WELLCOME TRUST SANGER INSTITUTE
THE UNIVERSITY OF CAMBRIDGE

Exploring mutational signatures in twenty-one breast cancers

Serena Nik-Zainal

Murray Edwards College

This dissertation is submitted for the degree of Doctor of Philosophy
on the
26th of September 2012

DECLARATION OF ORIGINALITY

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 Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given in the bibliography.

CONTENTS

SUMMARY

The set of somatic mutations observed in a cancer genome is the aggregate outcome of one or more biological processes that have been operative over the lifetime of a patient. Each biological process is characterised by the pattern of mutations that it leaves on the cancer genome or "signature" which is determined by the underlying mechanisms of DNA damage and of DNA repair that constitute the biological process.

In this thesis, I set out to extract the mutational signatures characterising the biological processes that have been operative in breast cancers. Catalogues of all classes of somatic mutation were generated from twenty-one whole-genome sequenced breast cancers using an integrated suite of bioinformatic algorithms which had been optimised for producing complete datasets with high sensitivity and specificity.

Mathematical methods were applied in order to extract underlying mutational signatures. Multiple distinct single-substitution, double-substitution and deletion signatures were unearthed by these analyses. Remarkably, these signatures were able to distinguish breast cancers from women with germline mutations in *BRCA1* and *BRCA2*, indicating how defects in homologous recombination leaves its mutagenic imprint on cancer genomes. Furthermore, an intriguing phenomenon of localised hypermutation characterised almost exclusively by cytosine mutations at TpC dinucleotides, demonstrating marked co-localisation with somatic rearrangements, was uncovered. These clusters of regional hypermutation were a frequent observation, occurring in thirteen out of the twenty-one breast cancers studied and have been termed *kataegis* (greek for showers/thunderstorms/towards the earth). The mechanism underlying this mutational signature is unknown. However, a role for the *APOBEC* family of cytidine deaminases in kataegis is proposed. Finally, integrated analysis of substitution mutations and expression data revealed the past operation of transcription-dependent mechanisms in generating the mutational profiles in these cancers.

This study harnesses the full scale of whole-genome sequencing demonstrating how detailed analyses of genomic data can provide biological understanding into hitherto unrecognised mutational signatures present in breast cancer genomes. In the future, the analyses of vast numbers of catalogues of somatic mutation from numerous worldwide cancer sequencing projects may herald further insights into mutational processes that underpin the development of cancers.

ACKNOWLEDGEMENTS

I would like to thank Professor Mike Stratton for welcoming me into his laboratory and for the patient manner in which he has supervised this work and guided me through research training. I would also like to acknowledge the advice and encouragement from the members of the Cancer Genome Project, in particular, Dr Peter Campbell and Dr Andy Futreal. I am indebted to the large number of people who have helped with technical wet-bench matters, provided informatic and administrative support, both within the Cancer Genome Project and the Wellcome Trust Sanger Institute. I would also like to acknowledge several key people who have helped me to in the intellectual development of this project: Ludmil B. Alexandrov, David Wedge and Peter Van Loo for informatic development and Michael Neuberger of the MRC Laboratory of Molecular Biology, Cambridge, for guidance on mutational signatures. Finally, I would like to thank my family, in particular, my husband who has been endlessly supportive, patient and cheerful, and who has counselled me on so many occasions, my children who have kept me smiling and my mother who has been ceaseless in her affection and her assistance.

This work is supported by the Wellcome Trust Clinical Research Training Programme in Cambridge.

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.

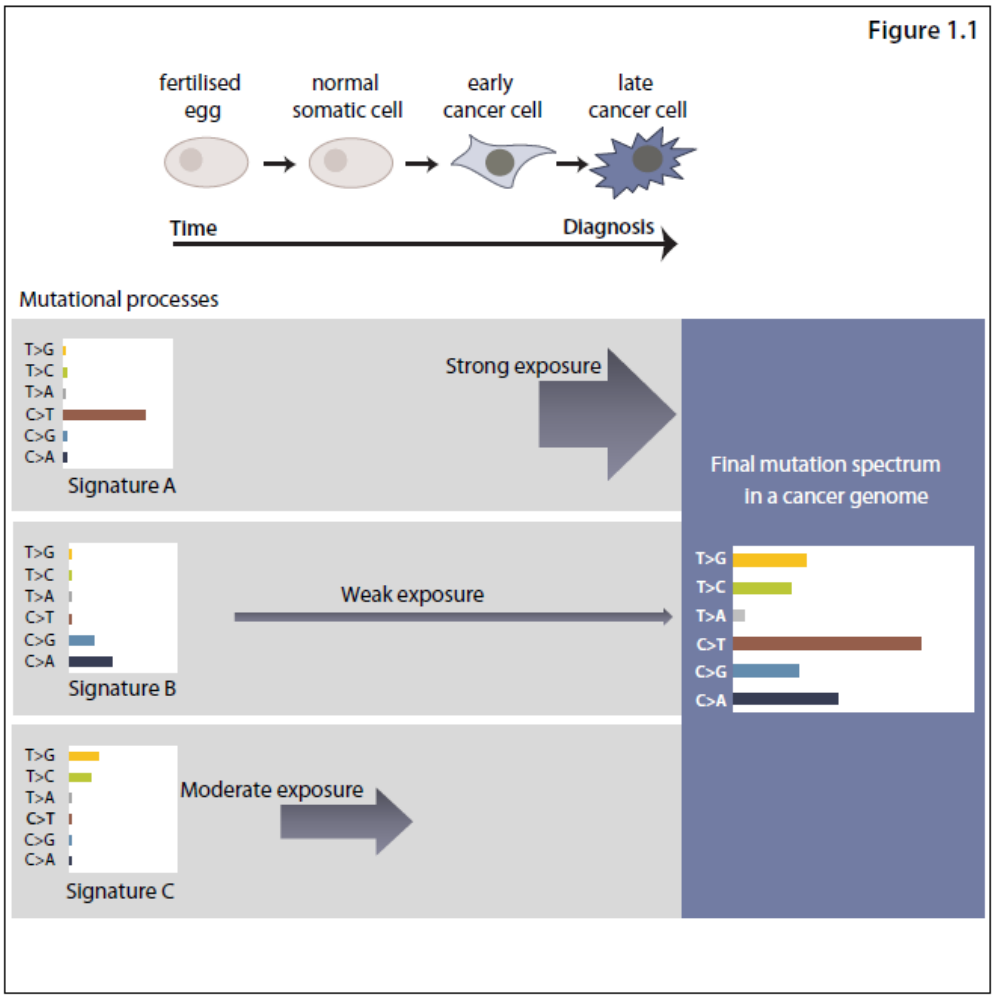


Figure 1.1: Mutational signatures in cancer genomes. From the time of the fertilised egg through to the development of an invasive cancer, multiple mutational processes are likely to be operative with each process producing its own characteristic signature. At the point of diagnosis and of sequencing 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.

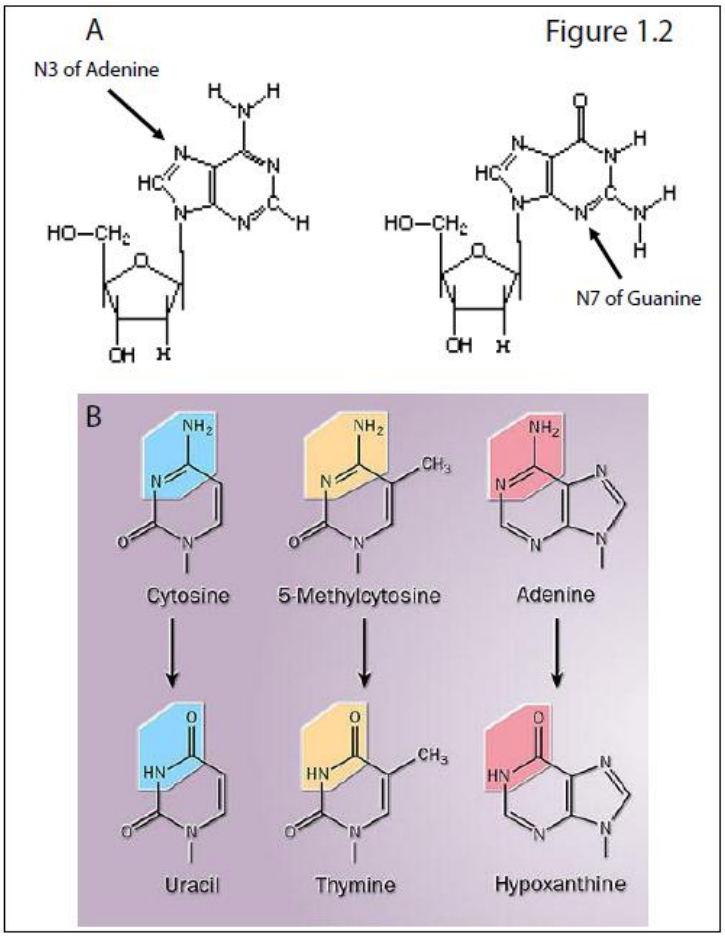


Figure 1.2: Base susceptibility to damage. The basic chemical unit of DNA is the nucleotide which comprises a phosphate group esterified to a pentose sugar, which is joined by a labile N-glycosidic bond to a base. However, structural properties of certain bases make them susceptible to DNA damage. (A) 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

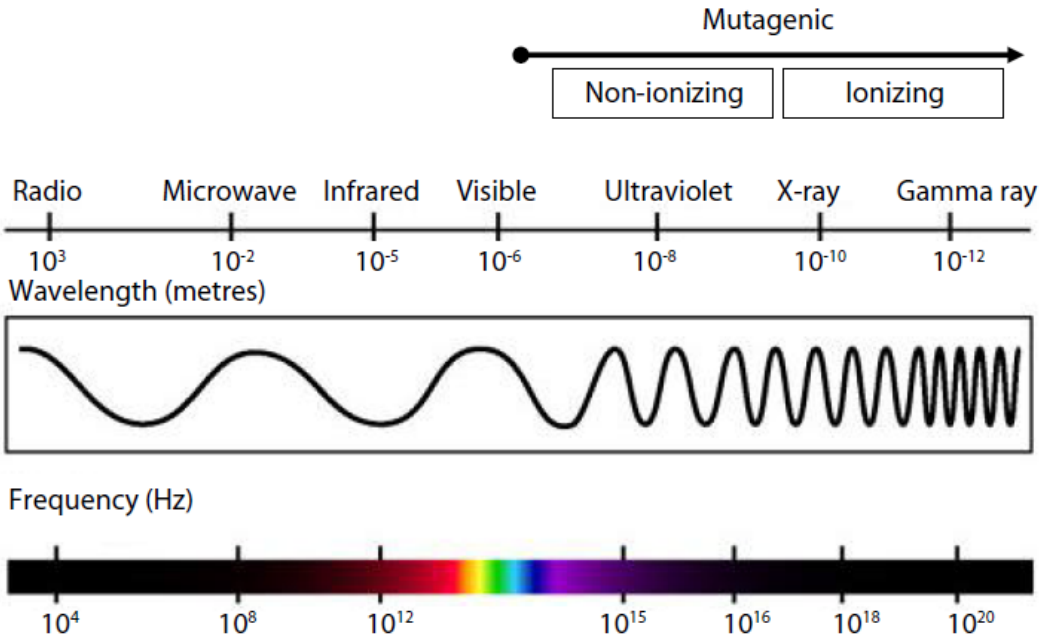


Figure 1.3: The range of mutagenic radiation in the electromagnetic spectrum

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

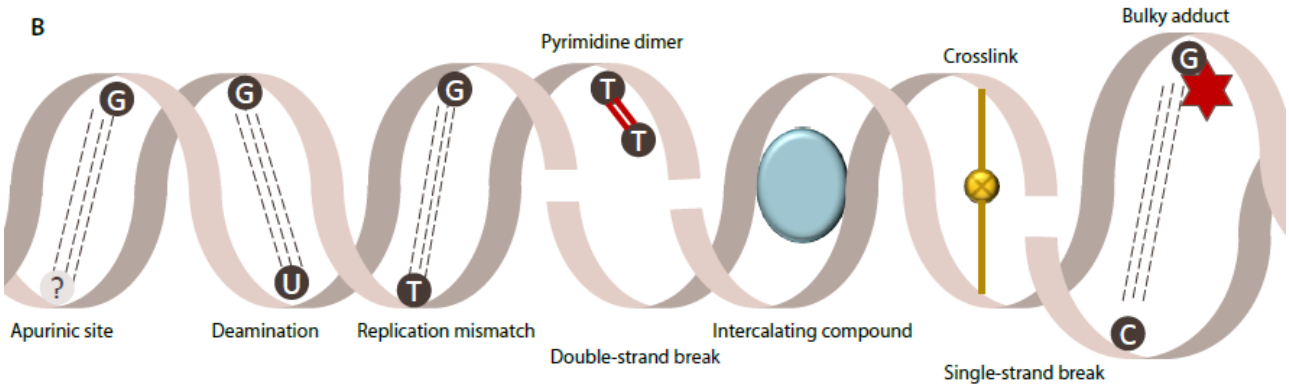


Figure 1.4: (A) Classification of DNA damaging agents in this thesis. (B) Mutagenic effects of various DNA damaging agent.

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

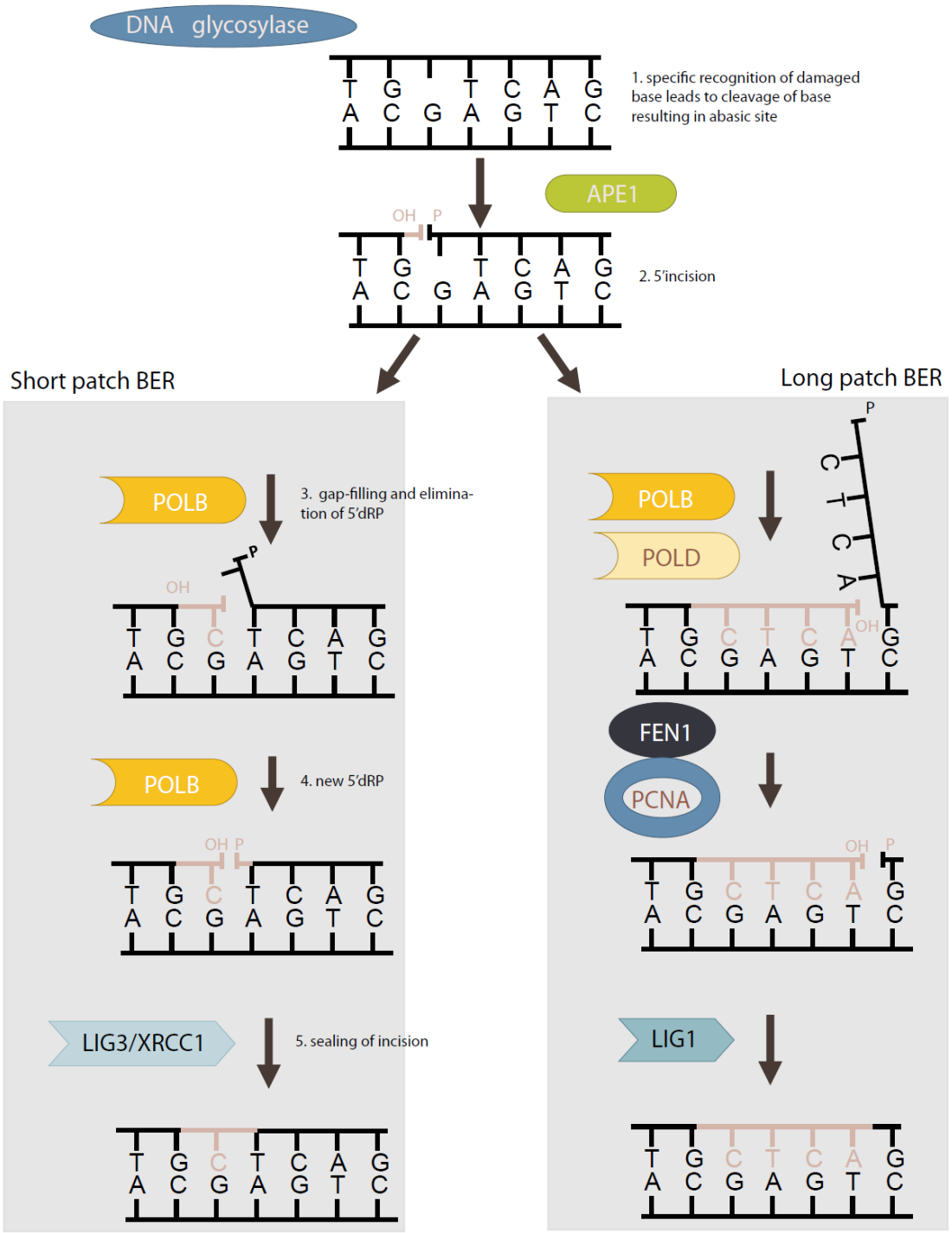


Figure 1.5: An outline of short-patch versus long-patch base excision repair (BER). BER begins with the recognition of a damaged base by a DNA glycosylase and removes the altered base, giving rise to an abasic site. The abasic site is then processed by an apurinic/apyrimidinic (AP) endonuclease (APE1), 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.

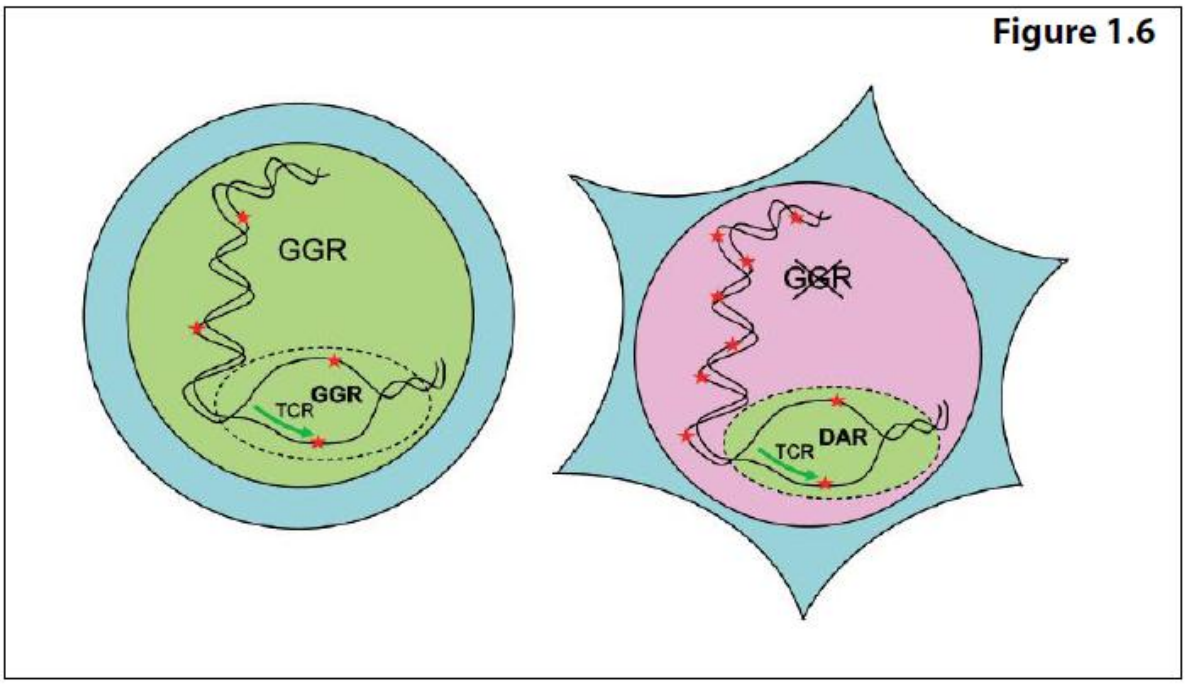


Figure 1.6: An outline of nucleotide excision repair (NER). Global genome repair (GGR) occurs in replicating cells throughout the genome. Transcription-coupled repair (TCR) occurs preferentially within transcription factories (dashed oval). In fully differentiated cells, GGR is down-regulated but the 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).

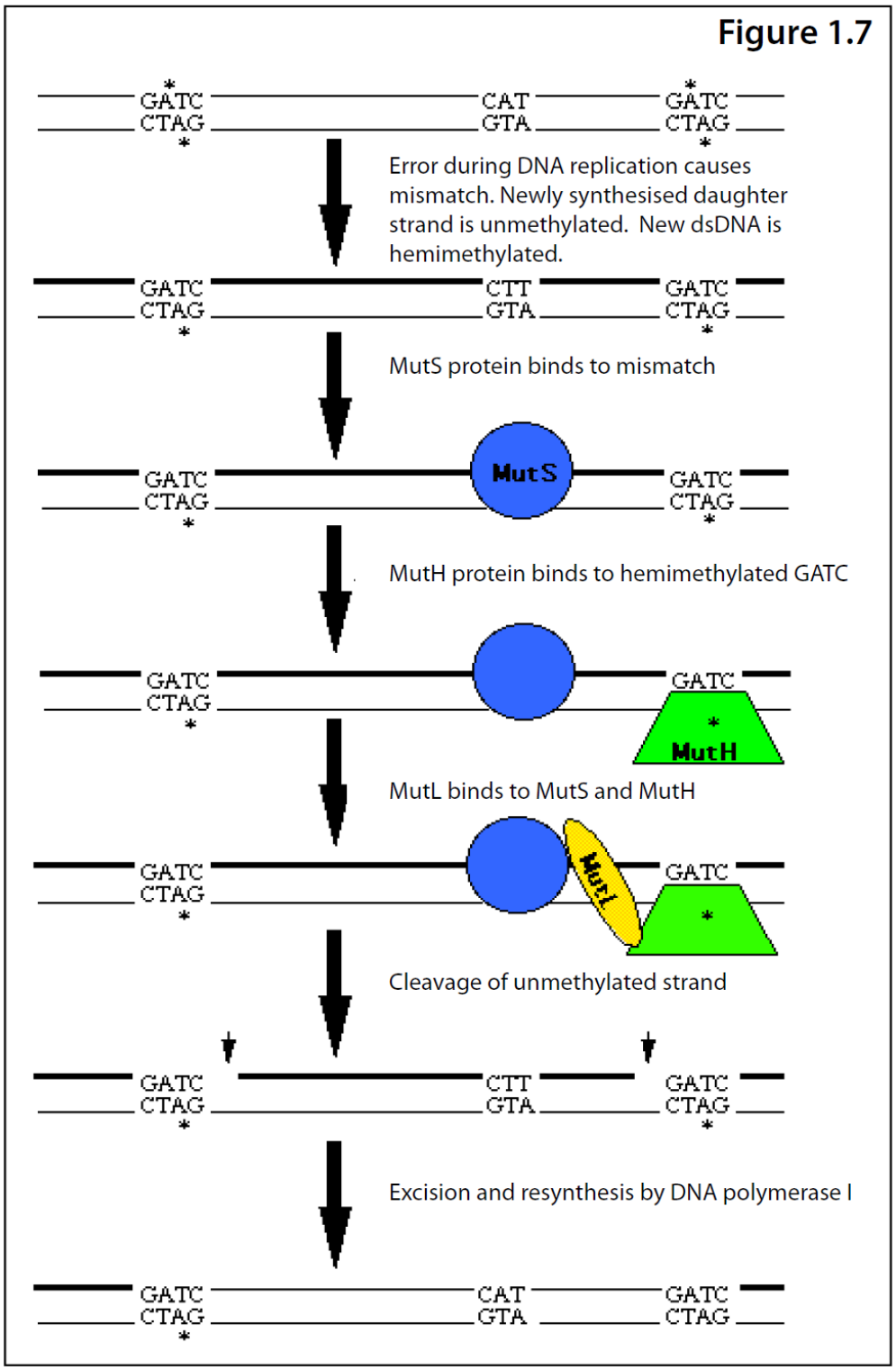


Figure 1.7: The key steps involved in mismatch repair which is able to discriminate between newly-synthesised daughter strand and the parental strand. This ensures that the newly-synthesised strand is preferentially repaired by this key pathway. Image taken from Maloy laboratory, San Diego State University,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

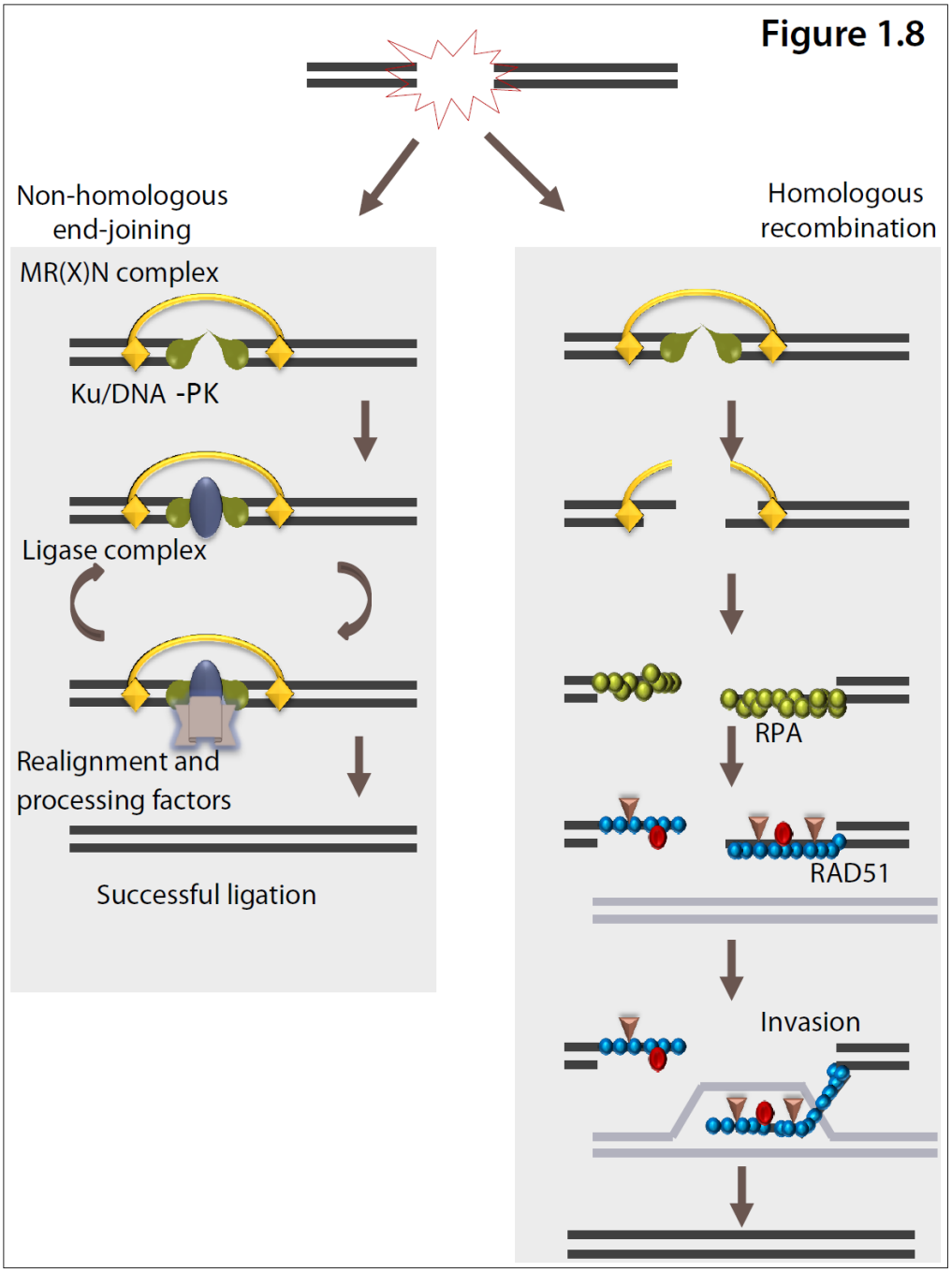


Figure 1.8: Repair of double-strand breaks: The options of non-homologous end-joining (NHEJ) or homologous recombination. In NHEJ, MR(X)N and Ku complexes are believed to bind to double-strand breaks shortly after double-strand break formation, bridging, tethering and stabilising the two broken ends, 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

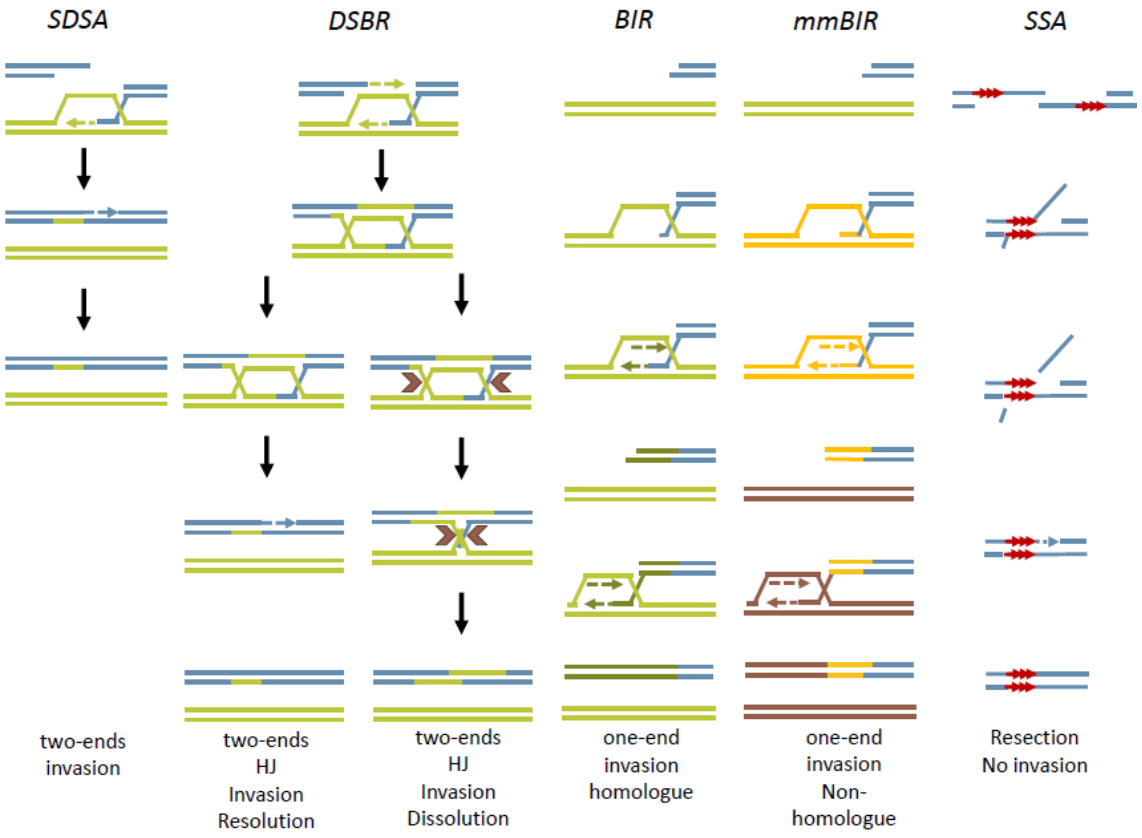


Figure 1.9: Different possible modes of resolving double-strand breaks within the homologous recombination (HR) pathway. The most conservative method of repair in HR is synthesis-dependent strand annealing (SDSA). Alternative sub-pathways differ in the following ways. If both ends of a double-strand 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

