Signatures of Mutational Processes in Human Cancer

Ludmil B. Alexandrov

Darwin College
University of Cambridge

Thesis submitted for the degree of Doctor of Philosophy

July 2014

DECLARATION OF ORIGINALITY

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given in the bibliography. The dissertation does not exceed the stipulated word limit of 60 000 words.

SUMMARY

All cancers originate from a single cell that starts to behave abnormally due to acquired somatic mutations in its genome. These somatic mutations may be the consequence of the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA, or defective DNA repair. In some cancer types, a substantial proportion of somatic mutations are known to be generated by exposures, for example tobacco smoking in lung cancers and ultraviolet light in skin cancers, or by abnormalities of DNA maintenance, for example defective DNA mismatch repair in some colorectal cancers. However, our understanding of the mutational processes that cause somatic mutations in most cancer classes has been remarkably limited.

Different mutational processes often generate different combinations of mutation types, termed "signatures." There is strong evidence from analyses of known cancer genes in lung cancers and skin cancers that the classes of mutations found and their characteristics match those induced experimentally by tobacco carcinogens and ultraviolet light respectively, the known carcinogenic influences in these cancer types. Thus, the analysis of mutational signatures found in human cancers can provide clues to the processes that have been operative during their development.

In this thesis, I create a theoretical model describing the signatures of mutational processes operative in cancer genomes and develop a systematic computational framework to decipher mutational signatures from mutational catalogues of cancer genomes. The approach is extensively evaluated with simulated data and initially applied to 119 breast cancer whole-genome sequences and 844 breast cancer whole-exome sequences. Novel and known breast cancer mutational signatures are revealed and the contribution of each signature to each cancer sample is estimated.

After this initial application, I use the developed computational framework to perform a comprehensive analysis of cancer genomics data. The approach is applied to 4,938,362 somatic substitutions and insertion/deletions from 7,042 human cancers of 30 classes revealing more than 20 distinct mutational signatures. Some are present in many cancer types, notably a signature attributed to the *APOBEC* family of cytidine deaminases, whereas others are confined to a single cancer class. For some of these processes the underlying biological mechanism is unknown. However, some of

the identified mutational signatures associate to age of cancer diagnosis, smoking, UV light, anticancer drug exposure, presence of *BRCA1* and *BRCA2* mutations, and inactivation of mismatch repair genes.

This thesis provides both a basis for characterizing mutational signatures from cancer-derived somatic mutational catalogues and the first large-scale examination of mutational signatures across multiple cancer types. The results reveal the diversity of mutational processes underlying the development of cancer, with potential implications for understanding of cancer etiology, prevention, and therapy.

ACKNOWLEDGEMENTS

I would like to thank my PhD supervisor, Professor Sir Mike Stratton, for giving me the opportunity to work on this extremely exciting topic and the freedom to explore it from various different angles. Thank you Mike for patiently enduring all my whims and for your support throughout the years! I would have never been able to complete this PhD without your continuous guidance and encouragement. Working with you has been a privilege and an honor!

I would also like to especially thank Dr. Serena Nik-Zainal for the many fruitful discussions about mutational signatures and their biological meaning as well as Dr. Peter Campbell and Dr. David Wedge for all the interesting discussions about mathematical modeling of mutational processes and statistical analysis.

The completion of this PhD project would not have been possible without the advice and encouragement from the members of the Cancer Genome Project. I am also indebted to the hundreds of collaborators throughout the world that were interested in mutational signatures and were willing to share their data with me. I would also like to thank the cancer genomics community (and especially the members of The Cancer Genome Atlas and the International Cancer Genome Consortium) for making all their somatic mutations freely available.

This PhD would not have been possible without the love and support of friends and family. I would not been able to list everyone, but many thanks to Stefan, Elizabeth, Jimmy, Alison, Nicole, and Tessa. I would like to especially thank Oakleigh for her incredible help with proofreading the whole thesis and all her love, care, and patience throughout the past two years. Thank you for being there for me!

I would like to thank my sisters Iliana and Stoyana for always brightening my day and for bringing excitement to my life. I would also like to thank my mother Dora for believing in me and for her love, support, and understanding throughout the past four years. I would like to thank my grandma Lyudmila and my late grandma Lubka for their love and support. Special thanks to my father Boian for showing me the amazing world of science, for enduring me throughout the hardest times, for being there for me no matter what! Гръб до гръб двамина с тебе, бихме се с цял екипаж.

Lastly, I would like to thank the Wellcome Trust for generously funding the last four years and for allowing me to explore the world of cancer genomics.

Table of contents

Chapter 1

Overview of the literature and a historical perspective

1.1 Introduction

The first known historical record in which cancer is described as a disease dates back to *c.* 2600 BCE and it is attributed to Imhotep, the high priest of the Sun god Ra during the rule of king Djoser of the Third Dynasty of ancient Egypt. Evidence of early attempts for surgical treatments of malignancies can be found in the records of the ancient Greek historian Herodotus around the fifth century BCE (Mukherjee, 2010). Throughout the last 4,600 years, our understanding of cancer has evolved and changed numerous times. Many hypotheses proposing the causes of cancer and potential ways to treat cancer have been put forward only to be rejected, and later re-proposed, and then rejected once again (Mukherjee, 2010). In the past 50 years, cancer research has become both a national and, more recently, an international priority. Perhaps the most famous on-going national initiative is the so-called "War on Cancer" – a federal law signed by the former United States president Richard Nixon in 1971 with the goal "to more effectively carry out the national effort against cancer" – resulting in billions of U.S. dollars for funding for cancer research every year. While significant scientific advances have been made in understanding cancer, the general public has perceived these initiatives as "lacking progress" (Rettig, 2006) and consider cancer one of its biggest fears (Roberts, 2010). This fear is, perhaps, well-grounded as ~8 million deaths worldwide each year are attributed to cancer and it is expected that this number will significantly rise with the anticipated increase of human life expectancy (Jemal et al., 2011).

Currently, the term “cancer” encompasses a broad group of over two hundred different diseases characterized by abnormal cellular growth. It is generally agreed that all cancers progress from a single cell that starts to behave abnormally, to divide uncontrollably, and (eventually) to invade adjacent tissues (Hanahan and Weinberg, 2000). It is also believed that the reason this single cell begins to behave abnormally is because of acquired changes to its genetic material, known as somatic DNA mutations.

In this thesis, I will examine patterns of somatic DNA mutations from cancer genomes in order to provide a better understanding of the processes that have caused these mutations and, as such, are the origins of cancer. The aim of this first chapter is to provide a general overview of the state of cancer genetics and cancer genomics as well as to summarize the current knowledge of DNA damage and repair processes. It should be noted that this chapter does not review any of the articles that have been published as part of this thesis as these will be presented in the next few chapters. A complete list of publications associated with this thesis can be found in Appendix VII.

1.1.1 The somatic mutation theory of cancer

The somatic mutation theory of cancer research was initially proposed in the late nineteen century. In 1890, David von Hansemann examined 13 different carcinoma samples and observed an asymmetric distribution of 'chromatin loops' (von Hansemann, 1890). He proposed that aberrant cell divisions are responsible for cellular defects that result in the development of cancer cells. This idea was largely ignored, but 25 years later the German biologist Theodor Boveri revived it and speculated that ‘a malignant cell [should be regarded] as one that carries an irreparable defect’ and that ‘this defect is located in the nucleus’ (Boveri, 2008; Manchester, 1995). Boveri’s and von Hansemann’s work came in a time before DNA was identified as the molecule of inheritance (Avery et al., 1944) and, as such, the defects they were referring to were anomalous chromosomes following aberrant cellular divisions. New observations allowed refinement of Boveri’s theory and, in 1953, Carl Nordling published his multi-mutation “theory on cancer-inducing mechanism” (Nordling, 1953). Nordling observed that in the United States, the United Kingdom, France, and Norway cancer death rates increased according to the sixth power of the age of the patient. He speculated that cancer development requires an accumulation of at least six consecutive mutations. While Nordling’s hypothesis

appealed to medical statisticians (Armitage and Doll, 1954), it was not widely accepted at the time.

Two decades later, Alfred Knudson refined Nordling's theory by examining retinoblastomas. Knudson observed that the heritable form of retinoblastoma occurred at a much earlier age than the non-heritable form, and he explained this observation by speculating that at least two mutational events were necessary for the development of this cancer (Knudson, 1971). Patients that present with the heritable form of retinoblastoma harbour a germline mutation since conception and require only one DNA mutation in a somatic cell to develop the cancer. In contrast, in the nonhereditary type of retinoblastoma, two DNA mutations need to occur in a somatic cell in order to initiate oncogenesis.

Further work on retinoblastoma revealed that the gene harbouring germline mutations is the retinoblastoma gene *RB1* (Murphree and Benedict, 1984). One of the functions of *RB1* is to inhibit cell cycle progression and, as such, to prevent excessive cellular growth. This was the first discovery of a *tumour suppressor gene* (also known as *anti-oncogene*) as *RB1* was directly inhibiting neoplastic development. In principle, most tumour suppressor genes are recessive since even one copy of the gene is sufficient to produce the correct protein and suppress tumorigenesis.

The discovery of the structure of deoxyribonucleic acid (Watson and Crick, 1953) and the experimental work that followed from it reinforced the notion that cancer has a genetic etiology. Early cytogenetic examinations of chromosomal abnormalities demonstrated that specific translocations are associated with particular cancer types. Perhaps the best-known example is that of the "Philadelphia chromosome," a translocation between chromosomes 9 and 22 found in approximately 95% of chronic myelogenous leukaemias (Nowell, 1962; Rowley, 1973). Subsequently, seminal studies in the 1970s and 1980s revealed that mutated genes could cause neoplastic transformation. Most notably, Harold Varmus and J. Michael Bishop demonstrated that the oncogene of the Rous sarcoma virus is required to transform infected chicken cells into neoplastic cells (Parker et al., 1984; Stehelin et al., 1976). A few years later, by transferring genomic DNA from tumour cell lines of mouse and human origin, Robert Weinberg and colleagues established that mouse fibroblasts could be converted into neoplastic cells (Shih et al., 1981).

Further studies demonstrated that the transformation of a normal cell to a neoplastic cell is due to mutated genes responsible for cellular growth control

(Perucho et al., 1981; Pulciani et al., 1982). Such genes were termed *proto-oncogenes* since they are able to induce oncogenesis when mutated. *HRAS* is generally considered to be the first discovered "naturally occurring" oncogene since it was shown that in the NIH/3T3 cell line a single point mutation, which results in an amino acid change of glycine to valine in codon 12 of *HRAS*, is sufficient for tumour initiation (Reddy et al., 1982). In principle, most oncogenes are dominant, as even a single malfunction copy of the gene may be able to provide clonal growth advantage.

The seminal findings summarized in this section have had colossal implications that have shaped the last 30 years of cancer research and underpinned the on-going hunt for mutated genes that cause human cancer.

1.1.2 Acquiring somatic mutations: drivers and their passengers

The somatic mutation theory postulates that cancer is due to the accumulation of somatic mutations, where a somatic mutation is defined as the change of the nucleotide sequence of the genome of a somatic cell since the first division of the zygote. These mutations are the by-product of the endogenous or exogenous DNA damaging processes (reviewed in section 1.2 of this chapter) and are affected by the activity of the operative DNA repair processes (reviewed in section 1.3). I will refer to the combination of DNA damaging and repair processes, operating together and resulting in the generation of somatic mutations, as a "mutational process".

In general, it is accepted that somatic mutations occur somewhat randomly across the genome and that they can be broadly separated into two categories – (i) mutations that provide selective advantage for clonal expansion and (ii) mutations that do not result in growth advantage (Stratton et al., 2009). The latter have been termed *passenger mutations*, while the former are referred to as *driver mutations*. It is widely believed that the number of driver mutations in a cancer sample is limited to a handful, usually two or more but less than ten (Hanahan and Weinberg, 2000). In contrast, the genome of a cancer can harbour more than a million somatic mutations (Alexandrov et al., 2013a) most of which are considered to be passengers. Passenger mutations are not *per se* involved in cancer development but are rather the residual molecular fingerprints of the operative mutational processes.

1.1.3 Mutational catalogues of cancer genomes

Even before the official start of the Human Genome Project, it was hypothesized that systematically analysing the genetic information of cancer cells at a single base resolution could give significant insights into the mechanisms of cancer development (Dulbecco, 1986). While previous approaches allowed identification of large genomic events (*e.g.*, copy number changes, chromosomal translocations, *etc.*) examining cancer genes by interrogating their sequence held the promise of observing previously unseen mutational events. At first, such sequencing examinations were performed using polymerase chain reaction (PCR)-based capillary sequencing for a targeted set of genes; however, the development of next-generation sequencing methods allowed rapidly sequencing of the complete set of exons in a cancer genome and even, at a low cost, the whole cancer genome of a patient.

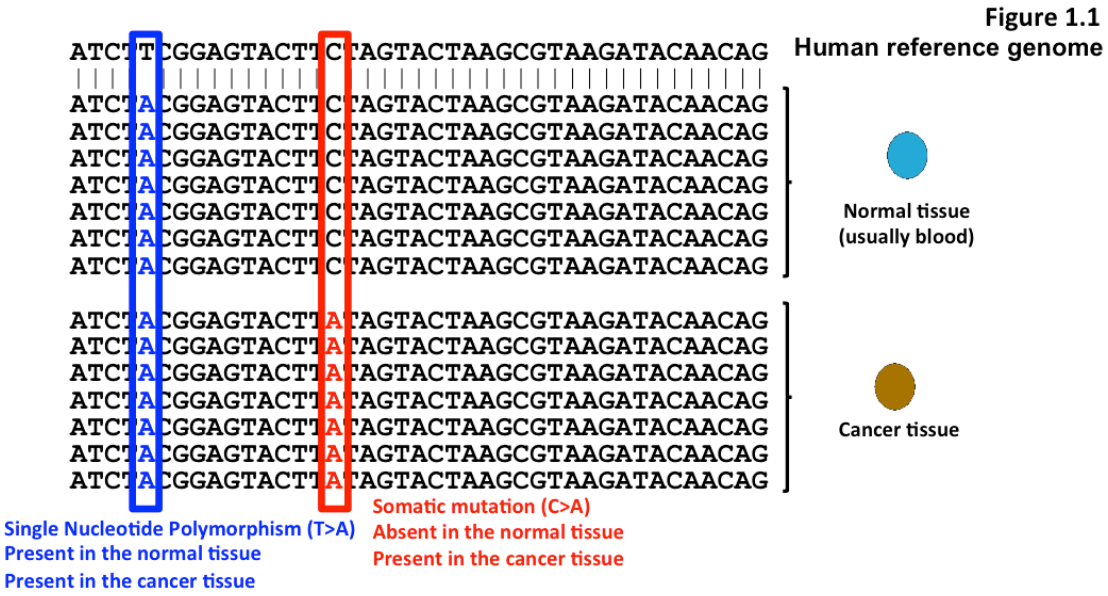


Figure 1.1: Somatic mutations in cancer versus nucleotide polymorphisms in the germline. Illustrated example demonstrating the identification of germline polymorphisms and somatic mutations from sequencing data.

Regardless of the experimental approach, the idea behind sequencing cancer genomes (or parts of these genomes) is simple. Genomic DNA is extracted from both the cancer and from normal tissue (which is usually but not always blood) and then these genomic DNAs are sequenced separately. The identified normal and cancer nucleotide sequences are aligned to the reference human genome, are compared to it, and are then compared to each other. The nucleotide differences found in both the normal and the cancer tissues are attributed to germline polymorphisms while DNA sequence changes identified only in the cancer tissues are attributed to somatic

mutations. The DNA changes identified only in the cancer tissue constitute the mutational catalogue of the cancer genomes. These can be single-base substitutions, small insertion or deletions (usually referred to as *indels*), copy number changes, intra-chromosomal rearrangements, or inter-chromosomal rearrangements. An illustrative example of the identification of a somatic base substitution and a single nucleotide polymorphism from next generation sequencing reads is shown in Figure 1.1.

The majority of somatic mutations identified in the mutational catalogues of cancer genomes are passenger mutations (Stratton et al., 2009). The ability to examine hundreds and even thousands of mutational catalogues of cancer genomes has resulted in the development of advanced statistical methods that allow pinpointing a handful of driver mutations from an ocean of passenger mutations. In simple terms, these algorithms evaluate which genes are mutated more often than purely expected by chance while correcting for multitude of different factors (Garraway and Lander, 2013).

Using targeted capillary sequencing, an early cancer genomics sequencing study demonstrated that mutations in the *BRAF* gene are found in ~70% of melanomas (Davies et al., 2002). This was followed by later studies identifying *PIK3CA* (Samuels et al., 2004) and *EGFR* (Lynch et al., 2004; Paez et al., 2004; Pao et al., 2004) as genes commonly mutated in human cancer. These early successes and their clinical significance (Antoniu, 2011; Chapman et al., 2011b) made the identification of cancer genes through the systematic sequencing of cancer genomes, one of the main topics of cancer research. The emergence of next generation sequencing technologies allowed rapid and cheap examination of the genetic material of cancer cells. This led to the formation of the International Cancer Genome Consortium (ICGC) (Hudson et al., 2010). The goal of the ICGC is the identification of novel cancer genes through the molecular characterization of tumours of 50 types (and their adjacent normal tissues) from more than 25,000 patients. Nowadays, large-scale initiatives, such as the ICGC, continue to identify genes causally implicated with tumorigenesis and the census of human cancer genes gets updated on nearly a monthly basis.

Figure 1.2

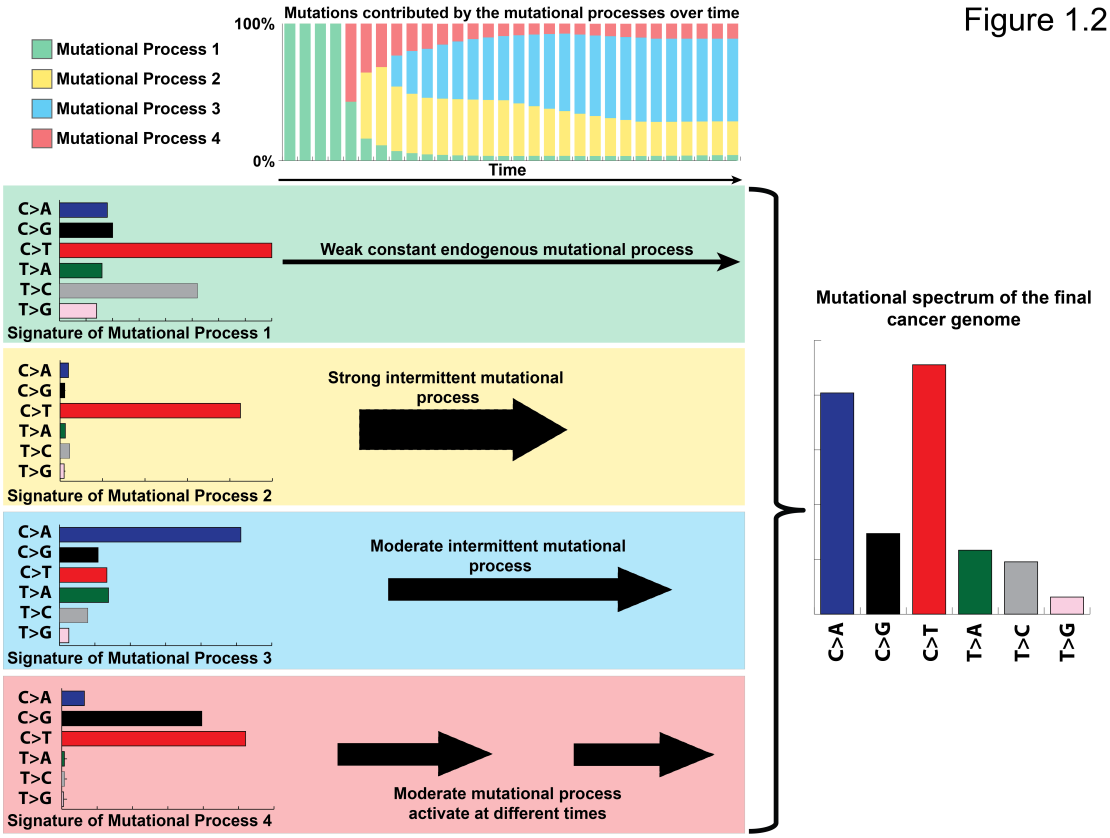


Figure 1.2: Illustration of mutational processes operative in a cancer. This simulated example illustrates four distinct mutational processes with variable strengths operative at different times throughout the lifetime of a patient. Each of these processes has a unique mutational signature exemplified by the six classes of somatic substitutions. At the beginning, all mutations in the cell (from which the cancer eventually developed) were due to the activity of the endogenous mutational process 1. As time progresses, other mutational processes get activated and the spectrum of the mutational catalogue continues to change. Note that the final sequenced cancer genome does not resemble any of the operative mutational signatures.

1.1.4 Mutational signatures - the fingerprints of mutational processes

The somatic mutations in a cancer genome are the cumulative result of the mutational processes that have been operative since the very first division of the fertilized egg from which the cancer cell was derived (Stratton, 2011; Stratton et al., 2009). Each of these mutations was caused by the activity of endogenous and/or exogenous mutational processes with different strengths. A mutational process can leave a characteristic imprint of mutation types, termed *mutational signature*, on the genome of a cancer cell. Some of these processes have been active throughout the whole lifetime of the cancer patient while others have been sporadically triggered, for example, due to lifestyle choices. As multiple mutational processes are operative at different times, multiple mutational signatures have been imprinted on the genome of a cancer cell (Figure 1.2). Thus, the mutational catalogues of a sequenced cancer

genome can be examined as an archaeological record moulded by the many different mutational processes operative since the very first division of the zygote. As such, the pattern of mutations found in the genome of a cancer cell may not resemble the signatures of any single individual operative mutational process; rather, it will be a mixture of these signatures (Figure 1.2). An exception from this rule will be when one of the mutational processes is dominant and generates the large majority of somatic mutations in a cancer sample (*e.g.*, ultraviolet light in skin cancer or tobacco smoking in some types of lung cancer).

1.2 Molecular processes that damage or mutate DNA

DNA damage plays a key role in the gradual decline of cellular functionality over time and it has significant implications for both neoplastic development (Stratton, 2011; Stratton et al., 2009) and ageing (Park and Gerson, 2005). A significant proportion of known DNA damage has been attributed to mutagens generated by normal cellular processes (De Bont and van Larebeke, 2004; Jackson and Loeb, 2001), while some DNA damage is due to the activity of exogenous mutagens (Morley and Turner, 1999). Damaged DNA can be repaired by the cellular machinery, trigger cellular senescence, activate apoptosis mechanisms, or result in a somatic mutation (Hoeijmakers, 2009). Although DNA damage is very common throughout the lifetime of a cell, it is widely believed that most of this damage is repaired and only a very small proportion results in subsequent somatic mutations (Sancar et al., 2004). In the next section I will discuss the most common types of DNA damage and the types of somatic mutations they may cause if unrepaired or repaired incorrectly. Summary of the known patterns of somatic mutations due to DNA damage is provided in Table 1.1. This list is in no way exhaustive as it is most probable that the current knowledge of DNA damage is incomplete.

DNA damage	Type of damage	Mutational pattern
Generation of apurinic/apyrimidinic sites	Spontaneous or enzymatic conversions	C>T substitutions
Deamination of methylated cytosine	Spontaneous or enzymatic conversions	C>T substitutions at CpG dinucleotides

Deamination of cytosine	Spontaneous or enzymatic conversions	C>T substitutions at TpC dinucleotides C>G substitutions at TpC dinucleotides
Deamination of adenine	Spontaneous or enzymatic conversions (extremely rare in humans)	T>C substitutions
Deamination of guanine	Spontaneous or enzymatic conversions	C>T substitutions in some rare cases
Ionizing radiation	Physical agents	Rearrangements due to double strand breaks
Non-ionizing radiation	Physical agents	C>T substitutions and CC>TT double substitutions at dipyrimidines
Oxidative damage	Spontaneous conversions, enzymatic conversions, or physical agents	Many different types but best-described spectrum of mutations for 8-oxoG: C>A with a preference for CpCpC trinucleotides
Alkylating agents	Chemical compound	C>T substitutions
Psoralen	Chemical compound	T>X substitutions
Polycyclic aromatic hydrocarbons	Chemical compound	C>A substitutions
Mineral fibres	Chemical compound	C>A substitutions

Table 1.1: Known mutational signatures due to DNA damage. All substitutions are referred to by the pyrimidine of the mutated Watson–Crick base pair. Mutated bases are underlined when the mutation depends on the immediate sequence context.

1.2.1 Spontaneously occurring endogenous DNA lesions and mutations

Perhaps the best-described endogenous DNA damaging processes are those due to spontaneous reactions (mostly hydrolysis), chemicals generated by cellular metabolic processes (*viz.*, reactive oxygen species, lipid peroxidation products, endogenous alkylating agents, *etc.*), errors during cellular division and misincorporation by DNA polymerases. Naturally and spontaneously occurring DNA damage and its consequent somatic mutations are continuously eroding the genome of every cell in the human body throughout the person's lifetime. It has been estimated that spontaneous DNA damage arises with an average rate of ~70,000 lesions and/or strand breaks per day per mammalian cell (most of which get repaired by the cellular machinery) with these ranging from 50,000 up to 200,000 between different cell types (Bernstein et al., 2013). In the next few paragraphs, I will briefly review some of the best-known DNA damaging processes.

1.2.1.1 Double-strand and single-strand DNA breaks

Double-strand and single-strand DNA breaks occur endogenously in mammalian cells and the cell employs different mechanisms to repair them. Non-homologous end joining, microhomology-mediated end joining, and homologous recombination are used by the cell to repair double-strand DNA breaks; in contrast single-strand breaks are repaired by the cellular excision repair mechanisms: base excision repair, nucleotide excision repair, or mismatch repair (see section 1.3 for more details). Endogenous double-strand breaks are particularly damaging for the cell and are generally driven by single-strand lesions. It has been estimated that ~1% of all single-strand lesions result in double-strand breaks after every cellular division (Vilenchik and Knudson, 2003). This results in approximately 50 double-strand breaks per cell per cell cycle. In contrast, endogenous single-strand breaks are believed to be more ubiquitous and it has been estimated that thousands (and even tens of thousands) of single-strand breaks occur in each human cell every single day (Tice and Setlow, 1985). Single-strand breaks can be caused by a variety of damaging agents such as oxidation, alkylation, formation of pyrimidine dimers, deamination, *etc.* The majority of single-strand breaks are repaired by the cellular repair mechanisms (Tice and Setlow, 1985).

1.2.1.2 Oxidative DNA damage

Oxidative DNA damage can be generated as both a product of normal activity of cellular metabolism and as a result of exogenous agents such as radiation exposure or air pollutants (Cooke et al., 2003). It is estimated that spontaneous oxidative DNA damage results in at least 12,000 lesions per cell per day in human cells (Helbock et al., 1998). In principle, reactive oxygen species (ROS) and reactive nitrogen species (RNS) are the intermediates responsible for the majority of oxidative damage (Wiseman and Halliwell, 1996). ROS is a collective term used to include O_2-derived free radicals as well as O_2-derived non-radical species that easily convert to radicals or that can act as oxidizing agents (Circu and Aw, 2010). Similarly, RNS is a very broad term that encompasses all oxides of nitrogen (Patel et al., 1999). Currently, more than 25 distinct DNA lesions have been described and associated with the activity of ROS/RNS. However, the exact chemistry of somatic mutations potentially arising from these lesions has only been well characterized for a few of these ROS/RNS (Evans et al., 2004).

The variety of ROS/RNS accounts for the plethora of DNA lesions that these substrates can induce on the deoxyribonucleic acids: generation of apurinic and apyrimidinic sites, single-strand and double-strand DNA breaks, deamination, *etc.* (Hori et al., 2011; Wang et al., 2012). The wide variety of DNA lesions that can be generated by RNS/ROS challenges the development of a comprehensive characterization of the spectrum of oxidation-arising somatic mutations. Perhaps the best-described spectrum of mutations is 7,8-dihydro-8-oxoguanine (8-oxoG), an oxidatively damaged form of guanine. 8-oxoG can lead to the misincorporation of adenine opposite the 8-oxoG resulting in a higher prevalence for C:G>A:T transversions upon replication (Michaels et al., 1992). It has been speculated that somatic mutations due to 8-oxoG might be dependent on the immediate sequence context with preference for C>A transversions at Cp<u>C</u>pC sequences, (the mutated base is underlined; all substitutions are referred to by the pyrimidine of the mutated Watson–Crick base pair) (Oikawa and Kawanishi, 1999; Oikawa et al., 2001).

1.2.1.3 Depurination and depyrimidination

Depurination and depyrimidination are some of the most common hydrolytic reactions that cleave the N-glycosidic bond of a nucleic acid base and damage DNA by respectively resulting in an apurinic or an apyrimidinic site (also known as abasic

site). The rate of generation of depurination is estimated to be ~10,000 per cell per day (Lindahl, 1993), while depyrimidination arises with a rate of about 700 lesions per cell per day (Tice and Setlow, 1985). While abasic sites lack genetic information and the majority of them are repaired by base excision repair (BER), some (especially the ones present during the DNA synthesis phase of the cell cycle) can present a challenge for the replicative polymerases during cellular division and cause replication fork stalling (Obeid et al., 2010). It has been previously demonstrated in yeast that the joint actions of DNA polymerases δ and ζ allow bypassing of abasic lesions and continuation of DNA replication (Haracska et al., 2001); however, the cost of continuing the replication process is the misincorporation of a nucleotide opposite the abasic site. This nucleotide is most commonly an adenine (also referred as the "A-rule") but in rare cases it can also be cytosine, guanine, or thymine (Haracska et al., 2001).

1.2.1.4 Methylation of DNA nucleotides

The addition of a methyl group to adenine or cytosine is referred to as DNA methylation. Methylation of a cytosine results in either N4-methylcytosine or 5-methylcytosine, whereas adenine methylation leads to the formation of N6-methyladenine (Ratel et al., 2006). Early examination of mammalian DNA revealed the widespread nature of 5-methylcytosine (Ehrlich et al., 1982). In contrast, N4-methylcytosine and N6-methyladenine are found almost exclusively in bacteria, although it has been speculated that they might exist at extremely low levels (less than a hundred nucleotides) in the genomic DNA of some human cells (Ratel et al., 2006).

In somatic mammalian cells, 5-methylcytosine occurs predominantly at a cytosine followed by a 3' guanine (*i.e.,* CpG dinucleotide), while cytosine methylation at non-CpG sites is ubiquitous in embryonic stem cells (Dodge et al., 2002; Haines et al., 2001; Lister et al., 2009). Interestingly, 5-methylcytosine plays the role of a double-edged sword. On the one hand, it carries epigenetic information that is leveraged by the cell, for example, in regard to regulating gene expression in different tissue types (Jones, 2012b); on the other hand, a 5-methylcytosine can easily be hydrolytically deaminated to a thymine, resulting in perhaps the best-described mutational pattern: C>T mutations at CpG dinucleotides (see below for details about spontaneous deamination).

Recently, it was shown that in mammalian tissues the *ten-eleven translocation methylcytosine dioxygenase* (TET) family of enzymes could facilitate the oxidation of 5-methylcytosine resulting in 5–hydroxymethylcytosine (Tahiliani et al., 2009). Further, studies have demonstrated that 5–hydroxymethylcytosine is widespread in embryonic stem cells as well as somatic brain tissue in mice and humans (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). The implications of these findings in regard to cancer and somatic mutagenesis are currently unknown (Pfeifer et al., 2013).

1.2.1.5 Deamination of DNA nucleotides

Deamination is an endogenously occurring molecular process that results in the removal of an amine group from a molecule. In the genome of eukaryotic cells, it is has been demonstrated that cytosine, 5-methylcytosine, 5-hydroxymethylcytosine, guanine, and adenine can be spontaneously deaminated.

1.2.1.5.1 Deamination of cytosine

Enzymes deaminate cytosine and convert it to uracil ~500 times per human cell per day (Lindahl and Nyberg, 1974). As uracil has the aptitude to pair with adenine, this DNA damage can give rise to C>T mutations. In general, the *activation-induced cytosine deaminase* (*AID*) and the family of *apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like* (*APOBEC*) enzymes have been associated with cytosine deamination. *AID* has been exhaustively studied in regards to somatic hypermutation, a process that mutates antibody genes in order for the immune system to respond to an invasion of foreign molecular agents (Liu and Schatz, 2009), and its pattern of somatic mutations has been well described. *AID* predominantly deaminates cytosine that is flanked by a 5' purine (Pham et al., 2003).

The *APOBEC* family of enzymes, which includes *APOBEC1*, *APOBEC2*, *APOBEC3A*, *APOBEC3B*, *APOBEC3C*, *APOBEC3D/E*, *APOBEC3F*, *APOBEC3G*, *APOBEC3H*, and *APOBEC4*, can also deaminate cytosine. Note that some classifications include *AID* in the *APOBEC* family of deaminases while others refer to it as the *AID/APOBEC* family (Conticello, 2008). With the exception of *APOBEC4*, which has been inferred only bioinformatically (Rogozin et al., 2005), all other *APOBEC* enzymes have a known nucleotide-editing capability at least in relation to mutating RNA (Conticello, 2008; Teng et al., 1993). The activities of these enzymes exhibit a characteristic set of base changes but different members of the enzyme

family act at different sequence contexts. Importantly, previous *in vitro* cell line studies have demonstrated that *APOBEC1*, *APOBEC3A*, and *APOBEC3B* are capable of mutating DNA by the deamination of cytosine flanked by a 5' thymine and thus result in C>T mutations at TpCpN trinucleotides (Harris et al., 2002; Hultquist et al., 2011; Suspene et al., 2011; Taylor et al., 2013). Furthermore, it has been shown that the activation of *APOBEC3A* and *APOBEC3B* in yeast can also result in C>G at TpCpN trinucleotides (Taylor et al., 2013). This mutational pattern was attributed to replication over an abasic site, formed when an *APOBEC* deaminated cytosine is excised by uracil-DNA glycosylase, which is catalysed by *REV1* (Taylor et al., 2013).

1.2.1.5.2 Deamination of 5-methylcytosine

In contrast to the spontaneous deamination of cytosine, which results in the formation of uracil, the methylated form of cytosine (*viz.*, 5-methylcytosine) is hydrolytically deaminated to thymine. In addition to hydrolytic deamination, deamination of 5-methylcytosine has also been attributed to the activity of AID and APOBEC1 (Morgan et al., 2004). The overall rate of 5-methylcytosine deaminations is approximately 1,500 deaminations per human cell per day with the majority of mutations occurring at CpG dinucleotides (Shen et al., 1994). This DNA damaging process has a very well-documented mutational profile, resulting in C:G>T:A mutations at CpG dinucleotides, and plays an important role in both evolution (Zemach et al., 2010) and neoplastic development (Laird and Jaenisch, 1996).

1.2.1.5.3 Deamination of 5-hydroxymethylcytosine

Deamination of 5-hydroxymethylcytosine results in the production of 5-hydroxymethyluracil, which is generally removed by the activity of base excision repair (Rusmintratip and Sowers, 2000). The rate of this deamination as well as the implications of the formation of 5-hydroxymethyluracil in regards to somatic mutations and cancer development are currently unknown (Pfeifer et al., 2013).

1.2.1.5.4 Deamination of adenine

Adenine is oxidatively deaminated to hypoxanthine with a rate of ~50 deaminations per human cell per day (Lindahl, 1993). During DNA replication, hypoxanthine preferentially pairs with guanine resulting in the formation of T:A>C:G mutations (Lindahl, 1993).

1.2.1.5.5 Deamination of guanine

Guanine can also be spontaneously deaminated and the resulting product is xanthine (Fernandez et al., 2009). Xanthine preferentially pairs with cytosine and, as such, in the majority of cases this product is not mutagenic. Nevertheless, it has been shown that xanthine can also pair (albeit less frequently) with thymine resulting in the C:G>T:A mutations after replication (Fernandez et al., 2009).

1.2.1.6 DNA mutations due to cellular replication

DNA replication is an essential biological process that occurs in all living organisms and underlies the basic inheritance of genetic information. In human beings, the mitosis of a cell involves accurately copying approximately six billion base pairs and, as such, DNA replication has been evolutionarily optimized to have an astonishing fidelity and to produce only a very limited number of errors during each cellular division (Masai et al., 2010).

DNA replication starts simultaneously from multiple specific locations of the genome, termed origins of replication. Between 30,000 and 50,000 such origins of replication are activated in a human cell during each cellular division (Mechali, 2010). In eukaryotic cells, prior to the initiation of replication, the double-stranded DNA is opened by DNA helicases to form the so-called “replication fork", which contains the two separated single strands of DNA – known as the leading and the lagging strand. Replication is a complex molecular process, recently reviewed in (Masai et al., 2010), that entails the coordinated activity of three distinct types of DNA polymerases: polymerase α, polymerase δ, and polymerase ε. Briefly, polymerase α is the enzyme that starts DNA replication by playing the role of a replicative primase. The closely related polymerases δ and ε are responsible for the synthesis of respectively the lagging and leading strands. Both polymerase δ and polymerase ε have intrinsic proofreading mechanisms and their probability for making a mistake has been estimated to be approximately 10^{-7} for each nucleotide (McCulloch and Kunkel, 2008). This error probability is further reduced to about 10^{-9} by the post-replicative activity of mismatch repair (McCulloch and Kunkel, 2008). Thus, theoretically, replicating the genome of a human cell that does not contain any damaged DNA will result in only ~6 somatic mutations. However, in practice, it is rare (if ever) for a cell to have a completely damage-free genome.

Replication is a sophisticated and fine-tuned molecular process that can be affected by the presence of most types of DNA damage (Sale et al., 2012). The existence of DNA damage presents a conundrum to a mitotic cell since it needs to replicate its damaged genome. The task of performing replication of a damaged genomic segment is referred to as DNA damage tolerance and attributed to a set of DNA polymerases that are members of the Y-family of polymerases. These polymerases are able to replicate damaged DNA but they lack any proofreading capabilities and, as such, have a probability for making an error between 10^{-1} and 10^{-4} (Sale et al., 2012). Nevertheless, it is generally believed that only very short stretches of DNA are being synthesized due to DNA damage tolerance, thus keeping the number of newly generated somatic mutations to a minimum (Sale et al., 2012).

The synthesis of a new genome is heavily dependent on the availability of substrates for the use of the DNA polymerizing enzymes, *viz.*, deoxynucleoside triphosphates (dNTPs). Changes in the levels of dNTPs have been associated with significant variation in mutagenesis. In eukaryotes, it has been demonstrated that imbalances (mostly reduction) of the dNTP pools result in decreased genome stability that increases the probability of somatic insertions and misalignments (Kumar et al., 2011). Interestingly, a recent study showed that, in *Escherichia coli,* decreasing the level of the dNTP pool is associated with improved accuracy of the DNA polymerases (Laureti et al., 2013). Thus, the interplay between DNA polymerases and dNTP pools might be more complex than was previously believed and it may result in both increased and decreased mutagenesis (Laureti et al., 2013). Nevertheless, analyses of somatic mutations in cancer genomes, as well as variation in the human germline, have shown that indels and point mutations are enriched in late replicating regions and this has been generally attributed to the reduced levels of dNTP (Koren et al., 2012).

Replication does not *per se* damage DNA but it does result in the generation of somatic mutations. While there is no comprehensive pattern of the mutations due to DNA replication, there are several known commonly occurring mutation types. Perhaps the best-described mutations are the ones due to “replication slippage”, where one of the strands forms a loop, which may result in the misincorporation of small insertions or the deletion of nucleotides. Specific regions (*viz.,* microsatellite and other repetitive regions) of the human genomes are more susceptible to replication slippage and, as such, are “hotspots” of mutations due to replication (Viguera et al.,

2001). Nevertheless, future studies are required to determine the precise patterns of all mutations induced by DNA replication.

1.2.2 Exogenous mutagens causing DNA damage and somatic mutations

In addition to endogenous DNA damage, the integrity of the double helix is constantly under attack by the activity of exogenous mutagens. These may be physical, chemical, and even biological agents. The list of external substances that are implicated in DNA mutagenesis is extensive and an exhaustive account is beyond the scope of this thesis.

Perhaps the most detailed catalogue of human carcinogens is the one provided under the auspices of the International Agency for Research on Cancer (IARC). The IARC catalogue includes over 100 confirmed human carcinogens as well as over 300 probable/possible human carcinogens, most recently reviewed in (Cogliano et al., 2011). The majority of these carcinogens have been identified by IARC via epidemiological studies. However, studies that used the *in vitro* Ames test have demonstrated that ~90% of known carcinogens are also mutagenic (McCann and Ames, 1976). In this section, I provide a concise overview of the DNA damage induced by exogenous mutagens that are of interest in regards to the subsequent chapters of this thesis. I will also discuss in detail the patterns of somatic mutations induced by known exogenous substances in human cancer in section 1.4 of this chapter.

1.2.2.1 Therapeutic agents inducing DNA damage

The majority of chemotherapeutic drugs work by damaging DNA (Kim et al., 2000). Notable examples of such chemotherapeutic drugs are alkylating agents and inorganic platinum-based compounds. Other types of therapeutics have also been known to cause DNA damage, *viz.,* psoralens and intercalating agents. It should be noted that cancer radiation therapy also results in DNA damage (Kim et al., 2000). DNA radiation damage will be examined in a wider context in section 1.2.2.3 of this chapter.

1.2.2.1.1 Alkylating agents

Alkylation of DNA is a molecular process in which an alkyl group is

transferred to a DNA nucleotide or the backbone of the double helix (Drablos et al., 2004). Monofunctional alkylating agents bind covalently to one side of DNA, whereas bifunctional alkylating agents create an inter-strand or an intra-strand DNA crosslink. Alkylating agents can arise from normal metabolic processes, environmental compounds, or be cytotoxic/cytostatic chemotherapy drugs. While there are many possible sources of endogenous DNA alkylation, currently their significance for cancer development or their rates of alkylation remain unknown (Drablos et al., 2004).

Although there is a lack of quantitative data in regards to environmental alkylation, it is generally believed that N-nitroso compounds formed in tobacco smoke are the most significant environmental alkylating agent for humans (Hecht, 1999). Nevertheless, a low concentration of N-nitroso compounds is also well established in some types of food such as cured meats (Goldman and Shields, 2003).

Chemotherapeutic anti-cancer drugs expose patients to extremely high doses of alkylation. Most commonly, these are chloroethylating drugs based on bifunctional alkylating compounds that result in the formation of either an inter-strand or an intra-strand DNA crosslink. This may affect a cancer cell in a wide range of ways: DNA breaks, S-phase arrest, accumulation of high levels of *TP53*, and apoptosis (Engelward et al., 1998). The somatic mutational pattern of treatment with alkylating agents has been characterized as C:G>T:A transitions exhibiting a specific immediate sequence context (Greenman et al., 2007; Parsons et al., 2008).

1.2.2.1.2 Inorganic platinum based compounds

Inorganic platinum-based compounds are commonly used as anti-cancer drugs. They form bulky adducts with DNA that result in inter-strand or intra-strand crosslinks. Platinum-based therapy is commonly described as "alkylating-like" due to the similar effects of these two types of antineoplastic drugs (Cruet-Hennequart et al., 2008). While the pattern of somatic mutations due to platinum treatment has not been yet characterized, it has been observed that the majority of platinum-based DNA adducts result in the formation of crosslinks via the coordination of two adjacent guanines (Poklar et al., 1996).

1.2.2.1.3 Intercalating agents

Molecules that may insert themselves between the two strands of the deoxyribonucleic acid (thus, effectively blocking DNA replication) are referred to as intercalating agents (Wakelin, 1986). Intercalating agents have found a wide-range of applications in human diseases and have been used for both antibacterial and anticancer treatment (Sissi and Palumbo, 2003). While these compounds damage DNA and block DNA synthesis, there is currently no known pattern of somatic mutations associated with treatment with intercalating agents.

1.2.2.1.4 Psoralen

Psoralen is a family of chemical compounds commonly used (in combination with ultraviolet light) for treatment of inflammatory conditions such as dermatitis and psoriasis (Stern, 2007). The interaction between ultraviolet light and psoralen compounds results in the formation of monoadducts as well as inter-strand crosslinks (Chiou and Yang, 1995). In human lymphoblasts treated with psoralen and ultraviolet light, examination of the mutational spectra of the *hprt* reporter locus revealed a high level of single base mutations exhibiting a preference for a (mutated) thymine followed by adenine (*i.e.*, T:A>X at TpA) (Papadopoulo et al., 1993).

1.2.2.2 Polycyclic aromatic hydrocarbons

Polycyclic aromatic hydrocarbons (PAHs) are fused aromatic rings usually produced by the burning of fuel. While there are at least a dozen known PAHs implicated in human carcinogenesis (Harvey, 1991), the best described polycyclic aromatic hydrocarbon (in regards to DNA damage and mutagenesis) is benzo[a]pyrene. Benzo[a]pyrene is the first discovered chemical carcinogen and it is one of the many carcinogens found in cigarette smoke (Harvey, 1991). The mutational pattern of benzo[a]pyrene is well described as this compound is able to form bulky adducts with a very high preference for guanines, thus resulting in C:G>T:A transversions. Examining the patterns of *TP53* mutations in lung cancers of tobacco smokers revealed a strong preference for mutations occurring on the untranscribed strand when compared to mutations occurring on the transcribed strand (Hollstein et al., 1999). This strand preference is known as transcriptional strand-bias and it is presumably due to the activity of transcription-coupled nucleotide excision

repair (see sections 1.3 and 1.4 for more details). It should be noted that a whole-genome examination of the mutational patterns of a tobacco smoker revealed that transcriptional strand-bias is present in all transcribed regions of the human genome (Pleasance et al., 2010b).

1.2.2.3 Mineral fibres

Early epidemiological studies have implicated mineral fibres in human and animal carcinogenesis (Barrett et al., 1989). Perhaps the most notable of these mineral fibres is asbestos as this mineral is believed to be "the leading cause of occupational related cancer death" (Tweedale, 2002). Asbestos is a carcinogen implicated in the development of the majority of mesotheliomas, cancers that usually arise in the outer lining of the lungs but could also be found in other organs (Tweedale, 2002). Using an *in vivo* mutagenesis assay based on transgenic rats with a *lacI* reporter gene, a distinct spectrum of somatic mutations was observed after exposure to asbestos (Unfried et al., 2002). This mutational pattern exhibits a combination of C:G>A:T transversions and small (1 to 3 bp long) deletions (Unfried et al., 2002).

1.2.2.3 DNA damage induced by exposure to radiation

Radiation is defined as a process in which an electromagnetic wave travels through a medium or through a vacuum (Vesley, 1999). Radiation can be broadly separated into two categories based on the spectrum of the electromagnetic wave: (i) ionizing radiation and (ii) non-ionizing radiation. The boundary between ionizing and non-ionizing radiation has not been clearly defined and different thresholds of photon energies have been suggested (most commonly either 10 electronvolts or 33 electronvolts). Nevertheless, it is generally agreed that the threshold falls somewhere in the spectrum of the ultraviolet light (Vesley, 1999).

By definition, ionizing radiation has sufficient energy to knock out an electron from its atom and thus to ionize the atom. In contrast, the photons of non-ionizing radiation do not have sufficient energy to ionize an atom. However, non-ionizing radiation may increase the temperature of a medium resulting in thermal-ionization (Vesley, 1999). In a living cell, an exposure to ionizing and (to a much lesser extent) non-ionizing radiation could also indirectly result in the generation of intermediate oxidants, such as reactive oxygen species, which can damage DNA (see section 1.2.1.2).

In the next subsections, I will discuss the different types of DNA damage that can be induced by ionizing and non-ionizing radiation while paying special attention to ultraviolet light.

1.2.2.3.1 DNA damage due to ionizing radiation

Ionizing radiation is an electromagnetic wave with a high frequency (and, thus, a short wavelength) that can break chemical bonds and ionize atoms. Due to its high energy, ionizing radiation is particularly damaging for biological matter (Vesley, 1999). In general there are three main types of ionizing radiation: (i) alpha particles, (ii) beta particles, and (iii) gamma rays. An alpha particle is similar to a helium nucleus as it contains two neutrons and two protons; a simple sheet of paper can absorb this type of radiation. A beta particle is a high-speed electron or positron and an aluminium sheet is required to stop this type of radiation. Lastly, gamma rays are a radiation with extremely high frequency and, thus, contain a very high energy per photon; thick lead walls are required for the complete absorption of high-energy gamma rays.

All three types of ionizing radiation have sufficient energy to break the sugar-phosphate backbone of DNA, disturb the hydrogen bonds in a DNA base pair, or damage a nucleotide (Ward, 1988). However, the best-described mutational signature due to ionizing radiation is the generation of single and double-strand DNA breaks resulting in the generation of small somatic insertions or deletions (Friedberg and Friedberg, 2006). Nevertheless, large numbers of single base substitutions have also been observed in mammalian cells exposed to ionizing radiation (Grosovsky et al., 1988). The spectrum of these mutations is heavily dependent on the type of ionizing radiation and this spectrum has been systematically characterized almost exclusively for ultraviolet light (see below).

1.2.2.3.2 DNA damage due to non-ionizing radiation

Non-ionizing radiation does not carry enough energy to ionize an atom but it can result in atom excitation – the movement of an electron from a ground energy state level to a higher (excited) energy state. DNA exposed to non-ionizing radiation results in excited molecular bonds that commonly form cyclobutane pyrimidine dimers (CPDs, including thymine dimers) and 6,4-photoproducts. These DNA lesions are generally repaired by nucleotide excision repair (Pfeifer et al., 2005) but (if left

unrepaired) they affect DNA base pairing and may result in replication stalling or mutagenesis. In general, 6,4-photoproducts are more mutagenic than CPDs but they occur only at a third of the rate of CPDs (Pfeifer et al., 2005).

In principle, non-ionizing radiation could result in a significant temperature increase and generate intermediate oxidants, such as reactive oxygen species, which can damage DNA (see section 1.2.1.2).

1.2.2.3.3 DNA somatic mutations due to exposure to ultraviolet light

The wavelength of ultraviolet light (UV) is situated between the wavelengths of ionizing radiation and non-ionizing radiation. Thus, exposure to UV light may result in DNA damage consistent with exposure to both types of radiation.

UV light is standardly separated into nine different categories based on the range of the length of the electromagnetic wave. However, with regard to biological organisms, the main interest is in three of these categories - ultraviolet A (UV-A), ultraviolet B (UV-B), and ultraviolet (UV-C) – as these types of UV light are emitted by the Sun and may reach the surface of the Earth. In general, all of UV-C and the majority of UV-B coming from the Sun are absorbed by either the ozone layer or the stratospheric oxygen. About 95% of the UV light reaching the Earth's surfaces is UV-A with the remaining 5% being UV-B. However, in places with a depleted ozone layer (such as Australia) these proportions vary and even some UV-C light may reach the planetary surface.

While UV-C has the highest energy, it has not been implicated in human cancer as, even if not completely stopped by the ozone layer, the outer dead layers of the epidermis easily absorb any residual UV-C (Campbell et al., 1993). UV-B is the ultraviolet light that has been implicated in skin reddening and sunburn. UV-B can penetrate the skin epidermis layer and it can reach (but it is usually absorbed by) the dermis layer. UV-A has been implicated in skin aging and wrinkling. This type of UV light can penetrate deeply in the skin reaching the subcutaneous layer. Both UV-A and UV-B are mutagenic and they have been implicated in cancer development.

I*n vitro* irradiation of mouse embryonic fibroblasts with UV-A and UV-B coupled with the examination of the *cII* transgene was used to characterize the patterns of somatic mutations induced by these two types of radiation. This analysis revealed that ~75% of all examined somatic mutations due to UV-B irradiation result in C:G>T:A transitions including significant numbers of CC:GG>TT:AA dinucleotide

substitutions (You et al., 2001). In contrast, only ~30% of all somatic mutations due to UV-A irradiation are C:G>T:A transitions and this type of irradiation generates only very few dinucleotide substitutions (Besaratinia et al., 2004). Further, UV-A radiation results in significant numbers of other types of somatic substitutions: ~25% C:G>A:T mutations, ~10% T:A>C:G mutations, and ~10% T:A>G:C mutations; and high numbers of small insertions and deletions (Besaratinia et al., 2004). These and other studies, reviewed in (Pfeifer et al., 2005), have demonstrated that the type of DNA damage and the arising spectrum of somatic mutations is highly dependent on the type of ultraviolet light irradiation.

1.2.2.3 Biological agents implicated in cancer development and their mutagenesis

In addition to chemical and physical agents, biological agents play an important role in cancer development. Oncoviruses have been implicated in approximately 12% of all human cancers and vaccination initiatives are on-going to reduce this rate (Schiller and Lowy, 2010). Bacterial infections have also been associated with oncogenesis due to the generation of bacterial metabolites and the initiation of chronic inflammation (Parsonnet, 1995). Nevertheless, currently there is no known type of DNA damage or pattern of somatic mutations due to either bacterial or viral infection.

1.3 Molecular processes responsible for DNA repair

The focus of the prior section was to review some of the most common types of DNA damage. The cell employs a variety of different defence mechanisms to alleviate DNA damage and reduce its effect on the genetic material. When these repair pathways are working properly only very few mutations accumulate in the genome of a cell. However, when one or more of these mechanisms goes awry the result is an increase in the mutational burden, which may produce (and thus it could be detected by) a specific mutational pattern.

In principle, DNA repair pathways can be separated into two categories based on the induced DNA damage. The first category encompasses processes that are operative on single-strand breaks and/or lesions. In contrast, the second type of repair processes has been evolutionary optimized to work on double-strand breaks. In this

section, I will briefly discuss the different repair pathways leveraged by the cell and their relationship with both the previously described types of DNA damage and human cancer. Summary of the known patterns of somatic mutations due to the activity of or the failure of DNA repair mechanisms is provided in Table 1.2.

DNA repair process	Repair activity	Mutational pattern
Base excision repair	Partial failure	C>T substitutions when *SMUG1* is mutated; C>A substitutions when *OGG1* is mutated
Transcription coupled base excision repair (very limited evidence)	Normal function	Transcriptional stand-bias with fewer mutations observed on the transcribed strand?
Transcription-coupled nucleotide excision repair	Normal function	Transcriptional stand-bias with fewer mutations observed on the transcribed strand
Transcription-coupled nucleotide excision repair	Failure	Lack of transcriptional strand bias for known exposure (*e.g.*, ultraviolet light)
DNA mismatch repair	Failure	Increase mutational burden with high prevalence for insertions/deletions at mononucleotide or polynucleotide repeats
Double strand break repair via non-homologous end joining (NHEJ)	Normal function	Increased numbers of insertions/deletions and translocations near microhomologies lengths <= 4bp
Double strand break repair via microhomology mediated end joining (MMEJ)	Normal function	Increased numbers of insertions/deletions and translocations near microhomologies lengths > 4bp

Double strand break repair via homologous recombination	Failure	Double strand breaks get repaired with either NHEJ or MMEJ resulting in a higher numbers of mutations with the mutational patterns of NHEJ/MMEJ

Table 1.2: Known mutational signatures due to the activity of DNA repair mechanisms. All substitutions are referred to by the pyrimidine of the mutated Watson–Crick base pair.

1.3.1 Repairing broken or damaged single strands of DNA

The repair mechanisms that operate on a damaged or broken single strand of DNA are nucleotide excision repair, base excision repair, and mismatch repair. Each of these processes gets activated due to different stimuli and will be reviewed in the next few sections.

1.3.1.1 Nucleotide excision repair

Nucleotide excision repair (NER) is arguably the most multipurpose repair pathway and acts on DNA distortions caused by biochemical modifications (Nouspikel, 2009). The ability of NER to repair a wide-range of DNA damage is based on a simple principle – this repair pathway does not leverage specific enzymes to recognize different DNA lesions but it rather detects any distortions of the DNA double helix (de Laat et al., 1999). When a DNA distortion is identified, a 25 to 30 bases long oligonucleotide (that includes the damage) is excised and replicative polymerases fill the gap by using the complementary undamaged DNA strand (de Laat et al., 1999). The versatility of NER allows it to act on a plethora of different types of DNA damage. Some examples are bulky adducts, aromatic amine compounds, photodimers, and any other lesion that distorts the DNA structure (Nouspikel, 2009). Defective NER in the germline has been associated with several human syndromes, most notably xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy (de Boer and Hoeijmakers, 2000).

NER is evolutionary conserved between eukaryotes and prokaryotes, albeit its molecular mechanisms are more complex in eukaryotic cells (Nouspikel, 2009). In eukaryotic cells, NER is generally separated into two subcategories as different proteins are responsible for the recognition of DNA distortion: (i) transcription

coupled nucleotide excision repair, recently reviewed in (Nouspikel, 2009), and (ii) global genomic nucleotide excision repair, recently reviewed in (Tornaletti, 2009). Additionally, it is also believed that there is a third type of NER, termed domain associated nucleotide excision repair, which has not yet been well described. Domain associated nucleotide excision has also been recently reviewed in (Nouspikel, 2009).

1.3.1.1.1 Global genome wide nucleotide excision repair

The global genome-wide nucleotide excision repair (GG-NER) is a molecular process that is constantly scanning the complete genome of a eukaryotic cell. This process leverages an *XPC-HR23B* protein complex to detect any structural modification of DNA and to bind to any such lesions (Nouspikel, 2009). The bound *XPC-HR23B* recruits a *TFIIH* complex that opens a denaturation bubble around the DNA damage and, in turn, it recruits the *ERCC1–XPF* heterodimer (McNeil and Melton, 2012). The *ERCC1–XPF* complex is a 5' to 3' structure specific endonuclease that excises the damaged DNA strand. The removed ~30 nucleotides are resynthesized by *PCNA* in combination with either DNA polymerase δ or DNA polymerase ε (Essers et al., 2005). Lastly, the chromosomal nicks are sealed by *XRCC1* in association with either DNA ligase I or DNA ligase III (Moser et al., 2007).

The ability of GG-NER to repair a wide variety of different types of DNA damage complicates its detection by the means of mutational patterns. Nevertheless, it is foreseeable that a cancer cell in which GG-NER has been disabled will accumulate somatic mutations at a higher rate and, as such, failure of GG-NER might be identifiable based on a higher mutational burden.

1.3.1.1.2 Transcription coupled nucleotide excision repair

The molecular mechanisms underlying transcription coupled nucleotide excision repair (TC-NER) are extremely similar to the ones of GG-NER (Tornaletti, 2009). The main difference is that TC-NER does not require the XPC-HR23B protein complex used by GG-NER to recognize a DNA lesion. Instead, it is believed that TC-NER is initiated due to stalling of RNA polymerase II (Pol II); such stalling is usually due to the polymerase encountering a damaged DNA base while transcribing a DNA sequence to an RNA sequence. Once Pol II recognizes the damaged DNA, the repair

process continues as previously described for global genome wide nucleotide excision repair (Tornaletti, 2009).

TC-NER repairs DNA damage that is exclusively occurring on the transcribed strand. Thus, when both TC-NER and GG-NER are active, damage occurring on the transcribed strand is more efficiently repaired than damage occurring on the untranscribed strand (Tornaletti, 2009). This has been initially observed in *in vitro* experiments and confirmed in more recent genomic studies. Notably, examining *TP53* mutational patterns from ultraviolet light associated skin cancers and tobacco associated lung cancers revealed the presence of mutational strand-bias (Greenblatt et al., 1994; Hollstein et al., 1999; Hollstein et al., 1991). Furthermore, analyses of the whole cancer genome of a small cell lung carcinoma and the whole cancer genome of malignant melanoma revealed that a mutational strand-bias is present on a genome wide scale (Pleasance et al., 2010a; Pleasance et al., 2010b). Thus, the activity of TC-NER can be evaluated based on the observed strand-bias in the transcribed regions of cancer genomes. Nevertheless, it is plausible that there are other mechanisms (in addition to TC-NER) that protect transcribed genomic regions and, thus, the observed strand-bias might not be exclusively due to the activity of TC-NER.

1.3.1.1.3 Domain associated nucleotide excision repair

The existence of a domain associated nucleotide excision repair (DA-NER) has been inferred based on experimental observations. Most notably, in terminally differentiated human neurons with attenuated GG-NER, it was observed that the DNA damage on the untranscribed strand of genic regions is efficiently repaired (Nouspikel and Hanawalt, 2000). Since there is almost no GG-NER activity in these cells and this type of repair cannot be performed by TC-NER, the existence of a third type of nucleotide excision repair has been proposed. Currently, the molecular mechanisms underlying DA-NER remain unclear. Further, there has been no genome scale mutational analysis associating a mutational pattern with the activity of DA-NER.

1.3.1.2 Base excision repair

Base excision repair (BER) is an evolutionary conserved molecular mechanism responsible for the repair of small lesions that do not distort the structural integrity of the double helix. This repair pathway has been recently extensively reviewed (Robertson et al., 2009; Wilson and Bohr, 2007). These lesions are most

commonly due to: oxidation, alkylation, deamination, depurination, or depyrimidination. In contrast to nucleotide excision repair, BER relies on a plethora of DNA glycosylases that recognize specific types of DNA damage and catalyse their removal (Robertson et al., 2009). The removal of a damaged DNA results in a creation of an abasic site, which subsequently is cleaved by the apurinic/apyrimidinic endonuclease (*APEX1*) thus forming a single-strand break.

In principle, BER can repair single-strand breaks in two distinct pathways (i) short patch base excision repair and (ii) long patch base excision repair. The former is activated most commonly when only a single nucleotide needs to be repaired, while the latter is leveraged when more than one nucleotide (usually between 2 and 10) must be replaced (Robertson et al., 2009). It should be noted that the decision of whether BER leverages short or long patch excision is poorly understood (Hashimoto et al., 2004; Robertson et al., 2009). Short patch and long patch repair will be briefly reviewed in the next two subsections.

Similarly to nucleotide excision repair, a complete failure of base excision repair in a cancer cell (provided that this cell remains viable) will be detectable due to a highly increased mutational burden. It should be noted that as BER is dependent on more than 20 distinct DNA glycosylases (Robertson et al., 2009), a partial failure of BER is also possible when one (or more) of these glycosylases are defective. *In vitro* experiments have demonstrated that a defect in *SMUG1* results in C:G>T:A mutations, while a defect in *OGG1* results in C:G>A:T (Robertson et al., 2009). Nevertheless, currently, there are no known *in vivo* mutational signatures due the failure of BER.

It should be noted that there has been some limited evidence for the existence of transcription coupled base excision repair (TC-BER) in regards to the repair of oxidative DNA damage (Hazra et al., 2007; Izumi et al., 2003). Nevertheless, the existence of TC-BER has not been widely accepted and it will not be reviewed in this thesis.

1.3.1.2.1 Short patch base excision repair

Short patch base excision repair (SP-BER) accounts for almost 90% of the DNA damage repaired by BER. In SP-BER, DNA polymerase β is responsible for catalysing the removal of the 5'-deoxyriboso-phosphate residue (generated by the *APEX1* cleaving) and re-synthesizing the previously removed damaged single

nucleotide (Robertson et al., 2009). Lastly, the residual chromosomal nick is sealed by *XRCC1* in association with either DNA ligase I or DNA ligase III (Robertson et al., 2009).

1.3.1.2.2 Long patch base excision repair

Long patch base excision repair (LP-BER) is generally recruited when more than one nucleotide needs to be repaired and LP-BER accounts for only ~10% of the DNA damage repair by BER (Robertson et al., 2009). After *APEX1* has catalysed the formation of a 5' nick to the abasic site, LP-BER recruits a set of DNA polymerases and ligases to replenish the previously excised nucleotide track (Robertson et al., 2009). In contrast to short patch base excision repair, in LP-BER the synthesis of nucleotides is mediated by DNA polymerases β, δ, and ε and it requires the availability of both *PCNA* and *FEN1* (Robertson et al., 2009; Wilson and Bohr, 2007).

1.3.1.3 DNA mismatch repair

DNA mismatch repair (MMR) is a molecular mechanism leveraged by both prokaryotes and eukaryotes to repair any insertions, deletions, or misincorporations of bases that have arisen during DNA replication or DNA recombination. MMR is a complex process that has been extensively reviewed in recent publications (Jiricny, 2006; Pena-Diaz and Jiricny, 2012). In principle, mismatch repair encompasses two essential tasks: (i) recognition of a mismatch of a DNA base pair and (ii) directing the repair mechanisms towards the newly synthesized strand that carries the erroneous genetic information. In bacteria, distinguishing between the two parental strands and the newly synthesized strand is done via hemimethylation as only the adenine on the parental strands is methylated at 5'-GATC-3' sequences (Jiricny, 2006; Pena-Diaz and Jiricny, 2012). The exact recognition mechanism in eukaryotes is currently unknown.

In bacteria, the *MutS* protein binds to the mismatch while the *MutH* protein binds to the hemimethylated 5'-GATC-3' sequence. The actions of *MutH* are latent until it gets activated upon contact with a *MutL* dimer, which binds the *MutS*-DNA complex (Jiricny, 2006). *MutH* recruits an *UvrD* helicase to separate the two strands and then the entire complex slides along the DNA in the direction of the mismatch.

This liberates the strand that needs to be excised and the molecular complex is followed by an exonuclease that digests the single-stranded DNA. The recruited exonuclease is dependent on whether the nick is on the 3' end of the mismatch or on the 5' end. The result from this process is excision of the mismatch and its surrounding nucleotides. DNA Polymerase III (in combination with a single-strand binding protein and a ligase) is used to repair the single-stranded gap using the remaining strand as a template (Jiricny, 2006). Lastly, a deoxyadenosine methylase is recruited to methylate the nascent strand.

In human beings, the exact molecular mechanisms of mismatch repair are not completely understood. The human *MSH* proteins are heterodimeric orthologs of *MutS*. *MSH2* dimerizes with *MSH6* to form the *MutSα* complex, while *MSH3* dimerizes with *MSH6* to form the *MutSβ* complex (Friedberg and Friedberg, 2006). These two complexes perform function similar to the one of the bacterial complex *MutS*. The functions of the bacterial *MutL* dimer are mimicked by its human orthologs *Mlh1* and *Pms1*, which form a heterodimer. This human heterodimer has three forms – *MutLα* made of *MLH1* and *PMS2*, *MutLβ* made of *MLH1* and *PMS1*, and *MutLγ* made of *MLH1* and *MLH3* – each with its own unique function (Friedberg and Friedberg, 2006). While there are no current known eukaryotic proteins that performed the roles of *MutH* or DNA helicase, recent studies have shown that MMR in eukaryotic organisms requires additional factors, *viz.*, *PCNA* and replication factor C (*RFC*) (Kadyrov et al., 2006).

DNA mismatch repair plays an essential role in reducing the number of replication-associated errors. When MMR is functioning correctly, no specific pattern of somatic mutations has been associated with its activity. However, defects in MMR increase the spontaneous mutation rate and they have been associated with hereditary and sporadic human cancers (Friedberg and Friedberg, 2006). In particular, a large proportion of human colorectal and uterine cancers (termed microsatellite unstable cancers) have been attributed to mutations in *MLH1* and/or *MSH2*. The mutational signature observed in this cancer types is highly reproducible and, in addition to an elevated base substitution mutational burden, contains a high number of small insertions and deletions at mononucleotide or polynucleotide repeats.

1.3.2 Repair of double-strand DNA breaks

Double-strand breaks are probably the most lethal type of DNA damage and even a single double-strand break may result in a cellular death. Three distinct molecular pathways can generally repair double-strand breaks: (i) homologous recombination, (ii) non-homologous end joining, and (ii) microhomology mediated end joining. Repair of DNA double-strand breaks by homologous recombination generally occurs between the late the S phase and the G2 phase of the cell cycle. In contrast, the cell uses non-homologous end joining predominantly during the early S phase and the G0/G1 phases, while microhomology mediated end joining occurs almost exclusively during the synthesis phase of the cell cycle (Friedberg and Friedberg, 2006). The cell attempts to repair a double-strand break as soon as the damage occurs preferentially relying, when possible, on homologous recombination instead of the alternative error-prone pathways (Boulton, 2010; Friedberg and Friedberg, 2006). The molecular mechanisms of the three double-stand repair pathways will be briefly reviewed in the next subsections.

1.3.2.1 Repair of DNA double-strand breaks by homologous recombination

Homologous recombination is the processes of exchanging DNA strands of identical (or extremely similar) nucleotide sequence. This pathway is widely used for accurately repairing the majority of double-strand breaks and interstrand crosslinks (San Filippo et al., 2008). Currently, there are at least four known models of the mechanisms underlying repair of DNA double-strand breaks by homologous recombination: classical double-strand break repair (DSBR), synthesis-dependent strand annealing (SDSA), break-induced replication (BIR) and single-strand annealing (SSA). These four molecular pathways are similar in their initial steps.

After the occurrence of a double-strand break, the *MRN/MRX* complex (*MRN* in human beings*; MRX* in *S. cerevisiae*) binds to the DNA on either side of the break and it performs a variety of functions: checkpoint signalling, tethering the ends of the double-strand break, and cleaving DNA nucleotide links. The actions of the *MRN/MRX* complex are followed by resection, a process in which sections of DNA around the 5' ends on either side of the break are removed by the *Sae2/CtIP* protein. Next, *Sgs1/YMR190C* helicase opens the double-stranded DNA and two nucleases (*Exo1/EXO1* and *Dna2/DNA2KL*) cut the single-stranded DNA produced by *Sgs1/YMR190C*. The formed single-stranded DNA is coated with the *Rad51/RAD51*

recombinase protein, which is dependent on *RPA* and *Rad52/BRCA2* (San Filippo et al., 2008). The final result of this molecular process are 3' single-stranded nucleoprotein filaments that can first search for a homologous DNA template and then can perform an invasion (San Filippo et al., 2008). In mitotic cells, the homologous template is usually a sister chromatid that is mostly identical to the damaged DNA. When a template is found, the invasive 3'end displaces one strand of a homologous duplex called a displacement-loop (D-loop) and pairs with the other to form a heteroduplex. After the strand invasion, a DNA polymerase is recruited to extend the end of the invading 3' strand changing the D-loop in a cross shaped structure commonly known as Holliday junction.

While the steps listed above are mostly shared by the four types of repair of DNA double-strand breaks by homologous recombination (*viz.*, SDSA, DSBR, BIR, and SSA), there are distinct differences between these molecular mechanisms, which are extensively reviewed in (Friedberg and Friedberg, 2006). Briefly, double-strand break repair relies on two-end invasion and it forms double Holliday junctions that may result in both crossover and (albeit rarely) non-crossover products. Due to its propensity to form crossover chromosomal products, DSBR is likely the mechanism that underlies homologous recombination occurring during meiosis (Friedberg and Friedberg, 2006).

Synthesis-dependent strand annealing also relies on two-end invasion, but SDSA produces only non-crossover recombinants. This process occurs in both mitotically and meiotically dividing cells.

Break-induced replication does not require two-end invasion, but it rather relies on the availability of a one-end invasion homologue. Most commonly, a cell undergoing replication makes use of BIR when a double-strand break is encountered by a DNA helicase at a replication fork (Friedberg and Friedberg, 2006). While the precise molecular mechanisms of BIR are still unclear, it is believe that a homologous sequence is invaded by the broken end resulting in the initiation of unidirectional DNA synthesis from the site of strand invasion. The DNA synthesis can lead to replicating up to a few hundred kilobases of the template chromosome and it is followed by repeated cycles of separation, reinvasion, and synthesis until the damaged DNA is repaired.

Single-stranded annealing is a special type of homologous repair that arises when no invasion occurs and it is used to repair breaks between repeat sequences

(Friedberg and Friedberg, 2006). During resection, SSA uncovers direct repeat sequences and repairs the double-strand break by annealing together both single-stranded ends. This type of homologous repair is mutagenic as any sequences that have existed between the two repeat sequences prior to the double-strand break will be lost.

In general, no specific and reproducible mutational signature has been identified for any of the types of DNA double-strand break repair by homologous recombination. Both, DSBR and SDSA are considered "highly faithful" repair pathways and it is unlikely that they result in the generation of any somatic mutations (Friedberg and Friedberg, 2006). In contrast, using yeast models, it was demonstrated that BIR is highly inaccurate but no specific mutational pattern was associated with this repair mechanism (Deem et al., 2011). SSA is potentially the most mutagenic of the four types of DNA double-strand break repair by homologous recombination. However, no specific mutational signature has been attributed to the activity of SSA.

Lastly, it should be noted that complete (or even partial) failure of DNA double-strand break repair by homologous recombination may result in a specific mutational signature as the cell starts predominantly relying on other, more mutagenic, molecular mechanisms for repairing the DNA double-strand breaks. These molecular mechanisms will be discussed in the next few sections.

1.3.2.2 Non-homologous end joining

Non-homologous end joining (NHEJ) repairs DNA double-strand breaks by ligating the two broken ends of the double helix. This molecular pathway does not require a long homologous sequence but rather the DNA repair is guided by short (less than four bases in *S. cerevisiae*) homologous sequences known as microhomologies (Friedberg and Friedberg, 2006). The single-stranded overhangs on the ends of the broken double-stranded DNA often contain these microhomologies. The NHEJ repair pathway is nonmutagenic in the rare cases when the overhangs are ideally matching; however, in the majority of NHEJ repairs, these overhangs are only partially compatible resulting in translocations or micro-insertions/micro-deletions at regions of microhomologies (Friedberg and Friedberg, 2006).

There are three molecular machineries involved in NHEJ: *MRN/MRX* (*MRN* in human beings*; MRX* in *S. cerevisiae*), *DNA-PK/Ku*, and *Ligase IV/ Lig4* complexes. Shortly after the double-strand break formation, the *MRN/MRX* and *DNA-PK/Ku*

complexes bind DNA to inhibit degradation by bridging and tethering the two broken ends. The *MRN/MRX* complex recruits the DNA ligases *Ligase IV/ Lig4*, while the *DNA-PK/Ku* is believe to stabilize DNA preventing repair based on homologous recombination (Friedberg and Friedberg, 2006). The *Ligase IV/ Lig4* complex facilitates the joining of the broken DNA strands. It should be noted that there is an intricate interaction between *Ligase IV/ Lig4* and *DNA-PK/Ku* providing NHEJ with significant flexibility that allows mismatch correction, gap-filling or removal of non-ligatable ends (Friedberg and Friedberg, 2006).

The activity of non-homologous end joining is associated with a specific pattern of somatic mutations: translocations and/or indels at regions of (or near) microhomologies (Friedberg and Friedberg, 2006). This mutational signature is thought to be especially prominent in samples where the molecular mechanisms of DNA double-strand break repair by homologous recombination have failed and the majority of double-strand breaks are repaired by NHEJ.

1.3.2.3 Microhomology mediated end joining

Microhomology mediated end joining (MMEJ) repairs a double-strand DNA break by relying on microhomologies with lengths between 5 and 20 nucleotides. The molecular mechanisms behind MMEJ are not precisely known but it is believed to reply to some extent on factors implicated both in repair based on homologous recombination (*viz*., *MRN/MRX*, *Rad51/RAD51*, and *Rad52/BRCA2*) as well as non-homologous end joining (*viz*., *MRN/MRX*, *DNA-PK/Ku*, and *Ligase IV/ Lig4*) (Friedberg and Friedberg, 2006). There is no known mutational signature associated with the activity of microhomology mediated end joining; however, it is foreseeable that the pattern of mutations generated by this error-prone repair process is very similar to the one of non-homologous end-joining, albeit with potentially longer microhomologous sequences near indels and/or translocations.

1.4 Mutational processes and patterns of somatic mutations

In the previous sections, I provided a literature review of the DNA damaging and repair processes. Here, I will review the known patterns of somatic mutations derived from examining cancer samples and put them in perspective of these damaging and repair processes.

As previously discussed, early studies have demonstrated that exposure to ultraviolet (UV) light can lead to the formation of dimers of any two adjacent pyrimidine bases on the same DNA strand with a preference for thymine-thymine dimers (Witkin, 1969). It was further shown that UV irradiation damage predominantly results in cytosine to thymine or cytosine-cytosine to thymine-thymine changes, preferentially occurring at these pyrimidine dimers (*i.e.*, C>T or CC>TT DNA mutations at dipyrimidine sites) (Howard and Tessman, 1964; Setlow and Carrier, 1966). This was the first detailed *in vitro* characterization of the pattern of DNA changes occurring due to the activity of an exogenous mutagen and, as such, the very first description of a signature of a mutational process.

