CHAPTER ONE: INTRODUCTION

1.1 THE WEALTH OF INFORMATION BURIED IN SOMATIC MUTATIONS

1.1.1 Cancer is a disease of the genome

Cancer is a disease of the genetic material of the cell. The earliest indication of a relationship between cancer and abnormalities of the genome was seen as far back as the turn of the twentieth century. David von Hansemann and Theodor Boveri both observed that, through erroneous cell division, cells could acquire an abnormal complement of genetic material with Boveri making early observations of aneuploidy. Through these studies, it was postulated that tumours could potentially arise from a progenitor cell that had acquired an anomalous complement of chromosomes following aberrant cell division (Boveri 1914).

DNA was identified as the constituent molecule of inheritance in the 1940s-1950s (Avery et al., 1944; Watson and Crick, 1953a) and this prompted an acceleration of discoveries that reinforced the belief that genetic pathology underpinned cancer. Increasing sophistication in chromosomal analyses of cancer cells showed specific and recurrent genomic abnormalities were associated with particular cancer types, such as the translocation between chromosomes 9 and 22 (the Philadelphia chromosome), in chronic myeloid leukaemia (Rowley, 1973). Subsequently, seminal work demonstrating that only a single *src* gene was required by Rous sarcoma virus to transform infected chicken cells into neoplastic cells (Bishop, 1985; Parker et al., 1984) paved the way to the earliest understanding of how transforming retroviruses were able to confer a cancer phenotype. Furthermore, the transfer of genomic DNA from a range of cancers into phenotypically normal NIH3T3 cells was shown to transform the recipient cells into neoplastic cells (Shih et al., 1981) and demonstrated that the cancer-causing genes found to underlie this transformation were mutant versions of normal growth-controlling genes, which were termed proto-oncogenes (Perucho et al., 1981; Pulciani et al., 1982). This transforming activity was eventually isolated to be due to the first naturally occurring, human cancer-causing sequence change—the single base G > T substitution that causes a glycine to valine substitution in codon 12 of the *HRAS* gene (Reddy et al., 1982). This discovery has essentially set the course for cancer research, where the enduring hunt for abnormal genes underlying the development of human cancer continues to the present day.

1.1.2 Multiple acquired mutations are required for the development of cancer

The multistep process of tumourigenesis was suggested as far back as 1958 (Foulds, 1958) and the molecular events punctuating cancer development unfolded over the next 30 years (Farber and Cameron, 1980; Weinberg, 1989). An appreciation of the complexity of the genetic path in cancer development has come from studies involving a series of colonic-tissue biopsy specimens representing the various histopathological stages from normal epithelium to frank colorectal cancer (Fearon and Vogelstein, 1990). They observed that the great majority of early adenomatous polyps carried inactivating mutations of the tumour-suppressor gene *APC*. Roughly half of the intermediate-sized carried activating mutations of *ras* oncogenes and about half of the advanced colorectal carcinomas had mutations in the tumour-suppressor gene *TP53* (Kinzler and Vogelstein, 1996). This study documents the genetic route to a neoplastic state in colorectal cancer. This scheme however, has not been reproduced in other cancers in such detail and cannot define the precise number nor the nature of key mutations required for normal cells to turn into tumour cells in humans.

1.1.3 Chronic chemical exposure leads to DNA damage, mutations and eventually cancer

Epidemiologic analyses have contributed to the understanding of how environmental and occupational chemicals cause cancer. For example, 18th century physicians reported an increased incidence of nasal polyps amongst users of snuff as well as scrotal cancer amongst English chimney sweeps [reviewed (Brown and Thornton, 1957)]. As the link between scrotal cancers and exposure to the polycyclic aromatic hydrocarbons in soot became apparent, European occupational authorities issued recommendations advising frequent bathing for chimney sweeps. A century later, this public health intervention saw a virtual eradication of scrotal cancers in chimney sweeps in Europe, but not in England, where bathing frequency remained low (Butlin, 1892). This epidemiologic observation reinforces a basic tenet of carcinogenesis: that there is a strong relationship between chemical exposure and tumour development. Many examples of chemical exposure leading to carcinogenesis are known including cigarette smoking and lung cancer, aniline dyes and bladder cancer, asbestos and mesothelioma, aflatoxin with liver cancer and benzene products with leukaemia (Pfeifer et al., 2002; Walker and Gerber, 1981; Yang, 2011) .

Despite the exposures, many of these tumours typically arise a long period of time after the exposure, usually in later life. It was postulated that this latent period represents the time required for early exposure-related DNA damage to become fixed as mutations and eventually evolve into a malignancy. This perception was underscored by the fact that in the general population, the

incidence of most cancers increases with increasing age reflecting the time taken to accumulate somatically acquired mutations in cancer cells (Armitage and Doll, 1954).

1.1.4 A critical accumulation of mutations prior to malignant transformation

Epidemiologic analyses of the incidence of cancer provide some measure of the number of distinct changes that must occur for tumourigenesis to reach completion. Fixed mutations in individual cells are transmitted from one generation of cells to another and whilst DNA damage by exogenous or endogenous chemicals occurs randomly, the gradual accumulation of somatic mutations eventually leads to the abnormal behaviour associated with cancer cells (Hanahan and Weinberg, 2000). The number of key somatic mutations required for the transformation of a normal cell into a cancerous state has been estimated to be in between 6 to 10 (Hahn and Weinberg, 2002; Renan, 1993).

1.1.5 Drivers and passengers

These key somatic mutations are thought to be "driver" mutations that confer selective clonal growth advantage, are causally implicated in oncogenesis and have been positively selected during the evolution of the cancer. The search for driver mutations has led to the discovery of many "cancer genes" providing insights into mechanisms of tumorigenesis and targets for therapeutic intervention (Stratton et al., 2009).

The vast majority of somatic mutations, however, are "passenger" events. These do not contribute to cancer development. Nevertheless, passenger mutations are a rich source of information. Despite not being the focus of selection, these bystander mutations bear the imprints of mutational mechanisms and DNA repair processes that have been operative during the development of the cancer (Stratton et al., 2009).

1.1.6 Historic analyses of mutation patterns in reporter genes unearthed the earliest signs of carcinogen-specific mutational processes in cancer

Historically, the analysis of mutation patterns to investigate underlying DNA damage and repair processes in human cancers has predominantly been restricted to reporter cancer genes, notably *RAS* oncogene and *TP53* tumour suppressor gene, which yield abundant mutations from case series (DeMarini et al., 2001; Giglia-Mari and Sarasin, 2003; Pfeifer, 2000). These studies have revealed that the overall mutational spectra, codon position, sequence context and DNA strand for the sequence-specific DNA binding-domain (amino acids 97 to 300) of the *TP53* gene, for example, can be tumour-type specific and related to exogenous carcinogens and repair processes.

For instance, benzo[a]pyrene diolepoxide (B[a]DPE) is a by-product of the polyaromatic hydrocarbons (PAH) from tobacco-smoke. The distribution of B[a]PDE adducts along the TP53 gene was mapped at nucleotide resolution level in PAH-treated normal human bronchial epithelial cells (Denissenko et al., 1996). Since then, remarkable correlations between benzo[a]pyrene adduct formation sites and the mutation spectrum in lung cancer (Pfeifer et al 2002), have been documented. Furthermore, the selective occurrence of these PAH-damage hotspots is related to patterns of cytosine methylation in the TP53 gene (Pfeifer, 2000). Guanines flanked by 5-methylcytosine were the preferentially adducted positions. In human lung cancers, 5 of the 6 most prominent mutation hotspots in the TP53 gene are represented by C>A/G>T transversion mutations at codons containing methylated CpG sequences, including codons 157, 158, 245, 248 and 273 (Pfeifer et al., 2002). Therefore, methylated CpGs in the TP53 gene represent a preferential target for exogenous carcinogens in smoking-associated lung cancer. This supports the role of by-products of tobacco-smoking in the aetiology of lung cancer. Additionally, these mutations exhibit a strong transcriptional strand bias with fewer C>A/G>T mutations on the transcribed than the non-transcribed strand. The latter is generally believed to reflect the past activity of transcription-coupled nucleotide excision repair on bulky adducts of guanine caused by tobacco carcinogens (Hainaut and Pfeifer, 2001).

Similarly, ultraviolet (UV) light associated damage has been shown to induce C>T/G>A and CC>TT/GG>AA transitions. These occur predominantly at dipyrimidines, reflecting the formation of pyrimidine dimers following exposure of DNA to ultraviolet light (Pfeifer et al., 2005). These mutations also show transcriptional strand bias, with fewer C>T/G>A mutations on the transcribed than non-transcribed strand, probably due to the action of transcription-coupled repair on impaired pyrimidines.

Insights have been gained through studies of other cancer types. For example, mouse embryonic fibroblasts, from a Hupki (exon 4-9 human p53 knock-in) mouse model, were treated with aristolochic acid, a plant extract implicated in Chinese herb nephropathy, leading to urothelial cancer development (Feldmeyer et al., 2006). A characteristic mutation spectrum of A>T/T>A transversions was seen mimicking the mutational spectra seen in urothelial tumours from patients with exposure to aristolochic acid, supporting the role of this compound in the aetiology of urothelial tumours (Nedelko et al., 2009). Other examples of exogenous mutagenic exposures leading to distinctive mutational patterns in human cancers include C>A/G>T transversions in aflatoxin B1-associated hepatocellular carcinomas (Mace et al., 1997).

Although these studies have been highly informative, they are limited by the fact that only a single mutation from each cancer sample is usually incorporated into each dataset. Moreover, because they depend upon driver mutations in cancer genes, the effects of selection have been superimposed upon the mutational patterns initially generated by the DNA damage and repair processes. These studies have, therefore, been well placed to report strong exposures and dominant repair processes that are operative across most cases of a particular tumour type. Where there is heterogeneity of damage and repair process in a cancer class, however, an averaged spectrum generated by many different processes will be reported.

1.1.7 The wealth of information revealed by detailed analysis of complete catalogues of somatic mutation

In recent years, technological improvements in sequencing methods have seen a vast increase in scale. No longer is sequencing limited to PCR-based coding exons. The generation of 30 gigabases per sequencing experiment permits whole human genomes to be sequenced in a single experiment. Recent analyses of comprehensive mutational catalogues obtained from whole-genome sequencing of a single malignant melanoma and a single lung cancer illustrate the power of this approach (Pleasance et al., 2010a; Pleasance et al., 2010b). They clearly revealed the characteristic mutational spectra of ultraviolet light and tobacco carcinogens respectively and provided strong evidence for the past activity of transcription-coupled repair. In addition, analysis of C>A/G>T mutations in the lung cancer showed a strong preference for CpG dinucleotides outside CpG islands, suggesting a role for methylated cytosine in fostering such mutations as CpG islands are usually unmethylated. Conversely, C>G/G>C mutations, which also preferentially occurred at CpG dinucleotides, were more prevalent within CpG islands suggesting that the mutagen(s) underlying these mutations preferentially acted on unmethylated DNA (Pleasance et al., 2010b). In the melanoma, at least one

additional mutational process characterised by C>A/G>T changes and which appeared to be independent of ultraviolet light exposure was shown to have been operative. In both cancers, mutations were discovered to be more common in poorly expressed genes than in highly expressed genes, both on the transcribed and non-transcribed strands. The mechanism underlying this expression-related phenomenon is unknown (Pleasance et al., 2010a).

In summary, these studies demonstrated how the global and unbiased depiction of these individual cancers provided by whole genome sequencing permitted more refined insights into mutational processes of known carcinogenic exposures and their relationship with genomic features. However, the nature of the underlying mutagenic and repair processes in most other cancer types is much less well understood than for melanoma and lung cancer. Following the lead of these individual genomes, in this thesis, essentially the full repertoire of somatic mutations in twenty-one breast cancers will be documented in order to investigate the mutational mechanisms and repair processes that have shaped these cancer genomes.

1.2 MUTATIONAL PROCESSES LEAVE CHARACTERISTIC IMPRINTS OR *MUTATIONAL SIGNATURES* IN CANCER GENOMES

A genome-wide archive of somatic mutations provides a panoramic view of the resulting mutational landscape. At the point of a patient's cancer diagnosis, the set of somatic mutations that is revealed through sequencing of the cancer is the aggregate outcome of one or more mutational processes. Each process leaves a characteristic imprint or *mutational signature* on the cancer genome, defined by the mechanisms of DNA damage and DNA repair that constitute it.

Whatever the nature of the mutagenic or repair mechanisms in operation, the final catalogue of mutations is also determined by the strength and duration of exposure to each mutational process (Figure 1.1). Some exposures may be weak or moderate in intensity, while others may be very strong in their assertion. Similarly, some exposures may be on-going through the entire lifetime of the patient, even preceding the formation of the cancer, and some may commence late or become dominant later in tumourigenesis.

Additionally, cancers are likely to comprise of populations of cells including subclonal populations which may have been variably exposed to each mutational process, promoting the complexity of the final landscape of somatic mutations in a cancer genome. Because there are so many potential exogenous and endogenous DNA damaging agents as well as a plethora of intrinsic DNA repair pathways, in the next section, mutagenic and repair pathways will be reviewed in brief and attention will be paid to documenting characteristic signatures associated with each mechanism. The purpose of this exercise is to build a framework of known signatures (Table 1.1) from past analyses of experimental systems. Throughout this thesis, mutational signatures identified in the cancers will be compared to this framework and matched to known signatures in order to gain insights into the nature of mutational and repair processes that have been operative on the cancers.