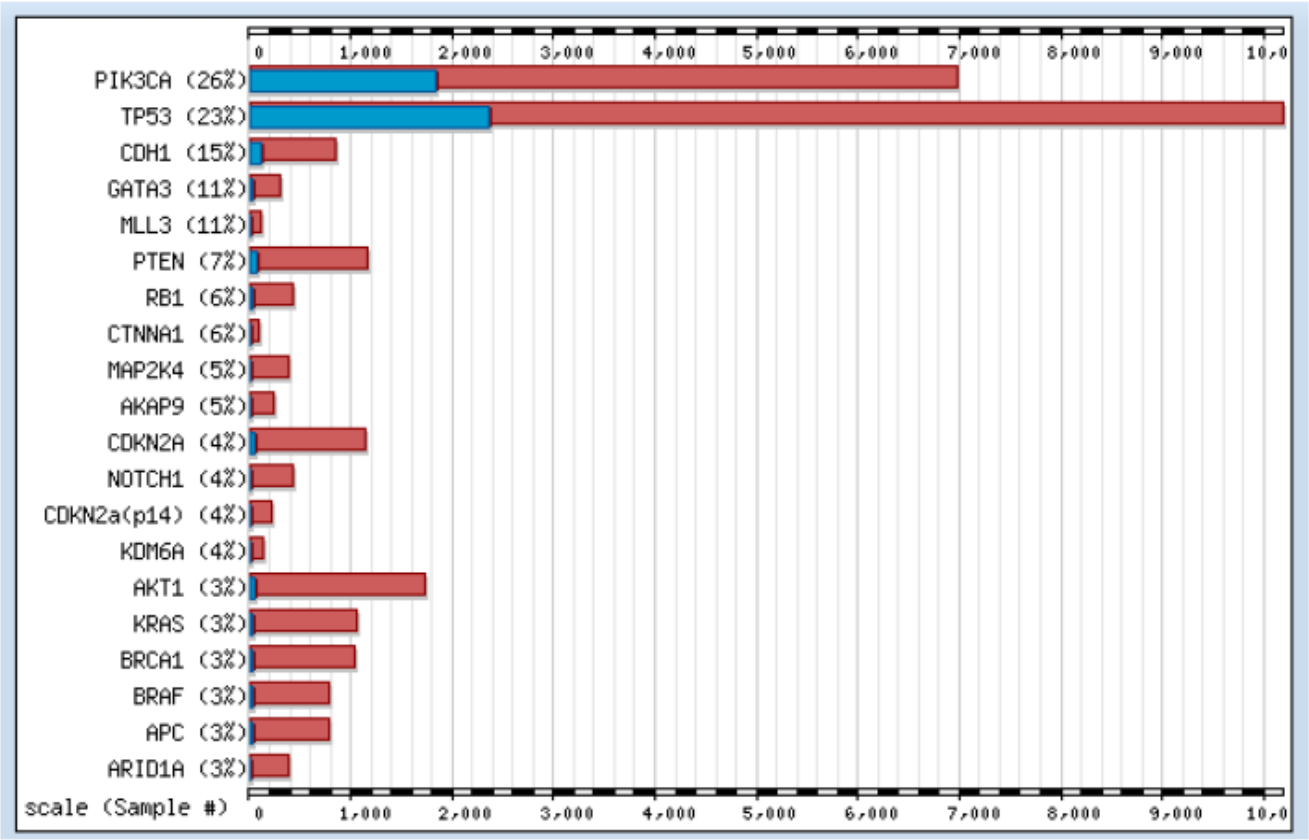
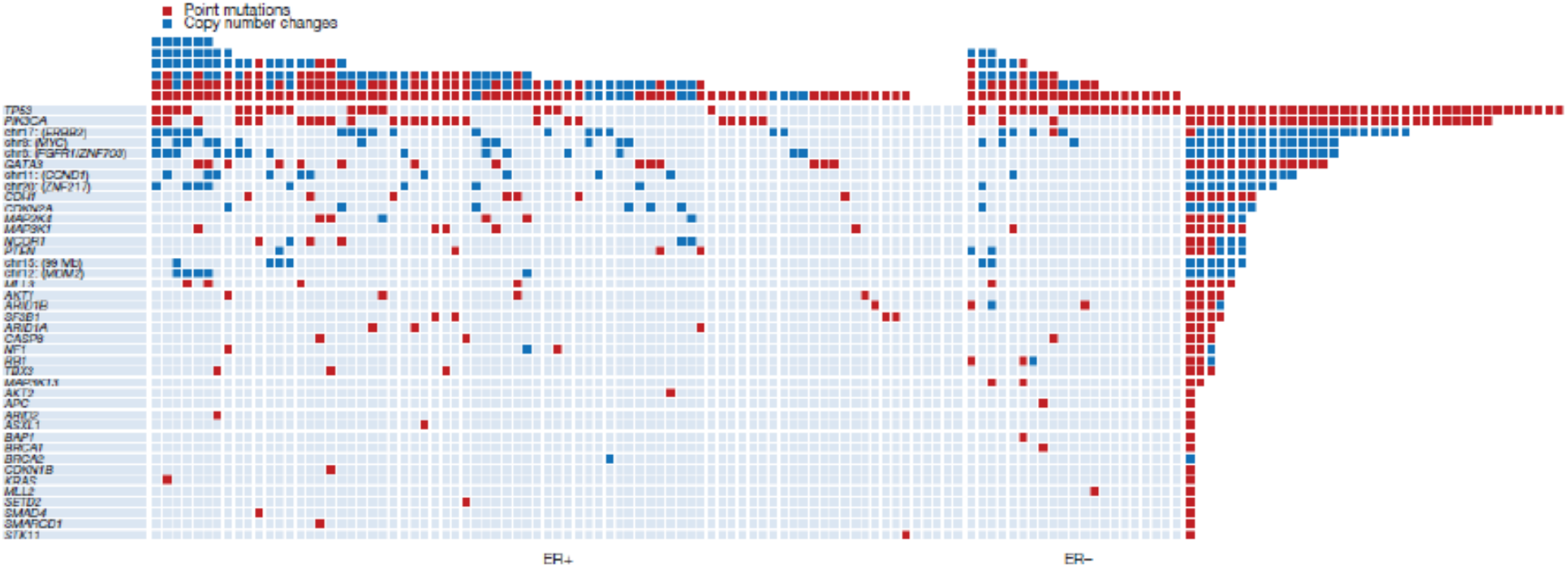


Figure 1.10: Somatic mutations in breast cancer. (A) This image is taken from http://www.sanger.ac.uk/genetics/CGP/cosmic/ and depicts the top 20 most mutated genes from 2354 breast cancer samples, which have been curated in Cosmic from many publications over many years. Preceding the advent of next-generation 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.

CHAPTER TWO: EXPERIMENTAL PROCEDURES

2.1 INTRODUCTION

Cancer is the ultimate genetic pathology; defined and characterised by an accumulation of somatic genomic aberrations, noted since the turn of the twentieth century. The relationship between exposure to DNA damaging agents and the subsequent accrual of mutations over an interval of time has been clearly documented. That a mutagen may attack individual sequence motifs and leave its imprint on a genome which may then be mitigated by the plethora of repair pathways present in the human cell has been acknowledged and will, here, be exploited as a mutational signature. These mutational signatures are inscribed layer upon layer on the cancer genome through the lifetime of the cancer patient.

Utilising the force and scale presented by next-generation sequencing technology, evidence for mutational signatures were explored from the wealth of data proffered by whole genome sequencing strategies in this thesis. However, in order to explore those mutational signatures, a set of high-confidence somatic mutations of all mutation classes for each breast cancer was essential for the nature of the downstream analysis intended. In order to obtain this dataset, a series of wet-bench and informatic procedures were performed and summarised in Figure 2.1. A more detailed description of each of these steps will be provided in the rest of this chapter.

An overview of the overall strategy was as follows:

- **Systematic re-sequencing using high-coverage paired-end next-generation sequencing technology**

DNA was obtained from twenty-one breast cancers and matched normal DNA from women diagnosed with breast cancers and systematic re-sequencing was performed from each of these samples using high-coverage, paired-end second-generation sequencing technology. The samples were obtained from the International Cancer Genome Consortium Breast Cancer Working Group according to local ethical approval. The full spectrum of histopathological subtypes of breast cancers were targeted and compared and contrasted to each other.

- **Employ bespoke bioinformatic algorithms to call all classes of somatic mutation and**

Bioinformatic algorithms were employed to map sequences back to the reference genome, call all variants in tumour and normal with subtraction of normal variation to generate comprehensive catalogues of somatic mutations. Further informatic tools were required to analyse and interpret all classes of somatic variants including substitutions, indels, somatic rearrangements and copy number aberrations. Post-processing filters were developed in order to obtain a dataset with high specificity and sensitivity. An orthogonal method (PCR, capillary sequencing, Roche pyrosequencing) was used to validate subsets of variants as being truly somatic in order to ensure high quality datasets for further analysis.

- **Extract and characterise patterns of somatic mutation and integrated analyses**

Having secured high-quality datasets, patterns of somatic mutation were sought taking types of mutation, sequence context and genomic architecture into consideration, in order to extract understanding regarding processes involved in initiating mutation and insights into DNA repair mechanisms. Excavation of these genomes included analyses of mutational rates and the timing of mutations through the evolution of the cancers. Transcriptomic profiling by expression arrays were performed in order to allow consideration of factors such as expression levels. Integrated analyses of different classes of mutation and transcriptomic data were performed to explore these relationships.

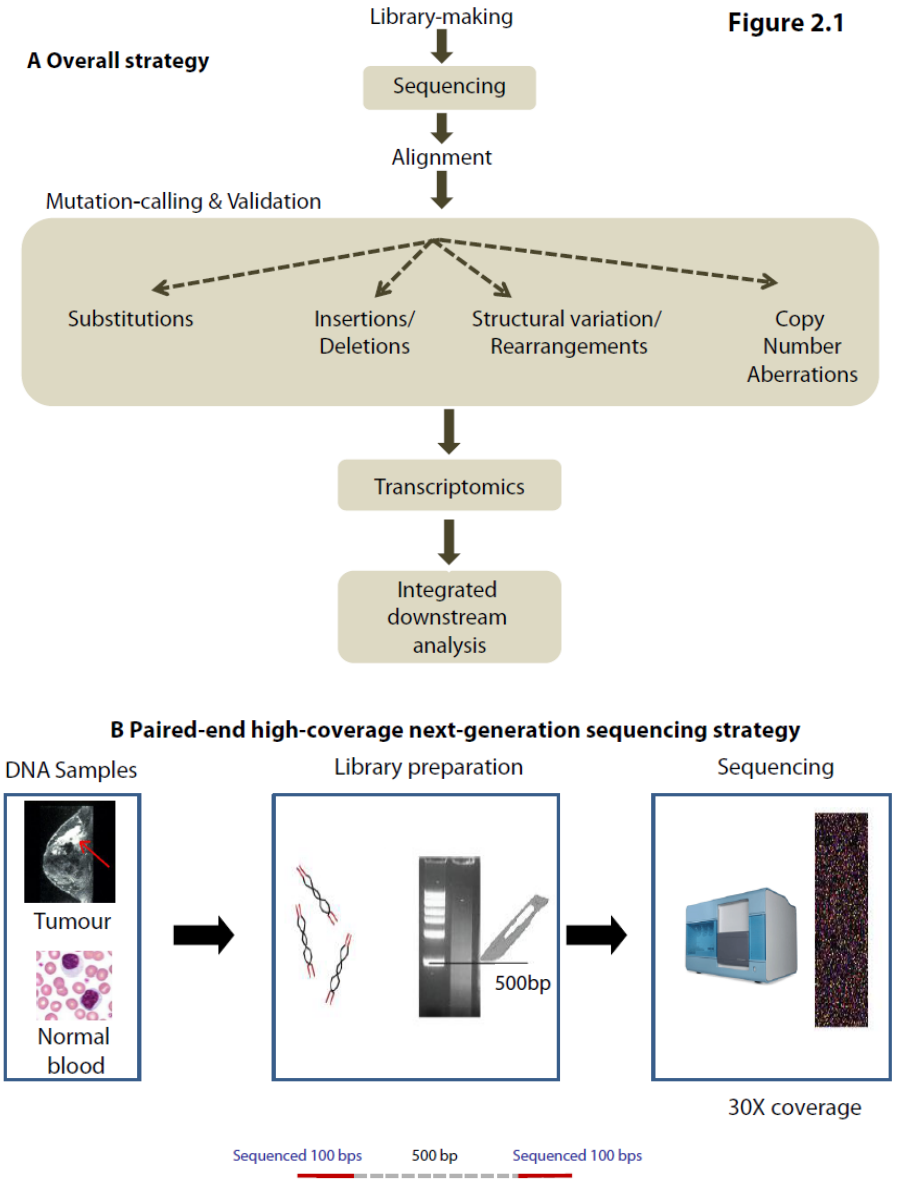


Figure 2.1: A flowchart of the full whole-genome sequencing and analysis strategy for twenty-one breast cancers. DNA obtained from collaborators was used to construct Illumina no-PCR libraries prior to sequencing on Illumina HiSeq 2000 sequencers. Raw sequences of the tumour sample and raw sequences of the normal sample were aligned back to the reference genome build 37 independently. All classes of somatic mutation including substitutions, insertions/deletions, somatic rearrangements and copy number aberrations were sought using a range of bioinformatic tools. Transcriptomics by expression arrays were also performed. A high quality dataset was obtained following post-processing or curation of the datasets, including validation on an orthogonal sequencing platform of a subset of substitutions and all insertions/deletions and rearrangements. The finalised dataset was used for all downstream analyses described in subsequent chapters of this thesis. (B) Paired-end next-generation sequencing strategy. A DNA sample was obtained from the breast cancer and from matched peripheral blood lymphocytes for each patient, fragmented using a Covaris Sonicator separately and following DNA preparation (end-repair, A-tailing and adaptor ligation), gel size-selected 500bp fragments to make a next-generation sequencing library. Each gel slice (library) contained billions of fragments of DNA and was representative for the entire genome of the population of cells in each cancer/matched normal sample. 100bp at both ends of each ~ 500bp fragment was sequenced. Each library was sequenced to generate enough raw sequence to ensure an average coverage of 30-fold per reference base in the genome, hence the term paired-end, high coverage next-generation sequencing strategy.

2.2 THE GENERATION OF ILLUMINA NO-PCR NEXT-GENERATION SEQUENCING LIBRARIES

Breast cancer samples included in this study had previously been subjected to pathology review by two pathologists independently scoring each sample, and only samples with >70% tumour cellularity were accepted for the project. DNA from tumour and matched normal samples were provided by collaborators of the International Cancer Genome Consortium (ICGC) Breast Cancer Working Group. The DNA samples provided were subject to local ethical approval of individual ICGC members. Illumina no-PCR libraries were generated from the DNA samples and a flow diagram of the principles of the library-making process is provided in Figure 2.2.

2.2.1 Starting quantity and fragmentation

Short insert 500bp library construction, flowcell preparation and cluster generation was in accordance with the Illumina no-PCR library protocol (Kozarewa et al., 2009). In brief, 5ug of DNA was brought to 120ul of T0.1E, transferred to a 150ul AFA Covaris vial and sealed. DNA was fragmented using a Covaris AFA DNA Sonicator. Fragment sizes ranging between 300 and 600bp were generated using the following shearing conditions:

- Intensity = 5
- 20% duty cycle
- 200 cycles/burst
- duration 30s
- at 4 ^{o}C.

This was followed by a purification step using a QIAquick protocol to result in 30ul of fragmented DNA. In brief, 600ul of PB buffer was added to the 120ul of fragmented DNA sample and the mixture (720ul) was added to a QIAquick column within a QIAquick tube. Centrifugation at 13,000RPM was performed for 1 minute in a benchtop centrifuge. Flow-through was discarded and the column holding the filter containing the fragmented DNA was replaced into the same tube. 750ul of PE buffer was added and centrifuged for 1 minute at 13,000 RPM. Flow-through was discarded and the column was replaced into the QIAquick tube again. An additional 1 minute of centrifugation was performed in order to remove excess fluid. The QIAquick column was now placed into a fresh tube and 32ul of EB buffer was placed onto the centre of the QIAquick membrane. Following a two minute wait, the

column was centrifuged for a further minute at 13,000RPM. The eluate containing the DNA was retained.

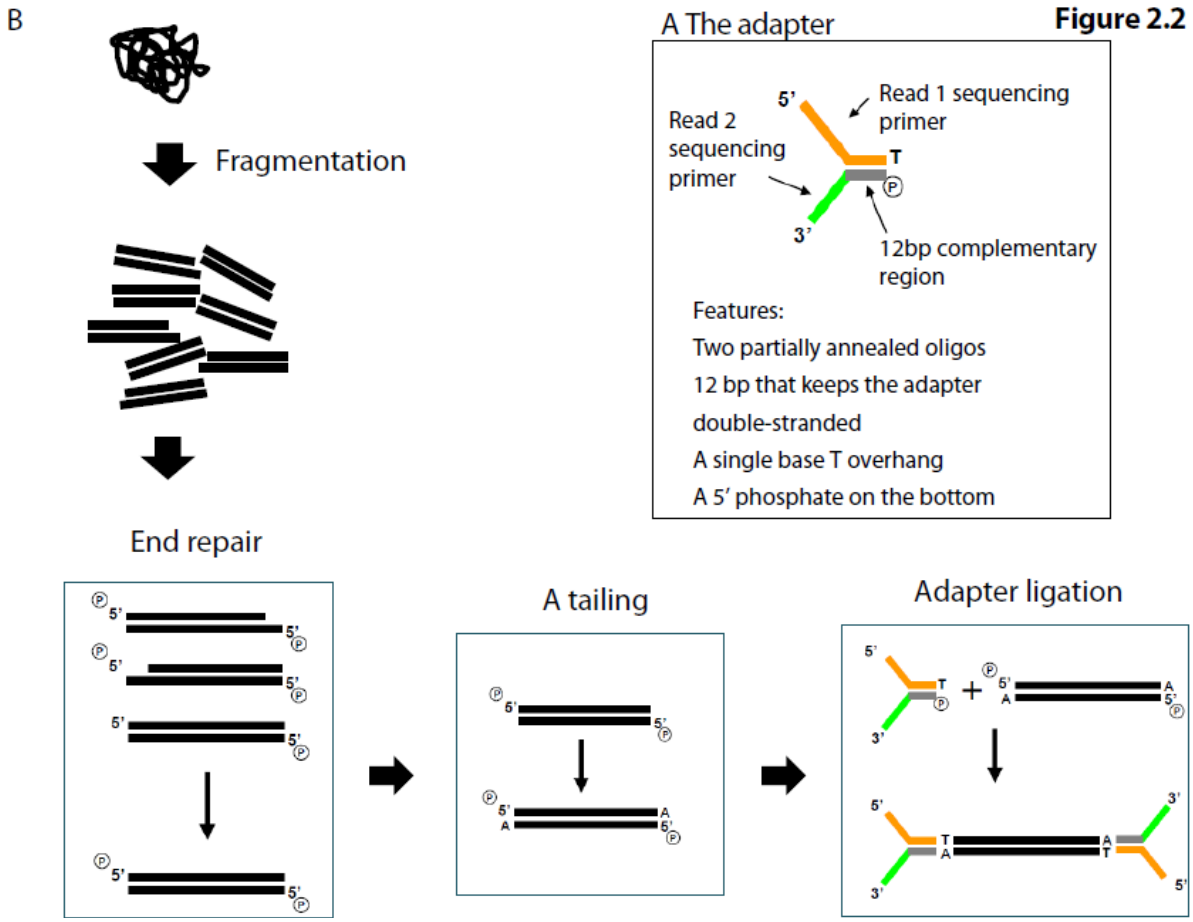


Figure 2.2: Flow diagram of the principles of the DNA preparation process. (A) The Illumina adaptor comprises two partially annealed oligos with a 12bp complementary region which keeps the adapter double-stranded. Sequencing primers for Read 1 and Read 2 of the sequencing process are embedded in the unannealed sections of the adapter. A single T base overhang is present at the 3' end with a 5' phosphate group present on the bottom. (B) DNA preparation of the library-making process. Genomic DNA sonically fragmented and the ragged ends of fragmented DNA are end-repaired to generate blunt ends (end-repair). An A base is added (A-tailing) to increase the efficiency of ligation prior to the adapter ligation step. The Illumina no-PCR library which is finally sent for sequencing therefore comprises billions of ~500bp fragments of DNA with adapters at both ends of each fragment.

2.2.2 End-repair, A-tailing and adaptor ligation

The fragmentation step generated double-stranded fragments which could have ragged edges and overhangs that could reduce the overall efficiency of ligation of the Illumina no-PCR adaptors. Therefore, end-repair and phosphorylation of fragmented DNA was performed using NEB reagents in the following quantities for each library:

No of samples	1
Water (ul)	45
T4 DNA ligase buffer (ul)	10
10mM dNTP Mix (ul)	4
T4 DNA Polymerase (ul)	5
Klenow DNA Polymerase (ul)	1
T4 PNK (ul)	5
Total (ul)	70
DNA (ul)	30
Total volume (ul)	100

QIAquick purification (as described previously) at this stage resulted in 32ul of end-repaired DNA, and was followed by A-tailing, a process which adds an "A" at the 3' end of the double-stranded fragments. This process was developed by Illumina and thought to increase the efficiency of the subsequent ligation step. NEB reagents were used in the following quantities:

No of samples	1
Klenow buffer (ul)	5
1mM dATP (ul)	10
Klenow exo- (ul)	3
Total (ul)	18
DNA(ul)	32
Total volume (ul)	50

A purification step was performed using MinElute columns. In brief, 250ul of Buffer PB was added to the 50ul of sample. The mixture (300ul) was added to a MinElute column, centrifuged at 13,000 RPM for 1 minute and flow-through discarded. 750ul of Buffer PE was added to the MinElute column and centrifuged for 1 min at 13,000 rpm. Flow-through was discarded again and an additional centrifugation was performed to remove excess fluid. Columns were placed into fresh tubes and 12ul of Buffer EB was placed onto the center of the MinElute membrane followed by another centrifugation. Eluate of 10ul of A-tailed DNA was retained

Ligation was carried out with the standard preparation of Illumina no PCR adapter oligo mix using the following quantities:

No of samples	1
Quick ligase buffer (ul)	25
Quick T4 DNA ligase (ul)	5
Total (ul)	30
DNA(ul)	10
Indexed Adaptor (ul)	10
Total volume (ul)	50

Purification was performed using Agencourt Ampure Magnetic beads (SPRI). In brief, 4ul of SPRI beads was added to each 50ul sample. The mixture was vortexed and left to stand for 5 minutes.

Tubes were placed in magnetic racks and left for 3 minutes or until the solution cleared. The clear solution was removed taking care not to displace beads (containing DNA). 200ul of 70% ethanol was added without disturbing beads, allowed to stand for 30 seconds and then gently aspirated and discarded. This was repeated once and the beads were then left to dry on a heated block for 5 minutes at 37^{o}C. 32ul of EB Qiagen solution was added, vortexed and the mixture spun gently. This was left to stand for a further 5 minutes. Tubes were replaced into magnetic racks and solutions left to clear. Clear fluid containing DNA was carefully aspirated and retained in a fresh tube.

Checks were performed by electrophoresis using a DNA 1000 chip on an Agilent Bioanalyser, at each step of the library preparation process to ensure recovery of library and to check overall distribution of fragment sizes obtained (Figure 2.3).

2.2.3 Gel-size selection

50ml of a 2% agarose gel in 1X TAE (1g agarose) suitable for 1 mini-gel, was prepared for each library. Libraries were loaded after mixing with 6X loading dye and run according to the following parameters:

- 60V
- Duration 2 hours
- Chilled at 4 ^{o}C and replaced at 1 hour
- In 1X TAE buffer

For a 500bp library, gels were size-selected at ~700-750bp. Gel slices immediately above and below this bandwidth were also archived in case they were required for the future. Extraction of DNA was performed using Qiagen gel extraction kit. Electrophoresis using a High-Sensitivity chip on an Agilent Bioanalyser was performed to ensure that the library was captured and to ensure that the modal fragment size was in the order of 400-500bp.

Figure 2.3

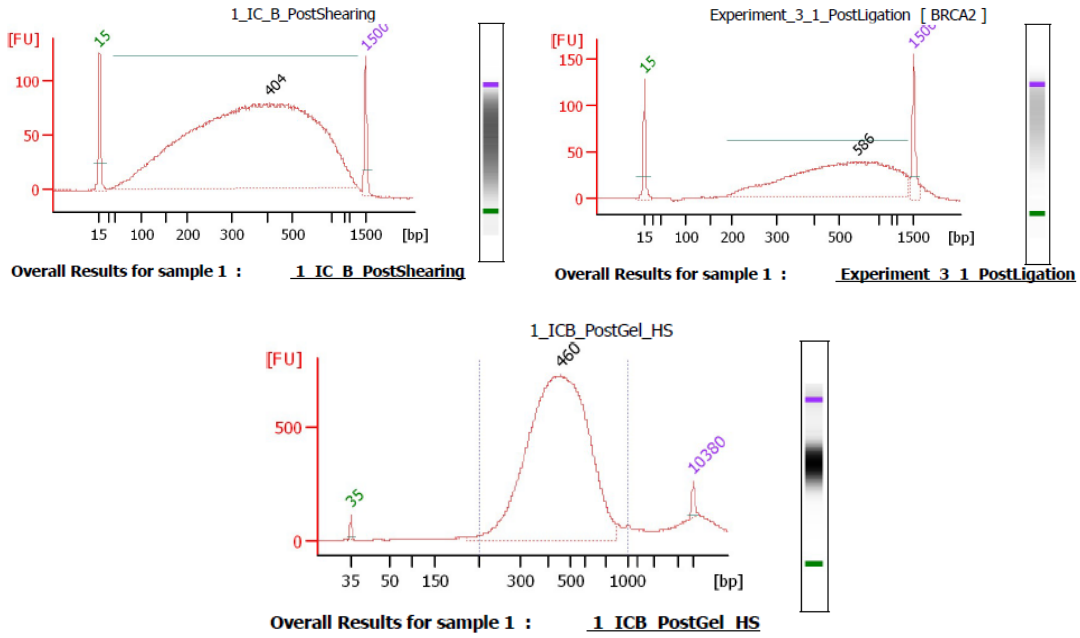


Figure 2.3: Typical Agilent bioanalyser traces (A) post-shearing (B) post end-repair, A-tailing and post adaptor-ligation (C) post gel size selection. Desirable features sought included the correct range of fragment sizes and the absence of a shoulder of remaining adapter or adapter-dimers.

2.2.4 Library quantification using quantitative PCR and sequencing

Illumina library quantification was performed using a real-time PCR assay in order to measure the quantity of fragments which were properly adapter-ligated and the result was used to determine the quantity of library necessary to produce the appropriate quantity of clusters on a single lane of the Illumina GAIIx or Illumina HiSeq. Flow-cell preparation was performed according to the manufacturer's protocol within the sequencing facility. Whole-genome sequencing was performed by the Wellcome Trust Sanger Institute core sequencing facility.

2.3 NEXT-GENERATION SEQUENCING

2.3.1 The principle of Illumina-based next-generation sequencing technology

The principle underlying next-generation sequencing technology (NGS) is not dissimilar to capillary sequencing. The bases of a small fragment of DNA are sequentially identified from signals emitted as each fragment is re-synthesised from a DNA template strand. However, in capillary electrophoresis sequencing, one averaged signal is obtained as a representation of a single sequencing reaction from many hundreds of DNA molecules which are mixed in solution. In contrast, NGS obtains many millions of signals across millions of reactions in a massively parallel fashion, with each reaction fixed to a single location on a sequencing chip. This advance enables rapid sequencing entire genomes, with the latest instruments capable of producing hundreds of gigabases of data in a single sequencing run.

2.3.1.1 The modified nucleotide

Standard Sanger capillary sequencing depends on the incorporation of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators. The chain-terminating nucleotides lacks a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, and results in termination of DNA strand extension and DNA fragments of varying length. The four types of dideoxynucleotide chain terminators are labeled with fluorescent dyes, each of which emits light at different wavelengths and are thus detectable in an automated fashion following separation by gel electrophoresis.

A development that permits the massively paralleled sequencing approach is the advent of the *reversible* terminator nucleotide. Here, the extension of each DNA molecule occurs base by base. Laser image detection of each fluorescently-labelled nucleotide is followed by cleavage of the fluorescent dye and reversal of the 3’-terminator. Addition of 3’-OH group then allows the extension of the same molecule. This is called sequencing by synthesis and is summarised in Figure 2.4.

