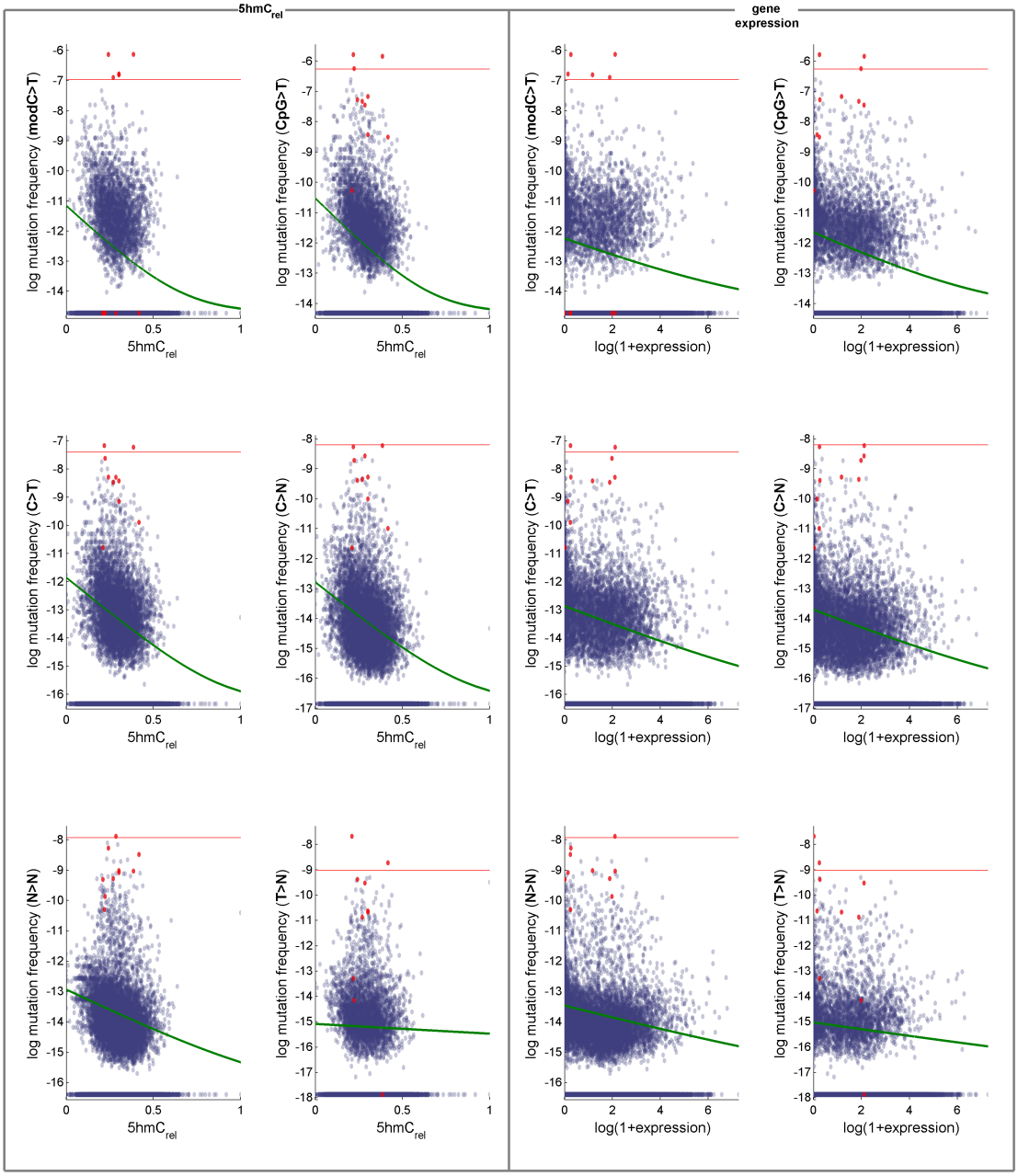

Figure 3.21. Effects of $5hmC_{rel}$ levels on gene mutability. Data for GLM with Poisson distribution (the fitted curve is in green). Genes defined as outliers in at least one definition of mutation frequency (above the red line) are plotted in red. For convenience, the mutation frequency is plotted on log-scale.

GLM, with mutation frequency modelled with two predictor variables: average 5hmC$_{rel}$ per gene and log_e-transformed gene expression. The following response variables computed in exons of each gene were compared:

- modC>T: number of C>T mutations in modified C sites / number of modified C sites
- CpG>TpG: number of C>T mutations in CpG sites / number of CpG sites
- C>T: number of C>T mutations / number of C sites
- C>N: number of mutations from C / number of C sites
- N>N: number of mutations / number of sites
- T>N: number of mutations from T / number of T sites

Genes with missing values in at least one of the predictors and genes classified as outliers in at least one response variable were excluded from the analysis. Outliers were classified as genes with response variable y with the following property:

$$y \geq \text{quantile}(y, 0.999) + 2.5 \cdot (\text{quantile}(y, 0.999) - \text{quantile}(y, 0.001)) \tag{3.3}$$

Out of 17,605 genes, 10 were classified and removed as outliers: ASPN, BBOX1, CCL4, ESPN, FOLH1, HLA-DPB1, IDH1, NLRP6, S100P, and TP53. To calculate the relative contribution of one predictor variable over the other, two models were fitted with either one or both predictor variables and F-test and the difference in D^2 were used to compare the two nested models. The individual models are shown in Fig. 3.21 with genes classified as outliers plotted in red.

In line with our earlier results, we found that 5hmC$_{rel}$ significantly contributes to the deviance explained by the model, beyond covariation with gene expression (Fig. 3.22; F-test p < 2e-100). We hypothesised that this effect should be most pronounced when using modC>T and CpG>TpG as the response variable, whereas it should decrease when using definitions of mutations that include a progressively wider range of loci (C>T, C>N, N>N). Indeed, the unique contribution of 5hmC$_{rel}$ to the explained gene

mutation frequency decreased as the mutation sets became larger and more distant from modC>T, as both the improvement of explained variance decreased and the model p-value increased (Fig. 3.22, Fig. 3.21, Fig. 3.23). Nevertheless, in all of the cases, $5hmC_{rel}$ significantly improved the fit of the model. Conversely, we confirmed that $5hmC_{rel}$ had no significant predictive power for the frequency of T>N mutations (Fig. 3.22; column T>N), supporting the hypothesis that $5hmC_{rel}$ selectively affects mutations in modified cytosines.

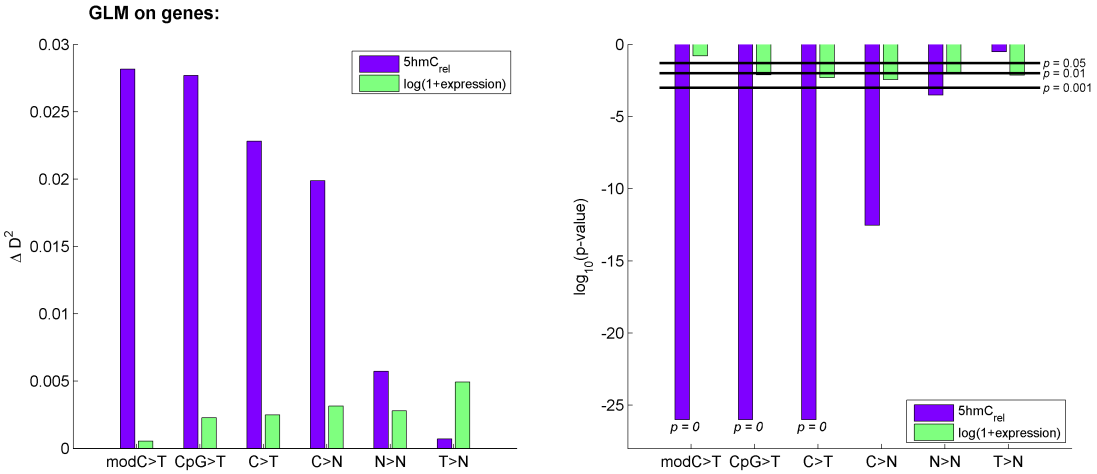

Figure 3.22. Predictors of mutations: $5hmC_{rel}$ compared to gene expression. Left: Prediction of different types of mutation frequency in genes. Increase in D^2 of a generalised linear model including $5hmC_{rel}$ over gene expression (violet) or gene expression over $5hmC_{rel}$ (green). Right: Significance of observations in left.

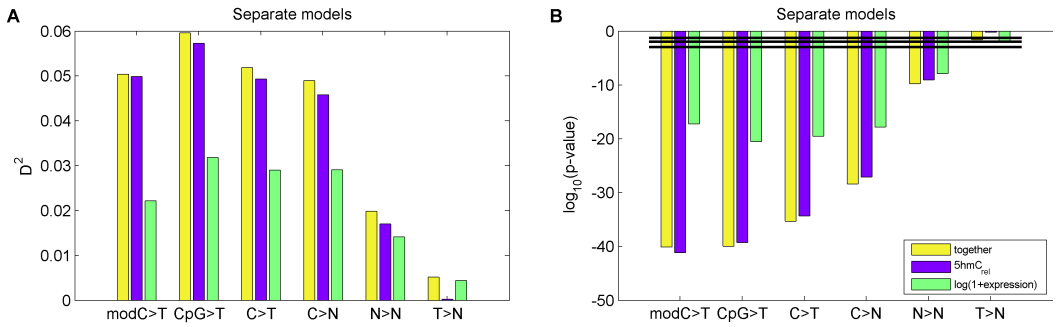

Figure 3.23. Effects of $5hmC_{rel}$ levels on gene mutability. **A–B:** GLM results fitted separately for $5hmC_{rel}$ (violet) and gene expression (green) and both of them together (yellow).

The comparison of $5hmC_{rel}$ and gene expression in their ability to predict genic C>T mutation frequency in modified sites is illustrated in the left panel of Fig. 3.24. For all values of gene expression (rows in the figure), there is a gradient of high mutation frequency in genes with low $5hmC_{rel}$ (left) to low mutation frequency in genes with high $5hmC_{rel}$ (right). On the contrary, such a gradient is much less apparent in the opposite direction. In summary, although $5hmC_{rel}$ and gene expression are correlated, these results suggest that the effect of DNA modifications on CpG>TpG mutations is greater than the effect of gene expression.

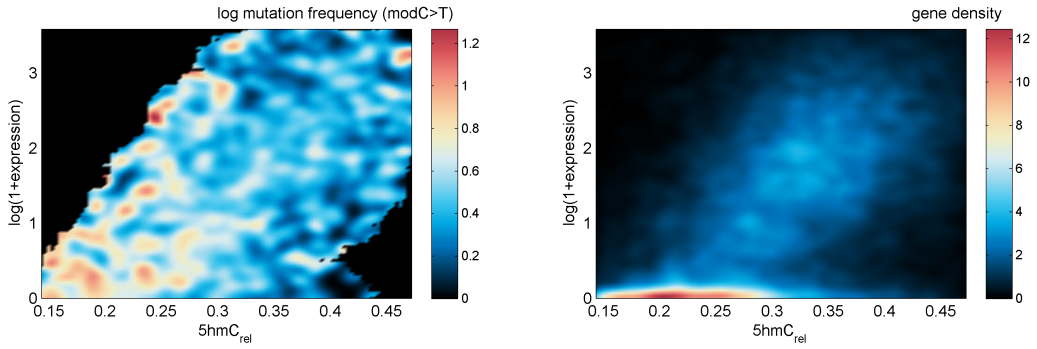

Figure 3.24. Effects of $5hmC_{rel}$ levels on gene mutability. Frequency of modC>T mutations of all genes (left) and gene density (right) in the space of $5hmC_{rel}$ and gene expression. The space was limited to $[\text{quantile}(x, 0.05), \text{quantile}(x, 0.95)]$ on both axes and then binned into 100 × 100 bins. In each bin, the average mutation frequency, in the form of $\log(\text{mutFreq} + \min(\text{mutFreq}(\text{mutFreq} > 0)))$, and gene density were computed. The resulting matrix were smoothed by applying a Gaussian filter (radius 5 bins, sigma 2) weighted by the number of genes in each bin (bins with $\geq 2/3$ missing values in their neighbourhood were set to NaN) and plotted with `pcolor` (NaN bins are shown in black).

3.3.6 Decreased CpG>TpG mutation frequency in 5hmC is not limited to brain tissue

The results from brain cancers showed that positions with 5hmC in normal brain are associated with decreased frequency of mutations in brain tumours compared to positions with 5mC. These results had two major limitations: they were based on maps of 5mC and 5hmC from only a single individual and they cannot answer whether this is a specific characteristic of the brain tissue, or a general property of 5mC and 5hmC in all somatic cells. However, two newly published BS-seq and TAB-seq datasets allowed us to address the question of mutational properties of 5mC and 5hmC also in two other tissues: kidney (Chen et al., 2015) and blood (Pacis et al., 2015). For blood we used 174 sequencing samples from Acute Myeloid Leukaemia (AML) as the cancer type closest to the blood dendritic cells in which the BS-seq and TAB-seq measurements were performed. For kidney we combined 585 samples from four distinct sequencing projects, covering Kidney Clear Cell, Kidney Papillary and Kidney Chromophobe carcinomas.

These three tissue types show a wide range of average 5hmC levels in CpG sites: from 2.5 % in blood to 19.9 % in brain (Fig. 3.1). The two biological replicates in kidney allowed us to compare inter-tissue and inter-individual differences in 5hmC levels. Of the four samples, the two kidney samples were best correlated (Pearson coefficient r = 0.83; 10 kbp windows), whereas the worst correlated pair consisted of the blood and brain samples (r = 0.39) (Fig. 3.25).

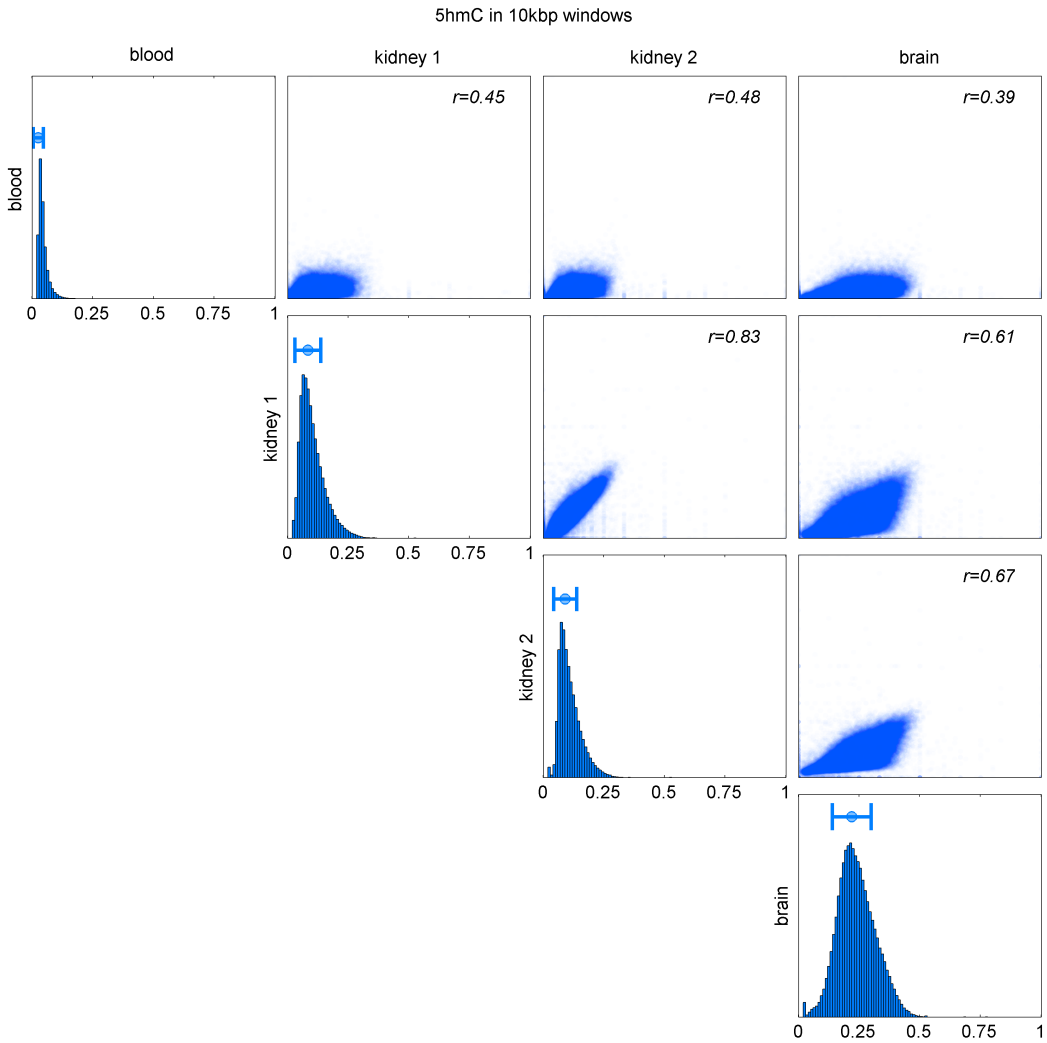

Figure 3.25. Comparison of 5hmC in 10 kbp **windows in blood, kidney (2 replicates), and brain:** distribution of 5hmC values in each map and Pearson correlation of pairs of maps.

Matching our findings in brain, 5hmC sites were mutated significantly less frequently than 5mC sites in both tissue types (Fig. 3.26), irrespective of whether genome or exome sequencing data were used. Moreover, a similar difference was present in all available replicates of the BS-seq and TAB-seq measurements (6 for blood, 2 for kidney, Fig. 3.27).

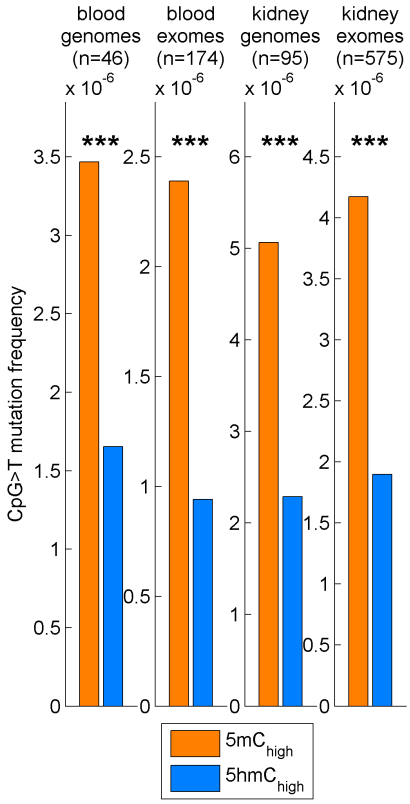

Figure 3.26. Decreased CpG>TpG mutation frequency in 5hmC is not limited to brain tissue. CpG>TpG mutation frequency in 5mC vs. 5hmC in kidney and blood.

Genomic distribution of 5hmC differs substantially between the three tissue types (Fig. 3.25). Consequently, we hypothesised that the association between mutation frequency and 5hmC could be highest when mutation and modification data are derived from matching tissue types. To test this hypothesis, we used a GLM on genomic windows of 100 kbp to predict CpG>TpG mutation rate from a combination of $5hmC_{rel}$ maps of all three tissues. In line with the hypothesis, for each cancer type, the best predictor came from the same tissue type (Fig. 3.28), suggesting that tissue differences

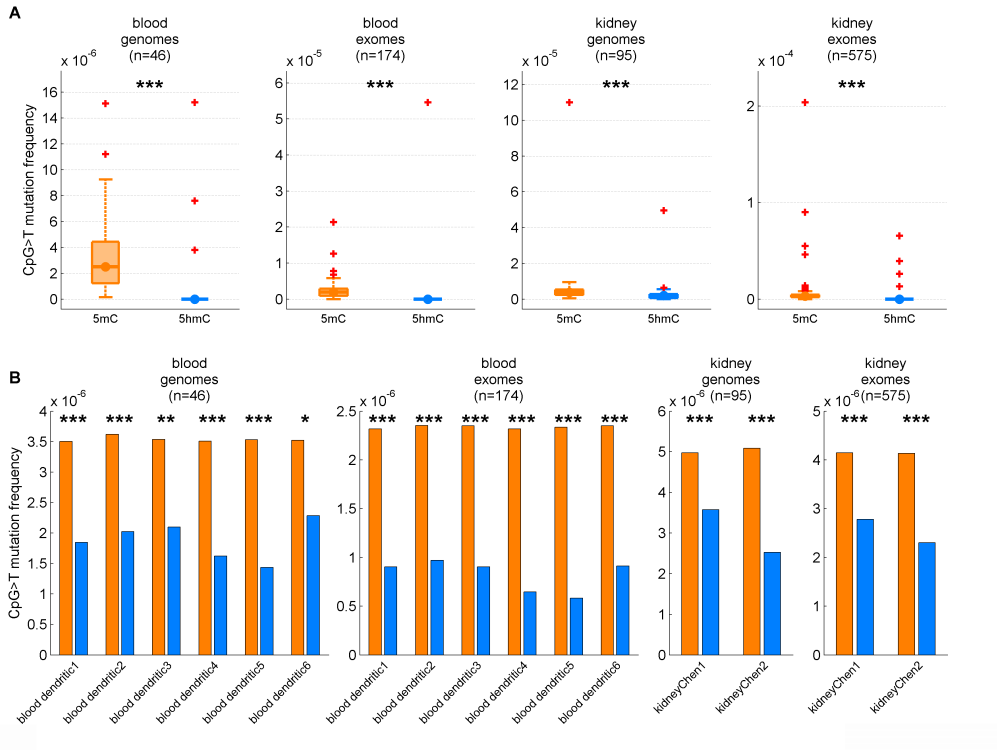

Figure 3.27. Decreased CpG>TpG mutation frequency in 5hmC is present in three tissues consistently for different replicates of modification maps. **A–B:** CpG>TpG mutation frequency in 5mC compared to 5hmC in blood and kidney using modification maps from different replicates merged together (A) and used separately (B).

in 5hmC are reflected in the CpG>TpG mutation landscape. The same results were obtained in all available replicates of the $5hmC_{rel}$ maps (Fig. 3.29). Finally, we added a $5hmC_{rel}$ map derived from embryonic stem cells (ESC) as an additional predictor, to compare our findings to previously reported results (Supek et al., 2014a). The ESC-derived 5hmC levels have substantially lower predictive power on CpG>TpG mutation rate than any of the tissue-derived maps, likely reflecting the substantial differences of 5hmC in ESC compared to the other tissues. These results highlight the importance of matching tissues, when comparing DNA modifications with other genomic properties, such as mutability of DNA.

While base-resolution maps of 5hmC for human tissue are still rare, there is a wide range of BS-seq data sets available in public databases. We therefore decided to measure global levels of 5mC and 5hmC in different tissues and combine these estimates with the single-base resolution mod maps and somatic mutations. The total levels of 5mC

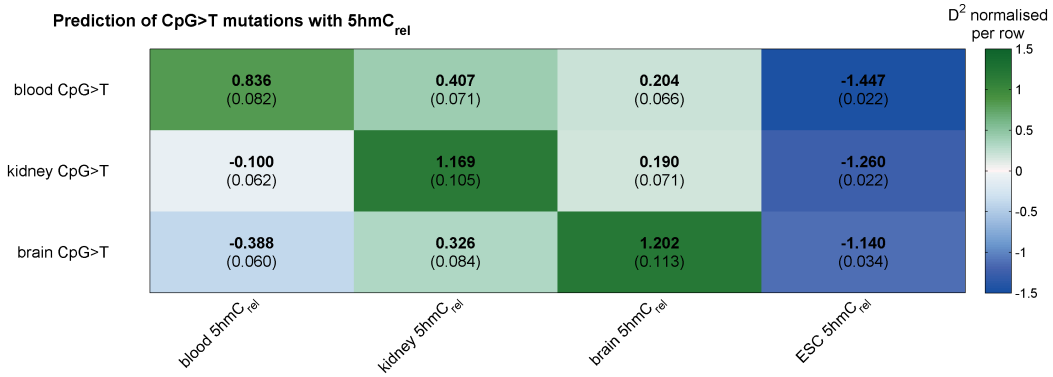

Figure 3.28. Decreased CpG>TpG mutation frequency in 5hmC is not limited to brain tissue. Predictions of CpG>TpG mutation frequency in whole genome cancers in blood (AML), kidney and brain using $5hmC_{rel}$ maps from blood, kidney, brain and embryonic stem cells (ESC) in 100 kbp genomic windows. The values are z-score normalised per rows in order to normalise for different number of patients and mutations in each cancer type (the original D^2 values are in parentheses); the higher values of D^2 (green), the better predictions.

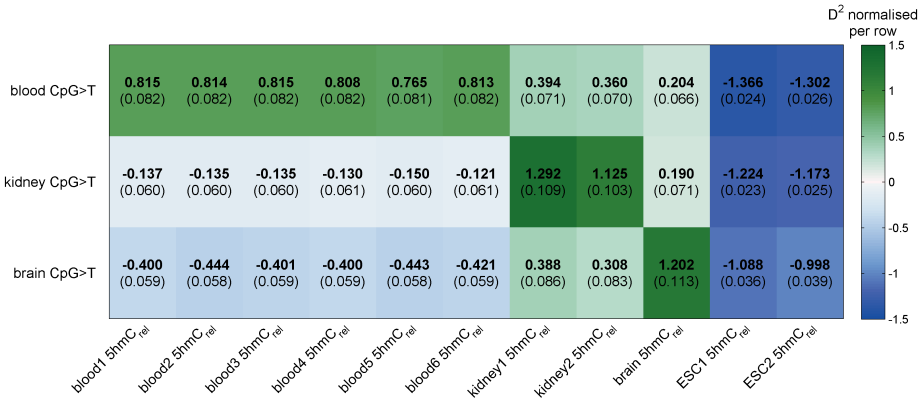

Figure 3.29. Decreased CpG>TpG mutation frequency in 5hmC is present in three tissues consistently for different replicates of modification maps. Predictions of CpG>TpG mutation frequency in whole genome cancers in blood (AML), kidney and brain using different replicates of $5hmC_{rel}$ maps from blood, kidney, brain and embryonic stem cells (ESC) in 100 kbp genomic windows. The values are z-score normalised per rows in order to normalise for different number of patients and mutations in each cancer type (the original D^2 values are in parentheses); the higher values of D^2 (green), the better predictions.

and 5hmC were measured using High Pressure Liquid Chromatography (HPLC-UV) by Michael McClellan; details are explained in the methods of Tomkova et al. (2016). As expected, brain contained the highest levels of 5hmC (1.50 $\pm$ 0.07 % of all cytosines) and blood was on the other end of the spectrum, with ca. 20-fold lower levels of 5hmC (0.07 $\pm$ 0.10 % of all cytosines) (Fig. 3.30). The levels of methylation were very similar in these two tissues (4.56 $\pm$ 0.46 in brain; 4.24 $\pm$ 0.22 in blood) and generally showed smaller relative differences among all tissues than the hydroxymethylated levels.

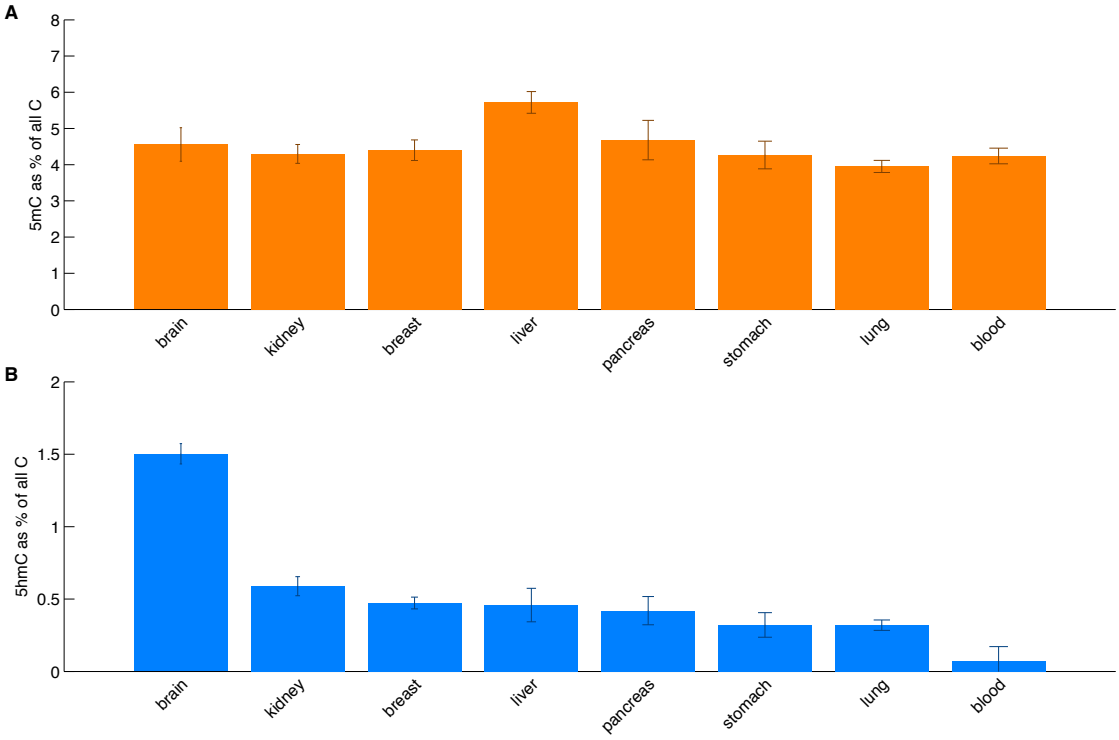

Figure 3.30. HPLC measurements of total 5hmC and 5mC in eight tissues: average values with standard deviation of 5mC and 5hmC (as a percentage of total cytosine). Measured by Michael McClellan.

Given our findings thus far, we predicted that tissues with high levels of 5hmC relative to 5mC would exhibit fewer CpG>TpG mutations in modified sites than tissues with low total 5hmC. To test this hypothesis, we compared total levels of 5mC and 5hmC in DNA of eight human tissue types for which BS-seq maps are publicly available(Table 9.1). As the total measurements are from the entire DNA, we included only WGS samples for estimates of the CpG>TpG mutability. In order to account for unrelated cancer-type specific differences in CpG>TpG mutability, we normalised the mutation

frequency in modified sites by the mutation frequency in unmodified sites. In each tissue, we computed the average of (C>T mutation frequency in modified CpGs / C>T mutation frequency in unmodified CpGs) and plotted these values against the global estimates of 5hmC/(5hmC+5mC) per tissue.

The analysis of association between genomic relative 5hmC and enrichment of CpG>TpG mutations revealed a strong negative correlation in nearly all tissues (Fig. 3.31). For instance brain had both high relative 5hmC levels and low relative C>T mutations in modified CpGs, whereas blood contained low relative 5hmC levels and high relative C>T mutations in modified CpGs.

Lung was the only outlier tissue distant from the linear fit, having a markedly lower frequency of CpG>TpG mutations in modified relative to unmodified sites. We explored whether this could be affected by smoking, which is a known strong mutagen affecting both mutations and modifications (Alexandrov et al., 2016). Interestingly, when splitting the samples by smoking status of the individuals (heavy smokers vs. not heavy smokers), only the group of heavy smokers showed to be a true outlier. It has been reported that methylation increases the formation of BPDE-dG bulky adducts (reviewed in Introduction 1.4.3). Moreover, direct changes in the methylation status of a number of CpGs have been observed in smokers (Breitling et al., 2011; Rakyan et al., 2011; Joehanes et al., 2016; Gao et al., 2017). Therefore, our data indicate that either CpG>TpG mutations might also be differentially affected by smoking-related mutagens, or that the actual modification maps are substantially affected by smoking.

3.3.7 Exploration of potential protective function of 5hmC

The finding that $5hmC_{rel}$ affects overall mutability of genes led us to speculate that protection against mutations could be a function of "long-lived" 5hmC. Proving or disproving this hypothesis is outside the scope of this thesis, however we performed a simple exploration for observations that might support this hypothesis. Known cancer "driver genes" —genes that are able to cause an abnormal growth phenotype when mutated— constitute a class of genes for which their disruptive potential upon mutation is well documented. We thus hypothesised that the levels of 5hmC could be

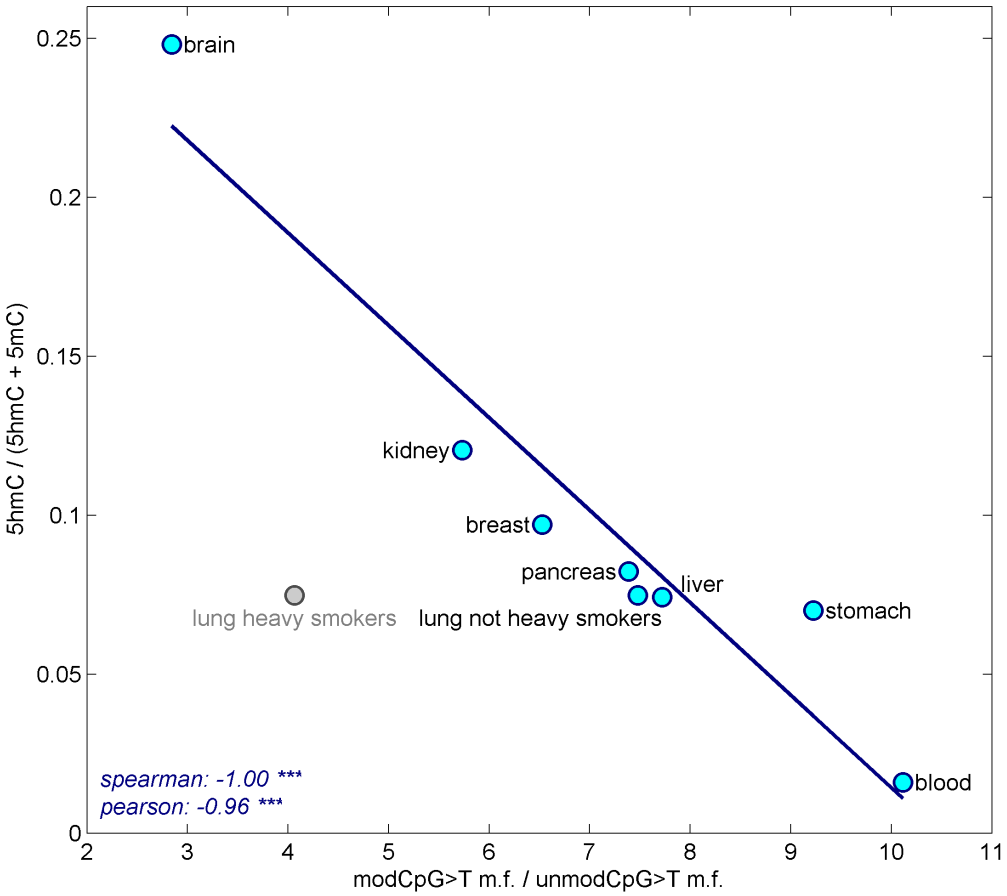

Figure 3.31. Decreased CpG>TpG mutation frequency in 5hmC is not limited to brain tissue. Correlation of total $5hmC_{rel}$ levels (measured with HPLC) with frequency of CpG>TpG mutations in modified cytosines normalised by the frequency in unmodified cytosines in different tissues. The correlation values are shown for the eight non-outlier tissues, i.e., without heavy smoking lung cancer patients.

higher in the driver genes, compared to other genes, in order to protect the cells from deleterious mutations.

We calculated the level of $5hmC_{rel}$ in 19 brain cancer driver genes (Parsons et al., 2008; TCGA, 2008; Pugh et al., 2012) and in 17 494 "non-driver" genes (see Methods 3.2). We found that the average $5hmC_{rel}$ level in the brain cancer driver genes was significantly higher than in non-driver genes (Fig. 3.32A).

However, also the average gene expression of brain cancer driver genes was higher than in the non-driver genes. We therefore repeated the analysis with a set of 152 non-driver genes with similar expression levels as the driver genes (see Methods 3.2) (Fig.

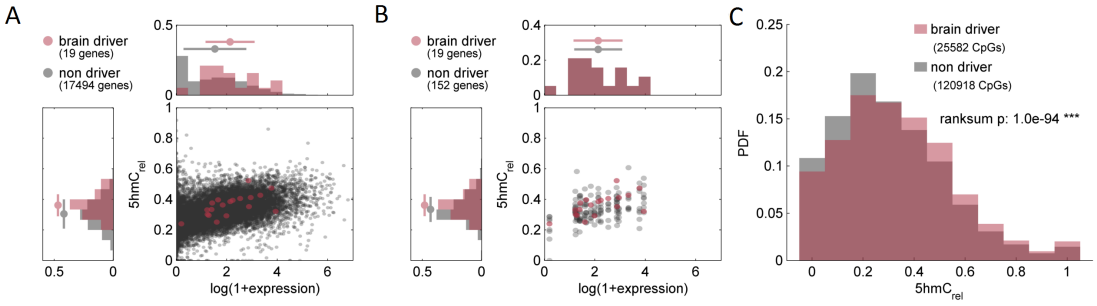

Figure 3.32. 5hmC is enriched in driven genes. A: Scatter plot of gene expression against $5hmC_{rel}$ for brain cancer driver genes (red) and all non-driver genes (black). Histograms above and to the left show expression and $5hmC_{rel}$ distribution differences for the two classes of genes, respectively. **B:** Same as E but only plotting a subset of expression-matched non-driver genes (see Methods). **C:** Detailed histogram of $5hmC_{rel}$ in individual CpG sites, illustrating elevated $5hmC_{rel}$ in driver genes compared to expression-matched non-driver genes.

3.32B) and observed higher levels of $5hmC_{rel}$ in brain cancer driver genes compared to similarly expressed non-driver genes (ranksum test $p < 10^{-94}$, Fig. 3.32C). The results therefore support the hypothesis of protective function of 5hmC in the genome. However, further research is need to prove (or disprove) the hypothesis.

3.4 Discussion

3.4.1 Results summary

Here we have established a link between the landscape of DNA modifications and the mutational profile of somatic human cells. Our measurements indicate that positions with high 5hmCs carry between 28 and 53% fewer mutations than methylated cytosines in brain. The mutation load of CpG positions without 5mC is comparable in unmodified cytosines and different levels of 5hmC. This differential mutagenicity in 5mC vs. 5hmC sites is not only observable in brain, but also in kidney cancers and myeloid leukaemias. The relationship between 5hmC and CpG>TpG mutation rate can be detected at the scale of the exome as well as genome-wide and is independent of other region-specific influences on mutation frequency. We show that the relative impact of hydroxymethylation on mutagenesis decreases proportionally to the level of relative 5hmC in the tissue, suggesting that it represents a general property of this DNA modification.

This is the first comparison of 5mC vs. 5hmC in the terms of mutability in cancer patients, using tissue-matched single-base resolution modification maps. Since the time of these results being published (Tomkova et al., 2016), the decrease of CpG>TpG mutation frequency in 5hmC compared to 5mC has also been confirmed in an independent data set of somatic mutations in a tumour biopsy and 5mC and 5hmC measurements in a matched normal sample in a Glioblastoma patient (Raiber et al., 2017).

3.4.2 Comparison of our results with the literature

It has previously been suggested that 5hmC levels increase the frequency of C>G mutations (Supek et al., 2014a). As part of their analysis, only a very small (albeit statistically significant) decrease of C>T mutations in 5hmC sites in both SNPs and cancer SNVs was observed. There are two factors that could explain why we observe very different effects sizes for C>T and C>G mutations in 5hmC sites. Firstly, Supek et al. consider all sites with as little as one 5hmC read to be hydroxymethylated, whereas we require the level of 5hmC to exceed 5mC. In fact, when examining the effect of

variation in these thresholds, we noticed that the results for C>G fluctuate substantially across the range of tested cut-off values (Fig. 3.13). Secondly, we present evidence that tissue-specific changes in 5hmC patterns have great influence on the extent of correlation between 5hmC and mutability (Fig. 3.28). Specifically, 5hmC genomic localisation in embryonic stem cells was a poor predictor of CpG>TpG mutations in brain, kidney and blood, compared to the respective tissue-specific 5hmC patterns.

Compared to the results by Supek et al. (2014a), 5hmC sites have recently been found depleted also in C>G mutations also in an independent (tissue-matched) study in Glioblastoma (Raiber et al., 2017).

3.4.3 Discussion of the potential mechanisms underlying the observed results

Two possible scenarios could explain the striking difference in mutability between 5mC and 5hmC. Firstly, spontaneous and enzymatic deamination reactions of 5hmC could be less favourable than 5mC. As a consequence, fewer new mutation events would be expected at 5hmC sites. Indeed, cytosine deaminases (namely, AID and APOBEC1-3) have 4.4–38x lower activity on sites with 5hmC compared to 5mC, supporting this possibility (Nabel et al., 2012; Rangam et al., 2012). The spontaneous deamination rate of 5hmC compared to 5mC has not been published, but unpublished *in vitro* measurements in single-stranded DNA performed by Michael McClellan and Pijus Brazauskas in Prof. Kriaucionis lab show that the deamination rate of 5mC is ca. 2-fold higher than that of 5hmC and C, in line with our observations of mutation rates in methylated, hydroxymethylated, and unmodified cytosines in cancer samples.

Secondly, deamination of 5mC produces thymine, whereas 5hmC deaminates to 5-hydroxymethyluracil (5hmU). This atypical base in DNA could be more efficiently recognised and replaced by the DNA glycosylases initiating base-excision repair (BER). Determining the relative contribution of DNA glycosylases to the lower mutation rate would be challenging, since some of these enzymes recognise several types of mismatches. TDG and MBD4 excise both T and 5hmU when mis-paired with G (Hardeland et al., 2003; Cortellino et al., 2011; Guo et al., 2011; Hashimoto et al., 2012a;

Moréra et al., 2012), whereas SMUG1 does not repair T:G but has a robust activity for 5hmU:G (Nilsen et al., 2001; Kemmerich et al., 2012). The 5hmU:G mismatches are also known substrates for the DNA glycosylases UNG2, NEIL1, and NTHL1 (Jacobs and Schär, 2012; Zhang et al., 2005).

Importantly, mass spectrometry-based isotope tracing of all major oxidized pyrimidine and purine bases in mouse ESCs showed that the steady-state levels of 5hmU reside in 5hmU:A base pairs and are derived from TET-induced oxidation of T, instead of deamination of 5hmC (Pfaffeneder et al., 2014). Therefore, the deamination of 5hmC is either not a frequent event, or the resulting 5hmU:G mismatches are very rapidly repaired. Further genome sequencing efforts might identify patients with rare inactivating mutations in the BER pathway that could be valuable for future investigations of the relationship between DNA repair and cytosine mutability.

3.4.4 Discussion of the generality of the observed results

The best predictor of CpG>TpG mutations in any of the three tested tissues was the $5hmC_{rel}$ map from the corresponding anatomical site. This provides evidence that the slow accumulation of CpG>TpG mutations in the pre-cancerous tissue was strongly influenced by the DNA modification landscape. However, any bulk tissue sample encompasses a mixture of different cell types. Mounting evidence suggests that solid tumours originate from a defined subset of cells within any one tissue type. For example, glioblastomas were proposed to originate from stem or progenitor cell types enriched in the subventricular zone, while medulloblastomas have mixed cells of origin (Visvader, 2011). Those cell types are of low abundance in normal tissue biopsies. The fact that we observe a clear inverse relationship between CpG>TpG mutations and the location of 5hmC in multiple tissue types suggests that the DNA modification landscape in cancer-progenitor cells is sufficiently similar to the tissue average to be informative about the mutation frequencies in cancer.

Under this assumption we predict that the impact of DNA modifications on the frequency of CpG>TpG mutations is likely to be bigger than measured here, since the terminally differentiated cells that make up the bulk of the tissue may have diverged

further from cancer-progenitors cells. Advancements in the identification of cancer origins and isolation of single cells, combined with single-cell bisulfite sequencing, will enable an improved assessment of the impact of DNA modifications on mutability.

The strong correlation between relative 5hmC levels in a tissue and the mutability of modified cytosine also points towards a shared underlying mutagenic process. The notable deviation of smoking-induced lung-cancers supports this hypothesis. We speculate that the deviation in smokers could have three reasons: substantial smoking-induced changes in the DNA modification maps, smoking-linked protection of modified CpG sites against C>T mutations (such as by reduced deamination rate of the 5mC paired with BPDE-dG adduct), or a yet undefined smoking-induced mutagenic mechanism that preferentially affects unmethylated CpG sites. More experimental work will be needed to elucidate the biochemical causes for this phenomenon. In the future, the linear relationship between 5hmC levels and CpG>TpG mutation rate could be used to identify other environmental mutagens with a differential effect on modified cytosines.

3.4.5 Discussion of potential evolutionary advantage of lower mutagenicity in 5hmC

Since the discovery of 5hmC eight years ago, the potential roles of this DNA modification have been extensively researched and discussed (Pfeifer et al., 2013; Hill et al., 2014; Ficz and Gribben, 2014; Brazauskas and Kriaucionis, 2014; Cimmino and Aifantis, 2016; Wu and Zhang, 2017). As reviewed in the Introduction 1.2.2, 5hmC is elevated in actively transcribed genes, in exons and enhancers, a substantial fraction of 5hmC is a stable ("long-lived") modification in tissue that undergoes little cell division, and finally 5hmC is depleted in tumours. Therefore, the fact that we observe a markedly decreased mutability of this base compared to 5mC and the fact that this has a substantial effect on the overall mutability of genes, raises the possibility that protection against mutations could be one of the functions of long-lived 5hmC.

It could conceivably be advantageous for the cell to use this DNA modification, which carries both a different signal to the unmodified C as well as protects important

regions of the genome against the mutagenic effect of 5mC. Unfortunately, proving or disproving this hypothesis is very challenging. Nevertheless, we at least attempted to obtain indirect evidence by comparing levels of 5hmC in regions that would benefit from protection against mutations with 5hmC in other regions. We thus compared the 5hmC levels in known brain cancer driver genes with non-driver genes (not reported as driver genes in any cancer type) of similar expression. We observed that the brain driver genes contain significantly higher levels of 5hmC (albeit with limited size effect), supporting the hypothesis about 5hmC protectivity.

It is worth noting that genes with high mutation frequency were overall associated with lower relative 5hmC levels. The group of driver genes therefore represents an exception of this general relationship, suggesting a non-random/functional role of the increased 5hmC in the driver genes. This supports the possibility that the function of 5hmC in the genome is not restricted to gene regulation but that 5hmC could also have a role in maintaining genome stability by protecting against the harmful mutagenic effect of 5mC. On the other hand, other explanations are possible, including a third confounding variable (which is correlated with 5hmC levels and is increased in driver genes), or involvement of regulatory functions of 5hmC in the driver genes. Nevertheless, since the time of this analysis and publication of this chapter (Tomkova et al., 2016), the concept of potential protectivity of 5hmC has been also suggested in (Cimmino and Aifantis, 2016).

Il se lève, c'est l'heure, écrase son mégot
Dans sa tasse de café, éteint la stéréo
Eteint le lampadaire, éteint le plafonnier
Eteint dans la cuisine, met la sécurité

— Bernard Lavilliers, Jean-Paul Drand, Catherine Ringer
Idées noires

Barra barra, noujoum t'fate derquéte chéms
Barra barra, ma b'qa kheir la saada la z'har
Barra barra, ma b'qa z'djour sektou lé tiour
Barra barra, ma b'qa lil ka n'har ghir dalma

— Rachid Taha *Barra barra*

4 The role of DNA modifications in different mutational processes

4.1 Introduction

Spontaneous deamination of 5mC is thought to be the main reason for the high frequency of CpG>TpG mutations observed in cancer and genetic disorders (Alexandrov et al., 2013a; Cooper and Youssoufian, 1988), healthy tissue (Blokzijl et al., 2016), and germline (Kong et al., 2012; Rahbari et al., 2015). It is also thought to be the cause of the mutational signature 1, the most common of all mutational signatures (Alexandrov et al., 2013a) and one of the only two signatures with clock-like properties, correlating with the age of patients, and therefore likely operating in normal somatic cells throughout the entire life (Alexandrov et al., 2015). However, many other signatures (1, 4, 6, 7, 8, 10, 11, 15, 18) show either increased or decreased frequency of mutations in CpG dinucleotides, after normalising for the frequency of trinucleotides in the genome (Fig. 4.1). Moreover, processes like UV-damage and tobacco smoking have known links to methylation, as reviewed in the Introduction 1.4. In this chapter, I focus on the individual mutational processes linked to these signatures and explore the role of DNA modifications in these processes. I use the publicly available maps of DNA modifications (from BS-seq and where available also from TAB-seq or oxBS-seq), publicly available data sets of somatic

mutations of patients strongly influenced by one of the mutational processes, and link the results to the existing experimental knowledge about these processes.

The role of DNA modifications in four mutational processes is researched in this chapter: replication, APOBECs, UV light, and tobacco smoking. The structure of this chapter is: a shared materials and methods section for all four mutational processes (section 4.2), results and discussion sections for each of the four processes (two are in the main text: 4.3 and 4.4.4, two are in the Appendix: 10.1 and 10.2), concluded with shared concluding remarks of the entire chapter (section 4.5).

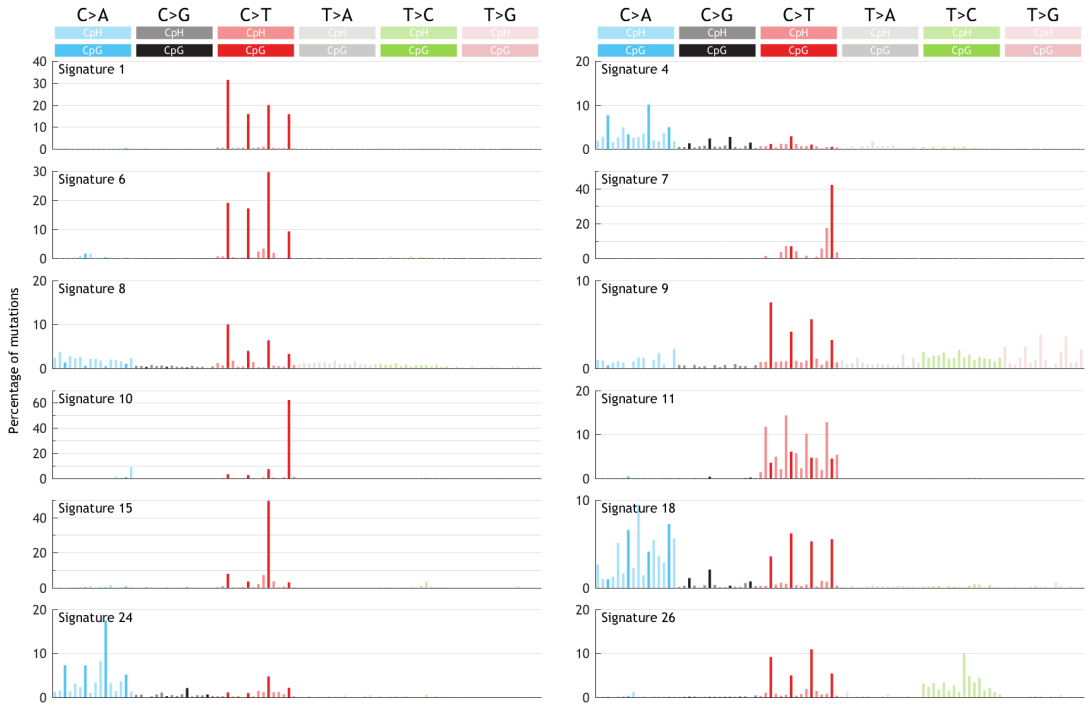

Figure 4.1. Selected mutational signatures. The signatures are normalised for the frequency of trinucleotides in the human genome. Mutations in a NCG context are shown in dark colours.

4.2 Materials and methods

4.2.1 Somatic mutations

Cancer somatic mutations in 3442 whole-genome sequencing samples were obtained from publicly available data sets (Table 9.4, only one sample per patient was included). MSI and *POLE*-MUT samples were combined from previous studies (Haradhvala et al.,

2016; Shlien et al., 2015; Shinbrot et al., 2014). Somatic mutations in autosomes only were taken into account.

4.2.2 DNA modification maps

Maps of cytosine modifications (Table 9.2) were obtained from BS-seq data sets from the data portals of The Cancer Genome Atlas (TCGA), Roadmap Epigenome, Blueprint, and from previously published data in peer-reviewed journals (Wen et al., 2014; Chen et al., 2015; Pidsley et al., 2016; Vandiver et al., 2015) and, where needed, converted to the genome build hg19 using liftOver tool (Hinrichs, 2006). For brain, kidney, and prostate maps, raw reads were processed with bsQC (see 2.2.1) and only sites covered with at least 5 reads were taken into account and only CpGs on autosomes were analysed.

4.2.3 Mutation frequency with respect to modification levels

All cytosines in the CpG context were divided into 10 right-open intervals according to their modification levels (the number of unconverted reads divided by the number of all reads in BS-seq): [0-0.1), [0.1-0.2), ..., [0.9-1]. In each bin, the frequency of mutations was computed and plotted for each sample. A linear regression model was fitted to the data (function `fitlm` in MatLab) and the offset, slope, and last value, and fold-change from first to last value were measured. When comparing CpG sites with low vs. intermediate vs. high modification levels, the thresholds (0.8 and 0.95) were chosen such that the three groups have approximately similar numbers of CpG sites in most tissues.

4.2.4 Direction of replication

Left- and right-replicating domains were taken from (Haradhvala et al., 2016). Each domain (called territory in the original source code and data) is 20 kbp wide and annotated with the direction of replication and with replication timing.

4.2.5 Mutation frequency with respect to the direction of replication

First, transitions between left- and right-replicated domains were computed as in (Haradhvala et al., 2016). These transitions represent regions rich for replication

origins. We computed the CpG>TpG mutation frequency in the 20 kbp domains distant 0 to 1 Mbp from the closest left-/right- transition, with respect to the strand (plus=Watson vs. minus=Crick) of the cytosine of the CpG. Template for the leading strand then corresponds to the plus strand in the left direction and minus strand in the right direction and vice versa for the lagging strand template. Finally, we annotated all cytosines in a CpG context whether they are on the leading or lagging strand, and computed CpG>TpG mutation frequency for the leading and lagging strand separately. Signtest was used for evaluating significance of CpG>TpG mutation frequency difference between the two strands.

4.2.6 Nucleosome maps

A map of nucleosome dyads was downloaded from supplementary materials of Yazdi et al. (2015a), sample GSM1194220. For each CpG in the genome, the closest nucleosome dyad was computed using `bedtools closest` command.

4.2.7 5hmC maps in skin and lung

For skin, MeDIP and hMeDIP measurements from benign skin naevus (Lian et al., 2012) were used: GSM937079 and GSM937084. For each CpG in the genome, the number of reads in the MeDIP and hMeDIP experiments were computed using `bedtools map` command.

For lung, oxBS-seq derived regional estimates of 5hmC from normal lung were downloaded from the supplementary materials of Li et al. (2016), Table S3. For each CpG in the genome, the regional estimates of 5hmC were computed using `bedtools closest` command (CpGs in regions without a significant amount of 5hmC have the value of 5hmC set to zero).

The consensus 5hmC and 5mC maps were computed from the only four existing whole genome TAB-seq measurements (Wen et al., 2014; Pacis et al., 2015; Chen et al., 2015) (1x brain, 2x kidney, 1x blood) and the respective BS-seq measurements (the first (GSM1565940) of the 6 BS-seq blood measurements was used). For each CpG, the average of the mod values (unconverted/coverage in BS-seq) and the average of 5hmC

values (unconverted/coverage in TAB-seq) were computed. Consensus 5hmC$_{rel}$map was computed on the consensus mod and 5hmC maps as min(1, 5hmC/mod).

4.3 Replication-related mutagenesis in modified cytosines

4.3.1 Motivation

Accurate replication and maintenance of the genome is essential for the normal function of cells and to avoid diseases, including cancer. The fidelity of DNA replication depends on the accurate incorporation of bases, on proofreading by the major replicative polymerases Pol ε and Pol δ, and on post-replicative DNA mismatch-repair (MMR) which removes errors from the newly synthesised DNA strand (Rayner et al., 2016). Deficiency in any of these protective mechanisms leads to an increase in the number of mutations. In particular, defects in MMR genes lead to "hypermutability" (10^4–10^5 mutations per Gbp), and mutations in the proofreading domain of Pol ε lead to "ultra-hypermutability", often exceeding 10^5 mutations per Gbp (Shinbrot et al., 2014; Zhao et al., 2014b; Shlien et al., 2015; Nowak et al., 2017). Moreover, defects in Pol ε and Pol δ proofreading cause tumours in mice (Albertson et al., 2009) and germline mutations in POLE and POLD1 (encoding the catalytic subunits of Pol ε and δ, respectively) and genes of the MMR pathway predispose to cancer in humans (Rayner et al., 2016).

Failure to correct the mismatch before the subsequent replication results in a mutation in one daughter cell due to semiconservative DNA replication. Thus replication of, e.g., a T:G mismatch leads to a C:G pair on one strand, but a T:A pair on the other strand, i.e., a C:G>T:A mutation. This mechanism means that the DNA polymerase proofreading and post-replicative MMR (in their canonical, replication-linked functions) are highly unlikely to play a role in repair of 5mC deamination induced mutations, as they operate after parental strands have been separated during replication. Therefore, although the total frequency of mutations due to unrepaired errors introduced during replication increases drastically in polymerase proofreading/MMR deficient samples, it would be expected that the number of CpG>TpG mutations should remain similar in both groups.

4.3.2 *POLE*-MUT and MSI samples exhibit unexpectedly high rates of CpG>TpG mutations

Contrary to the expectation, mutational signatures associated with Pol ε proofreading deficiency (signature 10) and MMR-deficiency (signatures 6, 15, 26) show high frequency of C>T mutations in a CpG context (Fig. 4.1). We therefore investigated possible explanations of this surprising observation. We explored the mutation spectra of 14 tumour samples with a mutation in Pol ε (*POLE*-MUT samples), 19 samples with microsatellite-instability (MSI) deficient in MMR, and 3409 other cancer samples (proficient; PROF). The median overall mutation frequency per base was 1.5×10^{-6} (interquartile range (IQR) $0.6 \times 10^{-6} - 3.5 \times 10^{-6}$) in PROF samples, 36.9×10^{-6} (IQR $18.0 \times 10^{-6} - 47.4 \times 10^{-6}$) in MSI samples, and 267.4×10^{-6} (IQR $99.9 \times 10^{-6} - 300.5 \times 10^{-6}$) in *POLE*-MUT samples (Fig. 4.2). In PROF samples, the median CpG>TpG mutation frequency (i.e., the number of CpG>TpG mutations relative to the number of CpGs in the genome) was 7.4×10^{-6} (IQR $3.7 \times 10^{-6} - 16.8 \times 10^{-6}$), approximately 5-fold higher than the overall mutation frequency (i.e., the number of all mutations relative to the number of all positions in the genome). Notably, the CpG>TpG mutation frequency also increased in MSI and *POLE*-MUT samples, compared to the overall mutation frequency (MSI: median 247.7×10^{-6} per CpG, IQR $162.7 \times 10^{-6} - 367.3 \times 10^{-6}$; *POLE*-MUT: median 1559.8×10^{-6} per CpG, IQR $707.9 \times 10^{-6} - 2574.2 \times 10^{-6}$) (Fig. 4.2, Fig. 4.3). This observation is surprising, since neither MMR nor proofreading during DNA replication by Pol ε are thought to be essential for effective repair of deamination induced T:G mismatches (Bellacosa and Drohat, 2015).

4.3.3 CpG>TpG mutations in POLE-MUT and MSI samples correlate with modification levels

We next used BS-seq derived DNA modification maps from normal tissue of the same organ as each cancer sample to explore whether DNA modifications play a role in the occurrence of CpG>TpG mutations in *POLE*-MUT and MSI samples. These maps represent levels of both the more frequent 5mC as well as the less frequent 5hmC, since BS-seq alone cannot distinguish between these two modifications. The global

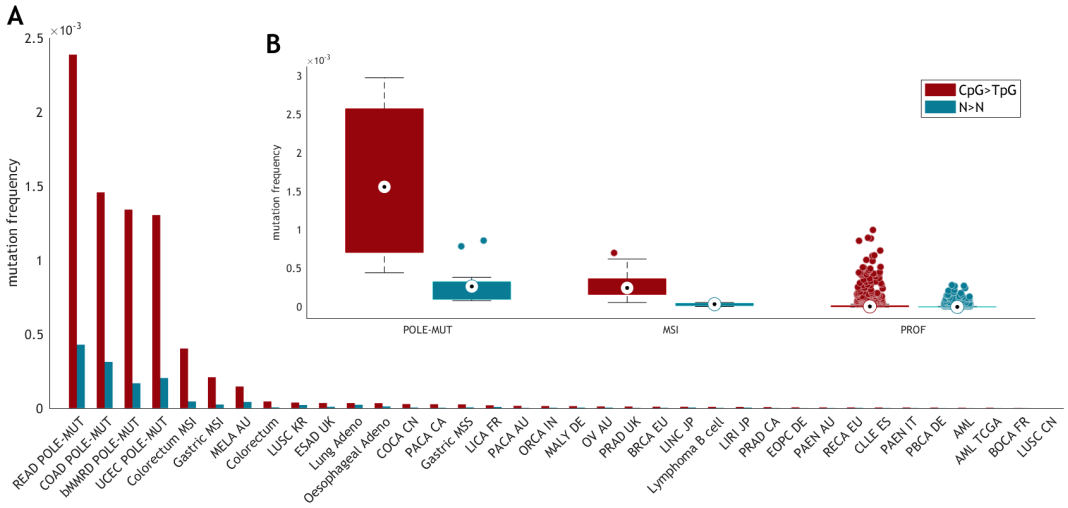

Figure 4.2. Frequency of C to T mutations in a CpG context is unexpectedly high in *POLE*-MUT and MSI samples. **A:** Mean CpG>TpG and N>N (overall) mutation frequency in each cancer type separately. **B:** Distribution of CpG>TpG and N>N mutation frequency in *POLE*-MUT, MSI, and PROF (other) samples. The white circle with the black dot inside denotes the median.

levels of 5hmC range between 1.6–24.8 % of mod (based on HPLC measurements from 8 tissues in Fig. 3.30) and are below 13 % in all the measured tissues apart from brain. Using 5mC-specific maps would be undoubtedly superior; however since such maps are currently not available, the BS-seq measurements should represent reasonable

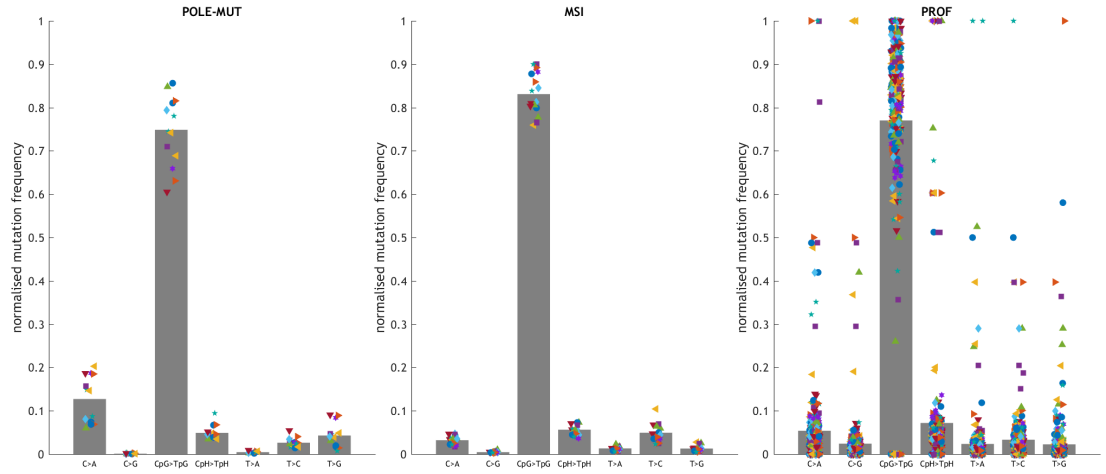

Figure 4.3. Frequency of C to T mutations in a CpG context is unexpectedly high in *POLE*-MUT and MSI samples. Frequency of individual types of mutations in *POLE*-MUT, MSI, and tissue-matched PROF samples, normalised by the total sum in each sample. The bars denote mean over samples and individual samples are shown as markers in different shapes and colours.

approximation. Moreover, instead of methylation levels, we use the term modification levels, referring to 5mC and 5hmC levels together.

In all *POLE*-MUT and MSI samples, the CpG>TpG mutation frequency was positively correlated with modification levels (Fig. 4.4A–E). We fitted a linear model through this correlation for each sample in the *POLE*-MUT, MSI, and PROF samples (Fig. 4.5). The slope of the correlation was significantly higher in *POLE*-MUT than in MSI, and in MSI than in tissue-matched PROF samples (Fig. 4.4F), showing that the increased mutability is not driven by the CpG sequence context, but also by the presence of modified cytosines. Also the offset was significantly higher in *POLE*-MUT and MSI than in PROF samples (Fig. 4.6B), suggesting that there is a general increase of mutability in all cytosines (including unmodified cytosines) in *POLE*-MUT and MSI. However, the fold-change from unmodified to modified cytosines was also significantly higher in *POLE*-MUT and MSI than in PROF samples (Fig. 4.6). These results support the notion that the presence of cytosine modifications is linked to the strong increase of the frequency of C>T mutations in CpG sites in *POLE*-MUT and MSI samples.

Brain is the only of the four tissues, for which there is a single-base resolution map of 5hmC. We therefore used this map of normal human brain (same as in chapter 3.2 and (Tomkova et al., 2016)) to measure mutation frequency separately for 5mC and 5hmC in brain. In the two *POLE*-MUT brain cancers, we observed a moderate gradual decrease of mutation frequency in hydroxymethylated CpGs compared to methylated CpGs (Fig. 4.7), in line with our previous report from PROF brain cancers (3.2 and (Tomkova et al., 2016)), albeit with smaller effect size.

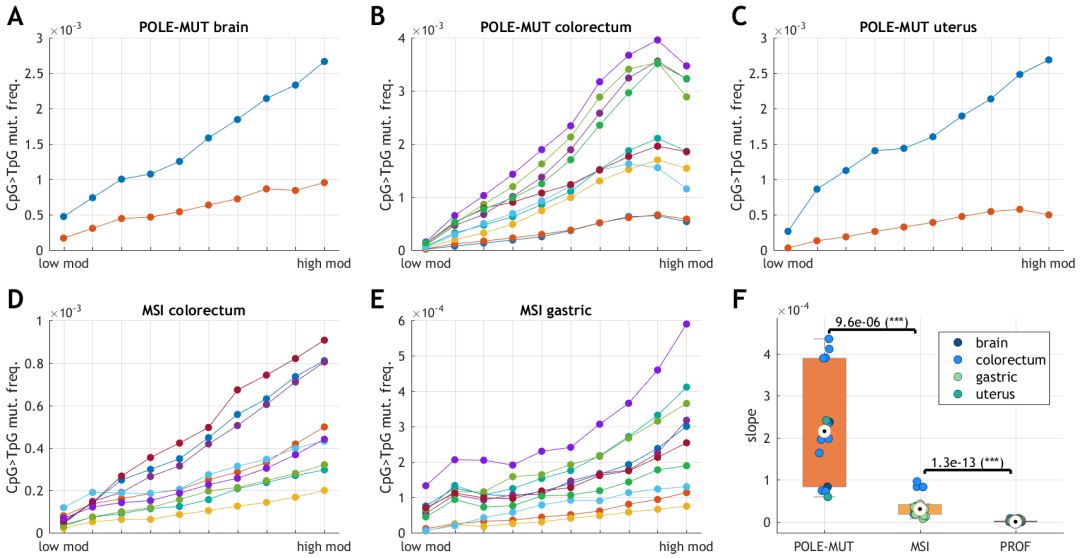

Figure 4.4. Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI samples correlates with DNA modification levels. A-E: Fraction of mutated CpG sites as a function of modification levels. The x-axis represents CpG sites grouped into 10 bins by their modification levels (0-0.1, …, 0.9-1.0). The y-axis represents C>T mutation frequency in each bin. Individual samples are plotted in different colours. **F:** Distribution of the slope of the linear relationship between DNA modification levels and CpG>TpG mutation frequency in four tissues (brain, colorectum, gastric, and uterus). The Wilcoxon ranksum test was used to evaluate differences between the groups (*POLE*-MUT, MSI, and PROF) of samples. See the distribution of offsets in 4.6.

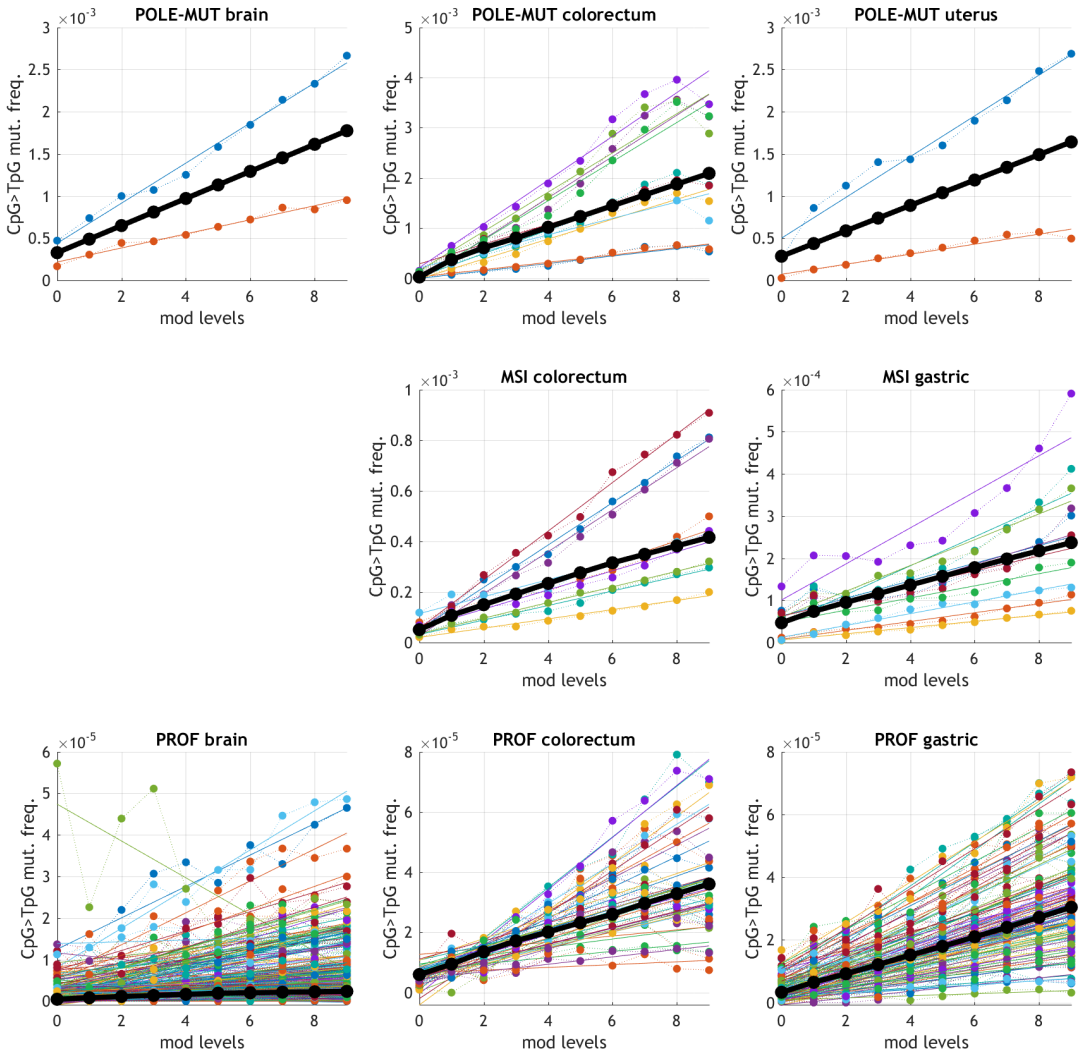

Figure 4.5. Frequency of C to T mutations in a CpG context in *POLE*-MUT, MSI and PROF samples correlates with DNA modification levels: linear models. C>T mutation frequency in CpG sites binned by their tissue-matched modification levels (0-0.1, …, 0.9-1.0). A linear model is fitted on data in each sample. Individual samples are plotted in different colours and the median is shown in black.

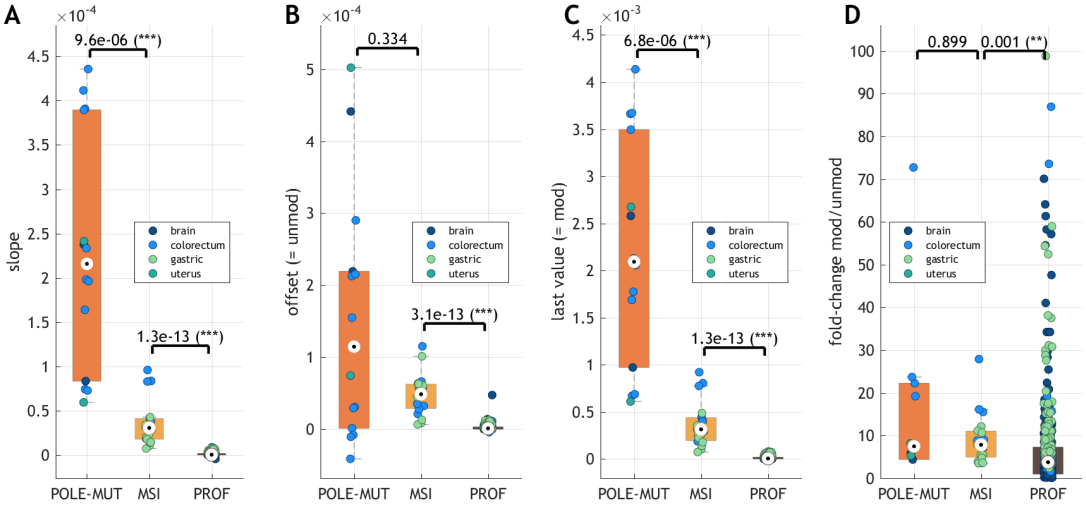

Figure 4.6. Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI samples correlates with DNA modification levels: comparison of linear models. In each sample, a linear model was fitted on the data, representing CpG>TpG mutation frequency in different bins of cytosine modification levels. The distribution of their parameters is compared: slope (A), offset, i.e., the value in unmodified cytosines (B), the last values, i.e., the value in fully modified cytosines (C), the fold-change from unmodified to fully modified cytosines (D) in MSI, POLE, and PROF samples in four tissues (brain, colorectum, gastric, and uterus). The Wilcoxon ranksum test was used to evaluate differences between the groups of samples.

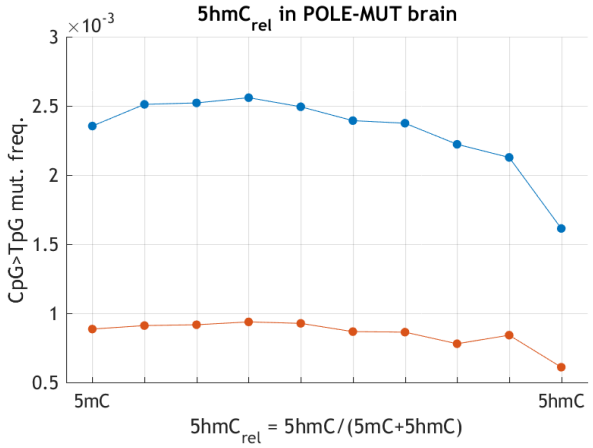

Figure 4.7. Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI samples negatively correlates with 5hmC$_{rel}$. C>T mutation frequency in CpG sites binned by their tissue-matched 5hmC$_{rel}$ levels (0-0.1, …, 0.9-1.0), with mostly methylated sites in the first bin and mostly hydroxymethylated sites in the last bin (only CpGs with mod>0.1). Individual samples are plotted in different colours.

4.3.4 Two independent observations suggest that the mechanism of CpG>TpG mutagenesis in *POLE*-MUT and MSI samples is linked to replication

Due to the semiconservative nature of DNA replication, it is unlikely that Pol ε or MMR, through their canonical, replication-linked activity, are used for the repair of deamination-induced T:G mismatches that happened before replication. However, it is possible that their non-canonical, replication unrelated, activity is involved in the repair of deamination induced mismatches. Conversely, the CpG>TpG mutations could be replication related, but independent of spontaneous deamination of 5mC. We therefore performed two independent analyses to distinguish between the two possibilities: a potential replication-unrelated repair of spontaneous deamination, and a potential replication-related source of CpG>TpG mutations.

First, we estimated the number of years needed to reach the observed frequency of C>T mutations in modified CpGs observed in *POLE*-MUT and MSI samples, assuming zero-efficient repair of these mutations. This calculation was motivated by the theoretical possibility that MMR and *POLE* are in fact essential in the repair of deamination-induced T:G mismatches.

We combined the spontaneous deamination rate of 5mC in double-stranded DNA ($5.8{\cdot}10^{-13}s^{-1}$) reported by Shen et al. (1994), the number of seconds in a year (31 556 736), the observed frequency of GCG>GTG mutations[1] (i.e., GmCG>T/GmCG; for mC with a modification level of at least 0.9) in MSI ($5.133 \cdot 10^{-4}$) and *POLE*-MUT ($1.785 \cdot 10^{-3}$) samples. The number of years needed to reach the observed mutation frequencies can be then computed as:

$$\text{MSI: } \frac{5.133 \cdot 10^{-4}}{5.8 \cdot 10^{-13}\text{s}^{-1} \cdot 31556736\,\text{s years}^{-1}} = 28.05 \text{ years} \tag{4.1}$$

$$\textit{POLE}\text{-MUT: } \frac{1.785 \cdot 10^{-3}}{5.8 \cdot 10^{-13}\text{s}^{-1} \cdot 31556736\,\text{s years}^{-1}} = 97.53 \text{ years} \tag{4.2}$$

[1] We focused on the GCG context, as it showed most consistent mutational properties across the samples, as explained later.

The results show that spontaneous deamination alone is highly unlikely to account for the mutation burden observed in *POLE*-MUT samples, because the age range of patients was 3 years to 81 years. The 28 years needed to reach the mutation frequencies observed in MSI samples does not completely rule out the possibility of spontaneous deamination as a sole source of CpG>TpG mutations in MSI. However, the number of 28 years is based on an assumption of zero-efficient repair, i.e., all spontaneously deaminated 5mC being fixated into a mutation. It is highly unlikely that the ability to repair T:G mismatches is completely disrupted in the MSI samples[2]. With increasing efficiency of repair of the deamination events, the number of years needed to reach the observed mutations grows dramatically, for instance reaching 93.5 years for 70 % efficiency (Fig. 4.8). The results in summary suggest that spontaneous deamination is not the sole source of the observed CpG>TpG mutations in these cohorts.

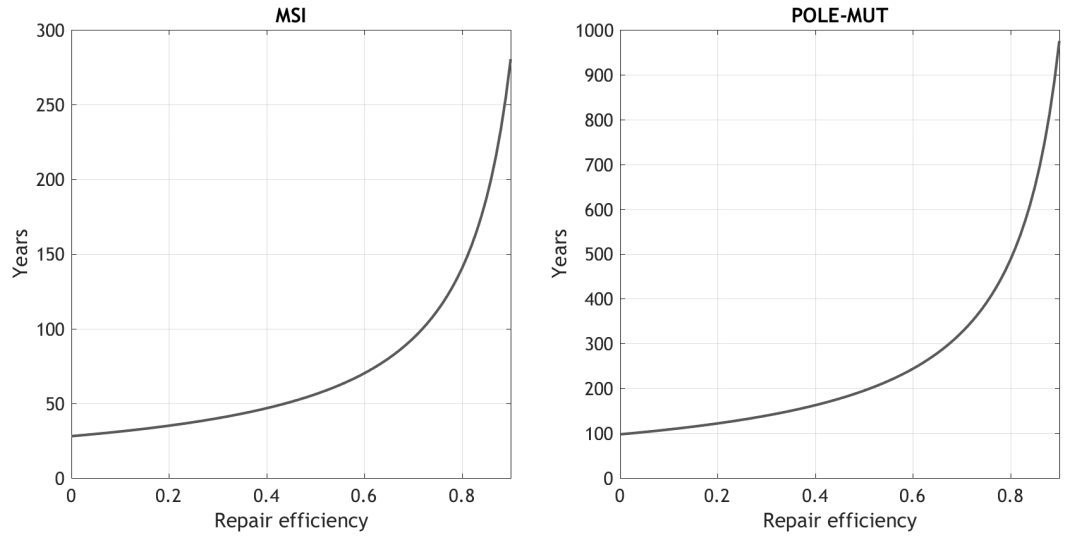

Figure 4.8. Estimated duration of mutation accumulation as a function of deamination repair efficiency. The x-axis represents repair efficiency of T:G mismatches resulting from 5mC deamination. The y-axis represents the number of years needed to reach the mutation frequencies observed in MSI and *POLE*-MUT samples. This relationship is computed as $y =$ mutation frequency$/($deamination rate $\cdot\ (1 - x) \cdot$ seconds in year$)$.