While these early examinations established the mutational signature of UV light, it was unclear whether UV induced mutations are present and involved in the neoplastic expansion of human cancers. The development of the DNA sequencing technique with chain-terminating inhibitors by Fred Sanger (Sanger et al., 1977) allowed rapid examination of the genetic material contained in cancer cells. In the early 1990s, two studies sequenced exons of the gene *TP53* (Brash et al., 1991; Ozturk, 1991; Bressac et al., 1991) from several patients and provided experimental evidence that aflatoxin and UV light leave distinct patterns (consistent with the ones observed in experimental systems) of DNA mutations respectively in hepatocellular and squamous-cell carcinomas. These studies confirmed that the mutational signatures of carcinogens are left as "evidence" in the genomes of cancer cells (Vogelstein and Kinzler, 1992) thus spawning research which first examined the mutations across *TP53* and later across multiple genes and even whole cancer genomes in order to provide a better understanding of the mutational processes involved in human carcinogenesis. In the next few sections, I summarize the current knowledge of the patterns of somatic mutations identified in human cancer.

1.4.1 Patterns of somatic mutations in TP53

Multiple independent studies used Sanger sequencing of some (or all) exons of a cancer gene to provide clues to the etiology of both endogenous and exogenous factors of human carcinogenesis. *TP53* was usually selected for this analysis due to its relatively small size of only 11 exons, high conservation in vertebrates, and its high prevalence of somatic mutations in almost all tumour classes (Greenblatt et al., 1994).

Further, the observed *TP53* mutations are predominantly missense thus subject to less restricted sets of mutated bases and sequence contents when compared to nonsense mutations. Commonly, each of these studies involved multiple samples of a cancer type that were examined for somatic mutations in *TP53,* studies reviewed in refs (Greenblatt et al., 1994; Hollstein et al., 1999; Hollstein et al., 1991). The *TP53* somatic mutations were aggregated, their spectrum was reported as specific for the given cancer type, and this spectrum was then compared to mutations generated experimentally in *in vitro* or *in vivo* systems (Greenblatt et al., 1994; Hollstein et al., 1999). It should be noted that the mutational spectra of other genes, albeit only occasionally, were also used for such analysis (Capella et al., 1991).

These early studies revealed a significant heterogeneity of the *TP53* spectra across different cancer types, which allowed associating some patterns of mutation to known carcinogens. Here, I provide a concise summary of some of the more important findings while details could be found in refs (Greenblatt et al., 1994; Hollstein et al., 1999; Hollstein et al., 1991). The *TP53* spectrum of skin carcinomas exhibited C>T and CC>TT mutations at dipyrimidines with a strong transcriptional strand-bias (all substitutions and dinucleotide substitutions are referred to by the pyrimidine(s) of the mutated Watson-Crick base pair). This was consistent with the *in vitro* described mutational signature of UV light. The *TP53* mutational spectrum derived from lung cancers in tobacco smokers was overwhelmed by C>A substitutions with a strong transcriptional strand-bias, which coincided with the class of mutation produced experimentally as a result of bulky adduct formation by tobacco carcinogens on guanine (Rodin and Rodin, 2005). In other tobacco associated cancers, such as oesophageal and head and neck tumours, C>A mutations (while still ubiquitous) were less common while there was a significant increase of T>C mutations. Interestingly, in both smokers and non-smokers, C>T and C>G mutations at non-CpG sites were elevated when compared to all other cancer types, with bladder tumours harbouring the most C>G mutations (Greenblatt et al., 1994). Additionally, it was demonstrated that C>A transversions were common in hepatocellular cancers and these mutations were believed to be associated with aflatoxin, a known carcinogen commonly found in food from southern Africa and Asia (Wogan, 1992). Lastly, all cancer types harboured at least some C>T mutations at CpG dinucleotides, a process

attributed to the normal cellular event of deamination of 5-methylcytosine (Greenblatt et al., 1994).

The analyses of *TP53* spectra were the first attempts to bridge the gap between molecular cancer genetics and epidemiology (Hainaut et al., 2001). The large number of studies examining *TP53* spectra required a computational resource to facilitate and retrieve the already identified somatic mutations. At first these data were managed by the researchers that were generating it but in 1994 the International Agency for Research on Cancer stepped in and started to maintain a database while providing a free access to it (Hainaut et al., 2001). The first release of the IARC *TP53* database contained ~3,000 somatic mutations while the most recent version (R17) released in November of 2013, which can be found at http://p53.iarc.fr/, contains over 28,000 somatic mutations in *TP53*.

Though extremely informative, the data gathered from single gene studies have significant limitations. In these studies, the spectrum of a cancer type is reported by aggregating mutations from multiple samples. This may be adequate when a single mutational process generates the majority of mutations in the particular cancer (*e.g.*, UV light is the predominant mutational process in melanoma (Alexandrov et al., 2013a)). However, usually multiple mutational processes are operative in a single cancer sample, and combining their mutations generates a mixed composition of the patterns of somatic mutations. In most cases, reporting this jumbled spectrum is uninformative for the diversity of the mutational processes operative in a single cancer type or even in a single cancer sample (Alexandrov et al., 2013a). Moreover, the examined *TP53* exons are both under selection and also have a specific nucleotide sequence. This affects the opportunity for observing a somatic mutation and as such, in addition to the processes of mutation, the reported spectrum can be a reflection of the processes of selection and/or the nucleotide architecture of the *TP53* gene (Stratton, 2011; Stratton et al., 2009).

Two studies tried to overcome some of the single gene limitations by leveraging a targeted capillary sequencing approach of large number of genes. A survey of the 518 protein kinase genes in 25 human breast cancer samples revealed 92 somatic mutations (90 substitutions and 2 indels) in which C>T transitions and C>G transversions preceded by thymine (*i.e.*, C>T and C>G at TpC) occurred with a higher

than expected frequency (Stephens et al., 2005). This survey was later expanded to 210 cancer samples and it revealed more than 1,000 somatic mutations with significant variations in their patterns across the examined twelve cancer types (Greenman et al., 2007). Only a small fraction of the mutations reported in these screens are likely to be affected by selection (Rubin and Green, 2009), thus indicating that the observed mutational patterns reflect the operative mutational processes in the analysed samples and not the processes of negative or positive selection.

1.4.2 Mutational patterns identified in next generation sequencing data

The development of second-generation sequencing technologies allowed examination of cancer exomes (*i.e.*, the combined protein coding exons) and even whole cancer genomes. Sequencing cancer exomes has been generally preferred as the majority of known cancer-causing driver somatic substitutions, indels, and copy number changes (although generally not rearrangements) (Stratton, 2011) are located in protein coding genes. As the nucleotide sequence of protein coding genes is ~1% of the whole genome, analysis of exomes is considered an advantageous and cost effective methodology for discovering the genes involved in neoplastic development. As a result, many studies have focused predominantly on the generation and analysis of exome sequences (Hudson et al., 2010).

Early next generation sequencing studies started revealing patterns of somatic substitutions in different cancer types. In 2010, two back-to-back studies in *Nature* reported the patterns of somatic mutations in a malignant melanoma (Pleasance et al., 2010a) and a small cell lung carcinoma (Pleasance et al., 2010b). As expected, a strong signature of tobacco carcinogens was found in the genome of the lung cancer, while the mutational signature of ultraviolet light overwhelmed the melanoma genome. These studies demonstrated the value of whole genome sequencing for evaluating signatures of mutational processes by providing greater resolution and mechanistic insight into mutational signatures due to known carcinogens, for example through the identification of a lower prevalence of mutations over the footprints of genes.

Multiple independent studies and international consortiums started sequencing large numbers of samples from both cancer genomes and exomes (Hudson et al.,

2010). An integrated genomic characterization was reported for many different cancer types including: acute lymphoblast leukaemia (De Keersmaecker et al., 2013; Holmfeldt et al., 2013; Zhang et al., 2012), acute myeloid leukaemia (Govindan et al., 2012), breast cancer (Nik-Zainal et al., 2012; Shah et al., 2012; Stephens et al., 2012), chronic lymphocytic leukaemia (Puente et al., 2011; Quesada et al., 2012), colorectal cancer (Cancer Genome Atlas, 2012; Seshagiri et al., 2012), oesophageal cancer (Dulak et al., 2013), glioblastoma (Parsons et al., 2008), cancers of the head and neck (Agrawal et al., 2011; Stransky et al., 2011), kidney cancer (Cancer Genome Atlas, 2013; Guo et al., 2012; Pena-Llopis et al., 2012), liver cancer (Fujimoto et al., 2012; Kan et al., 2013), lung cancer (Ding et al., 2008; Govindan et al., 2012; Imielinski et al., 2012; Peifer et al., 2012; Rudin et al., 2012; Seo et al., 2012), lymphomas (Love et al., 2012; Morin et al., 2011), melanoma (Berger et al., 2012; Hodis et al., 2012; Huang et al., 2013; Stark et al., 2012), multiple myeloma (Chapman et al., 2011a), ovarian cancer (Jones et al., 2010a), pancreatic cancer (Jiao et al., 2011; Wu et al., 2011), prostate cancer (Baca et al., 2013; Barbieri et al., 2012; Berger et al., 2011; Grasso et al., 2012), stomach cancer (Nagarajan et al., 2012; Wang et al., 2011; Zang et al., 2012), uterine cancer (Cancer Genome Atlas, 2013), and several different types of paediatric tumours (Jones et al., 2012a; Pugh et al., 2013; Pugh et al., 2012; Rausch et al., 2012; Robinson et al., 2012; Sausen et al., 2013; Zhang et al., 2013). While these studies focused on the identification of novel cancer genes, mutational spectra were usually reported for each of the examined samples and some studies even tried to associate certain types of somatic mutations with the activity of mutagens or the failure of polymerases and/or DNA repair mechanisms. A brief summary of the mutational patterns identified in these cancer genomics studies is provided in the next paragraph.

In lung cancer, comparison between tobacco smokers and non-smokers revealed that smokers have on average 10-fold increase in the burden of somatic mutations in their cancer genomes (Govindan et al., 2012; Imielinski et al., 2012). Consistent with the experimental evidence for tobacco carcinogens, this elevation is mainly due to the increase of the number of C>A transversions (Rodin and Rodin, 2005). Examination of the cancer genomes of melanomas confirmed that the majority of mutations are C>T and CC>TT at dipyrimidines in the ultraviolet-associated tumours, while acral melanomas exhibit predominantly C>T transitions at CpG sites

(Berger et al., 2012; Hodis et al., 2012). In glioblastoma multiforme, it was demonstrated that treatment with an alkylating agent, such as temozolomide, significantly elevates the numbers of somatic mutations and results in a distinct mutational pattern of C>T transitions (Parsons et al., 2008). In chronic lymphocytic leukaemia, it was observed that samples with mutations in the immunoglobulin genes have a higher proportion of T>G transversions (Puente et al., 2011). This mutational pattern and its immediate sequencing context are consistent with the activity of the error-prone polymerase η during somatic hypermutation (Puente et al., 2011; Spencer and Dunn-Walters, 2005). In endometrial and colorectal tumours, a set of ultra-hypermutators with increased mutational frequency of transversions was associated with somatic mutations in polymerase ε (Cancer Genome Atlas, 2012; Cancer Genome Atlas, 2013). Microsatellite unstable gastric cancer were observed to have a higher mutation prevalence of both C>T transitions and C>A transversions (Nagarajan et al., 2012). Examining the cancer exomes of patients with urothelial carcinoma (of the upper urinary tract) revealed a large number of somatic mutations with an unique pattern of T>A transversions predominantly located at CpTpG sites and possessing a very strong transcription strand-bias (Hoang et al., 2013; Poon et al., 2013). This pattern of mutations was associated with exposure to aristolochic acid. In oesophageal cancer, a high prevalence of T>G transversions was observed (Dulak et al., 2013) while certain breast cancer genomes were found to be overwhelmed with C>T and C>G mutations at TpC sites (Stephens et al., 2012).

These next generation sequencing studies provided an unbiased look into the patterns of DNA changes across cancer genomes. While they resolved some of the previous limitations from *TP53* studies (mostly by examining large portions of the human genome which are usually not under selection and which have a nucleotide context that is representative of the whole human genome) they still did not address the important issue of disentangling mixtures of mutations generated by different mutational processes.

1.5 Summary

In this chapter, I have provided a literature review encompassing cancer genetics, DNA damaging and mutational processes, DNA repair processes, and the patterns of somatic mutations observed in cancer genomes. In the next few chapters, I

will use the reviewed information to first introduce a theoretical model describing the activity of a set of mutational processes operative in cancer genomes as well as to develop a computational approach that can extract the signatures of these mutational processes from mutational catalogues of cancer genomes. The approach will be extensively evaluated with simulated data and, in the first instance, will be applied to genome and exome sequences from breast cancer. Further, I will perform a global analysis of mutational signatures across human cancer using the majority of common cancer classes and samples from more than seven thousand cancer patients. Lastly, using statistical analysis, I will propose etiology for some of the identified mutational signatures and discuss the implications of the performed analysis in the context of cancer research and cancer treatment.

Chapter 2

Deciphering signatures of mutational processes from mutational catalogues of cancer genomes

2.1 Introduction

The first chapter of this thesis defined *somatic mutations* as any change of DNA that is present in the genome of a somatic cell and has occurred after conception. Building upon this well-known definition, the chapter introduced several important concepts. A *somatic mutational process* was defined as a mixture of DNA damaging and repair mechanisms that act collectively and have the ability to cause mutations in somatic cells. A *mutational signature* was described as a characteristic pattern of somatic mutations exhibited by an operative mutational process in a genome of a cell. Lastly, a *mutational catalogue* of a cancer genome was defined as the conglomeration of all detected somatic mutations.

The main focus of the present chapter is to mathematically connect these biological terms and provide both the theoretical model and computational approach for examining and deciphering mutational signatures from sets of mutational catalogues of cancer genomes. The approach is evaluated extensively with simulated data, demonstrating that the developed computational framework is robust to a large range of different parameters and can be applied to both genome and exome sequences.

2.2 Theoretical model of mutational processes operative in cancer genomes

The mutational catalogue of a cancer genome is the cumulative result of all somatic mutational mechanisms, including DNA damage and repair processes, which have been operative during the cellular lineage of the cancer cell. Since the cellular lineage of the cancer cell can be traced back to the zygote, the mutational catalogue reflects the activity of all processes operative from the very first division of the fertilized egg (Stratton, 2011). The large majority of mutations in cancer genomes are believed to be passengers, and by definition their patterns are largely unmodified by selection (Rubin and Green, 2009). Thus, the mutational catalogue derived from a cancer cell may be treated as a representative archaeological record bearing the combined imprints (or signatures) of the mutational processes that have been operative.

2.2.1 Alphabets of mutation types

A mutational catalogue can include a diverse set of mutation classes including base substitutions, insertions/deletions, structural rearrangements and copy number changes. Each class of mutation can then be further subclassified. For example, base substitutions can be subclassified according to the six types of single base substitutions (using the pyrimidine of the Watson-Crick base pair as the reference, C>T, C>A, C>G, T>A, T>C, T>G) or the classification can be further elaborated to include a variety of mutational features such as the sequence context of the mutated base and the transcriptional strand on which the substitution has arisen.

For the purpose of mathematical modelling, a limited number of features of a mutational catalogue need to be selected. The choice of features may be influenced by prior biological knowledge. The choice is also often constrained by statistical considerations and the available data. Mathematically, a set of mutational features can be expressed as a finite alphabet Ξ with K letters, where each letter corresponds to a mutation feature. The simplest alphabet in this case, Ξ_6 contains $K = 6$ letters, and is based on the 6 types of single base substitution. The letters of this Ξ_6 alphabet are C>A, C>T, C>G, T>A, T>C, and T>G. It should be noted that this alphabet of mutation types could be easily extended by, for example, including other mutation types such as double substitutions.

In this thesis, mutational catalogues as well as the mutational signatures that contribute to these catalogues are examined predominantly using five distinct alphabets termed Ξ_6, Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536}. These five alphabets are discussed in further detail below as well as in Appendix I.

The Ξ_6 alphabet is perhaps the simplest possible alphabet as it considers only the six types of somatic substitutions. This alphabet will not be used in any analysis but, rather, its simplicity will be leveraged to provide examples and visual representations clarifying the developed mathematical model and computational approach.

The Ξ_{96} alphabet provides greater resolution for examining the six types of single nucleotide variants (*i.e.,* the Ξ_6 alphabet) by including the immediate sequence context of each mutated base. In this alphabet, a mutation type contains a somatic substitution and both the 5' and 3' base next to the somatic mutation. For example, a C>T mutation can be characterized as …TpCpG…>…TpTpG… (mutated base underlined and presented as the pyrimidine partner of the mutated base pair) generating 96 possible mutation types – (6 types of substitutions) * (4 types of 5' bases) * (4 types of 3' bases).

The Ξ_{1536} further extends Ξ_{96} by including two bases 5' and 3' to the mutated base resulting in 1,536 possible mutated pentanucleotides - (6 types of substitutions) * (16 types of the two immediate 5' bases) * (16 types of the two immediate 3' bases). For example, using the Ξ_{1536} alphabet, one of the 256 subclasses of a C>T mutation is …ApTpCpGpC… > …ApTpTpGpC…

The Ξ_{99} alphabet extends Ξ_{96} by including three additional mutation types, *viz.,* (i) double nucleotide substitutions, (ii) small insertions or deletions at short tandem repeats, and (iii) small insertions or deletions overlapping with microhomologies at breakpoints.

Lastly, Ξ_{192} elaborates Ξ_{96} by considering the transcriptional strand on which a substitution resides. In contrast to all previously discussed alphabets, Ξ_{192} is defined only in the regions of the genome where transcription occurs, which in these analyses has been limited to the genomic footprints of protein coding genes. Thus, the previously defined 96 substitution types are extended to 192 mutation types. For

example, the C>T mutations at TpCpA are split into two categories: the C>T mutations at TpCpA occurring on the untranscribed strand of a gene and the C>T mutations at TpCpA occurring on the transcribed strand. In general, one would expect that these two numbers are approximately the same unless the mutational processes are influenced by the activity of the transcriptional machinery. This could happen, for example, due to recruitment of the transcription-coupled component of nucleotide excision repair (NER). For example, if a mutational process has a higher number of C>A substitutions on the transcribed strand compared to C>A substitutions on the untranscribed strand (note that a C>A mutation on the untranscribed strand is the same as a G>T mutation on the transcribed strand), this could indicate that the mutations caused by this process are being repaired by NER, although other explanations are not excluded. A known example of such strand-bias due to interplay between a mutational process and a repair mechanism is the formation of photodimers due to ultraviolet light exposure that are repaired by NER resulting in a higher number of C>T mutations on the untranscribed strand (van Zeeland et al., 2005).

2.2.2 Mathematical definition of a signature of a mutational process

A signature of a mutational process is mathematically defined in the context of a pre-selected mutational alphabet. A mutational signature is defined as a discrete probability density function with a domain of mutation features based on a pre-selected alphabet Ξ, $P: \Xi \rightarrow \mathbb{R}_{+}^{K}$. Thus, by definition, a mutational signature P is a lexicographically ordered k-tuple; $P = [p^1, p^2, ..., p^K]^{\mathrm{T}}$, where p^i is the probability of process P to cause the mutation feature corresponding to the i-th letter of the pre-selected alphabet Ξ, and since p^i are probabilities:

$$\sum_{i=1}^{K} p^i = 1 \text{ and } p^i \geq 0, i = 1 \dots K \tag{2.1}$$

Examples of four mutational signatures defined over Ξ_6 and two mutational signatures defined over Ξ_{96} are given respectively in panels A and B of Figure 2.1. In the four examples of mutational signatures defined over Ξ_6, the mutational probability for each alphabet letter is displayed. For example, it can be seen that 35% of the mutations attributed to Signature 1 are C>G while only 3% of the mutations are T>G.

Further, while Signatures 1 through 4 are defined over Ξ_6, Signature 2 does not generate any C>T, T>A, and T>C mutations as the probability for each of these mutation types is equal to zero. Signatures 1 and 4 are defined both over Ξ_6 and Ξ_{96} to illustrate that, while a mutation type based on a given alphabet can be similar in two signatures (*e.g.*, C>A mutations are respectively 12% and 14% in these two signatures). Extending this alphabet may reveal an intrinsic internal structure making these mutation types significantly different.

Geometrically, a mutational signature can be examined as a vector in a K dimensional space. Since a mutational signature is modelled as a discrete probability density function defined over a given alphabet (see equation 2.1), all its components

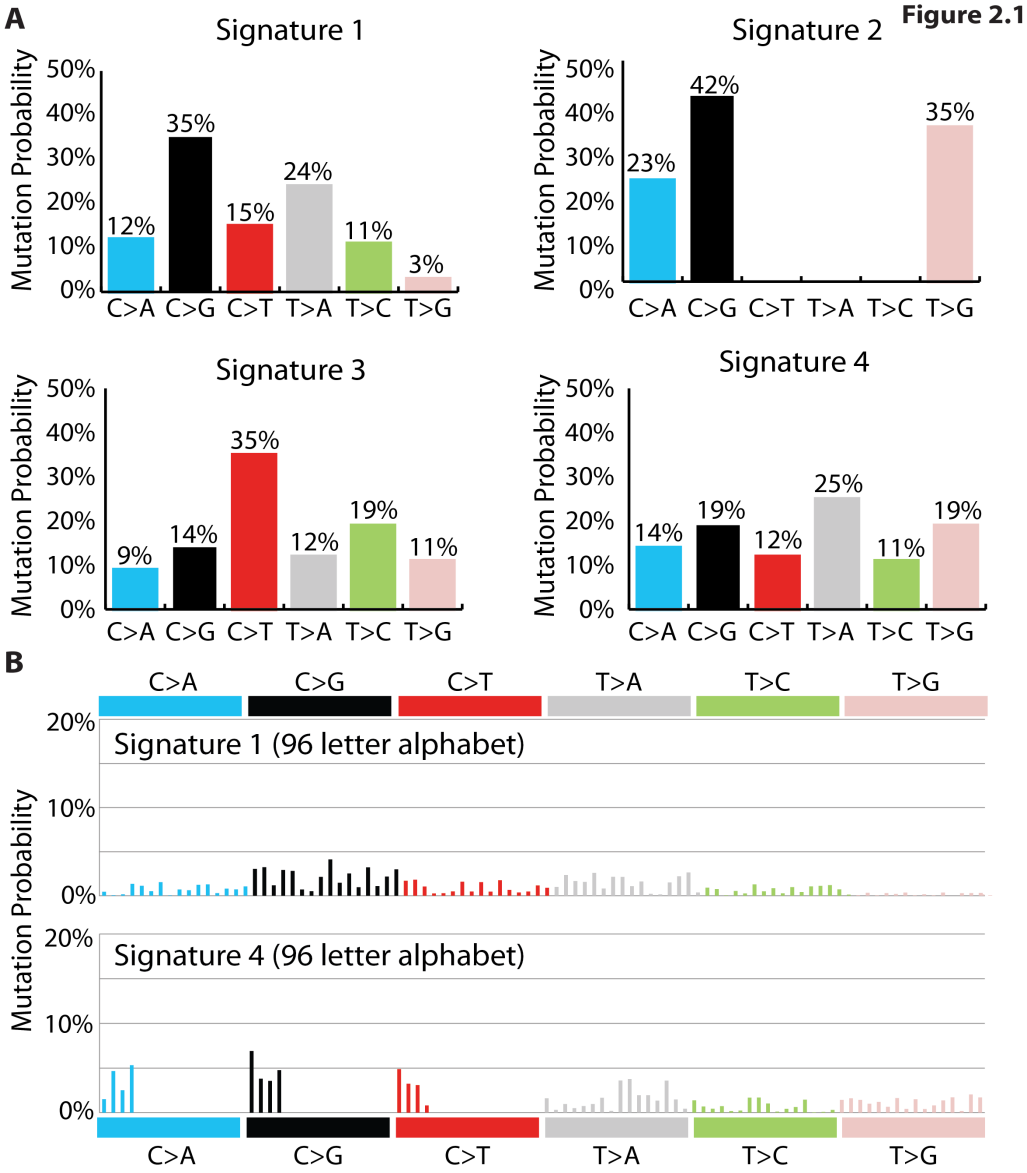


Figure 2.1

Figure 2.1: Simulated examples of mutational signatures defined over different mutational alphabets. (A) Four mutational signatures defined over Ξ_6 and **(B)** two mutational signatures defined over Ξ_{96}.

are nonnegative and this vector belongs to the first hyperoctant of this K dimensional space, $\mathbb{R}_{+}^{K}$. Further, as the sum of the vector components equals one, this vector is constrained by *K-1* dimensional hyperplane. Examining two mutational signatures as vectors in a high dimensional space allows a convenient way for comparing these signatures based on the angle between the vectors. Thus, comparison between two mutational signatures $\vec{A}$ and $\vec{B}$, each defined over an alphabet Ξ with K mutation types, is done using a cosine similarity:

$$sim(\vec{A},\vec{B}) = \frac{\sum_{i=1}^{K} \vec{A}_i \, \vec{B}_i}{\sqrt{\sum_{q=1}^{K}(\vec{A}_q)^2}\sqrt{\sum_{j=1}^{K}(\vec{B}_j)^2}} \tag{2.2}$$

Since the elements of $\vec{A}$ and $\vec{B}$ are nonnegative, the cosine similarity has a range between 0 and 1. When a cosine similarity between two signatures is 1, these signatures are exactly the same. In contrast, when the similarity is 0, the mutation types of these signatures are completely independent. The cosine similarity is a commutative function as $sim(\vec{A},\vec{B}) = sim(\vec{B},\vec{A})$. Two signatures should be compared only if they are defined over the same mutational alphabet. For example, one cannot compare a signature defined over Ξ_6 with a signature defined over Ξ_{192}. Lastly, one can also define a cosine distance between two mutational signatures as $dist(\vec{A},\vec{B}) = 1 - sim(\vec{A},\vec{B})$.

	Signature 1	Signature 2	Signature 3	Signature 4
Signature 1	1.00	0.65	0.75	0.88
Signature 2	0.65	1.00	0.43	0.71
Signature 3	0.75	0.43	1.00	0.78
Signature 4	0.88	0.71	0.78	1.00

Table 2.1: Similarities between simulated mutational signatures. The values of the cosine similarities between the signatures displayed in panel A of Figure 2.1 are shown in this table.

Table 2.1 contains the similarities between the simulated mutational processes displayed in Figure 2.1A. The two signatures that are most similar are Signatures 1 and Signature 4 with a cosine similarity of 0.88 while the signatures that are most different are Signatures 2 and 3 with a similarity of only 0.43. As expected, the similarity of Signatures 1 and Signature 4 is not the same when the signatures are defined and compared over different mutational alphabets. While Signatures 1 and Signature 4 have a similarity of 0.88 when defined over Ξ_6, they have a similarity of only 0.53 when defined over Ξ_{96}. As previously mentioned, this is due to the existence of an internal structure. In this simulated example, all C>X mutations

belonging to Signature 4 are in ApCpN sequence context while Signature 1 has no specific sequence context (Figure 2.1).

2.2.3 Mathematical definition of a mutational catalogue of a cancer genome

Quantitatively, a mutational catalogue of a cancer genome is a vector, m, containing the number of somatic mutations of a genome, g, defined over a finite alphabet of mutation types Ξ. Mathematically, a mutational catalog is a morphism between the pre-defined finite alphabet, Ξ, and a set of K nonnegative integers, $\mathbb{N}_0^K$, *i.e.*, $m: \Xi \rightarrow \mathbb{N}_0^K$. Thus, for a given genome, its mutational catalogue can be expressed as a K-tuple of natural numbers, $m = [m^1, m^2, \ldots, m^K]^{\mathrm{T}}$. A simulated example of a cancer genome defined over the mutational alphabet Ξ_6 is provided in Figure 2.2. This cancer genome has a total of 3,315 somatic substitutions and does not have any specific mutational features.

Figure 2.2

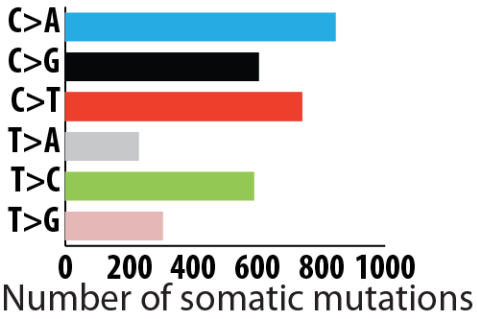


Figure 2.2: Simulated example of a mutational catalogue of cancer genome. The mutational catalogue is defined over the Ξ_6 alphabet.

Comparing the mutational catalogues of two cancer genomes, $m_1 = [m_1^1, m_1^2, \ldots m_1^K]^{\mathrm{T}}$ and $m_2 = [m_2^1, m_2^2, \ldots m_2^K]^{\mathrm{T}}$, requires that both mutational catalogues are defined over the same mutational alphabet Ξ. The similarity of two mutational catalogues can be evaluated in two distinct ways. The first comparison is based on Euclidean distance and examines whether mutational catalogues m_1 and m_2 are exactly the same:

$$dist(m_1, m_2) = \sqrt{\sum_{i=1}^{K} (m_1^i - m_2^K)^2} \tag{2.3}$$

With this comparison, a distance of zero is equivalent to the two mutational catalogues being exactly the same. Further, the larger the distance the more different the mutational catalogues.

While two mutational catalogues can have different numbers of somatic mutations (and therefore a large Euclidean distance between them) they can have exactly the same patterns of somatic mutations. Thus, a correlation distance is used to compare whether the patterns of mutations of two mutational catalogues are similar. The simplest correlation distance is based on the Pearson product-moment correlation coefficient. However, this correlation coefficient is very sensitive to outliers and it might be misleading if a small subset of mutation types have significantly larger values when compared to the rest of the mutation types (Abdullah, 1990). More robust measurements of correlations are Spearman's rank correlation coefficient and Kendal's rank correlation coefficient (Croux and Dehon, 2010). These two rank correlations usually produce very similar results and rarely is there a reason to choose one over the other (Croux and Dehon, 2010). In this work, I make use of Spearman's correlation coefficient to compare the patterns of mutations in two mutational catalogues. Spearman's correlation is defined as the Pearson's correlation coefficient between the ranked variables. Thus, the patterns of mutations in mutational catalogues m_1 and m_2 can be compared by the formula:

$$\rho = \frac{\sum_{i=1}^{K}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{K}(x_i - \bar{x})^2 \sum_{i=1}^{K}(y_i - \bar{y}\,)^2}} \tag{2.4}$$

where x_i is the rank of the i-th letter of the pre-selected alphabet Ξ in m_1, y_i is the rank of the i-th letter in m_2, $\bar{x}$ is the mean of $x_i, i = 1 \ldots K$, and $\bar{y}$ is the mean of $y_i, i = 1 \ldots K$.

In general, the Euclidean distance will be used to compare two mutational catalogues when one wants them to be as similar as possible. For example, a Euclidean distance will be used when extracting mutational signatures and evaluating the accuracy of the extraction. In contrast, the correlation distance will be used to compare the similarity of the patterns of somatic mutations between two mutational catalogues. For example, a correlation distance will be used when performing clustering of cancer genomes in order to identify distinct groups of mutational patterns.

2.2.4 Modelling mutational processes operative in a cancer genome

In the previous sections of this chapter, I provided mathematical definitions for mutation types, mutational signatures, and mutational catalogues. In this section, I make use of these definitions to provide a linear model of mutational processes operative in cancer genomes.

Different cancer genomes can be exposed to a particular mutational process at different intensities. For example, a mutational process could cause 1,000 mutations in one cancer genome while causing 20,000 in another. I will refer to this number of mutations as a *mutational exposure* (or simply *exposure*) of a signature of a mutational process in a cancer genome. Hence, one may say that a mutational process with a signature P has an exposure e, corresponding to the number of mutations caused by this process, in a mutational catalogue m of a given cancer genome.

Multiple mutational processes can be operative in a single cancer genome (Stratton, 2011) and each of these processes can have a distinct mutational exposure. In this section, I model a cancer somatic mutational catalogue as a linear combination of the signatures and intensities of the exposure of the mutational processes active at some point in the lineage of cells leading to the cancer cell, plus added noise vector accounting for non-systematic sequencing or analysis errors. Thus, the mutational catalogue of a cancer genome $m = [m^1, m^2, \dots m^K]^{\mathrm{T}}$, defined over the mutation alphabet Ξ with K letters, is a superposition of the signatures of the N operative mutational processes $P_i = [p_i^1, p_i^2, \dots p_i^K]^{\mathrm{T}}, i = 1 \dots N$, each defined over the mutation alphabet Ξ, with their respective exposures $e^i, i = 1 \dots N$, and non-systematic noise n. In particular, the number of the j-the mutation type in m is:

$$m^j = \sum_{i=1}^{N} P_i^j e^i + n^j \tag{2.5}$$

Note that in this definition, m, e, and n are vectors, while P is expressed as a matrix. Indeed, a set of signatures of N mutational processes, can be represented by a nonnegative matrix $P = \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix}$ with size $K \times N$, where K is the

number of mutation types and N is the number of signatures. The subscript index indicates the signature, while the superscript index corresponds to the mutation type.

Figure 2.3

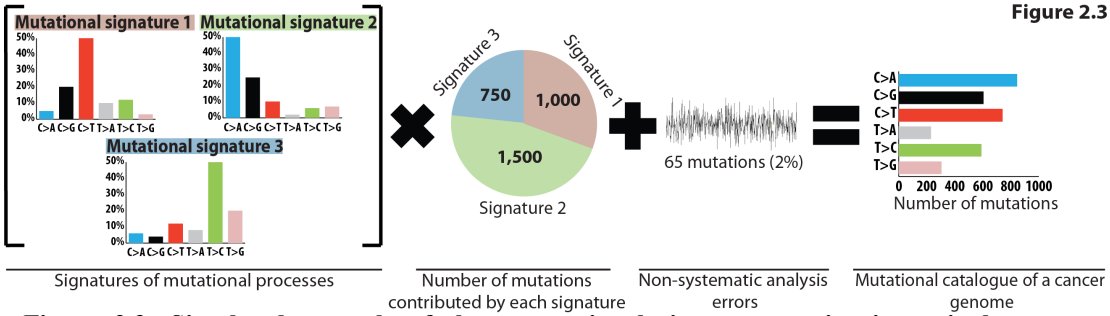


Figure 2.3: Simulated example of three mutational signatures active in a single cancer genome. The three mutational signatures were defined over the Ξ_6 alphabet. The mutational catalogue of the cancer genome is modelled as a linear superposition of the signatures of three processes and the respective number of mutations contributed by each signature, plus added non-

A simulated example illustrating this model is provided in Figure 2.3. Each of the signatures has a specific pattern over the six base substitutions. The first signature has a substantial proportion of C>T mutations and contributes, in total, 1,000 mutations to the cancer genome. The second process has a high proportion of C>A mutations while contributing 1,500 mutations. The third process generates substantial numbers of T>C mutations and contributes 750 mutations (Figure 2.3). The mutational catalogue of the cancer genome formed by these three processes, however, does not have any notable or specific features and does not obviously resemble any of the mutational signatures that are operative in it. This simulated mutational catalogue contains, in total, 3,315 mutations, 3,250 (~98%) contributed by the three mutational processes and the remaining 65 (~2%) by white noise corresponding to minor processes or experimental errors in generating the mutation catalogue of the genome.

2.3 Deciphering mutational signatures from a set of cancer genomes

In the previous section of this chapter, I described a mutational catalogue of a cancer genome as a linear combination of the signatures of the underlying mutational processes active in this cancer genome. A single mutational catalogue does not allow identification of the operative mutational signatures since there are many ways to decompose a single mutational catalogue into multiple mutational signatures. However, the availability of hundreds and even thousands of mutational catalogues of cancer genomes can address this limitation, as mutational signatures will have

different exposures in different catalogues, constraining the number of solutions and thus allowing deconvolution of the signatures.

In summary, the approach developed here is used to identify the signatures of mutational processes from a large number of mutational catalogues. In order to do this, I will start by introducing matrix notations for mutational signatures, mutational catalogues, exposures of mutational signatures, and noise terms. These matrix notations are necessary to alleviate and shorten the description of the developed algorithm for deciphering mutational signatures.

2.3.1 Matrix notations for deciphering mutational signatures

The signature P_1 of a mutational process, defined over an alphabet Ξ with K letters, can be expressed as a nonnegative K-tuple, $P_1 = [p_1^1, p_1^2, \ldots p_1^K]^{\mathrm{T}}$, where $\sum_{i=1}^{K} p_1^i = 1$ and p_1^i is the probability of the mutational processes P_1 to cause the mutation type corresponding to the i-th letter of the alphabet Ξ. As previously described, a set of N mutational signatures can be expressed as a nonnegative mutational signature matrix $P = \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix}$ with size $K \times N$, where K is the number of mutation types and N is the number of signatures. The subscript index indicates the signature, while the superscript index corresponds to the mutation type.

The mutational catalogue of a cancer genome, defined over the alphabet of mutation types Ξ, is represented by a morphism m, where $m: \Xi \to \mathbb{N}_0^K$. For a given genome, i, its mutational catalogue can be expressed as a nonnegative K-tuple, $m_i = [m_i^1, m_i^2, \ldots m_i^K]^{\mathrm{T}}$. Hence, the mutational catalogues of G cancer genomes can be expressed as a nonnegative matrix of mutational catalogues $M = \begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix}$ of size $K \times G$. In this case, the mutational catalogues form the columns of the matrix, where K is the number of mutation types and G is the number of genomes. The subscript index indicates the mutational catalogue while the superscript index corresponds to the mutation type.

The exposure to a mutational process with a signature $P_1 = [p_1^1, p_1^2, \ldots p_1^K]^{\mathrm{T}}$ is a scalar describing the number of mutations, $e^1 \in \mathbb{N}_0$, attributed to that signature in a given mutational catalogue. In this notation, the product $p_1^2 \times e_g^1$ is the number of mutations of type corresponding to the 2nd letter of alphabet Ξ caused by the mutational process P_1 in a cancer genome with number g. Hence, one can define a set of exposures of G genomes to a set of N processes as a nonnegative matrix $E = \begin{bmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{bmatrix}$ with size $N \times G$. Here, the subscript index indicates the genome while the superscript index corresponds to the signature.

In addition to the signatures of the operative mutational processes, the mutational catalogue of a cancer genome also reflects the effect of random error processes, which may occur due to the used experimental approach (*e.g.*, DNA sequencing) and/or bioinformatics methods (*e.g.,* algorithms for identifying somatic mutations from next-generation sequencing data). To reflect the existence of such errors, a random noise term is introduced in equation 2.5. This noise term n reflects an additive white Gaussian noise that occurs due to non-systematic errors. The noise term is specific to each mutational catalogue and it is defined over the alphabet Ξ of the mutational catalog, where n: $\Xi \rightarrow \mathbb{R}^K$. Hence, for a set of mutational catalogues of G cancer genomes, the noise term can be expressed as a matrix $N = \begin{bmatrix} n_1^1 & n_2^1 & \cdots & n_{G-1}^1 & n_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_1^K & n_2^K & \cdots & n_{G-1}^K & n_G^K \end{bmatrix}$ of real numbers with size $K \times G$. The subscript index indicates the noise term for the mutational catalogue while the superscript index corresponds to the mutation type. It should be noted that systematic sequencing and analysis errors are considered as "synthetic mutational processes" with specific profiles present in some (or all) genomes. A whole subsection in chapter 4 is devoted to examining such systematic sequencing and analysis errors across a large set of cancer genomes.

2.3.2 Defining the mutational signatures deciphering problem

The signatures of N different mutational processes and their respective exposures need to be extracted from the mutational catalogues of M cancer genomes (Figure 2.4). Using the introduced matrix notation, this could be expressed as:

$$\begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix} =$$

$$\begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix} \times \begin{bmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{bmatrix} + \begin{bmatrix} n_1^1 & n_2^1 & \cdots & n_{G-1}^1 & n_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_1^K & n_2^K & \cdots & n_{G-1}^K & n_G^K \end{bmatrix} \quad (2.6)$$

or one can simplify equation 2.6 in a matrix form as:

$$M = P \times E + N \quad (2.7)$$

In practice, one knows only the mutational catalogues in the matrix M and the goal is to identify P and E such that these matrices best describe the original matrix M without over-fitting the data. Figure 2.4 provides a graphic representation of the problem for deciphering signatures of mutational processes from a set of mutational catalogues.

2.3.3 Examining the problem as a blind source separation

The examined problem can be considered as a specific case of the classic "cocktail party" problem, where multiple people attending a party are speaking simultaneously while several microphones placed at different locations are recording

Figure 2.4

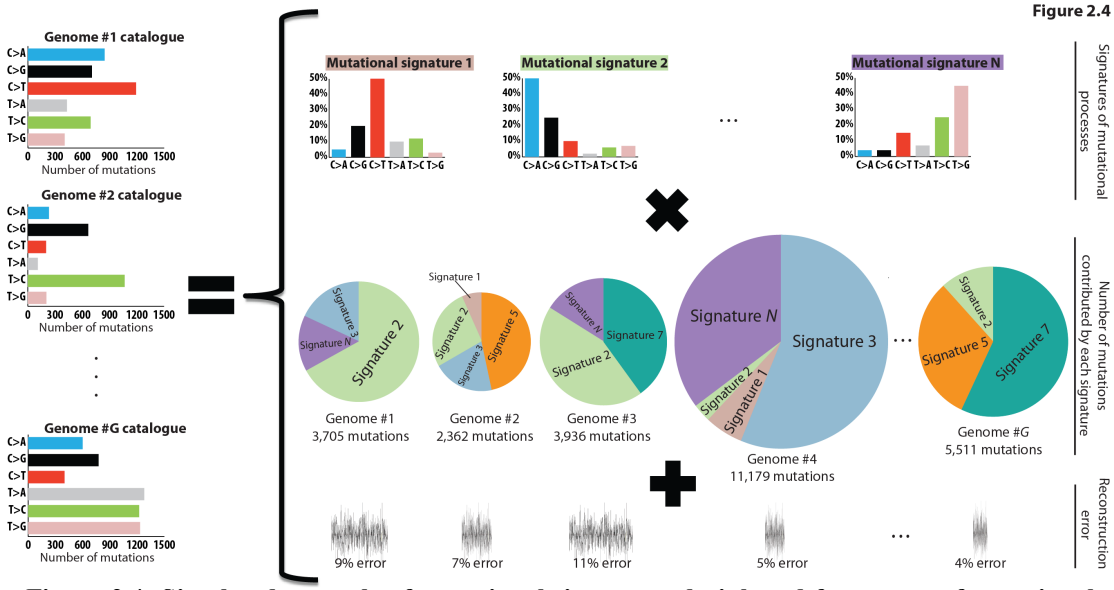


Figure 2.4: Simulated example of mutational signatures deciphered from a set of mutational catalogues. The mutational catalogues of G cancer genomes are used to decipher the signatures of N mutational processes as well as the number of mutations caused by each of the processes in each of the genomes. The extracted signatures and contributions do not allow an exact reconstruction of the original set, thus resulting in genome-specific reconstruction error.

the conversations. Each microphone captures a mixture of all sounds and the problem is to decipher the individual conversations from the recordings. This becomes possible because each microphone captures each conversation with a different intensity depending on the distance between the microphone and the conversation. Analogously, the provision of a catalogue of somatic mutations from a cancer genome only provides the final mixture of the signatures of all mutational processes operative in a cancer sample, and the goal is to decipher these signatures from a set of available mixtures (Figure 2.4). Thus, the mutational processes and their signatures are the "conversations," the exposure to a process is the "loudness of the conversation," the cancers themselves are the "microphones," and the final mutational catalogues are the "recordings."

The "cocktail party" problem is a type of blind source separation (BSS) problem that involves unscrambling latent (not observed) signals from a set of mixtures of these signals, without knowing anything about the mixing. To be able to "unmix" and reconstruct the original sources from the records, a BSS algorithm is needed for best possible extraction of the original signals from the mixtures. These BSS algorithms are capable of revealing hidden features and dependencies in large sets of observed data, and, based on these features, building a representation of the data that can contribute to understanding the biological mechanisms behind these data. The unmixing and reconstruction of the original signals is usually based on some constrained and/or regularized optimization procedure minimizing an objective (cost) function together with a few imposed constraints, such as: maximum variability, statistical independence, nonnegativity, smoothness, sparsity, simplicity, *etc*. The choice of the optimization constraints is usually based on *a priori* knowledge about the processed data, and hence the constraints could be different for every particular case.

The main difficulty in solving a BSS problem is that it is usually an under-determined (ill-posed) problem. There are two main/widely-used methods for resolving the under-determination of BSS: Independent Component Analysis (ICA) and Nonnegative Matrix factorization (NMF) (Comon, 2010; Roberts and Everson, 2001). Below, I briefly describe the basic principles of ICA and NMF.

ICA estimates the source and the mixing matrices by maximizing the statistical independence of the retrieved source signals (*i.e.,* the matrix columns are expected to be statistically independent). Typically, the source independence is achieved by maximizing some high-order statistics for each source signal, such as the kurtosis or negentropy (negative entropy). The main idea behind ICA is that while the probability distribution of a linear mixture of sources is expected to be close to a Gaussian (according to the Central Limit Theorem), the probability distribution of the original independent sources is expected to be non-Gaussian. As a result, ICA aims to maximize the non-Gaussian characteristics in the estimated sources with the goal of finding statistically independent non-Gaussian sources that reproduce the experimental data.

In contrast to ICA, NMF does not seek statistical independence or constrain any other statistical property. Thus, nonnegative matrix factorization allows the estimated sources to be partially or entirely correlated. Instead, NMF enforces a nonnegativity constraint on the original sources and their mixing components (*i.e.*, all the estimated matrix elements are greater than or equal to zero).

The differences between NMF and ICA have important implications for choosing one method over another. In general, ICA is used when one is looking for statistically independent signals. However, in practice, there are many cases where the ICA assumption of statistical independence contradicts the biological reality. For example, two distinct mutational processes may be reliant on the same components of the cellular machinery making them (at least partially) statistically dependent and, as such, these signals cannot be deciphered with an algorithm whose basis is to seek statistical independence. In contrast to ICA, NMF focuses purely on part-based decomposition (Lee and Seung, 1999). The part-based decomposition is particularly useful as it allows describing the original data only by additive signals that cannot cancel one another. This part-based decomposition results in "natural sparseness" of the underlying processes and it has been shown to extract meaningful components from complex datasets (Lee and Seung, 1999).

The nonnegative nature of the developed model in equation 2.7 requires a method that assumes (at the very least) nonnegativity of the original sources. The elegance, simplicity, and ability to extract meaningful processes make NMF the

method of choice in this thesis. It should be noted that there are different algorithms that can be used for nonnegative matrix factorization. The results presented in this study are exclusively based on the multiplicative update algorithm (Lee and Seung, 1999).

2.3.4 Approach for deciphering signatures of mutational processes

For a given mutational catalogue M that contains G cancer genomes defined over an alphabet Ξ with K letters corresponding to mutation types (*i.e.,* M has a size $K \times G$), the algorithm extracts N mutational signatures defined over the same alphabet Ξ. The algorithm has the following steps:

STEP 1 (Dimension Reduction): Reduce the dimensions of the original matrix M by removing any mutation types that together account for $\leq 1\%$ of the mutations in all genomes, *i.e.* remove the maximum set of rows R in M for which:

$$\sum_{r \in R} \sum_{g=1}^{G} m_g^r \leq 0.01 \times \sum_{i=1}^{K} \sum_{j=1}^{G} m_j^i$$

and the cardinality of the set R, $|R|$, is maximized. The matrix M is transformed into a new matrix $\dot{M}$ with dimensions $\dot{K} \times G$, where $\dot{K} = K - |R|$.

STEP 2 (Bootstrap): Apply Monte Carlo bootstrap resampling to avoid over-fitting the extracted mutational signatures. The dimensionally reduced matrix $\dot{M}$ resulting in a new matrix $\breve{M}$, where the probability for getting a mutation of type corresponding to the q^{th} letter in the alphabet Ξ in a genome g is $Pr(\breve{m}_g^q) = \frac{\dot{m}_g^q}{\sum_{i=1}^{K} \dot{m}_g^i}$ while the total number of mutations in each genome g remains unaffected, *i.e.,* $\sum_{i=1}^{K} \breve{m}_g^i = \sum_{j=1}^{K} m_g^j$.

STEP 3 (NMF): Apply the multiplicative update algorithm (Lee and Seung, 1999) for nonnegative matrix factorization to the bootstrapped data by finding the solution to $\min_{P \in \mathbb{M}_{\mathbb{R}_+}^{(\dot{K},N)}} \min_{E \in \mathbb{M}_{\mathbb{R}_+}^{(N,G)}} ||\breve{M} - P \times E||_F^2$:

I. Initialize matrices P and E as random nonnegative matrices with respective sizes $\dot{K} \times N$ and $N \times G$, where N is the number of signatures.

II. Iterate until convergence, defined as 10,000 iterations without change, or until the maximum number of 1,000,000 iterations is reached:

$$e_G^N \leftarrow e_G^N \frac{[P^T \breve{\mathrm{M}}]_{N,G}}{[P^T P E]_{N,G}}$$

$$p_N^{\dot{K}} \leftarrow p_N^{\dot{K}} \frac{[\breve{M} E^T]_{\dot{K},N}}{[P E E^T]_{\dot{K},N}}$$

The notation $[AB]_{x,y}$ is equivalent to the $(x,y)^{-\text{th}}$ element of the matrix C, where $C = A \times B$.

III. Store the identified signatures P and their respective exposures E.

STEP 4 (Iterate): Perform steps 2 and 3 for I iterations. I is determined by evaluating the convergence of the iteration-averaged signature matrix $\bar{P}$ (see below for deriving $\bar{P}$). I is selected in such a way that performing $2 * I$ iterations (*i.e.,* doubling the iterations) does not significantly change $\bar{P}$. In most cases between 400 and 500 iterations are needed, however, in some cases solutions could be found for $I \leq 100$ while in rare cases more than 1,000 iterations might be required. In general, the value of I is strongly dependent on the size and type of the initial matrix M.

STEP 5 (Cluster): The iterations performed in step 4 result in two sets of matrices, $S_P \in \mathbb{M}_{\mathbb{R}_+}^{(\dot{K},N)}$ and $S_E \in \mathbb{M}_{\mathbb{R}_+}^{(N,G)}$, that correspond respectively to the mutational signatures and their exposures generated over the I iterations. A partition-clustering algorithm is applied to the set of matrices S_P to cluster the data into N clusters. A variation of k-means (Jain, 2010), where each signature for $\forall P \in S_P$ is assigned to exactly one cluster, is used to partition the data. Similarities between mutational signatures are evaluated using a cosine similarity while the N centroids are calculated by averaging the signatures belonging to each cluster. The iteration-averaged matrix $\bar{P}$ is formed by combining the N centroid vectors ordered by their reproducibility (see Step 6). The error bars reported for each mutation type in each signature in $\bar{P}$ are calculated as the standard deviations of the corresponding mutation type in each centroid over the I iterations. Note that clustering the data in S_P effectively results in clustering S_E as each signature unambiguously corresponds to exactly one exposure, thus allowing derivation of $\bar{E}$.

STEP 6 (Evaluate): The reproducibility of the derived average signatures $\bar{P}$ is evaluated by examining the tightness and separation of the clusters used to form the centroids in $\bar{P}$ (see Step 5). More specifically, using cosine similarity, the average silhouette width for each of the N clusters is calculated. An average silhouette width of 1.00 is equivalent to consistently deciphering the same mutational signature, while a low silhouette width indicates a lack of reproducibility of the solution. The average silhouette width (Rousseeuw, 1987) of the N clusters is used as a measure of reproducibility for the whole solution. In addition to reproducibility, the average Frobenius reconstruction error is used to evaluate the accuracy with which the deciphered mutational signatures and their respective exposures describe the original matrix M, *i.e.,* $||M - \bar{P} \times \bar{E}||_F^2$, where a lower Frobenius reconstruction error corresponds to a better description of the original matrix. There is some association between the reproducibility of a solution and its reconstruction error. For example, solutions with very low reproducibility usually have high Frobenius reconstruction errors.

The developed framework for deciphering signatures of mutational processes relies on two input parameters, the original matrix M (size $K \times G$) and the number of mutational signatures N to be deciphered from M. However, in most cases, the value of N is unknown and needs to be determined from M. The model selection framework relies on applying the framework for deciphering signatures of mutational processes for values of N between 1 and $\min(K, G) - 1$. The reproducibility and average Frobenius reconstruction error is evaluated for each N, and the value of N is selected such that the extracted mutational signatures are reproducible and the reconstruction error is low.

2.3.5 Computational implementation of the algorithm

The framework for deciphering signatures of mutational processes — including its source code, brief documentation, and several examples of applying it to mutational catalogues — is freely available for download from:

http://www.mathworks.com/matlabcentral/fileexchange/38724

2.4 Evaluating the computational framework using simulated data

In the previous section of this chapter, I introduced a theoretical model of signatures of mutational processes operative in a cancer genome. Based on this model, I mathematically introduced the problem of deciphering mutational signatures from a set of mutational catalogues of cancer genomes. Further, I proposed an algorithm and developed a computational framework that allows to decipher these signatures. In this section, I focus on evaluating the developed approach with simulated data. The application of the approach to experimental data is performed in chapters 3 and 4.

2.4.1 Generating the simulation data

Signatures of mutational processes with different exposures are randomly generated and used to simulate mutational catalogues of cancer genomes. The simulated mutational catalogues are leveraged to assess the ability of the developed approach to decipher the mutational signatures with which the data are simulated. In most cases (unless specified otherwise in the text), the signatures of mutational processes are stochastically generated over the alphabet Ξ_{96} with similarities between them comparable to those previously observed in breast cancer genomes (Nik-Zainal et al., 2012). Similarly, unless specified otherwise, the exposures to mutational processes are uniformly distributed across the set of simulated cancer genomes while the total number of mutations in each mutational catalogue is drawn from a distribution comparable to the distribution of the total substitutions found in many human cancer genomes (Greenman et al., 2007; Nik-Zainal et al., 2012; Stratton, 2011; Wood et al., 2007). For every mutational process with signature $P_1 = [p_1^1, p_1^2, \ldots p_1^K]^{\mathrm{T}}$, defined over an alphabet Ξ with K letters, contributing e_g^1 mutations in a cancer genome g, each mutation is assigned to one of the K mutation types according to the discrete probability density function of P_1. Poisson noise and additive white Gaussian noise are added to every simulated mutational catalogue. Lastly, each simulation scenario is repeated 100 times and the standard deviations of the results over these 100 repeats are reported as error bars in the respective figures.

2.4.2 Extracting mutational signatures from 100 simulated cancer genomes

An example of applying the developed theoretical approach to a set of 100 simulated mutational catalogues of cancer genomes is shown in Figure 2.5. Similar to many human cancer genomes (Greenman et al., 2007; Nik-Zainal et al., 2012; Stratton, 2011; Wood et al., 2007), every simulated genome contains between 500 and

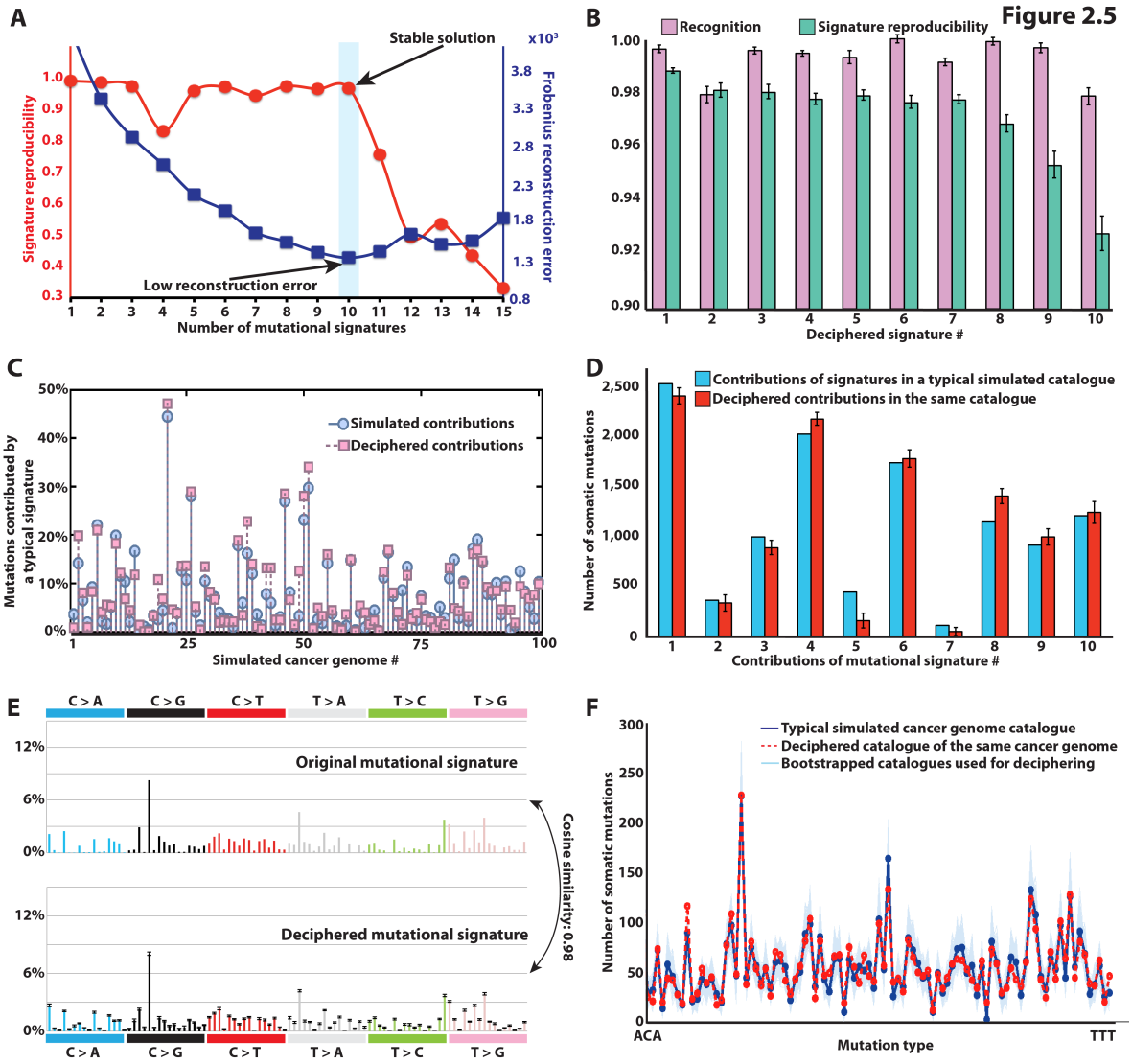


Figure 2.5: Deciphering mutational signatures from a set of 100 simulated mutational catalogues. ***(A)*** Identifying the number of processes operative in a set of 100 simulated cancer genomes based on reproducibility of their signatures and low error for reconstructing the original catalogues. ***(B)*** Comparison between the ten deciphered signatures and the ten signatures used to simulate the catalogues. Signature recognition, measured using cosine similarity, and signature reproducibility, measured using average silhouette width, is given for each mutational signature. ***(C)*** Comparison between deciphered and simulated contributions of one of the ten mutational processes in all cancer genomes. ***(D)*** Comparison between deciphered and simulated contributions of all signatures in a typical cancer genome. ***(E)*** Comparison between the profiles of typical deciphered and simulated signature. ***(F)*** Comparison between the mutational catalogues of a typical deciphered (red line) and simulated (dark blue line) cancer genome.

50,000 substitutions. The simulated mutations are generated using 10 mutational processes with distinct signatures each with 96 mutation types (*i.e.,* signatures are defined over Ξ_{96}).

Identifying the number, *N*, of mutational processes operative in a set of cancer genomes is required prior to deciphering their signatures. The developed model selection approach identifies *N* by applying the method for different values of *N* (see section 2.3.4). For every *N*, the similarity between the extracted processes (*i.e.*, process reproducibility) is evaluated from the stochastically initialized iterations. Further, for every *N*, the model selection approach assesses the average Frobenius reconstruction error of the averaged deciphered signatures $\bar{P}$ and their exposures $\bar{E}$, *i.e.,* $|| M - \bar{P}\times\bar{E}||_F^2$. Low reconstruction error is indicative of an accurate description of the original cancer genome catalogues. *N* is selected such that the extracted processes are reproducible and the reconstruction error is low. Over-fitting the mutational signatures is avoided by bootstrapping the data (in each iteration) before applying NMF to them (see section 2.3.4).

For the 100 simulated cancer genomes, the approach is able to identify reproducible solutions for *N* between 2 and 10 (Figure 2.5A). Increasing the number of signatures from 2 to 10 substantially reduces the reconstruction error, but increasing beyond 10 does not further reduce it (Figure 2.5A). This indicates that the computational approach can optimally distinguish the signatures of 10 mutational processes, precisely the number originally used to simulate the mutational catalogues of these 100 cancer genomes. The 10 deciphered signatures are very reproducible (average silhouette width > 0.96) as well as extremely similar (average cosine similarity > 0.98) to the ones used to generate the 100 mutational catalogues (Figure 2.5B). Further, the computational approach is able to accurately identify the number of mutations contributed by each of the 10 processes in each of the genomes. Comparison between original and deciphered exposures of one of the signatures in all genomes is shown in Figure 2.5C and a comparison of the contributions of all 10 signatures in a single genome is shown in Figure 2.5D. A typical comparison between an original and deciphered signature is shown in Figure 2.5E and a typical comparison between an original and reconstructed mutational catalogue of a genome is depicted in Figure 2.5F. In summary, the applied approach is able to accurately

identify the underlying mutational signatures and their respective exposures in this set of 100 simulated mutational catalogues.

2.4.3 Identifying factors that influence extraction of mutational signatures

To identify factors that affect the ability to extract mutational signatures, signatures of mutational processes and their respective exposures are simulated under a number of different scenarios. The original signatures used to simulate the data are compared to the deciphered signatures in order to evaluate both the limitations and robustness of the developed computational framework. All comparisons between mutational signatures are done using a cosine similarity as previously described in section 2.2.2.

Figure 2.6

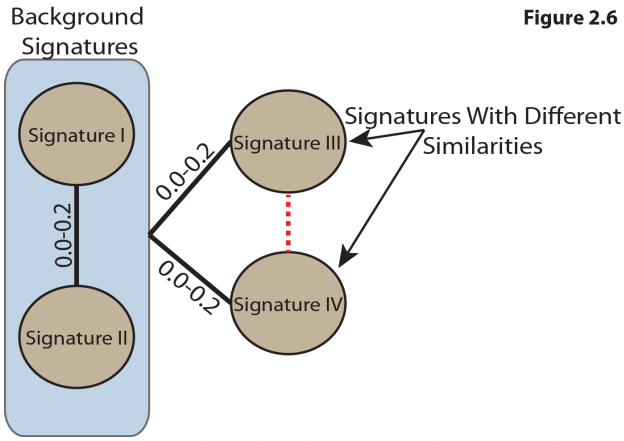


Figure 2.6: Design for simulating four mutational signatures with different similarities between them. Signatures I and II differ significantly from each other as well as from the other two Signatures (cosine similarity between 0.00 and 0.20). Signatures III and IV are simulated with varying similarities between them.

To evaluate how the degree of similarity between mutational signatures affects their extraction, sets of four randomly generated signatures are simulated; two of the signatures are very different from any of the other signatures, while the similarity of the remaining two to each other is varied (Figure 2.6). Hence, Signatures I and II are

Figure 2.7

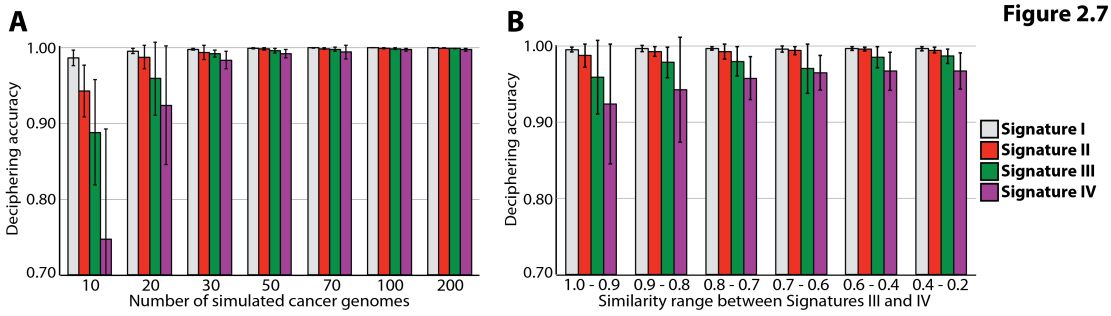


Figure 2.7: Deciphering mutational signatures with different similarities between them. ***(A)*** Different numbers of mutational catalogues are examined while Signatures III and IV are simulated with very similar profiles. ***(B)*** The mutational catalogues of 20 cancer genomes are simulated while the similarity between Signatures III and IV is varied.

simulated such as the cosine similarity between each of these signatures and any other signature is always within the range of 0.00 and 0.20 (*i.e.,* signatures with very

Figure 2.8

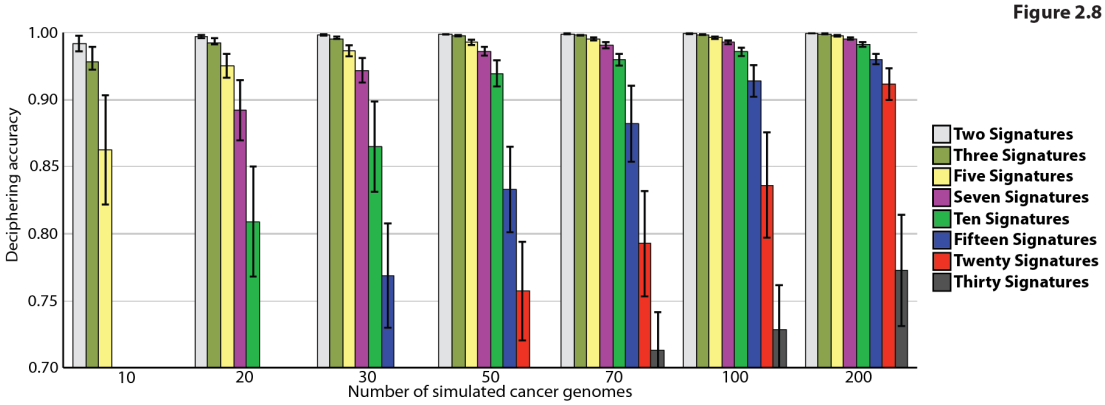


Figure 2.8: Deciphering mutational signatures from different sets of cancer genomes. Evaluating the effect of deciphering between two and thirty mutational signatures from sets of mutational catalogues derived from 10, 20, 30, 50, 70, 100, and 200 cancer genomes.

different mutational profiles) while the similarity range between Signatures III and IV is varied, as described below.

Sets of Signatures III and IV are simulated with a cosine similarity ranging between 0.90 and 1.00 (*i.e.,* signatures with extremely similar profiles). In addition, different numbers of mutational catalogues are examined (Figure 2.7A). The performed simulations indicate that 30 mutational catalogues are sufficient for adequately identifying the four mutational signatures, while 50 or more cancer genomes allow to perfectly decipher signatures that are extremely similar (Figure 2.7A). Further simulations are carried out in which sets of mutational catalogues of 20

Figure 2.9

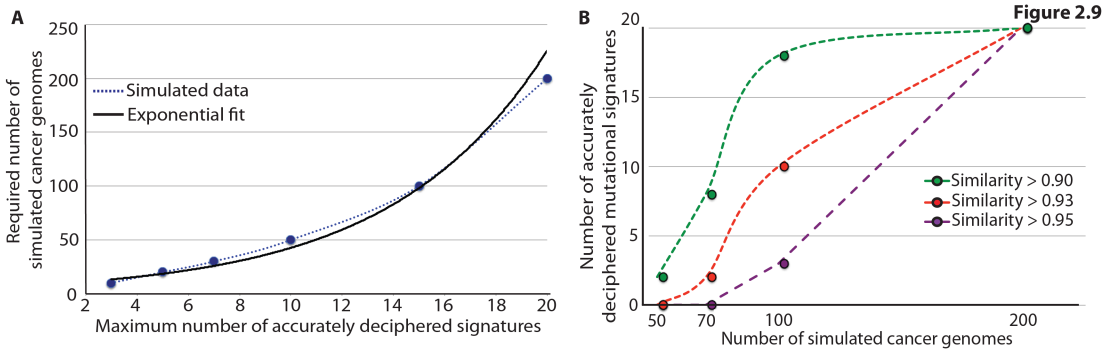


Figure 2.9: Dependencies between mutational signatures and mutational catalogues of cancer genomes. ***(A)*** Exponential dependency between accurately deciphered signatures (*i.e.,* cosine similarity between simulated and deciphered signature ≥ 0.95) and the number of mutational catalogs needed to decipher these signatures. ***(B)*** Identification of the maximum number of accurately deciphered signatures (cosine similarity between simulated and deciphered signature shown in the legend) from sets of mutational catalogues simulated using the signatures of 20 mutational processes.

cancer genomes are evaluated with a varied distance range between Signatures III and IV. Interestingly, even though 20 mutational catalogues are insufficient to decipher the profiles of very similar looking signatures, they are suitable for effectively extracting signatures that have similarities ≤ 0.70 (Figure 2.7B).

The number of available cancer genomes mathematically limits the number of signatures that can be extracted from the mutational catalogues of these genomes. For example, accurately deconvoluting signatures of 15 mutational processes from the mutational catalogues of only 10 cancer genomes is ineffective. To evaluate the effect of the number of mutational catalogues on extracting mutational signatures, simulations with different numbers of cancer genomes generated using a varying number of mutational signatures are performed. Between 10 and 200 sets of mutational catalogues are simulated using up to thirty mutational signatures (Figure 2.8). Interestingly, the number of mutational catalogues required to accurately decipher the signatures operative in them increases exponentially with the number of signatures (Figure 2.9A). Thus, while mutational catalogues from 100 cancer genomes are necessary to extract the signatures of fifteen mutational processes, at least 200 cancer genome catalogues are required to deconvolute twenty signatures (Figure 2.8). Nevertheless, it is possible to decipher at least some of the 20 mutational signatures from a set of 100 or fewer mutational catalogues (Figure 2.9B).

The number of somatic mutations in each cancer genome affects the ability to decipher the signatures of the operative mutational processes. In all previous

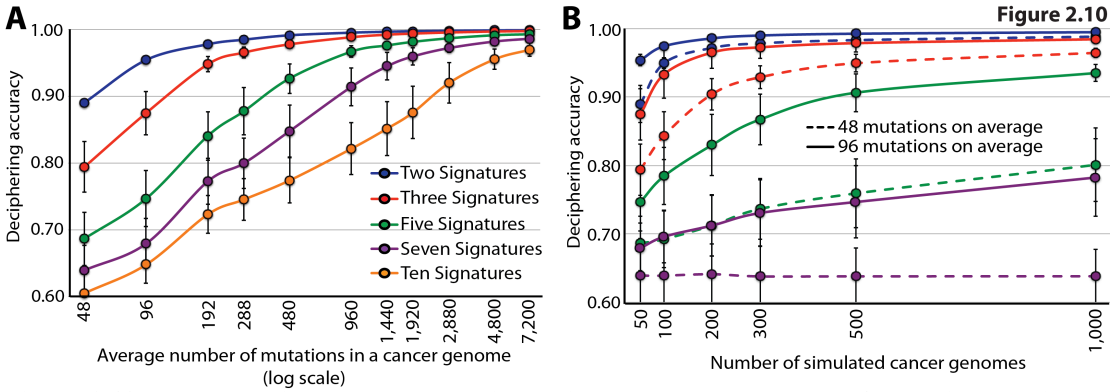


Figure 2.10: Dependencies between mutational signatures and numbers of somatic mutations. ***(A)*** Evaluating the effect of deciphering different number of mutational signatures from sets of mutational catalogues derived from 50 cancer genomes. The catalogues are simulated with different average number of mutations in a cancer genome. ***(B)*** Evaluating the effect of deciphering 2, 3, 5, or 7 mutational signatures from large sets of mutational catalogues containing small number of average mutations per cancer genome. The line colours correspond to the ones in the legend of panel A.

simulations, it is assumed that the distributions used to simulate the number of somatic mutations in cancer genomes are similar to those of some common cancers such as breast and prostate cancer. However, recent studies have demonstrated that there is substantial heterogeneity between the mutational burdens across major cancer types (Alexandrov et al., 2013a; Lawrence et al., 2013). In this section, simulations of 50 mutational catalogues with different average numbers of somatic mutations are performed. Each mutational catalogue is simulated using between two and ten mutational signatures. Obviously, having more somatic mutations (*i.e.,* more data for each sample) allows to better distinguish the profiles of the mutational signatures. As such, the focus of these simulations is to examine how lower average numbers of mutations (*i.e.,* between 48 and 7,200 mutations) affect the ability of the approach to identify mutational signatures. The results indicate that two or three signatures can be effectively extracted from catalogues with less than a hundred somatic mutations (Figure 2.10A). In contrast, extracting seven or more mutational signatures requires an average of at least 1,000 mutations per catalogue.

Figure 2.11

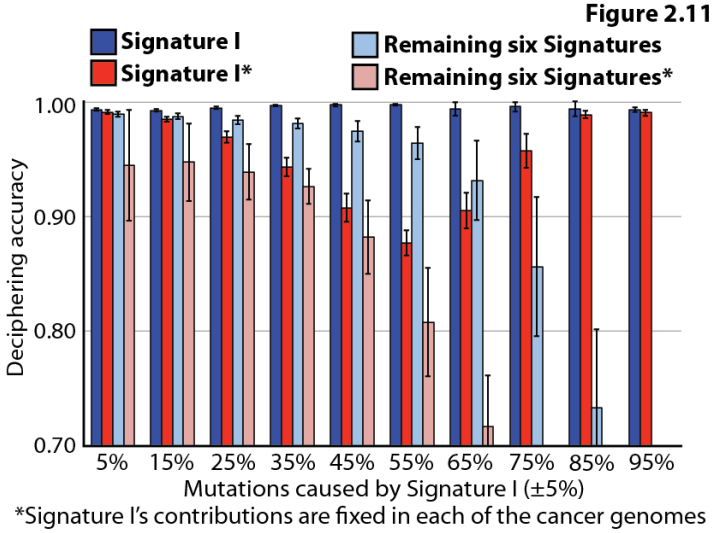


Figure 2.11: Deciphering mutational signatures with different contributions in mutational catalogues. Fifty mutational catalogues are simulated using mutational signatures with different contributions. Signature I's contributions are fixed to contribute a fixed percentage of all mutations in either the whole set of mutational catalogues (*i.e.*, the overall contribution is fixed but different genomes can have different contributions of Signature I; blue bars) or in each individual cancer genome (*i.e.,* Signature I's contributions are fixed in every single mutational catalogue; red bars).

The combined protein coding exons (the "exome") constitute only ~1% of the human genome. The analysis of exomes compared to whole-genome sequences is often perceived as advantageous because of lower costs and because a substantial proportion of cancer-causing driver somatic mutations may be found using this strategy. As a result, many more exome sequences of cancers have currently being generated than whole-genomes. To further evaluate the applicability of the

approach to only parts of the genome (and more specifically exome sequences), large sets of mutational catalogues simulated with small average numbers of somatic mutations are examined. The results reveal that at least 500 mutational catalogues with an average of 96 mutations per catalogue (a total of ~50,000 mutations) are needed to decipher five mutational processes (Figure 2.10A), but these five mutational processes can be more easily deciphered from 50 cancer genomes containing an average of 480 mutations (a total of ~25,000 mutations, Figure 2.10B). This result indicates that it is more effective to decipher mutational signatures from a small number of catalogues containing many mutations than from many catalogues containing few mutations.