Figure 2.4

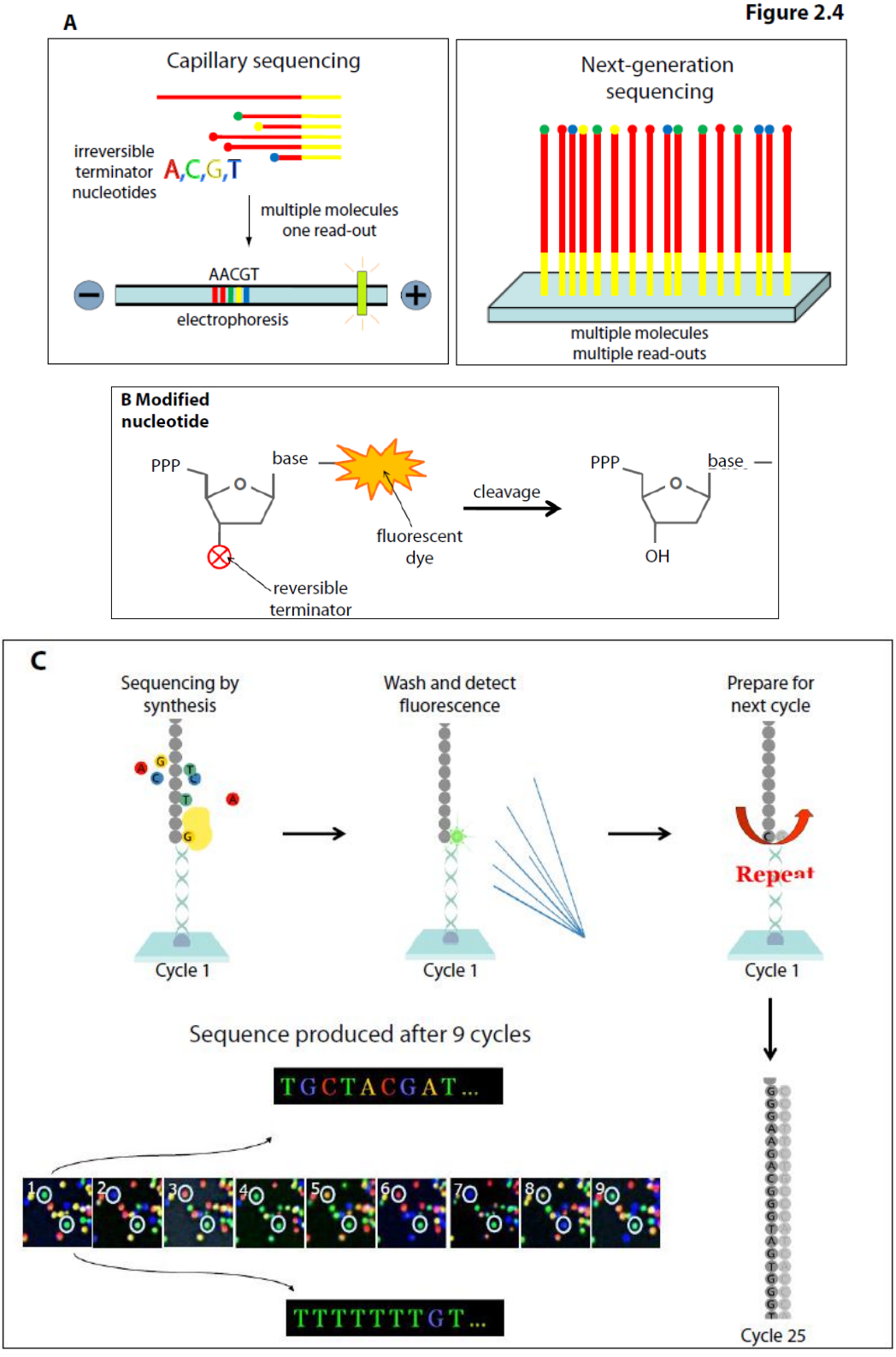


Figure 2.4: The principles of next-generation sequencing (A) In capillary sequencing, many thousands of DNA molecules in solution produce an averaged single read-out after electrophoretic separation. In NGS, many molecules immobilised on a flowcell produce independent read-outs thus increasing the scale of sequencing considerably. (B) The Illumina modified nucleotide contains a reversible terminator at the 3' end and a fluorescent dye which can be cleaved, to allow sequencing-by-synthesis. (C) Sequencing by synthesis: The appropriate and complementary reversible terminator nucleotide attaches at the start of cycle 1. After a wash step, photo-detection of the fluorescence allows identification of the nucleotide which has attached. Cleavage of the reversible terminator and the dye occurs and the cycle is repeated. After multiple cycles, a string of sequence is obtained. (D) Fluorescence colour sequence read-out of clusters of individual fragments on a flowcell. Images adapted from Harold Swerdlow, with thanks.

2.3.1.2 Variations on next-generation sequencing: targeted strategies

The ability to obtain hundreds of gigabases of sequencing data allows re-sequencing of whole genomes in a single experiment. However, variations of this approach can allow defined regions in a genome to be sequenced. This targeted approach involves steps that enrich a library for the regions of interest. The steps involved in making a NGS library for target enrichment are virtually identical to the steps involved in library generation for whole-genome sequencing, with the addition of a target enrichment step where labeled custom-designed oligonucleotide baits (for the desired sequence) is hybridized in solution to fragmented genomic DNA and pulled-down using magnetic beads, capturing the targeted sequence. This approach is commonly used, particularly for sequencing the coding sequences of the human genome and is referred to as exome sequencing. Although, exome-sequencing is not a central part of this thesis, four of the breast cancers in this study were also exome-sequenced and the data were used as a comparison dataset in Chapter 3.

2.3.2 Platforms used in this study

Cluster amplification and 108 or 100 base paired-end sequencing was performed on Illumina GAIIx genome analysers or Illumina Hiseq 2000 analysers respectively, as described in the Illumina Genome Analyser operating manual. Standard quality control metrics including error rates, percentage of purity-filter reads and the total number of bases sequenced were used to characterize process performance prior to alignment. The Core Sequencing pipeline generated data files that contained the sequenced reads and associated qualities (*qseq* files).

2.3.3 Quality control measures

Further quality control metrics for each library and each lane of sequencing were determined within the Cancer Genome Project and were as follows:

- The modal peak of fragment insert sizes for each library was required to be in the region of 400-500bp in size (Figure 2.5a)
- The GC plot (Figure 2.5b) should not show any skewing for or against GC
- The base qualities per cycle plots (Figure 2.5c) should show a high proportion of bases of a good minimum base quality of preferably 25 (see chapter 3 introduction for definition) and above

Figure 2.5

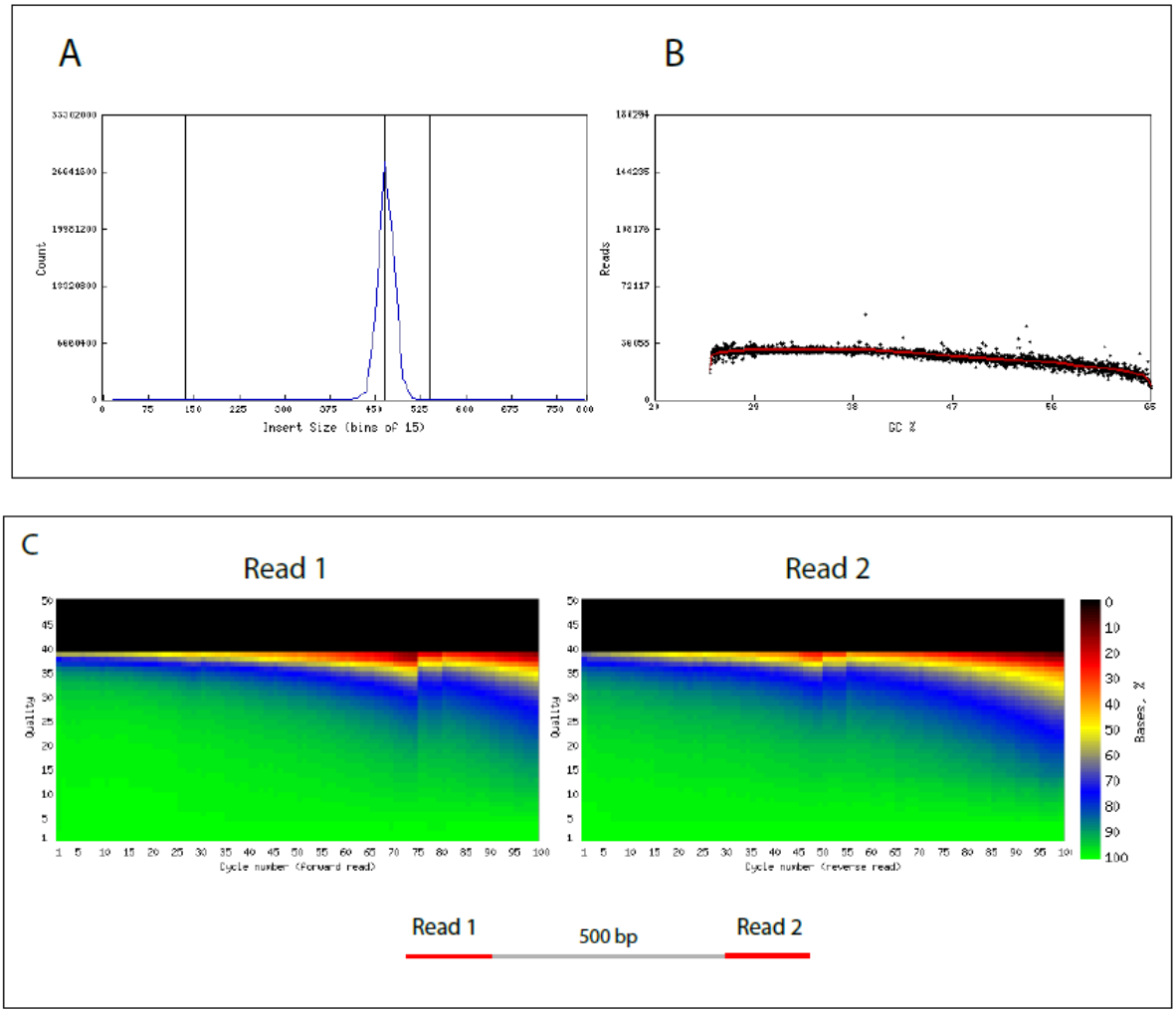


Figure 2.5: Library and sequencing QC metrics used within the Cancer Genome Project. (A) The modal peak of insert sizes should lie between 400-500bp with no evidence of "shoulders" suggestive of adapter contamination or smaller contaminating fragments in the library. (B) The horizontal axis shows the proportion of GC and the vertical axis demonstrates the number of reads. Because the genome differs in its GC content, a good library should have representation from all such regions showing a relatively uniform distribution of reads for all GC fractions. A pro-GC library would show a marked incline of the fitted slope and an anti-GC library would show a steep decline. (C) The base qualities per cycle of sequencing for both reads are demonstrated here. Green corresponds to 100% of bases with a minimum base quality which is on the vertical axis. The vast majority of both plots show green. The general decline towards the ends of reads is a well-known issue with Illumina sequencing. The step-wise change at cycle 75 in Read 1 and cycle 50 in Read 2 represents the laser-detector correction during the sequencing of both reads. This happens at a consistent time with the sequencing of each read.

2.3.4 The alignment of raw sequences to the reference human genome

The genomic DNA from a cancer sample is first fragmented into a library of small segments that can be uniformly and accurately sequenced in millions of parallel reactions. The newly identified strings of bases, called reads, are then reassembled using a known reference genome as a scaffold (resequencing), or in the absence of a reference genome (de novo sequencing). The full set of aligned reads reveals the entire sequence of each chromosome in the genomic DNA sample (Figure 2.8).

The raw data produced by the sequencers are contained in a *qseq* file which contains the quality scores, the precise location on the flow cell (lane and tile per lane), the sequencing run and the name of the sequencing machine used for each 100bp or 108bp sequence. Each *qseq* file was converted into a format that was more amenable to downstream manipulation called a *fastq* file. The *fastq* format essentially stored sequence information as concisely as possible but also included quality values for each of the bases sequenced in each read.

Figure 2.6

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Figure 2.6: The *fastq* format contains four lines per sequence: the first line begins with a '@' character and is a sequence identifier containing details regarding run, lane and position on the flowcell. The second line contains the raw sequence letters. The third line begins with a '+' character and is optionally followed by the same sequence identifier and any other optional description. The fourth line encodes the quality values in ASCII characters (character-encoding format) for the sequence in the second line, and contains the same number of symbols as letters in the sequence.

The sequencing data processing pipeline developed by the Cancer Genome Project starts with the short insert 108bp or 100bp paired-end reads and qualities in *fastq* format for all lanes and libraries generated for a single sample (either tumour or normal) and produces, after alignment to the reference human genome (NCBI37) using standard alignment software called Burrows-Wheeler Aligner (BWA), a single BAM file (http://samtools.sourceforge.net/SAM1.pdf) which represents the sample. The final binary-coded BAM file therefore stores all reads with well-calibrated qualities together with their successful alignments to the genome.

Figure 2.7

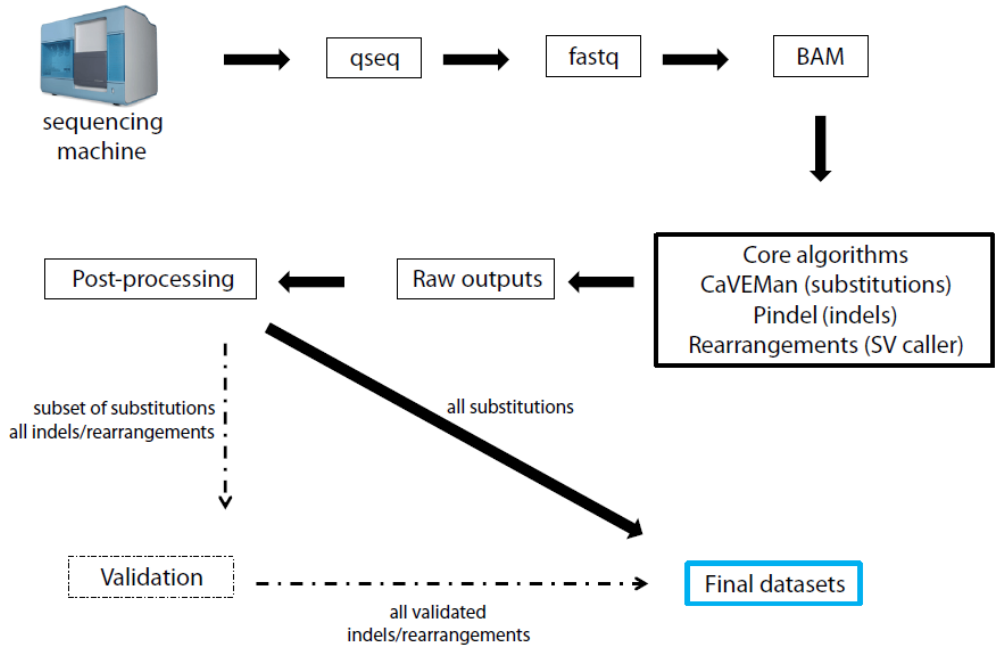


Figure 2.7: The generation of individual sample BAM files. Raw sequence data and quality scores were stored in *qseq* files. These were converted into *fastq* files which contained the sequence reads and qualities stored in a more concise format, amenable to economical storage and efficient computational manipulation. Data that passed QC were aligned back to the reference genome using BWA and a final BAM file was constructed for each library. BAM files were the input file for the subsequent step of calling somatic mutations. SV= structural variation.

2.4 THE PROCESS OF CALLING SOMATIC MUTATIONS

The sequenced reads from a cancer sample were aligned to the reference genome, and the sequenced reads from the matched normal sample were aligned separately to the reference genome. Therefore, two BAM files were generated per patient. The principle of calling somatic mutations involved identifying all variation in the cancer genome and the normal genome independently when compared to the reference genome, subtracting the normal variation (which would include all germline polymorphisms) to generate a final catalogue of somatic variation (Figure 2.8).

Figure 2.8

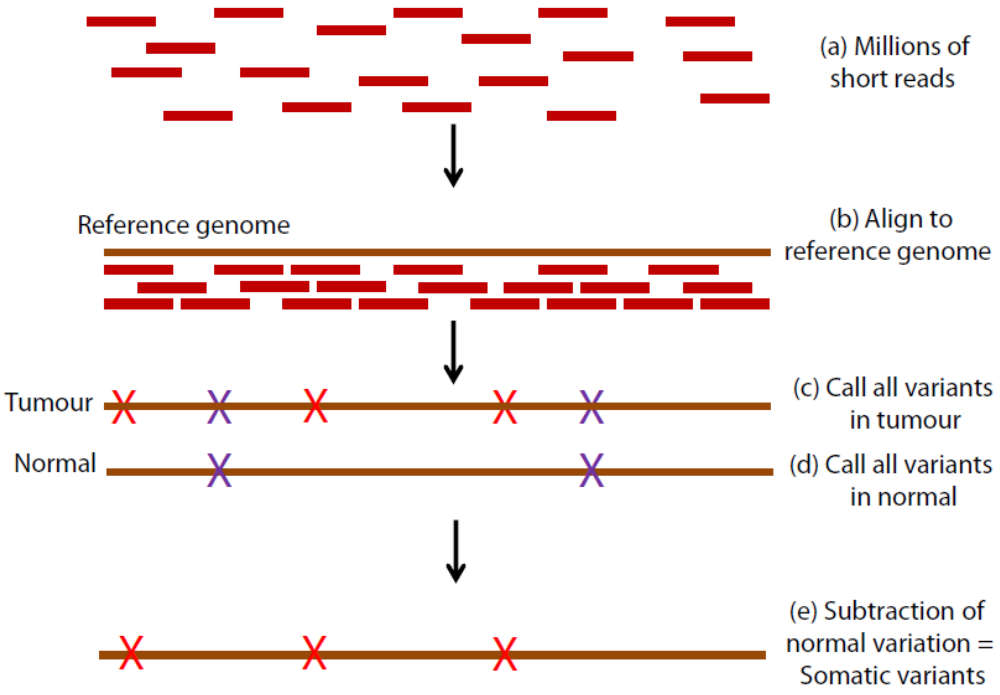


Figure 2.8: The principle of calling somatic mutations in cancer genomes. (a) Millions of short reads generated by sequencers were (b) aligned back to the reference genome separately in tumour and normal genomes. (c) All differences detected when comparing the tumour with the reference genome included somatic (red crosses) and germline variants (purple crosses). (d) All differences in the normal genome relative to reference genome were identified independently (purple crosses). (e) The germline polymorphisms in the normal genome were subtracted from the tumour genome to generate the catalogue of somatic variants for each breast cancer of each patient.

2.4.1 Genomic somatic mutation-calling

For the three mutation classes, substitutions, insertions/deletions and rearrangements, individual calling-algorithms were used. The input files for each of the calling algorithms were BAM files described in the previous section 2.3.4. Each mutation-caller generated a very large set of raw variant calls which comprised true somatic variants as well as a large proportion of false positive calls. In this section of the thesis, a very brief description of the principles of how each mutation-caller works is provided.

2.4.1.1 Substitutions

A bespoke algorithm, CaVEMan (unpublished) was used for calling somatic substitutions. Substitutions were identified, in principle, as alleles called in the tumour genome and not in the germline. Calls were made only from reads that mapped as linked pairs. Post-processing filters were developed to improve the specificity of mutation-calling and will be described in more detail in the next chapter. Copy number status (ploidy) and estimates of normal contamination from SNP6 data processed using ASCAT were used to enhance sensitivity and positive predictive value of substitution detection. CaVEMan will be described in more detail in the following chapter.

2.4.1.2 Indels

Insertions and deletions (indels) in the tumour and normal genomes were called using a modified version of Pindel (https://trac.nbic.nl/pindel/) 0.2.0 on the NCBI37 genome build (Ye et al., 2009). Pindel is a mutation-calling algorithm designed for the detection of small insertions/deletions, the breakpoints of large deletions, medium-sized insertions, inversions, tandem duplications and other structural variants at single-base resolution from next-generation sequencing data. It uses a pattern growth approach to identify the breakpoints of these variants from paired-end short reads. In this thesis, only the ability of Pindel to call small insertions/deletions was exploited, given that an alternative structural-variant caller was available and optimised for calling rearrangements.

During the preparation of a BAM file, all reads were mapped back to the reference genome. A subset of reads, however, mapped with either an indel within a read or could not be mapped although its paired-mate mapped correctly (unmapped singleton). This subset of reads was therefore potentially informative for indels. Pindel identified clusters of these informative reads, and used the mapped paired-mate to determine an anchor point in the reference genome. Having determined the anchor point and using *a priori* knowledge of the fragment insert size, Pindel worked out the orientation and the expected distance from the anchored read where the unmapped reads/reads containing the indel should be mapped. Pindel was able to split these informative reads into two (deletion) or 3 (insertion) smaller fragments, and attempted to align these in independent portions (Figure 2.9). Pindel called variants in tumour and normal separately but did not do a formal comparison. Post-processing filters were put in place to formally assist in the identification of somatic variants.

Somatic indels were required to be present in 5 reads or more in the tumour and not present in the matched normal sample. Variants were also screened against a panel of normal samples and were excluded if present in at least 5% of reads in at least 2 samples from this panel. Despite additional optimisation, the false positive rate in Pindel-called insertions/deletions remained high ~30%.

Therefore, all indels called by Pindel were validated and only validated variants were reported in this study.

2.4.1.3 Copy number

Copy number was determined using the Affymetrix SNP6.0 array for each of the twenty-one breast cancer samples. An informatic tool called "ASCAT" or allele-specific copy number analysis of tumours was used to estimate the fraction of aberrant cells and the tumour ploidy, as well as whole-genome allele-specific copy number profiles. ASCAT is an algorithm (Van Loo et al., 2010) that has considered and modeled the following two properties in cancer; that tumours often deviate from a diploid state (Holland and Cleveland, 2009; Rajagopalan and Lengauer, 2004) and that cancers are likely to comprise multiple populations of both tumour and non-tumour cells (Witz and Levy-Nissenbaum, 2006). ASCAT was therefore able to provide these estimates (Table 7.2) in the twenty-one breast cancers. These estimates were also used to optimise substitution-calling by CaVEMan (see section 3.3). Whole-genome allele-specific copy number profiles allowed regions of gains, losses, amplification and loss of heterozygosity (LOH) to be identified.

Figure 2.9

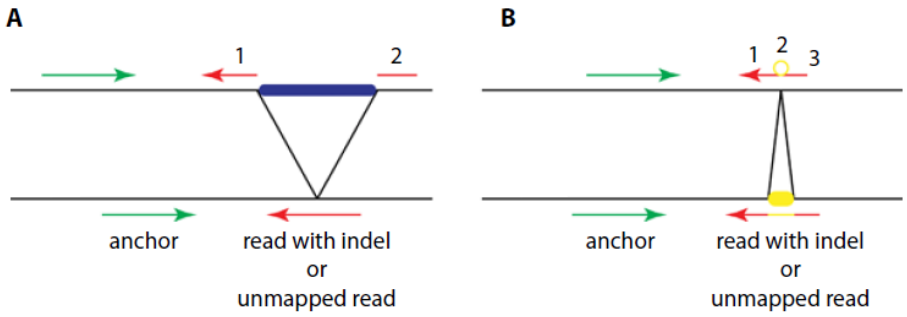


Figure 2.9: The basic principle underlying indel detection. Pindel detects simple deletions (A) and insertions (B) at nucleotide-level resolution (obtained from https://trac.nbic.nl/pindel/). Pindel identifies paired reads that are mapped but contain indels or paired reads with only one end mapped. Pindel uses the mapped read of the pair (green arrows) to determine an anchor point on the reference genome. A sub-region can then be located in the reference genome relative to the anchor read, where Pindel breaks the informative reads into 2 (deletion) or 3 (short insertion) fragments and maps these terminal fragments separately.

2.4.1.4 Rearrangements

Structural variants were called from discordantly mapping paired-end reads from short insert data using MAQ (Mapping and Assembly with Quality) alignments (Campbell et al., 2008; Stephens et al., 2009) (Figure 2.10 for summary). A set of optimisation filters and validation reduced the dataset considerably. Therefore, in order to improve sensitivity of detection, additional candidate structural variants were sought from within the proximity of copy number changes in the following way. All non-telomeric and non-centromeric coordinates of copy number changes were obtained from SNP6 data processed via ASCAT (Van Loo et al., 2010). Rearrangements close to copy number segmentation breakpoints were considered to be somatic if:

- copy number changes were identified for both rearrangement breakpoints and the sum of the distances between the rearrangement breakpoint and the copy number change was below 400kb or if
- a copy number change was identified in conjunction with only one rearrangement, then the distance between the rearrangement breakpoints and the copy number changes was less than 20kb.

If multiple rearrangements were identified for any copy number change, rearrangements closer to a copy number change were preferred over rearrangements further away.

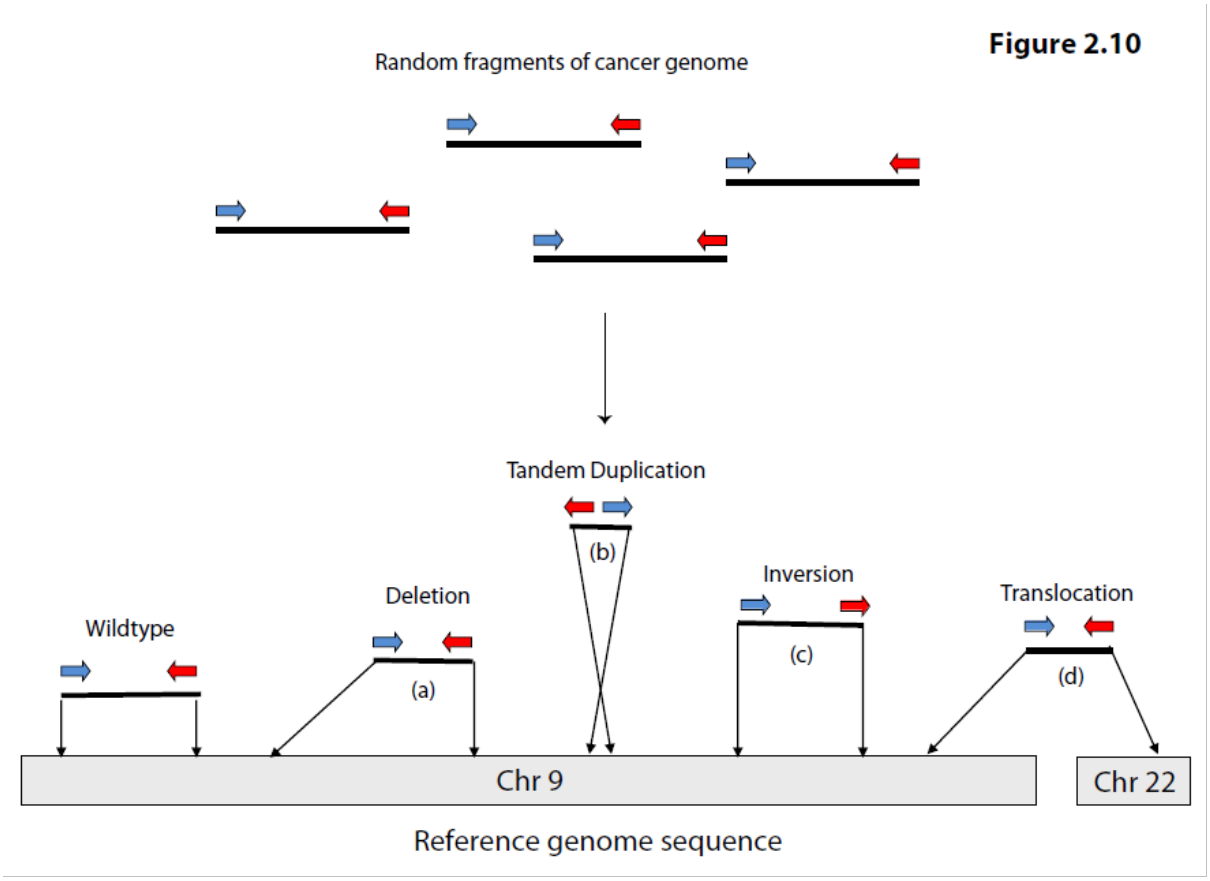


Figure 2.10: The classification of somatic rearrangements. Reads which were mapped at unexpected distances or in the wrong orientation were identified as discordantly mapping reads. Rearrangements were classified according to whether they were a (a) deletion: reads were closer together than expected and there was an associated copy-number change for variants > 200kb (b) tandem duplication: reads were further apart than expected and in the wrong orientation and associated with a copy number change for variants > 200kb (c) inversion: reads were not mapping in the appropriate orientation (d) translocation: reads were mapping to different chromosomes. *Amplicon-associated rearrangements involved reads within a region of high copy number, but are not depicted in this figure.

2.5 VALIDATION

In order to gain insight into the positive predictive value of the mutation-calling algorithms, validation experiments of a subset of putative somatic substitutions and all insertion/deletions and rearrangements were performed. Validation of putative substitutions was performed via Roche pyrosequencing (see section 2.5.2) in 20 tumour-normal pairs and capillary sequencing in 1 tumour-normal pair (PD3890a). All coding substitution variants and a random assortment of intronic and intergenic variants were selected for validation to make up to ~400 PCR products per sample. In addition to the set of variants selected for validation genome-wide, validation was also targeted to several hundred substitutions involved in regions of hypermutation and dinucleotides. The positive predictive value of the calling of substitution variants from the Illumina sequence reads was determined from the proportion of calls confirmed as somatic when sequenced on this orthogonal platform.

2.5.1 Capillary sequencing

Validation of variants was attempted by capillary re-sequencing of the tumour and normal pair. Capillary sequencing failed in ~20% variants. Two attempts at PCR validation for each variant were attempted. Somatic variants were required to be present in the tumour sample and absent in the normal sample. As mentioned previously, a capillary sequencing trace represents an averaged signal obtained from many thousands of DNA molecules in solution. It is believed that a variant has to be present at a sufficient proportion (> 10% of tumour cells) to be detectable by this method. It is therefore acknowledged that variants which are present at a low mutant burden may escape detection by this method of validation and may represent false negative calls.

2.5.2 Roche 454 pyrosequencing

Due to the large number of substitution variants, an alternative large-scale sequencing approach was favoured over the time-consuming, variant-by-variant approach of capillary sequencing. Similar to Illumina Sequencing, 454 sequencing involved a large-scale parallel sequencing approach and was able to generate roughly 400-600Mb of DNA per 10-hour run on a Genome Sequencer FLX using the GS FLX Titanium reagent series. 454 sequencing, also known as pyrosequencing, relies on fixing nebulised and adapter-ligated DNA fragments to small DNA-capture beads in a water-in-oil emulsion. The DNA fixed to these beads was then amplified by PCR. Each DNA-bound bead was placed into a ~29µm well together with a mix of sequencing reagents on a 454 PicoTiterPlate, which was essentially a fibre-optic chip. As a validation strategy, 454 pyrosequencing provided an alternative sequencing platform for targeted regions in the genome, allowing sequencing to high coverage. This alternative meant that variants were not subjected to the same systematic biases of Illumina sequencing during the validation step.

2.5.2.1 Library-preparation

Primers were designed to generate PCR amplicons for pyrosequencing of approximately 275-425bp. The PCR reaction was prepared in the following way:

No of samples	1
Whole-genome amplified DNA at 8ng/ul(ul)	4.5
Mixed primers at 4ng/ul (ul)	3
Buffer 10X (ul)	3.58
Taq polymerase(ul)	0.35
dNTPs at 1mM (ul)	3.58
Total volume (ul)	15

The PCR program was as follows:

- 95°C for 15 minutes
- 95°C for 30 seconds
- 60°C for 30 seconds
- 72°C for 30 seconds
- Repeated for 30 cycles
- 72°C for 10 minutes
- 4°C forever

Following the PCR reaction, the enzymes were inactivated using the standard protocol (ExoSAP-IT (Affymetrix)):

No of samples	1
PCR product (ul)	15
Reaction buffer (ul)	3
Dilution buffer (ul)	3.58
Exonuclease 20,000 U/ml(ul)	0.05
Antarctic phosphatase 25,000 U/ml (ul)	0.04
Water (ul)	8.9
Total volume (ul)	26

Using the following programme:

- 37°C for 30 minutes
- 80°C for 15 minutes
- 10°C forever

DNA was purified using Agencourt AMPure magnetic beads (using a similar protocol to that described in Section 2.2.2) and submitted to the 454 sequencing facility for adaptor ligation and sequencing on a Roche 454 Genome Sequencer FLX.

2.5.2.2 Raw data handling

Raw pyrosequencing files for tumour and normal samples were aligned to the reference human genome (NCBI37) using the genome alignment software Burrows-Wheeler Alignment with the addition of Smith-Waterman alignment to allow for longer read lengths (BWA-SW). Similar to Illumina reads, raw 454 pyrosequencing data was converted into pileup files for analysis (Figure 2.11).