[2]Only seven of the 19 MSI samples contained a variant in *TDG* or *MBD4* of at least a moderate consequence (based on Variant Effect Predictor (McLaren et al., 2016) and the GDC data portal).

Second, we explored whether the CpG>TpG mutagenicity in *POLE*-MUT and MSI samples shows any replication-linked characteristics. As summarised in the Introduction 1.1.3, DNA replication in eukaryotic cells is initiated around replication origins (ORI) from where it proceeds in both directions, synthesizing the leading strand continuously and the lagging strand discontinuously. As Pol ε is the main leading strand DNA polymerase (Stillman, 2008; Georgescu et al., 2015), mutations in *POLE*-MUT samples are distributed asymmetrically on the leading and lagging strands (Shinbrot et al., 2014; Haradhvala et al., 2016). MSI samples also display replication strand bias across several types of mutations (Haradhvala et al., 2016), presumably because MMR is involved in balancing the differences in fidelity of the leading and lagging polymerases (Lujan et al., 2012). In order to determine whether CpG>TpG mutations in *POLE*-MUT and MSI samples happened during or before replication, we computed the frequency of CpG>TpG mutations on the plus (Watson) and minus (Crick) strand around transitions between left- and right-replicating regions, as defined in (Haradhvala et al., 2016). The transitions correspond to regions enriched for replication origins.

In the *POLE*-MUT and MSI samples, we observed a strong enrichment of CpG>TpG mutations on the leading strand template (plus strand in the left direction, minus strand in the right direction) (Fig. 4.9). Moreover, the strand asymmetry was at least as strong or stronger in highly modified CpGs (top tertile) than in lowly modified CpGs (bottom tertile) (Fig. 4.9C–D). This effect was furthermore observed across cancer types and across modification levels (Fig. 4.10). It thus appears that DNA repair deficient cells accumulate more CpG>TpG mutations in cytosines that were modified on the template for the leading strand, suggesting that they are related to replication.

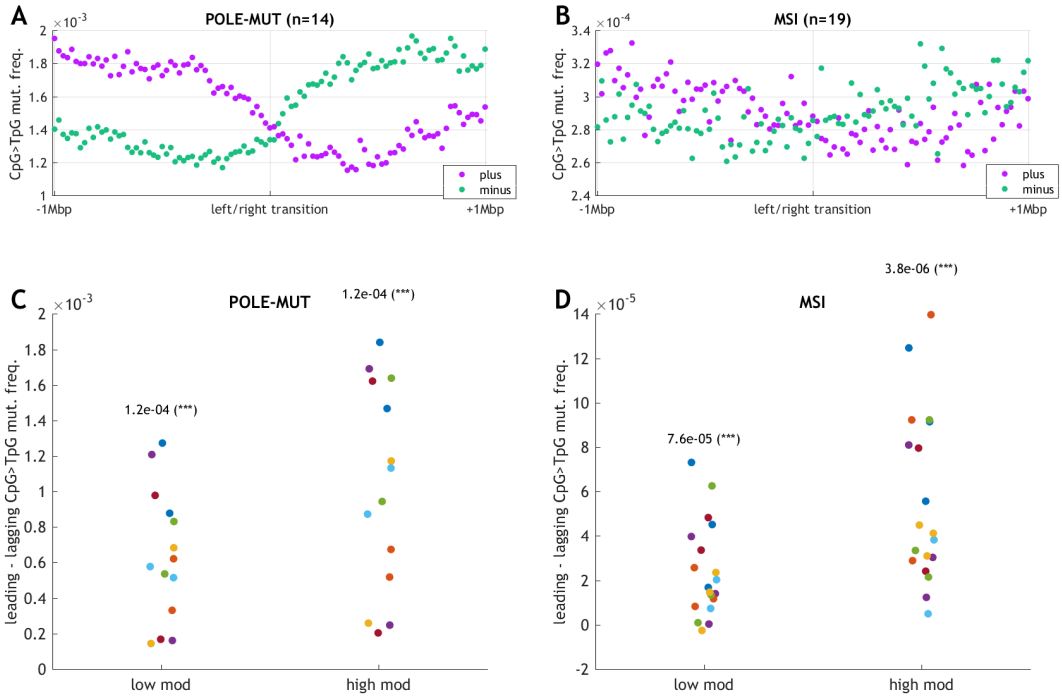

Figure 4.9. Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI samples is higher on the leading strand than on the lagging strand, especially in modified CpG sites. A–B: Mean CpG>TpG mutation frequency on the plus (Watson) and minus (Crick) strand around transitions between left- and right-replicating regions. The transitions correspond to regions enriched for replication origins. The leading strand template corresponds to the plus strand in the left direction and the minus strand in the right direction, whereas the lagging strand template corresponds to the minus strand in the left direction and the plus strand in the right direction. **C–D:** Difference in the leading and lagging CpG>TpG mutation frequency in each sample (signtest was used for evaluating significance between leading and lagging strand).

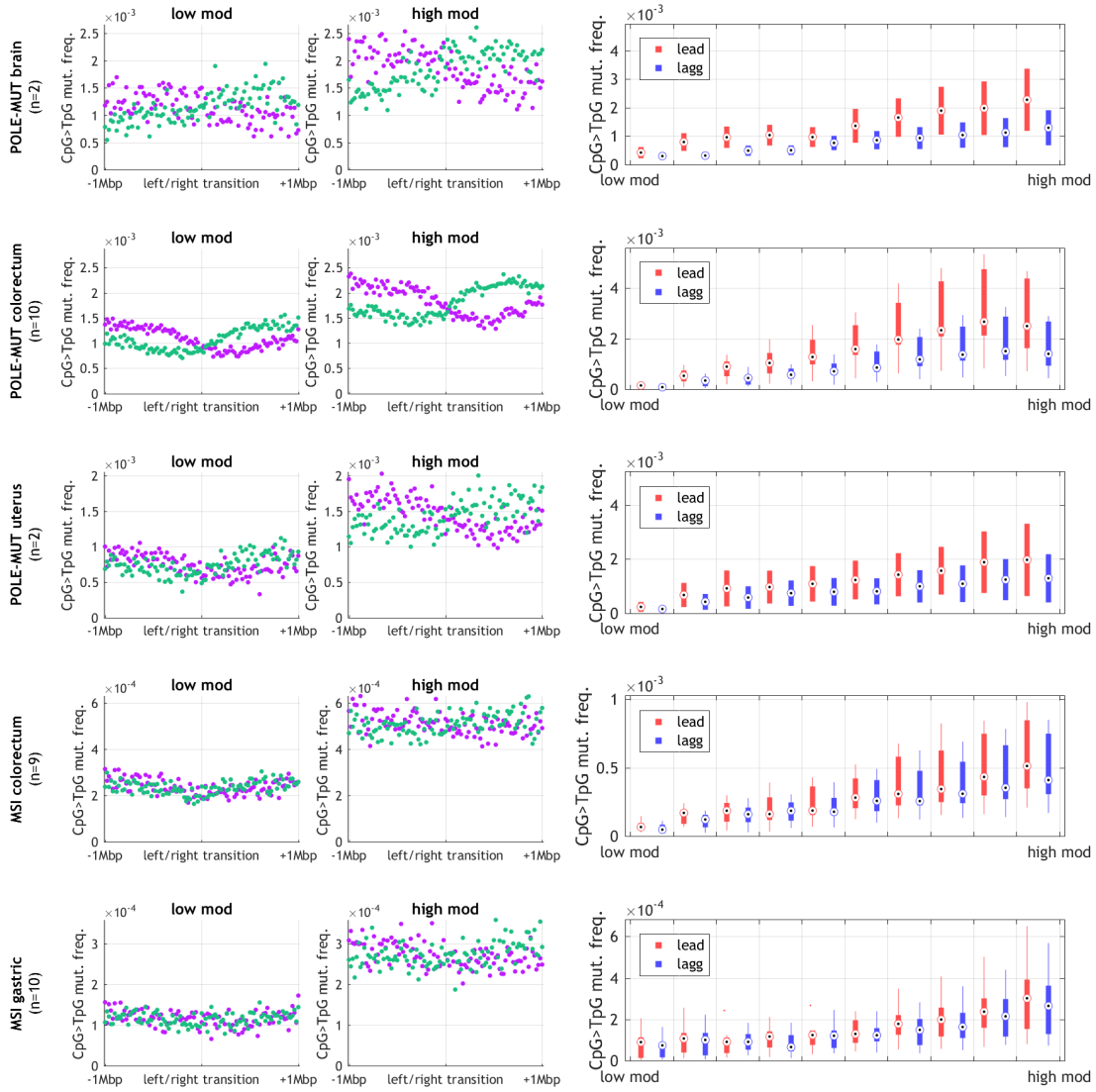

Figure 4.10. Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI samples is higher on the leading strand than on the lagging strand, especially in modified CpG sites. Left column: Mean CpG>TpG mutation frequency on the plus (Watson) and minus (Crick) strand around transitions between left- and right-replicating regions. The transitions correspond to regions enriched for replication origins. Comparison of CpG sites with low modification levels ($\leq$0.8) and high modification levels (>0.95) in different tissue types (rows) is shown. Right column: C>T mutation frequency in CpG sites in the leading and lagging strand binned by their tissue-matched modification levels (0-0.1, …, 0.9-1.0).

4.3.5 The effect of different variants, age, and sequence context

The link between C>T mutagenicity in modified CpG sites and replication could have two possible underlying mechanisms. It could either be a unique feature of *POLE*-MUT and MSI samples, actively causing the mutagenicity. Or it could be present also in proficient samples, but suppressed by the combination of Pol ε proofreading and MMR.

To explore the first option, we tested the observed POLE and MMR mutations for signs of a "gain of function" mutation. A range of 9 different variants in the proofreading domain of POLE were present in the 14 *POLE*-MUT samples, all of them showing an increase of CpG>TpG mutations in modified cytosine (Fig. 4.11A).

The *POLE*-MUT samples seem to separated into three groups according to their slope of CpG>TpG correlation with modification levels. We explored this observation, but did not find any characteristics that would explain the separation. The most mutated group with four samples contains two variants (P286R and V411L), which are known to be the most common deleterious POLE variants (Rayner et al., 2016). However, the same variants are also present in the least mutated samples in the middle mutated group in this cohort. Also the age of the patients at the time of diagnosis did not underlie the groupings. In summary, the positive correlation of CpG>TpG mutagenicity with modification levels seems to be independent of the type of POLE mutation, cancer type or age at diagnosis, and is present in both *POLE*-MUT and MSI samples (Fig 4.11A). A gain-of-function mutation therefore seems unlikely, as the altered function is usually mediated by an altered protein structure due to a very specific change in the sequence of amino acids.

Interestingly, the frequency of C>T mutations was not only affected by the 3' sequence context, but also the 5' base of cytosine. We noticed that while C>T mutations in a TCG context (TCG>TTG) dominate in colorectal *POLE*-MUT samples, both MSI and all *POLE*-MUT exhibited high levels of C>T mutations in a GCG context (GCG>GTG) (Fig. 4.11B, 4.12). GCG>GTG mutations also showed particularly strong strand asymmetry and correlation with modification levels in all MSI and *POLE*-MUT samples (Fig. 4.11C–D, 4.13).

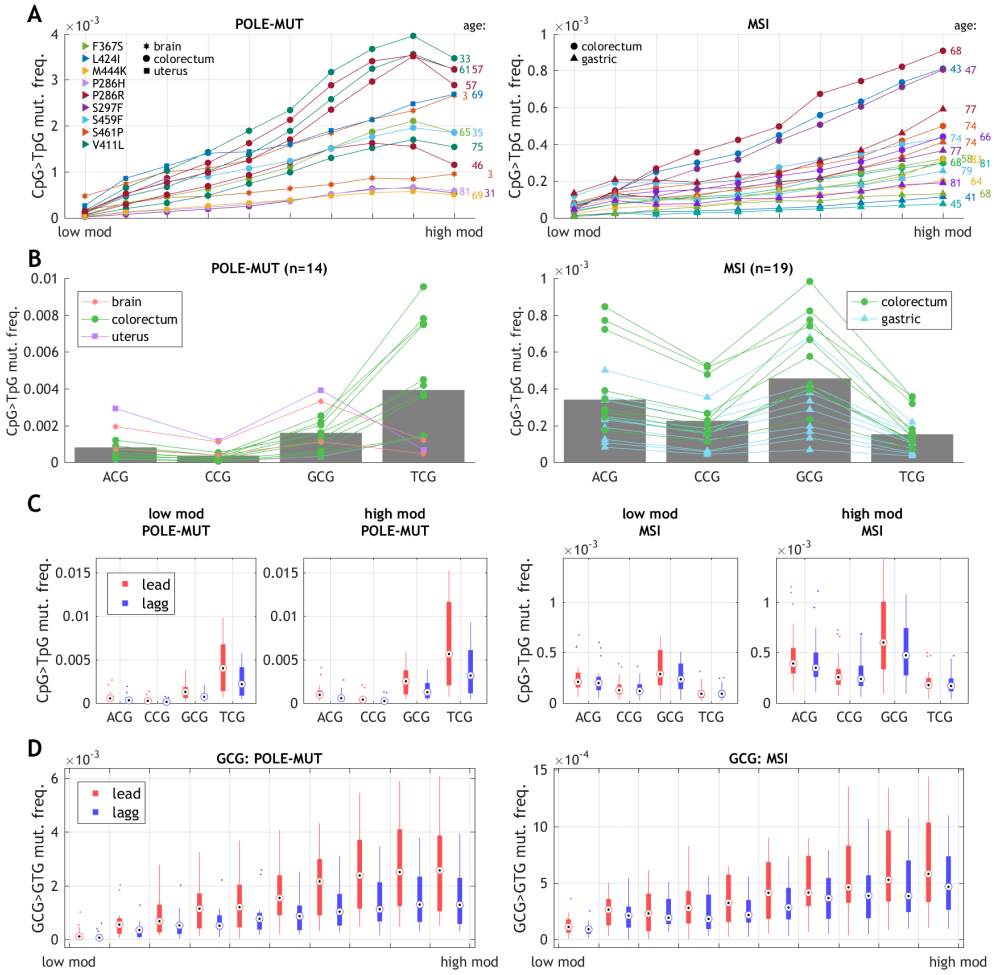

Figure 4.11. Increase of C to T mutations in modified cytosine on the leading strand is most consistent in a GCG sequence context in *POLE*-MUT and MSI samples. **A:** C>T mutation frequency in CpG context binned by the tissue-matched modification levels (0-0.1, …, 0.9-1.0). In *POLE*-MUT samples, the colour represents different variants of the POLE mutation. In both *POLE*-MUT and MSI samples, the marker represents different tissues. The age at diagnosis is shown next to the last value of the sample. **B:** CpG>TpG mutation frequency stratified by the 5' flanking sequence context. The bars denote mean over samples and individual samples are shown as markers with shape and colour distinguishing the tissue type. **C:** C>T mutation frequency in CpG sites in the leading and lagging strands, in low mod ($\leq$0.8) vs high mod (>0.95), and stratified by the 5' sequence context: ACG, CCG, GCG, and TCG. **D:** C>T mutation frequency in GCG context in leading and lagging strand binned by the tissue-matched modification levels (0-0.1, …, 0.9-1.0).

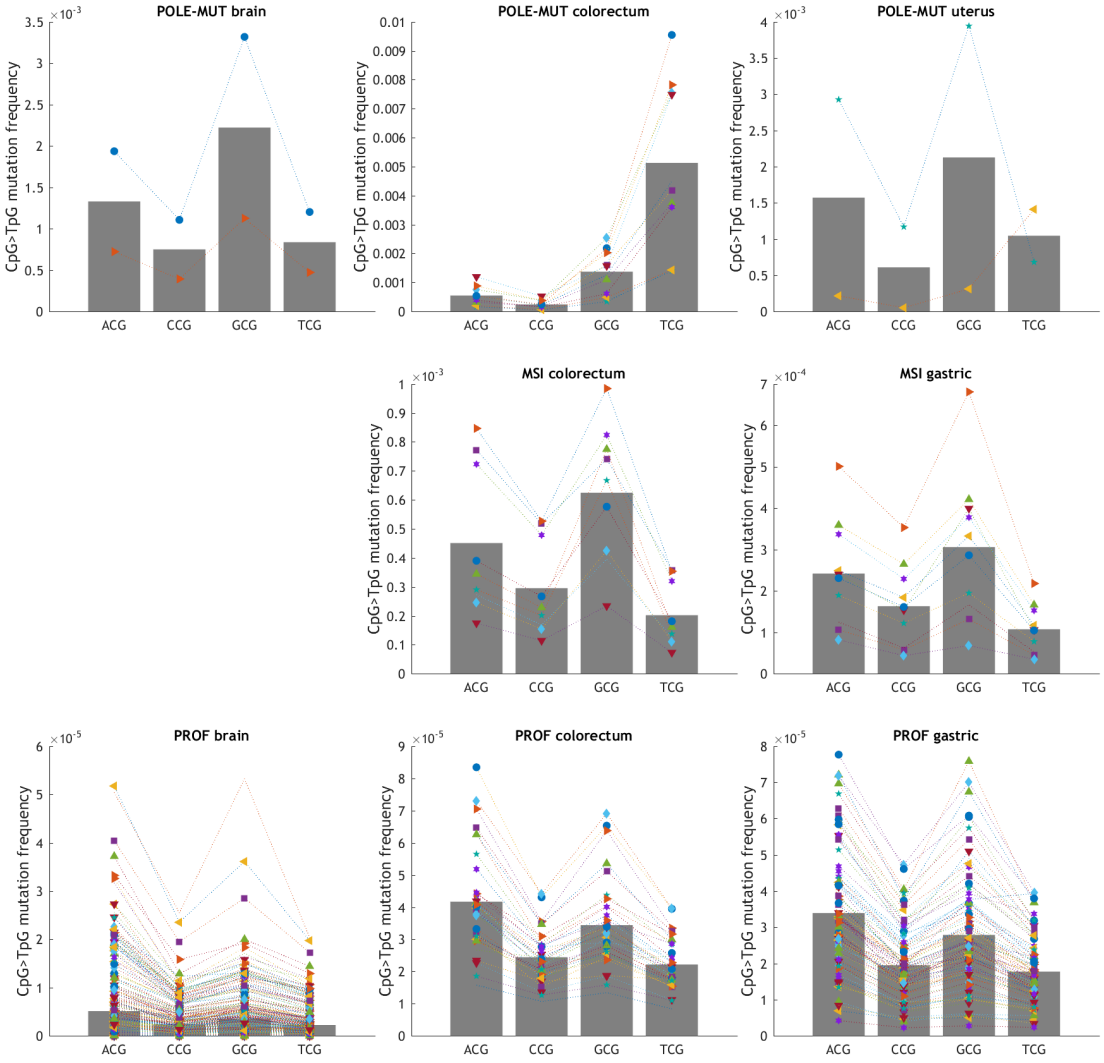

Figure 4.12. CpG>TpG mutation frequency in different sequence contexts. CpG>TpG mutation frequency stratified by the 5' flanking sequence context and tissue type. The bars denote mean over samples and individual samples are shown as markers. PROF samples are shown for a reference.

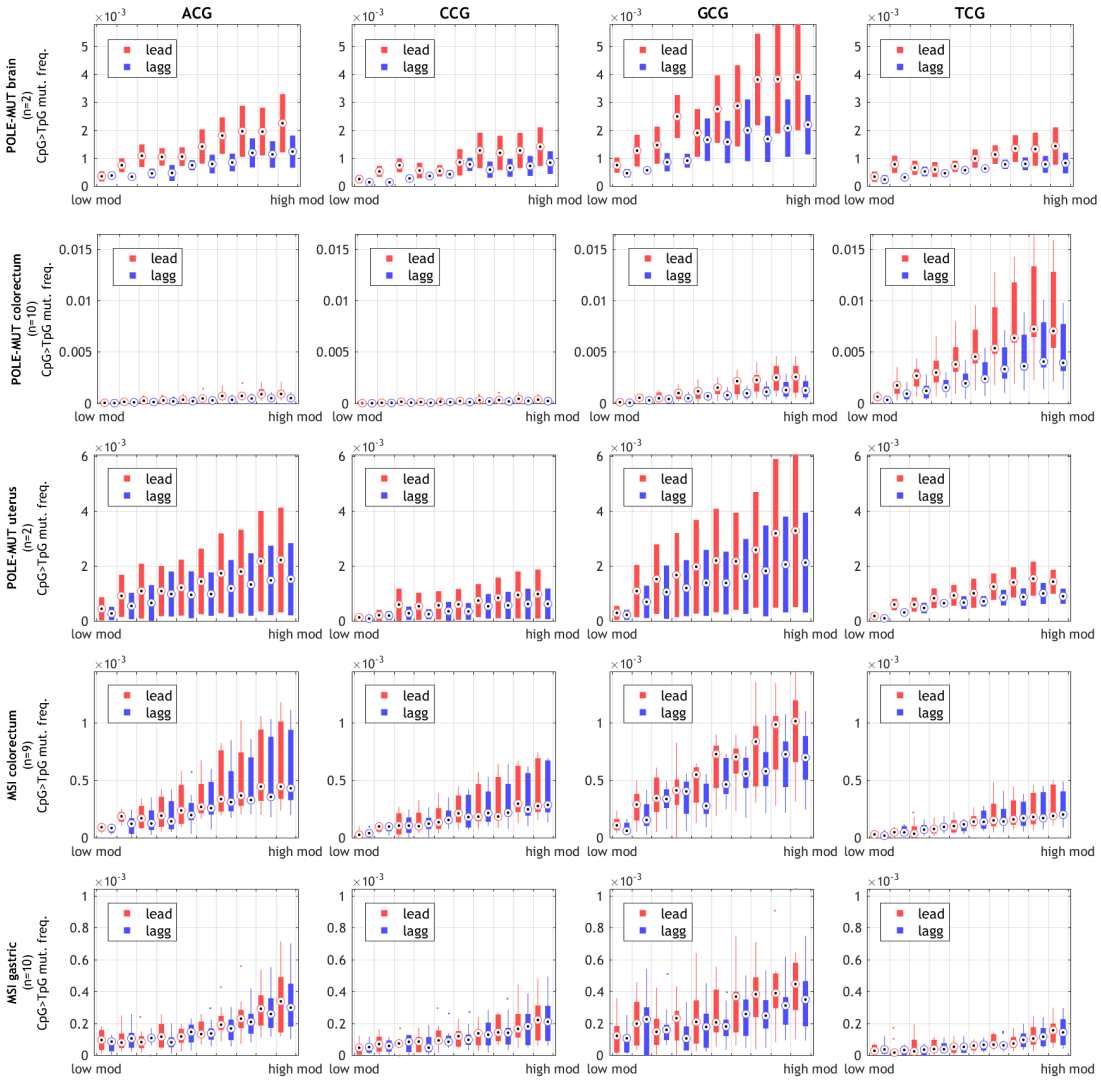

Figure 4.13. Increase of C to T mutations in modified cytosine on the leading strand is most consistent in a GCG sequence context in *POLE*-MUT and MSI samples. C>T mutation frequency in CpG sites in leading and lagging strand binned by their tissue-matched modification levels (0-0.1, 0.1-0.2, …, 0.9-1.0) and sequence context: ACG (first column), CCG (second column), GCG (third column), and TCG (fourth column).

4.3.6 A model of replication-linked mutagenicity in 5mC

Our observations could be explained by a model of CpG>TpG mutagenesis (Fig. 4.14), in which 5mC is occasionally incorrectly paired with adenine by Pol ε during replication of the leading strand. This decreased fidelity could potentially be enhanced by the structural similarity of 5mC and thymine. If such mismatches were not detected by the polymerase proofreading machinery, MMR, or BER, they would result in CpG>TpG mutations in the leading strand template.

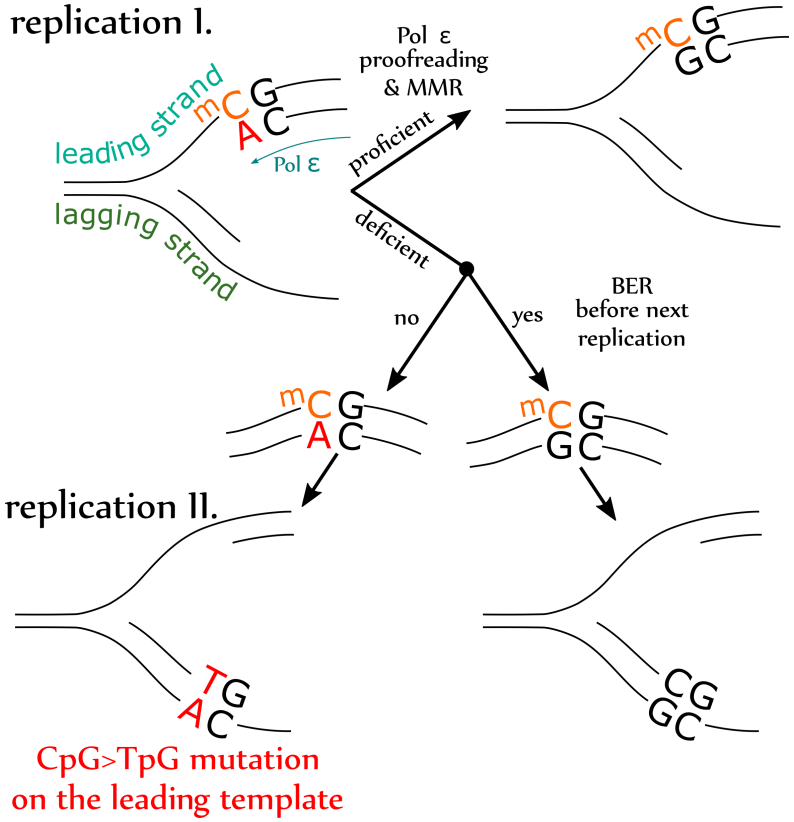

Figure 4.14. A hypothesised model of CpG>TpG mutagenesis in methylated cytosine due to replication. In the model, the leading strand polymerase Pol ε has an increased error-rate of incorporating adenine opposite 5mC. In cells with proficient Pol ε proofreading and MMR, most of these errors will be detected repaired. However, if they escape or in case of MMR/Pol ε proofreading deficiency, and if not detected and repaired by base excision repair, they will be fixated into a CpG>TpG mutation on the leading strand template in the next DNA replication.

Under this model of decreased fidelity of wild-type Pol ε in replication of 5mC, we would expect that such errors could sometimes escape the polymerase proofreading and MMR even in POLE-WT and MMR proficient samples, resulting in a mild strand asymmetry of CpG>TpG mutations. To test this, we grouped PROF samples by tissue, and in each tissue measured the percentage of samples with a higher CpG>TpG mutation frequency on the leading than the lagging strand, while also distinguishing between all four sequence contexts. The majority of samples exhibited leading strand bias for GCG>GTG mutations in 13 out of 16 tissue types in lowly and middle modified CpGs (Fig. 4.15). This effect was even more ubiquitous (16 out of 16 tissues) when restricting the analysis to highly modified CpGs only (Fig. 4.16), supporting the hypothesis that CpG>TpG mutations can also be caused by errors during the replication of methylated cytosine by Pol ε. The other sequence contexts did not show a consistent replication strand asymmetry, but ACG>ATG and CCG>CTG mutations were slightly enriched on the lagging strand in the highly modified CpGs (p-value < 0.05). We discuss possible causes of this observation in the discussion.

4.3.7 Discussion of replication-related mutagenesis in modified cytosines

The increased rate of C>T mutations at CpG dinucleotides across tissue types has been thought to primarily stem from spontaneous deamination of methylated cytosine. The fact that *POLE*-MUT and MSI samples exhibit high CpG>TpG mutation frequency is therefore surprising, since neither MMR nor proofreading by Pol ε are thought to be required for the repair of deamination damage.

A similar increase of CpG>TpG mutations in MSI and *POLE*-MUT colorectal cancer samples has also been observed in another study that was published very recently (Poulos et al., 2017). In this study, the CpG>TpG mutations also correlated with methylation levels and the slope of this correlation was higher in late-replicating regions in MSS and *POLE*-MUT samples, but not MSI samples. Such observation is in line with the expected enhanced activity of MMR in the early-replicated regions (Supek and Lehner, 2015), but not providing other mechanistic insight, apart from supporting

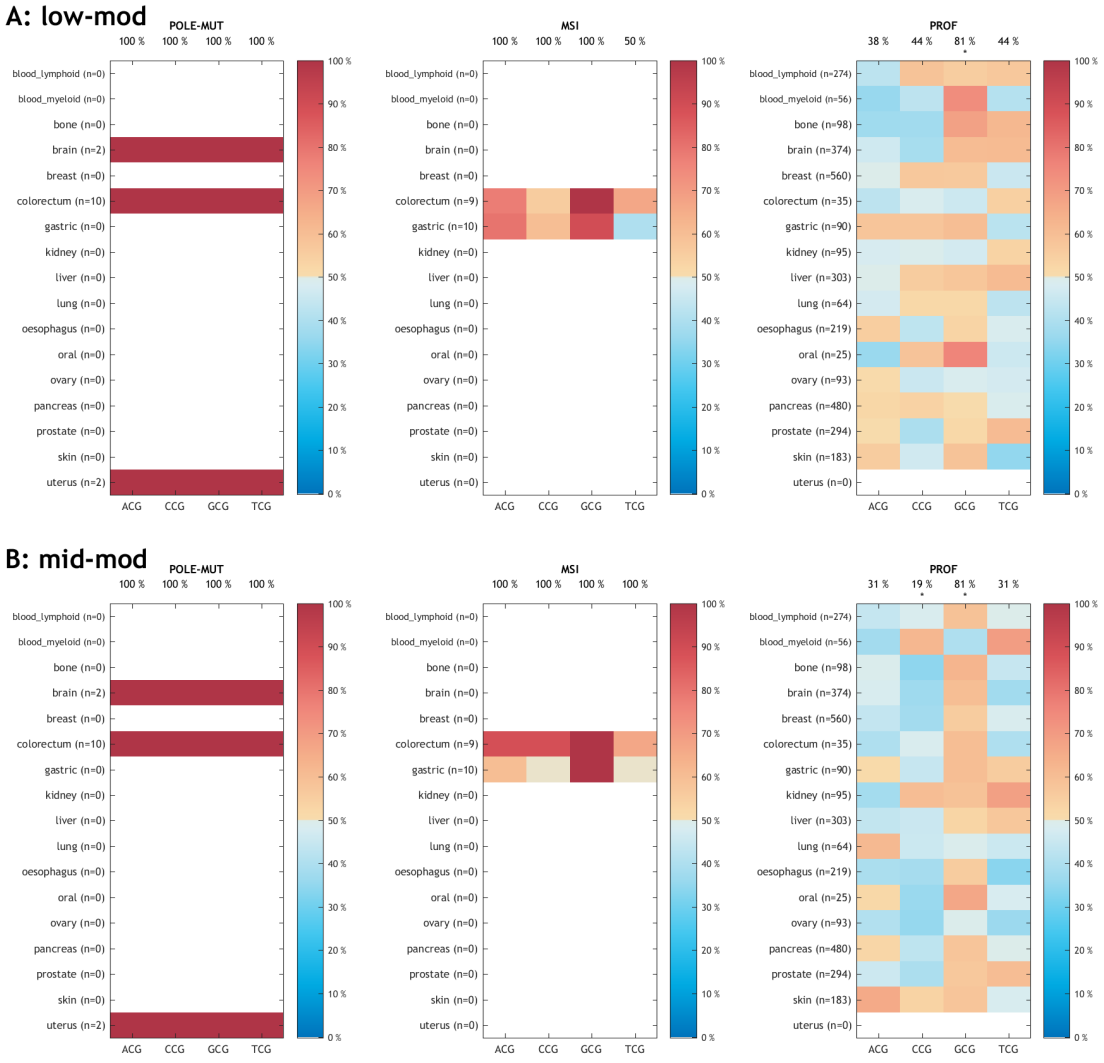

Figure 4.15. GCG>GTG mutations are more frequent on the leading strand than on the lagging strand, even in Pol ε and MMR proficient samples. Percentage of samples with higher C>T mutation frequency on the leading strand than on the lagging strand for CpG sites with low ($\leq$0.8) modification levels (A), and for sites with intermediate (between 0.8 and 0.95) modification levels (B), using tissue-matched modification maps. White colour denotes no data, blue colour denotes more frequent lagging strand bias, and red denotes more frequent leading strand bias. Asterisks represent significance of the bias (signtest; ***P < 0.001; **P < 0.01; *P < 0.05).

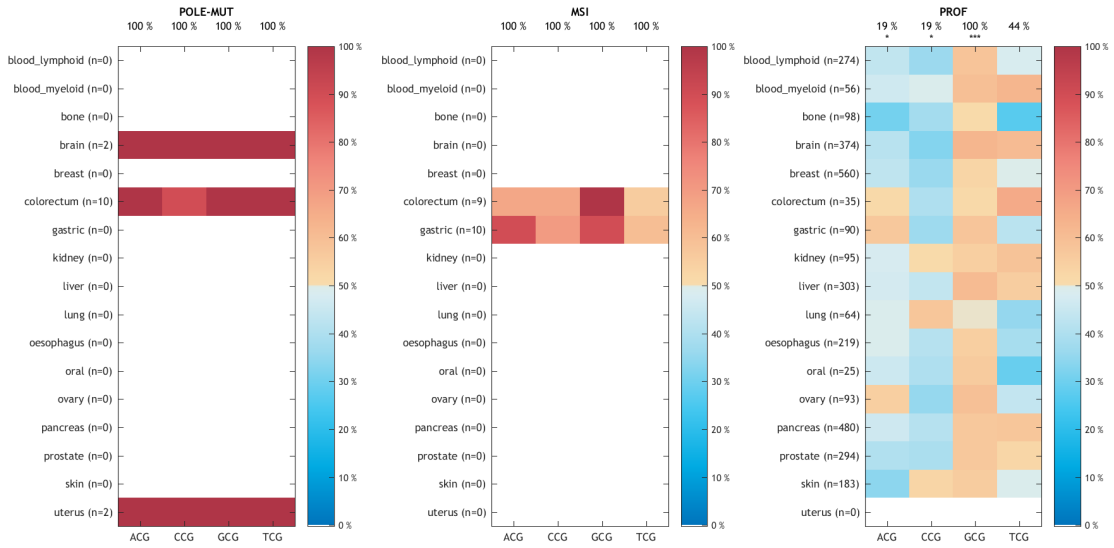

Figure 4.16. GCG>GTG mutations are more frequent on the leading strand than on the lagging strand, even in Pol ε and MMR proficient samples. The heatmap shows the percentage of samples with higher C>T mutation frequency on the leading strand than on the lagging strand (only C>T mutations in highly modified (>0.95) CpG sites, using tissue-matched modification maps): white colour denotes no data, blue colour denotes more frequent lagging bias, and red denotes more frequent leading bias. Asterisks represent significance of the bias (signtest; ***P < 0.001; **P < 0.01; *P < 0.05).

a link to replication. Compared to our genome-wide analysis of replication strand asymmetry in *POLE*-MUT, MSI, and PROF samples, Poulos et al. (2017) focused on ca. 600 kbp region around two known ORI loci and also observed strand-specific mutations in the *POLE*-MUT samples, in correspondence with our results.

Three theoretical models could explain our observations. In the first model, MMR and Pol ε —through a non-canonical, replication-unrelated mechanism— are in fact essential for the repair of T:G mismatches created by spontaneous deamination of 5mC. For MMR, this is the model proposed in the recent study by Poulos et al. (2017). However, the observed number of CpG>TpG mutations in MSI and *POLE*-MUT samples are difficult to reconcile with the known deamination kinetics of methylated cytosine in double-stranded DNA, even under the unrealistic assumption that no repair mechanisms at all are active in these samples. At 5.8×10^{-13} mutations per 5mC per second (Shen et al., 1994), it would take 28 years to reach the observed C>T mutation frequency in modified GCG sites of MSI samples, and 98 years for *POLE*-MUT samples. These time-scales are unlikely to represent the real time between the acquisition of the MMR

or Pol ε mutation and the collection of the sample. Moreover, the observed enrichment of CpG>TpG mutations on the leading strand also does not support this first model.

The second possible explanation is that the Pol ε and MMR mutations are gain of function mutations, causing a mutator phenotype that actively increases CpG>TpG mutagenicity during replication. This mechanism has been suggested by Poulos et al. (2017) for the *POLE*-MUT samples and by Kane and Shcherbakova (2014) in *S. cerevisiae*, where an analogue of the human P286R variant (but not other variants) in the yeast Pol ε produced a strong mutator phenotype, increasing the mutation rate beyond that of the proofreading-null allele. However, we observed a marked increase of C>T mutation frequency in modified CpG sites in a wide range of Pol ε variants (Fig. 4.11A). Furthermore, a strong correlation of GCG>GTG mutations with DNA modification levels was observed across *POLE*-MUT and MSI samples from multiple cancer types. It therefore seems unlikely that multiple different Pol ε and MMR mutations all result in the same mutator phenotype.

The third model posits that wildtype Pol ε has a slightly decreased fidelity when encountering 5mC, particularly in a GCG context, on the template strand and incorrectly pairs it with A, leading to 5mC:A mismatches (Fig. 4.14). This could potentially be a consequence of the high structural similarity between 5mC and T, both of which present a methyl group at the same position of pyrimidine ring. If the resulting 5mC:A mismatches were not repaired before the next round of replication, for example because of a lack of mismatch repair in MSI tumours, one would expect an enrichment of GCG>GTG mutations on the leading strand, as we observe in our data. Similarly, a lack of proofreading by Pol ε itself might overwhelm the capacity of downstream repair pathways and thus, too, lead to an increased CpG>TpG mutation rate. The fact that we also detected a leading strand bias for GCG>GTG mutations in a majority of Pol ε and MMR proficient tumours hints at the possibility that the mechanism described above does contribute to the overall CpG>TpG mutation burden. This model is also consistent with observations from other data sets. While Pol ε-deficient samples contain a large amount of CpG>TpG mutations, samples deficient in Pol δ are also highly mutated, but CpG>TpG mutations form only a small percentage of the

mutation burden (Shlien et al., 2015). This observation supports the notion that the CpG>TpG mutagenesis is linked to the leading strand synthesis. Moreover, mutation calls from single normal neurons (which are largely non-dividing cells) show relatively low percentage of CpG>TpG mutations[3] and similar mutation frequencies in 15 years-old and 42 years-old individuals (Lodato et al., 2015), in line with the possibility of a replication-linked component of CpG>TpG mutagenesis. Finally, this model could also explain why cancers from tissues with higher turnover rates exhibit an increased rate of CpG>TpG mutations (Alexandrov et al., 2015).

It is worth noting that we also observed a less consistent but significant (p-value < 0.05) enrichment of ACG>ATG and CCG>CTG mutations on the lagging strand template in proofreading proficient samples, especially in highly modified CpGs (Fig. 4.15, 4.16). This might be caused by the fact that the template of lagging strand is thought to be single-stranded for a longer period of time than the leading strand template, due to the discontinuous nature of lagging strand DNA synthesis (Okazaki et al., 1968; Seplyarskiy et al., 2016b; Hoopes et al., 2016). Single-stranded DNA is not only more prone to APOBEC-induced deamination, but also spontaneous deamination (with up to three orders of magnitude fold-difference) (Shen et al., 1994). Similarly as proposed in the mutagenesis caused by APOBECs, the mutations resulting from spontaneous deamination of 5mC could be expected to show an enrichment on the lagging strand. It is therefore possible that the observed replication strand asymmetries for CpG>TpG mutations are resulting from two opposing processes: spontaneous deamination enriched on the lagging strand and replication-induced mutagenesis enriched on the leading strand. This could mean that the effect of replication-induced CpG>TpG mutagenesis is not limited to the GCG context, but only in this context it outweighs the lagging-strand enrichment resulting from the spontaneous deamination. Further experimental research will be needed to evaluate the strand asymmetry of spontaneous deamination-induced mutagenesis and to validate the hypothesised replication-linked source of CpG>TpG mutations on the leading strand.

[3]The CpG>TpG mutations account only for 6–7 % of all mutations on average in the WGS samples, based on the mutation list from the supplementary data in (Lodato et al., 2015).

Since BS-seq does not distinguish between 5mC and 5hmC, we cannot make separate predictions about the role of these two modifications in the hypothesised replication source of CpG>TpG mutagenesis. 5mC is markedly more abundant than 5hmC (HPLC measurements of bulk $5hmC_{rel}$ are 6.3 % in colon, 7.0 % in stomach, and 8.5 % in rectum). It is therefore likely that the mutagenesis is mostly driven by 5mC. The results in brain, where 5hmC-specific maps are available, indicate that the replication-linked source of CpG>TpG mutations is indeed caused by 5mC rather than 5hmC, as the mutation frequencies in *POLE*-MUT samples decrease with increasing $5hmC_{rel}$ levels (Fig. 4.7). However, it is unclear how well the 5hmC maps correspond to the profiles in *POLE*-MUT samples at the time when most of the mutations were acquired, and therefore further data and experimental validation are needed to determine the fidelity of replicating 5mC and 5hmC.

In summary, the presented results suggest a possibility that part of the CpG>TpG mutations originate from erroneous replication instead of spontaneous deamination. It is unknown what might be the replication-linked proportion of CpG>TpG mutations in most somatic cells, nor whether this mechanism could also influence germline cells. CpG>TpG mutations are frequent also in the germline mutational spectra (Kong et al., 2012; Rahbari et al., 2015) and they correlate with methylation levels measured in human sperm cells (Mugal and Ellegren, 2011). Single nucleotide differences between closely related species have been used to estimate timing of species divergence during evolution (Arnheim and Calabrese, 2009). It has been suggested that this is most reliably estimated using CpG transitions, as they are caused by clock-like spontaneous 5mC deamination and are not affected by replication, as opposed to other types of mutations (Moorjani et al., 2016). This was supported by male bias α (male-to-female mutation ratio), which was high ($\alpha \sim 7–8$) in non-CpG sites and CpGs within CpG islands, but low in CpGs outside CpG islands ($\alpha \sim 2$), in line with the ongoing cell division in the male but not female germline (Taylor et al., 2006). However, the used methods of measuring male bias were very indirect and they did not take into account potential differences in methylation, deamination rate, nor repair efficiency between males and females, as is discussed in (Arnheim and Calabrese, 2009). Importantly, a newer and

more direct approach to measure male bias recently showed very similar values of α at CpG sites ($\alpha \sim 5.3$) and non-CpG sites ($\alpha \sim 5.6$) (Venn et al., 2014). Therefore, the existing knowledge does not contradict a potential involvement of replication in methylated CpG>TpG mutations in the germline. If this was confirmed, these results would have also important implications for the accuracy of the used methods to estimate divergence times.

4.4 UV-induced mutagenesis in modified cytosines

4.4.1 Motivation

It is known that methylation enhances the formation of CPDs (Tommasi and Pfeifer, 1997; Mitchell, 2007; Rochette et al., 2009; Martinez-Fernandez et al., 2017), the main mutagenic lesion formed after UV irradiation. C and 5mC in CPDs deaminate within hours into U and T, respectively, and during replication they are both paired with A by Pol η, creating a C>T mutation (Song et al., 2014). The deamination rate is highest in a TCG sequence context, which is the most commonly mutated trinucleotide in skin cancers (Cannistraro and Taylor, 2009). Given the enhancement of CPD formation by methylation, it could be therefore expected that the UV-induced C>T mutation frequency in skin cancers is positively correlated with methylation levels. We have tested this hypothesis in 183 WGS melanoma cancers and BS-seq maps from normal skin exposed to sunlight.

4.4.2 C>T mutations in melanoma show parabolic relationship with DNA modification levels

We first binned the CpG positions by their modification level (0-0.1, ..., 0.9-1.0) and computed the frequency of C>T mutations separately for each sequence context and each skin cancer sample. As expected, the TCG context was an order of magnitude more frequently mutated than ACG, CCG, or GCG (Fig. 4.17). Surprisingly, the relationship between TCG>TTG mutations and modification levels was non-monotonic, with a shape of a negative parabola and maximum in the middle modified positions. In total, 86 % of the skin cancer samples had the middle mod levels more modified than the low and high mod levels; moreover, the remaining 14 % of samples had only low numbers of mutations (Fig. 4.17B).

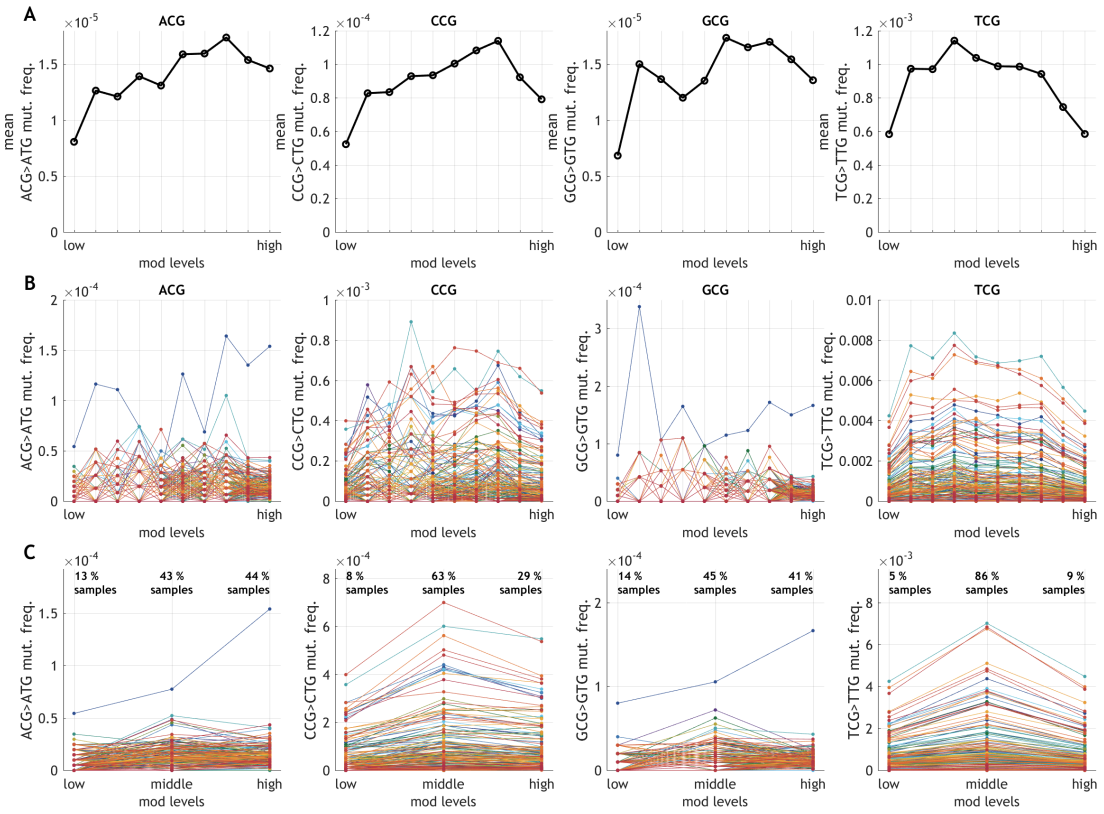

Figure 4.17. TCG>TTG mutations in skin cancer are highest in intermediate skin modification levels. All CpGs were binned according to their BS-seq measured mod levels (0-0.1, …, 0.9-1.0) from normal skin exposed to sun. The first bin represents unmodified sites and the last bin represents fully modified sites. C>T mutation frequency was computed in each bin, separately for each sequence context (columns). **A:** Mean over samples. **B:** One trace per sample. **C:** Only the low mod (first bin), high mod (last bin), and middle mod (mean of the two middle bins) values are shown. The percentage of samples with the highest mutation frequency in the low mod, middle mod, and high mod are written at the top of the figure. For example in TCG context, 86 % of samples have the middle mod value higher than the two extreme values.

We explored whether this parabolic relationship could be a property of the used modification map. For example, the positions of cytosine modifications in skin could be different from other tissues, or perhaps the particular used map might be inaccurate due to a technical bias. We therefore computed the TCG>TTG frequency of skin mutations in positions binned by modification levels using a number of other BS-derived maps: skin protected from sun, and other tissues. We observed a similar relationship in all the maps (Fig. 4.18), suggesting that it is a property of the mutations rather than the modification map.

The observed non-monotonic relationship is surprising. It is important to recall

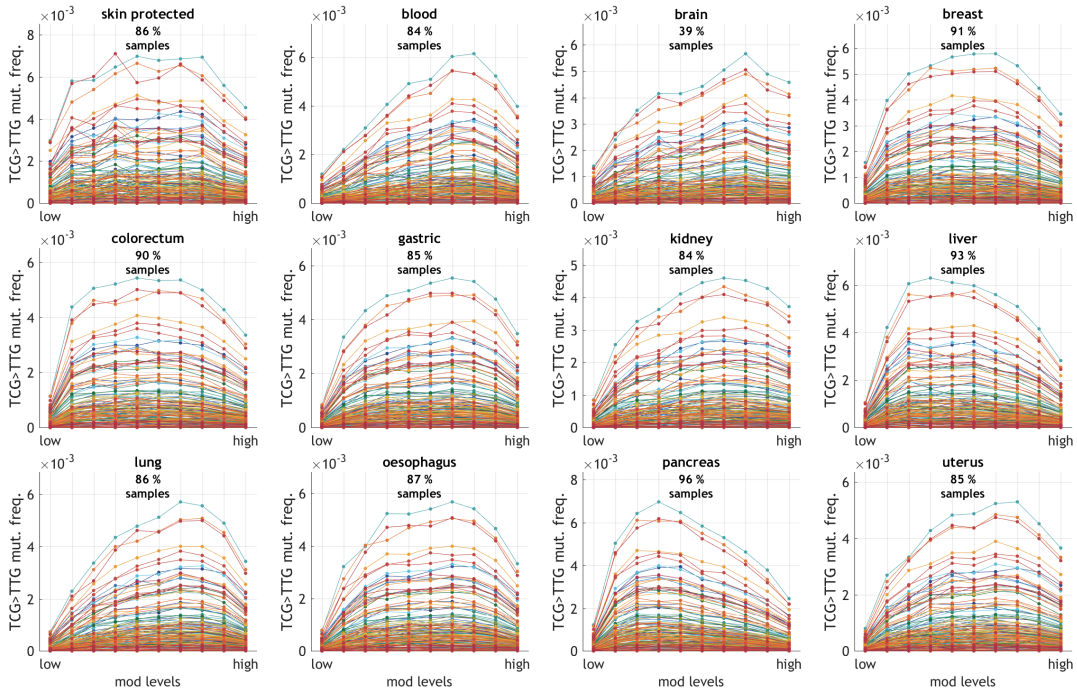

Figure 4.18. TCG>TTG mutations in skin cancer are highest in intermediate modification levels in a number of BS-derived modification maps. All CpGs in a TCG context were binned according to their BS-seq measured mod levels (0-0.1, …, 0.9-1.0) from normal skin protected from sun (first figure) and other normal tissues. The first bin represents unmodified sites and the last bin represents fully modified sites. C>T mutation frequency of melanoma mutations was computed in each bin separately for each sample (one trace per sample). The number at the top of each figure represent the percentage of samples with TCG>TTG mutation frequency higher in the middle modified positions than in lowly and highly modified positions.

that each position in one cell (and one allele) is either modified, or unmodified. The intermediate levels of modifications come from a mix of cells, some modified and others unmodified in this position. The mutation frequency in an individual position in one cell should be independent of the modification state of the same position in other cells. We would therefore expect to see either no relationship between mutagenesis and modification levels, or a monotonic relationship. However, a third variable might be involved: positively correlated with consistently modified cytosines (in most of the cells) and negatively correlated with UV-induced mutation frequency. Alternatively, the parabolic relationship might result from the combined nature of BS-seq measurements, i.e., that the modification levels represent a combination of 5mC and 5hmC levels.

4.4.3 5hmC is negatively correlated with C>T melanoma mutations

We first explored the possibility that a part of the highly modified CpGs are hydroxymethylated and that 5hmC has an opposing impact on the UV-induced mutagenesis compared to 5mC, resulting in a parabolic relationship when the combined BS-seq measurements are used. Direct single-base resolution maps of 5hmC are unfortunately currently not available for skin. However, regional estimates of 5hmC from hMeDIP-seq in primary benign naevus have been measured (Lian et al., 2012).

In hMeDIP-seq, hydroxymethylated DNA is enriched and the reads are then sequenced. Regions with higher 5hmC levels are therefore more covered by the aligned reads than low-5hmC regions. We binned CpGs according to their coverage in the hMeDIP-seq and computed mutation frequency in each bin. Interestingly, TCG>TTG mutations exhibited a steep decrease with increasing levels of 5hmC (i.e., coverage in hMeDIP-seq) (Fig. 4.19). The mean mutation frequency dropped 5-fold between uncovered CpGs (low 5hmC) and CpGs with at least 10 reads (high 5hmC).

We next combined the hMeDIP-seq measurements with MeDIP-seq measurements (i.e., enrichment for 5mC) from the same study and sample. We binned all CpGs by the combination of coverage in hMeDIP-seq and MeDIP-seq (0, 1, 2, or at least 3 reads). We observed that the decrease of C>T mutation frequency with increasing 5hmC levels is present in each MeDIP-seq coverage (columns in the figures) (Fig. 4.20). The rows of the figures (i.e., increasing MeDIP-seq coverage with fixed hMeDIP-seq coverage) were in most cases increasing (such as in ACG and CCG contexts, and in higher 5hmC levels of TCG context). In summary, although the hMeDIP and MeDIP measurements are not ideally quantitative and with sufficient resolution, the results support the hypothesis that 5hmC is negatively correlated with UV-induced mutagenesis.

In low 5hmC positions (up to 1x hMeDIP-seq coverage) and the TCG context, the parabolic relationship of C>T mutations with increasing 5mC levels was still present (Fig. 4.20). This suggests that either the parabolic relationship is not fully driven by the combined nature of BS-seq, or that hMeDIP-seq and MeDIP-seq measurements are not

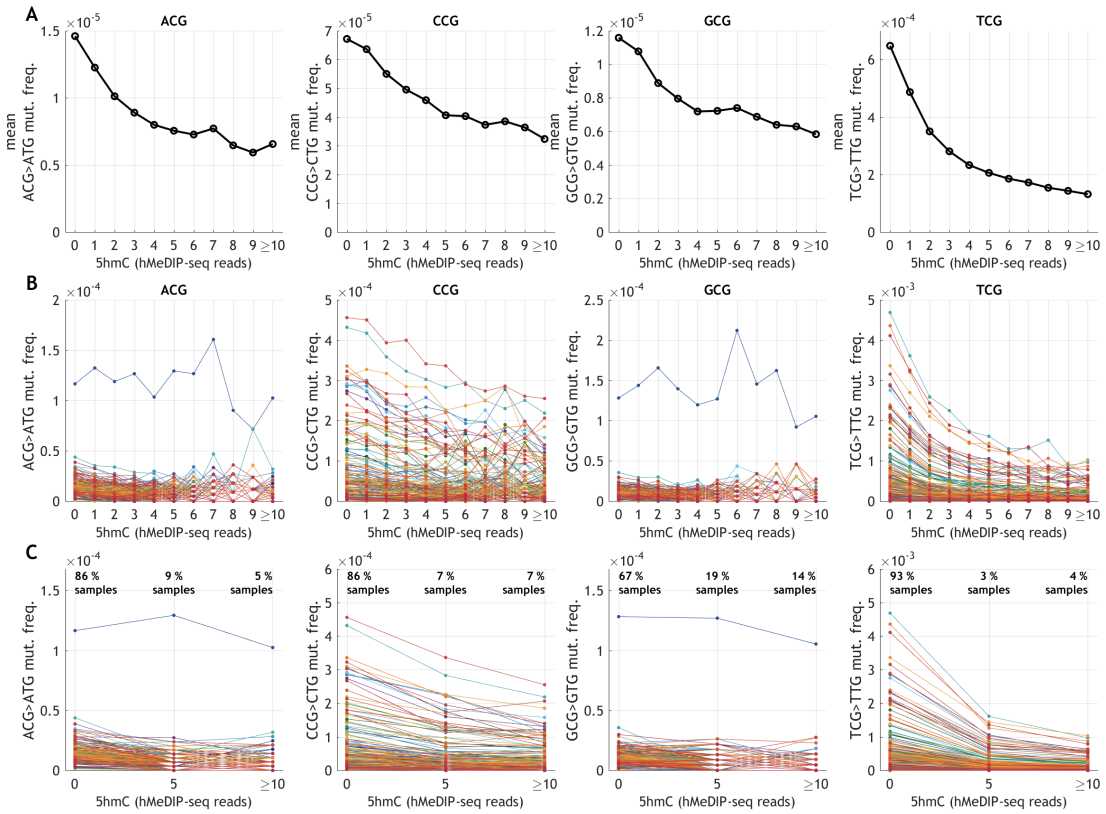

Figure 4.19. 5hmC estimates from hMeDIP negatively correlate with C>T skin mutations. All CpGs were binned according to their coverage in hMeDIP-seq measurements from benign skin naevus. Higher coverage represents higher 5hmC levels. C>T mutation frequency was computed in each bin, separately for each sequence context (columns). **A:** Mean over samples. **B:** One trace per sample. **C:** Quantification which samples have highest mutation frequency in low 5hmC (0 reads) vs. intermediate 5hmC (5 reads) vs. high 5hmC (at least 10 reads).

sufficiently accurate and quantitative for this question. Additional data from a complementary approach are therefore needed to determine the cause of the non-monotonicity.

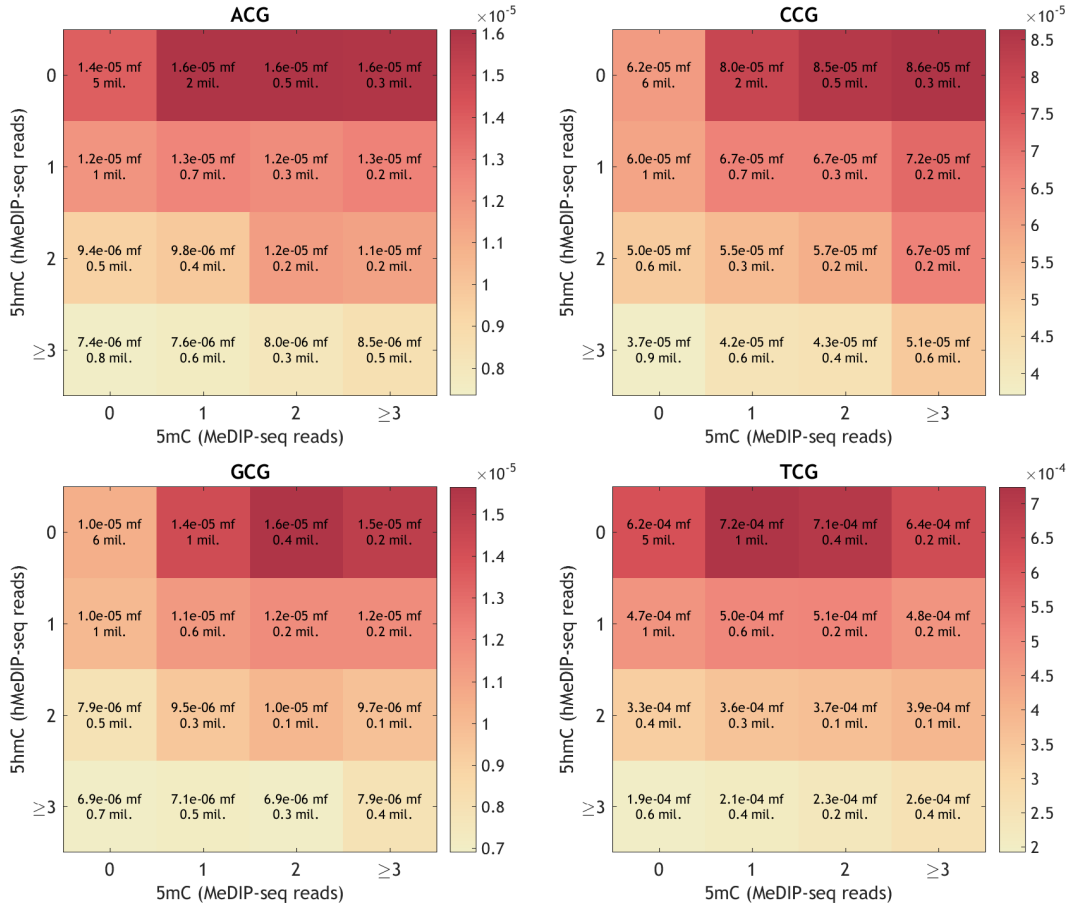

Figure 4.20. 5hmC estimates from hMeDIP negatively correlate with C>T skin mutations even after stratification by MeDIP coverage. All CpGs were binned according to their coverage in hMeDIP-seq and MeDIP-seq measurements from benign skin naevus. Higher coverage represents higher 5hmC levels in hMeDIP-seq (rows) and 5mC levels in MeDIP-seq (columns). C>T mutation frequency was computed in each bin, separately for each sequence context, and plotted as a heatmap. The numbers inside the heatmaps represent the mutation frequency and number of CpGs in each bin.

In order to try to explore the cause of non-monotonicity using a complementary approach, we used the available whole-genome BS-seq and TAB-seq measurements (available for brain, kidney, and blood) and made consensus mod and 5hmC maps from them (see Methods 4.2.7). We made an assumption that the consensus could represent a reasonable approximation of measurements for other tissues, including skin.

We used the consensus maps to compute the frequency of C>T skin mutations with respect to 5mC estimated as the difference of the consensus mod and 5hmC in each position in these tissues. Compared to the parabolic shape when mod maps were used (Fig. 4.18), the direct 5mC consensus estimates showed a nearly linear increasing relationship (Fig. 4.21). This would support the notion that 5hmC part of BS-seq measurements was driving the original non-linearity.

These results suggest that 5mC and 5hmC have an opposing effect on the UV-induced mutagenesis (enhancement by 5mC and protection by 5hmC; although the presented results show only correlations, not direct mechanism), resembling the relationship observed in brain, kidney, and blood C>T mutations in Chapter 3. In order to compare the size effects, we computed C>T mutation frequency in skin with respect to $5hmC_{rel}$, i.e., 5mC/mod, using the consensus mod and 5hmC maps and melanoma mutations (Fig. 4.22). In line with the previous observations in this section, the consensus $5hmC_{rel}$ is negatively correlated with C>T mutation frequency, showing a striking 7.3-fold decrease from fully methylated to fully hydroxymethylated positions in a TCG context. Since the decrease in other tissues was approximately two-fold, this suggests that the difference between 5mC and 5hmC on UV-induced mutagenesis is even larger than other sources of mutagenesis in CpG sites, such as spontaneous deamination.

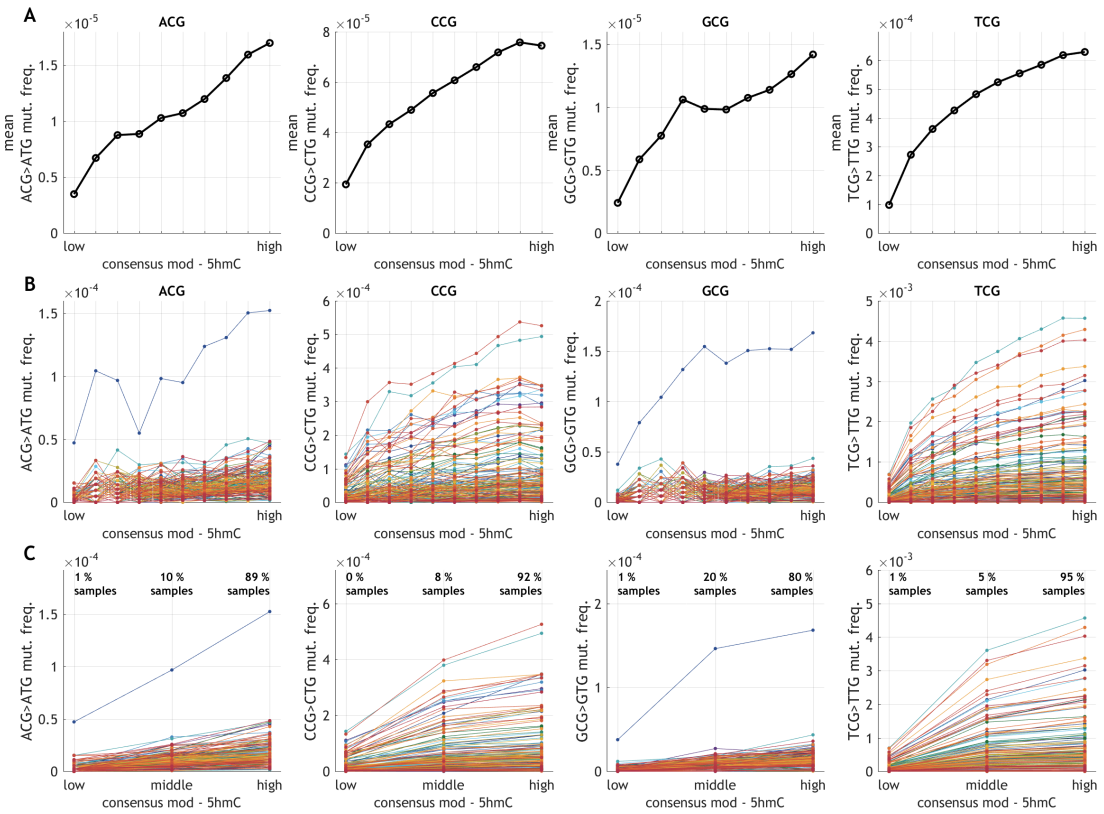

Figure 4.21. C>T mutations in skin cancer positively correlate with consensus estimates of 5mC (mod-5hmC). All CpGs were binned according to the consensus 5mC = mod - 5hmC, using a consensus from brain, kidney, and blood BS-seq and TAB-seq measurements. The first bin represents unmethylated sites and the last bin represents fully methylated sites. C>T mutation frequency was computed in each bin, separately for each sequence context (columns). **A:** Mean over samples. **B:** One trace per sample. **C:** Only the low 5mC (first bin), high 5mC (last bin), and middle 5mC (mean of the two middle bins) values are shown. The percentage of samples with the highest mutation frequency in the low 5mC, middle 5mC, and high 5mC are written at the top of the figure. For example in TCG context, 95 % of samples have the high 5mC value higher than the low 5mC and middle 5mC values.

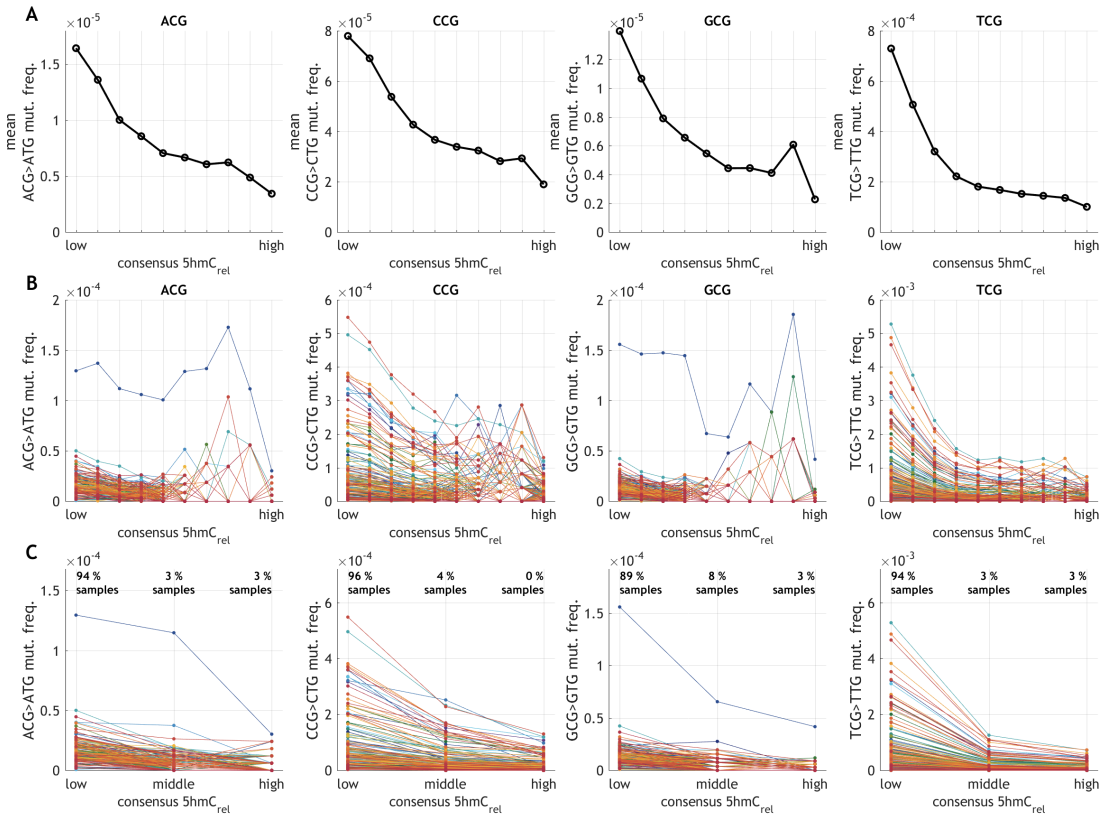

Figure 4.22. C>T mutations in skin cancer steeply decrease with increasing consensus estimates of $\mathbf{5hmC_{rel}}$. All CpGs were binned according to the consensus $5hmC_{rel}$ = 5hmC/mod, using a consensus from brain, kidney, and blood BS-seq and TAB-seq measurements. The first bin represents methylated sites and the last bin represents hydroxymethylated sites. C>T mutation frequency was computed in each bin, separately for each sequence context (columns). **A:** Mean over samples. **B:** One trace per sample. **C:** Only the low $5hmC_{rel}$ (first bin), high $5hmC_{rel}$ (last bin), and middle $5hmC_{rel}$ (mean of the two middle bins) values are shown. The percentage of samples with the highest mutation frequency in the low $5hmC_{rel}$, middle $5hmC_{rel}$, and high $5hmC_{rel}$ are written at the top of the figure. For example in TCG context, 94 % of samples have the low $5hmC_{rel}$ value higher than the middle $5hmC_{rel}$ and high $5hmC_{rel}$ values.

4.4.4 Nucleosome positioning affects melanoma mutation profiles

One of the factors known to have an influence on UV-induced mutagenesis is the positioning of nucleosomes. Three types of influence have been observed: formation of CPDs, deamination within CPDs, and repair by NER. The positions with the backbone furthest away from the histone surface ("OUT positions") were observed with enhanced formation of CPD *in vitro* by a factor of two, compared to free DNA and positions with the backbone closest to the histone surface ("IN positions") (Song et al., 2011). Moreover, the deamination rate in OUT positions was 8.9-fold increased, while it was 4.7-fold decreased in IN positions, compared to free DNA (Song et al., 2011, 2014), resulting in a 42-fold increased deamination rate in OUT compared to IN positions. On the other hand, the presence of nucleotides impairs different stages of repair of the lesions: detection, excision, and DNA resynthesis (Bell et al., 2011). Quantification of the impairment differs between conditions and experiments (Thoma, 2005), but for example in *Xenopus* oocyte nuclear extracts, nucleosomes decreased the NER rate by 2–3-fold and IN positions had 1.5-fold lower NER rate than OUT positions (Svedružić et al., 2005).

The combined effects of rotational and translational nucleosome positioning[4] on mutations observed in melanoma cancer patients have not yet been determined. We therefore compared the frequency of C>T melanoma mutations in CpGs with respect to the distance of the nearest nucleosome dyad, using nucleosome maps from sequencing of MNase digested H1 human ESCs (Yazdi et al., 2015a).

In spite of the expected variability in the nucleosome positioning between cells and tissues, we observed a remarkable periodicity of the mutation frequency signal around aligned nucleosome dyads (Fig. 4.23). The signal peaked at a distance of ca. 120 bp from the dyad (Fig. 4.23A). The periodicity was clearest in positions up to $\pm$ 50 bp from the dyad, with an average period of 10 bp (as visible in the Fig. 4.23B–C and in the Fourier transform of the signal 4.23D–E). Computing the mutation frequency separately for

[4]The terminology of "rotational positioning" refers to orientations of DNA relative to the histone surface and 'translational positioning" refers to the DNA sequence positions relative to the dyad (Mao et al., 2017).

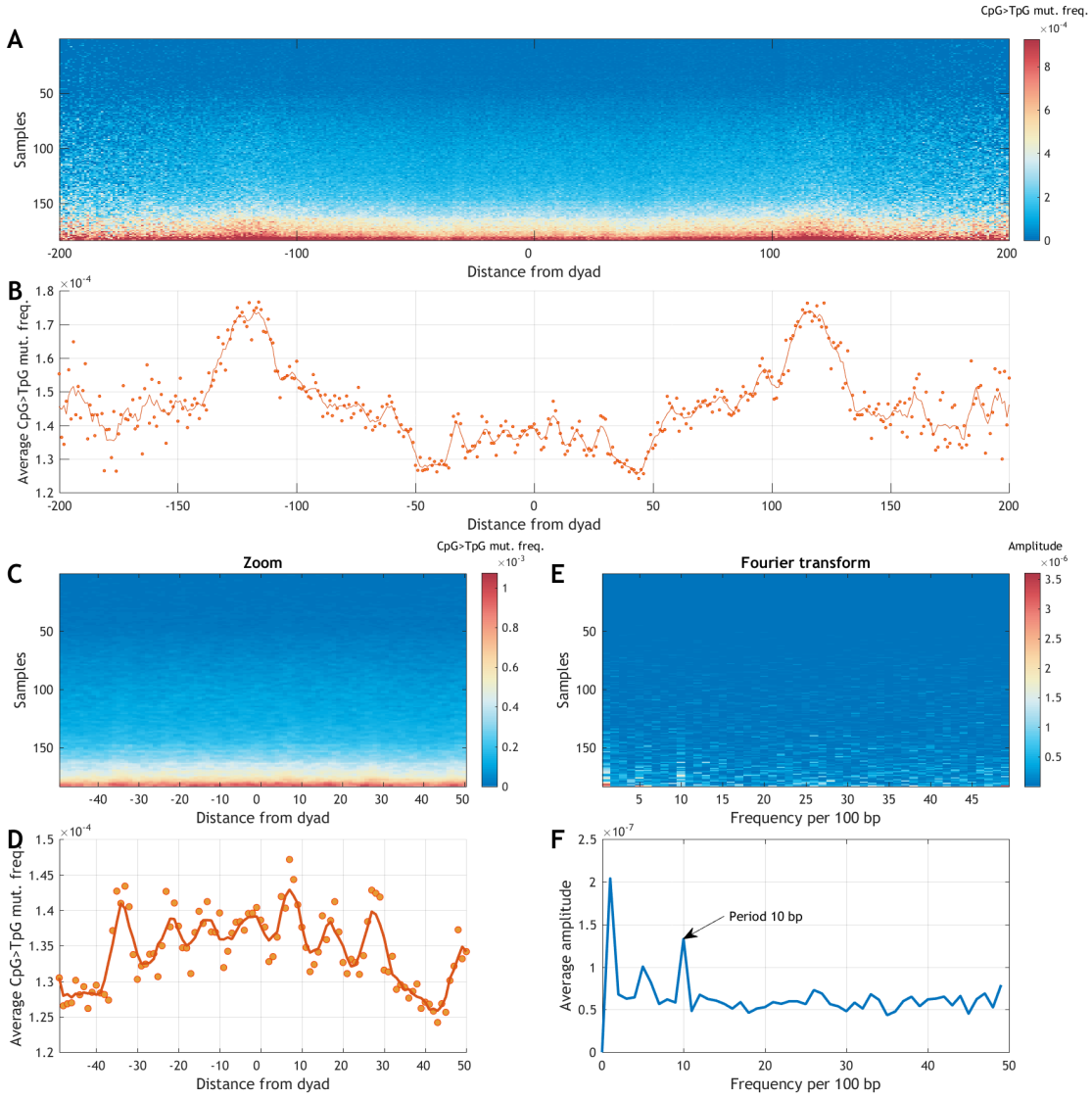

Figure 4.23. CpG>TpG mutation frequency in skin is influenced by nucleosome rotational positioning. All CpG positions were binned according to their distance to the nearest nucleosome dyad. C>T mutation frequency was then computed in each bin and sample. **A:** C>T mutation frequency in CpG positions in the distance up to 200 bp from nucleosome dyad. Samples (rows of the heatmap) are sorted by their mean value. **B:** Average of samples from the heatmap in (A). A line is fit through the data points using Matlab function smooth. **C:** A zoom of (A) up to 50 bp distance from the dyad. **C:** Average of samples from the heatmap in (C). **E:** Fourier transform, separately for each sample (row) from (D). **F:** Average of samples from (E). The peak at 10 corresponds to a period of 10 bp.

the four 5' sequence contexts revealed that the signal is most consistent in the TCG context, with a sharp peak at ca. 10 bp from the dyad (Fig. 4.24).

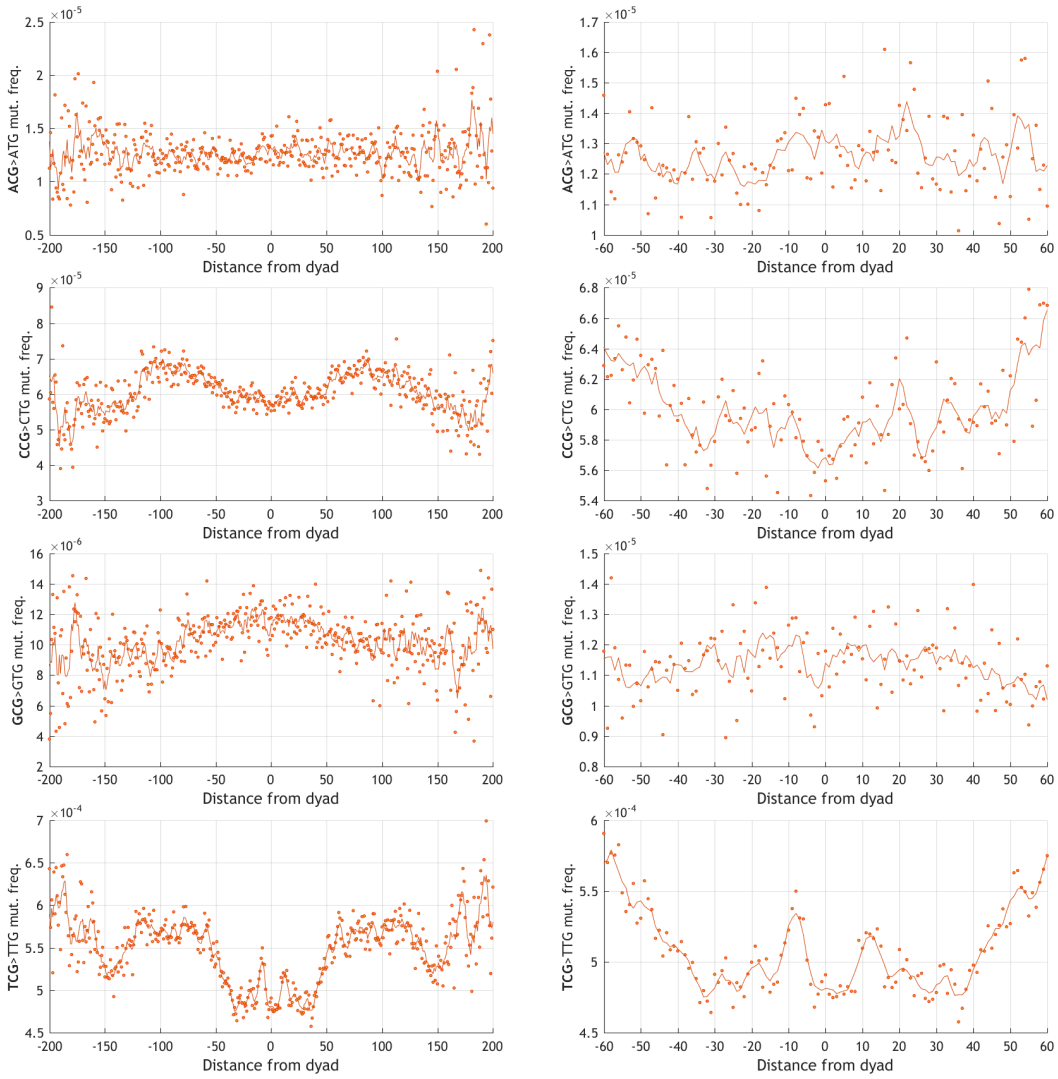

Figure 4.24. The influence of nucleosome positioning on skin C>T mutations is strongest in a TCG context. Average CpG>TpG mutation frequency with respect to distance from a nucleosome dyad, computed separately for each 5' sequence context (rows). The second column represents a zoomed view of the first column.