The strength of exposure of a mutational process in a set of genomes also influences the ability to decipher its signature. Two types of simulations of seven signatures operating with different strengths in 50 mutational catalogues are performed. In the first type, the percentage of exposures of Signature I in *all samples* is simulated as a constant parameter with values between 5% and 95% of all mutations (Figure 2.11). In contrast, in the second type of simulation, the exposures to Signature I are kept as a constant parameter in *every sample*, again, with values between 5% and 95% of all mutations (Figure 2.11). The results demonstrate that signatures contributing <5% of all mutations can be difficult to distinguish. Similarly, deciphering the members of a set of mutational signatures that have similar exposures with respect to each other over a set of cancer genomes is challenging (Figure 2.11). To overcome this problem, it may be advantageous to combine sets of mutational catalogues in which mutational processes are

Figure 2.12

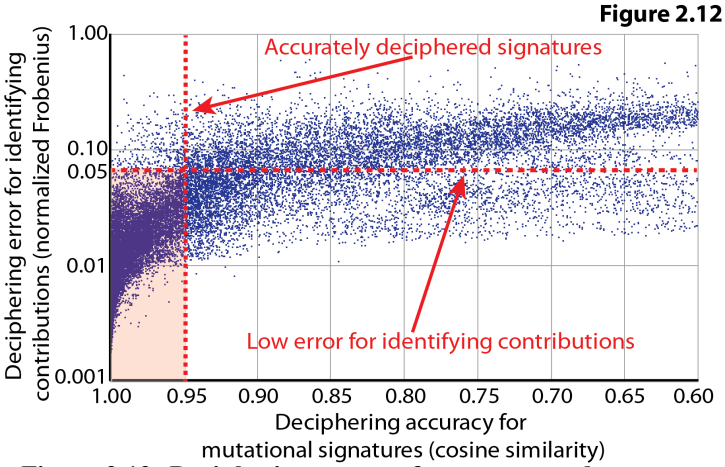


Figure 2.12: Deciphering errors of exposures and accuracy of mutational signatures. Comparison, across all previously performed simulations, between the accuracy of the deciphered mutational signatures and the deciphering error for identifying the contributions of these signatures. The deciphering Frobenius reconstruction error is calculated and averaged for each contribution and normalized based on the numbers of mutations in the respective mutational catalogue.

more likely to be active in different proportions (*e.g.,* from different cancer types). However, combining sets of mutational catalogues in this way ought to be considered with caution as the number of cancer genomes required for the extraction of signatures increases exponentially with the number of operative signatures and more cancer types may well entail more signatures (Figure 2.8 and Figure 2.9).

In addition to deciphering mutational signatures, the developed computational approach identifies the number of somatic mutations that each signature contributes to each mutational catalogue. In general, one would expect that the developed algorithm is, at least to some degree, symmetrical. Thus, when the algorithm correctly identifies the mutational signatures, it should also accurately estimate the contributions of these signatures (see section 2.3.4 in regards to the symmetric clustering of the data extracted in the sets of signatures, S_P, and the sets of exposures, S_E). Evaluating the average deciphering error for identifying contributions, for all previously performed simulations, confirms that the majority of accurately deciphered mutational signatures (*i.e.,* cosine similarity between simulated and extracted signatures ≥ 0.95) are associated with a low error (*i.e.,* normalized Frobenius error rate $\leq 5\%$) for their respective signature contributions (Figure 2.12). Further examination reveals that only very few of the accurately extracted signatures are associated with a normalized Frobenius error rate $\geq 5\%$ (Figure 2.13A). Interestingly, the analysis indicates that the contributions of signatures generating large numbers of mutations (> 200) are generally associated with lower error rates (Figure 2.13B).

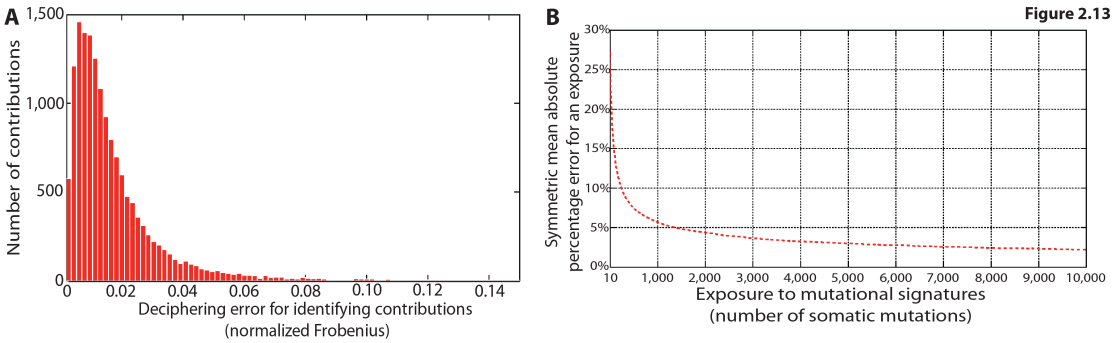


Figure 2.13: Evaluating the error rate of identified contributions of mutations signatures. ***(A)*** Distribution of the normalized Frobenius error for identifying the contributions of accurately deciphered signatures of mutational processes (*i.e.,* cosine similarity between simulated and deciphered signature ≥ 0.95). ***(B)*** Average symmetric mean absolute percentage error for identifying the contributions of accurately deciphered signatures of mutational processes (*i.e.,* cosine similarity between simulated and deciphered signature ≥ 0.95) based on the number mutations contributed by the signature.

2.5 Discussion

In this chapter, I have modelled the signatures of somatic mutational processes in cancer genomes as a blind source separation problem and introduced a computational framework that extracts these mutational signatures from the mutational catalogues obtained from cancer genome sequences. To identify these signatures, the intrinsic nonnegativity of mutations mandates employment of a method incorporating a nonnegative constraint. The extensive evaluations of the approach with simulated data demonstrate that the developed algorithm is effective in deciphering mutational signatures from mutational catalogues.

The efficiency of the algorithm could be further improved by incorporating additional constraints. For example, the current implementation of the computational framework relies on nonnegative matrix factorization, which has a natural weak sparsity constraint; however, a strong sparsity constraint could be applied to the exposure matrix E. This would guarantee that the mutational catalogue of a cancer genome is described by a minimum number of processes. Algorithms implementing this and other constraints have been previously developed (Berry et al., 2007; Gao and Church, 2005; Peharz and Pernkopf, 2012; Zheng et al., 2006) and could be applied to cancer genomics data. Nonetheless, this study demonstrates that an approach based on the simplest (*i.e.,* without additional constraints) NMF algorithm is sufficient to decipher both the signatures of the mutational processes operative in a set of cancer genomes as well as the number of mutations each signature contributes to the mutational catalogue of each cancer genome.

Parameters to which solutions are sensitive include the number of operative mutational processes, the strength of their exposures, the degree of difference between mutational signatures, the number of analysed cancer genomes, the number of mutations per cancer genome, and the number of mutation types that are incorporated into the model (Figures 2.6 through 2.13). These factors will determine the manner in which the method will be applied in the next chapters of this thesis. Importantly, the results show that, despite relatively few mutations present in each case, the approach can be applied to exome data, extracting at least some of the signatures of the operative processes.

It should be noted that when the number of samples in a dataset is too low or when the mutational burden is insufficient, the developed approach will lack the power to decipher the signatures of all operative mutational processes. Thus, in some cases, the extracted signatures will represent mixtures of multiple independent patterns of mutations and only additional samples will allow further differentiating these mutational signatures.

Diverse mutation classes can be included in this type of analysis. Thus the application of the developed approach can, if desired, be limited to single base substitutions or be widened to include double nucleotide substitutions, insertions, deletions, geographically localized forms of mutation and mutation features such as transcriptional strand-bias. Following this principle, rearrangements and copy number changes (and potentially even epigenetic modifications) could be incorporated in order to derive a comprehensive overview of operative mutational processes.

The complexity of the mutational processes operative in some cancers and the inherent challenges in extracting their attendant mutational signatures should not be underestimated. For example, tobacco smoke contains around 7,000 chemicals from which over 60 are known to be mutagenic (Rodgman and Perfetti, 2008). Thus, the mutational pattern of a lung cancer in a tobacco smoker will reflect the activity and potency of (at least) several of these chemicals. Each of these chemicals may have its unique mutational signature. A group of smokers loyal to the same brand will be simultaneously exposed to the same combination of mutagens. Analysis of tumours from this group of individuals therefore may not allow the mutagens to be distinguished from one another and the developed computational approach will extract only a single signature that encompasses the combined mutational activity of the most mutagenically potent chemicals. However, as different cigarette brands may contain different combinations and amounts of mutagens, analysis of mutational catalogues from cancers due to different tobacco brands could allow differentiation between the signatures of each of the different chemicals. An ambitious aspiration of this nature would, however, probably only be feasible with data from thousands of cases, coupled to the statistical power and resolution provided by whole-genome mutational catalogues.

Chapter 3

Signatures of mutational processes operative in breast cancer

3.1 Introduction

The previous chapter introduced a novel mathematical model of mutational processes operative in cancer genomes and a computational framework that allows deciphering of the signatures of these processes from a set of mutational catalogues. The newly developed computational approach was extensively evaluated with simulated data demonstrating its applicability to mutational catalogues derived from sequencing both cancer genomes and cancer exomes. Further, the performed simulations demonstrated that the method is robust to a wide range of different parameters. In this chapter, I present and discuss the application of the developed framework to experimentally generated data. The framework is used to examine the mutational catalogues derived from the sequences of 844 breast cancer exomes and 119 breast cancer whole-genomes. The aim of this chapter is to describe the signatures of the mutational processes operative in breast cancer as well as to serve as a prelude to chapter 4 in which analogous analysis will be performed for another 29 different types of human cancer.

3.2 Data generation and filtering of mutational catalogues

It should be noted that none of the examined data are generated for the purposes of this thesis. Rather, the analysis relies on previously identified somatic mutations by curating freely available published data as well as data that was unpublished at the time. Any unpublished breast cancer data were generated internally at the Cancer Genome Project (CGP) for the purposes of other projects. The majority of breast cancer exomes are taken from The Cancer Genome Atlas (TCGA) data

portal as well as from peer-reviewed publications. In contrast, the majority of breast cancer whole-genomes are previously unpublished data. Summary of the numbers of samples based on their data source is provided in Table 3.1, whereas a complete list including all samples, all examined cancer types, and their respective data sources is provided in Appendix II.

Sample types and data source	Total
Exome	**844**
doi:10.1038/nature10933	63
doi:10.1038/nature11017	9
New unpublished samples	5
TCGA data portal	767
Whole genome	**119**
doi:10.1016/j.cell.2012.04.024	21
New unpublished samples	98
Grand Total	**963**

Table 3.1: Summary of breast cancer samples and their data sources.

The somatic mutations of the 844 breast cancer exomes and the 119 breast cancer whole-genomes are curated, filtered, and mutational catalogues are generated for each sample based on the Ξ_6, Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536} alphabets. It should be noted that there is no sample overlap between the breast cancer genomes and exomes (*i.e.*, breast cancer whole-genomes are not included twice as exomes and genomes).

As these data are retrieved from many different sources and generated using different next-generation sequencing platforms and bioinformatics approaches, quality control is performed in order to remove any germline contamination and technology specific sequencing artefacts. Germline mutations are filtered out from the list of reported mutations using the data from dbSNP (Sherry et al., 2001), 1000 genomes project (Abecasis et al., 2012), NHLBI GO Exome Sequencing Project (Fu et al., 2013), and 69 Complete Genomics panel (http://www.completegenomics.com/public-data/69-Genomes/). Any mutation at a position of a previously identified germline variant in any of these datasets is removed from the signatures analysis. Furthermore, technology specific sequencing artefacts are filtered out by using panels of (unmatched) BAM files for normal tissue containing 137 normal genomes and 532 normal exomes. Any somatic mutation present in at least three well-mapping reads in at least two normal BAM files is discarded. The remaining somatic mutations are used for the generation of mutational catalogues and the extraction of mutational signatures.

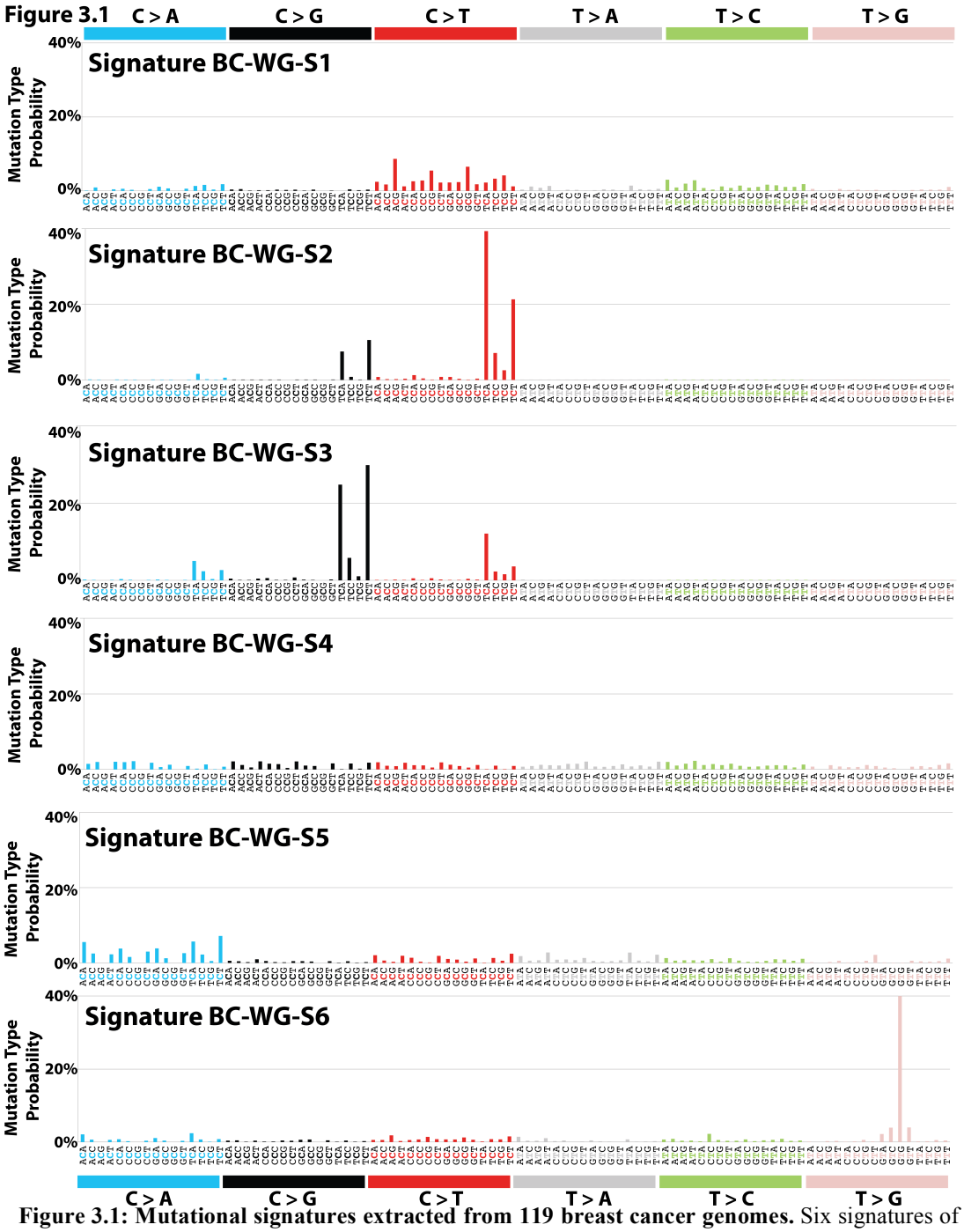


Figure 3.1: Mutational signatures extracted from 119 breast cancer genomes. Six signatures of mutational processes are deciphered from the base substitutions (including their immediate 5' and 3' sequence context) identified in the examined 119 breast cancer genomes. Each signature is depicted on an independent panel, where each type of substitution is displayed in a different colour. Mutational signatures are plotted based on the genome trinucleotide frequency.

The immediate 5' and 3' sequence context is extracted using the ENSEMBL Core APIs for human genome build GRCh37. Curated data originally mapped to an older version of the human genome is re-mapped using UCSC's freely available lift genome annotations tool. Dinucleotide substitutions are identified when two substitutions are present in consecutive bases on the same chromosome (sequence context is ignored). The immediate 5' and 3' sequence content of all small insertions

and deletions (indels) is examined and the ones present at mono/polynucleotide repeats or microhomologies are included in the analysed mutational catalogues as their respective types. Strand-bias catalogues are derived for each sample using only substitutions identified in the transcribed regions of well-annotated protein coding genes. Mutational signatures are independently derived from the mutational catalogues of breast cancer exomes and breast genomes (see below).

3.3 Deciphering the signatures of mutational processes from whole-genome sequencing of breast cancers

The developed computational approach presented in chapter 2 is applied to the mutational catalogue of 119 breast cancer whole-genomes that contain 654,308 somatic substitutions and indels. Mutational signatures are extracted based on the Ξ_{96}, Ξ_{99}, $\Xi_{192,}$ and Ξ_{1536} alphabets. The approach reveals six consistent and reproducible mutational signatures for all four alphabets – termed Signatures BC-WG-S1, BC-WG-S2, BC-WG-S3, BC-WG-S4, BC-WG-S5, and BC-WG-S6 (BC-WG-S stands here for breast cancer whole-genome signature).

The patterns of somatic substitutions for the signatures extracted using Ξ_{96} are depicted in Figure 3.1. Signature BC-WG-S1 is characterized by 50% C>T substitutions predominantly occurring at CpG dinucleotides and 25% T>C mutations with peaks at ApTpN trinucleotides. Signature BC-WG-S has predominantly (~76%) C>T mutations at TpCpN trinucleotides and (~20%) C>G mutations occurring at TpCpN trinucleotides. In contrast, Signature BC-WG-S3 is mirroring Signature BC-WG-S2 with ~65% of its substitutions being C>G at TpCpN trinucleotides, ~22% being C>T at TpCpN trinucleotides, and ~11% C>A at TpCpN trinucleotides. Signature BC-WG-S4 has a rather flat mutational pattern including all types of somatic mutations. While this mutational signature does not exhibit any strong features based on the immediate 5' or 3' sequence context, such as Signatures BC-WG-S2 or BC-WG-S3, the pattern of its substitutions is not completely uniform. Rather, the mutational pattern of Signature BC-WG-S4 has subtle trinucleotide features. Similar to BC-WG-S4, Signature BC-WG-S5 has a generally flat mutational pattern with subtle sequence context features. However, in addition, Signature BC-WG-S5 exhibits a predominance of C>A mutations (~40%) compared to the other

types of substitutions. Lastly, Signature BC-WG-S6 has a very strong sequence context with ~40% of all mutations being T>G at GpTpG.

As previously demonstrated, the developed computational framework can be applied to a wider repertoire of mutation types than the 96 mutated trinucleotides. The

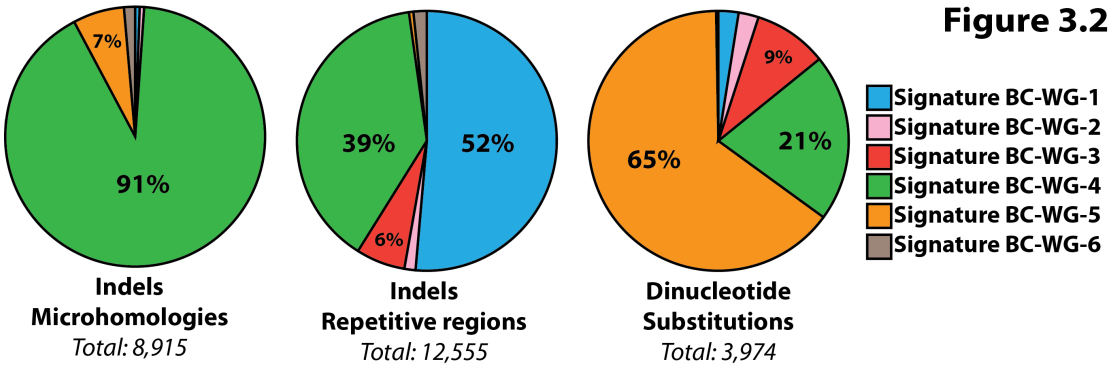


Figure 3.2: Breast cancer whole-genome mutational signatures with indels and dinucleotides. Mutational signatures analysis of the 119 breast cancer whole-genomes is extended to incorporate indels at microhomologies, indels at repetitive regions, and dinucleotide substitutions (*i.e.,* Ξ_{99} alphabet). The percentage of mutations attributed to these three additional mutation types is displayed for all signatures that contribute at least 5%. Each signature is displayed in a different colour.

Ξ_{96} alphabet can be extended to the Ξ_{99} alphabet by including three additional mutational subclasses: double nucleotide substitutions, indels at microhomologies, and indels at mono/polynucleotide repeats. This analysis reveals that Signature BC-WG-S4 is associated with 91% of the 8,915 indels at microhomologies found in the 119 whole breast genomes, 39% of the 12,555 indels found at mono/polynucleotide repeats, and 21% of the 3,974 dinucleotide substitutions (Figure 3.2). The activity of Signature BC-WG-S1 is associated with 52% of indels found at mono/polynucleotide repeats, whereas Signature BC-WG-S5 accounts for 65% of all dinucleotide substitutions. It should be noted that a significant proportion of the dinucleotide substitutions associated with Signature BC-WG-S5 are CC>AA. Signatures BC-WG-S2, BC-WG-S3, and BC-WG-S6 do not have a strong association with any type of indels or dinucleotide substitutions.

Previous examination of the mutational catalogues of 21 breast cancer genome showed a weak transcriptional strand-bias for all C>A mutations (Nik-Zainal et al., 2012). This bias results in C>A mutations being more common on the transcribed than the untranscribed strands of genes (and vice versa for G>T mutations). To investigate whether a particular mutational signature is associated with this (or any other) transcriptional strand-bias, the Ξ_{96} substitution alphabet is extended to include information on whether a substitution is on the transcribed or non-transcribed strand,

Figure 3.3

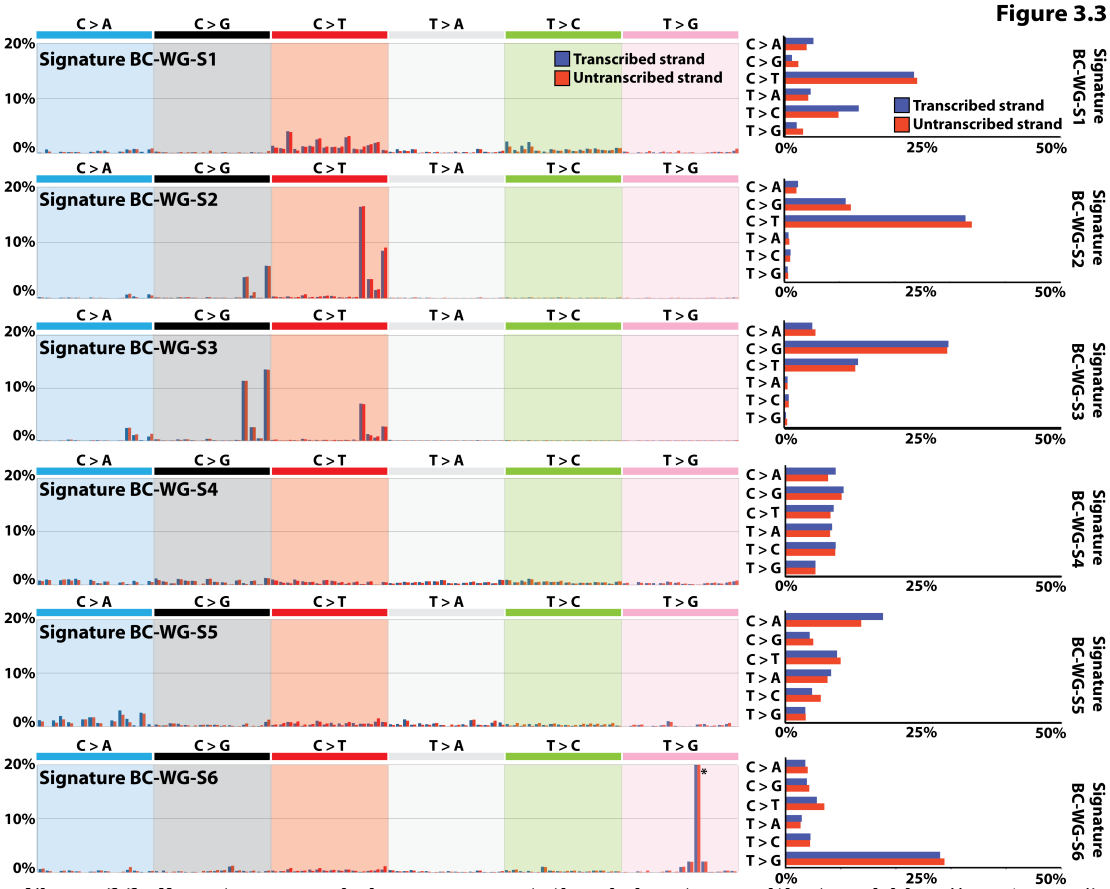


Figure 3.3: Breast cancer whole-genome mutational signatures with strand-bias. Signatures of mutational processes with strand-bias are extracted from the mutational catalogues of 119 breast cancer genomes. Six mutational signatures deciphered from the base substitutions (including their immediate 5' and 3' sequence context) identified in the transcribed regions of 119 breast cancer genomes. Each signature is depicted on an independent panel, where each type of substitution is highlighted in a different colour. The probability of a mutation to occur on a transcribed strand is depicted in blue, while red is used to display the probability of a mutation to occur on the untranscribed strand. Mutational signatures are plotted based on the genome trinucleotide frequency. Asterisk indicates mutation type exceeding 20%.

thus increasing the 96 trinucleotide substitutions to 192. The developed model selection approach again reveals the signature of six reproducible mutational processes (Figure 3.3) with patterns resembling the ones based on the Ξ_{96} alphabet (Figure 3.1). Examining the mutational signatures based on the Ξ_{192} alphabet reveals that Signature BC-WG-S2, Signature BC-WG-S3, Signature BC-WG-S4, and Signature BC-WG-S6 do not have statistically significant strand-bias (Figure 3.3). In contrast, Signature BC-WG-S1 exhibits a weak T>C strand-bias ($Q = 1.4 \times 10^{-3}$; in all cases Q refers to a q-value, see chapter 7), while Signature BC-WG-S5 is associated with a C>A strand-bias ($Q = 5.2 \times 10^{-7}$). The nature of the mutational process(es) underlying these transcription strand-biases is currently unknown, but it could be due to past activity of transcription-coupled nucleotide excision repair.

The previous assessment of the impact of sequence context on classification of mutational processes is limited to the bases immediately 5' and 3' to each mutated base. However, other sequence motifs close to or distant from the mutant base could be important in defining a mutational process. Here, I extend the sequence context to include the two bases 5' and 3' to each mutated base, which results in 1,536 possible mutated pentanucleotides (*i.e.,* mutational signatures are examined based on the Ξ_{1536} alphabet). The model selection approach is able to find six reproducible mutational signatures based on the 1,536 mutation types. New sequence context dependencies are found in several of the previously identified mutational signatures. Signature BC-WG-2 substitutions at TpCpN trinucleotides are dependent on the next base 5', which

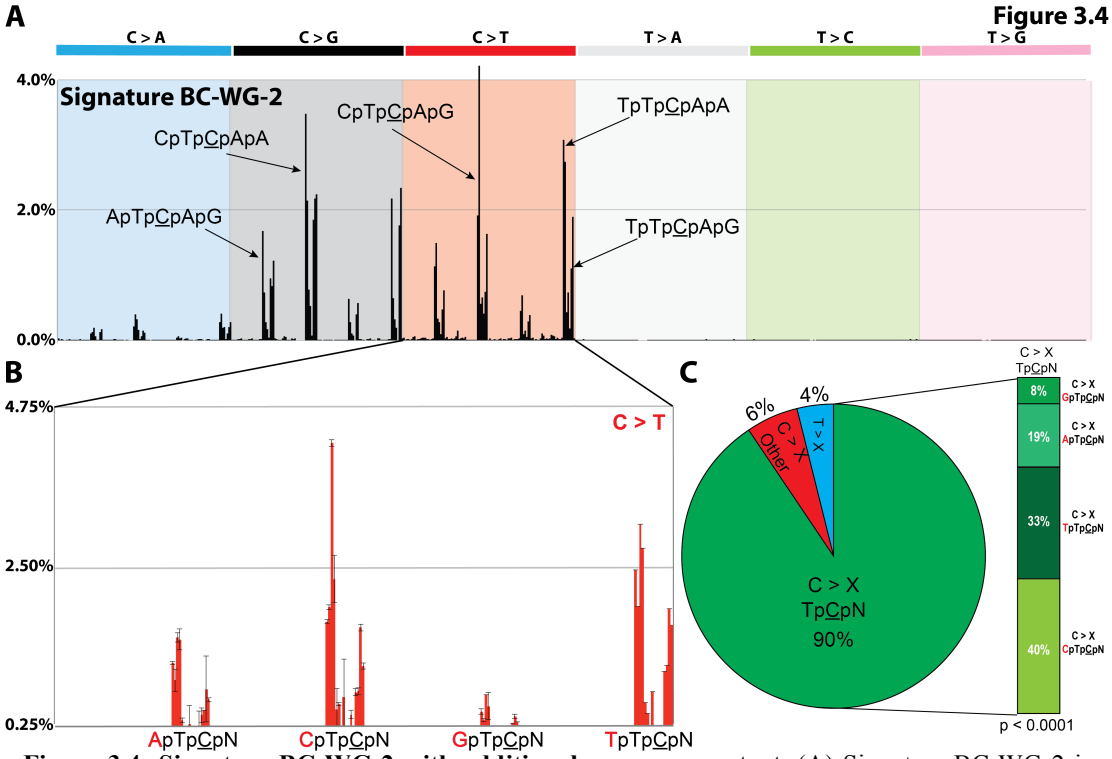


Figure 3.4: Signature BC-WG-2 with additional sequence context. (A) Signature BC-WG-2 is deciphered from the base substitutions (including the two bases 5' and 3' to each mutated base resulting in 1,536 possible mutated pentanucleotides) identified in 119 breast cancer genomes. (B) Detailed view of C>T mutation types in Signature BC-WG-2. (C) Summary of all mutation types caused by Signature BC-WG-2.

is predominantly a pyrimidine (Figure 3.4A and 3.4B). Of all C>X at TpCpN mutations caused by Signature 2, 40% are at CpTpCpN, 33% at TpTpCpN and the remaining 27% are either G or A 5' to the TpCpN trinucleotide (Figure 3.4C). Such a tetranucleotide distribution is highly unlikely to happen purely by chance in the human genome ($Q = 7.1 \times 10^{-14}$). Exactly the same set of observations can be made for Signature BC-WG-3 when additional sequence context is included (data not shown).

In addition to Signatures BC-WG-2 and BC-WG-3, Signature BC-WG-6 also exhibits a strong context dependency when it is examined based on the Ξ_{1536} alphabet (Figure 3.5). Approximately 20% of all somatic mutations due to this mutational signature are T>G at GpGpTpGpG pentanucleotides ($Q = 2.7 \times 10^{-31}$). It should be noted that, when extracted based on the Ξ_{1536} alphabet, Signatures BC-WG-1, BC-WG-4, and BC-WG-5 do not show any specific pentanucleotide patterns.

Figure 3.5

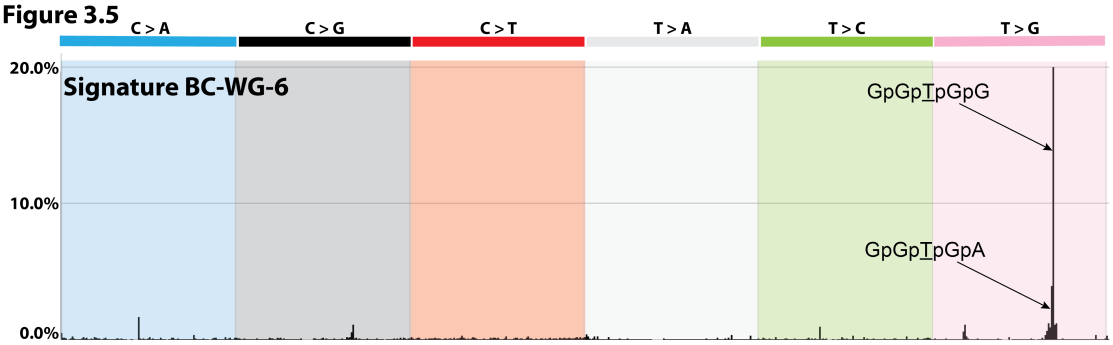


Figure 3.5: Signature BC-WG-6 with additional sequence context. Signature BC-WG-5 is deciphered from the base substitutions (including the two bases 5' and 3' to each mutated base resulting in 1,536 possible mutated pentanucleotides) identified in 119 breast cancer genomes.

3.4 Deciphering the signatures of mutational processes from exome sequencing of breast cancers

The developed computational approach presented in chapter 2 is applied to the mutational catalogues of 884 breast cancer exomes that contain 39,480 somatic substitutions and indels. Mutational signatures are extracted based on the Ξ_{96}, Ξ_{99}, and Ξ_{192} alphabets. The approach reveals three reproducible mutational signatures for all alphabets – termed Signatures BC-EX-S-1, BC-EX-S-2, and BC-EX-S-3 (BC-EX-S stands here for breast cancer exome signature). The numbers of somatic mutations in these exome data (average ~45 somatic mutations per sample) are found to be too low to perform signature analysis using 1,536 mutation types and, as such, no mutational signatures are derived based on the Ξ_{1536} alphabet.

The patterns of somatic substitutions for the signatures extracted from the breast cancer exomes using Ξ_{96} are depicted in Figure 3.6. Signature BC-EX-S-1 is characterized by 60% C>T substitutions predominantly occurring at CpG dinucleotides and 17% T>C mutations with peaks at ApTpN trinucleotides. The pattern of mutations of Signature BC-EX-S-1 (Figure 3.6) closely resembles the one of Signature BC-WG-S1 (Figure 3.1). In fact, these two mutational signatures have a

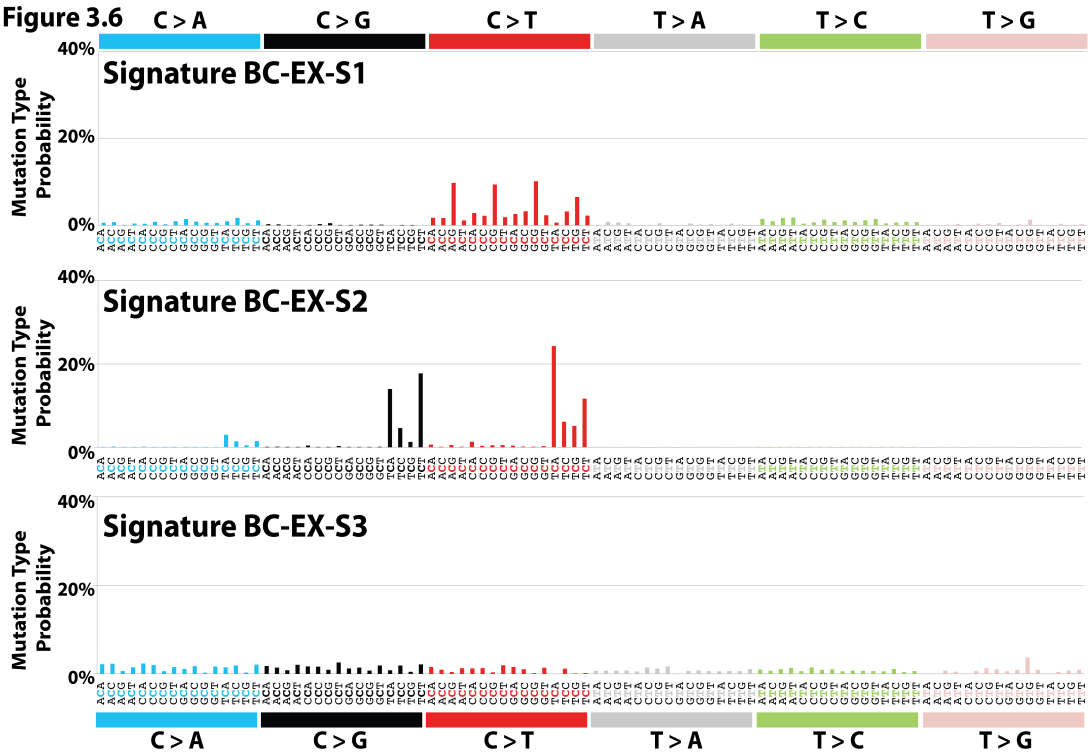


Figure 3.6: Mutational signatures extracted from 884 breast cancer exomes. Signatures of mutational processes are extracted from the mutational catalogues of 884 breast cancer exomes. Three mutational signatures deciphered from the base substitutions (including their immediate 5' and 3' sequence context) identified in the 884 breast cancer exomes. Each signature is depicted on an independent panel, where each type of substitution is displayed in a different colour. Mutational signatures are plotted based on the exome trinucleotide frequency.

Pearson correlation of 0.91. It should be noted that Signature BC-EX-S-1 is extracted from exome sequencing data while Signature BC-WG-S1 is extracted from whole-genome sequencing data. As exome sequencing samples only ~1.5% of the human genome, the examined trinucleotide frequencies in exomes is different than the one found in whole-genome sequencing. Correcting for the trinucleotide frequencies in the exome derived mutational signatures improves the correlation between Signatures BC-WG-S1 and BC-EX-S1 to 0.95.

The pattern of somatic substitutions of Signature BC-EX-S-2 is predominantly C>T, C>G, and C>A mutations at TpCpN trinucleotides. This exome-extracted signature resembles Signature BC-WG-S2, which is extracted from the mutational catalogues of whole-genomes. Nevertheless, Signature BC-EX-S-2 exhibits a strong preference of C>G mutations at TpCpN trinucleotides which is not as pronounced as the one in Signature BC-WG-S2. Thus, Signature BC-EX-S-2 is most likely a linear combination between Signatures BC-WG-S2 and BC-WG-S3 (Figure 3.1 and 3.6).

Signature BC-EX-S-3 is characterized by a flat mutational pattern with only subtle features based on the immediate sequence context. This subtle pattern of

Figure 3.7

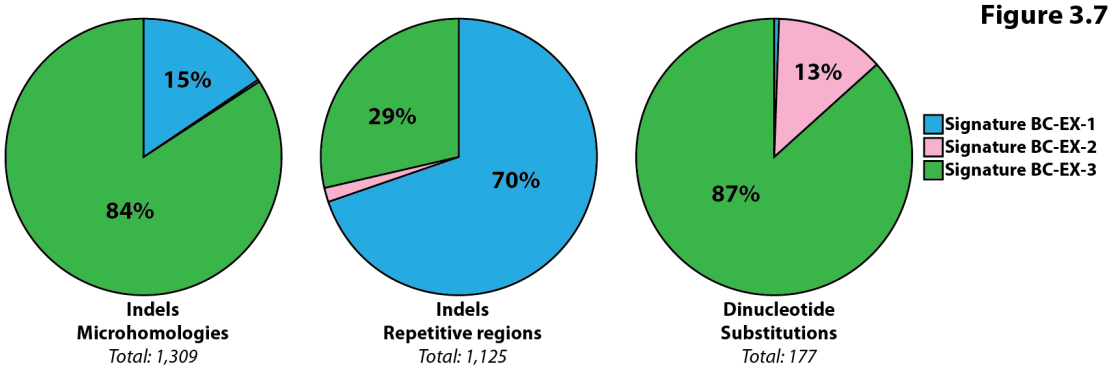


Figure 3.7: Breast cancer exome mutational signatures with indels and dinucleotides. The mutational signatures analysis is extended to incorporate indels at microhomologies, indels at repetitive regions, and dinucleotide substitutions (*i.e.*, Ξ_{99} alphabet). The percentage of mutations attributed to these three additional mutation types is displayed for all signatures that contribute at least 5%. Each signature is displayed in a different colour.

mutations resembles to some degree two of the signatures extracted from whole-genome sequencing data: Signature BC-WG-S-4 (Pearson correlation 0.65, after correcting for trinucleotide context) and Signature BC-WG-S-5 (Pearson correlation 0.49, after correcting for trinucleotide context). Signature BC-EX-S-3 has almost no correlation with any of the other mutational signatures extracted from whole-genome sequencing data. Thus, Signature BC-EX-S-3 is likely a combination of at least two previously identified signatures: Signature BC-WG-S-4 and Signature BC-WG-S-5.

The whole-genome signatures analysis is based on 654,308 somatic mutations and it reveals 6 distinct mutational signatures. In contrast, the exome signatures analysis is based on only 39,480 somatic substitutions and indels (~6% of the whole-genome data) and it reveals only 3 mutational signatures. The performed analyses demonstrate that mutational catalogues from exomes can be used to extract signatures of mutational processes. Furthermore, regardless of the fact that the DNA sequencing and initial bioinformatics analysis of these data were performed by different sequencing centres, the mutational signatures deciphered using exome sequencing are very similar to the ones extracted from whole-genome sequencing data. This illustrates the overall reproducibility of the results together with some vulnerability, particularly when the amount of data are limited or some of the mutational signatures are similar to each other. While using whole-genome sequencing data provides a great resolution for examining common mutational signatures, analysis of smaller, exome derived mutational catalogues (or catalogues from other subcomponents of the genome) may be beneficial as thousands of samples will allow sampling for the activity of mutational processes that are present only in rare cancer cases.

The mutational signatures analysis of breast cancer exomes is extended to evaluate double nucleotide substitutions, indels at microhomologies, and indels at mono/polynucleotide repeats. The results from this analysis are consistent with the indel/dinuc mutational signatures analysis of whole breast cancer genomes (Figure 3.2). Signature BC-EX-3 (which appears to be a mixture of Signatures BC-WG-S-4 and BC-WG-S-5) associated with the majority (>80%) of indels at microhomologies and dinucleotide substitutions as well as with some (~29%) indels at repetitive elements (Figure 3.7). Furthermore, Signature BC-EX-1 accounted for ~70% of indels at repetitive elements (Figure 3.7).

Figure 3.8

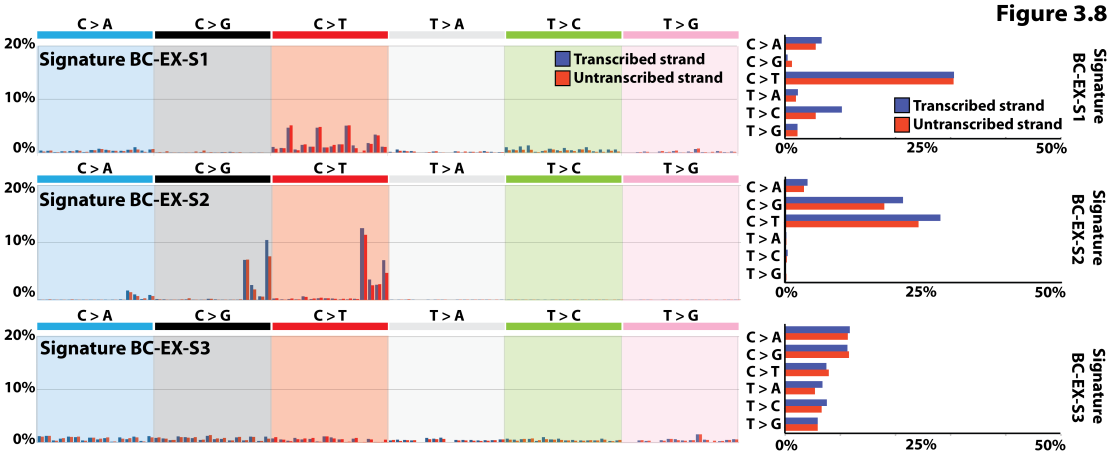


Figure 3.8: Breast cancer exome mutational signatures with strand-bias. Signatures of mutational processes with strand-bias are extracted from the mutational catalogues of 884 breast cancer exomes. Three mutational signatures are deciphered from the base substitutions (including their immediate 5' and 3' sequence context). Each signature is depicted on an independent panel, where each type of substitution is highlighted in a different colour. The probability of a mutation to occur on the transcribed strand is depicted in blue, while red is used to display the probability of a mutation to occur on the untranscribed strand. Mutational signatures are plotted based on the exome trinucleotide frequency.

Analysis of smaller, exome derived mutational catalogues (or catalogues from other subcomponents of the genome) may also be useful in detecting biologically revealing features of mutational processes that are particular to coding, transcribed, non-transcribed, or other functionally distinct regions. Consistent with the strand-bias analysis of whole-genome cancer samples, Signature BC-EX-S1 exhibited a weak T>C strand-bias ($Q = 7.2 \times 10^{-4}$). In contrast, no C>A strand-bias is observed in any of the mutational signatures derived from exome sequences (Figure 3.8). This could be due to the lack of somatic mutations to definitively separate Signature BC-EX-S-3 into two distinct mutational signatures. Further, incorporating transcriptional strand in the analysis of the 884 breast cancer exomes reveals strand-bias in BC-EX-S-2 for C>T and C>G mutations with a preference for specific trinucleotide context, *i.e.,*

TpCpT (Figure 3.8). However, this strand-bias is not observed in the versions of Signature BC-EX-S-3 (*i.e.,* Signatures BC-WG-S-2 and BC-WG-S-3) extracted from whole cancer genome sequences, which include complete footprints (including introns and untranslated exons) of protein coding genes, suggesting that the underlying mechanism generating strand-bias is restricted to exons (Figures 3.8 and 3.3). Examining only the exon compartments of the whole cancer genome sequences reveals the presence of this strand-bias in samples with substantial exposure to Signature BC-WG-S-2 and/or Signature BC-WG-S-3, supporting this conclusion. This result is biologically surprising and the mechanism underlying this difference in strand-bias between exons and introns is currently unknown.

3.5 Deriving and validating consensus mutational signatures in breast cancer

Figure 3.9

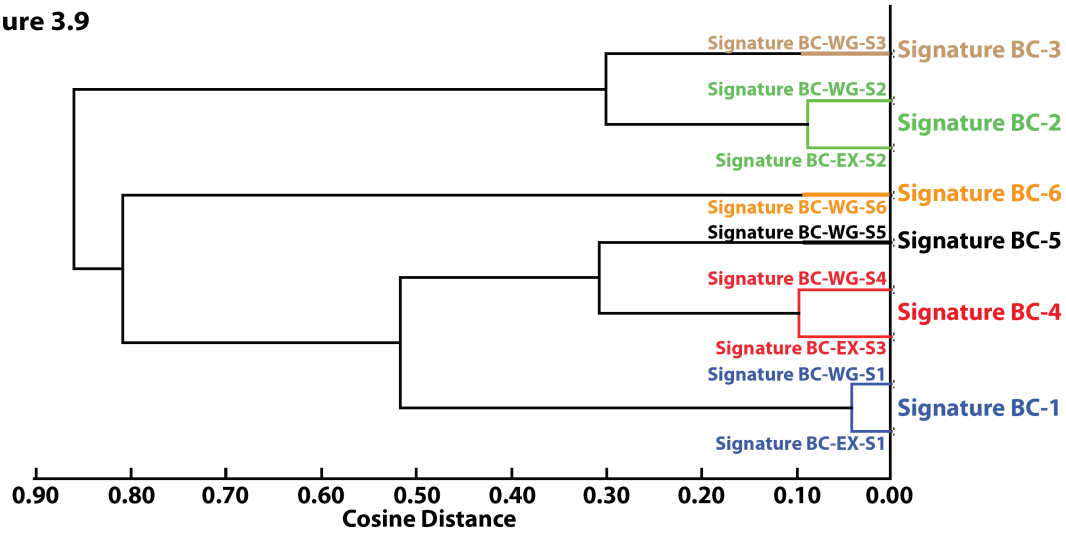


Figure 3.9: Clustering of breast cancer signatures derived from whole-genome and exome data. The originally deciphered mutational signatures are displayed inside the dendrogram near their respective branches. The consensus mutational signatures are displayed on the right-hand side of the dendrogram. Each of the six unique clusters is displayed in a distinct colour. Cosine distance threshold for separating the signatures into clusters is set at 0.09. Note that any threshold between 0.09 and 0.29 results in exactly the same clusters.

In the previous two sections, the signatures of the operative mutational processes in breast cancer are extracted by performing two independent analyses. One encompasses 654,308 somatic substitutions and indels derived from the mutational catalogues of 119 whole breast cancer genomes and reveals the existence of 6 mutational signatures. The second analysis examines only 39,480 somatic mutations from the mutational catalogues of 884 breast cancer exomes and it reveals the existence of 3 mutational signatures. While the patterns of somatic mutations between the signatures extracted from genomes and exomes are very similar, in this section, I

use the previous two analyses and leverage unsupervised hierarchical clustering to derive consensus mutational signatures that are operative in breast cancer. The previously extracted 9 mutational signatures (3 from exome mutational catalogues and 6 from genome mutational catalogues) are clustered using a cosine distance (Figure 3.9). The exome derived mutational signatures are re-normalized towards the genome trinucleotide frequency prior to clustering and a threshold of 0.09 is used to separate the original 9 mutational signatures into 6 unique consensus clusters (Figure 3.9).

The value of 0.09 is selected as a conservative measure for the different mutational signatures operative in breast cancer. This threshold is low enough to not cluster mutational signatures with different characteristics (*e.g.*, Signature BC-WG-S5 which exhibits C>A strand-bias and Signature BC-WG-S4 which is associated with indels at microhomologies) and it is high enough to cluster together extremely similar mutational signatures (*e.g.*, Signature BC-EX-S1 and Signature BC-WG-S1, which have a Pearson correlation of 0.95). Nevertheless, this threshold may result in a conservative estimate of the consensus mutational signatures as it may be clustering and mixing together distinct mutational patterns.

Each consensus mutational signature is derived using a weighted average of the signatures belonging to its respective cluster. For example, Signature BC-2 is constructed as a weighted average of genome Signature BC-WG-S2, which accounts for 152,762 somatic mutations, and exome Signature BC-EX-S2, which accounts only for 19,922 somatic mutations (Figure 3.9). As the majority of somatic mutations are found in the whole-genome sequencing data, in this case, the patterns of somatic mutations in the consensus mutational signatures are visually indistinguishable from the ones derived from whole-genome sequencing data (Figure 3.1). Thus, the pattern of mutations of the consensus Signature BC-2 is very similar to the one of Signature BC-WG-S2. It should be noted that the number of mutations attributed to a consensus mutational signature in a sample is set to the number of mutations of the original mutational signature identified in this sample and belonging to the cluster used to derive the consensus mutational signature. For example, Signature BC-2 contributes 69 somatic mutations in exome sample PD6042a as this is the number of somatic mutations attributed to Signature BC-EX-S2 in this sample. In total, Signature BC-2 accounts for 172,684 somatic mutations in the exome and genome breast cancer data (~24.9% of all mutations used in this breast cancer analysis).

In addition to deriving the consensus mutational signatures, in this section, I validate these signatures to check whether any of them might be due to sequencing artefacts or bioinformatics analysis. Validating a mutational signature requires ensuring that a large set of somatic mutations attributed to its pattern is genuine in at least one sample. Validation is complicated as multiple mutational processes are usually operative in most cancer samples, and thus every individual somatic mutation can be probabilistically assigned to several mutational signatures. To overcome this limitation, when possible, I examine the curated dataset for samples that are predominantly generated by one mutational signature (*i.e.*, more than 50% of the somatic mutations in the sample belong to an individual mutational signature) and for which validation data were available. The optimal sample for validating each of the six mutational signatures is identified and a subset of somatic mutations characteristic for this signature (*e.g.*, C>T and C>G substitutions at TpC dinucleotides for Signature BC-2) are chosen for validation through re-sequencing with an orthogonal sequencing technology.

Mutational Signature	Validation Status	Total Mutations in Sample	Total Mutations by Signature	Examined Mutations	Validated Mutations
Signature BC-1	PASS	58	55	58	56 (97%)
Signature BC-2	PASS	76	75	76	72 (95%)
Signature BC-3	PASS	8,612	5,697	200	190(95%)
Signature BC-4	PASS	70	65	70	69 (99%)
Signature BC-5	PASS	4,514	1,558	250	227 (91%)
Signature BC-6	FAIL	11,869	7,955	100	2(2%)

Table 3.2: Validating consensus mutational signatures found in breast cancer. Validation is performed with an orthogonal sequencing approach. The precise validation approach is outlined in the text.

The results reveal that Signatures BC-1, BC-2, BC-3, BC-4 and BC-5 are most likely genuine biological patterns of somatic mutations as they have validation rates of more than 90% (Table 3.2). In contrast, Signature BC-6 is probably due to a sequencing artifact as 98% of the mutations characteristic for this signature (*i.e.*, T>G at GpTpG trinucleotides) failed to validate using an alternative orthogonal sequencing approach. Further investigation into this signature reveals that it is an artifact specific to the configuration of some Illumina sequencing machines at the Wellcome Trust Sanger Institute.

3.6 Prevalence of mutational processes in breast cancer samples

In the previous sections of this chapter, I extract mutational signatures separately from exome and genome sequencing data, and identified the consensus mutational signatures operative in breast cancer. However, the developed computational approach (chapter 2) also allows quantifying the number of somatic mutations attributed to each mutational signature in each cancer sample.

Figure 3.10

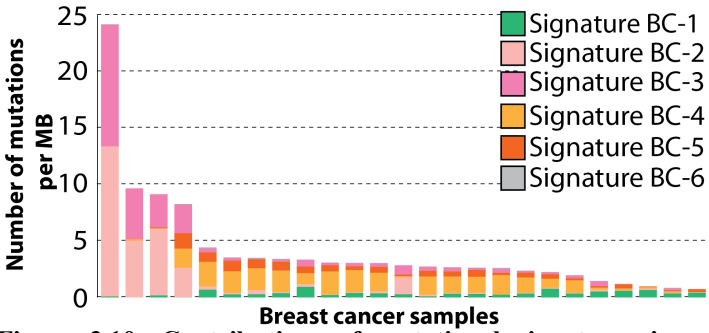


Figure 3.10: Contributions of mutational signatures in a selected set of 25 breast cancer samples. Each sample is displayed as a column with a height corresponding to the number of somatic mutations per megabase found in this sample. Every column is proportionately coloured to reflect the percentage of mutations attributed to different mutational signatures.

An example of a selected set of 25 cancer samples is displayed in Figure 3.10 (note that the contributions of all mutational signatures in all examined cancer samples is provided in Appendix V). This plot reveals the diversity of the activity of the mutational processes underlying the signatures identified in these breast cancer samples. For example, a small minority of samples exhibit a hypermutator phenotype with somatic mutational patterns best explained by Signatures BC-2 and BC-3 (Figure 3.10). A further subset of samples seems to be overwhelmed by the activity of the mutational process underlying Signature BC-4. In contrast, Signature BC-1 is

Figure 3.11

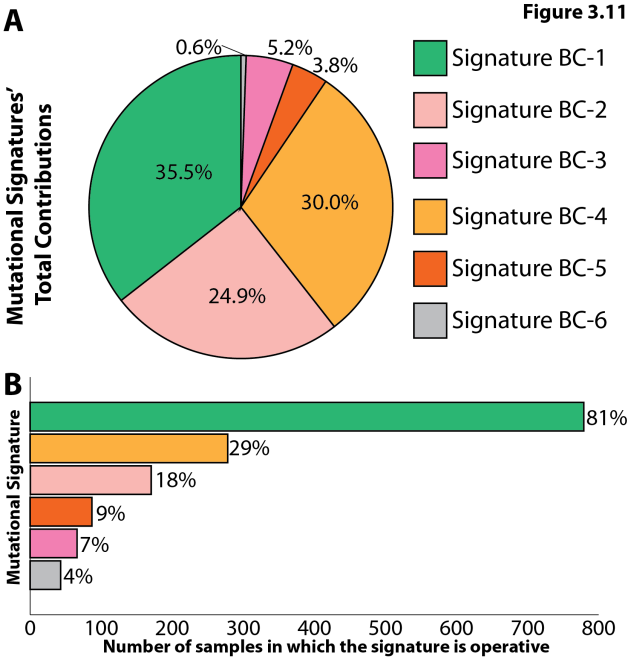


Figure 3.11: Summary of the contributions of the mutational signatures in breast cancer. ***(A)*** Percentage of total mutations contributed by each of the operative mutational signatures. ***(B)*** Percentage and number of samples in which each mutational signature contributes significant number of somatic mutations. For most signatures, significant number of mutations in a sample is defined as more than 100 substitutions or more than 25% of all mutations in that sample. Mutational signatures are displayed in distinct colours.

ubiquitously found at low levels in almost every examined sample (Figure 3.10).

In addition to examining the contributions of mutational signatures at the level of individual samples, one can evaluate the contributions of these signatures across all breast cancer samples and thus provide a mutational signature summary (Figure 3.11). Such an evaluation reveals that while Signature BC-1 accounts for only ~35% of all somatic mutations, it is the most prevalent mutational signature in breast cancer as it is found in 81% of all examined samples (Figure 3.11). In contrast, the next most prevalent signature is Signature BC-4, which is found in only 29% of the samples. Examining the prevalence of mutational signatures across breast cancer samples provides the means to propose etiologies underlying these mutational signatures based on statistical associations.

3.7 Etiology of the consensus mutational signatures in breast cancer

The analysis of breast cancer samples reveals the signatures of 6 distinct mutational processes. However, no molecular mechanisms or etiologies are proposed here for the identified mutational signatures. In principle, several approaches can be leveraged to make propositions for the mechanisms of the underlying mutational mechanisms. In this section, I consider potential mechanisms or underlying causes by comparing signatures with mutation patterns of known causation in the scientific literature or by associating contributions of mutational signatures with epidemiological and biological features specific for breast cancer.

The mutational pattern of Signature BC-1 is predominantly C>T mutations occurring at CpG dinucleotides. This signature is likely due to deamination of 5-methylcytosine, a relatively well-characterized endogenous mutational process present in most normal and neoplastic cells (chapter 1).

Signature BC-2 exhibits predominantly C>T mutations occurring at TpC dinucleotides, while Signature BC-3 generates mostly C>G substitutions occurring at TpC dinucleotides. On the basis of similarities in the sequence context of cytosine mutations caused by *APOBEC* deaminases in experimental systems, these two mutational signatures may be attributable to the activity of *APOBEC1*, *APOBEC3A* and/or *APOBEC3B* (chapter 1). Previous experimental studies have demonstrated that the activity of these proteins results in enzymatic deamination of cytosine to thymine at TpC dinucleotides and it has been speculated that these C>T mutations arise through replication across the uracil. Furthermore, it has been shown that these

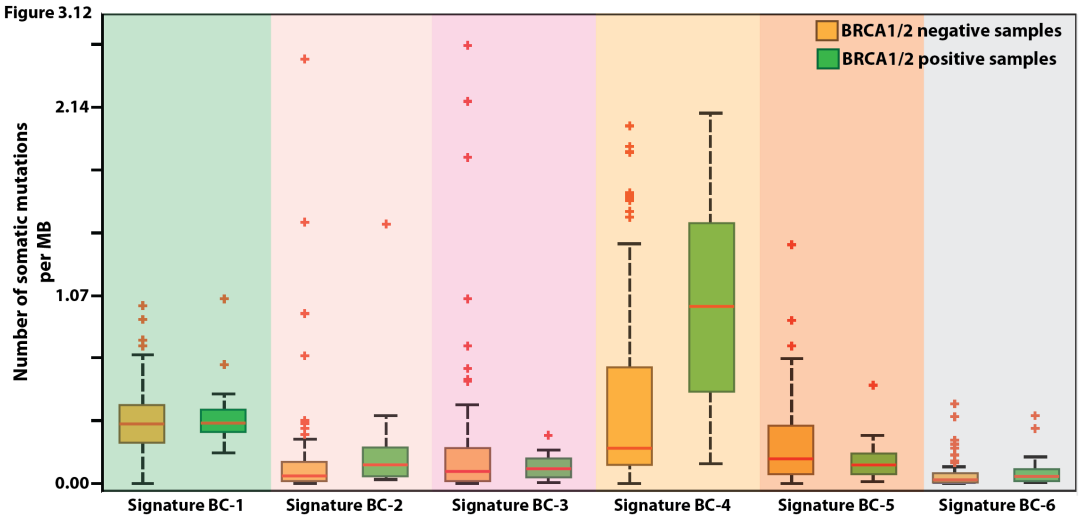


Figure 3.12: Samples harbouring *BRCA1/2* mutations and contributions of mutational signatures. Samples are separated into two sets: *BRCA1/2* positive samples (*i.e.*, with BRCA1/2 mutations, green) and *BRCA1/2* negative samples (*i.e.*, without BRCA1/2 mutations orange). A box plot of the mutations contributed by each mutational signature is displayed for each of the two sets. Outliers with more than 2.5 mutations per megabase are not shown but they are included in the statistical analysis. The only statistically significant difference in signature's contributions between the *BRCA1/2* positive and negative sets is the one due to Signature BC-4 ($Q = 1.6 \times 10^{-8}$).

deaminases can also generate C>G substitutions at TpC dinucleotides and it has been suggested that this mutational pattern is generated when an *APOBEC* deaminated cytosine is excised by uracil-DNA glycosylase with subsequent non-templated DNA replication across the abasic site by *REV1* (Taylor et al., 2013). Thus, Signature BC-2 is likely due to the activity of the *APOBEC* family of deaminases, while Signature BC-3 encompasses an interaction between *APOBEC* enzymes and *REV1*.

Substantial numbers of larger deletions (up to 50 bp) with overlapping microhomology at breakpoint junctions are found in some breast cancer samples with major contributions from Signature BC-4 (Figure 3.2). A subset of breast cancer cases is known to be due to inactivating mutations in *BRCA1* and *BRCA2*, and the presence of Signature BC-4 is strongly associated ($Q = 1.6 \times 10^{-8}$) with *BRCA1* and *BRCA2* mutations (Figure 3.12). No other mutational signature associated with the numbers of mutations in samples harbouring *BRCA1* and/or *BRCA2* mutations (Figure 3.12). *BRCA1* and *BRCA2* are implicated in homologous-recombination-based DNA double-strand break repair. Abrogation of their functions results in recruitment of non-homologous end-joining mechanisms, which can use microhomology at rearrangement junctions to re-join double-strand breaks, to take over DNA double-strand break repair. Indeed, almost all cases with *BRCA1* and *BRCA2* mutations showed a large contribution from Signature BC-4. However, some cases with a substantial contribution from Signature BC-4 do not have *BRCA1* and *BRCA2*

mutations, suggesting that other mechanisms of *BRCA1* and *BRCA2* inactivation or abnormalities of other genes may also generate this mutational signature.

Evaluating the enrichment of mutational signatures based on the molecular subtypes of breast cancer reveals that estrogen receptor negative breast cancer samples have significantly higher numbers of mutations due to Signature BC-4, $Q = 7.9 \times 10^{-5}$, and Signature BC-5, $Q = 1.6 \times 10^{-6}$ (Figure 3.13). No other molecular subtype associated with the numbers of somatic mutations attributed to any other mutational signature. Estrogen receptor negative breast cancer samples are enriched for *BRCA1* and *BRCA2* mutations. To evaluate whether the differences of contributions of mutational signatures are due to *BRCA1/2* mutations, these samples are re-examined after stratification. *BRCA1/2* wild-type samples do not show statistically significant differences based on their estrogen receptor status for Signature BC-4 ($Q = 0.09$). However, estrogen receptor negative *BRCA1/2* wild-type samples have significantly higher numbers of mutations attributable to Signature BC-5 when compared to estrogen receptor positive *BRCA1/2* wild-type breast cancers ($Q = 3.8 \times 10^{-3}$).

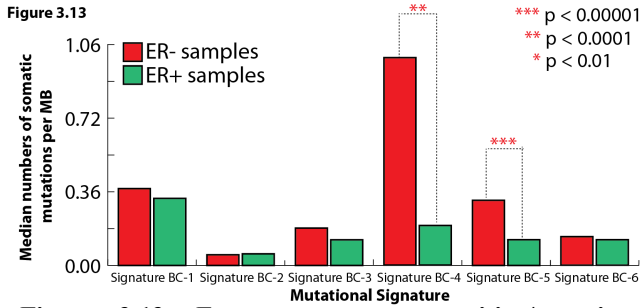


Figure 3.13: Estrogen receptor positive/negative samples and contributions of mutational signatures. Samples are separated into two sets: estrogen receptor negative samples (red) and estrogen receptor negative samples (green). The distributions of somatic mutations between the two sets are compared for each of the mutational signatures.

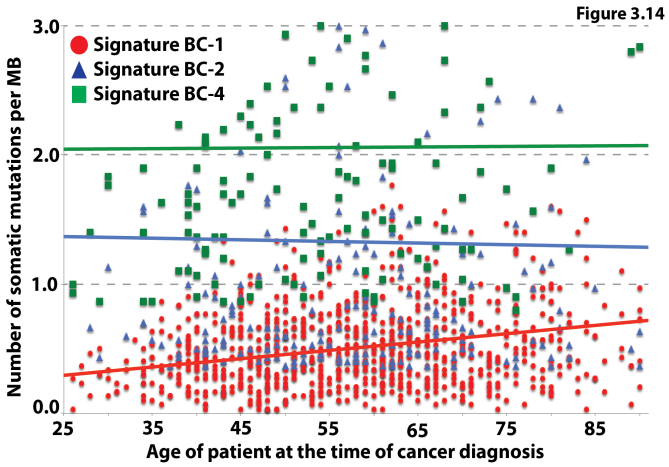


Figure 3.14: Age of diagnosis and mutations due to different mutational signatures. Each sign corresponds to a contribution of a given mutational signature for a patient at a given age. From the six mutational signatures identified in breast cancer, only Signature BC-1 (shown in red) correlates with age of diagnosis. Signatures BC-2 (blue) and BC-4 (green) are shown to illustrate the lack of correlation of other mutational signatures.

The performed validation experiments (Table 3.2) indicate that Signature BC-6 is most likely a centre specific sequencing artifact.

Lastly, I evaluate the correlations between age of diagnosis and the number of mutations attributable to each

signature in each sample. Only Signature BC-1 exhibited a strong positive correlation with age of diagnosis, Q = 1.5×10^{-8} (Figure 3.14). The mutations in a cancer genome may be acquired at any stage in the cellular lineage from the fertilized egg to the sequenced cancer cell. The correlation with age of diagnosis is consistent with the hypothesis that a substantial proportion of Signature BC-1 mutations in cancer genomes have been acquired over the lifetime of the cancer patient, at a relatively constant rate that is similar in different people, probably in normal somatic tissues.

3.8 Discussion

In this chapter of the thesis, I examine the mutational catalogues of 119 breast cancer genomes as well as 884 breast cancer exomes. Mutational signatures are deciphered separately from genome and exome sequencing data. The signatures analysis incorporated somatic single base substitutions and their immediate sequencing context as well as indels at mono/polynucleotide repeats, indels at microhomologies, and dinucleotide substitutions.

The identified genome-based and exome-based mutational signatures are used to derive the 6 consensus breast cancer signatures. Validation using an orthogonal sequencing technology reveals that one of these mutational signatures is most likely due to a sequencing artifact, while the remaining five are most likely genuine. An etiology is proposed for each of these five mutational signatures based on similarities of the mutational patterns with experimental data previously reported in the literature or a statistical association with a specific molecular phenotype.

Lastly, it should be noted that one of the objectives of this chapter is to serve as an exemplar for performing mutational signatures analysis in a cancer type. The next chapter presents analogous analyses performed for another 29 types of human cancer.

Chapter 4

Signatures of mutational processes in human cancer

4.1 Introduction

In the previous chapter of this thesis, I applied a newly developed computational approach to somatic mutational data derived from breast cancer genome and exome sequences, which revealed multiple signatures with distinct patterns of somatic mutations. Comparing these mutational patterns with the scientific literature as well as statistically associating them with molecular phenotypes provided an indication for the etiology of the mutational processes responsible for these signatures. In this chapter, I expand the scope of the mutational signatures analysis and apply the developed computational framework to 30 distinct cancer types. The approach taken in this chapter is analogous to the one used for breast cancer in the previous chapter; mutational signatures are extracted from mutational catalogues based on the Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536} alphabets separately for each cancer type (with further separation for samples derived from whole-genome and exome sequencing in a single cancer type). The deciphered mutational signatures are hierarchically clustered, as demonstrated in the previous chapter, to derive the consensus mutational signatures in human cancer. In this chapter, I will focus on extracting the signatures of the operative mutational processes in 7,042 samples across 30 cancer classes, examining their patterns of somatic mutations, and discussing them in the context of the different cancer types in which they are found. It should be noted that this chapter does not discuss the potential etiologies of the identified consensus mutational signatures since these will be the focus of chapter 5.

4.2 Data generation and filtering of mutational catalogues

Similarly to breast cancer, no data were generated solely for the purposes of this thesis. Rather, I curate already identified somatic mutations from freely available previously published and (at the time) unpublished data. The curated freely available data are taken from three distinct sources:

- The data portal of The Cancer Genome Atlas (TCGA)
- The data portal of the International Cancer Genome Consortium (ICGC)
- Previously published in peer-review journals cancer genomics mutational data: (Agrawal et al., 2011; Barbieri et al., 2012; Berger et al., 2011; Biankin et al., 2012; Dulak et al., 2013; Fujimoto et al., 2012; Govindan et al., 2012; Grasso et al., 2012; Gui et al., 2011; Imielinski et al., 2012; Jiao et al., 2011; Jones et al., 2012a; Jones et al., 2010; Krauthammer et al., 2012; Le Gallo et al., 2012; Liu et al., 2012a; Liu et al., 2012b; Love et al., 2012; Morin et al., 2011; Nik-Zainal et al., 2012; Peifer et al., 2012; Puente et al., 2011; Pugh et al., 2013; Rudin et al., 2012; Sausen et al., 2013; Seo et al., 2012; Seshagiri et al., 2012; Shah et al., 2012; Stephens et al., 2012; TCGA, 2012; Wang et al., 2011; Wei et al., 2011; Wiegand et al., 2010; Wu et al., 2011; Zang et al., 2012; Zhang et al., 2013)

The unpublished data are generated internally by the Cancer Genome Project (CGP) or donated by collaborating investigators that were willing to participate in the performed large-scale pan-cancer mutational signatures analysis. The majority of exome data are taken from the ICGC data portal, TCGA data portal, or from the abovementioned published peer-reviewed publications. In contrast, the majority of whole-genomes are previously unpublished data. A summary of the number of

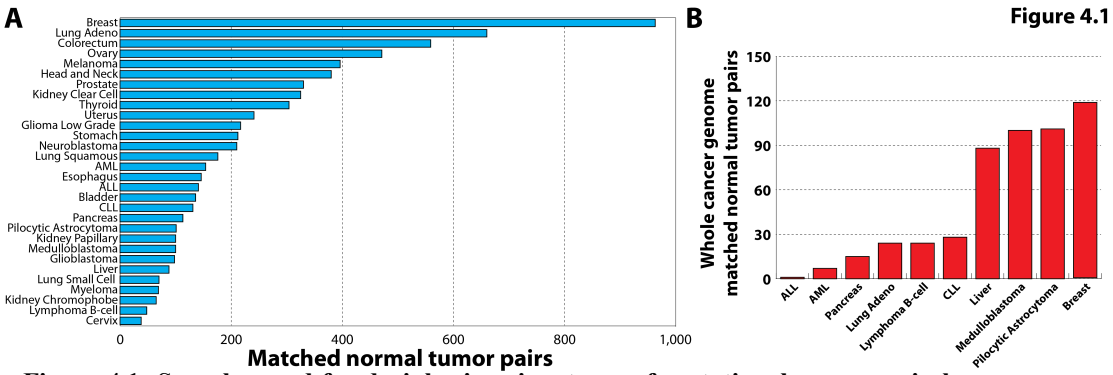


Figure 4.1

Figure 4.1: Samples used for deciphering signatures of mutational processes in human cancer. Mutational catalogues of ***(A)*** 7,042 primary cancers derived from 30 different cancer types are examined for mutational signatures, including ***(B)*** 507 whole cancer genomes with matched normal pairs.

samples based on cancer types is shown in Figure 4.1; in addition, a complete list including all samples, all examined cancer types, and their respective data sources is provided in Appendix II.

In total, I compiled the mutational catalogues of 7,042 primary cancers of 30 different classes: 507 from whole-genome and 6,535 from exome sequences (Figure 4.1). In all cases, normal DNAs from the same individuals have been sequenced to establish the somatic origin of the variants. The somatic mutations are extensively filtered to remove germline polymorphisms and sequencing artefacts as previously described for breast cancer (see chapter 3) and the final filtered dataset contains 4,938,362 somatic substitutions and small insertions/deletions (indels). The somatic mutations found in these 7,042 matched normal tumour pairs are used to decipher the mutational signatures from catalogues based on the Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536} alphabets (see below).

Examining the mutational catalogues of the 7,042 primary cancers revealed that the prevalence of somatic substitutions and indels is highly variable between and within cancer classes, ranging from about 0.001 somatic mutations per megabase to more than 400 somatic mutations per megabase (Figure 4.2). Certain childhood cancers carried fewest mutations whereas cancers related to chronic mutagenic exposures such as lung (tobacco smoking) and malignant melanoma (exposure to ultraviolet light) exhibited the highest prevalence. This variation in mutation prevalence is attributable to differences between cancers in the duration of the cellular

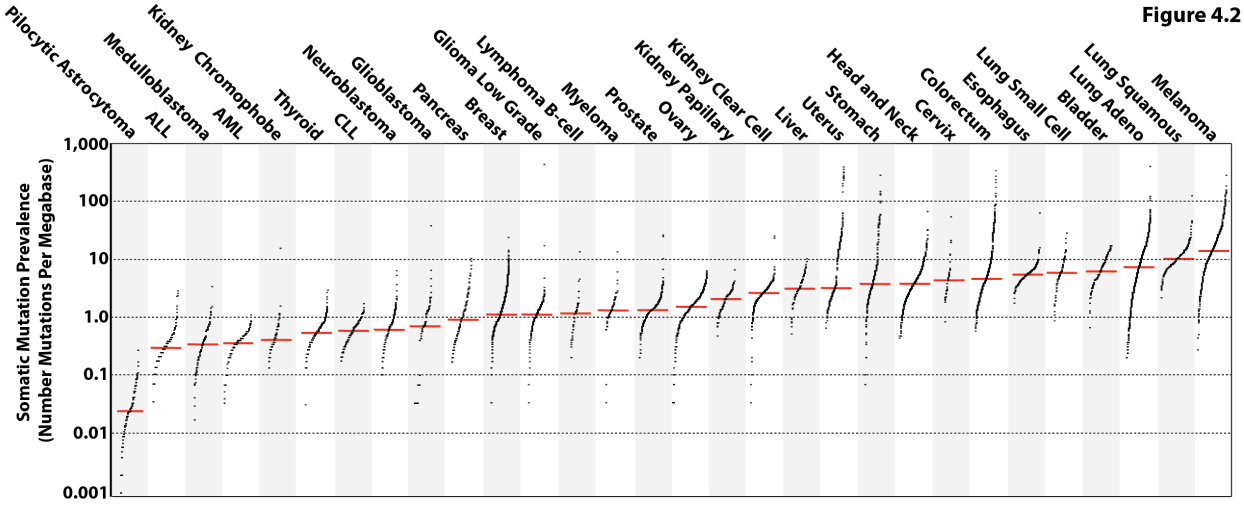


Figure 4.2: Mutational burden in human cancer. Every dot represents a sample whereas the red horizontal lines are the median numbers of mutations in the respective cancer types. The vertical axis (log scaled) shows the number of mutations per megabase whereas the different cancer types are ordered on the horizontal axis based on their median numbers of somatic mutations. ALL stands for acute lymphoblastic leukaemia; AML for acute myeloid leukaemia; CLL for chronic lymphocytic leukaemia.

lineage between the fertilized egg and the sequenced cancer cell and/or to differences in somatic mutation rates during the whole or parts of that cellular lineage (Stratton et al., 2009).