Figure 2.11

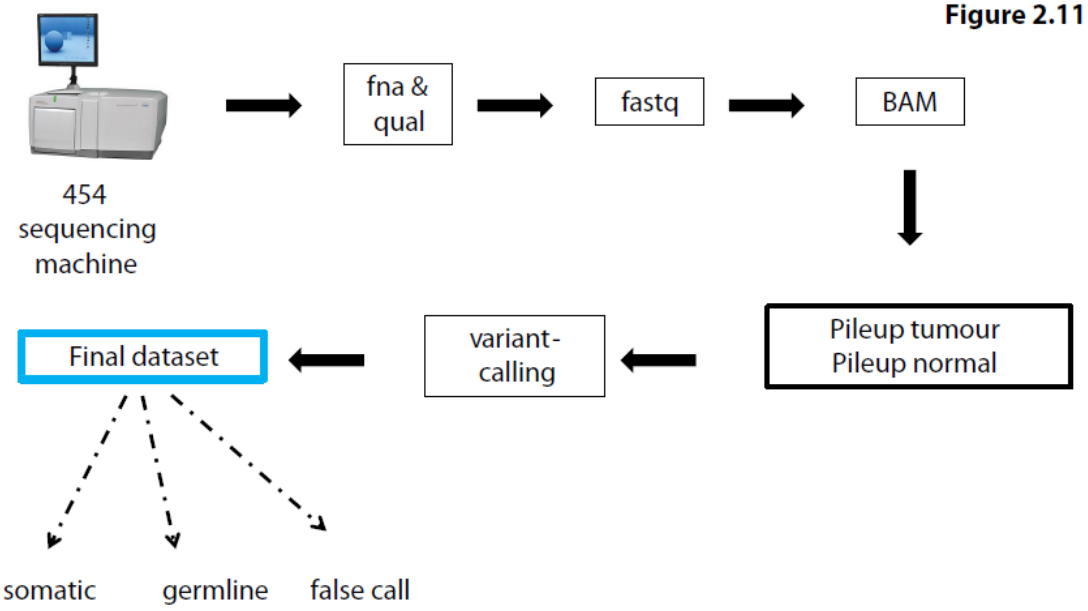


Figure 2.11: Data handling of 454 pyrosequencing output files. Sequence (fna) and quality (qual) files are converted into fastq and then BAM files, similar to the workflow for Illumina sequencing. Two separate pileup files are generated and variant-calling is performed across these two pileup files.

Pileup files were generated for each tumour and matched normal sample separately and variants were called as somatic if they were present in tumour and not in the normal. At least 25 reads of mapping quality of 20 and above and base quality of 25 and above were required to report each variant. To be considered as somatic, variants were required to be present in at least 5% of the reads in the tumour and not in the normal, or if present at a low mutation burden of < 5%, required chi-squared testing to assist in confirmation of somatic status. This imposition of relatively strict criteria could potentially generate false negative calls (true somatic variants called as tumour wild-type) resulting in an underestimation of the specificity of substitution-calling.

For pyrosequencing data, an average coverage of ~657X was achieved for each validated variant. A total of 6334 variants were amplified, of which 5561 met the aforementioned criteria. 4120 variants were found to be somatic, 26 were germline SNPs and 1395 did not show any evidence of the variant in the tumour or normal. The relationship between the variant allele fraction of the 454 experiment mirrored the variant allele fraction of the Illumina experiment in general (Figure 2.4).

Figure 2.12

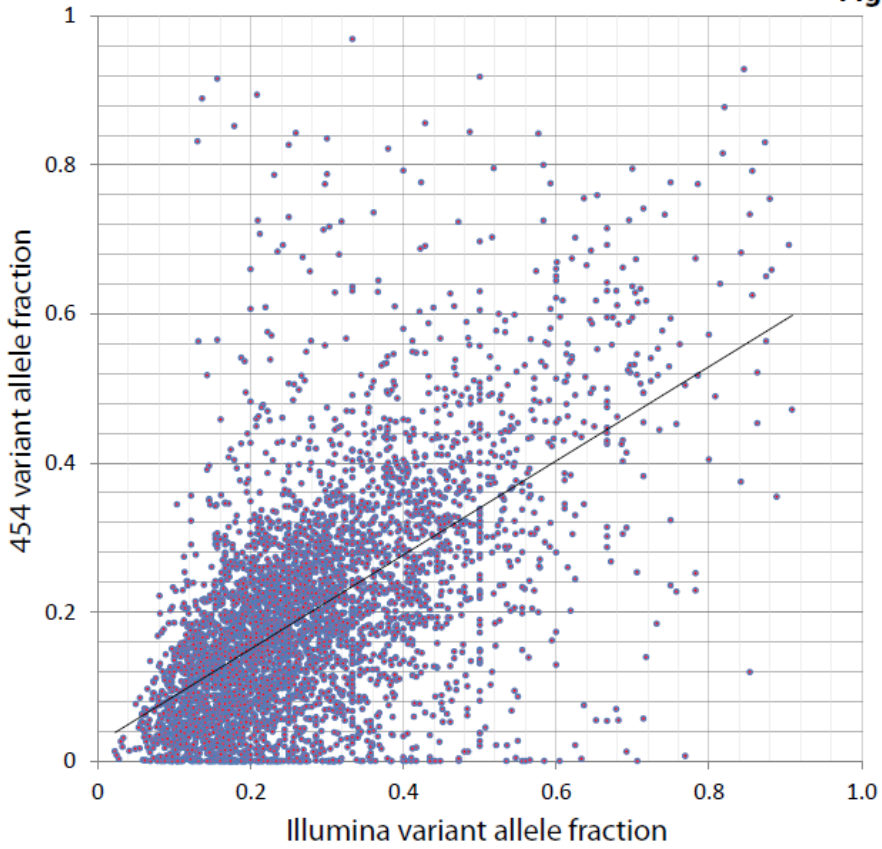


Figure 2.12: The variant allele fractions of the variants validated by 454 pyrosequencing were plotted against the variant allele fractions of the Illumina sequencing experiment to demonstrate that in general the representation of each variant in both experiments were correlated and therefore likely to represent the biological fraction of cells carrying the reported variant in the cancer (r=0.77).

2.5.3 Validation of somatic rearrangement

Structural variants were confirmed by custom-designed PCR across the rearrangement breakpoint (Campbell et al., 2008) or by local reassembly. Structural variants which were PCR-amplified were identified as putative somatic structural variants if a band on gel electrophoresis was seen in the tumour and not in the normal, in duplicate (Figure 2.13). Putative somatic structural variants were then capillary sequenced. Amplicons which were successfully sequenced were aligned back to the reference genome using Blat, in order to identify breakpoints to basepair resolution (http://genome.ucsc.edu/cgi-bin/hgBlat).

Figure 2.13

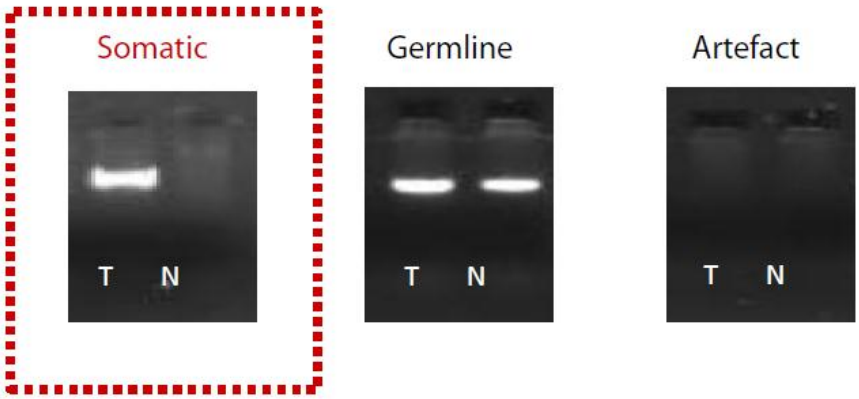


Figure 2.13: The validation step for putative somatic structural variation involved custom-designing primers to putative breakpoints and amplifying both tumour and normal DNA. A band appearing in the tumour (T) lane and not in the normal (N) lane, in duplicate, was taken for capillary sequencing.

For local reassembly, candidate rearrangements in regions of interest had been previously identified as rearrangements in close proximity to copy number changes. Discordantly mapping read pairs that were likely to span breakpoints, as well as a selection of nearby properly-paired reads, were grouped for each region of interest. Using the Velvet de novo assembler (Zerbino and Birney, 2008), reads were locally assembled within each of these regions to produce a contiguous consensus sequence of each region (Figure 2.14). Nearby properly-paired reads were added to increase coverage and to enlarge the resulting contigs. Heterozygous rearrangements, represented by reads from the rearranged derivative as well as the corresponding non-rearranged allele (Figure 2.14D), were instantly recognisable from a particular pattern of five vertices in the de Bruijn graph (a mathematical method used in de novo assembly of (short) read sequences) of component of Velvet (Figure 2.14C). Exact coordinates and features of junction sequence (e.g. microhomology or non-templated sequence) were derived from this. The exact breakpoints were identified by aligning to

the reference genome as though they were split reads. This local reassembly method continues to be under development at the present time.

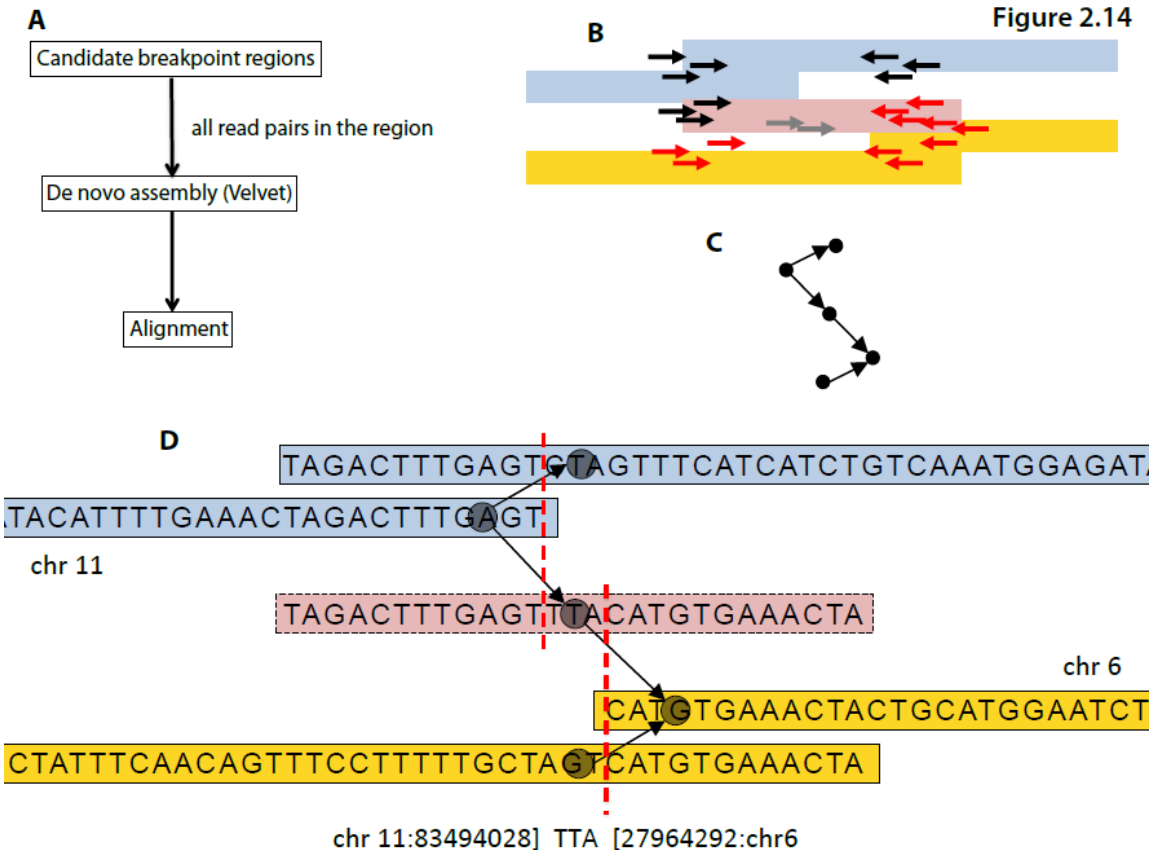


Figure 2.14

Figure 2.14: Validation of somatic rearrangement by local reassembly (Brass Phase II, under development). (A) Workflow of process of validation by de novo assembly. (B) Principle of local reassembly using a somatic interchromosomal translocation as an example. Read pairs are represented by two arrows facing each other on the same horizontal line. Black/red pairs in the centre represent pairs with ends on different chromosomes, i.e. are informative for a rearrangement. Nearby properly-paired reads were included for each region of interest (exclusively black pairs and exclusively red pairs). Reads with one end unmapped (grey arrows) spanning the breakpoints were also included. The coloured rectangles represent the contigs that Velvet was able to decipher: the two blue contigs represent one chromosome (11), and the yellow contigs represent another chromosome (6). The middle pink rectangle is the contig that reports the rearrangement. (C) The pattern of 5 vertices expected from the De Bruijn graph for successfully mapped rearrangement breakpoints. (D) Deciphering the rearrangement. Blue and yellow contigs were reported by properly paired reads. The pink contig reports the rearrangement. A successfully reassembled rearrangement breakpoint shows the pattern of 5 vertices. The breakpoint coordinates can be read from the pink contig. The lateral ends of the pink contig can be mapped back to the reference genome (chromosome 11 on the left and chromosome 6 on the right) until ambiguity is reached towards the middle of the pink contig. In this case, there is a stretch of non-templated sequence (TTA) in the middle of the breakpoint. The other possibility is that the two highlighted contigs meet in the middle and overlap by a few bases, which corresponds to a microhomology at the breakpoint.

2.6 STATISTICAL MEASURES

2.6.1 MONTE CARLO SIMULATION OF SUBSTITUTIONS

In order to assess the likelihood of some of the features or mutational patterns identified in the analyses which will be described in the following chapters, Monte Carlo simulations were performed for each cancer genome. The mutation prevalence of each mutation type (C>A, C>G, C>T, T>A, T>C, T>G) was obtained for each chromosome of each cancer genome. For each genome, 1000 simulations were then performed by generating mutations *in silico,* at the observed mutation rates. For each simulation, a variety of *in silico* parameters could be obtained and compared to observed features in each cancer genome. None of the simulations yielded mutational features according to the observed patterns, hence $p<0.001$ for the observed enrichment of each of those observed phenomena for each cancer genome.

2.6.2 GENERALISED LINEAR MIXED EFFECTS MODEL

Generalised linear models represent a class of fixed effects regression models for different types of dependent variables (including count data). Fixed effects models assume that all observations are independent of each other and are not appropriate for analysis of several types of correlated data structures, in particular, clustered data (where observed subjects are nested within larger units). Random effects can be added into the regression model to account for the variation in the correlation of the data. The resulting model is referred to as a Generalised Linear Mixed Effects Model which includes the usual fixed effects plus the random effects.

In order to examine correlations between mutation prevalence and gene expression as well as to consider transcriptional strand biases, a Generalised Linear Mixed Effects model was used for the analysis. For each mutation-type, the mixed effects comprised

- fixed effects: properties that were always present which here were the transcriptional strands and the expression levels
- random effects: the inter-sample variation.

The overall fitted curve for each mutation-type represents the combined effects across all seventeen cases for which there was expression data. The relationships described in Chapter 6 are based on the model performed here.

2.6.3 KOLMOGOROV-SMIRNOV TEST

The Kolmogorov–Smirnov test (K–S test) is a general nonparametric test which is sensitive to differences in shape of distribution functions between two groups and was therefore used to compare the empirical distribution functions of two groups. The null distribution of this statistic was calculated under the null hypothesis that the groups were drawn from the same distribution. The distributions considered under the null hypothesis were unrestricted, continuous distributions. The K-S test was used to compare differences in distribution functions between observed outcomes and expected outcomes assuming the latter occurred due to random chance (see chapter 7, sections 7.2 and 7.4).

CHAPTER THREE: OPTIMISATION OF MUTATION-CALLING IN ORDER TO OBTAIN A CURATED CATALOGUE OF SOMATIC SUBSTITUTIONS FOR DOWNSTREAM ANALYSES

3.1 INTRODUCTION

Obtaining raw sequence data for twenty-one breast cancer genomes was only the beginning of a complex process that required multiple iterations of computational processing in order to translate raw sequence data into a comprehensive list of somatic variants. In order to excavate the cancer genome for patterns in somatic substitutions, insertions/deletions and rearrangements, it was critical to obtain a set of high-confidence mutations with high specificity i.e. a low false positive rate.

Calling single nucleotide substitution and insertion/deletion variants from short-read sequencing data can be problematic in general but is particularly so in cancer genome sequences. A general issue associated with sequencing of short read data includes decline in sequencing qualities at lattermost cycles of the sequencing-by-synthesis process. In addition, certain sequencing motifs (for example strings of G bases (–GGGG)) have been known to cause an increase in polymerase errors resulting in sequencing errors immediately following such motifs (Abnizova et al., 2012). Inaccuracies in base assignment following photo-laser capture can also occur. The confidence in a base call made during the sequencing process is simply the probability estimate of that base call being a true nucleotide. The likelihood of the accuracy of a base call is reflected in the base quality score or Phred score. Base qualities can therefore be taken into account when considering variant calls. Finally, accurate mapping of short-read data to the reference genome can be hampered by the large proportion of repetitive sequence in the human genome. Errors in mapping can be seen, for example, as excessive coverage in certain regions of the genome due to inaccurately assembled reference genome sequence given by sites of low complexity. Non-unique mapping is reflected in mapping qualities and like base qualities, can be taken into consideration when curating catalogues of variants.

In whole-genome sequencing family-based studies of the germline, relatives enable the efficient elimination of errors based on Mendelian inheritance patterns and knowledge of parental haplotype blocks. This has, in fact, permitted successful identification of genes underlying a host of inherited disorders despite sequencing very few individuals, for example Miller syndrome and Freeman-Sheldon Syndrome (Ng et al., 2009; Roach et al., 2010). Furthermore, the digital nature of next-generation sequencing technology provides additional means of supporting variant-calling in the germline. For every base in the genome, coverage of 40-fold would mean that sequencing information from 40 DNA molecules is available at that particular genomic coordinate. A heterozygous mutation in the germline would be expected to be present in approximately 50% of

reads for a diploid genome and a homozygous mutation should be present in 100% of reads (Figure 3.2). Using this reasoning, sequencing artefacts that arise in just a small proportion of reads, for example, could be filtered from the variant dataset.

Figure 3.1

(a) Q = -10 * log(E)

(b) mE = 10 ^ (-mQ / 10.0)

Figure 3.1 Phred score and mapping qualities. (a) A base quality score or Phred score is a score of an estimate of a base call being the true nucleotide. The probability that a base call is wrong is called an error probability. If the error probability of a base call is E, then the Phred base quality score is ***Q*** where is as seen in the figure. If the quality of a base call is 30, the probability that it is wrong is 0.001. Therefore, on average 1 in every 1000 base calls with Q=30 is erroneous. (b) A similar principle applies for mapping qualities. Each read alignment is a probabilistic estimate of the true alignment. If the mapping quality of a read alignment is ***mQ***, the probability ***mE*** that the alignment is wrong is as above. Once again, one in every 1000 read alignments with mapping quality of 30 will be wrong on average.

The approach of using Mendelian-based elimination of errors cannot be applied directly to cancer genome sequencing. On top of the general problems associated with the next-generation sequencing process and mapping of short-read sequencing data, digital calling of variants in cancers is plagued further by issues such as intra-tumour heterogeneity, contamination by normal cells and marked abnormalities of ploidy. Unlike calling mutations in the diploid human germline genome, calling of variants in cancer requires consideration of these additional parameters in order to maximize the likelihood of detection (Figure 3.2). This however, may come at a cost on the specificity or the false positive rate of variant-calling.

In the last few years, multiple substitution-calling algorithms have been published although many of these result in an extremely large numbers of variants which turn out to be errors or false positive calls. Given the high false positive rate in studies utilizing short-read sequencing technology for the detection of somatic single-nucleotide variants, independent mid- to large-scale validation experiments has been obligatory, (preferably) on an orthogonal platform in order to avoid reproducing systematic sequencing artifacts. For instance, more than 500 somatic substitutions in a lung cancer were validated using mass spectrometry (Lee et al., 2010) whereas other studies re-sequenced hundreds of substitution variants using Sanger sequencing (Pleasance et al., 2010a; Pleasance et al., 2010b). These validation experiments rapidly become as costly as the initial discovery experiment and are labour-intensive.

Across the cancer genomics community, filters have been developed and applied to raw variant-called datasets in order to reduce the false positive rate. However, there is little consensus on what

filters should be used and at what threshold applied. Additionally, the extent to which filters discard true variants has not been formally assessed.

To the best of my knowledge, there is only one example of a study which has documented the challenges of variant-calling from short-read data and provided some detail on additional processing of raw data in order to obtain a set of high-confidence substitutions (Reumers et al., 2012). In that study, post-processing filter optimisation was performed on whole genome sequences in the germline obtained from a pair of identical twins. The authors reasoned that shared variants were more likely to be real whereas discordant variants were more likely to be false positive calls. Filters were optimised to remove as many discordant single nucleotide variants and as few shared ones as possible. There were drawbacks with this analysis. First, systematic sequencing artifacts with a predilection for certain sequence motifs were precisely the sort of systematic false positives that could be shared between twin genomes. Their metric for measuring the effectiveness of filters, which was based on the ratio of shared versus discordant variants, was therefore systematically overestimated. Second, aggregation of the fraction of the genome removed across the filters meant that up to 32% of the genome could be removed. Third, and acknowledged by the authors, calling of variants in tumour-normal pairs of ovarian cancer was attempted, and although generally was able to call variants, was plagued by difficulties in over-calling mutations at regions of extremes of ploidy (zones of amplification and loss-of-heterozygosity). Furthermore, validation of their method concentrated on coding regions of these cancer genomes. Coding sequences are generally more unique, show more sequence complexity and are less troubled by false positives than intronic/intergenic regions, and again this validation step is likely to have overestimated the effectiveness of their filters.

In this chapter, the challenge of distinguishing true mutations from errors in whole genome sequences is deliberated using substitution-calling as a foremost example, and the solutions that have been created, in the form of post-processing filters, are described.

Figure 3.2

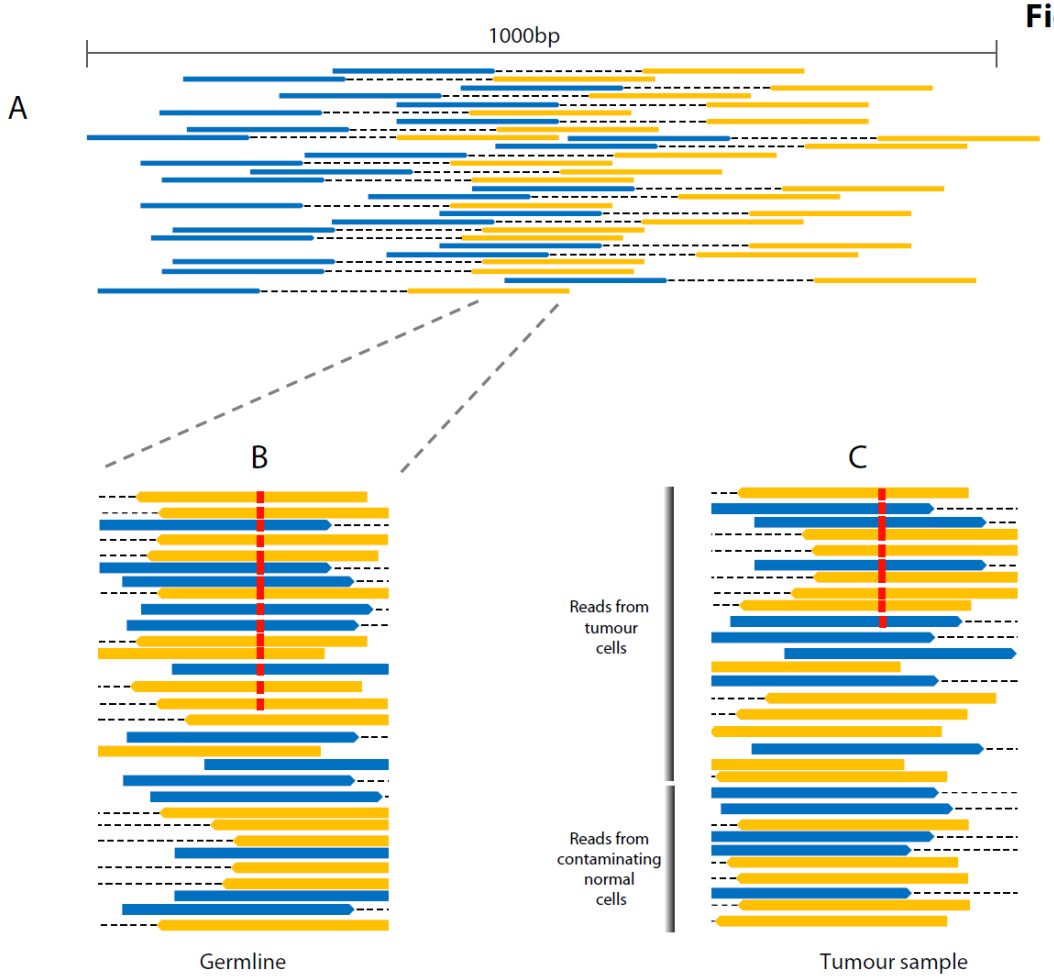


Figure 3.2: Differences in calling variants in a germline genome and a tumour genome. (A) Blue and yellow reads joined by a dotted line represent forward and reverse reads respectively of a 500bp fragment. (B) A higher resolution depiction of (A), which is a germline sample, showing 30-fold coverage of reads in the region of interest. The red marks represent a variant allele which is different to the reference genome. This heterozygous SNP in the diploid germline genome is seen in approximately 50% of reads or has a variant allele fraction of 0.5. (C) This higher resolution schematic of a tumour sample also has 30-fold coverage but has 1/3 of reads originating from *contaminating normal cells*. In this region which is diploid in the tumour, the somatic variant is a heterozygous mutation and is present at a lower variant allele fraction (when compared to the germline genome) of 0.33. However, if the variant allele fraction of a true variant is lower than expected for the level of ploidy and normal contamination, then this may be taken as evidence of a somatic mutation in a subclonal population (*intra-tumoural heterogeneity*). In contrast, a *polyploid* region where a somatic variant is present on only 1 of multiple alleles will be present at a much lower variant allele fraction.

3.2 THE METRICS USED FOR THIS ANALYSIS

In order to track the improvements in the performance of the mutation-calling and post-processing procedure, some statistical measures of the performance of a binary classification test (where a mutation is called as somatic or not) was required. Sensitivity, or the recall rate, measures the proportion of true positives which are correctly identified (e.g. the percentage of affected people who are correctly identified as having the condition). Specificity measures the proportion of negatives which are correctly identified (e.g. the percentage of healthy people who are correctly identified as not having the condition). An alternative metric which is easier to calculate for the purposes of this analysis is the positive predictive value (PPV). This metric measures the proportion of positives which are correctly identified. Specificity and the positive predictive value are sometimes used interchangeably although in theory reflect subtly different concepts.

The two measures of sensitivity and specificity are closely related to the concepts of type I and type II errors. The perfect algorithm would have 100% sensitivity and specificity. However, for any test, there is usually a trade-off between the measures. In this thesis, it was in theory impossible to measure the sensitivity given that a priori knowledge of mutations in any given cancer was not known. However, an attempt was made to infer sensitivity from a cross-comparison with a high-confidence set of mutations produced by an alternative substitution-calling algorithm produced by Illumina© as well as a cross-comparison with whole exome sequences for 3 samples. The positive predictive value (PPV) was the metric that was used to track the progress and improvements in mutation-calling and post-processing.

3.3 CaVEMan IS A BESPOKE SUBSTITUTION-CALLING ALGORITHM

An in-house bespoke substitution-calling algorithm, CaVEMan (Cancer Variants Through Expectation Maximization) was used for calling somatic substitutions. CaVEMan is a naïve Bayesian probabilistic classifier which utilizes the expectation maximization (EM) algorithm and is designed for calling substitution variants in new sequencing technology reads. Given prior information regarding reference and variant alleles, copy number status or ploidy, fraction of aberrant tumour cells present in each cancer sample and quality scores relating to sequencing and mapping, CaVEMan generates a probability score for potential genotypes at each genomic position. CaVEMan requires mapped, paired-end reads in the form of a sorted and indexed BAM file for the tumour and matched normal samples. An indexed reference sequence in FASTA format is also a prerequisite.

There are two main steps in the core CaVEMan algorithm. The first *maximization* or *M*-step generates a prior depiction of each genomic position by gathering data from all valid reads (reads that are properly paired and not marked as duplicates) that are available at that coordinate. These data or covariates include read information (1^{st} or 2^{nd} read of a pair), mapping qualities of the reads, lane information, base qualities, the expected reference allele (A, C, G or T), the variant allele (A, C, G or T) and the position of the variant in the read. CaVEMan iterates through each genomic position generating a multi-dimensional array of information in order to build an "error profile" for each coordinate.

The second *expectation* or *E*-step uses this profile to generate a probability for each possible genotype at this position, again iterating through each position in the genome. A number of parameters can be set to enhance the accuracy of the probability estimates in cancer. The degree of contamination from normal cells as well as the ploidy of each section of the genome (both obtained from SNP6 copy number analysis) can be provided to CaVEMan in order to enhance mutation-calling. In order to produce a set of raw variants, other parameters that are factored into this step include mutation rate (6e-6), SNP rate (1e-3), reference bias (0.95), a SNP probability cutoff (0.95) and a mutation probability cut-off (0.8). At the end of the *E*-step, a list of potential genotypes at each base is obtained. Three output files are generated following this process; a "raw substitutions" output file for those variants in which the sum of the genotype probabilities exceeds the mutation probability cut-off, a "raw SNPs" file if the sum of the SNP genotype probabilities exceeds the SNP probability cut-off and an "uncategorised" file for variants which meet neither of these criteria.

On average, tens of thousands of variants per raw output of breast cancer sample were obtained (Table 3.2). However, these variants were unlikely to all be true somatic variants. In the following section, the development of filters in order to remove false positive calls (post-processing) will be described.