The observed periodicity resembled some of the experimental measurements of CPD formation and deamination rate. The positions with high mutation frequency in our results correspond to the OUT positions (outward rotation setting), which were reported to have both higher formation of CPDs and higher deamination rate (Mao et al., 2017). In particular, the C>T mutation frequency in a TCG context resembles the

positions with high CPD density, as measured by CPD-seq *in vivo* in UV irradiated yeast cells (Fig. 2B 0h in (Mao et al., 2016)).

We next computed the mutation frequency separately for the plus (Watson) and minus (Crick) strands. Strikingly, the mutation frequency in a TCG context was strongly asymmetrical on the two strands around the nucleosome dyad (Fig. 4.25). Compared to the expected periodicity of several base pairs, the observed asymmetry switched at a distance ca. 90 bp from the dyad. The asymmetry was highly significant (signtest p-value $2 \cdot 10^{-24}$) and the predominant direction (higher in minus strand on the left of dyad and higher in plus strand on the right of the dyad) showed in 88 % of the samples.

As the mutation frequency in each bin is normalised by the number of positions with the given sequence context, the asymmetry is not a simple consequence of asymmetrical sequence context composition[5]. Moreover, the asymmetry was present in lowly modified, middle modified and highly methylated CpGs (Fig. 4.26A). The asymmetry was however slightly lower in the lowly methylated CpGs (consensus 5mC $\leq$ 0.7) than highly methylated CpGs (5mC > 0.85), suggesting that methylation is involved in the mechanism causing the asymmetry.

We also ascertained that the nucleosome strand asymmetry is not driven by replication strand asymmetry (Fig. 4.26B) or transcription strand asymmetry (Fig. 4.26D) and that it is also present in non-transcribed regions (Fig. 4.26C). The asymmetry was also slightly decreased in the transcribed strand compared to the non-transcribed strand, indicating that TC-NER is not the driving mechanism of the asymmetry, but instead the faster repair of CPDs in the transcribed strand might be decreasing the difference between the strands.

[5] Moreover, the sequence context is the same is all tissues, but other tissues —such as lung, oesophagus, or pancreas— did not show the nucleosome strand asymmetry.

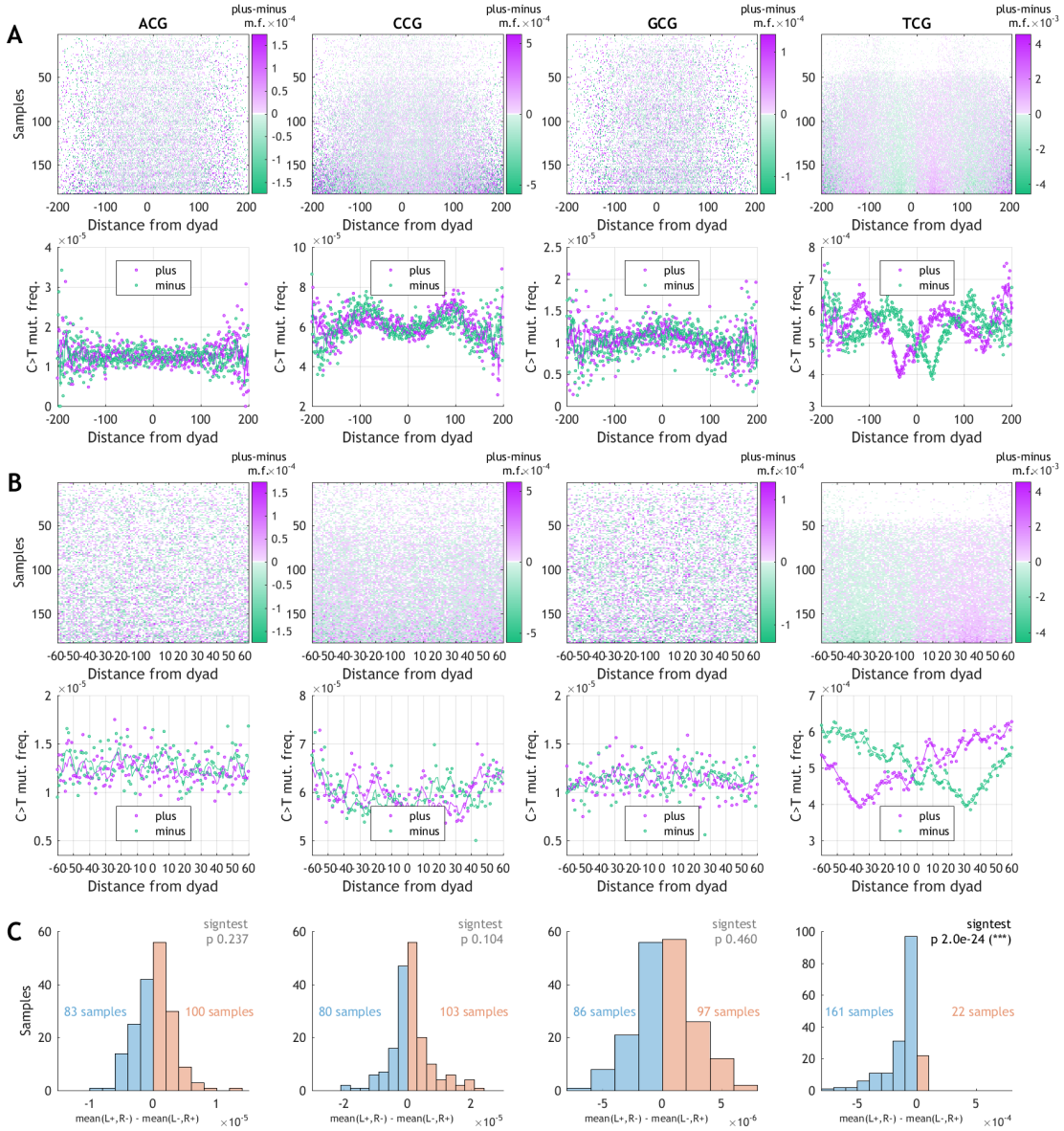

Figure 4.25. TCG>TTG mutations in skin exhibit nucleosome strand asymmetry. The CpG>TpG frequency was computed separately for the CpGs on the plus strand and the minus strand. **A:** Difference of mutation frequency in the plus and minus strand shown as a heatmap for all samples and an average of all samples below. **B:** Zoomed view of (A). **C:** Distribution of the asymmetry in the 183 skin cancer samples computed as mean(L+, R-)-mean(L-, R+), where L and R mean left and right from the dyad, respectively, in plus (+) and (-) strands. Numbers of negative and positive samples are printed on the sides of the histogram.

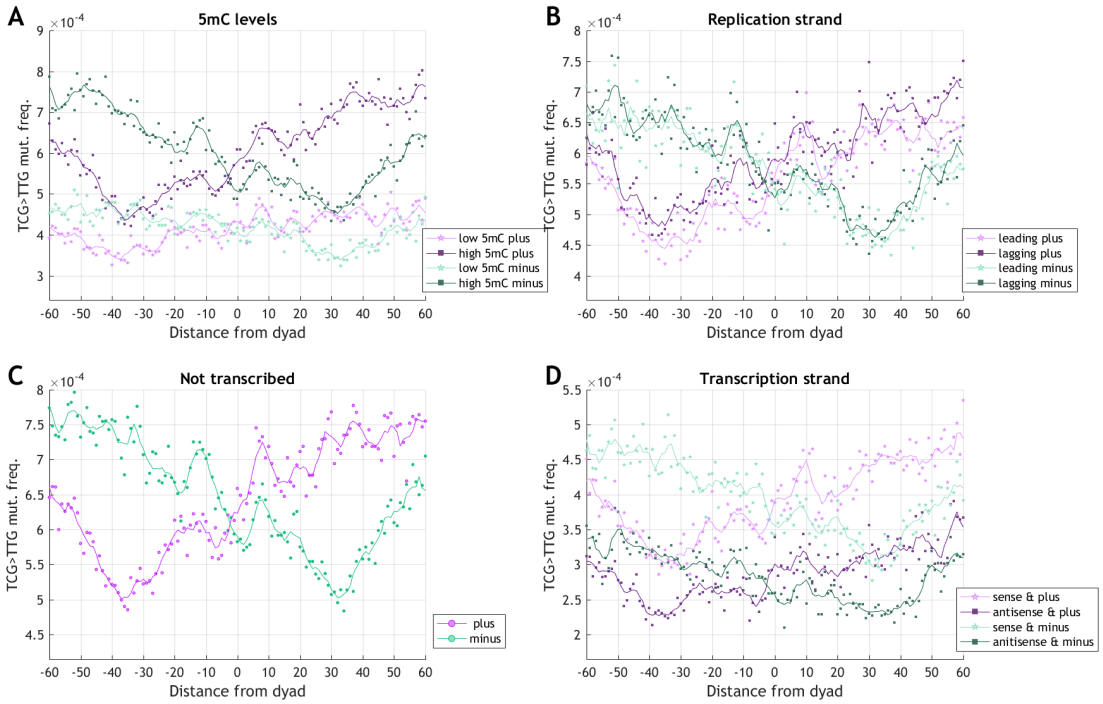

Figure 4.26. The skin nucleosome strand asymmetry is not explained by methylation, replication strand bias, or transcription strand bias. **A:** TCG>TTG mutation frequency around nucleosome dyad computed separately for CpG positions with low consensus 5mC ($\leq$ 0.7) and high consensus 5mC (>0.85). **B:** TCG>TTG mutation frequency around nucleosome dyad computed separately for replication leading strand template and lagging strand template. **C:** TCG>TTG mutation frequency around nucleosome dyad computed for CpGs in non-transcribed regions only. **D:** TCG>TTG mutation frequency around nucleosome dyad computed in transcribed regions shown separately for sense (i.e., the non-transcribed strand) and antisense (i.e., the transcribed template) strands.

4.4.5 Discussion of UV-induced mutagenesis in CpG sites

4.4.5.1 The impact of DNA modifications on UV mutagenesis

The enhancement of CPD formation by cytosine methylation is well documented (Tommasi and Pfeifer, 1997; Mitchell, 2007; Rochette et al., 2009; Martinez-Fernandez et al., 2017). However, the ultimate influence of cytosine modifications on UV-induced mutagenesis has not been explored in great depth. UV-induced mutation frequency is highest for C>T mutation in a TCG context. This has been linked both to methylation, as well as the sequence context itself, which was shown *in vitro* to be more prone to spontaneous deamination than other sequence contexts (Cannistraro and Taylor, 2009).

Here, we show that the UV-induced mutations are non-monotonically related to BS-seq derived measurements of cytosine modification levels in the individual positions: the mutation frequency first increases with increasing modification levels, but then decreases back to a similar level in fully modified sites as in unmodified sites. This non-linear relationship of skin mutagenesis with BS-seq levels was recently reported also by Poulos et al. (2017). As they used different cancer samples and BS-seq map and different analysis pipeline (developed simultaneously with our pipeline), our and their results represent an independent confirmation of the observed phenomenon. Poulos et al. (2017) suggested that this non-linearity is caused by decreased NER in highly modified sites in late-replicating regions, as NER-deficient $XPC^{-/-}$ squamous cell carcinoma samples show a slight increase of the vertex of the mutation-modification parabola: from 0.51 to 0.64 (Poulos et al., 2017). Although the vertex shifted, the shape of the relationship remained parabolic, suggesting that other mechanisms might be involved.

We explored whether the parabolic relationship could be also affected by the fact that BS-seq is a combined measurement of 5mC and 5hmC. A consensus mod (BS-seq) and 5hmC (TAB-seq) maps from three other available tissues showed that the parabolic relationship disappears when consensus 5mC is used instead of mod. The resulting average curve was gradually increasing (Fig. 4.21) and 95 % of samples had maximal mutation frequency in highly methylated CpGs compared to lowly and middle methylated CpGs. Moreover, we observed a steep decrease of TCG>TTG mutation frequency with increasing consensus $5hmC_{rel}$ and with increasing 5hmC levels in independent regional estimates of 5hmC from hMeDIP-seq in normal skin. Altogether, our results suggest that 5hmC is strongly protective against UV-induced mutagenesis, with an estimated 7.3-fold decrease compared to 5mC. Direct single-base resolution genome-wide measurements of 5hmC in skin are needed to confirm this prediction.

These results are interesting in the context of experimental measurements of CPD formation in a thesis by Liu (2014). Hydroxymethylation compared to methylation decreased the formation of CPDs in UV irradiated single-stranded oligonucleotides, duplex oligonucleotides, and *in vivo* (using TET2-overexpressing melanoma cell-line

compared to TET2-mutant cell line).[6] Notably, the decrease was stronger in a TCG sequenced context than in a CCG context. Very similar results were also observed in a different study, where formation of CPDs at dipyrimidines containing 5hmC after UV irradiation was largely reduced in a TCG context, but the reduction was smaller in a CCG context (Kim et al., 2013). This could help to explain why we observe a parabolic mutation-modification relationship in a TCG context, but more monotonic shape in a CCG context (Fig. 4.17), in line with the potentially lower protection by 5hmC in the CCG context.

The dynamics of 5hmC during skin carcinogenesis are also of interest. The levels of 5hmC, as well as TET1, TET2, and TET3 mRNA expression were observed to increase after UV exposure of normal skin cells (Wang et al., 2017a; Liu, 2014). The high content of 5hmC in normal melanocytes is however gradually lost during progression from benign naevus to malignant melanoma and the loss is mostly accompanied by decrease in TET/IDH expression (Lian et al., 2012; Larson et al., 2014; Lee et al., 2015b; Pavlova et al., 2016). The restoration of the 5hmC levels was observed to decrease invasiveness and is therefore actively researched as a potential therapeutic approach (Gustafson et al., 2015; Mustafi et al., 2017; Prasad et al., 2017).

The increased 5hmC levels in UV-exposed skin highlight the importance of the impact of 5hmC on CPD formation, the major mutagenic UV-induced lesion. Our results together with the summarised experimental evidence in the literature suggest that 5hmC could have a strong protective effect against UV-induced TCG>TTG mutations, the most frequent mutation type observed in melanoma patients. Further research is therefore needed to both validate this possibility and explore its impact on the loss of 5hmC during tumour progression.

4.4.5.2 The impact of nucleosomes on UV mutagenesis

Our results show that C>T mutations in skin are strongly affected by the nucleosome rotational positioning. This is to our knowledge the first report of such observation in skin

[6] In contrast with CPD, hydroxymethylation enhanced UV induction of 6-4PP lesions (Liu, 2014). However, 6-4PPs are rapidly repaired and though to little contribute to the mutation spectra observed in skin cancers (see Introduction 1.3.3.6).

cancer mutagenesis. Nucleosome occupancy on a broader scale has been shown to affect mutagenesis both in positive and negative directions (Hodgkinson et al., 2012; Schuster-Böckler and Lehner, 2012; Yazdi et al., 2015b). In particular, melanoma mutations were found enriched in regions with higher nucleosome occupancy, presumably due to decreased repair of UV lesions in these regions (Hodgkinson et al., 2012; Hara et al., 2000; Yazdi et al., 2015b; Polak et al., 2014; Zheng et al., 2014). However the impact of positions of DNA in the nucleosome on the cancer mutagenesis has not yet been explored.

Here we observed that CpG>TpG mutations in skin show a periodicity of 10 bp around nucleosome dyad, with higher frequency at OUT positions compared to IN positions. This is in line with formation of CPDs, the major mutagenic UV lesions, which are also found most frequently in OUT positions and least frequently in IN setting (Gale et al., 1987; Smerdon and Conconi, 1999; Liu, 2015; Mao et al., 2016, 2017). This is thought to be caused by the variation in mobility of DNA in the nucleosome, as it is minimal where the minor groove faces toward the histone octamer, and maximal for the bases with phosphate groups on the outside of the nucleosome particle (Mao et al., 2017) (Fig. 4.27). The increased mobility of dipyrimidines with their minor groove facing away from the histone octamer should make these regions the most favourable sites for CPD formation in the nucleosome core particles (Mao et al., 2017).

Not only the formation of CPDs, but also the deamination of 5mC within a CPD in a TCG context was previously shown to be affected by nucleosome positioning, being inhibited for the CPDs closest to the histone surface and enhanced for the outermost CPDs near the dyad (Song et al., 2014, 2011). Several mechanisms have been discussed as a potential cause of the difference: DNA flexibility, water accessibility, and a "flip-out" mechanism (Song et al., 2014). Differential deamination rate is therefore a second mechanism that might contribute to the differences in mutation frequency with respect to the distance from the dyad, as we observe a most consistent signal in a TCG context.

However, the most striking signal revealed by our analysis is the asymmetry of TCG>TTG mutation frequency between the two strands, flipping at the dyad. The difference was strongest at ca. 30–50 bp from the dyad, i.e., opposite the centre of the DNA in the nucleosome particle (bottom part in Fig. 4.27). The region contains several

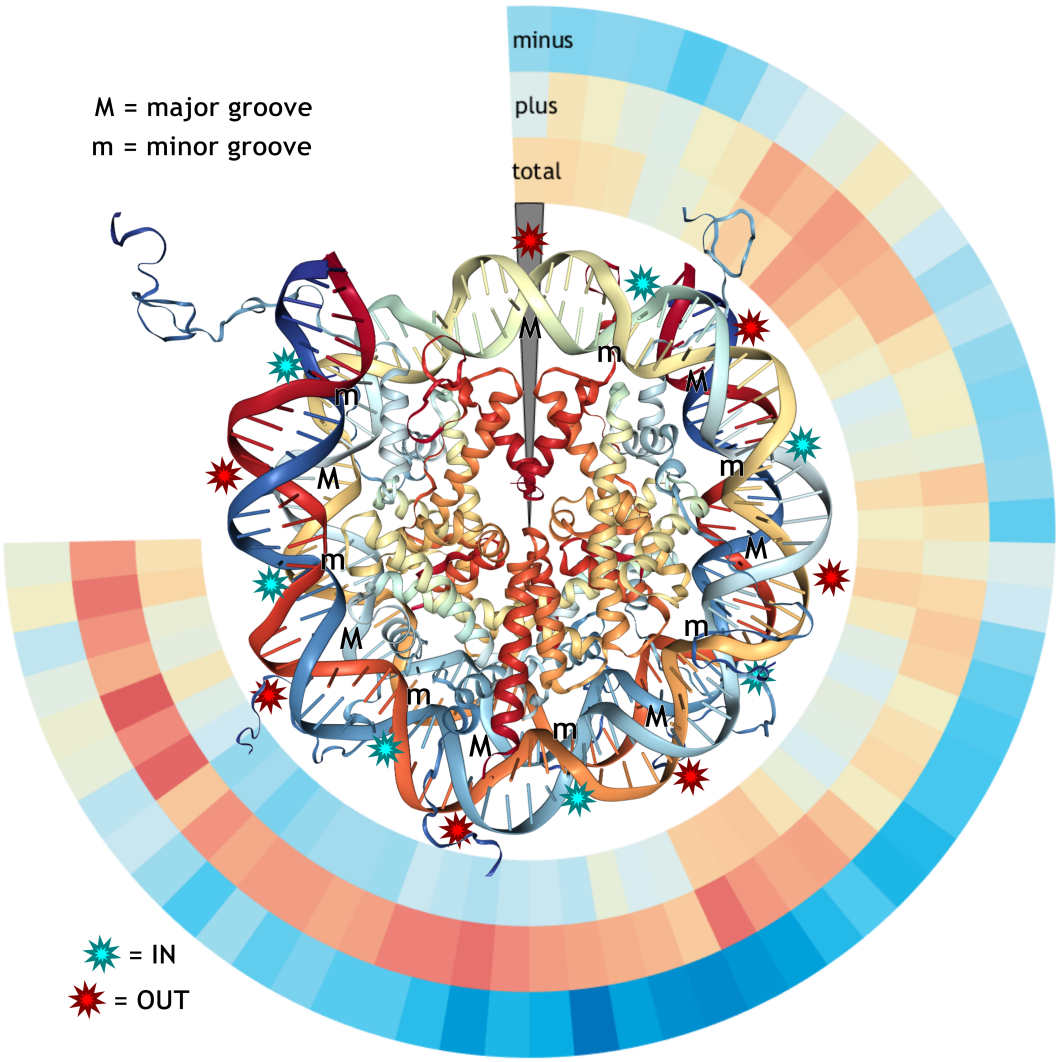

Figure 4.27. UV-induced mutagenesis in nucleosome. Crystal structure of the nucleosome core particle (PDB ID: 1KX5), rendered with NGL Viewer (Rose et al., 2016). Formation of UV-induced CPD dimers occurs more frequently at "OUT" rotational settings (indicated by red stars) than at "IN" rotational settings (blue stars) in nucleosomal DNA (Mao et al., 2017). The positions at which the major and minor grooves face the histone surface, are indicated by M and m, respectively (Wang and Taylor, 2017). The average C>T mutation frequency in melanoma cancers is visualised with approximate positions around the nucleosome core particle, with colour representing the mutation frequency (red: high, blue: low); *total* trace is from Fig. 4.23D, plus and *minus* are from Fig. 4.25A, last column.

DNA-histone interactions, which might affect the CPD formation, deamination, and CPD repair. The tail of histone H2A has been speculated to greatly inhibit deamination in measurements of *in vivo* deamination rates of CPDs placed in different translational and rotational positions along the nucleosome DNA (Cannistraro et al., 2015). Moreover, on one side of the dyad, in the OUT positions always one of the two strands faces the other DNA gyre around the nucleosome core particle and the other strand faces out from the other DNA gyre. On the other side of the dyad, the two strands swap. This can be seen in the schematic 3D structure of the nucleosome in Fig. 4.27, where the red strand always faces the other DNA gyre in OUT positions and blue strand faces out from the other gyre, but the colour swaps at the dyad. Maybe the limited space between the two gyres makes the positions facing towards the other gyre less accessible for lesion detection or repair (such as if the repair requires flipping-out of the lesion).

It will be very interesting to extend this analysis with sequencing data of CPDs and their repair, alongside the mutation spectra from NER deficient samples, to elucidate the mechanisms of the observed asymmetry and determine the proportional contributions of CPD formation, deamination, and differential repair. We also believe that applying the presented methodology to different cancer types and sequence contexts (or even mutational signatures) will be a useful approach to determine the impact of nucleosomes in different mutational processes, or even link unexplained mutational processes to their likely potential mechanisms.

4.5 Concluding remarks

The results presented in this chapter (and associated supplementary results in 10.1 and 10.2) demonstrate that the role of DNA modifications in mutagenesis goes far beyond the well-known spontaneous deamination. Here we have shown that levels of DNA modifications in normal tissues correlate with mutation frequency of the same positions in tissue-matched cancers in a number of mutational processes. In fact, most of the major processes causing mutations in the C:G pair seem to be impacted by the presence of DNA modifications. The results summarised in Table 4.1 show that the effect of cytosine modifications is not always to increase the mutagenesis. The impact is moreover not always trivially the same as predicted by the in vitro measurements, such as in the case of parabolic relationship between modification levels and TCG>TTG melanoma mutations, or unexpectedly low proportion (50 %) of samples with decreasing relationship for C>T mutations (in breast samples with a strong APOBEC signature).

Mutagenic process	Tissues	Sequence context	Mutation type	Correlation with mod	Correlation with $5hmC_{rel}$	Suggested mechanism
Tobacco	Lung	all	C>A	increasing	decreasing	BPDE formation
APOBECs	Breast	TCG	C>G	decreasing	unknown	APOBEC preference
APOBECs	Breast	TCG	C>T	mixed	unknown	APOBEC preference
UV light	Skin	TCG	C>T	parabolic; increasing for 5mC*	decreasing*	CPD formation
Replication	Colon, Rectum, Brain, Uterus, Brain	GCG, TCG	C>T	increasing	unknown; decreasing in brain	Pol ε fidelity

Table 4.1. Summary of results in this chapter. Asterisk denotes that a consensus 5hmC map was used, due to unavailability of tissue-matched map.

Interestingly, 5hmC negatively correlated with mutation frequency in all the mutational processes apart for APOBEC-induced mutagenesis, which is present in tissues, for which the 5hmC maps were not available. However, the *in vitro* measurements predict strongly decreased mutagenesis in 5hmC induced by APOBEC enzymes. Together with the previous chapter, the results thus suggest that 5hmC is generally protecting the genome from a range of mutagenic sources.

Finally, results in this chapter show an unexpected link between replication and mutagenesis in methylated positions. The results are in line with a model, in which 5mC in replicated with a slightly decreased fidelity by Pol ε, the leading strand polymerase. Although this surprising possibility first needs to be carefully validated and explored experimentally (see Conclusions 6 for the planned experiments), it opens a possibility for a novel and ubiquitous source of mutagenesis in CpGs, which is however efficiently repaired in cells proficient for post-replicative proofreading and repair. The relative contribution of spontaneous deamination compared to the hypothesised replication source of CpG>TpG mutations would also need to be researched. Deficiency of Pol ε proofreading increases the CpG mutation rate 210-fold in human cancers, while Mbd4 deficient mice exhibit an increase in mutation frequency by 3-fold (Millar, 2002). As MBD4 is one of the two main glycosylases repairing mismatches caused by spontaneous deamination of 5mC, this hints at the possibility that replication might be even more mutagenic at methylated CpGs than deamination, unless TDG plays a dominant role in repair of deaminated 5mC.

They're alive, they're awake
While the rest of the world is asleep
Below the mine shaft roads, it will all unfold
There's a world going on underground

— Tom Waits *Underground*

Oh, let the sun beat down upon my face
And stars to fill my dream
I'm a traveller of both time and space
To be where I have been

— Led Zeppelin *Kashmir*

5 The role of replication in different mutational processes

5.1 Introduction

Mounting evidence suggests replication itself contributes to cancer risk (Tomasetti and Vogelstein, 2015). However, the extent to which DNA replication influences distinct mutational mechanisms, with their manifold possible causes, remains incompletely understood.

Copying of DNA is intrinsically asymmetrical, with leading and lagging strands being processed by distinct sets of enzymes (Lujan et al., 2016), and different genomic regions replicating at defined times during S phase (Fragkos et al., 2015). Previous analyses have focused either on the genome-wide distribution of mutation rate or on the strand specificity of individual base changes. These studies revealed that the average mutation frequency is increased in late-replicating regions (Stamatoyannopoulos et al., 2009; Lawrence et al., 2013), and that the asymmetric synthesis of DNA during replication leads to strand-specific frequencies of base changes (Shinbrot et al., 2014; Lujan et al., 2012; Reijns et al., 2015; Haradhvala et al., 2016).

A very useful framework for detection of replication strand bias in single base changes was presented by Haradhvala et al. (2016). The authors binned the human

genome into 20 kbp windows and annotated each window with the direction of replication based on replication timing measurements from lymphoblastoid cell lines of six individuals (Koren et al., 2012) as described in the Methods (section 2.3.2). Regions with constant timing (i.e., the valleys and peaks) were excluded from the analysis, because they do not present a clear direction of replication and they are the source of most tissue-specific variation in replication timing (Rhind and Gilbert, 2013; Ryba et al., 2010). Using this framework, Haradhvala et al. (2016) showed that a significant replication strand asymmetry is present in *POLE*-MUT samples (C>A vs. G>T), samples with APOBEC-associated mutations (C>G vs. G>C), and MSI samples (A>G vs. T>C), while a strong transcription strand asymmetry is present in liver samples (A>G vs. T>C), samples with smoking-associated mutations (C>A vs. G>T) and UV-associated mutations (C>T vs. G>A).

The limitation of this approach is that it does not take into account several types of strand asymmetry. First, mutational processes that exhibit high mutation frequencies in only few very specific sequences contexts might not be detected. Second, mutational processes, such as one that enhances CCA>CAA mutations on the leading strand but GCA>GAA on the lagging strand, will be very possibly missed. Third, samples with multiple processes that cause the same base change but are enhanced on the opposite strands can cause inaccurate results.

In order to overcome these limitations, we decided to compute strand asymmetry using mutational signatures. This approach has the important advantage of being able to distinguish between processes that have the same major mutation type (such as C>T transitions), but differ in their sequence context and possibly also the two strands. A methodology to compute strand-specific mutational signatures was presented by Alexandrov et al. (2013a) and applied on replication in breast cancers by Morganella et al. (2016). This method extends the basic approach (described in the General methods 2.1.1) such that the vector of each signature is doubled, i.e., containing 96 values for one strand and 96 values for the opposite strand. The mutation matrix M therefore contains 192 elements for each sample, but the exposure matrix E contains one value for each sample and mutational signature.

The disadvantage of such a method is that it does not allow to quantify the magnitude of the asymmetry in individual samples, but the asymmetry is quantified for the entire cohort instead. Mutational processes with a small size effect in the asymmetry but high consistency across samples can thus be easily missed. Moreover, the direction of the asymmetry has to be unified for all samples within the cohort.

We therefore developed a method for quantification of replication strand asymmetry in individual samples, allowing for different directions in individual samples, and applied the method on 3056 WGS samples from 19 cancer types.

5.2 Materials and methods

5.2.1 Methods overview

We used two independent data sets to describe replication direction (Fig. 5.1A) relative to the reference sequence, one derived from high-resolution replication timing data (Haradhvala et al., 2016) and the other from direct detection of ORIs by short nascent strands sequencing (SNS-seq) (Besnard et al., 2012), corrected for technical artefacts (Foulk et al., 2015). The former provides information for more genomic loci, while the latter is of higher resolution. As a third measure of DNA replication, we compared regions replicating early during S phase to regions replicating late (Haradhvala et al., 2016).

We calculated strand-specific signatures (Morganella et al., 2016) of length 192, based on the direction of DNA replication (Fig. 5.1B). We further condensed the strand-specific signatures into directional signatures consisting of 96 mutation types and a binary value in each type, representing the dominant direction (leading or lagging) of the mutation type in the strand-specific signature (Fig. 5.1C).

We next designed an algorithm (section 5.2.8) to quantify presence of each signature on the leading and lagging strand in individual samples, which we call the exposure to the signature in a sample. Depending on whether the strand bias matches the consensus of the directional signature, the exposure can be *matching* or *inverse* (Fig. 5.1D). The output of the algorithm gives thus two values (matching and inverse exposure) for each sample and each signature present in the sample.

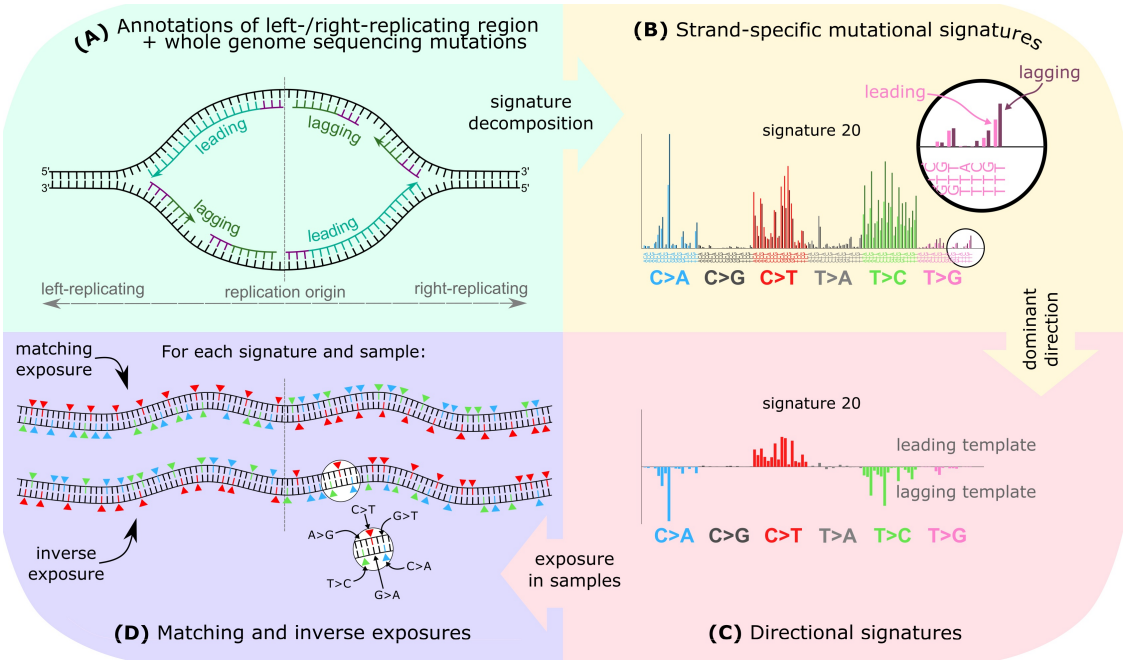

Figure 5.1. Methods overview. **A:** Mutation frequency on the leading and lagging strand is computed using annotated left/right-replicating regions and somatic single-nucleotide mutations oriented according to the strand of the pyrimidine in the base-pair. **B:** Leading and lagging strand-specific mutational signatures are extracted (signature 20 is shown as an example). **C:** Each of the 96 mutation types is annotated according to its dominant direction (upwards-facing bars for leading, downwards-facing bars for lagging template preference). **D:** Exposures to the directional signatures are separately quantified for the leading and lagging strand of each patient. The exposure in the matching orientation reflects the extent to which mutations in pyrimidines on the leading (and lagging) strand can be explained by the leading (and lagging) component of the signature, respectively. Conversely, the exposure in the inverse orientation reflects how mutations in pyrimidines on the leading strand can be explained by the lagging component of the signature (or vice-versa) (Methods). Top part of 1D shows an example of a sample with completely matching exposure, given the signature in 1C, with C>T mutations on the leading template and C>A and T>C mutations on the lagging template, whereas bottom part of 1D shows an example of a sample with completely inverse exposure.

5.2.2 Somatic mutations

Cancer somatic mutations in 3056 whole-genome sequencing samples were obtained from publicly available data sets (Table 9.4). MSI samples (gastric, colorectum, and oesophageal adenocarcinoma) and *POLE*-MUT samples (colorectum, uterus, and brain) were treated as (two) separate groups, since they are associated with specific mutational processes.

5.2.3 Direction of replication and replication origins

Left- and right-replicating domains were taken from (Haradhvala et al., 2016). Each domain (called territory in the original source code and data) is 20 kbp wide and annotated with the direction of replication and with replication timing. This was the major replication direction data set used in the analyses.

The left/right transitions of the replication domains represent regions with on average higher density of replication origins. In order to get better resolution of the replication origins, and to validate the results using independent estimates of left- and right-replicating domains, genome-wide maps of human replication origins from SNS-seq by (Besnard et al., 2012) were used. Eight Fastq files (HeLa, iPS, hESC, IMR; each with two replicates) were downloaded and mapped to hg19 using bowtie2 (version 2.1.0). To control for the inefficient digestion of λ-exo step of SNS-seq, reads from non-replicating genomic DNA (LexoG0) were used as a control (Foulk et al., 2015). Peaks were called using "macs callpeak" with parameters `–gsize=hs –bw=200 –qvalue=0.05 –mfold 5 50` and LexoG0 mapped reads as a control. Only peaks covered in at least seven of the eight samples were used. In total 1 000 1 kbp bins were generated to the left and right of each origin, as long as they did not reach half the distance to the next origin.

5.2.4 Excluded regions

We excluded gencode protein-coding genes from the major analysis in order to prevent potential confounding of the results by transcription strand asymmetry or selection. We tested that this exclusion does not bias the results and that exclusion of all (not just protein-coding) genes leads to similar results. We also excluded regions with low unique mappability of sequencing reads (positions with mean mappability in 100 bp sliding windows below 0.99 from UCSC mappability track) and blacklisted regions defined by Anshul Kundaje (Encode Consortium 2012):

- Anshul_Hg19UltraHighSignalArtifactRegions.bed,
- Duke_Hg19SignalRepeatArtifactRegions.bed,

- wgEncodeHg19ConsensusSignalArtifactRegions.bed,
- http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg19-human/.

5.2.5 Mutation frequency analysis

All variants were classified by the pyrimidine of the mutated Watson-Crick base pair (C or T), strand of this base pair (C or T), and the immediate 5' and 3' sequence context into 96 possible mutation types as described by Alexandrov et al. (2013a). The frequency of trinucleotides on each strand was computed for each replication domain. Then the mutation frequency of each mutation type in each replication domain on the leading (plus = Watson strand in left replicating domains; minus = Crick strand in right replicating domains) and lagging strand (*vice versa*) was computed for each sample.

5.2.6 Extraction of mutational signatures

Matlab code (Alexandrov et al., 2013a) was used for extraction of strand-specific mutational signatures. The input data were the mutation counts on the leading and lagging strands (summed from all replicating domains together, but without the excluded regions) in each sample. The 192-elements-long mutational signatures (example in Fig. 5.1B) were extracted in each cancer type separately (for K number of signatures between 2 and 7). The best K with minimal error and maximal stability (minimising $error_K/max(error) + (1 - stability_K)$ and with a stability of at least 0.8) was selected for each cancer type. Signatures present in only a small number of samples with very low exposures were excluded ((95th percentile of exposures of this signature) / (mean total exposure per samples) < 0.2). The remaining signatures were then normalised by the frequency of trinucleotides in the leading and lagging strand and subsequently multiplied by the frequency of trinucleotides in the genome. This made them comparable with the 30 previously identified whole-genome-based COSMIC signatures (http://cancer.sanger.ac.uk/cosmic/signatures).

Signatures extracted in each cancer type and COSMIC signatures were all pooled together (with equal values in the leading and lagging part in the COSMIC signatures) and were clustered using unsupervised hierarchical clustering (with cosine distance

and complete linkage). A threshold was selected to identify clusters of similar signatures. Mis-clustering was avoided by manual examination (and whenever necessary re-assignment) of all signatures in all clusters. The resulting 29 signatures (representing the detected clusters) contained 25 previously observed (COSMIC) and 4 new signatures. For the subsequent analysis, the signatures were converted back to 96 values: the 25 previously observed signatures were used in their original form and average of the leading and lagging part was used for the 4 newly identified signatures.

5.2.7 Annotation of signatures with leading and lagging direction

Each signature was annotated with strand direction: which of the 96 mutation types were higher on the leading strand and which on the lagging strand (Fig. 5.1C). This was based on the dominant strand direction within the signature's cluster. Mutation types (such as C>T) with small values (maximum < 0.1 or sum relative to the other strand < 1/20) or similar values on both strands (absolute maximum difference < 0.1, while sum relative to the other strand < 1/2) were assigned according to the predominant direction of other trinucleotides of the same mutation type.

5.2.8 Calculating strand-specific exposures in individual samples

Exposures to the leading and lagging parts of the signatures on the leading and lagging strands in individual samples were quantified using non-negative least squares regression with the Matlab function $e = \mathtt{lsqnonneg}(S, m)$, where

$$S = \begin{pmatrix} S_{LD} & S_{LG} \\ S_{LG} & S_{LD} \end{pmatrix} \tag{5.1}$$

$$m = \begin{pmatrix} m_{LD} \\ m_{LG} \end{pmatrix} \tag{5.2}$$

$$e = \begin{pmatrix} e_{matching} \\ e_{inverse} \end{pmatrix} \tag{5.3}$$

The matrix S_{LD} has 96 rows and 29 columns and represents the leading parts of the signatures, i.e., the elements of the lagging parts contain zeros in this matrix. Similarly, S_{LG} has the same size, but contains zeros in the leading parts. The vector m_{LD} of length 96 contains mutations on the leading strand (normalised by trinucleotides in leading strand/whole genome), and similarly m_{LG} contains mutations from the lagging strand. Finally, $e = \mathtt{lsqnonneg}(S, m)$ finds a non-negative vector of exposures e such that it minimizes a function

$$|m - S \cdot e|. \tag{5.4}$$

A similar approach has been used in Rosenthal et al. (2016) for finding exposures to a given set of signatures. Our extension includes the strand-specificity of the signatures. The interpretation of the model is that the matching exposure $e_{matching}$ represents exposure of the leading part of the signature on the leading strand and exposure of the lagging part of the signature on the lagging strand, whereas $e_{inverse}$ represents the two remaining options. It is important to note that the direction of the mutation is relative to the nucleotide in the base pair chosen as the reference, i.e., mutations of a pyrimidine on the leading strand correspond to mutations of a purine on the lagging strand.

In order to minimize the number of spurious signature exposures, the least exposed signature was incrementally removed (in both leading and lagging parts) while the resulting error did not exceed the original error by 0.5%. The resulting reported values in each sample and signature were the difference (or fold change) of $e_{matching}$ and $e_{inverse}$. In each signature, the signtest was used to compare matching and inverse exposures across samples with sufficient minimal exposure (at least 10) to the signature. Bonferroni correction was applied to correct for multiple testing.

5.2.9 Quantification of exposures with respect to replication timing, left/right transitions, and replication origins

First, we computed the exposures for the entire genomes only. Next, we computed the exposures separately for different regions:

- 4 replication timing quartiles (we computed both the strand-specific and strand-unspecific exposures in the quartiles)
- 100 bins around aligned left/right transitions ($\pm$ 50 bins, each of 20 kbp)
- 2 000 bins around aligned SNS-seq derived replication origins ($\pm$ 1 000 bins, each of 1 kbp).

In replication timing plots, a linear regression model (function `fitlm` in Matlab) was fitted to the mean exposure in each quartile (separately for matching and inverse exposures) and the significance of the linear coefficient was tested using F-test for the hypothesis that the regression coefficient is zero (function `coefTest` in Matlab).

5.3 Results

In total, we detected 25 mutational signatures that each corresponded to one of the COSMIC signatures (Supplementary Fig. 10.11–10.15) and 4 novel signatures, which were primarily found in samples that had not been previously used for signature extraction (N1 and N2 in myeloid blood cancers, N3 in melanoma cancers, and N4 in MSI and ovarian cancers) (Fig. 5.2).

We quantified the strand asymmetry for all samples and signatures present in each sample. The results confirmed that both APOBEC signatures (2 and 13) exhibit a clear strand asymmetry, with signature 13 being the most significantly asymmetric signature (p-value = $8 \cdot 10^{-100}$). The asymmetry was present in most samples and was detected both when using the replication timing-derived maps and SNS-seq derived maps (Fig. 5.3). This is in line with previous studies, where the activity of the APOBEC class of enzymes was linked to a selective editing of exposed single-stranded cytosines on the lagging strand (Morganella et al., 2016; Hoopes et al., 2016; Haradhvala et al., 2016; Green et al., 2016; Seplyarskiy et al., 2016b). We also observed differences in these signatures with respect to replication timing: signature 2 shows clear enrichment in late replicating regions ($\log_2$ fold-change 0.91 from early to late), whereas signature 13 shows only a mild increase in late replicating regions ($\log_2$ fold-change 0.18; Fig. 5.3), which is consistent with previous reports (Morganella et al., 2016). These results validate that our approach is able to correctly identify strand and timing asymmetries of mutagenic processes.

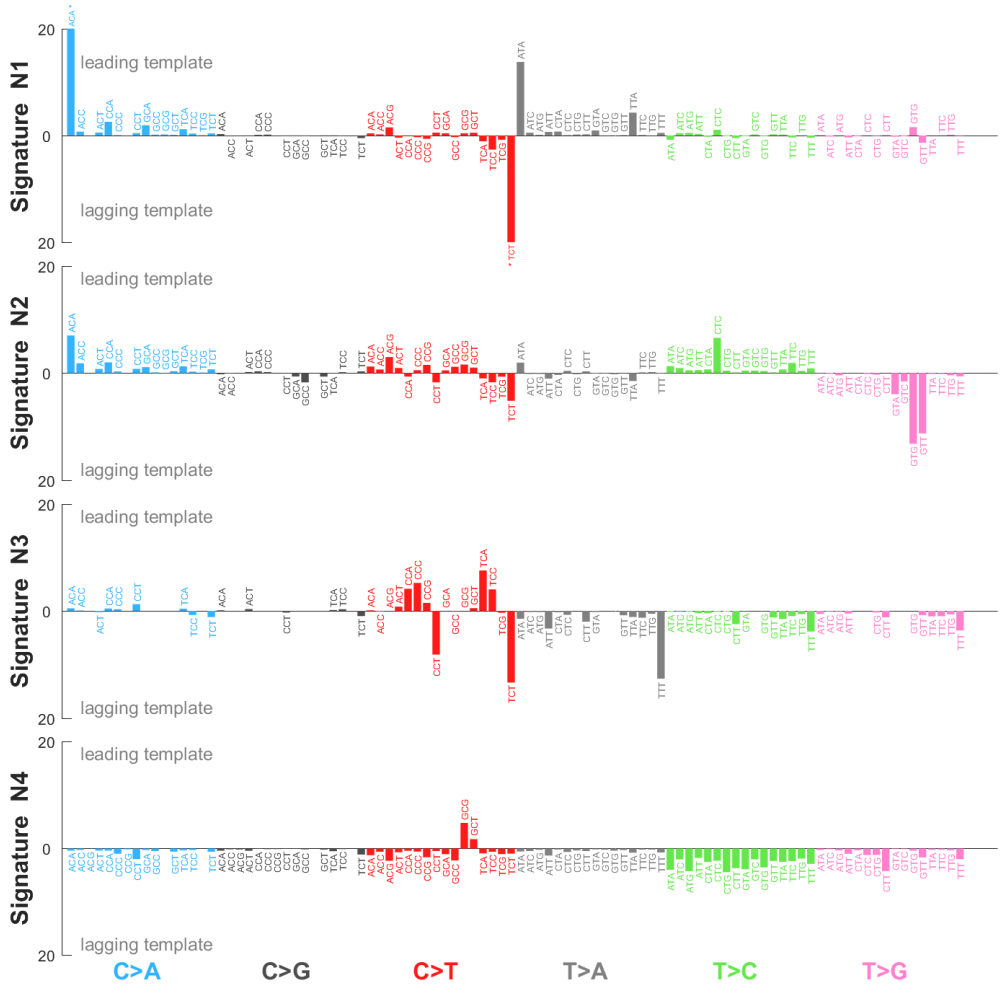

Figure 5.2. Directional signatures N1-N4 (newly detected signatures). Each of the 96 mutation types is annotated with a dominant direction: leading (pointing up), or lagging (pointing down). Asterisks indicate mutation types exceeding 20 %.

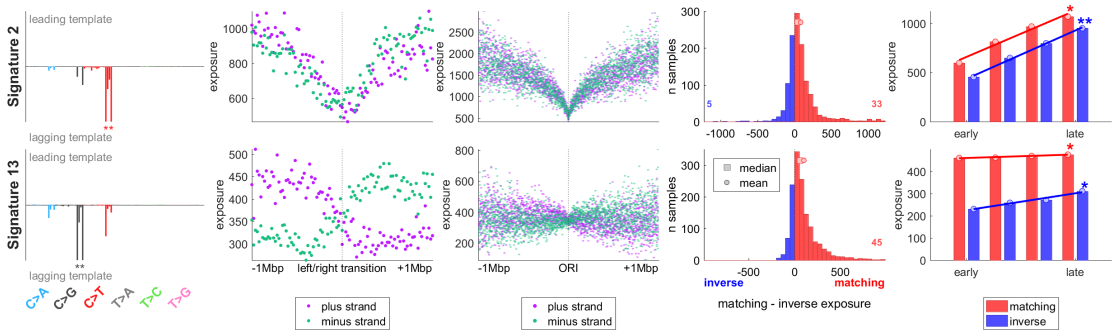

Figure 5.3. The two APOBEC signatures show strong but distinct effects of replication. Column 1: directional signatures for the two APOBEC signatures. Column 2: mean exposure on the plus (Watson) and minus (Crick) strand around transitions between left- and right-replicating regions. The transition corresponds to a region enriched for replication origins. Column 3: mean exposure on the plus and minus strand around directly ascertained replication origins. Column 4: distribution of differences between matching and inverse exposure amongst patients with sufficient exposure. Number of outliers is denoted by the small numbers on the sides. Column 5: mean matching and inverse exposure in four quartiles of replication timing; asterisks represent significance of the fit (F-test for coefficient of deviation from 0; ***P < 0.001; **P < 0.01; *P < 0.05). The leading and lagging strand annotations used in columns 4 and 5 are based on the direction of replication derived from replication timing data.

In total, 22 out of 29 signatures exhibited significant replication strand asymmetry or significant correlation with replication timing (signtest p-value < 0.05, with Bonferroni correction; Fig. 5.4) and some of the remaining signatures were significant in the main tissue associated with the signature (Fig. 5.5). Such widespread replication bias across the mutational landscape is surprising, considering that previous reports documented strand bias for only a few mutational processes (Haradhvala et al., 2016). Interestingly, the individual signatures differed in the terms of size effect of mean asymmetry, consistency among samples, slope of replication timing, and asymmetry with regards to the distance from left/right transitions and SNS-seq derived replication origins (examples of distinct signatures in Fig. 5.6).

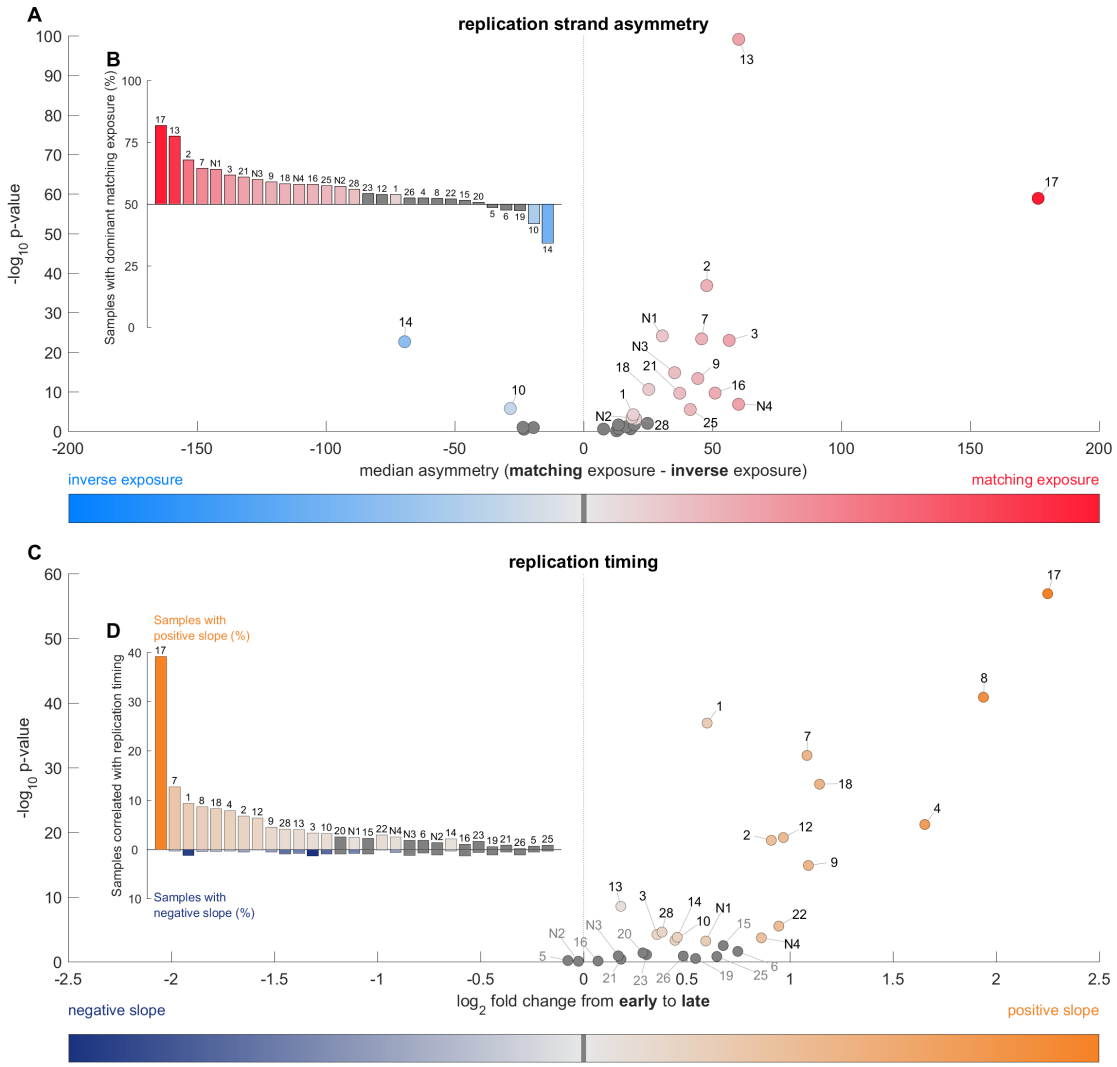

Figure 5.4. Most mutational signatures exhibit a significant replication strand asymmetry and/or correlation with replication timing. **A:** The difference of matching and inverse exposure is computed for each sample and signature. For each signature, the median value of these differences (in samples exposed to this signature) is plotted against -$\log_{10}$ p-value (signtest of strand asymmetry per sample; with Bonferroni correction). **B:** Percentage of samples that have higher matching than inverse exposure to the signature (denoted above/below each bar). **C:** X-axis: $\log_2$-transformed fold change from average exposure in early (first quartile) to late (last quartile). Values on the left denote more mutations in early-replicated regions, while on the right are later-enriched signatures. Y-axis: significance of the direction of the correlation of signature with replication timing in individual samples (signtest of correlation sign per sample: 0 for non-significant correlation, -1 for negative correlation, 1 for positive correlation; with Bonferroni correction). **D:** Percentage of samples with a significantly positive and negative correlation with exposure, respectively.

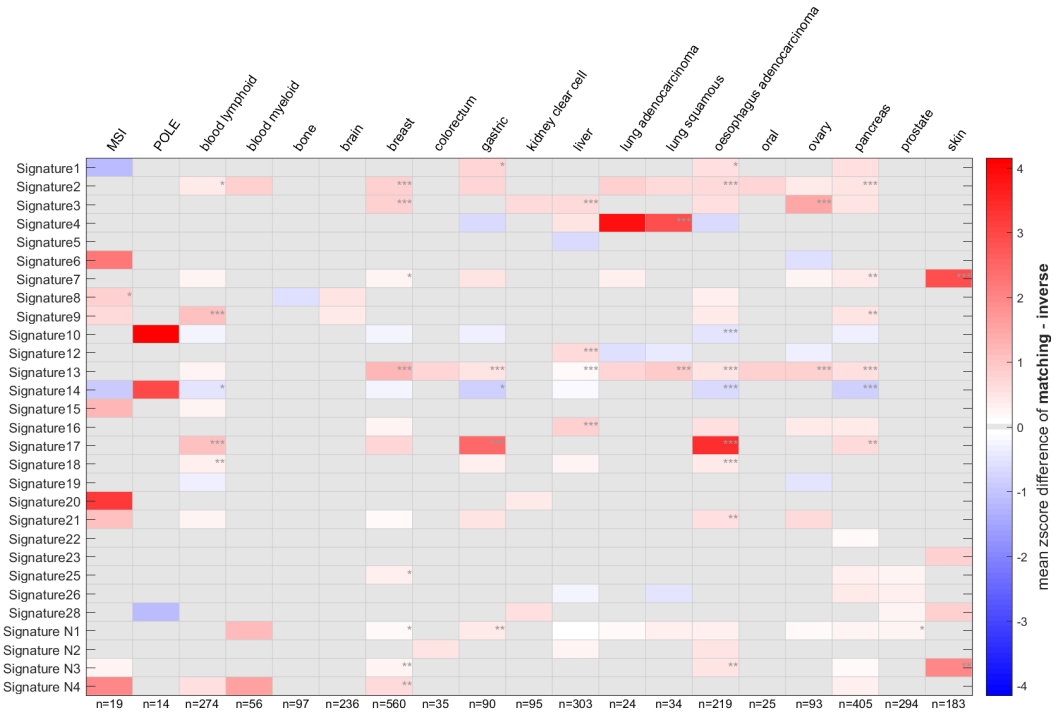

Figure 5.5. Mean replication strand asymmetry per signature and cancer type (z-score normalised per sample). Red represents matching strand asymmetry between signature and sample, blue represents inverse asymmetry. Only significant values are shown (non-significant are in grey). Asterisks represent values that also pass Bonferroni correction for multiple testing.

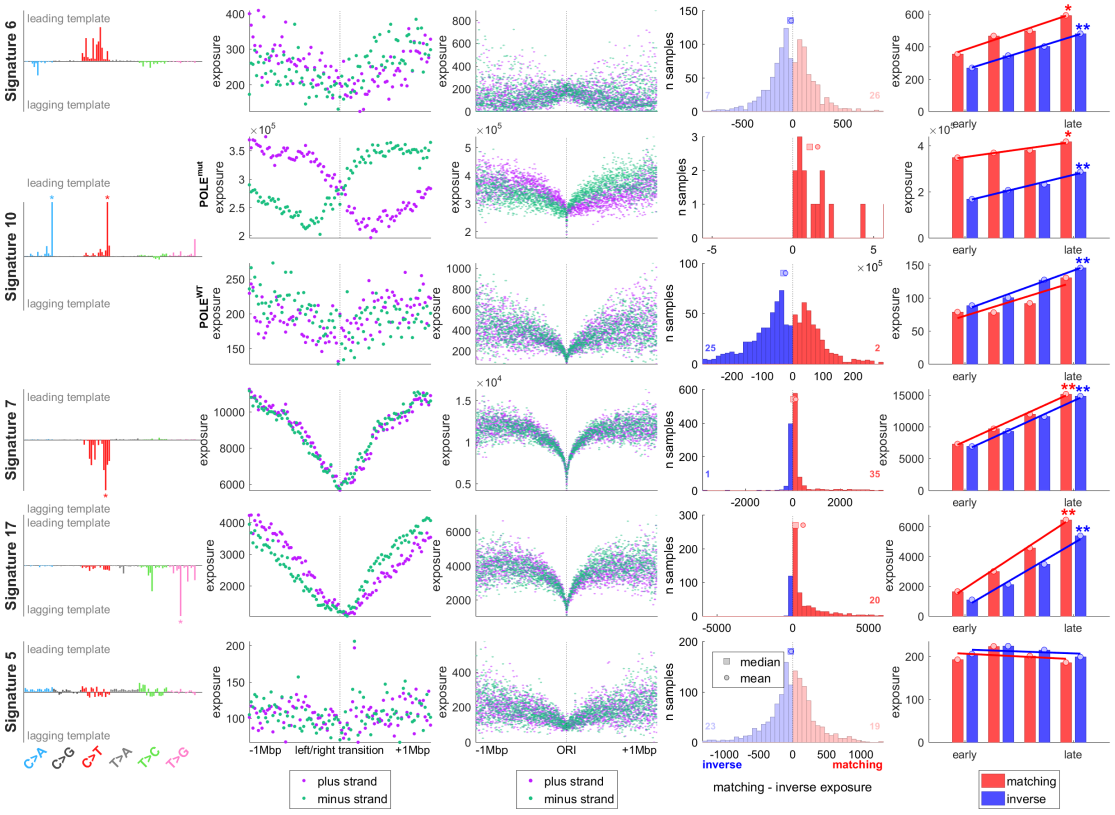

Figure 5.6. Different mutational signatures exhibit characteristic timing and strand asymmetry profiles. Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig 5.3. Row 1: Signature 6, associated with mismatch-repair deficiency. Row 2–3: signature 10, associated with deficiency in Pol ε proofreading, shown for patients with known *POLE* mutations (row 2), and those without (row 3). Row 4: signature 7, representing UV-induced damage. Row 5: signature 17, characteristic of gastric and oesophageal cancers. Row 6: Signature 5, of unknown aetiology, is not discernibly affected by replication.

Including protein coding genes did not qualitatively change the results (Supplementary Fig. 10.16,10.18), nor did the exclusion of non-coding in addition to protein-coding genes (Supplementary Fig. 10.17,10.19). Moreover, using SNS-seq data to determine replication strand direction leads to highly similar findings (Fig. 5.7 and Supplementary Fig. 10.20), confirming that the results are not specific to one tissue type or one used replication direction map.

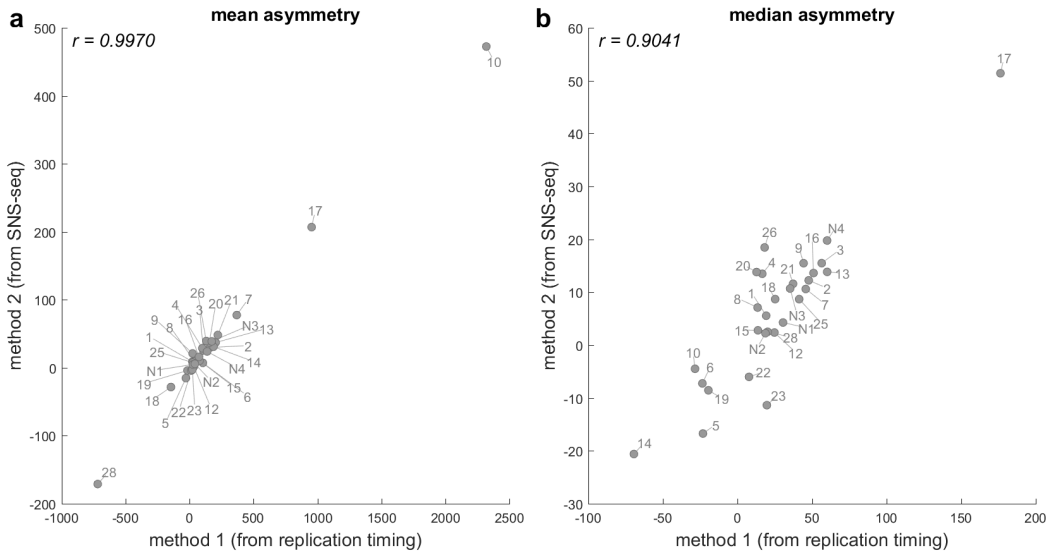

Figure 5.7. Both methods of estimating direction of replication result in very similar mutation strand asymmetries. Comparison of resulting mean (a) and median (b) replication strand asymmetry per signature in the two methods of measuring replication direction: from replication timing (20 kbp bins annotated as in (Haradhvala et al., 2016) vs. from measurements of ORIs using SNS-seq (1 kbp bins, see Methods). The absolute values of exposures are different between the two methods since regions around ORIs cover fewer bases (and therefore also fewer mutations).

The next sections contain detailed results of selected signatures, while results of other signatures are shown in 10.3.1.

5.3.1 Signatures associated with MMR

In MSI samples, all 5 signatures previously associated with MMR (signatures 6, 15, 20, 21, 26) and the novel N4 exhibit replication strand asymmetry, generally with enrichment of C>T mutations on the leading strand template and C>A and T>C mutations on the lagging strand template (Fig. 5.8). This is in line with the previous observation that MSI samples show a strand asymmetry (Haradhvala et al., 2016), extending the knowledge that it is a property of all the known mutational signatures associated with MMR.

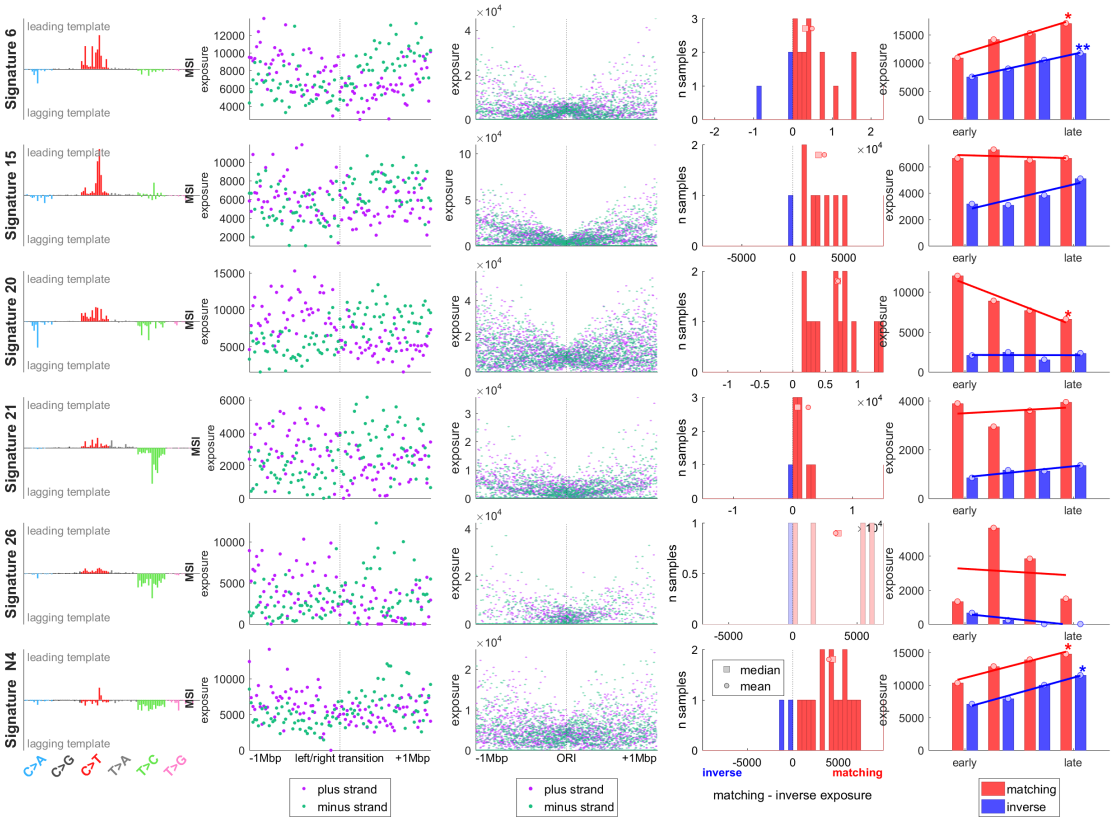

Figure 5.8. Replication strand asymmetry and replication timing in MMR signatures in MSI samples. Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig. 5.3.

It has previously been proposed that the correlation of overall mutation rate with replication timing (as shown in Fig 5.4B) is a direct result of the activity of MMR, because this correlation is diminished in MSI samples (Supek and Lehner, 2015). In contrast, we observed a more complex relationship. Some MMR signatures in MSI samples do not correlate with replication timing (signatures 15, 21, 26) or do so only in one direction of replication (such as the negative correlation in the leading direction in signature 20), whereas others show a steady significant correlation (signatures 6 and N4, Fig. 5.8), indicating that MMR might be only one of several factors influencing mutagenesis in a timing-dependent manner.

Unexpectedly, two MMR signatures (signatures 6 and N4) showed increased exposures around ORIs (Fig. 5.8). This increase was significant in both of these signatures (p-value $7.6 \cdot 10^{-6}$ in signature 6, p-value $1.4 \cdot 10^{-4}$ in signature N4; Fig. 5.9).

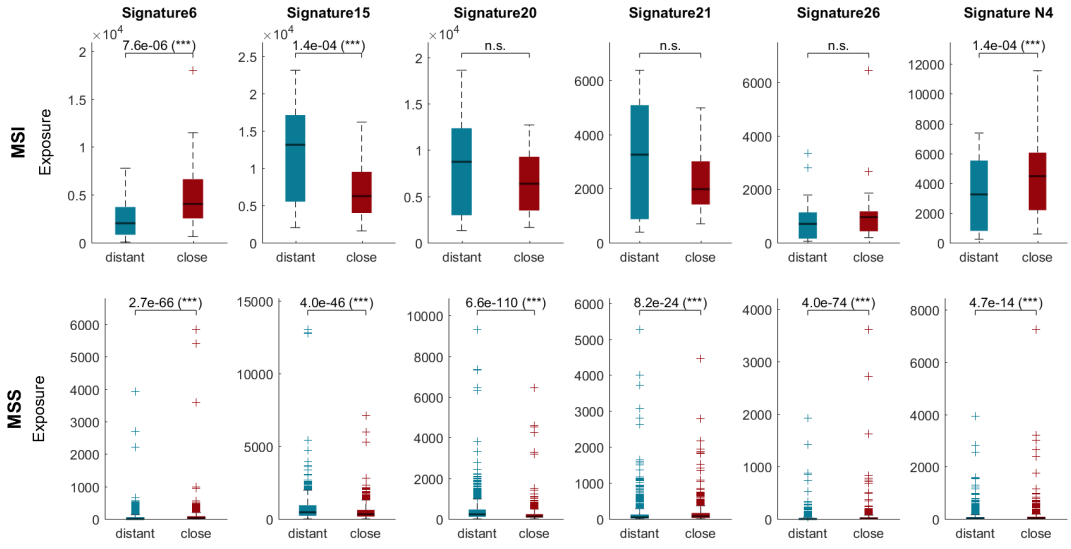

Figure 5.9. MMR signatures 6 and N4 have increased exposures around ORIs both in MSI and MSS samples. Based on SNS-seq of ORI, exposures to signatures were compared in regions close to ORI (at most 250 kbp) and distant from ORI (between 500 kbp and 1 Mbp). The difference was evaluated with signtest.

Notably, we also detected weaker but still significant exposure to MMR signatures in samples with seemingly intact mismatch repair (Fig. 5.10). Replication strand asymmetry in these samples was substantially smaller, but the higher exposure to signatures 6 and N4 around ORIs remained (Fig. 5.9).

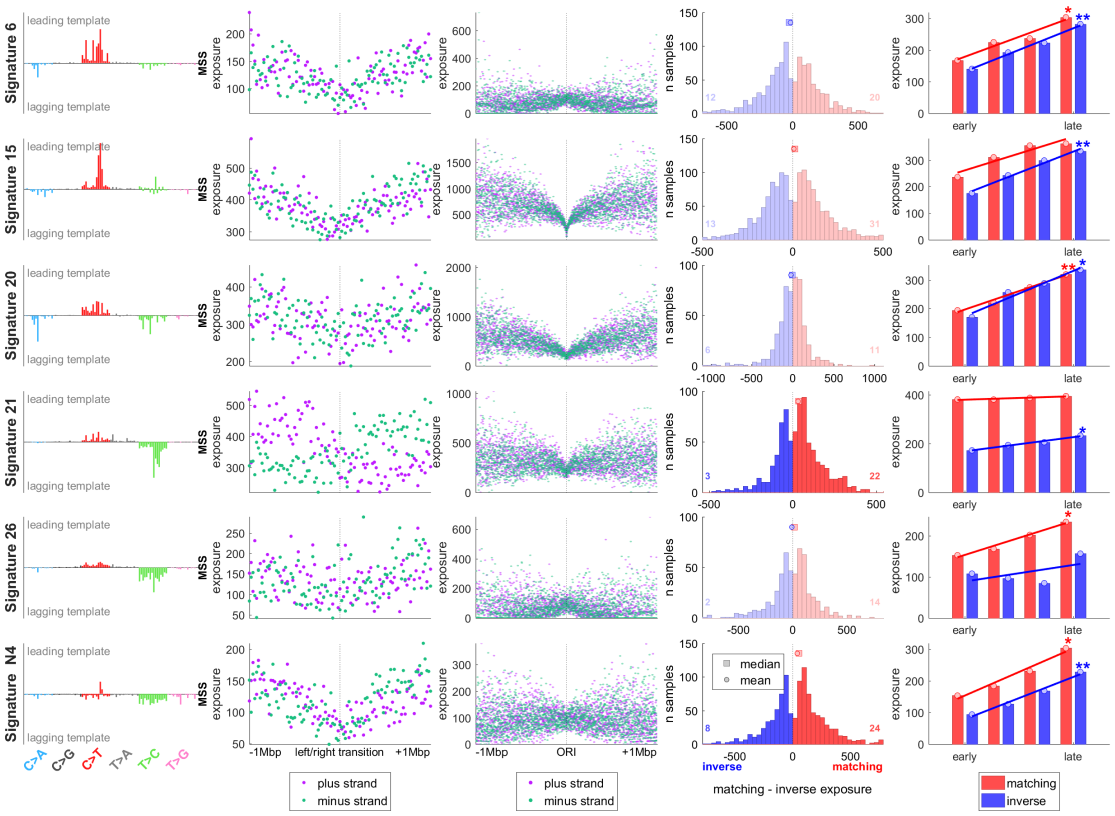

Figure 5.10. Replication strand asymmetry and replication timing in MMR signatures in microsatellite stable samples (MSS). Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig. 5.3.

5.3.2 Signatures associated with Pol ε

POLE-MUT samples were previously reported to be "ultra-hypermutated" with excessive C>A and C>T mutations on the leading strand (Haradhvala et al., 2016; Shinbrot et al., 2014; Shlien et al., 2015). Two mutational signatures (10 and 14) have been associated with Pol ε, the main leading strand polymerase (Stillman, 2008; Georgescu et al., 2015). As expected, we observed very strong strand asymmetry for these two signatures in all *POLE*-MUT samples, with an increase of C>A, C>T, and T>G mutations on the leading strand (Fig. 5.11), extending the results about a leading strand enrichment of NCG>NTG observed in the previous chapter (section 4.3).

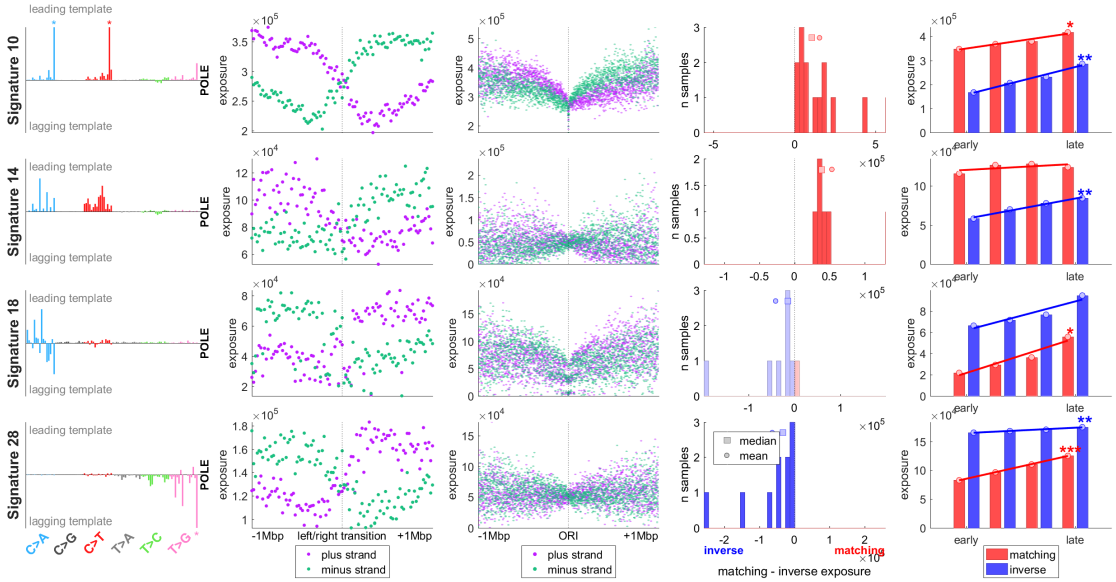

Figure 5.11. Replication strand asymmetry and replication timing in signatures detected in *POLE*-MUT samples. Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig. 5.3.

As with MMR signatures, we also found weak but significant evidence of signature 10 and 14 in samples without Pol ε defects (*POLE*-WT). Strikingly, however, in these samples the strand asymmetry was in the inverse orientation compared to the *POLE*-MUT samples, i.e., more C>A, C>T, and T>G mutations on the lagging strand (Fig. 5.12). Conversely, we detected the presence of two signatures of unknown aetiology, signatures 18 and 28, in *POLE*-MUT samples, but in the inverse orientation compared to *POLE*-WT samples.

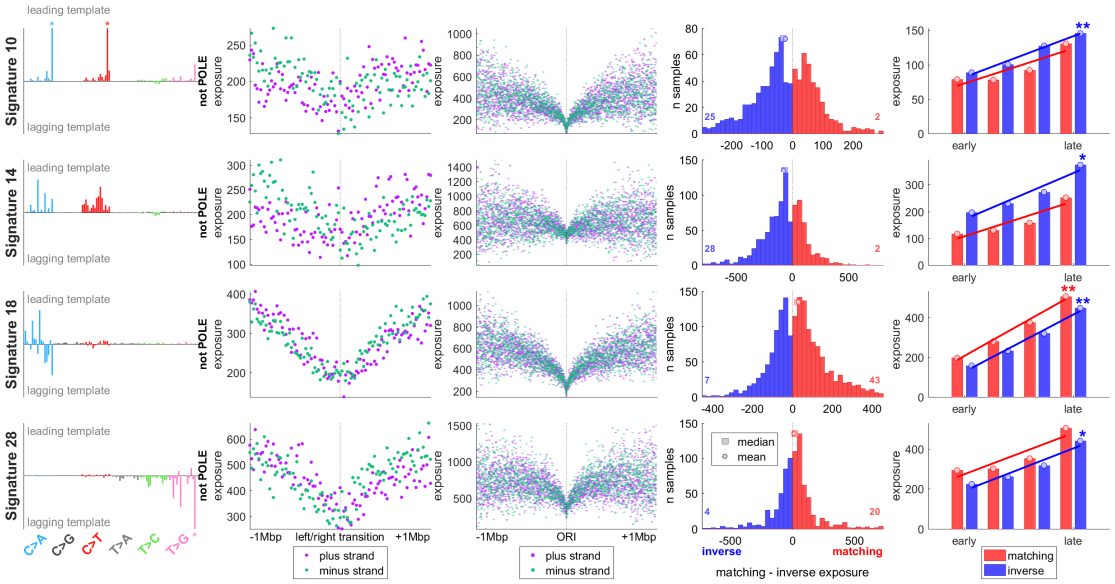

Figure 5.12. Replication strand asymmetry and replication timing in signatures 10, 14, 18, and 28, in *POLE*-WT samples. Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig. 5.3.

In order to validate that this is not an artefact of the signature exposures decomposition, we directly compared the frequencies of the most prominent mutation types for each of the four signatures (signatures 10, 14, 18, and 28) in *POLE*-MUT and *POLE*-WT samples on the leading and lagging strands. The inverse strand preference observed in the signatures was also detected for individual mutation types (Fig. 5.13, 10.21, 10.22, 10.23). For example, the frequency of mutations in TCT>TAT, TCG>TTG, and TTT>TGT, the three major components of signature 10, is higher on the lagging strand than on the leading strand in *POLE*-WT samples, whereas it is higher on the leading strand in *POLE*-MUT (Fig. 5.13).