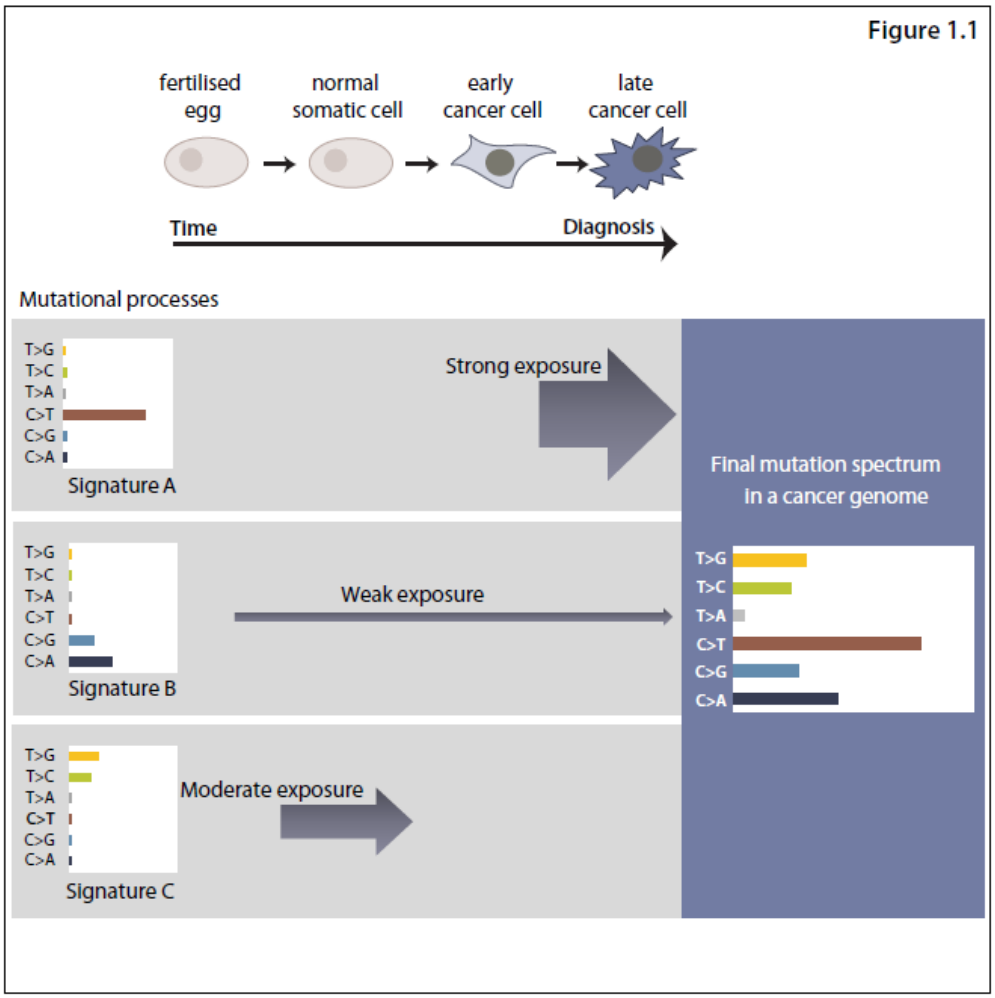


Figure 1.1: Mutational signatures in cancer genomes. From the time of the fertilised egg through to the development of an invasive cancer, multiple mutational processes are likely to be operative with each process producing its own characteristic signature. At the point of diagnosis and of sequencing the cancer genome, the final mutation spectrum is a composite of the multiple mutational processes that have been operative which may show variation in the intensity (size of arrow) and duration (length of arrow) of exposure to each mutational process.

1.3 PROCESSES OF DNA DAMAGE AND THEIR CHARACTERISTIC SIGNATURES

DNA is under a constant stream of attack from a variety of exogenous and endogenous sources. Each of these mutagens may cause damage directly or indirectly to the nucleotides in the genome. The ensuing damage may be in the form of biochemical covalent modifications or spontaneous/enzymatic alteration of the nucleotides. Here, the causes of DNA damage have been classified in the following way; (i) spontaneous or enzymatic conversions, (ii) physical agents, (iii) free radical species (iv) chemical agents. Each class of DNA damaging agent will be discussed in the following sections.

1.3.1 Spontaneous or enzymatic conversions

Mutations in DNA can occur without exposure of cells to chemicals or irradiation and may accumulate simply as part of the natural rate of endogenous errors in the human genome. Those errors that are known to be due to an enzymatic reaction are regarded as such, whereas those errors for which there is no causative enzyme known, may historically have been termed “spontaneous”. However, the possibility of a yet unknown aberrant enzymatic cause for such conversions cannot be excluded. In this section, spontaneous or endogenous enzyme-catalysed forces that drive DNA mutagenesis will be discussed.

1.3.1.1 Spontaneous generation of apurinic/apyrimidinic sites

The chemical bond linking a base and a pentose sugar in nucleotides is the N-glycosidic bond, which is particularly labile, and can lead to spontaneous base loss (~10^4/cell/day) (Lindahl, 1993) resulting in apurinic/apyrimidinic sites (AP site). Purines are believed to be more frequently affected. If an AP site remains uncorrected upon entering replication, there will be uncertainty regarding which base should be inserted opposite the AP site. This non-instructive lesion obstructs replicative polymerases during DNA synthesis and increases the likelihood for errors. It is thought that error-prone translesion synthesis polymerases are triggered by AP-site induced replication-blocking with a predilection for insertion of ‘A’ opposite AP sites, or the A-rule (Strauss, 2002). Furthermore, via DNA damage tolerance mechanisms, other translesion polymerases can also provide an escape route avoiding replication fork collapse at the expense of generating a C>T:G>A transition signature, resulting in a myriad of other mutation spectra (e.g. REV1 generates a C>G:G>C signature, Pol η generates mutations at A:T bases)(Kunz et al., 2000; Sale et al., 2012).

1.3.1.2 Deamination of bases

Deamination is a reaction which causes loss of an amine group from a molecule to generate a carboxyl group. It is thought that deamination reactions can occur spontaneously in all four bases in the human cell, albeit slowly (Figure 1.2b). There are endogenous enzyme families that exist which can catalyse the deamination process. In the following section, various types of deamination and the mutational signatures which they leave on the human genome are considered.

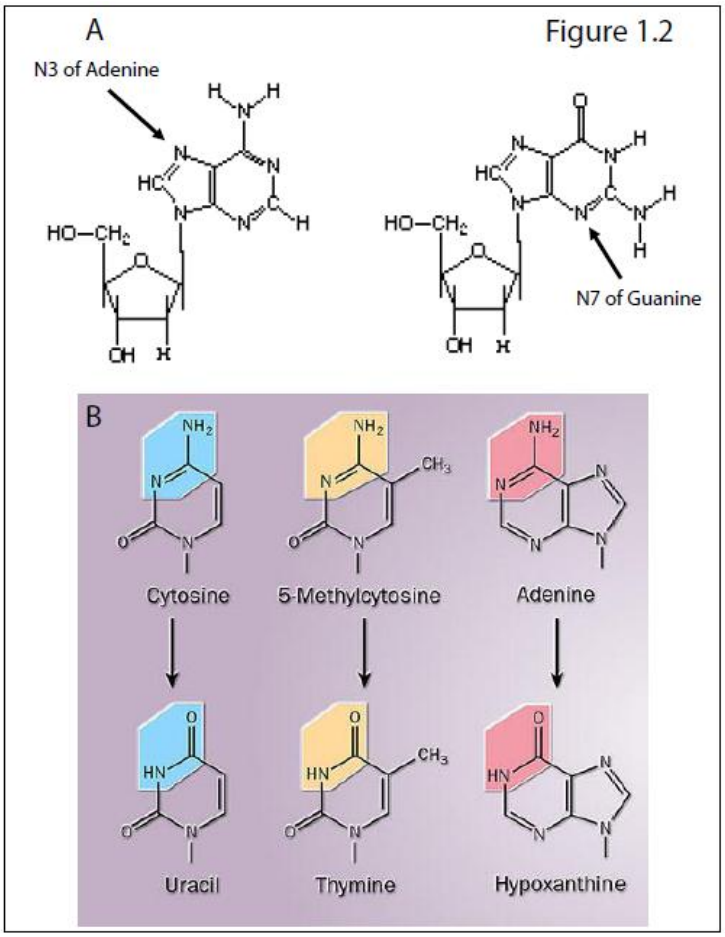


Figure 1.2: Base susceptibility to damage. The basic chemical unit of DNA is the nucleotide which comprises a phosphate group esterified to a pentose sugar, which is joined by a labile N-glycosidic bond to a base. However, structural properties of certain bases make them susceptible to DNA damage. (A) Nitrogenous hexose rings of adenine and guanine make them excellent targets for electrophilic attack by reactive compounds. These are called nucleophilic centers and N^7 on guanine and N^3 on adenine are (highlighted with arrows and) examples of such nucleophilic centers. (B) Common deamination reactions of bases in the human genome. Image from www.chemistrypictures.org.

i. Methylated cytosines at CpG dinucleotides

It has been observed in the human genome that CpG dinucleotides which are not within CpG islands are frequently methylated to form 5-methylcytosine. In the human genome, evolutionary loss of methylated CpG dinucleotides is believed to have resulted in the number of methylated CpGs to be a fifth of what is expected (Shen et al., 1994). This evolutionary decay at methylated CpGs coupled with approximately 23% of mutations in hereditary human diseases and 24% of mutations in the reporter gene *TP53* in human cancers shown to be C>T/G>A transitions at sites of cytosine methylation (Waters and Swann, 2000) has led to methylated CpG dinucleotides being considered "mutational hotspots" .

The propensity for this well-documented mutational phenomenon to result in C>T/G>A transitions has historically been hypothesised to be due to hydrolytic deamination of 5-methylcytosine to form thymine(Lutsenko and Bhagwat, 1999). More recently however, it has been thought to be attributed to failure of attempted maintenance of methylation of CpG dinucleotides by DNA-(cytosine-5) methyltransferase (Stojic et al., 2004). Whatever the true cause of this decay, the net observed effect is one of C>T/G>A transitions at methylated CpG dinucleotides.

ii. Cytosine to uracil deaminations

The spontaneous process of cytosine deamination to uracil is believed to occur slowly (~10^2-10^3/cell/day) but can be catalysed by members of the cytidine deaminase family. Uracil has a propensity to base pair with adenine instead of guanine, subsequently giving rise to a C>T/G>A transition. Although activation-induced cytosine deaminase (AID) is the enzyme that is most well-characterised from this family of DNA editors, all the family members will be discussed below in some detail.

a) Activation-induced cytidine deaminase (AID)

Activation-induced cytidine deaminase (AID), is a nucleotide-editing enzyme which deaminates cytosine residues within the immunoglobulin loci in B lymphocytes and triggers double-strand breaks, initiating both somatic hypermutation and class-switch recombination [reviewed (Longerich et al., 2006)]. Whilst AID primarily functions in antibody diversification, recent studies have revealed that AID has DNA editing abilities at non-immunoglobulin loci, like *bcl11a* (Staszewski et al., 2011*)*. Furthermore, this mutagenic activity is not restricted to B lymphocytes, occurring in non-lymphoid cells in experimental systems as well (Chen et al., 2012b; Jovanic et al., 2008). The mutational

signature of this DNA-editing enzyme is well-characterised. AID exhibits a strong preference for deaminating C residues flanked by a 5'-purine (Pham et al., 2003).

b) APOBEC1

APOBEC1 was first identified as an RNA-editing enzyme (Teng et al., 1993) with restricted expression to the small intestine, where it strictly deaminates a single cytosine on the apolipoprotein B mRNA (C6666), creating a premature translational stop codon. Of interest, this stringent editing fidelity can be overcome. When forcibly over-expressed in transgenic mice, APOBEC1 can lead to non-specific editing of apoB mRNA as well as other mRNAs (Petit et al., 2009). When editing DNA, APOBEC1 favours cytosine residues flanked by a 5'T (Harris et al., 2002; Hultquist et al., 2011).

c) APOBEC2

APOBEC2 was thought to be expressed exclusively in skeletal muscle and heart and its function, substrate and nucleotide-editing activity was essentially unknown (Conticello, 2008; Liao et al., 1999). Recently, APOBEC2 transgenic murine models were used to demonstrate that constitutive expression of APOBEC2 in the liver resulted in elevated RNA editing in *Eif4g2* and *PTEN* reporter genes. Furthermore, hepatocellular carcinoma developed in 2 of 20 APOBEC2 transgenic mice at 72 weeks of age and caused lung tumors in 7 of 20 transgenic mice analyzed (Okuyama et al., 2012). However, DNA-editing capacity has not been demonstrated and no known DNA mutational signatures have been attributed to this enzyme.

d) APOBEC3

The APOBEC3 family of enzymes is believed to have arisen from a gene duplication event of AID in placental mammals, which was subsequently followed by an expansion, and presently comprises seven APOBEC3 proteins in humans (APOBEC3A-H)(Conticello, 2008). The prototypical APOBEC3G, as well as several other APOBEC3s, act on lentiviral replication intermediates constituting an innate pathway of anti-retroviral defence (Hultquist et al., 2011; Sheehy et al., 2002).

APOBEC3 activity is not confined to restriction of viral genomes. *In vitro*, forced over-expression of APOBEC3A was shown to compromise genomic integrity of human cells, inducing double-strand breaks and triggering the DNA damage response (Landry et al., 2011). This process was further shown to be dependent on the specific glycosylase associated with base excision repair (BER), uracil-N-glycosylase (UNG) (Landry et al., 2011). More direct evidence of cytosine deamination on host DNA

was shown, where mitochondrial DNA amplified from peripheral blood mononuclear cells expressing APOBEC3A contained evidence of C>T/G>A transitions (Suspene et al., 2011). Similar hyper-editing was demonstrated in nuclear DNA from ung-/- human cell lines. Thus, APOBEC3A has at least been shown to have a direct effect on human mitochondrial genomes as well as nuclear DNA *in vitro*, including generating double-strand breaks.

The APOBEC DNA-editing enzymes leave distinctive mutational marks. A predilection for C>T transitions at TpC and CpC context was demonstrated *in vitro* in cell lines induced to over-express APOBEC3A (Suspene et al., 2011). The degree of editing was much greater in patients lacking the uracil DNA-glycosylase gene indicating that the observed levels of editing reflected the equilibrium between APOBEC3 deamination and excision by the glycosylase.

e) APOBEC4

APOBEC4 was inferred from informatic approaches given the orthologs which were identified in other mammals, chicken and frog species (Rogozin et al., 2005). It is expressed exclusively in the testis and no nucleotide editing signature is yet known.

iii. Adenine to hypoxanthine

At a deamination rate of one tenth the rate of cytosine deamination, adenine is capable of deaminating very slowly to hypoxanthine (Karran and Lindahl, 1980). The product pairs preferentially with cytosine during replication and can give rise to A>G/T>C transitions (Lindahl, 1993).

1.3.1.3 Replication errors

The size of the human genome, at ~$3X10^9$ nucleotides, makes even the smallest error rate during DNA synthesis potentially result in many mutations. During DNA synthesis, DNA polymerases use a template DNA strand to select nucleotides for incorporation into the nascent strand, whether it is in the context of DNA replication or synthesis associated with DNA repair. Replication mismatches are generated on the nascent strand by DNA polymerases during replication and DNA repair [reviewed (McCulloch and Kunkel, 2008)]. High-fidelity B family DNA polymerases, Pol δ and ε, have an error rate of one in 10^7 for every nucleotide synthesised due to intrinsic proofreading properties. The post-replicative mismatch repair pathway is thought to reduce that error rate one hundred-fold to one in 10^9 [reviewed (McCulloch and Kunkel, 2008)]. There exists a collection of low-fidelity error-prone polymerases that are able to replicate damaged DNA templates. These translesion polymerases have a higher error rate because they lack proof-reading capacity and are poor discriminators of mismatched, non-fitting nucleotides (error-rate 10^{-4} to 10^{-1}) [reviewed (Sale et al., 2012)]. Although they are thought to synthesise only very short stretches of DNA, this implicates internal replication machinery as a source of mutagenesis. Indeed, in order to avoid replication fork collapse, translesion polymerases are crucial in allowing completion of replication at the cost of errors which may be fixed later by excision repair pathways. This is called DNA damage tolerance and is extensively reviewed (Klarer and McGregor, 2011; Knobel and Marti, 2011; Sale et al., 2012).

An additional factor which affects the likelihood of nucleotide misincorporation by replicative DNA polymerases is the balance of the cellular dNTP pool. Perturbations of the dNTP pool can lead to insertion-deletion loops, erroneous base incorporation and can affect proofreading efficiency (Roberts and Kunkel, 1988) and be another source of replication-related errors.