4.3 Deciphering signatures of mutational processes in 30 human cancer types

Mutational signatures are extracted using the previously defined four mutational alphabets: Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536} (Appendix I). Briefly, Ξ_{96} examines all somatic substitutions and additionally includes information on the sequence context of each substitution. This classification has 96 possible mutations since there are six classes of base substitution C>A, C>G, C>T, T>A, T>C, T>G (all substitutions are referred to by the pyrimidine of the mutated Watson-Crick base pair) and the bases immediately 5' and 3' to each mutated base are incorporated. The Ξ_{99} alphabet extends the Ξ_{96} alphabet by incorporating three additional mutation types: dinucleotide substitutions, indels at repetitive elements, and indels at microhomologies. The Ξ_{1536} alphabet examines substitutions and their immediate sequence context; however, this alphabet incorporates two bases 5' and 3' to each mutated base instead of the one base used in the Ξ_{96} alphabet. Lastly, the Ξ_{192} alphabet examines all somatic mutations in transcribed regions of the human genome. This alphabet has all the features of Ξ_{96} but it also incorporates information on whether the mutation is occurring on the transcribed or the untranscribed strand of protein-coding genes. The 96 and 1,536 substitution classifications are particularly useful for distinguishing mutational signatures which cause the same substitutions but in different sequence contexts. In contrast, the Ξ_{99} alphabet allows the evaluation of the amount of indels and dinucleotide substitutions caused by different mutational processes, while the Ξ_{192} alphabet is leveraged to evaluate the activity of repair processes operative on the transcribed regions of the human genome.

Mutational signatures are deciphered independently for each of the 30 cancer types following the same analysis procedure as the one previously used in breast cancer (chapter 3). In total, 106 mutational signatures based on the Ξ_{96} alphabet are extracted from these 30 cancer types. These mutational signatures are clustered using an unsupervised hierarchical clustering, where a cosine distance is used as a measure

for comparing mutational signatures (Figure 4.3). Any signature derived from exome sequencing data is re-normalized towards the genome trinucleotide frequency prior to applying the clustering procedure.

A threshold of 0.18 is used to separate the original 106 mutational signatures into 27 unique clusters. This threshold is conservatively selected based on visual inspection and prior biological knowledge. More specifically, annotation 1 in Figure 4.3 shows the separation of two mutational patterns overwhelmed by C>T mutations with a difference in their immediate sequence context (later referred to as Signature 7 and Signature 11, Figure 4.5). The upper branch of annotation 1 contains patterns of mutations that are consistent with exposure to ultraviolet light, while the signatures in the lower branch are exclusively found in samples that are treated with an alkylating agent (see chapter 5). Since these two sets of mutational signatures have distinct

Figure 4.3

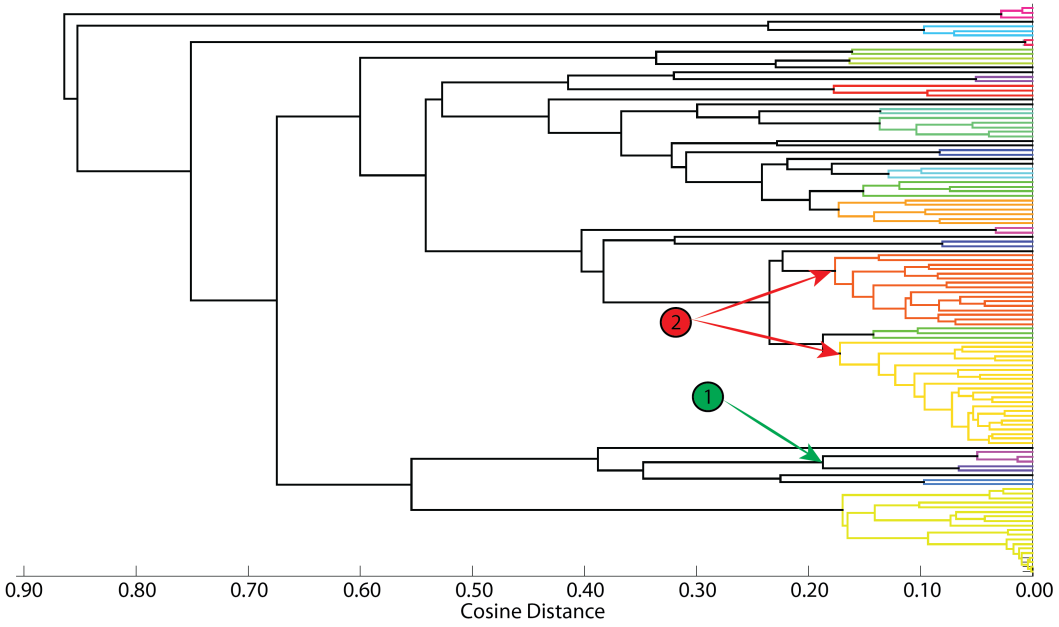


Figure 4.3: Clustering of mutational signatures. Clustering of 106 original mutational signatures deciphered from the mutational catalogues of 7,042 cancer samples. Each of the 27 unique clusters is displayed in a different colour. The cosine distance threshold for separating the signatures into clusters is set at 0.18 based on annotation 1 (green) and annotation 2 (red).

patterns and etiologies, the selected clustering threshold needs to separate them and as such it needs to be lower than 0.184. Visual inspection of clustering of the original mutational signatures (annotation 2, Figure 4.3) shows that all of these signatures possess similar patterns of somatic mutations (*e.g.,* C>T at CpG). However, these patterns are contaminated since, most probably, they cannot be extracted with the

same accuracy from different datasets (*e.g.,* less than 40 samples are used for signature analysis in cervical cancer versus the almost 1,000 samples used for signature analysis in breast cancer). To ensure that these visually similar mutational signatures cluster together, a threshold of 0.18 is selected. It should be noted that visual examination may be misleading and it may result in clustering mutational signatures that are different. Nevertheless, this analysis provides a conservative estimation of the mutational signatures found in human cancer and it is foreseeable that some of the reported mutational signatures are, in fact, mixtures of multiple distinct signatures. Only further samples across all types of human cancer will allow a further separation of these mutational signatures. As was previously performed for breast cancer, each consensus mutational signature is derived using a weighted average of the signatures belonging to its respective cluster and the number of somatic mutations attributed to a consensus mutational signature in a sample is set to the number of mutations of the original signature found in that sample.

In addition to deciphering mutational signatures using mutational catalogues based on the Ξ_{96} alphabet, an analysis is performed also for the Ξ_{99}, Ξ_{192}, and Ξ_{1536} alphabets. In all cases the consensus signatures results from the Ξ_{99} and Ξ_{192} catalogues are consistent with the previous observations based on the Ξ_{96} alphabet. However, deciphering mutational signatures for the Ξ_{1536} alphabet produced results only for a few of the cancer types (see below). The inability to decipher mutational signatures using the 1,536 mutation types is, most probably, due to the absence of sufficient numbers of somatic mutations in the examined mutational catalogues. This is perhaps unsurprising as ~93% of the examined mutational catalogues are derived from exome sequences, which harbour very few somatic mutations. Furthermore, the majority of whole-genome sequences are from childhood cancers and they have a low mutational burden (Figure 4.2).

Figure 4.4

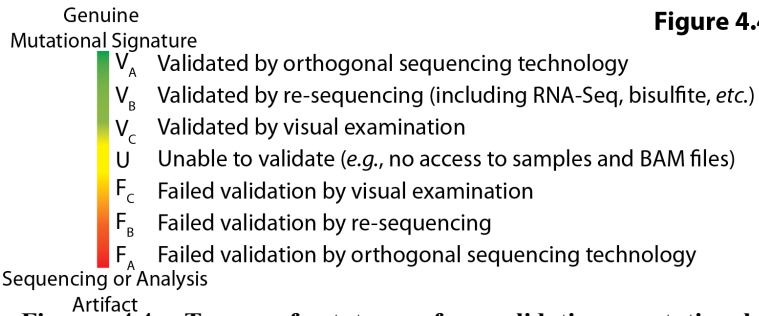


Figure 4.4: Types of statuses for validating mutational signatures.

4.4 Validating consensus mutational signatures

Validating a signature of a mutational process requires ensuring that a large set of somatic mutations, with a mutational spectrum resembling the one of the mutational signature of interest, is genuine in at least one sample in which this process is operative. As previously discussed with regard to breast cancer (chapter 3), validation is complicated as various mutational processes are found in a single cancer sample and, as such, every individual somatic mutation can be probabilistically assigned to several mutational signatures. In this analysis, I leveraged the same approach as the one used in validating mutational signatures in breast cancer: the dataset is examined for samples that are predominantly generated by one mutational signature (*i.e.,* more than 50% of the somatic mutations in the sample belong to an individual mutational signature). Since I did not have access to the biological samples, I mostly relied on previously performed validation experiments (*e.g.,* samples in TCGA sequenced by two different groups using two different next-generation sequencing technologies) as well as visual validation of BAM files by an experienced curator. Based on the data, I identified the optimal available sample for every mutational signature and attempted to validate a subset of somatic mutations attributed to this signature using one of three methods (Figure 3.3):

- Validation by re-sequencing with an orthogonal sequencing technology
- Validation by re-sequencing with the same sequencing technology (including RNA-Seq, bisulfide sequencing*, etc.*)
- Validation by visual examination of somatic mutations performed by an experienced curator using a genomic browser and BAM files for both the tumour and its matched normal

When possible, somatic mutations are validated by either re-sequencing with orthogonal technology or re-sequencing using the same sequencing technology. I resorted to visual validation only when there is no other possibility for validating a mutational signature. 22 of the 27 consensus mutational signatures were validated (Table 4.1 and Figure 4.4). Three of the mutational signatures failed validation (termed Signatures R1 to R3), while another two mutational signatures were not validated (termed Signatures U1 and U2) due to lack of access to biological samples and BAM files for the samples with sufficient numbers of somatic mutations generated by these two mutational signatures. A validation summary for all consensus

mutational signatures is provided in Table 4.1. The validated mutational signatures are depicted in Figure 4.5, while the signatures that failed validation and the signatures that remain with unknown validation status are shown respectively in Figure 4.6 and Figure 4.7.

Mutational Signature	Validation Type	Total Mutations in Sample	Total Mutations by Signature	Examined Mutations	Validated Mutations
Signature 1A	V_A	48	40	48	48 (100%)
Signature 1B	V_A	58	55	58	56 (97%)
Signature 2	V_A	76	75	76	72 (95%)
Signature 3	V_A	70	65	70	69 (99%)
Signature 4	V_A	196	192	196	182 (95%)
Signature 5	V_C	332	286	91	75 (82%)
Signature 6	V_A	598	440	598	540 (90%)
Signature 7	V_A	470	432	470	412 (88%)
Signature 8	V_A	4,514	1,558	250	227 (91%)
Signature 9	V_B	4,423	2,811	4,423	3,977 (90%)
Signature 10	V_A	12,848	10,558	12,848	9,420 (74%)
Signature 11	V_A	102	100	102	67 (66%)
Signature 12	V_C	2,808	2,327	100	93 (93%)
Signature 13	V_A	8,612	5,697	200	190 (95%)
Signature 14	V_C	12,984	12,984	100	86 (86%)
Signature 15	V_A	784	784	31	30 (97%)
Signature 16	V_A	793	678	73	69 (95%)
Signature 17	V_B	2,627	1,959	2,627	2,476 (94%)
Signature 18	V_A	158	156	158	142 (90%)
Signature 19	V_C	769	769	103	102 (99%)
Signature 20	V_A	885	488	198	198 (100%)
Signature 21	V_C	6,790	4,368	121	103(85%)
Signature U1	N/A	N/A	N/A	N/A	N/A
Signature U2	N/A	N/A	N/A	N/A	N/A
Signature R1	F_C	11,869	7,955	100	2(2%)
Signature R2	F_C	738	738	50	1(2%)
Signature R3	F_C	385	235	83	3(4%)

Table 4.1. Validating consensus mutational signatures found in human cancer. The precise validation approach is outlined in the text. The codes of validation types are explained in Figure 4.4.

4.5 The landscape of consensus mutational signatures in human cancer

Applying the developed computational approach to the 7,042 samples derived from 30 cancer types revealed 22 distinct and validated mutational signatures (Figure 4.5; an individual figure for each signature can be found in Appendix III). These 22 mutational signatures show substantial diversity in their patterns of somatic mutations. There are signatures characterized by the prominence of only one or two of the 96 possible substitution mutations, indicating a remarkable specificity of mutation type and sequence context. One such example is Signature 10, which is

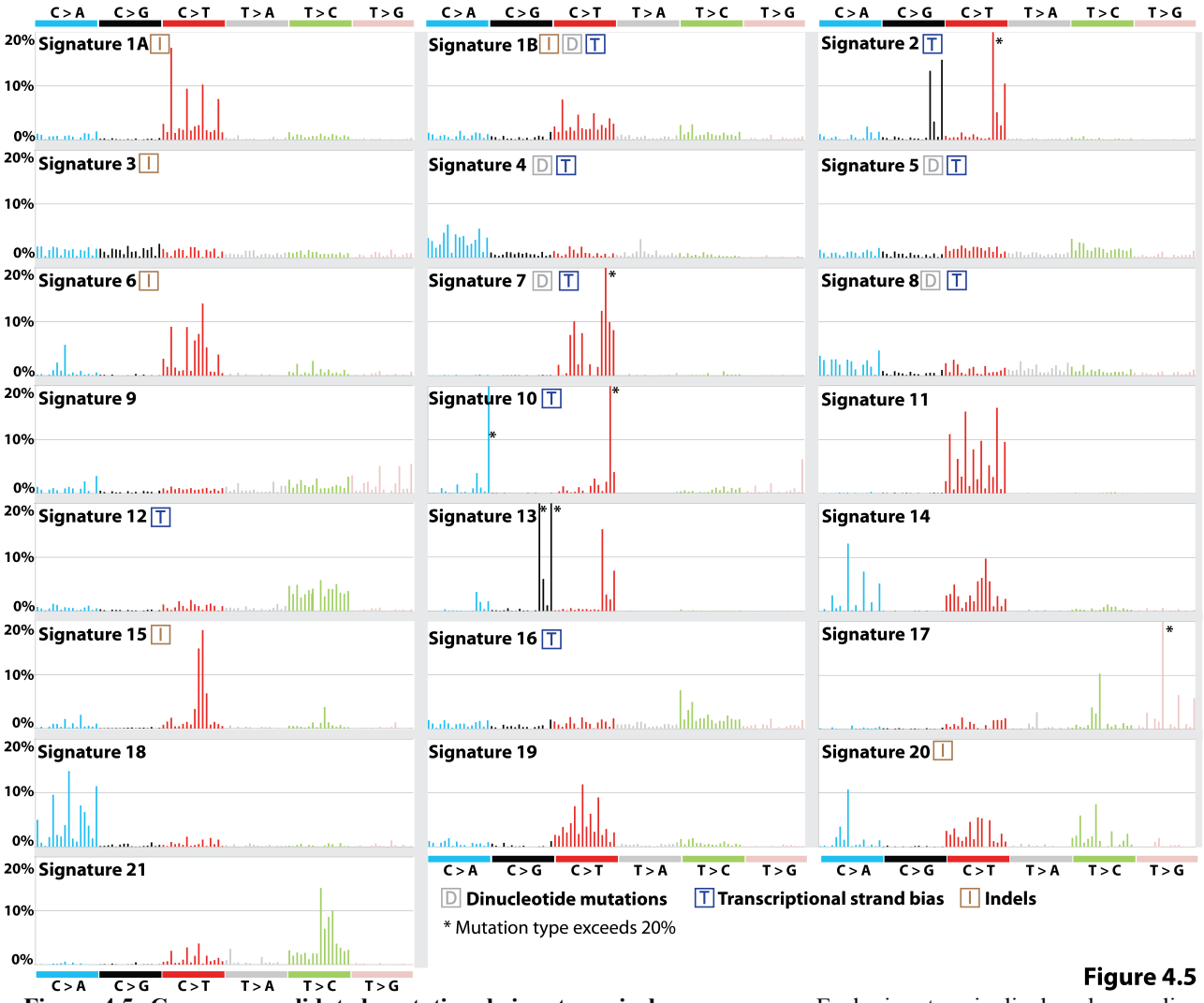


Figure 4.5

Figure 4.5: Consensus validated mutational signatures in human cancer. Each signature is displayed according to the 96 substitution classification defined by the substitution class and sequence context immediately 3' and 5' to the mutated base. The probability bars for the six types of substitutions are displayed in different colours. The mutation types are on the horizontal axes, whereas vertical axes depict the percentage of mutations attributed to a specific mutation type. All mutational signatures are displayed on the basis of the trinucleotide frequency of the human genome. A higher resolution of each panel is found Appendix III. Asterisk indicates mutation type exceeding 20%.

predominantly characterized by C>A mutations at TpCpT and C>T mutations at TpCpG. At the other extreme, some mutational signatures exhibit a more-or-less equal representation of all 96 mutations. Examples of such mutational signatures are Signatures 3 and 8. A large proportion of the validated consensus mutational signatures are characterized predominantly by C>T substitutions at different trinucleotide sequence contexts: Signatures 1A, 1B, 6, 7, 11, 15, and 19. Signatures 4, 8, and 18 have a prevalence for C>A mutations, while Signatures 5, 12, 16, and 21 exhibit a preference for T>C substitutions. Signatures 9 and 17 exhibit a preference of T>G mutations at specific sequence contexts. Lastly, no mutational signatures in this series are dominated by T>A substitutions.

Signatures 1A and 1B are observed in 25 of the 30 cancer classes (Figure 4.9). Both are characterized by a prominence of C>T substitutions at NpCpG trinucleotides. Since they are almost mutually exclusive among tumour types (Figure 4.9) they probably represent the same underlying process, with Signature 1B representing a less efficient separation from other signatures in some cancer types. Signature 1A/B is most likely related to the relatively elevated rate of spontaneous deamination of 5-methyl-cytosine which results in C>T transitions and which predominantly occurs at NpCpG trinucleotides (Pfeifer, 2006). This mutational process operates in the germline, where it has resulted in substantial depletion of NpCpG sequences, as well as in normal somatic cells (Welch et al., 2012).

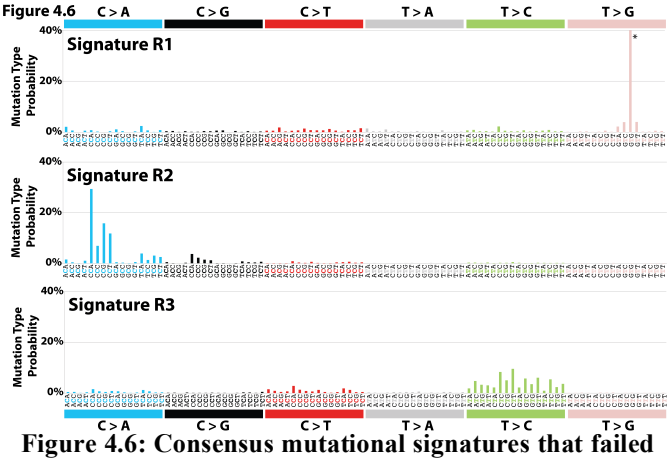


Figure 4.6: Consensus mutational signatures that failed validation. Each signature is displayed according to the 96 substitution classification defined by the substitution class and sequence context immediately 3' and 5' to the mutated base. The probability bars for the six types of substitutions are displayed in different colours. The mutation types are on the horizontal axes, whereas vertical axes depict the percentage of mutations attributed to a specific mutation type. All mutational signatures are displayed on the basis of the trinucleotide frequency of the human genome. A higher resolution of each panel is found Appendix III. Asterisk indicates mutation type exceeding 40%.

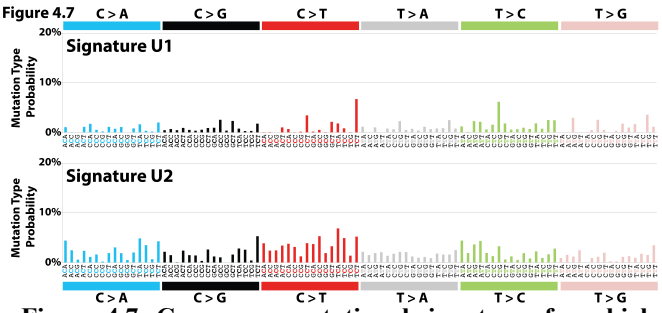


Figure 4.7: Consensus mutational signatures for which it is not possible to perform validation. Each signature is displayed according to the 96 substitution classification defined by the substitution class and sequence context immediately 3' and 5' to the mutated base. The probability bars for the six types of substitutions are displayed in different colours. The mutation types are on the horizontal axes, whereas vertical axes depict the percentage of mutations attributed to a specific mutation type. All mutational signatures are displayed on the basis of the trinucleotide frequency of the human genome. A higher resolution of each panel is found Appendix III.

In addition to the 22 consensus mutational signatures that validated (Table 4.1), three signatures failed validation, and thus most likely reflect technology specific sequencing artefacts (Figure 4.6). Signature R1 is previously described in chapter 3 and is predominantly characterized by T>G mutations at GpGpTpGpG. Signature R2 exhibits a C>A pattern of

mutations with a preference for CpC and TpC dinucleotides. Finally, Signature R3 is predominantly composed of T>C mutations with a specific trinucleotide pattern (Figure 4.6). Interestingly, these mutational signatures are confined to samples from specific sequencing centres. Signature R1 is found in samples analysed by the Sanger Institute, Signature R2 in samples sequenced at the Broad Institute, and Signature R3 is found only in data generated by the Baylor College of Medicine. This observation further confirms the suspicion that these three mutational processes reflect technical/analysis artefacts rather than real biological processes.

For three of the 27 consensus mutational signatures, I was unable to identify available samples that could be used to validate these signatures (Figure 4.7). Both Signatures U1 and U2 exhibit a rather uniform pattern of mutations across the six types of substitutions without any mutation type exceeding 10%. It should be noted that the patterns of these two mutational signatures are different from the previously identified and validated uniform mutational signatures: Signature 3 and Signature 8 (Figure 4.5).

Lastly, all of the previously identified breast cancer mutational signatures are found by this pan-cancer analysis. Breast cancer Signature BC-1 (chapter 3) has the same pattern of mutations as the global consensus Signature 1B, Signature BC-2 corresponds to Signature 2, Signature BC-3 corresponds to Signature 13, Signature BC-4 corresponds to Signature 3, Signature BC-5 corresponds Signature 8, and Signature BC-6 corresponds to Signature R1.

4.5.1 Consensus mutational signatures with transcriptional strand-bias

The efficiency of DNA damage and DNA maintenance processes can differ between the transcribed and untranscribed strands of genes. The most celebrated cause of this phenomenon is transcription-coupled nucleotide excision repair (NER) that operates exclusively on the transcribed strand of genes and is recruited by RNA polymerase II when it encounters bulky DNA helix-distorting lesions (Hanawalt and Spivak, 2008). Evaluation of the efficiency of transcription-coupled DNA repair is done analogously to the analysis performed for breast cancer (chapter 3). Briefly,

mutational signatures are re-extracted incorporating the transcriptional strand on which each mutation has taken place.

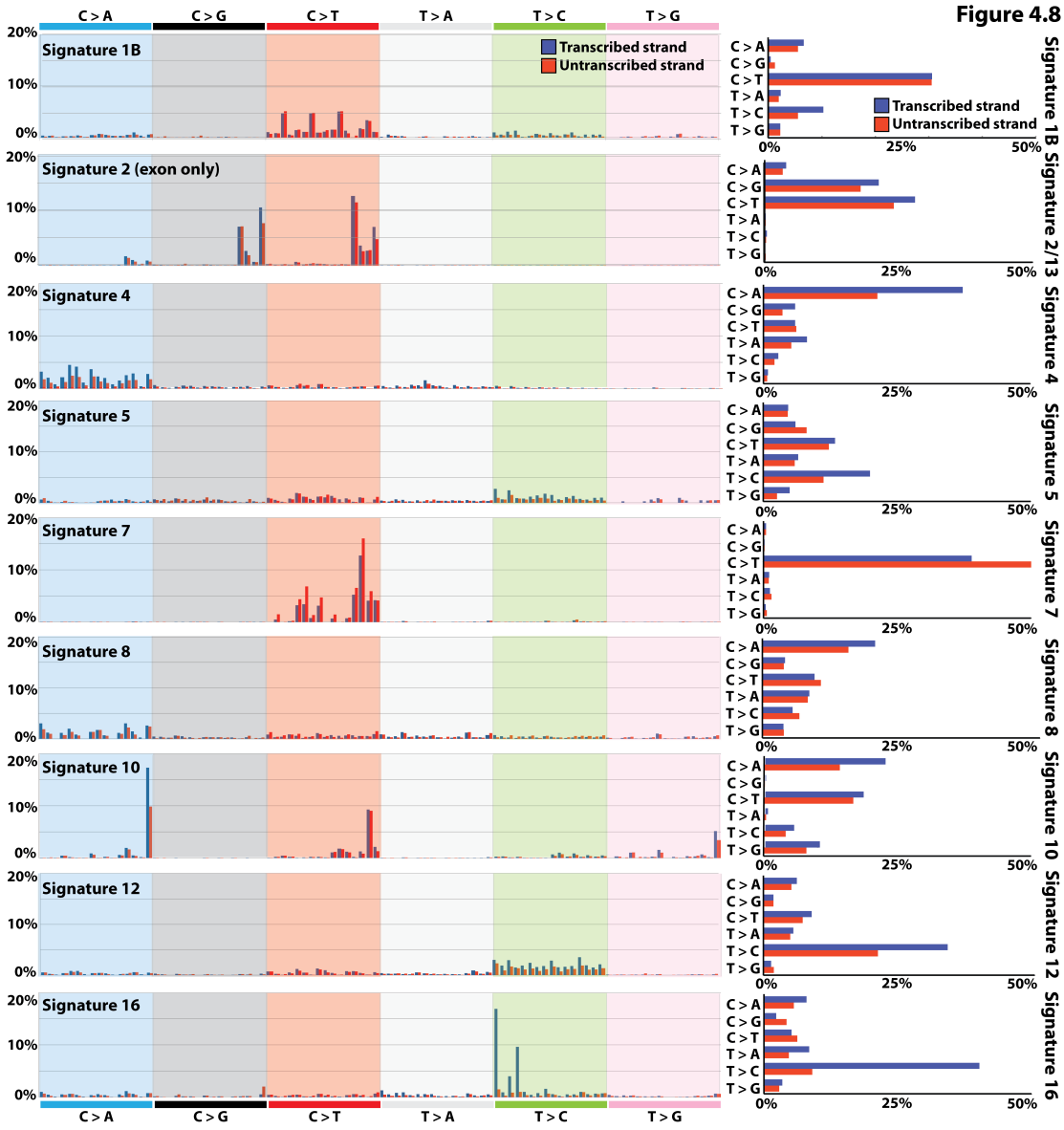


Figure 4.8: Consensus mutational signatures with strand-bias. Mutations are shown according to the 192 mutation classification incorporating the substitution type, the sequence context immediately 5' and 3' to the mutated base and whether the mutated pyrimidine is on the transcribed or untranscribed strand. The mutation types are displayed on the horizontal axis, whereas the vertical axis depicts the percentage of mutations attributed to a specific mutation type. A higher resolution version of all mutational signatures with transcriptional strand-bias is found in Appendix IV.

Nine consensus signatures showed substantial differences in mutation prevalence between transcribed and untranscribed strands, known as transcriptional strand-bias (Figure 4.8; an individual figure for each signature can be found in Appendix IV). This strand-bias is observed only for validated mutational signatures (Figure 4.5) and it is absent in the signatures that failed validation (Figure 4.6) or for

which validation is not possible (Figure 4.7). In eight of these nine signatures the strand-bias is observed across the complete footprints of transcribed protein coding genes. In contrast, the strand-bias in Signature 2 is observed only in exons and it is lacking in intronic regions.

Two of the nine mutational signatures likely implicate activity of transcription-coupled nucleotide excision repair. Signature 4 shows transcriptional strand-bias for C>A mutations (Figure 4.8). Signature 4 is observed in lung adenocarcinoma, squamous and small cell carcinomas, head and neck squamous, and liver cancers (Figure 4.9), most of which are caused by tobacco smoking. Therefore, Signature 4 is probably an imprint of the bulky DNA adducts generated by polycyclic hydrocarbons found in tobacco smoke and their removal by transcription-coupled NER (Pfeifer et al., 2002). The higher prevalence of C>A mutations on transcribed compared to untranscribed strands is consistent with the propensity of many tobacco carcinogens to form adducts on guanine.

Similarly, Signature 7, mainly found in malignant melanoma, shows a higher prevalence of C>T mutations on the untranscribed compared to the transcribed strands consistent with the formation, through ultraviolet light exposure, of pyrimidine dimers and other lesions which are known to be repaired by transcription-coupled NER (Pfeifer et al., 2005).

Beyond these known examples of DNA damage processed by transcription-coupled NER, other signatures show strong transcriptional strand-bias: Signatures 1B, 2, 5, 8, 10, 12, and 16. Notably, Signature 16, which is characterized by T>C mutations at ApTpA, ApTpG, and ApTpT trinucleotides and is observed in hepatocellular carcinomas, shows the strongest transcriptional strand-bias of any signature, with T>C mutations occurring almost exclusively on the transcribed strand (Figure 4.8). Similarly, Signature 12, which features T>C mutations at NpTpN trinucleotides, also found in hepatocellular carcinomas, shows strong transcriptional strand-bias with more T>C mutations on the transcribed than untranscribed strands (Figure 4.8). Based on the assumption that the transcriptional strand-biases in Signatures 12 and 16 are introduced by transcription-coupled NER, these currently unexplained signatures might be the result of bulky DNA helix distorting adducts on adenine. However, there is no prior basis for invoking transcription-coupled NER in

the genesis of these signatures (or any of the other mutational signatures) and other causes of transcriptional strand-bias may exist.

4.5.2 Mutational signatures with dinucleotide substitutions and indels

Mutational signatures are re-extracted including, in addition to the 96 substitution types, three further classes of mutation: dinucleotide substitutions, indels at short nucleotide repeats, and indels with overlapping microhomology at breakpoint junctions. This analysis also revealed 27 consensus mutational signatures (annotated on Figure 4.5). No indels or dinucleotide substitutions are found in the signatures that are not validated. Six of the validated mutational signatures are associated with indels, while five of the validated mutational signatures are associated with double nucleotide substitutions.

Signature 1A and Signature 1B both associate with indels at repetitive elements. Interestingly, these mutational signatures do not contribute large amounts of indels (or substitutions) in any given sample but, rather, these mutational signatures are present at low background levels in almost all samples in which they are found.

Four of the 22 base substitution signatures associated with large numbers of indels. Signature 6, which is characterized predominantly by C>T at NpCpG mutations, but is distinct from Signature 1A/B, contributes very large numbers of substitutions and small indels (mostly of 1bp) at nucleotide repeats to subsets of colorectal, uterine, liver, kidney, prostate, oesophageal and pancreatic cancers.

Signature 15 and Signature 20 also contribute very large numbers of substitutions and small indels at nucleotide repeats but, compared to Signature 6, Signature 15 exhibits greater prominence of C>T at GpCpN trinucleotides, whereas Signature 20 contains C>A and T>C mutations. Signature 15 is found in several samples of lung and stomach cancer, whereas Signature 20 is found only in few gastric carcinomas (Figure 4.9). The origin of both mutational signatures is currently unknown.

By contrast, substantial numbers of larger deletions (up to 50 bp) with overlapping microhomology at breakpoint junctions are found in breast, ovarian and

pancreatic cancer cases with major contributions from Signature 3. In the chapter 3, I associated this particular mutational signature with inactivating mutations in *BRCA1* and/or *BRCA2* in breast cancer. This association will be further elaborated upon in the next chapter for ovarian and pancreatic cancers.

Signatures 1B, 5, 4, 7, and 8 are associated with double nucleotide substitutions. Samples with Signature 1B, 5, or 8 have low numbers of dinucleotide substitutions. In contrast, overwhelming numbers of dinucleotide substitutions are present in samples in which Signature 4 or Signature 7 is found. CC>AA/GG>TT or TC>AA/GA>TT are the predominant types of dinucleotide substitutions caused by Signatures 1B and 5. Signature 4 and 8 generate mostly CC>AA/GG>TT mutations, while Signature 7 is characterized by CC>TT/GG>AA mutations occurring predominantly at dipyrimidines.

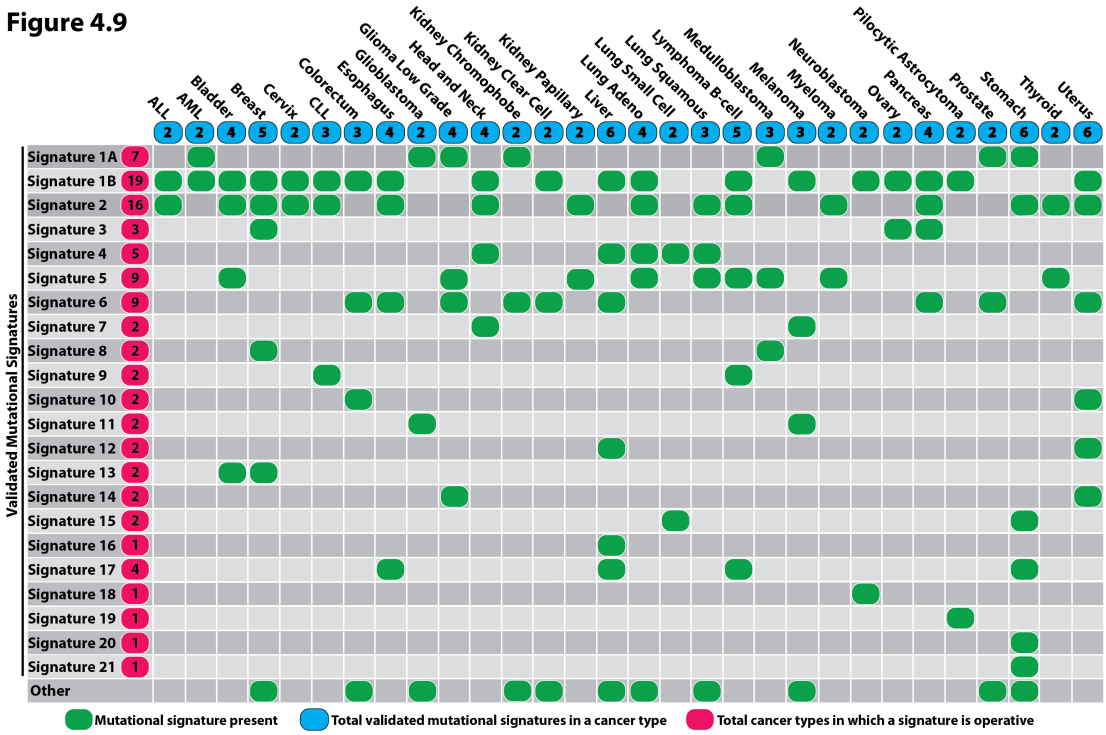


Figure 4.9: Signatures of mutational processes and the cancer types in which they are found. Cancer types are ordered alphabetically as columns, whereas mutational signatures are displayed numerically as rows. 'Other' indicates mutational signatures for which validation was not performed or for which validation failed.

4.5.3 Mutational signatures with additional sequence context

Mutational signatures are further extracted using mutational catalogues based on the Ξ_{1536} alphabet. Unfortunately, the majority of the examined cancer types are

derived from exome sequencing data and, as such, they harbour too few somatic mutations for this analysis. The examination of low numbers of somatic mutations based on a classification system that contains 1,536 types of mutations resulted in predominantly binary matrix data and, for the majority of cancer types, the analysis either fails or it does not reveal any further elaborations of the consensus mutational signatures.

Nevertheless, there are four mutational signatures that are refined by this analysis. As previously demonstrated for breast cancer, Signature 2 and Signature 13 exhibit a preference for a pyrimidine prior to the mutated TpC dinucleotide while the majority of Signature R1's T>G substitutions occur at T>G at GpGpTpGpG pentanucleotides (chapter 3). Further, this analysis demonstrated that the T>X peaks at CpT dinucleotides characteristic for Signature 17 are, in fact, dependent on the presence of an adenine located 5' prior to the dinucleotide; thus these peaks occur at ApCpTpN tetranucleotides ($Q = 1.3 \times 10^{-11}$; in all cases Q refers to a q-value, see chapter 7). Lastly, Signature 10 also displays a pentanucleotide pattern different than the one expected purely by chance ($Q = 4.5 \times 10^{-42}$). The three large peaks of Signature 10 are highly dependent on either an adenine or thymine two bases 5' to the somatic mutation.

4.6 Prevalence of consensus mutational signatures in human cancer

The previous sections of this chapter discussed the identified consensus mutational signatures. In this section, I will examine and summarize their prevalence across the analysed 30 human cancer types. In most cancer classes at least two mutational signatures are observed, with a maximum of six in cancers of the liver,

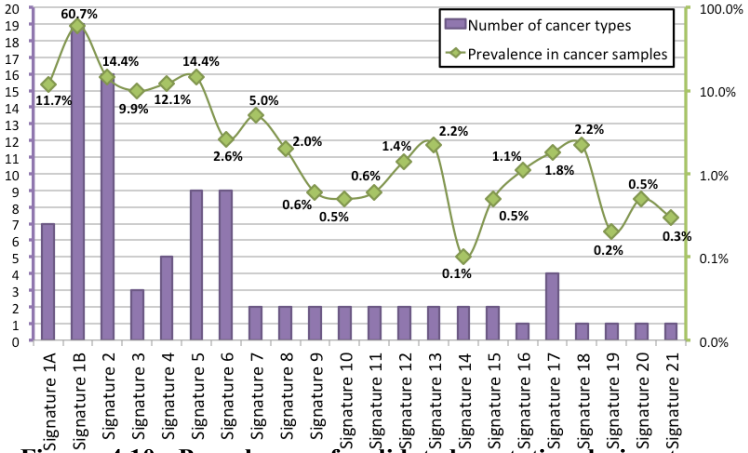


Figure 4.10: Prevalence of validated mutational signatures across all cancer types. The X-axis depicts the mutational signatures. The right Y-axis reflects the number of cancer types in which the validated consensus signature has been identified, while the left Y-axis indicates the percentage of samples from the data set of 7,042 cancers in which the signature contributed a significant number of somatic mutations.

uterus, and stomach (Figure 4.9 and Figure 4.10). Although these differences may, in part, be due to the available data in each cancer type, it seems likely that some cancers have a more complex repertoire of mutational processes than others. Signature 1A/B is found in the majority of the samples (Figure 4.10), while Signatures 2, 3, 4, 5, and 7 are present in at least ~5% of the samples. Notably Signature 2 is found in 16 of the 30 cancer types and in ~14% of all samples.

Most individual cancer genomes exhibit more than one mutational signature and many different combinations of signatures are observed (Figure 4.11). An individual figure for each cancer type depicting the contributions of the mutational signatures in each sample of that cancer type can be found in Appendix V. Further, an individual figure for each cancer type depicting the summary of the signatures' contributions in that cancer type can be found in Appendix VI. Liver cancers have the richest mutational landscape since the average liver cancer sample has at least 5 signatures imprinted by different mutational processes (Appendix V).

The patterns of contribution to individual cancer samples vary markedly between signatures. Signature 1A/B contributes relatively similar numbers of mutations to most cancer cases whereas most other signatures contribute overwhelming numbers of mutations to some cancer samples but very few to others of the same cancer class. Examples of such mutational signatures are Signatures 2, 3, 4, 6, 7, 9, 10, 11, and 13 (Figure 4.11).

Some mutational signatures are found in significant proportions of samples in some cancer types, while contributing only to a subset of samples in other cancer types. Notably, Signature 2 is identified in the majority of cervical (79%), thyroid (52%), and bladder (51%) samples but it is found only in a limited set of multiple myelomas (6%), B-cell lymphomas (11%), and breast cancers (18%) (Appendix V). Other examples include: Signature 6, identified in 20% of colorectal samples but only present in 0.6% of prostate cancer samples; Signature 13, identified in 67% of bladder samples but only present in 7% of breast cancer samples; Signature 17, found in 44% of oesophageal cancers but only in 14% of stomach cancers; Signature 3, found in 30% of breast cancers but only in 12% of pancreatic cancers (Appendix V). The reasoning behind the tissue specificity of the identified mutational signatures remains elusive. However, it is possible that some (or even most) mutational signatures are not as variable by cancer type as currently appears to be the case and examination of more

genomics data will reveal the presence of these mutational signatures in the majority of cancer types (albeit with low prevalence). Nevertheless, there are at least some mutational signatures that are most likely specific to a set of cancer types. For

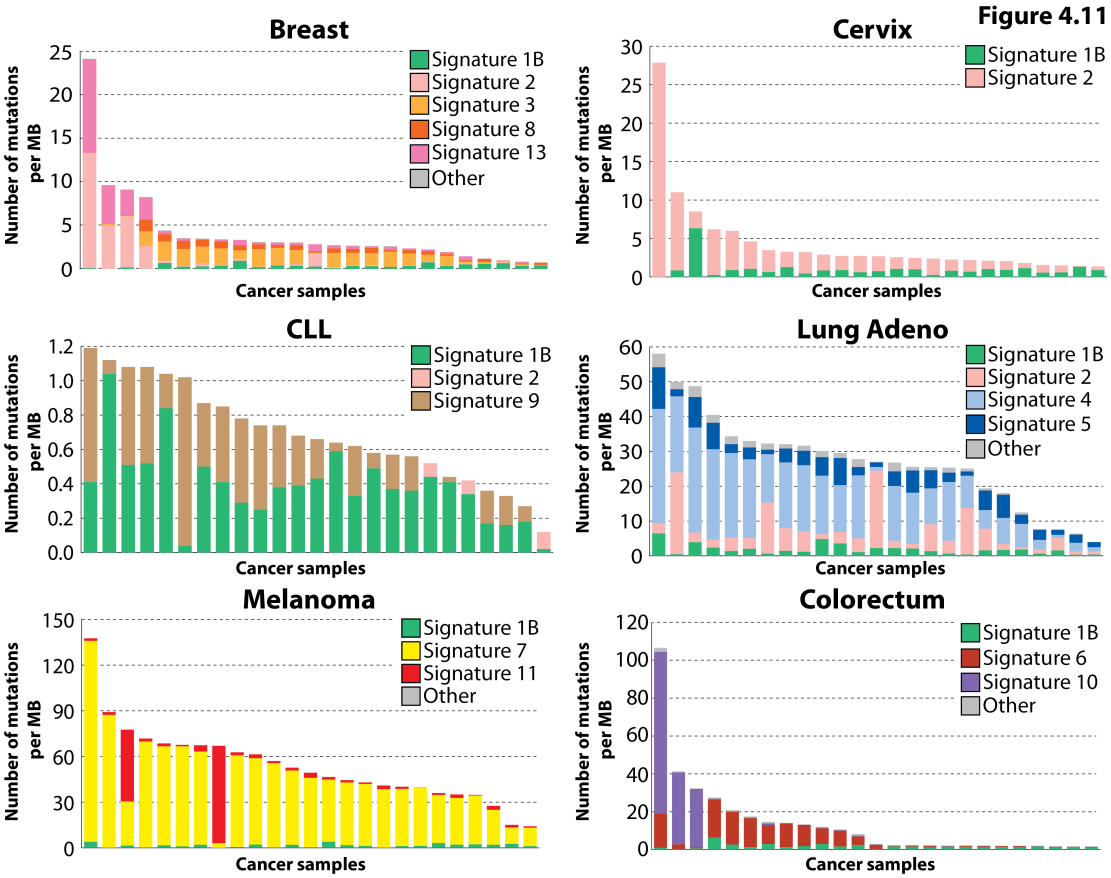


Figure 4.11: Contributions of mutational signatures in a selected set of cancer types. 25 samples are displayed for each cancer type. Each sample is displayed as a column with a height corresponding to the number of somatic mutations per megabase found in this sample. Every column is proportionately coloured to reflect the percentage of mutations attributed to different mutational signatures. 'Other' indicates mutational signatures for which validation is not performed or for which validation failed.

example, one would not expect to find the mutational signature of ultraviolet light in a primary colorectal cancer.

4.7 Discussion

In this chapter, I presented and discussed mutational signatures analysis encompassing 7,042 samples derived from 30 human cancers. The results revealed more than 20 consensus mutational signatures with a complex landscape across the different cancer types and, even, across individual cancer samples.

It should be noted that as in any computational analysis, the extraction of mutational signatures is not a perfect process. In chapter 2, I described in detail the

factors that influence the extraction of mutational signatures. These included the number of available samples, the mutation prevalence in samples, the number of mutations contributed by different mutational signatures, the similarity between the signatures of mutational processes operative in cancer samples, as well as the limitations of the developed computational approach.

In this chapter, I examined datasets with varying sizes from 30 different cancer types and great care has been taken to report only validated mutational signatures. However, the developed approach identified two similar patterns most likely representing the same biological process, viz., Signature 1A and 1B. The reasons for this is, for some cancer types, sufficient numbers of samples and/or mutations are available (i.e., statistical power) to decipher the cleaner version (i.e., Signature 1A) while for other cancer types there are not sufficient data and the approach extracts a version of the signature which is more contaminated by other, likely partially correlated, signatures present in that cancer type (i.e., Signature 1B). Nevertheless, the two signatures are visually very similar and they have been named 1A and 1B. Being almost mutually exclusive amongst cancer types (i.e., finding either Signature 1A or Signature 1B in each cancer type but not usually both) is supportive of the notion that they represent the same underlying process as is the fact that Signatures 1A and 1B have the same overall pattern of contributions to individual cancer genomes. Indeed, it is likely that if there were sufficient data, Signature 1B would disappear and the algorithm would extract only Signature 1A.

In summary, through examination of the mutational patterns buried within cancer genomes, this analysis revealed the diversity and complexity of somatic mutational processes underlying carcinogenesis in human beings. It is likely that more mutational signatures will be extracted, together with more precise definition of their features, as the number of whole-genome sequenced cancers increases and analytic methods are further refined.

Chapter 5

Etiology of mutational processes operative in human cancer

5.1 Introduction

The previous chapter of this thesis presented 27 consensus mutational signatures that were extracted from the cancer genomes of 7,042 patients across 30 distinct classes of human cancer. The chapter discussed the mutational patterns of the derived consensus signatures; however, no propositions were made about potential endogenous or exogenous mutational processes associated with any of these patterns. The aim of the present chapter is to suggest etiology for the molecular and/or environmental processes underlying at least some of these mutational signatures. These suggestions will be based on either comparing the spectrum of a mutational signature with mutational patterns of known causation or by statistically associating a signature with epidemiological, biological, or molecular features specific for each of the cancer types in which the signature has been identified.

5.2 Associating cancer etiology and mutational signatures based on mutational patterns with known causation

Each mutational signature is the imprint left on a cancer genome by a mutational process that may include one or more DNA damage and/or DNA maintenance mechanisms, with the latter either functioning normally or abnormally. Here, I consider probable mechanisms or underlying causes of the identified signatures by comparing signatures with mutation patterns of known causation in the scientific literature.

Signature 1A and Signature 1B exhibit a very similar mutational pattern. This pattern is likely related to the relatively elevated rate of spontaneous deamination of 5-methylcytosine which results in C>T transitions and which predominantly occurs at NpCpG trinucleotides (Pfeifer, 2006). As discussed in chapter 4, this mutational process operates both in the germline and in somatic cells (Welch et al., 2012). Thus, Signature 1A/B is probably due to spontaneously occurring endogenous mutational processes present in most normal and neoplastic cells that are initiated by deamination of 5-methylcytosine (Pfeifer, 2006). Other signatures are likely attributable to exogenous mutagenic exposures or failure of cellular molecular mechanisms.

The mutational patterns of Signature 2 and 13 are similar as they are both composed of C>A, C>T, and C>G substitutions at TpC dinucleotides. In chapter 3, I proposed that Signature 2 could be attributed to the activity of the *AID/APOBEC* family of cytidine deaminases, while Signature 13 encompasses an interaction between *APOBEC* enzymes and the DNA repair protein *REV1*. On the basis of similarities in the sequence context of cytosine mutations caused by *APOBEC* enzymes in experimental systems, a role for *APOBEC1*, *APOBEC3A* and/or *APOBEC3B* in human cancer appears more likely than for other members of the family (Burns et al., 2013; Harris et al., 2002; Taylor et al., 2013). Furthermore, recent studies have demonstrated that there is an association between the observed patterns of somatic mutations and the expression of *APOBEC3B* (Burns et al., 2013; Taylor et al., 2013). However, the reason for extreme activation of this mutational process, such as Signatures 2 and/or 13 hypermutated samples with up to 25 somatic mutations per megabase, remains unknown. Since *APOBEC* activation constitutes part of the innate immune response to viruses and retrotransposons (Koito and Ikeda, 2013) it may be that these mutational signatures represent collateral damage on the human genome from a response originally directed at retrotransposing DNA elements or exogenous viruses. Confirmation of this hypothesis would establish an important new mechanism for initiation of human carcinogenesis. However, it is plausible that entirely different mechanisms (both endogenous and/or exogenous) are activating the *APOBEC* enzymes.

In smoking-associated lung cancer, C>A transversions are the predominant known mutational pattern induced by tobacco carcinogens (Pfeifer et al., 2002). It is

believed that this type of substitutions is due to the formation of bulky adducts on guanine. Furthermore, previous studies have shown that the tobacco carcinogenic lesions occurring on the transcribed strand are correctly identified and removed by transcription-coupled nucleotide excision repair resulting in strong transcriptional strand-bias on a genomic scale (Pfeifer et al., 2002; Pleasance et al., 2010b). In the previous chapter, I demonstrated that Signature 4 generates predominantly C>A substitutions and that it possesses a strong transcriptional strand-bias (chapter 4). Furthermore, this signature is present in cancer types with a well-known association to tobacco smoking: lung adenocarcinoma, lung squamous, small cell lung carcinomas, head and neck squamous, and liver cancers (Figure 4.9). Thus, it is reasonable to causally associate Signature 4 with tobacco smoking. This association will be further refined using statistical analysis in the next section (see below).

Signature 7 is the predominant mutational signature found in malignant melanoma. This signature bears a mutational pattern that is expected from ultraviolet light: C>T and CC>TT mutations at dipyrimidines (chapter 4). Moreover, as expected from a mutational pattern of ultraviolet light, Signature 7 exhibits a strong transcriptional strand-bias indicating that mutations occur at pyrimidines (*viz.*, by formation of pyrimidine-pyrimidine photodimers) and these mutations are being effectively repaired by transcription-coupled nucleotide excision repair. In addition to malignant melanoma, this mutational pattern is also found in two cases of squamous carcinoma of the head and neck. Further examination revealed that both these head and neck cases are the only two cancers of the lip in the dataset. Indeed, lip cancers have been previously associated with exposure to ultraviolet light (Pfeifer et al., 2002). Based on the similarity of the mutational pattern to the one observed in experimental systems exposed to ultraviolet light and the presence of Signature 7 in ultraviolet associated cancers (*viz.*, lip cancer and malignant melanoma), Signature 7 is most likely due to exposure to ultraviolet light.

Some anticancer drugs are mutagens that have specific patterns of somatic mutations (Hunter et al., 2006). Signature 11 has mutational features very similar to those previously reported in experimental studies of alkylating agents (Hunter et al., 2006). Further analysis will be performed in the next section to statistically associate Signature 11 with a specific cancer treatment.

Abnormalities in DNA maintenance may also be responsible for mutational signatures. Previous studies have demonstrated that defective DNA mismatch repair results in highly elevated numbers of somatic mutations and exhibits significant numbers of small (1bp and 2bp long) insertion and/or deletions (indels) predominantly found at repetitive elements (Tomita-Mitchell et al., 2000). Further, microsatellite unstable tumours are characteristic for colorectal, uterine, and stomach cancers. Taken together, these observations are consistent with the behaviours and patterns of three of the identified mutational signatures: Signature 6, Signature 15, and Signature 20. Thus, it is plausible that Signatures 6, 15, and 20 are due to the failure of one or more of the molecular mechanisms of DNA mismatch repair. In the next section, I will statistically demonstrate that at least one of these mutational signatures is highly elevated in microsatellite unstable samples.

Defective repair of DNA double-strand breaks based on homologous recombination has also been known to cause an elevated numbers of large indels with overlapping microhomology at breakpoint junctions (chapter 1). This pattern of mutations is consistent with the behaviour of Signature 3. Further, in chapter 3, Signature 3 is statistically associated with failure of homologous recombination in breast cancer due to mutations in *BRCA1* and/or *BRCA2*. In a latter section, I will demonstrate that this statistical association also holds for pancreatic and ovarian cancers.

Mutational signatures may also result from the abnormal function of enzymes that modify DNA or the activity of error-prone polymerases. Previous studies have demonstrated that the activity *POL η*, an error prone polymerase involved in processing *AID* induced cytidine deamination, results in an excess of T>G transversions at ApTpN and TpTpN trinucleotides in chronic lymphocytic leukaemias with mutated immunoglobulin genes (Di Noia and Neuberger, 2007; Puente et al., 2011). This pattern of mutations is consistent with Signature 9, which is found in chronic lymphocytic leukaemia and malignant B-cell lymphomas (Figure 4.9).

Similarly, previous studies have associated recurrent somatic mutations altering the functions of the error-prone polymerase *POL ε* (*POLE*) with a subset of colorectal and uterine tumours that exhibit an ultra-hypermutator phenotype. This

behaviour is consistent with Signature 10, which is found in cancers of the colorectum and uterus with an extremely high prevalence of somatic mutations.

Many of the validated mutational signatures do not, however, have an established or proposed underlying mutational process or etiology. Some, for example Signatures 8, 12 and 16, show strong transcriptional strand-bias (Figure 4.8) and possibly reflect the involvement of transcription-coupled nucleotide excision repair acting on bulky DNA adducts due to exogenous carcinogens. Others, for example Signatures 14 and 21, show an overwhelming activity in a small number of cancer cases and are perhaps due to currently uncharacterized defects in DNA maintenance or abnormal activity of DNA polymerases.

In addition to the 22 validated consensus mutational signatures, there are another 5 consensus signatures identified through extraction of mutational signatures. The mutational patterns of Signature U1 and Signature U2 (Figure 4.7) are too uniform and unspecific to unambiguously associate them with any previously published patterns of somatic mutations. In contrast, the mutational patterns of Signatures R1, R2, and R3 are extremely specific and sequence-context dependent (Figure 4.6). Further, as discussed in chapter 4, these artifactual mutational signatures seem to be confined to data generated within specific sequencing centres. Signature R1 is associated with the next generation sequencing protocol used at the Wellcome Trust Sanger Institute. This protocol has been optimized to avoid the generation of this signature. Signature R2 is present in data from the Broad Institute and in-depth investigation revealed its pattern of mutations is due to the generation of 8-oxoguanine during DNA shearing (Costello et al., 2013). Lastly, Signature R3 is confined to colorectal data generated by the Baylor College of Medicine. After investigation, this pattern is attributed to the settings of the used bioinformatics pipelines, which are set to call somatic mutations from only a few reads in genes previously associated with colorectal cancer.

5.3 Associating cancer etiology and mutational signatures based on statistical analysis

In the previous section, a mutational signature is causally associated with a potential etiology based on the similarity of its pattern to mutational patterns of

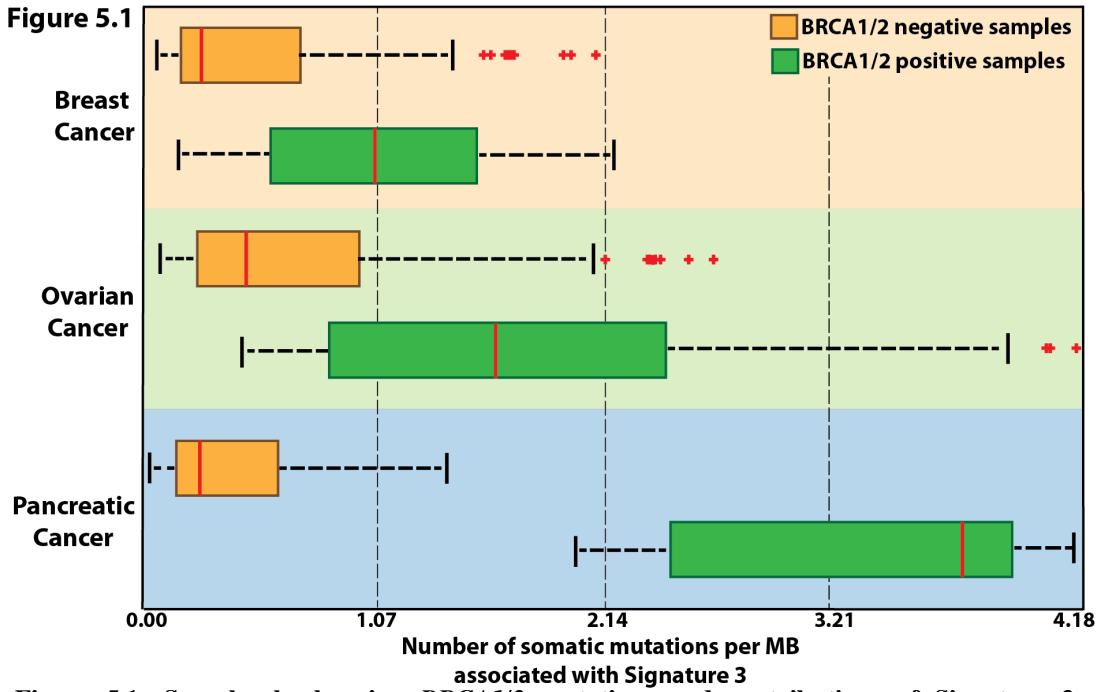


Figure 5.1: Samples harbouring *BRCA1/2* mutations and contributions of Signature 3. Signature 3 is examined in breast, ovarian, and pancreatic cancers. In each cancer type samples are separated into two sets: *BRCA1/2* positive samples (green) and *BRCA1/2* negative samples (orange). A box plot of the mutations contributed by Signature 3 in each cancer type is displayed for each of the two sets. Outliers with more than 4.18 mutations per megabase are not shown but they are included in the statistical analysis. All consensus mutational signatures are evaluated for statistical association with *BRCA1/2* in their respective cancer types. The only statistically significant difference in signatures' contributions between the *BRCA1/2* positive and negative sets is the one due to Signature 3 ($Q = 1.6 \times 10^{-8}$ for breast cancer; $Q = 2.3 \times 10^{-7}$ for ovarian cancer; $Q = 0.02$ for pancreatic cancer).

known causation in the scientific literature. This section will focus on re-confirming (or identifying new) associations via statistical analysis. Briefly, a cancer type is split based on a feature of interest (*e.g.*, smoking status separating lung adenocarcinomas in smokers and non-smokers) and statistical analysis is performed for all signatures found in that cancer type. The analysis checks whether mutations attributed to the signature in question are statistically different between the set of samples possessing the feature (*e.g.,* smokers) and the set of samples without the feature (*e.g.,* non-smokers). Any samples with missing information about a selected feature (*e.g.*, when the smoking status is unknown) are ignored. In all cases, q-values are reported for all statistically significant associations between a signature and a feature of interest. In most cases, only a single mutational signature associates with a particular feature. Features of interest are selected based on prior biological knowledge or based on advice from collaborators who are experts in a specific cancer type.

Previous analysis of breast cancer data demonstrated that samples harbouring *BRCA1* and/or *BRCA2* mutations have an elevated numbers of somatic mutations attributable to Signature 3 (Figure 3.12). Mutations associated to other mutational

signatures found in breast cancer are not statistically different between *BRCA1/2* wild type samples and *BRCA1/2* mutants (Figure 3.12). Analogous analysis is performed for the two additional cancer types in which Signature 3 is found: ovarian and pancreatic cancer (Figure 4.9). The subset of cases from these three cancer classes, known to be due to inactivating mutations in *BRCA1* and *BRCA2*, is strongly associated with the presence of Signature 3 ($Q = 1.6 \times 10^{-8}$ for breast cancer; in all cases Q refers to a q-value, see chapter 7; $Q = 2.3 \times 10^{-7}$ for ovarian cancer; $Q = 0.02$ for pancreatic cancer; Figure 5.1 and Figure 5.3). Similarly to breast cancer, no other mutational signature associated with the BRCA1/2 status in pancreatic and ovarian cancers. Interestingly, every single pancreatic cancer that harboured *BRCA1/2* mutations exhibited an extremely elevated mutational burden for Signature 3. Indeed, almost all cases with *BRCA1* and *BRCA2* mutations in breast and ovarian cancers also showed a large contribution from Signature 3. However, some ovarian and breast cancers with a substantial contribution from Signature 3 do not have *BRCA1/2* mutations, which suggests that other mechanisms of *BRCA1/2* inactivation or abnormalities of other genes may also generate the mutational pattern.

BRCA1 and *BRCA2* are implicated in homologous recombination-based DNA double-strand break repair (Thompson, 2012). The abrogation of their functions results in non-homologous end-joining mechanisms, which can utilize microhomology at rearrangement junctions to re-join double-strand breaks, taking over DNA double-strand break repair. The results show that, in addition to the genomic structural instability conferred by defective double-strand break repair, a base substitution mutational signature is associated with *BRCA1/2* deficiency in three distinct cancer types.

The statistical analysis performed in chapter 3 associated Signature 8 with estrogen receptor negative breast cancer samples. Signature 8 is also found in medulloblastoma (Figure 4.9); however, the mutations attributed to this mutational signature do not associate with any molecular subtype of medulloblastoma.

In the previous section, a causal association is proposed between tobacco smoking and Signature 4 based on the similarity between the mutational pattern of the signature and the mutational pattern observed in experimental systems exposed to tobacco carcinogens. This relationship is supported by a strong elevation of the

Figure 5.2

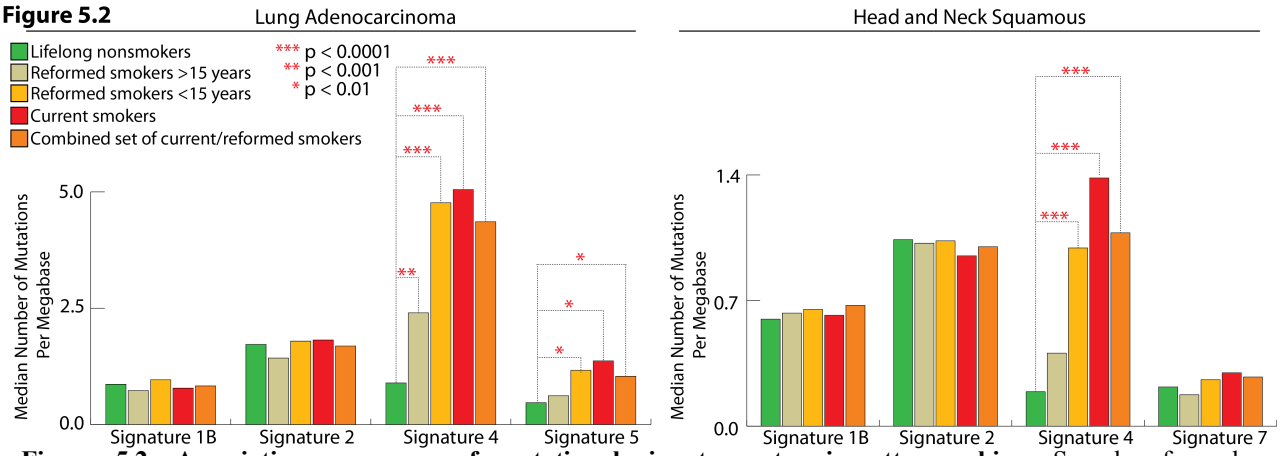


Figure 5.2: Associating exposures of mutational signatures to cigarette smoking. Samples from lung adenocarcinomas and head and neck squamous are examined. Each of the two cancer types is separated in 5 categories: lifelong non-smokers (dark green); reformed smokers for more than 15 years (light green); reformed smokers for less than 15 years (yellow); current smokers (red); a combined set containing all current and reformed smokers (orange). Statistical analysis is performed for every mutational signature by comparing the set of non-smokers with the other four sets. All reported p-values have been adjusted for multiple hypothesis testing.
The X-axis depicts the mutational signatures operative in the respective cancer types, while the Y-axis reflects the median numbers of somatic mutations attributed to each signature in each of the five categories. Note that the two Y-axes have a different scale.

mutations attributed to Signature 4 in current smokers when compared to non-smokers ($Q = 1.1 \times 10^{-7}$ for lung adenocarcinomas; $Q = 2.4 \times 10^{-5}$ for head and neck squamous; Figure 5.2). Further, there is even a statistically significant difference between the numbers of mutations attributed to Signature 4 in lung adenocarcinomas from non-smokers when compared to the mutations found in adenocarcinomas from people who stopped smoking more than fifteen years prior to their tumour diagnosis (Figure 5.2). This association is not found in head and neck cancers; however, that might be partly explained by the low number of head and neck squamous cancers from patients that stopped smoking more than 15 years prior to their diagnosis. At the very least, this result confirms that tobacco smoking leaves a strong and long lasting mutational imprint on the genome of a lung cancer.

Cigarette smoke contains over 60 carcinogens (Pfeifer et al., 2002) and it is possible that this complex mixture may initiate other mutational processes. Signature 1B, 2, and 7 are identified in head and neck squamous but they do not associate with the smoking statuses of the examined patients (Figure 5.2). However, Signature 5, but not Signatures 1A/B and 2, also showed a positive correlation between smoking history and mutation contribution in lung adenocarcinomas ($Q = 8.0 \times 10^{-3}$, Figure 5.2). Thus, in lung cancer, Signature 5 may also be generated by tobacco carcinogens.

From the carcinogens present in tobacco smoke, vinyl chloride and ethyl carbamate have been reported to generate the T>C mutations characteristic of Signature 5 (Pfeifer et al., 2002). However, Signature 5 is also present in nine other cancer types, most of which are not strongly associated with tobacco consumption, and therefore its overall etiology remains unclear (Figure 4.9).

The mutational pattern of Signature 6's indels, often termed "microsatellite instability", is characteristic of cancers with defective DNA mismatch repair (Boland and Goel, 2010). Consistent with this explanation, the presence of Signature 6 is strongly associated with the inactivation of DNA mismatch repair genes in colorectal cancer ($Q = 3.3 \times 10^{-5}$ for colorectal cancers; Figure 5.3).

Signature 9 is observed in chronic lymphocytic leukaemia and malignant B-cell lymphomas. This signature is characterized by a pattern of mutations that has been attributed to polymerase η, which is implicated with the activity of *AID* during somatic hypermutation (Puente et al., 2011). Examining chronic lymphocytic leukaemias that possess immunoglobulin gene hypermutation (IGHV-mutated) reveals a statistically significant elevation of Signature 9 ($Q = 2.5 \times 10^{-4}$; Figure 5.3). This analysis is not performed for B-cell lymphomas due to the lack of sufficient number of IGHV-mutated samples. Nevertheless, only one of the B-cell lymphomas is IGHV-mutated and this sample exhibits an extremely high level of Signature 9 (Appendix V).

Signature 10 generates huge numbers of mutations in subsets of colorectal and uterine cancers. It has been proposed that the mutational process underlying this signature is due to the altered activity of the error-prone polymerase *POLE*. To support this hypothesis, a high number of recurrent function modifying somatic mutations, *viz.*, Pro286Arg and Val411Leu, have been observed in *POLE* in colorectal and uterine samples with high mutational burden (Kandoth et al., 2013; TCGA, 2012). Statistical analysis reveals an extremely strong association between these recurrent somatic mutations and the contributions of Signature 10 ($Q = 3.1 \times 10^{-22}$ for colorectal cancer; $Q = 8.8 \times 10^{-9}$ for uterine cancer; Figure 5.3).

Signature 11 exhibits a mutational pattern resembling the one of an alkylating agent and this signature is identified in malignant melanoma and glioblastoma

multiforme. Examining information from the patients' histories revealed a statistical association between treatments with the alkylating agent temozolomide in both cancer

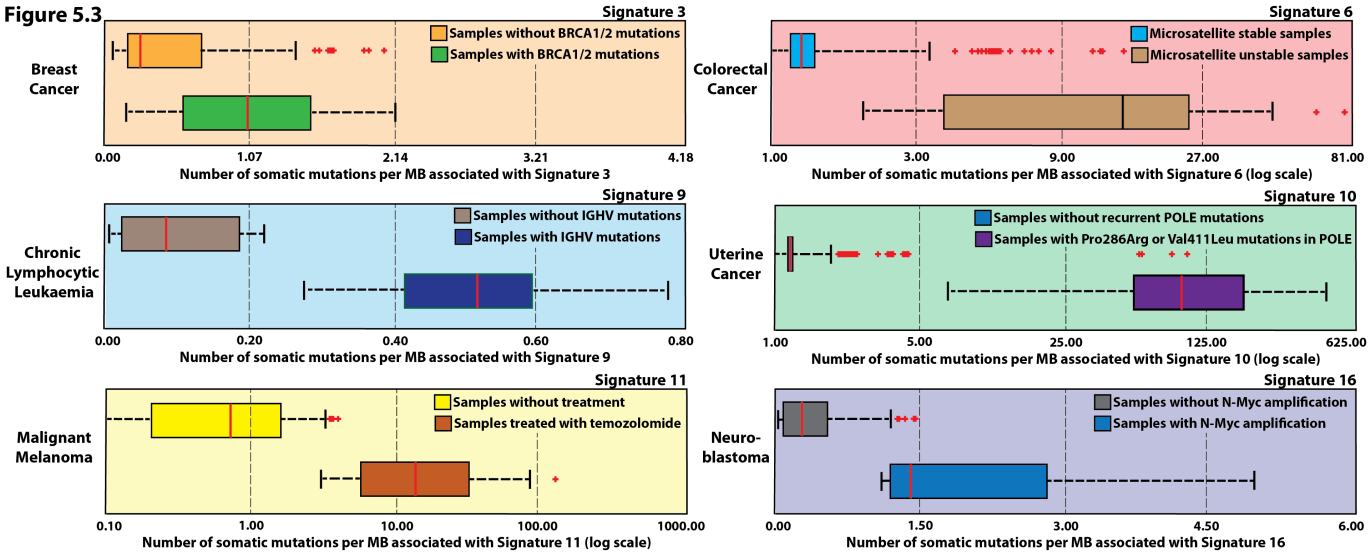


Figure 5.3: **Associating molecular or clinical features with the activity of mutational signatures**. In each case, all signatures found in a given cancer type are evaluated for a potential association with a selected feature. Only a single mutational signature associates with a selected feature in each of the examined cases. Each boxplot represents numbers of somatic mutations for a given signature for samples possessing or lacking a specific feature in a specific cancer type. The examined cancer type is annotated on the left of each panel, while the evaluated mutational signature is displayed in the upper right corner of each panel. In all cases, the X-axis depicts the number of mutations per megabase attributable to a give signature. Note that a logarithmic scale is used for the X-axes of Signatures 6, 10, and 11. For clarity, some outliers are not displayed but all data are included in the statistical analysis.

types ($Q = 4.0 \times 10^{-3}$ for malignant melanomas; $Q = 2.1 \times 10^{-3}$ for glioblastoma multiforme; Figure 5.3).

Signature 18 has a very specific mutational pattern of C>A transversions which is observed only in neuroblastomas. *N-Myc* amplification is a common feature of neuroblastomas (Brodeur et al., 1984) and statistical analysis reveals that samples with *N-Myc* amplification exhibit a significantly higher numbers of somatic mutations attributed to Signature 18 when compared to samples without this amplification ($Q = 1.2 \times 10^{-7}$; Figure 5.3).

5.4 Activity of mutational signatures and association with age of diagnosis

The origin of a cancer cell can be traced back to the zygote and, hence, the accumulation of somatic mutations identified by cancer genome sequencing can be roughly separated into mutations occurring prior to neoplastic development and mutations occurring after tumour initiation. The mutations occurring prior to

neoplastic development can be further separated as spontaneous somatic mutations occurring due to the activity of normal cellular processes and sporadic somatic mutations triggered by environmental exposures or lifestyle choices. Assuming that the accumulation of spontaneous mutations is (on average) the same across different people and that spontaneous pre-neoplastic mutations can be separated from all other somatic mutations found in a cancer, one would expect to see a strong correlation between the numbers of spontaneous pre-neoplastic somatic mutations and the age of cancer diagnosis in a large cohort of people.

A first order of approximation of this logic entails using cancer genomics data

Cancer Type	Samples with age information	Mutational Signature	P-value (FDR corrected)
ALL	**106**	**Signature 1B**	2.13E-04
AML	**151**	**Signature 1B**	6.81E-06
Breast	**879**	**Signature 1B**	7.23E-04
Colorectum	**488**	**Signature 1B**	2.89E-02
Glioma Low Grade	**154**	**Signature 1A**	1.50E-07
Head and Neck	**299**	**Signature 1B**	4.54E-03
Kidney Chromophobe	**21**	**Signature 1A**	3.53E-02
Kidney Clear Cell	**294**	**Signature 1B**	7.34E-12
Kidney Papillary	**95**	**Signature 5**	3.10E-03
Lymphoma B-cell	**24**	**Signature 1B**	1.06E-02
Medulloblastoma	**100**	**Signature 1A**	2.83E-10
Melanoma	**216**	**Signature 1B**	1.33E-03
Melanoma	**216**	**Signature 7**	2.00E-03
Neuroblastoma	**192**	**Signature 1B**	2.84E-05
Ovary	**425**	**Signature 1B**	7.18E-09
Pilocytic Astrocytoma	**63**	**Signature 1B**	4.76E-02
Stomach	**148**	**Signature 1A**	3.43E-02
Thyroid	**157**	**Signature 5**	2.95E-03

Table 5.1: Mutational signatures and age of diagnosis. All statistically significant correlations between exposures of consensus mutational signatures and age of cancer diagnosis are shown.

and attempting to correlate the age of cancer diagnosis with the mutational burden of the previously identified mutational signatures. Thus, examination is performed in

each cancer type for correlations between the age of diagnosis and the number of mutations attributable to each signature in each sample.

Signature 1A/B exhibits strong positive correlations with the age of diagnosis in the majority of cancer types of both childhood and adulthood (Table 5.1). No other mutational signature shows a consistent correlation with the age of diagnosis. Exposure to Signature 5 also correlates with the age of diagnosis in kidney papillary and thyroid cancers. However, in both cancer types, Signature 1A/B is not detected/extracted due to low number of mutations in their samples and it is likely that Signatures 1A/B and Signature 5 are mixed together. Further studies involving whole-genome sequences will be needed to validate this hypothesis. Interestingly, in melanoma, the age of diagnosis also correlates with exposure to Signature 7, which has been associated with exposure to ultraviolet light. Presumably this exposure is due to the relatively uniform chronic exposure to ultraviolet light throughout a person's lifetime.

The mutations in a cancer genome may be acquired at any stage in the cellular lineage from the fertilized egg to the sequenced cancer cell. The correlation with age of diagnosis is consistent with the hypothesis that a substantial proportion of Signature 1A/B mutations in cancer genomes have been acquired over the lifetime of the cancer patient, at a relatively constant rate that is similar in different people, probably in normal somatic tissues. The absence of consistent correlation of all other signatures with age of diagnosis suggests that mutations associated with these signatures have been generated at different rates in different people, possibly as a consequence of different mutagenic exposures or after neoplastic change has been initiated.

5.5 Summary

In this chapter, I examined the mechanistic basis of the signatures of the mutational processes operative in 30 distinct types of human cancer. An etiology is proposed either by performing a statistical comparison between sets of samples with and without specific characteristics or by comparing the observed mutational patterns with the ones in the scientific literature.

This chapter provides an indication of the processes underlying the observed patterns of somatic mutations for at least some of the mutational signatures. However, for many of the processes their etiology remains speculative or unknown.

Further elucidating the underlying mutational processes will depend upon two major streams of investigation. First, compilation of mutational signatures from model systems exposed to known mutagens or perturbations of the DNA maintenance machinery and comparing those to the ones found in human cancers. Second, correlating the contributions of mutational signatures with other biological characteristics of each cancer through diverse approaches ranging from molecular profiling to epidemiology. Collectively, these studies will advance understanding of cancer etiology with potential implications for prevention and treatment.