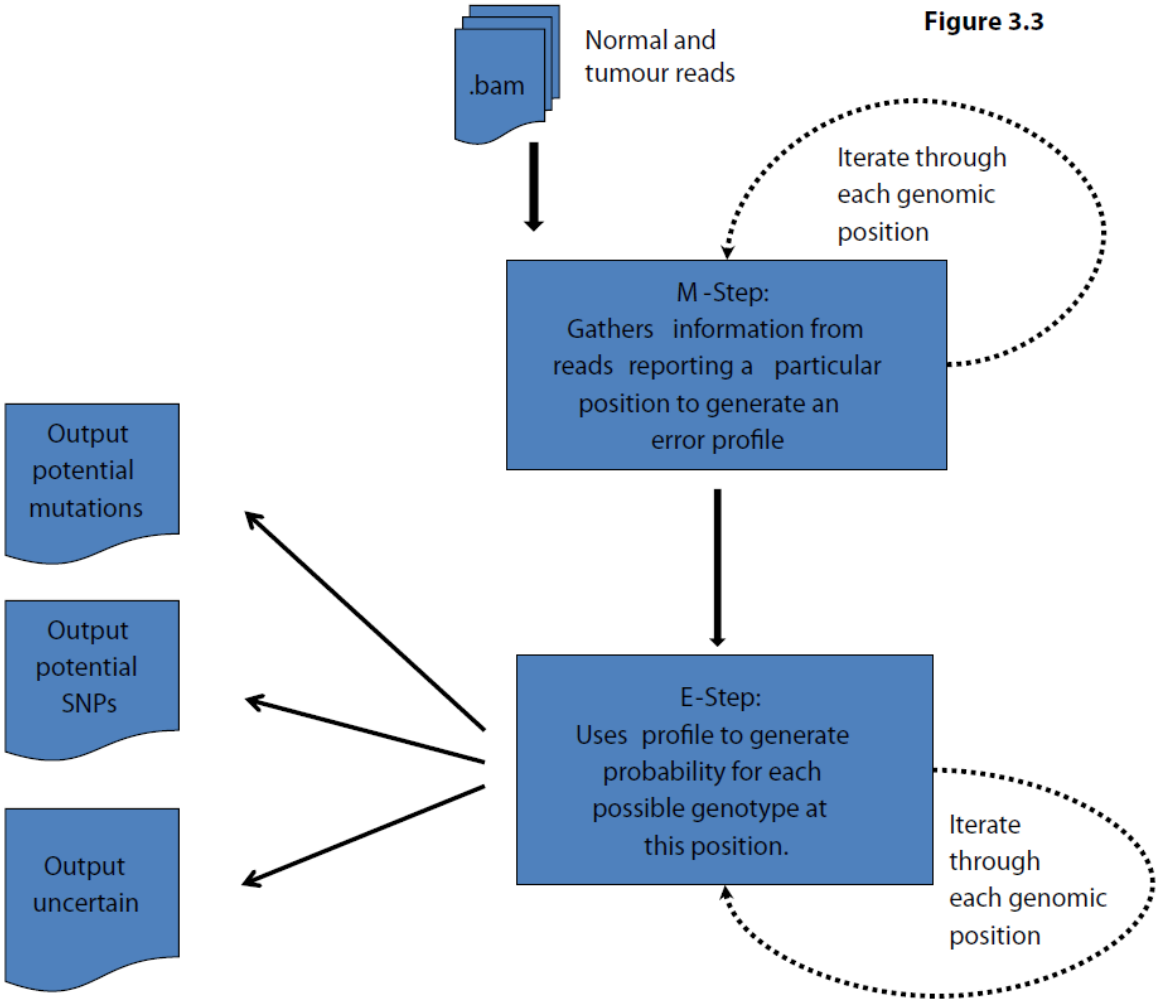


Figure 3.3: The CaVEMan workflow. CaVEMan takes a BAM file as an input file and performs two main steps, the M-step and E-step before generating three output files, a file of potential somatic substitutions, a file of possible SNPs and a file for variants that meet neither of the criteria for the other two files.

3.4 A FIRST COMPARISON BETWEEN DATA FROM CaVEMan AND THE ILLUMINA SUBSTITUTION-CALLING ALGORITHM REVEALED GOOD SENSITIVITY AND ALLOWED IDENTIFICATION OF FALSE POSITIVES FOR THE DEVELOPMENT OF EARLY FILTERS

The first and only breast cancer sample to be sequenced at Illumina© was PD3890a. 4836 highly-filtered high-confidence substitution variants were identified using the Illumina© substitution-calling algorithm. 201 variants were selected for validation by Sanger sequencing, comprising all coding variants and a random selection of non-coding variants. 168 were confirmed as somatic (83.6%) and 33 were found to be false positive calls (16.4%). The PPV of the Illumina substitution-calling process was 83.6%. This PPV was, however, possibly an overestimate of the true PPV of the Illumina © substitution-calling algorithm. Variant selection for validation was targeted to the coding exons where genomic sequence shows higher complexity. These variants were more likely to be called correctly and to therefore be true somatic variants, given the favourable mapping characteristics of the coding sequence.

On the first iteration of CaVEMan, 76235 raw variants were called in PD3890a. 100% of the 4836 variants identified by Illumina were present in this raw list of CaVEMan variants. All of the 168 confirmed somatic variants were identified demonstrating that the sensitivity or the ability to recall true variants was high. However, the total number of variants called by CaVEMan was vastly more than Illumina, likely to be overwhelmed by a variety of mis-calls and unlikely to reflect the true mutation burden in the cancer. Therefore, some early intrinsic filters were used to remove potential false positive variants whilst maintaining the number of true somatic variants.

3.5 EARLY POST-PROCESSING FILTERS

The earliest thresholds used were relatively simple. Firstly, only variants with a high likelihood (of 0.95 and above) were retained (*Mutation Probability Threshold*). Secondly, it was reasoned that a variant reported in the tumour had to be appropriately represented in the tumour. Substitution variants were identified as mismatches relative to the reference genome (Figure 3.4). However, true substitution variants were usually of a good base quality. In contrast, false positive calls arising from sequencing artifacts could also present as mismatches but were frequently at lower base qualities. With this knowledge, putative somatic variants were required to be appropriately represented in the tumour with at least a third of the reads carrying the variant allele showing a base quality score of more than or equal to 25 (*Read Depth*). Thirdly, it was considered that any putative somatic variant should not be present in the matched normal sample as well. A variant present in 5% of reads or more in the matched normal sample at base qualities of 15 or more would fail this filter and be excluded from further analysis (*Matched normal*). Using these three main criteria, the total number of variants fell to 21659 from 76235 variants.

Figure 3.4

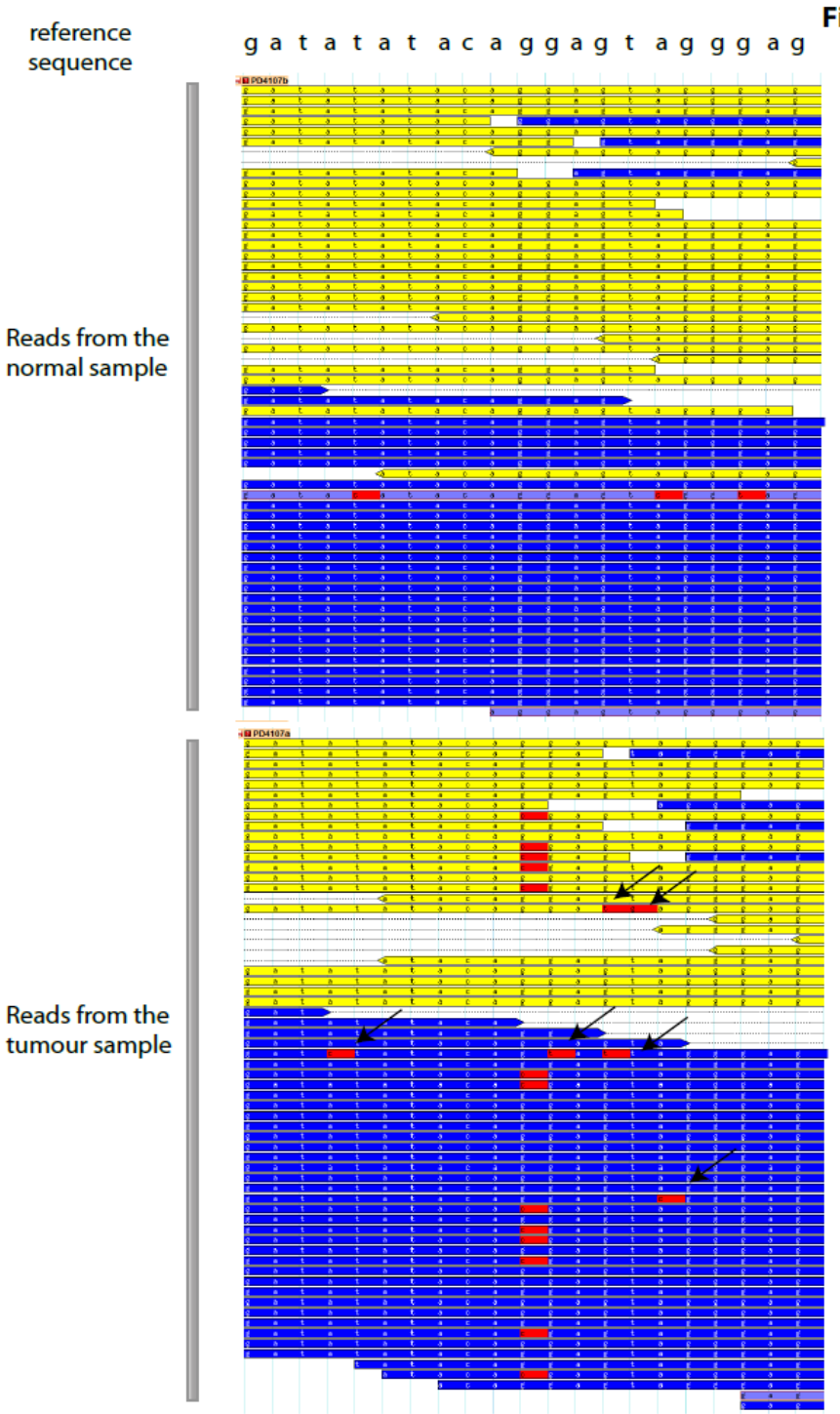


Figure 3.4: Reads in G-browse, the genome browser used to view short read sequences. Blue and yellow reads represent next-generation sequencing reads in the forward and reverse directions respectively. Each base in the reference genome is re-sequenced many-fold. The intensity of the colour reflects the mapping quality of each read. The dotted line joins each read to its read-mate. The reads on top represent reads from the matched normal and the reads below represent the tumour sample. Each read represents sequencing information from a single DNA molecule. A sequenced base which correctly matches the reference genome is not highlighted. In contrast, a base which is different to the reference genome appears (a mismatch) appears red. Here, 13 of the 57 reads in the tumour carry a G>C mismatch at the same genomic coordinate whilst no reads in the normal carry the same mismatch, corresponding to a somatic heterozygous change at this location. Note that 6 other mismatches can be seen within the same screenshot (arrows) in the tumour which represent mismatches arising as random sequencing errors or arising from mismapped reads. However, the mutation probability estimates of these randomly distributed errors are not sufficient to being called as a somatic variant.

250 variants were selected for validation by Sanger sequencing at this stage in order to identify the true PPV of CaVEMan and to identify the nature of the false positive variants that remained. Of these, 58% were confirmed as somatic (Figure 3.5a). 42% showed no evidence of the somatic variant by Sanger sequencing and were declared false positives. Of these false positive calls, 3% fell within the vicinity of germline indels, 7% were within or immediately adjacent to repeat tracts, 7% were germline single-nucleotide polymorphisms (SNPs) and 12% showed a systematic sequencing artifact characterised by unidirectionality of reads on which variants were called (Figure 3.5b). The identification of false positives and subsequent determination of causes of mis-calls were critical for development of more post-processing filters and will be described in more detail in the following sections. A further 13% showed no immediately discernible pattern initially, but as the dataset improved in its specificity, more subtle patterns emerged and became amenable to post-processing.

Figure 3.5

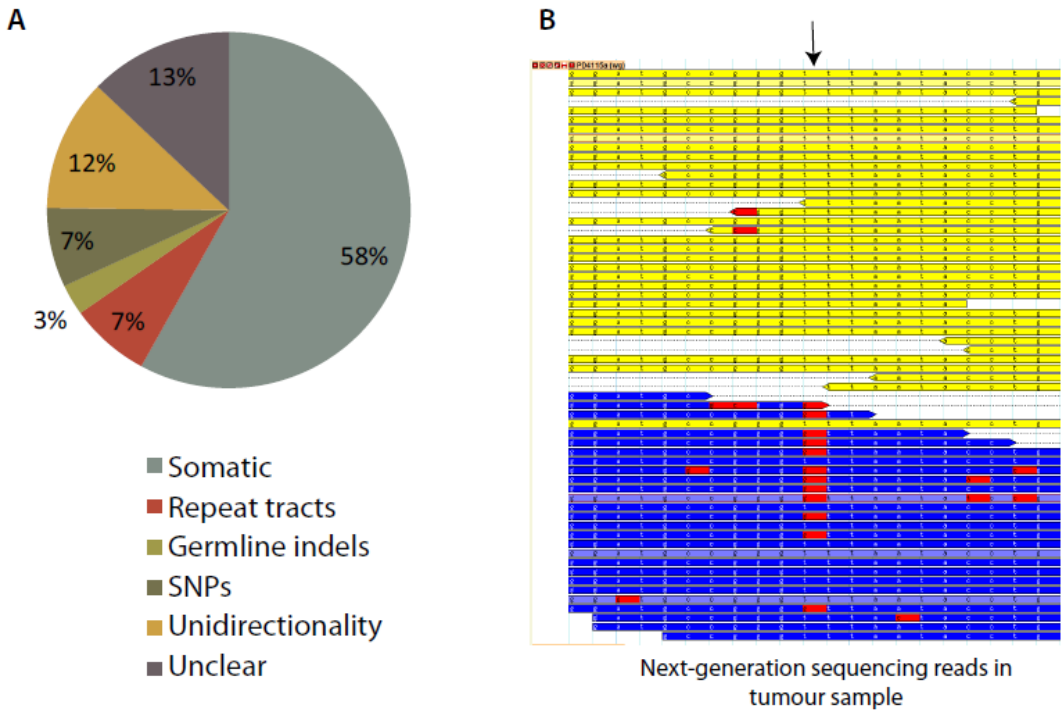


Figure 3.5: False positive calls revealed. (A) A breakdown of the false positive variants for the first iteration of validation of CaVEMan variants. (B) An example of a false positive call appearing in a unidirectional manner (only systematically on blue or forward reads) and present in tumour as well as normal reads (not shown). The variant was always the same as the preceding base in the reference genome in the direction of sequencing of the read. Here, a T>G variant following a string of G's.

3.6 THE PRINCIPLE OF DEVELOPING FURTHER POST-PROCESSING FILTERS

From the false positive variants identified in the above experiment, it was possible to classify variants that showed recurrent patterns. Some false positive variants, for example, occurred near homo-polymer or microsatellite repeat tracts, in regions of excessively low or high sequence coverage, at particular sequence motifs, at particular positions in sequencing reads (at the very ends) or near germline indels.

A post-processing filter was developed for each of these reasons and tested individually. For each filter, the reason for the filter was decided, a boolean relationship outlined and the code tested. For each test, it was necessary to ensure:

- That the expected false positives were removed
- That the known true somatic variants remained
- That there were no other unexpected changes due to errors in writing the code.

If a filter was deemed to be appropriate, it was implemented and the next filter was introduced. This procedure of "training" of filters was also performed on several other genomes in order to not over-fit filters to one sample (Figure 3.6).

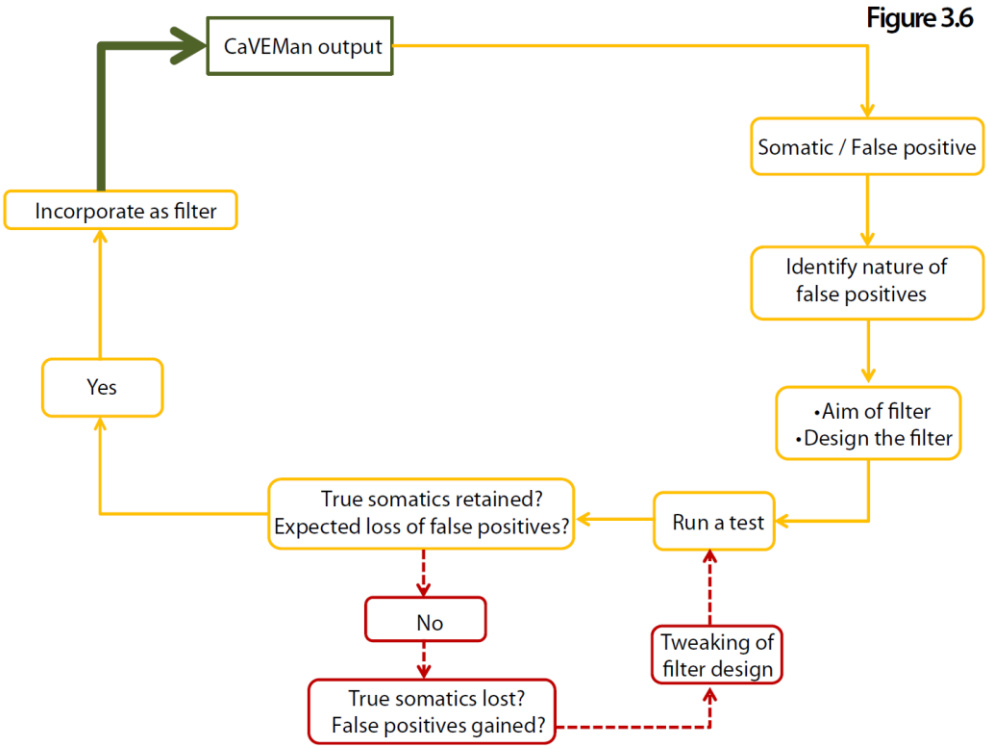


Figure 3.6: The principle of developing post-processing filters.

3.7 THE DEFINITIONS OF INDIVIDUAL POST-PROCESSING FILTERS FOR SUBSTITUTIONS

The final list of post-processing filters comprised twelve filters altogether. These could broadly be classified into three main categories.

- Filters dependent on intrinsic thresholds of sequencing/mutation-calling
- Filters for removal of systematic sequencing artifacts caused by the next-generation sequencing reaction
- Filters for genomic features that result in errors of mis-mapping

The table below provides a more detailed description of all of the filters (Table 3.1). Many of the filters take base qualities or mapping qualities into account which were described in the introduction (Figure 3.1).

Table 3.1: The reasons for and the definitions of each post-processing filter used in substitution-calling of the twenty-one breast cancer genomes

CATEGORY	NAME OF FILTER	DEFINITION	RATIONALE
Intrinsic threshold	***Mutation probability threshold***	The mutant allele probability score based on the core algorithm was equal to or above 0.95	Variants with a lower probabilistic score were simply less likely to be true somatic variants
Intrinsic threshold	***Read depth***	At least a third of bases in the tumour sample reporting the mutant allele had to exceed or equal a base quality of 25	Randomly erroneous variant bases due to the occasional fall in sequencing efficiency produced lower base qualities. True somatic variant bases had the same base qualities as other bases representing the reference allele. For a variant allele to be considered a true somatic variant, it had to be well-represented in the tumour sample, with good base qualities on several reads.
Intrinsic threshold	***Average mapping quality***	The mean mapping quality of reads reporting the mutant allele had to exceed 20	Some reads, particularly those where a germline SNP was present somewhere in the read mate, could map erroneously in highly homologous regions. If a read could map with equal or almost equivalent likelihood in more than one locus in the genome, then the mapping quality of the read reflected this lack of uniqueness. In essence, these were likely to be a cluster of mismapped reads.
Systematic sequencing artifacts	***Read Position***	The mutant allele failed this flag if it was present in less than 8 reads AND only represented on the last third of a read or only last third and first 8% of any read	Sequencing qualities and the reliability of base calls were known to fall towards the ends of reads. As a result, mismatches appeared to be more common towards the end of reads. This flag was designed to detect recurrent mismatches at the very

			ends of reads.
Systematic sequencing artifacts	***Matched normal***	The mutant allele failed this flag if it was present at base qualities exceeding 15 in more than 5% of reads in the matched normal sample	This flag was intended for removing remaining germline SNPs which had escaped initial exclusion.
Systematic sequencing artifacts	***Panel of other normals***	The mutant allele failed this flag if it was present in at least 5% of reads in at least 2 samples from the panel of randomly selected normal samples	Systematic sequencing artifacts should not discriminate between tumour and normal samples. However, they may only happen in a small fraction of reads. This flag was designed to identify those recurrent sequencing artifacts that arose intermittently in Illumina next-generation sequencing. In order to avoid the possibility of removing recurrent somatic events occurring in a subclonal population in a cancer, a randomly selected panel of normals was used to screen out recurrent sequencing artifacts.
Systematic sequencing artefacts	***Pentameric motif***	The mutant allele failed this flag if all reads carrying the variant but one were unidirectional (on forward or reverse strands only) AND the variants were only present in the last half of the read AND The reads carrying the mutant allele contained the motif GGC[A/T]G in the same sequencing direction as the variant AND the mean base quality for every base after the variant was calculated for each read and was less than 20	A systematic sequencing artefact was occurring following a specific sequencing motif characterised by GGC[A/T]G. Furthermore, the base qualities for all the bases following the putative variant usually fell well below expected. This pattern was exploited for the purposes of removing this sequencing artifact which was inexplicably worse for some tumours than others.
Systematic	***Phasing***	The mutant allele itself was required	Systematic sequencing artefacts that

sequencing artefacts		to have a mean variant base quality of more than or equal to 21 and was not unidirectionally represented.	resulted in next-generation sequencing polymerases going out of phase at some sequencing cycles. This was particularly predisposed at certain sequence motifs (-GGGG). The result was usually mutant alleles represented unidirectionally and of the same base as the immediately preceding allele in the direction of sequencing, in the reference genome. These variant alleles were usually of low base quality.
Genomic features	***Simple repeat***	The mutant variant call was failed if it fell within a simple repeat or within the immediate 5bp flanking the boundaries of a simple repeat as defined by UCSC	Mismapping of reads frequently occurred in and around simple repeats generating miscalls within or immediately flanking simple repeats.
Genomic features	***Centromeric microsatellite***	The mutant variant call was failed if it fell within the boundaries of a centromeric repeat as defined by UCSC.	Mismapping of reads frequently occurred in centromeric microsatellites generating miscalls.
Genomic features	***HiSeq coverage***	The mutant variant call was failed if it fell within a genomic window where the coverage in 2 or more genomes in a panel of normal genomes, exceeded 8 SD of the average of the coverage for those genomes *or* if it fell within parts of the genome which were consistently in the top 5% of coverage of HiSeq sequenced genomes as defined by UCSC (Pickrell et al., 2011).	Some repetitive sequences which are polymorphic in number of copies have been collapsed into a single copy in the human reference genome. When individual genomes are sequenced and mapped back to the collapsed reference genome, this results in excessively high coverage, increasing the likelihood for the accumulation of sequencing artifacts.
Genomic features	***Germline indels***	The mutant allele must not fall within the boundaries or be within ± 4bp of a germline indel as detected	Reads which ended in indels were more likely to map the very tip of the read within the indel and erroneously call it a

		by the indel-detecting algorithm.	mismatch than to map it correctly with a gap. If this occurred in multiple reads, this was effectively called as a substitution variant.

Figure 3.7

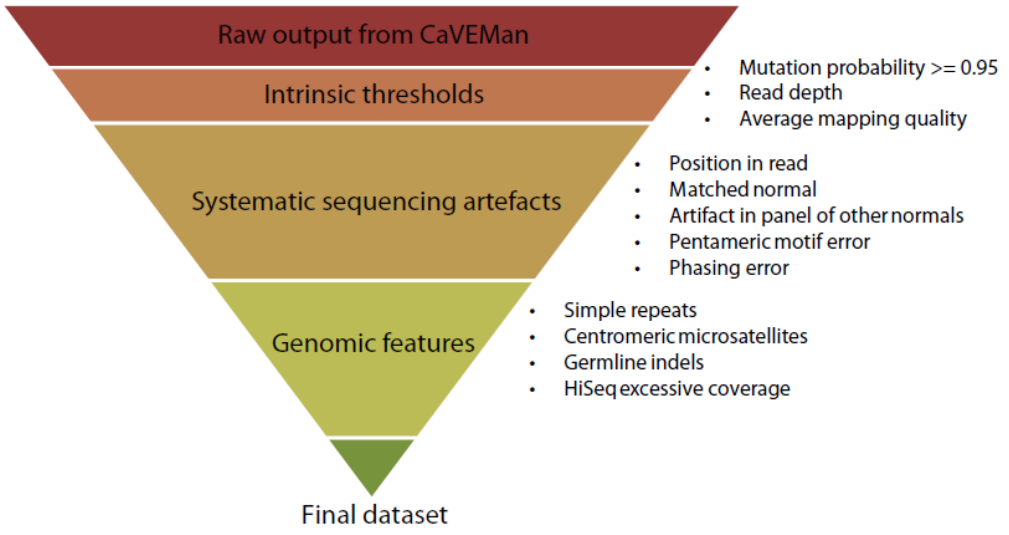


Figure 3.7: A schematic of the number of substitution variants following post-processing. The final curated dataset was always a small fraction of the total number of substitutions called.

Because each filter was applied independently for each variant, some variants could fail on multiple filters. In fact, the majority of raw substitution variants failed on multiple filters, attesting to the low likelihood of these variants being true somatic variants (Table 3.2). The final tally of substitution variants was always substantially fewer than the original raw output of the core CaVEMan substitution-calling algorithm for each genome (Figure 3.7, Table 3.2).

A revealing analysis of the effectiveness of each filter was seen in the number of variants that were removed exclusively by each filter (Figure 3.8). This demonstrated that the *Panel of other normals* was one of the most effective filters, removing the largest number of variants uniquely. This was followed by the *Matched normal* filter and the *Read Position* filter.

Sample	Raw calls	Failed more than one filter	Failed one filter	Final number of variants	Fraction of somatic variants from raw output
PD3851a	67917	58909	7226	1782	0.03
PD3890a	76235	58649	11462	6124	0.08
PD3904a	61665	49753	6304	5608	0.09
PD3905a	100027	82520	12920	4587	0.05
PD3945a	61668	44899	6461	10308	0.17
PD4005a	76186	61313	8769	6104	0.08
PD4006a	89525	69808	10523	9194	0.10
PD4085a	94504	84875	6956	2673	0.03
PD4086a	86594	77697	6698	2199	0.03
PD4088a	46420	41964	2751	1705	0.04
PD4103a	81750	70576	5814	5360	0.07
PD4107a	103870	86902	6677	10291	0.10
PD4109a	81007	65815	5304	9888	0.12
PD4115a	81136	63866	7316	9954	0.12
PD4116a	76191	59506	8659	8026	0.11
PD4192a	100127	85638	10570	3919	0.04
PD4194a	46466	40507	4475	1484	0.03
PD4198a	106246	89756	11938	4552	0.04
PD4199a	85204	68122	10150	6932	0.08
PD4248a	138435	120443	15456	2536	0.02

Table 3.2: Summary of substitution variants: From raw output to final datasets. Note that PD4120a, the deep-sequenced cancer, has not been included in this analysis of samples sequenced to 30-40-fold coverage.

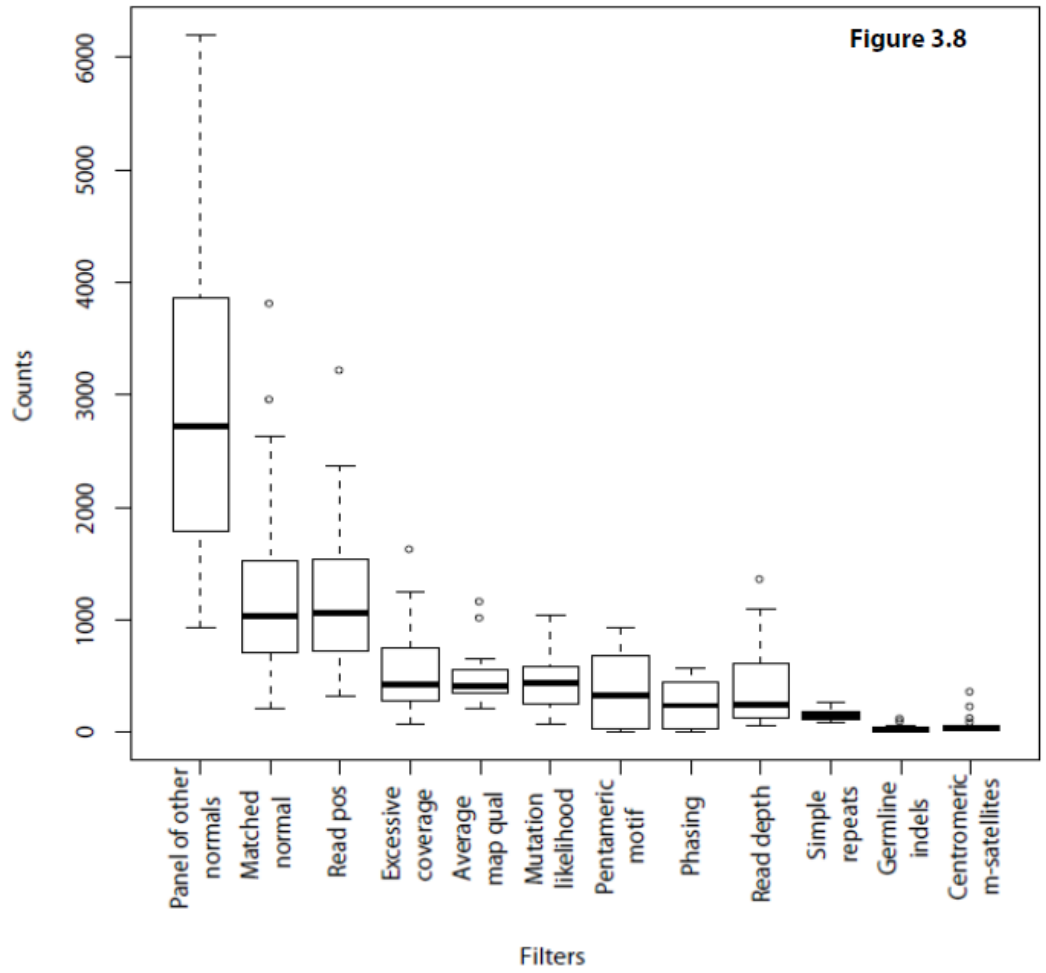


Figure 3.8: Variants removed exclusively by each filter. Total number of substitution variants removed by each filter exclusively on the vertical axis. Bottom and top of boxes in boxplots represent 25th and 75th percentiles with middle thick band at 50th percentile. Whiskers represent lowest and highest datapoints within 1.5 of the interquartile range. Small circles are outliers.

3.8 THE FRACTION OF THE GENOME WHERE MUTATIONS CAN NEVER BE CALLED

There were regions in the genome which were filtered out by virtue of being in zones of automatic exclusion. The fraction of the genome that was potentially filtered out did not simply represent the number of variants removed but was informative for the non-variant sites in the reference genome where mutations could never be called. The fraction of the genome affected by the relevant filters is documented in Table 3.3. The *germline indel* flag also contributed a proportion of genome in which no variants could be called. However, because germline indels vary between individuals, the coordinates involved in this filter was variable between cancer genomes. In general, ~1% of the genome was excluded by this filter.

Filter	Number of bases removed in the genome (bp)	Proportion of genome
Simple repeats	82,688,560	2.52
Centromeric repeats	1,660,347	0.06
HiSeq coverage	3,073,270	0.11

Table 3.3: Fraction of the genome effectively excluded by relevant filters

3.9 FINAL POSITIVE PREDICTIVE VALUE (PPV) OF CAVEMAN FOR THE DATASETS

To evaluate the improvement of the PPV of the substitution-calling process, ~400 substitution variants were re-sequenced using an orthogonal sequencing technology, in particular, Roche 454 pyrsequencing.