The spectrum of base mismatch generated by replication errors is varied. There is a propensity for certain sequence motifs. For example, microsatellites are prone to “slippage” with one strand creating a loop which may lead to deletions or insertions (indels) in new replicated DNA [reviewed (Eckert and Hile, 2009)]. Here, the signature is one of small indels occurring at microsatellite repeat tracts. This signature however, can also be attributed to failure of mismatch repair which performs as a safety net in replication, and will be dealt with in a separate section (section 1.4.3). Otherwise, no precise sequence motif is known to be associated with replication errors, although the final *spectrum* of mutations may be determined by the specificity of the translesion polymerase involved.

1.3.2 Physical agents

1.3.2.1 Ionizing radiation

Ionizing radiation is radiation composed of particles that can liberate an electron from an atom or molecule, producing *ions* or atoms/molecules with a net electric charge. These are highly chemically reactive, and the reactivity produces significant biological damage per unit of energy of ionizing radiation. This particularly injurious type of radiation includes electromagnetic radiation, comprising γ rays, X-rays and some ultraviolet radiation on the high-frequency and short wavelength end of the spectrum, or particle radiation (α- and β-) (Friedberg et al., DNA repair and mutagenesis, 2nd edition). Ionizing radiation deposits its energy directly on DNA with the potential to cause loss of a base, fragmentation of the sugar ring and strand breaks, often creating non-ligatable ends. As such, the best-described signature of direct ionizing radiation is the generation of double strand breaks (Friedberg et al., DNA repair and mutagenesis, 2nd edition). However, ionizing radiation can also indirectly produce excited or ionised biological molecules, such as reactive oxygen species, which can be damaging to nucleotides, and will be discussed later.

1.3.2.2 Non-ionizing radiation

Lower-energy radiation, such as visible light, infrared, microwaves, and radio waves, are not ionizing. This low-energy non-ionizing radiation may damage molecules, but the effect is generally indistinguishable from the effects of simple heating. Such heating does not produce free radicals unless extremely high temperatures are attained. However, there is a degree of overlap between ionizing radiation and the lower ultraviolet spectrum that contains a range of molecularly-damaging radiation that is not ionizing, but has somewhat similar biological effects (Friedberg et al., DNA repair and mutagenesis, 2nd edition).

Non-ionizing ultraviolet radiation carrying enough energy to excite molecular bonds in DNA molecules can form cyclobutane pyrimidine dimers (CPDs) and (6-4) pyrimidine-pyrimidone photoproducts [(6-4)PPs]. The signature that is associated with ultraviolet light damage is C>T/G>A transitions or CC>TT/GG>AA double nucleotide transitions (Pfeifer et al., 2005).

Figure 1.3

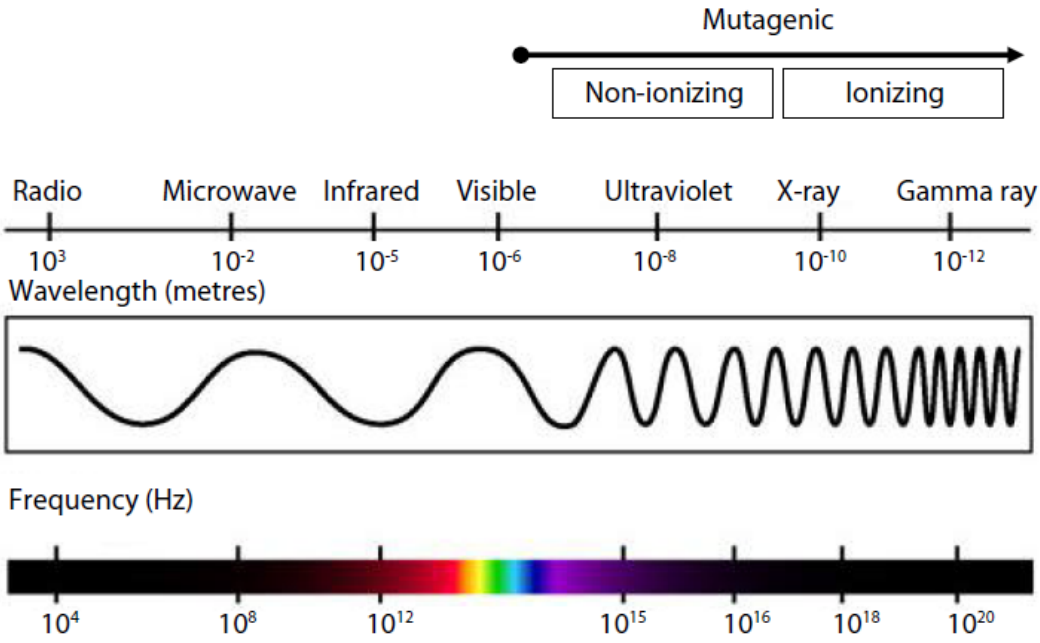


Figure 1.3: The range of mutagenic radiation in the electromagnetic spectrum

1.3.3 Free radical species

Free radical species include reactive oxygen species as well as reactive nitrogen oxide species. However, for the purposes of a description in this thesis, the following section will concentrate on reactive oxygen species. Reactive oxygen species are a type of free radical generated by cellular exposure to exogenous agents such as ionizing radiation, chemicals and metals as well as exposure to endogenous by-products of normal cellular metabolism, including apoptosis and the inflammatory response (Hussain et al., 2003). Irrespective of their origin, free radical species can interact with cellular molecules like DNA leading to a variety of modifications. One of the commonest or most well-studied oxidative DNA lesions of reactive oxygen species is 8-oxo-2'-deoxyguanosine (8-oxo-dG), although over 25 different oxidative DNA base lesions have been described (Evans et al., 2004). It is not possible to consider the multitude of oxidative DNA base lesions exhaustively here, although there are notable oxidative lesions worthy of mention. Cyclopurines, generated by hydroxyl radicals, are characterised by a covalent bond between the purine and the sugar moiety of the sugar-phosphate backbone resulting in bulky distortion of the double helix. Lipid peroxidation has also been known to yield a highly reactive product, malondialdehyde, which can also form bulky DNA adducts on guanine (Frosina et al., 1996; Voulgaridou et al., 2011).

The consequence of DNA interaction with reactive oxygen species include the generation of abasic sites, single-strand DNA breaks, deaminated bases and adducted bases (Hori et al., 2011). As such, although oxidative base lesions are predominantly repaired by base excision repair, the characteristics of some oxidative lesions, like cyclopurines and reactive by-products of malondialdehyde, challenges the effectiveness of base excision repair and poses the perfect substrate for nucleotide excision repair (Robertson et al., 2009; Slupphaug et al., 2003). Furthermore, two or more oxidative DNA lesions present within 10 base pairs of each other are termed oxidative clustered DNA lesion (OCDL) and can be more difficult to resolve (Eot-Houllier et al., 2005). These oxidative DNA lesions can also lead to secondary double-strand break formation (Bonner et al., 2008).

There is such a wide variety of potential oxidative DNA lesions that it is difficult to isolate any particular signature due to reactive oxygen species. However, 8-oxo-G has been shown to favour hydrogen-bonding with A which gives rise to G>T:C>A transversions upon replication across an uncorrected lesion. Furthermore, a specific sequence context has been associated with some oxidative damage. Evidence for DNA damage at site-specific GGG sequence by oxidative stress was shown in the context of telomere shortening in senescence (Oikawa and Kawanishi, 1999). Amino-1-methyl-6-phenylimidazo [4,5-*b*] pyridine (PhIP), a heterocyclic amine isolated from cooked meats, and known to generate increased 8-hydroxy-2'-deoxyguanosine (8-OH-dG) in rat mammary gland when administered orally (El-Bayoumy et al., 2000), was shown to cause site-specific oxidative damage to the 5' end guanine at GG and GGG sequences in a study using *HRAS* and *TP53* reporter assays (Oikawa et al., 2001).

1.3.4 Chemical agents

1.3.4.1 Oestrogens can form DNA adducts as well as generate oxidative DNA damage

An endogenous exposure that is somewhat overlooked but for which there exists a large body of epidemiologic evidence linking exposure and cancer incidence is oestrogen. Oestrogens are thought to have two roles in the induction of cancer: stimulating proliferation of cells by receptor-mediated processes, and generating electrophilic species that can covalently bind to DNA. The latter role is thought to proceed through catechol oestrogen metabolites, which can be oxidised into intermediates that bind to DNA. Stable oestrogen adducts can be formed through these 2,3-quinone oxidative species (Spencer et al., 2012), cause bulky distortion of the genome and are ideal candidates for nucleotide excision repair. Conversely, 3,4-quinone intermediates produce guanine adducts prone to depurination and base excision repair.

More recently, the capacity of the endogenous oestrogen, 17β-oestradiol, as well as the more well-studied equine oestrogen formulations in hormone replacement drugs, equilenin and equilin, to induce oxidatively generated DNA damage was demonstrated. This oxidatively generated DNA damage is believed to be the product of the attack of free radicals on DNA, rather than direct adduct formation (Spencer et al., 2012).

1.3.4.2 Alkylating agents

DNA contains several nucleophilic centers that are susceptible to attack from electrophilic agents resulting in alkylation. In particular, ring nitrogens are particularly susceptible as nucleophilic centers and alkylation-reactions and some of the positions most prone to attack are N7 in guanine (N^7G) and N3 in adenine (N^3A) (Figure 1.1b) (Denny, 2001).

Many alkylating agents are present as environmental compounds as well as intermediates of normal metabolism (Figure 1.3B). Monofunctional alkylating compounds such as methyl methane sulfonate (MMS), methyl nitrosurea (MNU) and ethyl nitrosurea (ENU) are directly-acting and can bind covalently to one site in DNA (Eisenbrand et al., 1986). In contrast, nitrosamines are not directly-acting and require activating by the P450 enzymes in the liver. Furthermore, bifunctional alkylating compounds such as mustard compounds contain two reactive centers and can therefore create highly cytotoxic inter-strand and intra-strand crosslinks (Hartley et al., 1988). As such, these make effective chemotherapeutic agents and have been developed as such (e.g. cyclophosphamide). In fact, the effects of such chemotherapeutic agents have been documented in a screen of protein-

kinase genes of gliomas which had been treated with alkylating agents, demonstrating a marked C>T/G>A predominance of mutations (Greenman et al., 2007).

1.3.4.3 Platinum-based compounds

Platinum-based compounds are used as chemotherapeutic agents in cancer. Platinum compounds have a propensity to bind DNA to cause bulky adducts, inter-strand and intra-strand crosslinks (Hofr et al., 2001; Knox et al., 1987). At present, no mutation signature has been documented with these compounds.

1.3.4.4 Poly-aromatic hydrocarbons

Benzo(a)pyrene is an example of a poly-aromatic hydrocarbon (PAH) class of tobacco smoke-related carcinogen. Compounds such as these are able to form bulky adducts particularly on guanines generating a signature of G>T/C>A transversions with a predilection for endogenously methylated CpG dinucleotides in *TP53* reporter studies (Pfeifer et al., 2002). In a genome-wide analysis of a smoking-related small cell lung cancer, G>T/C>A transversions were the dominant mutation type, also demonstrating a lack of mutations on the transcribed strand. This strand bias was attributed to the past activity of transcription-coupled repair in operation, a repair pathway known to remove tobacco-smoke related bulky adducts (Pleasance et al., 2010b).

1.3.4.5 Psoralens

Psoralens are a type of phototherapy agent used for inflammatory conditions like psoriasis. These compounds are found naturally in the environment. When exposed to ultraviolet light, psoralens bind covalently to nucleotide bases where they can form bulky monoadducts as well as inter-strand crosslinks (Chiou and Yang, 1995). Mutational spectra at endogenous *HPRT* reporter loci in studies of human lymphoblasts treated with psoralens and phototherapy revealed a high level of base substitutions with a preference for pyrimidines at a TpA dinucleotide sequence (Papadopoulo et al., 1993; Yang et al., 1994). Furthermore, more base substitutions were found in the non-transcribed strand of the HPRT gene suggesting that DNA distorting psoralen photolesions were preferentially removed from the transcribed strand (Laquerbe et al., 1995).

1.3.4.6 Intercalating agents

Intercalating drugs such as the antibiotic class which includes daunorubicin and actinomycin-D are able to slot between two DNA strands essentially blocking DNA synthesis [reviewed (Chaires, 1990, 1998)]. Such DNA perturbations are likely to block replication. No mutation signature has been assigned to this class of compound.

1.3.5 Summary of mutational processes

DNA is under a constant stream of attack from a variety of DNA damaging agents. Whether the DNA damaging agent causes direct or indirect damage to DNA, the mutagenic effect is often a biochemical conversion with a secondary stoichiometric consequence. Mutagenic effects may be the same even for different primary insults, and a summary of such mutagenic effects is shown in Figure 1.4.

Fundamentally, correct base pairing is integral to the structural and functional properties of DNA. The base moieties of nucleotides point inwards, or towards the axis of the double helix, there they come to lie within hydrogen-bonding distance of each other. Watson-Crick base-pairing follows the canonical rule of A pairing with T and C pairing with G, forming 2 and 3 hydrogen bonds respectively. The complementary nature of these interactions ensures that DNA strands are mirror-image replicas of each other, providing a template for faithful duplication of the genome as well as a source for accurate maintenance of the genome.

Base-base mismatches affect hydrogen-bonding to different extents and can affect the helical structure of DNA. Furthermore, additional and missing nucleotides can lead to one or more nucleotides being unpaired and form a small insertion/deletion loop. Finally, chemical modification of bases may alter hydrogen-bonding potential and therefore confer partiality to different bases. For example, 8-oxo-G tends to rotate the damaged base around the glycosidic bond, making bonding with A more favourable than C (Wang et al., 1998).

Figure 1.4

A

Spontaneous or enzymatic conversions	Reactive oxygen species	Physical agents	Chemical agents
• spontaneous AP sites • 5meCpG deamination • C->U deamination • A->hypoxanthine deamination • replication errors		• Ionising radiation • Non - ionising radiation	• PAH • Oestrogens • Alkylating agents • Psoralens • Platinum compounds • Intercalating agents

B

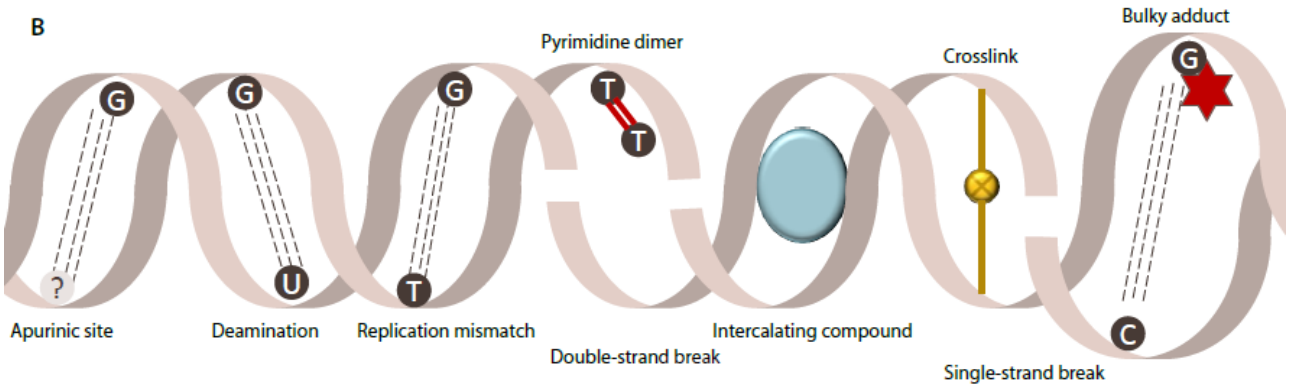


Figure 1.4: (A) Classification of DNA damaging agents in this thesis. (B) Mutagenic effects of various DNA damaging agent.