Chapter 6

Discussion and future explorations

6.1 Introduction

The main aim of this thesis was to improve our understanding of the mutational processes underlying cancer development by examining the molecular patterns of mutations imprinted on cancer genomes by these processes. This goal was achieved by the development of a novel theoretical model that conceptualized the idea of a *mutational signature* and mathematically connected mutational signatures with catalogues of somatic mutations identified in cancer genomes.

The developed mathematical model was used to create a computational approach to decipher the signatures of the mutational processes operative in a set of cancer genomes, based on the somatic mutations identified in the mutational catalogues of these cancers. The computational framework was extensively evaluated with a wide-range of simulated data and it was demonstrated that the framework is robust to a variety of distinct parameters and can be effectively applied to both genome and exome sequences.

The developed novel computational framework was applied to genomics data from 7,042 cancer patients to reveal the mutational processes operative across the spectrum of 30 distinct types of human cancers. This largest to date analysis of cancer genomics data has provided the first map of the signatures of the mutational processes moulding the genomes of human cancers. More than 20 distinct signatures were identified and an etiology was proposed for some of these signatures. Nevertheless, the underlying mechanisms for the majority of the mutational signatures remain mysterious and future studies will be needed to elucidate their true nature.

This chapter discusses the importance of the results presented throughout the thesis. It also provides a critical reflection on the analyses of mutational signatures and outlines potential future directions for improvement with regard to the development of novel methodologies for deciphering mutational signatures and further refining of the already identified signatures.

6.2 Implications of the identified mutational signatures

In this thesis, I report the first systematic computational analysis of large-scale cancer genomics data in order to reveal the signatures of the mutational processes underlying the development of human cancer. A brief summary of the main results of the thesis is provided in Table 6.1. The table emphasizes the characteristic mutational pattern of each mutational signature, the most common cancer types in which the signature is observed, as well as any potential etiology proposed for a mutational signature.

Signature name	Characteristic mutational pattern	Most common cancer types	Proposed etiology	Etiology proposed based on
Signature 1A	C>T at CpG	All cancer types	Deamination of 5-methylcytosine	Similarity of the mutational pattern
Signature 1B	C>T at CpG	All cancer types	Deamination of 5-methylcytosine	Similarity of the mutational pattern
Signature 2	C>T at TpC	Sixteen different cancer types	*APOBEC1*, *APOBEC3A*, or *APOBEC3B*	Similarity of the mutational pattern
Signature 3	Uniform mutational signature	Breast, ovarian, and pancreatic cancer	Defective repair of DNA double-strand breaks based on homologous recombination	Statistical association
Signature 4	C>A mutations with strong strand bias	Lung, head and neck, and liver cancer	Tobacco smoking	Similarity of the mutational pattern and statistical association
Signature 5	Mostly uniform mutational signature	Nine different cancer types	Mostly unknown but there is a weak	Some statistical association

	with some peaks of T>C mutations at ApT		association with tobacco smoking in lung cancer	
Signature 6	C>A mutations and C>T at GpC mutations	Nine different cancer types but most prevalent in colorectal and uterine cancers	Defective DNA mismatch repair	Similarity of the mutational pattern and statistical association
Signature 7	C>T at dipyrimidines	Malignant melanoma and lip cancers	Ultraviolet light	Similarity of the mutational pattern
Signature 8	C>A mutations with a moderate strand bias	Breast cancer and medulloblastoma	Higher prevalence in estrogen receptor negative breast cancers	Statistical association
Signature 9	T>G transversions at ApT and TpT	Chronic lymphocytic leukaemias and B-cell lymphomas	Polymerase η	Similarity of the mutational pattern and statistical association
Signature 10	C>A at TpCpT and C>T at TpCpG	Colorectal and uterine cancers	Polymerase ε	Statistical association
Signature 11	C>T substitutions	Malignant melanoma and glioblastoma multiforme	Treatment with temozolomide	Similarity of the mutational pattern and statistical association
Signature 12	T>C substitutions with strand bias	Liver and uterine cancer	Unknown	N/A
Signature 13	C>A and C>G at TpC	Bladder and breast cancer	*APOBEC1*, *APOBEC3A*, or *APOBEC3B* and *REV1*	Similarity of the mutational pattern
Signature 14	C>A mutations and C>T at GpC mutations	Low grade glioma and uterine cancer	Unknown	N/A
Signature 15	C>T at GpC	Stomach and lung	Defective DNA	

	mutations	cancer	mismatch repair	Similarity of the mutational pattern
Signature 16	T>C mutations at ApT with extremely strong strand-bias	Liver cancer	Unknown	N/A
Signature 17	T>G at TpT and T>C at CpT	Oesophagus cancer, liver cancer, stomach cancer, and B-cell lymphoma	Unknown	N/A
Signature 18	C>A mutations	Neuroblastoma	Amplification of *N-Myc*	Statistical association
Signature 19	C>T mutations	Pilocytic astrocytoma	Unknown	N/A
Signature 20	C>A and C>T mutations	Stomach cancer	Defective DNA mismatch repair	Similarity of the mutational pattern
Signature 21	T>C mutations	Stomach cancer	Unknown	N/A
Signature R1	T>G at GpTpG	Breast cancers generated by the Sanger Institute	Sequencing artifact	Fine-tuning a sequencing protocol
Signature R2	C>A mutations	Lung and kidney cancers generated by the Broad Institute	Sequencing artifact	Fine-tuning a sequencing protocol
Signature R3	T>C mutations	Colorectal cancers generated by the Baylor College of Medicine	Bioinformatics analysis artifact	Fine-tuning a bioinformatics analysis

Signature U1	Uniform mutational signature	Glioblastoma and prostate cancer	Unknown	N/A
Signature U2	Uniform mutational signature	Liver and kidney cancer	Unknown	N/A

Table 6.1: Summary of the deciphered signatures of mutational processes in human cancer.

This thesis has three potential implications for cancer research and cancer treatment. First, from a basic science perspective, the thesis provides the first roadmap of the mutational signatures underlying human cancer and it reveals that these signatures have a complex landscape both in an individual cancer type and across multiple cancer types.

Second, from a targeted therapeutics perspective, many of the described mutational signatures are believed to reflect failure of DNA repair mechanisms and, as such, they might be better predictors of clinical outcome when compared to mutations in genes. For example, Signature 3 is associated with mutations in *BRCA1* and *BRCA2* and it is believed to reflect failure of repair of DNA double-strand breaks based on homologous recombination (Table 6.1). This mutational signature is observed in many breast and ovarian samples lacking any *BRCA1/2* mutations and it could potentially be used for targeted treatment especially for cancers such as triple negative breast cancer. A similar logic may be applied to some of the other mutational signatures reflecting failure of DNA repair mechanisms; however, future studies will be required to reveal the applicability of mutational signatures in the clinic.

Third, some of the identified mutational signatures reflect exposures to exogenous mutagens. These signatures might be useful for the development of cancer prevention strategies. For example, Signature 4 is due to tobacco smoking while Signature 7 is associated with exposure to ultraviolet light (Table 6.1). It is foreseeable that some of the other deciphered mutational signatures might be due to or triggered by environmental exposures. For example, Signature 2 is found in 16 cancer types and it is believed that this signature is due to the activity of the APOBEC family of enzymes, which could get activated by viral infection. In support of this claim, Signature 2 is found overwhelmingly in cervical cancer, which is by far the most common HPV-related cancer. It is highly plausible that Signature 2 is indeed triggered by viral infection in cervical cancer and it is foreseeable that this might be the case in one or more of the other fifteen cancer types in which Signature 2 is

observed. While future analysis will be required to evaluate the validity of this hypothesis, confirming it will establish an important new mechanism for initiation of human carcinogenesis with significant potential for cancer prevention.

6.3 Limitations of the performed analyses of mutational signatures

The mutational signatures analyses have a number of shortcomings pertaining to the developed computational approach and the examined mutational data.

With regard to the data limitations, the majority of the work is restricted to certain classes of mutations, namely substitutions and small insertions/deletions (indels), with no attention to rearrangements and copy number changes. Further, the examined data are taken from a range of different sources (*e.g.*, publications, data portals, collaborators, *etc.*) in which the quality of DNA sequencing and mutation identification is highly variable. This is especially true for indels where the quality of the data allowed only limited exploration of indel-based mutational signatures.

Most of the analysed cancer cases are derived from exome sequencing data. Power calculations (chapter 2) and empirical observations indicate that, in general, a small number of whole-genome sequences are more powerful than a large number of exome sequences in extracting substitution and indel signatures. Indeed, in some cancer types the number of substitutions and indels available from exome sequences is so limited that only a very crude assessment of the landscape of mutational signatures is possible (*e.g.,* ovarian and thyroid cancers). Moreover, some cancer types with known patterns of mutations are not included at all in the analyses as data are either not freely available or non-existent (*e.g.*, cancer types due to exposure to aristolochic acid or aflatoxin).

While the developed computational approach is extensively evaluated with simulated data, this evaluation did not foresee the extreme variability of the numbers of somatic mutations found in cancer genomes. For example, an average cancer genome of a pilocytic astrocytoma has ~100 somatic mutations while a representative malignant melanoma harbours about 40,000 somatic mutations in its cancer genome (Figure 4.2). Extracting mutational signatures from a set containing equal numbers of mutational catalogues from melanomas and pilocytic astrocytomas will only result in finding the signatures of the mutational processes that are operative in malignant melanoma. In this example, pilocytic astrocytomas account for only 0.25% of all mutations in the dataset, well-below the 5% threshold used for optimizing and testing

the computational framework (chapter 2). These differences in mutational burdens across cancer types required performing independent mutational signatures analyses for each of the 30 cancer types as, otherwise, the highly mutated cancers would overwhelm the extraction of mutational signatures. Further, for each of the individual cancer types, great care is taken to perform the analyses with and without hypermutated samples that may be skewing the extracted mutational signatures. Improving the developed method to allow analysis of all mutational catalogues together would be extremely beneficial, for example, to decipher common mutational signatures that contribute only very few mutations to a large set of samples belonging to different cancer types. Such mutational signatures would be most likely associated with underlying spontaneous endogenous mutational processes.

Remarkably, despite all the listed obscuring factors, the analyses allowed identification and validation of more than 20 distinct mutational signatures. Nevertheless, future studies will be required to both improve and extend this compendium of mutational signatures.

6.4 Future explorations

The developed roadmap of mutational signatures is in no way final or exclusive, and future work will be required to further refine it. This will include both improvement of the computational approach as well as generation of more whole-genome sequences across the complete spectrum of human cancer types.

Briefly, the computational method will need to allow analysis, in a single run, of thousands of mutational catalogues (including hypermutators and ultra-hypermutators) from multiple distinct classes of human cancer rather than artificially separating samples by cancer types. This will most probably require extending the developed framework to a hierarchical nonnegative matrix approach, where the current method would be applied multiple consecutive times and well-explained samples would be removed from further analysis after each of the performed iterations. Moreover, minimizing the Frobenius norm between original and reconstructed samples (chapter 2) might not be optimal as outliers can affect this measure. A more robust measure (*i.e.,* average Spearman correlation) may prove to give better results with this highly variable dataset. No matter what improvements are made to the developed computational framework, extensive validation with simulated data will be require to confirm its ability to better decipher mutational signatures.

In the previous analysis, the majority of examined data are derived from cancer exomes and I heavily rely on somatic mutations of variable quality as these mutations are identified using different mutation bioinformatics algorithms. Using the same mutational-calling algorithm will provide consistent results and allow exploring indels in greater detail and including previously neglected mutation types (*viz.,* rearrangements and copy number changes). Using cancer exomes limited the extent to which the genome landscape is introduced into signature characterization. In principle, there could be many features of the landscape that can be used to distinguish between signatures (*e.g.,* origins of replications, regions of open or closed chromatin, *etc.*) and hence provide further insights into the etiology and mechanisms underlying each signature. Further studies using whole-genome sequencing would be required to perform this analysis.

It is highly likely that a future large-scale mutational signatures analysis will become a reality in the next year as part of the forthcoming International Cancer Genome Consortium's pan-cancer initiative. This analysis will encompass 2,000 to 3,000 whole-genome sequences and ~10,000 exome sequences across the complete spectrum of human cancer. The somatic mutations of these cancer samples will be identified by a predefined set of optimized mutation-calling algorithm and include all types of somatic mutations. I am currently working on improving the developed computational framework to address its current limitations and apply it to this set of cancer genomics data. This large dataset will allow substantial improvements to the biological insights into mutational signatures.

6.5 Thesis summary

In this thesis, I introduced and mathematically connected the concepts of mutational processes and mutational signatures. A *mutational process* was defined as a mixture of DNA damage and repair mechanisms that act together and have the ability to cause mutations in somatic cells. A *mutational signature* was described as a characteristic pattern of somatic mutations exhibited by an operative mutational process in a genome of a cell. The mutational catalogue of a cancer represents the aggregated outcome of the activity of all mutational processes that have been operative since the very first division of the fertilized egg. Thus, a mutational catalogue of a cancer genome is a linear mixture of mutational signatures and this

catalogue can be used as an archaeological record to identify the patterns of mutations exhibited by the mutational processes that have been operative in the cancer.

In this thesis, I developed a novel computational framework that allows extracting mutational signatures from a set of mutational catalogues, then exhaustively evaluated the developed method with simulated data, and applied it to 7,042 samples across 30 distinct classes of human cancer. This revealed more than 20 distinct signatures of mutational processes, for some of which I was able to propose an underlying mechanism.

In summary, this study examined a large scale of whole-genome and whole-exome sequencing data and provided insights into hitherto unrecognized mutational signatures present across the spectrum of human cancer. This study is the first of its kind and demonstrates the wealth of biological information that is hidden within the genomes of cancer cells.

Chapter 7

Materials and methods

7.1 Introduction

This chapter provides further details about the materials and methods used. As one of the main results of this thesis is the development of a novel method for analysing patterns of somatic mutations, the majority of materials and methods have already been presented in chapter 2. Thus, to avoid repetition, this chapter only discusses additional methods that were used through this thesis. It should be noted that I did not personally perform any DNA sequencing or mutation identification but I rather relied on somatic mutations previously identified by others. Thus, this chapter will not cover any experimental procedures for DNA sequencing or bioinformatics algorithms for identifying somatic mutations from next-generation sequencing data.

7.2 Deciphering signatures of mutational processes

Mutational signatures are deciphered independently for each of the 30 cancer types using the previously developed computational framework. The algorithm deciphers the minimal set of mutational signatures that optimally explains the proportion of each mutation type found in each catalogue and then estimates the contribution of each signature to each catalogue. Mutational signatures are also extracted separately for genomes and exomes. Mutational signatures extracted from exomes are normalized using the observed trinucleotide frequency in the human exome to the trinucleotide frequency of the human genome. All mutational signatures are clustered using unsupervised agglomerative hierarchical clustering and a threshold is selected to identify the set of consensus mutational signatures. Misclustering is avoided by manual examination and, whenever necessary, re-assignment of all

signatures in all clusters. 27 consensus mutational signatures are identified across the 30 cancer types. The computational framework for deciphering mutational signatures as well as all the data used in this study are freely available and can be downloaded from:

http://www.mathworks.com/matlabcentral/fileexchange/38724

7.3 Displaying mutational signatures

Mutational signatures are displayed using a 96 substitution classification defined by the substitution class and the sequence context immediately 5' and 3' to the mutated base. Mutational signatures are displayed in the main text (unless otherwise specified) based on the observed trinucleotide frequency of the human genome, *i.e.,* representing the relative proportions of mutations generated in each signature based on the actual trinucleotide frequencies of the reference human genome.

7.4 Filtering and generating mutational catalogues

In all examined samples, normal DNAs from the same individuals are sequenced to establish the somatic origin of variants. Extensive filtering is performed to remove any residual germline mutations and technology specific sequencing artefacts prior to analysing the data. Germline mutations are filtered out from the lists of reported mutations using the complete list of germline mutations from dbSNP (Sherry et al., 2001), 1000 genomes project (Abecasis et al., 2012), NHLBI GO Exome Sequencing Project (Fu et al., 2013), and 69 Complete Genomics panel (http://www.completegenomics.com/public-data/69-Genomes/). Technology specific sequencing artefacts are filtered out by using panels of BAM files of (unmatched) normal tissues containing more than 137 normal genomes and 532 normal exomes. Any somatic mutation present in at least three well mapping reads in at least two normal BAM files are discarded. The remaining somatic mutations are used for generating a mutational catalogue for every sample.

The immediate 5′ and 3′ sequence context is extracted using the ENSEMBL Core APIs for human genome build GRCh37. Curated somatic mutations that originally mapped to an older version of the human genome are re-mapped using UCSC's freely available lift genome annotations tool (any somatic mutations with ambiguous or missing mappings are discarded). Dinucleotide substitutions are

identified when two substitutions are present in consecutive bases on the same chromosome (sequence context is ignored). The immediate 5′ and 3′ sequence content of all indels is examined and the ones present at mono/polynucleotide repeats or microhomologies are included in the analysed mutational catalogues as their respective types. Strand bias catalogues are derived for each sample using only substitutions identified in the transcribed regions of well-annotated protein coding genes. Genomic regions of bidirectional transcription are excluded from the strand bias analysis.

7.5 Statistical evaluation of associations

Generalized linear models (GLMs) are used to fit signatures' exposures (*i.e.,* number of mutations assigned to a signature) and the age of cancer diagnoses. For each cancer type, all mutational signatures operative in it are evaluated using GLMs. The Benjamini–Hochberg false discovery rate (FDR) procedure is used to adjust for multiple hypothesis testing and in all cases q-values are reported.

Associations between all other etiologies and signature exposures are performed using two-sample Kolmogorov-Smirnov tests between two sets of samples. The first set encompasses the signature exposures of the samples with the "desired feature" (*e.g.,* samples that contain immunoglobulin gene hypermutation) and the second set encompasses the signature exposures of the samples without the "desired feature" (*e.g.,* samples that do NOT contain immunoglobulin gene hypermutation). Samples with unknown features status (*e.g.,* not knowing the hypermutation status of the immunoglobulin gene) are ignored. Kolmogorov-Smirnov tests are performed for all signatures and all examined "features" in a cancer type. Similarly, the Benjamini–Hochberg false discovery rate (FDR) procedure is used to adjust for multiple hypothesis testing in a particular cancer class and in all cases q-values are reported.

BIBLIOGRAPHY

Abdullah, M.B. (1990). On a Robust Correlation-Coefficient. J Roy Stat Soc D-Sta 39, 455-460.

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.

Agrawal, N., Frederick, M.J., Pickering, C.R., Bettegowda, C., Chang, K., Li, R.J., Fakhry, C., Xie, T.X., Zhang, J., Wang, J., et al. (2011). Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science 333, 1154-1157.

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

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013b). Deciphering signatures of mutational processes operative in human cancer. Cell reports 3, 246-259.

Antoniu, S.A. (2011). Crizotinib for EML4-ALK positive lung adenocarcinoma: a hope for the advanced disease? Evaluation of Kwak EL, Bang YJ, Camidge DR, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med 2010;363(18):1693-703. Expert opinion on therapeutic targets 15, 351-353.

Armitage, P., and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. British journal of cancer 8, 1-12.

Avery, O.T., Macleod, C.M., and McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. The Journal of experimental medicine 79, 137-158.

Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated evolution of prostate cancer genomes. Cell 153, 666-677.

Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., et al. (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nature genetics 44, 685-689.

Barrett, J.C., Lamb, P.W., and Wiseman, R.W. (1989). Multiple mechanisms for the carcinogenic effects of asbestos and other mineral fibers. Environmental health perspectives 81, 81-89.

Berger, M.F., Hodis, E., Heffernan, T.P., Deribe, Y.L., Lawrence, M.S., Protopopov, A., Ivanova, E., Watson, I.R., Nickerson, E., Ghosh, P., et al. (2012). Melanoma genome sequencing reveals frequent PREX2 mutations. Nature 485, 502-506.

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C., et al. (2011). The genomic complexity of primary human prostate cancer. Nature 470, 214-220.

Bernstein, C., Prasad, A.R., Nfonsam, V., and Bernstein, H. (2013). DNA Damage, DNA Repair and Cancer. In New Research Directions in DNA Repair, C. Chen, ed. (InTech).

Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., and Plemmons, R.J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis 52, 155-173.

Besaratinia, A., Synold, T.W., Xi, B., and Pfeifer, G.P. (2004). G-to-T transversions and small tandem base deletions are the hallmark of mutations induced by ultraviolet a radiation in mammalian cells. Biochemistry 43, 8169-8177.

Biankin, A.V., Waddell, N., Kassahn, K.S., Gingras, M.C., Muthuswamy, L.B., Johns, A.L., Miller, D.K., Wilson, P.J., Patch, A.M., Wu, J., et al. (2012). Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 491, 399-405.

Boland, C.R., and Goel, A. (2010). Microsatellite instability in colorectal cancer. Gastroenterology 138, 2073-2087 e2073.

Boulton, S.J. (2010). DNA repair: Decision at the break point. Nature 465, 301-302.

Boveri, T. (2008). Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. Journal of cell science 121 Suppl 1, 1-84.

Brash, D.E., Rudolph, J.A., Simon, J.A., Lin, A., McKenna, G.J., Baden, H.P., Halperin, A.J., and Ponten, J. (1991). A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. Proceedings of the National Academy of Sciences of the United States of America 88, 10124-10128.

Brodeur, G.M., Seeger, R.C., Schwab, M., Varmus, H.E., and Bishop, J.M. (1984). Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. Science 224, 1121-1124.

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., et al. (2013). APOBEC3B is an enzymatic source of mutation in breast cancer. Nature 494, 366-370.

Campbell, C., Quinn, A.G., Angus, B., Farr, P.M., and Rees, J.L. (1993). Wavelength specific patterns of p53 induction in human skin following exposure to UV radiation. Cancer research 53, 2697-2699.

Cancer Genome Atlas, N. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499, 43-49.

Cancer Genome Atlas, N. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330-337.

Capella, G., Cronauer-Mitra, S., Pienado, M.A., and Perucho, M. (1991). Frequency and spectrum of mutations at codons 12 and 13 of the c-K-ras gene in human tumors. Environmental health perspectives 93, 125-131.

Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.P., Ahmann, G.J., Adli, M., et al. (2011a). Initial genome sequencing and analysis of multiple myeloma. Nature 471, 467-472.

Chapman, P.B., Hauschild, A., Robert, C., Haanen, J.B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., et al. (2011b). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. The New England journal of medicine 364, 2507-2516.

Chiou, C.C., and Yang, J.L. (1995). Mutagenicity and specific mutation spectrum induced by 8-methoxypsoralen plus a low dose of UVA in the hprt gene in diploid human fibroblasts. Carcinogenesis 16, 1357-1362.

Circu, M.L., and Aw, T.Y. (2010). Reactive oxygen species, cellular redox systems, and apoptosis. Free radical biology & medicine 48, 749-762.

Cogliano, V.J., Baan, R., Straif, K., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Freeman, C., et al. (2011). Preventable exposures associated with human cancers. Journal of the National Cancer Institute 103, 1827-1839.

Comon, P. (2010). Handbook of blind source separation : independent component analysis and blind deconvolution, 1st edn (Boston, MA: Elsevier).

Conticello, S.G. (2008). The AID/APOBEC family of nucleic acid mutators. Genome biology 9, 229.

Cooke, M.S., Evans, M.D., Dizdaroglu, M., and Lunec, J. (2003). Oxidative DNA damage: mechanisms, mutation, and disease. FASEB journal : official publication of the Federation of American Societies for Experimental Biology 17, 1195-1214.

Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., et al. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic acids research 41, e67.

Croux, C., and Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. Stat Method Appl-Ger 19, 497-515.

Cruet-Hennequart, S., Glynn, M.T., Murillo, L.S., Coyne, S., and Carty, M.P. (2008). Enhanced DNA-PK-mediated RPA2 hyperphosphorylation in DNA polymerase eta-deficient human cells treated with cisplatin and oxaliplatin. DNA repair 7, 582-596.

Davies, H., Bignell, G.R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M.J., Bottomley, W., et al. (2002). Mutations of the BRAF gene in human cancer. Nature 417, 949-954.

de Boer, J., and Hoeijmakers, J.H. (2000). Nucleotide excision repair and human syndromes. Carcinogenesis 21, 453-460.

De Bont, R., and van Larebeke, N. (2004). Endogenous DNA damage in humans: a review of quantitative data. Mutagenesis 19, 169-185.

De Keersmaecker, K., Atak, Z.K., Li, N., Vicente, C., Patchett, S., Girardi, T., Gianfelici, V., Geerdens, E., Clappier, E., Porcu, M., et al. (2013). Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nature genetics 45, 186-190.

de Laat, W.L., Jaspers, N.G., and Hoeijmakers, J.H. (1999). Molecular mechanism of nucleotide excision repair. Genes & development 13, 768-785.

Deem, A., Keszthelyi, A., Blackgrove, T., Vayl, A., Coffey, B., Mathur, R., Chabes, A., and Malkova, A. (2011). Break-induced replication is highly inaccurate. PLoS biology 9, e1000594.

Di Noia, J.M., and Neuberger, M.S. (2007). Molecular mechanisms of antibody somatic hypermutation. Annu Rev Biochem 76, 1-22.

Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B., et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455, 1069-1075.

Dodge, J.E., Ramsahoye, B.H., Wo, Z.G., Okano, M., and Li, E. (2002). De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. Gene 289, 41-48.

Drablos, F., Feyzi, E., Aas, P.A., Vaagbo, C.B., Kavli, B., Bratlie, M.S., Pena-Diaz, J., Otterlei, M., Slupphaug, G., and Krokan, H.E. (2004). Alkylation damage in DNA and RNA--repair mechanisms and medical significance. DNA repair 3, 1389-1407.

Dulak, A.M., Stojanov, P., Peng, S., Lawrence, M.S., Fox, C., Stewart, C., Bandla, S., Imamura, Y., Schumacher, S.E., Shefler, E., et al. (2013). Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nature genetics 45, 478-486.

Dulbecco, R. (1986). A turning point in cancer research: sequencing the human genome. Science 231, 1055-1056.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic acids research 10, 2709-2721.

Engelward, B.P., Allan, J.M., Dreslin, A.J., Kelly, J.D., Wu, M.M., Gold, B., and Samson, L.D. (1998). A chemical and genetic approach together define the biological consequences of 3-methyladenine lesions in the mammalian genome. The Journal of biological chemistry 273, 5412-5418.

Essers, J., Theil, A.F., Baldeyron, C., van Cappellen, W.A., Houtsmuller, A.B., Kanaar, R., and Vermeulen, W. (2005). Nuclear dynamics of PCNA in DNA replication and repair. Molecular and cellular biology 25, 9350-9359.

Evans, M.D., Dizdaroglu, M., and Cooke, M.S. (2004). Oxidative DNA damage and disease: induction, repair and significance. Mutation research 567, 1-61.

Fernandez, J.R., Byrne, B., and Firestein, B.L. (2009). Phylogenetic analysis and molecular evolution of guanine deaminases: from guanine to dendrites. Journal of molecular evolution 68, 227-235.
Friedberg, E.C., and Friedberg, E.C. (2006). DNA repair and mutagenesis, 2nd edn (Washington, D.C.: ASM Press).

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-220.

Fujimoto, A., Totoki, Y., Abe, T., Boroevich, K.A., Hosoda, F., Nguyen, H.H., Aoki, M., Hosono, N., Kubo, M., Miya, F., et al. (2012). Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nature genetics 44, 760-764.

Gao, Y., and Church, G. (2005). Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics 21, 3970-3975.

Garraway, L.A., and Lander, E.S. (2013). Lessons from the cancer genome. Cell 153, 17-37.

Goldman, R., and Shields, P.G. (2003). Food mutagens. The Journal of nutrition 133 Suppl 3, 965S-973S.

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N.D., Kanchi, K.L., Maher, C.A., Fulton, R., Fulton, L., Wallis, J., et al. (2012). Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell 150, 1121-1134.

Grasso, C.S., Wu, Y.M., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J., Jing, X., Lonigro, R.J., Brenner, J.C., et al. (2012). The mutational landscape of lethal castration-resistant prostate cancer. Nature 487, 239-243.

Greenblatt, M.S., Bennett, W.P., Hollstein, M., and Harris, C.C. (1994). Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer research 54, 4855-4878.

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

Grosovsky, A.J., de Boer, J.G., de Jong, P.J., Drobetsky, E.A., and Glickman, B.W. (1988). Base substitutions, frameshifts, and small deletions constitute ionizing radiation-induced point mutations in mammalian cells. Proceedings of the National Academy of Sciences of the United States of America 85, 185-188.

Gui, Y., Guo, G., Huang, Y., Hu, X., Tang, A., Gao, S., Wu, R., Chen, C., Li, X., Zhou, L., et al. (2011). Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. Nature genetics 43, 875-878.

Guo, G., Gui, Y., Gao, S., Tang, A., Hu, X., Huang, Y., Jia, W., Li, Z., He, M., Sun, L., et al. (2012). Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. Nature genetics 44, 17-19.

Hainaut, P., Olivier, M., and Pfeifer, G.P. (2001). TP53 mutation spectrum in lung cancers and mutagenic signature of components of tobacco smoke: lessons from the IARC TP53 mutation database. Mutagenesis 16, 551-553; author reply 555-556.

Haines, T.R., Rodenhiser, D.I., and Ainsworth, P.J. (2001). Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. Developmental biology 240, 585-598.

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

Hanawalt, P.C., and Spivak, G. (2008). Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev Mol Cell Biol 9, 958-970.

Haracska, L., Unk, I., Johnson, R.E., Johansson, E., Burgers, P.M., Prakash, S., and Prakash, L. (2001). Roles of yeast DNA polymerases delta and zeta and of Rev1 in the bypass of abasic sites. Genes & development 15, 945-954.

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

Harvey, R.G. (1991). Polycyclic aromatic hydrocarbons : chemistry and carcinogenicity (Cambridge ; New York: Cambridge University Press).

Hashimoto, K., Tominaga, Y., Nakabeppu, Y., and Moriya, M. (2004). Futile short-patch DNA base excision repair of adenine:8-oxoguanine mispair. Nucleic acids research 32, 5928-5934.

Hazra, T.K., Das, A., Das, S., Choudhury, S., Kow, Y.W., and Roy, R. (2007). Oxidative DNA damage repair in mammalian cells: a new perspective. DNA repair 6, 470-480.

Hecht, S.S. (1999). DNA adduct formation from tobacco-specific N-nitrosamines. Mutation research 424, 127-142.

Helbock, H.J., Beckman, K.B., Shigenaga, M.K., Walter, P.B., Woodall, A.A., Yeo, H.C., and Ames, B.N. (1998). DNA oxidation matters: the HPLC-electrochemical detection assay of 8-oxo-deoxyguanosine and 8-oxo-guanine. Proceedings of the National Academy of Sciences of the United States of America 95, 288-293.

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

Hodis, E., Watson, I.R., Kryukov, G.V., Arold, S.T., Imielinski, M., Theurillat, J.P., Nickerson, E., Auclair, D., Li, L., Place, C., et al. (2012). A landscape of driver mutations in melanoma. Cell 150, 251-263.

Hoeijmakers, J.H. (2009). DNA damage, aging, and cancer. The New England journal of medicine 361, 1475-1485.

Hollstein, M., Hergenhahn, M., Yang, Q., Bartsch, H., Wang, Z.Q., and Hainaut, P. (1999). New approaches to understanding p53 gene tumor mutation spectra. Mutation research 431, 199-209.

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

Holmfeldt, L., Wei, L., Diaz-Flores, E., Walsh, M., Zhang, J., Ding, L., Payne-Turner, D., Churchman, M., Andersson, A., Chen, S.C., et al. (2013). The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nature genetics 45, 242-252.

Hori, M., Suzuki, T., Minakawa, N., Matsuda, A., Harashima, H., and Kamiya, H. (2011). Mutagenicity of secondary oxidation products of 8-oxo-7,8-dihydro-2'-deoxyguanosine 5'-triphosphate (8-hydroxy-2'- deoxyguanosine 5'-triphosphate). Mutation research 714, 11-16.

Howard, B.D., and Tessman, I. (1964). Identification of the Altered Bases in Mutated Single-Stranded DNA. Ii. In Vivo Mutagenesis by 5-Bromodeoxyuridine and 2-Aminopurine. Journal of molecular biology 9, 364-371.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science 339, 957-959.

Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I., Gerhard, D.S., et al. (2010). International network of cancer genome projects. Nature 464, 993-998.

Hultquist, J.F., Lengyel, J.A., Refsland, E.W., LaRue, R.S., Lackey, L., Brown, W.L., and Harris, R.S. (2011). Human and rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H demonstrate a conserved capacity to restrict Vif-deficient HIV-1. Journal of virology 85, 11220-11234.

Hunter, C., Smith, R., Cahill, D.P., Stephens, P., Stevens, C., Teague, J., Greenman, C., Edkins, S., Bignell, G., Davies, H., et al. (2006). A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. Cancer research 66, 3987-3991.

Imielinski, M., Berger, A.H., Hammerman, P.S., Hernandez, B., Pugh, T.J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., et al. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell 150, 1107-1120.

Izumi, T., Wiederhold, L.R., Roy, G., Roy, R., Jaiswal, A., Bhakat, K.K., Mitra, S., and Hazra, T.K. (2003). Mammalian DNA base excision repair proteins: their interactions and role in repair of oxidative DNA damage. Toxicology 193, 43-65.

Jackson, A.L., and Loeb, L.A. (2001). The contribution of endogenous sources of DNA damage to the multiple mutations in cancer. Mutation research 477, 7-21.

Jain, A.K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31, 651-666.
Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. CA: a cancer journal for clinicians 61, 69-90.

Jiao, Y., Shi, C., Edil, B.H., de Wilde, R.F., Klimstra, D.S., Maitra, A., Schulick, R.D., Tang, L.H., Wolfgang, C.L., Choti, M.A., et al. (2011). DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. Science 331, 1199-1203.

Jiricny, J. (2006). The multifaceted mismatch-repair system. Nature reviews Molecular cell biology 7, 335-346.

Jones, D.T., Jager, N., Kool, M., Zichner, T., Hutter, B., Sultan, M., Cho, Y.J., Pugh, T.J., Hovestadt, V., Stutz, A.M., et al. (2012a). Dissecting the genomic complexity underlying medulloblastoma. Nature 488, 100-105.

Jones, P.A. (2012b). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews Genetics 13, 484-492.

Jones, S., Wang, T.L., Shih Ie, M., Mao, T.L., Nakayama, K., Roden, R., Glas, R., Slamon, D., Diaz, L.A., Jr., Vogelstein, B., et al. (2010). Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science 330, 228-231.

Kadyrov, F.A., Dzantiev, L., Constantin, N., and Modrich, P. (2006). Endonucleolytic function of MutLalpha in human mismatch repair. Cell 126, 297-308.

Kan, Z., Zheng, H., Liu, X., Li, S., Barber, T.D., Gong, Z., Gao, H., Hao, K., Willard, M.D., Xu, J., et al. (2013). Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome research.

Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature 497, 67-73.

Kim, T.K., Kim, T., Kim, T.Y., Lee, W.G., and Yim, J. (2000). Chemotherapeutic DNA-damaging drugs activate interferon regulatory factor-7 by the mitogen-activated protein kinase kinase-4-cJun NH2-terminal kinase pathway. Cancer research 60, 1153-1156.

Knudson, A.G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America 68, 820-823.

Koito, A., and Ikeda, T. (2013). Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. Front Microbiol 4, 28.

Koren, A., Polak, P., Nemesh, J., Michaelson, J.J., Sebat, J., Sunyaev, S.R., and McCarroll, S.A. (2012). Differential relationship of DNA replication timing to different forms of human mutation and variation. American journal of human genetics 91, 1033-1040.

Krauthammer, M., Kong, Y., Ha, B.H., Evans, P., Bacchiocchi, A., McCusker, J.P., Cheng, E., Davis, M.J., Goh, G., Choi, M., et al. (2012). Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nature genetics 44, 1006-1014.

Kriaucionis, S., and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science 324, 929-930.

Kumar, D., Abdulovic, A.L., Viberg, J., Nilsson, A.K., Kunkel, T.A., and Chabes, A. (2011). Mechanisms of mutagenesis in vivo due to imbalanced dNTP pools. Nucleic acids research 39, 1360-1371.

Laird, P.W., and Jaenisch, R. (1996). The role of DNA methylation in cancer genetic and epigenetics. Annual review of genetics 30, 441-464.

Laureti, L., Selva, M., Dairou, J., and Matic, I. (2013). Reduction of dNTP levels enhances DNA replication fidelity in vivo. DNA repair 12, 300-305.

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., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214-218.

Le Gallo, M., O'Hara, A.J., Rudd, M.L., Urick, M.E., Hansen, N.F., O'Neil, N.J., Price, J.C., Zhang, S., England, B.M., Godwin, A.K., et al. (2012). Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. Nature genetics 44, 1310-1315.

Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature 401, 788-791.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. Nature 362, 709-715.

Lindahl, T., and Nyberg, B. (1974). Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry 13, 3405-3410.

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., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322.

Liu, J., Lee, W., Jiang, Z., Chen, Z., Jhunjhunwala, S., Haverty, P.M., Gnad, F., Guan, Y., Gilbert, H.N., Stinson, J., et al. (2012a). Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. Genome research 22, 2315-2327.

Liu, P., Morrison, C., Wang, L., Xiong, D., Vedell, P., Cui, P., Hua, X., Ding, F., Lu, Y., James, M., et al. (2012b). Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. Carcinogenesis 33, 1270-1276.

Liu, M., and Schatz, D.G. (2009). Balancing AID and DNA repair during somatic hypermutation. Trends in immunology 30, 173-181.

Love, C., Sun, Z., Jima, D., Li, G., Zhang, J., Miles, R., Richards, K.L., Dunphy, C.H., Choi, W.W., Srivastava, G., et al. (2012). The genetic landscape of mutations in Burkitt lymphoma. Nature genetics 44, 1321-1325.

Lynch, T.J., Bell, D.W., Sordella, R., Gurubhagavatula, S., Okimoto, R.A., Brannigan, B.W., Harris, P.L., Haserlat, S.M., Supko, J.G., Haluska, F.G., et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. The New England journal of medicine 350, 2129-2139.

Manchester, K.L. (1995). Theodor Boveri and the origin of malignant tumours. Trends in cell biology 5, 384-387.

Masai, H., Matsumoto, S., You, Z., Yoshizawa-Sugata, N., and Oda, M. (2010). Eukaryotic chromosome DNA replication: where, when, and how? Annual review of biochemistry 79, 89-130.

McCann, J., and Ames, B.N. (1976). Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals: discussion. Proceedings of the National Academy of Sciences of the United States of America 73, 950-954.

McCulloch, S.D., and Kunkel, T.A. (2008). The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell research 18, 148-161.

McNeil, E.M., and Melton, D.W. (2012). DNA repair endonuclease ERCC1-XPF as a novel therapeutic target to overcome chemoresistance in cancer therapy. Nucleic acids research 40, 9990-10004.

Mechali, M. (2010). Eukaryotic DNA replication origins: many choices for appropriate answers. Nature reviews Molecular cell biology 11, 728-738.

Michaels, M.L., Cruz, C., Grollman, A.P., and Miller, J.H. (1992). Evidence that MutY and MutM combine to prevent mutations by an oxidatively damaged form of guanine in DNA. Proceedings of the National Academy of Sciences of the United States of America 89, 7022-7025.

Morgan, H.D., Dean, W., Coker, H.A., Reik, W., and Petersen-Mahrt, S.K. (2004). Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. The Journal of biological chemistry 279, 52353-52360.

Morin, R.D., Mendez-Lago, M., Mungall, A.J., Goya, R., Mungall, K.L., Corbett, R.D., Johnson, N.A., Severson, T.M., Chiu, R., Field, M., et al. (2011). Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. Nature 476, 298-303.

Morley, A.A., and Turner, D.R. (1999). The contribution of exogenous and endogenous mutagens to in vivo mutations. Mutation research 428, 11-15.

Moser, J., Kool, H., Giakzidis, I., Caldecott, K., Mullenders, L.H., and Fousteri, M.I. (2007). Sealing of chromosomal DNA nicks during nucleotide excision repair requires XRCC1 and DNA ligase III alpha in a cell-cycle-specific manner. Molecular cell 27, 311-323.

Mukherjee, S. (2010). The emperor of all maladies : a biography of cancer, 1st Scribner hardcover edn (New York: Scribner).

Murphree, A.L., and Benedict, W.F. (1984). Retinoblastoma: clues to human oncogenesis. Science 223, 1028-1033.

Nagarajan, N., Bertrand, D., Hillmer, A.M., Zang, Z.J., Yao, F., Jacques, P.E., Teo, A.S., Cutcutache, I., Zhang, Z., Lee, W.H., et al. (2012). Whole-genome reconstruction and mutational signatures in gastric cancer. Genome biology 13, R115.

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., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell 149, 979-993.

Nordling, C.O. (1953). A new theory on cancer-inducing mechanism. British journal of cancer 7, 68-72.

Nouspikel, T. (2009). DNA repair in mammalian cells : Nucleotide excision repair: variations on versatility. Cellular and molecular life sciences : CMLS 66, 994-1009.

Nouspikel, T., and Hanawalt, P.C. (2000). Terminally differentiated human neurons repair transcribed genes but display attenuated global DNA repair and modulation of repair gene expression. Molecular and cellular biology 20, 1562-1570.

Nowell, P.C. (1962). The minute chromosome (Phl) in chronic granulocytic leukemia. Blut 8, 65-66.

Obeid, S., Blatter, N., Kranaster, R., Schnur, A., Diederichs, K., Welte, W., and Marx, A. (2010). Replication through an abasic DNA lesion: structural basis for adenine selectivity. The EMBO journal 29, 1738-1747.

Oikawa, S., and Kawanishi, S. (1999). Site-specific DNA damage at GGG sequence by oxidative stress may accelerate telomere shortening. FEBS letters 453, 365-368.

Oikawa, S., Tada-Oikawa, S., and Kawanishi, S. (2001). Site-specific DNA damage at the GGG sequence by UVA involves acceleration of telomere shortening. Biochemistry 40, 4763-4768.

Ozturk, M. (1991). p53 mutation in hepatocellular carcinoma after aflatoxin exposure. Lancet 338, 1356-1359.

Paez, J.G., Janne, P.A., Lee, J.C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F.J., Lindeman, N., Boggon, T.J., et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 304, 1497-1500.

Pao, W., Miller, V., Zakowski, M., Doherty, J., Politi, K., Sarkaria, I., Singh, B., Heelan, R., Rusch, V., Fulton, L., et al. (2004). EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proceedings of the National Academy of Sciences of the United States of America 101, 13306-13311.

Papadopoulo, D., Laquerbe, A., Guillouf, C., and Moustacchi, E. (1993). Molecular spectrum of mutations induced at the HPRT locus by a cross-linking agent in human cell lines with different repair capacities. Mutation research 294, 167-177.

Park, Y., and Gerson, S.L. (2005). DNA repair defects in stem cell function and aging. Annual review of medicine 56, 495-508.

Parker, R.C., Varmus, H.E., and Bishop, J.M. (1984). Expression of v-src and chicken c-src in rat cells demonstrates qualitative differences between pp60v-src and pp60c-src. Cell 37, 131-139.

Parsonnet, J. (1995). Bacterial infection as a cause of cancer. Environmental health perspectives 103, 263-268.

Parsons, D.W., Jones, S., Zhang, X., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.M., Gallia, G.L., et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. Science 321, 1807-1812.

Patel, R.P., McAndrew, J., Sellak, H., White, C.R., Jo, H., Freeman, B.A., and Darley-Usmar, V.M. (1999). Biological aspects of reactive nitrogen species. Biochimica et biophysica acta 1411, 385-400.
Peharz, R., and Pernkopf, F. (2012). Sparse nonnegative matrix factorization with _0-constraints. Neurocomputing 80, 38-46.

Peifer, M., Fernandez-Cuesta, L., Sos, M.L., George, J., Seidel, D., Kasper, L.H., Plenker, D., Leenders, F., Sun, R., Zander, T., et al. (2012). Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nature genetics 44, 1104-1110.

Pena-Diaz, J., and Jiricny, J. (2012). Mammalian mismatch repair: error-free or error-prone? Trends in biochemical sciences 37, 206-214.

Pena-Llopis, S., Vega-Rubin-de-Celis, S., Liao, A., Leng, N., Pavia-Jimenez, A., Wang, S., Yamasaki, T., Zhrebker, L., Sivanand, S., Spence, P., et al. (2012). BAP1 loss defines a new class of renal cell carcinoma. Nature genetics 44, 751-759.

Perucho, M., Goldfarb, M., Shimizu, K., Lama, C., Fogh, J., and Wigler, M. (1981). Human-tumor-derived cell lines contain common and different transforming genes. Cell 27, 467-476.

Pfeifer, G.P. (2006). Mutagenesis at methylated CpG sequences. Curr Top Microbiol Immunol 301, 259-281.

Pfeifer, G.P., Denissenko, M.F., Olivier, M., Tretyakova, N., Hecht, S.S., and Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene 21, 7435-7451.

Pfeifer, G.P., Kadam, S., and Jin, S.G. (2013). 5-hydroxymethylcytosine and its potential roles in development and cancer. Epigenetics & chromatin 6, 10.

Pfeifer, G.P., You, Y.H., and Besaratinia, A. (2005). Mutations induced by ultraviolet light. Mutation research 571, 19-31.

Pham, P., Bransteitter, R., Petruska, J., and Goodman, M.F. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature 424, 103-107.

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

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

Poklar, N., Pilch, D.S., Lippard, S.J., Redding, E.A., Dunham, S.U., and Breslauer, K.J. (1996). Influence of cisplatin intrastrand crosslinking on the conformation, thermal stability, and energetics of a 20-mer DNA duplex. Proceedings of the National Academy of Sciences of the United States of America 93, 7606-7611.

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

Puente, X.S., Pinyol, M., Quesada, V., Conde, L., Ordonez, G.R., Villamor, N., Escaramis, G., Jares, P., Bea, S., Gonzalez-Diaz, M., et al. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 475, 101-105.

Pugh, T.J., Morozova, O., Attiyeh, E.F., Asgharzadeh, S., Wei, J.S., Auclair, D., Carter, S.L., Cibulskis, K., Hanna, M., Kiezun, A., et al. (2013). The genetic landscape of high-risk neuroblastoma. Nature genetics 45, 279-284.

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., et al. (2012). Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature 488, 106-110.

Pulciani, S., Santos, E., Lauver, A.V., Long, L.K., Robbins, K.C., and Barbacid, M. (1982). Oncogenes in human tumor cell lines: molecular cloning of a transforming gene from human bladder carcinoma cells. Proceedings of the National Academy of Sciences of the United States of America 79, 2845-2849.

Quesada, V., Conde, L., Villamor, N., Ordonez, G.R., Jares, P., Bassaganyas, L., Ramsay, A.J., Bea, S., Pinyol, M., Martinez-Trillos, A., et al. (2012). Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nature genetics 44, 47-52.

Ratel, D., Ravanat, J.L., Berger, F., and Wion, D. (2006). N6-methyladenine: the other methylated base of DNA. BioEssays : news and reviews in molecular, cellular and developmental biology 28, 309-315.

Rausch, T., Jones, D.T., Zapatka, M., Stutz, A.M., Zichner, T., Weischenfeldt, J., Jager, N., Remke, M., Shih, D., Northcott, P.A., et al. (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell 148, 59-71.

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

Rettig, R.A. (2006). The war on cancer: An anatomy of failure, a blueprint for the future. Health Affair 25, 1446-1447.

Roberts, M. (2010). Cancer 'is nation's biggest fear' (BBC News: BBC).

Roberts, S., and Everson, R. (2001). Independent component analysis : principles and practice (Cambridge ; New York: Cambridge University Press).

Robertson, A.B., Klungland, A., Rognes, T., and Leiros, I. (2009). DNA repair in mammalian cells: Base excision repair: the long and short of it. Cellular and molecular life sciences : CMLS 66, 981-993.

Robinson, G., Parker, M., Kranenburg, T.A., Lu, C., Chen, X., Ding, L., Phoenix, T.N., Hedlund, E., Wei, L., Zhu, X., et al. (2012). Novel mutations target distinct subgroups of medulloblastoma. Nature 488, 43-48.

Rodin, S.N., and Rodin, A.S. (2005). Origins and selection of p53 mutations in lung carcinogenesis. Seminars in cancer biology 15, 103-112.

Rodgman, A., and Perfetti, T.A. (2008). The chemical components of tobacco and tobacco smoke (Boca Raton: CRC Press).

Rogozin, I.B., Basu, M.K., Jordan, I.K., Pavlov, Y.I., and Koonin, E.V. (2005). APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. Cell cycle 4, 1281-1285.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53-65.

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

Rubin, A.F., and Green, P. (2009). Mutation patterns in cancer genomes. Proc Natl Acad Sci U S A 106, 21766-21770.

Rudin, C.M., Durinck, S., Stawiski, E.W., Poirier, J.T., Modrusan, Z., Shames, D.S., Bergbower, E.A., Guan, Y., Shin, J., Guillory, J., et al. (2012). Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nature genetics 44, 1111-1116.

Rusmintratip, V., and Sowers, L.C. (2000). An unexpectedly high excision capacity for mispaired 5-hydroxymethyluracil in human cell extracts. Proceedings of the National Academy of Sciences of the United States of America 97, 14183-14187.

Sale, J.E., Lehmann, A.R., and Woodgate, R. (2012). Y-family DNA polymerases and their role in tolerance of cellular DNA damage. Nature reviews Molecular cell biology 13, 141-152.

Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., Szabo, S., Yan, H., Gazdar, A., Powell, S.M., Riggins, G.J., et al. (2004). High frequency of mutations of the PIK3CA gene in human cancers. Science 304, 554.

San Filippo, J., Sung, P., and Klein, H. (2008). Mechanism of eukaryotic homologous recombination. Annual review of biochemistry 77, 229-257.

Sancar, A., Lindsey-Boltz, L.A., Unsal-Kacmaz, K., and Linn, S. (2004). Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. Annual review of biochemistry 73, 39-85.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74, 5463-5467.

Sausen, M., Leary, R.J., Jones, S., Wu, J., Reynolds, C.P., Liu, X., Blackford, A., Parmigiani, G., Diaz, L.A., Jr., Papadopoulos, N., et al. (2013). Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. Nature genetics 45, 12-17.

Schiller, J.T., and Lowy, D.R. (2010). Vaccines to prevent infections by oncoviruses. Annual review of microbiology 64, 23-41.

Seo, J.S., Ju, Y.S., Lee, W.C., Shin, J.Y., Lee, J.K., Bleazard, T., Lee, J., Jung, Y.J., Kim, J.O., Shin, J.Y., et al. (2012). The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome research 22, 2109-2119.

Seshagiri, S., Stawiski, E.W., Durinck, S., Modrusan, Z., Storm, E.E., Conboy, C.B., Chaudhuri, S., Guan, Y., Janakiraman, V., Jaiswal, B.S., et al. (2012). Recurrent R-spondin fusions in colon cancer. Nature 488, 660-664.

Setlow, R.B., and Carrier, W.L. (1966). Pyrimidine dimers in ultraviolet-irradiated DNA's. Journal of molecular biology 17, 237-254.

Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature 486, 395-399.

Shen, J.C., Rideout, W.M., 3rd, and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic acids research 22, 972-976.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29, 308-311.

Shih, C., Padhy, L.C., Murray, M., and Weinberg, R.A. (1981). Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. Nature 290, 261-264.

Sissi, C., and Palumbo, M. (2003). The quinolone family: from antibacterial to anticancer agents. Current medicinal chemistry Anti-cancer agents 3, 439-450.

Spencer, J., and Dunn-Walters, D.K. (2005). Hypermutation at A-T base pairs: the A nucleotide replacement spectrum is affected by adjacent nucleotides and there is no reverse complementarity of sequences flanking mutated A and T nucleotides. Journal of immunology 175, 5170-5177.

Stark, M.S., Woods, S.L., Gartside, M.G., Bonazzi, V.F., Dutton-Regester, K., Aoude, L.G., Chow, D., Sereduk, C., Niemi, N.M., Tang, N., et al. (2012). Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing. Nature genetics 44, 165-169.

Stehelin, D., Varmus, H.E., Bishop, J.M., and Vogt, P.K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature 260, 170-173.

Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., Bignell, G., Teague, J., Smith, R., Stevens, C., et al. (2005). A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nature genetics 37, 590-592.

Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature 486, 400-404.

Stern, R.S. (2007). Psoralen and ultraviolet a light therapy for psoriasis. The New England journal of medicine 357, 682-690.

Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A., et al. (2011). The mutational landscape of head and neck squamous cell carcinoma. Science 333, 1157-1160.

Stratton, M.R. (2011). Exploring the genomes of cancer cells: progress and promise. Science 331, 1553-1558.

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

Suspene, R., Aynaud, M.M., Guetard, D., Henry, M., Eckhoff, G., Marchio, A., Pineau, P., Dejean, A., Vartanian, J.P., and Wain-Hobson, S. (2011). Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. Proceedings of the National Academy of Sciences of the United States of America 108, 4858-4863.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324, 930-935.

Taylor, B.J., Nik-Zainal, S., Wu, Y.L., Stebbings, L.A., Raine, K., Campbell, P.J., Rada, C., Stratton, M.R., and Neuberger, M.S. (2013). DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. eLife 2, e00534.

Teng, B., Burant, C.F., and Davidson, N.O. (1993). Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science 260, 1816-1819.

Thompson, L.H. (2012). Recognition, signaling, and repair of DNA double-strand breaks produced by ionizing radiation in mammalian cells: the molecular choreography. Mutat Res 751, 158-246.

Tice, R.R., and Setlow, R.B. (1985). DNA repair and replication in aging organisms and cells. In Handbook of the Biology of Aging, F. E.E., and E.L. Schneider, eds. (New York), pp. 173-224.

Tomita-Mitchell, A., Kat, A.G., Marcelino, L.A., Li-Sucholeiki, X.C., Goodluck-Griffith, J., and Thilly, W.G. (2000). Mismatch repair deficient human cells: spontaneous and MNNG-induced mutational spectra in the HPRT gene. Mutat Res 450, 125-138.

Tornaletti, S. (2009). DNA repair in mammalian cells: Transcription-coupled DNA repair: directing your effort where it's most needed. Cellular and molecular life sciences : CMLS 66, 1010-1020.

Tweedale, G. (2002). Asbestos and its lethal legacy. Nature reviews Cancer 2, 311-315.

Unfried, K., Schurkes, C., and Abel, J. (2002). Distinct spectrum of mutations induced by crocidolite asbestos: clue for 8-hydroxydeoxyguanosine-dependent mutagenesis in vivo. Cancer research 62, 99-104.

van Zeeland, A.A., Vreeswijk, M.P., de Gruijl, F.R., van Kranen, H.J., Vrieling, H., and Mullenders, L.F. (2005). Transcription-coupled repair: impact on UV-induced mutagenesis in cultured rodent cells and mouse skin tumors. Mutation research 577, 170-178.

Vesley, D. (1999). Ionizing and Nonionizing Radiation. In, H.H.a.t. Environment, ed. (Springer), pp. 65-74.

Viguera, E., Canceill, D., and Ehrlich, S.D. (2001). Replication slippage involves DNA polymerase pausing and dissociation. The EMBO journal 20, 2587-2595.

Vilenchik, M.M., and Knudson, A.G. (2003). Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer. Proceedings of the National Academy of Sciences of the United States of America 100, 12871-12876.

Vogelstein, B., and Kinzler, K.W. (1992). Carcinogens leave fingerprints. Nature 355, 209-210.

von Hansemann, D. (1890). Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. Virchow's Arch Path Anat 119.

Wakelin, L.P. (1986). Polyfunctional DNA intercalating agents. Medicinal research reviews 6, 275-340.

Wang, K., Kan, J., Yuen, S.T., Shi, S.T., Chu, K.M., Law, S., Chan, T.L., Kan, Z., Chan, A.S., Tsui, W.Y., et al. (2011). Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. Nature genetics 43, 1219-1223.

Wang, Q.Q., Begum, R.A., Day, V.W., and Bowman-James, K. (2012). Sulfur, oxygen, and nitrogen mustards: stability and reactivity. Organic & biomolecular chemistry 10, 8786-8793.

Ward, J.F. (1988). In Progress in Nucleic Acid Research and Molecular Biology, W.E. Cohn, and K. Moldave, eds. (Academic Press, Inc.), pp. 95-121.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737-738.

Wei, X., Walia, V., Lin, J.C., Teer, J.K., Prickett, T.D., Gartner, J., Davis, S., Stemke-Hale, K., Davies, M.A., Gershenwald, J.E., et al. (2011). Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nature genetics 43, 442-446.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell 150, 264-278.

Wiegand, K.C., Shah, S.P., Al-Agha, O.M., Zhao, Y., Tse, K., Zeng, T., Senz, J., McConechy, M.K., Anglesio, M.S., Kalloger, S.E., et al. (2010). ARID1A mutations in endometriosis-associated ovarian carcinomas. The New England journal of medicine 363, 1532-1543.

Wilson, D.M., 3rd, and Bohr, V.A. (2007). The mechanics of base excision repair, and its relationship to aging and disease. DNA repair 6, 544-559.

Wiseman, H., and Halliwell, B. (1996). Damage to DNA by reactive oxygen and nitrogen species: role in inflammatory disease and progression to cancer. The Biochemical journal 313 (Pt 1), 17-29.

Witkin, E.M. (1969). Ultraviolet-induced mutation and DNA repair. Annual review of microbiology 23, 487-514.

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

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. Science 318, 1108-1113.

Wu, J., Jiao, Y., Dal Molin, M., Maitra, A., de Wilde, R.F., Wood, L.D., Eshleman, J.R., Goggins, M.G., Wolfgang, C.L., Canto, M.I., et al. (2011). Whole-exome sequencing of neoplastic cysts of the pancreas reveals recurrent mutations in components of ubiquitin-dependent pathways. Proceedings of the National Academy of Sciences of the United States of America 108, 21188-21193.

You, Y.H., Lee, D.H., Yoon, J.H., Nakajima, S., Yasui, A., and Pfeifer, G.P. (2001). Cyclobutane pyrimidine dimers are responsible for the vast majority of mutations induced by UVB irradiation in mammalian cells. The Journal of biological chemistry 276, 44688-44694.

Zang, Z.J., Cutcutache, I., Poon, S.L., Zhang, S.L., McPherson, J.R., Tao, J., Rajasegaran, V., Heng, H.L., Deng, N., Gan, A., et al. (2012). Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nature genetics 44, 570-574.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328, 916-919.

Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., et al. (2012). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature 481, 157-163.

Zhang, J., Wu, G., Miller, C.P., Tatevossian, R.G., Dalton, J.D., Tang, B., Orisme, W., Punchihewa, C., Parker, M., Qaddoumi, I., et al. (2013). Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. Nature genetics 45, 602-612.

Zheng, C.-H., Huang, D.-S., Sun, Z.-L., Lyu, M.R., and Lok, T.-M. (2006). Nonnegative independent component analysis based on minimizing mutual information technique. Neurocomputing 69, 878-883.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

8-oxoG	7,8-dihydro-8-oxoguanine
AID	Activation-induced cytosine deaminase
APEX1	Apurinic/apyrimidinic endonuclease
API	Application programming interface
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
BAM	Binary sequence alignment and mapping
BER	Base excision repair
BIR	Break-induced replication
BSS	Blind source separation
CGP	Cancer Genome Project
CPD	Cyclobutane pyrimidine dimer
D-loop	Displacement-loop
DA-NER	Domain associated nucleotide excision repair
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
DSBR	Classical double-strand break repair
FDR	False discovery rate
GG-NER	Global genome-wide nucleotide excision repair
GLM	Generalized linear model
GRCh37	Genome Reference Consortium human genome (build 37)
IARC	International Agency for Research on Cancer
ICA	Independent component analysis
ICGC	International Cancer Genome Consortium
IGHV	Immunoglobulin gene hypermutation
Indel	Small insertion/deletion
LP-BER	Long patch base excision repair
MMEJ	Microhomology mediated end joining
MMR	DNA mismatch repair
NER	Nucleotide excision repair
NHEJ	Non-homologous end joining

NHLBI	National Heart, Lung, and Blood Institute
NMF	Nonnegative matrix factorization
PAH	Polycyclic aromatic hydrocarbons
PCR	Polymerase chain reaction
Pol	Polymerase
POL II	RNA polymerase II
POLE	DNA polymerase epsilon catalytic subunit A
RFC	Replication factor C
RNA	Ribonucleic acid
RNS	Reactive nitrogen species
ROS	Reactive oxygen species
SDSA	Synthesis-dependent strand annealing
SNP	Single nucleotide polymorphism
SP-BER	Short patch base excision repair
SSA	Single-strand annealing
TC-BER	Transcription coupled base excision repair
TC-NER	Transcription coupled nucleotide excision repair
TCGA	The Cancer Genome Atlas
TET	Ten-eleven translocation methylcytosine dioxygenase
TP53	Tumour protein p53
UCSC	University of California, Santa Cruz
UV	Ultraviolet
UV-A	Ultraviolet A
UV-B	Ultraviolet B
UV-C	Ultraviolet B

APPENDIX I: Alphabets of mutational types

This appendix contains information for the alphabets of mutation types used throughout the course of this thesis. These alphabets were termed Ξ_6, Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536} in chapter 2 and each one will be discussed in more details in the next few sections.

Mutational alphabet Ξ_6

The Ξ_6 alphabet is perhaps the simplest possible alphabet as it considers only the six types of somatic substitutions: **C>A**, **C>G**, **C>T**, **T>A**, **T>C**, **T>G**. All mutations are denoted using the pyrimidine of the Watson-Crick base pair as the reference and, in this appendix, these substitutions are coloured consistently with the way they are plotted in the majority of figures throughout this thesis.

Mutational alphabet Ξ_{96}

The Ξ_{96} alphabet provides greater resolution for examining the six types of single nucleotide variants (*i.e.,* the Ξ_6 alphabet) by including the immediate sequence context of each mutated base. In this alphabet, a mutation type contains a somatic substitution and both the 5' and 3' base next to the somatic mutation. For example, a **C>T** mutation can be characterized as …**TpCpG**…>…**TpTpG**… (mutated base underlined and presented as the pyrimidine partner of the mutated base pair) generating 96 possible mutation types – (6 types of substitutions) * (4 types of 5' bases) * (4 types of 3' bases). Table listing each of the 96 substitution types, the reference trinucleotide, and the mutated trinucleotide is provided below.

Sub	Ref	Mut	Sub	Ref	Mut
C>A	ApCpA	ApApA	T>A	ApTpA	ApApA
C>A	ApCpC	ApApC	T>A	ApTpC	ApApC
C>A	ApCpG	ApApG	T>A	ApTpG	ApApG
C>A	ApCpT	ApApT	T>A	ApTpT	ApApT
C>A	CpCpA	CpApA	T>A	CpTpA	CpApA
C>A	CpCpC	CpApC	T>A	CpTpC	CpApC
C>A	CpCpG	CpApG	T>A	CpTpG	CpApG
C>A	CpCpT	CpApT	T>A	CpTpT	CpApT
C>A	GpCpA	GpApA	T>A	GpTpA	GpApA
C>A	GpCpC	GpApC	T>A	GpTpC	GpApC
C>A	GpCpG	GpApG	T>A	GpTpG	GpApG
C>A	GpCpT	GpApT	T>A	GpTpT	GpApT

C>A	TpCpA	TpApA	T>A	TpTpA	TpApA
C>A	TpCpC	TpApC	T>A	TpTpC	TpApC
C>A	TpCpG	TpApG	T>A	TpTpG	TpApG
C>A	TpCpT	TpApT	T>A	TpTpT	TpApT
C>G	ApCpA	ApGpA	T>C	ApTpA	ApCpA
C>G	ApCpC	ApGpC	T>C	ApTpC	ApCpC
C>G	ApCpG	ApGpG	T>C	ApTpG	ApCpG
C>G	ApCpT	ApGpT	T>C	ApTpT	ApCpT
C>G	CpCpA	CpGpA	T>C	CpTpA	CpCpA
C>G	CpCpC	CpGpC	T>C	CpTpC	CpCpC
C>G	CpCpG	CpGpG	T>C	CpTpG	CpCpG
C>G	CpCpT	CpGpT	T>C	CpTpT	CpCpT
C>G	GpCpA	GpGpA	T>C	GpTpA	GpCpA
C>G	GpCpC	GpGpC	T>C	GpTpC	GpCpC
C>G	GpCpG	GpGpG	T>C	GpTpG	GpCpG
C>G	GpCpT	GpGpT	T>C	GpTpT	GpCpT
C>G	TpCpA	TpGpA	T>C	TpTpA	TpCpA
C>G	TpCpC	TpGpC	T>C	TpTpC	TpCpC
C>G	TpCpG	TpGpG	T>C	TpTpG	TpCpG
C>G	TpCpT	TpGpT	T>C	TpTpT	TpCpT
C>T	ApCpA	ApTpA	T>G	ApTpA	ApGpA
C>T	ApCpC	ApTpC	T>G	ApTpC	ApGpC
C>T	ApCpG	ApTpG	T>G	ApTpG	ApGpG
C>T	ApCpT	ApTpT	T>G	ApTpT	ApGpT
C>T	CpCpA	CpTpA	T>G	CpTpA	CpGpA
C>T	CpCpC	CpTpC	T>G	CpTpC	CpGpC
C>T	CpCpG	CpTpG	T>G	CpTpG	CpGpG
C>T	CpCpT	CpTpT	T>G	CpTpT	CpGpT
C>T	GpCpA	GpTpA	T>G	GpTpA	GpGpA
C>T	GpCpC	GpTpC	T>G	GpTpC	GpGpC
C>T	GpCpG	GpTpG	T>G	GpTpG	GpGpG
C>T	GpCpT	GpTpT	T>G	GpTpT	GpGpT
C>T	TpCpA	TpTpA	T>G	TpTpA	TpGpA
C>T	TpCpC	TpTpC	T>G	TpTpC	TpGpC
C>T	TpCpG	TpTpG	T>G	TpTpG	TpGpG
C>T	TpCpT	TpTpT	T>G	TpTpT	TpGpT

Mutational alphabet Ξ_{99}

The Ξ_{99} alphabet extends Ξ_{96} by including three additional mutation types, *viz.,* (i) double nucleotide substitutions, (ii) small insertions or deletions at short tandem repeats, and (iii) small insertions or deletions overlapping with microhomologies at breakpoints.

Mutational alphabet Ξ_{192}

The Ξ_{192} alphabet elaborates Ξ_{96} by considering the transcriptional strand on which a substitution resides. In contrast to all other alphabets, Ξ_{192} is defined only in the regions of the genome where transcription occurs, which in these analyses has been limited to the genomic footprints of protein coding genes. For example, the **C>T** mutations at **TpCpA** are split into two categories: the **C>T** mutations at **TpCpA** occurring on the untranscribed strand of a gene and the **C>T** mutations at **TpCpA** occurring on the transcribed strand. Similarly, all 96 mutations types from Ξ_{96} are extended to form the Ξ_{192} alphabet.

Mutational alphabet Ξ_{1536}

The Ξ_{1536} further extends Ξ_{96} by including two bases 5' and 3' to the mutated base resulting in 1,536 possible mutated pentanucleotides - (6 types of substitutions) * (16 types of the two immediate 5' bases) * (16 types of the two immediate 3' bases). For example, using the Ξ_{1536} alphabet, one of the 256 subclasses of a **C>T** mutation is …**ApTpCpGpC**… > …**ApTpTpGpC**… For brevity, the complete list of mutation types included in Ξ_{1536} is not provided here.

APPENDIX II: List of analysed samples

This appendix contains a summary list of all samples analysed throughout the course of this thesis. Summarized information is provided for all 7,042 separated by sequencing types (exome sequencing versus whole-genome sequencing), cancer types, and respective data sources. It should be noted that the pilocytic astrocytomas dataset contains a small number of other paediatric low-grade gliomas and paediatric low-grade glioneuronal tumours. Information for each individual sample including its mutational catalogues and somatic mutations (both before and after filtering) could be found at ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl.

Exome sample types and data sources	Total
ALL	**140**
doi:10.1038/nature10725	15
doi:10.1038/ng.2508	29
doi:10.1038/ng.2532	42
New unpublished samples	54
AML	**147**
TCGA data portal	147
Bladder	**136**
TCGA data portal	136
Breast	**844**
doi:10.1038/nature10933	63
doi:10.1038/nature11017	9
New unpublished samples	5
TCGA data portal	767
Cervix	**38**
TCGA data portal	38
CLL	**103**
doi:10.1038/ng.1032	80
ICGC data portal	23
Colorectum	**559**
doi:10.1038/nature11282	70
TCGA data portal	489
Oesophageal	**146**
doi:10.1038/ng.2591	146
Glioblastoma	**98**
ICGC data portal	50
TCGA data portal	48
Glioma Low Grade	**217**
TCGA data portal	217

Whole-genome sample types and data sources	Total
ALL	**1**
New unpublished samples	1
AML	**7**
doi:10.1038/nature10738	7
Breast	**119**
doi:10.1016/j.cell.2012.04.024	21
New unpublished samples	98
CLL	**28**
doi:10.1038/nature10113	4
New unpublished samples	24
Liver	**88**
ICGC data portal	66
New unpublished samples	22
Lung Adenocarcinoma	**24**
doi:10.1016/j.cell.2012.08.029	24
Lymphoma B-cell	**24**
doi:10.1038/ng.2468	1
New unpublished samples	23
Medulloblastoma	**100**
New unpublished samples	100
Pancreas	**15**
New unpublished samples	15
Pilocytic Astrocytoma	**101**
doi:10.1038/ng.2611	38
New unpublished samples	63
Grand Total	**507**

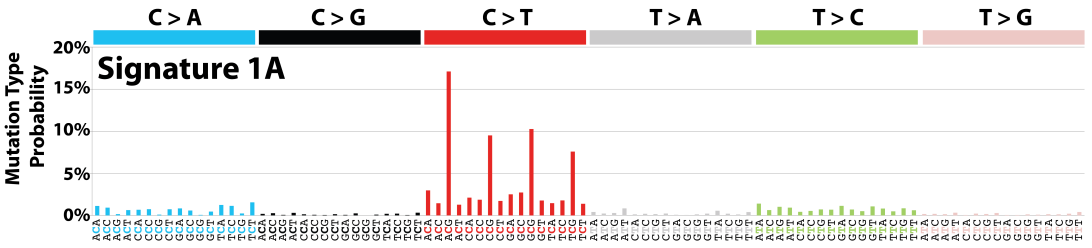
Head and Neck	**380**
doi:10.1126/science.1206923	12
doi:10.1126/science.1208130	68
TCGA data portal	300
Kidney Chromophobe	**65**
TCGA data portal	65
Kidney Clear Cell	**325**
doi:10.1038/ng.1014	10
doi:10.1038/ng.2323	7
TCGA data portal	308
Kidney Papillary	**100**
TCGA data portal	100
Lung Adenocarcinoma	**636**
doi:10.1016/j.cell.2012.08.029	150
doi:10.1038/nature07423	30
doi:10.1101/gr.145144.112	75
TCGA data portal	381
Lung Small Cell	**70**
doi:10.1038/ng.2396	29
doi:10.1038/ng.2405	40
ICGC data portal	1
Lung Squamous	**176**
TCGA data portal	176
Lymphoma B-cell	**24**
doi:10.1038/nature10351	16
doi:10.1038/ng.2468	8
Melanoma	**396**
doi:10.1016/j.cell.2012.06.024	92
doi:10.1038/nature11071	28
doi:10.1038/ng.1041	8
ICGC data portal	1
New unpublished samples	17
TCGA data portal	250
Myeloma	**69**
New unpublished samples	69
Neuroblastoma	**210**
doi:10.1038/ng.2493	13
doi:10.1038/ng.2529	197
Ovary	**471**
doi:10.1126/science.1196333	8
TCGA data portal	463
Pancreas	**98**
doi:10.1073/pnas.1118046108	22
doi:10.1126/science.120060	10
ICGC data portal	37

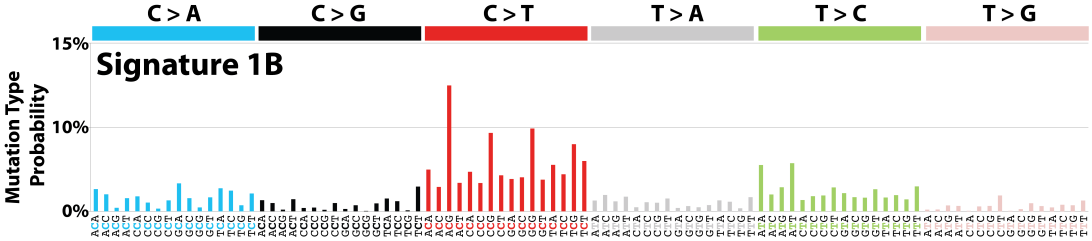
TCGA data portal	29
Prostate	**330**
doi:10.1038/nature09744	7
doi:10.1038/nature11125	61
doi:10.1038/ng.2279	112
TCGA data portal	150
Stomach	**212**
doi:10.1038/ng.2246	14
doi:10.1038/ng.982	22
ICGC data portal	10
TCGA data portal	166
Thyroid	**304**
TCGA data portal	304
Uterus	**241**
TCGA data portal	241
Grand Total	**6,535**

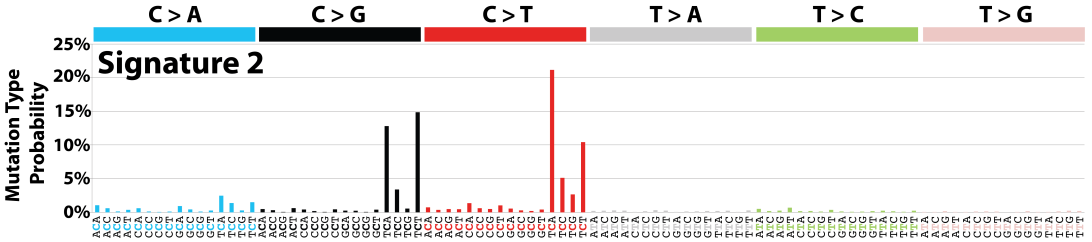
APPENDIX III: Mutational signatures in human cancer

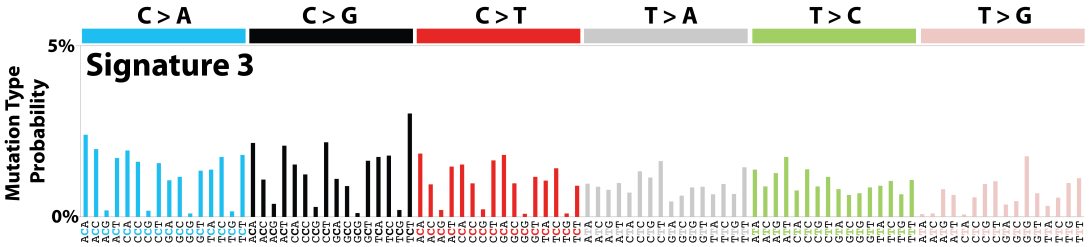
This appendix contains high-resolution figures for the twenty-seven consensus mutational signatures that were deciphered by applying the developed computational approach across the spectrum of human cancer (chapter 4). Each mutational signature is shown using the same plot. Signatures are displayed based on the trinucleotide frequency of the human genome. The probability bars for each of the six types of substitutions as well as the mutated bases are displayed in different colours. The mutation types are displayed on the horizontal axes, while vertical axes depict the percentages of mutations attributed to specific mutation types. The plots are ordered by signature validation types: validated mutational signatures (Signatures 1A, 1B, 2 through 21), mutational signatures that failed validation (Signatures R1, R2, and R3), and mutational signatures for which it was not possible to perform validation (Signatures U1 and U2).

