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.

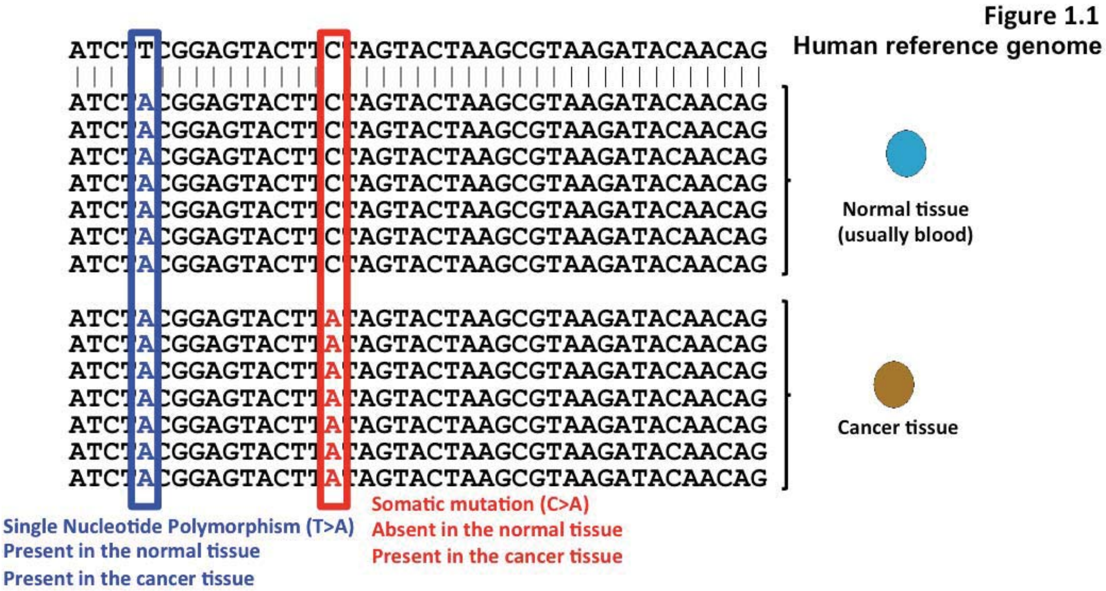
Figure 1.1
Human reference genome
ATCTTCGGAGTACTTCTAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTCTAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTCTAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTCTAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTCTAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTCTAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTCTAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTCTAGTACTAAGCGTAAGATACAACAG
Normal tissue
(usually blood)
ATCTACGGAGTACTTATAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTATAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTATAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTATAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTATAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTATAGTACTAAGCGTAAGATACAACAG
ATCTACGGAGTACTTATAGTACTAAGCGTAAGATACAACAG
Cancer tissue
Somatic mutation (C>A)
Absent in the normal tissue
Present in the cancer tissue
Single Nucleotide Polymorphism (T>A)
Present in the normal tissue
Present in the cancer tissue

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.

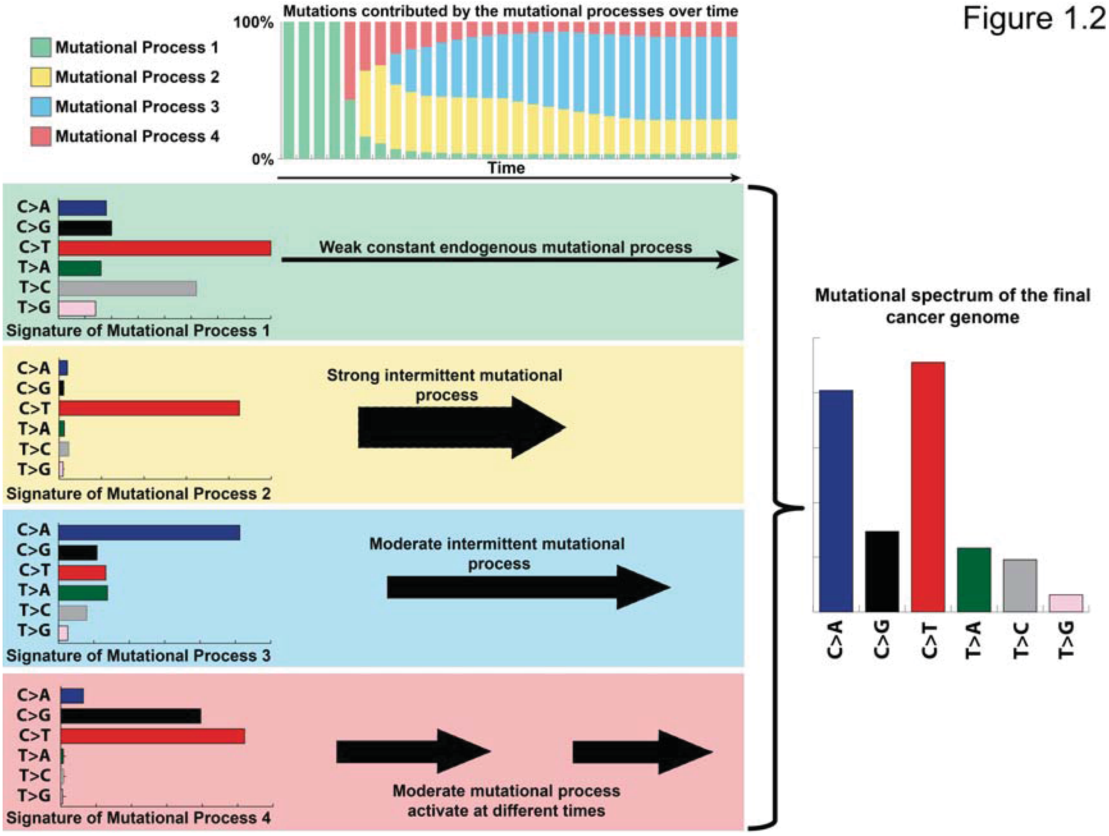

Figure 1.2: Illustration of mutational processes operative in a cancer. This simulated example illustrates four distinct mutational processes with variable strengths operative at different times throughout the lifetime of a patient. Each of these processes has a unique mutational signature exemplified 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 Tp$\underline{C}$ dinucleotides C>G substitutions at Tp$\underline{C}$ 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 Cp$\underline{C}$pC 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 CpCpC 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 Tp<u>C</u>) 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.