The PPV for each cancer genome at the point of having the first three filters and later when all twelve filters were in place is shown in Figure 3.9. The average positive predictive value for twenty cancer genomes was in the region of 92.1%.

Figure 3.9

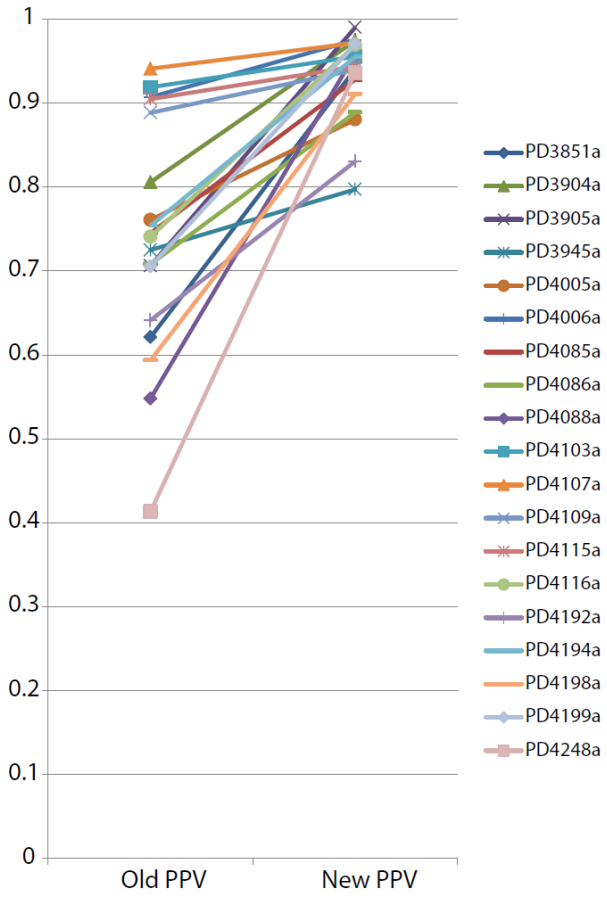


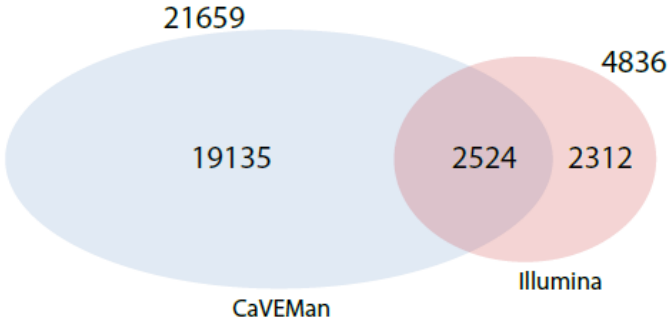
Figure 3.9: Improvement in positive predictive value for each cancer genome at the start of the experiment with three filters in place (*Mutation Probability Threshold*, *Read Depth* and *Matched Normal* filters) and later in the experiment with twelve filters in place (PPV is positive predictive value). Only 19 of 20 samples are shown here as PD3890a, was used as the sample for training many of the filters and so was excluded. The 21st sample, PD4120a, was sequenced to ultra-high depth and was therefore also excluded.

The fine-tuning of this large-scale process is expected to result in a trade-off between the gain in specificity and the loss in sensitivity. A comparison of these two parameters can be seen in PD3890a, which was sequenced at Illumina© and in which substitutions were called by an alternative caller. For the marked enhancement of the positive predictive value (56% to 90%), there was a loss of sensitivity (97% to 94.7%), at least for PD3890a.

Figure 3.10

Minimal filters

Sensitivity	Positive predictive value
0.970	0.58



All filters

Sensitivity	Positive predictive value
0.947	0.90

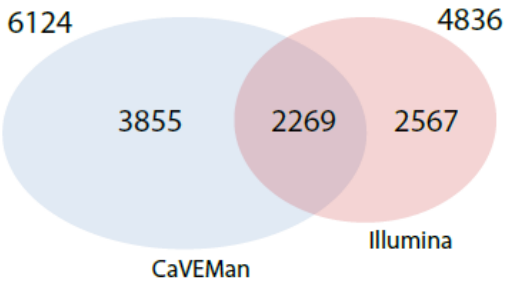


Figure 3.10: A comparison of the sensitivity and positive predictive value of PD3890a before and after development of all post-processing filters.

3.9.1 Positive predictive value does not correlate with sequencing coverage but correlates with degree of normal tissue contamination as predicted by the ASCAT (copy number algorithm)

The breast cancer genomes were assessed for whether the final PPV correlated with sequence coverage in tumour or normal. Neither of these appeared to show a correlation with the specificity of variant calling (Figure 3.11). Instead, the PPV of CaVEMan did appear to correlate with the degree of normal tissue contamination as predicted by ASCAT (the copy number algorithm used for this study). The general trend was that as aberrant cell fraction increased (and the normal contamination decreased), the PPV also increased.

Figure 3.11

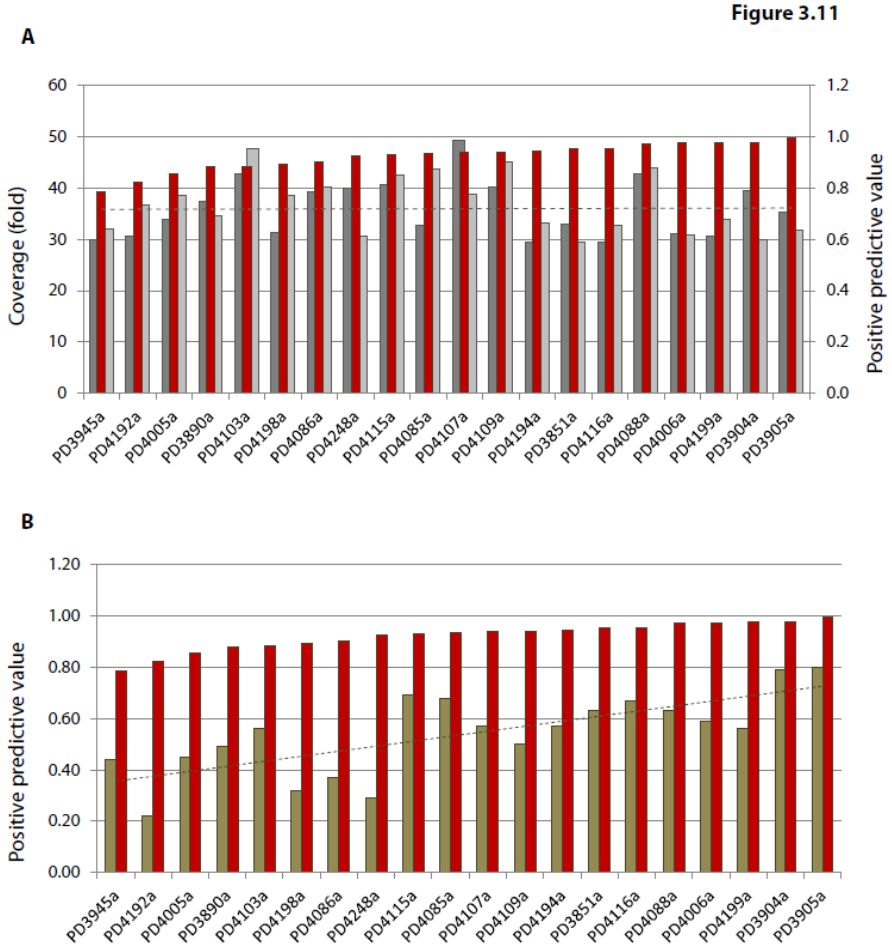


Figure 3.11: (A) No correlation was seen between positive predictive value (PPV) and tumour/normal sequencing coverage. Dotted line represents linear trend for tumour coverage (R^2=0.0002). (B) A correlation was appreciable from the comparison between PPV and aberrant cell fraction (R^2=0.5328). Dark grey = tumour coverage, light grey = normal coverage, red = PPV, tan = aberrant cell fraction. Only 20 of the 21 breast cancers were included in this analysis as PD4120a was sequenced to ultra-high coverage, and not all the filters designed were applied to this cancer.

3.10 SENSITIVITY OF DETECTION OF VARIANTS RELATIVE TO EXOMES

Four of the 21 breast cancer genomes were involved in a high-coverage (~100-fold) screen of coding sequences (exome screen, see section 2.3.1.2 for description) of 100 breast cancers (PD4103a, PD4107a, PD4109a and PD4120a). In order to gauge the sensitivity of mutation-detection in coding regions, the intersection between genome variants and exome variants was sought in three of the four cancers (PD4120a was an outlier having been whole genome sequenced to ~188-fold coverage and thus was not included in this analysis). For the three genomes, on average, 76.6% of variants detected through exome sequencing were detected in the whole genome sequences of the same cancers (range 68-82%).

The converse comparison was also performed. In each breast cancer, a proportion of variants in the coding sequence were called in the genome and missed in the exome screen. On average 22.3% of variants were missed by the exome screen ranging from 11.6-36.2%. Those variants that were missed in the exome screen were almost always due to a lack of coverage by the pull-down experiment in that region of the exome-sequenced cancer.

3.11 COMPARING CAVEMAN TO OTHER AVAILABLE MUTATION CALLERS

Although several mutation callers are available, none provides the level of (publicly available) post-processing that has been developed for these 21 breast cancer genomes. Comparing the dataset here with the raw output from other mutation callers does not therefore constitute a fair comparison. A version of Somatic Sniper was used to call variants in PD4107a but generated an enormous number of mutations (~450,000) as no post-processing was available at the time (http://gmt.genome.wustl.edu/somatic-sniper/current/). An alternative somatic single nucleotide variant caller which did have some post-processing options, MuTect (https://confluence.broadinstitute.org/display/CGATools/MuTect) generated an excess of 2.5 to 8 fold more variants for 3 breast cancers tested (Table 3.4). This was despite adopting the most stringent of post-processing filters available.

Table 3.4: Comparison between MuTect and CaVEMan, using three genomes as examples.

Sample	MuTect variants	CaVEMan variants	Overlap-ping variants	Variants missed by MuTect	Proportion of variants missed by MuTect	Average variant allele fraction of overlapping variants	Average variant allele fraction of variants missed by MuTect	Aberrant cell fraction	Tumour ploidy	Variants missed by CaVEMan which appear real	Variants missed by MuTect which appear real
PD4192a	20307	3919	3078	841	0.21	0.17	0.16	0.22	4.68	0	0.34
PD4198a	14618	4552	4142	410	0.09	0.21	0.18	0.32	3.05	0	0.48
PD4199a	17499	6932	6542	390	0.06	0.28	0.21	0.56	1.69	0	0.54

In order to evaluate the performance of MuTect and CaVEMan relative to each the other, a cohort of variants were sampled and visually assessed. In an ideal situation, these cohorts would have been validated. Of the variants missed by CaVEMan but were present in MuTect, none were real. Interestingly, between 17-28% of these were previously seen in CaVEMan but filtered out on the *Panel of Normals* filter alone. It is therefore likely that the vast majority of the excess of variant calls made by MuTect are false positive calls.

Assessing the variants present in CaVEMan and missed by MuTect, between 34-54% of variants looked real on visual inspection with many of the true variants being present at a lower variant allele fraction both in regions which were diploid as well as regions that were polyploid. This suggests that the sensitivity of variant detection by CaVEMan was higher for subclonal variants as well as variants which occurred on a single allele in of a multi-allele region in the clonal population.

3.12 INSERTIONS/DELETIONS AND REARRANGEMENTS

A similar methodical process of elimination of potential false positives was performed on the insertions/deletions. However, the indel-calling algorithm, Pindel, worked in a relatively simple way in its method of detecting variants. Pindel does not work on a probabilistic model and does not perform a comparison between tumour and normal. Therefore, a set of crude filters were designed in order to reduce the total number of variants.

Validation experiments on this filtered dataset revealed that the positive predictive value was still relatively low (40%-60%). As a result, only validated indels have been presented for downstream investigation, leaving a smaller but purer cohort of variants for analysis. The same principle applied to the detection of structural variants.

3.13 SUMMARY OF THE ANALYSIS PROCESS USED TO GENERATE THE FINAL DATASET

Following multiple iterations of validation and post-processing, the final analysis process was one which showed a high degree of interdependency (Figure 3.12). The final datasets used and described in the subsequent chapters therefore comprised:

- all the filtered substitutions with a subset of variants which were validated (Appendix 1)
- validated insertions/deletions (Appendix 2)
- validated rearrangements (Appendix 3)

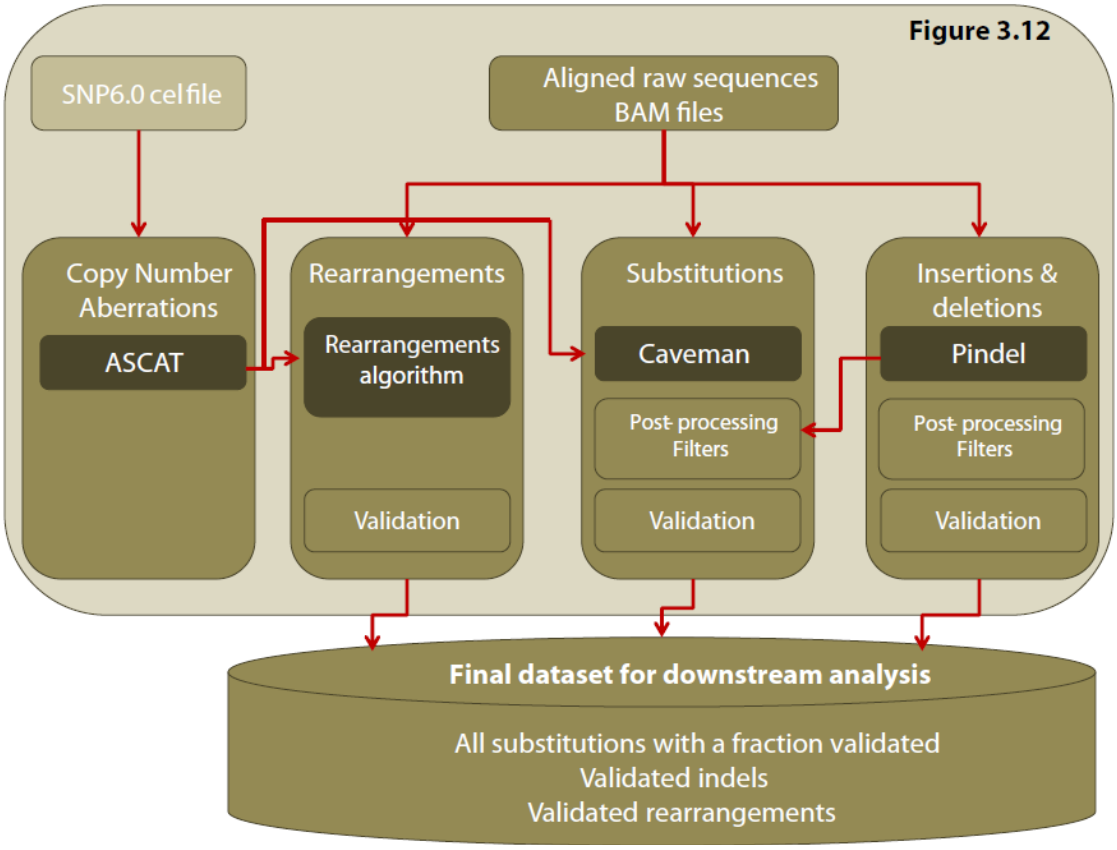


Figure 3.12: A schematic of the final analysis pipeline used to obtain a list of variants for downstream analysis in this thesis

3.14 DISCUSSION

This chapter was dedicated towards the development of post-processing filters required to obtain a final curated dataset that was essential for the detailed analysis performed later in this thesis, particularly for substitutions. Here, a systematic approach of identification of false positives, the reasons why they occur and the development of a collection of post-processing filters, were described. The positive predictive value (PPV) was used as a measure of the effectiveness of each of the post-processing filters.

In all, twelve post-processing filters were designed, reducing the dataset substantially and increasing the positive predictive value remarkably, with a minor cost to sensitivity. These filters could be classified into three main categories: those which involved intrinsic thresholds of the algorithm, those which were designed to remove systematic sequencing artifacts and those which were necessary to remove erroneous calls due to genomic features which caused mapping errors of short-read data. The final average positive predictive value for this cohort of breast cancers was ~92%.

3.14.1 A fair comparison between different mutation callers would involve comparing datasets *after* post-processing

Although other substitution-callers exist, none are known to consider complicating factors associated with the complexity of cancer: tumour heterogeneity, degree of contaminating normal cells and abnormalities of ploidy. Inclusion of these additional parameters in probability estimates in CaVEMan allowed base-by-base adjustments and in theory, increased the likelihood of calling true somatic substitution variants, particularly those which were present in a minor subclone in a cancer or those occurring on only one allele in a polyploid region of a cancer. This increased sensitivity is reflected in the variants missed by other callers but present by CaVEMan substitution-calling, which were all at a low variant allele fraction. Furthermore, depending on the nature of the biological specimen studied (e.g. cell lines), these parameters could be tuned in order to maximize the likelihood of detection of somatic variants.

Presently, despite application of the highest stringency filters (of which there are very few if any for some callers), the total number of mutations called by alternative callers are markedly more than by CaVEMan. Given that the curated dataset obtained here had twelve post-processing filters applied, a fairer comparison to other substitution-callers would require application of equivalent post-processing filters. Furthermore, a more comprehensive comparison between the performance of

CaVEMan relative to other substitution-callers would possibly require a degree of validation of those variants missed by CaVEMan. This has not been performed as part of this thesis due to time constraints.

3.14.2 Balance between sensitivity and specificity: two sorts of datasets

The set of mutations obtained from any large-scale genomic experiment will always comprise a set of true somatic variants and a larger set of false positive calls. The degree to which a dataset is filtered will depend entirely on the question sought. In exome-sequencing experiments of cancers, targeted enrichment of the coding sequence and higher sequencing coverage in these protein-coding exons is primarily aimed at identification of driver events and demands as high a measure of sensitivity as possible. Although this may result in a high number of false positive calls, the total burden of mutations is still relatively low and amenable to validation in order to isolate true somatic events. The same approach would overwhelm a genome-sequencing experiment. Because the focus in this genome-sequencing project was on seeking genome-wide signatures and related less to detection of cancer genes, it was imperative for specificity to be set as high as possible in order to reduce the likelihood of detection of false positive signatures.

In the near future and for large-scale genome sequencing projects in cancer, it may be necessary for some combination of both approaches to be used. Perhaps, the core algorithm could be run with a set of “high sensitivity” filters concentrating on the coding sequence in order to detect all important coding mutations, as well as “high specificity” filters for the whole genome, in order to obtain a complete catalogue of variants from a single sequencing experiment.

3.14.3 Scope for improvement of individual filters

There is likely to be scope for improvement of some filters. First, there was considerable overlap in the variants removed by some filters, particularly between the “Read Position” and “Germline indel” filters. However, each also removed a definite and mutually exclusive subset. Hence, it was difficult to justify removing either as a filter. Second, more time could have been spent on improving the sensitivity lost with each filter. This would have required several more iterations of each filter and for this thesis, had to be balanced with the timeline of getting an adequately curated dataset. Nevertheless, enhancements to the current set of filters are expected in the near future.

Furthermore, post-processing filters developed here had been trained to accommodate cancer genomes sequenced to 30X-50X coverage of 100bp reads, with equivalent depth in the matched

normal. The efficacy of these filters is likely to be affected by genomes with significantly different levels of coverage between tumour and normal. The use of proportions was favoured over the use of absolute values particularly when defining read depth in the post-processing filters, but this was not always possible (e.g. *Read Position* filter). Therefore, filters which are sensitive to variation in coverage may become less effective if the coverage in the tumour is not at 30-50X. Distinctions based on proportional distance along each read were also made in some filters and this could be adversely affected by shorter read lengths of 50 or 75bp reads. Therefore, subtle differences in experimental approach may affect the application of these filters and could possibly be factored into the design of each filter, in the future.

3.14.4 The moving target: future optimization will be necessary

Any improvements to the core algorithm will necessitate further optimization of the substitution-calling process. In addition, changes in sequencing technology and chemistry resulting in vastly increased yields per lane of sequencing is likely to give rise to other novel sequencing artifacts and will require thoughtful application of new filters, or adaptations to old ones, in order to manage new problems.

3.14.5 The performance of callers on indels and rearrangements

This chapter has focused on developing post-processing filters for calling substitutions. The performance of the core algorithms and current filters for indels and rearrangements was much less desirable, with poorer specificity for both of these mutation classes. As a result, confidence can only be placed on validated variants and only these validated indels and rearrangements were used for downstream analysis.

Other approaches could be considered for the near future. Local reassembly is a feature used by the Broad Institute (GATK) to improve the mapping of reads overlying indels. This is an approach that has not been explored in this thesis. Because suitably stringent post-processing filters are not available for GATK, one possibility would be to perform primary indel-calling using Pindel and then perform local reassembly across these indels to improve mapping characteristics of the informative reads before post-processing. Another approach that is described is to use multiple callers on the same dataset and to simply use the variants which are overlapping.

CHAPTER FOUR: EXPLORING MUTATIONAL SIGNATURES FROM BASE SUBSTITUTIONS IN TWENTY-ONE BREAST CANCER GENOMES

4.1 INTRODUCTION

In the introduction chapter of this thesis, the concept of a *mutational signature* as a characteristic imprint left on the cancer genome by a *mutational process* which comprises some combination of DNA damaging and DNA reparative mechanism was introduced. However, each cancer genome could have multiple mutational processes acting through the lifespan of the cancer. When a cancer is diagnosed, removed at surgery and is sequenced, the final mutational portrait that we come to see of each cancer, therefore, is a composite of multiple mutational signatures that have been added layer upon layer through the development of the cancer (Figure 1.1). Each complex and multidimensional cancer genome bears the inscription of its biological history including that of mutagenic damage from environmental or endogenous sources and bears the hallmarks of repair processes that have been operative as well.

In addition, excavation of the biological history borne by mutations across not one, but multiple cancers of the same tissue-type may highlight processes that are shared in common. Some exogenous and many endogenous mutagenic processes are likely to be mutual between different individuals as each person will be subjected to by-products of cellular metabolism alike or be exposed to background levels of radiation, for example. The sequencing of twenty-one cancer genome datasets therefore offers an opportunity to explore and tease apart the underlying processes that are present collectively across these breast cancers.

Furthermore, the vast numbers of somatic mutations provided by a pooled analysis gives us an opportunity to unravel processes that are superficially similar but in fact, distinct. For example, historically, many cancers have an over-representation of C>T/G>A mutations. However, C>T/G>A mutations occurring at CpG dinucleotides, which are not in CpG islands and are more likely to be methylated, are likely to be attributed to the well-described phenomenon of deamination of methylated cytosines. In contrast, C>T/G>A and CC>TT/GG>AA mutations occurring at dipyrimidines in malignant melanomas or other sun-induced cancers are believed to be due to ultraviolet-radiation damage. Therefore, additional facets of mutation such as sequence context can be explored in order to derive biological insights.

In this chapter, common mutational signatures from the complex multidimensional dataset of 21 breast cancer genomes will be sought. The development and refinement of the mathematical algorithm used in the extraction of mutational signatures is the subject of the doctoral thesis of another graduate student, Ludmil B. Alexandrov. Here, the focus is on developing a conceptual understanding and biological framework of the data produced by the algorithm. Mutational signatures identified in the cancers will be compared and matched to known signatures in order to gain insights into the biology of mutational and repair processes that have been operative on the cancers.

4.2 THE SERIES OF BREAST CANCERS USED IN THIS STUDY

The initial intention was to sequence 20 breast cancers across the spectrum of histopathological breast cancer subtypes and to include breast cancers derived from individuals with germline mutations in the cancer predisposition genes, *BRCA1* and *BRCA2*. Subsequently, a breast cancer known to harbour a very large number of mutations (more than 600 substitutions in the coding sequence alone (Stephens et al., 2012)) was sequenced to very high coverage and included in this analysis. The final series of breast cancers used in this study were:

- five cases that were estrogen receptor (ER) positive and HER2 negative;
- two cases that were ER positive and HER2 positive;
- two cases that were ER negative and HER2 positive;
- three cases that were ER negative, progesterone receptor (PR) negative and HER2 negative (triple negative);
- five cases with germline mutations in the high-risk breast cancer predisposition gene *BRCA1* and
- four cases with germline mutations in *BRCA2*.

Verification of germline mutation status was sought in those breast cancers reported as being derived from germline *BRCA1* and *BRCA2* mutation carriers. In addition, CaVEMan, Pindel and rearrangement outputs were screened for potential previously unidentified germline *BRCA1* and *BRCA2* mutation, in all the breast cancers (Table 4.1). Via this method, PD4107a, a breast cancer initially included in the study as a sporadic triple negative breast cancer was found to harbour a cryptic germline frame-shifting insertion in *BRCA1*, essentially diagnosing *BRCA1* carrier status in the patient.

Sample	Age at first diagnosis	Previous histopath-ological diagnosis	Histo patho-logical Grade	ER Status	PR Status	HER2 Status	Germline mutation status			
							Genomic	Gene	cDNA	Protein change
PD3851	61	Ductal	III	+ve	+ve	-ve				
PD3890	41	Ductal	III	-ve	-ve	-ve	chr17:g.41245047delC	BRCA1	c.2501delG	p.G834fs*12
PD3904	39	Ductal	III	+ve	+ve	-ve	chr13:g.32914974_32914977delACAA	BRCA2	c.6482_6485delACAA	p.K2162fs*5
PD3905	34	Ductal	III	-ve	-ve	-ve	chr17:g.41232400_41236234del3835	BRCA1	c.4186-1642_4357+2021del3835	p.?
PD3945	59	Ductal	III	+ve	-ve	-ve	chr13:g.32914557C>G	BRCA2	c.6065C>G	p.S2022*
PD4005	39	Ductal	III	-ve	-ve	-ve	chr17:g.41243838delA	BRCA1	c.3710delT	p.I1237fs*27
PD4006	39	Ductal	III	-ve	-ve	-ve	chr17:g.41245861G>A	BRCA1	c.1687C>T	p.Q563*
PD4085	64	Ductal	III	+ve	+ve	-ve				
PD4086	58	Ductal	III	-ve	-ve	-ve				
PD4088	32	Ductal	III	+ve	-ve	-ve				
PD4103	46	Ductal	III	+ve	+ve	-ve				
PD4107	33	Ductal	III	-ve	-ve	-ve	chr17:g.41246538_41246539insT	BRCA1	c.1009_1010insA	p.V340fs*6
PD4109	67	Ductal	III	-ve	-ve	-ve				
PD4115	54	Ductal	III	+ve	+ve	-ve	chr13:g.32968863C>A	BRCA2	c.9294C>A	p.Y3098*
PD4116	32	Ductal	III	+ve	+ve	-ve	chr13:g.32911947T>G	BRCA2	c.3455T>G	p.L1152*
PD4120	60	Ductal	II	+ve	+ve	-ve				
PD4192	70	Ductal	III	-ve	-ve	+ve				
PD4194	43	Lobular	III	+ve	+ve	+ve				
PD4198	59	Ductal	III	+ve	-ve	+ve				
PD4199	59	Ductal	II	-ve	-ve	+ve				
PD4248	48	Ductal	II	-ve	-ve	-ve				

Table 4.1: Demographic information regarding breast cancers, histopathological diagnosis, and germline mutation status where relevant

4.2.1 Coverage

An average of 135 gigabases of sequence data was generated for each tumour or normal library to achieve average sequence coverage of 30X for each library. One breast cancer, PD4120a, was sequenced to achieve ~188X coverage (Table 4.2).

Sample	Coverage Tumour (X)	Coverage Matched Normal (X)
PD3851a	33.02	29.40
PD3890a	37.46	34.61
PD3904a	39.42	30.03
PD3905a	35.33	31.68
PD3945a	30.03	32.08
PD4005a	34.00	38.57
PD4006a	31.10	30.85
PD4085a	32.73	43.65
PD4086a	39.25	40.13
PD4088a	42.84	43.86
PD4103a	42.84	47.63
PD4107a	49.21	38.79
PD4109a	40.13	44.98
PD4115a	40.70	42.46
PD4116a	29.45	32.76
PD4120a	188.07	32.50
PD4192a	30.68	36.78
PD4194a	29.46	33.13
PD4198a	31.24	38.57
PD4199a	30.58	33.80
PD4248a	39.98	30.52

Table 4.2: Final sequencing metrics of whole-genome sequenced breast cancers.

4.3 PUTATIVE SOMATIC DRIVER EVENTS IN TWENTY-ONE BREAST CANCERS

In the last forty years, cancer research has focused on the discovery of cancer genes which carry the "driver" mutations that confer selective clonal growth advantage and are causally implicated in oncogenesis. 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).

Likely driver events have been sought and were documented briefly in this section, although the main thrust of this thesis is the genome-wide exploration of mutational signatures in twenty-one breast cancers. Putative driver substitutions and insertions/deletions in cancer genes were found in *TP53*, *GATA3*, *PIK3CA*, *MAP2K4*, *SMAD4*, *MLL2*, *MLL3,* and *NCOR1* (Table 7.5)(cross-referenced with known driver mutations in http://www.sanger.ac.uk/genetics/CGP/cosmic/). Amplification was observed over several cancer genes previously implicated in breast cancer development including *ERBB2, CCND1, MYC, MDM2, ZNF217* and *ZNF703* and a homozygous deletion involving *MAP2K4* was identified (Table 7.3 and Table 7.4). All tumours derived from *BRCA1* or *BRCA2* germline mutation carriers showed loss of the wild type haplotypes at 17q21 or 13q12 respectively, as expected of recessive cancer genes (Supplementary Table 7.1). As expected, no new cancer genes or fusion genes have been unearthed, given the well-studied disease and the relatively small sample size.