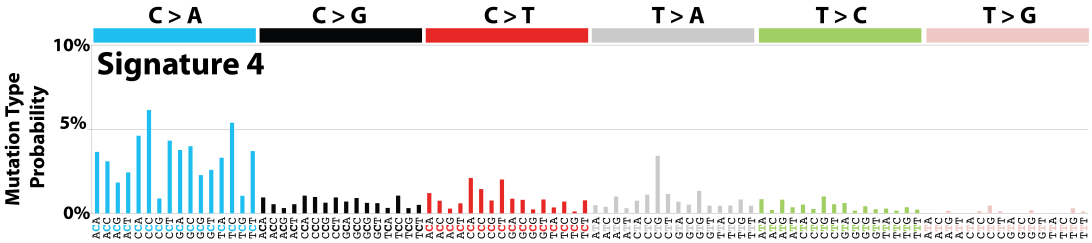
Validated mutational signatures

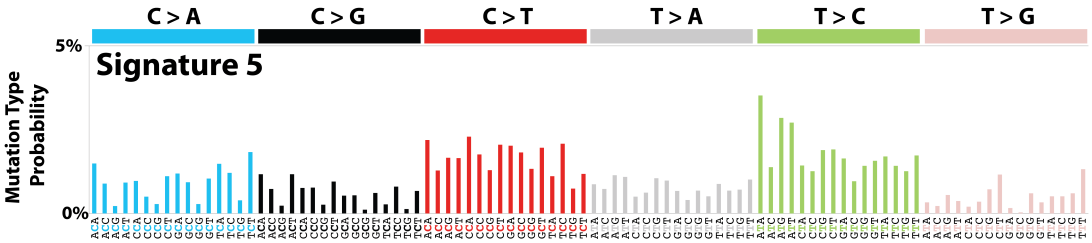


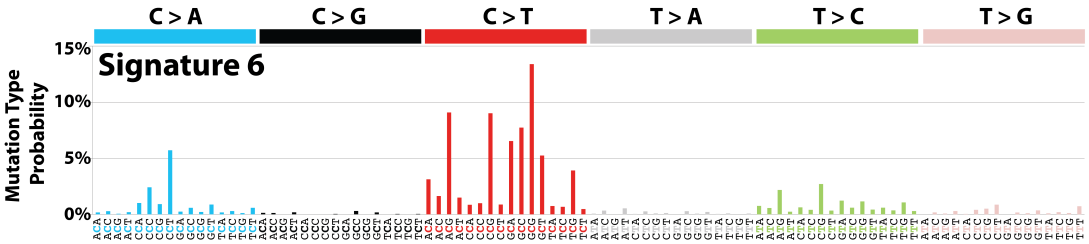


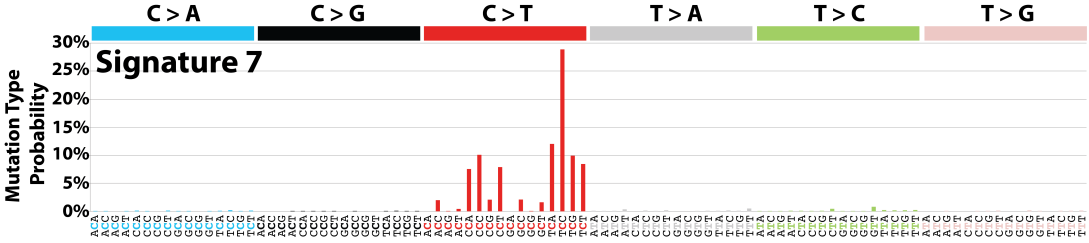


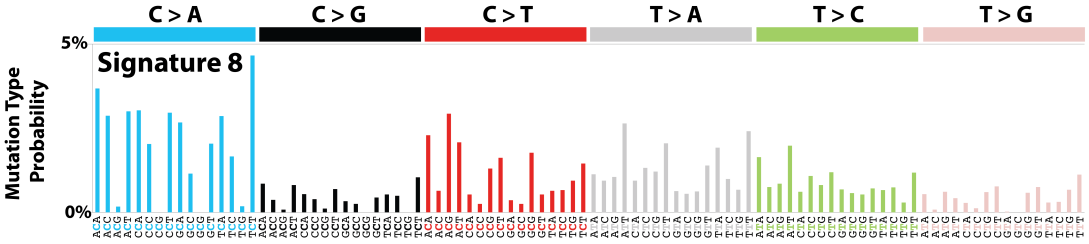


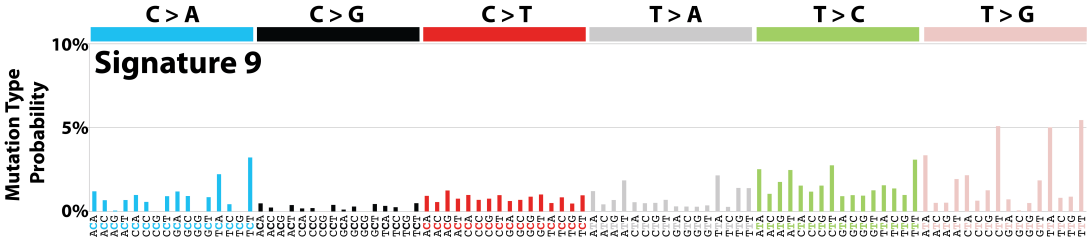


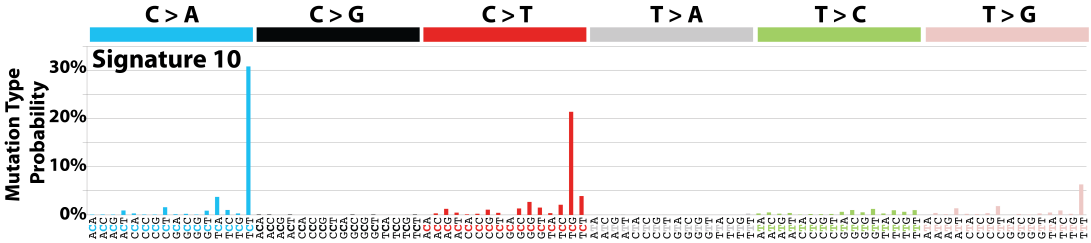


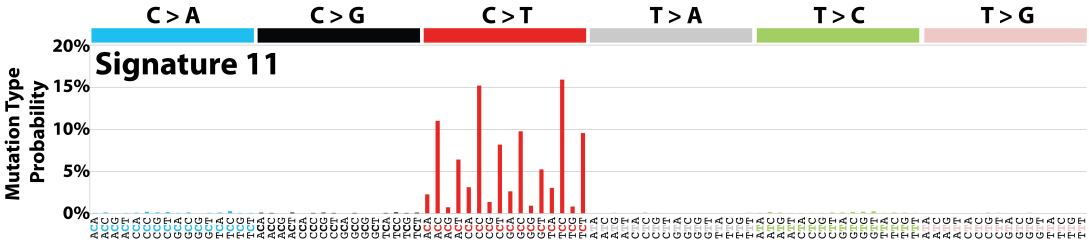

C > A
C > G
C > T
T > A
T > C
T > G
Signature 6
Mutation Type Probability
15%
10%
5%
0%

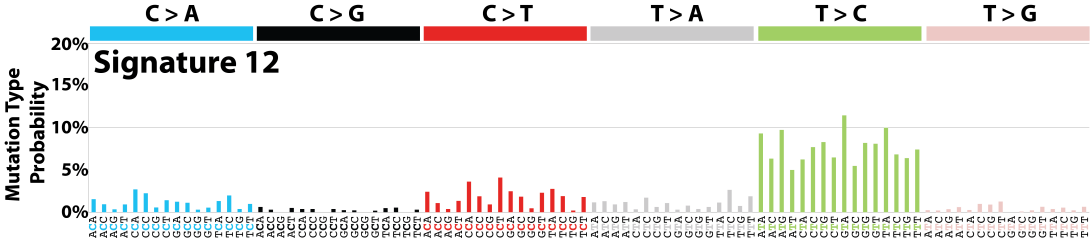

C > A
C > G
C > T
T > A
T > C
T > G
Signature 7
Mutation Type Probability
30%
25%
20%
15%
10%
5%
0%

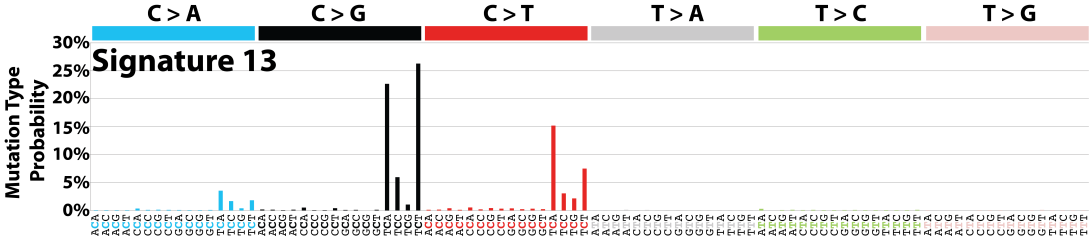

C > A
C > G
C > T
T > A
T > C
T > G
Signature 8
Mutation Type Probability
5%
0%

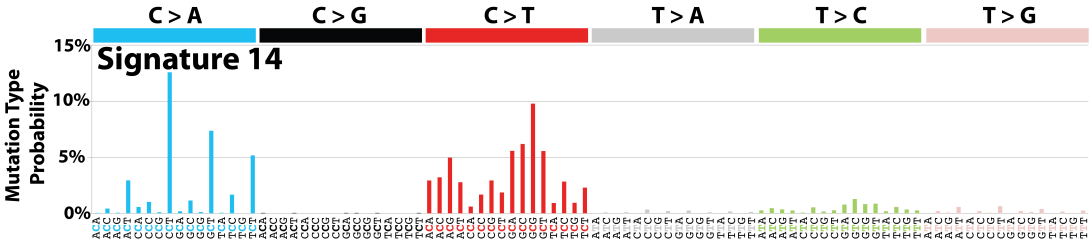

C > A
C > G
C > T
T > A
T > C
T > G
Signature 9
Mutation Type Probability
10%
5%
0%

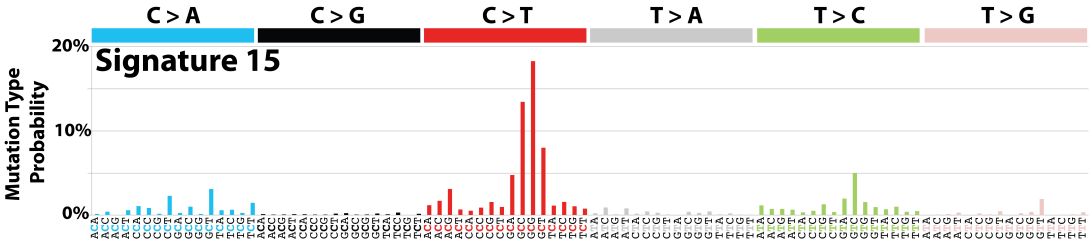

C > A
C > G
C > T
T > A
T > C
T > G
Signature 10
Mutation Type Probability
30%
20%
10%
0%

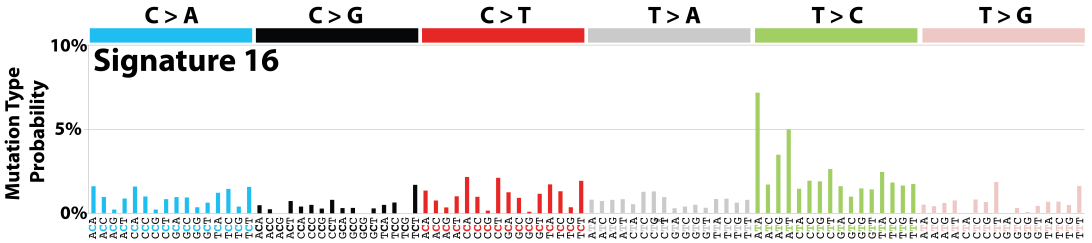

C > A
C > G
C > T
T > A
T > C
T > G
Signature 11
Mutation Type Probability
20%
15%
10%
5%
0%

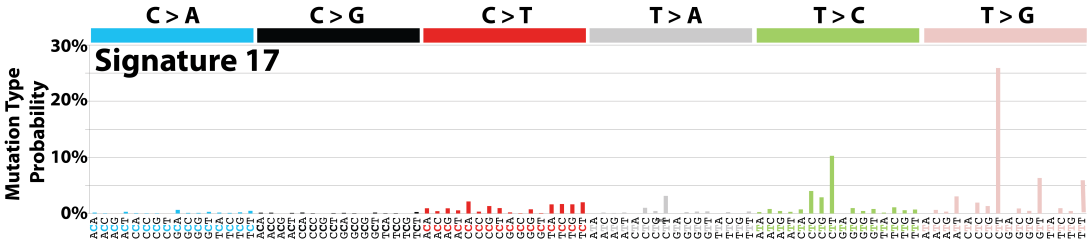

C > A
C > G
C > T
T > A
T > C
T > G
Signature 12
Mutation Type Probability
20%
15%
10%
5%
0%

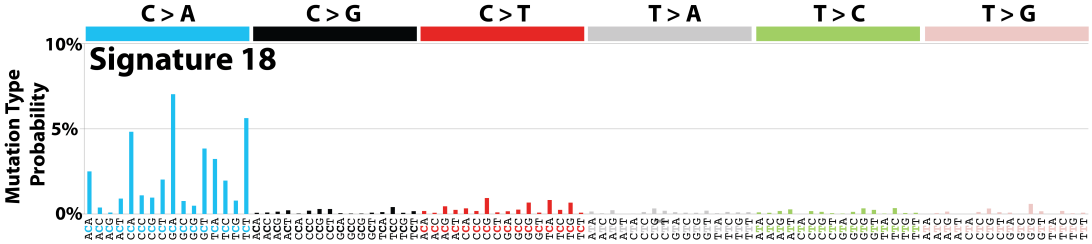

C > A
C > G
C > T
T > A
T > C
T > G
Signature 13
Mutation Type Probability
30%
25%
20%
15%
10%
5%
0%

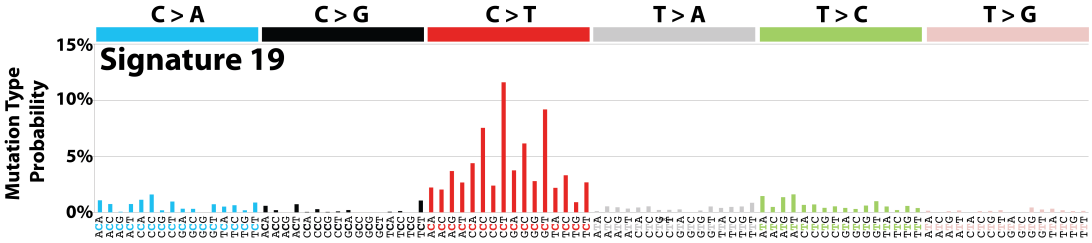

C > A
C > G
C > T
T > A
T > C
T > G
Signature 14
Mutation Type Probability
15%
10%
5%
0%

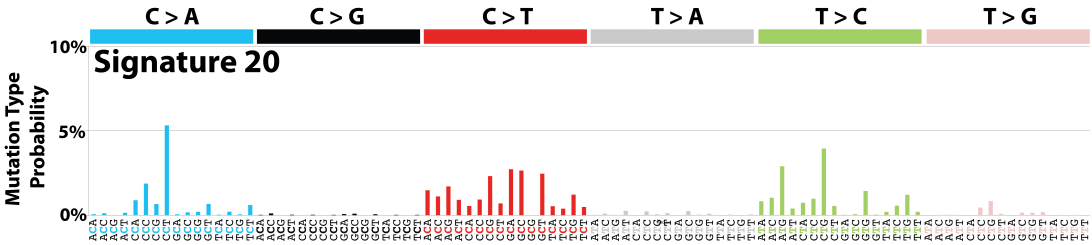

C > A
C > G
C > T
T > A
T > C
T > G
Signature 15
Mutation Type Probability
20%
10%
0%

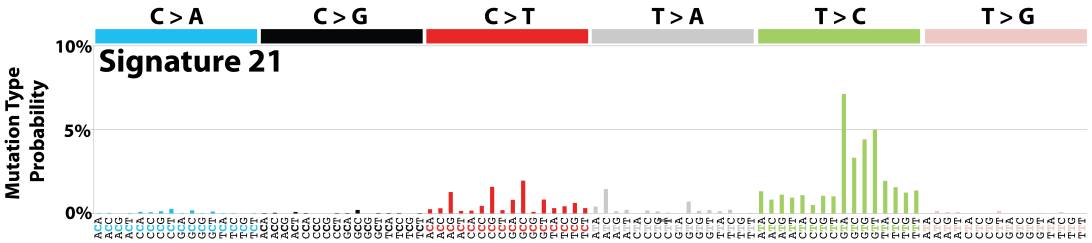

C > A
C > G
C > T
T > A
T > C
T > G
Signature 16
Mutation Type Probability
10%
5%
0%


C > A
C > G
C > T
T > A
T > C
T > G
Signature 17
Mutation Type Probability
30%
20%
10%
0%

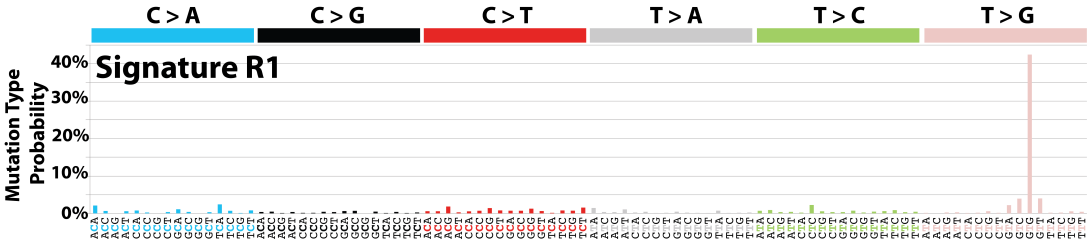

C > A
C > G
C > T
T > A
T > C
T > G
Signature 18
Mutation Type Probability
10%
5%
0%

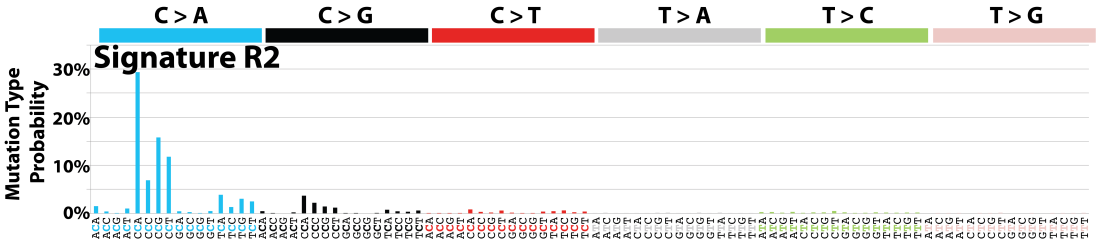

C > A
C > G
C > T
T > A
T > C
T > G
Signature 19
Mutation Type Probability
15%
10%
5%
0%

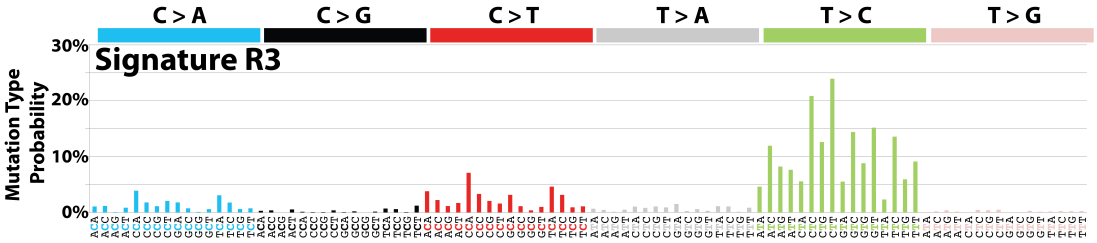

C > A
C > G
C > T
T > A
T > C
T > G
Signature 20
Mutation Type Probability
10%
5%
0%


C > A
C > G
C > T
T > A
T > C
T > G
Signature 21
Mutation Type Probability
10%
5%
0%

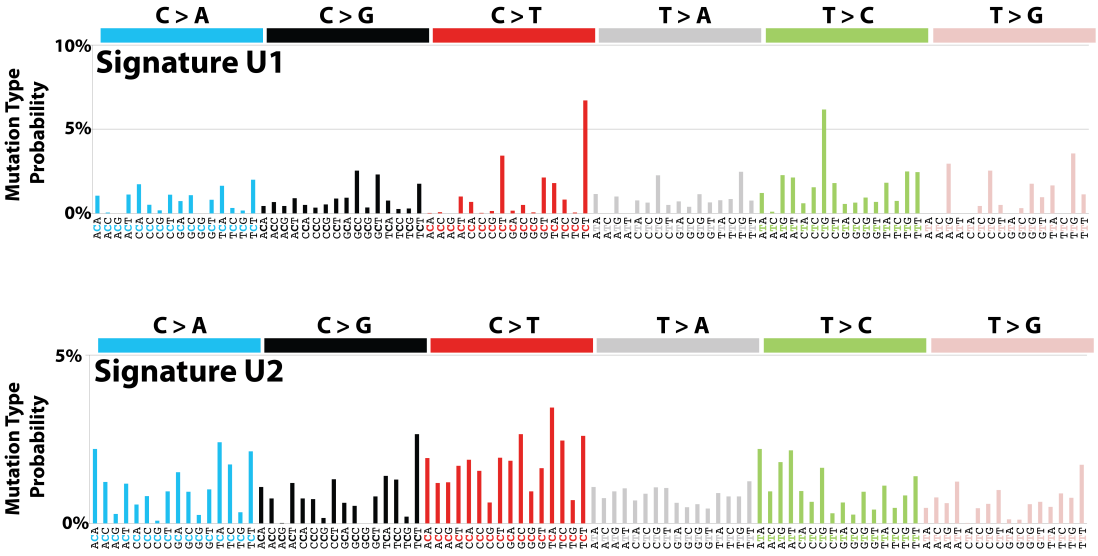
Mutational signatures that failed validation





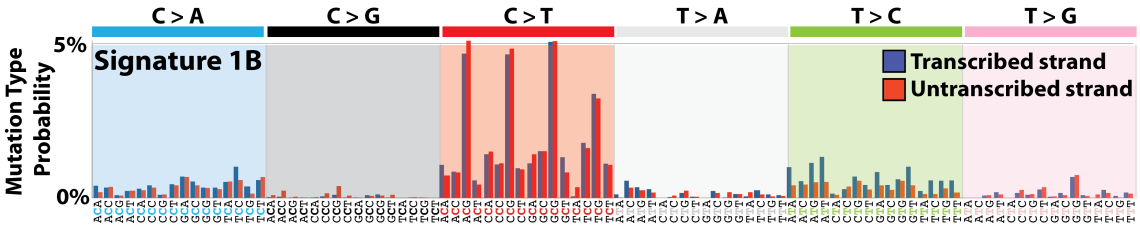


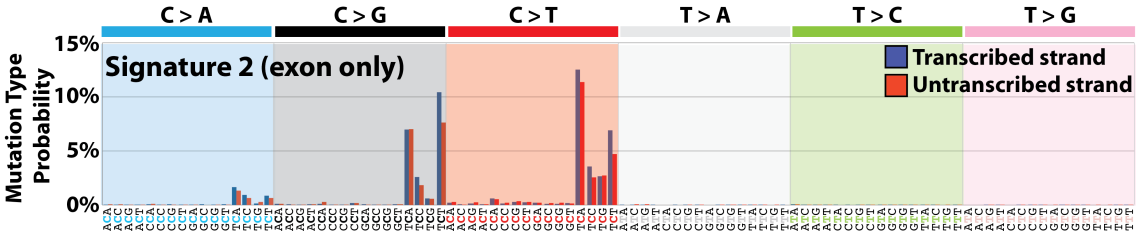
Mutational signatures for which validation was not performed

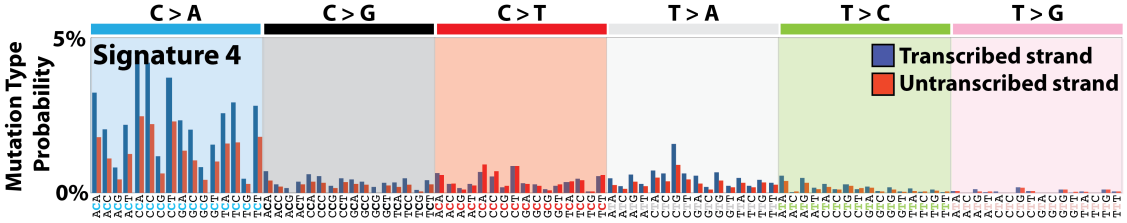


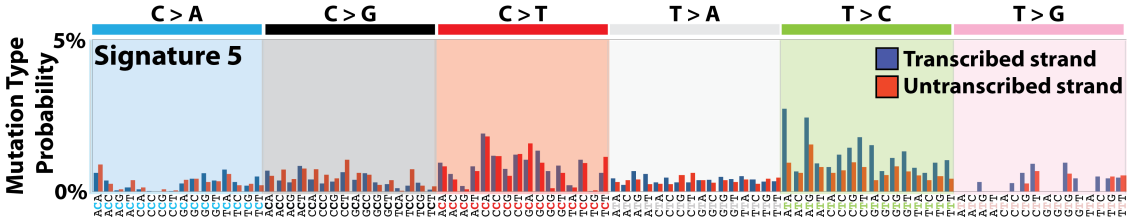
APPENDIX IV: Mutational signatures with transcriptional strand-bias

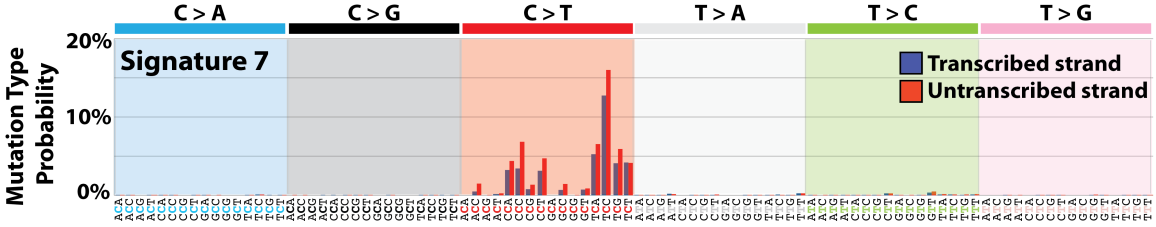
This appendix contains high-resolution figures for the nine consensus mutational signatures that exhibit transcriptional strand-bias. Each mutational signature is shown using the same figure format based on a 192 substitution classification incorporating the substitution type, the sequence context immediately 5' and 3' to the mutated base and whether the mutated base (in pyrimidine context) is on the transcribed or untranscribed strand. The panels for each of the six types of substitutions as well as the mutated bases are displayed in different colours. Mutations on the transcribed pyrimidine strand are displayed in blue while mutations on the untranscribed strand are displayed in red.

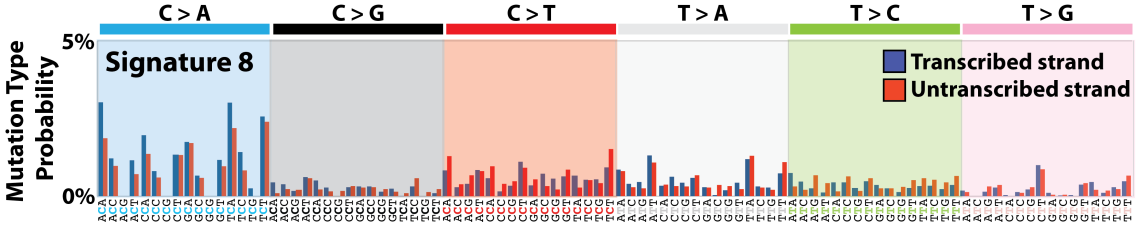

C > A
C > G
C > T
T > A
T > C
T > G
Signature 1B
Transcribed strand
Untranscribed strand
Mutation Type Probability
5%
0%

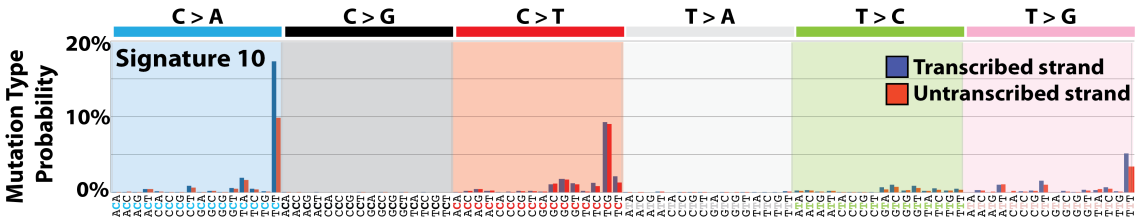

C > A
C > G
C > T
T > A
T > C
T > G
Signature 2 (exon only)
Transcribed strand
Untranscribed strand
Mutation Type Probability
15%
10%
5%
0%

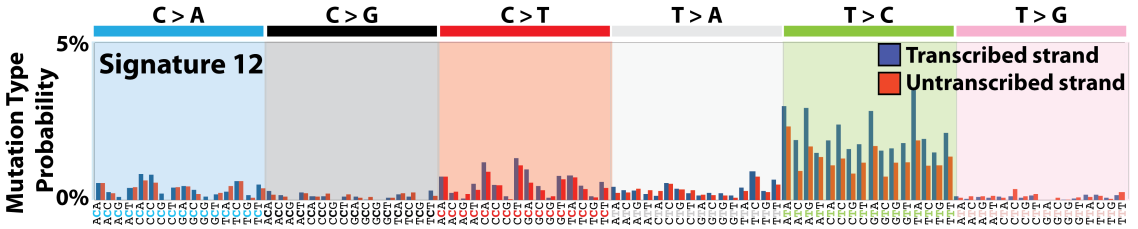

C > A
C > G
C > T
T > A
T > C
T > G
Signature 4
Transcribed strand
Untranscribed strand
Mutation Type Probability
5%
0%

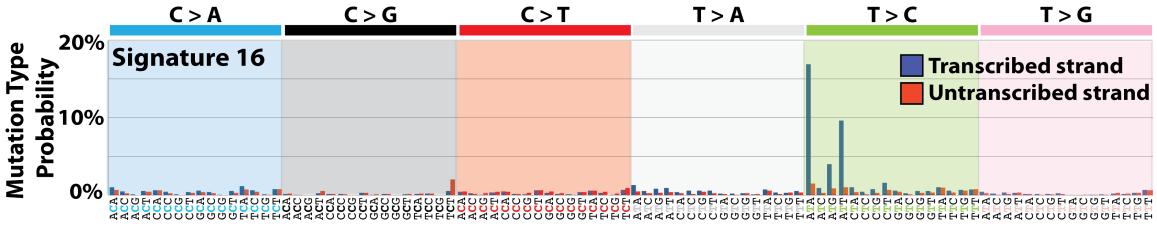

C > A
C > G
C > T
T > A
T > C
T > G
Signature 5
Transcribed strand
Untranscribed strand
Mutation Type Probability
5%
0%


C > A
C > G
C > T
T > A
T > C
T > G
Signature 7
Transcribed strand
Untranscribed strand
Mutation Type Probability
20%
10%
0%


C > A
C > G
C > T
T > A
T > C
T > G
Signature 8
Transcribed strand
Untranscribed strand
Mutation Type Probability
5%
0%


C > A
C > G
C > T
T > A
T > C
T > G
Signature 10
Mutation Type Probability
20%
10%
0%
Transcribed strand
Untranscribed strand

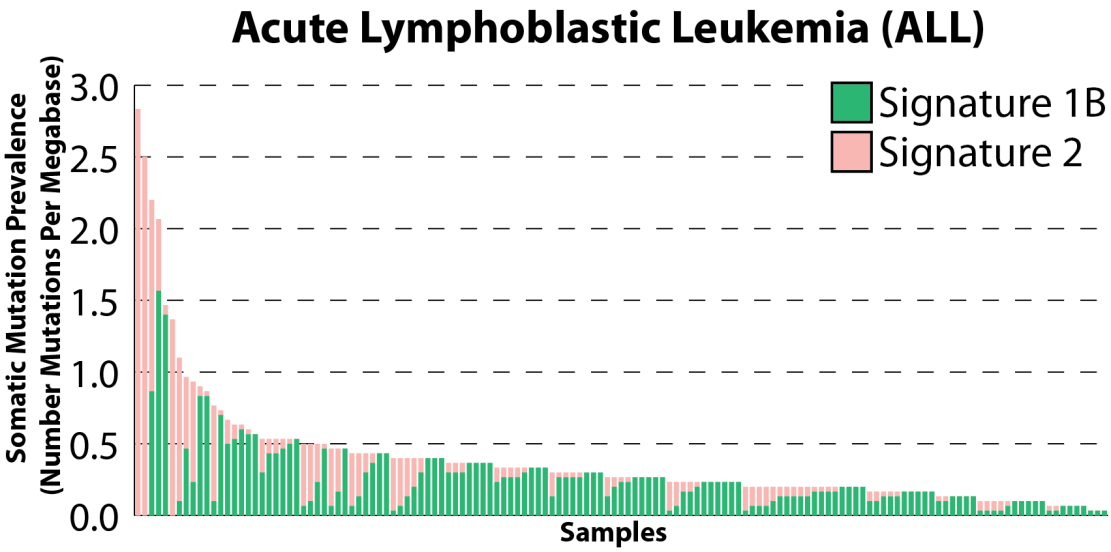

C > A
C > G
C > T
T > A
T > C
T > G
Signature 12
Mutation Type Probability
5%
0%
Transcribed strand
Untranscribed strand

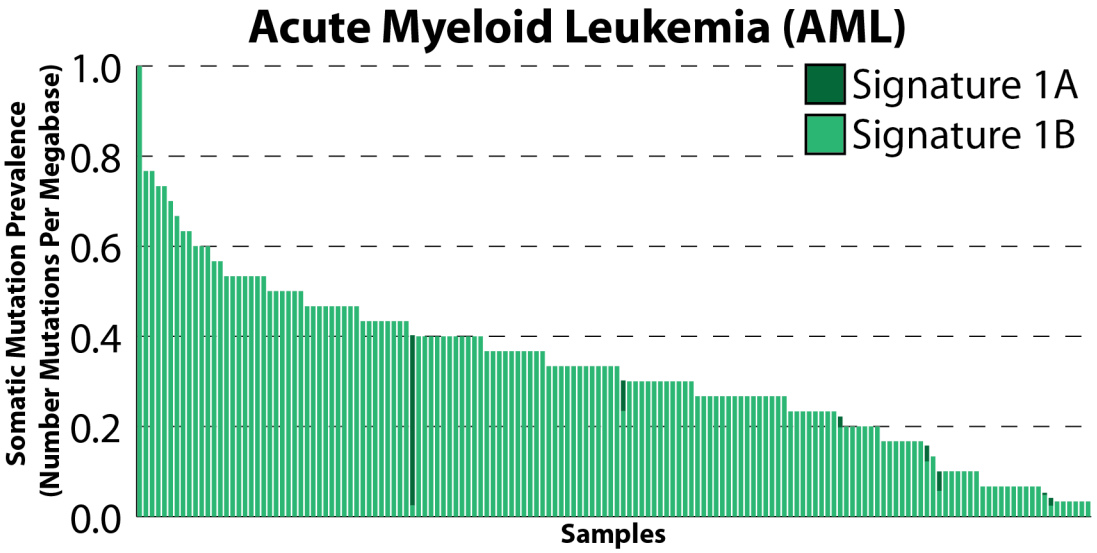

C > A
C > G
C > T
T > A
T > C
T > G
Signature 16
Mutation Type Probability
20%
10%
0%
Transcribed strand
Untranscribed strand

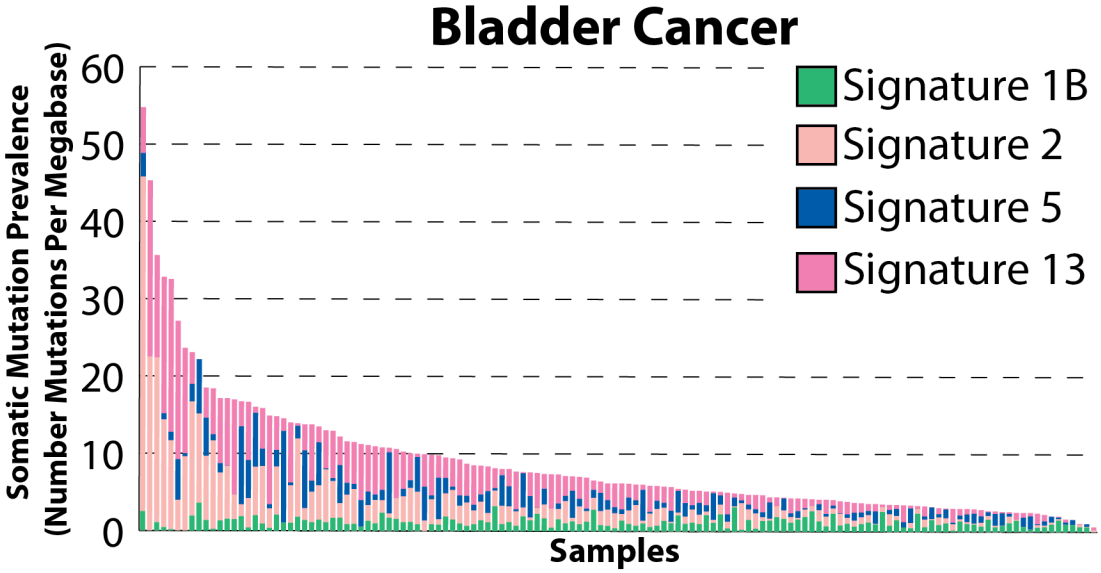
APPENDIX V: Contributions of mutational signatures in individual samples

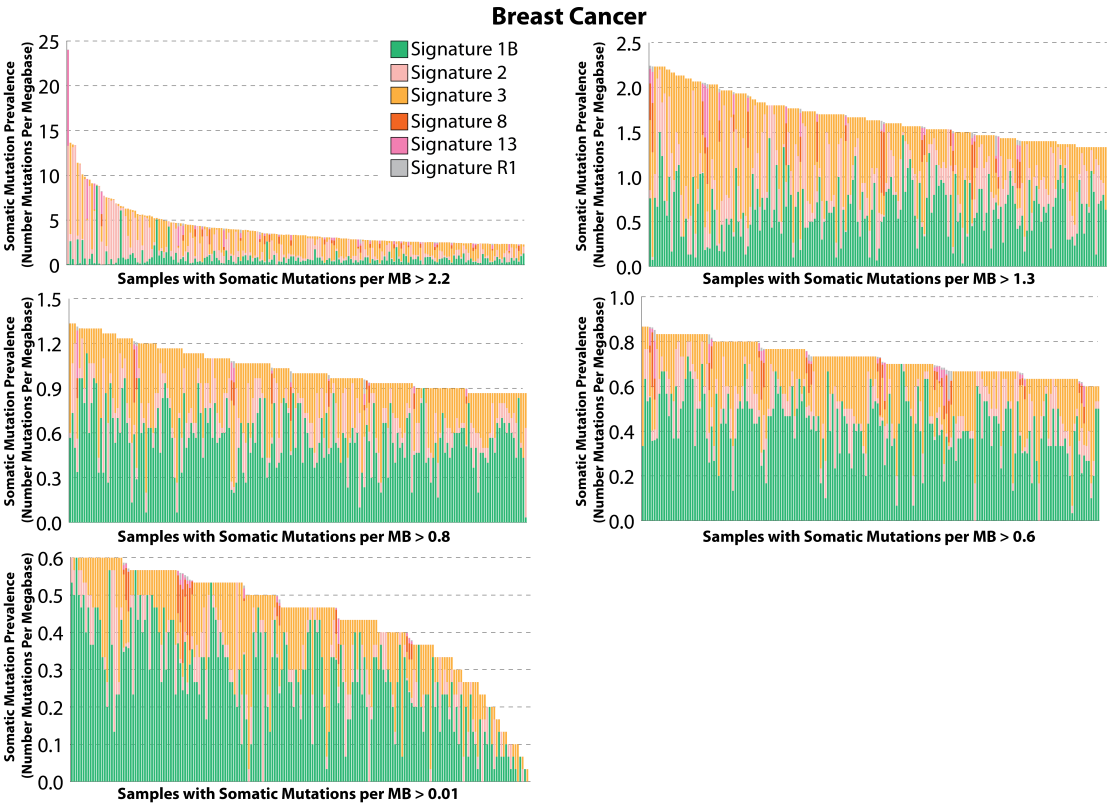
This appendix contains a high-resolution figure for each of the 30 examined cancer types (chapter 4). Each figure depicts all of the samples in a single cancer type and shows the contributions of the consensus mutational signatures (found in that cancer type) for each sample. All figures use the same format: samples are displayed on the horizontal axis, sorted in descending order based on the numbers of somatic mutations per megabase found in each sample, and the somatic mutation prevalence is displayed on the vertical axis. Mutational signatures are displayed in distinct colours, consistent in all figures. For clarity, several panels are provided (and clearly labelled) when the number of samples is too high or the somatic prevalence differs significantly between samples. Figures are displayed on individual pages, labelled to clearly show the names of the cancer types, and they are ordered alphabetically based on the names of these cancer types. In general, all samples are displayed in each cancer type and the two exceptions are denoted with an asterisk in the appropriate figures and listed below:

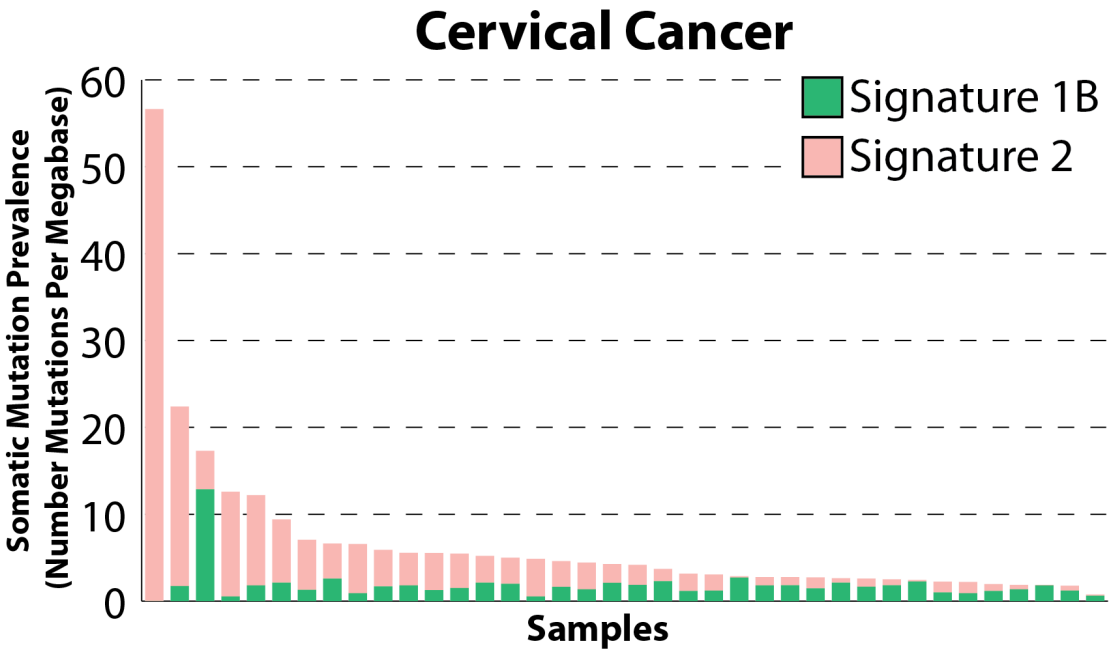
- For clarity, in glioma low grade, one hypermutator sample purely of Signature 14 (254 mutations per MB) is not displayed.
- In lung squamous, one hypermutator sample purely of Signature 7 (72 mutations per MB) is not displayed. Signature 7 is associated with exposure to ultraviolet light, an unlikely carcinogen for lung cancer. As such, this TCGA sample is most likely either a melanoma metastasis or a misannotated sample. Thus, the association between Signature 7 and lung squamous has not been discussed in chapter 4 and this association has not been displayed in Figure 4.9.

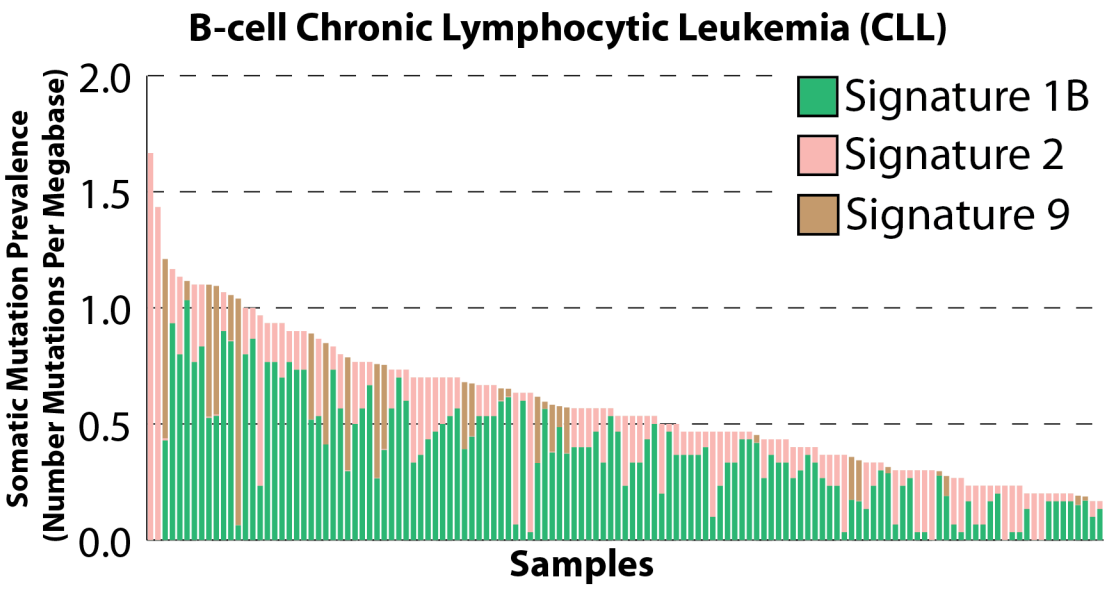

Acute Lymphoblastic Leukemia (ALL)
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Signature 1B
Signature 2
Samples

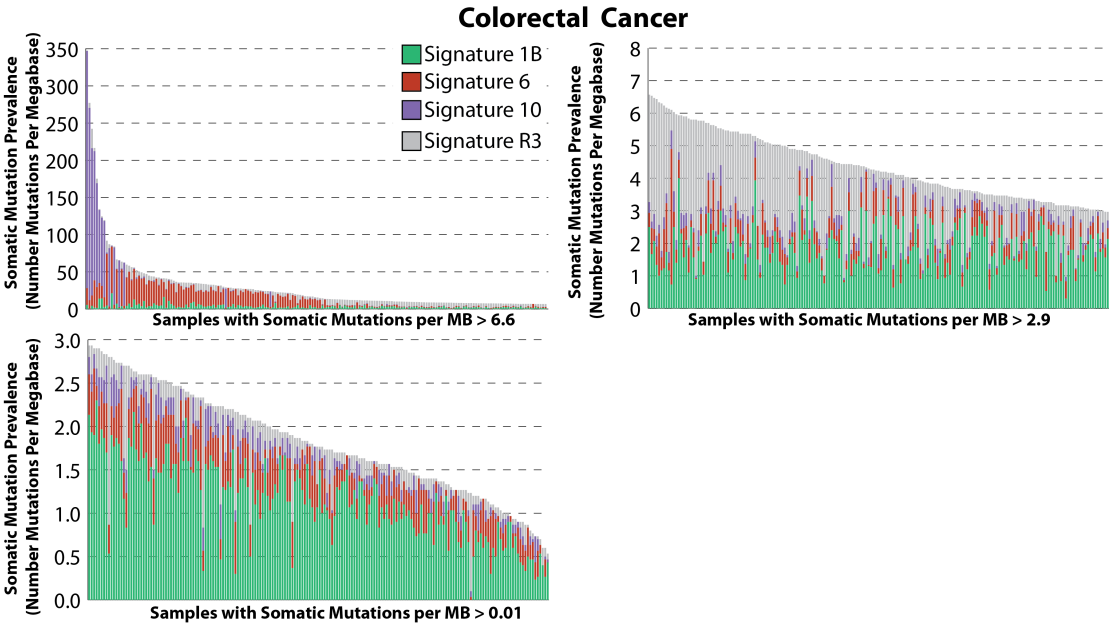

Acute Myeloid Leukemia (AML)
Signature 1A
Signature 1B
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
1.0
0.8
0.6
0.4
0.2
0.0
Samples

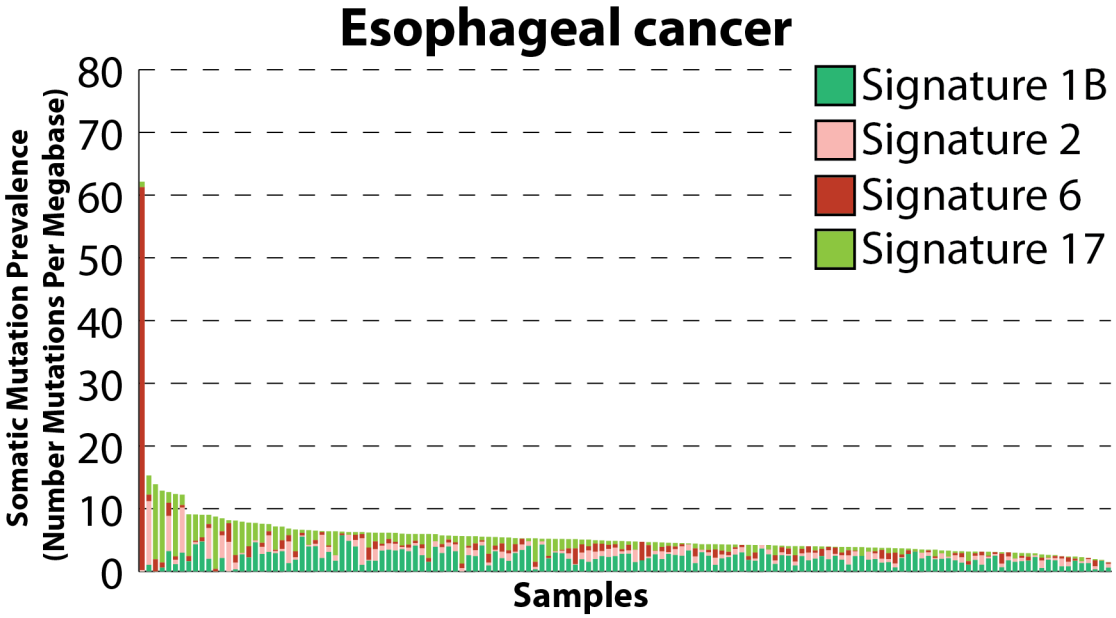

Bladder Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
60
50
40
30
20
10
0
Samples
Signature 1B
Signature 2
Signature 5
Signature 13

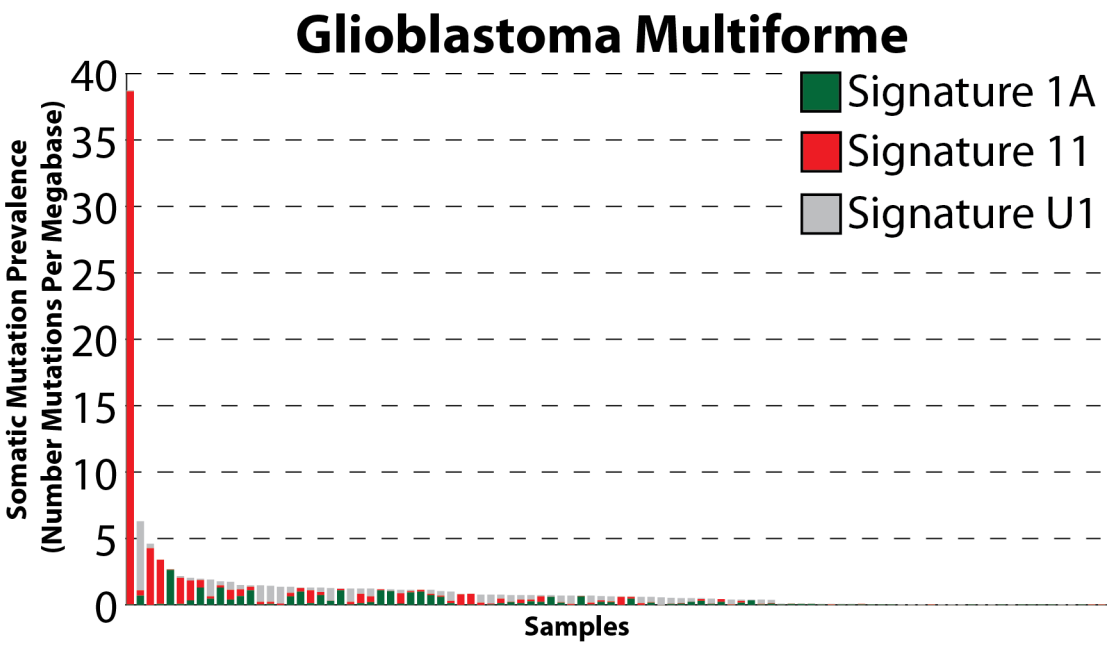

Breast Cancer
Signature 1B
Signature 2
Signature 3
Signature 8
Signature 13
Signature R1
Somatic Mutation Prevalence (Number Mutations Per Megabase)
Samples with Somatic Mutations per MB > 2.2
Samples with Somatic Mutations per MB > 1.3
Samples with Somatic Mutations per MB > 0.8
Samples with Somatic Mutations per MB > 0.6
Samples with Somatic Mutations per MB > 0.01

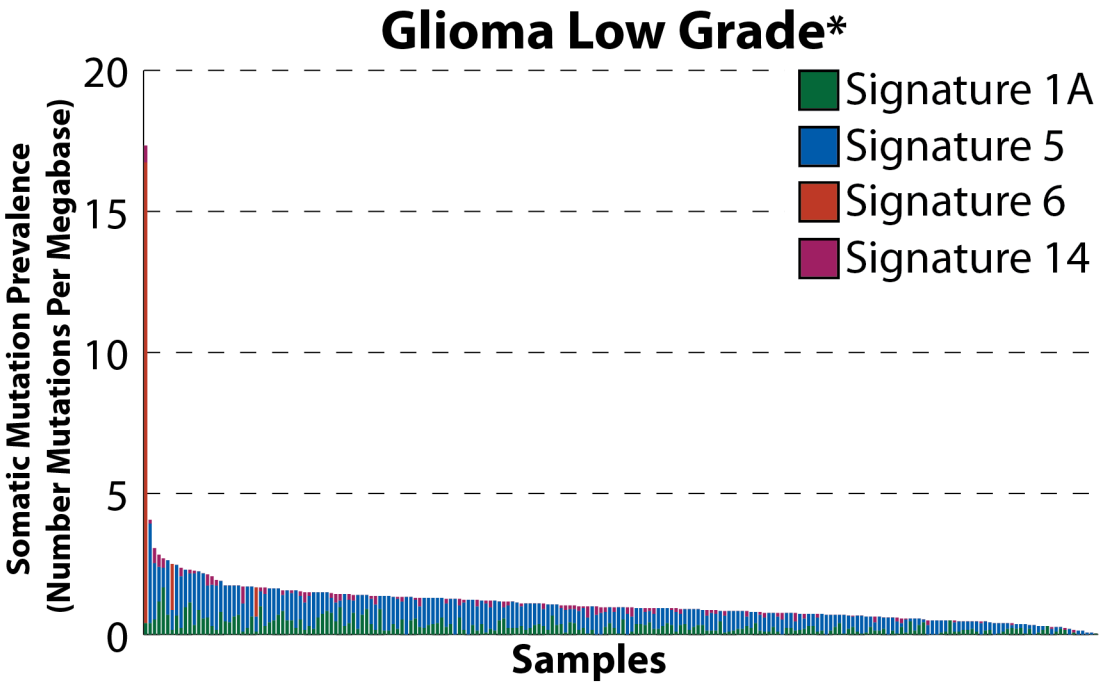

Cervical Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
60
50
40
30
20
10
0
Signature 1B
Signature 2
Samples

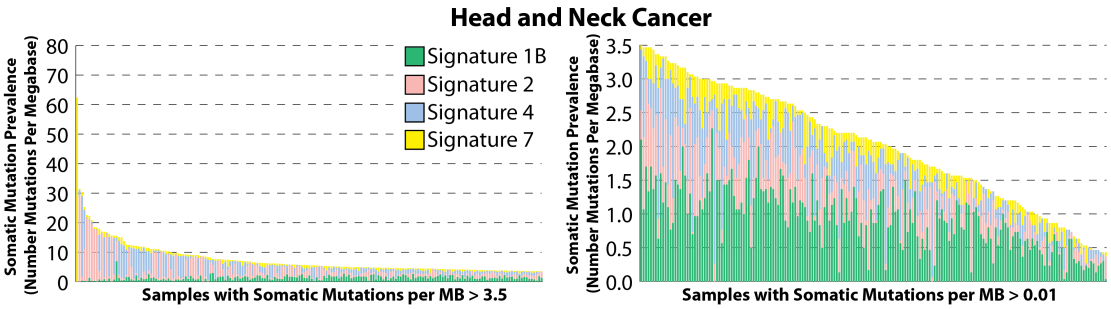

B-cell Chronic Lymphocytic Leukemia (CLL)
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
2.0
1.5
1.0
0.5
0.0
Signature 1B
Signature 2
Signature 9
Samples

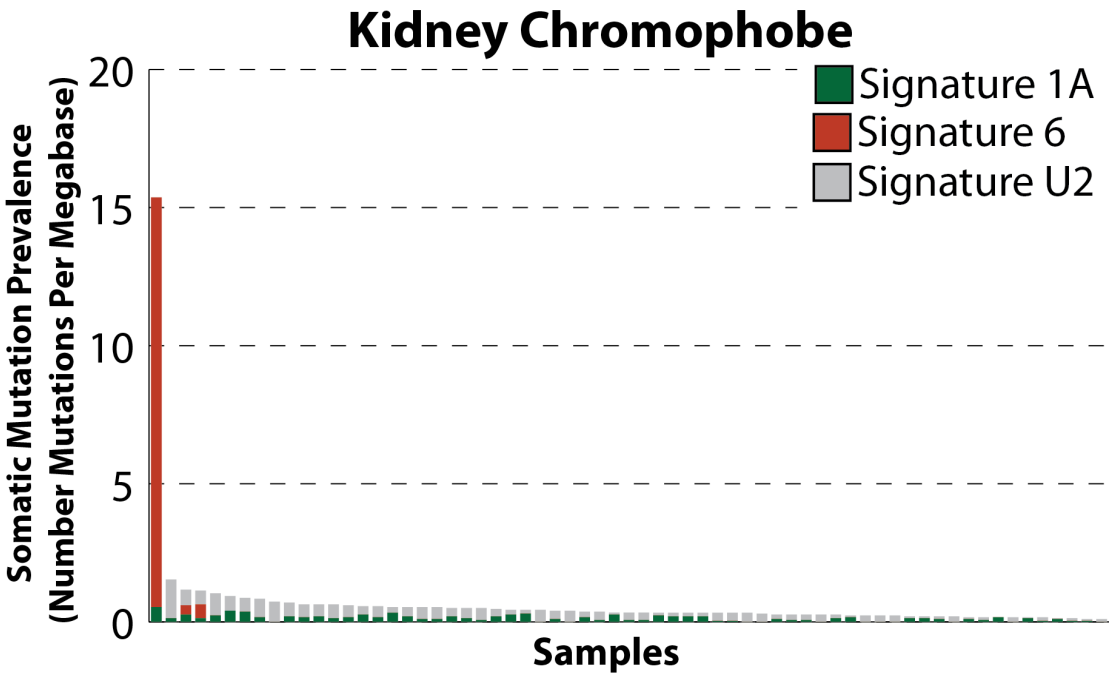

Colorectal Cancer
Somatic Mutation Prevalence (Number Mutations Per Megabase)
Signature 1B
Signature 6
Signature 10
Signature R3
Samples with Somatic Mutations per MB > 6.6
Samples with Somatic Mutations per MB > 2.9
Samples with Somatic Mutations per MB > 0.01

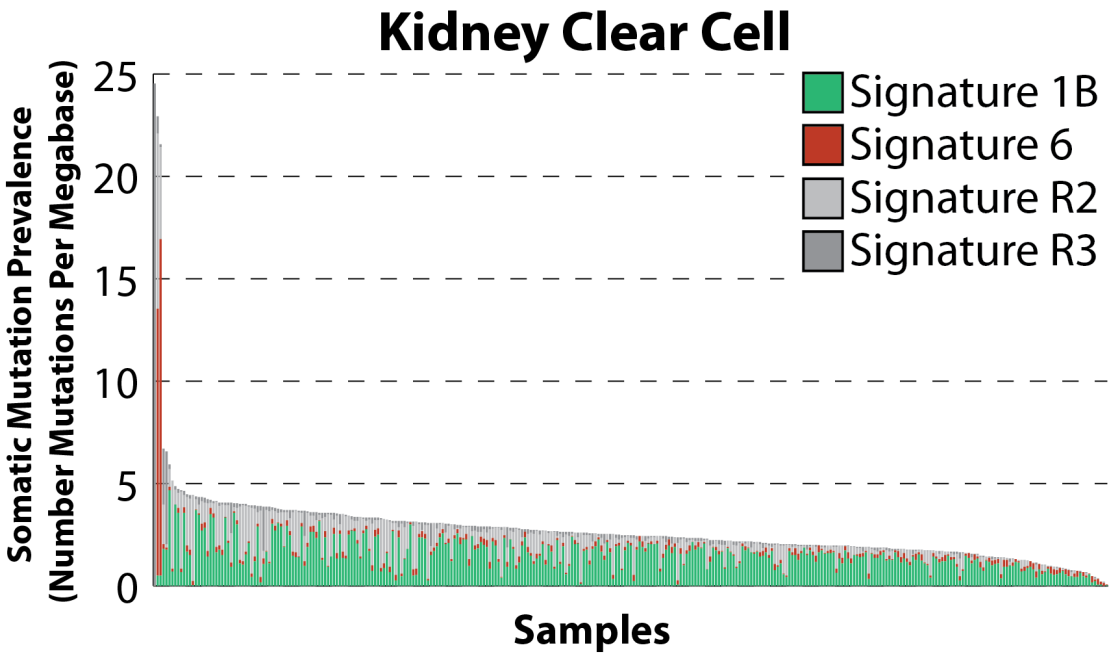

Esophageal cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
80
70
60
50
40
30
20
10
0
Samples
Signature 1B
Signature 2
Signature 6
Signature 17

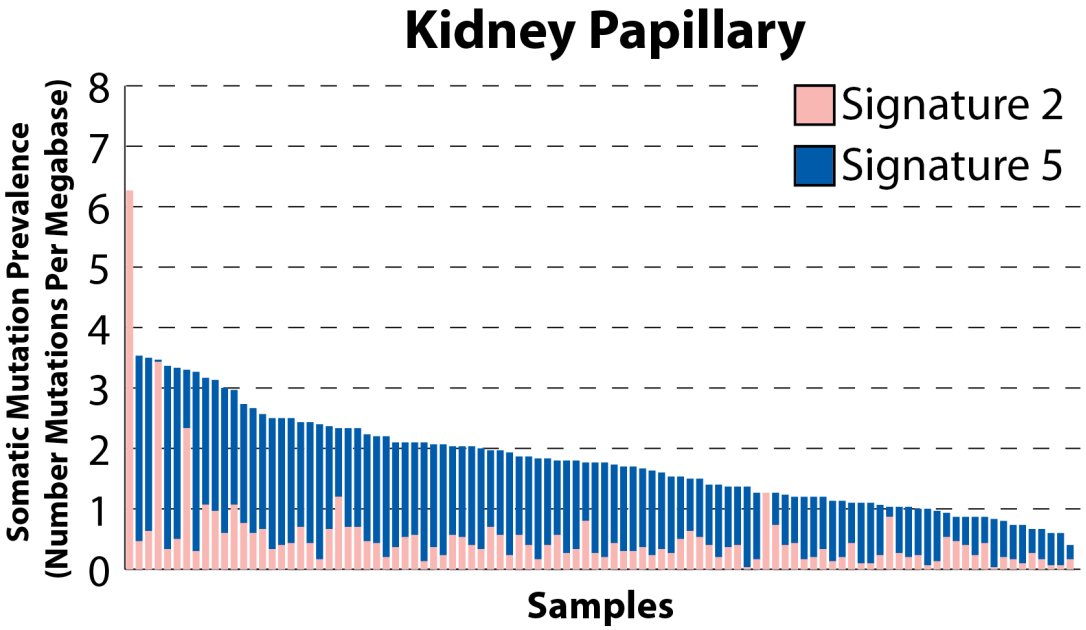

Glioblastoma Multiforme
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
40
35
30
25
20
15
10
5
0
Samples
Signature 1A
Signature 11
Signature U1

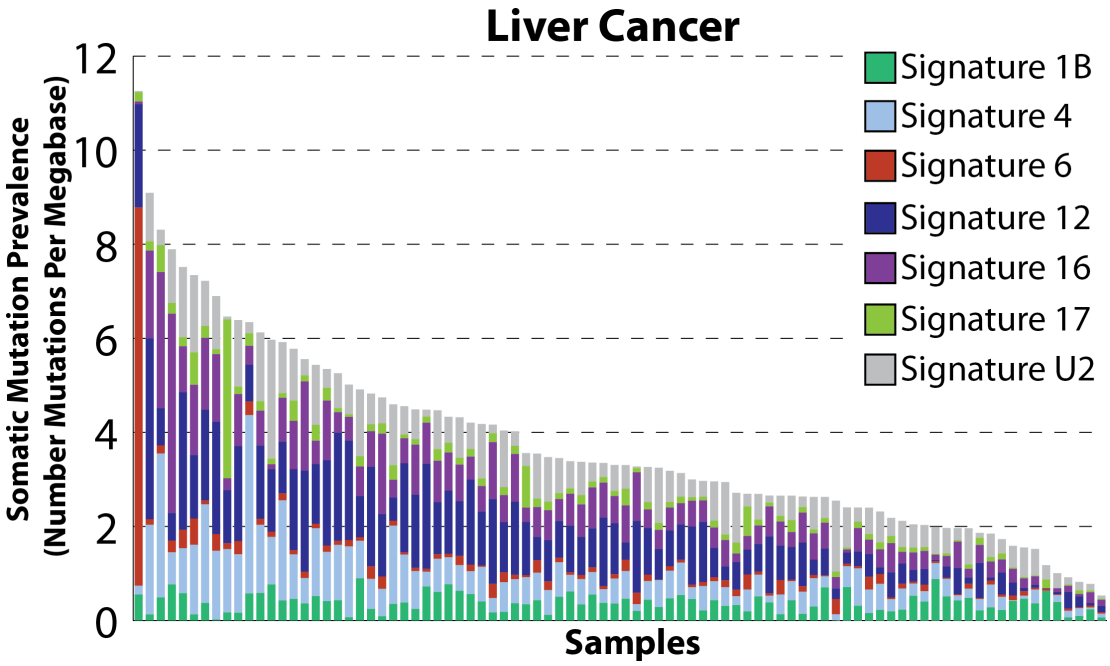

Glioma Low Grade*
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
20
15
10
5
0
Samples
Signature 1A
Signature 5
Signature 6
Signature 14


Head and Neck Cancer
Somatic Mutation Prevalence (Number Mutations Per Megabase)
80
70
60
50
40
30
20
10
0
Signature 1B
Signature 2
Signature 4
Signature 7
Samples with Somatic Mutations per MB > 3.5
Somatic Mutation Prevalence (Number Mutations Per Megabase)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Samples with Somatic Mutations per MB > 0.01

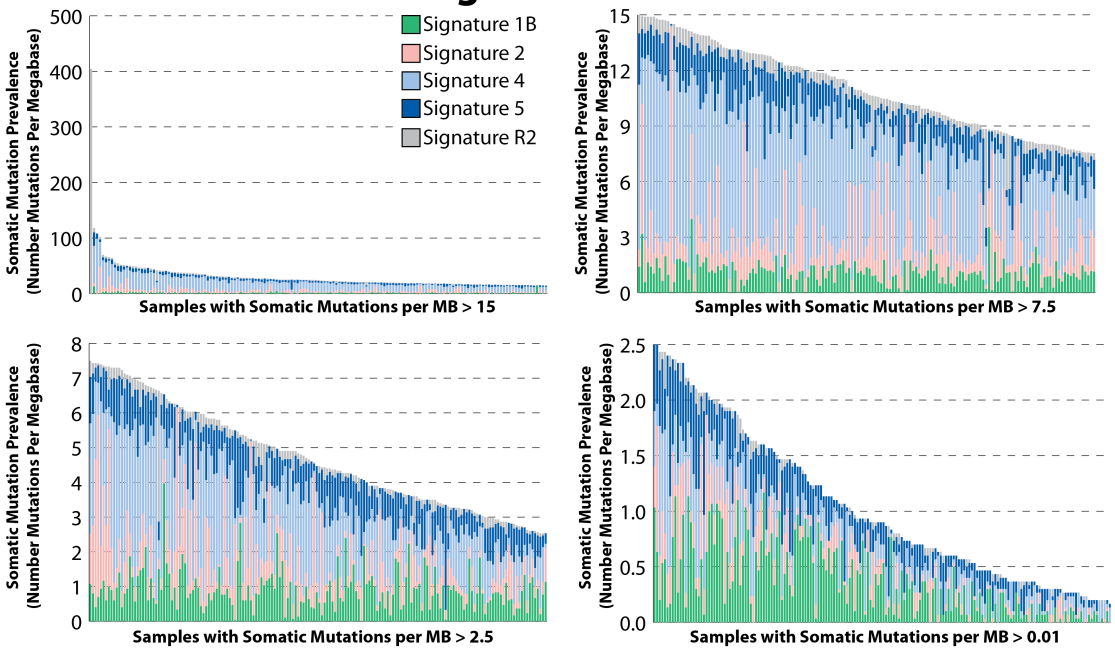

Kidney Chromophobe
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
20
15
10
5
0
Samples
Signature 1A
Signature 6
Signature U2

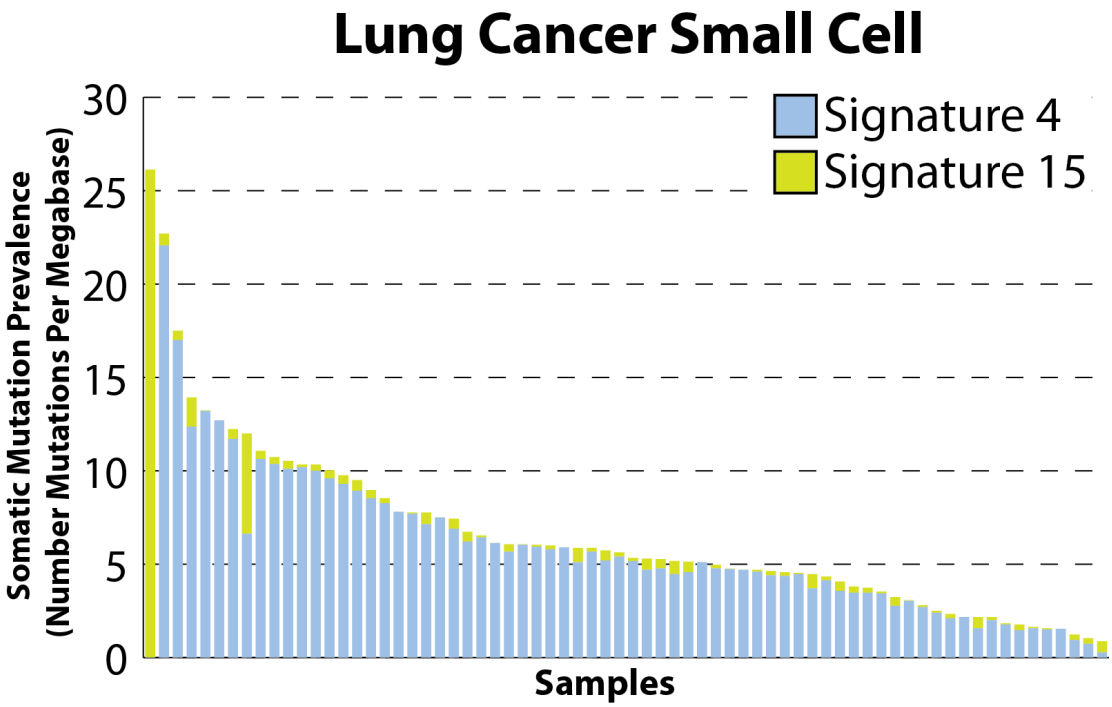

Kidney Clear Cell
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
25
20
15
10
5
0
Signature 1B
Signature 6
Signature R2
Signature R3
Samples

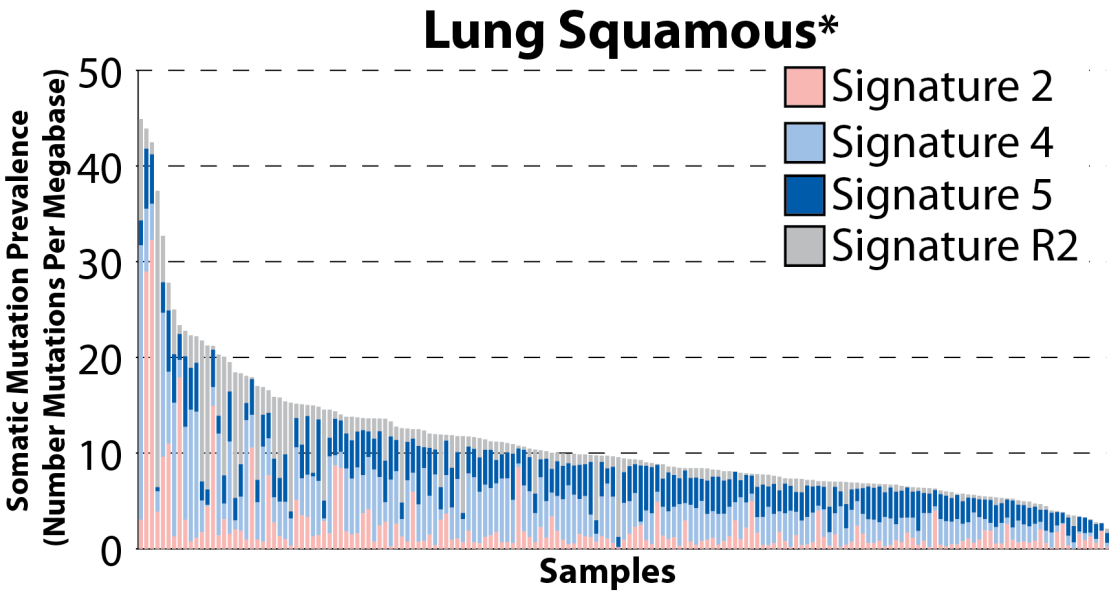

Kidney Papillary
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
8
7
6
5
4
3
2
1
0
Signature 2
Signature 5
Samples