However, the cell has developed a repertoire of repair mechanisms in order to maintain genomic integrity, in the face of a constant barrage of endogenous and exogenous damaging agents that can generate an array of potential mutagenic changes. In the following section, the various DNA repair processes known to be involved in correcting many of these DNA lesions will be discussed.

1.4 DNA REPAIR PROCESSES AND ASSOCIATED CHARACTERISTIC SIGNATURES

A vast literature exists documenting what is understood regarding both prokaryotic and eukaryotic repair pathways. In this thesis, it will not be possible to exhaustively describe all such repair pathways in all organisms (nor to quote all references). Therefore, a brief description focusing where possible on higher eukaryotes will be provided in each section, mainly in order to build a framework for each repair pathway and understand how each leaves its molecular mark on a genome, whether it is working correctly or has turned awry.

1.4.1 Base excision repair

DNA damage arising from a variety of sources including oxidative damage, alkylation and deamination events can cause non-Watson-Crick base-pairing. These non-canonical base-pairing situations call upon the core base excision repair pathways in order to maintain genomic integrity. Many of the mechanistic details regarding base excision repair have been extensively reviewed elsewhere (Robertson et al., 2009). Briefly, the key steps in humans, begins with the recognition of a damaged base by the appropriate and relatively specific DNA glycosylase that recognizes, hydrolytically cleaves and removes the altered base, giving rise to an abasic site. The abasic site is then processed by an apurinic/apyrimidinic (AP) endonuclease (APE1), which incises the DNA strand 5' to the abasic sugar. DNA polymerase β (POLB) catalyses the elimination of the 5'-deoxyriboso-phosphate residue, then fills the one-nucleotide gap. Finally, the nick is sealed by the DNA ligase III/XRCC1 complex in what is termed short-patch base excision repair (Figure 1.5).

An alternative within short-patch base excision repair involves bifunctional DNA glycosylases which contain intrinsic AP lyase activity that process oxidative DNA lesions and incises abasic sites 3' to the abasic sugar leaving a 3'(2,3-didehydro-2,3-dideoxyribose) terminus that is then removed by AP endonuclease (Demple and DeMott, 2002). As in the main pathway, the gap is filled by DNA polymerase and the nick is sealed by DNA ligase. Overall, short-patch base excision repair accounts for 80–90% of all base excision repair.

Long-patch base excision repair, which replaces 2–10 nucleotides of DNA, is utilised when an oxidised lesion is refractory to the AP lyase activity of DNA polymerase β. Long-patch base excision repair is dependent on the co-factor proliferating cell nuclear antigen (PCNA) and flap structure-specific endonuclease 1 (FEN1) enzyme and DNA synthesis is thought to be mediated by several DNA

polymerases including polymerases β, δ and ε (Frosina et al., 1996). The decision whether to proceed with short or long-patch repair is human cells is not understood (Figure 1.5).

DNA glycosylases crucially recognise specific lesions and excise them from the genome, hence initiating base excision repair (Robertson et al., 2009). There is an extensive list of known mammalian DNA glycosylases in base excision repair. Multiple mutation signatures associated with engineered defects of certain glycosylases in various experimental systems have been documented and are listed in Table 1.1.

Figure 1.5

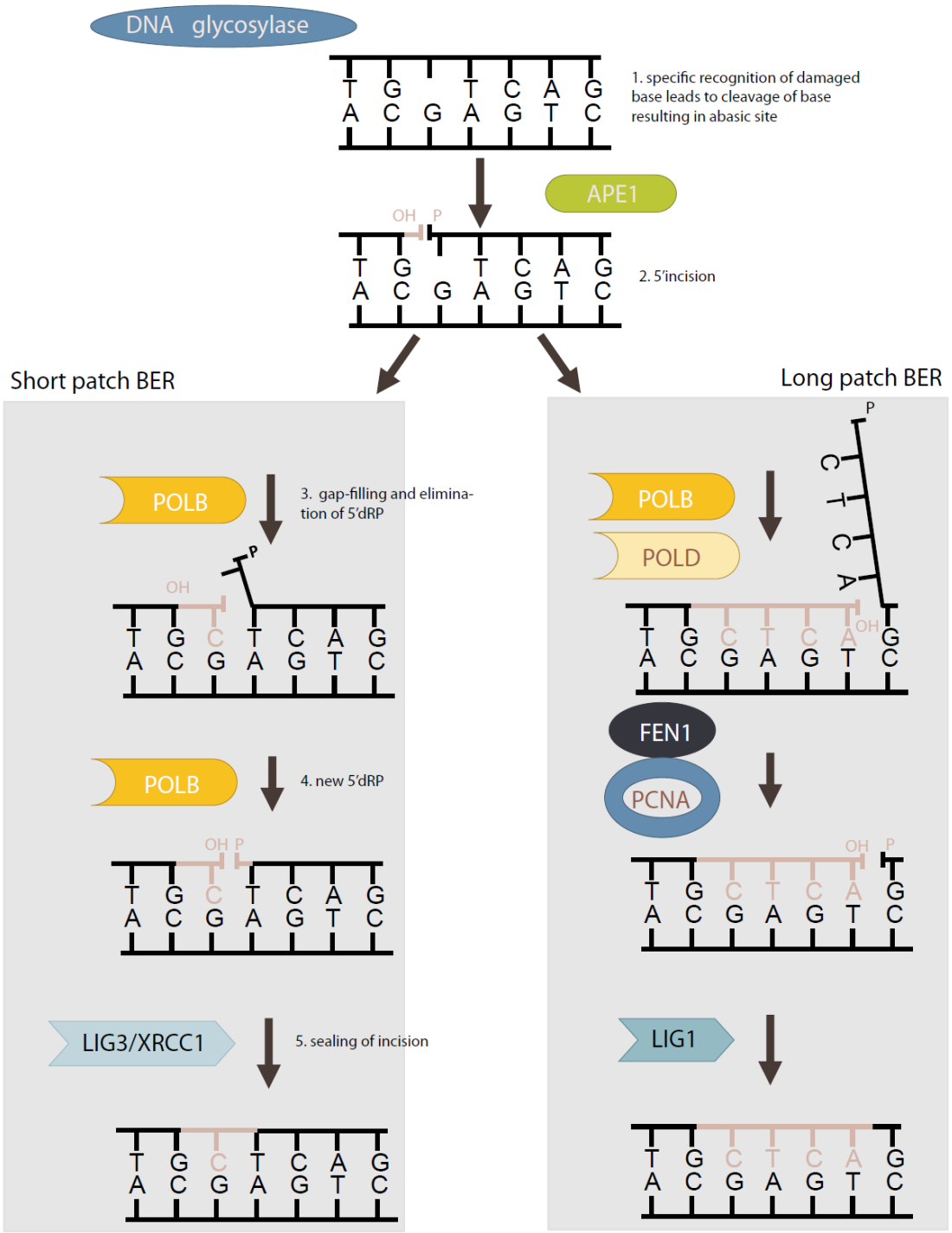


Figure 1.5: An outline of short-patch versus long-patch base excision repair (BER). BER begins with the recognition of a damaged base by a DNA glycosylase and removes the altered base, giving rise to an abasic site. The abasic site is then processed by an apurinic/apyrimidinic (AP) endonuclease (APE1), which incises the DNA strand 5' to the abasic sugar. POLB catalyses the elimination of the 5'-deoxyriboso-phosphate (5'dRP) residue, then fills the one-nucleotide gap. Finally, the nick is sealed by the DNA ligase III/XRCC1 complex. Long-patch BER replaces 2–10 nucleotides of DNA is dependent on the co-factor proliferating cell nuclear antigen (PCNA) and flap structure-specific endonuclease 1 (FEN1) enzyme and DNA synthesis is thought to be mediated by several DNA polymerases including polymerases β, δ and ε.

1.4.2 Nucleotide excision repair

Nucleotide excision repair (NER) is a non-specific repair process which is activated upon sensing of bulky DNA distortion caused by biochemical DNA modifications (Nouspikel, 2009). These biochemically-driven distortions include bulky adducts, such as exogenously occurring benzo[(a)pyrenes, aromatic amines compounds like aflatoxin and nitrosamines like MNNG as well as endogenously generated by-products like malondialdehyde and cyclopurines, modifications due to chemical compounds by platinum-based compounds, nitrogen mustards and psoralens, and non-chemical induced covalent modifications like UV-induced lesions (cyclobutane pyrimidine dimers (CPDs) and (6-4) pyrimidine-pyrimidone photoproducts [(6-4)PPs]) (Nouspikel, 2009).

Nucleotide excision repair is well-understood in mammalian cells (Aboussekhra et al., 1995). Firstly, distortion of the double-helical structure by biochemical modifications is sensed by the XPC protein complex, comprising XPC, HR23B and centrin 2. This results in the opening of a denaturation bubble around the damaged base via the TFIIH complex, which comprises no less than ten subunits with known and unknown functions. The damaged strand is incised at the 5'end by the XPF-ERCC1 complex and the 3' end by XPG endonuclease resulting in an oligonucleotide gap of approximately 25-30 nucleotides in length. This gap is filled in by DNA polymerase δ or DNA polymerase ε in association with the PCNA sliding clamp, and the nick is sealed by DNA ligase III or DNA ligase I in replicating cells, in association with XRCC1. In replicating cells, this series of steps is termed global genome repair (GGR) and occurs throughout the genome. However, a particular class of nucleotide excision repair exists that is coupled to transcription, called transcription-coupled repair (TCR) (Nouspikel, 2009).

In transcription-coupled repair, DNA lesion sensing is believed to be due to stalling of RNA polymerase II (RNAPII). Apart from this, repair proceeds in the same way as described for global genome repair (Figure 1.6). A consequence of transcription-coupled repair is that DNA damage on the transcribed strand is repaired more efficiently than damage on the non-transcribed strand. Thus, fewer mutations accumulate on the transcribed strand.

A less well-described phenomenon in nucleotide excision repair involves proficient repair of the non-transcribed strand of genic regions in cells where global genome repair is attenuated (Nouspikel and Hanawalt, 2000). This repair of the non-transcribed strand cannot be attributed to transcription-coupled repair which does not maintain the non-transcribed strand, includes regions of a gene that is not reached by RNAPII for which transcription-coupled repair is dependent and although is dependent on XPC, an integral feature of lesion-sensing in global genome repair, is not dependent on

CSB, a component crucial to transcription-coupled repair (Barnes et al., 1993). As such, this under-studied mechanism has been termed transcription domain-associated repair (DAR) and describes the persistence of pockets of repair akin to global genome repair, which does not discriminate between strands and occurs within sites of transcription, described as “transcription factories”(Nouspikel and Hanawalt, 2000). However, in genome-wide mutation analysis, evidence of preferential repair of actively or heavily transcribed regions particularly in the absence of bias between the transcribed or non-transcribed strands may be the indication of transcription domain-associated repair in operation.

The activity of transcription-coupled nucleotide excision repair in particular is one that has been well-described in the literature (Nouspikel, 2009). For example, DNA damage induced by short wavelength ultraviolet light can cause cyclobutane pyrimidine dimers (CPDs) and (6-4) pyrimidine-pyrimidone photoproducts [(6-4) PPs] which are ideal substrates for nucleotide excision repair. A genome-wide analysis of a malignant melanoma cell line, COLO-829, uncovered a strand bias where fewer C>T/G>A transitions were seen on the transcribed strand ($P < 0.0001$). This strand bias was attributed to preferential repair of the ultraviolet-induced pyrimidine dimers that underlie C>T /G>A mutations on the transcribed strand (Pleasance et al., 2010a). The results are therefore consistent with transcription-coupled nucleotide excision repair being operative on ultraviolet-light-induced DNA damage in COLO-829. Hence, strand bias of mutations within the genomic footprint may be an indicator or a signature of the past operation of transcription-coupled repair.

Other descriptions of strand bias have been attributed to transcription-coupled repair. However, the possibility remains that there exist other, currently uncharacterised forms of transcription-related DNA repair or transcription-related DNA damage processes.

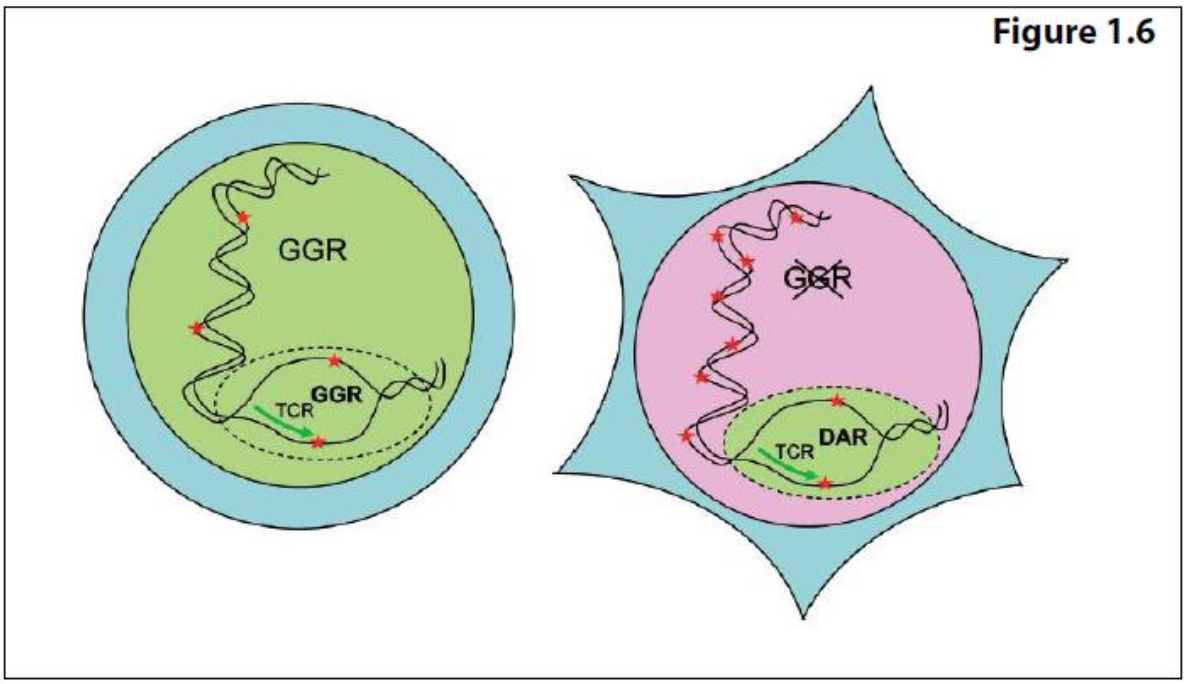


Figure 1.6: An outline of nucleotide excision repair (NER). Global genome repair (GGR) occurs in replicating cells throughout the genome. Transcription-coupled repair (TCR) occurs preferentially within transcription factories (dashed oval). In fully differentiated cells, GGR is down-regulated but the activity of TCR remains within transcription factories. The alternative domain associated repair (DAR) has been postulated to continue NER activity in fully-differentiate cells without regard to transcriptional strands. Figure adapted from Nouspikel et al 2009.