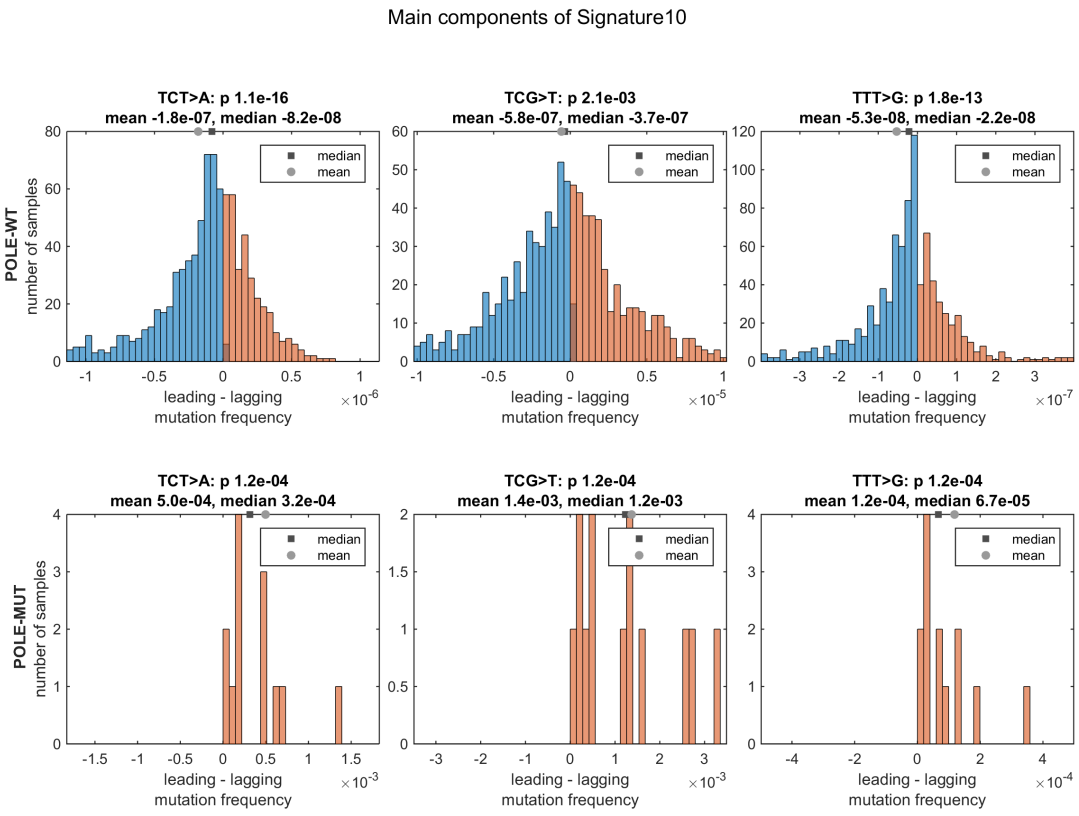

Figure 5.13. Inverse exposure of signature 10 in *POLE*-MUT vs. *POLE*-WT samples. Frequency of mutations in TCT>TAT, TCG>TTG, and TTT>TGT, the three major components of signature 10, is higher on the lagging strand than on the leading strand in *POLE*-WT samples, whereas it is higher on the leading strand in *POLE*-MUT. Only samples exposed to signature 10 (exposure above 10) are shown. Signtest was used to evaluate the mutation frequency difference between the leading and lagging strands.

5.3.3 Signatures due to environmental mutagens

We next focused on signatures that have not previously been reported to be connected to replication, or for which the causal mechanism is unknown. From the signatures present in the analysed samples, three mutational signatures have a known link to external mutagen: UV light (signature 7), tobacco smoke (signature 4), and aristolochic acid (AA) (signature 22) (Helleday et al., 2014).

All three environmental signatures 4, 7, and 22 show a strong significant correlation with replication timing (Fig. 5.14 and 10.24). We also observed weak but significant replication strand asymmetry in the mutagen-induced signatures in the tissues associated with the respective mutagen (Fig. 5.14). Interestingly, in all three cases the enrichment of mutations corresponds to the damaged DNA being placed on the lagging strand: adduct on guanine in signature 4, adduct on adenine in signature 22, and covalently linked cytosine with another neighbouring pyrimidine in signature 7.

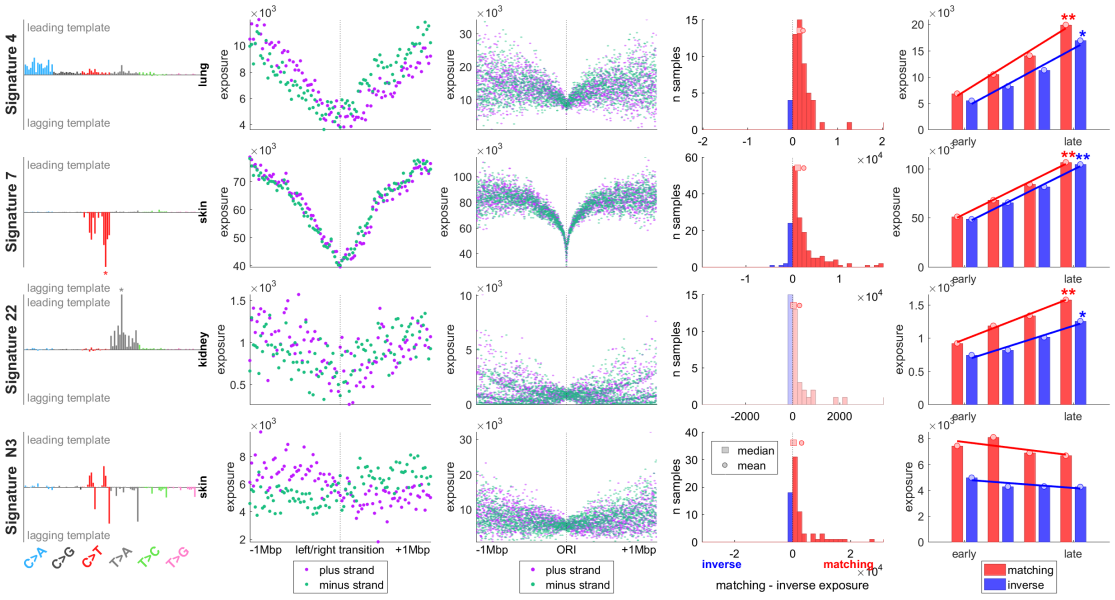

Figure 5.14. Replication strand asymmetry and replication timing in mutagen signatures: (4 in lung cancer samples, 7 in skin cancer samples, 22 in kidney cancer samples) and signature N3 of unknown aetiology (skin cancer samples). The format is as described in Fig. 5.3.

5.3.4 Signature 17

Signature 17 exhibits the largest median strand asymmetry (176.3, p-value < 10^{-59}), highest consistency across samples (81.8 % samples with matching exposure), the strongest ($\log_2$ fold-change 2.25 from early to late; p-value < 10^{-57}) and the most consistent (40 % positive and 0 % negative) correlation with replication timing from all explored mutational signatures (Fig. 5.4). In spite of this list of primacies, the relative increase of the matching compared to inverse exposure was relatively low (Fig. 5.15), compared to the other strongly asymmetrical signatures, such as signatures 10 or 13. This suggests that the process generating signature 17 is present on both strands, but one of the strands is faster repaired or slower fixated into mutations.

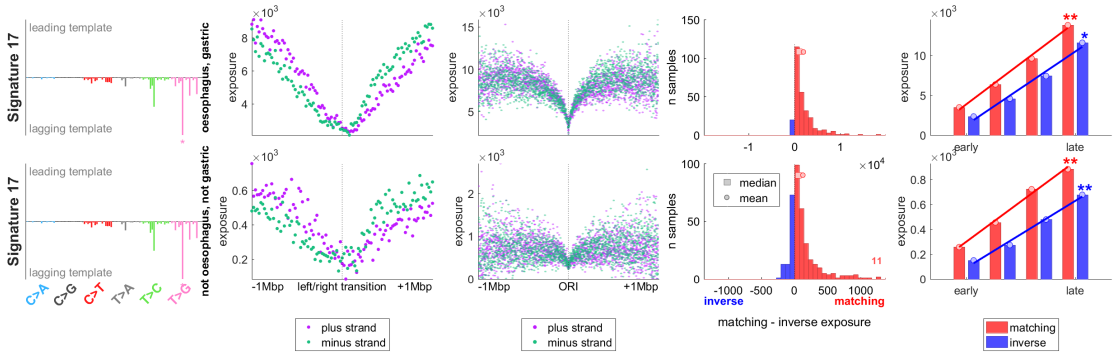

Figure 5.15. Replication strand asymmetry in signature 17. Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig. 5.3. Top row: EAC and gastric samples, bottom row: other samples.

This is the first time this signature is linked to DNA replication. Although it consists of very specific types of mutations (with a strong CTT>CGT component, accompanied by other NTT>NGT and CTT>CNT mutations) and relatively specific occurrence in gastric cancers and oesophageal adenocarcinoma (EAC), the mutational process causing this signature remains elusive (Wellcome Trust Sanger Institute, 2017).

Interestingly, signature 17 is present both in MSS and MSI samples, but it is enriched in MSS samples (Fig. 5.16A, (Wang et al., 2014)). This is in contrast with the other signatures in MSI samples, which are 1–2 orders of magnitude higher in MSI than MSS samples (Fig. 5.8 and 5.10). The difference of exposure to signature 17 between MSS and MSI samples is largest in early replicating regions and the difference decreases with the increasing replication timing (Fig. 5.16B).

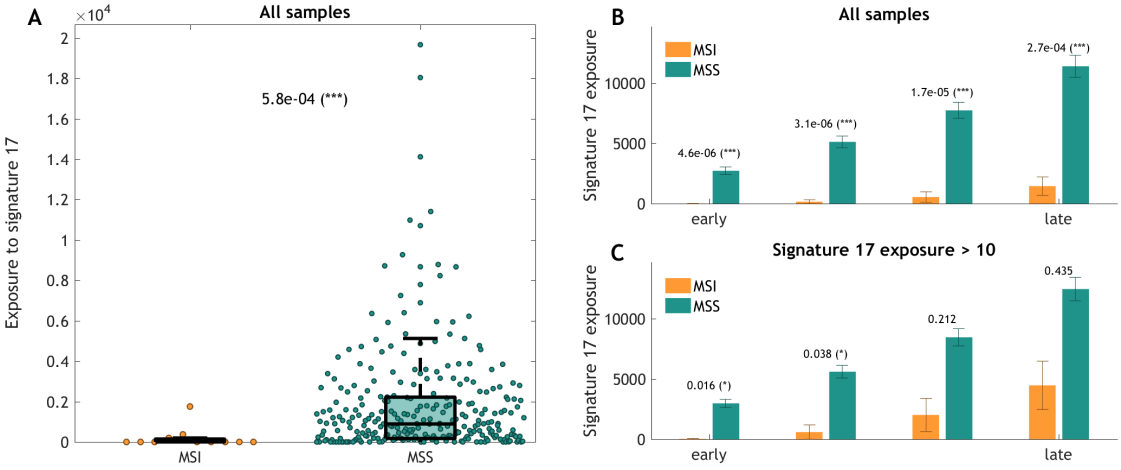

Figure 5.16. Signature 17 is increased in MSS compared to MSI gastric and EAC cancers. A: distribution of signature 17 in MSS vs. MSI gastric cancers (ranksum test). B: Replication strand asymmetry in the four quartiles of replication timing. **C:** As in (B) but only for samples sufficiently exposed to signature 17 (exposure > 10).

The most characteristic cancer type for signature 17 is EAC. Sequencing of EAC and its precursor Barrett's oesophagus (BO) showed that signature 17 is present already in the benign lesion (Ross-Innes et al., 2015). In additional exploratory analyses, we noticed that signature 17 is significantly enriched in EAC samples with a history of BO compared to EAC samples without known BO history (Fig. 5.17). Also samples with reported acid reflux exhibited increased signature 17, but interestingly the grouping by BO history resulted in larger difference.

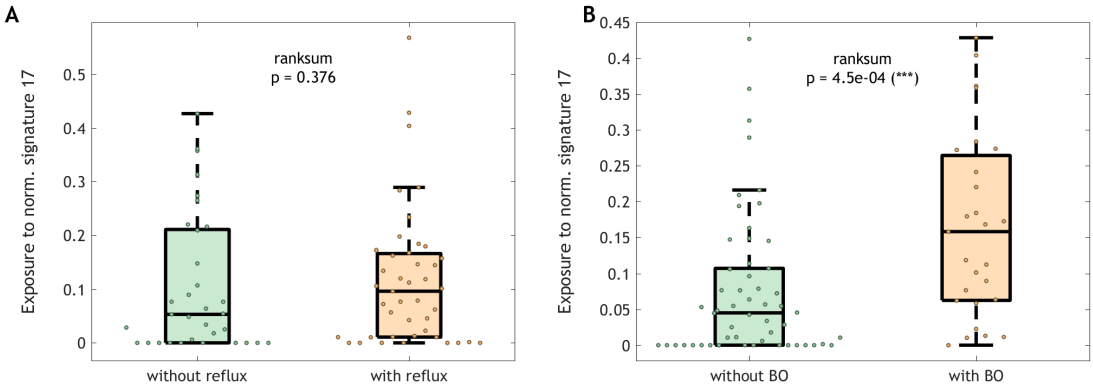

Figure 5.17. Signature 17 is increased in EAC patients with a history of Barrett's oesophagus. Exposure to signature 17 (normalised by mean mutation frequency) in adenocarcinoma patients from TCGA ESCA. **A:** Comparison of patients with and without reflux. **B:** Comparison of patients with and without a history of Barrett's oesophagus.

We next compared survival of samples with low and high exposure to signature 17 in this cohort, as tumours with high contribution of this signature showed previously (in a different cohort) a trend towards poor survival (Nones et al., 2015). Also in the TCGA cohort, samples with high exposure exhibited worse survival (logrank p-value 0.012) than samples less exposed to signature 17 (Fig. 5.18A) and this difference was even larger for samples with a history of BO (but less significant and limited by the samples size; p-value 0.032; Fig. 5.18B).

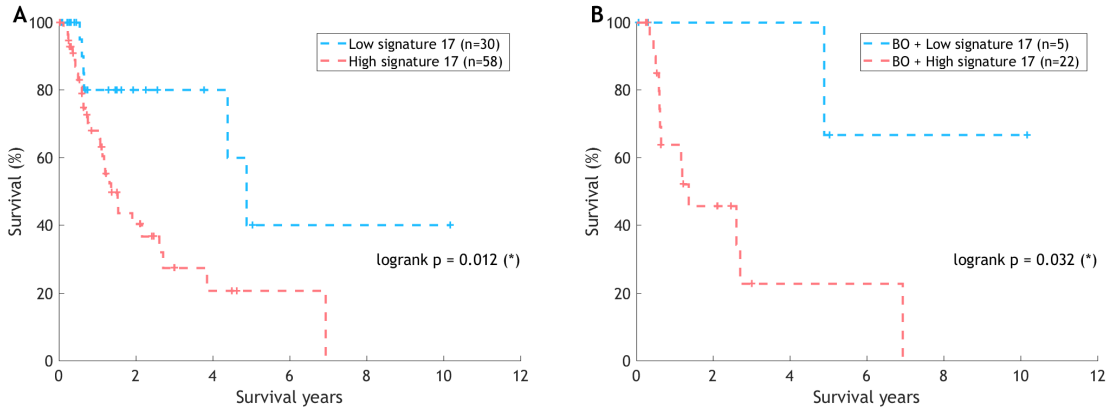

Figure 5.18. High signature 17 is associated with shortened survival. Kaplan-Meier survival analysis of adenocarcinoma patients from TCGA ESCA. **A:** Comparison of patients with low and high signature 17 (normalised exposure cutoff = 0.0285). **B:** As in (A), but only for patients with a history of Barrett's oesophagus.

Inspired by the nucleosomal influence on UV-induced mutagenesis in the previous chapter (section 4.4.4), we computed the frequency of CTT>CGT mutations (the major mutation type in signature 17) with respect to the distance of the nearest nucleosome dyad[1]. The mutation frequency exhibited a strikingly periodic signal around the nucleosome dyad with an increase towards the dyad. The signal was different in the two strands, shifted by approximately a half of the period and increased in the minus strand left to the dyad and in the plus strand right to the dyad (Fig. 5.19).

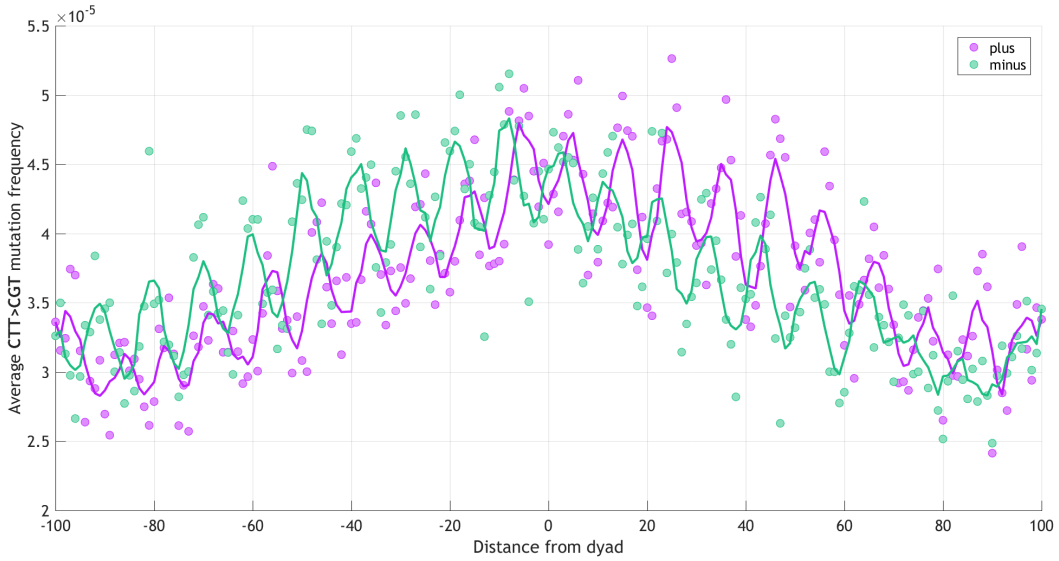

Figure 5.19. CTT>CGT, the main component of signature 17, around a nucleosome dyad. Frequency of T>G mutations in a CTT context with respect to distance from a nucleosome dyad, computed separately for Watson (+) and Crick (-) strands; average of 213 WGS ICGC EAC samples.

[1]We performed this analysis in an unbiased pan-cancer pan-signature approach, but the scope and length limitations of this thesis do not allow to include a proper description of all the results. However, even in this analysis, signature 17 (in all major components) was one of the most affected signatures. Compared to the replication asymmetry results, many signatures were not strongly influenced by the nucleosome positioning.

5.4 Discussion

In this chapter we showed that a large number of mutational signatures exhibit footprints of influence by replication. We believe that the characteristics of the footprints can help to improve the understanding of the underlying mutational processes of the signatures. Examples of such synthesis of observations from the replication analysis and literature are shown below.

5.4.1 Signatures associated with MMR

Single base changes in MMR-deficient patients have been previously shown to exhibit a replication strand asymmetry (Haradhvala et al., 2016). Here, we extend this knowledge by showing that the strand asymmetry is present in all MMR-associated SNV mutational signatures. In spite of the active research of MMR (Crouse, 2016; Zhao et al., 2014a; Cortes-Ciriano et al., 2017; Le et al., 2017), the exact mechanisms causing these signatures are poorly understood.

In two of the signatures, we observe an increase around ORIs. Based on experiments in yeast, it has been suggested that MMR is involved in balancing the differences in fidelity of the leading and lagging polymerases (Lujan et al., 2012). This balancing is the strongest for errors made by Pol α (Lujan et al., 2012), which primes the leading strand at ORIs and each Okazaki fragment (Stillman, 2008), and lacks intrinsic proofreading capabilities (McCulloch and Kunkel, 2008). It has been recently shown that error-prone Pol α-synthesised DNA is retained *in vivo*, causing an increase of mutations on the lagging strand (Reijns et al., 2015). Since regions around ORIs have a higher density of Pol α-synthesised DNA (as discussed e.g. in (Waisertreiger et al., 2012)), it is possible that increased exposure to signatures 6 and N4 around ORIs is caused by incomplete repair of Pol α-induced errors. Moreover, while Pol α-synthesised DNA in the Okazaki fragments is displaced by Pol δ (summarised in the Introduction 1.1.3), we did not find any reports studying a displacement of Pol α-synthesised DNA at the replication origins, leaving a possibility that these small regions are remained in the DNA. Finally, the most common Pol α-induced mismatches normally repaired by MMR are G-dT and

C-dT, leading to C>T mutations on the leading and C>A mutations on the lagging strand (Nick McElhinny et al., 2010), matching our observations in the MMR-linked signatures.

We detected a weak exposure to MMR signatures also in MSS samples, but without a significant asymmetry (signatures 6, 15, 20, and 26) or with only a mild asymmetry (in the terms of small increase in the median of samples compared to zero; signatures 21 and N4) (Fig. 5.10). This could be explained by a small amount of errors which escaped the correction by MMR, with a similarly small frequency on the two strands. Most of the MMR signatures in MSS samples were increased in late-replicating regions, in line with the assumed enrichment of MMR activity in the early-replicated regions (Supek and Lehner, 2015; Stamatoyannopoulos et al., 2009; Chen et al., 2010) and therefore greater probability to escape MMR repair in the late-replicated regions.

The increased exposure to signatures 6 and N4 around replication origins remained also in the MSS samples (Fig. 5.9), in line with the explanation that this increase is not caused by the MMR itself, but by a different process, which generates the — normally MMR-corrected— mutations enriched around replication origins, such as the hypothesised Pol α-synthesised DNA. The fact that other MMR signatures (signature 15, 20, 21, and possibly 26) did not show this increase around ORIs could be due to other roles of MMR, such as repairing errors made by Pol δ (Andrianova et al., 2017).

In summary, our results support a model, in which mismatch repair balances the effect of mis-incorporation of nucleotides by Pol α and escape or deficiency of such repair leads to MMR-associated signatures (possibly 6 and N4).

5.4.2 Signatures associated with Pol ε

Although the ultra-hypermutability of *POLE*-MUT samples has attracted a lot of attention in the recent years (Shinbrot et al., 2014; Palles et al., 2013; Shlien et al., 2015; Rayner et al., 2016), the mechanisms of the hypermutated phenotype are still unclear (Ganai and Johansson, 2016; Barbari and Shcherbakova, 2017; Mertz et al., 2017a). The current view is that other factors than a simple nucleotide misincorporation that escapes proofreading might play a role in the *POLE*-MUT mutagenesis, such as

altered DNA binding (Barbari and Shcherbakova, 2017), or imbalanced dNTP pools (Mertz et al., 2015; Williams et al., 2015).

We observed a strong replication strand asymmetry in all *POLE*-MUT samples and all signatures present in the *POLE*-MUT samples, in line with the major role of Pol ε in the leading-strand synthesis. Interestingly, we observed a weak exposure to the *POLE*-associated signatures also in *POLE*-WT tumours, but enhanced on the opposite strand. In fact, all the four signatures with a dominantly matching exposure in one cancer type and inverse exposure in another cancer type, contained the *POLE*-MUT cohort as one of the cancer types. This suggests that the *POLE*-linked signatures are originally caused by a process that affects both strands, and under normal circumstances is slightly enriched on the lagging strand. Such an output could be caused by certain types of DNA lesions which under normal circumstances are less accurately replicated when on the template of the lagging strand, e.g. due to a lower fidelity of Pol δ or Pol α compared to WT Pol ε when replicating these lesions. In *POLE*-MUT samples the lack of replication-associated proofreading would then lead to a strong relative increase in these mutations on the leading strand, explaining the flipped orientation of signatures.

The nature of which types of DNA damage could underlie these observations is unknown. However, for example bypass of AP sites was observed to be enhanced by suppression of exonuclease proofreading of Pol ε in human cells, and the exonuclease inactivation led to decreased sensitivity to H_2O_2 (Henninger, 2015; Henninger et al., 2015). One source of AP sites is depurination of guanine. While REV1-mediated insertion of C opposite the AP site would lead to restoration of the original C:G pair, insertion of A according to the A-rule would lead to a C:G>A:T mutations. It is theoretically possible that the latter is slightly more common on the lagging strand in *POLE*-WT cells, while it is markedly increased on the leading strand in *POLE*-MUT samples, as C>A mutations form a large proportion of signatures 10, 14, and 18 and are markedly enriched on the leading strand in *POLE*-MUT samples.

5.4.3 Signatures due to environmental mutagens

We observed a strong enrichment of signatures linked to environmental mutagens (signatures 4, 7, and 22) in the late-replicated regions and a small but significant replication strand asymmetry.

Previously, higher mutation frequency in late-replicating regions has been observed in mouse embryonic fibroblast (MEFs) treated with AA or Benzo[a]pyrene (B[a]P, a mutagen in tobacco smoke) (Nik-Zainal et al., 2015). Differences in chromatin accessibility could be responsible for the decreased mutagenicity in early-replicated regions. Open chromatin is on average replicated earlier and is also more accessible to repair enzymes which could contribute to the decreased mutation frequency in early-replicating regions (Adar et al., 2016).

Alternatively, this increased mutagenicity in late-replicating regions could be due to differences of DNA damage tolerance pathways active during early and late replication. Regions replicated early in S-phase are thought to prefer high-fidelity template switching, whereas regions replicated late are more likely to require translesion synthesis (TLS) which has a higher error rate (Waters and Walker, 2006; Lang and Murray, 2011; Karras et al., 2013; Gonzalez-Huici et al., 2014; Bi, 2015; Branzei and Szakal, 2016b; D'Souza et al., 2016). This is consistent with the observation in yeast that a disruption of TLS leads to decreased mutation frequency in late-replicating regions and therefore a more even distribution of mutation frequency between early and late-replicating regions (Lang and Murray, 2011). In particular, TLS has been observed to increase in activity and mutagenicity later in the cell cycle when replicating DNA damaged by B[a]P (Diamant et al., 2012).

The reason for the observed replication strand asymmetry is currently unknown. However, it matches a previously observed lower efficiency of bypass of DNA damage on the lagging strand (Cordeiro-Stone and Nikolaishvili-Feinberg, 2002) and a strong mutational strand asymmetry in cells lacking Pol η, the main TLS polymerase responsible for the replication of UV-induced photolesions (McGregor et al., 1999). Altogether, our data highlight the importance of replication in converting DNA damage into actual

mutations and suggest that bypass of DNA damage occurring on the lagging template results in detectably lower fidelity on this strand.

We also detected a novel mutational signature in melanoma cancers. Compared to the canonical skin signature 7, this novel signature N3 was very strongly asymmetrical between the two strands (both in the terms of size effect and consistency across samples) but uncorrelated with replication timing. As this signature consists mostly of mutations in T-containing dipyrimidines, it might be caused by erroneous replication of TT-CPDs or 6-4-TTs. The most frequent form of CPDs are TT-CPDs (Bryan et al., 2014), but they are very efficiently bypassed by Pol η in an error-free manner (Silverstein et al., 2010; Pfeifer and Besaratinia, 2012). However, TT-CPDs can be also bypassed by other polymerases, inserting T opposite the 3' TT (Wang et al., 2007), which could be a potential cause of the high TTT>TAT peak in signature N3. Moreover, Pol δ was shown to be able to bypass TT-CPD, causing T>A mutations, both *in vitro* and *in vivo* (Narita et al., 2010; Hirota et al., 2016). This suggests an interesting direction of future investigations, to determine whether the large enrichment of the TTT>TAT mutations on the lagging strand in a subgroup of melanoma cancers could be related to the TLS activity of Pol δ.

5.4.4 Signature 17

Signature 17 contained the highest percentage of samples with replication asymmetry and correlation with replication timing (Fig. 5.4). We also noticed that the timing asymmetry and exposure distribution around replication origins (Fig. 5.15) closely resemble that of the signatures of external mutagens (signatures 4 and 7; Fig. 5.14), suggesting a possible involvement of DNA damage in signature 17. We therefore explored the existing literature linking the signature 17 to processes potentially causing damage to the DNA.

5.4.4.1 Importance of acid, bile, and oxidative damage in EAC development

Signature 17 appears early during EAC development (Murugaesu et al., 2015), and it is also present in BO, a precursor to EAC (Ross-Innes et al., 2015). Gastro-esophageal reflux disease (GERD) is a key risk factor for EAC and BO (Erichsen et al., 2012; Schlottmann

et al., 2017). Although the symptoms of GERD have been treated with proton pump inhibitors in the last 30 years, it failed to produce a positive effect on the incidence of EAC, which has been one of the fastest rising cancers during that time (Pohl and Welch, 2005; Schlottmann et al., 2017). Moreover, both beneficial and detrimental impact of long-term use of proton pump inhibitors on BO and EAC have been observed (Hvid-Jensen et al., 2014; Hayakawa et al., 2016).

In addition to GERD, BO is also associated with increased duodeno-gastric reflux, and thus higher exposure to bile (Bernstein et al., 2005; Souza, 2010). Patients with increased duodeno-gastric reflux are more likely to have oesophagitis and BO (Fein et al., 2006). Expression of bile acid transporter proteins is increased in BO but decreases with progression to cancer, suggesting an adaptive mechanism in BO to protect cells from bile acids, which is gradually lost as BO progresses to EAC (Dvorak et al., 2009).

BO and EAC cells exhibit increased oxidative DNA damage and production of reactive oxygen species (Sihvo et al., 2002; Bernstein et al., 2005; Jimenez et al., 2005). In particular, significantly higher levels of 8-oxo-7,8-dihydro-2'-deoxyguanosine (8-oxo-dG) were found in BO and EAC and gastric adenocarcinoma, compared to normal tissue (Rasanen et al., 2007; Dvorak et al., 2007; Lagadu et al., 2010; Borrego et al., 2013; Kauppi et al., 2016). Interestingly, the 8-oxo-dG levels are highest in the stage of low-grade dysplasia and gradually decrease in the higher stages: high-grade dysplasia and EAC (Rasanen et al., 2007; Dvorak et al., 2007). Several studies show a causal link between acids and bile from the reflux and the observed oxidative damage (Bernstein et al., 2005; Jenkins et al., 2007; Bonde et al., 2007). Specifically, exposure to pH 4 and bile acid cocktail together (but not separately) leads to an increase of 8-oxo-dG in BO biopsies and oesophageal cells (Dvorak et al., 2007).

5.4.4.2 Incorporation of 8-oxo-dGTP into DNA by TLS causes T>G mutations

As described in the Introduction (sections 1.3.3.3 and 1.3.3.4), 8-oxoG, one of the most common oxidative DNA lesions, can cause C:G>A:T mutations (when guanine in the DNA is oxidised) or T:A>G:C mutations (when guanine precursor in the dNTP pool is oxidised and incorporated into DNA). In particular, Pol η (together with Rev1, and Pol

ζ) were shown to incorporate 8-oxo-dGTP opposite A, causing T:A>G:C mutations (Satou et al., 2009).

5.4.4.3 Strand asymmetric bypass of 8-oxo-dGTP

Importantly, the mismatch of 8-oxoG and A has been shown in yeast to be more efficiently repaired into G:C when 8-oxoG is on the lagging strand template (Pavlov et al., 2002) by two independent mechanisms: more efficient MMR repair of incorporated A opposite the template 8-oxoG (Pavlov et al., 2003) and by more efficient bypass activity of Pol η on the lagging strand (Mudrak et al., 2009). Oxidation of guanine on the DNA would thus lead to increased C>A mutations on the lagging strand (i.e., 8-oxoG on the leading strand). However, incorporation of 8-oxo-dGTP opposite adenine also leads to 8-oxoG:A pair. In the next replication, incorporation of cytosine opposite the template 8-oxoG, leads to a T:A>G:C mutation. The asymmetric repair/bypass of 8-oxoG would thus lead to an enrichment of T>G mutations on the lagging strand template (Fig. 5.20).

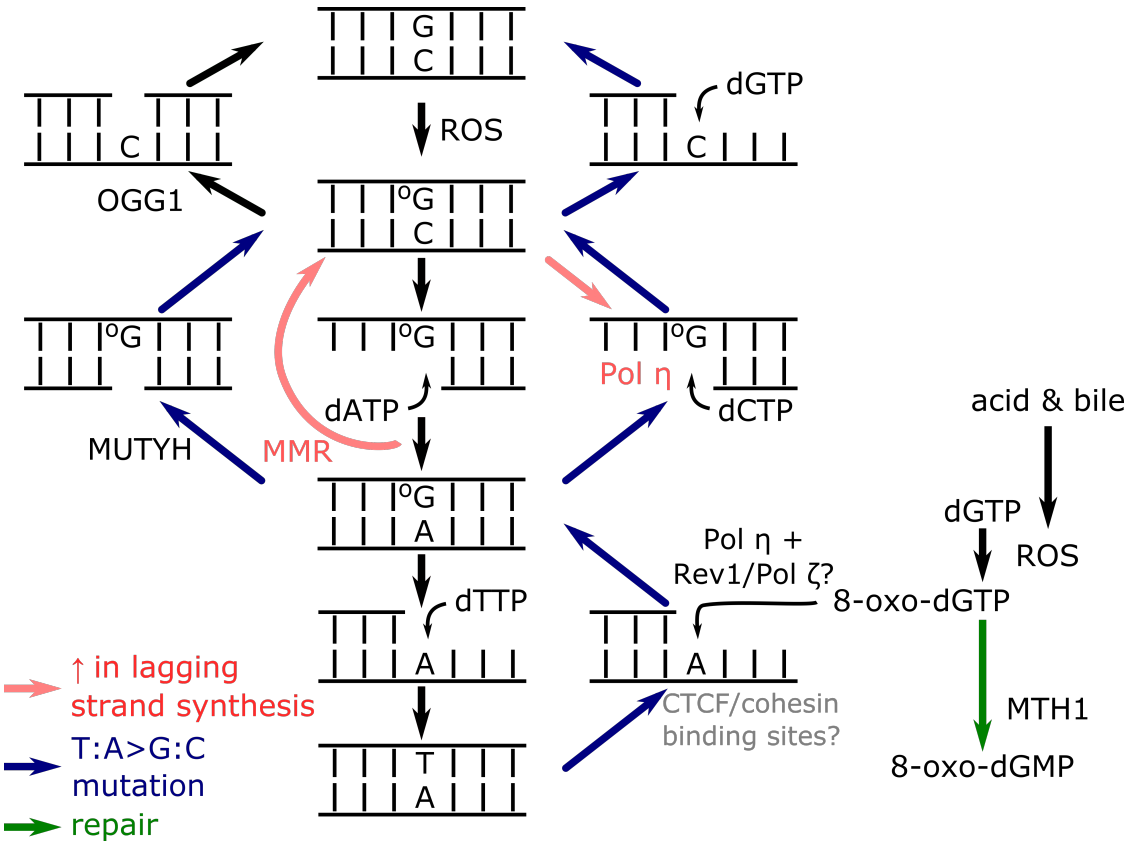

Figure 5.20. A model of oxidative damage that could cause signature 17.

5.4.4.4 Links of signature 17 with 8-oxoG

Our data show a strong lagging-strand bias of T>G mutations and overall higher exposure to signature 17 on the lagging strand, supporting the hypothesis that signature 17 is a by-product of oxidative damage. The correlation of signature 17 with replication timing would be in line with increased TLS in late-replicated regions. We observe an enrichment of signature 17 in EAC patients with acid reflux, an even stronger association with history of BO, in line with the suggested role of a combination of acid and bile reflux in the pathogenesis of BO. Moreover, a decrease of both the signature 17 (Murugaesu et al., 2015; Stachler et al., 2015) and 8-oxoG levels (Rasanen et al., 2007; Dvorak et al., 2007) were observed during the progression of BO, both possibly resulting from a treatment of reflux.

Signature 17 was observed also in other tissues, although with a lower frequency than in EAC and BO. The highest frequency was observed in gastric adenocarcinoma, the cancer type molecularly most similar to EAC (Kim et al., 2017). Moreover, the mutation types corresponding to signature 17 were higher in gastric MSS tumours arising from the antrum compared to other locations, but did not show any association with *Helicobacter pylori* infection status (Wang et al., 2014), suggesting that they are induced by other carcinogens in the antrum, such as the bile acids. The other cancer types with (less frequent) presence of signature 17 have a known component of oxidative stress: breast cancer, lung cancer, melanoma, B-cell lymphoma, pancreas (Ikehata and Ono, 2011; Peroja et al., 2012; Wang et al., 2016; Hecht et al., 2016; Martinez-Useros et al., 2017). Interestingly, signature 17 was also observed in *in vitro* expansion of cells from small-bowel organoids (Behjati et al., 2014), whole genome mutation profiles of clones derived from primary Hupki mouse embryo fibroblast (Nik-Zainal et al., 2015), and inflammatory bowel disease associated colorectal cancers (Robles et al., 2016), all of which could be influenced by oxidative stress (Parrinello et al., 2003; Colussi et al., 2002; Balmus et al., 2016; Rouhani et al., 2016; Moura et al., 2015; Pereira et al., 2016).

The role of MMR in the repair of oxidative damage is still under debate, but the current evidence suggests that it helps to prevent C>A mutations by detecting misincorporated A opposite template 8-oxoG, it helps to prevent indels resulting from

misincorporated 8-oxo-dGTP, but it does not prevent incorporation of 8-oxo-dGTP opposite A nor C (Larson et al., 2003; Macpherson et al., 2005; Russo et al., 2004). MMR would be therefore predicted to enhance fixation of the 8-oxo-dGTP opposite A into A:T>C:G mutations, via increased incorporation of C opposite 8-oxoG in the next round of replication. This is in line with our observation of decreased signature 17 in MMR-deficient MSI samples compared to MSS samples (Fig. 5.16). Moreover, this difference is largest in the early-replicated regions, in line with the assumed enriched activity of MMR early in the replication (Supek and Lehner, 2015; Stamatoyannopoulos et al., 2009; Chen et al., 2010).

Interestingly, we observed a strong influence of nucleosome positioning on the mutations of signature 17 in the two strands (Fig. 5.19). This observation supports the suggested involvement of DNA damage in the aetiology of signature 17, as the geometric orientation within the nucleosome core particle and other factors influence the efficiency of BER glycosylases to recognise and repair DNA damage (Olmon and Delaney, 2017; Bacolla et al., 2014; Menoni et al., 2012, 2017).

Both EAC and the subgroup of gastric cancers similar to EAC can be characterised by high chromosomal instability (Kim et al., 2017). Signature 17 could have two potential links with chromosomal instability. First, signature 17 was observed enriched in CTCF/cohesin binding sites (Katainen et al., 2015; Piraino and Furney, 2017), which are subject to double-strand breaks, and thereby a source of chromosomal instability (Canela et al., 2017). Moreover, Pol η promotes fragile site stability (Rey et al., 2009; Bergoglio et al., 2013) and was implicated in DNA synthesis in homology-directed repair of double-strand break (Buisson et al., 2014; Kawamoto et al., 2005; McIlwraith et al., 2005). These observations suggest a possibility that Pol η is recruited to the CTCF/cohesin binding sites to synthesise DNA during double-strand break, at the cost of introducing point mutations and misincorporation of 8-oxo-dGTP into the DNA (Fig. 5.20).

Second, the increased presence of 8-oxoG in the DNA, excised by MUTYH could lead to increased single-strand and double-strand breaks (the latter for proximal 8-oxoG

on opposite strands). It would be therefore interesting to explore whether single-/double-strand break repair inhibition (Srivastava and Raghavan, 2015; Helleday, 2011), potentially combined with MTH1 inhibition, could lead to an effective treatment in *MUTYH*-WT patients with high signature 17. Therapeutic potential of MTH1 inhibition has recently brought hopes (Gad et al., 2014; Huber et al., 2014) and hypes (Kettle et al., 2016; Ellermann et al., 2017). Along the research of how the individual inhibitors work (Wang et al., 2017b), how dNTP pool sanitation works (Rudd et al., 2016), and what are the implication of MTH1 inhibition for neurodegeneration (Nakabeppu et al., 2017; De Luca et al., 2008; Sheng et al., 2012), we propose that these activities should be complemented with a research of mutational signatures and identification of patients in which the incorporation of oxidised dNTPs is already present as a strong mutagen.

5.4.5 Concluding remarks

In conclusion, our findings demonstrate how the relationship between mutational signatures and DNA replication can help to illuminate the mechanisms underlying several currently unexplained mutational processes, as exemplified by Signature 17 in oesophageal cancer. Crucially, our computational analysis produces testable hypotheses which we anticipate to be experimentally validated in the future. Our results also add a new perspective to the recent debate regarding the correlation of tissue-specific cell division rates with cancer risk (Tomasetti and Vogelstein, 2015). It has been argued that this correlation is primarily attributable to "bad luck" in the form of random errors that are introduced during replication by DNA polymerases. However, the range of mutational signatures observed in cancer samples makes a purely replication-driven aetiology of cancer mutations unlikely (Gao et al., 2016; Crossan et al., 2015). Here, we show that most mutational signatures are themselves affected by DNA replication, including signatures linked to environmental mutagens. The presence of mutational signatures on the one hand and a strong relationship between replication and the risk of cancer on the other therefore need not be mutually exclusive. In summary, our results provide evidence that DNA replication interacts with most processes that introduce

mutations in the genome, suggesting that differences among DNA polymerases and post-replicative repair enzymes might play a larger part in the accumulation of mutations than previously appreciated.

Beneath the stains of time
The feelings disappear
You are someone else
I am still right here

— Nine Inch Nails, performed by Johnny Cash *Hurt*

Neste mě, neste, ptáci, do nebes
netřeba na jih, stačí Polárka
Neste mě, neste, ještě voní bez
tam, kde je Petřín a kde Vikárka, cigárka

— Zuzana Navarová *Do nebes*

6 Conclusion

6.1 Results summary

6.1.1 Aim 1, chapter 3

The first aim was to investigate how frequently hydroxymethylated positions are mutated in cancer. We integrated maps of 5mC and 5hmC in normal tissues with somatic mutations in cancer patients in the respective tissues. Since 5hmC has been shown to be most abundant in human brain (Li and Liu, 2011; Nestor et al., 2012), we have initially focussed on assessing the relationship between mutability and DNA modifications in brain cancers. Based on DNA sequencing data from five brain cancer types encompassing 665 patients, we show that the dominant mutational signature in brain cancers, CpG>TpG, is modulated by the modification state of cytosine. Strikingly, the CpG>TpG mutation frequency of 5hmC is reduced nearly two-fold compared to the methylated state. We find that the ratio of 5hmC to 5mC in genomic regions correlates with CpG>TpG mutation frequency even after accounting for confounding factors like gene density or CpG islands. When we expand our analysis to include mutations and 5hmC maps from kidney and myeloid lineage of blood cells, we observe a clear tissue-specific effect of 5hmC on mutagenicity. Finally, we measured global 5mC and 5hmC levels using a methodology of high accuracy in eight different human tissue types and show that reduced 5hmC levels associate with an increased proportion of

modified CpG>TpG mutations in cancers of the corresponding tissue. Together, our findings suggest that hydroxymethylation has a significant influence on the likelihood of mutations at CpG sites across many human tissue types.

A limitation of the used approach was that the modification maps were from different individuals than the cancer mutations (because individual-matched maps did not exist). However, since the time of these results being published (Tomkova et al., 2016), the same conclusions were confirmed in a single Glioblastoma patient using modification maps from a neighbouring healthy tissue of the same patient (Raiber et al., 2017).

6.1.2 Aim 2, chapter 4

The second aim was to explore the role of DNA modifications in other processes than spontaneous deamination; in particular we focused on mutational processes associated with replication, UV exposure, tobacco exposure, and APOBEC enzymes. The results suggest that all of these mutagenic processes are affected by DNA modifications. The correlation between mutation frequency and modification levels is in some cases positive (such as spontaneous deamination or tobacco-induced mutagenesis), in some cases negative (APOBECs-induced mutagenesis), and in other cases even non-linear (UV-induced mutagenesis; however the non-linearity seems to result from a positive correlation for 5mC and a strong negative correlation for 5hmC). Together with the results in the previous chapter, we show that individual DNA modifications can have different effects on the mutagenicity. Interestingly, 5hmC seems to be associated with decreased mutagenicity compared to 5mC in all these mutational processes; albeit this observation needs validation with high-quality tissue-matched 5hmC maps. Using consensus of the existing 5hmC maps predicts a particularly large decrease in UV-induced mutagenesis. In addition to DNA modifications, we observed that the C>T mutations in melanoma cancers are strongly affected by nucleosome positioning and exhibit an unexpected asymmetry on the two strands around the nucleosome dyad.

Finally, results in this chapter show a surprising link between DNA modifications and replication, affecting tumours with defective Pol ε or MMR. These tumours exhibit

extremely high numbers of mutations, which are thought to arise as errors during replication. It would be expected that these errors should quickly outnumber the mutations due to spontaneous deamination of 5mC. Contrary to this expectation, we observe that the frequency of CpG>TpG mutations in tumours with defective Pol ε or MMR is approximately six-fold higher than for other types of mutations. We show that the increased CpG>TpG mutation rate in Pol ε or MMR mutant cancers is linked to DNA methylation, has a clear replication strand asymmetry, being enriched on the leading strand, with a common preference for a GCG sequence context. We also detect a weaker but consistent replication strand asymmetry of GCG>GTG mutations in Pol ε and MMR proficient samples. Together, our results suggest that a substantial fraction of C>T mutations at methylated cytosines is independent of spontaneous deamination, instead arising during DNA replication. A surprising but theoretically possible explanation, most consistent with the observed data, is that the Pol ε has a decreased fidelity when replicating 5mC.

6.1.3 Aim 3, chapter 5

The third aim was to assess the role of DNA replication in individual mutational processes by analysing mutational signatures with respect to replication strand asymmetry and replication timing. We developed a method for quantification of replication strand asymmetry in individual samples and applied the method on 3056 WGS samples from 19 cancer types. We show that replication affects the distribution of most mutational signatures across the genome, including those that represent chemical mutagens. The unique strand-asymmetry and replication timing profiles of different signatures reveal novel aspects of the underlying mechanisms. For example, we discovered a strong lagging strand bias of T>G mutations in oesophageal adenocarcinoma, suggesting an involvement of oxidative damage to the nucleotide pool in the aetiology of the disease. Together, our results highlight the critical role of DNA replication and the associated repair in the accumulation of somatic mutations.

6.2 Future work

The results suggest a number of predictions about the mechanisms of mutagenesis, which can be tested experimentally. In this section I briefly summarise wet-lab experiments which we plan (and in one case have already started) to use for the validation of some of the predictions. I also suggest future directions of bioinformatics analyses, which form a natural follow-up of the obtained results.

6.2.1 Fidelity of Pol ε in 5mC using maximum-depth sequencing

Our results suggest an increased infidelity of Pol ε when replicating 5mC. If this surprising hypothesis proved to be true, it would change the paradigm of spontaneous deamination of 5mC being the only ubiquitous[1] source of CpG>TpG mutations, the most common mutation type in cancer, normal somatic cells, and germline.

The existing knowledge about *in vitro* fidelity of human Pol ε comes from lacZ forward mutation assay (Korona et al., 2011). Interestingly, insertion of adenine opposite a template cytosine was the second most common error type observed in the human Pol ε lacking the exonuclease domain. This could be the cause of the relatively high frequency of C>T in a CpH context or unmodified CpG context observed in our results. However, as the assay by Korona et al. (2011) did not contain methylated cytosines, any infidelity when replicating DNA modifications would not be detected. Stably introducing DNA methylation into lacZ assay would be difficult and this assay moreover does not allow measurements of mutations in all sequence contexts.

We are therefore designing a system that will allow accurate measurements of fidelity of Pol ε in methylated and unmethylated cytosine in any ca. 300 bp long genomic region, using maximum-depth sequencing (MDS), a modern technique for detecting extremely rare variants in a population of cells through error-corrected, high-throughput sequencing (Jee et al., 2016). The experiments are being performed by Michael McClellan and our collaborators, I have set-up a bioinformatics pipeline for analysis of the sequencing data, and we are in the process of iterative optimisation of the method.

[1] By ubiquitous we mean mutagenesis present in all cells, independent of external mutagens.

Briefly, the experimental design contains the following steps. A region of interest is inserted into pUC18, CpGs are methylated (or left unmethylated), the top strand is removed, and the single-stranded region is filled in by Pol ε. The region of interest is restricted and a unique molecular barcode (random 19mer), a pad (indexing the condition and improving library complexity) and Illumina adapter are added by PCR. Each uniquely barcoded DNA fragment is first linearly and then exponentially amplified using a high fidelity polymerase, and subsequently sequenced. The molecular barcoding and linear amplification enable to distinguish mutations that happened during (or before) the Pol ε synthesis from errors caused by amplification of the DNA fragment. Sequencing both the template and daughter strand will enable comparing a background mutation frequency with errors introduced by Pol ε. A control experiment without Pol ε, in which the whole plasmid and ROI are heated in water, but sequenced in an identical way, will allow us to evaluate the rate of 5mC>T mutations due to spontaneous deamination with Pol ε-induced mutations. Finally, although the primary purpose is to validate the hypothesis about Pol ε and 5mC, the same technique can be used to explore a number of other directions suggested by the observations made in this thesis (such as how the individual polymerases interact with different types of DNA lesions and changes to the nucleotide pool and whether any of these conditions result in the POLE-associated signatures).

6.2.2 The role of oxidative damage in mutational signature 17

Our results support the oxidative damage to the nucleotide pool as the cause of mutational signature 17. If this hypothesis proved to be true, it could have important impact on the prevention of the disease (by prioritising treatment of both gastro-oesophageal reflux and duodeno-gastric reflux) and maybe also the treatment of the disease (discussed in section 5.4.4.4). A number of combined wet-lab and dry-lab experiments could be performed to validate the hypothesis:

1. Does signature 17 in BO samples correlate with the amount of 8-oxoG in the DNA or dNTP pool? What is the relationship with MTH1 expression and other

proteins involved in the predicted pathway (POLH, REV1, REV3L, MTH2, NUDT5, NUDT15)?

2. What is the relationship between signature 17 and chromosomal rearrangements? We plan to explore this using publicly available (and in-house) data sets, correlating signature 17 with the amount of chromosomal breakpoints, and comparing the locations of breakpoints with mutations of signature 17.
3. Can pH 4 and bile acid cocktail (as used in Dvorak et al. (2007)) induce signature 17 in a cell-line that is otherwise free of signature 17? Would the same experiment, but with an overexpressed/knockdown of MTH1 lead to decreased/increased (respectively) signature 17?
4. What are the levels of 8-oxoG in the *in vitro* experimental settings that were previously observed to lead to signature 17 (organoids, MEFs)?

6.2.3 The role of nucleosomes in mutagenesis

Results in this thesis suggest nucleosome rotational positions as a novel genomic feature modulating the mutation landscape in cancer genomes. Moreover, we observed an unexpected novel type of mutation strand asymmetry: on the two strands around the nucleosome dyad. A number of computational and experimental future plans can be done to explore further context, the frequency and impact of these observations:

1. We plan to use sequencing data sets of CPDs, their repair, and mutations in NER-deficient samples to assess which of the components of the UV-induced mutational process (CPD formation, CPD repair, deamination within CPD) cause the nucleosome strand asymmetry and nucleosome-associated modulation of mutagenesis.
2. The nucleosome-associated hypotheses resulting from the first point could be validated experimentally, possibly also using the MDS-based assay.
3. One of the limitations of the current analysis is a use of only a single map of nucleosome positioning. Validation of the results using multiple, ideally tissue-matched maps, and grouping by strongly and weakly positioned nucleosomes, would allow for quantification of robustness and generality of the results.

4. We have also performed an analysis of effects of nucleosomes (distance from dyad, nucleosome occupancy, nucleosome strand asymmetry) on all mutational signatures, and plan to explore the results in the context of current knowledge about each of the signatures.

6.2.4 Modulation of mutational processes by DNA replication

The analysis of replication strand asymmetry in mutational signatures is limited by tissue-unmatched replication maps. Repeating the analysis with tissue-matched maps (when such maps are available) would enable to assess the tissue-specificity of the effects and possibly even stronger signal could be revealed. Moreover, the analysis could be extended with further techniques of measuring replication origins, such as OK-seq (Petryk et al., 2016) and ini-seq (Langley et al., 2016). However, since OK-seq gives similar ORI maps as replication timing domains (Petryk et al., 2016) and ini-seq as the SNS-seq (Langley et al., 2016), we do not expect dramatic differences in the results of mutation signatures analysis. Indeed, our very preliminary results using maps from OK-seq confirm similar results, including the increase of signatures 6 and N4 (but also several others) around the replication origins. Nevertheless, explorations of results using these complementary approaches is one of the possible future steps.

6.2.5 Modulation of mutational processes by DNA modifications

Maps of 5hmC in skin (ideally exposed to sun), lung (ideally from smokers), and one of the tissues with APOBEC-associated mutagenesis (ideally breast, a tissue with a high number of sequenced cancer samples) are needed for more accurate examination of the effects of 5hmC vs. 5mC on mutagenesis in the respective tissues. The highest priority could be given to the skin maps, as our results predict the strongest protectivity of 5hmC in this tissue, and as these results could be interesting to link with the increased 5hmC after UV exposure, but decreased 5hmC during cancer progression, as discussed in section 4.4.5.1.

6.3 Concluding remarks

Understanding the mechanisms of mutagenesis is essential for several aspects of applying the cancer research into practice. First, it can be used to understand which changes in the lifestyle are needed for cancer prevention. Second, it is important for understanding the molecular pathways in the disease: to distinguish which mutated positions give selective advantage to the cancer cells and are therefore driving the disease, from positions that are simply favoured by the mutational processes present in the cell. Finally, mutagenesis is the basis of many types of anti-cancer therapies: from the old types as radiation therapy and chemotherapy, to the targeted and more recent approaches based on DNA repair-mediated synthetic lethality or immunotherapy. Understanding the mutational processes is therefore needed for design of anti-cancer therapeutics, prediction of effective therapies for individual patients (such as suggested in Secrier and Fitzgerald, 2016), understanding resistance to existing therapies, and finding approaches to overcome the resistance.

Results in this thesis show that both DNA modifications and DNA replication play a larger part in the accumulation of mutations than previously appreciated. The results provide novel insights into the mechanisms of a number of mutational processes. We hope that this improved knowledge will help in the long-term efforts of finding effective and personalised cancer treatment.

In the days of my youth, I was told what it means to be a man.
Now I've reached that age, I've tried to do all those things the best I can.
No matter how I try, I find my way into the same old jam.
Good Times, Bad Times, you know I've had my share.

— Led Zeppelin *Good Times Bad Times*

7 Appendix: Publications

7.1 Publications directly associated with the thesis

1. **Marketa Tomkova**, Michael McClellan, Skirmantas Kriaucionis, Benjamin Schuster-Böckler. 5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA. *eLife*, 5(MAY2016):1–23, 2016.
2. **Marketa Tomkova**, Jakub Tomek, Skirmantas Kriaucionis, Benjamin Schuster-Böckler. Widespread impact of DNA replication on mutational mechanisms in cancer. *bioRxiv*, 2017. *In review*.
3. **Marketa Tomkova**, Michael McClellan, Skirmantas Kriaucionis, Benjamin Schuster-Böckler. DNA Replication and associated repair pathways are involved in the mutagenesis of methylated cytosine. *DNA Repair*, 62:1–7, 2018.
4. **Marketa Tomkova**, Skirmantas Kriaucionis, Benjamin Schuster-Böckler. bsQC: quality control pipeline of high-throughput bisulfite-sequencing data. *In preparation*.

7.2 Other publications

1. Chiara Bardella, Osama Al-Dalahmah, Daniel Krell, Pijus Brazauskas, Khalid Al-Qahtani, **Marketa Tomkova**, …, & Ian Tomlinson. Expression of *Idh1*^{R132H} in the Murine Subventricular Zone Stem Cell Niche Recapitulates Features of Early Gliomagenesis. *Cancer Cell*, 30(4):578–594, 2016.
2. Jana Rubáčková Popelová, Karel Kotaška, **Markéta Tomková**, Jakub Tomek. Usefulness of N-Terminal Pro-Brain Natriuretic Peptide to Predict Mortality in Adults With Congenital Heart Disease. *The American Journal of Cardiology*, 116(9):1425–1430, 2015.
3. Jana Rubáčková Popelová, Roman Gebauer, Štěpán Černý, Petr Pavel, Ferdinand Timko, Pavel Jehlička, Petr Plášil, Jakub Tomek, **Markéta Tomková**, Ivo Skalský. Operations of adults with congenital heart disease–Single center experience with 10 years results. *Cor et Vasa*, 58(3):e317–e327, 2016.
4. Jana Rubáčková Popelová, **Markéta Tomková**, Jakub Tomek. NT-proBNP predicts mortality in adults with transposition of the great arteries late after Mustard or Senning correction. *Congenital Heart Disease*, 12(4):448–457, 2017.
5. **Markéta Tomková**, Jakub Tomek, Ondřej Novák, Ondřej Zelenka, Josef Syka, Cyril Brom. Formation and disruption of tonotopy in a large-scale model of the auditory cortex. *Journal of computational neuroscience*, 39(2):131-153, 2015.

7.3 Conferences

1. Medical Sciences DPhil Day, Oxford, 2015, poster.
2. Epigenomics of Common Diseases, Cambridge, 2016, oral presentation.
3. DTC 4th Year Conference/Symposium, 2017, oral presentation, best talk award.
4. EMBL Conference: Cancer Genomics, Heidelberg, 2017, accepted for oral presentation.

Uh! Party over there
Ha! Hands in the air
No! We don't stop
Ha! We rock the spot
No! We don't quit, get ready ya'll this is it

— Captain Jack Lyrics *Dream A Dream*

8

Appendix: Supplementary introduction

8.1 History of epigenomics

The history of epigenetics dates back to the debate about how a single fertilised egg can give rise to a complex organism: is it by enlarging cells which contain preformed elements (*preformationism*) or by gradual developmental changes involving chemical reactions among cellular components (*epigenesis*)? Foundations of the idea of epigenesis[1] were laid already by Aristotle in his *De Generatione Animalium* (Van Speybroeck et al., 2002; Felsenfeld, 2014). In 1950s, Conrad Waddington defined the study of *epigenetics* by combining epigenesis and genetics, as "how genotypes give rise to phenotypes during development" (Waddington, 1957). The developmental focus later diverged into "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" (Russo et al., 1996). Although this definition is still sometimes used, the field of epigenetics also studies genomic marks which are short-lived and not necessarily transmittable between generations (Bird, 2007). It is rather viewed as an additional information beyond the DNA sequence, which is used for coding of states of the cell and in the regulation of gene expression. In 2007, Adrian Bird proposed an updated definition of epigenetic events as "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states"

[1] Although not the term itself.

with an emphasis on responsiveness compared to proactivity of the epigenetic marks (Bird, 2007). The terminology is further complicated by the term *epigenomics*, which has a relationship to epigenetics similar as *genomics* has to *genetics*; however epigenomics and epigenetics are more interchangable and often used in the same context.

8.2 Chromatin and other epigenomic modifications

Tails of the histones in the nucleosome can be post-translationally modified by acetylation, methylation, phosporylation, and other reactions. Depending on the histone core, site on the tail, and type of modification, the histone marks are associated with different states (Zhou et al., 2011; Berger, 2007; Barski et al., 2007; Shlyueva et al., 2014). For instance active promoters are often marked by histone H3 lysine 4 dimethylation (H3K4me2), H3K4me3, and acetylation (ac), and histone variant H2A.Z, whereas promoters of silenced genes commonly have elevated levels of H3K27me3 or H3K9me3. Similarly, active enhancers are associated with H3K4me1 and H3K27ac, whereas closed or poised enhancer regions are associated with H3K27me3 mark. Histone mark H3K36me3 is often found deposited on transcribed gene bodies, whereas H3K9me3 is a repressive mark.

Not only the marks on histone tails, but also the nucleosome positioning itself is viewed as an epigenetic mark. Nucleosomes are often depleted in active promoters, terminator regions, and enhancers, while they occupy genes and intergenic regions (Struhl and Segal, 2013). Nucleosomes tend to be especially well-positioned[2] around TSS, having a strongly positioned nucleosome downstream of the TSS ("+1 nucleosome"), but being depleted upstream of the TSS, form so-called nucleosome depleted regions (NDR; also called nucleosome free regions, NFRs) (Bai and Morozov, 2010). However, some of these established features can be confounded by technical artefacts of the used techniques to measure nucleosome positioning, such as those which employ micrococcal nuclease (MNase) digestion. MNase is known to preferentially cleave DNA at A/T-rich sites (Bai and Morozov, 2010). Indeed, a different technique based on

[2]

chemical mapping detects nucleosomes in regions upstream of the TSSs, which appear to be nucleosome-depleted when measured by MNase-seq (Voong et al., 2016).

Other epigenetic marks involve chromatin interactions and chromatin domains (such as topologically associated domains (TADs), lamina associated domains (LADs), large organized chromatin K9 modifications (LOCKs), long-range epigenetic activation domains (LREAs), long-range epigenetic silencing domains (LRESs), etc.), non-coding RNAs, and numerous modifications of the RNA bases (RNA modifications) (Stricker et al., 2016; Pope et al., 2014).

8.3 Functions of DNA modifications in normal cells

DNA methylation has multiple functions. It is important for X chromosome inactivation[3], where it helps to maintain the silent state of genes on the inactive X chromosome (Lock et al., 1987).

Similarly, methylation is used in imprinting[4] to maintain long-term silencing of the inactive allele. Disruption of methylation in these regions can lead to loss-of-imprinting (LOI), which is found in cancer, Beckwith–Wiedemann syndrome, and other diseases (Robertson, 2005; Peters, 2014).

Methylation is important for genome and chromosomal stability, especially by suppressing expression of transposable elements and preventing instability in repeat regions (Jones, 2012). Mutations in DNMT3B can lead to immunodeficiency, centromere instability and facial anomalies syndrome (ICF syndrome) (Moarefi and Chédin, 2011).

The most studied role of DNA methylation is the regulation of gene transcription. While most CpG sites are highly methylated, CpGs in CpG islands (CGIs) near TSSs are mostly unmethylated (Jones, 2012). *CGI shores* are defined as regions up to 2 kbp from CGIs and *CGI shelves* are regions 2–4 kbp from CGIs (Rechache et al., 2012).

[3]For dosage compensation, one of the X chromosomes is epigenetically inactivated during development of female mammalian cells. The choice of the chromosome is random, but is kept for the rest of the lifetime of the cell.

[4]Imprinting is a process in which several genes are expressed only on one of the two alleles and the other one is silenced. Currently, there are 84 known imprinted genes in the human genome (Morison et al., 2005; Peters, 2014).

Methylated CGIs are often found in repressed (inactive) genes, commonly in a tissue-specific manner (Bird, 2002). The timing and direction of causation of this correlation have been extensively debated. Does the methylation come first and directly cause gene inactivation? Or is the promoter CGI methylation a secondary consequence of the gene being silenced (by other mechanisms)? Once the methylated DNA in a CGI near TSS is assembled into nucleosome, the transcription cannot be initiated (Jones, 2012; Hashimshony et al., 2003; Kass et al., 1997; Venolia and Gartler, 1983). The assembled nucleosome is often marked by repressive H3K9me3, while the active acetylation marks are removed by histone deacetylases, which can be recruited by methylated-DNA binding proteins (Jones and Baylin, 2002; Wade and Wolffe, 2001). In the generally preferred and more supported model, silencing precedes CGI methylation, which serves as a "molecular mark" of the silenced state and demethylation is required for long-term reactivation, while short-term reactivation might be achieved by chromatin remodelling (Jones, 2012; Raynal et al., 2012).

Methylation appears to be important also in regulatory regions with low CpG density (in promoters without CGI, enhancers, and insulators[5]), where methylation also tends to be negatively correlated with expression, especially in tissue-specific genes (Farthing et al., 2008; Han et al., 2011; Gal-Yam et al., 2008). The mechanism is linked to transcription factors, which are often bound to these regions when unmethylated (Schübeler, 2015).

However, the exact mechanisms causing this correlation are also not simple to disentangle. The methylation status can have a direct effect on the transcription factor binding. For instance, the presence of 5mC inhibits binding of MYC (Prendergast and Ziff, 1991) and methylation at CpG-poor LAMB3 and RUNX3 promoters can directly lead to transcriptional silencing of these genes (Han et al., 2011). However, binding of other TFs is not affected by cytosine methylation, such as in the case of SP1 (Harrington et al., 1988). Moreover, systematic survey on the effects of DNA methylation on TF binding showed that some TFs bind specifically methylated CpG sites (Hu et al., 2013).

[5] *Insulators* can be defined as elements that block the interaction between an enhancer and a promoter (Jones, 2012).

The alternative explanation for decreased methylation of regulatory regions of active genes is that transcription factors bind methylated CpG-poor regulatory regions, leading to their demethylation (Schübeler, 2015; Jones, 2012). In this case, the methylation does not instructively cause changes in the gene expression, instead it is itself altered by the transcriptional regulation.

These models are further complicated by recent whole-genome sequencing studies, which show that many methylated CpG-island promoters in male germ cells are actively transcribed (Hammoud et al., 2014) and most differentially methylated regions are not in the promoters or CGIs, but rather in enhancers (Xie et al., 2013; Ziller et al., 2013) and regions adjacent to CGIs (Doi et al., 2009; Irizarry et al., 2009; Edgar et al., 2014). In summary, examples of different models of relationship of transcription and DNA methylation in regulatory regions have been observed (summarised in Spruijt et al. (2013)), but their quantification in different tissues and diseases and their regulation remain to be elucidated.

Finally, gene bodies are extensively methylated and the 5mC levels are often correlated with gene expression (Wolf et al., 1984; Hellman and Chess, 2007; Ogoshi et al., 2011; Aran et al., 2011; Maunakea et al., 2010; Kulis et al., 2012; Varley et al., 2013). This specific hypermethylation of the active gene bodies is linked to the activity of DNMT3B, as this correlation is severely disrupted in cells of DNMT3B-mutated patients with ICF syndrome (Aran et al., 2011). The gene body methylation might have translational implications, as treatment with DNA methylation inhibitor 5-aza-2'-deoxycytidine induces demethylation in gene bodies, altering expression of the genes (Yang et al., 2014).

Two main functions of gene body methylation have been proposed. In the first proposed function, gene body methylation prevents spurious transcription initiation that can stem from cryptic promoters or remnants of transposable elements (Yoder et al., 1997).

In particular, histone methyltransferase SETD2 is recruited by RNA Pol II during transcription elongation to deposit H3K36me3 marks (Wagner and Carpenter, 2012). The H3K36me3 marks are recognised by DNMT3B, which methylates the transcribed gene body to protect the gene body from spurious RNA polymerase II entry and

cryptic transcription initiation, as shown by recent experiments in mouse *Dnmt3B*$^{-/-}$ embryonic stem cells.(Neri et al., 2017; Teissandier and Bourc'his, 2017).

The second proposed function of the gene body methylation is to regulate alternative splicing. DNA methylation is enriched at exons relative to introns (Lister et al., 2009) even after normalisation by the number of CpG sites (Choi, 2010; Gelfman et al., 2013). Moreover, sharp spikes of methylation at 5' splice sites and sharp dips at 3' splice sites were observed (Laurent et al., 2010), supporting the hypothesis that DNA methylation aids the spliceosome in the process of exon definition, which may be possible because spliceosome assembly occurs co-transcriptionally (Pandya-Jones and Black, 2009). RNA sequencing has shown a higher level of DNA methylation in included exons than in excluded exons (Choi, 2010) and inhibition of DNA methylation resulted in aberrant splicing (Maunakea et al., 2013). It was estimated that the splicing of about 22 % of alternative exons is regulated by DNA methylation (Lev Maor et al., 2015). Two mechanisms of how DNA methylation can affect alternative splicing were described and other predicted to exist (Lev Maor et al., 2015). Inclusion or exclusion of exons can be regulated by modulation of RNA Pol II elongation rate via methylation-affected binding by CTCF Shukla et al. (2011) and MeCP2 (Maunakea et al., 2013). Alternatively, DNA methylation can directly affect mRNA alternative splicing by chromatin changes, binding of heterochromatin protein 1 (HP1), and recruitment of splicing factors (Yearim et al., 2015).

Also 5hmC has been proposed to regulate alternative splicing. In brain, 5hmC is slightly but significantly increased on the sense strand (Wen et al., 2014). All three TET proteins bind preferentially near TSS (Williams et al., 2011; Deplus et al., 2013) and the binding of TET2 correlates with gene expression (Chen et al., 2012). 5hmC is enriched especially at the 5' splice sites at the exon-intron boundary in human brain cells (Wen et al., 2014). Moreover, in human frontal cortex, constitutive exons contained higher levels of 5hmC relative to alternatively spliced exons (Khare et al., 2012).

Both 5mC and 5hmC have been implicated in the CTCF-regulated splicing. Methylation inhibits binding of CTCF to exon 5 in CD45, causing exclusion of exon 5 from the transcript, whereas in normal conditions, when exon 5 is not methylated, CTCF

binds to it and promotes inclusion of exon 5 in spliced mRNA through enforcing RNA polymerase II to pause (Shukla et al., 2011). Interestingly, the levels of 5mC negatively correlated with 5hmC levels in this CTCF-binding site and TET-catalysed oxidation of 5mC is required for CD45 exon inclusion, while low TET levels allow methylation of the binding site and subsequent exon exclusion (Marina et al., 2015). Moreover, the same scenario was observed also on genome-wide level: CTCF-binding sites with increased 5hmC and decreased 5mC are associated with upstream exon inclusion, while the opposite situation happens when the upstream exon is excluded (Marina et al., 2015; Marina and Oberdoerffer, 2016).

Similarly as for 5mC and 5hmC, a potential involvement of 5fC and 5caC in splicing regulation has been explored. CTCF preferentially interacts with 5caC-containing DNA and 5caC was detected within CTCF binding sites in the CD45 gene (Marina et al., 2015). Moreover, 5fC and 5caC reduce the rate of RNA Pol II elongation both *in vitro* (Kellinger et al., 2012) and *in vivo* (Wang et al., 2015), which may be also used in the regulation of alternative splicing.

8.4 History of cancer genomics

The link between DNA mutations and cancer dates back to the time of von Hansemann (Hansemann, 1890) and Boveri (Boveri, 1914), when a somatic mutation theory of cancer was proposed, inspired by observations of chromosomal aberrations of dividing cells under the microscope. This was followed by discoveries of other chromosomal abnormalities in cancer (Nowell and Hungerford, 1960; Rowley, 1973), in parallel with studies showing induction of cancer after exposure to various chemicals (reviewed in Loeb and Harris, 2008), such as coal tar (Yamagiwa and Ichikawa, 1918), benzo[a]pyrene (Kennaway, 1930), and aflatoxin (Adamson et al., 1973; Croy et al., 1978). Finally, in 1971 the first tumour suppressor genes were discovered (Knudson, 1971) and in 1982 the first naturally occurring human cancer-causing somatic point mutations were identified (Tabin et al., 1982; Reddy et al., 1982).

Well it's been a long day but I don't like to moan
It's the middle of summer and I'm chilled to the bone
There's holes in my shoes where the rain comes in
I'm sitting on top of the world

— The Pogues *Sitting On Top Of The World*

9 Appendix: Supplementary materials

Tissue	BS-seq		TAB-seq		Source	Link	
	CpGs	Average	CpGs	Average		BS-seq	TAB-seq
Brain	53388534	0.7912	53847986	0.221845	(Wen et al., 2014)	GSM1135081	GSM1135082
Kidney r1	54117976	0.759816	46303252	0.088314	(Chen et al., 2015)	GSM1546664	GSM1546660
Kidney r2	53861360	0.756454	54585341	0.094624	(Chen et al., 2015)	GSM1546666	GSM1546662
Kidney total	54857866	0.757601	54928295	0.092868	(Chen et al., 2015)		
Blood dendritic r1	24586388	0.791371	-	-	(Pacis et al., 2015)	GSM1565940	
Blood dendritic r2	23524704	0.786322	-	-	(Pacis et al., 2015)	GSM1565942	
Blood dendritic r3	24419613	0.782467	-	-	(Pacis et al., 2015)	GSM1565944	
Blood dendritic r4	24452547	0.787728	-	-	(Pacis et al., 2015)	GSM1565946	
Blood dendritic r5	24399654	0.774058	-	-	(Pacis et al., 2015)	GSM1565948	
Blood dendritic r6	24533745	0.790145	-	-	(Pacis et al., 2015)	GSM1565950	
Blood dendritic total	25586845	0.789083	27754454	0.029103	(Pacis et al., 2015)		GSM1565996
Breast	53222114	0.735273	-	-	Epigenome Roadmap	GSM1127125	
Pancreas	54341922	0.697484	-	-	Epigenome Roadmap	GSM983651	
Lung	54236520	0.77405	-	-	Epigenome Roadmap	GSM983647	
Liver	51884076	0.757123	-	-	Epigenome Roadmap	GSM916049	
Stomach	54054176	0.762082	-	-	Epigenome Roadmap	GSM1010984	
Blood (HMPC)	51822931	0.85217	-	-	Blueprint	FTP	

Table 9.1. DNA modification data sets used in chapter 3. HMPC = haematopoietic multipotent progenitor cell. Consortia: Blueprint (http://www.blueprint-epigenome.eu/, dcc.blueprint-epigenome.eu), Epigenome Roadmap (http://www.roadmapepigenomics.org/).

Tissue	Method	Source	Link
blood lymphoid	BS-seq	Blueprint	FTP
blood myeloid	BS-seq	Blueprint	FTP
blood HMPC	BS-seq	Blueprint	FTP
blood dendritic	BS-seq	(Pacis et al., 2015)	SRR1725812, SRR1725813, SRR1725814, SRR1725815
blood dendritic	TAB-seq	(Pacis et al., 2015)	SRR1725859, SRR1725860, SRR1725861
bone	BS-seq	Blueprint	FTP
brain	BS-seq	(Wen et al., 2014)	SRR847423, SRR847424
brain	TAB-seq	(Wen et al., 2014)	SRR847425, SRR847426, SRR847427, SRR847428
breast	BS-seq	Epigenome Roadmap	FTP
colorectum	BS-seq	TCGA	TCGA-AA-3518-11A-01D-1518-05
gastric	BS-seq	Epigenome Roadmap	FTP
kidney r1	BS-seq	(Chen et al., 2015)	SRR1654399, SRR1654400, SRR1654401
kidney r2	BS-seq	(Chen et al., 2015)	SRR1654404, SRR1654405, SRR1827571
kidney r1	TAB-seq	(Chen et al., 2015)	SRR1654388, SRR1654389, SRR1654390, SRR1654391
kidney r2	TAB-seq	(Chen et al., 2015)	SRR1654394, SRR1654395, SRR1827569
liver	BS-seq	Epigenome Roadmap	FTP
lung	BS-seq	Epigenome Roadmap	FTP
oesophagus	BS-seq	Epigenome Roadmap	FTP
oral	BS-seq	Blueprint	FTP
ovary	BS-seq	Epigenome Roadmap	FTP
pancreas	BS-seq	Epigenome Roadmap	FTP
prostate	BS-seq	(Pidsley et al., 2016)	FTP
skin	BS-seq	(Vandiver et al., 2015)	SRR1042910
uterus	BS-seq	TCGA	TCGA-AX-A1CI-11A-11D-A17H-05

Table 9.2. DNA modification data sets used in chapter 4. HMPC = haematopoietic multipotent progenitor cell. Consortia: Blueprint (http://www.blueprint-epigenome.eu/, dcc.blueprint-epigenome.eu), Epigenome Roadmap (http://www.roadmapepigenomics.org/), TCGA = The Cancer Genome Atlas (https://cancergenome.nih.gov/, https://portal.gdc.cancer.gov/).

Cohort	Tissue	WES		WGS		Source
		Patients	SNVs	Patients	SNVs	
Glioblastoma	brain	39	22954	-	-	(Alexandrov et al., 2013a)
Glioma Low Grade	brain	215	9678	-	-	(Alexandrov et al., 2013a)
Medulloblastoma	brain	-	-	100	139553	(Alexandrov et al., 2013a)
Neuroblastoma	brain	210	5027	-	-	(Alexandrov et al., 2013a)
Pilocytic Astrocytoma	brain	-	-	101	12989	(Alexandrov et al., 2013a)
Kidney Chromophobe	kidney	65	1646	-	-	(Alexandrov et al., 2013a)
Kidney Clear Cell	kidney	325	30273	-	-	(Alexandrov et al., 2013a)
Kidney Papillary	kidney	100	6479	-	-	(Alexandrov et al., 2013a)
ICGC RECA EU	kidney	-	-	95	488922	ICGC
Acute Myeloid Leukaemia	blood myeloid	147	2214	7	3659	(Alexandrov et al., 2013a)
Acute Myeloid Leukaemia	blood myeloid	-	-	49	176164	TCGA
Acute Lymphoblastic Leukaemia	blood other	140	1869	1	7881	(Alexandrov et al., 2013a)
Chronic Lymphoid Leukaemia	blood other	103	3998	28	54746	(Alexandrov et al., 2013a)
Lymphoma B-cell	blood other	24	824	24	142753	(Alexandrov et al., 2013a)
Myeloma	blood other	69	3973	-	-	(Alexandrov et al., 2013a)
Breast	breast	844	55731	119	647692	(Alexandrov et al., 2013a)
Liver	liver	-	-	88	899445	(Alexandrov et al., 2013a)
Lung Adeno	lung	636	248519	24	1505512	(Alexandrov et al., 2013a)
Lung Small Cell	lung	70	17639	-	-	(Alexandrov et al., 2013a)
Lung Squamous	lung	176	70485	-	-	(Alexandrov et al., 2013a)
Pancreas	pancreas	98	5093	15	122787	(Alexandrov et al., 2013a)
Stomach	stomach	212	102110	-	-	(Alexandrov et al., 2013a)
Stomach	stomach	-	-	100	1995618	(Wang et al., 2014)
Total sum		3473	588512	751	6197721	

Table 9.3. Mutation data sets used in chapter 3. WES = whole exome sequencing, WGS = whole genome sequencing, ICGC = International Cancer Genome Consortium (http://icgc.org/, https://dcc.icgc.org/), TCGA = The Cancer Genome Atlas (https://cancergenome.nih.gov/, https://portal.gdc.cancer.gov/).

Cohort	Tissue	Samples	Source
ICGC BOCA FR	Bone	97	ICGC
ICGC BRCA EU	Breast	560	ICGC
ICGC CLLE ES	Blood lymphoid	150	ICGC
ICGC COCA CN	Colorectum	26	ICGC
ICGC EOPC DE	Prostate	62	ICGC
ICGC ESAD UK	Oesophagus adenocarcinoma	203	ICGC
ICGC LICA FR	Liver	14	ICGC
ICGC LINC JP	Liver	31	ICGC
ICGC LIRI JP	Liver	258	ICGC
ICGC LUSC CN	Lung squamous	4	ICGC
ICGC LUSC KR	Lung squamous	30	ICGC
ICGC MALY DE	Blood lymphoid	100	ICGC
ICGC MELA AU	Skin	183	ICGC
ICGC ORCA IN	Oral	25	ICGC
ICGC OV AU	Ovary	93	ICGC
ICGC PACA AU	Pancreas	161	ICGC
ICGC PACA CA	Pancreas	159	ICGC
ICGC PAEN AU	Pancreas	48	ICGC
ICGC PAEN IT	Pancreas	37	ICGC
ICGC PBCA DE	Brain	236	ICGC
ICGC PRAD CA	Prostate	124	ICGC
ICGC PRAD UK	Prostate	108	ICGC
ICGC RECA EU	Kidney clear cell	95	ICGC
TCGA AML	Blood myeloid	49	TCGA
TCGA MSI	Colorectum MSI	9	TCGA
TCGA POLE COAD	Colon *POLE*-MUT	7	TCGA
TCGA POLE READ	Rectum *POLE*-MUT	3	TCGA
TCGA POLE UCEC	Uterus *POLE*-MUT	2	TCGA
AML	Blood myeloid	7	(Alexandrov et al., 2013a; Ding et al., 2012)
Lung Adeno	Lung adenocarcinoma	24	(Alexandrov et al., 2013a; Imielinski et al., 2012)
Lymphoma B cell	Blood lymphoid	24	(Alexandrov et al., 2013a)
Bass Colon	Colorectum	9	(Bass et al., 2011)
bMMRD	Brain *POLE*-MUT	2	(Shlien et al., 2015)
Dulak Oesophagus	Oesophagus adenocarcinoma	16	(Dulak et al., 2013)
Wang Gastric MSI	Gastric MSI	10	(Wang et al., 2014)
Wang Gastric MSS	Gastric MSS	90	(Wang et al., 2014)
Total sum		3056	

Table 9.4. WGS data sets used in chapter 4 and 5. For chapter 4, from each patient only one sample was taken. ICGC = International Cancer Genome Consortium (http://icgc.org/, https://dcc.icgc.org/), TCGA = The Cancer Genome Atlas (https://cancergenome.nih.gov/, https://portal.gdc.cancer.gov/).