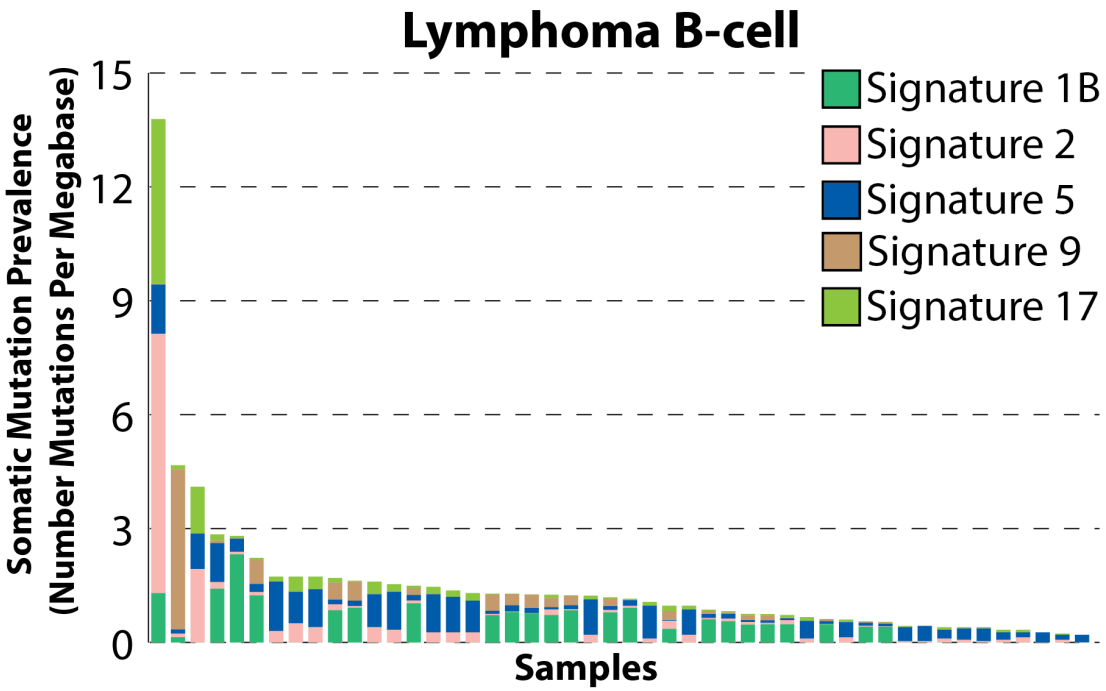

Liver Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
12
10
8
6
4
2
0
Samples
Signature 1B
Signature 4
Signature 6
Signature 12
Signature 16
Signature 17
Signature U2

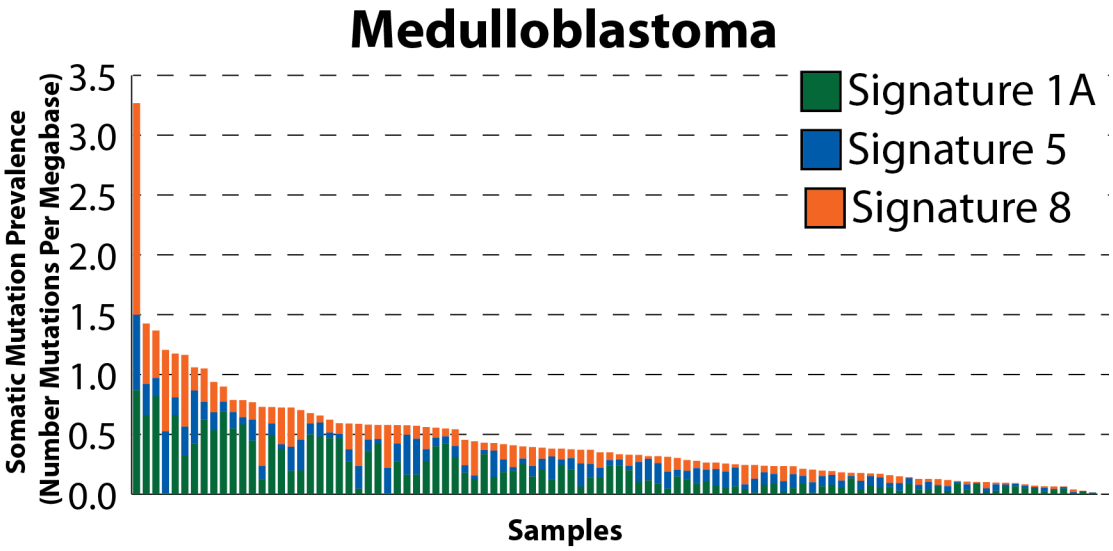
Lung Adenocarcinoma

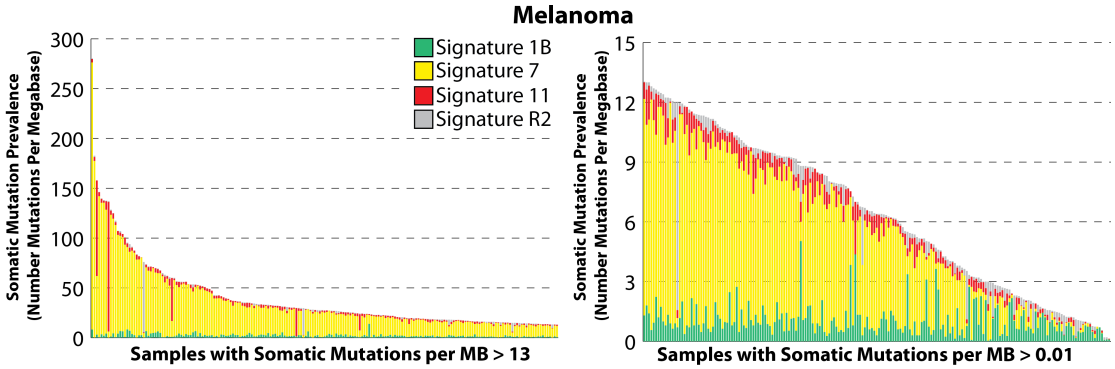


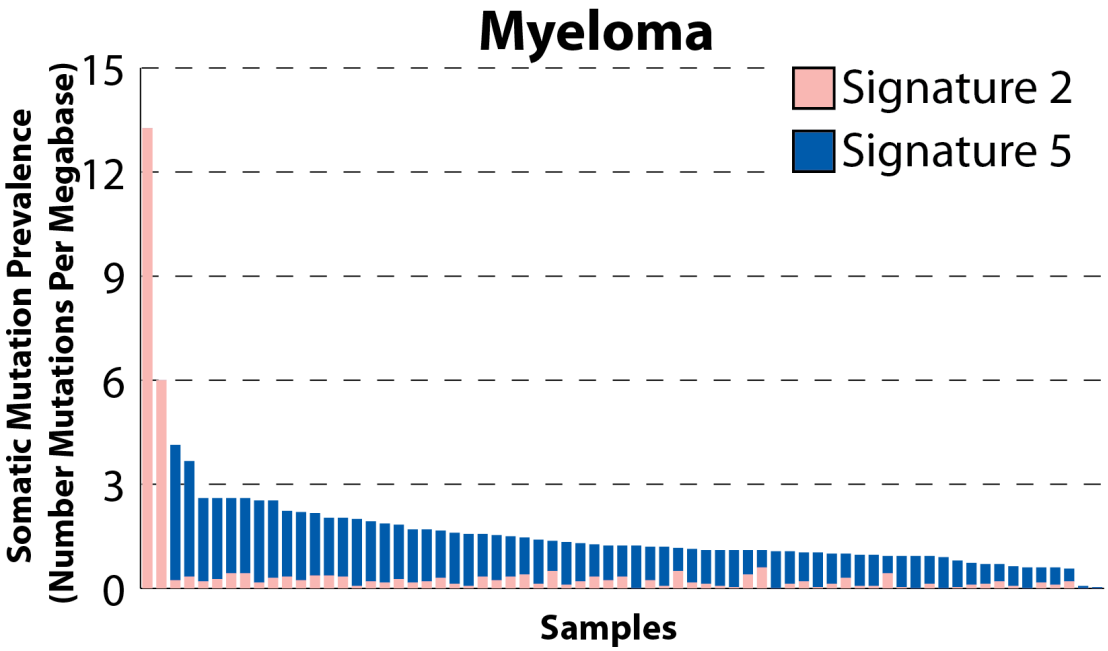

Lung Cancer Small Cell
Signature 4
Signature 15
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
30
25
20
15
10
5
0
Samples

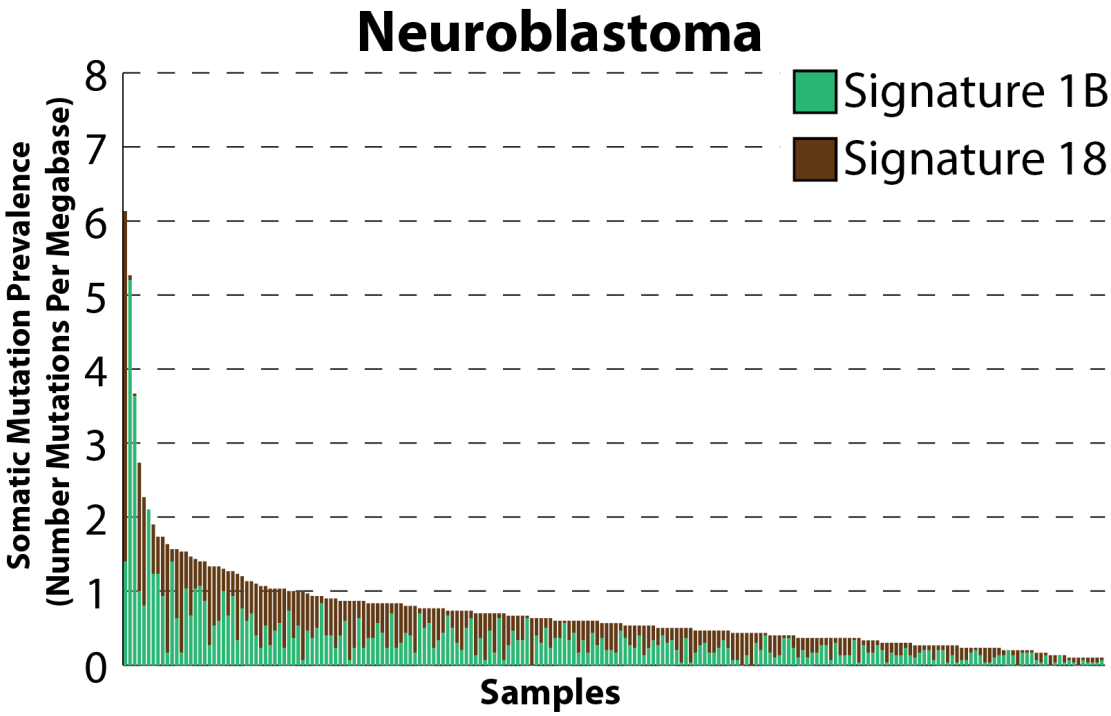

Lung Squamous*
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
50
40
30
20
10
0
Samples
Signature 2
Signature 4
Signature 5
Signature R2


Lymphoma B-cell
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
15
12
9
6
3
0
Samples
Signature 1B
Signature 2
Signature 5
Signature 9
Signature 17

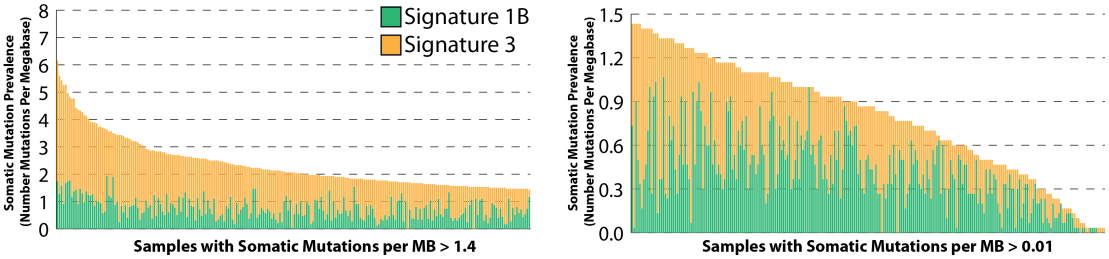

Medulloblastoma
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Signature 1A
Signature 5
Signature 8
Samples

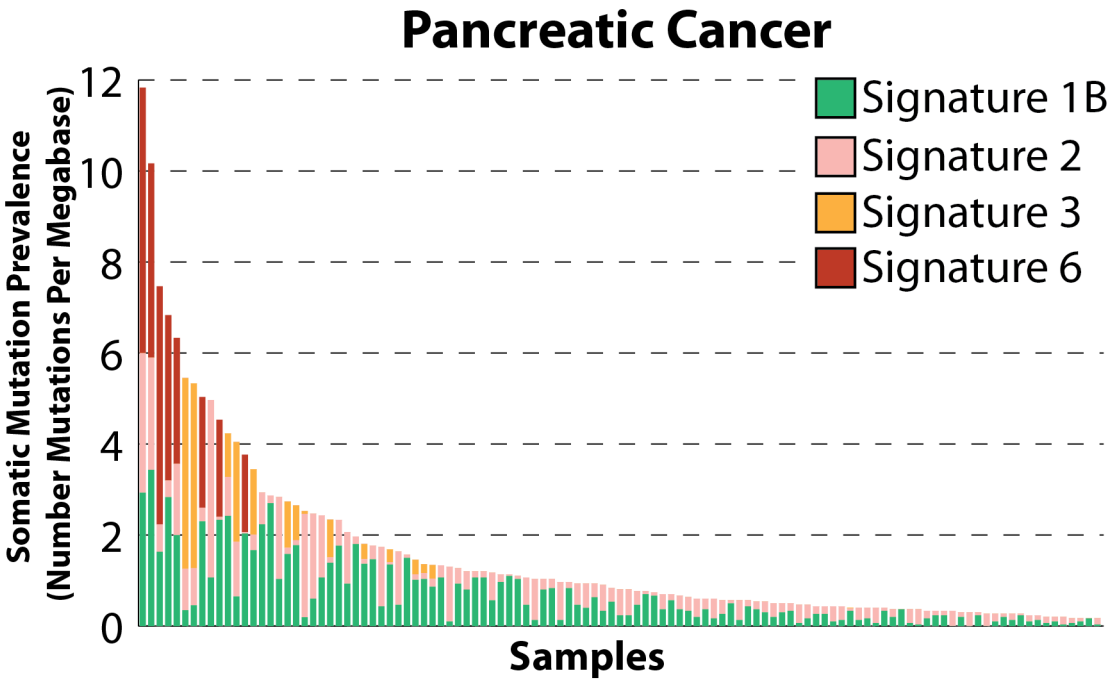

Melanoma
Signature 1B
Signature 7
Signature 11
Signature R2
Somatic Mutation Prevalence (Number Mutations Per Megabase)
300
250
200
150
100
50
0
Samples with Somatic Mutations per MB > 13
Somatic Mutation Prevalence (Number Mutations Per Megabase)
15
12
9
6
3
0
Samples with Somatic Mutations per MB > 0.01

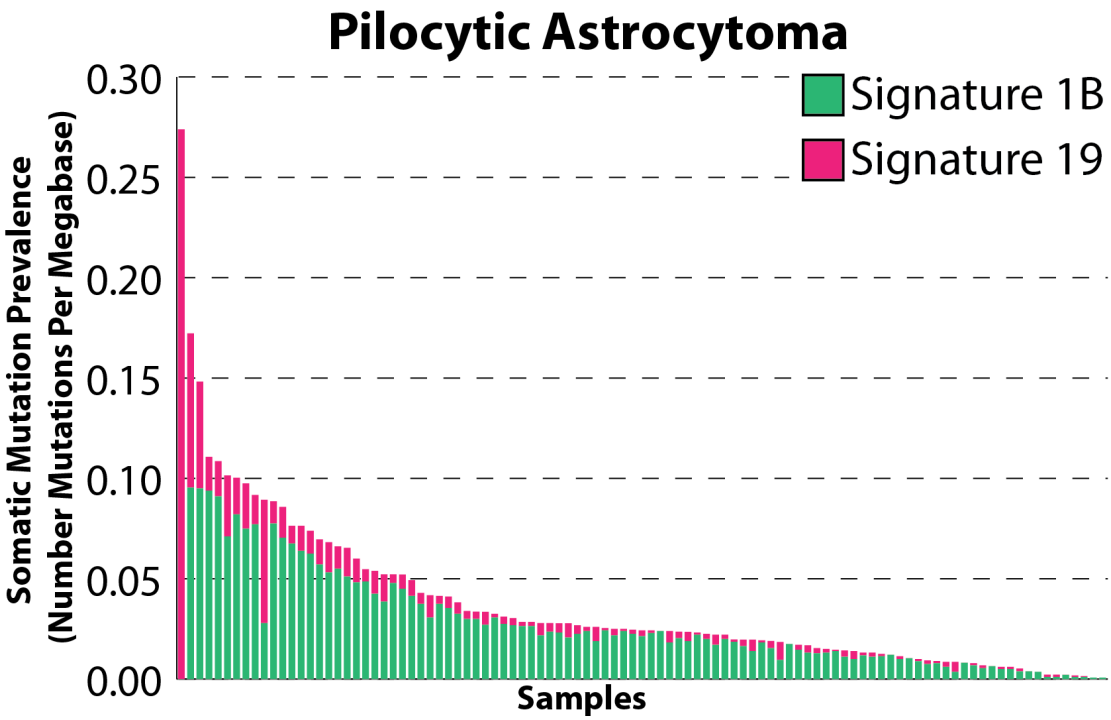

Myeloma
Signature 2
Signature 5
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
15
12
9
6
3
0
Samples

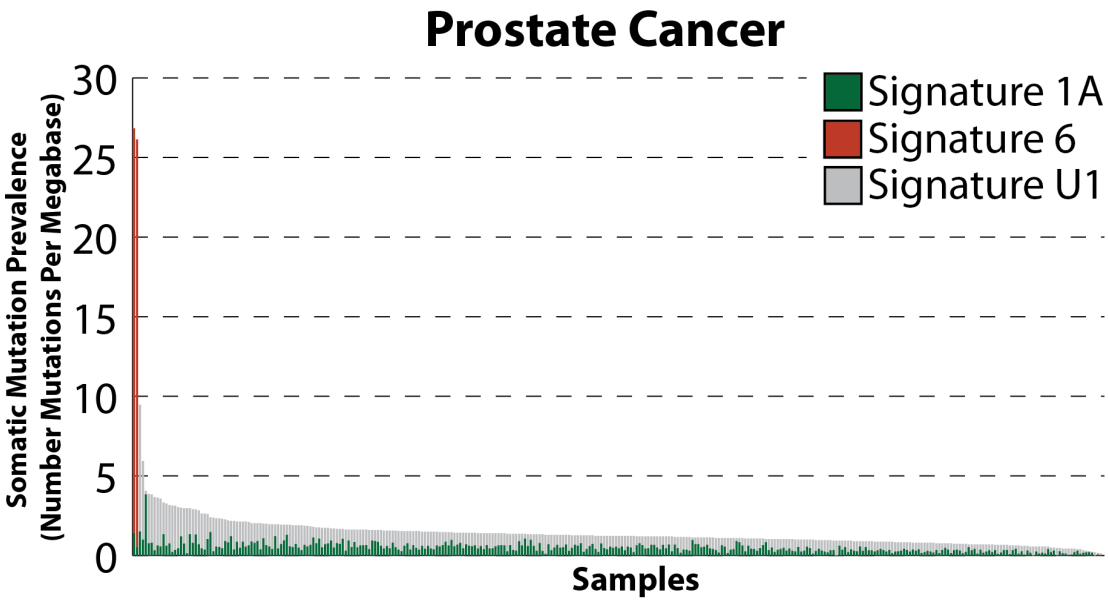

Neuroblastoma
Signature 1B
Signature 18
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
8
7
6
5
4
3
2
1
0
Samples

Ovarian Cancer

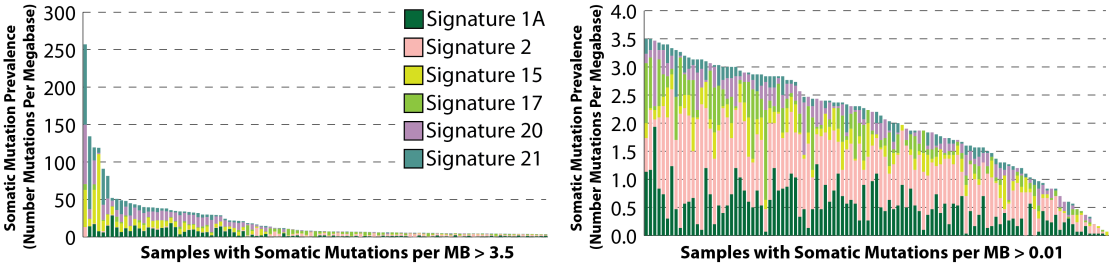


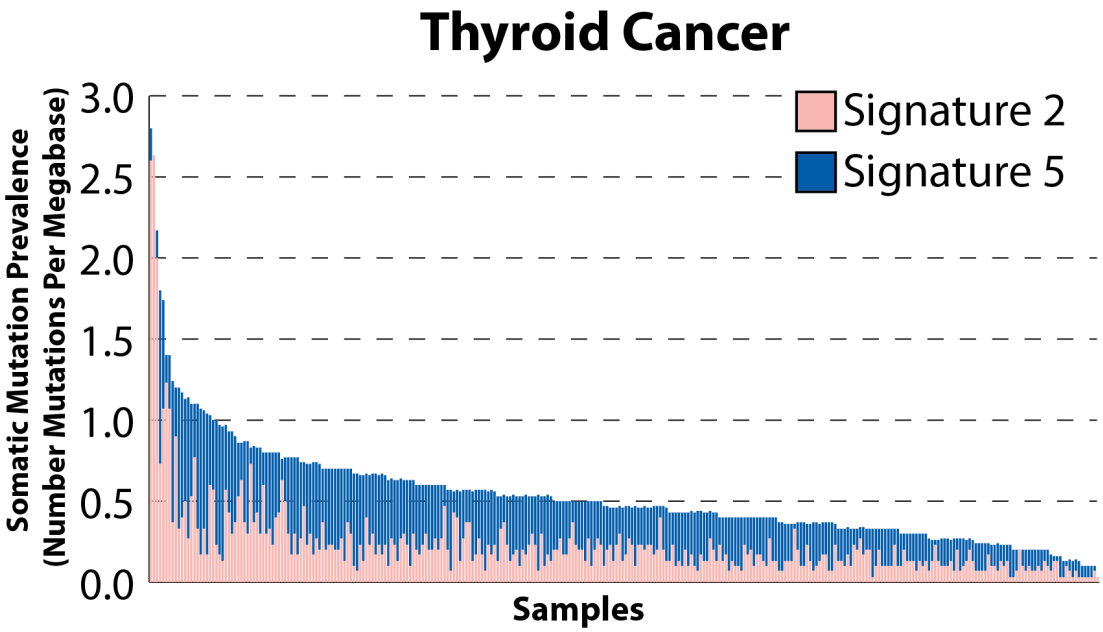

Pancreatic Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
12
10
8
6
4
2
0
Samples
Signature 1B
Signature 2
Signature 3
Signature 6

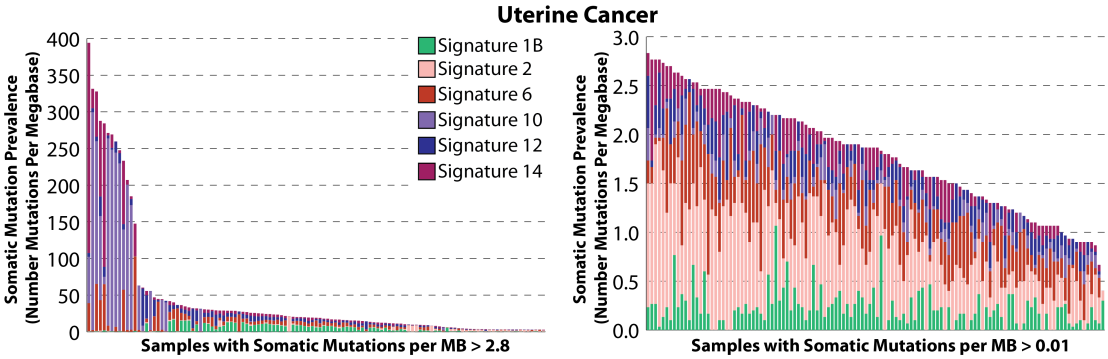

Pilocytic Astrocytoma
Signature 1B
Signature 19
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Samples


Prostate Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
30
25
20
15
10
5
0
Samples
Signature 1A
Signature 6
Signature U1

Stomach Cancer

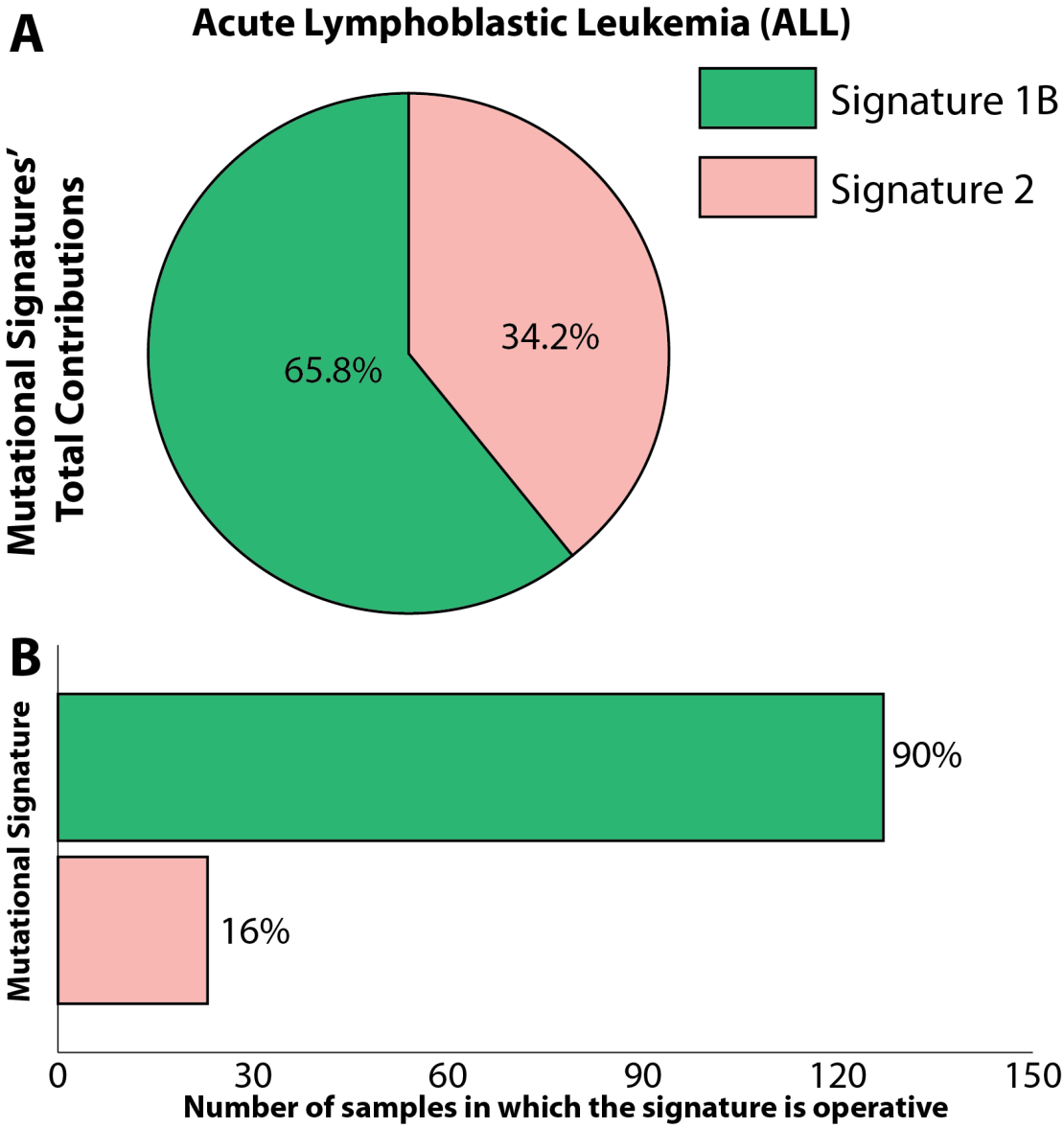


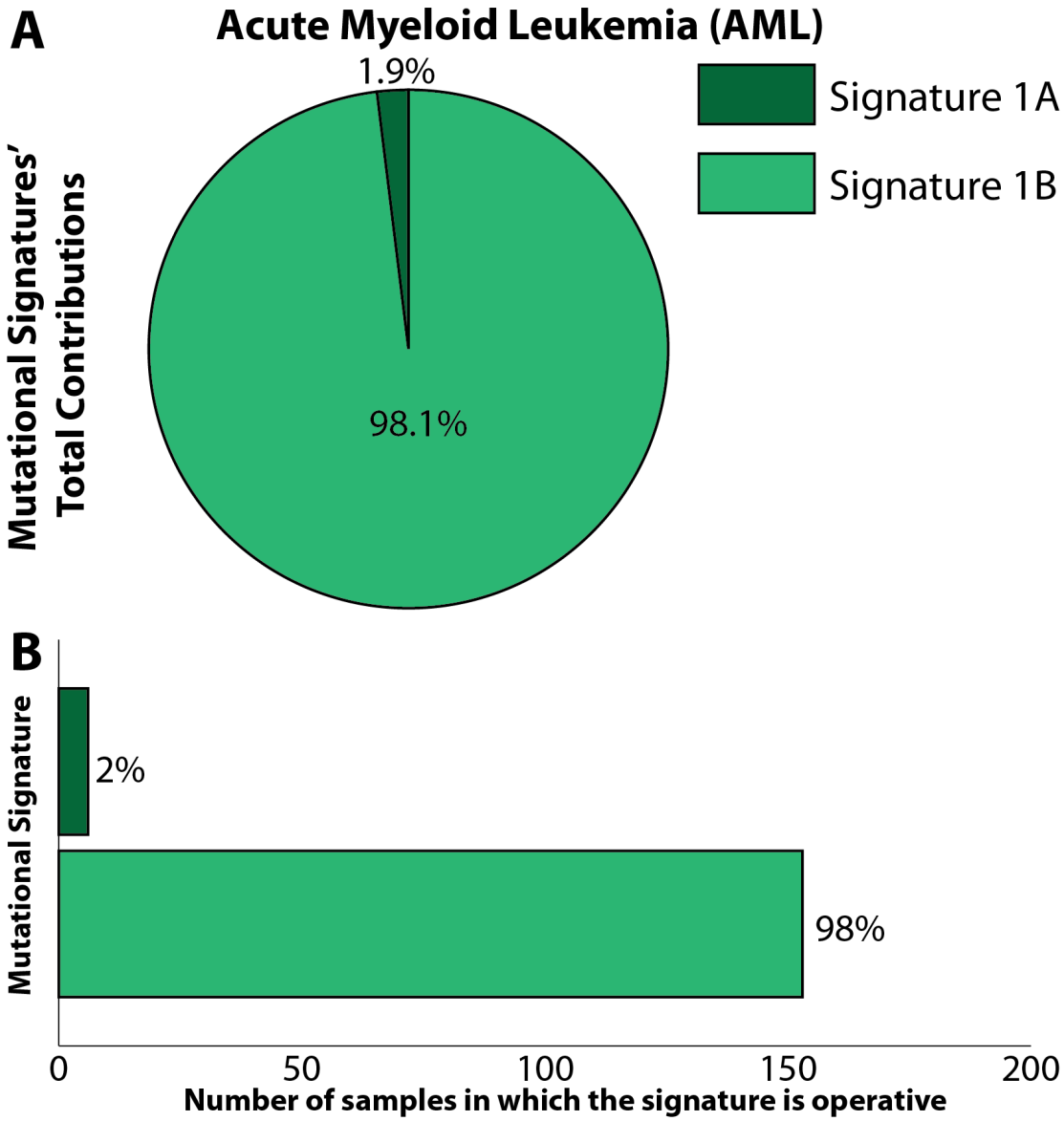

Thyroid Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Signature 2
Signature 5
Samples

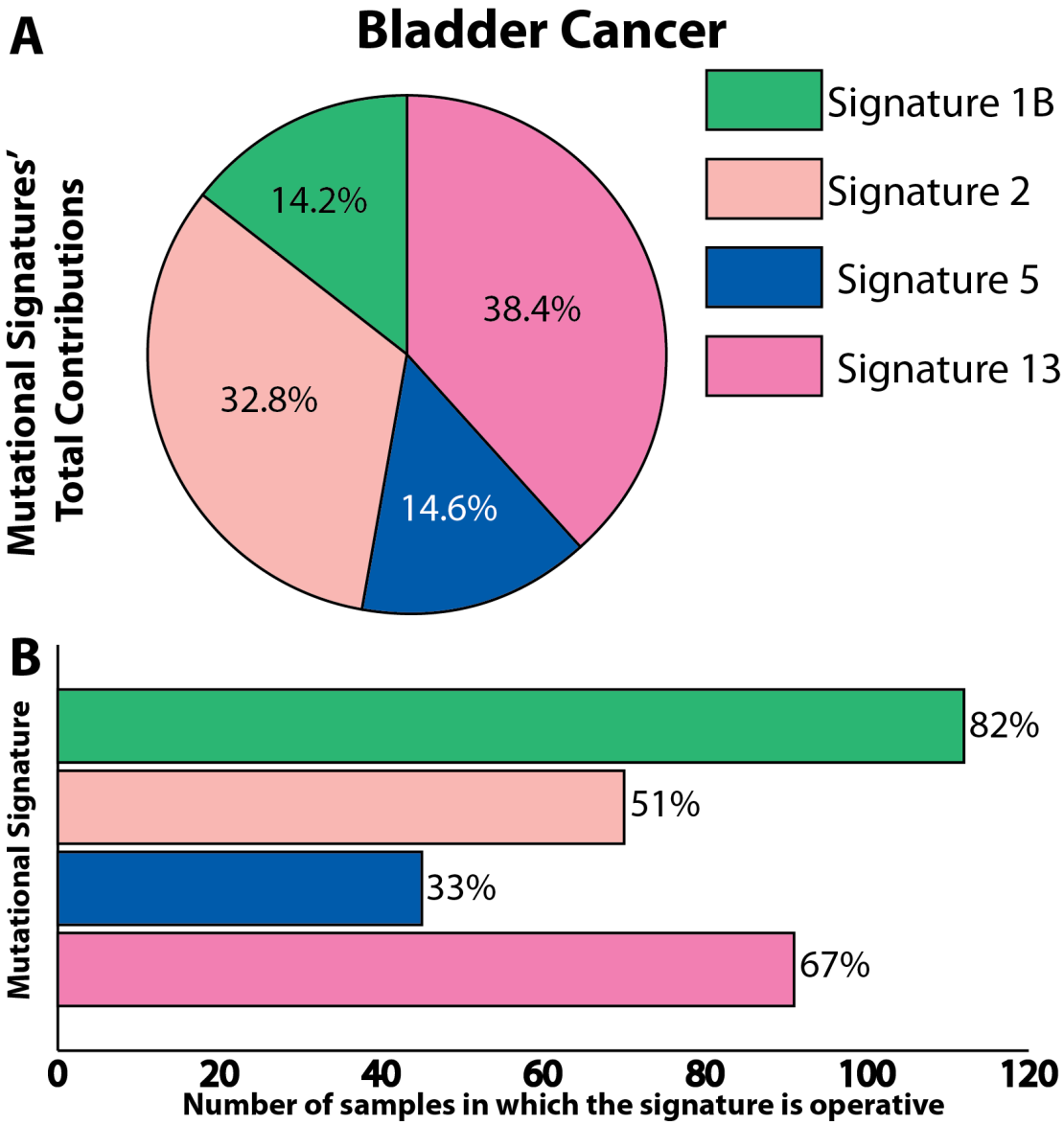

Uterine Cancer
Somatic Mutation Prevalence (Number Mutations Per Megabase)
400
350
300
250
200
150
100
50
0
Signature 1B
Signature 2
Signature 6
Signature 10
Signature 12
Signature 14
Samples with Somatic Mutations per MB > 2.8
Somatic Mutation Prevalence (Number Mutations Per Megabase)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Samples with Somatic Mutations per MB > 0.01

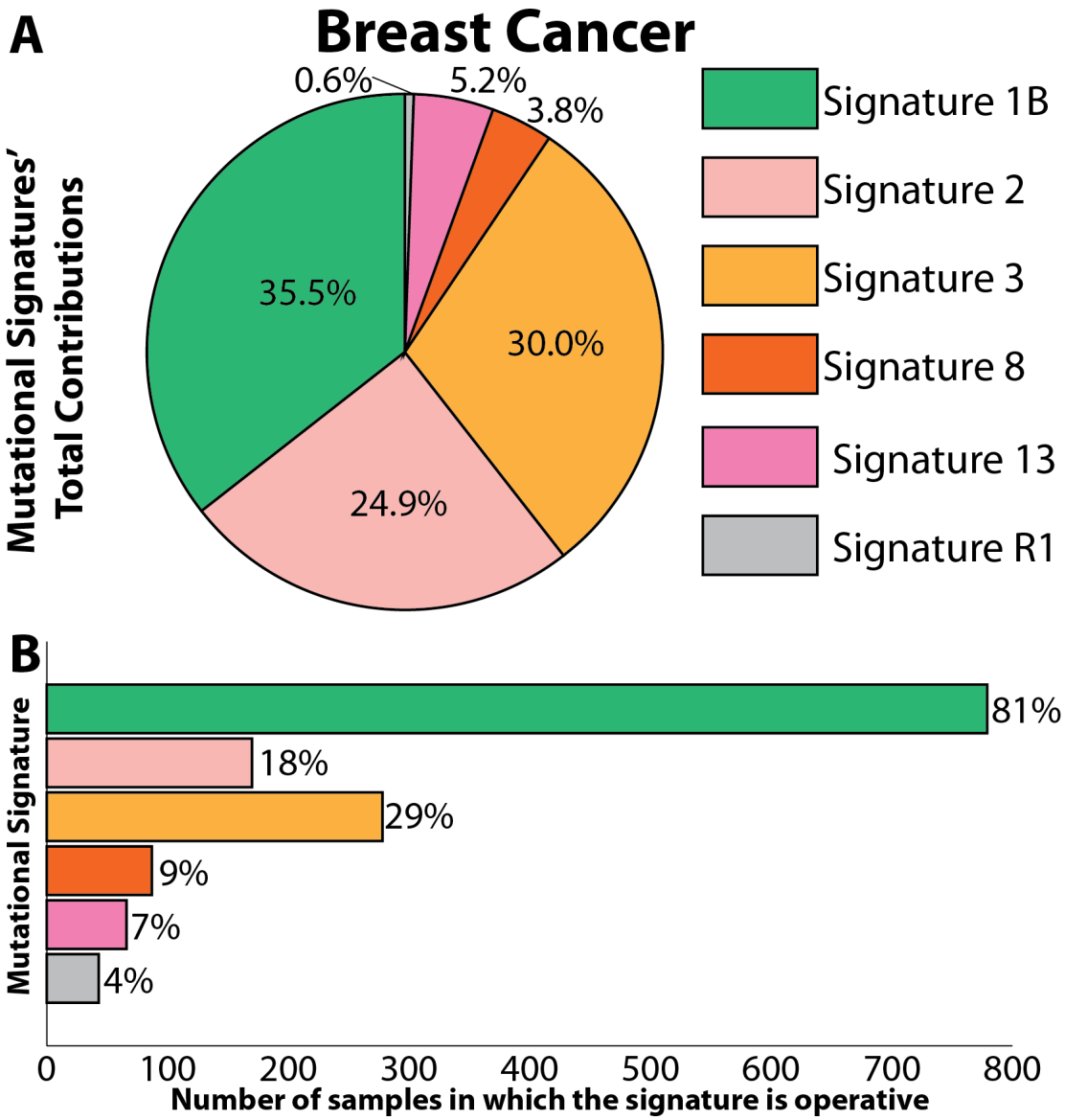
APPENDIX VI: Summary of signatures' contributions in cancer types

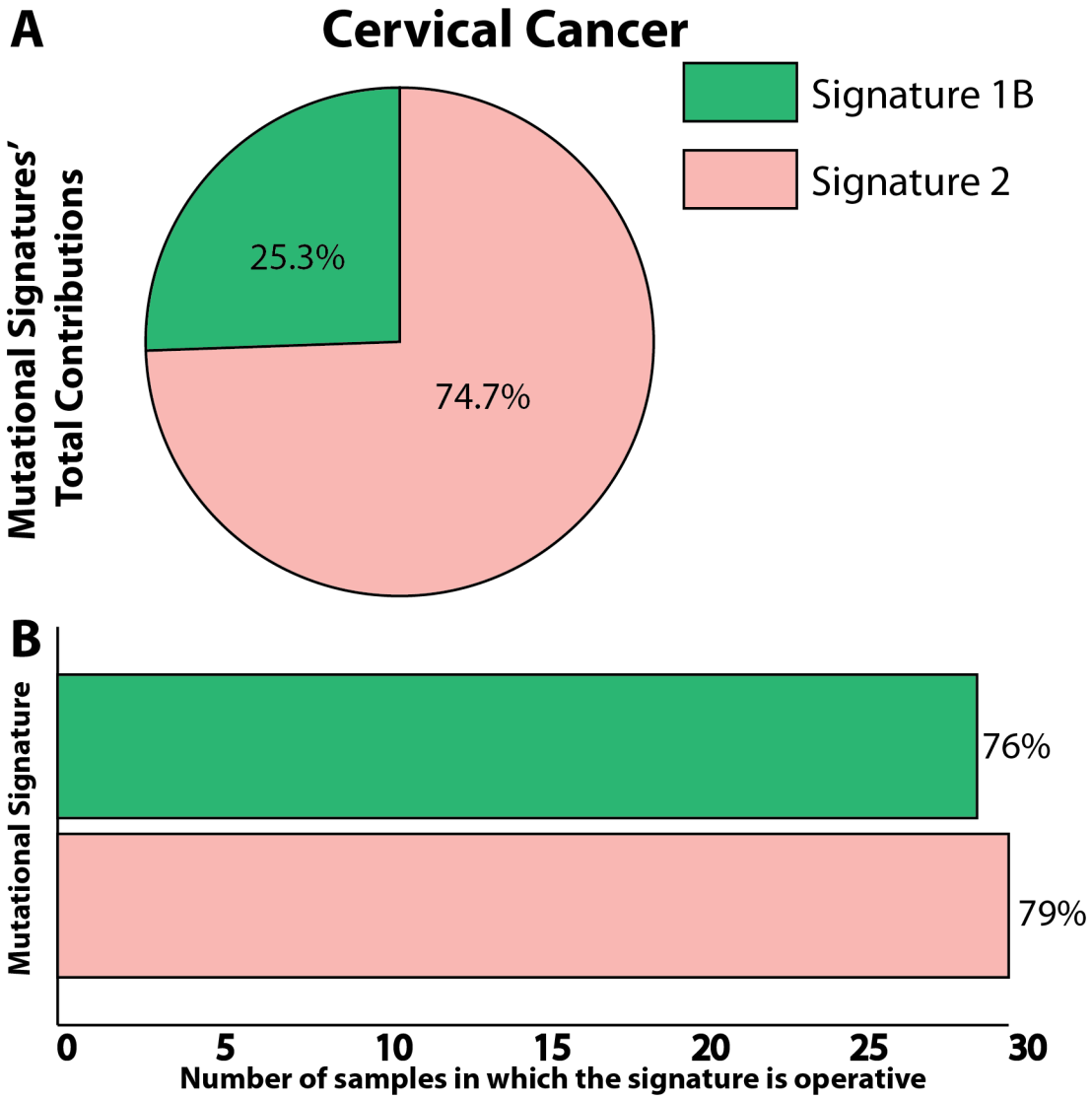
This appendix contains a high-resolution figure for each of the 30 examined cancer types (chapter 4). Each figure depicts a single cancer type and provides a summary of the contributions of the mutational signatures found in this cancer type. All figures have two panels: panel ***A*** depicting the percentage of total mutations contributed by each of the operative mutational signatures in that cancer type and panel ***B*** depicting the percentage and number of samples in which each mutational signature contributes significant number of somatic mutations. For most signatures, significant number of mutations in a sample is defined as more than 100 substitutions or more than 25% of all mutations in that sample. Mutational signatures are displayed in distinct colours, consistent in both panels of each figure as well as in all figures in Appendices V and VI. Figures are displayed on individual pages, labelled to clearly show the names of the cancer types, and they are ordered alphabetically based on the names of these cancer types. In general, all samples are included in the summary of each cancer type. The only exception (denoted with an asterisk in the appropriate figure) is one lung squamous hypermutator sample purely of Signature 7 (72 mutations per MB). Signature 7 is associated with exposure to ultraviolet light, an unlikely carcinogen for lung cancer. As such, this TCGA sample is most likely either a melanoma metastasis or a misannotated sample. Thus, the association between Signature 7 and lung squamous has not been discussed in chapter 4 and this association has not been displayed in Figure 4.9.

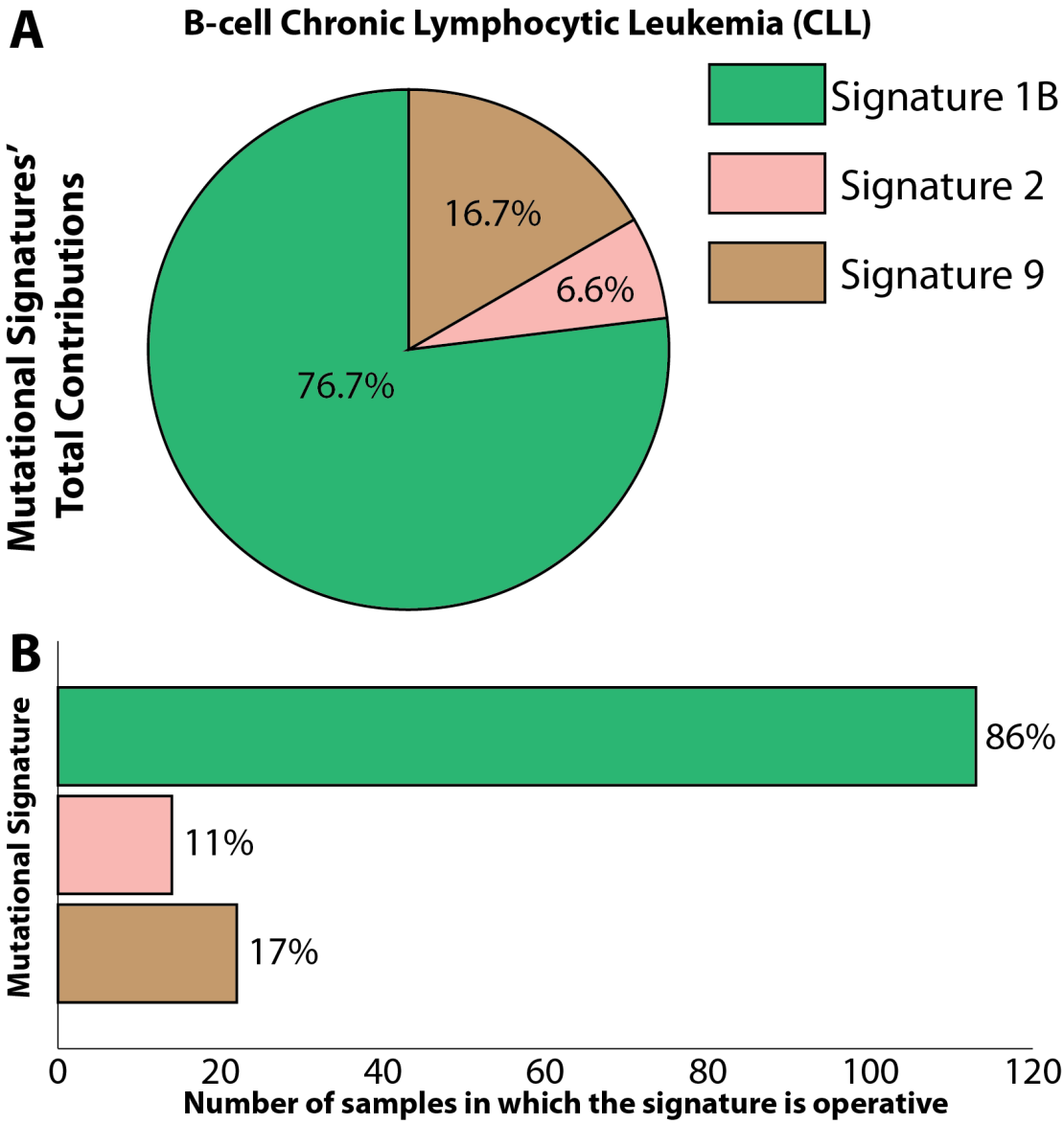

A
Acute Lymphoblastic Leukemia (ALL)
Mutational Signatures' Total Contributions
Signature 1B
Signature 2
65.8%
34.2%
B
Mutational Signature
90%
16%
0
30
60
90
120
150
Number of samples in which the signature is operative

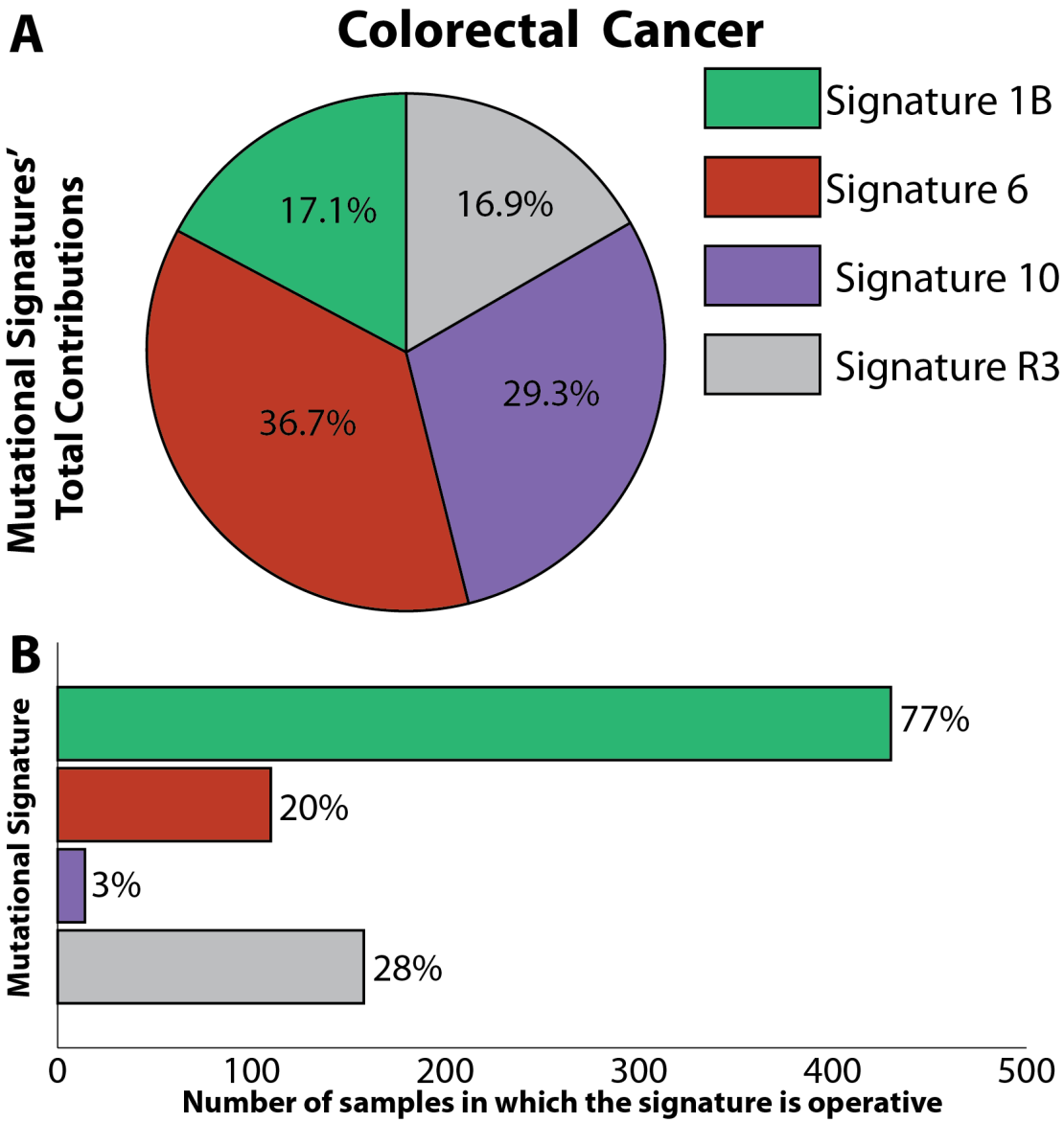

A
Acute Myeloid Leukemia (AML)
1.9%
98.1%
Signature 1A
Signature 1B
Mutational Signatures' Total Contributions
B
Mutational Signature
2%
98%
0
50
100
150
200
Number of samples in which the signature is operative

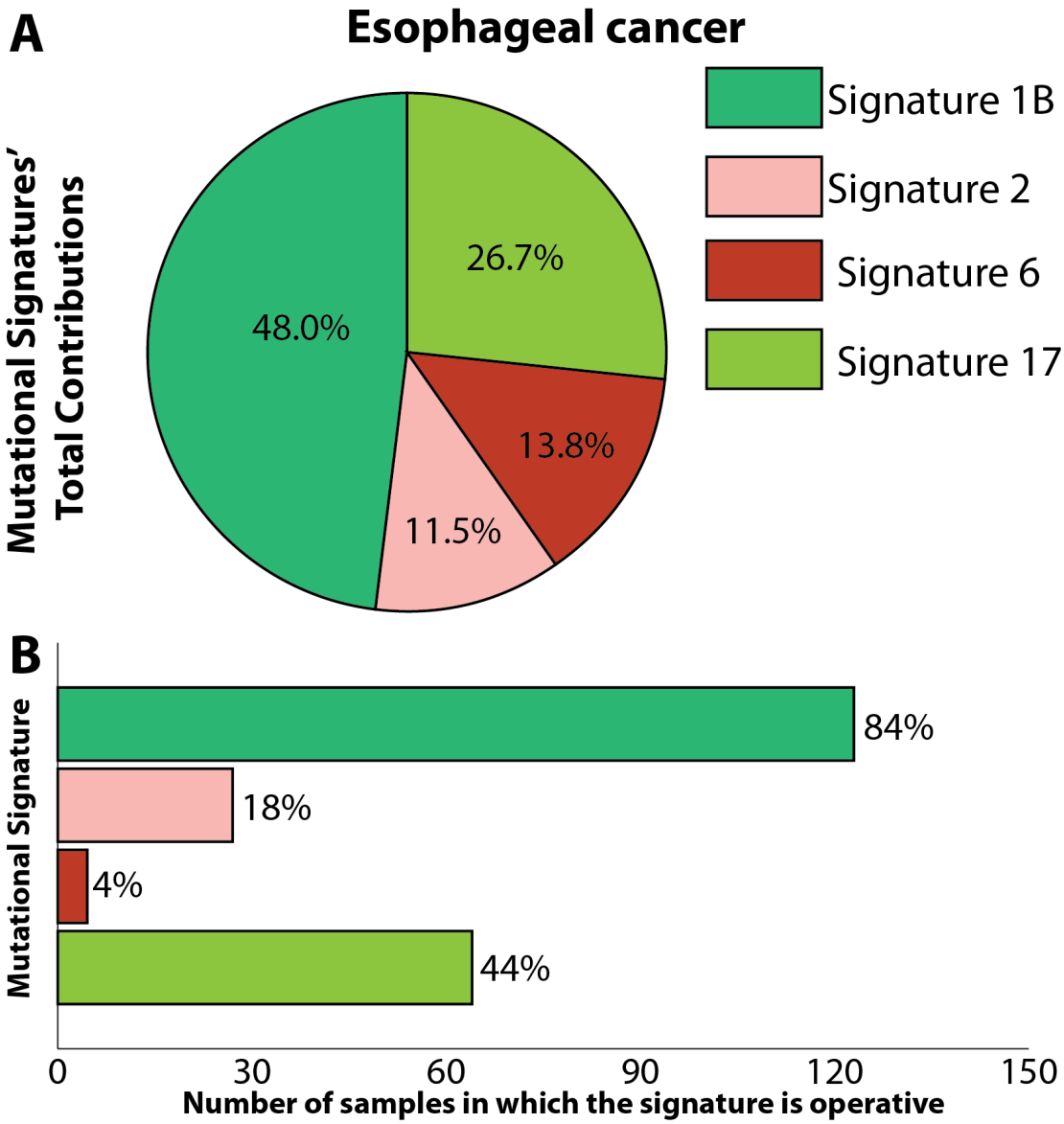

A
Bladder Cancer
Mutational Signatures' Total Contributions
14.2%
38.4%
32.8%
14.6%
Signature 1B
Signature 2
Signature 5
Signature 13
B
Mutational Signature
82%
51%
33%
67%
0
20
40
60
80
100
120
Number of samples in which the signature is operative

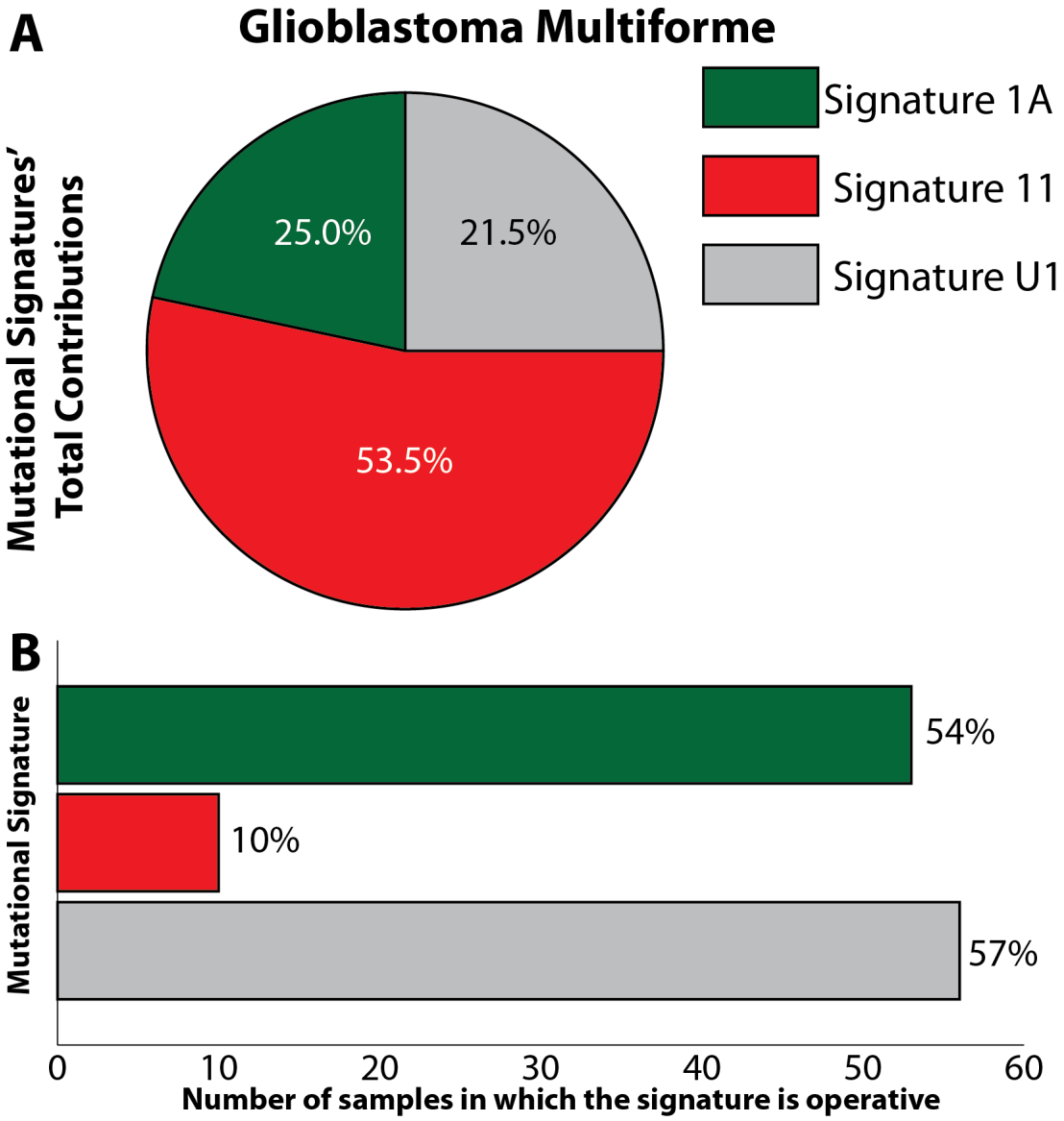

A
Breast Cancer
Mutational Signatures' Total Contributions
0.6%
5.2%
3.8%
35.5%
30.0%
24.9%
Signature 1B
Signature 2
Signature 3
Signature 8
Signature 13
Signature R1
B
Mutational Signature
81%
18%
29%
9%
7%
4%
0
100
200
300
400
500
600
700
800
Number of samples in which the signature is operative

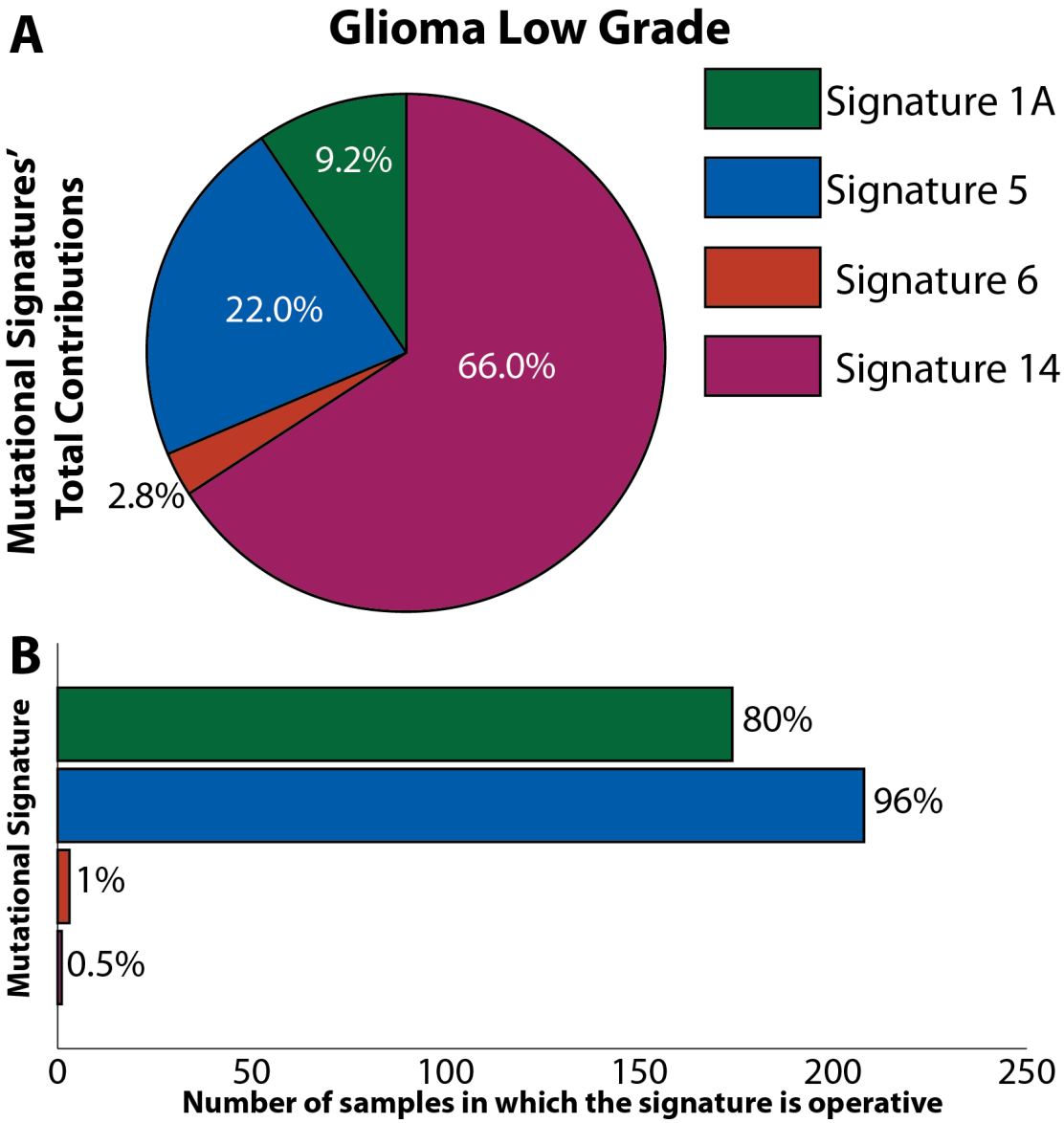

A
Cervical Cancer
Signature 1B
Signature 2
Mutational Signatures' Total Contributions
25.3%
74.7%
B
Mutational Signature
76%
79%
0
5
10
15
20
25
30
Number of samples in which the signature is operative

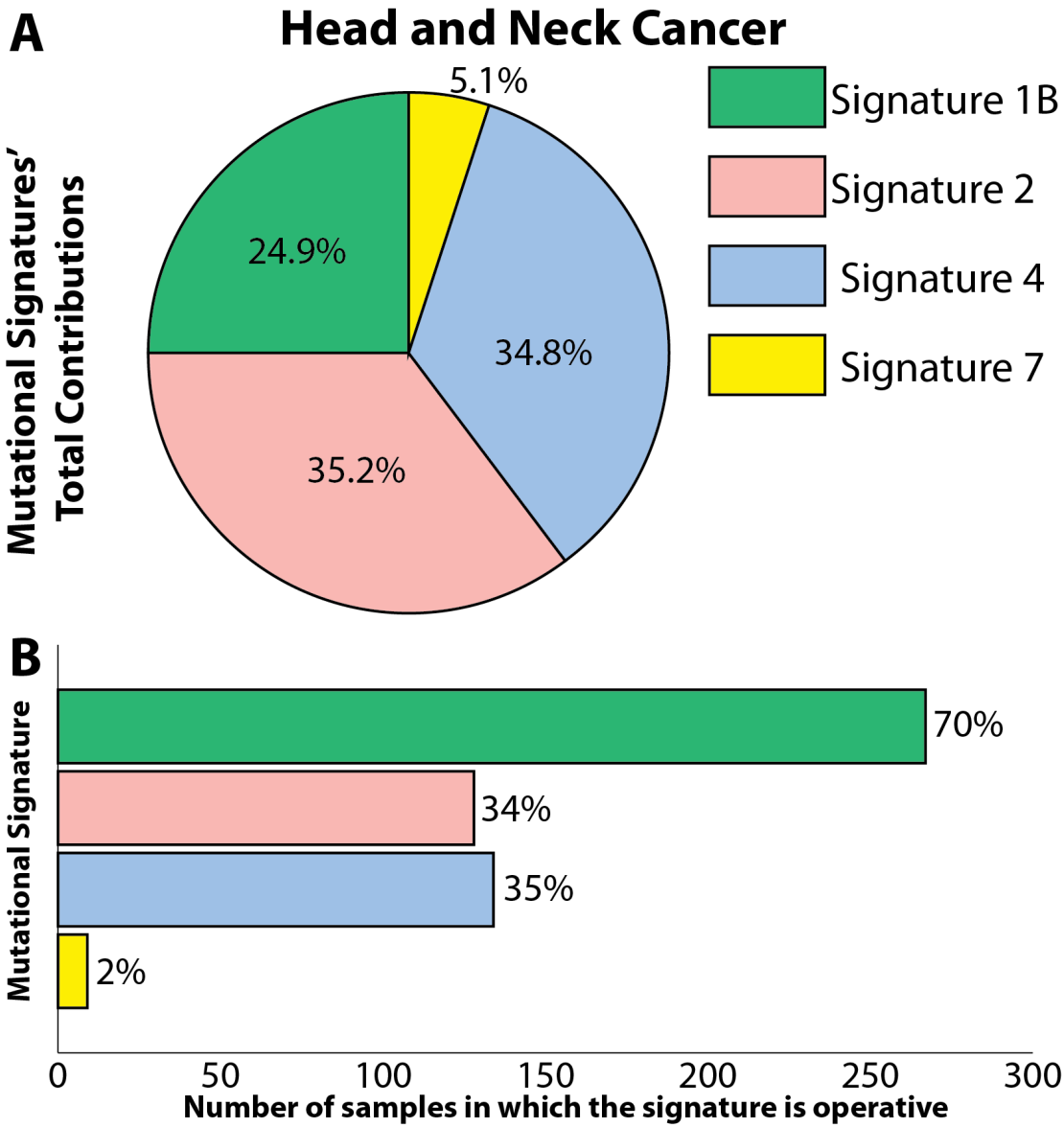

A
B-cell Chronic Lymphocytic Leukemia (CLL)
Mutational Signatures' Total Contributions
16.7%
6.6%
76.7%
Signature 1B
Signature 2
Signature 9
B
Mutational Signature
86%
11%
17%
0
20
40
60
80
100
120
Number of samples in which the signature is operative

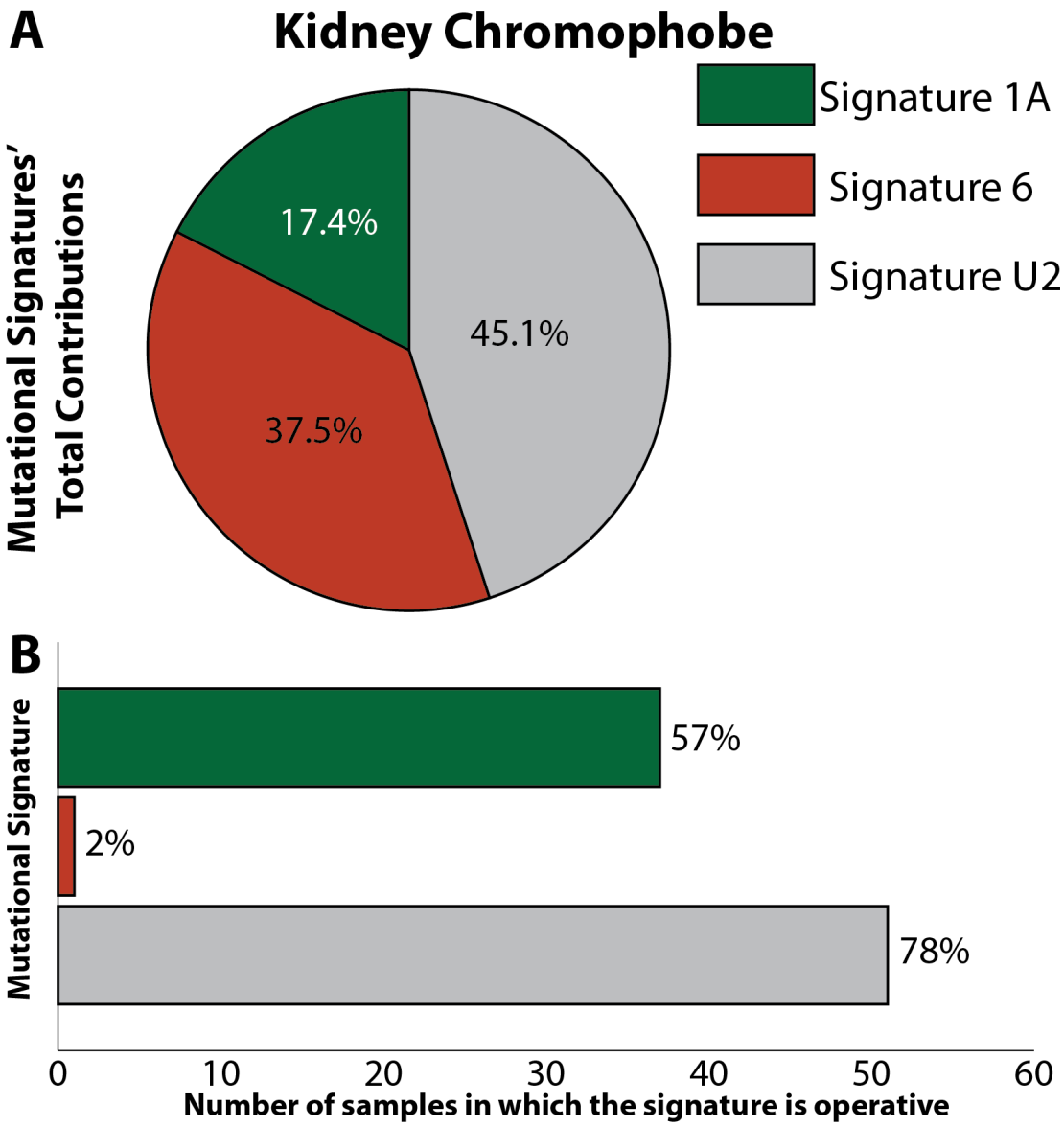

A
Colorectal Cancer
Mutational Signatures' Total Contributions
17.1%
16.9%
36.7%
29.3%
Signature 1B
Signature 6
Signature 10
Signature R3
B
Mutational Signature
77%
20%
3%
28%
0
100
200
300
400
500
Number of samples in which the signature is operative

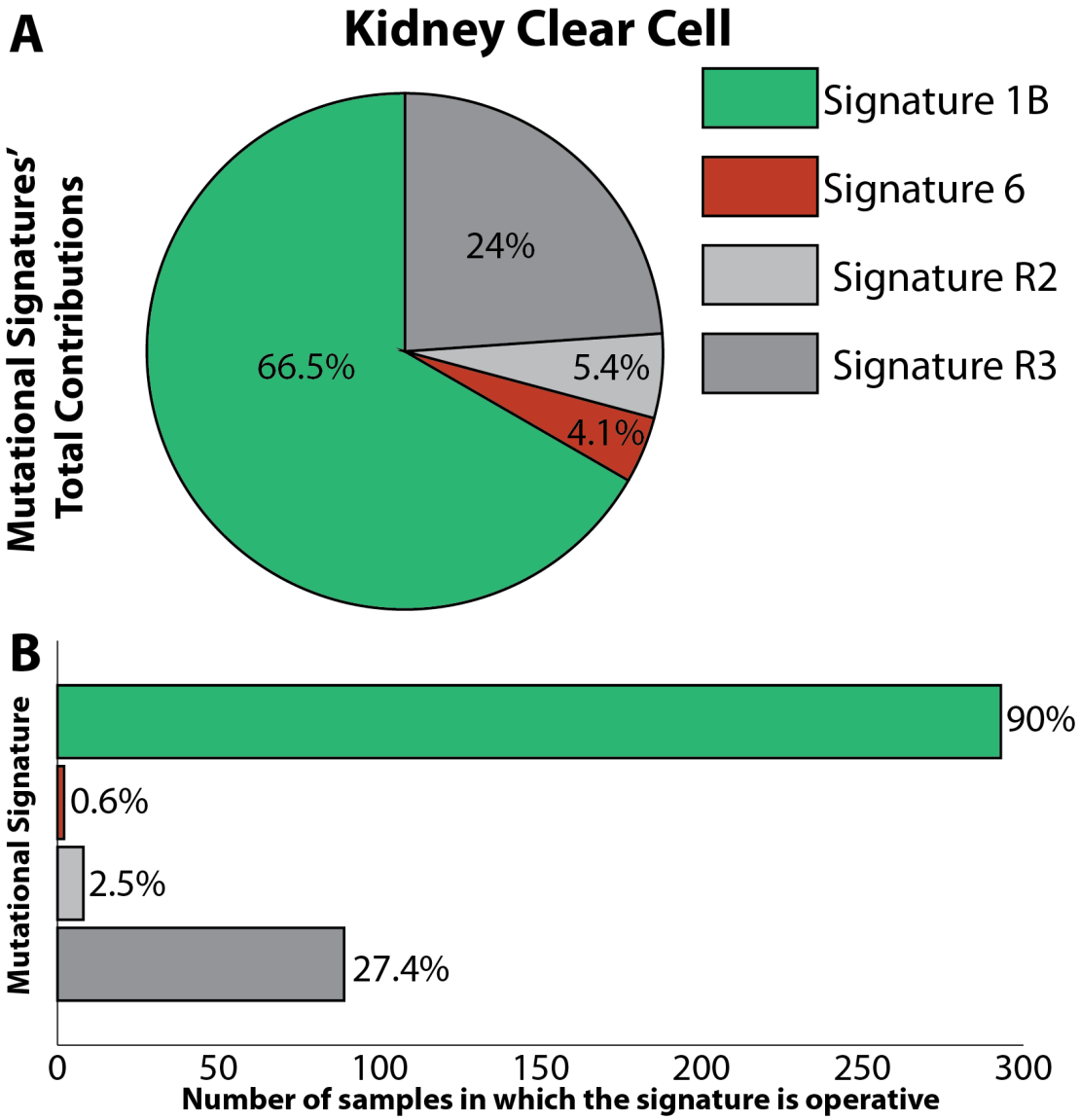

A
Esophageal cancer
Mutational Signatures' Total Contributions
48.0%
26.7%
13.8%
11.5%
Signature 1B
Signature 2
Signature 6
Signature 17
B
Mutational Signature
84%
18%
4%
44%
0
30
60
90
120
150
Number of samples in which the signature is operative