1.4.3 Mismatch repair

The mismatch repair system recognizes and repairs misincorporated bases as well as erroneous insertions/deletions that arise during DNA replication and DNA recombination repair activity [extensively reviewed (Jiricny, 2006; Pena-Diaz and Jiricny, 2012)]. The correction of the mismatches involves a series of steps that vary from one organism to another. The archetypal *Escherichia coli* mismatch repair pathway has been extensively studied and is well characterised. Thus, *E. coli* mismatch repair will be used as a framework for the rest of this description. First it is necessary to distinguish the two parental strands from the newly-synthesised daughter strand which contains the aberration. This is achieved by transient hemi-methylation where the parental strand is methylated at dGATC sequences and the nascent strand is not. The exact mechanism for distinguishing the strands is not clear in other organisms (Figure 1.7).

In *E. coli*, a series of Mut proteins is required to complete MMR. MutS forms a dimer, $MutS_2$, which recognises the mismatched base on the daughter strand and binds the mutated DNA. MutH binds

hemi-methylated sites along the daughter DNA, but is only activated upon contact with a MutL dimer ($MutL_2$) which binds the MutS-DNA complex. $MutL_2$ acts as a mediator between $MutS_2$ and MutH, activating the latter. MutH nicks the daughter strand near the hemi-methylated site and recruits a UvrD helicase (DNA Helicase II) to separate the two strands with a specific 3' to 5' polarity. The MutSHL complex slides along the DNA strands in the direction of the mismatch, liberating the strand to be excised as it goes. An exonuclease trails the complex and digests the single-stranded DNA tail. The exonuclease recruited is dependent on which side of the mismatch MutH incises the strand – 5' or 3'. If the nick made by MutH is on the 5' end of the mismatch, either RecJ or ExoVII (both 5' to 3' exonucleases) is used. If however the nick is on the 3' end of the mismatch, ExoI (a 3' to 5' enzyme) is used. The entire process ends past the mismatch site - i.e. both the site itself and its surrounding nucleotides are fully excised. The single-stranded gap created by the exonuclease can then be repaired by DNA Polymerase III (assisted by single-strand binding protein), which uses the other strand as a template, and finally sealed by DNA ligase. Dam methylase then rapidly methylates the daughter strand (Figure 1.7).

In humans, the MSH proteins are heterodimeric orthologs of MutS. MSH2 dimerizes with MSH6 or MSH3 to form two complexes MutSα and MutSβ respectively and perform similar functions to MutS in mismatch recognition and initiation of repair. There is no known MutH-type function or DNA helicase identified in eukaryotic cells. However, homologs of bacterial MutL do exist, and they do form heterodimers. hMLH1 heterodimerizes with hPMS2 and hPMS1 or hMLH3 to form MutLα, MutLβ and MutLγ complexes respectively. Whilst MutLα is involved in general mismatch recognition and nucleolytic processing, MutLγ in involved in IDL repair, whilst nothing is known regarding MutLβ. Eukaryotic organisms also require additional factors including PCNA and replication factor C (RFC) which plays a critical role in 3' nick-directed MMR involving EXO1 (Kadyrov et al., 2006).

Because MMR reduces the number of replication-associated errors, defects in MMR increase the spontaneous mutation rate (Tiraby et al., 1975). Inactivation of MMR in human cells is associated with hereditary and sporadic human cancers (Lynch and de la Chapelle, 1999), and the MMR system is required for cell cycle arrest and/or programmed cell death in response to certain types of DNA damage (Stojic et al., 2004). Mutations in the human homologues of the Mut proteins affect genomic stability, which result in microsatellite instability (Shibata et al., 1994). In particular, the overwhelming majority of hereditary non-polyposis colorectal cancers (HNPCC) are attributed to mutations in the genes encoding the MutS and MutL homologues MSH2 and MLH1 respectively (Lynch and de la Chapelle, 1999). The signature of insertions/deletions on a background of MMR deficiency is highly reproducible in experimental systems (Kuraguchi et al., 2000).

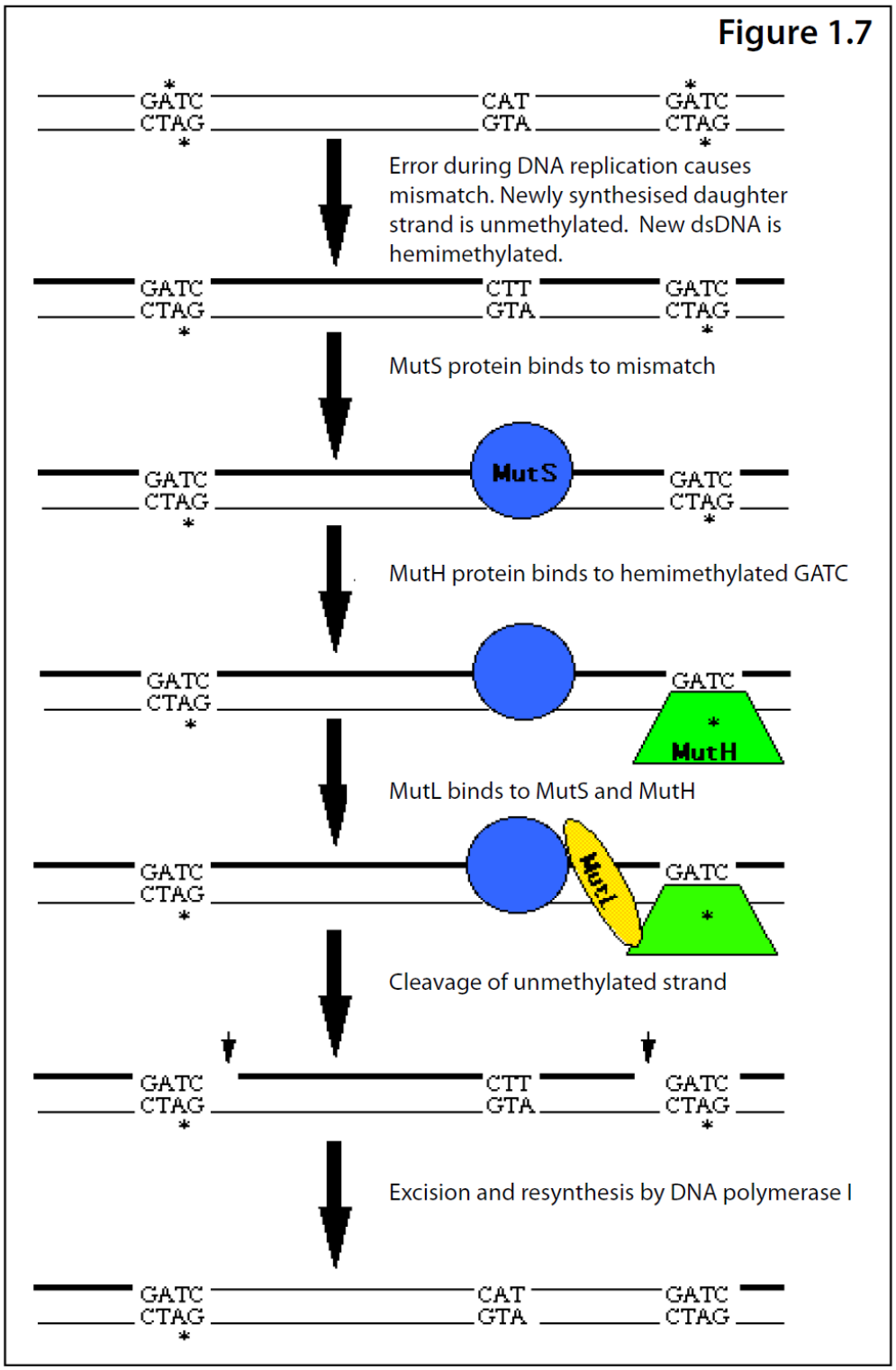


Figure 1.7: The key steps involved in mismatch repair which is able to discriminate between newly-synthesised daughter strand and the parental strand. This ensures that the newly-synthesised strand is preferentially repaired by this key pathway. Image taken from Maloy laboratory, San Diego State University, with minor adaptation.

1.4.4 Double-strand break repair

A single double-strand break is able induce cell death making it one of the most harmful DNA lesions in cells. Consequently, efficient repair mechanisms for double-strand breaks have evolved which can occur via two main pathways: non-homologous end-joining and homologous recombination. In the following sections, these pathways and the mutational signatures they leave in the genome will be discussed.

1.4.4.1 Non-homologous end-joining

Non-homologous end-joining (NHEJ) repairs double strand breaks by re-ligating two broken ends with no prior requirement for homologous sequence. NHEJ is thought to seek minimum base pairing of less than four bases in yeast, generating overhangs which increases the efficiency of repair (Daley and Wilson, 2005).

The core NHEJ machinery is composed of three complexes: MR(X)N, Ku and the DNA ligase complexes. MR(X)N and Ku complexes are believed to bind to double-strand breaks shortly after double-strand break formation, bridging and tethering the two broken ends and inhibiting degradation. They also recruit, stabilise and stimulate the ligase complexes. Following this, different alignments and base pairing overhangs take place and ligations attempted. If end-processing is required, the Ku and ligase complexes are able to recruit a large number of DNA-modifying enzymes, reattempting alignment and ligation until successful, demonstrating that non-homologous end-joining is a highly dynamic process (Friedberg et al., DNA repair and mutagenesis, 2nd edition).

MR(X)N comprises Rad50/RAD50, Mre11/MRE11 and Xrs2/NBS1 proteins in yeast/humans and is essential for tethering DSB ends together and recruiting the ligase complex. The Ku heterodimeric complex comprises yKu70/KU70 and yKu80/KU80. In vertebrates, Ku is part of a larger complex called DNA-dependent protein kinase (DNA-PK) which has a catalytic subunit (DNA-PKcs) with end-bridging capacity similar to that of MRX in yeast, perhaps explaining the redundancy of MRN in vertebrates which is not involved in NHEJ. Ku binds double-stranded DNA and makes contact with ligases and is thought to stabilise DNA ends preventing 5’ resection associated with HR. The NHEJ ligase complex comprises Lig4/Ligase IV and requires obligatory cofactor Lif1/XRCC4 and Nej1/XLF to perform ligation. However, incompatible double strand break ends may require some processing prior to ligation. In the presence of Ku, the NHEJ ligase complex has enormous flexibility allowing mismatch correction, gap-filling or removal of non-ligatable ends prior to NHEJ proceeding.

Mutational signatures associated with NHEJ include a preponderance of microhomology at junctional sequences involved in uniting two ends.

1.4.4.2 Microhomology mediated end-joining

A double strand break repair pathway using microhomology of approximately 5-20 nucleotides, observed in the absence of some core non-homologous end-joining factors and generating larger deletions has been termed microhomology-mediated end-joining. It appears to require some NHEJ factors (MRX, Ku, Lig4) and some factors associate with homologous recombination (MRX, Rad1-Rad10, Rad52). Little else is known about this pathway which has been based largely on experiments in *Saccharomyces cerevisiae*, apart from one study performed in chinese hamster ovary cells (Pulciani et al., 1982).

The mutational signature associated with microhomology-mediated end-joining is likely to be very similar to that of non-homologous end-joining. However, it is possible that the microhomologous sequences may be longer.

1.4.4.3 Homologous recombination

Classical HR requires three successive steps. First, resection of the 5'strand at the break ends, second, strand invasion into a homologous DNA duplex and strand exchange and third, resolution of recombination intermediates. It is within this third step of resolution of recombination intermediates where homologous recombination has further subgroups: synthesis-dependent strand annealing (SDSA), classical double-strand break repair (DSBR), break-induced replication (BIR) and single-strand annealing (SSA). Here, the shared initial steps in homologous recombination will be discussed first, concentrating on what is known regarding repair in mammals. This will be followed by a brief introduction into a reduction of the different subtypes of homologous recombination (Freidberg et al., DNA repair and mutagenesis, 2nd edition).

Following the occurrence of a double-strand break, the MR(X)N complex performs multiple functions including a checkpoint signalling role, double-strand break end tethering and nucleolytic cleaving. Efficient resection of the 5'ends at the double-strand break requires Sae2/CtIP and Exo1/EXO1 to

generate a 3’single-stranded DNA end that is competent for searching a homologous template and performing invasion. 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 or hybrid DNA by strand exchange. These reactions are mainly achieved by a nucleoprotein filament comprising the 3’single-stranded end coated with Rad51/RAD51 recombinase protein. Rad51/RAD51 loading is dependent on RPA which interacts with Rad52/BRCA2. Whilst these steps are a common pre-requisite for repair by homologous recombination, the final stages of resolution are subtly different.