You know it feels like you're going insane
And I've done everything you told me to take away the pain
Then you change my medication again
It's getting harder to tell just who or what's insane

— Levellers *The Fear*

10 Appendix: Supplementary results

10.1 Tobacco-induced mutagenesis in modified cytosines

Tobacco-induced mutagenesis is another major mutagenic process with a known link to DNA modifications. As reviewed in the Introduction (section 1.4.3), it is known that BPDE adducts from tobacco smoke preferentially bind guanines of methylated CpGs (Denissenko et al., 1997), in particular when the 5mC is opposite the guanine (Guza et al., 2011). We could therefore expect to see a linear relationship between the tobacco-induced C>A mutations and methylation levels. Such relationship has however not yet been verified in human cancer samples.

Here we explored the mutation spectra in CpG positions of 58 lung cancer patients with a history of smoking and with whole genomes sequenced, combining the data with BS-seq derived modification levels from normal lung tissue. We first binned the CpG positions by their modification level (0-0.1, ..., 0.9-1.0) and computed the frequency of C>A mutations separately for each sequence context in each lung cancer sample.

In all four sequence contexts, the average C>A mutation frequencies were positively correlated with DNA modification levels (Fig. 10.1A). This relationship was relatively consistent across the samples (Fig. 10.1B,C). Fitting a linear model for each sample showed that the slope of the correlation is in the vast majority of cases positive (Fig. 10.1D). In line with this observation, the overall slope of CpG>ApG correlation with

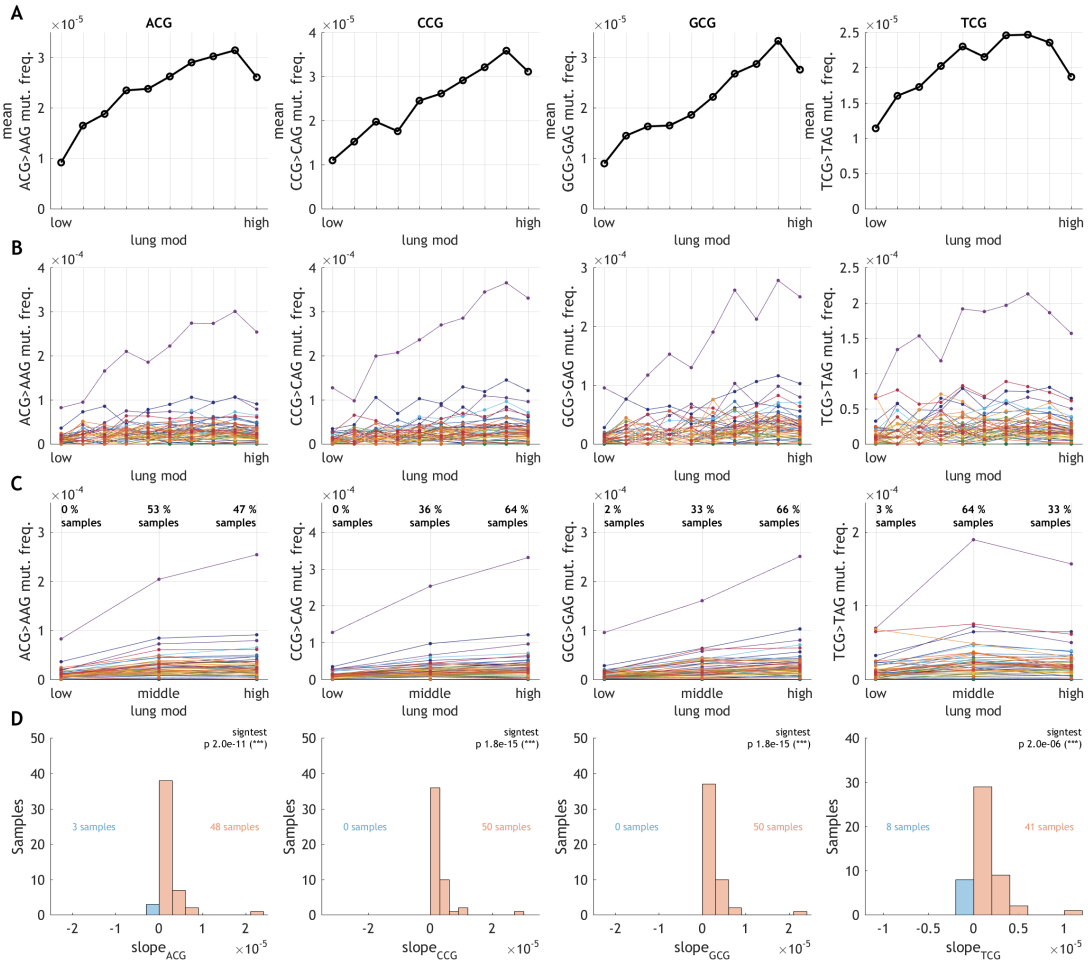

Figure 10.1. C>A mutations positively correlate with modification levels in lung cancer samples. All CpGs were binned according to normal lung BS-seq measured mod levels (0-0.1, …, 0.9-1.0). The first bin represents unmodified sites and the last bin represents fully modified sites. C>A mutation frequency was computed in each bin, separately for each sequence context (columns). **A:** Mean over samples. **B:** One trace per sample. **C:** Only the low mod (first bin), high mod (last bin), and middle mod (mean of the two middle bins) values are shown. The percentage of samples with the highest mutation frequency in the low mod, middle mod, and high mod are written at the top of the figure. For example in TCG context, 86 % of samples have the middle mod value higher than the two extreme values. **D:** Distribution of slopes of linear fits to data in (B) in individual samples. Numbers of samples with negative and positive slope are printed on the sides of the histogram.

modification levels (all contexts together) strongly correlated with the mutational signature 4, the signature associated with tobacco smoking, (Pearson correlation 0.99 with p-value of $4.3 \cdot 10^{-47}$; Fig. 10.2).

The highest number of samples with a negative correlation with modification levels was in a TCG context (8 samples; Fig. 10.1D). A possible explanation is via the activity

of APOBEC enzymes, which can also contribute to TCN>TAN mutations, albeit with lower frequency than to TCN>TGN and TCN>TTN mutations (Alexandrov et al., 2013a; Akre et al., 2016; Roberts et al., 2013; Burns et al., 2013b). All the 8 samples had a strong component of APOBEC-associated mutations in their mutation spectra (in all the 8 samples, exposure to signature 2 was $\geq$ 426, exposure to signature 13 was $\geq$ 723), suggesting that the negative slope is due to decreased APOBEC activity in methylated cytosine, rather than an effect of tobacco-induced mutagenesis.

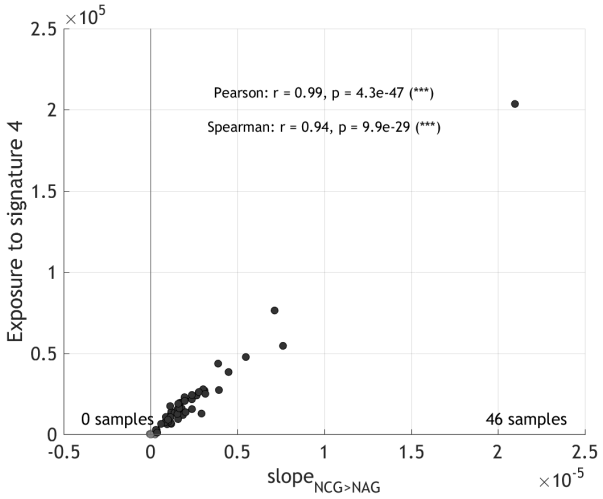

Figure 10.2. Exposure to signature 4 negatively correlates with the slope of CpG>ApG correlation with modification levels. Slope of the linear fits from Fig. 10.1 (all contexts together) is plotted against the exposure to signature 4 (tobacco smoke signature). Samples with exposure to signature 4 above 500 are shown in dark black.

Finally, we attempted to determine the impact of 5hmC on the tobacco-induced mutagenesis. To our knowledge, there are no available whole-genome TAB-seq measurements of 5hmC from lung samples. However, whole-genome oxBS-seq measurements have been recently published (Li et al., 2016). As oxBS-seq measures directly 5mC and BS-seq measures mod, the difference of BS-seq and oxBS-seq measurements in individual positions should represent levels of 5hmC. However, due to the noise in the measurements and other reasons summarised in the General methods (section 2.2), additional steps need to be performed to distinguish truly hydroxymethylated sites. Li et al. (2016) therefore applied regional smoothing, identified regions with 5hmC significantly higher than zero, and assigned the 5hmC values of CpGs in these

regions as the subtraction between the smoothed BS-seq and oxBS-seq values. Here, we used the 2 204 915 CpGs with assigned 5hmC estimates in normal lung and combined them with mod levels. Due to the relatively low number of CpGs and cancer samples, we grouped the CpGs into three bins only, according to their estimated $5hmC_{rel}$, such that each bin contains approximately the same number of CpGs ($5hmC_{rel}$: 0-0.0820, 0.0820-0.1139, 0.1139-1).

The frequency of CpG>ApG mutations showed a general decreasing trend with increasing $5hmC_{rel}$ levels (Fig. 10.3A). The decrease was only mild (1.4-fold in the third bin compared to the first bin) and the relationship was not highly consistent across the samples (17 % of samples had highest mutation frequency in the high $5hmC_{rel}$ bin; Fig. 10.3B), but this might be also related to the low statistical power and the generally low amount of 5hmC in the lung maps: the third bin contains CpGs with a broad range of $5hmC_{rel}$ values in between = 0.1139 and 1.

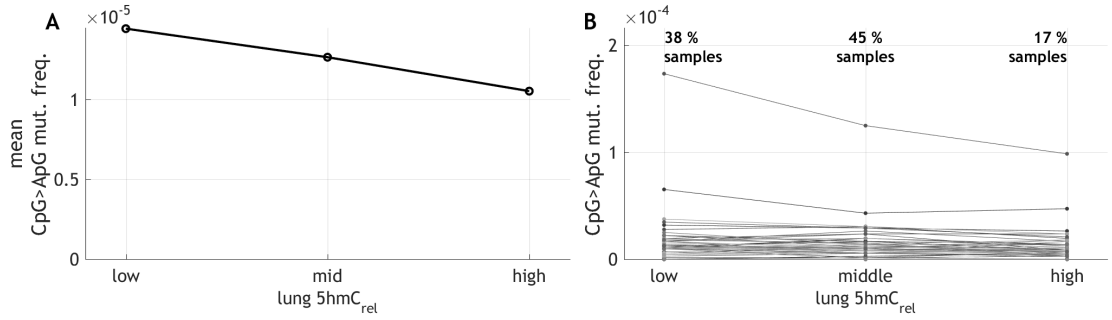

Figure 10.3. C>A mutations are more frequent in CpG sites with 5mC than 5hmC. All CpGs were binned according to normal lung $5hmC_{rel}$ levels (0-0.0820, 0.0820-0.1139, 0.1139-1), such that each bin contained ca. $722 \cdot 10^3$ CpGs (detail in text). The first bin represents fully methylated sites and the last bin represents sites with at least 11 % of 5hmC. C>A mutation frequency was computed in each bin. **A:** Mean over samples. **B:** One trace per sample. The percentage of samples with the highest mutation frequency in the low $5hmC_{rel}$, middle $5hmC_{rel}$, and high $5hmC_{rel}$ are written at the top of the figure.

In summary, our results support the model in which 5mC increases the probability of a BPDE adduct forming on the guanine opposite the 5mC. The BPDE-dG adduct can be then replicated in an error-free (such as by Pol κ (Avkin et al., 2004; Jha et al., 2016)), or error-prone (by Pol η (Zhao et al., 2006; Klarer et al., 2012)) manner, paired with adenine on the daughter strand, thus creating a C>A mutation. The effect of 5hmC on the tobacco-induced C>A mutagenesis is less clear (due to the limited statistical power

of the explored data sets), but our results suggest that also these mutations occur more frequently in methylated than hydroxymethylated CpGs.

10.2 APOBEC-induced mutagenesis in modified cytosines

Cytosines can deaminate spontaneously, or enzymatically, such as by one of the AID/APOBEC family enzymes. As reviewed in the Introduction (section 1.4.4), AID and APOBEC enzymes *in vitro* exhibit decreased activity on methylated and hydroxymethylated cytosine compared to unmodified cytosine, although the exact values of deamination rate in C, 5mC, and 5hmC differ between the different members of AID/APOBEC family. The APOBEC-induced mutation frequency would be therefore expected to decrease with increasing modification levels. Indeed, tumours with high APOBEC signature showed two-fold higher frequency of TCG mutation in lowly methylated than highly methylated cytosines (Seplyarskiy et al., 2016b). However, this relationship has never been explored into any further detail.

Here we used mutation spectra of 560 breast cancer whole-genome sequenced samples and BS-seq measurements from normal breast tissue to determine the CpG>TpG mutation frequency with respect to modification levels. First we binned the CpG positions by their modification level (0-0.1, ..., 0.9-1.0) and computed the frequency of C>T mutations separately for each sequence context in each breast cancer sample. While most samples in ACG, CCG, and GCG contexts showed a clear increase of C>T mutation frequency, TCG was the only context with a non-trivial proportion of samples exhibiting decreasing frequency of C>T mutations (Fig. 10.4). In order to quantify this proportion, we fitted a linear model through the TCG>TTG mutation frequency of each mod level (i.e., the data shown in in Fig. 10.4B) and defined $slope_{TCG>TTG}$ as the slope of the fitted model. The distributions of slopes for individual samples are shown in Fig. 10.4D separately for each sequence context. In a TCG context, 18.9% of samples had a negative slope, compared to less than 3% of samples with negative slope in the other three contexts 10.4D.

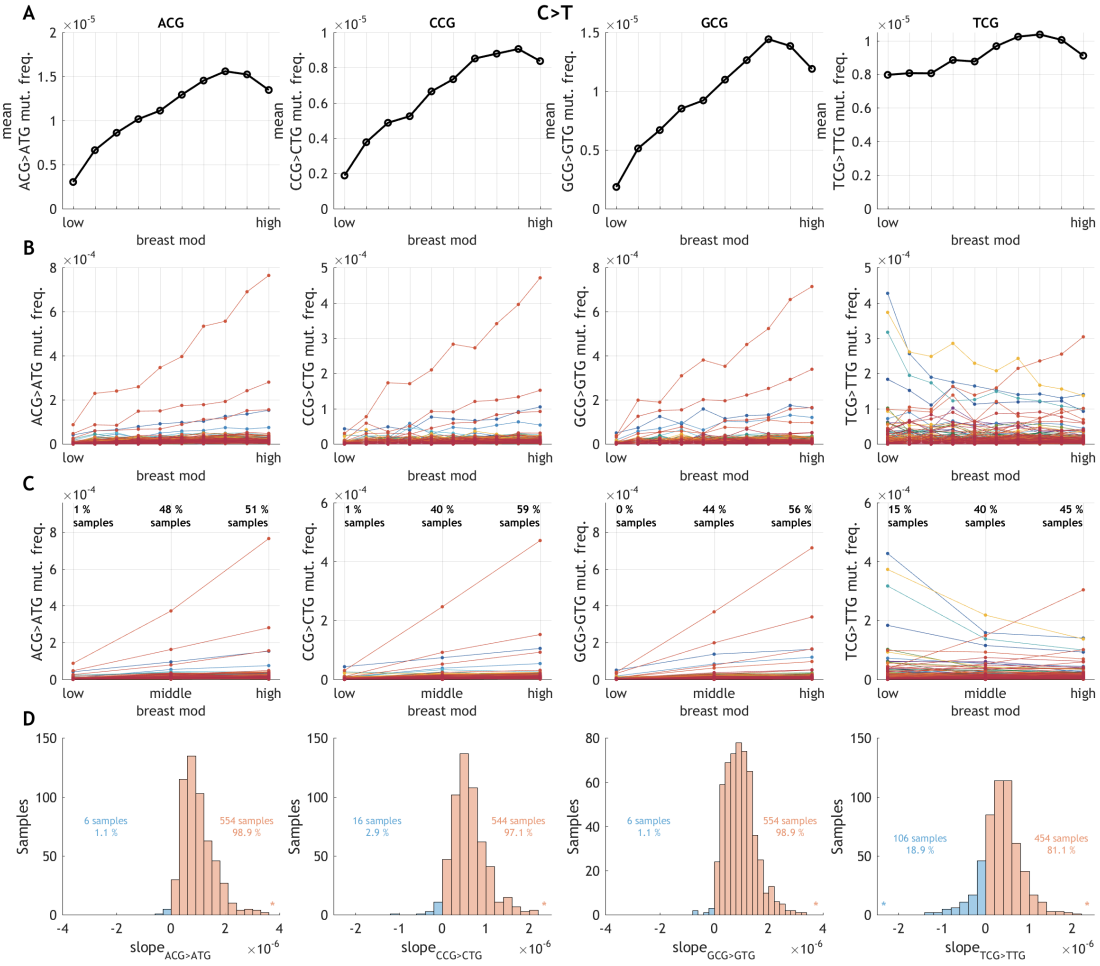

Figure 10.4. C>T mutations negatively correlate with modification levels in breast cancer samples. All CpGs were binned according to normal breast BS-seq measured mod levels (0-0.1, …, 0.9-1.0). The first bin represents unmodified sites and the last bin represents fully modified sites. C>T mutation frequency was computed in each bin, separately for each sequence context (columns). **A:** Mean over samples. **B:** One trace per sample. **C:** Only the low mod (first bin), high mod (last bin), and middle mod (mean of the two middle bins) values are shown. The percentage of samples with the highest mutation frequency in the low mod, middle mod, and high mod are written at the top of the figure. **D:** Distribution of slopes of linear fits to data in (B) in individual samples. Numbers of samples with negative and positive slope are printed on the sides of the histogram.

Based on the *in vitro* measurements of activity of APOBEC enzymes on 5mC, we would expect much larger proportion of APOBEC-exposed samples to exhibit negative $\text{slope}_{\text{TCG>TTG}}$. However, not all samples in this cohort have a dominant APOBEC component. When taking into account only samples strongly exposed to signature 2, the APOBEC signature dominated by C>T mutations (exposure > 500), the percentage of samples with negative $\text{slope}_{\text{TCG>TTG}}$ is exactly 50% (52 out of 104 samples; Fig 10.5).

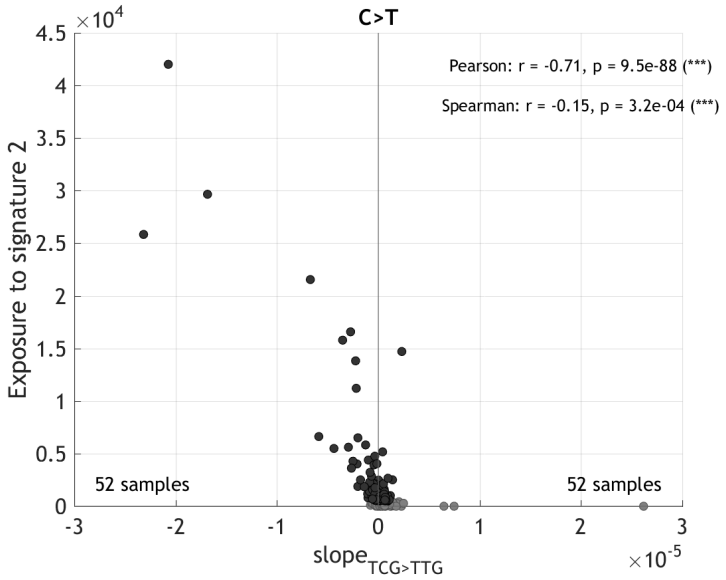

Figure 10.5. Exposure to signature 2 negatively correlates with the slope of CpG>TpG correlation with modification levels in breast. Slope of the linear fits from Fig. 10.4 (all contexts together) is plotted against the exposure to signature 2 (APOBEC-associated signature). Samples with exposure to signature 2 above 500 are shown in black.

Samples with the strongest signature 2 were generally those with the most negative $\text{slope}_{\text{TCG>TTG}}$, suggesting that the APOBEC-induced C>T mutagenesis is at least in some samples enhanced by the lack of cytosine modifications, in line with the *in vitro* measurements. We also explored what could be causing that 52 of samples strongly exposed to signature 2 show positive $\text{slope}_{\text{TCG>TTG}}$ (see details in Appendix 10.2.1) and conclude that this might be result of a combination of spontaneous deamination of 5mC and involvement of APOBEC3H, which exhibits similar activity on C and 5mC, as opposed to APOBEC3A/B which strongly prefer C.

Finally, when repeating the same analysis for C>G mutations, we observe dominance of negative slope in most samples (and interestingly also most contexts) (Fig. 10.7) and a significant correlation of the $\text{slope}_{\text{TCG>TGG}}$ with exposure to signature 13, the APOBEC-associated signature dominated by C>G mutations (Fig. 10.6), supporting a decreased APOBEC-induced C>G mutagenesis in modified cytosines. Two mechanisms might be contributing to the much larger proportion of negative $\text{slope}_{\text{TCG>TGG}}$ than negative $\text{slope}_{\text{TCG>TTG}}$. First, C>T mutations can be also caused by spontaneous deamination of 5mC. Second, the C>G mutations in APOBEC-induced mutagenesis are thought to arise

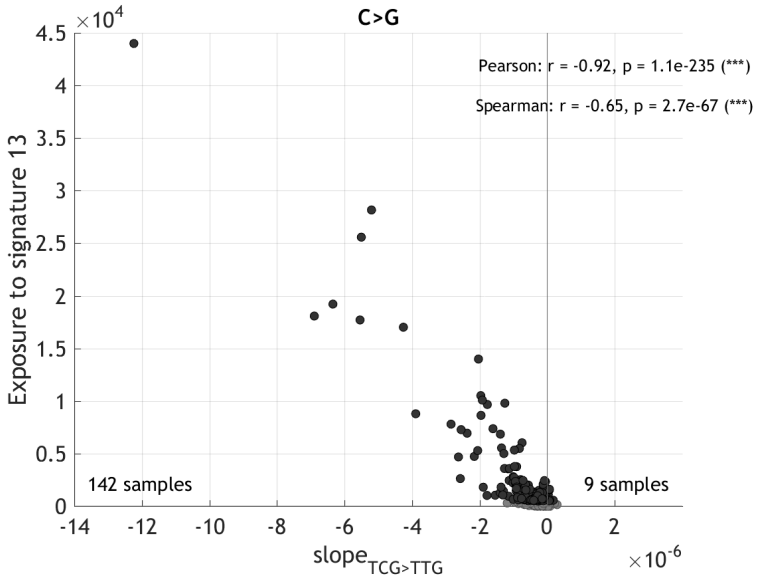

Figure 10.6. Exposure to signature 13 negatively correlates with the slope of CpG>GpG correlation with modification levels in breast. Slope of the linear fits from Fig. 10.7 (all contexts together) is plotted against the exposure to signature 13 (APOBEC-associated signature). Samples with exposure to signature 13 above 500 are shown in black.

from excision of the deaminated base and insertion of C opposite the AP site by REV1 (Morganella et al., 2016). While cytosine deaminates into uracil, 5mC deaminates into thymine. It is possible that uracil will get excised with higher efficiency, especially if the deamination happened on single-stranded DNA, as APOBECs prefer single-stranded DNA as their substrate (Chen et al., 2006; Roberts et al., 2012).

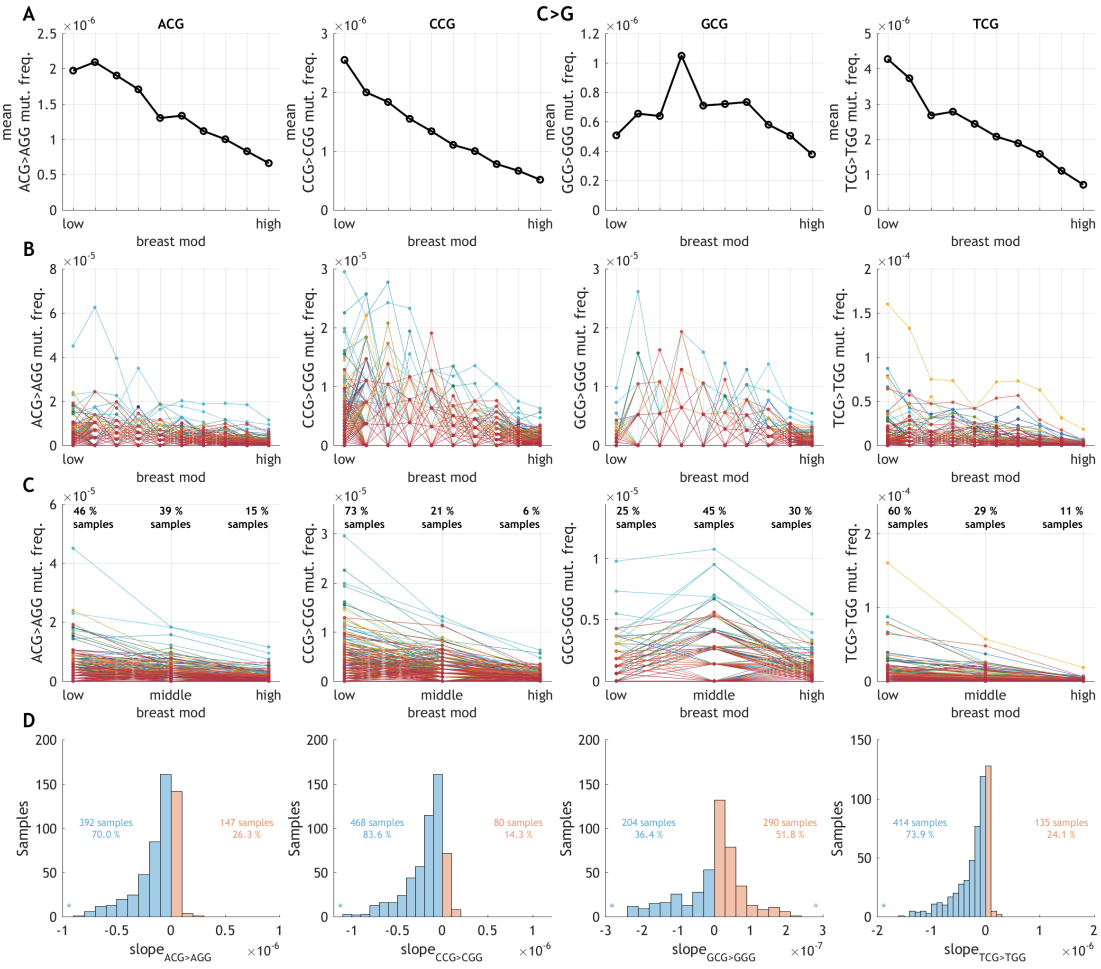

Figure 10.7. C>G mutations negatively correlate with modification levels in breast cancer samples. All CpGs were binned according to normal breast BS-seq measured mod levels (0-0.1, …, 0.9-1.0). The first bin represents unmodified sites and the last bin represents fully modified sites. C>G mutation frequency was computed in each bin, separately for each sequence context (columns). **A:** Mean over samples. **B:** One trace per sample. **C:** Only the low mod (first bin), high mod (last bin), and middle mod (mean of the two middle bins) values are shown. The percentage of samples with the highest mutation frequency in the low mod, middle mod, and high mod are written at the top of the figure. **D:** Distribution of slopes of linear fits to data in (B) in individual samples. Numbers of samples with negative and positive slope are printed on the sides of the histogram.

10.2.1 Samples with positive correlation of TCG>TTG with mod

Two possibilities could explain the high percentage of samples with positive $\text{slope}_{\text{TCG>TTG}}$. First, these samples are highly affected by signature 1, which is assumed to be caused by spontaneous deamination of 5mC to T, and therefore has a positive correlation with methylation levels. Alternatively, some of these samples are affected by APOBEC enzymes that have similar or higher efficiency on 5mC as on unmodified cytosine. Plotting the difference of exposures to signatures 1 (spontaneous deamination) and signature 2 (APOBEC) against the $\text{slope}_{\text{TCG>TTG}}$ shows a strong correlation between these two variables (Pearson correlation 0.8 with p-value of 10^{156}, Spearman correlation 0.5 with p-value of 0; Fig. 10.8), supporting the first scenario. Nevertheless, 26 samples exhibit a strong exposure to signature 2 (exposure > 500), higher exposure to signature 2 than to signature 1, and positive $\text{slope}_{\text{TCG>TTG}}$.

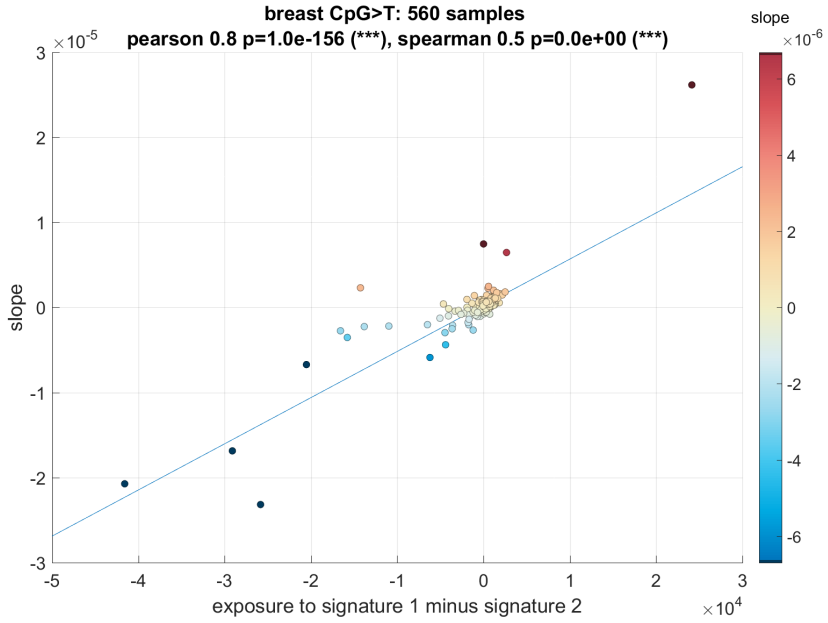

Figure 10.8. The slope of TCG>TTG correlation with modification levels is negative in samples exposed more to signature 2 than to signature 1, but positive in the opposite scenario. Slope of the linear fits from Fig. 10.4 in a TCG context ($\text{slope}_{\text{TCG>TTG}}$) on the y-axis, plotted against the difference of exposures to signatures 1 (spontaneous deamination) and signature 2 (APOBEC) on the x-axis.

The only APOBEC enzyme with efficiency on 5mC similar to efficiency on C is APOBEC3H (see Introduction 1.4.4). This enzyme has been recently implicated to

play a role in cancer mutagenesis alongside APOBEC3B and especially in APOBEC3B-null breast cancers (Starrett et al., 2016). It has been also observed that APOBEC3H deaminates 5mC in a TCG context with a ca. 2-fold higher efficiency than AID or APOBEC3B (Gu et al., 2016). We therefore explored whether the samples with positive correlation of TCG>T mutations with modification levels could in fact be more affected by APOBEC3H than APOBEC3A/B enzymes. As APOBEC3A/B have a preference for TCA (and TCT) sequence context (Mertz et al., 2017b), we compared the C>T mutation frequency in TCA vs. TCG in samples with a high exposure to signature 2 (exposure > 500) and higher exposure to signature 2 than to signature 1. In this group, 50 samples had a negative $\text{slope}_{\text{TCG>TTG}}$, whereas 26 samples showed a positive $\text{slope}_{\text{TCG>TTG}}$. Interestingly, the samples with positive $\text{slope}_{\text{TCG>TTG}}$ had an increased preference for the TCG context compared to TCA context (ranksum test, p = 0.002; Fig. 10.9), more so than the samples with negative $\text{slope}_{\text{TCG>TTG}}$ (ranksum test on TCG/(TCA+TCG) in the two groups, p = 0.017; Fig. 10.9). This supports the possibility of APOBEC3H-induced mutagenesis in the samples positively correlated with modification levels, as the positive $\text{slope}_{\text{TCG>TTG}}$ would be a combination of less steep negative $\text{slope}_{\text{TCG>TTG}}$ due to APOBEC3H (compared to APOBEC3A/B) and a positive $\text{slope}_{\text{TCG>TTG}}$ due to the spontaneous deamination of 5mC.

In order to separate these two factors, we also focused on TCG>TGG mutations, as those would not be confounded by the spontaneous deamination. Compared to C>T mutation, the C>G mutations were mostly negatively correlated with modification levels: e.g., for TCG context 73.9 % of the 560 breast samples had a negative $\text{slope}_{\text{TCG>TTGG}}$ (Fig. 10.7D).

Finally, we compared the $\text{slope}_{\text{TCG>TGG}}$ (Fig. 10.7D) in the same two groups of samples as in Fig. 10.9, i.e., samples strongly exposed to signature 2 and with either positive, or negative $\text{slope}_{\text{TCG>TTG}}$. In line with the previous results, the samples with positive $\text{slope}_{\text{TCG>TTG}}$ exhibit also less negative $\text{slope}_{\text{TCG>TGG}}$ compared to samples with negative $\text{slope}_{\text{TCG>TTG}}$ (ranksum test, p = 0.006, Fig. 10.10), supporting the potential involvement of APOBEC3H mutagenesis in these samples.

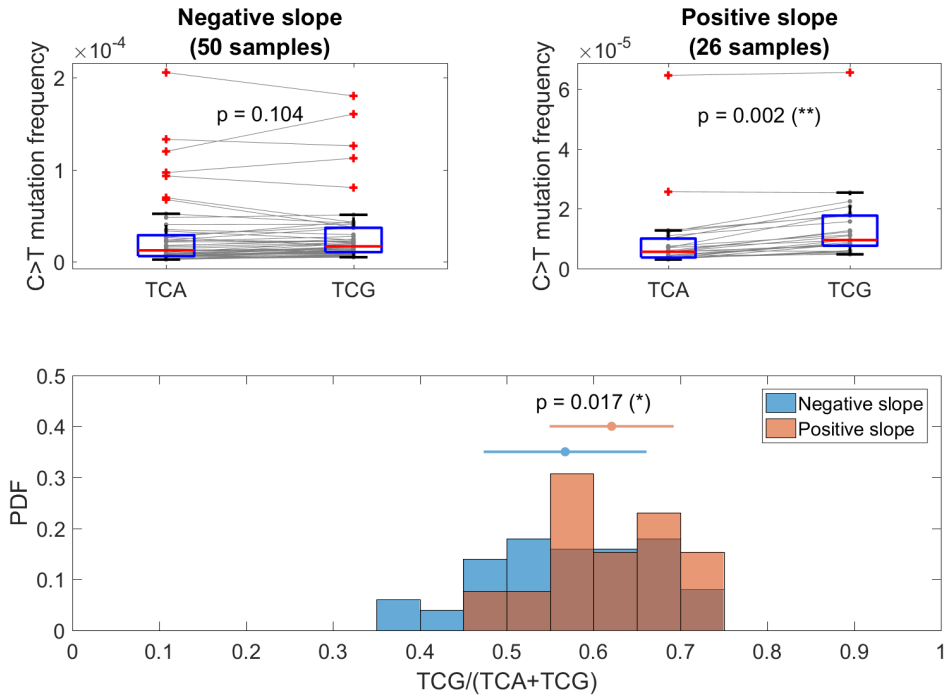

Figure 10.9. Samples with TCG>T mutations positively correlated with modification levels show an increased frequency of TCG>TTG mutations vs. TCA>TTA mutations, compared to negatively correlated samples, suggesting a potential involvement of APOBEC3H in the mutagenesis of the positively correlated samples. The breast cancer samples with exposure to signature 2 above 500 and above exposure to signature 1 were split into two groups: 50 samples had a negative $\text{slope}_{\text{TCG>TTG}}$ and 26 samples with a positive $\text{slope}_{\text{TCG>TTG}}$ (the slope of the linear fits is shown in Fig. 10.4D). **Top:** Frequency of C>T mutations in a TCA context vs. a TCG context is compared in the two groups with a ranksum test. **Bottom:** Distribution of TCG>TTG mutation frequency divided by (TCG>TTG mutation frequency + TCA>TTA mutation frequency) was compared between the two groups also using a ranksum test.

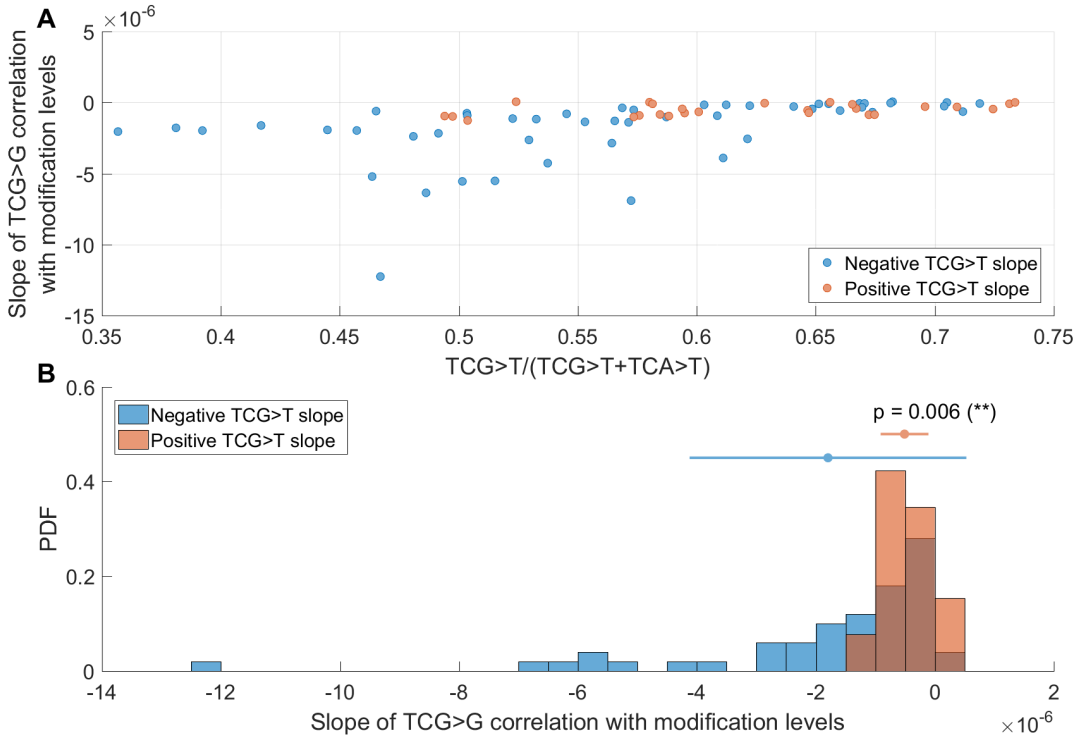

Figure 10.10. Samples with TCG>TTG mutations positively correlated with modification levels show less negative slope$_{TCG>TGG}$, compared to negatively correlated samples, in line with the suggested potential involvement of APOBEC3H in the mutagenesis of the positively correlated samples. The breast cancer samples with exposure to signature 2 above 500 and above exposure to signature 1 were split into two groups: 50 samples had a negative slope$_{TCG>TTG}$ and 26 samples with a positive slope (the slope of the linear fits is shown in Fig. 10.4). **Top:** TCG>TTG mutation frequency divided by (TCG>TTG mutation frequency + TCA>TTA mutation frequency) plotted against the slope of TCG>TGG mutation frequency correlation with modification levels. **Bottom:** Distribution of the slope of TCG>TGG mutation frequency correlation with modification levels was compared between the two groups also using a ranksum test.

10.2.2 Discussion of APOBEC-induced mutagenesis in modified cytosines

APOBEC-induced mutagenesis has attracted a lot of attention in the recent five years. They have been shown to play a role in drug resistance (Law et al., 2016), predict patient survival (Glaser et al., 2017), correlate with the expression of PD-L1 and a T-cell inflamed signature (Boichard et al., 2017; Rieke et al., 2017), and suggested as a therapeutic target (Wang and Taylor, 2017; Swanton et al., 2015). Understanding the mechanisms of APOBEC-induced mutagenesis and how to distinguish potential different modes of this mutagenicity in the mutation spectra is therefore important for translating the knowledge into the clinic.

Our results show decreased APOBEC-induced mutagenicity in modified cytosines in the TCG context, but the effect is smaller than expected. *In vitro*, the decrease of deamination efficiency in 5mC compared to C is 5–10-fold decrease in APOBEC3A and even 50-fold decrease in APOBEC3B (reviewed in the Introduction 1.4.4). The differences in sequence context and modification status preferences of individual AID/APOBEC enzymes suggest the possibility to use these features to infer which of the enzymes have likely been operating in individual cancer samples. In particular, the strongest negative correlation with modification levels would be expected for APOBEC3G, APOBEC3B and AID, less for APOBEC3A, and the least for APOBEC3H (especially hap II). This can be combined with the known preferences of sequence context of the different AID/APOBEC enzymes: TC[A/G/T] in APOBEC3H, [T/C]TC[A/T] for APOBEC3A, [A/G]TC[A/T] for APOBEC3B, CCN for APOBEC3G, and finally [A/T][A/G]C[C/T] for AID (Kamba et al., 2015; Rebhandl, 2015; Gu et al., 2016; Seplyarskiy et al., 2016a; Starrett et al., 2016).

Here, we applied such inference on a cohort of 560 WGS breast cancer samples. We identified 26 samples which have likely been affected by APOBEC3H mutagenesis, more than by other of the AID/APOBEC enzymes. Compared to the other samples with a strong APOBEC signature 2, they have "less negative" $\text{slope}_{\text{TCG>TGG}}$ (i.e., negative but in absolute values smaller than the other samples), corresponding to the only slightly decreased activity of APOBEC3H on 5mC than C (while APOBEC3A/B exhibit much larger decrease of activity on 5mC) (Gu et al., 2016). These samples have an enrichment

of C>T mutations in a TCG context compared to TCA context, in line with the observed context preference of APOBEC3H compared to APOBEC3A/B (Gu et al., 2016; Mertz et al., 2017b). Moreover, they have TCG>TTG mutations positively correlated with modification levels, possibly due to a combined effect of APOBEC-induced deamination and spontaneous deamination. Combining these observation with survival information, gene expression, and proteomics data in the future will help to elucidate the role of individual APOBEC enzymes in cancer mutagenesis.

10.3 Replication-strand asymmetry of mutational signatures

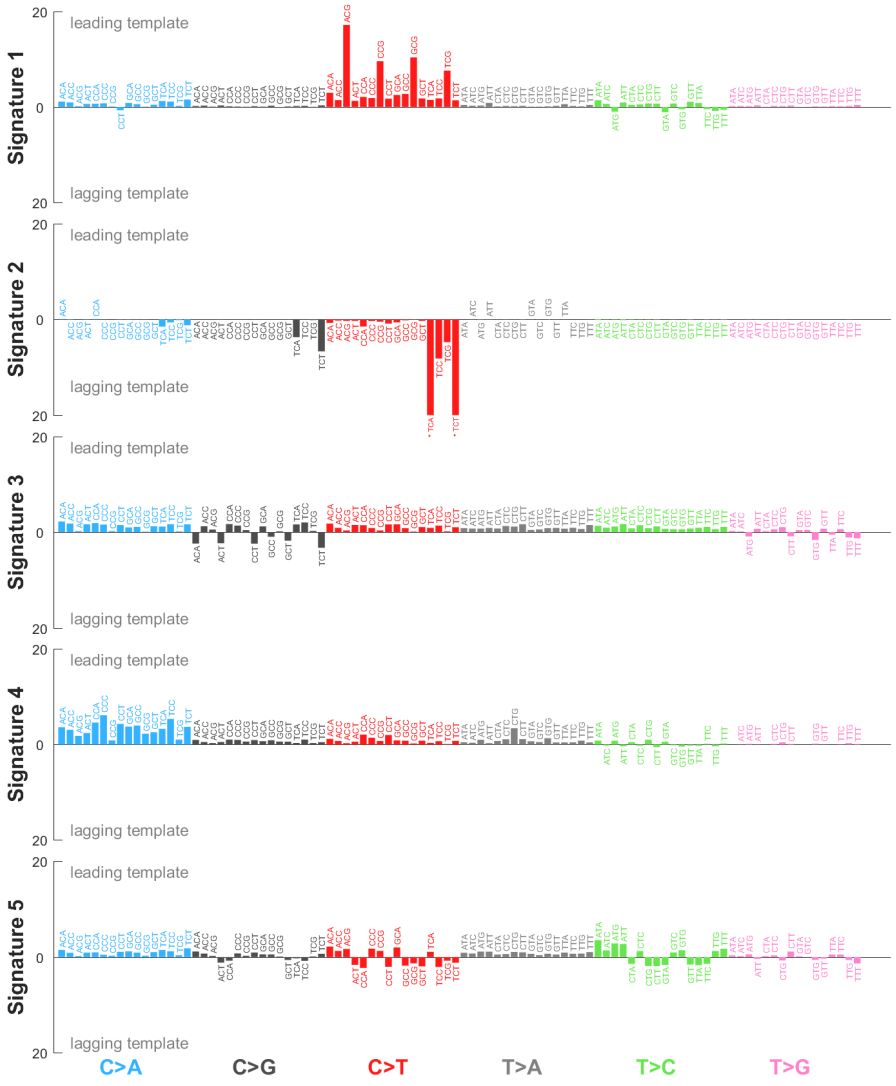

Figure 10.11. Directional signatures 1-5 Each of the 96 mutation types is annotated with a dominant direction: leading (pointing up), or lagging (pointing down). Asterisks indicate mutation types exceeding 20 %.

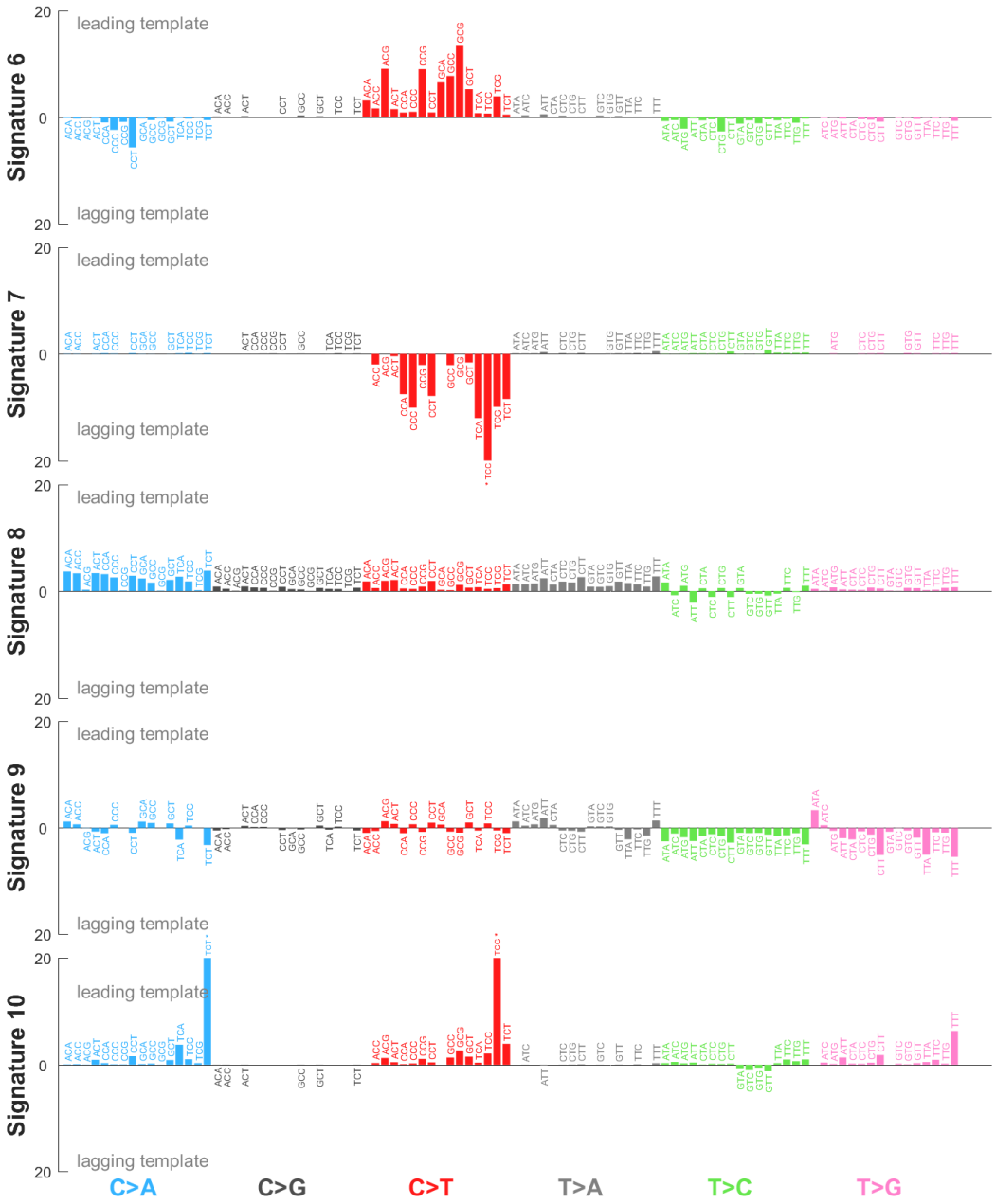

Figure 10.12. Directional signatures 6-10 Each of the 96 mutation types is annotated with a dominant direction: leading (pointing up), or lagging (pointing down). Asterisks indicate mutation types exceeding 20 %.

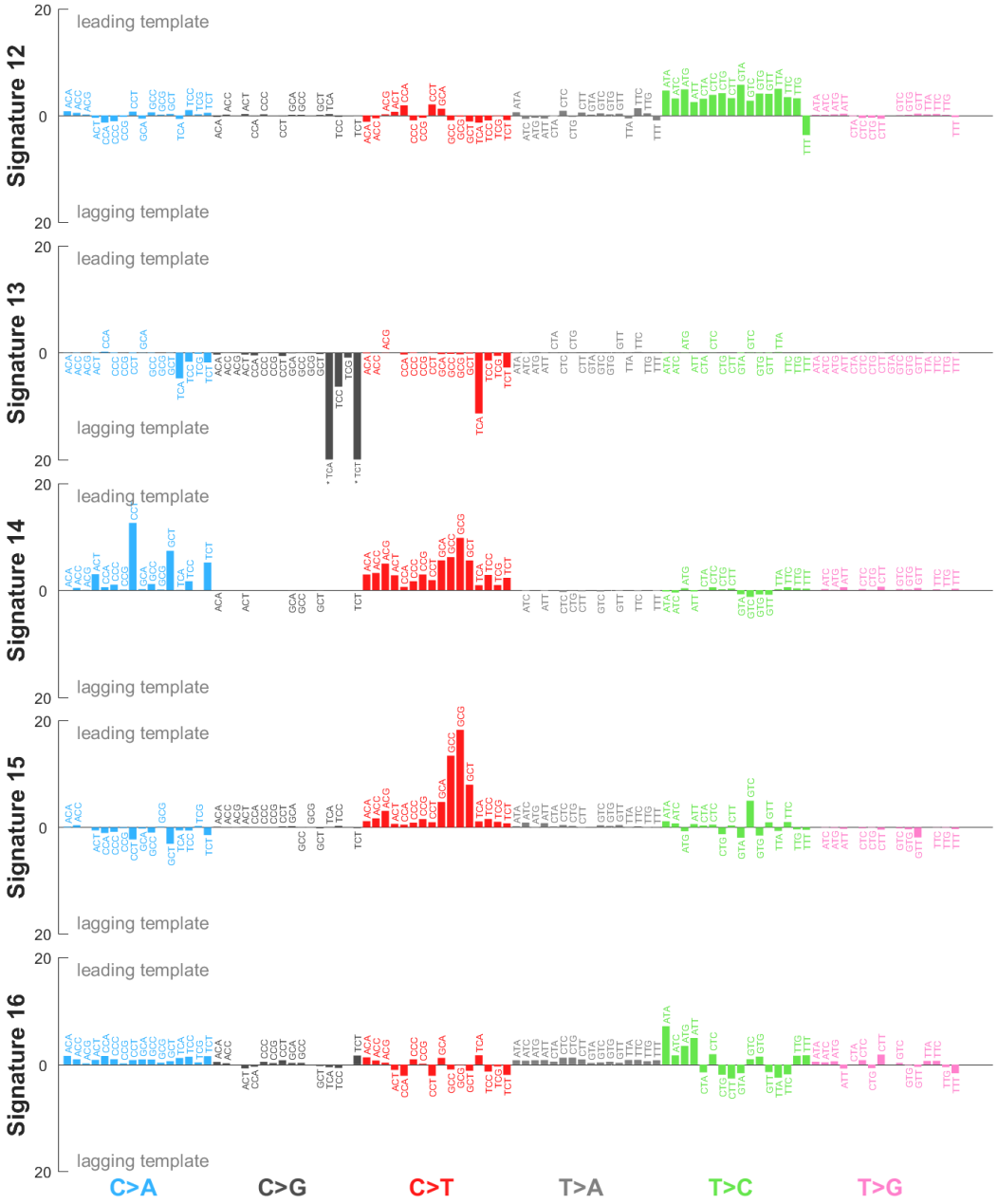

Figure 10.13. Directional signatures 12-16 Each of the 96 mutation types is annotated with a dominant direction: leading (pointing up), or lagging (pointing down). Asterisks indicate mutation types exceeding 20 %.

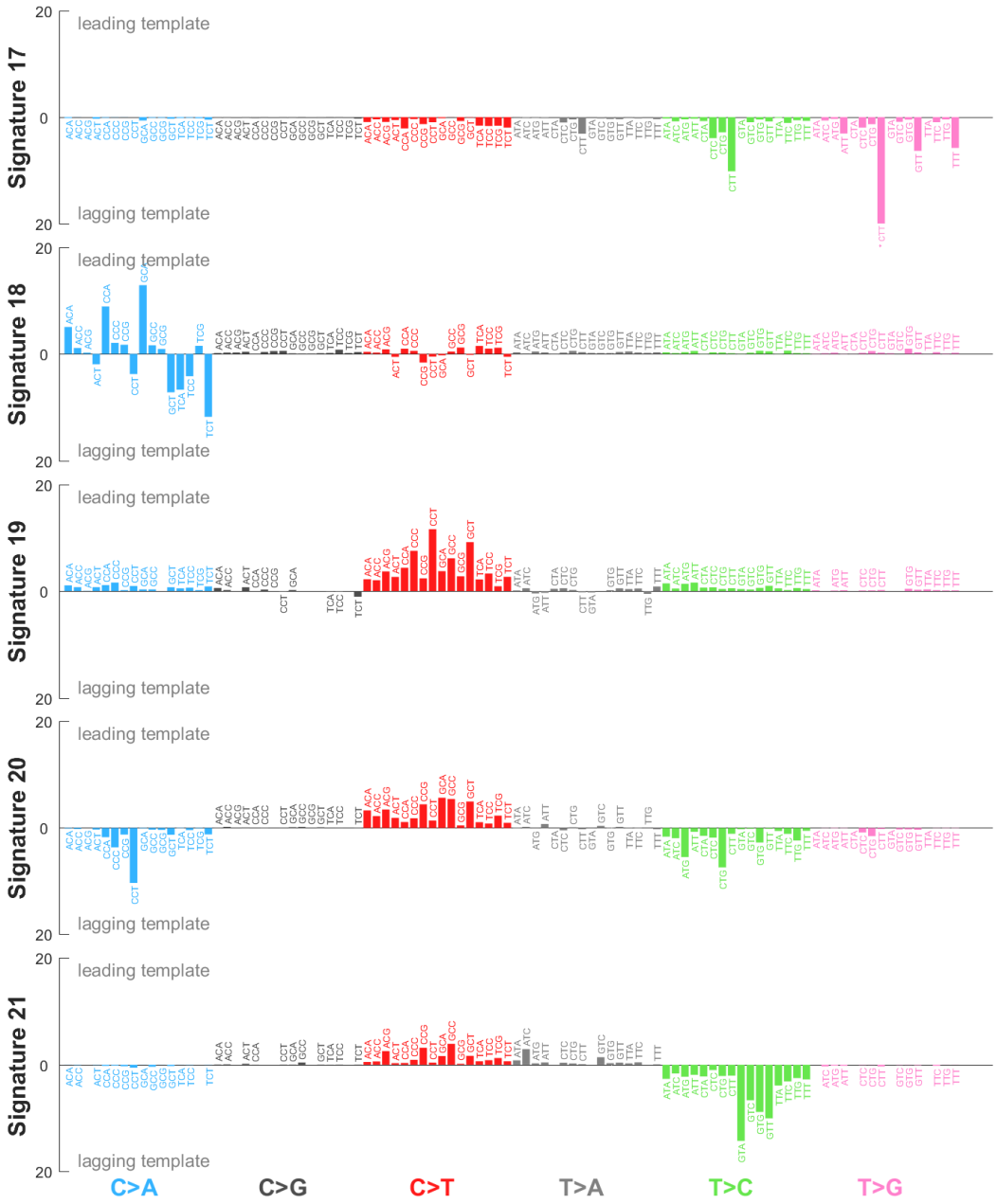

Figure 10.14. Directional signatures 17-21 Each of the 96 mutation types is annotated with a dominant direction: leading (pointing up), or lagging (pointing down). Asterisks indicate mutation types exceeding 20 %.

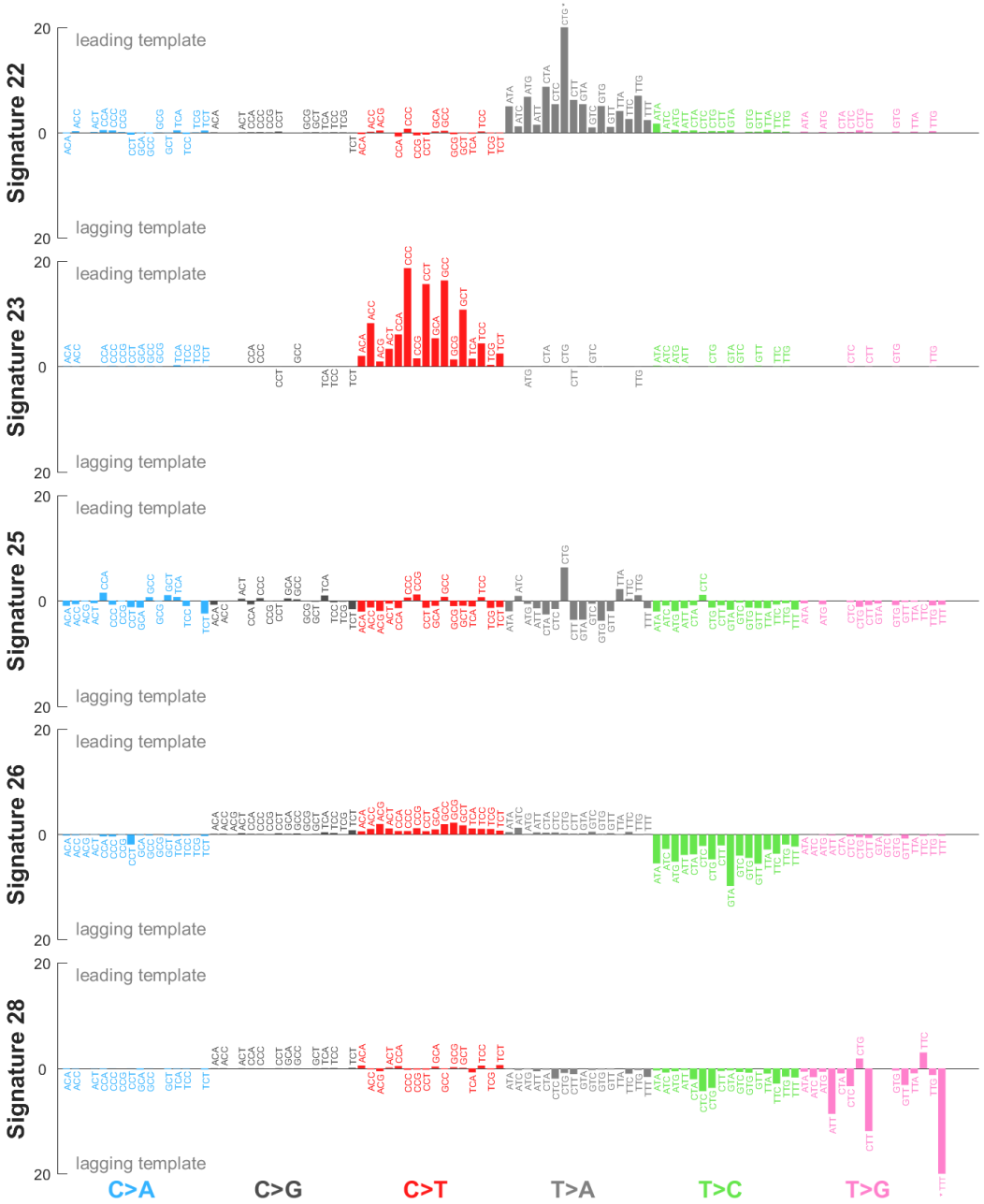

Figure 10.15. Directional signatures 22-28 Each of the 96 mutation types is annotated with a dominant direction: leading (pointing up), or lagging (pointing down). Asterisks indicate mutation types exceeding 20 %.

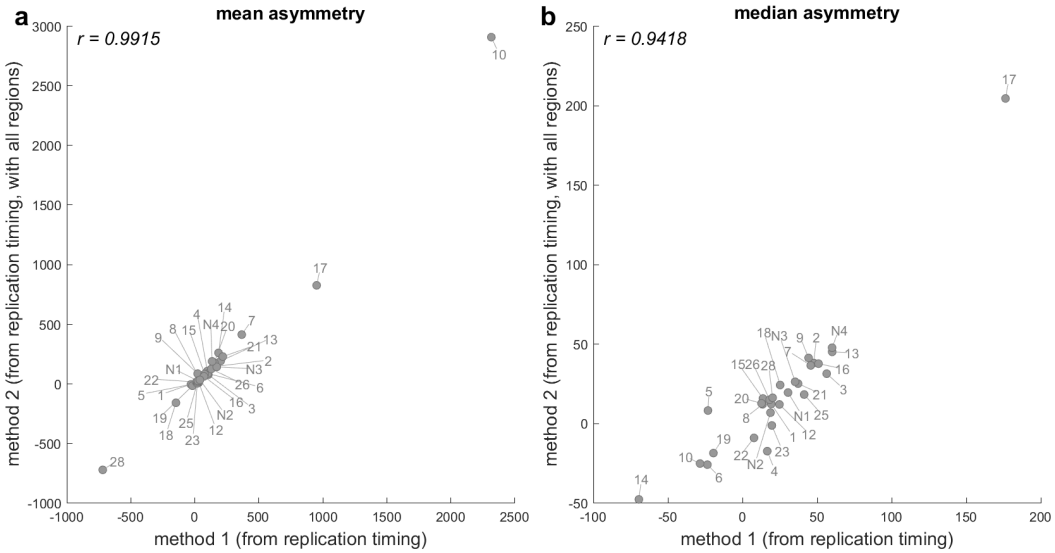

Figure 10.16. Inclusion of protein coding genes results in very similar mutation strand asymmetries. Comparison of resulting mean (a) and median (b) replication strand asymmetry per signature when all regions were taken into account (y axis) vs. when protein-coding genes were excluded (x axis).

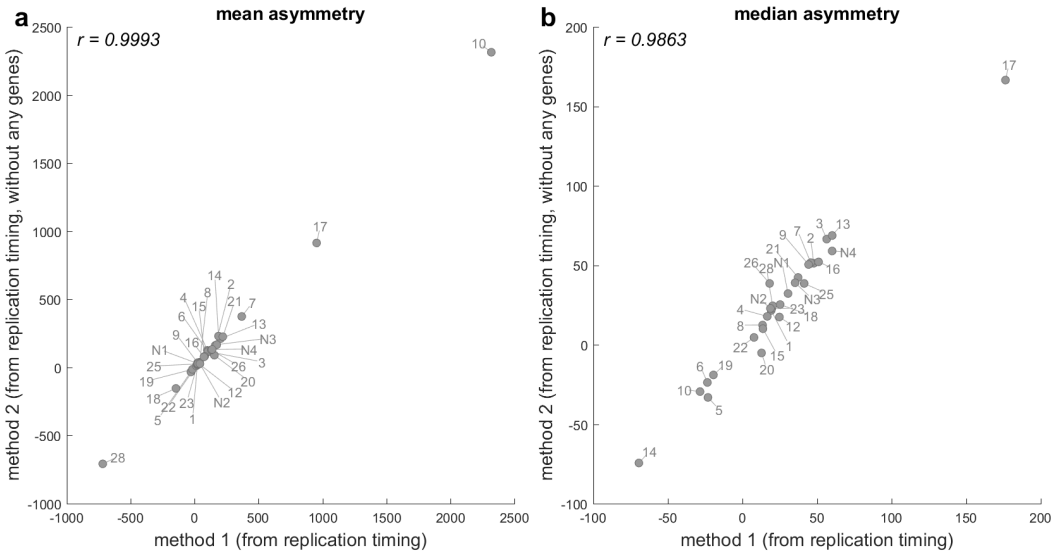

Figure 10.17. Exclusion of non-protein coding genes results in very similar mutation strand asymmetries. Comparison of resulting mean (a) and median (b) replication strand asymmetry per signature when all genes were excluded (y axis) vs. when protein-coding genes were excluded (x axis).

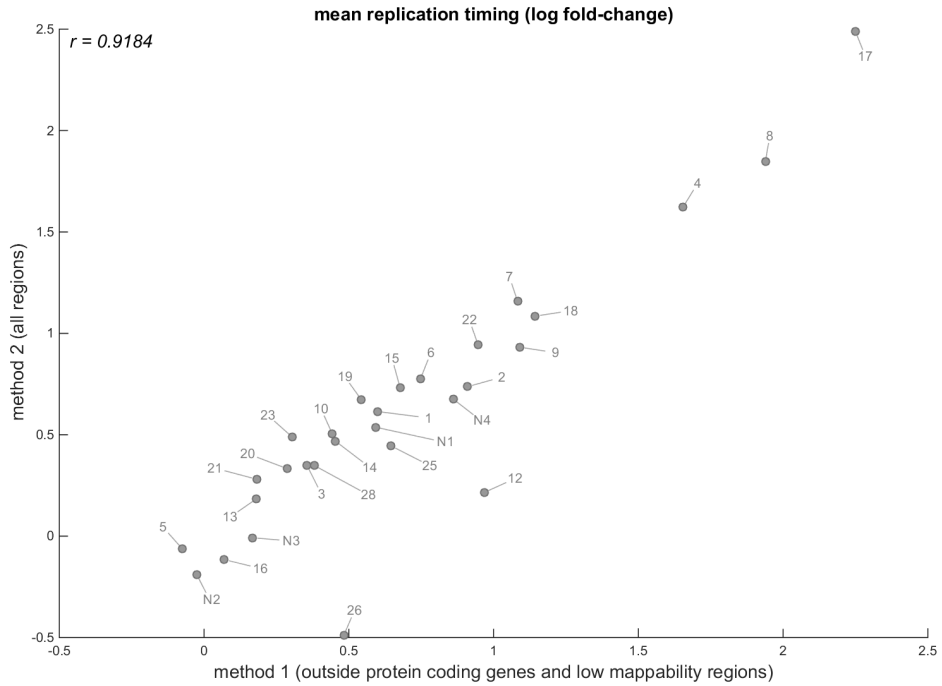

Figure 10.18. Inclusion of protein coding genes results in very similar correlation with replication timing. Comparison of resulting mean correlation with replication timing per signature when all regions were taken into account (y axis) vs. when protein-coding genes and regions with low mappability were excluded (x axis).

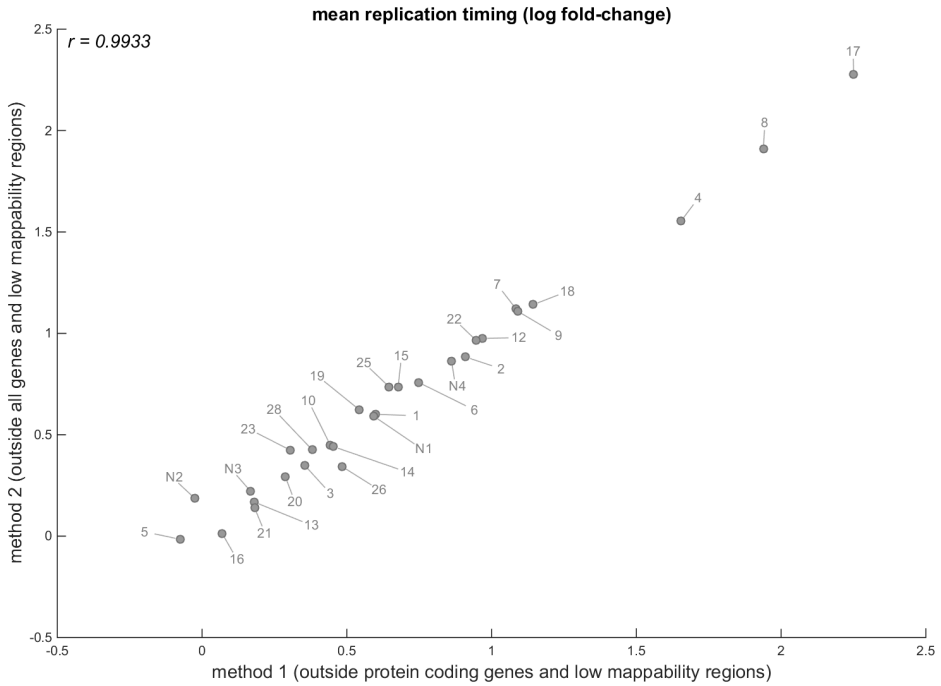

Figure 10.19. Exclusion of non-protein coding genes results in very similar correlation with replication timing. Comparison of resulting mean correlation with replication timing per signature per signature when all genes and regions with low mappability were excluded (y axis) vs. when protein-coding genes and regions with low mappability were excluded (x axis).

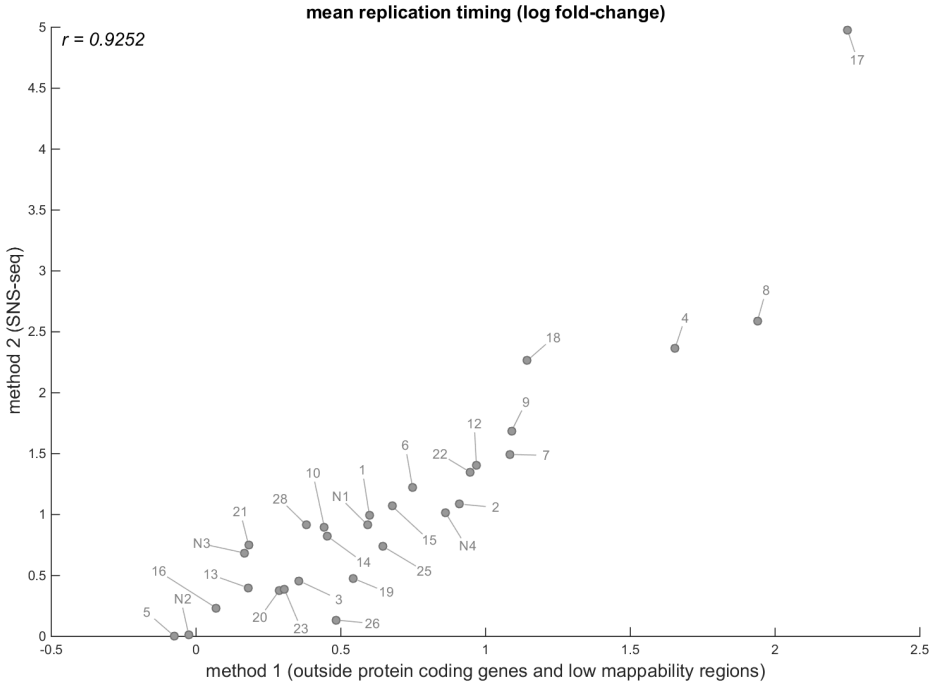

Figure 10.20. Both methods of estimating direction of replication result in very similar correlation with replication timing. Comparison of resulting mean correlation with replication timing per signature in the two methods of measuring replication direction: from replication timing (20 kbp bins annotated as in Haradhvala et al.) vs. from measurements of ORIs using NS-seq (1 kbp bins, see Methods). The absolute values of exposures are different between the two methods since regions around ORIs cover fewer bases (and therefore also fewer mutations).

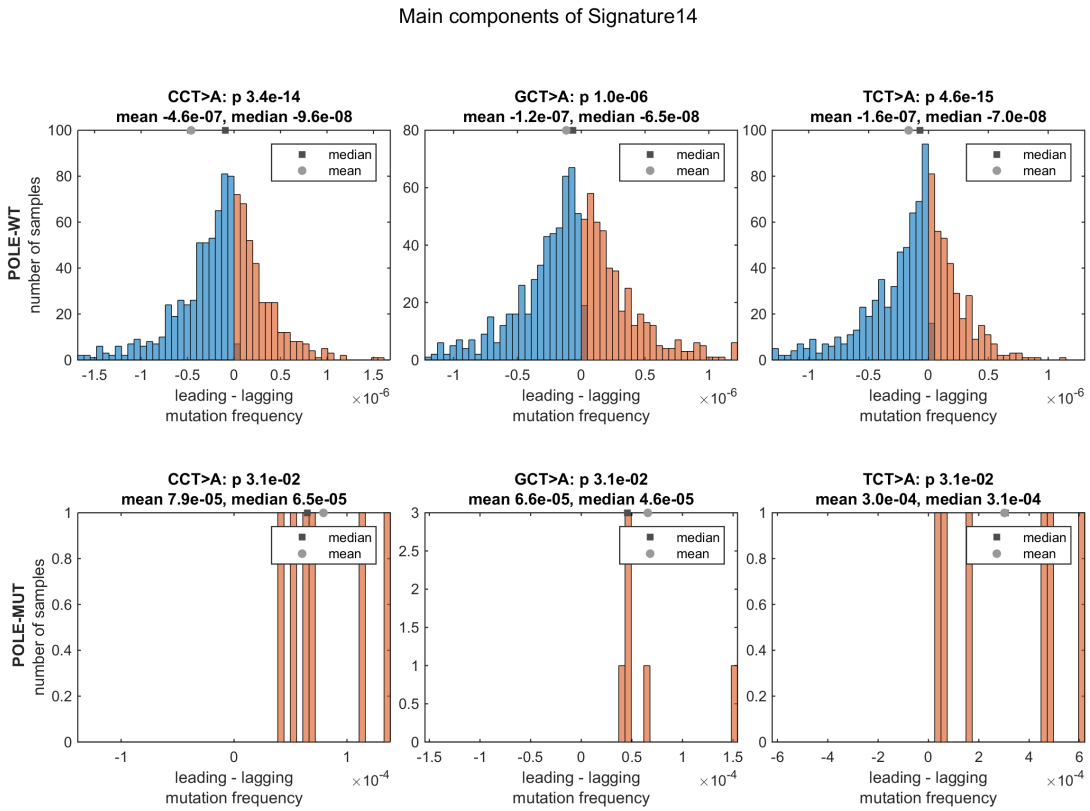

Figure 10.21. Inverse exposure of signature 14 in *POLE*-MUT vs. *POLE*-WT samples. Frequency of mutations in CCT>CAT, GCT>GAT, and TCT>TAT, the three major components of signature 14, is higher on the lagging strand than on the leading strand in *POLE*-WT samples, whereas it is higher on the leading strand in *POLE*-MUT. Only samples exposed to signature 14 (exposure above 10) are shown. Signtest was used to evaluate the mutation frequency difference between the leading and lagging strands.

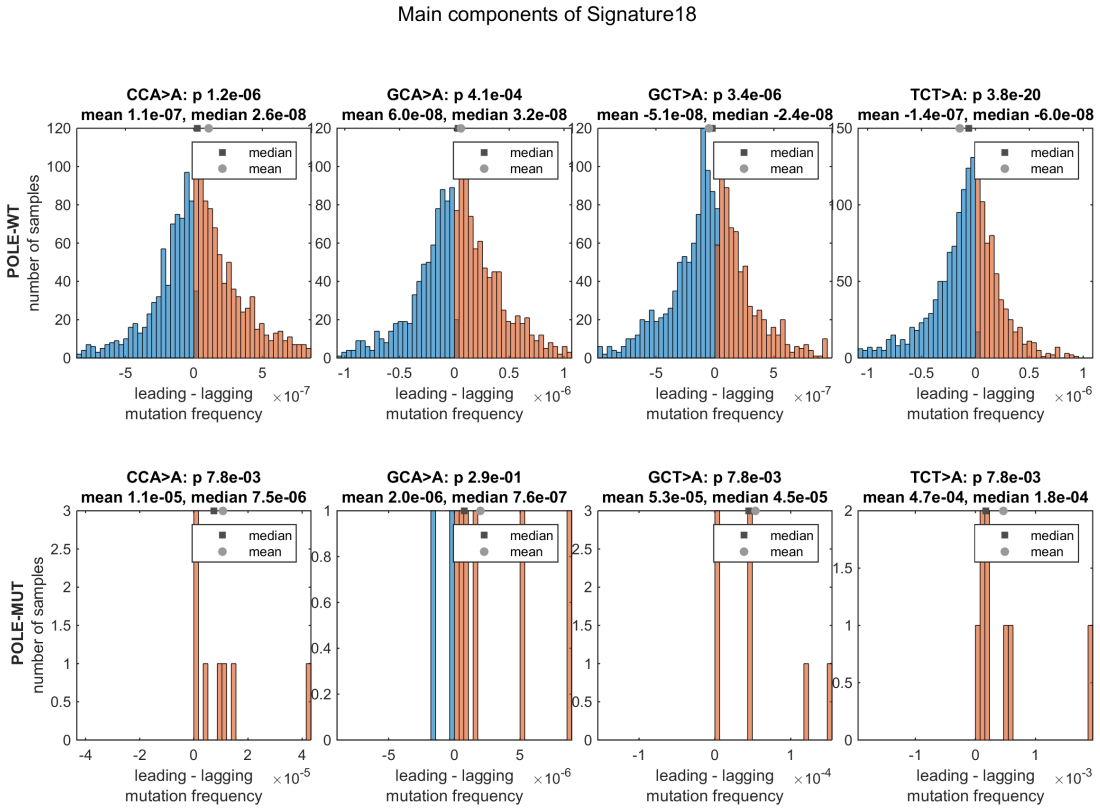

Figure 10.22. Inverse exposure of signature 18 in *POLE*-MUT vs. *POLE*-WT samples. Frequency of mutations in CCA>CAA, GCA>GAA, GCT>GAT, and TCT>TAT, the four major components of signature 18, in *POLE*-WT and *POLE*-MUT. Only samples exposed to signature 18 (exposure above 10) are shown. Signtest was used to evaluate the mutation frequency difference between the leading and lagging strands.