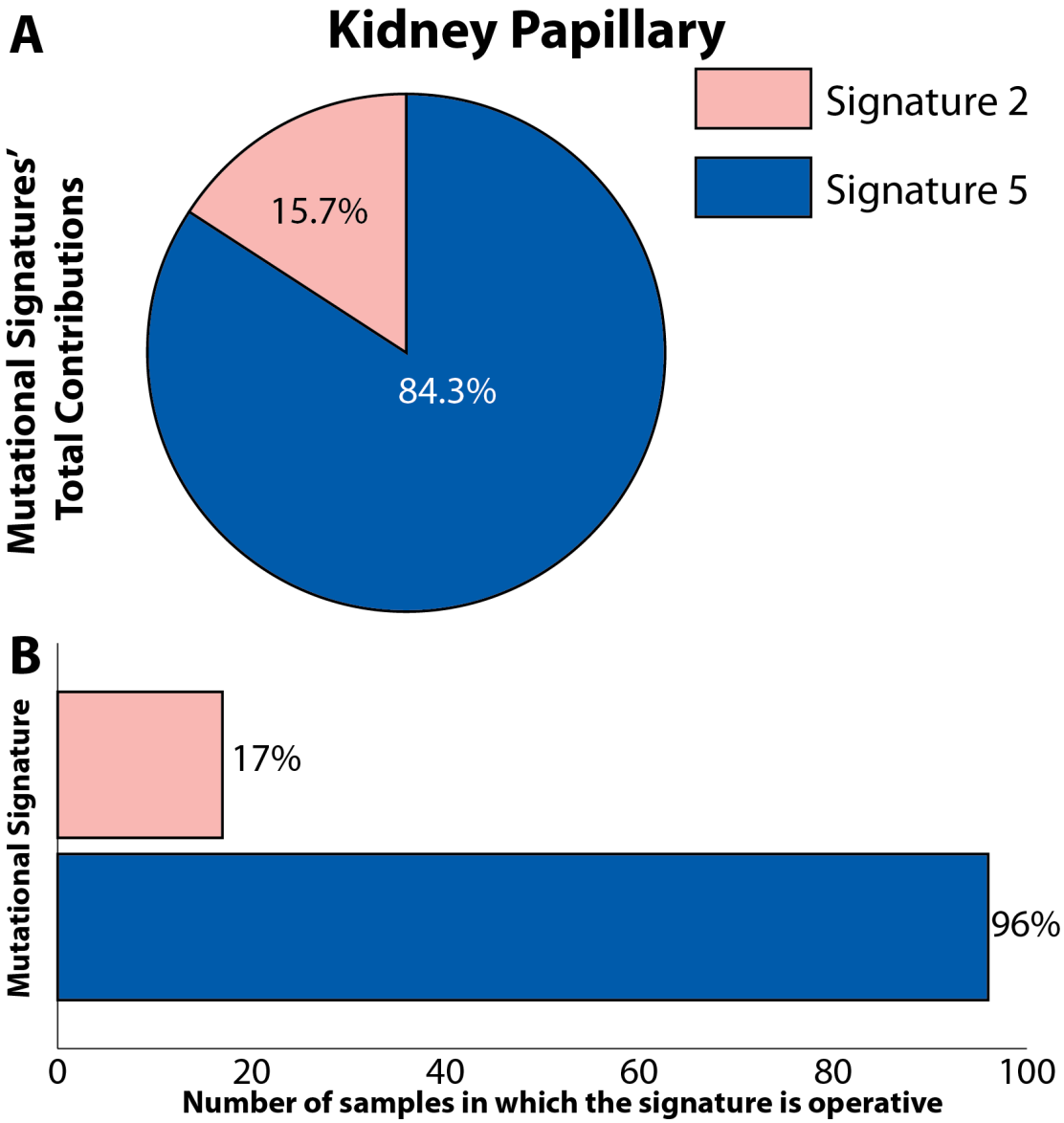

A
Glioblastoma Multiforme
Mutational Signatures' Total Contributions
25.0%
21.5%
53.5%
Signature 1A
Signature 11
Signature U1
B
Mutational Signature
54%
10%
57%
0
10
20
30
40
50
60
Number of samples in which the signature is operative

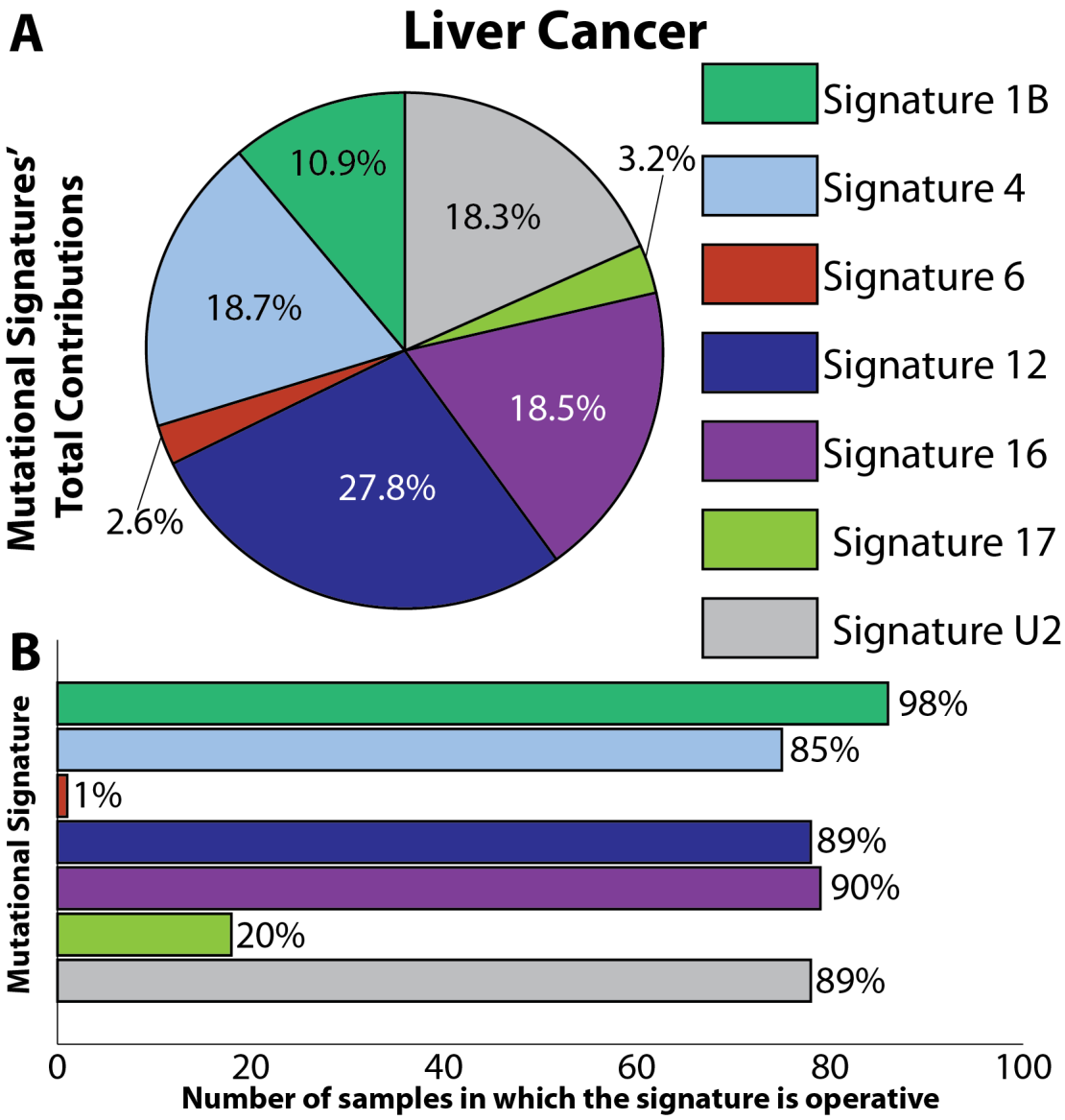

A
Glioma Low Grade
Mutational Signatures' Total Contributions
9.2%
22.0%
66.0%
2.8%
Signature 1A
Signature 5
Signature 6
Signature 14
B
Mutational Signature
80%
96%
1%
0.5%
0
50
100
150
200
250
Number of samples in which the signature is operative

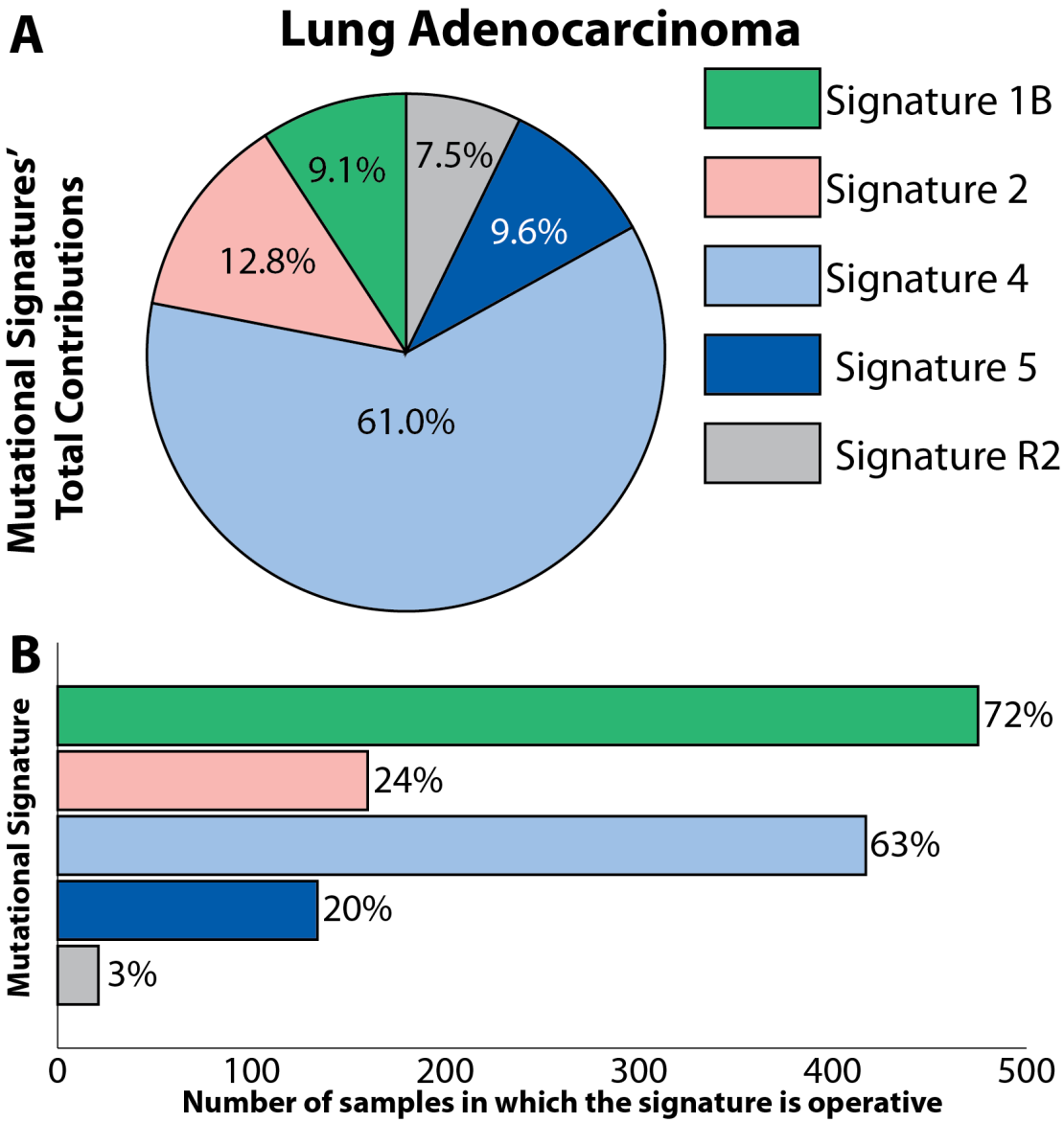

A
Head and Neck Cancer
Mutational Signatures' Total Contributions
5.1%
24.9%
34.8%
35.2%
Signature 1B
Signature 2
Signature 4
Signature 7
B
Mutational Signature
70%
34%
35%
2%
0
50
100
150
200
250
300
Number of samples in which the signature is operative

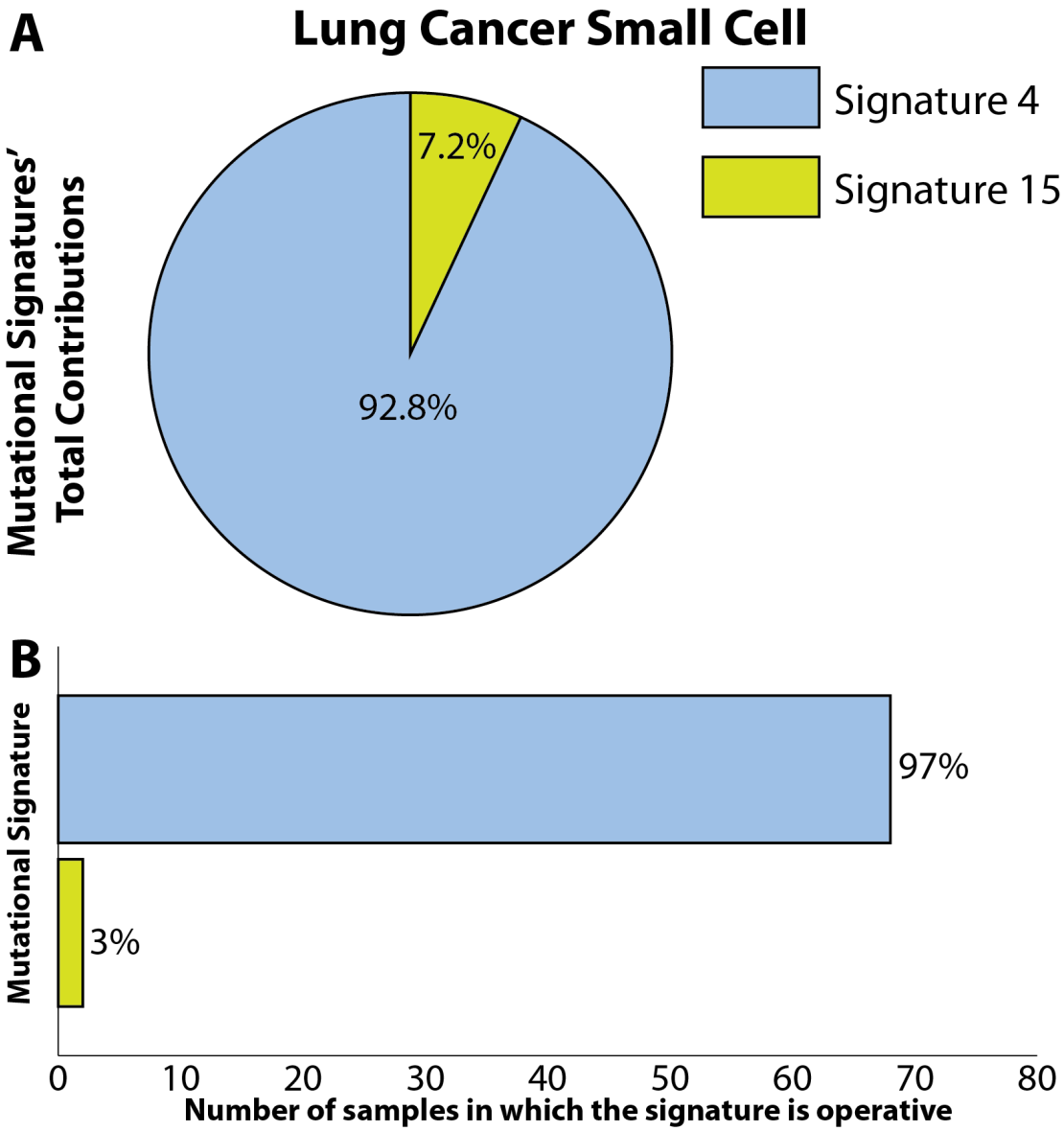

A
Kidney Chromophobe
Mutational Signatures' Total Contributions
17.4%
45.1%
37.5%
Signature 1A
Signature 6
Signature U2
B
Mutational Signature
57%
2%
78%
0
10
20
30
40
50
60
Number of samples in which the signature is operative

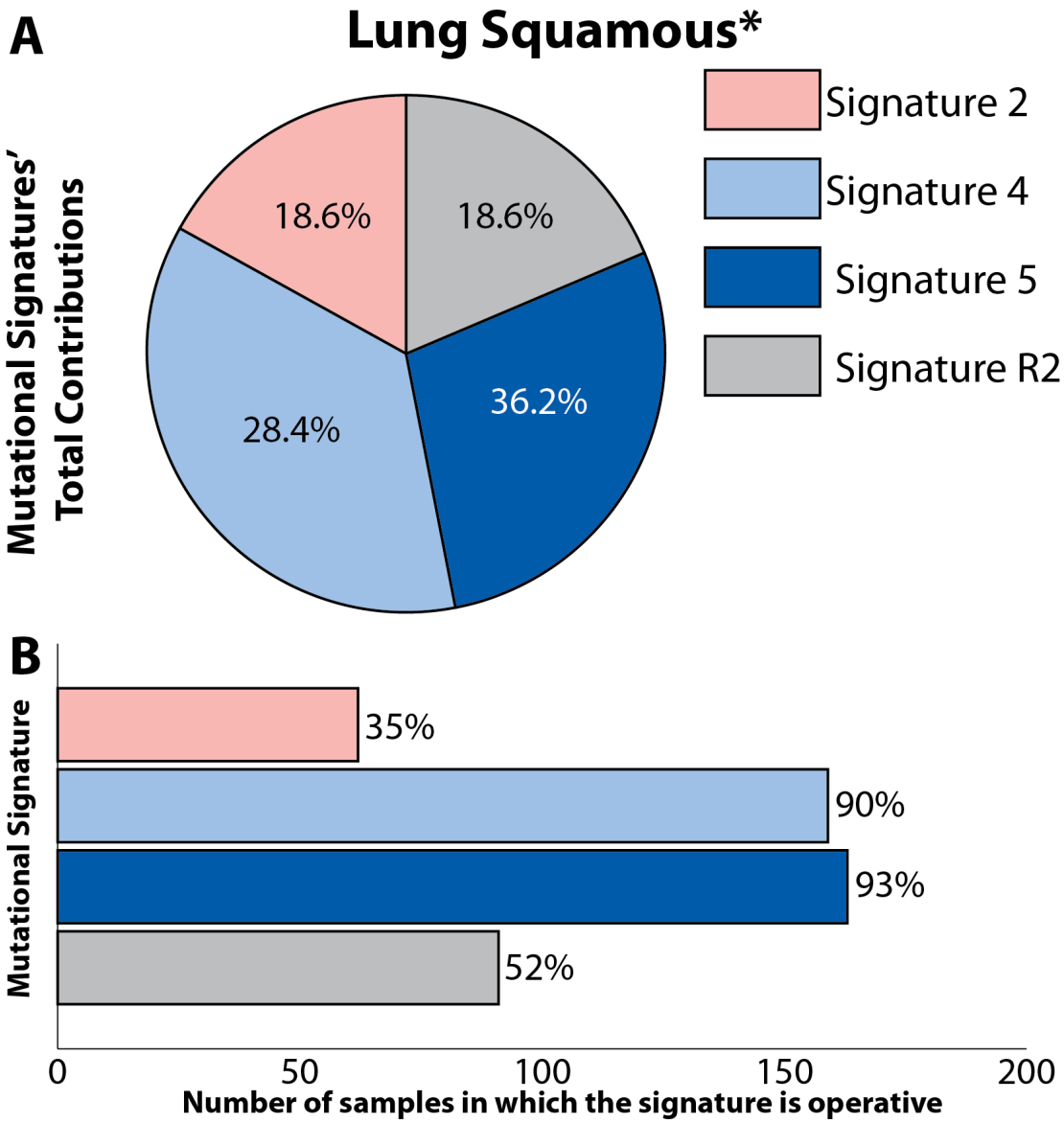

A
Kidney Clear Cell
Mutational Signatures' Total Contributions
Signature 1B
Signature 6
Signature R2
Signature R3
24%
66.5%
5.4%
4.1%
B
Mutational Signature
90%
0.6%
2.5%
27.4%
0
50
100
150
200
250
300
Number of samples in which the signature is operative

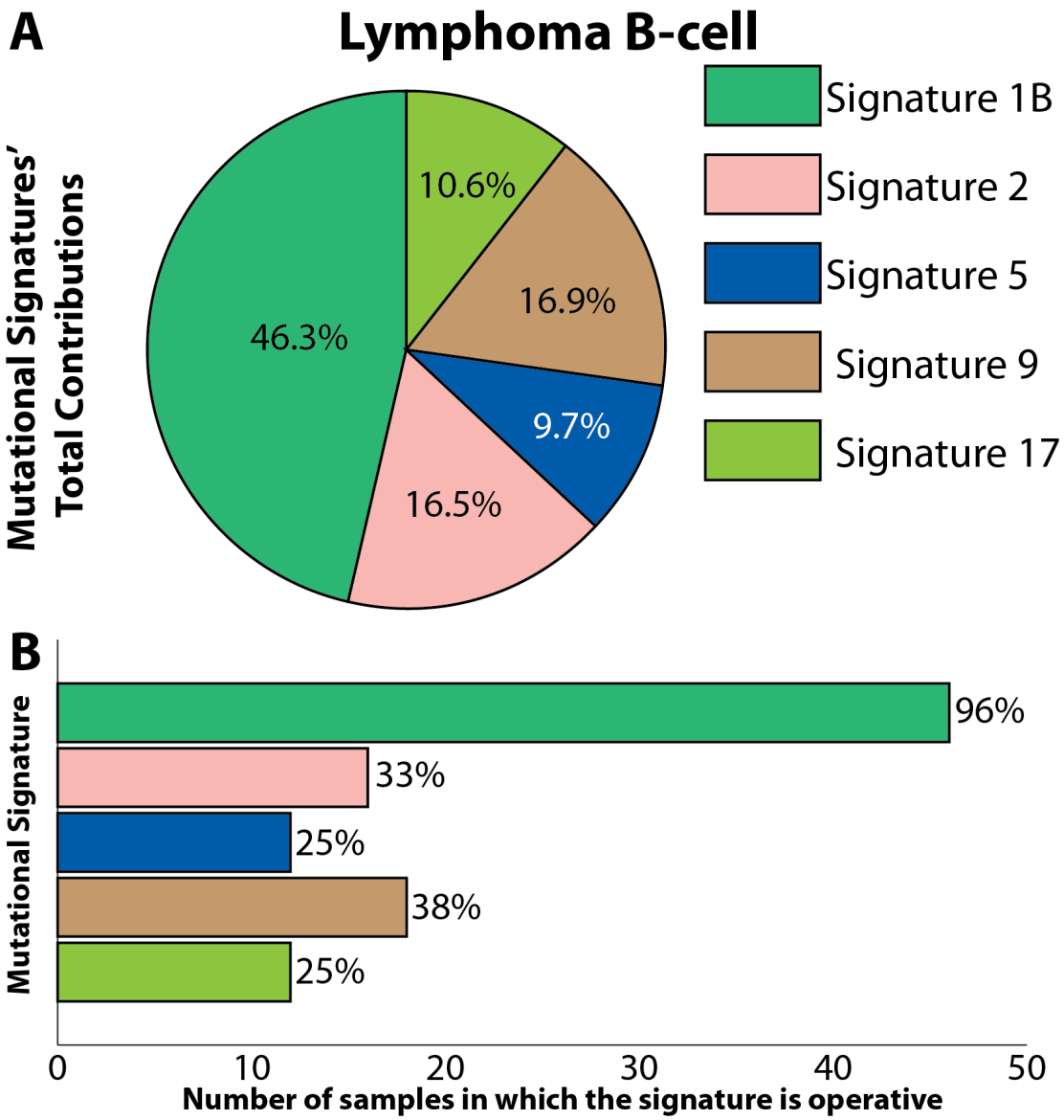

A
Kidney Papillary
Mutational Signatures' Total Contributions
Signature 2
Signature 5
15.7%
84.3%
B
Mutational Signature
17%
96%
0
20
40
60
80
100
Number of samples in which the signature is operative

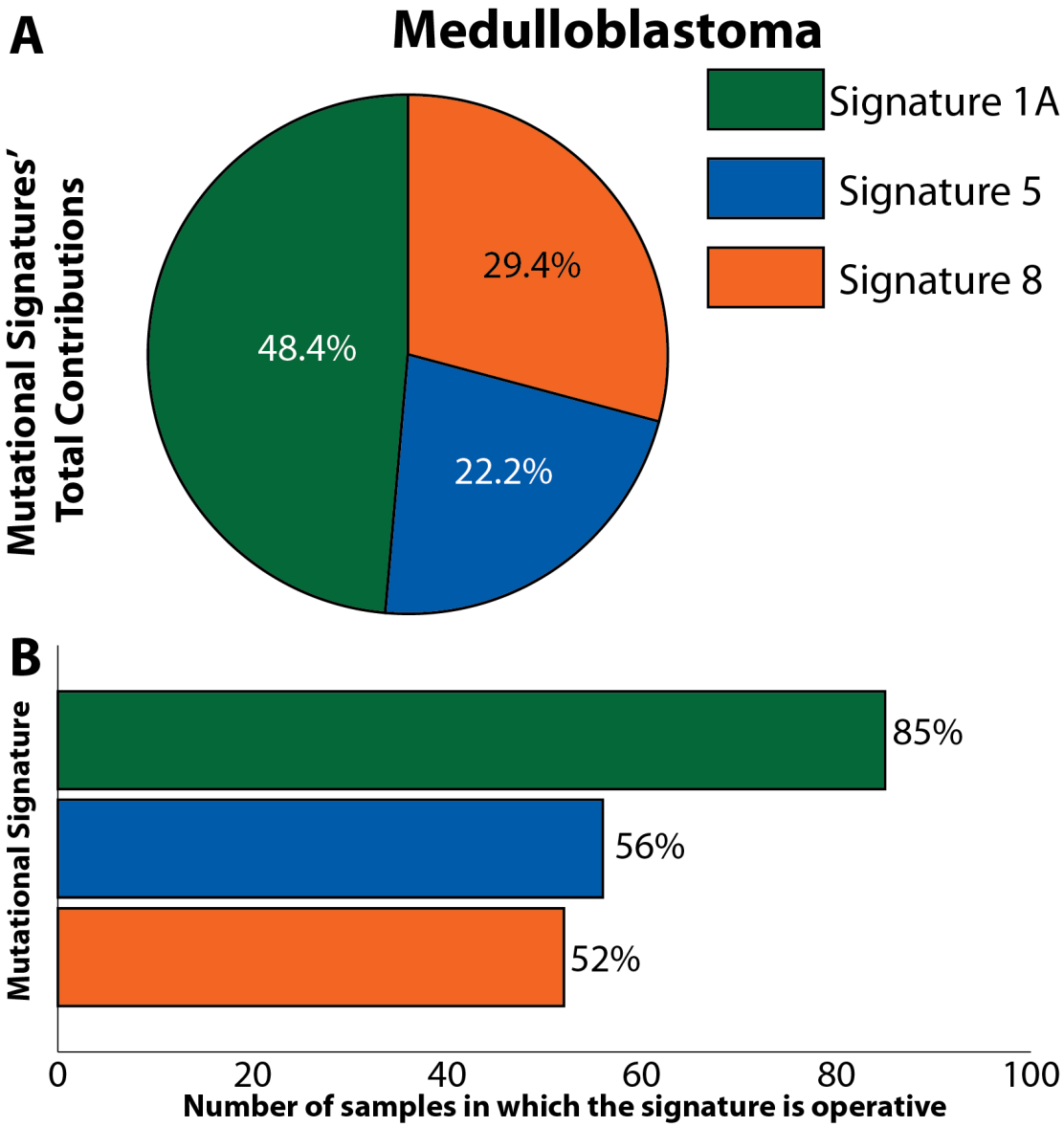

A
Liver Cancer
Mutational Signatures' Total Contributions
10.9%
18.3%
3.2%
18.7%
18.5%
27.8%
2.6%
Signature 1B
Signature 4
Signature 6
Signature 12
Signature 16
Signature 17
Signature U2
B
Mutational Signature
98%
85%
1%
89%
90%
20%
89%
0
20
40
60
80
100
Number of samples in which the signature is operative

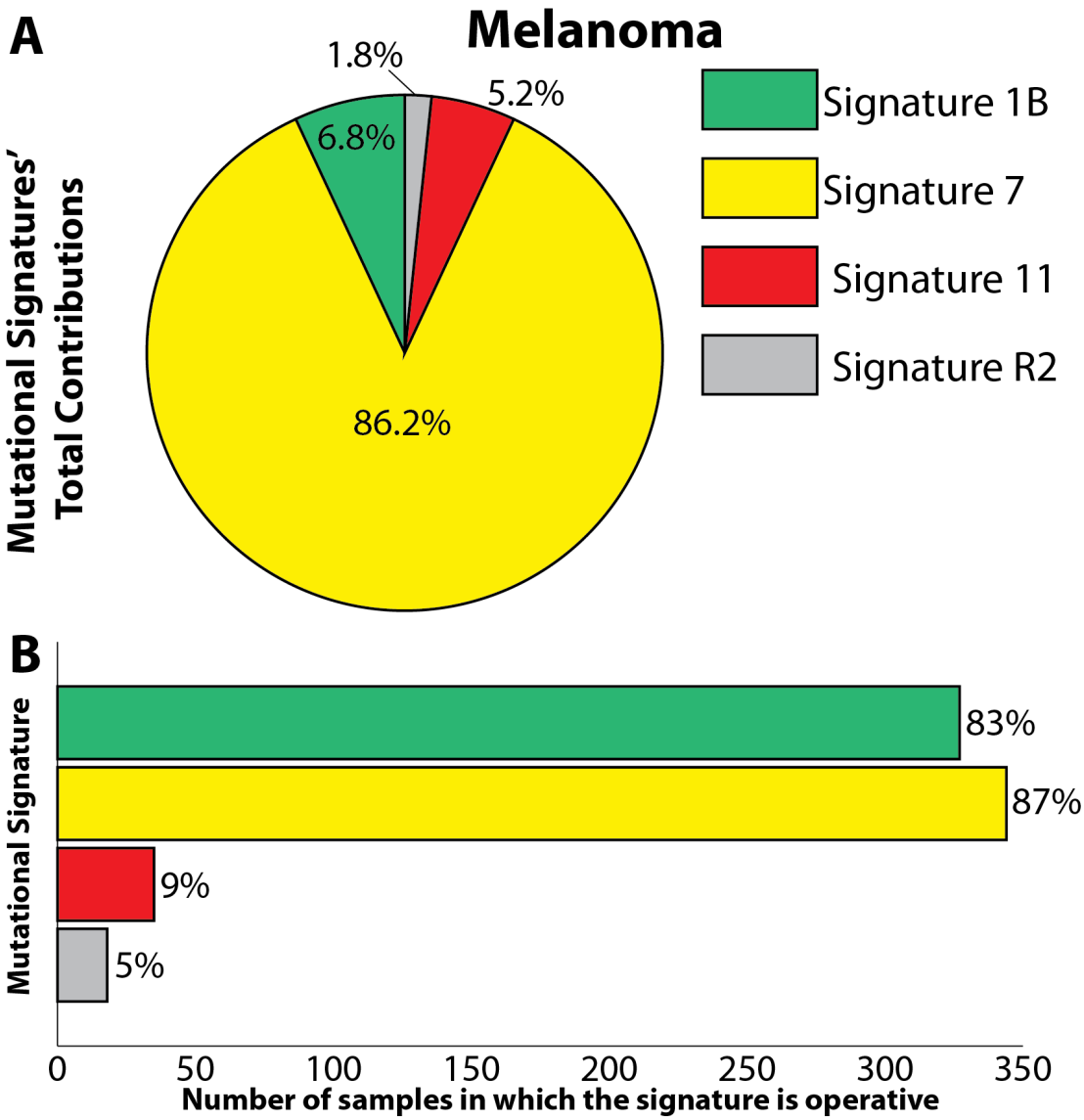

A
Lung Adenocarcinoma
Mutational Signatures' Total Contributions
9.1%
7.5%
9.6%
12.8%
61.0%
Signature 1B
Signature 2
Signature 4
Signature 5
Signature R2
B
Mutational Signature
72%
24%
63%
20%
3%
0
100
200
300
400
500
Number of samples in which the signature is operative

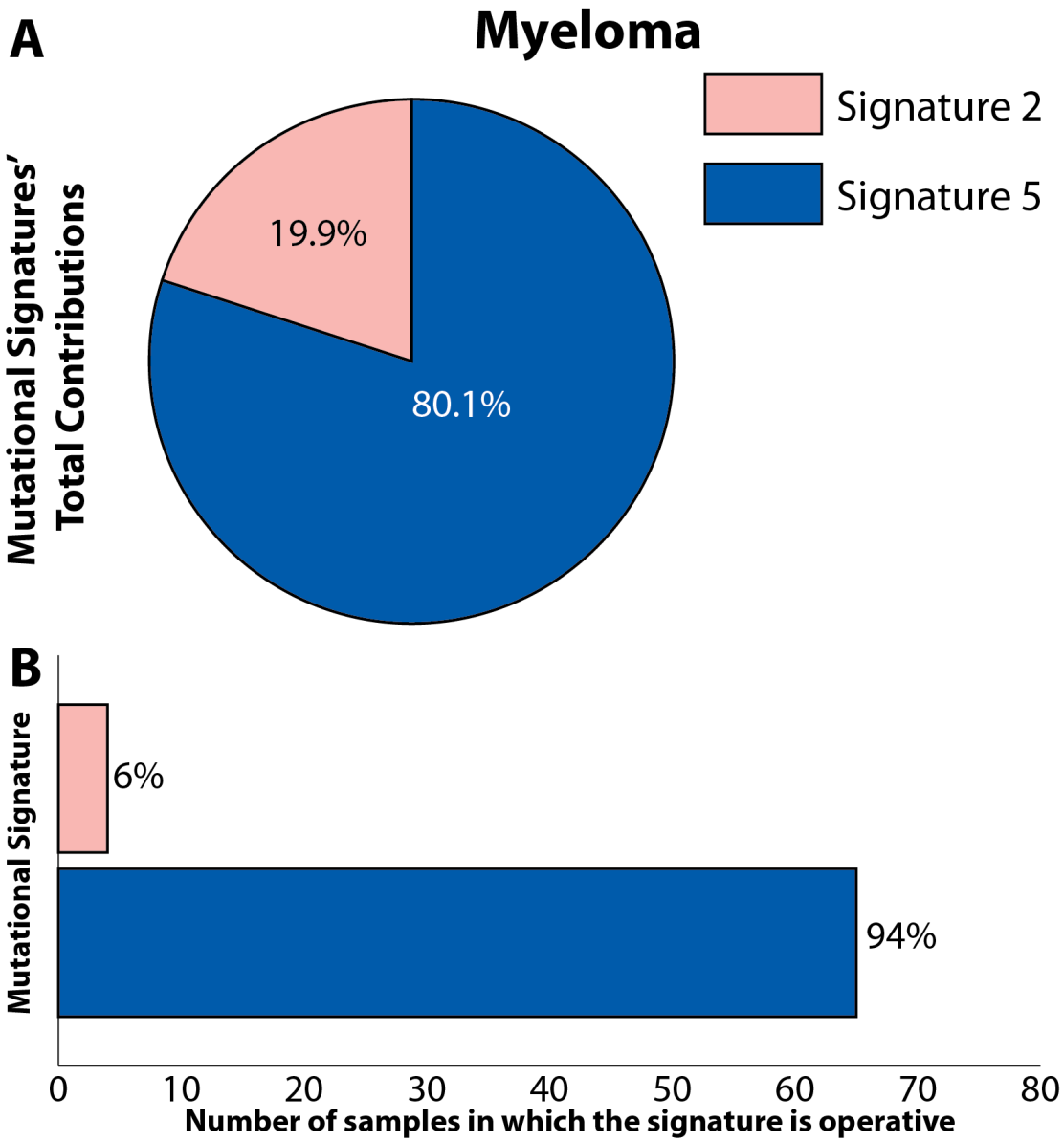

A
Lung Cancer Small Cell
Mutational Signatures' Total Contributions
Signature 4
Signature 15
7.2%
92.8%
B
Mutational Signature
97%
3%
0
10
20
30
40
50
60
70
80
Number of samples in which the signature is operative

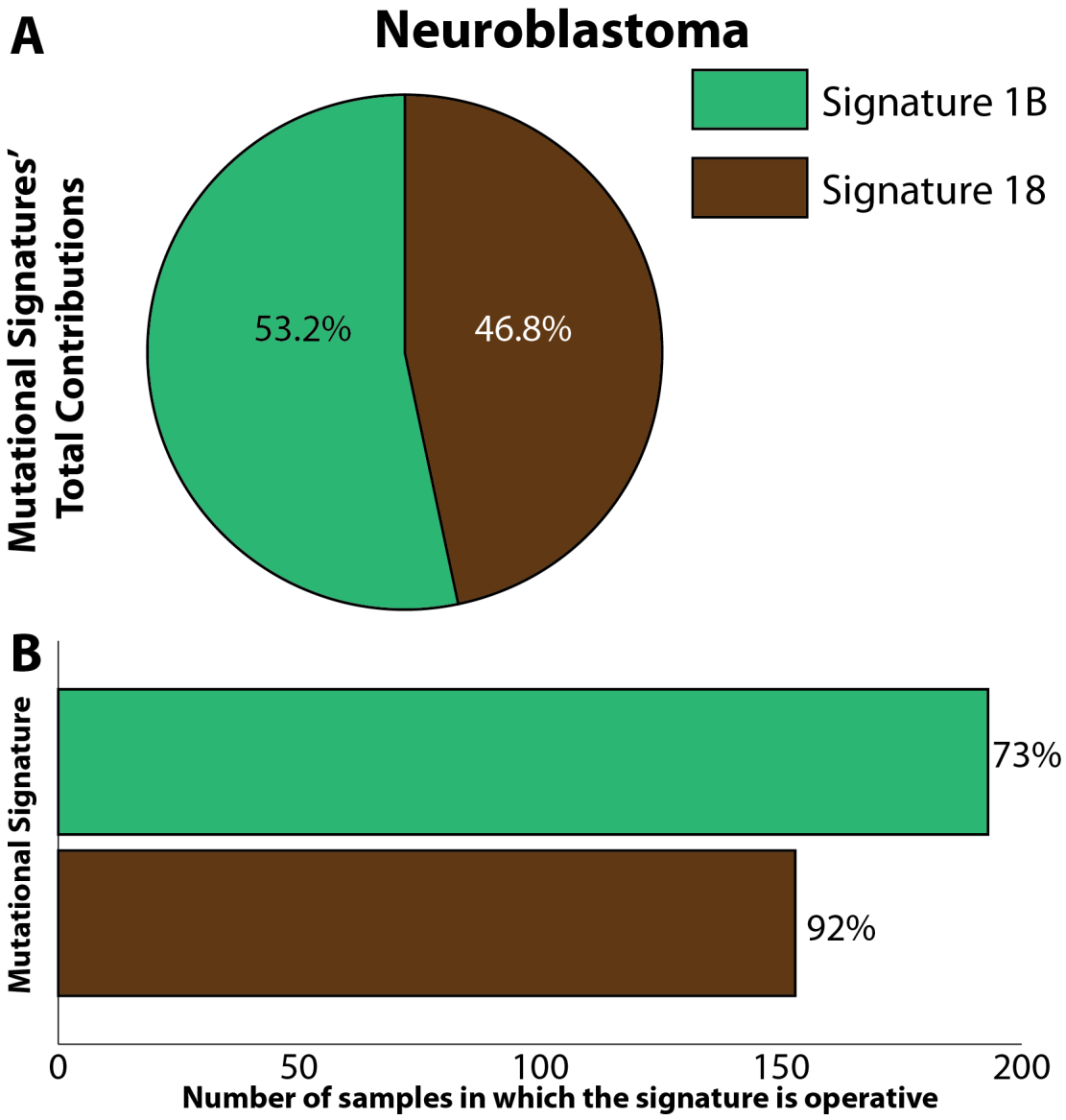

A
Lung Squamous*
Mutational Signatures' Total Contributions
18.6%
18.6%
36.2%
28.4%
Signature 2
Signature 4
Signature 5
Signature R2
B
Mutational Signature
35%
90%
93%
52%
0
50
100
150
200
Number of samples in which the signature is operative

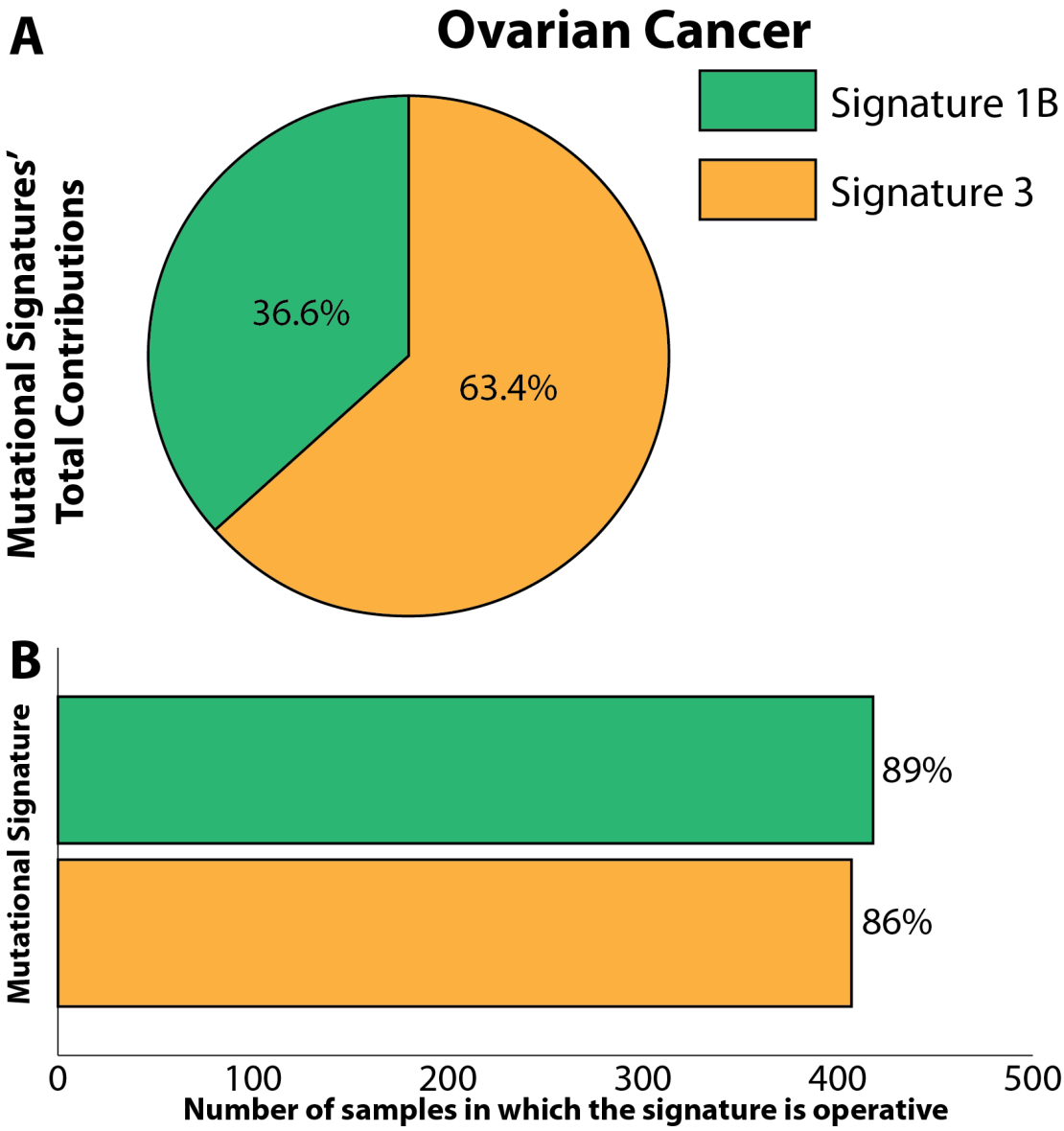

A
Lymphoma B-cell
Mutational Signatures' Total Contributions
10.6%
16.9%
46.3%
9.7%
16.5%
Signature 1B
Signature 2
Signature 5
Signature 9
Signature 17
B
Mutational Signature
96%
33%
25%
38%
25%
0
10
20
30
40
50
Number of samples in which the signature is operative

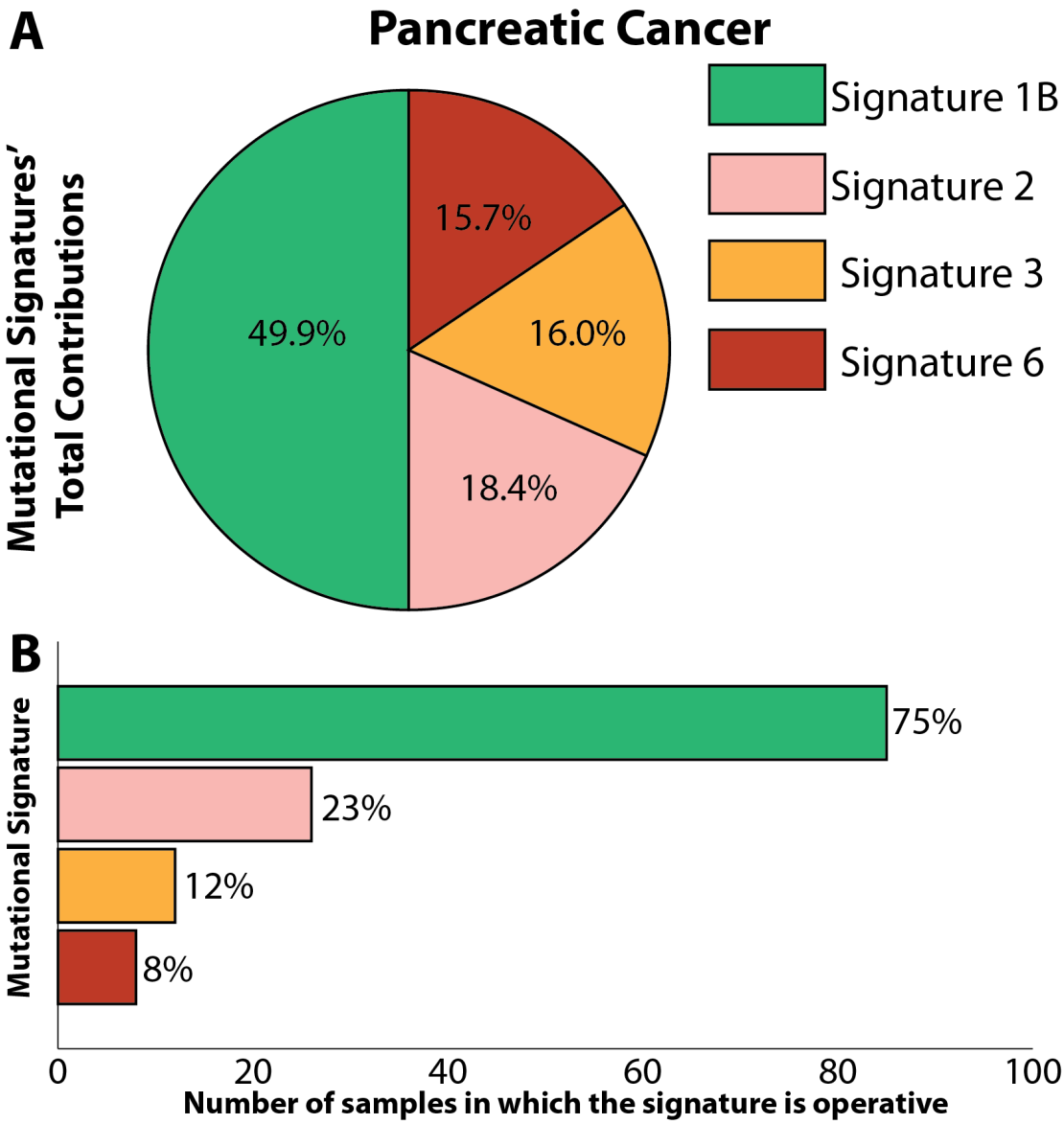

A
Medulloblastoma
Mutational Signatures' Total Contributions
Signature 1A
Signature 5
Signature 8
48.4%
29.4%
22.2%
B
Mutational Signature
85%
56%
52%
0
20
40
60
80
100
Number of samples in which the signature is operative

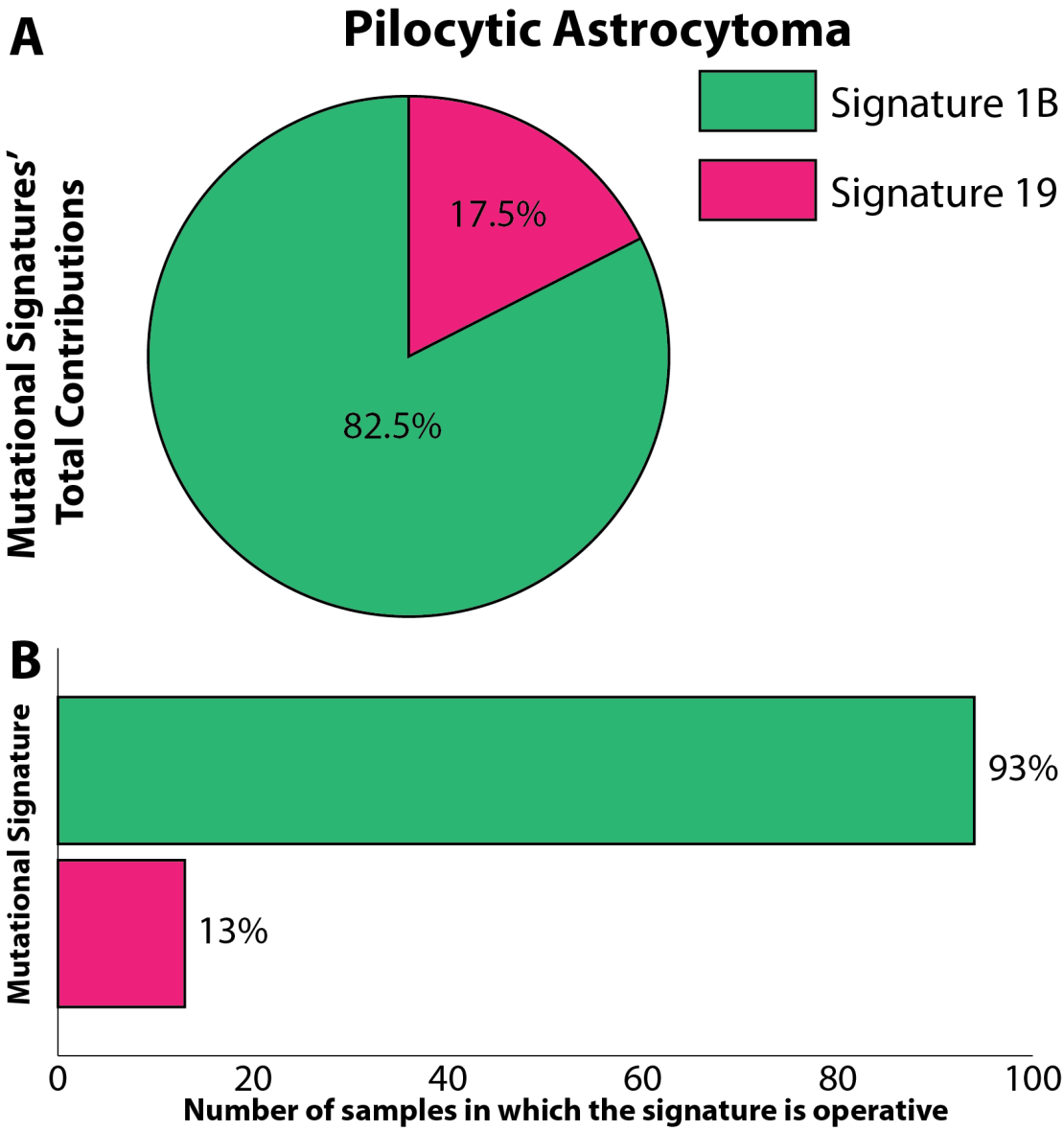

A
Melanoma
Mutational Signatures' Total Contributions
1.8%
5.2%
6.8%
86.2%
Signature 1B
Signature 7
Signature 11
Signature R2
B
Mutational Signature
83%
87%
9%
5%
0
50
100
150
200
250
300
350
Number of samples in which the signature is operative

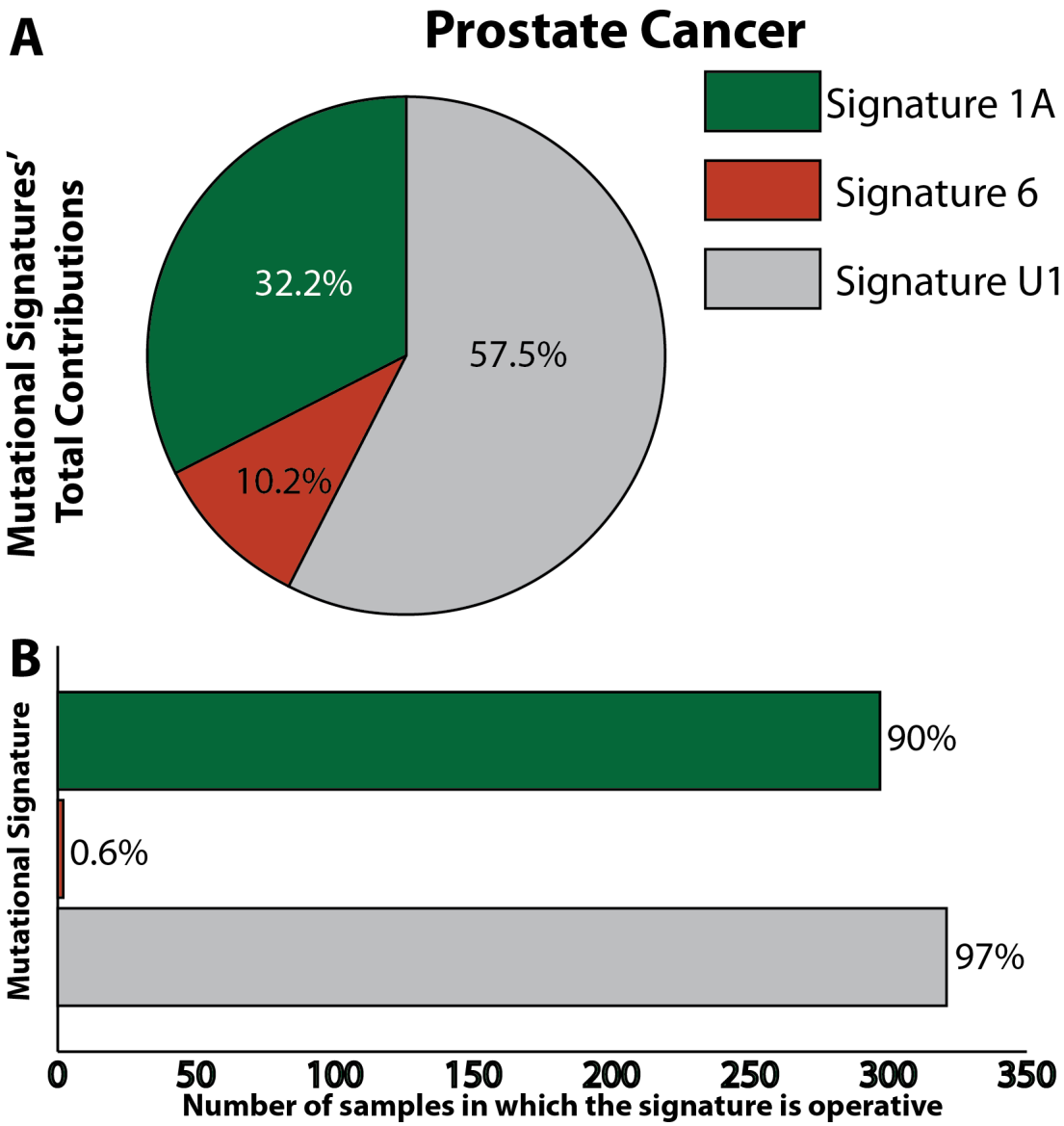

A
Myeloma
Mutational Signatures' Total Contributions
Signature 2
Signature 5
19.9%
80.1%
B
Mutational Signature
6%
94%
0
10
20
30
40
50
60
70
80
Number of samples in which the signature is operative

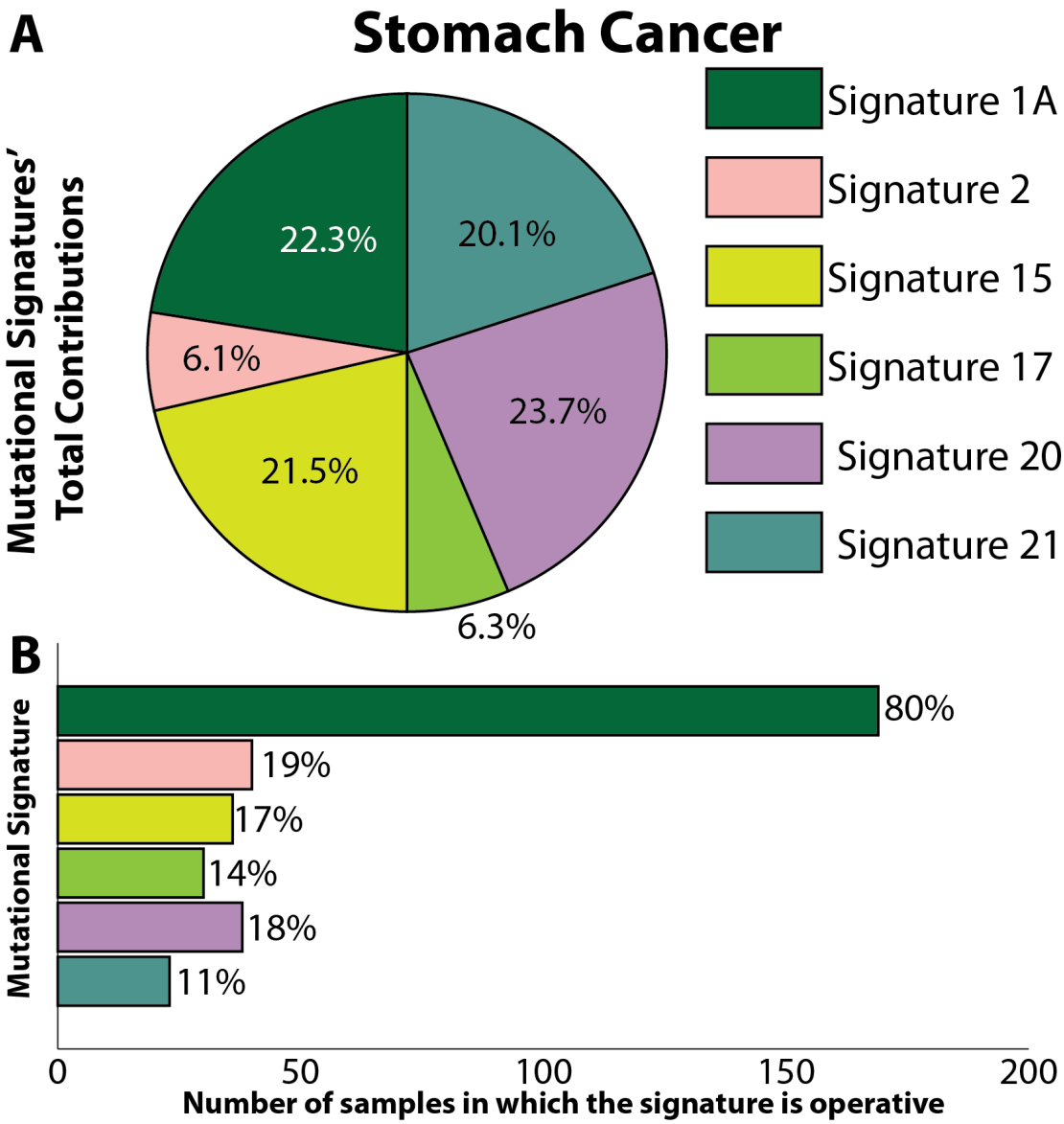

A
Neuroblastoma
Signature 1B
Signature 18
Mutational Signatures' Total Contributions
53.2%
46.8%
B
Mutational Signature
73%
92%
0
50
100
150
200
Number of samples in which the signature is operative

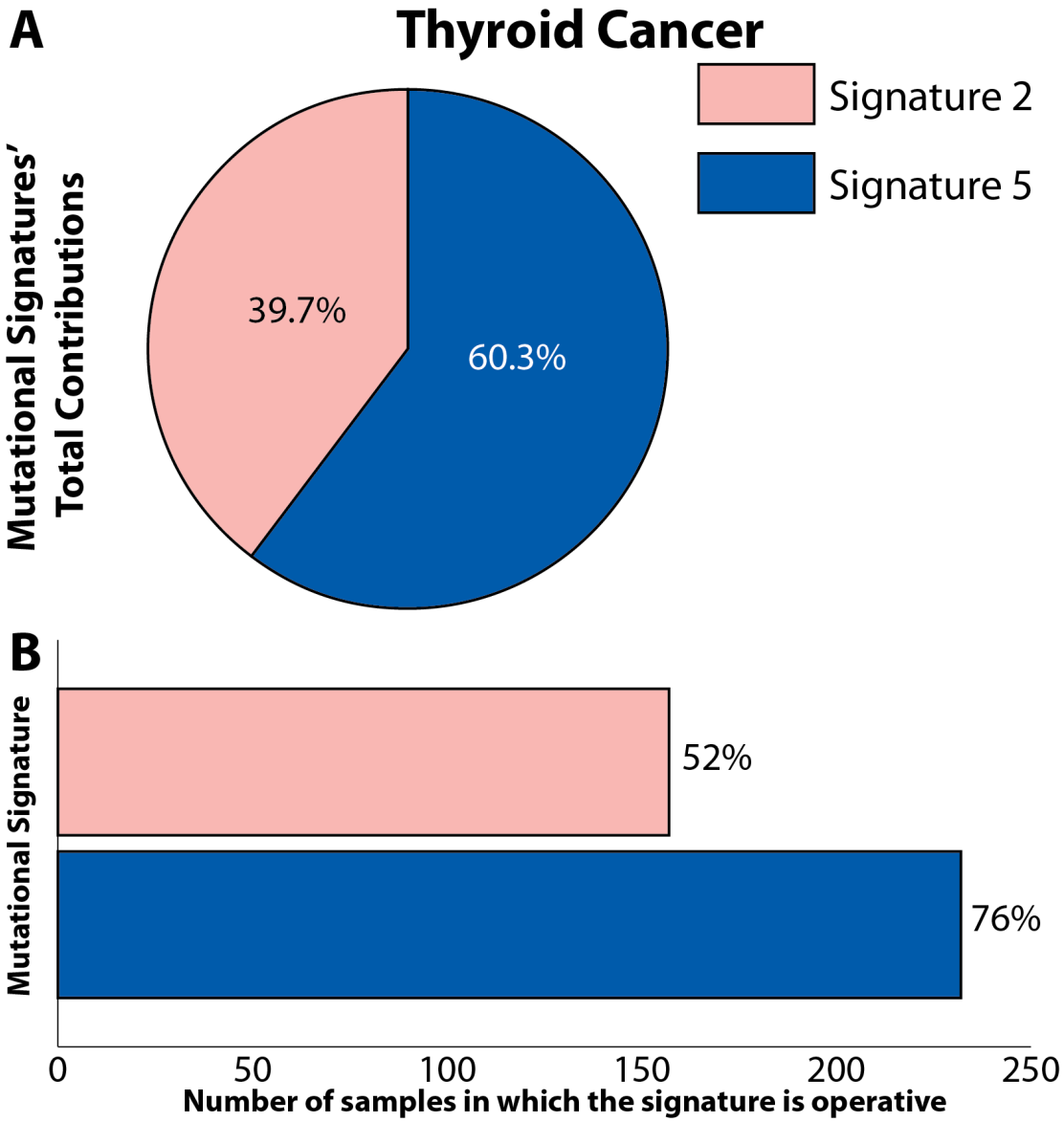

A
Ovarian Cancer
Mutational Signatures' Total Contributions
Signature 1B
Signature 3
36.6%
63.4%
B
Mutational Signature
89%
86%
0
100
200
300
400
500
Number of samples in which the signature is operative

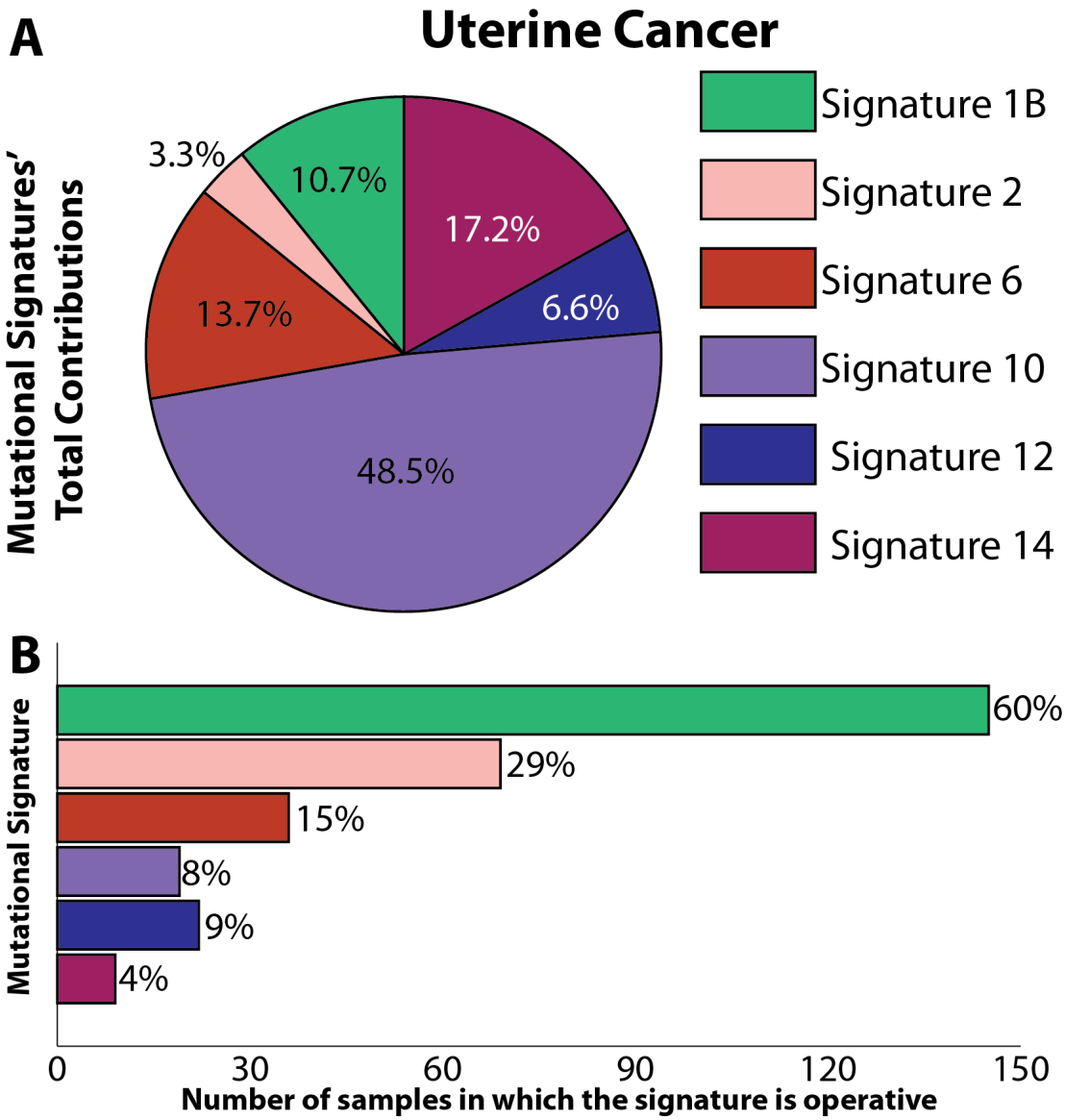

A
Pancreatic Cancer
Mutational Signatures' Total Contributions
Signature 1B
Signature 2
Signature 3
Signature 6
15.7%
49.9%
16.0%
18.4%
B
Mutational Signature
75%
23%
12%
8%
0
20
40
60
80
100
Number of samples in which the signature is operative


A
Pilocytic Astrocytoma
Signature 1B
Signature 19
Mutational Signatures' Total Contributions
17.5%
82.5%
B
Mutational Signature
93%
13%
0
20
40
60
80
100
Number of samples in which the signature is operative


A
Prostate Cancer
Mutational Signatures' Total Contributions
32.2%
10.2%
57.5%
Signature 1A
Signature 6
Signature U1
B
Mutational Signature
90%
0.6%
97%
0
50
100
150
200
250
300
350
Number of samples in which the signature is operative


A
Stomach Cancer
Mutational Signatures' Total Contributions
22.3%
20.1%
6.1%
23.7%
21.5%
6.3%
Signature 1A
Signature 2
Signature 15
Signature 17
Signature 20
Signature 21
B
Mutational Signature
80%
19%
17%
14%
18%
11%
0
50
100
150
200
Number of samples in which the signature is operative


A
Thyroid Cancer
Signature 2
Signature 5
Mutational Signatures' Total Contributions
39.7%
60.3%
B
Mutational Signature
52%
76%
0
50
100
150
200
250
Number of samples in which the signature is operative


A
Uterine Cancer
Mutational Signatures' Total Contributions
3.3%
10.7%
17.2%
13.7%
6.6%
48.5%
Signature 1B
Signature 2
Signature 6
Signature 10
Signature 12
Signature 14
B
Mutational Signature
60%
29%
15%
8%
9%
4%
0
30
60
90
120
150
Number of samples in which the signature is operative

APPENDIX VII: Publications associated with this thesis

This appendix contains the references of the articles that have been written and published as part of this thesis. The articles are separated into two categories: (i) main articles – four manuscripts directly related to developing and presenting the approach for deciphering mutational signatures and applying this approach to a large scale of whole-genome and whole-exome sequencing data, and (ii) supporting articles – seven manuscripts in which mutational signatures (and/or patterns of somatic mutations) have been examined. It is worth noting that this list of manuscripts does not include another six published articles unrelated to mutational signatures and/or cancer nor does it include another seven articles currently under review with which I have been involved during the course of my doctoral studies. Lastly, it should be noted that this thesis is almost entirely written based on the four main mutational signatures articles.

Main articles

Alexandrov LB and Stratton MR (2014) Mutational Signatures: The Patterns of Somatic Mutations Hidden in Cancer Genomes. **Current Opinion in Genetics & Development** 24, 52-60 (invited review article/corresponding author).

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinsk M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, and Stratton MR (2013) Signatures of mutational processes in human cancer. **Nature** 500:415-421.

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

Nik-Zainal S, **Alexandrov LB,** Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jonsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerod A, Tutt A, Martens JW, Aparicio SA, Borg A, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR (2012) Mutational processes molding the genomes of 21 breast cancers. **Cell** 149:979-993.

Supporting articles

Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, Martincorena I, Petljak M, **Alexandrov LB**, Gundem G, Tarpey PS, 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 MR (2014) Genome sequencing of normal cells reveals developmental lineages and mutational processes. **Nature (AOP)** doi:10.1038/nature13448

Murchison EP, Wedge DC, **Alexandrov LB**, Fu B, Martincorena I, Ning Z, Tubio J, Werner EI, Allen J, Barboza di Nardi A, Donelan EM, Marino G, Fassati A, Campbell PJ, Yang F, Burt A, Weiss RA, and Stratton MR (2014) Transmissible dog cancer genome reveals the origin and history of an ancient cell lineage. **Science** 343:437-440.

Bolli B, Avet-Loiseau H, Wedge D, Van Loo P, **Alexandrov LB**, Martincorena I, Dawson K, Iorio F, Nik-Zainal S, Bignell G, Hinton H, Li Y, Tubio J, McLaren S, O'Meara S, Butler AS, Teague J, Mudie L, Anderson E, Rashid N, Tai YT, Shammas M, Sperling A, Fulciniti M, Richardson P, Parmigiani G, Magrangeas F, Minvielle S, Moreau P, Attal M, Facon T, Futreal A, Anderson K, and Campbell PJ (2014) Heterogeneity of genomic architecture and evolution in multiple myeloma. **Nature Communications** 5 (2997).

Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, **Alexandrov LB**, Loo PV, Cooke SL, Marshall J, Martincorena I, Hinton J, Gundem G, van Delft FW, Nik-Zainal S, Jones DR, Ramakrishna M, Titley I, Stebbings L, Leroy C, Menzies A, Gamble K, Robinson B, Mudie L, Raine K, O'Meara S, Teague JW, Butler AP, Cazzaniga G, Biondi A, Zuna J, Kempski H, Muschen M, Ford AM, Stratton MR, Greaves M, and Campbell PJ (2014) RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. **Nature Genetics** 46 (2): 116-125.

Wong CC, Martincorena I, Rust AG, Rashid M, Alifrangis C, **Alexandrov LB**,

Tiffen JC, Kober C, Green AR, Massie CE, Nangalia J, Lempidaki S, Döhner H, Döhner K, Bray SJ, McDermott U, Papaemmanuil E, Campbell PJ, and Adams DJ (2013) Inactivating *CUX1* mutations promote tumorigenesis. **Nature Genetics** 46 (1): 33-38.

Nik-Zainal S, Van Loo P, Wedge DC, **Alexandrov LB**, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, Shlien A, Cooke SL, Hinton J, Menzies A, Stebbings LA, Leroy C, Jia M, Rance R, Mudie LJ, Gamble SJ, Stephens PJ, McLaren S, Tarpey PS, Papaemmanuil E, Davies HR, Varela I, McBride DJ, Bignell GR, Leung K, Butler AP, Teague JW, Martin S, Jonsson G, Mariani O, Boyault S, Miron P, Fatima A, Langerod A, Aparicio SA, Tutt A, Sieuwerts AM, Borg A, Thomas G, Salomon AV, Richardson AL, Borresen-Dale AL, Futreal PA, Stratton MR, Campbell PJ, and Breast Cancer Working Group of the International Cancer Genome C (2012) The life history of 21 breast cancers. **Cell** 149:994-1007

Murchison EP, Schulz-Trieglaff OB, Ning Z, **Alexandrov LB**, Bauer MJ, Fu B, Hims M, Ding Z, Ivakhno S, Stewart C, Ng BL, Wong W, Aken B, White S, Alsop A, Becq J, Bignell GR, Cheetham RK, Cheng W, Connor TR, Cox AJ, Feng ZP, Gu Y, Grocock RJ, Harris SR, Khrebtukova I, Kingsbury Z, Kowarsky M, Kreiss A, Luo S, Marshall J, McBride DJ, Murray L, Pearse AM, Raine K, Rasolonjatovo I, Shaw R, Tedder P, Tregidgo C, Vilella AJ, Wedge DC, Woods GM, Gormley N, Humphray S, Schroth G, Smith G, Hall K, Searle SM, Carter NP, Papenfuss AT, Futreal PA, Campbell PJ, Yang F, Bentley DR, Evers DJ, and Stratton MR (2012) Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. **Cell** 148:780-791