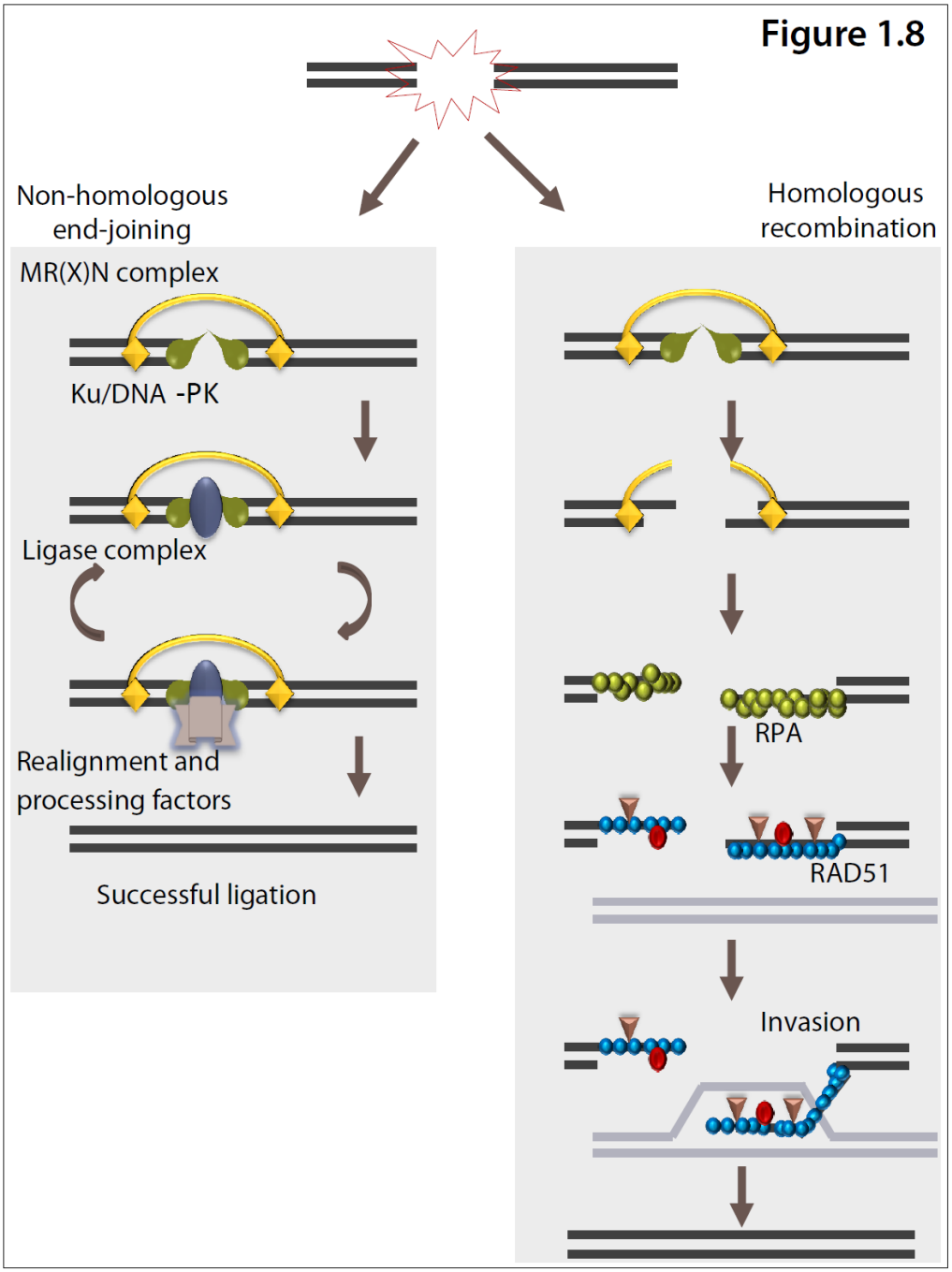


Figure 1.8: Repair of double-strand breaks: The options of non-homologous end-joining (NHEJ) or homologous recombination. In NHEJ, MR(X)N and Ku complexes are believed to bind to double-strand breaks shortly after double-strand break formation, bridging, tethering and stabilising the two broken ends, stimulating recruited ligase complexes. Different alignments and base pairing overhangs would take place and ligations attempted. If end-processing is required, the Ku and ligase complexes are able to recruit a large number of DNA-modifying enzymes, reattempting alignment and ligation until successful. In HR, MR(X)N and Ku also bind and tether the DSB ends, but further stimulate 5'end resection. The exposed 3'single-stranded end is initially coated with RPA (green circles) which enhances RAD51 (blue circles) loading generating a nucleofilament capable of strand invasion of a homologous duplex, forming a D-loop, following dependent interactions with RAD52 (brown triangles) and RAD54 (red ovals). These steps are a common pre-requisite for repair by homologous recombination. However, the final stages of resolution are subtly different and will be dealt with in Figure 1.9.

1.4.5 Synthesis-dependent strand annealing (SDSA)

The most conservative model for resolution of repair-intermediates of double-strand breaks is synthesis-dependent strand annealing. Here, the two 3' single-stranded ends of a double-strand break share homology to the repair template can engage regions of homology independently. It is thought that one end is more likely to perform invasion, forming a D-loop (see figure 1.8 for description) and performing DNA synthesis whilst extending the D-loop. Eventually, the newly-synthesised and elongated end is displaced from the D-loop. Re-annealing of the initially separated ends occurs via this newly synthesised complementary region. Synthesis-dependent strand annealing is highly faithful, does not result in crossovers and provides genome stability in mitotic cells (Nassif et al., 1994). Synthesis-dependent strand annealing is promoted by Sgs1& Srs2 helicases in yeast and BLM RecQ helicase in mammals (Wu and Hickson, 2003). A mutational signature is not associated with this highly faithful and most conservative form of double-strand break repair.

1.4.6 Double-strand break repair

An alternative model for the resolution of double-strand break intermediates involves elongation of the invasive strand and displacement of the homologous duplex strand which anneals to the second 3' end of the double-strand break. The second 3' end will also be elongated by DNA synthesis. This situation results in two branched structures called Holliday junctions. Differential ways of cleavage of these Holliday junctions can result in crossover or non-crossover products (see figure for details) by a process termed resolution. Double Holliday junction intermediates can also undergo dissolution which involves migration of the two Holliday junctions towards each other which is then unravelled by the action of DNA helicases and DNA topoisomerases. The "resolvases" are enzymes that are involved in resolving Holliday junctions by resolution or dissolution and have recently been under intense investigation. A specific mutational signature has not been attributed to this form of repair.

1.4.7 Break-induced replication

In some situations, only one end of a double strand break is available for repair, for example at telomeres that have lost their protective telomeric repeats or when a replication fork collapses. Here, the broken end invades a homologous sequence and initiates unidirectional DNA synthesis from the site of strand invasion and replicates the chromosome template for potentially very long stretches of up to hundreds of kilobases. Repeated cycles of separation and reinvasion can occur, but usually of the homologous template. This is called break-induced replication and in principle, is an accurate process that depends on recombination proteins and demands extensive homology for strand invasion. Nevertheless, it can lead to loss of heterozygosity and chromosomal rearrangements if the invading strand is paired with homologous allelic and non-allelic sequence (Smith et al., 2007). Indeed, break-induced repair-based mechanisms can explain the complexity of the chromosomal structural changes that occur in cancer cells (Smith et al., 2007). Furthermore, in a yeast model of break-induced replication, it was shown to be highly inaccurate over the entire path of the replication fork, as the rate of frameshift mutagenesis during break-induced replication was up to 2,800-fold higher than during normal replication (Deem et al., 2011). A specific and reproducible mutational signature has not been attributed to this repair mechanism.

1.4.8 Microhomology-mediated break-induced replication

A more recently elucidated pathway related to break-induced replication but dependent on a degree of microhomology-annealing is microhomology-mediated break induced repair. Although this mechanism is not very well-characterised, briefly, a one-ended double strand break attempts to pair with stretches of DNA which share microhomology with the 3’ strand of the break. The key difference with break-induced repair is that invasion can occur of completely unrelated DNA molecules as only minimal microhomology is required. Following a degree of replication, separation can occur with repeated reinvasion of other unrelated templates giving rise to complex genomic rearrangements. Microhomology-mediated break induced replication probably accounts for only a small fraction of DSB repair in yeast, whereas in mammalian cells it seems to be more efficient (Bentley *et al*, 2004). This repair mechanism is likely to show evidence of microhomology of bases at multiple adjoining bits of sequence in complex rearrangements (Lee et al., 2007).

1.4.9 Single-strand annealing

In a situation where no homologous template for repair is found, 5' to 3' end resection can extend for many kilobases. If resection uncovers direct repeat sequences, both single-stranded ends can anneal together to repair the break. This repair process is called single-strand annealing and can lead to deletions. As such, it is a potentially mutagenic pathway within homologous recombination. The expected molecular signature of single-strand annealing in operation would be loss of one DNA repeat plus the sequence located between the repeats.

Figure 1.9

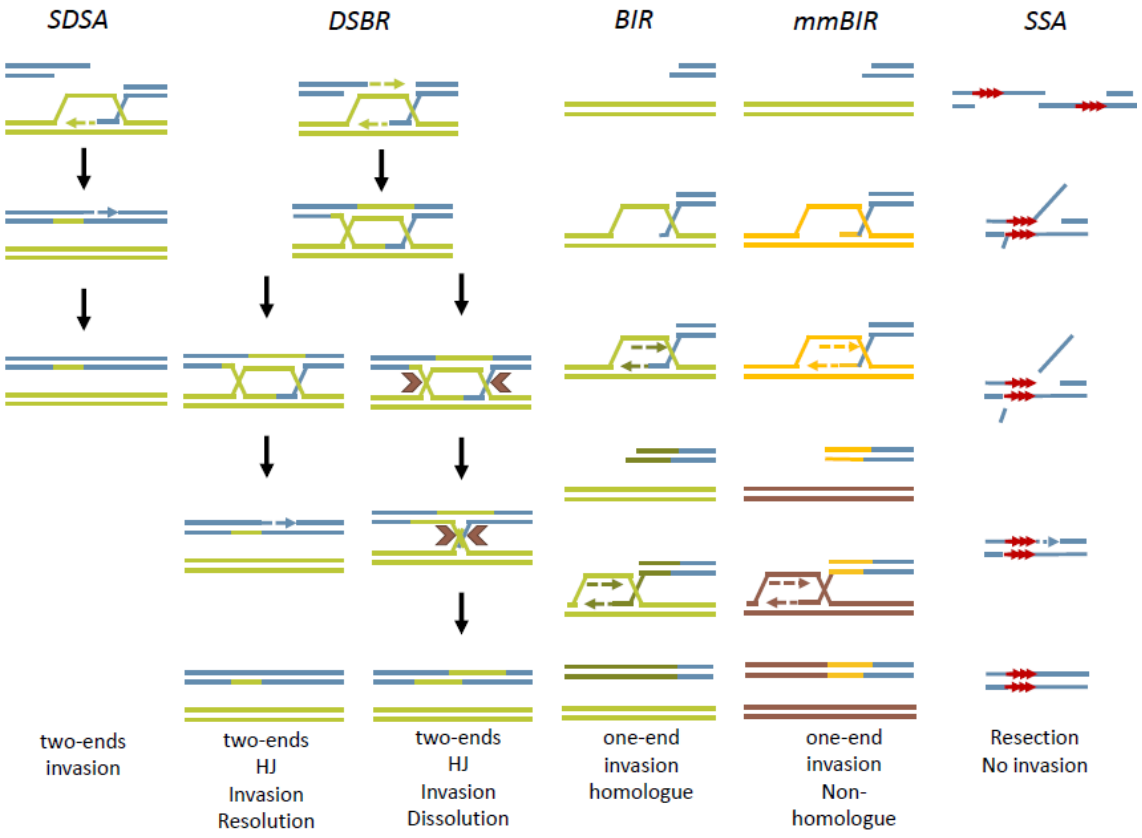


Figure 1.9: Different possible modes of resolving double-strand breaks within the homologous recombination (HR) pathway. The most conservative method of repair in HR is synthesis-dependent strand annealing (SDSA). Alternative sub-pathways differ in the following ways. If both ends of a double-strand break are available for repair, double-strand break repair (DSBR) can occur with subtle differences in resolving the Holliday junctions (HJ) resulting in resolution or dissolution. When only one end of the double-strand break is available for repair, if invasion occurs of a homologue, break-induced repair (BIR) ensues. Alternatively, invasion of non-homologous sequence could result in microhomology-mediated break-induced repair (mmBIR). If invasion does not occur, single-stranded annealing (SSA) may arise. Blue chromosomes represent broken double-stranded ends. Green, yellow and brown chromosomes represent stretches of invaded dsDNA of different chromosomes. Red arrows represent direct repeat sequences.

Table 1.1: Known mutational signatures of DNA damage and repair mechanisms			
Processes of DNA damage			**SIGNATURE**
Spontaneous or enzymatic conversions	Spontaneous generation of apurinic/apyrimidinic sites		C>T/G>A
	Deamination of bases	Methylated CpG dinucleotides	C>T/G>A at methylated CpG dinucleotides
		Cytosine to uracil deaminations AID APOBEC1 APOBEC2 APOBEC3 APOBEC4	C>T/G>A at ApC or GpC TpC context - TpC or CpC -
		Adenine to hypoxanthine	A>G/T>C
	Replication errors		
Physical agents	Ionizing radiation		Double-strand breaks
	Non-ionizing radiation		C>T/G>A or CC>TT/GG>AA
Free radical species			Mixed including OCDLs, double-strand breaks, G>T/C>A at GpG or GpGpG
Chemical agents	Oestrogens		G>X
	Alkylating agents		C>T/G>A
	Platinum-based compounds		
	Poly-aromatic hydrocarbons		G>T/C>A at methylated CpG
	Psoralens		Pyrimidine at TpA
	Intercalating agents		
Processes of DNA repair			
Base excision			C>T/G>A in defects

repair		of SMUG1; G>T/C>A for defects in OGG1
Nucleotide excision repair	Transcription-coupled repair	Strand bias with less mutations on the transcribed strand
Mismatch repair		Insertions/deletions at tandemly-repeating bases
Double-strand break repair	Non-homologous end-joining	Microhomology <= 5bp at places of double-strand breaks
	Microhomology mediated end-joining	Microhomology > 5bp at places of double-strand breaks

1.4.5 Summary of DNA repair processes

The requirement for correct base-pairing underlies the fundamental structural properties of the DNA double-helix. As described previously, base-base mismatches and gross biochemical modifications can both affect hydrogen-bonding potential resulting in a range of distortions to the double-helix. Multiple complex repair pathways exist in order to maintain genomic integrity. However, the choice of which repair system to use depends on the type of lesion and on the available template which is dependent on the cell-cycle phase of the cell.

The recruitment of repair proteins to damaged DNA is likely to involve post-translational modifications which tune the efficiency or the specificity of the repair machinery towards a certain type of lesion, facilitating repair in a specific cell-cycle phase. Regardless of the precise spatio-temporal orchestration of DNA repair, this section was dedicated to describing those repair pathways associated with damaged or non-fitting bases which are removed as a free base in base excision repair, removed as single-stranded oligonucleotides in nucleotide excision repair, removed in the context of transcription, and removed as mismatched bases by mismatch repair. In addition, repair mechanisms dealing with DNA breaks were also considered. These repair pathways were considered mainly for the mutational signatures they may leave whether operational or failing. A summary table of molecular signatures associated with the various repair pathways is provided in Table 1.1, and will be referenced through the rest of this thesis.

1.5 EXPLORING BREAST CANCER

In this thesis, twenty-one different breast cancers will be explored using whole genome sequencing in order to understand the mutagenic and repair processes that have been operative in these solid tumours. In this section, an overview of the epidemiology, classification and genetics of breast cancer will be provided, emphasising at the close, how the enormity of scale offered by second-generation sequencing technologies can assist in the detailed exploration sought in this thesis.

1.5.1 Epidemiology and risk factors in breast cancer

One of nine women in the United Kingdom will develop breast cancer in her lifetime. Breast cancer is the most common class of cancer in women worldwide, with 1.38 million new cancer cases diagnosed and it remains the most frequent cause of cancer death in women globally (Ferlay et al., 2010). The incidence of breast cancer increases with age. Recognition of risk factors has helped the identification of patients at high-risk of developing breast cancer and who may benefit from intense monitoring and allow modification of lifestyle factors. Recognised risk factors (Key et al., 2001) are summarised in Table 1.2 below.

	High risk	Moderate risk	Slight risk
Relative risk increase	>4X	2-4X	1-2X
Personal history	Prior breast cancer	Prior ovarian cancer	
Family history	Family history of bilateral premenopausal breast cancer or familial cancer syndrome	First degree relative with history of breast cancer	
Lifestyle factors		Upper socio-economic class Prolonged uninterrupted menses Post-menopausal obesity	Onset menarche < 12; Late menopause; Late first birth; Moderate alcohol intake; OCP/HRT exposure
Histological markers	Proliferative breast disease with atypia	Proliferative breast disease with no atypia	

Table 1.2: Risk factors for developing breast cancer. OCP = oral contraceptive pill, HRT = hormone replacement therapy

1.5.2 Sub-classification of breast cancer

Breast cancer is extremely heterogeneous with diversity in histology, immunohistochemistry (IHC) and gene expression profiles emphasising the multiple biological subtypes that constitute this disease.