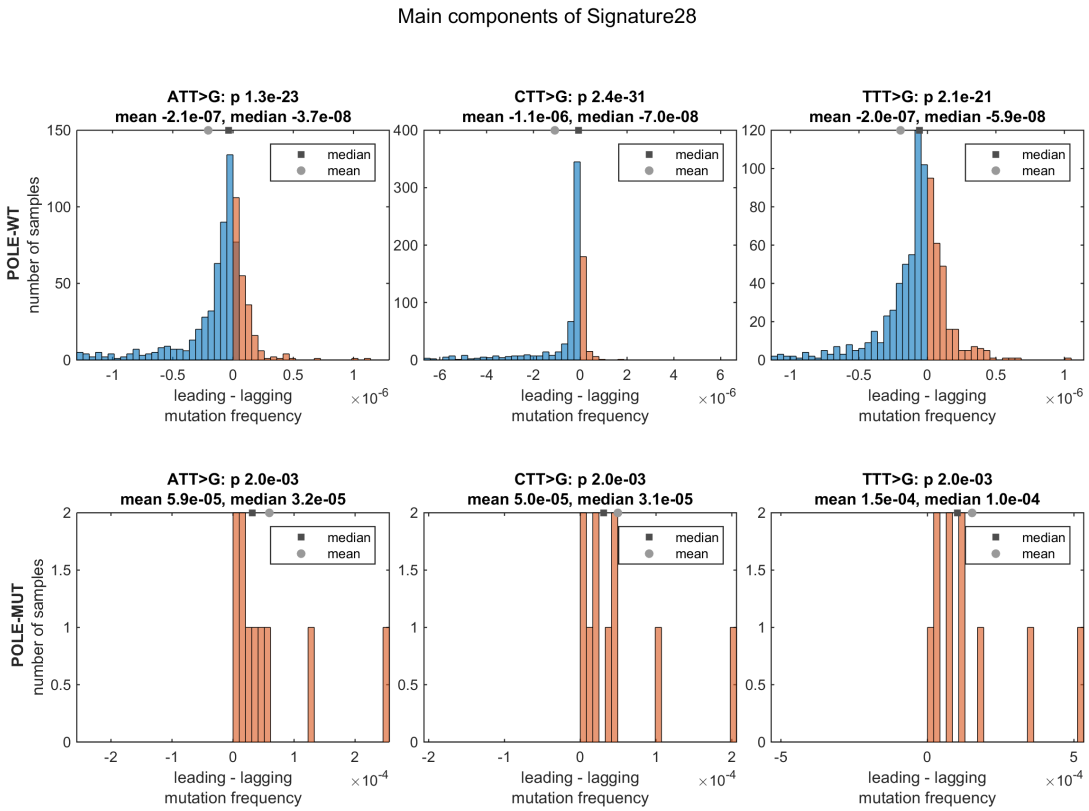

Figure 10.23. Inverse exposure of signature 28 in *POLE*-MUT vs. *POLE*-WT samples. Frequency of mutations in ATT>AGT, CTT>CGT, and TTT>TGT, the three major components of signature 28, is higher on the lagging strand than on the leading strand in *POLE*-WT samples, whereas it is higher on the leading strand in *POLE*-MUT. Only samples exposed to signature 28 (exposure above 10) are shown. Signtest was used to evaluate the mutation frequency difference between the leading and lagging strands.

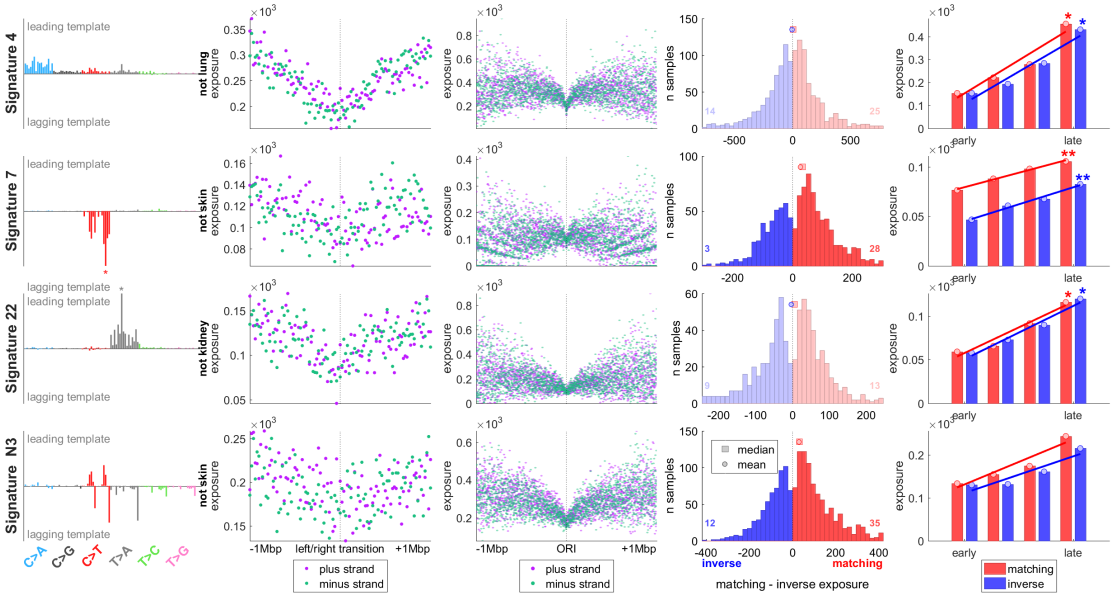

Figure 10.24. Replication strand asymmetry and replication timing in signatures 4, 7, 22, and N3 shown in other than their dominant tissue. Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig. 5.3.

10.3.1 Other mutational processes

A small but significant strand asymmetry was present also in signature 1 (Fig. 10.25). This is in line with our results in Chapter 4.3, supporting the notion that CpG>TpG mutations are slightly enriched on the leading strand even in samples with proficient post-replicative proofreading and repair.

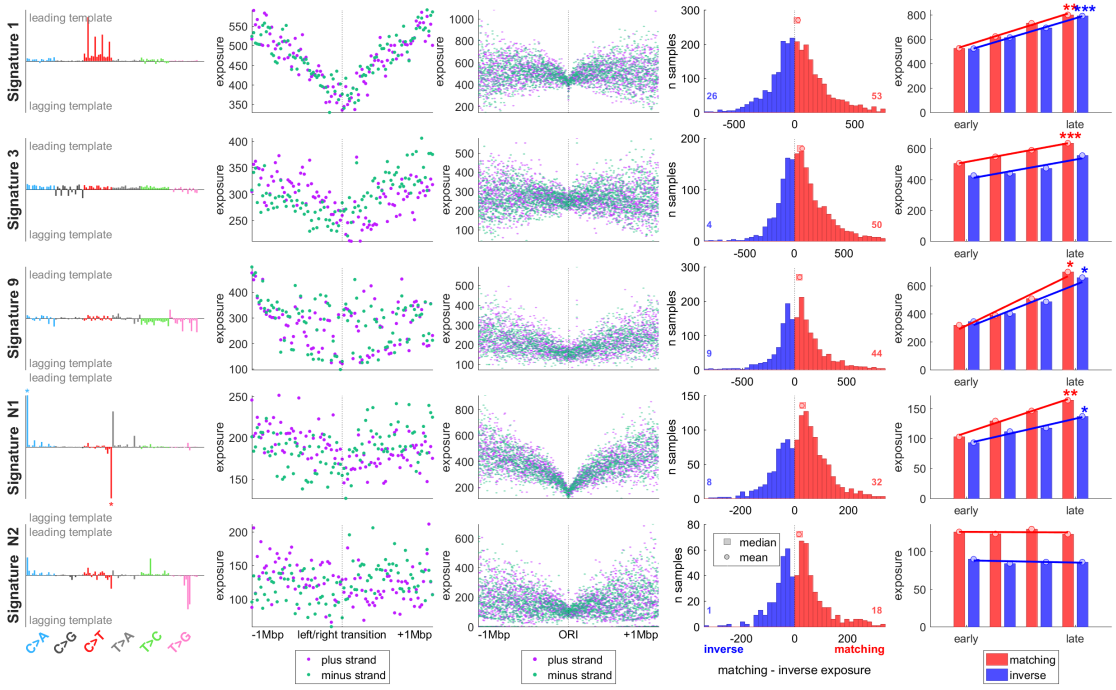

Figure 10.25. Replication strand asymmetry and replication timing in signatures 1, 3, 9, N1, N2. Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig. 5.3.

Signature 1 is also significantly correlated with replication timing. This has been observed previously for SNPs and Human-Chimpanzee substitutions (Stamatoyannopoulos et al., 2009) and suggested to be due to increased methylation in late replicated regions (Chen et al., 2010). The same reason could underlie the correlation with cancer somatic mutations. However, at least part of the correlation could also result from the role of MMR in repairing errors in CpGs introduced by Pol ε, as MMR is thought to be active primarily in the early-replicated regions (Supek and Lehner, 2015; Stamatoyannopoulos et al., 2009; Chen et al., 2010).

A significant replication strand asymmetry is present also in signatures 3 (associated with a failure of DSBR by HR) and 9 (associated with Pol η and AID-mediated somatic hypermutation). Interestingly, the T>G part of the signature 9 shares a substantial similarity with signature 17: both have high NTT>NGT mutations (and signature 9 contains moreover high TTA>TGA and ATA>AGA). It is therefore interesting that in both signatures the shared mutations are enriched on the lagging strand. As signature 9 has been associated with the error-prone activity of Pol η (in the context of immunoglobulin gene hypermutation), this supports our prediction about these mutations being caused by Pol η incorporating 8-oxo-dGTP into DNA 5.4.4.4.

All the four new signatures exhibited a strong replication strand asymmetry (Fig. 10.26). Both signatures N1 and N2 were detected in acute myeloid leukemia (AML), suggesting a novel replication-modulated source of mutagenesis in AML cancers.

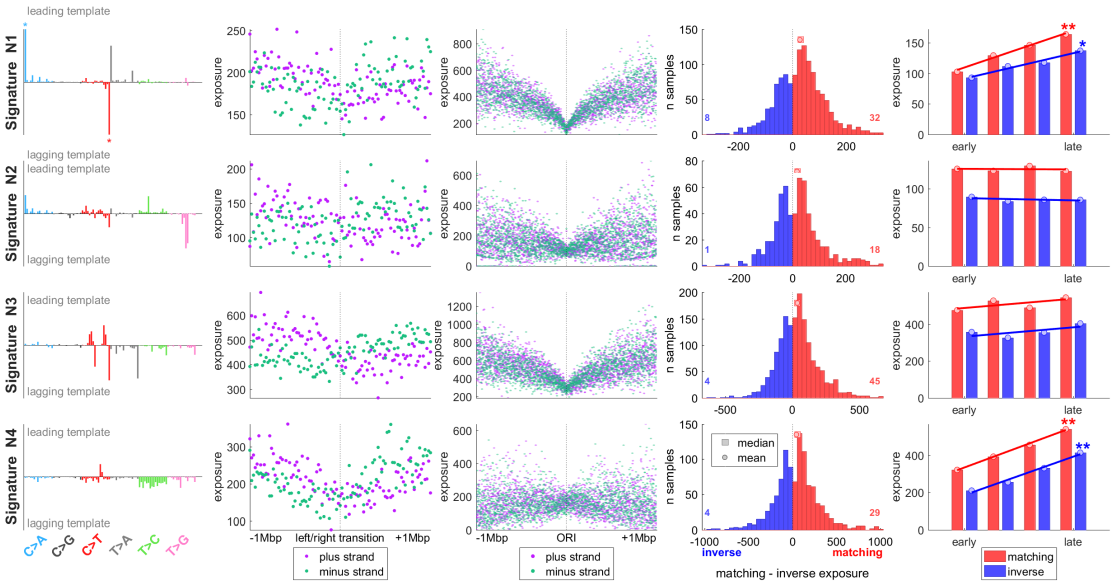

Figure 10.26. Replication strand asymmetry and replication timing in the four new signatures: N1, N2, N3, and N4. Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig. 5.3.

While one of the two liver signatures (signature 16) exhibits a weak but significant replication strand asymmetry, the other (signature 12) is significantly correlated with replication timing. Compared to the weak replication strand asymmetry, these two signatures were previously shown to have a very strong transcription strand bias (Alexandrov et al., 2013a). Interestingly, the T>C mutations were observed not only enriched on the transcribed strand, but also depleted on the non-transcribed strand, compared to intergenic regions; which was suggested to be caused by a transcription-coupled damage (Haradhvala et al., 2016).

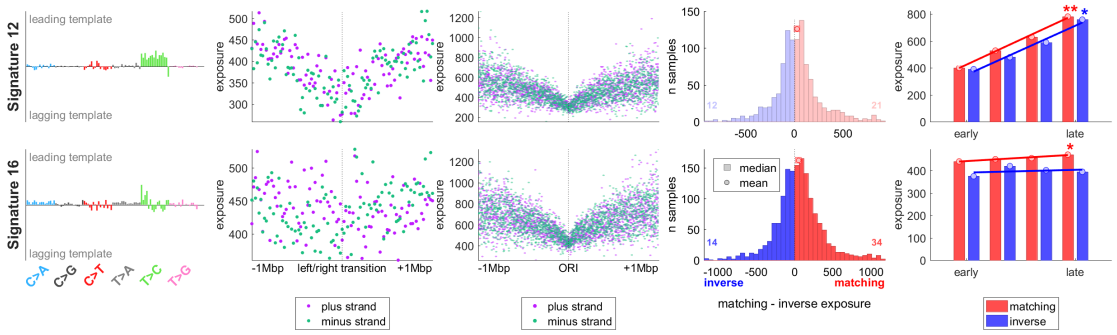

Figure 10.27. Replication strand asymmetry and replication timing in liver-associated signatures. Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig. 5.3.

Finally, only five signatures exhibited no (or weak) replication strand asymmetry (signatures 5, 8, 19, 23, and 25) (Fig 10.28). Even in these cases, an effect of replication cannot be entirely ruled out: signatures 19, 23, and 25 were only detected in a small number of samples, reducing the statistical power to detect replication bias. Notably, signature 8 was the second most correlated with replication timing (Fig. 5.4C). Only signature 5 clearly did not show any effects of replication, while being present in a sufficient number of samples (Fig 10.28). Signature 5 is frequent across different cancer types (Wellcome Trust Sanger Institute, 2017) and it is the second of the only two signatures correlated with age at diagnosis (Alexandrov et al., 2015). The biological processes leading to this clock-like signature are currently completely unknown. Our results suggest that replication is not likely to be involved in the aetiology.

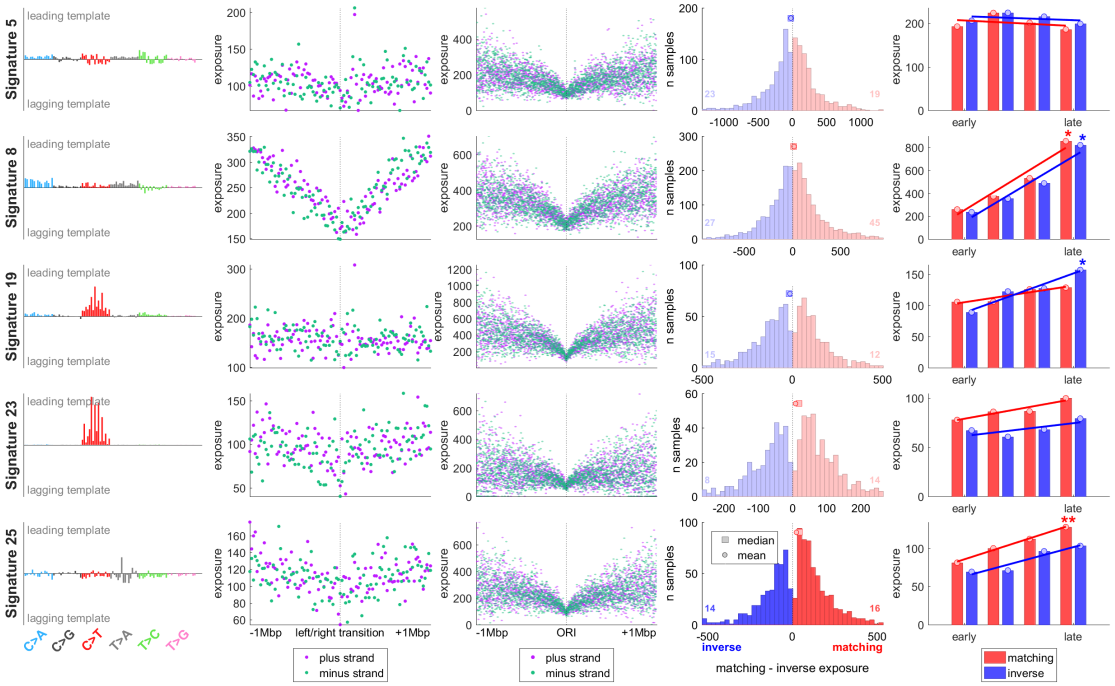

Figure 10.28. Replication strand asymmetry and replication timing in signatures with little or no asymmetry: 5, 8, 19, 23, and 25. Columns show directional signature (column 1), distribution around timing transition regions (column 2) and around replication origins (column 3), per-patient mutation strand asymmetry (column 4; non-significant asymmetry is shown in light-coloured histogram) and correlation with replication timing (column 5), as described in Fig. 5.3.

Baba yetu, yetu uliye
Mjina lako elitukuzwe.

— Christopher Tin *Baba Yetu*

11 Appendix: Abbreviations

- APOBEC: apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
- AML: acute myeloid leukemia
- AP: abasic site
- ATP: adenosine triphosphate
- BER: base excision repair
- BO: barrett's oesophagus
- BPDE: benzo[a]pyrene diol-epoxide adduct
- BS-seq: bisulfite sequencing
- CIMP: CpG island hypermethylation phenotype
- CLL: chronic lymphoid leukaemia
- CPD: cyclobutane pyrimidine dimer
- CRC: colorectal
- CS: Cockayne syndrome
- CSR: class switch recombination
- CTCF: CCCTC-binding factor
- DDT: DNA damage tolerance
- DMR: differentially methylated region
- DNA: deoxyribonucleic acid
- dNTP: deoxyribonucleoside triphosphate

- DSB: double-strand break
- DSBR: double-strand break repair
- EAC: oesophageal adenocarcinoma
- ESC: embryonic stem cell
- FPKM: fragments per kilobase per million sequenced reads
- GBM: glioblastoma
- GERD: gastro-esophageal reflux disease
- GG-NER: global genome nucleotide excision repair
- GLM: generalised linear model
- HDAC: histone deacetylase
- HMPC: hematopoietic multipotent progenitor cell
- HNPCC: hereditary nonpolyposis colorectal cancer
- HPLC: high-performance liquid chromatography
- HR: homologous recombination
- ICGC: International Cancer Genome Consortium
- IQR: interquartile range
- KO: knockout
- LGG: low grade glioma
- LOI: loss-of-imprinting
- MACS: Model-based Analysis for ChIP-Seq
- MAP: MUTYH-associated polyposis
- MDB: Medulloblastoma
- MDS: multidimensional scaling
- MEF: mouse embryonic fibroblasts
- MMR: mismatch repair
- MSI: (samples with) microsatellite instability
- MSS: microsatellite stable samples, samples without microsatellite instability
- NDR: nucleosome depleted region
- NER: nucleotide excision repair
- NGS: next-generation sequencing
- NHEJ: non-homologous end joining

- NMF: non-negative matrix factorisation
- NRB: Neuroblastoma
- NS: nascent strand
- OK-seq: Okazaki fragment sequencing
- ORC: origin recognition complex
- ORI: origin of replication
- oxBS-seq: oxidative bisulfite sequencing
- PA: Pilocytic astrocytoma
- PAH: polycyclic aromatic hydrocarbon
- PCA: principal component analysis
- PCNA: proliferating cell nuclear antigen
- PCR: polymerase chain reaction
- POLE-MUT: hypermutated samples with a mutation in POLE
- PPAP: polymerase proofreading-associated polyposis
- RNA: ribonucleic acid
- ROS: reactive oxygen species
- RPA: replication protein A
- SAM: S-adenosyl methionine
- SHM: somatic hypermutation
- siRNA: small interfering RNA
- SNP: single-nucleotide polymorphism
- SNS-seq: short nascent strand sequencing
- SNV: single-nucleotide variant
- SSB: single-strand break
- TAB-seq: TET-assisted bisulfite sequencing
- TCGA: The Cancer Genome Atlas
- TC-NER: transcription-coupled nucleotide excision repair
- TCR: transcription-coupled repair
- TDG: thymine DNA glycosylase
- TES: transcription end site
- TET: Ten-eleven translocation enzyme

- TF: transcription factor
- TFBS: transcription factor binding site
- TLS: translesion synthesis
- TMZ: Temozolomide
- TS: template switching
- TSS: transcription start site
- TTR: temporal transition region
- UDG, UNG: uracil DNA glycosylase
- UV: ultraviolet
- WES: whole-exome sequencing
- WXS: whole-genome sequencing
- XP: Xeroderma pigmentosum
- XPC: Xeroderma pigmentosum, complementation group C
- XPV: Xeroderma pigmentosum, variant
- 6-4PP: pyrimidine (6-4) pyrimidone photoproduct

In this library I could lose myself
Transports, gateways on every shelf
Dark words, bright words of ice and fire
As if an angel did descend and use the writer as a pen

— The Waterboys *Universal Hall*

References

Adamson R. H., Correa P., and Dalgard D. W. Occurrence of a primary liver carcinoma in a Rhesus monkey fed aflatoxin B 1 . *Journal of the National Cancer Institute*, 50(2): 549–53, feb 1973. ISSN 0027-8874.

Adar S., Hu J., Lieb J. D., and Sancar A. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, page 201603388, 2016. ISSN 1091-6490. doi: 10.1073/pnas.1603388113.

Äijö T., Huang Y., Mannerström H., Chavez L., Tsagaratou A., Rao A., and Lähdesmäki H. A probabilistic generative model for quantification of DNA modifications enables analysis of demethylation pathways. *Genome biology*, 17(1):49, mar 2016a. ISSN 1474-760X. doi: 10.1186/s13059-016-0911-6.

Äijö T., Yue X., Rao A., and Lähdesmäki H. LuxGLM: A probabilistic covariate model for quantification of DNA methylation modifications with complex experimental designs. *Bioinformatics*, 32(17):i511–i519, 2016b. ISSN 14602059. doi: 10.1093/bioinformatics/btw468.

Akiyama Y., Maesawa C., Ogasawara S., Terashima M., and Masuda T. Cell-type-specific repression of the maspin gene is disrupted frequently by demethylation at the promoter region in gastric intestinal metaplasia and cancer cells. *The American journal of pathology*, 163(5):1911–9, nov 2003. ISSN 0002-9440. doi: 10.1016/S0002-9440(10)63549-3.

Akre M. K., Starrett G. J., Quist J. S., Temiz N. A., Carpenter M. A., Tutt A. N. J., Grigoriadis A., and Harris R. S. Mutation processes in 293-based clones overexpressing the

DNA cytosine deaminase APOBEC3B. *PLoS ONE*, 11(5):1–17, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0155391.

Aladjem M. I. and Redon C. E. Order from clutter: selective interactions at mammalian replication origins. *Nature Reviews Genetics*, 18(2):101–116, 2016. ISSN 1471-0056. doi: 10.1038/nrg.2016.141.

Albertson T. M., Ogawa M., Bugni J. M., Hays L. E., Chen Y., Wang Y., Treuting P. M., Heddle J. A., Goldsby R. E., and Preston B. D. DNA polymerase epsilon and delta proofreading suppress discrete mutator and cancer phenotypes in mice. *Proceedings of the National Academy of Sciences of the United States of America*, 106(40):17101–4, 2009. ISSN 1091-6490. doi: 10.1073/pnas.0907147106.

Alexandrov L. B. Understanding the origins of human cancer. *Science*, 350(6265), 2015. ISSN 0036-8075. doi: 10.1126/science.aad7363.

Alexandrov L. B., Nik-Zainal S., Wedge D. C., Aparicio S. A. J. R., Behjati S., Biankin A. V., Bignell G. R., Bolli N., Borg A., Børresen-Dale A.-L., Boyault S., Burkhardt B., Butler A. P., Caldas C., Davies H. R., Desmedt C., Eils R., Eyfjörd J. E., Foekens J. A., Greaves M., Hosoda F., Hutter B., Ilicic T., Imbeaud S., Imielinski M., Imielinsk M., Jäger N., Jones D. T. W., Jones D., Knappskog S., Kool M., Lakhani S. R., López-Otín C., Martin S., Munshi N. C., Nakamura H., Northcott P. A., Pajic M., Papaemmanuil E., Paradiso A., Pearson J. V., Puente X. S., Raine K., Ramakrishna M., Richardson A. L., Richter J., Rosenstiel P., Schlesner M., Schumacher T. N., Span P. N., Teague J. W., Totoki Y., Tutt A. N. J., Valdés-Mas R., van Buuren M. M., van 't Veer L., Vincent-Salomon A., Waddell N., Yates L. R., Zucman-Rossi J., Futreal P. A., McDermott U., Lichter P., Meyerson M., Grimmond S. M., Siebert R., Campo E., Shibata T., Pfister S. M., Campbell P. J., and Stratton M. R. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–21, aug 2013a. ISSN 1476-4687. doi: 10.1038/nature12477.

Alexandrov L. B., Nik-Zainal S., Wedge D. C., Campbell P. J., and Stratton M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3 (1):246–59, jan 2013b. ISSN 2211-1247. doi: 10.1016/j.celrep.2012.12.008.

Alexandrov L. B., Jones P. H., Wedge D. C., Sale J. E., and Peter J. Clock-like mutational processes in human somatic cells. *Nature*, 47(12):1402–1407, 2015. ISSN 1061-4036. doi: 10.1038/ng.3441.

Alexandrov L. B., Ju Y. S., Haase K., Van Loo P., Martincorena I., Nik-Zainal S., Totoki Y., Fujimoto A., Nakagawa H., Shibata T., Campbell P. J., Vineis P., Phillips D. H., and Stratton M. R. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–22, 2016. ISSN 1095-9203. doi: 10.1126/science.aag0299.

Amary M. F., Bacsi K., Maggiani F., Damato S., Halai D., Berisha F., Pollock R., O'Donnell P., Grigoriadis A., Diss T., Eskandarpour M., Presneau N., Hogendoorn P. C., Futreal A., Tirabosco R., and Flanagan A. M. IDH1 and IDH2 mutations are frequent events in central chondrosarcoma and central and periosteal chondromas but not in other mesenchymal tumours. *The Journal of pathology*, 224(3):334–43, jul 2011a. ISSN 1096-9896. doi: 10.1002/path.2913.

Amary M. F., Damato S., Halai D., Eskandarpour M., Berisha F., Bonar F., McCarthy S., Fantin V. R., Straley K. S., Lobo S., Aston W., Green C. L., Gale R. E., Tirabosco R., Futreal A., Campbell P., Presneau N., and Flanagan A. M. Ollier disease and Maffucci syndrome are caused by somatic mosaic mutations of IDH1 and IDH2. *Nature genetics*, 43(12):1262–5, dec 2011b. ISSN 1546-1718. doi: 10.1038/ng.994.

Ames B. N., Shigenaga M. K., and Hagen T. M. Oxidants, antioxidants, and the degenerative diseases of aging. *Proceedings of the National Academy of Sciences of the United States of America*, 90(17):7915–7922, 1993. ISSN 0027-8424. doi: 10.1073/pnas.90.17.7915.

Andrianova M. A., Bazykin G. A., Nikolaev S. I., and Seplyarskiy V. B. Human mismatch repair system balances mutation rates between strands by removing more mismatches

from the lagging strand. *Genome Research*, 27(8):1336–1343, 2017. ISSN 15495469. doi: 10.1101/gr.219915.116.

Aoki Y., Hashimoto A., Sugawara Y., Hiyoshi-Arai K., Goto S., Masumura K., and Nohmi T. Alterations in the mutagenicity and mutation spectrum induced by benzo[a]pyrene instilled in the lungs of gpt delta mice of various ages. *Genes and Environment*, 37(1), 2015. ISSN 18807062 18807046. doi: 10.1186/s41021-015-0004-x.

Aran D., Toperoff G., Rosenberg M., and Hellman A. Replication timing-related and gene body-specific methylation of active human genes. *Human Molecular Genetics*, 20(4): 670–680, 2011. ISSN 09646906. doi: 10.1093/hmg/ddq513.

Arand J., Spieler D., Karius T., Branco M. R., Meilinger D., Meissner A., Jenuwein T., Xu G., Leonhardt H., Wolf V., and Walter J. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genetics*, 8(6), 2012. ISSN 15537390. doi: 10.1371/journal.pgen.1002750.

Arnheim N. and Calabrese P. Understanding what determines the frequency and pattern of human germline mutations. *Nature Reviews Genetics*, 10(7):478–488, 2009. ISSN 1471-0056. doi: 10.1038/nrg2529.

Avkin S., Goldsmith M., Velasco-Miguel S., Geacintov N., Friedberg E. C., and Livneh Z. Quantitative analysis of translesion DNA synthesis across a benzo[a]pyrene-guanine adduct in mammalian cells: The role of DNA polymerase. *Journal of Biological Chemistry*, 279(51):53298–53305, 2004. ISSN 00219258. doi: 10.1074/jbc.M409155200.

Bachman M., Uribe-Lewis S., Yang X., Williams M., and Murrell A. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nature Chemistry*, 6(12):1049–1055, 2014. ISSN 1755-4330. doi: 10.1038/nchem.2064.

Bachman M., Uribe-Lewis S., Yang X., Burgess H. E., Iurlaro M., Reik W., Murrell A., and Balasubramanian S. 5-Formylcytosine can be a stable DNA modification in mammals. *Nature chemical biology*, 11(8):1–4, 2015. ISSN 1552-4450. doi: 10.1038/nchembio.1848.

Bacolla A., Cooper D. N., and Vasquez K. M. Mechanisms of base substitution mutagenesis in cancer genomes. *Genes*, 5(1):108–146, 2014. ISSN 20734425. doi: 10.3390/genes5010108.

Baeissa H., Benstead-Hume G., Richardson C. J., and Pearl F. M. G. Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. *Oncotarget*, 8 (13):21290–21304, 2017. ISSN 1949-2553. doi: 10.18632/oncotarget.15514.

Bai L. and Morozov A. V. Gene regulation by nucleosome positioning. *Trends in Genetics*, 26(11):476–483, 2010. ISSN 01689525. doi: 10.1016/j.tig.2010.08.003.

Bak S. T., Sakellariou D., and Pena-Diaz J. The dual nature of mismatch repair as antimutator and mutator: For better or for worse. *Frontiers in Genetics*, 5(AUG):1–12, 2014. ISSN 16648021. doi: 10.3389/fgene.2014.00287.

Baker A., Audit B., Chen C. L., Moindrot B., Leleu A., Guilbaud G., Rappailles A., Vaillant C., Goldar A., Mongelard F., D'Aubenton-Carafa Y., Hyrien O., Thermes C., and Arneodo A. Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. *PLoS Computational Biology*, 8(4), 2012. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002443.

Balmus I. M., Ciobica A., Trifan A., and Stanciu C. The implications of oxidative stress and antioxidant therapies in Inflammatory Bowel Disease: Clinical aspects and animal models. *Saudi journal of gastroenterology : official journal of the Saudi Gastroenterology Association*, 22(1):3–17, 2016. ISSN 1998-4049. doi: 10.4103/1319-3767.173753.

Barbari S. R. and Shcherbakova P. V. Replicative DNA polymerase defects in human cancers: Consequences, mechanisms, and implications for therapy. *DNA Repair*, 56 (June):16–25, 2017. ISSN 15687856. doi: 10.1016/j.dnarep.2017.06.003.

Bardella C., Al-Dalahmah O., Krell D., Brazauskas P., Al-Qahtani K., Tomkova M., Adam J., Serres S., Lockstone H., Freeman-Mills L., Pfeffer I., Sibson N., Goldin R., Schuster-Böeckler B., Pollard P. J., Soga T., McCullagh J. S., Schofield C. J., Mulholland P., Ansorge O., Kriaucionis S., Ratcliffe P. J., Szele F. G., and Tomlinson I.

Expression of Idh1(R132H) in the Murine Subventricular Zone Stem Cell Niche Recapitulates Features of Early Gliomagenesis. *Cancer cell*, 30(4):578–594, oct 2016. ISSN 1878-3686. doi: 10.1016/j.ccell.2016.08.017.

Barski A., Cuddapah S., Cui K., Roh T. Y., Schones D. E., Wang Z., Wei G., Chepelev I., and Zhao K. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, 2007. ISSN 00928674. doi: 10.1016/j.cell.2007.05.009.

Bass A. J., Lawrence M. S., Brace L. E., Ramos A. H., Drier Y., Cibulskis K., Sougnez C., Voet D., Saksena G., Sivachenko A., Jing R., Parkin M., Pugh T., Verhaak R. G., Stransky N., Boutin A. T., Barretina J., Solit D. B., Vakiani E., Shao W., Mishina Y., Warmuth M., Jimenez J., Chiang D. Y., Signoretti S., Kaelin W. G., Spardy N., Hahn W. C., Hoshida Y., Ogino S., Depinho R. A., Chin L., Garraway L. A., Fuchs C. S., Baselga J., Tabernero J., Gabriel S., Lander E. S., Getz G., and Meyerson M. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nature genetics*, 43(10):964–8, 2011. ISSN 1546-1718. doi: 10.1038/ng.936.

Bauer N. C., Corbett A. H., and Doetsch P. W. The current state of eukaryotic DNA base damage and repair. *Nucleic acids research*, 43(21):10083–101, 2015. ISSN 1362-4962. doi: 10.1093/nar/gkv1136.

Beck S. and Rakyan V. K. The methylome: approaches for global DNA methylation profiling. *Trends in Genetics*, 24(5):231–237, 2008. ISSN 01689525. doi: 10.1016/j.tig.2008.01.006.

Behjati S., Huch M., van Boxtel R., Karthaus W., Wedge D. C., Tamuri A. U., Martincorena I., Petljak M., Alexandrov L. B., Gundem G., Tarpey P. S., Roerink S., Blokker J., Maddison M., Mudie L., Robinson B., Nik-Zainal S., Campbell P., Goldman N., van de Wetering M., Cuppen E., Clevers H., and Stratton M. R. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, 513(7518):422–425, 2014. ISSN 0028-0836. doi: 10.1038/nature13448.

Behjati S., Gundem G., Wedge D. C., Roberts N. D., Tarpey P. S., Cooke S. L., Van Loo P., Alexandrov L. B., Ramakrishna M., Davies H., Nik-Zainal S., Hardy C., Latimer C., Raine K. M., Stebbings L., Menzies A., Jones D., Shepherd R., Butler A. P., Teague J. W., Jorgensen M., Khatri B., Pillay N., Shlien A., Futreal P. A., Badie C., Cooper C. S., Eeles R. A., Easton D., Foster C., Neal D. E., Brewer D. S., Hamdy F., Lu Y.-J., Lynch A. G., Massi C. E., Ng A., Whitaker H. C., Yu Y., Zhang H., Bancroft E., Berney D., Camacho N., Corbishley C., Dadaev T., Dennis N., Dudderidge T., Edwards S., Fisher C., Ghori J., Gnanapragasam V. J., Greenman C., Hawkins S., Hazell S., Howat W., Karaszi K., Kay J., Kote-Jarai Z., Kremeyer B., Kumar P., Lambert A., Leongamornlert D., Livni N., Luxton H., Matthews L., Mayer E., Merson S., Nicol D., Ogden C., O'Meara S., Pelvender G., Shah N. C., Tavare S., Thomas S., Thompson A., Verrill C., Warren A., Zamora J., McDermott U., Bova G. S., Richardson A. L., Flanagan A. M., Stratton M. R., and Campbell P. J. Mutational signatures of ionizing radiation in second malignancies. *Nature Communications*, 7, 2016. ISSN 2041-1723. doi: 10.1038/ncomms12605.

Bell O., Tiwari V. K., Thomä N. H., and Schübeler D. Determinants and dynamics of genome accessibility. *Nature Reviews Genetics*, 12(8):554–564, 2011. ISSN 1471-0056. doi: 10.1038/nrg3017.

Bellacosa A. and Drohat A. C. Role of base excision repair in maintaining the genetic and epigenetic integrity of CpG sites. *DNA Repair*, 32:33–42, 2015. ISSN 15687856. doi: 10.1016/j.dnarep.2015.04.011.

Benigni R. and Bossa C. Mechanisms of chemical carcinogenicity and mutagenicity: A review with implications for predictive toxicology. *Chemical Reviews*, 111(4):2507–2536, 2011. ISSN 00092665. doi: 10.1021/cr100222q.

Berger S. L. The complex language of chromatin regulation during transcription. *Nature*, 447(7143):407–412, 2007. ISSN 0028-0836. doi: 10.1038/nature05915.

Bergoglio V., Boyer A. S., Walsh E., Naim V., Legube G., Lee M. Y., Rey L., Rosselli F., Cazaux C., Eckert K. A., and Hoffmann J. S. DNA synthesis by pol η promotes fragile

site stability by preventing under-replicated DNA in mitosis. *Journal of Cell Biology*, 201(3):395–408, 2013. ISSN 00219525. doi: 0.1083/jcb.201207066.

Bernstein H., Bernstein C., Payne C., Dvorakova K., and Garewal H. Bile acids as carcinogens in human gastrointestinal cancers. *Mutation Research/Reviews in Mutation Research*, 589(1):47–65, 2005. ISSN 13835742. doi: 10.1016/j.mrrev.2004.08.001.

Besnard E., Babled A., Lapasset L., Milhavet O., Parrinello H., Dantec C., Marin J.-M., and Lemaitre J.-M. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nature structural & molecular biology*, 19(8):837–844, aug 2012.

Bi X. Mechanism of DNA damage tolerance. *World Journal of Biological Chemistry*, 6(3): 48, 2015. ISSN 1949-8454. doi: 10.4331/wjbc.v6.i3.48.

Bird A. DNA methylation patterns and epigenetic memory. *Genes & development*, 16(1): 6–21, jan 2002. ISSN 0890-9369. doi: 10.1101/gad.947102.

Bird A. Perceptions of epigenetics. *Nature*, 447(7143):396–8, 2007. ISSN 1476-4687. doi: 10.1038/nature05913.

Bird A. P. and Taggart M. H. Variable patterns of total DNA and rDNA methylation in animals. *Nucleic Acids Research*, 8(7):1485–1497, apr 1980. ISSN 03051048. doi: 10.1093/nar/8.7.1485.

Blokzijl F., de Ligt J., Jager M., Sasselli V., Roerink S., Sasaki N., Huch M., Boymans S., Kuijk E., Prins P., Nijman I. J., Martincorena I., Mokry M., Wiegerinck C. L., Middendorp S., Sato T., Schwank G., Nieuwenhuis E. E. S., Verstegen M. M. A., van der Laan L. J. W., de Jonge J., IJzermans J. N. M., Vries R. G., van de Wetering M., Stratton M. R., Clevers H., Cuppen E., and van Boxtel R. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, 538(7624):260–264, 2016. ISSN 1476-4687. doi: 10.1038/nature19768.

Bock C., Halbritter F., Carmona F. J., Tierling S., Datlinger P., Assenov Y., Berdasco M., Bergmann A. K., Booher K., Busato F., Campan M., Dahl C., Dahmcke C. M., Diep D., Fernández A. F., Gerhauser C., Haake A., Heilmann K., Holcomb T., Hussmann D., Ito M., Kläver R., Kreutz M., Kulis M., Lopez V., Nair S. S., Paul D. S., Plongthongkum N., Qu W., Queirós A. C., Reinicke F., Sauter G., Schlomm T., Statham A., Stirzaker C., Strogantsev R., Urdinguio R. G., Walter K., Weichenhan D., Weisenberger D. J., Beck S., Clark S. J., Esteller M., Ferguson-Smith A. C., Fraga M. F., Guldberg P., Hansen L. L., Laird P. W., Martín-Subero J. I., Nygren A. O. H., Peist R., Plass C., Shames D. S., Siebert R., Sun X., Tost J., Walter J., and Zhang K. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nature Biotechnology*, 34(7):726–737, 2016. ISSN 1087-0156. doi: 10.1038/nbt.3605.

Boichard A., Tsigelny I. F., and Kurzrock R. High expression of PD-1 ligands is associated with kataegis mutational signature and APOBEC3 alterations. *OncoImmunology*, 6 (3):e1284719, 2017. ISSN 2162-402X. doi: 10.1080/2162402X.2017.1284719.

Bonde P., Gao D., Chen L., Miyashita T., Montgomery E., Harmon J. W., and Wei C. Duodenal Reflux Leads to Down Regulation of DNA Mismatch Repair Pathway in an Animal Model of Esophageal Cancer. *Annals of Thoracic Surgery*, 83(2):433–440, 2007. ISSN 1552-6259. doi: 10.1016/j.athoracsur.2006.06.090.

Booth M. J., Branco M. R., Ficz G., Oxley D., Krueger F., Reik W., and Balasubramanian S. Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science (New York, N.Y.)*, 336(6083):934–937, 2012. ISSN 0036-8075. doi: 10.1126/science.1220671.

Booth M. J., Marsico G., Bachman M., Beraldi D., and Balasubramanian S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nature chemistry*, 6 (5):435–40, 2014. ISSN 1755-4349. doi: 10.1038/nchem.1893.

Borrego S., Vazquez A., Dasí F., Cerdá C., Iradi A., Tormos C., Sánchez J. M., Bagán L., Boix J., Zaragoza C., Camps J., and Sáez G. Oxidative stress and DNA damage in human gastric carcinoma: 8-Oxo-7'8-dihydro-2'-deoxyguanosine (8-oxo-dG) as a

possible tumor marker. *International Journal of Molecular Sciences*, 14(2):3467–3486, 2013. ISSN 16616596. doi: 10.3390/ijms14023467.

Boveri T. Zur Frage der EntstehungMaligner Tumoren. *Science*, 1041:857–859, 1914.

Branzei D. and Szakal B. Priming for tolerance and cohesion at replication forks. *Nucleus*, 7(1):8–12, 2016a. ISSN 1949-1034. doi: 10.1080/19491034.2016.1149663.

Branzei D. and Szakal B. DNA damage tolerance by recombination: Molecular pathways and DNA structures. *DNA Repair*, 44:68–75, 2016b. ISSN 15687856. doi: 10.1016/j.dnarep.2016.05.008.

Brash D. E. UV signature mutations. *Photochemistry and Photobiology*, 91(1):15–26, 2015. ISSN 17511097. doi: 10.1111/php.12377.

Brazauskas P. and Kriaucionis S. DNA modifications: Another stable base in DNA. *Nature Chemistry*, 6(12):1031–1033, 2014. ISSN 1755-4330. doi: 10.1038/nchem.2115.

Breiling A. and Lyko F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics {&} Chromatin*, 8(1):24, 2015. ISSN 1756-8935. doi: 10.1186/s13072-015-0016-6.

Breitling L. P., Yang R., Korn B., Burwinkel B., and Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *American Journal of Human Genetics*, 88(4):450–457, 2011. ISSN 00029297. doi: 10.1016/j.ajhg.2011.03.003.

Bryan D. S., Ransom M., Adane B., York K., and Hesselberth J. R. High resolution mapping of modified DNA nucleobases using excision repair enzymes. *Genome Research*, 24(9):1534–1542, 2014. ISSN 15495469. doi: 10.1101/gr.174052.114.

Buisson R., Niraj J., Pauty J., Maity R., Zhao W., Coulombe Y., Sung P., and Masson J. Y. Breast cancer proteins PALB2 and BRCA2 stimulate polymerase η in recombination-associated DNA Synthesis At Blocked Replication Forks. *Cell Reports*, 6(3):553–564, 2014. ISSN 22111247. doi: 10.1016/j.celrep.2014.01.009.

Burgers P. M. and Kunkel T. A. Eukaryotic DNA Replication Fork. *Annu Rev Biochem*, 2017. doi: 10.1146/annurev-biochem-061516-044709.

Burgers P. M., Gordenin D., and Kunkel T. A. Who Is Leading the Replication Fork, Pol epsilon or Pol delta? *Molecular Cell*, 61(4):492–493, 2016. ISSN 10972765. doi: 10.1016/j.molcel.2016.01.017.

Burns K. H. Transposable elements in cancer. *Nature Reviews Cancer*, 17(7):415–424, 2017. ISSN 1474-175X. doi: 10.1038/nrc.2017.35.

Burns M. B., Lackey L., Carpenter M. A., Rathore A., Land A. M., Leonard B., Refsland E. W., Kotandeniya D., Tretyakova N., Nikas J. B., Yee D., Temiz N. A., Donohue D. E., McDougle R. M., Brown W. L., Law E. K., and Harris R. S. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*, 494(7437):366–370, 2013a. ISSN 0028-0836. doi: 10.1038/nature11881.

Burns M. B., Temiz N. A., and Harris R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nature Genetics*, 45(9):977–983, 2013b. ISSN 1061-4036. doi: 10.1038/ng.2701.

Campbell T. C. Nutrition and Cancer: An Historical Perspective—The Past, Present, and Future of Nutrition and Cancer. Part 2. Misunderstanding and Ignoring Nutrition. *Nutrition and Cancer*, 0(0):1–7, 2017. ISSN 0163-5581. doi: 10.1080/01635581.2017.1339094.

Canela A., Maman Y., Jung S., Wong N., Callen E., Day A., Kieffer-Kwon K. R., Pekowska A., Zhang H., Rao S. S., Huang S. C., Mckinnon P. J., Aplan P. D., Pommier Y., Aiden E. L., Casellas R., and Nussenzweig A. Genome Organization Drives Chromosome Fragility. *Cell*, 170(3):507–521.e18, 2017. ISSN 10974172. doi: 10.1016/j.cell.2017.06.034.

Cannistraro V. J. and Taylor J. S. Acceleration of 5-Methylcytosine Deamination in Cyclobutane Dimers by G and Its Implications for UV-Induced C-to-T Mutation

Hotspots. *Journal of Molecular Biology*, 392(5):1145–1157, 2009. ISSN 00222836. doi: 10.1016/j.jmb.2009.07.048.

Cannistraro V. J., Pondugula S., Song Q., and Taylor J. S. Rapid deamination of cyclobutane pyrimidine dimer photoproducts at TCG sites in a translationally and rotationally positioned nucleosome in Vivo. *Journal of Biological Chemistry*, 290(44):26597–26609, 2015. ISSN 1083351X. doi: 10.1074/jbc.M115.673301.

Cao S., Zhang C., and Xu Y. Somatic mutations may not be the primary drivers of cancer formation. *International Journal of Cancer*, pages n/a–n/a, 2015. ISSN 00207136. doi: 10.1002/ijc.29639.

Capuano F., Mülleder M., Kok R., Blom H. J., and Ralser M. Cytosine DNA methylation is found in drosophila melanogaster but absent in saccharomyces cerevisiae, schizosaccharomyces pombe, and other yeast species. *Analytical Chemistry*, 86(8): 3697–3702, 2014. ISSN 15206882. doi: 10.1021/ac500447w.

Carpenter M. A., Li M., Rathore A., Lackey L., Law E. K., Land A. M., Leonard B., Shandilya S. M. D., Bohn M. F., Schiffer C. A., Brown W. L., and Harris R. S. Methylcytosine and normal cytosine deamination by the foreign DNA restriction enzyme APOBEC3A. *Journal of Biological Chemistry*, 287(41):34801–34808, 2012. ISSN 00219258. doi: 10.1074/jbc.M112.385161.

Cayrou C., Coulombe P., Vigneron A., Stanojcic S., Ganier O., Peiffer I., Rivals E., Puy A., Laurent-Chabalier S., Desprat R., and Méchali M. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Research*, 21(9):1438–1449, 2011. ISSN 10889051. doi: 10.1101/gr.121830.111.

Chan K., Roberts S. A., Klimczak L. J., Sterling J. F., Saini N., Malc E. P., Kim J., Kwiatkowski D. J., Fargo D. C., Mieczkowski P. A., Getz G., and Gordenin D. A. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet*, 47(9):1067–1072, 2015. ISSN 1546-1718. doi: 10.1038/ng.3378.

Chan K., Resnick M. A., and Gordenin D. A. The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. *DNA repair*, 12(11):878–89, nov 2013. ISSN 1568-7856. doi: 10.1016/j.dnarep.2013.07.008.

Chang H. H. Y., Pannunzio N. R., Adachi N., and Lieber M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature Reviews Molecular Cell Biology*, 18(8):495–506, 2017. ISSN 1471-0072. doi: 10.1038/nrm.2017.48.

Chen C.-l., Rappailles A., Duquenne L., Huvet M., Guilbaud G., Farinelli L., Audit B., D'Aubenton-Carafa Y., Arneodo A., Hyrien O., Thermes C., Chen C.-l., Guilbaud G., Farinelli L., Audit B., Aubenton-carafa Y., Arneodo A., Hyrien O., and Thermes C. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Research*, 20(4):447–457, apr 2010. ISSN 1549-5469. doi: 10.1101/gr.098947.109.1.

Chen H., Lilley C. E., Yu Q., Lee D. V., Chou J., Narvaiza I., Landau N. R., and Weitzman M. D. APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. *Current Biology*, 16(5):480–485, 2006. ISSN 09609822. doi: 10.1016/j.cub.2006.01.031.

Chen K., Zhang J., Guo Z., Ma Q., Xu Z., Zhou Y., Xu Z., Li Z., Liu Y., Ye X., Li X., Yuan B., Ke Y., He C., Zhou L., Liu J., and Ci W. Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer. *Cell Research*, pages 103–118, 2015. ISSN 1001-0602. doi: 10.1038/cr.2015.150.

Chen Q., Chen Y., Bian C., Fujiki R., and Yu X. TET2 promotes histone O-GlcNAcylation during gene transcription. *Nature*, 493(7433):561–564, 2012. ISSN 0028-0836. doi: 10.1038/nature11742.

Cheng X. Structural and functional coordination of dna and histone methylation. *Cold Spring Harbor Perspectives in Biology*, 6(8):1–24, 2014. ISSN 19430264. doi: 10.1101/cshperspect.a018747.

Chiu Y.-L. and Greene W. C. The APOBEC3 Cytidine Deaminases: An Innate Defensive Network Opposing Exogenous Retroviruses and Endogenous Retroelements. *Annual Review of Immunology*, 26(1):317–353, 2008. ISSN 0732-0582. doi: 10.1146/annurev.immunol.26.021607.090350.

Cho M., Grabmaier K., Kitahori Y., Hiasa Y., Nakagawa Y., Uemura H., Hirao Y., Ohnishi T., Yoshikawa K., and Ooesterwijk E. Activation of the MN/CA9 gene is associated with hypomethylation in human renal cell carcinoma cell lines. *Molecular carcinogenesis*, 27(3):184–9, mar 2000. ISSN 0899-1987.

Choi J. K. Contrasting chromatin organization of CpG islands and exons in the human genome. *Genome biology*, 11(7):R70, jan 2010. ISSN 1465-6914. doi: 10.1186/gb-2010-11-7-r70.

Church D. N., Briggs S. E. W., Palles C., Domingo E., Kearsey S. J., Grimes J. M., Gorman M., Martin L., Howarth K. M., Hodgson S. V., Kaur K., Taylor J., and Tomlinson I. P. M. DNA polymerase ϵ and δ exonuclease domain mutations in endometrial cancer. *Human Molecular Genetics*, 22(14):2820–2828, jun 2013. ISSN 1460-2083. doi: 10.1093/hmg/ddt131.

Cimmino L. and Aifantis I. Alternative roles for oxidized mCs and TETs. *Current opinion in genetics & development*, 42:1–7, 2016. ISSN 1879-0380. doi: 10.1016/j.gde.2016.11.003.

Cohen I. S., Bar C., Paz-Elizur T., Ainbinder E., Leopold K., de Wind N., Geacintov N., and Livneh Z. DNA lesion identity drives choice of damage tolerance pathway in murine cell chromosomes. *Nucleic acids research*, 43(3):1–9, 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1398.

Colussi C., Parlanti E., Degan P., Aquilina G., Barnes D., Macpherson P., Karran P., Crescenzi M., Dogliotti E., and Bignami M. The Mammalian Mismatch Repair pathway removes DNA 8-oxodGMP incorporated from the oxidized dNTP pool. *Current Biology*, 12(11):912–918, 2002. ISSN 09609822. doi: 10.1016/S0960-9822(02)00863-1.

Cooper D. N. and Youssoufian H. The CpG dinucleotide and human genetic disease. *Human Genetics*, 78(2):151–155, 1988. ISSN 0340-6717. doi: 10.1007/BF00278187.

Cooper D. N., Mort M., Stenson P. D., Ball E. V., and Chuzhanova N. A. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Human Genomics*, 4(6):406–410, 2010. ISSN 1479-7364. doi: 10.1186/1479-7364-4-6-406.

Cordeiro-Stone M. and Nikolaishvili-Feinberg N. Asymmetry of DNA replication and translesion synthesis of UV-induced thymine dimers. *Mutation research*, 510(1-2): 91–106, 2002. ISSN 00275107. doi: 10.1016/S0027-5107(02)00255-5.

Cortellino S., Xu J., Sannai M., Moore R., Caretti E., Cigliano A., Le Coz M., Devarajan K., Wessels A., Soprano D., Abramowitz L. K., Bartolomei M. S., Rambow F., Bassi M. R., Bruno T., Fanciulli M., Renner C., Klein-Szanto A. J., Matsumoto Y., Kobi D., Davidson I., Alberti C., Larue L., and Bellacosa A. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell*, 146(1): 67–79, 2011. ISSN 00928674. doi: 10.1016/j.cell.2011.06.020.

Cortes-Ciriano I., Lee S., Park W.-Y., Kim T.-M., and Park P. J. A molecular portrait of microsatellite instability across multiple cancers. *Nature Communications*, 8:15180, 2017. ISSN 2041-1723. doi: 10.1038/ncomms15180.

Cortez D. Preventing replication fork collapse to maintain genome integrity. *DNA Repair*, 32:149–157, 2015. ISSN 15687856. doi: 10.1016/j.dnarep.2015.04.026.

Crossan G. P., Garaycoechea J. I., and Patel K. J. Do mutational dynamics in stem cells explain the origin of common cancers? *Cell Stem Cell*, 16(2):111–112, 2015. ISSN 18759777. doi: 10.1016/j.stem.2015.01.009.

Crouse G. F. Non-canonical actions of mismatch repair. *DNA Repair*, 38:102–109, 2016. ISSN 15687856. doi: 10.1016/j.dnarep.2015.11.020.

Croy R. G., Essigmann J. M., Reinhold V. N., and Wogan G. N. Identification of the principal aflatoxin B1-DNA adduct formed in vivo in rat liver. *Proceedings of the National Academy of Sciences of the United States of America*, 75(4):1745–9, apr 1978. ISSN 0027-8424.

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

Daley T. and Smith A. D. Predicting the molecular complexity of sequencing libraries. *Nature methods*, 10(4):325–7, 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2375.

David S. S., O'Shea V. L., and Kundu S. Base-excision repair of oxidative DNA damage. *Nature*, 447(7147):941–950, 2007. ISSN 0028-0836. doi: 10.1038/nature05978.

De Bont R. and van Larebeke N. Endogenous DNA damage in humans: A review of quantitative data. *Mutagenesis*, 19(3):169–185, 2004. ISSN 02678357. doi: 10.1093/mutage/geh025.

De Luca G., Russo M. T., Degan P., Tiveron C., Zijno A., Meccia E., Ventura I., Mattei E., Nakabeppu Y., Crescenzi M., Pepponi R., P??zzola A., Popoli P., and Bignami M. A role for oxidized DNA precursors in Huntington's disease-like striatal neurodegeneration. *PLoS Genetics*, 4(11):e1000266, nov 2008. ISSN 15537390. doi: 10.1371/journal.pgen.1000266.

Delhommeau F., Dupont S., Valle V. D., James C., Trannoy S., Massé A., Kosmider O., Couedic J.-p. L., Robert F., Alberdi A., Lecluse Y., Plo I., Dreyfus F. J., Marzac C., Casadevall N., Lacombe C., Romana S. P., Dessen P., Soulier J., Viguie F., Fontenay M., Vainchenker W., and Bernard O. A. Mutation in TET2 in Myeloid Cancers François. *The New England journal of medicine*, 360:2289–2301, 2009.

Dellino G. I., Cittaro D., Piccioni R., Luzi L., Banfi S., Segalla S., Cesaroni M., Mendoza-Maldonado R., Giacca M., and Pelicci P. G. Genome-wide mapping of human DNA-replication origins: Levels of transcription at ORC1 sites regulate origin selection

and replication timing. *Genome Research*, 23(1):1–11, 2013. ISSN 10889051. doi: 10.1101/gr.142331.112.

Denissenko M. F., Chen J. X., Tang M. S., and Pfeifer G. P. Cytosine methylation determines hot spots of DNA damage in the human P53 gene. *Proceedings of the National Academy of Sciences of the United States of America*, 94(8):3893–8, 1997. ISSN 0027-8424. doi: 10.1073/pnas.94.8.3893.

Denissenko M. F. and Pao A. Preferential Formation of Benzo[a]pyrene Adducts at Lung Cancer Mutational Hotspots in P53. *Science*, 274(5286):430–432, 1996. ISSN 0036-8075. doi: 10.1126/science.274.5286.430.

Deplus R., Delatte B., Schwinn M. K., Defrance M., Méndez J., Murphy N., Dawson M. A., Volkmar M., Putmans P., Calonne E., Shih A. H., Levine R. L., Bernard O., Mercher T., Solary E., Urh M., Daniels D. L., and Fuks F. TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. *The EMBO Journal*, 32(5): 645–655, 2013. ISSN 0261-4189. doi: 10.1038/emboj.2012.357.

Desprat R., Thierry-Mieg D., Lailler N., Lajugie J., Schildkraut C., Thierry-Mieg J., and Bouhassira E. E. Predictable dynamic program of timing of DNA replication in human cells. *Genome Research*, 19(12):2288–2299, dec 2009.

Di Noia J. M. and Neuberger M. S. Molecular mechanisms of antibody somatic hypermutation. *Annual review of biochemistry*, 76:1–22, jan 2007. ISSN 0066-4154. doi: 10.1146/annurev.biochem.76.061705.090740.

Diamant N., Hendel A., Vered I., Carell T., Reißner T., De Wind N., Geacinov N., and Livneh Z. DNA damage bypass operates in the S and G2 phases of the cell cycle and exhibits differential mutagenicity. *Nucleic Acids Research*, 40(1):170–180, 2012. ISSN 03051048. doi: 10.1093/nar/gkr596.

Ding L., Ley T. J., Larson D. E., Miller C. A., Koboldt D. C., Welch J. S., Ritchey J. K., Young M. A., Lamprecht T., McLellan M. D., McMichael J. F., Wallis J. W., Lu C., Shen D., Harris C. C., Dooling D. J., Fulton R. S., Fulton L. L., Chen K., Schmidt H.,

Kalicki-Veizer J., Magrini V. J., Cook L., McGrath S. D., Vickery T. L., Wendl M. C., Heath S., Watson M. A., Link D. C., Tomasson M. H., Shannon W. D., Payton J. E., Kulkarni S., Westervelt P., Walter M. J., Graubert T. A., Mardis E. R., Wilson R. K., and DiPersio J. F. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, 2012. ISSN 0028-0836. doi: 10.1038/nature10738.

Doi A., Park I. H., Wen B., Murakami P., Aryee M. J., Irizarry R., Herb B., Ladd-Acosta C., Rho J., Loewer S., Miller J., Schlaeger T., Daley G. Q., and Feinberg A. P. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nature genetics*, 41(12):1350–1353, 2009. ISSN 1546-1718. doi: 10.1038/ng.471.

Draizen E. J., Shaytan A. K., Mariño-Ramírez L., Talbert P. B., Landsman D., and Panchenko A. R. HistoneDB 2.0: A histone database with variants - An integrated resource to explore histones and their variants. *Database*, 2016(September 2017):1–10, 2016. ISSN 17580463. doi: 10.1093/database/baw014.

D'Souza S., Yamanaka K., and Walker G. C. Non mutagenic and mutagenic DNA damage tolerance. *Cell Cycle*, 15(3):314–315, 2016. ISSN 15514005. doi: 10.1080/15384101.2015.1132909.

Du J., Johnson L. M., Jacobsen S. E., and Patel D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nature Reviews Molecular Cell Biology*, 16 (9):519–532, 2015. ISSN 1471-0072. doi: 10.1038/nrm4043.

Dulak A. M., Stojanov P., Peng S., Lawrence M. S., Fox C., Stewart C., Bandla S., Imamura Y., Schumacher S. E., Shefler E., McKenna A., Carter S. L., Cibulskis K., Sivachenko A., Saksena G., Voet D., Ramos A. H., Auclair D., Thompson K., Sougnez C., Onofrio R. C., Guiducci C., Beroukhim R., Zhou Z., Lin L., Lin J., Reddy R., Chang A., Landrenau R., Pennathur A., Ogino S., Luketich J. D., Golub T. R., Gabriel S. B., Lander E. S., Beer D. G., Godfrey T. E., Getz G., and Bass A. J. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and

mutational complexity. *Nature genetics*, 45(5):478–86, 2013. ISSN 1546-1718. doi: 10.1038/ng.2591.

Duncan T., Trewick S. C., Koivisto P., Bates P. A., Lindahl T., and Sedgwick B. Reversal of DNA alkylation damage by two human dioxygenases. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26):16660–16665, 2002. ISSN 00278424. doi: 10.1073/pnas.262589799.

Duns G., van den Berg E., van Duivenbode I., Osinga J., Hollema H., Hofstra R. M. W., and Kok K. Histone Methyltransferase Gene SETD2 Is a Novel Tumor Suppressor Gene in Clear Cell Renal Cell Carcinoma. *Cancer Research*, 70(11):4287–4291, 2010. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-10-0120.

Dvorak K., Payne C. M., Chavarria M., Ramsey L., Dvorakova B., Bernstein H., Holubec H., Sampliner R. E., Guy N., Condon A., Bernstein C., Green S. B., Prasad A., and Dvorak K. Bile acids in combination with low pH induce oxidative stress and oxidative DNA damage: relevance to the pathogenesis of Barrett's oesophagus. *Gut*, 56:763–771, 2007. ISSN 0017-5749. doi: 10.1136/gut.2006.103697.

Dvorak K., Watts G. S., Ramsey L., Holubec H., Payne C. M., Bernstein C., Jenkins G. J., Sampliner R. E., Prasad A., Garewal H. S., and Bernstein H. Expression of bile acid transporting proteins in Barrett's esophagus and esophageal adenocarcinoma. *The American journal of gastroenterology*, 104(December 2007):302–309, 2009. ISSN 0002-9270. doi: 10.1038/ajg.2008.85.

Eden A., Waghmare A., and Jaenisch R. Chromosomal Instability and Tumors Promoted by DNA. *Science*, 300, 2003.

Edgar R., Tan P. P. C., Portales-Casamar E., and Pavlidis P. Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. *Epigenetics & chromatin*, 7(1):28, 2014. ISSN 1756-8935. doi: 10.1186/1756-8935-7-28.

Ehrlich M., Norris K. F., Wang R. Y., Kuo K. C., and Gehrke C. W. DNA cytosine methylation and heat-induced deamination. *Bioscience reports*, 6(4):387–93, apr 1986. ISSN 0144-8463.

Elinav E., Nowarski R., Thaiss C. A., Hu B., Jin C., and Flavell R. A. Inflammation-induced cancer: Crosstalk between tumours, immune cells and microorganisms. *Nature Reviews Cancer*, 13(11):759–771, 2013. ISSN 1474175X. doi: 10.1038/nrc3611.

Ellermann M., Eheim A., Rahm F., Viklund J., Guenther J., Andersson M., Ericsson U., Forsblom R., Ginman T., Lindström J., Silvander C., Trésaugues L., Giese A., Bunse S., Neuhaus R., Weiske J., Quanz M., Glasauer A., Nowak-Reppel K., Bader B., Irlbacher H., Meyer H., Queisser N., Bauser M., Haegebarth A., and Gorjánácz M. Novel Class of Potent and Cellularly Active Inhibitors Devalidates MTH1 as Broad-Spectrum Cancer Target. *ACS Chemical Biology*, page acschembio.7b00370, 2017. ISSN 1554-8929. doi: 10.1021/acschembio.7b00370.

Erichsen R., Robertson D., Farkas D. K., Pedersen L., Pohl H., Baron J. A., and Sørensen H. T. Erosive Reflux Disease Increases Risk for Esophageal Adenocarcinoma, Compared With Nonerosive Reflux. *Clinical Gastroenterology and Hepatology*, 10(5): 475–480.e1, 2012. ISSN 15423565. doi: 10.1016/j.cgh.2011.12.038.

Esteller M. and Herman J. G. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *The Journal of pathology*, 196(1):1–7, jan 2002. ISSN 0022-3417. doi: 10.1002/path.1024.

Ewels P., Magnusson M., Lundin S., and K??ller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btw354.

Farthing C. R., Ficz G., Ng R. K., Chan C. F., Andrews S., Dean W., Hemberger M., and Reik W. Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. *PLoS Genetics*, 4(6), 2008. ISSN 15537390. doi: 10.1371/journal.pgen.1000116.

Fein M., Maroske J., and Fuchs K. H. Importance of duodenogastric reflux in gastro-oesophageal reflux disease. *British Journal of Surgery*, 93(12):1475–1482, 2006. ISSN 00071323. doi: 10.1002/bjs.5486.

Felsenfeld G. A brief history of epigenetics. *Cold Spring Harbor perspectives in biology*, 6(1):15–22, jan 2014. ISSN 1943-0264. doi: 10.1101/cshperspect.a018200.

Feng Z., Hu W., Hu Y., and Tang M.-s. Acrolein is a major cigarette-related lung cancer agent: Preferential binding at p53 mutational hotspots and inhibition of DNA repair. *Proceedings of the National Academy of Sciences*, 103(42):15404–15409, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0607031103.

Ficz G. and Gribben J. G. Loss of 5-hydroxymethylcytosine in cancer: Cause or consequence? *Genomics*, 104(5):352–357, 2014. ISSN 10898646. doi: 10.1016/j.ygeno.2014.08.017.

Figueroa M. E., Abdel-Wahab O., Lu C., Ward P. S., Patel J., Shih A., Li Y., Bhagwat N., Vasanthakumar A., Fernandez H. F., Tallman M. S., Sun Z., Wolniak K., Peeters J. K., Liu W., Choe S. E., Fantin V. R., Paietta E., Löwenberg B., Licht J. D., Godley L. A., Delwel R., Valk P. J., Thompson C. B., Levine R. L., and Melnick A. Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell*, 18(6):553–567, 2010. ISSN 15356108. doi: 10.1016/j.ccr.2010.11.015.

Flood C. L., Rodriguez G. P., Bao G., Shockley A. H., Kow Y. W., and Crouse G. F. Replicative DNA Polymerase δ but Not ϵ Proofreads Errors in Cis and in Trans. *PLoS Genetics*, 11(3):1–32, 2015. ISSN 15537404. doi: 10.1371/journal.pgen.1005049.

Fontebasso A. M., Schwartzentruber J., Khuong-Quang D.-A., Liu X.-Y., Sturm D., Korshunov A., Jones D. T. W., Witt H., Kool M., Albrecht S., Fleming A., Hadjadj D., Busche S., Lepage P., Montpetit A., Staffa A., Gerges N., Zakrzewska M., Zakrzewski K., Liberski P. P., Hauser P., Garami M., Klekner A., Bognar L., Zadeh G., Faury D., Pfister S. M., Jabado N., and Majewski J. Mutations in SETD2 and genes affecting histone

H3K36 methylation target hemispheric high-grade gliomas. *Acta neuropathologica*, 125(5):659–69, may 2013. ISSN 1432-0533. doi: 10.1007/s00401-013-1095-8.

Foulk M. S., Urban J. M., Casella C., and Gerbi S. A. Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. *Genome Research*, 25:725–735, 2015. doi: 10.1101/gr.183848.114.

Fragkos M., Ganier O., Coulombe P., and Méchali M. DNA replication origin activation in space and time. *Nature reviews. Molecular cell biology*, 16(6):360–74, 2015. ISSN 1471-0080. doi: 10.1038/nrm4002.

Franklin R. E. and Gosling R. G. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–1, apr 1953. ISSN 0028-0836. doi: 10.1038/171740a0.

Fu Y., Ito F., Zhang G., Fernandez B., Yang H., and Chen X. S. DNA cytosine and methylcytosine deamination by APOBEC3B: enhancing methylcytosine deamination by engineering APOBEC3B. *The Biochemical Journal*, 471(1):25–35, 2015. ISSN 1470-8728. doi: 10.1042/BJ20150382.

Fujishita T., Okamoto T., Akamine T., Takamori S., Takada K., Katsura M., Toyokawa G., Shoji F., Shimokawa M., Oda Y., Nakabeppu Y., and Maehara Y. Association of MTH1 expression with the tumor malignant potential and poor prognosis in patients with resected lung cancer. *Lung Cancer*, 109(March):52–57, 2017. ISSN 18728332. doi: 10.1016/j.lungcan.2017.04.012.

Fuller R. S., Funnell B. E., and Kornberg A. The dnaA protein complex with the E. coli chromosomal replication origin (oriC) and other DNA sites. *Cell*, 38(3):889–900, oct 1984. ISSN 00928674. doi: 10.1016/0092-8674(84)90284-8.

Gad H., Koolmeister T., Jemth A.-S., Eshtad S., Jacques S. A., Ström C. E., Svensson L. M., Schultz N., Lundbäck T., Einarsdottir B. O., Saleh A., Göktürk C.,

Baranczewski P., Svensson R., Berntsson R. P.-A., Gustafsson R., Strömberg K., Sanjiv K., Jacques-Cordonnier M.-C., Desroses M., Gustavsson A.-L., Olofsson R., Johansson F., Homan E. J., Loseva O., Bräutigam L., Johansson L., Höglund A., Hagenkort A., Pham T., Altun M., Gaugaz F. Z., Vikingsson S., Evers B., Henriksson M., Vallin K. S. A., Wallner O. A., Hammarström L. G. J., Wiita E., Almlöf I., Kalderén C., Axelsson H., Djureinovic T., Puigvert J. C., Häggblad M., Jeppsson F., Martens U., Lundin C., Lundgren B., Granelli I., Jensen A. J., Artursson P., Nilsson J. A., Stenmark P., Scobie M., Berglund U. W., and Helleday T. MTH1 inhibition eradicates cancer by preventing sanitation of the dNTP pool. *Nature*, 508(7495):215–21, 2014. ISSN 1476-4687. doi: 10.1038/nature13181.

Gal-Yam E. N., Egger G., Iniguez L., Holster H., Einarsson S., Zhang X., Lin J. C., Liang G., Jones P. A., Tanay A., Gal-Yam E. N., Egger G., Iniguez L., Holster H., Lin J. C., Liang G., Jones P. A., Tanay A., Sssi M., Einarsson S., Zhang X., Lin J. C., Liang G., Jones P. A., and Tanay A. Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proceedings of the National Academy of Sciences of the United States of America*, 105(35):12979–12984, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0806437105.

Gale J. M., Nissen K. A., and Smerdon M. J. UV-induced formation of pyrimidine dimers in nucleosome core DNA is strongly modulated with a period of 10.3 bases. *Biochemistry*, 84:6644–6648, 1987. ISSN 0027-8424. doi: 10.1073/pnas.84.19.6644.

Ganai R. A. and Johansson E. DNA Replication-A Matter of Fidelity. *Molecular Cell*, 62 (5):745–755, 2016. ISSN 10974164. doi: 10.1016/j.molcel.2016.05.003.

Gao X., Thomsen H., Zhang Y., Breitling L. P., and Brenner H. The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes. *Clinical Epigenetics*, 9(1):87, 2017. ISSN 1868-7075. doi: 10.1186/s13148-017-0387-6.

Gao Z., Wyman M. J., Sella G., Przeworski M., Campbell P., and Nik-Zainal S. Interpreting the Dependence of Mutation Rates on Age and Time. *PLOS Biology*, 14(1): e1002355, jan 2016. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002355.

Gelfman S., Cohen N., Yearim A., and Ast G. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome research*, 23(5):789–99, may 2013. ISSN 1549-5469. doi: 10.1101/gr.143503.112.

Georgescu R. E., Schauer G. D., Yao N. Y., Langston L. D., Yurieva O., Zhang D., Finkelstein J., and O'Donnell M. E. Reconstitution of a eukaryotic replisome reveals suppression mechanisms that define leading/lagging strand operation. *eLife*, 2015(4): 1–20, 2015. ISSN 2050084X. doi: 10.7554/eLife.04988.

Glaser A. P., Fantini D., Rimar K. J., Meeks J. J., and Meeks J. J. APOBEC-mediated mutagenesis in urothelial carcinoma is associated with improved survival, mutations in DNA damage response genes , and immune response. *bioRxiv*, 2017. doi: 10.1101/123802.

Globisch D., Münzel M., Müller M., Michalakis S., Wagner M., Koch S., Brückl T., Biel M., and Carell T. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS one*, 5(12):e15367, jan 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0015367.

Gonzalez-Huici V., Szakal B., Urulangodi M., Psakhye I., Castellucci F., Menolfi D., Rajakumara E., Fumasoni M., Bermejo R., Jentsch S., and Branzei D. DNA bending facilitates the error-free DNA damage tolerance pathway and upholds genome integrity. *The EMBO Journal*, 33(4):327–340, 2014.

Gonzalez-Zulueta M., Bender C. M., Yang A. S., Nguyen T., Beart R. W., Van Tornout J. M., and Jones P. A. Methylation of the 5' CpG Island of the p16/CDKN2 Tumor Suppressor Gene in Normal and Transformed Human Tissues Correlates with Gene Silencing. *Cancer Res.*, 55(20):4531–4535, oct 1995.

Goodwin S., McPherson J. D., and McCombie W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016. ISSN 1471-0056. doi: 10.1038/nrg.2016.49.

Greaves M. and Maley C. C. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012. ISSN 0028-0836. doi: 10.1038/nature10762.

Green A. M., Landry S., Budagyan K., Avgousti D. C., Shalhout S., Bhagwat A. S., and Weitzman M. D. APOBEC3A damages the cellular genome during DNA replication. *Cell Cycle*, 15(7):998–1008, 2016. ISSN 15514005. doi: 10.1080/15384101.2016.1152426.

Gu J., Chen Q., Xiao X., Ito F., Wolfe A., and Chen X. S. Biochemical Characterization of APOBEC3H Variants: Implications for Their HIV-1 Restriction Activity and mC Modification. *Journal of Molecular Biology*, 428(23):4626–4638, 2016. ISSN 10898638. doi: 10.1016/j.jmb.2016.08.012.

Guilliam T. A. and Doherty A. J. Primpol-prime time to reprime. *Genes*, 8(1), 2017. ISSN 20734425. doi: 10.3390/genes8010020.

Guisan A. and Zimmermann N. E. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135:147–186, 2000. ISSN 03043800. doi: 10.1016/S0304-3800(00)00354-9.

Guo J. U., Su Y., Zhong C., Ming G.-l. L., and Song H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell*, 145(3):423–434, apr 2011. ISSN 00928674. doi: 10.1016/j.cell.2011.03.022.

Guo S., Diep D., Plongthongkum N., Fung H.-L., Zhang K., and Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nature genetics*, 49(4):635–642, apr 2017. ISSN 1546-1718. doi: 10.1038/ng.3805.

Gustafson C. B., Yang C., Dickson K. M., Shao H., Van Booven D., Harbour J. W., Liu Z.-J., and Wang G. Epigenetic reprogramming of melanoma cells by vitamin C treatment. *Clinical Epigenetics*, 7(1):1–11, 2015. ISSN 1868-7075. doi: 10.1186/s13148-015-0087-z.

Guza R., Kotandeniya D., Murphy K., Dissanayake T., Lin C., Giambasu G. M., Lad R. R., Wojciechowski F., Amin S., Sturla S. J., Hudson R. H. E., York D. M., Jankowiak R., Jones R., and Tretyakova N. Y. Influence of C-5 substituted cytosine and related

nucleoside analogs on the formation of benzo[a]pyrene diol epoxide-dG adducts at CG base pairs of DNA. *Nucleic Acids Research*, 39(9):3988–4006, 2011. ISSN 03051048. doi: 10.1093/nar/gkq1341.

Haffner M. C., Chaux A., Meeker A. K., Esopi D. M., Gerber J., Pellakuru L. G., Toubaji A., Argani P., Iacobuzio-Donahue C., Nelson W. G., Netto G. J., De Marzo A. M., and Yegnasubramanian S. Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. *Oncotarget*, 2(8): 627–37, aug 2011. ISSN 1949-2553.

Hammoud S. S., Low D. H. P., Yi C., Carrell D. T., Guccione E., and Cairns B. R. Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell Stem Cell*, 15(2):239–253, 2014. ISSN 18759777. doi: 10.1016/j.stem.2014.04.006.

Han H., Cortez C. C., Yang X., Nichols P. W., Jones P. A., and Liang G. DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Human Molecular Genetics*, 20(22):4299–4310, 2011. ISSN 09646906. doi: 10.1093/hmg/ddr356.

Hang B. Formation and repair of tobacco carcinogen-derived bulky DNA adducts. *Journal of nucleic acids*, 2010:709521, 2010. ISSN 2090-021X. doi: 10.4061/2010/709521.

Hansemann D. Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung. *Archiv für Pathologische Anatomie und Physiologie und für Klinische Medicin*, 119(2):299–326, feb 1890. ISSN 0945-6317. doi: 10.1007/BF01882039.

Hansen R. S., Thomas S., Sandstrom R., Canfield T. K., Thurman R. E., Weaver M., Dorschner M. O., Gartler S. M., and Stamatoyannopoulos J. A. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1):139–144, jan 2010.

Hara R., Mo J., and Sancar A. DNA damage in the nucleosome core is refractory to repair by human excision nuclease. *Molecular and cellular biology*, 20(24):9173–9181, 2000. ISSN 0270-7306. doi: 10.1128/MCB.20.24.9173-9181.2000.Updated.

Haradhvala N. J., Polak P., Stojanov P., Covington K. R., Shinbrot E., Hess J. M., Rheinbay E., Kim J., Maruvka Y. E., Braunstein L. Z., Kamburov A., Hanawalt P. C., Wheeler D. A., Koren A., Lawrence M. S., and Getz G. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell*, 164(3): 538–549, 2016. ISSN 00928674. doi: 10.1016/j.cell.2015.12.050.

Hardeland U., Bentele M., Jiricny J., and Schär P. The versatile thymine DNA-glycosylase: A comparative characterization of the human, Drosophila and fission yeast orthologs. *Nucleic Acids Research*, 31(9):2261–2271, 2003. ISSN 03051048. doi: 10.1093/nar/gkg344.

Hardwick S. A., Deveson I. W., and Mercer T. R. Reference standards for next-generation sequencing. *Nature Reviews Genetics*, 18(8):473–484, 2017. ISSN 1471-0056. doi: 10.1038/nrg.2017.44.

Harrington M. A., Jones P. A., Imagawat M., and Karint M. Cytosine methylation does not affect binding of transcription factor Spl. *Biochemistry*, 85(April):2066–2070, 1988. ISSN 0027-8424. doi: 10.1073/pnas.85.7.2066.

Hashimoto H., Zhang X., and Cheng X. Excision of thymine and 5-hydroxymethyluracil by the MBD4 DNA glycosylase domain: Structural basis and implications for active DNA demethylation. *Nucleic Acids Research*, 40(17):8276–8284, 2012a. ISSN 03051048. doi: 10.1093/nar/gks628.

Hashimoto K., Cho Y., Yang I. Y., Akagi J. I., Ohashi E., Tateishi S., De Wind N., Hanaoka F., Ohmori H., and Moriya M. The vital role of polymerase ζ and REV1 in mutagenic, but not correct, DNA synthesis across Benzo[a]pyrene-dG and recruitment of polymerase ζ by REV1 to replication-stalled site. *Journal of Biological Chemistry*, 287(12):9613–9622, 2012b. ISSN 00219258. doi: 10.1074/jbc.M111.331728.

Hashimoto K., Bonala R., Johnson F., Grollman A. P., and Moriya M. Y-family DNA polymerase-independent gap-filling translesion synthesis across aristolochic acid-derived adenine adducts in mouse cells. *DNA Repair*, 46:55–60, 2016. ISSN 15687856. doi: 10.1016/j.dnarep.2016.07.003.

Hashimshony T., Zhang J., Keshet I., Bustin M., and Cedar H. The role of DNA methylation in setting up chromatin structure during development. *Nature genetics*, 34(2): 187–192, 2003. ISSN 10614036. doi: 10.1038/ng1158.