Table 4.3: Putative somatic substitution and insertion/deletion driver events in twenty-one breast cancers

Insertions and deletions

CGP Variant ID	Sample	Chr	Start	End	Deleted sequence	Indel type	Default gene	Transcript ID	CDS mut syntax	AA mut syntax
53377626	PD4085a	10	8111433	8111434	CA	deletion	GATA3	ENST00000379328	c.925-3_925-2delca	p.?
52848859	PD4085a	17	11984671	11984672	AG	deletion	MAP2K4	ENST00000353533	c.219-2_219-1delag	p.?
27976289	PD4107a	17	7578263	7578263	G	deletion	TP53	ENST00000269305	c.586delC	p.R196fs*51

Substitutions

CGP Variant ID	Sample	Chr	Position	WT base	MT base	Default mut type	Default gene	Transcript ID	CDS mut syntax	AA mut syntax
22791325	PD4120a	3	178916946	C	G	misssense	PIK3CA	ENST00000263967	c.333G>C	p.K111N
27104511	PD3905a	3	178936082	C	T	misssense	PIK3CA	ENST00000263967	c.1624G>A	p.E542K
22791336	PD4120a	3	178952085	T	C	misssense	PIK3CA	ENST00000263967	c.3140A>G	p.H1047R
28357778	PD4085a	3	178952085	T	C	misssense	PIK3CA	ENST00000263967	c.3140A>G	p.H1047R
27351862	PD4192a	3	178952085	T	C	misssense	PIK3CA	ENST00000263967	c.3140A>G	p.H1047R
28279236	PD4103a	7	151876918	C	T	essential splice	MLL3	ENST00000262189	c.7442+1G>A	p.?
27469358	PD4109a	12	49415846	C	T	nonsense	MLL2	ENST00000301067	c.16501C>T	p.R5501*
27705761	PD4199a	17	7576852	C	T	essential splice	TP53	ENST00000269305	c.993+1G>A	p.?
22400333	PD4120a	17	7577127	C	G	misssense	TP53	ENST00000269305	c.811G>C	p.E271Q
27639366	PD3890a	17	7577539	C	T	misssense	TP53	ENST00000269305	c.742C>T	p.R248W
27506790	PD4109a	17	7578190	T	C	misssense	TP53	ENST00000269305	c.659A>G	p.Y220C
28169984	PD4005a	17	7578212	C	T	nonsense	TP53	ENST00000269305	c.637C>T	p.R213*
22400335	PD4120a	17	7578380	C	G	misssense	TP53	ENST00000269305	c.550G>C	p.D184H
22402355	PD4120a	17	16046958	C	A	nonsense	NCOR1	ENST00000268712	c.1135G>T	p.E379*
22353347	PD4120a	18	48575671	C	G	nonsense	SMAD4	ENST00000342988	c.431C>G	p.S144*
22353349	PD4120a	18	48591837	C	T	nonsense	SMAD4	ENST00000342988	c.1000C>T	p.Q334*

4.4 SUBSTANTIAL VARIATION IN THE NUMBERS AND CLASSES OF SOMATIC SUBSTITUTION MUTATIONS IS FOUND IN BREAST CANCER

In aggregate, there were 183,916 substitution variants from 21 breast cancers with an average of 8758 variants per genome. The 21 breast cancers exhibited substantial variation in the total number of somatic substitution mutations ranging from 1,484 substitutions in PD4194a, the solitary lobular ER positive, PR positive and HER2 positive breast cancer in the group, to 70,690 substitutions in PD4120a, a ductal, ER positive, PR positive, HER2 negative breast cancer (Table 4.4). Although there did not appear to be a direct relationship between histopathological status and the total number of substitution variants, the breast cancers with germline defects in *BRCA1* and *BRCA2*, genes involved in the homologous recombination repair of double-strand breaks, did have more mutations on average per genome (when PD4120a, the outlier hypermutated breast cancer was excluded) ($p<2.2e-16$).

Sample	Age at first diagnosis	ER Status	PR Status	HER2 Status	Germline mutation	Total number of substitutions
					Gene	
PD4194a	43	+ve	+ve	+ve		1484
PD4088a	32	+ve	-ve	-ve		1705
PD3851a	61	+ve	+ve	-ve		1782
PD4086a	58	-ve	-ve	-ve		2199
PD4248a	48	-ve	-ve	-ve		2536
PD4085a	64	+ve	+ve	-ve		2673
PD4192a	70	-ve	-ve	+ve		3919
PD4198a	59	+ve	-ve	+ve		4552
PD3905a	34	-ve	-ve	-ve	BRCA1	4587
PD4103a	46	+ve	+ve	-ve		5360
PD3904a	39	+ve	+ve	-ve	BRCA2	5608
PD4005a	39	-ve	-ve	-ve	BRCA1	6104
PD3890a	41	-ve	-ve	-ve	BRCA1	6124
PD4199a	59	-ve	-ve	+ve		6932
PD4116a	32	+ve	+ve	-ve	BRCA2	8026
PD4006a	39	-ve	-ve	-ve	BRCA1	9194
PD4109a	67	-ve	-ve	-ve		9888
PD4115a	54	+ve	+ve	-ve	BRCA2	9954
PD4107a	33	-ve	-ve	-ve	BRCA1	10291
PD3945a	59	+ve	-ve	-ve	BRCA2	10308
PD4120a	60	+ve	+ve	-ve		70690

Table 4.4: Breast cancer series and total number of substitutions

In protein coding regions, there were 1,372 missense, 117 nonsense, 2 stop-lost, 37 essential splice-site and 521 silent mutations. The majority of mutations fell in intergenic regions as would be expected (Table 4.5).

		Count	%
Intergenic		111358	0.61
Genomic footprint			0.39
	Intronic	68734	
	Missense	1372	
	Nonsense	117	
	Essential splice-site	37	
	Stop-lost	2	
	Start-gained	12	
	Silent	521	
	UTR	1763	
	Total	183916	1.00

Table 4.5: Breakdown of the different types of (predicted) substitution mutations identified in this series of 21 breast cancers.

Substantial variation was observed in the relative contributions of each of the six classes of base substitution (C>A/G>T, C>G/G>C, C>T/G>A, T>A/A>T, T>C/A>G and T>G/A>C) (Figure 4.1a). In general, although there was a predominance of C>T/G>A in almost all the breast cancers, there were differences in the shape of the distribution of the mutational spectra (Figure 4.1a). PD4120a (which has an alternative x-axis in Figure 4.1a) has an order of magnitude more mutations than the rest of the cancers. Despite having significantly more mutations, the shape of the distribution of the mutation spectrum of PD4120a closely resembles that of PD4199a, with C>T/G>A mutations exceeding C>G/G>C mutations, but both dominating the spectra over and above any other mutation type. In contrast, PD3851a, a ductal carcinoma with ER positive, PR positive and HER2 negative status sharing the same histopathological status as PD4120a, has far fewer mutations than PD4120a at only 1782 substitutions and has C>T/G>A mutations as the modal mutation-type but is followed by C>A/G>T mutations instead. Contrast that again with PD4116a, a germline BRCA2 cancer with the same histopathological status as PD3851a and PD4120a of ER positivity, PR positivity and HER2 negativity, contains 8026 mutations and essentially equivalent numbers of C>A/G>T, C>G/G>C and C>T/G>A mutations, and considerable contribution from T>A/A>T, T>C/A>G and T>G/A>C mutations

as well. In summary, clear variation in the shape of the mutational spectra or distribution of mutations were seen which were unrelated to the histopathological statuses of these twenty-one breast cancers.

4.5 EXPLORING THE SEQUENCE CONTEXT OF SOMATIC SUBSTITUTIONS IN BREAST CANCER

Sequence context is known to have an impact on mutation rates in the genome. For example, the process of deamination at methylated cytosines at CpG dinucleotides is believed to be the cause for general depletion of CpG dinucleotides in the human genome over evolutionary time. In order to explore mutational signatures and gain greater depth of insight into mutational processes that may be operative, the sequence context of the bases immediately 5' and 3' to each mutated base was taken into consideration. Since there are six classes of base substitution and 16 possible sequence contexts for each mutated base (A, C, G or T at the 5' base and A, C, G or T at the 3' base), there are 96 possible mutated trinucleotides for each cancer. Henceforth, the following convention will be taken to describe mutations: For example, a C to T mutation occurring at a 5' thymine and a 3' guanine will be described as TpCpG > TpTpG with the mutated base underlined.

The human genome shows asymmetric GC/AT content throughout. Therefore, a correction or normalisation for the true prevalence of each trinucleotide was included. To ensure that bias was not introduced by either properties of the library (pro- or anti-GC bias) or by the mutation-caller, the prevalence of each trinucleotide was counted for bases that were examined by the substitution-caller for each individual cancer genome. The observed fraction of mutations at each trinucleotide has therefore been normalised according to the prevalence of each trinucleotide in individual cancer genomes.

To facilitate visualisation of the mutational patterns present, for each cancer, the fraction of mutations at each of the 96 mutated trinucleotides was represented in a heatmap. A log (10)-transformation of the normalised values was plotted in a heatmap (Figure 4.1c). The heatmap therefore highlights the presence of mutational processes that favour particular classes of mutation and/or particular sequence contexts in which they occur.

Figure 4.1

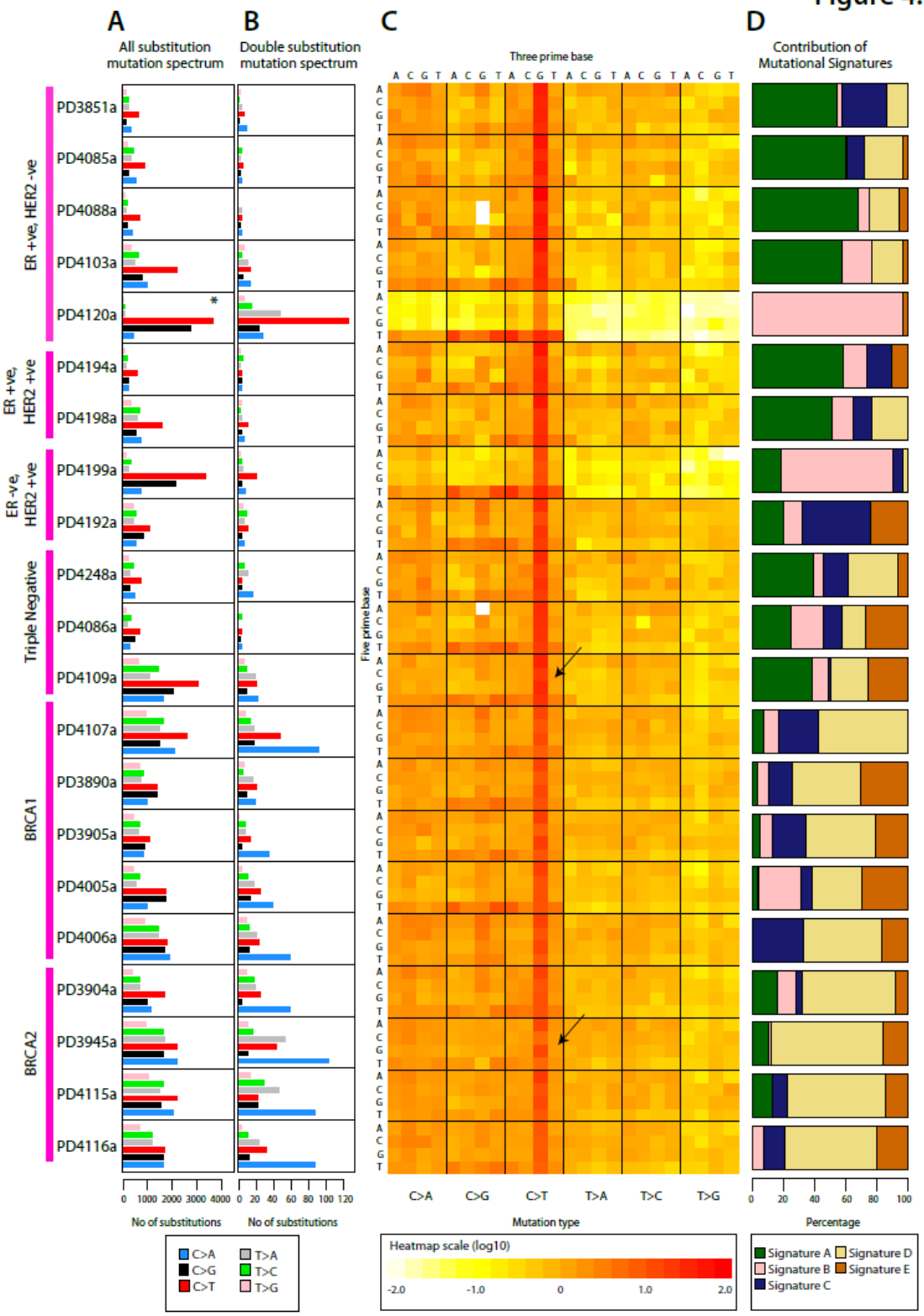


Figure 4.1: Somatic mutation profiles of 21 breast cancers. Breast cancers grouped according to subtype on the far left. (A) Base substitution mutation spectra. *Ultra-deep sequenced PD4120a has an alternative scale on the x axis (0 to 45,000). (B) Mutation spectra of double substitutions from all 21 samples. (C) Genomic heat map constructed from counts of each mutation-type at each mutation context corrected for the frequency of each trinucleotide in the reference genome. Log-transformed values of these ratios have been plotted in the heatmap. The 5' base to each mutated base is shown on the vertical axis and 3' base on the horizontal axis. The log (10) scale of the genomic heatmap is presented at the bottom. (D) Proportion of the total substitutions contributed by each of the five mutational signatures, as identified by NMF analysis, for all 21 cancer genomes. This is discussed later in section 4.7.4.

4.6 VISUAL IDENTIFICATION OF MUTATION PATTERNS

Visual inspection of the 21 heatmaps provided evidence for the presence of multiple independent mutational processes and indicated that, in many cancers, more than one process has been operative. Furthermore, the heatmaps highlighted how several mutational processes were ubiquitously present in many of the different cancer genomes albeit operating to differing degrees in each. A more detailed account of apparent mutation signatures is provided in the following section.

4.6.1 C>T at XpCpG is a dominant mutation signature in all breast cancers

An ostensible feature of the heatmap was the over-representation compared to chance of C>T substitutions at XpCpG triplets which was observed in all the cancers, albeit to different extents (arrows in Figure 4.1 highlighting variation in this signature between PD4109a and PD3945a). Additionally, subtler features of this mutational process were also apparent. The base 5' to the mutated cytosine also influenced the C>T mutation rate with an A being associated with a higher rate than a G, which had a higher rate than a C, which had a higher rate than a T (for example see PD3905a). It should be stressed that the absolute number of C>T mutations at XpCpG trinucleotides in all the breast cancer genomes is relatively modest but the normalised heat map representation emphasises the ubiquitous elevation of the C>T mutation rate at XpCpG trinucleotides because of the general depletion of XpCpGs from the human genome due to the activity of the same, or a similar, mutational process in the germline over evolutionary time.

When compared to the framework of biological signatures constructed in the introductory chapter (Table 1.1), the markedly universal nature of the elevated C>T mutation rate at XpCpG triplets is plausibly due to an endogenous and well-recognised mutational mechanism that is likely attributable to the high rate of deamination to thymine of methylated cytosines, which are usually at XpCpGs (Waters and Swann, 2000).

To support this conclusion, an analysis was performed of where C>T transitions at XpCpG triplets occur. Accordingly, these transitions are occurring at higher frequency outside CpG islands (where most CpGs are methylated) than inside CpG islands (where most CpGs are unmethylated) (OR 9.95; 95% CI 7.17-13.8; $p < 0.0001$).

4.6.2 C>X at TpCpX is over-represented and variable

There was also an over-representation of C>T, C>G and C>A mutations at TpCpX triplets which appears to be present in many breast cancers but particularly pronounced in some. Two cancers in particular, PD4199a and P4120a, show an overwhelming predominance of this mutational signature. In addition to the high proportion of T immediately 5' to the mutated cytosine in this signature, the base immediately 3' to the mutated C also appears to influence the mutational process with greater overrepresentation of TpCpA, TpCpT and TpCpG than of TpCpC. This mutational signature has previously been reported in breast cancer and might also be present to some extent in other cancer types (Greenman et al., 2007; Stephens et al., 2005; Stephens et al., 2012). It is notable that PD4199a and PD4120a are most similar to each other in the shape of the distribution of the 6-bar mutational spectra as well as in the genomic heatmap despite the difference in scale of mutations, where PD4120a has an order of magnitude more mutations than PD4199a and is different in histopathological subtype.

Given the propensity for specific sequence context and relatively ubiquitous nature of this signature, the most likely candidate for the underlying process when compared to known signatures in Table 1.1 is the endogenous DNA deamination enzyme family of AID/APOBECs. However, further evidence supporting this hypothesis will be discussed later.

4.6.3 Subtle mutation signatures and internal correlations may not be appreciable via this visual approach

This approach of using mutations at different sequence contexts for exploring the presence of mutational processes has been useful for demonstrating the presence of hypothesised signatures left behind by different mutational processes. It has also been notable for emphasizing the ubiquitous nature of some mutational processes and for highlighting the variation in the intensity of each mutational process. However, there are limitations to this purely visual approach.

Whilst the stripe of C>T transitions at XpCpG trinucleotides is instantly appreciable, other subtler features exist, for example C>G mutations at XpCpG trinucleotides, in PD3851a, PD4192a, PD4107a, PD4006a and PD4116a. Furthermore, subtle internal correlations between different mutational processes could also be pervasive but difficult to appreciate using this method.

4.7 APPLICATION OF A MATHEMATICAL APPROACH TO EXTRACT MUTATION PATTERNS

Although some major mutational processes can be discerned by visual inspection, a formal mathematical approach to extract these signatures was required in order to detect subtle processes, to provide better definition of the mutational features that define each process and to assess the relative contribution of each mutational process to the mutation set in each cancer. This application of a mathematical approach was used as a proof-of-principle to demonstrate that existing mutational signatures seen in the heatmap could be extracted and quantified, and to see if other subtler signatures were discernible. Detailed mathematical development and application of this approach was performed by Ludmil B Alexandrov and further refinements to this approach are the subject of his doctoral thesis. Here, the focus is on interpretation of the features extracted by comparison to the framework of signatures built in the Introduction.

4.7.1 Non-negative matrix factorization is a method of extracting mutational signatures from multidimensional and complex datasets

Fundamentally, the pooled somatic substitutions from the 21 breast cancer genomes was a complex, multi-dimensional dataset made up of 96 features of mutation counts of each mutation type (C>A, C>G, C>T, T>A, T>C, T>G) at each 5' and 3' base context. Within this pool of substitutions, the aim was to identify underlying mutational signatures that make up this pooled dataset. The process of extracting multiple independent signals from a pool of data is one described as a blind source separation problem with multiple known and applicable methods to achieve a solution to this problem.

Non-negative matrix factorization (NMF) and model selection is one such approach that has been previously developed to factorize or decompose complex multi-dimensional datasets in order to identify common, defining underlying signatures that make up the pooled dataset (Berry et al., 2007). To use an analogy, each human face is a complex assembly of features but which is instantly recognizable as an individual face. NMF applied to a pool of images of faces yields interpretable underlying "features" shared across the group of faces such as the eyes, nose and mouth. The aggregate of somatic substitutions of each cancer is essentially the "face" of a cancer, with each extracted "feature" equivalent to an individual mutational process.

In contrast, the application of other methods of extracting signal from noise produces components lacking obvious visual meaning (Berry et al., 2007). Furthermore, for individual faces, NMF is able to derive the contribution or the amount of exposure of each of those meaningful features. The desire to extract biologically meaningful mutational processes, as well as the intrinsic non-negativity of the

mutation spectrum data, renders NMF an appropriate choice for decomposing the mutational spectra of the 21 cases.

4.7.2 At least five mutational processes are identified across 21 breast cancer genomes

In brief, a matrix A was considered to be the complex, pooled, multi-dimensional dataset made up of 96 features (N) comprising mutation counts of each mutation type (C>A, C>G, C>T, T>A, T>C, T>G) at each 5' and 3' base context, from 21 (M) breast cancer cases. Thus, matrix *A* has a size of *96 X 21.* This dataset can be decomposed into two matrices – *W* with size *96 X k* and *H* with size *k X 21* where *k* was the number of signatures which we were trying to model and identify. NMF was performed and a model selection approach for *k* = 2 .. 20 was used to identify the optimal value of k or the ideal number of mutational processes. An optimal decomposition and value of *k* was chosen based on the cophenetic correlation coefficient (a measure of how faithfully clustering approaches preserve pairwise distances and therefore dendrogram structures) (Berry et al., 2007) and the average reconstruction error (Brunet et al., 2004).

NMF was performed using a modified version of the publicly-available implementation (Brunet et al., 2004; Lee and Seung, 1999) and was repeated 1,000 times for each value of *k* where *k* is the number of putative signatures. The cophenetic correlation coefficient indicated reproducibility and stability for k values between 2 and 6 (Figure 4.2a). The cophenetic correlation fell sharply for *k* > 6 (less than 0.95) indicating a lack of robustness when a decomposition exceeded 6 signatures for this dataset. Given a value of *k*, each sample was reconstructed and compared to the observed data. Error in reconstruction for each value of *k* was plotted (Figure 4.2b), and a dramatic reduction in the slope of the reconstruction error revealed that the model stabilised at five mutational signatures. At present, various simulation experiments are being explored in order to assess the stability and accuracy of this method. For the purposes of this study, a typical comparison between the reconstructed and observed mutation profile was sought (Figure 4.2c). The concordance indicated that five signatures were sufficient to describe the general behaviour of mutation profiles of the 21 breast cancer samples.

Figure 4.2

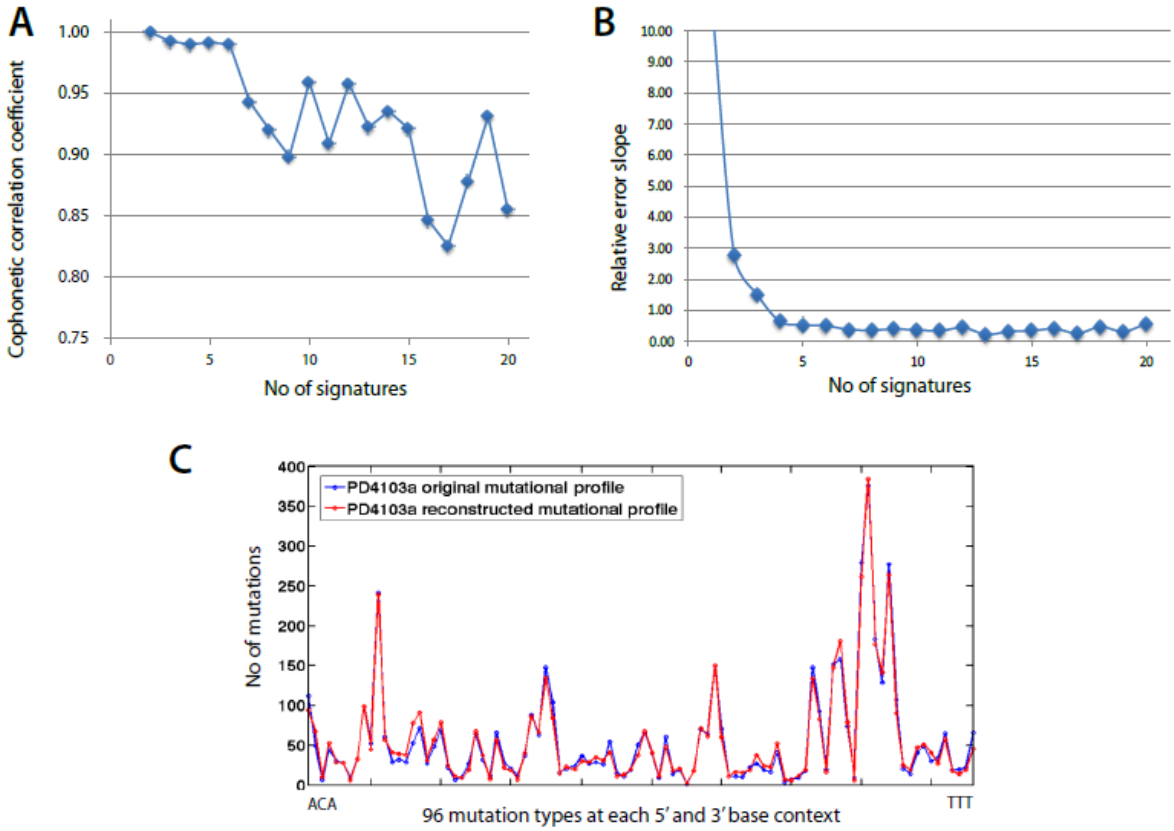


Figure 4.2: Selection of the optimal number of signatures via the NMF model selection framework. (A) The x axis depicts the number of signatures while the y axis shows the cophenetic coefficient. As an indicator of stable reproducibility, the cophenetic correlation coefficient is at its highest points between 2 and 6 processes. Given that there are no further peaks after 6 for this dataset, the number of signatures recognised by the NMF algorithm here is up to six. (B) The error in reconstruction for each number of potential signatures, k, showed a marked reduction in the slope of the reconstruction error until k = 5, suggesting that the model was stable at five mutational signatures. (C) A typical comparison between the reconstructed and original mutation profile demonstrating how well the extracted signatures and their exposures describe the original data for five signatures.

An evaluation of the decompositions by NMF suggested that a best estimate of five biologically distinct mutational processes were operative across the 21 cancers (termed Signatures A-E, Figure 4.3). Each signature was characterised by a different profile of the 96 potential trinucleotide mutations and contributed to a different extent to each of the 21 cancers, and each will be described in more detail in the following section.

Signature A was primarily characterised by C>T mutations at XpCpG trinucleotides but also included several other mutation classes making smaller contributions (Figure 4.3). This signature mirrored the dominant and ubiquitously present signature identified in the genomic heatmap in section 4.4.

Signature B was composed predominantly of C>T mutations at TpCpX, C>G mutations at TpCpA, TpCpC and TpCpT and C>A mutations at TpCpA and TpCpT trinucleotides. This signature was also visually significant in the heatmap described earlier.

Apart from reassuringly recognising the two apparent signatures in the heatmap, NMF was able to extract three additional mutational signatures. Two of the three signatures termed Signature C and Signature D both exhibited a rather small and relatively uniform distribution of mutations across the 96 trinucleotides and at first glance were rather similar. However, subtle differences were noticeable with Signature C being moderately enriched for C>T, C>G and to a lesser extent, C>A mutations at XpCpG trinucleotides (Figure 4.4a). In contrast, Signature D did not show enrichment for any particular trinucleotide and did appear to have a small and relatively uniform contribution from all 96 trinucleotides. In hindsight, an enrichment of C>G and C>A mutations at XpCpG trinucleotides can be discerned in some cancers in the heat map (Figure 4.1C). Moreover, the strength of this enrichment does not appear to be well correlated with enrichment of C>T mutations at XpCpG trinucleotides, suggesting that they are due to different processes, providing the rationale for NMF to separate Signature C from Signature A (compare, for example, PD4006a and PD3945a in Figure 1C). Finally, NMF also extracted Signature E which had a dominant feature of C>G mutations at TpCpX trinucleotides. Signature E is therefore similar to Signature B, but lacks the C>T mutations at TpCpX trinucleotides characteristic of Signature B. This extraction highlighted a subtle process not easily distinguished by visual inspection of the heatmap.

Different combinations of the five processes can account for the observed variation in the 21 mutational catalogues from the tumour set (Figure 4.1D). Biologically, this translates into varying degrees of exposure to each mutational process. NMF is also able to estimate the contributions of each mutational process for each cancer genome and this will be dealt with in section 4.10.

Figure 4.3

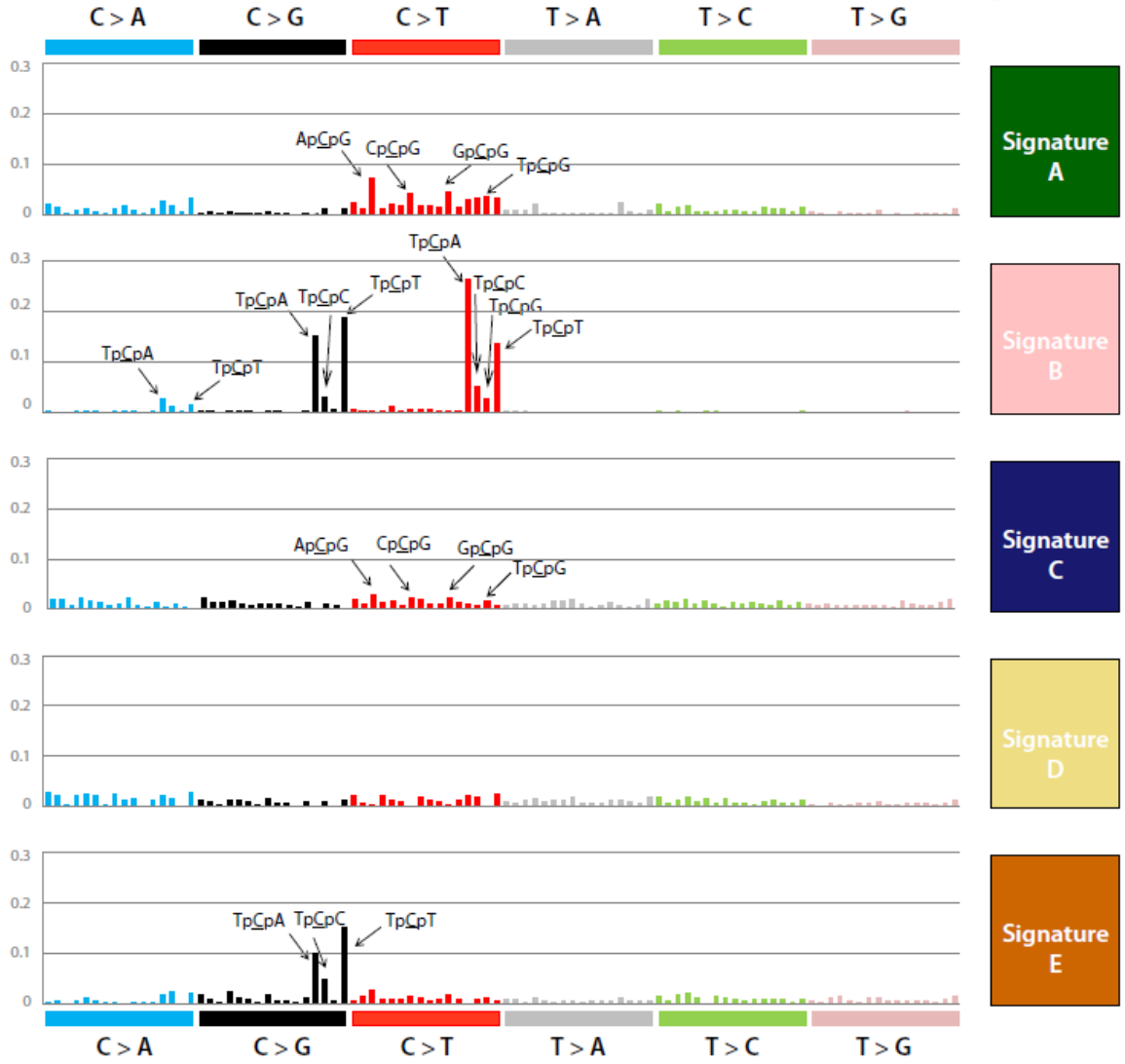


Figure 4.3: Five mutational signatures extracted by NMF in 21 breast cancers. The fraction of contribution of each mutation-type at each context for the five mutational signatures identified by NMF analysis is presented. The major components contributing to each signature are highlighted with arrows.