1.5.2.1 Histopathology and immunohistochemistry

Classification of breast cancer has been reviewed extensively elsewhere and only a brief description will be provided here (Weigelt et al., 2010). The most common type of breast cancer is ductal carcinoma which has a stellate or spiculated appearance on mammography. The tumour usually has an infiltrating edge which extends beyond what is grossly visible, warranting ample excision of surrounding normal tissue. The histological grading of the tumour is based on mitotic count, cytological atypia and degree of tubule formation. Invasive lobular carcinoma comprises only 5-10% of primary breast cancers, tends to be multi-centric within the same breast and is diffusely infiltrating. Other histological variants exist including medullary carcinoma, colloid (mucinous) carcinoma, tubular carcinoma and papillary carcinoma.

The advent of mammographic screening has led to an increase in the number of cases of ductal carcinoma *in situ* (DCIS) diagnosed over the last 30 years. DCIS consists of a malignant population of epithelial cells that are confined by the basement membrane. These cells can spread throughout a regional ductal system, producing extensive segmental lesions or develop into invasive cancer. Lobular carcinoma *in situ* (LCIS) is usually an incidental finding in breast tissue removed for other reasons. Lobules are distended and filled by relatively uniform, round, small- to medium-sized cells. Marked atypia, pleomorphism and mitotic activity are usually absent.

Immunohistochemistry (IHC) permitted early informative classification of breast cancer. Based on the degree of cell surface expression of human epidermal growth factor receptor 2 (HER2) or hormone receptors (oestrogen-receptor (ER) and progesterone receptor (PR)), a taxonomy of breast cancer was derived which correlated with clinical outcome and assisted in decision-making for therapeutic intervention. For example, HER2-positive cancers were recognisable for their intermediate outcome and sensitivity to HER2-inhibitors whilst triple negative cancers were associated with a poorer outcome.

1.5.2.2 Gene expression profiling

Microarray-based gene expression profiling studies provided confirmation of the heterogeneity of this disease and showed how breast cancer could be defined based on the intrinsic molecular expression characteristics and not determined simply by anatomical factors such as tumour size or nodal status. Seminal early work (Perou et al., 2000; Sorlie et al., 2001b) revealed the existence of at least four molecular subtypes of breast cancer— luminal epithelial-like (subtypes A and B), HER2-enriched, basal-like, and normal breast-like (Table 1.3) which showed a degree of correlation with IHC characteristics. Subsequently, further distinctions were demonstrated within some of these subtypes including directed efforts at defining expression signatures that predict disease recurrence/survival (Paik et al., 2004; van 't Veer et al., 2002) and it is anticipated that the complexity of classification will continue to increase (reviewed extensively (Reis-Filho and Pusztai, 2011)(Table 1.3)). At the last iteration, some ten subtypes of breast cancer were posited (Curtis et al., 2012).

Breast cancer subtype	IHC markers*	Histological grade*	Other markers	Outcome*	Benefit from chemotherapy*
Luminal A	ER+: 91–100% PR+: 70–74% HER2+: 8–11% Ki67: low Basal markers: –	GI/II: 70–87% GIII: 13–30%	*FOXA1* high	Good	Low (0–5% pCR)
Luminal B	ER+: 91–100% PR+: 41–53% HER2+: 15–24% Ki67: high Basal markers: –	GI/II: 38–59% GIII: 41–62%	*FGFR1* and *ZIC3* amp	Intermediate or poor‡	Intermediate (10–20% pCR)
Basal-like	ER+: 0–19% PR+: 6–13% HER2+: 9–13% Ki67: high Basal markers: +	GI/II: 7–12% GIII: 88–93%	*RB1* : low/– *CDKN2A* : high *BRCA1* : low/– *FGFR2* : amp	Poor	High (≥40% pCR)
HER2-enriched	ER+: 29–59% PR+: 25–30% HER2+: 66–71% Ki67: high Basal markers: –/+	GI/II: 11–45% GIII: 55–89%	*GRB7* : high	Poor	Intermediate (25–40% pCR)
Normal breast-like	ER+: 44–100% PR+: 22–63% HER2+: 0–13% Ki67: low/intermediate Basal markers: –/+	GI/II: 37–80% GIII: 20–63%	..	Intermediate	Low (0–5% pCR)
Claudin-low	ER+: 12–33% PR+: 22–23% HER2+: 6–22% Ki67: intermediate Basal markers: +/–	GI/II: 62–23% GIII: 38–77%	*CDH1* : low/– Claudins: low/–§	Intermediate	Intermediate (25–40% pCR)
Molecular apocrine	ER– PR– HER2 +/– Ki67: high‡ Basal markers: –/+	Predominantly GII/GIII	Androgen receptor: +	Poor	Not examined

Table 1.3: Gene expression classification taken from (Reis-Filho and Pusztai, 2011) with minor adaptation. IHC= immunohistochemistry, ER=oestrogen receptor, PR=progesterone receptor, G=histological grade, pCR=pathological complete response to neoadjuvant chemotherapy. * Conventional chemotherapy regimens; information about ER, PR, HER2, histological grade, outcome, and response to chemotherapy retrieved from a reference for luminal A, luminal B, basal-like, HER2-enriched, claudin-low, and normal breast-like subtypes (Prat et al., 2010); information about molecular apocrine subtype extracted from two references (Doane et al., 2006; Farmer et al., 2005). ‡ Outcome of luminal B varies according to the definition used.

1.5.3 Germline susceptibility alleles in breast cancer

Approximately 10-15% of cases of breast cancer have a family history of breast or ovarian cancer ((Thompson, 1994). Through linkage analysis, mutational screening of candidate genes and genome-wide association studies (GWAS), genetic predisposition factors have been identified of three distinct risk prevalence profiles: rare high-penetrance alleles, rare intermediate-penetrance alleles, and common low-penetrance alleles (reviewed (Turnbull and Rahman, 2008)).

1.5.3.1 Rare high-penetrance germline predisposing alleles

Linkage analysis of high-penetrance early-onset breast cancer families led to the identification of rare breast cancer susceptibility genes, *BRCA1* and *BRCA2* on chromosomes 17 and 13 respectively (Miki et al., 1994; Wooster et al., 1995), providing the earliest evidence of germline predisposition alleles. Loss-of-function mutations reported in these large genes were frequently private to individual families although founder mutations were reported amongst the Ashkenazim (BRCA1_185delAG, BRCA1_5382insC and BRCA2_6174delT) and the Icelandic population (BRCA2_999del5).

Germline mutations associated with *BRCA1* mutations confer an elevated lifetime risk of developing breast cancer of up to 80%, while the lifetime risk associated with *BRCA2* mutations is 40-50% (Antoniou et al., 2003). In addition, carriers of *BRCA2* germline mutations also have an increased risk of developing other cancers including pancreatic, melanoma and gastric cancers. Both genes confer elevated risks of ovarian cancer, with the risks for BRCA1 carriers exceeding those of BRCA2 mutation carriers, particularly for early-onset ovarian cancer (Antoniou et al., 2003).

Histologically, BRCA1 tumours resemble 'basal-like' breast tumours which demonstrate high histological grade, high mitotic index, central necrotic zones and lymphocytic infiltrates. They frequently lack IHC evidence of ER, PR or HER2 expression (Palacios et al., 2008), thus being triple-negative tumours. Gene expression profiles of BRCA1 tumours are similar to those associated with basal myoepithelial cells and breast cancers with a basal-like phenotype, showing expression of cytokeratins 5/6, 14, 17, vimentin, p-cadherin, fascin, caveolins 1 and 2 (Hedenfalk et al., 2001). In contrast, BRCA2 tumours have no distinguishing histopathological features and exhibit a pattern of ER expression similar to those of sporadic breast cancers.

Li-Fraumeni Syndrome is a cancer predisposition syndrome characterised by a high frequency of early onset breast cancer, sarcoma and childhood-onset cancers of the adrenal cortex and medulloblastoma (Birch et al., 2001). Early mortality associated with this syndrome makes it

reproductively limiting and thus, rare. p53 is a transcription factor integral to signal transduction in cells and has frequently been shown to be somatically mutated in cancers.

Several genes have been associated with an increased risk of breast cancer although the magnitude of associated risks remains uncertain. *CDH1* encodes a transmembrane protein, E-cadherin. Germline mutations in *CDH1* cause Hereditary Diffuse Gastric Cancer syndrome and has been associated with an increased risk of lobular breast cancer (Masciari et al., 2007). *PTEN*, a gene known to cause a multiple hamartoma syndrome, is characterised by a predisposition to benign and malignant lesions of the breast, thyroid gland and endometrium (Chen et al., 1998). *STK11*, a serine/threonine kinase, is a gene responsible for Peutz-Jegher Syndrome, characterised by hamartomatous intestinal polyps and mucocutaneous pigmentation. There is an increased incidence of different cancers in Peutz-Jegher Syndrome, including breast cancer (Bignell et al., 1998). Collectively, the attributable risk of mutations in these genes to familial breast cancer is low (Turnbull and Rahman, 2008) and many women with a family history of breast cancer do not carry mutations in any of the genes described in this section.

1.5.3.2 Rare intermediate-penetrance germline predisposing alleles

Intermediate-penetrance breast cancer genes confer a relative risk of 2 to 4 and are rare. *CHEK2* encodes CHK2, a mediator in the DNA damage response to double-strand breaks. The 1100delC mutation was reported to be present at approximately 1% population frequency and was shown to be significantly enriched in breast cancer families (Meijers-Heijboer et al., 2002). *ATM* was sought as a potential predisposition gene based on the observation that female relatives of patients with ataxia telangiectasia, an autosomal recessive condition caused by mutations in this gene characterised by progressive cerebellar ataxia, showed an excess of breast cancer (Thompson et al., 2005). ATM is involved in the DNA damage response to double-strand breaks, initiating a signal cascade upstream of p53, CHK2 and BRCA1. Truncating mutations in *BRIP1* (or *BACH1*) were found to be enriched in breast cancer families negative for *BRCA1* and *BRCA2* mutations ((Seal et al., 2006)). BRIP1 has a BRCA1-dependent role in DNA repair. Bi-allelic mutations in *BRIP1* result in Fanconi anaemia type J which is not associated with childhood tumours (Litman et al., 2005). PALB2 was identified as a novel protein in precipitated BRCA2-related complexes. Truncating mutations were enriched in probands of breast cancer families negative for BRCA1 and BRCA2 mutations when compared to controls (Rahman et al., 2007). Bi-allelic mutations result in a Fanconi anaemia type N with marked childhood predisposition to tumours such as Wilms tumour of the kidney and medulloblastoma (Reid et al., 2007). Founder mutations have been reported in Finnish and Canadian populations. A 657delT

truncating mutation has been identified in RAD50 but is presently restricted to Finnish breast cancer families (Heikkinen et al., 2006).

1.5.3.3 Common low-penetrance alleles

Eight common low-penetrance alleles have been shown recurrently in multiple genome-wide association studies to be associated with breast cancer (Cox et al., 2007; Easton et al., 2007; Hunter et al., 2007; Stacey et al., 2007; Stacey et al., 2008) conferring a relative risk of less than 1.5. These have been summarised in Table 1.4.

	Gene/Locus	Relative Risk of breast cancer	Carrier Frequency†	Breast cancer subtype	Other cancers in monoallelic carriers	Syndrome in biallelic carriers	Method of identification
High penetrance	BRCA1	>10	0.10%	basal-like	Ovarian		Linkage study
	BRCA2	>10	0.10%		Ovarian prostate	Fanconi anaemia D1	Linkage study
	TP53	>10	rare		Sarcomas adrenal brain		Candidate resequencing study
Uncertain penetrance	PTEN	2–10	rare		Thyroid endometrium		Linkage study
	STK11	2–10	rare		Gasto-intestinal		Linkage study
	CDH1	2–10	rare	lobular	Gastric (diffuse)		Linkage study
Intermediate penetrance	ATM	2–3	0.40%			Ataxia telangiectasia	Candidate resequencing study
	CHEK2	2–3	0.40%				Candidate resequencing study
	BRIP1	2–3	0.10%			Fanconi anaemia J	Candidate resequencing study
	PALB2	2–4	rare			Fanconi anaemia N	Candidate resequencing study
Low penetrance	10q26, 16q12, 2q35, 8q24, 5p12	1.08–1.26	24–50%	ER-positive			Genome-wide association studies
	11p15, 5q11	1.07–1.13	28–30%				Genome-wide association study
	2q33	1.13	0.87				Candidate association study

Table 1.4: Summary of known genetic cancer-predisposing alleles obtained from review (Turnbull and Rahman, 2008). †estimated carrier frequency of mutations/risk allele in the UK; where 'rare', the carrier frequency is unlikely to be >0.1%.

1.5.4 Somatic genetics in breast cancer

Historic analyses of somatic genetics in breast cancer were restricted to lower resolution genome-wide technologies such as karyotyping and array-CGH initially (Hicks et al., 2006; Hicks et al., 2005), and more recently high-resolution SNP6 arrays (Ching et al., 2011; Fang et al., 2011). These highly informative copy number analyses have been complemented by the increasing throughput of sequencing technologies (Greenman et al., 2007; Wood et al., 2007). Very recently, in a striking testament to the power of modern genome-wide sequencing technologies, five back-to-back publications on breast cancer demonstrated further intricacies in this highly heterogeneous disease, (Banerji et al., 2012; Curtis et al., 2012; Ellis et al., 2012; Shah et al., 2012; Stephens et al., 2012) providing a more thorough view of the molecular foundations of breast cancers. The detailed analysis described in this thesis has as well, generated two publications which provided insights into the mutagenic and repair processes that have been operative in breast cancers (Nik-Zainal et al., 2012a) and highlighted the sobering clonal heterogeneity and complexity of individual breast cancers (Nik-Zainal et al., 2012b).

1.5.4.1 Copy number aberrations

DNA copy number aberrations (CNA) in cancer lead to altered expression and function of genes within the affected regions of the genome. Affected segments are thought to harbour oncogenes or tumour suppressor genes depending on whether the regions involve gains or losses of copy number.

The most notable copy number aberration in breast cancer is the amplification of the HER2 locus, present in 10% to 15% of all breast tumours (King et al., 1985). Since then, however, no other similarly amplified ERBB2-like oncogene has been conclusively identified. In fact, other genome-wide profiling studies combining high-resolution copy number analyses and matched gene expression data had suggested candidate oncogenes in regions of recurrent amplification (e.g. 8p12, 8q24, 11q13-14, 17q21-24, and 20q13)(Chin et al., 2006; Chin et al., 2007). However, the amplification profiles were complex, multi-modal and not clearly focused at a specific genomic location, suggesting that multiple targets co-existed within such regions. Subsequent higher resolution SNP array studies were able to enlarge the repertoire of copy-number amplifications and homozygous deletions in breast cancer, with some of these changes within regions smaller than 250 Kb. Thus, in addition to identifying focal aberrations encompassing known oncogenes (such as *CCND1*, *CCNE1*, and *FGFR2*) and tumour suppressor genes (*CDKN2A* and *PTEN*), these analyses unveiled a number of other genes with potential oncogenic or tumour suppressor roles (*PCDH8*, *MRE11A*, and *HOXA3*) (Leary et al., 2008).