Hayakawa Y., Sethi N., Sepulveda A. R., Bass A. J., and Wang T. C. Oesophageal adenocarcinoma and gastric cancer: should we mind the gap? *Nature Reviews Cancer*, 16(5):305–318, 2016. ISSN 1474-175X. doi: 10.1038/nrc.2016.24.

He Y. and Ecker J. R. Non-CG Methylation in the Human Genome. *Annual Review of Genomics and Human Genetics*, 16(1):150615185749007, 2015. ISSN 1527-8204. doi: 10.1146/annurev-genom-090413-025437.

Hecht F., Pessoa C. F., Gentile L. B., Rosenthal D., Carvalho D. P., and Fortunato R. S. The role of oxidative stress on breast cancer development and therapy. *Tumor Biology*, 37(4):4281–4291, 2016. ISSN 14230380. doi: 10.1007/s13277-016-4873-9.

Helleday T. The underlying mechanism for the PARP and BRCA synthetic lethality: Clearing up the misunderstandings. *Molecular Oncology*, 5(4):387–393, 2011. ISSN 18780261. doi: 10.1016/j.molonc.2011.07.001.

Helleday T., Eshtad S., and Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nature reviews. Genetics*, 15(9):585–598, 2014. ISSN 1471-0064. doi: 10.1038/nrg3729.

Hellman A. and Chess A. Gene body-specific methylation on the active X chromosome. *Science (New York, N.Y.)*, 315(5815):1141–3, 2007. ISSN 1095-9203. doi: 10.1126/science.1136352.

Hemerly J. P., Bastos A. U., and Cerutti J. M. Identification of several novel non-p.R132 IDH1 variants in thyroid carcinomas. *European journal of endocrinology / European Federation of Endocrine Societies*, 163(5):747–55, nov 2010. ISSN 1479-683X. doi: 10.1530/EJE-10-0473.

Henninger E., LeCompte K., McBride C., Bunnell B., and Pursell Z. *Somatic mutant alleles of POLE found in human cancers suppress proofreading and replication fidelity in vitro (LB121)*, volume 28. Federation of American Societies for Experimental Biology, apr 2015.

Henninger E. E. *Understanding human DNA polymerase epsilon funtions: Cancer-assocaited mutator variants, proofreading defects, and post-translational modifications*. PhD thesis, Tulane University, 2015.

Herman J. G., Latif F., Weng Y., Lerman M. I., Zbar B., Liu S., Samid D., Duan D. S., Gnarra J. R., and Linehan W. M. Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proceedings of the National Academy of Sciences*, 91(21):9700–9704, oct 1994. ISSN 0027-8424. doi: 10.1073/pnas.91.21.9700.

Herr A. J., Kennedy S. R., Knowels G. M., Schultz E. M., and Preston B. D. DNA replication error-induced extinction of diploid yeast. *Genetics*, 196(3):677–691, 2014. ISSN 19432631. doi: 10.1534/genetics.113.160960.

Hewish M., Lord C. J., Martin S. A., Cunningham D., and Ashworth A. Mismatch repair deficient colorectal cancer in the era of personalized treatment. *Nature Reviews Clinical Oncology*, 7(4):197–208, 2010. ISSN 1759-4774. doi: 10.1038/nrclinonc.2010.18.

Heyn H. and Esteller M. DNA methylation profiling in the clinic: applications and challenges. *Nature Reviews Genetics*, 13(10):679–692, 2012. ISSN 1471-0056. doi: 10.1038/nrg3270.

Hidaka K., Yamada M., Kamiya H., Masutani C., Harashima H., Hanaoka F., and Nohmi T. Specificity of mutations induced by incorporation of oxidized dNTPs into DNA by

human DNA polymerase η. *DNA Repair*, 7(3):497–506, 2008. ISSN 15687864. doi: 10.1016/j.dnarep.2007.12.005.

Hill P. W. S., Amouroux R., and Hajkova P. DNA demethylation, Tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: An emerging complex story. *Genomics*, 104(5):324–333, 2014. ISSN 10898646. doi: 10.1016/j.ygeno.2014.08.012.

Hiltunen M. O., Alhonen L., Koistinaho J., Myöhänen S., Pääkkönen M., Marin S., Kosma V. M., and Jänne J. Hypermethylation of the APC (adenomatous polyposis coli) gene promoter region in human colorectal carcinoma. *International journal of cancer. Journal international du cancer*, 70(6):644–8, mar 1997. ISSN 0020-7136.

Hinrichs A. S. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34(90001):D590–D598, 2006. ISSN 0305-1048. doi: 10.1093/nar/gkj144.

Hirota K., Tsuda M., Mohiuddin, Tsurimoto T., Cohen I. S., Livneh Z., Kobayashi K., Narita T., Nishihara K., Murai J., Iwai S., Guilbaud G., Sale J. E., and Takeda S. In vivo evidence for translesion synthesis by the replicative DNA polymerase δ. *Nucleic Acids Research*, 44(15):7242–7250, 2016. ISSN 13624962. doi: 10.1093/nar/gkw439.

Hoang M. L., Kinde I., Tomasetti C., Mcmahon K. W., Rosenquist T. A., Grollman A. P., Kinzler K. W., Vogelstein B., and Papadopoulos N. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proceedings of the National Academy of Sciences*, 113(35):9846–9851, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1607794113.

Hodgkinson A., Chen Y., and Eyre-Walker A. The large-scale distribution of somatic mutations in cancer genomes. *Human Mutation*, 33(1):136–143, 2012. ISSN 10597794. doi: 10.1002/humu.21616.

Holliday R. and Pugh J. E. DNA modification mechanisms and gene activity during development. *Science (New York, N.Y.)*, 187(4173):226–32, jan 1975. ISSN 0036-8075.

Hoopes J. I., Cortez L. M., Mertz T. M., Malc E. P., Mieczkowski P. A., and Roberts S. A. APOBEC3A and APOBEC3B Preferentially Deaminate the Lagging Strand Template during DNA Replication. *Cell Reports*, pages 1–10, 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.01.021.

Hori M., Satou K., Harashima H., and Kamiya H. Suppression of mutagenesis by 8-hydroxy-2'-deoxyguanosine 5'-triphosphate (7,8-dihydro-8-oxo-2'-deoxyguanosine 5'-triphosphate) by human MTH1, MTH2, and NUDT5. *Free Radical Biology and Medicine*, 48(9):1197–1201, 2010. ISSN 08915849. doi: 10.1016/j.freeradbiomed.2010.02.002.

Houseman E. A., Johnson K. C., and Christensen B. C. OxyBS: Estimation of 5-methylcytosine and 5-hydroxymethylcytosine from tandem-treated oxidative bisulfite and bisulfite DNA. *Bioinformatics*, 32(16):2505–2507, 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btw158.

Hu J., Adar S., Selby C. P., Lieb J. D., and Sancar A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes and Development*, 29(9):948–960, 2015. ISSN 15495477. doi: 10.1101/gad.261271.115.4.

Hu J., Lieb J. D., Sancar A., and Adar S. Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, 113(41):11507–11512, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1614430113.

Hu S., Wan J., Su Y., Song Q., Zeng Y., Nguyen H. N., Shin J., Cox E., Rho H. S., Woodard C., Xia S., Liu S., Lyu H., Ming G. L., Wade H., Song H., Qian J., and Zhu H. DNA methylation presents distinct binding sites for human transcription factors. *eLife*, 2013(2):1–16, 2013. ISSN 2050084X. doi: 10.7554/eLife.00726.

Huber K. V. M., Salah E., Radic B., Gridling M., Elkins J. M., Stukalov A., Jemth A.-S., Göktürk C., Sanjiv K., Strömberg K., Pham T., Berglund U. W., Colinge J., Bennett K. L., Loizou J. I., Helleday T., Knapp S., and Superti-Furga G. Stereospecific targeting of MTH1 by (S)-crizotinib as an anticancer strategy. *Nature*, 508(7495):222–7, 2014. ISSN 1476-4687. doi: 10.1038/nature13194.

Huberman J. A. and Riggs A. D. On the mechanism of DNA replication in mammalian chromosomes. *Journal of molecular biology*, 32(2):327–41, mar 1968. ISSN 0022-2836. doi: 10.1101/cshperspect.a010116.

Hughes L. A. E., Melotte V., de Schrijver J., de Maat M., Smit V. T. H. B. M., Bovée J. V. M. G., French P. J., van den Brandt P. A., Schouten L. J., de Meyer T., van Criekinge W., Ahuja N., Herman J. G., Weijenberg M. P., and van Engeland M. The CpG island methylator phenotype: what's in a name? *Cancer research*, 73(19):5858–68, oct 2013. ISSN 1538-7445. doi: 10.1158/0008-5472.CAN-12-4306.

Hvid-Jensen F., Pedersen L., Funch-Jensen P., and Drewes A. M. Proton pump inhibitor use may not prevent high-grade dysplasia and oesophageal adenocarcinoma in Barrett's oesophagus: A nationwide study of 9883 patients. *Alimentary Pharmacology and Therapeutics*, 39(9):984–991, 2014. ISSN 13652036. doi: 10.1111/apt.12693.

Ikehata H. and Ono T. The Mechanisms of UV Mutagenesis. *Journal of Radiation Research*, 52(2):115–125, 2011. ISSN 0449-3060. doi: 10.1269/jrr.10175.

Ikehata H., Chang Y., Yokoi M., Yamamoto M., and Hanaoka F. Remarkable induction of UV-signature mutations at the 3'-cytosine of dipyrimidine sites except at 5'-TCG-3' in the UVB-exposed skin epidermis of xeroderma pigmentosum variant model mice. *DNA Repair*, 22:112–122, 2014. ISSN 15687856. doi: 10.1016/j.dnarep.2014.07.012.

Ikehata H., Mori T., and Yamamoto M. In Vivo Spectrum of UVC-induced Mutation in Mouse Skin Epidermis May Reflect the Cytosine Deamination Propensity of Cyclobutane Pyrimidine Dimers. *Photochemistry and Photobiology*, 91(6):1488–1496, 2015. ISSN 17511097. doi: 10.1111/php.12525.

Imielinski M., Berger A. H., Hammerman P. S., Hernandez B., Pugh T. J., Hodis E., Cho J., Suh J., Capelletti M., Sivachenko A., Sougnez C., Auclair D., Lawrence M. S., Stojanov P., Cibulskis K., Choi K., De Waal L., Sharifnia T., Brooks A., Greulich H., Banerji S., Zander T., Seidel D., Leenders F., Ansén S., Ludwig C., Engel-Riedel W., Stoelben E., Wolf J., Goparju C., Thompson K., Winckler W., Kwiatkowski D., Johnson B. E.,

Jänne P. A., Miller V. A., Pao W., Travis W. D., Pass H. I., Gabriel S. B., Lander E. S., Thomas R. K., Garraway L. A., Getz G., and Meyerson M. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6):1107–1120, 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.08.029.

Inoue M., Kamiya H., Fujikawa K., Ootsuyama Y., Murata-Kamiya N., Osaki T., Yasumoto K., and Kasai H. Induction of chromosomal gene mutations in Escherichia coli by direct incorporation of oxidatively damaged nucleotides: New evaluation method for mutagenesis by damaged dna precursors in vivo. *Journal of Biological Chemistry*, 273(18):11069–11074, 1998. ISSN 00219258. doi: 10.1074/jbc.273.18.11069.

Irizarry R. A., Ladd-Acosta C., Wen B., Wu Z., Montano C., Onyango P., Cui H., Gabo K., Rongione M., Webster M., Ji H., Potash J. B., Sabunciyan S., and Feinberg A. P. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet*, 41(2):178–186, 2009. ISSN 1546-1718. doi: 10.1038/ng.298.

Ito S., Shen L., Dai Q., Wu S. C., Collins L. B., Swenberg J. A., He C., and Zhang Y. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–3, sep 2011. ISSN 1095-9203. doi: 10.1126/science.1210597.

Iurlaro M., McInroy G. R., Burgess H. E., Dean W., Raiber E.-A., Bachman M., Beraldi D., Balasubramanian S., and Reik W. In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome biology*, 17(1):141, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1001-5.

Iyama T. and Wilson D. M. DNA repair mechanisms in dividing and non-dividing cells. *DNA Repair*, 12(8):620–636, 2013. ISSN 15687864. doi: 10.1016/j.dnarep.2013.04.015.

Jacobs A. L. and Schär P. DNA glycosylases: In DNA repair and beyond. *Chromosoma*, 121(1):1–20, 2012. ISSN 00095915. doi: 10.1007/s00412-011-0347-4.

Jee J., Rasouly A., Shamovsky I., Akivis Y., R. Steinman S., Mishra B., and Nudler E. Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature*, 534(7609):693–696, 2016. ISSN 0028-0836. doi: 10.1038/nature18313.

Jenkins G. J. S., D'Souza F. R., Suzen S. H., Eltahir Z. S., James S. A., Parry J. M., Griffiths P. A., and Baxter J. N. Deoxycholic acid at neutral and acid pH, is genotoxic to oesophageal cells through the induction of ROS: The potential role of anti-oxidants in Barrett's oesophagus. *Carcinogenesis*, 28(1):136–142, 2007. ISSN 01433334. doi: 10.1093/carcin/bgl147.

Jha V., Bian C., Xing G., and Ling H. Structure and mechanism of error-free replication past the major benzo[a]pyrene adduct by human DNA polymerase kappa. *Nucleic Acids Research*, 44(10):4957–4967, 2016. ISSN 13624962. doi: 10.1093/nar/gkw204.

Jimenez P., Piazuelo E., Sanchez M. T., Ortego J., Soteras F., and Lanas A. Free radicals and antioxidant systems in reflux esophagitis and Barrett's esophagus. *World journal of gastroenterology : WJG*, 11(18):2697–2703, 2005.

Jin S.-G., Jiang Y., Qiu R., Rauch T. A., Wang Y., Schackert G., Krex D., Lu Q., and Pfeifer G. P. 5-Hydroxymethylcytosine is strongly depleted in human cancers but its levels do not correlate with IDH1 mutations. *Cancer research*, 71(24):7360–5, dec 2011. ISSN 1538-7445. doi: 10.1158/0008-5472.CAN-11-2023.

Joehanes R., Just A. C., Marioni R. E., Pilling L. C., Reynolds L. M., Mandaviya P. R., Guan W., Xu T., Elks C. E., Aslibekyan S., Moreno-Macias H., Smith J. A., Brody J. A., Dhingra R., Yousefi P., Pankow J. S., Kunze S., Shah S. H., McRae A. F., Lohman K., Sha J., Absher D. M., Ferrucci L., Zhao W., Demerath E. W., Bressler J., Grove M. L., Huan T., Liu C., Mendelson M. M., Yao C., Kiel D. P., Peters A., Wang-Sattler R., Visscher P. M., Wray N. R., Starr J. M., Ding J., Rodriguez C. J., Wareham N. J., Irvin M. R., Zhi D., Barrdahl M., Vineis P., Ambatipudi S., Uitterlinden A. G., Hofman A., Schwartz J., Colicino E., Hou L., Vokonas P. S., Hernandez D. G., Singleton A. B., Bandinelli S., Turner S. T., Ware E. B., Smith A. K., Klengel T., Binder E. B., Psaty B. M., Taylor K. D., Gharib S. A., Swenson B. R., Liang L., Demeo D. L., O'Connor G. T.,

Herceg Z., Ressler K. J., Conneely K. N., Sotoodehnia N., Kardia S. L., Melzer D., Baccarelli A. A., Van Meurs J. B., Romieu I., Arnett D. K., Ong K. K., Liu Y., Waldenberger M., Deary I. J., Fornage M., Levy D., and London S. J. Epigenetic Signatures of Cigarette Smoking. *Circulation: Cardiovascular Genetics*, 9(5):436–447, 2016. ISSN 19423268. doi: 10.1161/CIRCGENETICS.116.001506.

Johnson K. C., Houseman E. A., King J. E., von Herrmann K. M., Fadul C. E., Christensen B. C., Ostrom Q. T., Ostrom Q. T., Verhaak R. G., Eckel-Passow J. E., Christensen B. C., Noushmehr H., Zheng S., Ceccarelli M., Hegi M. E., Kreth S., Thon N., Kreth F. W., Brennan C. W., Tahiliani M., Bachman M., Vasanthakumar A., Godley L. A., Takai H., Booth M. J., Field S. F., Stewart S. K., Houseman E. A., Johnson K. C., Christensen B. C., Uribe-Lewis S., Jones P. A., Williams K., Song C. X., He C., Ivanov M., Stroud H., Feng S., Kinney S. M., Pradhan S., Jacobsen S. E., Lunnon K., Wen L., Tang F., Hnisz D., Neri F., Yu M., Sheffield N. C., Bock C., Gustems M., Schnetz M. P., Ryan M., Orr B. A., Haffner M. C., Nelson W. G., Yegnasubramanian S., Eberhart C. G., Langevin S. M., Koestler D. C., Marsit C. J., Jin S. G., Kraus T. F., Kraus T. F., Muller T., Ahsan S., Ziller M. J., Hansen K. D., Meissner A., Aryee M. J., Matsubara K., Kafer G. R., Jin S. G., Wu X., Li A. X., Pfeifer G. P., Rampal R., Friedmann-Morvinski D., Hon G. C., Taylor S. E., Jackson M., Hassiotou F., Nowak A., Moen E. L., Stark A. L., Zhang W., Dolan M. E., Godley L. A., Aryee M. J., Chen Y. A., Dedeurwaerder S., McLean C. Y., Houseman E. A., Molitor J., Marsit C. J., Houseman E. A., Akalin A., Franke V., Vlahovicek K., Mason C. E., Schubeler D., and Suva M. L. 5-Hydroxymethylcytosine localizes to enhancer elements and is associated with survival in glioblastoma patients. *Nature Communications*, 7:13177, 2016. ISSN 2041-1723. doi: 10.1038/ncomms13177.

Johnson R. E., Prakash L., and Prakash S. Distinct mechanisms of cis-syn thymine dimer bypass by Dpo4 and DNA polymerase eta. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12359–12364, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0504380102.

Johnson R. E., Klassen R., Prakash L., and Prakash S. A Major Role of DNA Polymerase delta in Replication of Both the Leading and Lagging DNA Strands. *Molecular Cell*, 59(2):163–175, 2015. ISSN 10974164. doi: 10.1016/j.molcel.2015.05.038.

Jones P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, 2012. ISSN 1471-0056. doi: 10.1038/nrg3230.

Jones P. A. and Baylin S. B. The fundamental role of epigenetic events in cancer. *Nature reviews. Genetics*, 3(6):415–28, jun 2002. ISSN 1471-0056. doi: 10.1038/nrg816.

Jones P. A. and Liang G. Rethinking how DNA methylation patterns are maintained. *Nature reviews. Genetics*, 10(11):805–11, 2009. ISSN 1471-0064. doi: 10.1038/nrg2651.

Kamba K., Nagata T., and Katahira M. Catalytic analysis of APOBEC3G involving real-time NMR spectroscopy reveals nucleic acid determinants for deamination. *PLoS ONE*, 10(4):1–16, 2015. ISSN 19326203. doi: 10.1371/journal.pone.0124142.

Kamiya H. Mutations Induced by Oxidized DNA Precursors and Their Prevention by Nucleotide Pool Sanitization Enzymes. *Genes and Environment*, 29(4):133–140, 2007.

Kamiya H., Tsuchiya H., Karino N., Ueno Y., Matsuda A., and Harashima H. Mutagenicity of 5-Formylcytosine, an Oxidation Product of 5-Methylcytosine, in DNA in Mammalian Cells. *Journal of Biochemistry*, 132(4):551–555, oct 2002. ISSN 0021-924X. doi: 10.1093/oxfordjournals.jbchem.a003256.

Kane D. P. and Shcherbakova P. V. A common cancer-associated DNA polymerase ϵ mutation causes an exceptionally strong mutator phenotype, indicating fidelity defects distinct from loss of proofreading. *Cancer Research*, 74(7):1895–1901, 2014. ISSN 15387445. doi: 10.1158/0008-5472.CAN-13-2892.

Kanu N., Grönroos E., Martinez P., Burrell R. A., Yi Goh X., Bartkova J., Maya-Mendoza A., Mistrík M., Rowan A. J., Patel H., Rabinowitz A., East P., Wilson G., Santos C. R., McGranahan N., Gulati S., Gerlinger M., Birkbak N. J., Joshi T., Alexandrov L. B., Stratton M. R., Powles T., Matthews N., Bates P. A., Stewart A., Szallasi Z., Larkin J., Bartek J., and Swanton C. SETD2 loss-of-function promotes renal cancer branched

evolution through replication stress and impaired DNA repair. *Oncogene*, 34(46): 5699–5708, 2015. ISSN 0950-9232. doi: 10.1038/onc.2015.24.

Karran P. and Lindahl T. Hypoxanthine in Deoxyribonucleic Acid: Generation by Heat-Induced Hydrolysis of Adenine Residues and Release in Free Form by a Deoxyribonucleic Acid Glycosylase from Calf Thymus. *Biochemistry*, 19(26):6005–6011, 1980. ISSN 15204995. doi: 10.1021/bi00567a010.

Karras G. I., Fumasoni M., Sienski G., Vanoli F., and Branzei D. Article Noncanonical Role of the 9-1-1 Clamp in the Error-Free DNA Damage Tolerance Pathway. *Molecular Cell*, 49(3):536–546, 2013. ISSN 1097-2765. doi: 10.1016/j.molcel.2012.11.016.

Kass S. U., Landsberger N., and Wolffe A. P. DNA methylation directs a time-dependent repression of transcription initiation. *Current biology*, 7(3):157–165, 1997. ISSN 09609822. doi: 10.1016/S0960-9822(97)70086-1.

Katafuchi A. and Nohmi T. DNA polymerases involved in the incorporation of oxidized nucleotides into DNA: Their efficiency and template base preference. *Mutation Research - Genetic Toxicology and Environmental Mutagenesis*, 703(1):24–31, 2010. ISSN 13835718. doi: 10.1016/j.mrgentox.2010.06.004.

Katainen R., Dave K., Pitkänen E., Palin K., Kivioja T., Välimäki N., Gylfe A. E., Ristolainen H., Hänninen U. A., Cajuso T., Kondelin J., Tanskanen T., Mecklin J.-P., Järvinen H., Renkonen-Sinisalo L., Lepistö A., Kaasinen E., Kilpivaara O., Tuupanen S., Enge M., Taipale J., and Aaltonen L. A. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics*, advance on(7):818–21, 2015. ISSN 1061-4036. doi: 10.1038/ng.3335.

Kauppi J., Räsänen J., Sihvo E., Nieminen U., Arkkila P., Ahotupa M., and Salo J. Increased Oxidative Stress in the Proximal Stomach of Patients with Barrett's Esophagus and Adenocarcinoma of the Esophagus and Esophagogastric Junction. *Translational oncology*, 9(4):336–339, 2016. ISSN 1936-5233. doi: 10.1016/j.tranon.2016.06.004.

Kawamoto T., Araki K., Sonoda E., Yamashita Y. M., Harada K., Kikuchi K., Masutani C., Hanaoka F., Nozaki K., Hashimoto N., and Takeda S. Dual roles for DNA polymerase η in homologous DNA recombination and translesion DNA synthesis. *Molecular Cell*, 20(5):793–799, 2005. ISSN 10972765. doi: 10.1016/j.molcel.2005.10.016.

Kazanov M. D., Roberts S. A., Polak P., Stamatoyannopoulos J., Klimczak L. J., Gordenin D. A., and Sunyaev S. R. APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions. *Cell Reports*, 13(6):1103–1109, 2015. ISSN 22111247. doi: 10.1016/j.celrep.2015.09.077.

Kellinger M. W., Song C.-X., Chong J., Lu X.-Y., He C., and Wang D. 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nature Structural & Molecular Biology*, 19(8):831–833, 2012. ISSN 1545-9993. doi: 10.1038/nsmb.2346.

Kemmerich K., Dingler F. A., Rada C., and Neuberger M. S. 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 Research*, 40(13):6016–6025, 2012. ISSN 03051048. doi: 10.1093/nar/gks259.

Kennaway E. L. Further experiments on cancer-producing substances. *The Biochemical journal*, 24(2):497–504, 1930. ISSN 0264-6021.

Keshet I., Schlesinger Y., Farkash S., Rand E., Hecht M., Segal E., Pikarski E., Young R. A., Niveleau A., Cedar H., and Simon I. Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nature Genetics*, 38(2):149–153, 2006. ISSN 1061-4036. doi: 10.1038/ng1719.

Kettle J. G., Alwan H., Bista M., Breed J., Davies N. L., Eckersley K., Fillery S., Foote K. M., Goodwin L., Jones D. R., Käck H., Lau A., Nissink J. W. M., Read J., Scott J. S., Taylor B., Walker G., Wissler L., and Wylot M. Potent and Selective Inhibitors of MTH1 Probe Its Role in Cancer Cell Survival. *Journal of Medicinal Chemistry*, 59(6):2346–2361, 2016. ISSN 15204804. doi: 10.1021/acs.jmedchem.5b01760.

Khare T., Pai S., Koncevicius K., Pal M., Kriukiene E., Liutkeviciute Z., Irimia M., Jia P., Ptak C., Xia M., Tice R., Tochigi M., Morera S., Nazarians A., Belsham D., Wong A. H. C., Blencowe B. J., Wang S. C., Kapranov P., Kustra R., Labrie V., Klimasauskas S., and Petronis A. 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. *Nature structural & molecular biology*, 19(10):1037–43, oct 2012. ISSN 1545-9985. doi: 10.1038/nsmb.2372.

Kim J., Bowlby R., Mungall A. J., Robertson A. G., Odze R. D., Cherniack A. D., Shih J., Pedamallu C. S., Cibulskis C., Dunford A., Meier S. R., Kim J., Raphael B. J., Wu H.-T., Wong A. M., Willis J. E., Bass A. J., Derks S., Garman K., McCall S. J., Wiznerowicz M., Pantazi A., Parfenov M., Thorsson V., Shmulevich I., Dhankani V., Miller M., Sakai R., Wang K., Schultz N., Shen R., Arora A., Weinhold N., Sánchez-Vega F., Kelsen D. P., Zhang J., Felau I., Demchok J., Rabkin C. S., Camargo M. C., Zenklusen J. C., Bowen J., Leraas K., Lichtenberg T. M., Curtis C., Seoane J. A., Ojesina A. I., Beer D. G., Gulley M. L., Pennathur A., Luketich J. D., Zhou Z., Weisenberger D. J., Akbani R., Lee J.-S., Liu W., Mills G. B., Zhang W., Reid B. J., Hinoue T., Laird P. W., Shen H., Piazuelo M. B., Schneider B. G., McLellan M., Taylor-Weiner A., Cibulskis C., Lawrence M., Cibulskis K., Stewart C., Getz G., Lander E., Gabriel S. B., Ding L., McLellan M. D., Miller C. A., Appelbaum E. L., Cordes M. G., Fronick C. C., Fulton L. A., Mardis E. R., Wilson R. K., Schmidt H. K., Fulton R. S., Ally A., Balasundaram M., Bowlby R., Carlsen R., Chuah E., Dhalla N., Holt R. A., Jones S. J. M., Kasaian K., Brooks D., Li H. I., Ma Y., Marra M. A., Mayo M., Moore R. A., Mungall A. J., Mungall K. L., Robertson A. G., Schein J. E., Sipahimalani P., Tam A., Thiessen N., Wong T., Cherniack A. D., Shih J., Pedamallu C. S., Beroukhim R., Bullman S., Cibulskis C., Murray B. A., Saksena G., Schumacher S. E., Gabriel S., Meyerson M., Hadjipanayis A., Kucherlapati R., Pantazi A., Parfenov M., Ren X., Park P. J., Lee S., Kucherlapati M., Yang L., Baylin S. B., Hoadley K. A., Weisenberger D. J., Bootwalla M. S., Lai P. H., Van Den Berg D. J., Berrios M., Holbrook A., Akbani R., Hwang J.-E., Jang H.-J., Liu W., Weinstein J. N., Lee J.-S., Lu Y., Sohn B. H., Mills G., Seth S., Protopopov A., Bristow C. A., Mahadeshwar H. S., Tang J., Song X., Zhang J., Laird P. W., Hinoue T., Shen H., Cho J., Defrietas T.,

Frazer S., Gehlenborg N., Heiman D. I., Lawrence M. S., Lin P., Meier S. R., Noble M. S., Voet D., Zhang H., Kim J., Polak P., Saksena G., Chin L., Getz G., Wong A. M., Raphael B. J., Wu H.-T., Lee S., Park P. J., Yang L., Thorsson V., Bernard B., Iype L., Miller M., Reynolds S. M., Shmulevich I., Dhankani V., Abeshouse A., Arora A., Armenia J., Kundra R., Ladanyi M., Lehmann K.-V., Gao J., Sander C., Schultz N., Sánchez-Vega F., Shen R., Weinhold N., Chakravarty D., Zhang H., Radenbaugh A., Hegde A., Akbani R., Liu W., Weinstein J. N., Chin L., Bristow C. A., Lu Y., Penny R., Crain D., Gardner J., Curley E., Mallery D., Morris S., Paulauskis J., Shelton T., Shelton C., Bowen J., Frick J., Gastier-Foster J. M., Gerken M., Leraas K. M., Lichtenberg T. M., Ramirez N. C., Wise L., Zmuda E., Tarvin K., Saller C., Park Y. S., Button M., Carvalho A. L., Reis R. M., Matsushita M. M., Lucchesi F., de Oliveira A. T., Le X., Paklina O., Setdikova G., Lee J.-H., Bennett J., Iacocca M., Huelsenbeck-Dill L., Potapova O., Voronina O., Liu O., Fulidou V., Cates C., Sharp A., Behera M., Force S., Khuri F., Owonikoko T., Pickens A., Ramalingam S., Sica G., Dinjens W., van Nistelrooij A., Wijnhoven B., Sandusky G., Stepa S., Crain D., Paulauskis J., Penny R., Gardner J., Mallery D., Morris S., Shelton T., Shelton C., Curley E., Juhl H., Zornig C., Kwon S. Y., Kelsen D., Kim H. K., Bartlett J., Parfitt J., Chetty R., Darling G., Knox J., Wong R., El-Zimaity H., Liu G., Boussioutas A., Park D. Y., Kemp R., Carlotti C. G., da Cunha Tirapelli D. P., Saggioro F. P., Sankarankutty A. K., Noushmehr H., dos Santos J. S., Trevisan F. A., Eschbacher J., Dubina M., Mozgovoy E., Carey F., Chalmers S., Forgie I., Godwin A., Reilly C., Madan R., Naima Z., Ferrer-Torres D., Vinco M., Rathmell W. K., Dhir R., Luketich J., Pennathur A., Ajani J. A., McCall S. J., Janjigian Y., Kelsen D., Ladanyi M., Tang L., Camargo M. C., Ajani J. A., Cheong J.-H., Chudamani S., Liu J., Lolla L., Naresh R., Pihl T., Sun Q., Wan Y., Wu Y., Demchok J. A., Felau I., Ferguson M. L., Shaw K. R. M., Sheth M., Tarnuzzer R., Wang Z., Yang L., Zenklusen J. C., Hutter C. M., Sofia H. J., and Zhang J. Integrated genomic characterization of oesophageal carcinoma. *Nature*, 2017. ISSN 0028-0836. doi: 10.1038/nature20805.

Kim S.-i., Jin S.-G., and Pfeifer G. P. Formation of cyclobutane pyrimidine dimers at dipyrimidines containing 5-hydroxymethylcytosine. *Photochemical & Photobiological Sciences*, 12(8):1409, 2013. ISSN 1474-905X. doi: 10.1039/c3pp50037c.

Kinde B., Gabel H. W., Gilbert C. S., Griffith E. C., and Greenberg M. E. Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. *Proceedings of the National Academy of Sciences of the United States of America*, 112(22):6800–6, 2015. ISSN 1091-6490. doi: 10.1073/pnas.1411269112.

Klarer A. C., Stallons L. J., Burke T. J., Skaggs R. L., and McGregor W. G. DNA polymerase eta participates in the mutagenic bypass of adducts induced by benzo[a]pyrene diol epoxide in mammalian cells. *PLoS ONE*, 7(6):4–8, 2012. ISSN 19326203. doi: 10.1371/journal.pone.0039596.

Klaunig J. E. and Kamendulis L. M. The role of oxidative stress in carcinogenesis. *Annual review of pharmacology and toxicology*, 44(1):239–67, 2004. ISSN 0362-1642. doi: 10.1146/annurev.pharmtox.44.101802.121851.

Klutstein M., Nejman D., Greenfield R., and Cedar H. DNA methylation in cancer and aging. *Cancer Research*, 76(12):3446–3450, 2016. ISSN 15387445. doi: 10.1158/0008-5472.CAN-15-3278.

Klutstein M., Moss J., Kaplan T., and Cedar H. Contribution of epigenetic mechanisms to variation in cancer risk among tissues. *Proceedings of the National Academy of Sciences*, page 201616556, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1616556114.

Knudson A. G. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4):820–3, apr 1971. ISSN 0027-8424.

Kohli R. M. and Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, 502(7472):472–9, oct 2013. ISSN 1476-4687. doi: 10.1038/nature12750.

Kong A., Frigge M. L., Masson G., Besenbacher S., Sulem P., Magnusson G., Gudjonsson S. A., Sigurdsson A., Jonasdottir A., Jonasdottir A., Wong W. S. W., Sigurdsson G., Walters G. B., Steinberg S., Helgason H., Thorleifsson G., Gudbjartsson D. F., Helgason A., Magnusson O. T., Thorsteinsdottir U., and Stefansson K. Rate of de novo

mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–5, 2012. ISSN 1476-4687. doi: 10.1038/nature11396.

Koren A., Polak P., Nemesh J., Michaelson J. J., Sebat J., Sunyaev S. R., and McCarroll S. A. Differential relationship of DNA replication timing to different forms of human mutation and variation. *American Journal of Human Genetics*, 91(6):1033–1040, 2012. ISSN 00029297. doi: 10.1016/j.ajhg.2012.10.018.

Korona D. A., Lecompte K. G., and Pursell Z. F. The high fidelity and unique error signature of human DNA polymerase epsilon. *Nucleic Acids Research*, 39(5):1763–1773, 2011. ISSN 03051048. doi: 10.1093/nar/gkq1034.

Koziol M. J., Bradshaw C. R., Allen G. E., Costa A. S. H., Frezza C., and Gurdon J. B. Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nature Structural & Molecular Biology*, 23(1):24–30, 2015. ISSN 1545-9993. doi: 10.1038/nsmb.3145.

Kraus T. F. J., Kolck G., Greiner A., Schierl K., Guibourt V., and Kretzschmar H. A. Loss of 5-hydroxymethylcytosine and intratumoral heterogeneity as an epigenomic hallmark of glioblastoma. *Tumor Biology*, 2015. ISSN 1010-4283. doi: 10.1007/s13277-015-3606-9.

Krause L., Nones K., Loffler K. A., Nancarrow D., Oey H., Tang Y. H., Wayte N. J., Patch A. M., Patel K., Brosda S., Manning S., Lampe G., Clouston A., Thomas J., Stoye J., Hussey D. J., Watson D. I., Lord R. V., Phillips W. A., Gotley D., Mark Smithers B., Whiteman D. C., Hayward N. K., Grimmond S. M., Waddell N., and Barbour A. P. Identification of the CIMP-like subtype and aberrant methylation of members of the chromosomal segregation and spindle assembly pathways in esophageal adenocarcinoma. *Carcinogenesis*, 37(4):356–365, 2016. ISSN 14602180. doi: 10.1093/carcin/bgw018.

Kriaucionis S. and Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, 324(5929):929–30, may 2009. ISSN 1095-9203. doi: 10.1126/science.1169786.

Krokan H. E., Sætrom P., Aas P. A., Pettersen H. S., Kavli B., and Slupphaug G. Error-free versus mutagenic processing of genomic uracil-Relevance to cancer. *DNA Repair*, 19: 38–47, 2014. ISSN 15687856. doi: 10.1016/j.dnarep.2014.03.028.

Krueger F. and Andrews S. R. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btr167.

Kulis M., Heath S., Bibikova M., Queirós A. C., Navarro A., Clot G., Martínez-Trillos A., Castellano G., Brun-Heath I., Pinyol M., Barberán-Soler S., Papasaikas P., Jares P., Beà S., Rico D., Ecker S., Rubio M., Royo R., Ho V., Klotzle B., Hernández L., Conde L., López-Guerra M., Colomer D., Villamor N., Aymerich M., Rozman M., Bayes M., Gut M., Gelpí J. L., Orozco M., Fan J.-B., Quesada V., Puente X. S., Pisano D. G., Valencia A., López-Guillermo A., Gut I., López-Otín C., Campo E., and Martín-Subero J. I. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nature Genetics*, 44(11):1236–1242, 2012. ISSN 1546-1718. doi: 10.1038/ng.2443.

Lackey L., Law E. K., Brown W. L., and Harris R. S. Subcellular localization of the APOBEC3 proteins during mitosis and implications for genomic DNA deamination. *Cell Cycle*, 12(5):762–772, 2013. ISSN 15514005. doi: 10.4161/cc.23713.

Lagadu S., Lechevrel M., Sichel F., Breton J., Pottier D., Couderc R., Moussa F., and Prevost V. 8-oxo-7,8-dihydro-2'-deoxyguanosine as a biomarker of oxidative damage in oesophageal cancer patients: lack of association with antioxidant vitamins and polymorphism of hOGG1 and GST. *Journal of experimental & clinical cancer research : CR*, 29(1):157, 2010. ISSN 1756-9966. doi: 10.1186/1756-9966-29-157.

Landry S., Narvaiza I., Linfesty D. C., and Weitzman M. D. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. *EMBO reports*, 12(5):444–450, 2011. ISSN 1469-221X. doi: 10.1038/embor.2011.46.

Lang G. I. and Murray A. W. Mutation rates across budding yeast chromosome VI Are correlated with replication timing. *Genome Biology and Evolution*, 3(1):799–811, 2011. ISSN 17596653. doi: 10.1093/gbe/evr054.

Lange S. S., Takata K.-i., and Wood R. D. DNA polymerases and cancer. *Nature Reviews Cancer*, 11(2):96–110, 2011. ISSN 1474-175X. doi: 10.1038/nrc2998.

Langemeijer S. M. C., Kuiper R. P., Berends M., Knops R., Aslanyan M. G., Massop M., Stevens-Linders E., van Hoogen P., van Kessel A. G., Raymakers R. A. P., Kamping E. J., Verhoef G. E., Verburgh E., Hagemeijer A., Vandenberghe P., de Witte T., van der Reijden B. A., and Jansen J. H. Acquired mutations in TET2 are common in myelodysplastic syndromes. *Nature genetics*, 41(7):838–42, jul 2009. ISSN 1546-1718. doi: 10.1038/ng.391.

Langley A. R., Gräf S., Smith J. C., and Krude T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic acids research*, 44(21):10230–10247, sep 2016. ISSN 1362-4962. doi: 10.1093/nar/gkw760.

Larson A. R., Dresser K. A., Zhan Q., Lezcano C., Woda B. A., Yosufi B., Thompson J. F., Scolyer R. A., Mihm M. C., Shi Y. G., Murphy G. F., and Lian C. G. Loss of 5-hydroxymethylcytosine correlates with increasing morphologic dysplasia in melanocytic tumors. *Modern Pathology*, 27(7):936–944, 2014. ISSN 0893-3952. doi: 10.1038/modpathol.2013.224.

Larson E. D., Iams K., and Drummond J. T. Strand-specific processing of 8-oxoguanine by the human mismatch repair pathway: Inefficient removal of 8-oxoguanine paired with adenine or cytosine. *DNA Repair*, 2(11):1199–1210, 2003. ISSN 15687864. doi: 10.1016/S1568-7864(03)00140-X.

Laurent L., Wong E., Li G., Huynh T., Tsirigos A., Ong C. T., Low H. M., Kin Sung K. W., Rigoutsos I., Loring J., and Wei C.-L. Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–31, mar 2010. ISSN 1549-5469. doi: 10.1101/gr.101907.109.

Law E. K., Sieuwerts A. M., LaPara K., Leonard B., Starrett G. J., Molan A. M., Temiz N. A., Vogel R. I., Meijer-van Gelder M. E., Sweep F. C. G. J., Span P. N., Foekens J. A., Martens J. W. M., Yee D., and Harris R. S. The DNA cytosine deaminase APOBEC3B promotes tamoxifen resistance in ER-positive breast cancer. *Science advances*, 2(10): e1601737, 2016. ISSN 2375-2548. doi: 10.1126/sciadv.1601737.

Lawrence M. S., Stojanov P., Polak P., Kryukov G. V., Cibulskis K., Sivachenko A., Carter S. L., Stewart C., Mermel C. H., Roberts S. A., Kiezun A., Hammerman P. S., McKenna A., Drier Y., Zou L., Ramos A. H., Pugh T. J., Stransky N., Helman E., Kim J., Sougnez C., Ambrogio L., Nickerson E., Shefler E., Cortés M. L., Auclair D., Saksena G., Voet D., Noble M., DiCara D., Lin P., Lichtenstein L., Heiman D. I., Fennell T., Imielinski M., Hernandez B., Hodis E., Baca S., Dulak A. M., Lohr J., Landau D.-A., Wu C. J., Melendez-Zajgla J., Hidalgo-Miranda A., Koren A., McCarroll S. A., Mora J., Lee R. S., Crompton B., Onofrio R., Parkin M., Winckler W., Ardlie K., Gabriel S. B., Roberts C. W. M., Biegel J. A., Stegmaier K., Bass A. J., Garraway L. A., Meyerson M., Golub T. R., Gordenin D. A., Sunyaev S., Lander E. S., and Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–8, jul 2013. ISSN 1476-4687. doi: 10.1038/nature12213.

Le T. D., Durham N., Smith N., Wang H., and Bartlett B. Mismatch-repair deficiency predicts response of solid tumors to PD-1 blockade. *Science*, 6733(June), 2017. doi: 10.1126/science.aan6733.

Le Page F., Guy A., Cadet J., Sarasin A., and Gentil A. Repair and mutagenic potency of 8-oxoG:A and 8-oxoG:C base pairs in mammalian cells. *Nucleic Acids Research*, 26(5): 1276–1281, 1998. ISSN 03051048. doi: 10.1093/nar/26.5.1276.

Lee D. D. and Seung H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, 1999. ISSN 0028-0836. doi: 10.1038/44565.

Lee D. H., Liu Y., Lee H. W., Xia B., Brice A. R., Park S. H., Balduf H., Dominy B. N., and Cao W. A structural determinant in the uracil DNA glycosylase superfamily for

the removal of uracil from adenine/uracil base pairs. *Nucleic Acids Research*, 43(2): 1081–1089, 2015a. ISSN 13624962. doi: 10.1093/nar/gku1332.

Lee D. H. and Pfeifer G. P. Deamination of 5-methylcytosines within cyclobutane pyrimidine dimers is an important component of UVB mutagenesis. *Journal of Biological Chemistry*, 278(12):10314–10321, 2003. ISSN 00219258. doi: 10.1074/jbc.M212696200.

Lee J. J., Cook M., Mihm M. C., Xu S., Zhan Q., Wang T. J., Murphy G. F., and Lian C. G. Loss of the epigenetic mark, 5-Hydroxymethylcytosine, correlates with small cell/nevoid subpopulations and assists in microstaging of human melanoma. *Oncotarget*, 6(35):37995–8004, 2015b. ISSN 1949-2553. doi: 10.18632/oncotarget.6062.

Leonard A. C. and Méchali M. DNA replication origins. *Cold Spring Harbor perspectives in biology*, 5(10):a010116, oct 2013. ISSN 1943-0264. doi: 10.1101/cshperspect.a010116.

Lev Maor G., Yearim A., and Ast G. The alternative role of DNA methylation in splicing regulation, 2015. ISSN 13624555.

Ley T. J., Ding L., Walter M. J., McLellan M. D., Lamprecht T., Larson D. E., Kandoth C., Payton J. E., Baty J., Welch J., Harris C. C., Lichti C. F., Townsend R. R., Fulton R. S., Dooling D. J., Koboldt D. C., Schmidt H., Zhang Q., Osborne J. R., Lin L., O'Laughlin M., McMichael J. F., Delehaunty K. D., McGrath S. D., Fulton L. A., Magrini V. J., Vickery T. L., Hundal J., Cook L. L., Conyers J. J., Swift G. W., Reed J. P., Alldredge P. A., Wylie T., Walker J., Kalicki J., Watson M. A., Heath S., Shannon W. D., Varghese N., Nagarajan R., Westervelt P., Tomasson M. H., Link D. C., Graubert T. A., DiPersio J. F., Mardis E. R., and Wilson R. K. DNMT3A Mutations in Acute Myeloid Leukemia. *New England Journal of Medicine*, 363(25):2424–2433, dec 2010. ISSN 0028-4793. doi: 10.1056/NEJMoa1005143.

Li E., Bestor T. H., and Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, 1992. ISSN 00928674. doi: 10.1016/0092-8674(92)90611-F.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., and Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp352.

Li M. M. H. and Emerman M. Polymorphism in Human APOBEC3H Affects a Phenotype Dominant for Subcellular Localization and Antiviral Activity. *Journal of Virology*, 85 (16):8197–8207, 2011. ISSN 0022-538X. doi: 10.1128/JVI.00624-11.

Li W. and Liu M. Distribution of 5-hydroxymethylcytosine in different human tissues. *Journal of nucleic acids*, 2011:870726, jan 2011. ISSN 2090-021X. doi: 10.4061/2011/870726.

Li X., Liu Y., Salz T., Hansen K. D., and Feinberg A. Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. *Genome research*, 26(12):1730–1741, 2016. ISSN 1549-5469. doi: 10.1101/gr.211854.116.

Lian C. G., Xu Y., Ceol C., Wu F., Larson A., Dresser K., Xu W., Tan L., Hu Y., Zhan Q., Lee C. W., Hu D., Lian B. Q., Kleffel S., Yang Y., Neiswender J., Khorasani A. J., Fang R., Lezcano C., Duncan L. M., Scolyer R. A., Thompson J. F., Kakavand H., Houvras Y., Zon L. I., Mihm M. C., Kaiser U. B., Schatton T., Woda B. A., Murphy G. F., and Shi Y. G. Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of Melanoma. *Cell*, 150 (6):1135–1146, 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.07.033.

Libertini E., Heath S. C., Hamoudi R. A., Gut M., Ziller M. J., Herrero J., Czyz A., Ruotti V., Stunnenberg H. G., Frontini M., Ouwehand W. H., Meissner A., Gut I. G., and Beck S. Saturation analysis for whole-genome bisulfite sequencing data. *Nature Biotechnology*, 34(7):11–13, 2016. ISSN 1087-0156. doi: 10.1038/nbt.3524.

Lindahl T. Instability and decay of the primary structure of DNA. *Nature*, 362(6422): 709–15, apr 1993. ISSN 0028-0836. doi: 10.1038/362709a0.

Lindahl T. and Nyberg B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry*, 13(16):3405–3410, jul 1974. ISSN 0006-2960. doi: 10.1021/bi00713a035.

Lister R., Pelizzola M., Dowen R. H., Hawkins R. D., Hon G., Tonti-Filippini J., Nery J. R., Lee L., Ye Z., Ngo Q.-M., Edsall L., Antosiewicz-Bourget J., Stewart R., Ruotti V., Millar A. H., Thomson J. A., Ren B., and Ecker J. R. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–22, nov 2009. ISSN 1476-4687. doi: 10.1038/nature08514.

Lister R., Mukamel E. A., Nery J. R., Urich M., Puddifoot C. A., Johnson N. D., Lucero J., Huang Y., Dwork A. J., Schultz M. D., Yu M., Tonti-Filippini J., Heyn H., Hu S., Wu J. C., Rao A., Esteller M., He C., Haghighi F. G., Sejnowski T. J., Behrens M. M., and Ecker J. R. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146):1237905, aug 2013. ISSN 1095-9203. doi: 10.1126/science.1237905.

Liu J. *Epigenetic shielding: 5-hydroxymethylcytosine and 5-carboxylcytosine modulate UV induction of DNA photoproducts*. PhD thesis, Harvard University, 2014.

Liu S., Wang J., Su Y., Guerrero C., Zeng Y., Mitra D., Brooks P. J., Fisher D. E., Song H., and Wang Y. Quantitative assessment of Tet-induced oxidation products of 5-methylcytosine in cellular and tissue DNA. *Nucleic Acids Research*, 41(13):6421–6429, 2013. ISSN 03051048. doi: 10.1093/nar/gkt360.

Liu X. In vitro chromatin templates to study nucleotide excision repair. *DNA Repair*, 36: 68–76, 2015. ISSN 15687856. doi: 10.1016/j.dnarep.2015.09.026.

Lock L. F., Takagi N., and Martin G. R. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell*, 48(1):39–46, jan 1987. ISSN 0092-8674. doi: 10.1016/0092-8674(87)90353-9.

Lodato M. A., Woodworth M. B., Lee S., Evrony G. D., Mehta B. K., Karger A., Lee S., Chittenden T. W., Gama A. M. D., Cai X., Luquette L. J., Lee E., Park P. J., and Walsh C. A. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*, 350(6256):94–97, 2015. ISSN 0036-8075. doi: 10.1126/science.aab1785.

Loeb L. A. Mutator Phenotype May Be Required for Multistage Carcinogenesis. *Cancer Research*, 51(12):3075–3079, jun 1991. ISSN 15387445.

Loeb L. A. and Harris C. C. Advances in chemical carcinogenesis: A historical review and prospective. *Cancer Research*, 68(17):6863–6872, 2008. ISSN 00085472. doi: 10.1158/0008-5472.CAN-08-2852.

Lord C. J. and Ashworth A. BRCAness revisited. *Nature Reviews Cancer*, 16(2):110–120, 2016. ISSN 1474-175X. doi: 10.1038/nrc.2015.21.

Lu X., Song C. X., Szulwach K., Wang Z., Weidenbacher P., Jin P., and He C. Chemical modification-assisted bisulfite sequencing (CAB-seq) for 5-carboxylcytosine detection in DNA. *Journal of the American Chemical Society*, 135(25):9315–9317, jun 2013. ISSN 00027863. doi: 10.1021/ja4044856.

Luger K., Mäder A. W., Richmond R. K., Sargent D. F., and Richmond T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389(6648): 251–260, 1997. ISSN 0028-0836. doi: 10.1038/38444.

Lujan S. A., Williams J. S., Pursell Z. F., Abdulovic-Cui A. A., Clark A. B., Nick McElhinny S. A., and Kunkel T. A. Mismatch Repair Balances Leading and Lagging Strand DNA Replication Fidelity. *PLoS Genetics*, 8(10):e1003016, 2012. ISSN 15537390. doi: 10.1371/journal.pgen.1003016.

Lujan S. A., Clausen A. R., Clark A. B., MacAlpine H. K., MacAlpine D. M., Malc E. P., Mieczkowski P. A., Burkholder A. B., Fargo D. C., Gordenin D. A., and Kunkel T. A. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Research*, 24(11):1751–1764, 2014. ISSN 15495469. doi: 10.1101/gr.178335.114.

Lujan S. A., Williams J. S., and Kunkel T. A. DNA Polymerases Divide the Labor of Genome Replication. *Trends in Cell Biology*, 26(9):640–654, 2016. ISSN 18793088. doi: 10.1016/j.tcb.2016.04.012.

Luo C. and Ecker J. R. Exceptional epigenetics in the brain. *Science*, 348(6239):1094–1095, 2015. ISSN 0036-8075. doi: 10.1126/science.aac5832.

Luo G.-Z., Blanco M. A., Greer E. L., He C., and Shi Y. DNA N6-methyladenine: a new epigenetic mark in eukaryotes? *Nature Reviews Molecular Cell Biology*, 16(12):705–710, 2015. ISSN 1471-0072. doi: 10.1038/nrm4076.

Lynch H. T., Lynch P. M., Lanspa S. J., Snyder C. L., Lynch J. F., and Boland C. R. Review of the Lynch syndrome: History, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clinical Genetics*, 76(1):1–18, 2009. ISSN 00099163. doi: 10.1111/j.1399-0004.2009.01230.x.

Mack S. C., Witt H., Piro R. M., Gu L., Zuyderduyn S., Stutz A. M., Wang X., Gallo M., Garzia L., Zayne K., Zhang X., Ramaswamy V., Jager N., Jones D. T., Sill M., Pugh T. J., Ryzhova M., Wani K. M., Shih D. J., Head R., Remke M., Bailey S. D., Zichner T., Faria C. C., Barszczyk M., Stark S., Seker-Cin H., Hutter S., Johann P., Bender S., Hovestadt V., Tzaridis T., Dubuc A. M., Northcott P. A., Peacock J., Bertrand K. C., Agnihotri S., Cavalli F. M., Clarke I., Nethery-Brokx K., Creasy C. L., Verma S. K., Koster J., Wu X., Yao Y., Milde T., Sin-Chan P., Zuccaro J., Lau L., Pereira S., Castelo-Branco P., Hirst M., Marra M. A., Roberts S. S., Fults D., Massimi L., Cho Y. J., Van Meter T., Grajkowska W., Lach B., Kulozik A. E., von Deimling A., Witt O., Scherer S. W., Fan X., Muraszko K. M., Kool M., Pomeroy S. L., Gupta N., Phillips J., Huang A., Tabori U., Hawkins C., Malkin D., Kongkham P. N., Weiss W. A., Jabado N., Rutka J. T., Bouffet E., Korbel J. O., Lupien M., Aldape K. D., Bader G. D., Eils R., Lichter P., Dirks P. B., Pfister S. M., Korshunov A., and Taylor M. D. Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature*, 506(7489):445–450, 2014. ISSN 1476-4687. doi: 10.1038/nature13108.

Macpherson P., Barone F., Maga G., Mazzei F., Karran P., and Bignami M. 8-Oxoguanine incorporation into DNA repeats vitro and mismatch recognition by MutSα. *Nucleic Acids Research*, 33(16):5094–5105, 2005. ISSN 03051048. doi: 10.1093/nar/gki813.

Maga G., Villani G., Crespan E., Wimmer U., Ferrari E., Bertocci B., and Hübscher U. 8-oxo-guanine bypass by human DNA polymerases in the presence of auxiliary proteins. *Nature*, 447(7144):606–608, 2007. ISSN 0028-0836. doi: 10.1038/nature05843.

Mahfoudhi E., Talhaoui I., Cabagnols X., Della Valle V., Secardin L., Rameau P., Bernard O. A., Ishchenko A. A., Abbes S., Vainchenker W., Saparbaev M., and Plo I. TET2-mediated 5-hydroxymethylcytosine induces genetic instability and mutagenesis. *DNA Repair*, 43:78–88, 2016. ISSN 15687856. doi: 10.1016/j.dnarep.2016.05.031.

Mailand N., Gibbs-Seymour I., and Bekker-Jensen S. Regulation of PCNA–protein interactions for genome stability. *Nature Reviews Molecular Cell Biology*, 14(5):269–282, 2013. ISSN 1471-0072. doi: 10.1038/nrm3562.

Maiti A. and Drohat A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. *Journal of Biological Chemistry*, 286(41):35334–35338, 2011. ISSN 00219258. doi: 10.1074/jbc.C111.284620.

Malla S., Kadimisetty K., Fu Y.-J., Choudhary D., Schenkman J. B., and Rusling J. F. Methyl-Cytosine-Driven Structural Changes Enhance Adduction Kinetics of an Exon 7 fragment of the p53 Gene. *Scientific Reports*, 7(January):40890, 2017. ISSN 2045-2322. doi: 10.1038/srep40890.

Mao P., Smerdon M. J., Roberts S. A., and Wyrick J. J. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, 113(32):9057–9062, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1606667113.

Mao P., Wyrick J. J., Roberts S. A., and Smerdon M. J. UV-Induced DNA Damage and Mutagenesis in Chromatin. *Photochemistry and Photobiology*, 93(1):216–228, 2017. ISSN 17511097. doi: 10.1111/php.12646.

Mardis E. R., Ding L., Dooling D. J., Larson D. E., McLellan M. D., Chen K., Koboldt D. C., Fulton R. S., Delehaunty K. D., McGrath S. D., Fulton L. A., Locke D. P., Magrini V. J., Abbott R. M., Vickery T. L., Reed J. S., Robinson J. S., Wylie T., Smith S. M., Carmichael L., Eldred J. M., Harris C. C., Walker J., Peck J. B., Du F., Dukes A. F., Sanderson G. E., Brummett A. M., Clark E., McMichael J. F., Meyer R. J., Schindler J. K., Pohl C. S.,

Wallis J. W., Shi X., Lin L., Schmidt H., Tang Y., Haipek C., Wiechert M. E., Ivy J. V., Kalicki J., Elliott G., Ries R. E., Payton J. E., Westervelt P., Tomasson M. H., Watson M. A., Baty J., Heath S., Shannon W. D., Nagarajan R., Link D. C., Walter M. J., Graubert T. A., DiPersio J. F., Wilson R. K., and Ley T. J. Recurring mutations found by sequencing an acute myeloid leukemia genome. *The New England journal of medicine*, 361(11):1058–66, sep 2009. ISSN 1533-4406. doi: 10.1056/NEJMoa0903840.

Margueron R. and Reinberg D. Chromatin structure and the inheritance of epigenetic information. *Nature Reviews Genetics*, 11(4):285–296, 2010. ISSN 1471-0056. doi: 10.1038/nrg2752.

Marina R. J. and Oberdoerffer S. Epigenomics meets splicing through the TETs and CTCF. *Cell Cycle*, 15(11):1397–1399, 2016. ISSN 15514005. doi: 10.1080/15384101.2016.1171650.

Marina R. J., Sturgill D., Bailly M. A., Thenoz M., Varma G., Prigge M. F., Nanan K. K., Shukla S., Haque – N., and Oberdoerffer S. TET-catalyzed oxidation of intragenic 5-methylcytosine regulates CTCF-dependent alternative splicing. *The EMBO Journal*, 35(3):1–21, 2015. ISSN 1460-2075. doi: 10.15252/embj.

Markkanen E., Dorn J., and Hübscher U. MUTYH DNA glycosylase: The rationale for removing undamaged bases from the DNA. *Frontiers in Genetics*, 4(FEB):1–20, 2013. ISSN 16648021. doi: 10.3389/fgene.2013.00018.

Marnett L. J. and Plastaras J. P. Endogenous DNA damage and mutation. *Trends in Genetics*, 17(4):214–221, 2001. ISSN 01689525. doi: 10.1016/S0168-9525(01)02239-9.

Marteijn J. A., Lans H., Vermeulen W., and Hoeijmakers J. H. J. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature Reviews Molecular Cell Biology*, 15(7):465–481, 2014. ISSN 1471-0072. doi: 10.1038/nrm3822.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200.

Martincorena I. and Campbell P. J. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015. ISSN 0036-8075. doi: 10.1126/science.aab4082.

Martinez-Fernandez L., Banyasz A., Esposito L., Markovitsi D., and Improta R. UV-induced damage to DNA: effect of cytosine methylation on pyrimidine dimerization. *Signal Transduction and Targeted Therapy*, 2(January):17021, 2017. ISSN 2059-3635. doi: 10.1038/sigtrans.2017.21.

Martinez-Useros J., Li W., Cabeza-Morales M., and Garcia-Foncillas J. Oxidative Stress: A New Target for Pancreatic Cancer Prognosis and Treatment. *Journal of Clinical Medicine*, 6(3):29, 2017. ISSN 2077-0383. doi: 10.3390/jcm6030029.

Mathews C. K. Deoxyribonucleotide metabolism, mutagenesis and cancer. *Nature Publishing Group*, 15(9):528–539, 2015. ISSN 1474-175X. doi: 10.1038/nrc3981.

Maunakea A. K., Nagarajan R. P., Bilenky M., Ballinger T. J., D'Souza C., Fouse S. D., Johnson B. E., Hong C., Nielsen C., Zhao Y., Turecki G., Delaney A., Varhol R., Thiessen N., Shchors K., Heine V. M., Rowitch D. H., Xing X., Fiore C., Schillebeeckx M., Jones S. J. M., Haussler D., Marra M. A., Hirst M., Wang T., and Costello J. F. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303): 253–7, 2010. ISSN 1476-4687. doi: 10.1038/nature09165.

Maunakea A. K., Chepelev I., Cui K., and Zhao K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell research*, 23(11):1256–69, nov 2013. ISSN 1748-7838. doi: 10.1038/cr.2013.110.

Mayer W., Niveleau A., Walter J., Fundele R., and Haaf T. Embryogenesis: Demethylation of the zygotic paternal genome. *Nature*, 403(6769):501–502, 2000. ISSN 0028-0836. doi: 10.1038/35000656.

McAuley-Hecht K. E., Leonard G. A., Gibson N. J., Thomson J. B., Watson W. P., Hunter W. N., and Brown T. Crystal Structure of a DNA Duplex Containing 8-Hydroxydeoxyguanine-adenine Base Pairs. *Biochemistry*, 33(34):10266–10270, 1994. ISSN 0006-2960. doi: 10.1021/bi00200a006.

McCulloch S. D. and Kunkel T. A. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Research*, 18:148–161, 2008. doi: 1001-0602/08.

McCulloch S. D., Kokoska R. J., Masutani C., Iwai S., Hanaoka F., and Kunkel T. A. Preferential cis–syn thymine dimer bypass by DNA polymerase η occurs with biased fidelity. *Nature*, 428(6978):97–100, 2004. ISSN 0028-0836. doi: 10.1038/nature02352.

McGinty R. K. and Tan S. Nucleosome structure and function. *Chemical Reviews*, 115(6): 2255–2273, 2015. ISSN 15206890. doi: 10.1021/cr500373h.

McGregor W. G., Wei D., Maher V. M., and McCormick J. J. Abnormal, Error-Prone Bypass of Photoproducts by Xeroderma Pigmentosum Variant Cell Extracts Results in Extreme Strand Bias for the Kinds of Mutations Induced by UV Light. *Molecular and Cellular Biology*, 19(1):147–154, 1999. ISSN 0270-7306.

McLaren W., Gil L., Hunt S. E., Riat H. S., Ritchie G. R. S., Thormann A., Flicek P., and Cunningham F. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0974-4.

McIlwraith M. J., Vaisman A., Liu Y., Fanning E., Woodgate R., and West S. C. Human DNA polymerase η promotes DNA synthesis from strand invasion intermediates of homologous recombination. *Molecular Cell*, 20(5):783–792, 2005. ISSN 10972765. doi: 10.1016/j.molcel.2005.10.001.

Méchali M. Eukaryotic DNA replication origins: many choices for appropriate answers. *Nature reviews. Molecular cell biology*, 11(10):728–738, 2010. ISSN 1471-0072. doi: 10.1038/nrm2976.

Mellén M., Ayata P., Dewell S., Kriaucionis S., and Heintz N. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell*, 151(7):1417–1430, 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.11.022.

Menck C. F. M. and Munford V. DNA repair diseases: What do they tell us about cancer and aging? *Genetics and Molecular Biology*, 37(1 SUPPL. 1):220–233, 2014. ISSN 16784685. doi: 10.1590/S1415-47572014000200008.

Menoni H., Shukla M. S., Gerson V., Dimitrov S., and Angelov D. Base excision repair of 8-oxoG in dinucleosomes. *Nucleic Acids Research*, 40(2):692–700, 2012. ISSN 03051048. doi: 10.1093/nar/gkr761.

Menoni H., Di Mascio P., Cadet J., Dimitrov S., and Angelov D. Chromatin associated mechanisms in base excision repair - nucleosome remodeling and DNA transcription, two key players. *Free Radical Biology and Medicine*, 107(September 2016):159–169, 2017. ISSN 18734596. doi: 10.1016/j.freeradbiomed.2016.12.026.

Mertz T. M., Baranovskiy A. G., Wang J., Tahirov T. H., and Shcherbakova P. V. Nucleotide selectivity defect and mutator phenotype conferred by a colon cancer-associated DNA polymerase δ mutation in human cells. *Oncogene*, 36(31):4427–4433, 2017a. ISSN 0950-9232. doi: 10.1038/onc.2017.22.

Mertz T. M., Sharma S., Chabes A., and Shcherbakova P. V. Colon cancer-associated mutator DNA polymerase δ variant causes expansion of dNTP pools increasing its own infidelity. *Proceedings of the National Academy of Sciences of the United States of America*, 112(19):E2467–76, may 2015. ISSN 1091-6490. doi: 10.1073/pnas.1422934112.

Mertz T. M., Harcy V., and Roberts S. A. Risks at the DNA Replication Fork: Effects upon Carcinogenesis and Tumor Heterogeneity. *Genes*, 8(1):46, 2017b. ISSN 2073-4425. doi: 10.3390/genes8010046.

Mi R., Dong L., Kaulgud T., Hackett K. W., Dominy B. N., and Cao W. Insights from Xanthine and Uracil DNA Glycosylase Activities of Bacterial and Human SMUG1: Switching SMUG1 to UDG. *Journal of Molecular Biology*, 385(3):761–778, 2009. ISSN 00222836. doi: 10.1016/j.jmb.2008.09.038.

Millar C. B. Enhanced CpG Mutability and Tumorigenesis in MBD4-Deficient Mice. *Science*, 297(5580):403–405, 2002. ISSN 00368075. doi: 10.1126/science.1073354.

Ming X., Matter B., Song M., Veliath E., Shanley R., and Jones R. Mapping Structurally Defined Guanine Oxidation Products along. *Journal of the American Chemical Society*, 136:4223–4235, 2014.

Mitchell D. L. Effects of Cytosine Methylation on Pyrimidine Dimer Formation in DNA. *Photochemistry and Photobiology*, 71(2):162–165, may 2007. ISSN 00318655. doi: 10.1562/0031-8655(2000)0710162EOCMOP2.0.CO2.

Mittlböck M. and Heinzl H. Pseudo R-squared measures for generalized linear models. In *Proceedings of the 1st European Workshop on the Assessment of Diagnostic Performance*, pages 71–80, Milan, Italy, 2004.

Miyabe I., Mizuno K., Keszthelyi A., Daigaku Y., Skouteri M., Mohebi S., Kunkel T. A., Murray J. M., and Carr A. M. Polymerase δ replicates both strands after homologous recombination–dependent fork restart. *Nature Structural & Molecular Biology*, 22 (October):1–8, 2015. ISSN 1545-9993. doi: 10.1038/nsmb.3100.

Moarefi A. H. and Chédin F. ICF syndrome mutations cause a broad spectrum of biochemical defects in DNMT3B-mediated de novo DNA methylation. *Journal of Molecular Biology*, 409(5):758–772, 2011. ISSN 00222836. doi: 10.1016/j.jmb.2011.04.050.

Moldovan G. L., Pfander B., and Jentsch S. PCNA, the Maestro of the Replication Fork. *Cell*, 129(4):665–679, 2007. ISSN 00928674. doi: 10.1016/j.cell.2007.05.003.

Moore P. S. and Chang Y. Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nat Rev Cancer*, 10(12):878–889, 2010. ISSN 1474-175X. doi: 10.1038/nrc2961.

Moorjani P., Amorim C. E. G., Arndt P. F., and Przeworski M. Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences*, 113(38):10607–10612, 2016. ISSN 0027-8424. doi: 10.1101/036434.

Moran S., Martinez-Cardús A., Boussios S., and Esteller M. Precision medicine based on epigenomics: the paradigm of carcinoma of unknown primary. *Nature Reviews Clinical Oncology*, 2017. ISSN 1759-4774. doi: 10.1038/nrclinonc.2017.97.

Moréra S., Grin I., Vigouroux A., Couvé S., Henriot V., Saparbaev M., and Ishchenko A. A. Biochemical and structural characterization of the glycosylase domain of MBD4 bound to thymine and 5-hydroxymethyuracil-containing DNA. *Nucleic Acids Research*, 40(19):9917–9926, 2012. ISSN 03051048. doi: 10.1093/nar/gks714.

Morganella S., Alexandrov L. B., Glodzik D., Zou X., Davies H., Staaf J., Sieuwerts A. M., Brinkman A. B., Martin S., Ramakrishna M., Butler A., Kim H.-Y., Borg Å., Sotiriou C., Futreal P. A., Campbell P. J., Span P. N., Van Laere S., Lakhani S. R., Eyfjord J. E., Thompson A. M., Stunnenberg H. G., van de Vijver M. J., Martens J. W. M., Børresen-Dale A.-L., Richardson A. L., Kong G., Thomas G., Sale J., Rada C., Stratton M. R., Birney E., and Nik-Zainal S. The topography of mutational processes in breast cancer genomes. *Nature Communications*, 7(May 2016):11383, 2016. ISSN 2041-1723. doi: 10.1038/ncomms11383.

Morison I. M., Ramsay J. P., and Spencer H. G. A census of mammalian imprinting. *Trends in Genetics*, 21(8):457–465, 2005. ISSN 01689525. doi: 10.1016/j.tig.2005.06.008.

Moura F. A., de Andrade K. Q., dos Santos J. C. F., Araújo O. R. P., and Goulart M. O. F. Antioxidant therapy for treatment of inflammatory bowel disease: Does it work? *Redox Biology*, 6:617–639, 2015. ISSN 22132317. doi: 10.1016/j.redox.2015.10.006.

Mudrak S. V., Welz-Voegele C., and Jinks-Robertson S. The polymerase eta translesion synthesis DNA polymerase acts independently of the mismatch repair system to limit mutagenesis caused by 7,8-dihydro-8-oxoguanine in yeast. *Molecular and cellular biology*, 29(19):5316–26, 2009. ISSN 1098-5549. doi: 10.1128/MCB.00422-09.

Mugal C. F. and Ellegren H. Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. *Genome Biology*, 12(6): R58, 2011. ISSN 1465-6906. doi: 10.1186/gb-2011-12-6-r58.

Münzel M., Lischke U., Stathis D., Pfaffeneder T., Gnerlich F. A., Deiml C. A., Koch S. C., Karaghiosoff K., and Carell T. Improved synthesis and mutagenicity

of oligonucleotides containing 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine. *Chemistry - A European Journal*, 17(49):13782–13788, 2011. ISSN 09476539. doi: 10.1002/chem.201102782.

Murugaesu N., Wilson G. A., Birkbak N. J., Watkins T. B. K., McGranahan N., Kumar S., Abbassi-Ghadi N., Salm M., Mitter R., Horswell S., Rowan A., Phillimore B., Biggs J., Begum S., Matthews N., Hochhauser D., Hanna G. B., and Swanton C. Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discovery*, 5(8):821–832, 2015. ISSN 21598290. doi: 10.1158/2159-8290.CD-15-0412.

Murugan A. K., Bojdani E., and Xing M. Identification and functional characterization of isocitrate dehydrogenase 1 (IDH1) mutations in thyroid cancer. *Biochemical and biophysical research communications*, 393(3):555–9, mar 2010. ISSN 1090-2104. doi: 10.1016/j.bbrc.2010.02.095.

Mustafi S., Sant D. W., Liu Z.-J., and Wang G. Ascorbate induces apoptosis in melanoma cells by suppressing Clusterin expression. *Scientific Reports*, 7(1):3671, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-03893-5.

Nabel C. S., Jia H., Ye Y., Shen L., Goldschmidt H. L., Stivers J. T., Zhang Y., and Kohli R. M. AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nature Chemical Biology*, 8(9):751–758, 2012. ISSN 1552-4450. doi: 10.1038/nchembio.1042.

Nakabeppu Y., Ohta E., and Abolhassani N. MTH1 as a nucleotide pool sanitizing enzyme: Friend or foe? *Free Radical Biology and Medicine*, 107(November 2016):151–158, 2017. ISSN 18734596. doi: 10.1016/j.freeradbiomed.2016.11.002.

Narita T., Tsurimoto T., Yamamoto J., Nishihara K., Ogawa K., Ohashi E., Evans T., Iwai S., Takeda S., and Hirota K. Human replicative DNA polymerase δ can bypass T-T (6-4) ultraviolet photoproducts on template strands. *Genes to cells : devoted to molecular & cellular mechanisms*, 15(12):1228–39, dec 2010. ISSN 1365-2443. doi: 10.1111/j.1365-2443.2010.01457.x.

Neeley W. L. and Essigmann J. M. Mechanisms of formation, genotoxicity, and mutation of guanine oxidation products. *Chemical Research in Toxicology*, 19(4):491–505, 2006. ISSN 0893228X. doi: 10.1021/tx0600043.

Neri F., Rapelli S., Krepelova A., Incarnato D., Parlato C., Basile G., Maldotti M., Anselmi F., and Oliviero S. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, 543(7643):72–77, 2017. ISSN 0028-0836. doi: 10.1038/nature21373.

Nestor C. E., Ottaviano R., Reddington J., Sproul D., Reinhardt D., Dunican D., Katz E., Dixon J. M., Harrison D. J., and Meehan R. R. Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Research*, pages 467–477, 2012. doi: 10.1101/gr.126417.111.

Nick McElhinny S. A., Kissling G. E., and Kunkel T. A. Differential correction of lagging-strand replication errors made by DNA polymerases α and δ. *Proceedings of the National Academy of Sciences of the United States of America*, 107(49):21070–21075, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1013048107.

Nielsen R., Paul J. S., Albrechtsen A., and Song Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011. ISSN 1471-0056. doi: 10.1038/nrg2986.

Nik-Zainal S., Alexandrov L. B., Wedge D. C., Van Loo P., Greenman C. D., Raine K., Jones D., Hinton J., Marshall J., Stebbings L. A., Menzies A., Martin S., Leung K., Chen L., Leroy C., Ramakrishna M., Rance R., Lau K. W., Mudie L. J., Varela I., McBride D. J., Bignell G. R., Cooke S. L., Shlien A., Gamble J., Whitmore I., Maddison M., Tarpey P. S., Davies H. R., Papaemmanuil E., Stephens P. J., McLaren S., Butler A. P., Teague J. W., Jönsson G., Garber J. E., Silver D., Miron P., Fatima A., Boyault S., Langerød A., Tutt A., Martens J. W. M., Aparicio S. A. J. R., Borg Å., Salomon A. V., Thomas G., Børresen-Dale A.-L., Richardson A. L., Neuberger M. S., Futreal P. A., Campbell P. J., Stratton M. R., and Breast Cancer Working Group of the International

Cancer Genome Consortium. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–93, may 2012a. ISSN 1097-4172. doi: 10.1016/j.cell.2012.04.024.

Nik-Zainal S., Van Loo P., Wedge D. C., Alexandrov L. B., Greenman C. D., Lau K. W., Raine K., Jones D., Marshall J., Ramakrishna M., Shlien A., Cooke S. L., Hinton J., Menzies A., Stebbings L. A., Leroy C., Jia M., Rance R., Mudie L. J., Gamble S. J., Stephens P. J., McLaren S., Tarpey P. S., Papaemmanuil E., Davies H. R., Varela I., McBride D. J., Bignell G. R., Leung K., Butler A. P., Teague J. W., Martin S., Jönsson G., Mariani O., Boyault S., Miron P., Fatima A., Langerød A., Aparicio S. A. J. R., Tutt A., Sieuwerts A. M., Borg Å., Thomas G., Salomon A. V., Richardson A. L., Børresen-Dale A.-L., Futreal P. A., Stratton M. R., Campbell P. J., and Breast Cancer Working Group of the International Cancer Genome Consortium. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, may 2012b. ISSN 1097-4172. doi: 10.1016/j.cell.2012.04.023.

Nik-Zainal S., Kucab J. E., Morganella S., Glodzik D., Alexandrov L. B., Arlt V. M., Weninger A., Hollstein M., Stratton M. R., and Phillips D. H. The genome as a record of environmental exposure. *Mutagenesis*, 30(October):763–770, 2015. ISSN 14643804. doi: 10.1093/mutage/gev073.

Nilsen H., Haushalter K. A., Robins P., Barnes D. E., Verdine G. L., and Lindahl T. Excision of deaminated cytosine from the vertebrate genome: Role of the SMUG1 uracil-DNA glycosylase. *EMBO Journal*, 20(15):4278–4286, 2001. ISSN 02614189. doi: 10.1093/emboj/20.15.4278.

Nishigaki M., Aoyagi K., and Danjoh I. Discovery of Aberrant Expression of R-RAS by Cancer-Linked DNA Hypomethylation in Gastric Cancer Using Microarrays DNA Hypomethylation in Gastric Cancer Using Microarrays. *Cancer Research*, 65(6):2115–2124, 2005.

Nones K., Waddell N., Wayte N., Patch A.-M., Bailey P., Newell F., Holmes O., Fink J. L., Quinn M. C. J., Tang Y. H., Lampe G., Quek K., Loffler K. A., Manning S., Idrisoglu S., Miller D., Xu Q., Waddell N., Wilson P. J., Bruxner T. J. C., Christ A. N., Harliwong I.,

Nourse C., Nourbakhsh E., Anderson M., Kazakoff S., Leonard C., Wood S., Simpson P. T., Reid L. E., Krause L., Hussey D. J., Watson D. I., Lord R. V., Nancarrow D., Phillips W. A., Gotley D., Smithers B. M., Whiteman D. C., Hayward N. K., Campbell P. J., Pearson J. V., Grimmond S. M., and Barbour A. P. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nature Communications*, 5:1–9, 2015. ISSN 2041-1723. doi: 10.1038/ncomms6224.

Nordentoft I., Lamy P., Birkenkamp-Demtr??der K., Shumansky K., Vang S., Hornsh??j H., Juul M., Villesen P., Hedegaard J., Roth A., Thorsen K., H??yer S., Borre M., Reinert T., Fristrup N., Dyrskj??t L., Shah S., Pedersen J. S., and ??rntoft T. F. Mutational context and diverse clonal development in early and late bladder cancer. *Cell Reports*, 7(5):1649–1663, 2014. ISSN 22111247. doi: 10.1016/j.celrep.2014.04.038.

Noushmehr H., Weisenberger D. J., Diefes K., Phillips H. S., Pujara K., Berman B. P., Pan F., Pelloski C. E., Sulman E. P., Bhat K. P., Verhaak R. G., Hoadley K. A., Hayes D. N., Perou C. M., Schmidt H. K., Ding L., Wilson R. K., Van Den Berg D., Shen H., Bengtsson H., Neuvial P., Cope L. M., Buckley J., Herman J. G., Baylin S. B., Laird P. W., and Aldape K. Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell*, 17(5):510–522, 2010. ISSN 15356108. doi: 10.1016/j.ccr.2010.03.017.

Nowak J. A., Yurgelun M. B., Bruce J. L., Rojas-Rudilla V., Hall D. L., Shivdasani P., Garcia E. P., Agoston A. T., Srivastava A., Ogino S., Kuo F. C., Lindeman N. I., and Dong F. Detection of Mismatch Repair Deficiency and Microsatellite Instability in Colorectal Adenocarcinoma by Targeted Next-Generation Sequencing. *The Journal of Molecular Diagnostics*, 19(1):84–91, 2017. ISSN 15251578. doi: 10.1016/j.jmoldx.2016.07.010.

Nowell P. and Hungerford D. A minute chromosome in human chronic granulocytic leukemia. *Science*, 132, 1960.

O'Brien J. M., Beal M. A., Yauk C. L., and Marchetti F. Next generation sequencing of benzo(a)pyrene-induced lacZ mutants identifies a germ cell-specific mutation

spectrum. *Scientific Reports*, 6(October):36743, 2016. ISSN 2045-2322. doi: 10.1038/srep36743.

Ogoshi K., Hashimoto S.-i., Nakatani Y., Qu W., Oshima K., Tokunaga K., Sugano S., Hattori M., Morishita S., and Matsushima K. Genome-wide profiling of DNA methylation in human cancer cells. *Genomics*, 98(4):280–287, 2011. ISSN 08887543. doi: 10.1016/j.ygeno.2011.07.003.

Okano M., Bell D. W., Haber D. A., and Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99 (3):247–257, 1999. ISSN 00928674. doi: 10.1016/S0092-8674(00)81656-6.

Okazaki R., Okazaki T., Sakabe K., Sugimoto K., and Sugino A. Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. *Proceedings of the National Academy of Sciences of the United States of America*, 59(2):598–605, feb 1968. ISSN 0027-8424. doi: 10.1073/pnas.59.2.598.

Olins A. L. and Olins D. E. Spheroid chromatin units (v bodies). *Science (New York, N.Y.)*, 183(4122):330–2, jan 1974. ISSN 0036-8075. doi: 10.1126/science.183.4122.330.