Genome-wide copy number patterns have shown modest correlations with gene expression-based classification of breast cancer (Bergamaschi et al., 2006). The 'simple' genomic profile characterised by a relative paucity of CNAs and defined by a gain of 1q, 16p and loss of 16q was associated with ER-positive/luminal-A breast cancers. The 'simple amplifier' usually consisted of amplifications at 11q13-14 or 17q11-13, and was most often ER-positive/luminal-A cancers or ER-negative, HER2-positive cancers. The 'complex amplifier' showed a large degree of genomic instability, with a lot of complex rearrangements and amplifications at 8q24 and 8p12. These correlated with triple-negative cancers and ER-positive/luminal B cancers. Finally, some triple negative cancers had a relatively quiet or 'flat' copy number profile (Vincent-Salomon et al., 2008). However, the direct relevance of this copy number based classification remains uncertain.

Recently, an integrative analysis of copy number, gene expression and clinical outcome of ~2000 primary breast tumours, revealed novel putative cancer genes in *PPP2R2A*, *MTAP* and *MAP2K4*. Furthermore, prognostic stratification was derived from unsupervised analysis of paired genome-transcriptome profiles, which revealed subgroups with distinct clinical outcomes including a high-risk, oestrogen-receptor-positive 11q13/14 cis-acting subgroup and a favourable prognosis subgroup devoid of somatic copy number aberrations (Curtis et al., 2012).

1.5.4.2 Point mutations, insertions/deletions and rearrangements

Landmark sequencing studies revealed the complexity of the somatic point mutation and insertion/deletion landscape in breast cancers, highlighting high-frequency somatic mutations in *TP53* (53%), *PIK3CA* (8-26%), *CDH1* (21%), *AKT1* (8%) and *GATA3* (4%) (Carpten et al., 2007; Greenman et al., 2007; Samuels et al., 2004; Sjoblom et al., 2006; Usary et al., 2004; Wood et al., 2007) (Figure 1.10a for landscape of curated cancer gene mutations), but also hinting at a remarkably large number of other genes which were more frequently mutated than what could be accounted for by chance, albeit at much lower frequencies than *TP53* or *PIK3CA* (Greenman et al., 2007; Wood et al., 2007) (Figure 1.10b for complexity somatic mutations in breast cancer to date).

A

Figure 1.10

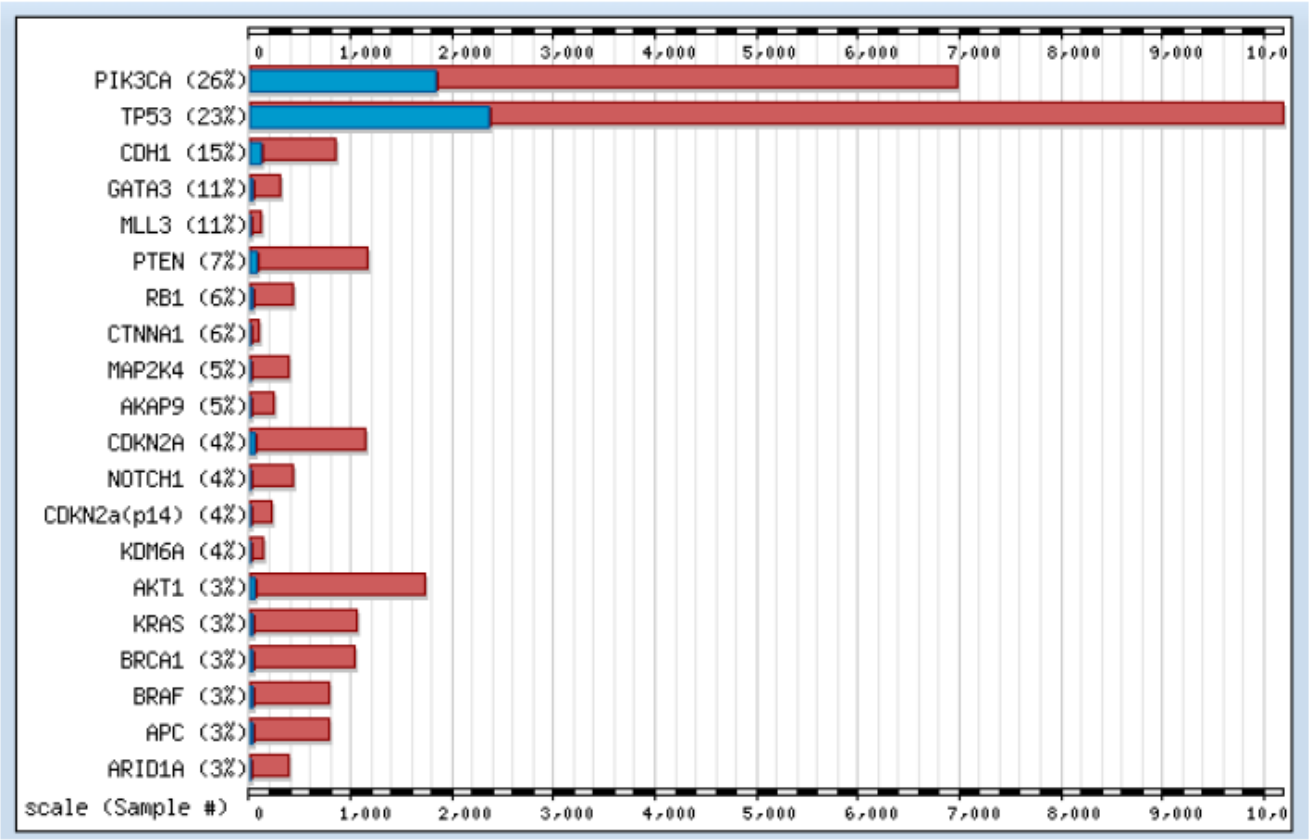
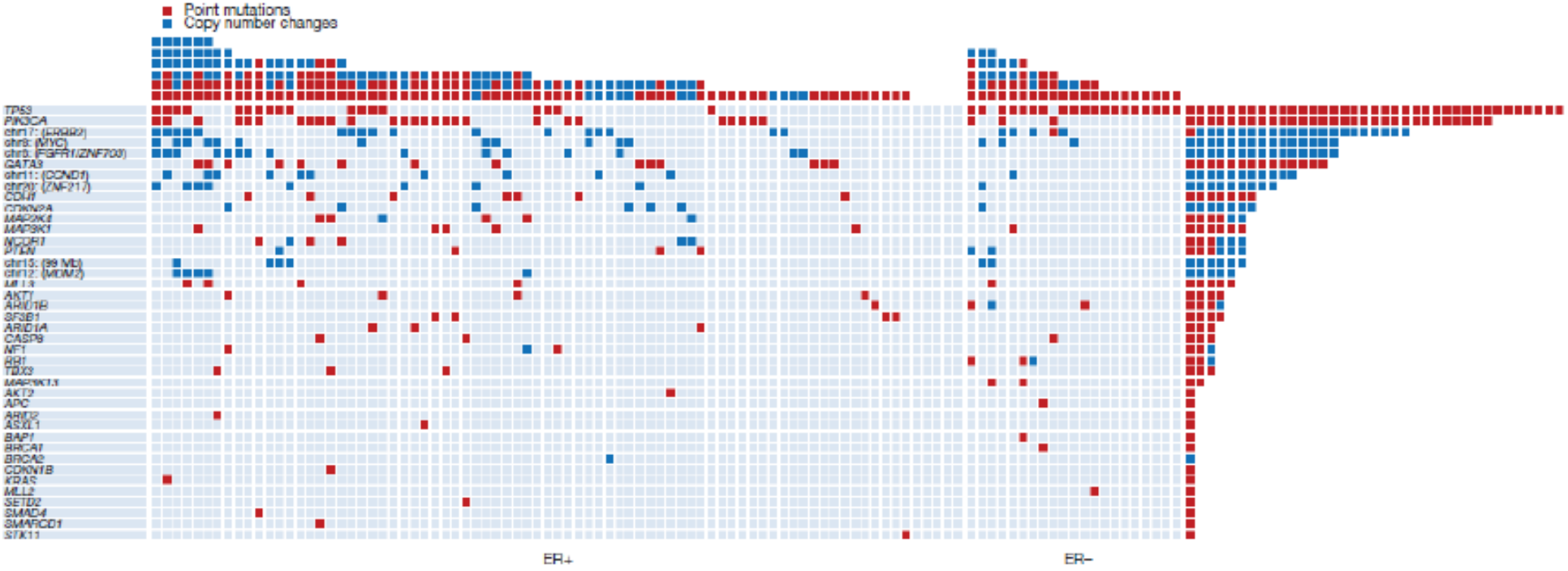


Figure 1.10: Somatic mutations in breast cancer. (A) This image is taken from http://www.sanger.ac.uk/genetics/CGP/cosmic/ and depicts the top 20 most mutated genes from 2354 breast cancer samples, which have been curated in Cosmic from many publications over many years. Preceding the advent of next-generation sequencing technology, one sample could contribute one or only a few mutations. (B, overleaf) This image is taken from a single publication, Stephens et al 2012, and depicts up-to-date complexity and marked variability between breast cancers, obtained from one large-scale next-generation sequencing experiment of 100 breast cancer samples. Each of the 40 cancer genes mutated in this experiment are documented on the left. The number of mutations in each gene in the 100 tumours is shown (rows), as is the number of driver mutations in each breast cancer (columns). Point mutations and copy number changes are coloured red and blue, respectively.

B


Point mutations
Copy number changes
TP53
PIK3CA
chr17: (ERBB2)
chr8: (MYC)
chr8: (FGFR1/ZNF703)
GATA3
chr11: (CCND1)
chr20: (ZNF217)
CDH1
CDKN2A
MAP2K4
MAP3K1
NCOR1
PTEN
chr15: (99 Mb)
chr12: (MDM2)
MLL3
AKT1
ARID1B
SF3B1
ARID1A
CASP8
NF1
RB1
TBX3
MAP3K13
AKT2
APC
ARID2
ASXL1
BAP1
BRCA1
BRCA2
CDKN1B
KRAS
MLL2
SETD2
SMAD4
SMARCD1
STK11
ER+
ER−

Exploiting the increase in scale afforded by NGS technology, targeted exome sequencing and copy number analysis of 100 breast cancers revealed nine new cancer genes (Stephens et al., 2012). These genes were rarely mutated but more so than would be expected by chance and many of the acquired mutations in these genes were predicted to lead to protein truncations. In a separate targeted exome experiment of 103 breast cancers and whole-genome sequencing experiment of 22 breast cancers of diverse subtypes from patients in Mexico and Vietnam, recurrent mutations in the *CBFB* transcription factor gene and its partner *RUNX1* was reported beyond confirmation of recurrent somatic mutations in *PIK3CA*, *TP53*, *AKT1*, *GATA3* and *MAP3K1* (Banerji et al., 2012).

In a study of 104 triple-negative breast cancers, striking inter-tumoural and intra-tumoural heterogeneity was seen in the frequencies of copy-number abnormalities and mutations. Although high-frequency somatic mutations like *TP53*, *PIK3CA* and *PTEN* were involved in the early stages of breast-cancer development, only one-third of the low-prevalence mutated genes identified in this analyses were expressed, suggesting that many of these were simply passenger events (Shah et al., 2012). There have been some efforts correlating somatic mutation profiles with clinical outcomes (Ellis et al., 2012). Focusing on ER-positive pre-treatment breast cancer biopsies from patients treated with a drug called aromatase inhibitors, it was demonstrated that tumours that had a high frequency of cells expressing Ki67, a protein associated with resistance to aromatase inhibitors, contained an elevated frequency of somatic mutations and copy number changes compared with tumours with a low frequency of Ki67-positive cells. This implicates acquired genetic/genomic modifications in the development of resistance to this drug in this subtype of breast cancer (Ellis et al., 2012), although most mutations were not recurrent.

Apart from the *ETV6-NTRK3* gene fusion associated with secretory breast carcinoma (Lae et al., 2009), recurrent gene fusions are not a common feature in breast cancer. Using low-coverage second-generation sequencing technology to assess 24 breast cancers/cell lines, 21 out of 29 somatic rearrangements predicted to generate in-frame gene fusions were found to be expressed although none were recurrent in the cohort (Stephens et al., 2009). Furthermore, 3 rearrangements of potential biological interest (*ETV6–ITPR2*, *NFIA–EHF* and *SLC26A6–PRKAR2A*) were screened across 288 additional breast cancer cases and were also not found to be recurrent (Stephens et al., 2009). Recently however, a *MAGI3–AKT3* fusion predicted to lead to a combined loss of function of PTEN and activation of the AKT3 oncogene was found to be enriched in triple-negative breast cancers (5 out of 72 examined) (Banerji et al., 2012). Perhaps as more whole genome sequences of breast cancers become available in the near future, rarer recurrent gene fusions will come to light.

In summary, breast cancer is a common and complex malignancy. Epidemiological risk factors and germline predisposition alleles are well-recognised, open to monitoring and intervention, and provide some insight into disease pathogenesis. Although a spectrum of tumour phenotypes is known and is informative for clinical outcome and treatment options, somatic genome-wide characterisation of this disease has shown marked inter-tumoural and intra-tumoural heterogeneity by genomic copy number analyses, gene expression profiling and by scrutiny of the landscape of known somatic mutations. As the resolution of genome-wide profiling continues to increase, it is expected that more detailed multi-dimensional analyses will increase the transparency of how somatic mutation is linked to tumour development and biology.

In this thesis, five breast cancers were obtained from patients with germline mutations in *BRCA1* and four from germline *BRCA2* mutation carriers. Twelve breast cancers were derived from women who developed sporadic breast cancers. A spectrum of breast cancers was sought in order to gain insights into potentially distinguishing variation in genomic patterns particularly as this cohort of samples included cancers with a known defect in a repair pathway, homologous recombination.

1.5.5 Using second-generation sequencing technology to study breast cancer in this thesis

This thesis will exploit the increasing resolution afforded by new sequencing technologies. The ability to sequence entire breast cancer genomes rests on the marked improvements in sequencing technology and the completion of the human genome sequence which has allowed systematic re-sequencing of cancer genomes to identify all classes somatic mutations. Historic limitations in technology restricted early studies to PCR-based sequencing of exons of protein-coding genes (Greenman et al., 2007; Wood et al., 2007). The recent advent of second-generation sequencing technology (Bentley et al., 2008) has permitted large-scale sequencing of whole cancer genomes for identification of all classes of somatic mutations. While many studies have focused on cancer gene discovery and/or analysis of mutations in coding regions, detailed analyses of the entire catalogue of somatic mutations in a malignant melanoma and a small cell lung cancer (Pleasance et al., 2010a; Pleasance et al., 2010b) laid the foundations for how genome-wide signatures of environmental mutagenic insults and endogenous repair mechanisms could be appreciated.

The primary aim of this thesis is to exploit the advances in sequencing technology so as to archive full catalogues of somatic mutations from twenty-one different breast cancers, in order to explore whether evidence of mutational processes comprising DNA damaging activity and DNA repair mechanisms may be identifiable across these breast cancers. The experimental and informatics steps involved in achieving the final catalogues of somatic mutations will be described. The mutational processes which have shaped these breast cancers are anticipated to leave distinguishing imprints or mutational signatures which will be extracted and characterised. The wealth of biological information that is buried within this rich dataset will be discussed.