Olmon E. D. and Delaney S. Differential Ability of Five DNA Glycosylases to Recognize and Repair Damage on Nucleosomal DNA. *ACS Chemical Biology*, 12(3):692–701, 2017. ISSN 15548937. doi: 10.1021/acschembio.6b00921.

Oshimo Y., Nakayama H., Ito R., Kitadai Y., Yoshida K., Chayama K., and Yasui W. Promoter methylation of cyclin D2 gene in gastric carcinoma. *International Journal of Oncology*, 23(6):1663–1670(8), 2003.

Oswald J., Engemann S., Lane N., Mayer W., Olek A., Fundele R., Dean W., Reik W., and Walter J. Active demethylation of the paternal genome in the mouse zygote. *Current Biology*, 10(8):475–478, 2000. ISSN 09609822. doi: 10.1016/S0960-9822(00)00448-6.

Pacis A., Tailleux L., Morin A. M., Lambourne J., MacIsaac J. L., Yotova V., Dumaine A., Danckaert A., Luca F., Grenier J.-c., Hansen K. D., Gicquel B., Yu M., Pai A., He C., Tung J., Pastinen T., Kobor M. S., Pique-Regi R., Gilad Y., and Barreiro L. B. Bacterial

infection remodels the DNA methylation landscape of human dendritic cells. *Genome research*, 25(12):1801–11, 2015. ISSN 1549-5469. doi: 10.1101/gr.192005.115.

Palles C., Cazier J.-B., Howarth K. M., Domingo E., Jones A. M., Broderick P., Kemp Z., Spain S. L., Guarino E., Guarino Almeida E., Salguero I., Sherborne A., Chubb D., Carvajal-Carmona L. G., Ma Y., Kaur K., Dobbins S., Barclay E., Gorman M., Martin L., Kovac M. B., Humphray S., Lucassen A., Holmes C. C., Bentley D., Donnelly P., Taylor J., Petridis C., Roylance R., Sawyer E. J., Kerr D. J., Clark S., Grimes J., Kearsey S. E., Thomas H. J. W., McVean G., Houlston R. S., and Tomlinson I. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature genetics*, 45(2):136–44, 2013. ISSN 1546-1718. doi: 10.1038/ng.2503.

Pandya-Jones A. and Black D. L. Co-transcriptional splicing of constitutive and alternative exons. *RNA (New York, N.Y.)*, 15(10):1896–908, oct 2009. ISSN 1469-9001. doi: 10.1261/rna.1714509.

Pansuriya T. C., van Eijk R., D'Adamo P., van Ruler M. A. J. H., Kuijjer M. L., Oosting J., Cleton-Jansen A.-M., van Oosterwijk J. G., Verbeke S. L. J., Meijer D., van Wezel T., Nord K. H., Sangiorgi L., Toker B., Liegl-Atzwanger B., San-Julian M., Sciot R., Limaye N., Kindblom L.-G., Daugaard S., Godfraind C., Boon L. M., Vikkula M., Kurek K. C., Szuhai K., French P. J., and Bovée J. V. M. G. Somatic mosaic IDH1 and IDH2 mutations are associated with enchondroma and spindle cell hemangioma in Ollier disease and Maffucci syndrome. *Nature genetics*, 43(12):1256–61, dec 2011. ISSN 1546-1718. doi: 10.1038/ng.1004.

Parker M., Mohankumar K. M., Punchihewa C., Weinlich R., Dalton J. D., Li Y., Lee R., Tatevossian R. G., Phoenix T. N., Thiruvenkatam R., White E., Tang B., Orisme W., Gupta K., Rusch M., Chen X., Li Y., Nagahawhatte P., Hedlund E., Finkelstein D., Wu G., Shurtleff S., Easton J., Boggs K., Yergeau D., and Vadodaria B. C11orf95–RELA fusions drive oncogenic NF-kB signalling in ependymoma. *Nature*, 506:451–554, 2014. ISSN 0028-0836. doi: 10.1038/nature13109.

Parrinello S., Samper E., Krtolica A., Goldstein J., Melov S., and Campisi J. Oxygen sensitivity severely limits the replicative lifespan of murine fibroblasts. *Nature cell biology*, 5(8):741–747, 2003. ISSN 14657392. doi: 10.1038/ncb1024.

Parsons D. W., Jones S., Zhang X., Lin J. C.-H., Leary R. J., Angenendt P., Mankoo P., Carter H., Siu I.-M., Gallia G. L., Olivi A., McLendon R., Rasheed B. A., Keir S., Nikolskaya T., Nikolsky Y., Busam D. A., Tekleab H., Diaz L. A., Hartigan J., Smith D. R., Strausberg R. L., Marie S. K. N., Shinjo S. M. O., Yan H., Riggins G. J., Bigner D. D., Karchin R., Papadopoulos N., Parmigiani G., Vogelstein B., Velculescu V. E., and Kinzler K. W. An integrated genomic analysis of human glioblastoma multiforme. *Science*, 321(5897):1807–1812, sep 2008. ISSN 1095-9203. doi: 10.1126/science.1164382.

Patro J. N., Urban M., and Kuchta R. D. Interaction of human DNA polymerase α and DNA polymerase I from Bacillus stearothermophilus with hypoxanthine and 8-oxoguanine nucleotides. *Biochemistry*, 48(34):8271–8278, 2009. ISSN 00062960. doi: 10.1021/bi900777s.

Pavlov Y. I., Newlon C. S., and Kunkel T. A. Yeast origins establish a strand bias for replicational mutagenesis. *Molecular Cell*, 10(1):207–213, 2002. ISSN 10972765. doi: 10.1016/S1097-2765(02)00567-1.

Pavlov Y. I., Mian I. M., and Kunkel T. A. Evidence for Preferential Mismatch Repair of Lagging Strand DNA Replication Errors in Yeast. *Current Biology*, 13:744–748, 2003. doi: 10.1016/S0960.

Pavlova O., Fraitag S., and Hohl D. 5-Hydroxymethylcytosine Expression in Proliferative Nodules Arising within Congenital Nevi Allows Differentiation from Malignant Melanoma. *Journal of Investigative Dermatology*, 136(12):2453–2461, 2016. ISSN 15231747. doi: 10.1016/j.jid.2016.07.015.

Pellegrini L. The Pol α-primase complex. *Sub-cellular biochemistry*, 62:157–69, 2012. ISSN 0306-0225. doi: 10.1007/978-94-007-4572-8_9.

Peña-Diaz J. and Jiricny J. Mammalian mismatch repair: Error-free or error-prone? *Trends in Biochemical Sciences*, 37(5):206–214, 2012. ISSN 09680004. doi: 10.1016/j.tibs.2012.03.001.

Peña-Diaz J., Bregenhorn S., Ghodgaonkar M., Follonier C., Artola-Borán M., Castor D., Lopes M., Sartori A. A., and Jiricny J. Noncanonical mismatch repair as a source of genomic instability in human cells. *Molecular Cell*, 47(5):669–680, sep 2012.

Pereira C., Coelho R., Grácio D., Dias C., Silva M., Peixoto A., Lopes P., Costa C., Teixeira J. P., Macedo G., and Magro F. DNA damage and oxidative DNA damage in inflammatory bowel disease. *Journal of Crohn's and Colitis*, 10(11):1316–1323, 2016. ISSN 18764479. doi: 10.1093/ecco-jcc/jjw088.

Peroja P., Pasanen A., Haapasaari K.-M., Jantunen E., Soini Y., Turpeenniemi-Hujanen T., Bloigu R., Lilja L., Kuittinen O., and Karihtala P. Oxidative stress and redox state-regulating enzymes have prognostic relevance in diffuse large B-cell lymphoma. *Experimental Hematology & Oncology*, 1(1):2, 2012. ISSN 2162-3619. doi: 10.1186/2162-3619-1-2.

Peters J. The role of genomic imprinting in biology and disease: an expanding view. *Nature reviews. Genetics*, 15(8):517–530, 2014. ISSN 1471-0064. doi: 10.1038/nrg3766.

Petryk N., Kahli M., D'Aubenton-Carafa Y., Y J., Shen Y, Maud S., Thermes C., Chen C.-L., Hyrien O., and (*co coreponding). Replication landscape of the human genome. *Nature communications*, 7, 2016. ISSN 2041-1723. doi: 10.1038/ncomms10208.

Pfaffeneder T., Spada F., Wagner M., Brandmayr C., Laube S. K., Eisen D., Truss M., Steinbacher J., Hackner B., Kotljarova O., Schuermann D., Michalakis S., Kosmatchev O., Schiesser S., Steigenberger B., Raddaoui N., Kashiwazaki G., Müller U., Spruijt C. G., Vermeulen M., Leonhardt H., Schär P., Müller M., and Carell T. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nature chemical biology*, 10(7):574–81, jul 2014. ISSN 1552-4469. doi: 10.1038/nchembio.1532.

Pfeifer G. P. p53 mutational spectra and the role of methylated CpG sequences. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 450(1-2):155–166, 2000. ISSN 00275107. doi: 10.1016/S0027-5107(00)00022-1.

Pfeifer G. P. An elusive DNA base in mammals. *Nature*, 532:319–320, 2016. doi: 10.1038/nature17315.

Pfeifer G. P. and Besaratinia A. UV wavelength-dependent DNA damage and human non-melanoma and melanoma skin cancer. *Photochemical & photobiological sciences : Official journal of the European Photochemistry Association and the European Society for Photobiology*, 11(1):90–7, 2012. ISSN 1474-9092. doi: 10.1039/c1pp05144j.

Pfeifer G. P., Denissenko M. F., Olivier M., Tretyakova N., Hecht S. S., and Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene*, 21(48):7435–51, oct 2002. ISSN 0950-9232. doi: 10.1038/sj.onc.1205803.

Pfeifer G. P., You Y.-H., and Besaratinia A. Mutations induced by ultraviolet light. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 571(1): 19–31, 2005. ISSN 00275107. doi: 10.1016/j.mrfmmm.2004.06.057.

Pfeifer G. P., Kadam S., and Jin S.-G. 5-Hydroxymethylcytosine and Its Potential Roles in Development and Cancer. *Epigenetics & Chromatin*, 6(1):10, 2013. ISSN 1756-8935. doi: 10.1186/1756-8935-6-10.

Pidsley R., Zotenko E., Peters T. J., Lawrence M. G., Risbridger G. P., Molloy P., Van Djik S., Muhlhausler B., Stirzaker C., Clark S. J., Jones P., Baylin S., Ko Y., Mohtat D., Suzuki M., Park A., Izquierdo M., Han S., Dayeh T., Volkov P., Salo S., Hall E., Nilsson E., Olsson A., Pidsley R., Viana J., Hannon E., Spiers H., Troakes C., Al-Saraj S., Stirzaker C., Taberlay P., Statham A., Clark S., Clark S., Harrison J., Paul C., Frommer M., Lister R., Pelizzola M., Dowen R., Hawkins R., Hon G., Tonti-Filippini J., Bibikova M., Le J., Barnes B., Saedinia-Melnyk S., Zhou L., Shen R., Hinoue T., Weisenberger D., Lange C., Shen H., Byun H., Berg D., Breitling L., Yang R., Korn B., Burwinkel B.,

Brenner H., Rakyan V., Down T., Maslau S., Andrew T., Yang T., Beyan H., Bibikova M., Barnes B., Tsan C., Ho V., Klotzle B., Le J., Morris T., Beck S., Chen Y., Choufani S., Grafodatskaya D., Butcher D., Ferreira J., Weksberg R., Chen Y., Lemire M., Choufani S., Butcher D., Grafodatskaya D., Zanke B., Naeem H., Wong N., Chatterton Z., Hong M., Pedersen J., Corcoran N., Peters T., Buckley M., Statham A., Pidsley R., Samaras K., Lord R. V., Wang D., Yan L., Hu Q., Sucheston L., Higgins M., Ambrosone C., Warden C., Lee H., Tompkins J., Li X., Wang C., Riggs A., Lizio M., Harshbarger J., Shimoji H., Severin J., Kasukawa T., Sahin S., Siggens L., Ekwall K., Dedeurwaerder S., Defrance M., Calonne E., Denis H., Sotiriou C., Fuks F., Pidsley R., CC Y. W., Volta M., Lunnon K., Mill J., Schalkwyk L., Teschendorff A., Marabita F., Lechner M., Bartlett T., Tegner J., Gomez-Cabrero D., Touleimat N., Tost J., Thurman R., Rynes E., Humbert R., Vierstra J., Maurano M., Haugen E., Andersson R., Gebhard C., Miguel-Escalada I., Hoof I., Bornholdt J., Boyd M., Kundaje A., Meuleman W., Ernst J., Bilenky M., Yen A., Ritchie M., Phipson B., Wu D., Hu Y., Law C., Shi W., Stadler M., Murr R., Burger L., Ivanek R., Lienert F., Schöler A., Ziller M., Gu H., Müller F., Donaghey J., Tsai L.-Y., Kohlbacher O., Huang S., Bao B., Hour T., Huang C., Yu C., Liu C., Neuhausen S., Slattery M., Garner C., Ding Y., Hoffman M., Brothman A., Reams R., Kalari K., Wang H., Odedina F., Soliman K., Yates C., Song J., Stirzaker C., Harrison J., Melki J., Clark S., Coolen M., Stirzaker C., Song J., Statham A., Kassir Z., Moreno C., Makrides M., Gibson R., McPhee A., Yelland L., Quinlivan J., Ryan P., Lawrence M., Taylor R., Toivanen R., Pedersen J., Norden S., Pook D., Clark S., Statham A., Stirzaker C., Molloy P., Frommer M., Auton A., Brooks L., Durbin R., Garrison E., Kang H., and Kent W. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17(1):208, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1066-1.

Piraino S. W. and Furney S. J. Identification of coding and non-coding mutational hotspots in cancer genomes. *BMC genomics*, 18(1):17, 2017. ISSN 1471-2164. doi: 10.1186/s12864-016-3420-9.

Pleasance E. D., Stephens P. J., O'Meara S., McBride D. J., Meynert A., Jones D., Lin M. L., Beare D., Lau K. W., Greenman C., Varela I., Nik-Zainal S., Davies H. R., Ordonez G. R., Mudie L. J., Latimer C., Edkins S., Stebbings L., Chen L., Jia M., Leroy C., Marshall J., Menzies A., Butler A., Teague J. W., Mangion J., Sun Y. A., McLaughlin S. F., Peckham H. E., Tsung E. F., Costa G. L., Lee C. C., Minna J. D., Gazdar A., Birney E., Rhodes M. D., McKernan K. J., Stratton M. R., Futreal P. A., and Campbell P. J. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278):184–190, 2010. ISSN 0028-0836. doi: 10.1038/nature08629.

Plongthongkum N., Diep D. H., and Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nature Reviews Genetics*, 15(10): 647–661, 2014. ISSN 1471-0056. doi: 10.1038/nrg3772.

Pohl H. and Welch H. G. The role of overdiagnosis and reclassification in the marked increase of esophageal adenocarcinoma incidence. *Journal of the National Cancer Institute*, 97(2):142–146, 2005. ISSN 00278874. doi: 10.1093/jnci/dji024.

Polak P., Lawrence M. S., Haugen E., Stoletzki N., Stojanov P., Thurman R. E., Garraway L. A., Mirkin S., Getz G., Stamatoyannopoulos J. A., and Sunyaev S. R. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nature biotechnology*, 32(1):71–5, 2014. ISSN 1546-1696. doi: 10.1038/nbt.2778.

Poon S. L., Pang S.-T., McPherson J. R., Yu W., Huang K. K., Guan P., Weng W.-H., Siew E. Y., Liu Y., Heng H. L., Chong S. C., Gan A., Tay S. T., Lim W. K., Cutcutache I., Huang D., Ler L. D., Nairismägi M.-L., Lee M. H., Chang Y.-H., Yu K.-J., Chan-On W., Li B.-K., Yuan Y.-F., Qian C.-N., Ng K.-F., Wu C.-F., Hsu C.-L., Bunte R. M., Stratton M. R., Futreal P. A., Sung W.-K., Chuang C.-K., Ong C. K., Rozen S. G., Tan P., and Teh B. T. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Science translational medicine*, 5(197):197ra101, 2013. ISSN 1946-6242. doi: 10.1126/scitranslmed.3006086.

Pope B. D., Ryba T., Dileep V., Yue F., Wu W., Denas O., Vera D. L., Wang Y., Hansen R. S., Canfield T. K., Thurman R. E., Cheng Y., Gülsoy G., Dennis J. H., Snyder M. P., Stamatoyannopoulos J. A., Taylor J., Hardison R. C., Kahveci T., Ren B., and Gilbert D. M. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405, 2014. ISSN 0028-0836. doi: 10.1038/nature13986.

Poulogiannis G., Frayling I. M., and Arends M. J. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. *Histopathology*, 56(2):167–179, 2010. ISSN 03090167. doi: 10.1111/j.1365-2559.2009.03392.x.

Poulos R. C., Olivier J., and Wong J. W. The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. *Nucleic Acids Research*, pages 1–10, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx463.

Prasad R., Singh T., and Katiyar S. K. Honokiol inhibits ultraviolet radiation-induced immunosuppression through inhibition of ultraviolet-induced inflammation and DNA hypermethylation in mouse skin. *Scientific Reports*, 7(1):1657, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-01774-5.

Prendergast G. C. and Ziff E. B. Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science (New York, N.Y.)*, 251(4990):186–9, 1991. ISSN 0036-8075. doi: 10.1126/science.1987636.

Pugh T. J., Weeraratne S. D., Archer T. C., Pomeranz Krummel D. A., Auclair D., Bochicchio J., Carneiro M. O., Carter S. L., Cibulskis K., Erlich R. L., Greulich H., Lawrence M. S., Lennon N. J., McKenna A., Meldrim J., Ramos A. H., Ross M. G., Russ C., Shefler E., Sivachenko A., Sogoloff B., Stojanov P., Tamayo P., Mesirov J. P., Amani V., Teider N., Sengupta S., Francois J. P., Northcott P. A., Taylor M. D., Yu F., Crabtree G. R., Kautzman A. G., Gabriel S. B., Getz G., Jäger N., Jones D. T. W., Lichter P., Pfister S. M., Roberts T. M., Meyerson M., Pomeroy S. L., and Cho Y.-J. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature*, 488(7409):106–10, aug 2012. ISSN 1476-4687. doi: 10.1038/nature11329.

Quinet A., Martins D. J., Vessoni A. T., Biard D., Sarasin A., Stary A., and Menck C. F. M. Translesion synthesis mechanisms depend on the nature of DNA damage in UV-irradiated human cells. *Nucleic Acids Research*, 44(12):5717–5731, 2016. ISSN 13624962. doi: 10.1093/nar/gkw280.

Rahbari R., Wuster A., Lindsay S. J., Hardwick R. J., Alexandrov L. B., Al Turki S., Dominiczak A., Morris A., Porteous D., Smith B., Stratton M. R., and Hurles M. E. Timing, rates and spectra of human germline mutation. *Nature Genetics*, 48(December):1–11, 2015. ISSN 1061-4036. doi: 10.1038/ng.3469.

Raiber E.-A., Murat P., Chirgadze D. Y., Beraldi D., Luisi B. F., and Balasubramanian S. 5-Formylcytosine alters the structure of the DNA double helix. *Nature structural & molecular biology*, 22(1):44–9, 2015. ISSN 1545-9985. doi: 10.1038/nsmb.2936.

Raiber E.-A., Beraldi D., Martínez Cuesta S., McInroy G. R., Kingsbury Z., Becq J., James T., Lopes M., Allinson K., Field S., Humphray S., Santarius T., Watts C., Bentley D., and Balasubramanian S. Base resolution maps reveal the importance of 5-hydroxymethylcytosine in a human glioblastoma. *npj Genomic Medicine*, 2(1):6, 2017. ISSN 2056-7944. doi: 10.1038/s41525-017-0007-6.

Rakyan V., Down T., Balding D., and Beck S. Epigenome-wide association studies for common human diseases. *Nature reviews. Genetics*, 12(8):529–541, 2011. ISSN 1471-0056. doi: 10.1038/nrg3000.

Rangam G., Schmitz K. M., Cobb A. J. A., and Petersen-Mahrt S. K. AID enzymatic activity is inversely proportional to the size of cytosine c5 orbital cloud. *PLoS ONE*, 7 (8):3–8, 2012. ISSN 19326203. doi: 10.1371/journal.pone.0043279.

Rasanen J. V., Sihvo E. I. T., Ahotupa M. O., Färkkilä M. A., and Salo J. A. The expression of 8-hydroxydeoxyguanosine in oesophageal tissues and tumours. *European Journal of Surgical Oncology*, 33(10):1164–1168, 2007. ISSN 07487983. doi: 10.1016/j.ejso.2007.03.003.

Rashid M., Fischer A., Wilson C. H., Tiffen J., Rust A. G., Stevens P., Idziaszczyk S., Maynard J., Williams G. T., Mustonen V., Sampson J. R., and Adams D. J. Adenoma development in familial adenomatous polyposis and MUTYH-associated polyposis: Somatic landscape and driver genes. *Journal of Pathology*, 238(1):98–108, 2016. ISSN 10969896. doi: 10.1002/path.4643.

Raynal N. J.-M., Si J., Taby R. F., Gharibyan V., Ahmed S., Jelinek J., Estecio M. R. H., and Issa J.-P. J. DNA Methylation Does Not Stably Lock Gene Expression but Instead Serves as a Molecular Mark for Gene Silencing Memory. *Cancer Research*, 72(5): 1170–1181, 2012. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-11-3248.

Rayner E., van Gool I. C., Palles C., Kearsey S. E., Bosse T., Tomlinson I., and Church D. N. A panoply of errors: polymerase proofreading domain mutations in cancer. *Nature Reviews Cancer*, 16(2):71–81, 2016. ISSN 1474-175X. doi: 10.1038/nrc.2015.12.

Ream T. S., Haag J. R., and Pikaard C. S. Nucleic Acid Polymerases. *Book*, 30:289–308, 2014. doi: 10.1007/978-3-642-39796-7.

Rebhandl S. AID / APOBEC deaminases and cancer. *Oncoscience*, 2(4), 2015. doi: 10.18632/oncoscience.155.

Rechache N. S., Wang Y., Stevenson H. S., Killian J. K., Edelman D. C., Merino M., Zhang L., Nilubol N., Stratakis C. A., Meltzer P. S., and Kebebew E. DNA methylation profiling identifies global methylation differences and markers of adrenocortical tumors. *Journal of Clinical Endocrinology and Metabolism*, 97(6):1004–1013, 2012. ISSN 0021972X. doi: 10.1210/jc.2011-3298.

Reddy E. P., Reynolds R. K., Santos E., and Barbacid M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature*, 300(5888):149–52, nov 1982. ISSN 0028-0836.

Reid-Bayliss K. S., Arron S. T., Loeb L. A., Bezrookove V., and Cleaver J. E. Why Cockayne syndrome patients do not get cancer despite their DNA repair deficiency. *Proceedings*

of the National Academy of Sciences, 113(36):201610020, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1610020113.

Reijns M. A. M., Kemp H., Ding J., de Procé S. M., Jackson A. P., and Taylor M. S. Lagging-strand replication shapes the mutational landscape of the genome. *Nature*, 518(7540): 502–506, 2015. ISSN 0028-0836. doi: 10.1038/nature14183.

Rey L., Sidorova J. M., Puget N., Boudsocq F., Biard D. S. F., Monnat R. J., Cazaux C., and Hoffmann J.-S. Human DNA polymerase eta is required for common fragile site stability during unperturbed DNA replication. *Molecular and cellular biology*, 29(12): 3344–54, jun 2009. ISSN 1098-5549. doi: 10.1128/MCB.00115-09.

Rhind N. and Gilbert D. M. DNA Replication Timing. *Cold Spring Harb Perspect Med*, 3: 1–26, 2013. ISSN 2157-1422. doi: 10.1101/cshperspect.a010132.

Rieke D., Messerschmidt C., and Ochsenreither S. Association of an APOBEC mutational signature, mutational load, and BRCAness with inflammation and PD-L1 expression in HNSCC. *Journal of Clinical Oncology*, 2017.

Riggs A. D. X inactivation, differentiation, and DNA methylation. *Cytogenetics and cell genetics*, 14(1):9–25, 1975. ISSN 0301-0171.

Rivera-Mulia J. C., Buckley Q., Sasaki T., Zimmerman J., Didier R. A., Nazor K., Loring J. F., Lian Z., Weissman S., Robins A. J., Schulz T. C., Menendez L., Kulik M. J., Dalton S., Gabr H., Kahveci T., and Gilbert D. M. Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. *Genome*, 25:1091–1103, 2015. doi: 10.1101/gr.187989.114.

Roberts S. A. and Gordenin D. A. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat Rev Cancer*, 14(12):786–800, 2014. ISSN 1474-1768. doi: 10.1038/nrc3816.

Roberts S. A., Sterling J., Thompson C., Harris S., Mav D., Shah R., Klimczak L. J., Kryukov G. V., Malc E., Mieczkowski P. A., Resnick M. A., and Gordenin D. A. Clustered

Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. *Molecular Cell*, 46(4):424–435, 2012. ISSN 10972765. doi: 10.1016/j.molcel.2012.03.030.

Roberts S. A., Lawrence M. S., Klimczak L. J., Grimm S. A., Fargo D., Stojanov P., Kiezun A., Kryukov G. V., Carter S. L., Saksena G., Harris S., Shah R. R., Resnick M. A., Getz G., and Gordenin D. A. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genetics*, 45(9):970–976, 2013. ISSN 1061-4036. doi: 10.1038/ng.2702.

Robertson K. D. DNA methylation and human disease. *Nature Reviews Genetics*, 6: 597–610, 2005. ISSN 1471-0056. doi: 10.1038/nrg1655.

Robles A. I., Traverso G., Zhang M., Roberts N. J., Khan M. A., Joseph C., Lauwers G. Y., Selaru F. M., Popoli M., Pittman M. E., Ke X., Hruban R. H., Meltzer S. J., Kinzler K. W., Vogelstein B., Harris C. C., and Papadopoulos N. Whole-Exome Sequencing Analyses of Inflammatory Bowel Disease-Associated Colorectal Cancers. *Gastroenterology*, 150 (4):931–943, 2016. ISSN 15280012. doi: 10.1053/j.gastro.2015.12.036.

Rochette P. J., Lacoste S., Therrien J. P., Bastien N., Brash D. E., and Drouin R. Influence of cytosine methylation on ultraviolet-induced cyclobutane pyrimidine dimer formation in genomic DNA. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 665(1-2):7–13, 2009. ISSN 00275107. doi: 10.1016/j.mrfmmm.2009.02.008.

Rodriguez G. P., Song J. B., and Crouse G. F. In Vivo Bypass of 8-oxodG. *PLoS Genetics*, 9(8), 2013. ISSN 15537390. doi: 10.1371/journal.pgen.1003682.

Rose A. S., Bradley A. R., Valasatava Y., Duarte J. M., Prlić A., and Rose P. W. Web-based molecular graphics for large complexes. In *Proceedings of the 21st International Conference on Web3D Technology - Web3D '16*, pages 185–186, 2016. ISBN 9781450344289. doi: 10.1145/2945292.2945324.

Rose N. R. and Klose R. J. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochimica et Biophysica Acta - Gene Regulatory*

Mechanisms, 1839(12):1362–1372, 2014. ISSN 18764320. doi: 10.1016/j.bbagrm.2014.02.007.

Rosenthal R., McGranahan N., Herrero J., Taylor B. S., and Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1):31, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0893-4.

Ross-Innes C. S., Becq J., Warren A., Cheetham R. K., Northen H., O'Donovan M., Malhotra S., di Pietro M., Ivakhno S., He M., Weaver J. M. J., Lynch A. G., Kingsbury Z., Ross M., Humphray S., Bentley D., Fitzgerald R. C., Hayes S. J., Ang Y., Welch I., Preston S., Oakes S., Save V., Skipworth R., Tucker O., Davies J., Crichton C., Schusterreiter C., Underwood T., Noble F., Stacey B., Kelly J., Byrne J., Haydon A., Sharland D., Owsley J., Barr H., Lagergren J., Gossage J., Davies A., Mason R., Chang F., Zylstra J., Sanders G., Wheatley T., Berrisford R., Bracey T., Harden C., Bunting D., Roques T., Nobes J., Loo S., Lewis M., Cheong E., Priest O., Parsons S. L., Soomro I., Kaye P., Saunders J., Pang V., Welch N. T., Catton J. A., Duffy J. P., Ragunath K., Lovat L., Haidry R., Miah H., Kerr S., Eneh V., Butawan R., Roques T., Lewis M., Cheong E., Kumar B., Igali L., Walton S., Dann A., Safranek P., Hindmarsh A., Sudjendran V., Scott M., Cluroe A., Miremadi A., Mahler-Araujo B., Nutzinger B., Peters C., Abdullahi Z., Crawte J., MacRae S., Noorani A., Elliott R. F., Bower L., Edwards P., Tavare S., Eldridge M., Bornschein J., Secrier M., Yang T.-P., O'Neill J. R., Adamczuk K., Lao-Sirieix P., Grehan N., Smith L., Lishman S., Beardsmore D., and Dawson S. Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nature Genetics*, 47(July):1–11, 2015. ISSN 1061-4036. doi: 10.1038/ng.3357.

Rouhani F. J., Nik-Zainal S., Wuster A., Li Y., Conte N., Koike-Yusa H., Kumasaka N., Vallier L., Yusa K., and Bradley A. Mutational History of a Human Cell Lineage from Somatic to Induced Pluripotent Stem Cells. *PLoS Genetics*, 12(4):1–15, 2016. ISSN 15537404. doi: 10.1371/journal.pgen.1005932.

Rowley J. D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, 243 (5405):290–3, jun 1973. ISSN 0028-0836. doi: 10.1038/243290a0.

Rudd S. G., Bianchi J., and Doherty A. J. PrimPol—A new polymerase on the block. *Molecular & Cellular Oncology*, 1(2):e960754, 2014. ISSN 2372-3556. doi: 10.4161/23723548.2014.960754.

Rudd S. G., Valerie N. C. K., and Helleday T. Pathways controlling dNTP pools to maintain genome stability. *DNA repair*, 44:193–204, 2016. ISSN 1568-7856. doi: 10.1016/j.dnarep.2016.05.032.

Russo M. T., Blasi M. F., Chiera F., Fortini P., Degan P., Macpherson P., Furuichi M., Nakabeppu Y., Karran P., Aquilina G., and Bignami M. The Oxidized Deoxynucleoside Triphosphate Pool Is a Significant Contributor to Genetic Instability in Mismatch Repair-Deficient Cells. *Molecular and cellular biology*, 24(1):465–474, 2004. ISSN 0270-7306. doi: 10.1128/MCB.24.1.465.

Russo V. E. A. V. E. A., Martienssen R. A., and Riggs A. D. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, 1996. ISBN 0879694904.

Ryba T., Hiratani I., Lu J., Itoh M., Kulik M., Zhang J., Schulz T. C., Robins A. J., Dalton S., and Gilbert D. M. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Research*, 20(6):761–770, jun 2010.

Ryba T., Battaglia D., Pope B. D., Hiratani I., and Gilbert D. M. Genome-scale analysis of replication timing: from bench to bioinformatics. *Nature Protocols*, 6(6):870–895, 2011. ISSN 1754-2189. doi: 10.1038/nprot.2011.328.

Sakai T., Toguchida J., Ohtani N., Yandell D. W., Rapaport J. M., and Dryja T. P. Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *American journal of human genetics*, 48(5):880–8, may 1991. ISSN 0002-9297.

Sale J. E., Lehmann A. R., and Woodgate R. Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nature reviews. Molecular cell biology*, 13(3): 141–52, 2012. ISSN 1471-0080. doi: 10.1038/nrm3289.

Sanjiv K., Gad H., Rudd S., Hurley R., Herr P., Montaño J. M. C., Mortusewicz O., Koolmeister T., Jaques S., Morón E. B., Hoglund A., Lee T.-C., Scobie M., Kaufmann S., Weroha J., Berglund U. W., Hendrickson A. W., and Helleday T. Abstract 1260: Polymerase kappa determines the sensitivity of MTH1 inhibitors to cisplatin-resistant cell. *Cancer Research*, 76(14 Supplement), 2016.

Saparbaev M. and Laval J. Excision of hypoxanthine from DNA containing dIMP residues by the Escherichia coli, yeast, rat, and human alkylpurine DNA glycosylases. *Proc Natl Acad Sci USA*, 91(June):5873–5877, 1994. ISSN 0027-8424. doi: 10.1073/pnas.91.13.5873.

Sato N., Maitra A., Fukushima N., van Heek N. T., Matsubayashi H., Iacobuzio-Donahue C. A., Rosty C., and Goggins M. Frequent Hypomethylation of Multiple Genes Overexpressed in Pancreatic Ductal Adenocarcinoma. *Cancer Res.*, 63(14): 4158–4166, jul 2003.

Satou K., Kawai K., Kasai H., Harashima H., and Kamiya H. Mutagenic effects of 8-hydroxy-dGTP in live mammalian cells. *Free Radical Biology and Medicine*, 42(10): 1552–1560, 2007. ISSN 08915849. doi: 10.1016/j.freeradbiomed.2007.02.024.

Satou K., Hori M., Kawai K., Kasai H., Harashima H., and Kamiya H. Involvement of specialized DNA polymerases in mutagenesis by 8-hydroxy-dGTP in human cells. *DNA Repair*, 8(5):637–642, 2009. ISSN 15687864. doi: 10.1016/j.dnarep.2008.12.009.

Saunders C. T., Wong W. S. W., Swamy S., Becq J., Murray L. J., and Cheetham R. K. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts271.

Schlottmann F., Patti M. G., and Shaheen N. J. From Heartburn to Barrett's Esophagus, and Beyond. *World Journal of Surgery*, 41(7):1–7, 2017. ISSN 14322323. doi: 10.1007/s00268-017-3957-z.

Schmeiser H. H., Schoepe K.-B., and Wiessler M. DNA adduct formation of aristolochic acid I and II in vitro and in vivo. *Carcinogenesis*, 9(2):297–3, 1988.

Schmidt B., Rinke M., and Güsten H. Photophysical properties of 5-methylcytosine. *Journal of Photochemistry and Photobiology A: Chemistry*, 49(0):131 – 135, 2006. ISSN 10106030. doi: 10.1016/j.jphotochem.2006.03.020.

Schübeler D. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, 2015. ISSN 0028-0836. doi: 10.1038/nature14192.

Schultz M. D., He Y., Whitaker J. W., Hariharan M., Mukamel E. A., Leung D., Rajagopal N., Nery J. R., Urich M. A., Chen H., Lin S., Lin Y., Jung I., Schmitt A. D., Selvaraj S., Ren B., Sejnowski T. J., Wang W., and Ecker J. R. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, 523(7559):212–216, 2015. ISSN 0028-0836. doi: 10.1038/nature14465.

Schuster-Böckler B. and Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488(7412):504–7, aug 2012. ISSN 1476-4687. doi: 10.1038/nature11273.

Schutsky E. K., Nabel C. S., Davis A. K. F., DeNizio J. E., and Kohli R. M. APOBEC3A efficiently deaminates methylated, but not TET-oxidized, cytosine bases in DNA. *Nucleic Acids Research*, 45(13):7655–7665, may 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx345.

Schwartz R. and Schäffer A. A. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, 2017. ISSN 1471-0056. doi: 10.1038/nrg.2016.170.

Secrier M. and Fitzgerald R. C. Signatures of Mutational Processes and Associated Risk Factors in Esophageal Squamous Cell Carcinoma: A Geographically Independent Stratification Strategy? *Gastroenterology*, 150(5):1080–1083, 2016. ISSN 15280012. doi: 10.1053/j.gastro.2016.03.017.

Sedgwick B., Bates P. A., Paik J., Jacobs S. C., and Lindahl T. Repair of alkylated DNA: Recent advances. *DNA Repair*, 6(4):429–442, 2007. ISSN 15687864. doi: 10.1016/j.dnarep.2006.10.005.

Seplyarskiy V. B., Andrianova M. A., and Bazykin G. A. APOBEC3A/B-induced mutagenesis is responsible for 20% of heritable mutations in the TpCpW context. *Genome Research*, page gr.210336.116, 2016a. ISSN 1088-9051. doi: 10.1101/gr.210336.116.

Seplyarskiy V. B., Soldatov R. A., Popadin K. Y., Antonarakis S. E., Bazykin G. A., and Nikolaev S. I. APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Research*, 26(2):174–182, 2016b. ISSN 15495469. doi: 10.1101/gr.197046.115.

Shen J.-C., Rideout W. M., and Jones P. A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Research*, 22(6):972–976, 1994. ISSN 03051048. doi: 10.1093/nar/22.6.972.

Sheng Z., Oka S., Tsuchimoto D., Abolhassani N., Nomaru H., Sakumi K., Yamada H., and Nakabeppu Y. 8-Oxoguanine causes neurodegeneration during MUTYH-mediated DNA base excision repair. *Journal of Clinical Investigation*, 122(12):4344–4361, 2012. ISSN 00219738. doi: 10.1172/JCI65053.

Shi K., Carpenter M. A., Banerjee S., Shaban N. M., Kurahashi K., Salamango D. J., McCann J. L., Starrett G. J., Duffy J. V., Demir Ö., Amaro R. E., Harki D. A., Harris R. S., and Aihara H. Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. *Nature Structural & Molecular Biology*, 24(2):131–139, 2016a. ISSN 1545-9993. doi: 10.1038/nsmb.3344.

Shi X., Yu Y., Luo M., Zhang Z., Shi S., Feng X., Chen Z., and He J. Loss of 5-hydroxymethylcytosine is an independent unfavorable prognostic factor for esophageal squamous cell carcinoma. *PLoS ONE*, 11(4):1–12, 2016b. ISSN 19326203. doi: 10.1371/journal.pone.0153100.

Shibutani S., Takeshita M., and Grollman A. P. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature*, 349(6308):431–4, jan 1991. ISSN 0028-0836. doi: 10.1038/349431a0.

Shibutani T., Ito S., Toda M., Kanao R., Collins L. B., Shibata M., Urabe M., Koseki H., Masuda Y., Swenberg J. A., Masutani C., Hanaoka F., Iwai S., and Kuraoka I. Guanine-5-carboxylcytosine base pairs mimic mismatches during DNA replication. *Scientific reports*, 4:5220, 2014. ISSN 2045-2322. doi: 10.1038/srep05220.

Shinbrot E., Henninger E. E., Weinhold N., Covington K. R., Göksenin A. Y., Schultz N., Chao H., Doddapaneni H., Muzny D. M., Gibbs R. A., Sander C., Pursell Z. F., and Wheeler D. A. Exonuclease mutations in DNA Polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome research*, pages 1740–1750, 2014. ISSN 1549-5469. doi: 10.1101/gr.174789.114.

Shlien A., Campbell B. B., de Borja R., Alexandrov L. B., Merico D., Wedge D., Van Loo P., Tarpey P. S., Coupland P., Behjati S., Pollett A., Lipman T., Heidari A., Deshmukh S., Avitzur N., Meier B., Gerstung M., Hong Y., Merino D. M., Ramakrishna M., Remke M., Arnold R., Panigrahi G. B., Thakkar N. P., Hodel K. P., Henninger E. E., Göksenin A. Y., Bakry D., Charames G. S., Druker H., Lerner-Ellis J., Mistry M., Dvir R., Grant R., Elhasid R., Farah R., Taylor G. P., Nathan P. C., Alexander S., Ben-Shachar S., Ling S. C., Gallinger S., Constantini S., Dirks P., Huang A., Scherer S. W., Grundy R. G., Durno C., Aronson M., Gartner A., Meyn M. S., Taylor M. D., Pursell Z. F., Pearson C. E., Malkin D., Futreal P. A., Stratton M. R., Bouffet E., Hawkins C., Campbell P. J., and Tabori U. Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nature Genetics*, 47(3):257–262, 2015. ISSN 1061-4036. doi: 10.1038/ng.3202.

Shlyueva D., Stampfel G., and Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews. Genetics*, 15(4):272–86, 2014. ISSN 1471-0064. doi: 10.1038/nrg3682.

Shukla S., Kavak E., Gregory M., Imashimizu M., Shutinoski B., Kashlev M., Oberdoerffer P., Sandberg R., and Oberdoerffer S. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371):74–9, nov 2011. ISSN 1476-4687. doi: 10.1038/nature10442.

Sihvo E. I. T., Salminen J. T., Rantanen T. K., Rämö O. J., Ahotupa M., Färkkilä M., Auvinen M. I., and Salo J. A. Oxidative stress has a role in malignant transformation in Barrett's oesophagus. *International journal of cancer*, 102(6):551–5, 2002. ISSN 0020-7136. doi: 10.1002/ijc.10755.

Silverstein T. D., Johnson R. E., Jain R., Prakash L., Prakash S., and Aggarwal A. K. Structural basis for the suppression of skin cancers by DNA polymerase η. *Nature*, 465(7301):1039–1043, 2010. ISSN 0028-0836. doi: 10.1038/nature09104.

Simpson V. J., Johnson T. E., and Hammen R. F. Caenorhabditis elegans DNA does not contain 5-methylcytosine at any time during development or aging. *Nucleic acids research*, 14(16):6711–9, aug 1986. ISSN 0305-1048.

Siriwardena S. U., Guruge T. A., and Bhagwat A. S. Characterization of the Catalytic Domain of Human APOBEC3B and the Critical Structural Role for a Conserved Methionine. *Journal of Molecular Biology*, 427(19):3042–3055, 2015. ISSN 10898638. doi: 10.1016/j.jmb.2015.08.006.

Sjöblom T., Jones S., Wood L. D., Parsons D. W., Lin J., Barber T. D., Mandelker D., Leary R. J., Ptak J., Silliman N., Szabo S., Buckhaults P., Farrell C., Meeh P., Markowitz S. D., Willis J., Dawson D., Willson J. K. V., Gazdar A. F., Hartigan J., Wu L., Liu C., Parmigiani G., Park B. H., Bachman K. E., Papadopoulos N., Vogelstein B., Kinzler K. W., and Velculescu V. E. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797):268–74, oct 2006. ISSN 1095-9203. doi: 10.1126/science.1133427.

Skvortsova K., Zotenko E., Luu P.-L., Gould C. M., Nair S. S., Clark S. J., and Stirzaker C. Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. *Epigenetics & Chromatin*, 10(1):16, 2017. ISSN 1756-8935. doi: 10.1186/s13072-017-0123-7.

Smerdon M. J. and Conconi A. Modulation of DNA damage and DNA repair in chromatin. *Prog Nucleic Acid Res Mol Biol*, 62:227–255, 1999. ISSN 0079-6603 (Print).

Smith D. J. and Whitehouse I. Intrinsic coupling of lagging-strand synthesis to chromatin assembly. *Nature*, 483(7390):434–8, 2012. ISSN 1476-4687. doi: 10.1038/nature10895.

Smith H. C., Bennett R. P., Kizilyer A., McDougall W. M., and Prohaska K. M. Functions and regulation of the APOBEC family of proteins. *Seminars in Cell and Developmental Biology*, 23(3):258–268, 2012. ISSN 10849521. doi: 10.1016/j.semcdb.2011.10.004.

Snedeker J., Wooten M., and Chen X. The Inherent Asymmetry of DNA Replication. *Annual review of cell and developmental biology*, 33(1):1–28, aug 2017. ISSN 1530-8995. doi: 10.1146/annurev-cellbio-100616-060447.

Song C. X., Szulwach K. E., Dai Q., Fu Y., Mao S. Q., Lin L., Street C., Li Y., Poidevin M., Wu H., Gao J., Liu P., Li L., Xu G. L., Jin P., and He C. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*, 153(3):678–691, 2013. ISSN 00928674. doi: 10.1016/j.cell.2013.04.001.

Song Q., Cannistraro V. J., and Taylor J. S. Rotational position of a 5-methylcytosine-containing cyclobutane pyrimidine dimer in a nucleosome greatly affects its deamination rate. *Journal of Biological Chemistry*, 286(8):6329–6335, 2011. ISSN 00219258. doi: 10.1074/jbc.M110.183178.

Song Q., Sherrer S. M., Suo Z., and Taylor J. S. Preparation of site-specific T = mCG cis-syn cyclobutane dimer-containing template and its error-free bypass by yeast and human polymerase η. *Journal of Biological Chemistry*, 287(11):8021–8028, 2012. ISSN 00219258. doi: 10.1074/jbc.M111.333591.

Song Q., Cannistraro V. J., and Taylor J. S. Synergistic modulation of cyclobutane pyrimidine dimer photoproduct formation and deamination at a TmCG site over a full helical DNA turn in a nucleosome core particle. *Nucleic Acids Research*, 42(21): 13122–13133, 2014. ISSN 13624962. doi: 10.1093/nar/gku1049.

Souza R. F. The role of acid and bile reflux in oesophagitis and Barrett's metaplasia. *Biochemical Society transactions*, 38(2):348–52, 2010. ISSN 1470-8752. doi: 10.1042/BST0380348.

Spruijt C. G., Gnerlich F., Smits A. H., Pfaffeneder T., Jansen P. W. T. C., Bauer C., Münzel M., Wagner M., Müller M., Khan F., Eberl H. C., Mensinga A., Brinkman A. B., Lephikov K., Müller U., Walter J., Boelens R., van Ingen H., Leonhardt H., Carell T., and Vermeulen M. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell*, 152(5):1146–59, feb 2013. ISSN 1097-4172. doi: 10.1016/j.cell.2013.02.004.

Srivastava M. and Raghavan S. DNA Double-Strand Break Repair Inhibitors as Cancer Therapeutics. *Chemistry & Biology*, 22(1):17–29, 2015. ISSN 10745521. doi: 10.1016/j.chembiol.2014.11.013.

Stachler M. D., Taylor-Weiner A., Peng S., McKenna A., Agoston A. T., Odze R. D., Davison J. M., Nason K. S., Loda M., Leshchiner I., Stewart C., Stojanov P., Seepo S., Lawrence M. S., Ferrer-Torres D., Lin J., Chang A. C., Gabriel S. B., Lander E. S., Beer D. G., Getz G., Carter S. L., and Bass A. J. Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nature genetics*, 47(9):1047–55, 2015. ISSN 1546-1718. doi: 10.1038/ng.3343.

Stamatoyannopoulos J. A., Adzhubei I., Thurman R. E., Kryukov G. V., Mirkin S. M., and Sunyaev S. R. Human mutation rate associated with DNA replication timing. *Nature genetics*, 41(4):393–395, 2009. ISSN 1061-4036. doi: 10.1038/ng.363.

Starrett G. J., Luengas E. M., McCann J. L., Ebrahimi D., Temiz N. A., Love R. P., Feng Y., Adolph M. B., Chelico L., Law E. K., Carpenter M. A., and Harris R. S. The DNA cytosine deaminase APOBEC3H haplotype I likely contributes to breast and lung

cancer mutagenesis. *Nature communications*, 7(May):12918, 2016. ISSN 2041-1723. doi: 10.1038/ncomms12918.

Stephens P. J., Tarpey P. S., Davies H., Van Loo P., Greenman C., Wedge D. C., Zainal S. N., Martin S., Varela I., Bignell G. R., Yates L. R., Papaemmanuil E., Beare D., Butler A., Cheverton A., Gamble J., Hinton J., Jia M., Jayakumar A., Jones D., Latimer C., Lau K. W., McLaren S., McBride D. J., Menzies A., Mudie L., Raine K., Rad R., Spencer Chapman M., Teague J., Easton D., Langerød A., OSBREAC, Lee M. T. M., Shen C.-Y., Tee B. T. K., Huimin B. W., Broeks A., Vargas A. C., Turashvili G., Martens J., Fatima A., Miron P., Chin S.-F., Thomas G., Boyault S., Mariani O., Lakhani S. R., van de Vijver M., van 't Veer L., Foekens J., Desmedt C., Sotiriou C., Tutt A., Caldas C., Reis-Filho J. S., Aparicio S. A. J. R., Salomon A. V., Børresen-Dale A.-L., Richardson A., Campbell P. J., Futreal P. A., Stratton M. R., Karesen R., Schlichting E., Naume B., Sauer T., and Ottestad L. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400–404, 2012. ISSN 0028-0836. doi: 10.1038/nature11017.

Stillman B. DNA Polymerases at the Replication Fork in Eukaryotes. *Molecular Cell*, 30 (3):259–260, 2008. ISSN 10972765. doi: 10.1016/j.molcel.2008.04.011.

Stirzaker C., Taberlay P. C., Statham A. L., and Clark S. J. Mining cancer methylomes: Prospects and challenges. *Trends in Genetics*, 30(2):75–84, 2014. ISSN 01689525. doi: 10.1016/j.tig.2013.11.004.

Stith C. M., Sterling J., Resnick M. A., Gordenin D. A., and Burgers P. M. Flexibility of eukaryotic Okazaki fragment maturation through regulated strand displacement synthesis. *Journal of Biological Chemistry*, 283(49):34129–34140, 2008. ISSN 00219258. doi: 10.1074/jbc.M806668200.

Stratton M. R., Campbell P. J., and Andrew F P. The cancer genome. *Nature*, 458(7239): 719–724, 2009. ISSN 0028-0836. doi: 10.1038/nature07943.

Stricker S. H., Köferle A., and Beck S. From profiles to function in epigenomics. *Nature Reviews Genetics*, 18(1):51–66, 2016. ISSN 1471-0056. doi: 10.1038/nrg.2016.138.

Stroud H., Feng S., Morey Kinney S., Pradhan S., and Jacobsen S. E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biology*, 12(6):R54, 2011. ISSN 1465-6906. doi: 10.1186/gb-2011-12-6-r54.

Struhl K. and Segal E. Determinants of nucleosome positioning. *Nature structural & molecular biology*, 20(3):267–73, 2013. ISSN 1545-9985. doi: 10.1038/nsmb.2506.

Su Z., Han L., and Zhao Z. Conservation and divergence of DNA methylation in eukaryotes: New insights from single base-resolution DNA methylomes. *Epigenetics*, 6(2):134–140, 2011. ISSN 15592308. doi: 10.4161/epi.6.2.13875.

Supek F. and Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*, 521(7550):81–84, may 2015.

Supek F., Lehner B., Hajkova P., and Warnecke T. Hydroxymethylated cytosines are associated with elevated C to G transversion rates. *PLoS genetics*, 10(9):e1004585, sep 2014a. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004585.

Supek F., Miñana B., Valcárcel J., Gabaldón T., and Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6):1324–35, mar 2014b. ISSN 1097-4172. doi: 10.1016/j.cell.2014.01.051.

Suzuki M. M. and Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews. Genetics*, 9(6):465–76, 2008. ISSN 1471-0064. doi: 10.1038/nrg2341.

Suzuki T. and Kamiya H. Mutations induced by 8-hydroxyguanine (8-oxo-7,8-dihydroguanine), a representative oxidized base, in mammalian cells. *Genes and environment*, 39:2, 2017. ISSN 1880-7046. doi: 10.1186/s41021-016-0051-y.

Suzuki T., Harashima H., and Kamiya H. Effects of base excision repair proteins on mutagenesis by 8-oxo-7,8-dihydroguanine (8-hydroxyguanine) paired with cytosine and adenine. *DNA repair*, 9(5):542–50, may 2010. ISSN 1568-7856. doi: 10.1016/j.dnarep.2010.02.004.

Svedruzić Z. M., Wang C., Kosmoski J. V., and Smerdon M. J. Accommodation and repair of a UV photoproduct in DNA at different rotational settings on the nucleosome surface. *The Journal of biological chemistry*, 280(48):40051–7, dec 2005. ISSN 0021-9258. doi: 10.1074/jbc.M509478200.

Swanton C., McGranahan N., Starrett G. J., and Harris R. S. APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. *Cancer discovery*, 5(7): 704–712, 2015. ISSN 21598290. doi: 10.1158/2159-8290.CD-15-0344.

Szwagierczak A., Bultmann S., Schmidt C. S., Spada F., and Leonhardt H. Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic acids research*, 38(19):e181, oct 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq684.

Tabin C. J., Bradley S. M., Bargmann C. I., Weinberg R. A., Papageorge A. G., Scolnick E. M., Dhar R., Lowy D. R., and Chang E. H. Mechanism of activation of a human oncogene. *Nature*, 300(5888):143–9, nov 1982. ISSN 0028-0836. doi: 10.1038/300143a0.

Tahiliani M., Koh K. P., Shen Y., Pastor W. A., Bandukwala H., Brudno Y., Agarwal S., Iyer L. M., Liu D. R., Aravind L., and Rao A. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324(5929): 930–5, may 2009. ISSN 1095-9203. doi: 10.1126/science.1170116.

Takai H., Masuda K., Sato T., Sakaguchi Y., Suzuki T., Suzuki T., Koyama-Nasu R., Nasu-Nishimura Y., Katou Y., Ogawa H., Morishita Y., Kozuka-Hata H., Oyama M., Todo T., Ino Y., Mukasa A., Saito N., Toyoshima C., Shirahige K., and Akiyama T. 5-Hydroxymethylcytosine Plays a Critical Role in Glioblastomagenesis by Recruiting the CHTOP-Methylosome Complex. *Cell Reports*, 9(1):48–60, oct 2014. ISSN 22111247. doi: 10.1016/j.celrep.2014.08.071.

Taylor J., Tyekucheva S., Zody M., Chiaromonte F., and Makova K. D. Strong and weak male mutation bias at different sites in the primate genomes: Insights from the human-chimpanzee comparison. *Molecular Biology and Evolution*, 23(3):565–573, 2006. ISSN 07374038. doi: 10.1093/molbev/msj060.

TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8, oct 2008. ISSN 1476-4687. doi: 10.1038/nature07385.

Teissandier A. and Bourc'his D. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. *The EMBO Journal*, 36(11):e201796812, 2017. ISSN 0261-4189. doi: 10.15252/embj.201796812.

Teperek-Tkacz M., Pasque V., Gentsch G., and Ferguson-Smith A. C. Epigenetic reprogramming: Is deamination key to active DNA demethylation? *Reproduction*, 142(5):621–632, 2011. ISSN 14701626. doi: 10.1530/REP-11-0148.

Terato H., Masaoka A., Asagoshi K., Honsho A., Ohyama Y., Suzuki T., Yamada M., Makino K., Yamamoto K., and Ide H. Novel repair activities of AlkA (3-methyladenine DNA glycosylase II) and endonuclease VIII for xanthine and oxanine, guanine lesions induced by nitric oxide and nitrous acid. *Nucleic Acids Res.*, 30(22):4975–4984, 2002. ISSN 1362-4962.

Thoma F. Repair of UV lesions in nucleosomes - Intrinsic properties and remodeling. *DNA Repair*, 4(8):855–869, 2005. ISSN 15687864. doi: 10.1016/j.dnarep.2005.04.005.

Thomson J. P. and Meehan R. R. The application of genome-wide 5-hydroxymethylcytosine studies in cancer research. *Epigenomics*, 9(1):77–91, jan 2017. ISSN 1750-192X. doi: 10.2217/epi-2016-0122.

Tomasetti C., Vogelstein B., and Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the National Academy of Sciences*, 110(6):1999–2004, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1221068110.

Tomasetti C. and Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science (New York, N.Y.)*, 347(6217):78–81, 2015. ISSN 1095-9203. doi: 10.1126/science.1260825.

Tomkova M., McClellan M., Kriaucionis S., and Schuster-Boeckler B. 5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA. *eLife*, 5(MAY2016):1–23, 2016. ISSN 2050084X. doi: 10.7554/eLife.17082.

Tommasi S. and Pfeifer G. P. Sunlight Induces Pyrimidine Dimers Preferentially at 5-Methylcytosine Bases. *Cancer Research*, 57:4727–4730, 1997.

Topal M. and Baker M. DNA precursor pool: a significant target for N-methyl-N-nitrosourea in C3H/10T1/2 clone 8 cells. *Proceedings of the National Academy of Sciences of the United States of America*, 79(7):2211–2215, 1982. ISSN 00278424. doi: 10.1073/pnas.79.7.2211.

Toyota M., Ahuja N., Ohe-Toyota M., Herman J. G., Baylin S. B., and Issa J.-P. J. CpG island methylator phenotype in colorectal cancer. *Medical Sciences*, 96(July):8681–8686, 1999. ISSN 10093079. doi: 10.11569/wcjd.v24.i4.558.

Tu Y., Wang Z., Wang X., Yang H., Zhang P., Johnson M., Liu N., Liu H., Jin W., Zhang Y., and Cui D. Birth of MTH1 as a therapeutic target for glioblastoma: MTH1 is indispensable for gliomatumorigenesis. *American Journal of Translational Research*, 8 (6):2803–2811, 2016. ISSN 19438141.

Tubbs A. and Nussenzweig A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell*, 168(4):644–656, 2017. ISSN 10974172. doi: 10.1016/j.cell.2017.01.002.

Turcan S., Rohle D., Goenka A., Walsh L. A., Fang F., Yilmaz E., Campos C., Fabius A. W. M., Lu C., Ward P. S., Thompson C. B., Kaufman A., Guryanova O., Levine R., Heguy A., Viale A., Morris L. G. T., Huse J. T., Mellinghoff I. K., and Chan T. A. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*, 483(7390):479–83, mar 2012. ISSN 1476-4687. doi: 10.1038/nature10866.

Urban J. M., Foulk M. S., Casella C., and Gerbi S. A. The hunt for origins of DNA replication in multicellular eukaryotes. *F1000prime reports*, 7(March 2015):30, 2015. ISSN 2051-7599. doi: 10.12703/P7-30.

Van Speybroeck L., De Waele D., and Van de Vijver G. Theories in early embryology: close connections between epigenesis, preformationism, and self-organization. *Annals of the New York Academy of Sciences*, 981:7–49, dec 2002. ISSN 0077-8923.

Vandiver A. R., Irizarry R. A., Hansen K. D., Garza L. A., Runarsson A., Li X., Chien A. L., Wang T. S., Leung S. G., Kang S., and Feinberg A. P. Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. *Genome Biology*, 16(1):80, 2015. ISSN 1465-6914. doi: 10.1186/s13059-015-0644-y.

Varley K. E., Gertz J., Bowling K. M., Parker S. L., Reddy T. E., Pauli-Behn F., Cross M. K., Williams B. A., Stamatoyannopoulos J. A., Crawford G. E., Absher D. M., Wold B. J., and Myers R. M. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome research*, 23(3):555–67, mar 2013. ISSN 1549-5469. doi: 10.1101/gr.147942.112.

Venkatesan S., Swanton C., Taylor B. S., and Costello J. F. Treatment-Induced Mutagenesis and Selective Pressures Sculpt Cancer Evolution. *Cold Spring Harbor perspectives in medicine*, 7(8), aug 2017. ISSN 2157-1422. doi: 10.1101/cshperspect.a026617.

Venn O., Turner I., Mathieson I., de Groot N., Bontrop R., and McVean G. Strong male bias drives germline mutation in chimpanzees. *Science*, 344(6189):1272–1275, 2014. ISSN 0036-8075. doi: 10.1126/science.344.6189.1272.

Venolia L. and Gartler S. M. Comparison of transformation efficiency of human active and inactive X-chromosomal DNA. *Nature*, 302(5903):82–3, mar 1983. ISSN 0028-0836.

Versteeg R. Cancer: Tumours outside the mutation box. *Nature*, 506(7489):438–9, 2014. ISSN 1476-4687. doi: 10.1038/nature13061.

Vieira V. C. and Soares M. A. Review Article The Role of Cytidine Deaminases on Innate Immune Responses against Human Viral Infections. *BioMed Research International*, 2013, 2013.

Vineis P., Schatzkin A., and Potter J. D. Models of carcinogenesis: An overview. *Carcinogenesis*, 31(10):1703–1709, 2010. ISSN 14602180. doi: 10.1093/carcin/bgq087.

Visnes T., Doseth B., Pettersen H. S., Hagen L., Sousa M. M., Akbari M., Otterlei M., Kavli B., Slupphaug G., and Krokan H. E. Uracil in DNA and its processing by different DNA glycosylases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1517):563–568, 2009. ISSN 0962-8436. doi: 10.1098/rstb.2008.0186.

Visvader J. E. Cells of origin in cancer. *Nature*, 469(7330):314–322, 2011. ISSN 0028-0836. doi: 10.1038/nature09781.

Vogelstein B., Papadopoulos N., Velculescu V. E., Zhou S., Diaz L. A., and Kinzler K. W. Cancer genome landscapes. *Science*, 339(6127):1546–58, 2013. ISSN 1095-9203. doi: 10.1126/science.1235122.

Vongchampa V., Dong M., Gingipalli L., and Dedon P. Stability of 2'-deoxyxanthosine in DNA. *Nucleic acids research*, 31(3):1045–51, feb 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg177.

Voong L. N., Xi L., Sebeson A. C., Xiong B., Wang J. P., and Wang X. Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell*, 167(6):1555–1570.e15, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.10.049.

Vu B., Cannistraro V. J., Sun L., and Taylor J. S. DNA synthesis past a 5-methylc-containing cis-syn-cyclobutane pyrimidine dimer by yeast Pol ?? is highly nonmutagenic. *Biochemistry*, 45(30):9327–9335, 2006. ISSN 00062960. doi: 10.1021/bi0602009.

Waddington C. H. The strategy of the genes., 1957.

Wade P. A. and Wolffe A. P. ReCoGnizing methylated DNA. *Nature structural biology*, 8 (7):575–7, 2001. ISSN 1072-8368. doi: 10.1038/89593.

Wagner E. J. and Carpenter P. B. Understanding the language of Lys36 methylation at histone H3. *Nature Reviews Molecular Cell Biology*, 13(2):115–126, 2012. ISSN 1471-0072. doi: 10.1038/nrm3274.

Waisertreiger I. S.-R., Liston V. G., Menezes M. R., Kim H.-M., Lobachev K. S., Stepchenkova E. I., Tahirov T. H., Rogozin I. B., and Pavlov Y. I. Modulation of mutagenesis in eukaryotes by DNA replication fork dynamics and quality of nucleotide pools. *Environmental and molecular mutagenesis*, 53(9):699–724, dec 2012. ISSN 1098-2280. doi: 10.1002/em.21735.

Wall J. D., Tang L. F., Zerbe B., Kvale M. N., Kwok P. Y., Schaefer C., and Risch N. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Research*, 24(11):1734–1739, 2014. ISSN 15495469. doi: 10.1101/gr.168393.113.

Walsh E. and Eckert K. A. Eukaryotic Replicative DNA Polymerases. In *Nucleic Acid Polymerases*, volume 30, pages 17–41. Springer, Berlin, Heidelberg, 2014. ISBN 978-3-642-39795-0. doi: 10.1007/978-3-642-39796-7.

Wang D., Huang J. H., Zeng Q. H., Gu C., Ding S., Lu J. Y., Chen J., and Yang S. B. Increased 5-hydroxymethylcytosine and ten-eleven translocation protein expression in ultraviolet B-irradiated HaCaT cells. *Chinese Medical Journal*, 130(5):594–599, 2017a. ISSN 03666999. doi: 10.4103/0366-6999.200539.

Wang H. T., Weng M. W., Chen W. C., Yobin M., Pan J., Chung F. L., Wu X. R., Rom W., and Tang M. S. Effect of CpG methylation at different sequence context on acrolein- and BPDE-DNA binding and mutagenesis. *Carcinogenesis*, 34(1):220–227, 2013. ISSN 01433334. doi: 10.1093/carcin/bgs323.

Wang K., Yuen S. T., Xu J., Lee S. P., Yan H. H. N., Shi S. T., Siu H. C., Deng S., Chu K. M., Law S., Chan K. H., Chan A. S. Y., Tsui W. Y., Ho S. L., Chan A. K. W., Man J. L. K., Foglizzo V., Ng M. K., Chan A. S., Ching Y. P., Cheng G. H. W., Xie T., Fernandez J., Li V. S. W., Clevers H., Rejto P. A., Mao M., and Leung S. Y. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature genetics*, 46(6):573–82, 2014. ISSN 1546-1718. doi: 10.1038/ng.2983.

Wang K. and Taylor J.-S. A. Modulation of cyclobutane thymine photodimer formation in T11-tracts in rotationally phased nucleosome core particles and DNA minicircles.

Nucleic Acids Research, 45(12):7031–7041, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx427.

Wang L., Zhou Y., Xu L., Xiao R., Lu X., Chen L., Chong J., Li H., He C., Fu X.-D., and Wang D. Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature*, 2015. ISSN 0028-0836. doi: 10.1038/nature14482.

Wang M., Zhou S., Chen Q., Wang L., Liang Z., and Wang J. Understanding the molecular mechanism for the differential inhibitory activities of compounds against MTH1. *Scientific Reports*, 7(August 2016):40557, 2017b. ISSN 2045-2322. doi: 10.1038/srep40557.

Wang Y., Woodgate R., McManus T. P., Mead S., McCormick J. J., and Maher V. M. Evidence that in xeroderma pigmentosum variant cells, which lack DNA polymerase η, DNA polymerase ι causes the very high frequency and unique spectrum of UV-induced mutations. *Cancer Research*, 67(7):3018–3026, 2007. ISSN 00085472. doi: 10.1158/0008-5472.CAN-06-3073.

Wang Z., Li Z., Ye Y., Xie L., and Li W. Oxidative Stress and Liver Cancer: Etiology and Therapeutic Targets. *Oxidative medicine and cellular longevity*, 2016. ISSN 1942-0994. doi: 10.1155/2016/7891574.

Waters L. S. and Walker G. C. The critical mutagenic translesion DNA polymerase Rev1 is highly expressed during G(2)/M phase rather than S phase. *Proceedings of the National Academy of Sciences of the United States of America*, 103(24):8971–8976, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0510167103.

Watson J. D. and Crick F. H. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–7, may 1953. ISSN 0028-0836. doi: 10.1038/ng0403-431.

Watson J. D. and Crick F. H. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 248(5451):765, apr 1974. ISSN 0028-0836. doi: 10.1038/171737a0.

Wellcome Trust Sanger Institute. COSMIC: Signatures of Mutational Processes in Human Cancer, 2017.

Wen L., Li X., Yan L., Tan Y., Li R., Zhao Y., Wang Y., Xie J., Zhang Y., Song C., Yu M., Liu X., Zhu P., Li X., Hou Y., Guo H., Wu X., He C., Li R., Tang F., and Qiao J. Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome biology*, 15(3):R49, mar 2014. ISSN 1465-6914. doi: 10.1186/gb-2014-15-3-r49.

Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), 2016.

Wijesinghe P. and Bhagwat A. S. Efficient deamination of 5-methylcytosines in DNA by human APOBEC3A, but not by AID or APOBEC3G. *Nucleic Acids Research*, 40(18): 9206–9217, 2012. ISSN 03051048. doi: 10.1093/nar/gks685.

Williams J. S., Lujan S. A., and Kunkel T. A. Processing ribonucleotides incorporated during eukaryotic DNA replication. *Nature Reviews Molecular Cell Biology*, 17(6): 350–363, 2016. ISSN 1471-0080. doi: 10.1038/nrm.2016.37.

Williams K., Christensen J., Pedersen M. T., Johansen J. V., Cloos P. A. C., Rappsilber J., and Helin K. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, 473(7347):343–8, may 2011. ISSN 1476-4687. doi: 10.1038/nature10066.

Williams L. N., Marjavaara L., Knowels G. M., Schultz E. M., Fox E. J., Chabes A., and Herr A. J. dNTP pool levels modulate mutator phenotypes of error-prone DNA polymerase ϵ variants. *Proceedings of the National Academy of Sciences of the United States of America*, 112(19):E2457–66, 2015. ISSN 1091-6490. doi: 10.1073/pnas.1422948112.

Wolf S. F., Jolly D. J., Lunnen K. D., Friedmann T., and Migeon B. R. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. *Proceedings of the National Academy of Sciences of the United States of America*, 81(9):2806–10, 1984. ISSN 0027-8424. doi: 10.1073/pnas.81.9.2806.

Wongtrakoongate P. Epigenetic therapy of cancer stem and progenitor cells by targeting DNA methylation machineries. *World J Stem Cells January World J Stem Cells*, 26(71): 137–148, 2015. ISSN 1948-0210. doi: 10.4252/wjsc.v7.i1.137.

Wu H. and Zhang Y. Tet1 and 5-hydroxymethylation: A genome-wide view in mouse embryonic stem cells. *Cell Cycle*, 10(15):2428–2436, aug 2011. ISSN 1538-4101. doi: 10.4161/cc.10.15.16930.

Wu H., Wu X., Shen L., and Zhang Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nature Biotechnology*, 32(12):1231–1240, 2014. ISSN 1087-0156. doi: 10.1038/nbt.3073.

Wu S., Powers S., Zhu W., and Hannun Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature*, 529(7584):43–47, 2015. ISSN 0028-0836. doi: 10.1038/nature16166.

Wu S. C. and Zhang Y. Active DNA demethylation: many roads lead to Rome. *Nature Reviews Molecular Cell Biology*, 11(9):607–620, 2010. ISSN 1471-0072. doi: 10.1038/nrm2950.

Wu T. P., Wang T., Seetin M. G., Lai Y., Zhu S., Lin K., Liu Y., Byrum S. D., Mackintosh S. G., Zhong M., Tackett A., Wang G., Hon L. S., Fang G., Swenberg J. A., and Xiao A. Z. DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature*, 532(7599):329–333, 2016. ISSN 0028-0836. doi: 10.1038/nature17640.

Wu X. and Zhang Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nature Reviews Genetics*, 2017. ISSN 1471-0056. doi: 10.1038/nrg.2017.33.

Xie W., Schultz M. D., Lister R., Hou Z., Rajagopal N., Ray P., Whitaker J. W., Tian S., Hawkins R. D., Leung D., Yang H., Wang T., Lee A. Y., Swanson S. A., Zhang J., Zhu Y., Kim A., Nery J. R., Urich M. A., Kuan S., Yen C. A., Klugman S., Yu P., Suknuntha K., Propson N. E., Chen H., Edsall L. E., Wagner U., Li Y., Ye Z., Kulkarni A., Xuan Z., Chung W. Y., Chi N. C., Antosiewicz-Bourget J. E., Slukvin I., Stewart R., Zhang M. Q., Wang W., Thomson J. A., Ecker J. R., and Ren B. Epigenomic analysis of multilineage

differentiation of human embryonic stem cells. *Cell*, 153(5):1134–1148, 2013. ISSN 00928674. doi: 10.1016/j.cell.2013.04.022.

Xing X. W., Liu Y. L., Vargas M., Wang Y., Feng Y. Q., Zhou X., and Yuan B. F. Mutagenic and Cytotoxic Properties of Oxidation Products of 5-Methylcytosine Revealed by Next-Generation Sequencing. *PLoS ONE*, 8(9), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0072993.

Xu W., Yang H., Liu Y., Yang Y., Wang P. P., Kim S.-H. H., Ito S., Yang C., Wang P. P., Xiao M.-T. T., Liu L.-x. X., Jiang W.-q. Q., Liu J., Zhang J.-y. Y., Wang B., Frye S., Zhang Y., Xu Y.-h. H., Lei Q.-y. Y., Guan K.-L. L., Zhao S.-m. M., and Xiong Y. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. *Cancer cell*, 19(1):17–30, jan 2011. ISSN 1878-3686. doi: 10.1016/j.ccr.2010.12.014.

Xu Z., Taylor J. A., Leung Y.-K. K., Ho S.-M. M., and Niu L. oxBS-MLE: An efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated DNA. *Bioinformatics*, 32(August):btw527, 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btw527.

Yamagiwa K. and Ichikawa K. Experimental Study of the Pathogenesis of Carcinoma. *The Journal of Cancer Research*, 3(1), 1918.

Yan H., Parsons D., and Jin G. IDH1 and IDH2 mutations in gliomas. *The New England journal of medicine*, 2009.

Yang H., Liu Y., Bai F., Zhang J.-Y., Ma S.-H., Liu J., Xu Z.-D., Zhu H.-G., Ling Z.-Q., Ye D., Guan K.-L., and Xiong Y. Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene*, 32(5):663–9, jan 2013a. ISSN 1476-5594. doi: 10.1038/onc.2012.67.

Yang Q., Wu K., Ji M., Jin W., He N., Shi B., and Hou P. Decreased 5-hydroxymethylcytosine (5-hmC) is an independent poor prognostic factor in gastric cancer patients. *Journal of biomedical nanotechnology*, 9(9):1607–16, sep 2013b. ISSN 1550-7033.

Yang X., Lay F., Han H., and Jones P. A. Targeting DNA methylation for epigenetic therapy. *Trends in pharmacological sciences*, 31(11):536–46, nov 2010. ISSN 1873-3735. doi: 10.1016/j.tips.2010.08.001.

Yang X., Han H., DeCarvalho D. D., Lay F. D., Jones P. A., and Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell*, 26(4):577–590, 2014. ISSN 18783686. doi: 10.1016/j.ccr.2014.07.028.

Yazdi P. G., Pedersen B. A., Taylor J. F., Khattab O. S., Chen Y.-H., Chen Y., Jacobsen S. E., and Wang P. H. Nucleosome Organization in Human Embryonic Stem Cells. *Plos One*, 10(8):e0136314, 2015a. ISSN 1932-6203. doi: 10.1371/journal.pone.0136314.

Yazdi P. G., Pedersen B. A., Taylor J. F., Khattab O. S., Chen Y.-H., Chen Y., Jacobsen S. E., and Wang P. H. Increasing Nucleosome Occupancy Is Correlated with an Increasing Mutation Rate so Long as DNA Repair Machinery Is Intact. *Plos One*, 10(8):e0136574, 2015b. ISSN 1932-6203. doi: 10.1371/journal.pone.0136574.

Yearim A., Gelfman S., Shayevitch R., Melcer S., Glaich O., Mallm J.-P., Nissim-Rafinia M., Cohen A.-H. S., Rippe K., Meshorer E., and Ast G. HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. *Cell reports*, 10(7):1122–34, feb 2015. ISSN 2211-1247. doi: 10.1016/j.celrep.2015.01.038.

Yi C., Chen B., Qi B., Zhang W., Jia G., Zhang L., Li C. J., Dinner A. R., Yang C.-G., and He C. Duplex interrogation by a direct DNA repair protein in search of base damage. *Nature Structural & Molecular Biology*, 19(7):671–676, 2012. ISSN 1545-9993. doi: 10.1038/nsmb.2320.

Yoder J. A., Walsh C. P., and Bestor T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 13(8):335–340, 1997. ISSN 01689525. doi: 10.1016/S0168-9525(97)01181-5.

You J. S. and Jones P. A. Cancer Genetics and Epigenetics: Two Sides of the Same Coin? *Cancer Cell*, 22(1):9–20, 2012. ISSN 15356108. doi: 10.1016/j.ccr.2012.06.008.

Yu M., Hon G. C., Szulwach K. E., Song C. X., Zhang L., Kim A., Li X., Dai Q., Shen Y., Park B., Min J. H., Jin P., Ren B., and He C. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, 149(6):1368–1380, 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.04.027.

Yu S.-L., Johnson R. E., Prakash S., and Prakash L. Requirement of DNA Polymerase for Error-Free Bypass of UV-Induced CC and TC Photoproducts. *Molecular and Cellular Biology*, 21(1):185–188, 2001. ISSN 0270-7306. doi: 10.1128/MCB.21.1.185-188.2001.

Zauri M., Berridge G., Thézénas M.-L., Pugh K. M., Goldin R., Kessler B. M., and Kriaucionis S. CDA directs metabolism of epigenetic nucleosides revealing a therapeutic window in cancer. *Nature*, 524(7563):114–118, 2015. ISSN 0028-0836. doi: 10.1038/nature14948.

Zhang L., Lu X., Lu J., Liang H., Dai Q., Xu G.-L., Luo C., Jiang H., and He C. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nature chemical biology*, 8(4):328–30, apr 2012. ISSN 1552-4469. doi: 10.1038/nchembio.914.

Zhang Q. M., Yonekura S. I., Takao M., Yasui A., Sugiyama H., and Yonei S. DNA glycosylase activities for thymine residues oxidized in the methyl group are functions of the hNEIL1 and hNTH1 enzymes in human cells. *DNA Repair*, 4(1):71–79, 2005. ISSN 15687864. doi: 10.1016/j.dnarep.2004.08.002.

Zhang Y., Wu K., Shao Y., Sui F., Yang Q., Shi B., Hou P., and Ji M. Decreased 5-Hydroxymethylcytosine (5-hmC) predicts poor prognosis in early-stage laryngeal squamous cell carcinoma. *American Journal of Cancer Research*, 6(5):1089–1098, 2016. ISSN 21566976.

Zhang Y., Liu T., Meyer C. A., Eeckhoute J., Johnson D. S., Bernstein B. E., Nussbaum C., Myers R. M., Brown M., Li W., and Liu X. S. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008. ISSN 1465-6906. doi: 10.1186/gb-2008-9-9-r137.

Zhao B., Wang J., Geacintov N. E., and Wang Z. Polη, Polζ and Rev1 together are required for G to T transversion mutations induced by the (+)- and (-)-trans-anti-BPDE-N2-dG DNA adducts in yeast cells. *Nucleic Acids Research*, 34(2):417–425, 2006. ISSN 03051048. doi: 10.1093/nar/gkj446.

Zhao H., Thienpont B., Yesilyurt B. T., Moisse M., Reumers J., Coenegrachts L., Sagaert X., Schrauwen S., Smeets D., Matthijs G., Aerts S., Cools J., Metcalf A., Spurdle A., ANECS, Amant F., and Lambrechts D. Mismatch repair deficiency endows tumors with a unique mutation signature and sensitivity to DNA double-strand breaks. *eLife*, 3: e02725, 2014a.

Zhao L. and Todd Washington M. Translesion synthesis: Insights into the selection and switching of DNA polymerases. *Genes*, 8(1):1–25, 2017. ISSN 20734425. doi: 10.3390/genes8010024.

Zhao Y., Yu H., and Hu W. The regulation of MDM2 oncogene and its impact on human cancers. *Acta biochimica et biophysica Sinica*, 46(January):180–189, 2014b. doi: 10.1093/abbs/gmt147.Advance.

Zhen A., Du J., Zhou X., Xiong Y., and Yu X. F. Reduced APOBEC3H variant anti-viral activities are associated with altered RNA binding activities. *PLoS ONE*, 7(7):1–10, 2012. ISSN 19326203. doi: 10.1371/journal.pone.0038771.

Zheng C. L., Wang N. J., Chung J., Moslehi H., Sanborn J. Z., Hur J. S., Collisson E. A., Vemula S. S., Naujokas A., Chiotti K. E., Cheng J. B., Fassihi H., Blumberg A. J., Bailey C. V., Fudem G. M., Mihm F. G., Cunningham B. B., Neuhaus I. M., Liao W., Oh D. H., Cleaver J. E., LeBoit P. E., Costello J. F., Lehmann A. R., Gray J. W., Spellman P. T., Arron S. T., Huh N., Purdom E., and Cho R. J. Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. *Cell Reports*, 9(4):1228–1234, nov 2014.

Zheng L. and Shen B. Okazaki fragment maturation: Nucleases take centre stage. *Journal of Molecular Cell Biology*, 3(1):23–30, 2011. ISSN 16742788. doi: 10.1093/jmcb/mjq048.

Zhou V. W., Goren A., and Bernstein B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews. Genetics*, 12(1):7–18, 2011. ISSN 1471-0064. doi: 10.1038/nrg2905.

Ziller M. J., Gu H., Müller F., Donaghey J., Tsai L. T.-Y., Kohlbacher O., De Jager P. L., Rosen E. D., Bennett D. A., Bernstein B. E., Gnirke A., and Meissner A. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463): 477–81, 2013. ISSN 1476-4687. doi: 10.1038/nature12433.

Zong L., Abe M., Ji J., Zhu W.-G., and Yu D. Tracking the Correlation Between CpG Island Methylator Phenotype and Other Molecular Features and Clinicopathological Features in Human Colorectal Cancers: A Systematic Review and Meta-Analysis. *Clinical and translational gastroenterology*, 7(3):e151, 2016. ISSN 2155-384X. doi: 10.1038/ctg.2016.14.

This is the end, beautiful friend
This is the end, my only friend, the end
Of our elaborate plans, the end
Of everything that stands, the end
— The Doors *The End*