4.7.3 Caution with interpretation: Non-negative matrix factorization is able to detect true and artefactual mutational processes

It should be noted that application of NMF will extract mutational patterns that are due to systematic sequencing artefacts. On an earlier exploratory iteration of NMF, a signature characterised by T>G mutations at GpTpX trinucleotides was identified (Figure 4.4b). These variants did not have the hallmarks associated with true somatic variants when next-generation sequencing reads were visually inspected. They occurred after poly-T tracts, were unidirectional (present only on forward reads or only on reverse reads) and were not experimentally reproduced on verification of somatic mutations using an orthogonal sequencing methodology (Figure 4.4c). This signature has turned out to be a systematic artefact of aberrant Illumina sequence phasing at Ts following runs of Gs in the genome. It does, however, demonstrate that despite comprising less than 3% of the total mutation burden in the affected cancers, any systematic mutational process whether biological or artificial, is detectable by this analysis. This reemphasises the requirement for directed verification of each signature.

Figure 4.4

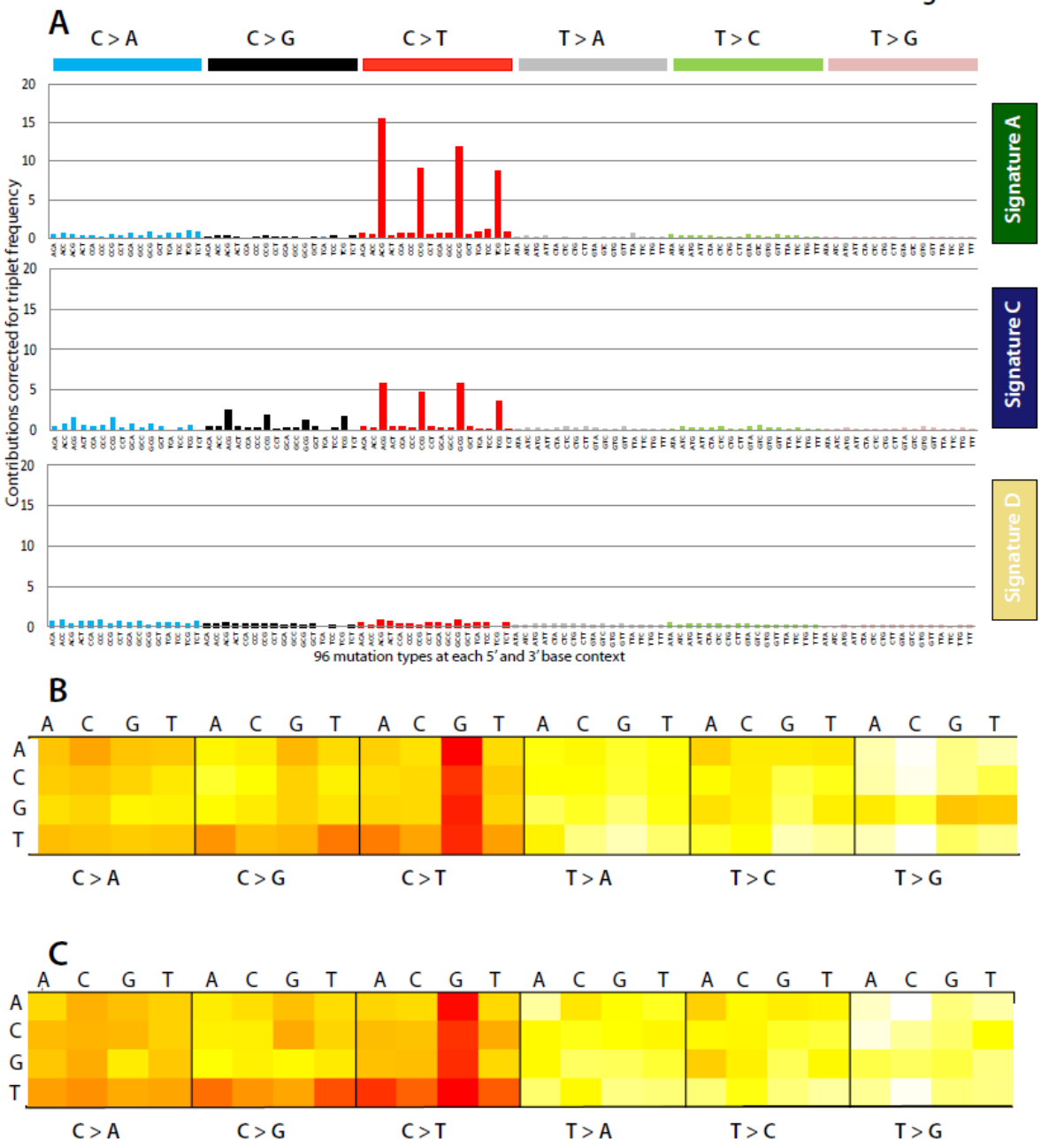


Figure 4.4: Contrasting and validating signatures. (A) Signatures A,C and D with contributions from each of the 96 trinucleotides corrected for the frequency of trinucleotides in the genome. This form of representation highlights the contrast between Signature A and C, as well as demonstrates the differences between Signatures C and D. Note the absence of C>T transitions at XpCpG in Signature D. (B) A heatmap of the combined genomes containing false positives generated by a systematic sequencing artefact of HiSeq 2000 sequencing of T>G at GpTpX dinucleotides. 5'base on the left hand vertical axis and 3'base on the top horizontal axis. Mutation type provided on the lower horizontal axis. (C) A heatmap of all variants that were successfully validated (from the same genomes as in B) shows that this signature is not reproducible in the validated variants.

4.7.4 The contribution of each mutational process for each cancer is identifiable.

For each process, NMF allowed estimation of the relative contribution of each mutational process to the final mutational catalogue of each of the 21 breast cancers and is presented as proportional barcharts in Figure 4.1D. The results indicate that most cancers have contributions from multiple mutational processes.

Several cancers, PD3851a, PD4085a, PD4108a, PD4103a, PD4194a, PD4198a, PD4248a and PD4194a showed Signature A as the modal or predominant signature, and this inclination did not appear to be restricted to any histopathological subtype. Furthermore, two breast cancers of different subtypes PD4120a (ER positive, HER2 negative) and PD4199a (ER negative, HER2 positive) were dominated largely by Signature B. In contrast, the *BRCA1* and *BRCA2* germline mutant breast cancers demonstrated modal contributions from Signature D. Signature E appeared to be present in most of the *BRCA1* and *BRCA2* germline mutant cancers and ER negative cancers but was absent from PD4107a and PD4199a, as well as PD4198a and PD3851a. Signature E made only a minor contribution to the rest of the ER positive breast cancers. There did not appear to be a mutational process that was restricted to any particular histopathological subtype.

4.8 UNSUPERVISED HIERARCHICAL CLUSTERING USING INFORMATION EXTRACTED FROM NMF CLUSTERS BREAST CANCERS WITH DEFECTS IN HOMOLOGOUS RECOMBINATION FROM OTHER BREAST CANCERS

Unsupervised hierarchical clustering was performed using the relative contributions of each of the five signatures to the mutational catalogues of the 21 genomes. Here, a priori knowledge regarding histopathological subtype was not provided to the clustering algorithm. Interestingly, all nine breast cancers with *BRCA1* or *BRCA2* mutations clustered together in one of the two major branches of the tree, whereas the remaining 12 cancers were in the alternative branch (Figure 4.5). The clustering of *BRCA1* and *BRCA2* mutant cases appeared to be predominantly due to a relatively substantial contribution by mutational process D and a relative deficiency of process A in these cancers. Notably, unsupervised hierarchical clustering did not cluster the breast cancers according to histopathological subtype.

Biologically, this is indicative of the underlying defect in homologous recombination resulting in distinguishing somatic mutational signatures. Evidence to support this comes from forcing changes in NMF parameters. Even when forced to decompose to four main mutational processes, unsupervised hierarchical clustering based on these four processes continued to result in a persistent separation of germline mutant breast cancers from sporadic breast cancers.

Furthermore, previous exploration of the dataset using other mathematical approaches such as principal components analysis and factor analysis showed that germline mutant breast cancers were separating from sporadic breast cancers on identifiable components from the 96 features. The use of different methods of mathematical decomposition resulted in a similar marked separation suggested that distinguishing mutational features were an inherent characteristic of the full catalogue of somatic mutations in the 21 genomes and not simply restricted by the choice of mathematical model used.

Figure 4.5

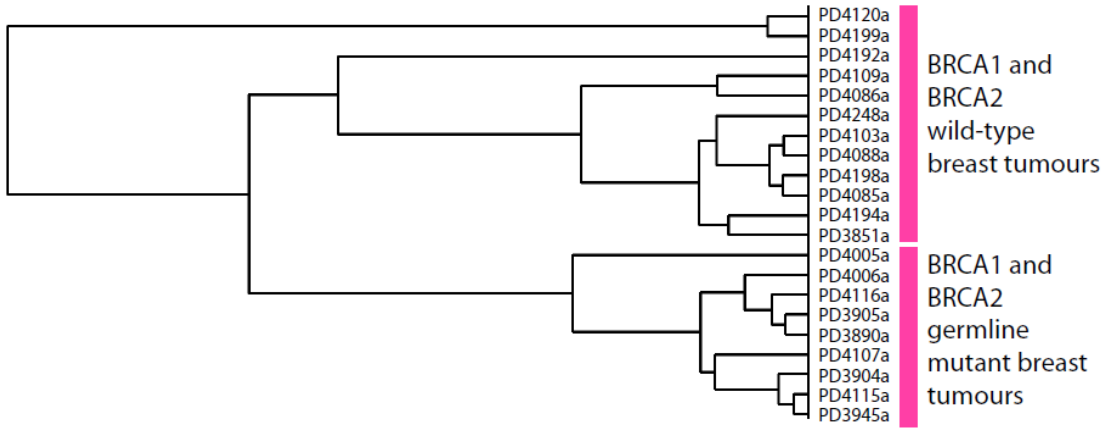


Figure 4.5: Cluster dendrogram generated by unsupervised hierarchical clustering based on contributions of the five mutational signatures identified by NMF for the 21 breast cancer genomes.

4.9 THE IDENTIFICATION OF A *BRCA1* GERMLINE MUTATION AND PREDISPOSITION TO CANCER USING THIS APPROACH

Clinical history including germline mutation statuses were obtained from the respective collaborators who provided samples. One particular breast cancer, PD4107a was initially thought to be a sporadic triple negative breast cancer. The patient was a 33 year old woman at diagnosis, relapsed within 15 months of surgery and died of aggressive metastatic breast cancer shortly after. There was no family history of breast or ovarian cancer.

The unsupervised hierarchical clustering approach clustered PD4107a with other cancers carrying defects in *BRCA1* and *BRCA2*, genes involved in DNA double-strand break repair by homologous recombination. Given this result, cryptic germline mutations in the *BRCA1* and *BRCA2* genes were sought in all the samples. Surprisingly, a 1 bp indel was identified in exon 11 of the *BRCA1* gene in PD4107a, predicted to result in a p.V340fs*6 change, and is a reported deleterious variant in HGMD (Human Gene Mutation Database).

This approach of clustering somatic mutation signatures provides independent verification of the biological effects of the germline indel identified in this patient. Indeed, as the germline mutation status of this patient was not known prior to this study, it appears that the somatic mutation profiling and clustering approach used here was able to predict germline BRCA1/BRCA2 status, thereby predicting germline predisposition to cancer for this family.

Apart from this connection with germline BRCA status, no correlation was found between the presence of a particular somatically mutated gene and any of these processes. It is worthy of note that both PD4120a and PD4199a are dominated by Signature B and globally mutated by C>T mutations at TpC context and both have *TP53* mutations. However, many other breast cancers also carry somatically-acquired *TP53* mutations and do not demonstrate this phenotype. The number of samples in this study is likely to be too small to draw any conclusions on this issue but it would be interesting to explore a permissive state for global hypermutation provided by a defective TP53 pathway.

4.10 THE CONTRIBUTIONS OF MUTATIONAL SIGNATURES CHANGE OVER EVOLUTIONARY TIME

Apart from identifying individual mutational signatures in each breast cancer and the contributions of individual signatures to each cancer, the processes that generate the different signatures may vary in temporality, with some mutational processes occurring early in the evolution of a cancer, and others occurring later. In this section, integration of other somatic changes with base substitutions is used to seek insight into timing of mutational processes.

4.10.1 Integration of copy number with base substitutions to inform temporality of mutation events

Copy number changes are a common feature of many cancers. In breast cancers, several genomic regions show loss of one parental chromosome (loss of heterozygosity) followed by re-duplication of the remaining copy. In such regions, mutations which occurred early or before the re-duplication event will be homozygous, whereas those arising late or after re-duplication will be heterozygous (Figure 4.6A). Furthermore, the presence of distinct clusters of mutations at variant allele fractions lower than expected for the estimated ploidy and degree of normal contamination suggests the presence of subclonal populations.

4.10.2 The rationale for interrogating timing of mutational processes

Comparisons of somatic substitutions that occurred relatively early in the evolution of the cancer with those that occurred later in such informative regions, have revealed differences in their mutational spectra in the past (Pleasance et al., 2010a). For example, examination of the spectra of a metastatic malignant melanoma cell line following the integration of copy number data with base substitutions, revealed that C>T mutations related to ultraviolet light exposure accounted for a higher proportion of early compared to late mutations, contrasting with C>A changes which accounted for a higher proportion of late mutations (19% to 2%). The authors hypothesised that this was consistent with early mutational processes driven by exposure to ultraviolet light resulting in the C>T mutational signature whilst another unrelated mutational process was likely to be underlying the late C>A mutations.

4.10.3 The temporality of mutational processes

Previously, non-negative matrix factorization (NMF) identified five separate processes from the pooled dataset across the 21 breast cancer genomes. By classifying whether mutations were early, late or subclonal in regions of copy number gains (Figure 4.6A), the relative contributions of these five processes at different times during a cancer's evolution (Figure 4.6B) could be assessed.

However, this analysis is restricted to breast cancer samples that have a sufficient number of mutations present in such regions to generate a stable NMF solution. This was possible in eight patients (Figure 4.6C). In these eight cancers, Signature A characterised by C>T mutations at CpG dinucleotides, contributed a relatively large proportion of the early mutations in all cancers compared to late in the evolution of the tumours. In contrast, Signature E, denoting C>G mutations at TpCpA, TpCpC and TpCpT trinucleotides, was a late onset mutational signature, contributing a large fraction of subclonal mutations in many patients. Hence, the data indicated that the mutational processes moulding the breast cancer genomes vary over evolutionary time.

Figure 4.6

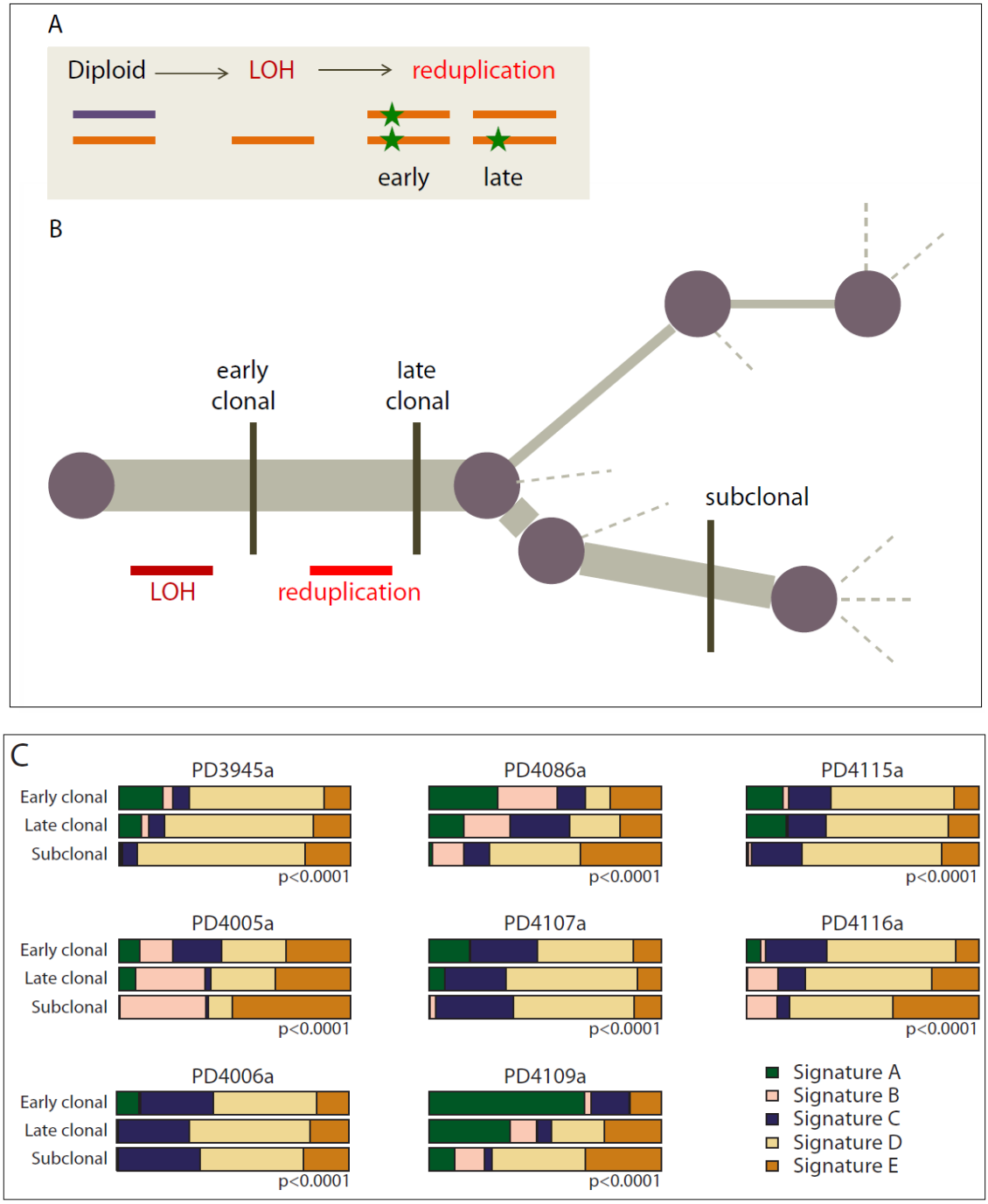


Figure 4.6: Temporality of mutational processes (A) Starting from a diploid state, loss of one parental allele will lead to a state of loss of heterozygosity and a ploidy of 1. However, reduplication of this single allele can occur. If a mutation occurs on the single allele early in the evolution of the cancer, prior to the reduplication event, then the mutation will appear homozygous. Conversely, a mutation occurring later in the evolution of the cancer, after the reduplication event, will be heterozygous. Subclonal mutations are identified as mutations occurring at a variant allele fraction that is less than what would be expected for the level of ploidy for that chromosome and the degree of normal contamination in the cancer sample. (B) The groups of mutations classed as early clonal, late clonal and subclonal depicted within the phylogenetic evolution of the cancer. (C) Stacked bar charts showing comparison of mutational processes identified by non-negative matrix factorization. The comparison is across early clonal mutations (ploidy > 1), late fully clonal mutations (ploidy = 1) and subclonal mutations (ploidy < 1) for 8 samples. Signature A describes C>T mutations at XpCpG trinucleotides. Signature B was composed predominantly of C>T, C>G mutations and C>A mutations in a TpC context. Signature C and Signature D were relatively uniform processes across all 96 possible mutated trinucleotides. Signature E specifically identifies C>G mutations at TpCpA, TpCpC and TpCpT trinucleotides.

4.11 DISCUSSION

Analysis of the catalogues of somatic mutation from 21 breast cancers has yielded several insights into the nature of the underlying mutational processes that have shaped the cancer genomes. By considering the flanking sequence context of each mutation, multiple mutational patterns were visually appreciable using a genomic heatmap which also highlighted the variation in intensity of each mutational pattern across the 21 genomes. Reinforcing this observation, application of Non-negative Matrix Factorization (NMF) suggested that five independent single nucleotide substitution processes had been operating to different extents across the cancers, generating the observed variation in mutation numbers and patterns. It is possible, however, that additional subtle processes exist and will become apparent with refinements in the design and application of the algorithm. The processes generally appeared to have been acting in combination in each breast cancer case and could vary in temporality through the development of the cancer.

4.11.1 High-quality data with low false positive rates were essential for these analyses

In order to examine the catalogues of somatic mutations for mutational signatures, considerable effort was put into generating clean datasets with low false positive rates. The necessity for accurate mutation-calling was reinforced by the detection of a mutational signature by NMF characterised by T>G mutations at a GpT dinucleotide context. This systematic sequencing artefact was one of several known systematic sequencing artefacts which arose during Illumina sequencing. Despite the smallest amount of this artefact in only 4 or 5 samples, it was detectable by the mathematical approach used to extract mutational signatures emphasizing the potential sensitivity of NMF but also the potential for misinterpretation. A systematic sequencing artefact is arguably a mutational process, albeit one which occurred during sequencing rather than a biological mutational process which had occurred during the development of a cancer. As sequencing technology and chemistry improves and brings greater yields per lane of sequencing, it is anticipated that novel sequencing artefacts are likely to arise. Intermittent surveys or curation of whole genome sequencing datasets will continue to be required in order to maintain specificity of mutation-calling for accurate interrogation of mutational signatures.

4.11.2 Limitations of these analyses: More samples and refinements to the Non-negative matrix factorisation approach

This study utilised data from only twenty-one whole genome sequenced breast cancers. It is anticipated that as more breast and other cancer genomes come to be sequenced, more signatures will come to light. Expected signatures include those already recognised as being causal with exogenous mutagenic damage like smoking, ultraviolet radiation, alkylating agents and aristolochic acid consumption. In addition, cancers with known epidemiological correlates may reveal specific signatures in association with distinct aetio-pathogenesis, for example, hepatocellular carcinoma and alcohol versus virus-driven cancer. However, it is hoped that other signatures associated with perhaps reactive oxygen species, other endogenous mutagens and repair defects may reveal themselves. Furthermore, when more cancers become available for analysis, closer examination of clustering relationships may reveal sub-clustering of cancers which were not appreciable from an analysis of just twenty-one breast cancers. Insights may also be gained in the near future from cancers derived from people with other germline mutations (e.g. *PTEN*, *TP53, VHL*) and correlations between signatures and somatically mutated genes may become informative with increasing numbers of sequenced cancer samples, pending refinements to the NMF model.

4.11.3 Comparing cancer-detected signatures with known mutational signatures curated from the literature

One of these processes bears a strong resemblance to the familiar mutational mechanism that results in C>T transitions and is mediated by the elevated rates of deamination of 5-methylcytosine usually found at XpCpG trinucleotides (see introduction). Furthermore, this mutational process appears to occur early in the evolution of the cancer and may reflect a background mutagenic process possibly occurring in the breast cell before the transformation into cancer.

The mechanisms underlying the remainder are currently unknown. The most distinctive of these signatures, Signature B, is characterised by C>T, C>G, and to a lesser extent, C>A substitutions at TpCpX trinucleotides, is responsible for the overwhelming majority of mutations in two cancer samples, PD4120a and PD4199a. These two cancers are most similar to each other and most dissimilar to the other breast cancers, despite having an order of magnitude difference in mutation burden (70690 versus 6932 total substitutions). These two cancers are also of divergent histopathological subtypes; PD4120a is an ER positive, PR positive and HER2 negative breast cancer, whilst PD4199a is an ER negative, PR negative and HER2 positive cancer suggesting that the underlying mutational process generating this striking signature is independent of and unrelated to

expression-based profiling. Signature B has similarities with the mutational signature produced by the endogenous deaminating enzyme superfamily described in the introductory chapter, the APOBEC family.

Although off-target deamination by AID is likely responsible for the mutations and translocations seen in many B cell tumours [reviewed in (Nussenzweig and Nussenzweig, 2010)], AID is unlikely to be the enzyme responsible for the mutational processes described here since it exhibits a strong preference for deaminating C residues flanked by a 5'-purine (Pham et al., 2003). In contrast, the Cs targeted in Signature B in the breast cancer genomes are nearly all preceded by a 5'-T. However, both APOBEC1 (when acting on DNA) as well as all the APOBEC3 enzymes (apart from APOBEC3G) favour C residues flanked by a 5'-T (Harris et al., 2002; Hultquist et al., 2011). Furthermore, transgenic overexpression of APOBEC1 is associated with cancer (Yamanaka et al., 1995) and although most APOBEC3s are thought to function in the cytoplasm, recent results (Landry et al., 2011; Stenglein et al., 2010) indicate that enforced overexpression of APOBEC3A can result in genomic damage and mutation (Suspene et al., 2011). Thus APOBEC1 as well as some of the APOBEC3s constitute attractive candidates for being responsible for Signature B.

Thus far, it has not been possible to demonstrate a clear correlation between over-expression of any member of the AID/APOBEC family and Signature B. This is confounded first by a relatively small dataset and second by absence of expression data from key samples. Notwithstanding, an absence of over-expression at the time of cancer diagnosis would not preclude activity of a member of the AID/APOBEC family earlier in evolution of the cancer. The features characterising Signatures C, D and E have not been previously described.

4.11.4 Somatic mutational signatures of breast cancers with *BRCA1* and *BRCA2* germline mutations

The similarity between the mutational profiles of *BRCA1* and *BRCA2* mutant cancers contrasts with the differences observed in their histological characteristics, immunohistochemical features and mRNA expression profiles. *BRCA1* mutant cancers have characteristic high grade histology, are ER, PR, HER2 negative and locate with basal-like breast cancers in hierarchical clustering of expression levels (Hedenfalk et al., 2001; Palacios et al., 2008; Perou et al., 2000; Sorlie et al., 2001a). Conversely, *BRCA2* cancers have histology that is overall similar to age matched cases, are generally ER positive and cluster with luminal A or B cancers (Palacios et al., 2008). Thus the mutational patterns, which are plausibly more closely related to the underlying biological defect, appear to be

reporting the similarities in underlying disease pathogenesis between *BRCA1* and *BRCA2* mutant cancers better than analysis of cellular phenotype.

BRCA1 and *BRCA2* wild type cancers, including the three triple negative cases, did not show these mutational features. It remains to be seen, however, from more extensive series whether other modes of inactivation of BRCA1 or BRCA2, for example by methylation, have similar mutational patterns. *BRCA1* and *BRCA2* cancers are particularly responsive to certain DNA damaging agents and inhibitors of other DNA repair processes, notably PARP inhibitors (Fong et al., 2009). Since there are reports of cancers without mutations in *BRCA1* and *BRCA2* responding to these treatments (Harris et al., 2002), it will be interesting to explore whether the presence of the mutational patterns characteristic of *BRCA1* and *BRCA2* cancers, which are indicators of the critical defects in DNA repair, are better predictors of response to these therapies than the presence of mutations in the two genes.

Intriguingly, *BRCA1* and *BRCA2* are different genes which generate different proteins and have differing roles in the repair of double-strand breaks. They do, however, converge on the unifying principle of homologous recombination repair and despite arising from a variety of germline defects in two different genes, appear to produce similar mutational signatures in this analysis. This observation may serve as an early clue that mutational signatures may be informative of an abrogated pathway even without knowledge of the precise gene defect. Perhaps, as we sequence more cancers, informative mutational signatures will serve as an indicator of which pathways cancers are also addicted to and these may become targets of therapeutic intervention.

Why BRCA1 and BRCA2 cancers have greater representation from Signature D, a fairly non-specific and uniform signature, is uncertain. It is notable that BRCA1 has been shown to have a role in post-replication repair, contributing to the response to UV irradiation. It is recruited to UV-damaged sites in a replication-dependent but nucleotide excision repair independent way. At replication forks stalled by UV-induced damage, it has a number of roles including promoting excision of the damaged base, localization and activation of replication factor C complex (RFC) subunits which triggers checkpoint activation, post-replicative repair and suppression of translesion synthesis (Pathania et al., 2011). These functions are distinct to those observed in double-strand break repair. It is possible that the overall increase in background mutations resulting in Signature D may be due to the increased impact of translesion polymerases given defective BRCA1/BRCA2.

4.11.5 The temporal variation in mutational processes may reflect normal processes and tumour-specific processes that have occurred over the phylogenetic development of the cancer

These data also indicate that mutational processes shaping the breast cancer genome vary over time. The mutational process of deamination of methylated cytosines plays a significant role in the early acquisition of mutations. It is possible that this is a default mutation spectrum, given that it is seen in many tumour types such as blood, pancreatic and brain cancer (Greenman et al., 2007; Jones et al., 2008; Papaemmanuil et al., 2011; Puente et al., 2011) and is a feature of germline nucleotide substitutions (Hwang and Green, 2004). Indeed, it is possible that it is a mutational signature that may well represent processes occurring in normal tissues. The higher proportional contribution of other variant-types among late mutations in most of these breast cancers could be explained by an increase in the rate of other mutation types which may reflect tumour-specific mutagenic signatures.

CHAPTER FIVE: LOCALISED HYPERMUTATION OR *KATAEGIS* IS PRESENT IN THIRTEEN OF TWENTY-ONE BREAST CANCER GENOMES

5.1 INTRODUCTION

In the previous chapters, patterns of somatic substitution were sought from the dataset generated by whole genome sequencing of breast cancers. However, these analyses did not explore the possibility that mutations in cancer genomes are non-randomly distributed and may show regional clustering.

There is some evidence of geographic clustering of base substitution mutations in experimental systems. For example, it has been shown that multiple mutations occurred at an unexpectedly high frequency within the *lacI* mutation target in the Big Blue transgenic mouse system (Wang et al., 2007). It was demonstrated statistically that clustered mutations in this system were likely to be the result of "mutation showers", giving rise to an average mutation rate of one mutation per 3kb (Wang et al., 2007). Such clustered mutations imply transiently hypermutable moments during cell division.

Transient hypermutability is implicit in two examples of large-scale structural variation seen in cancer. A phenomenon called *chromothripsis,* characterised by tens to hundreds of chromosomal rearrangements, localised to a limited genomic region was described recently and the rearrangements were believed to be acquired in a single catastrophic event (Stephens et al., 2011). Furthermore, gene amplification events which are a relatively common occurrence in cancer arise through cycles of breakage-fusion-bridge and are also locoregional. Both of these gross genomic mutational events show topographic clustering and are thought to be triggered by a stochastic insult.

Substitution mutation showers have not, to the best of my knowledge been reported in cancer genomes. This is because historical analyses of mutation spectra in cancer genomes have been mainly restricted to the use of cancer genes in gene reporter assays. In this chapter, expounding the benefits of whole genome sequencing, the possibility of variation in mutation prevalence across twenty-one cancer genomes, is explored.

5.2 THE EXPLORATION OF VARIATION IN MUTATION RATE IN CANCER GENOMES USING "RAINFALL PLOTS"

Each cancer genome explored in this thesis produced many thousands of substitution variants. A method of visualising the variation in mutation rate was required. In order to avoid bias by the

introduction of “genomic bins” in presenting mutation rates across the genome, the possibility of regional variation in mutation rate and potential clustering of substitutions was investigated by calculating an intermutation distance, or the distance between each somatic substitution and the substitution immediately prior to it on the reference genome (Figure 5.1a). Intermutation distances were plotted on the vertical axis on a log base 10 scale with mutations ranked and ordered on the *x* axis from the first variant on the short arm of chromosome 1 to the last variant on the long arm of chromosome X, in what have been termed “rainfall plots”. The advantage of these genome-wide rainfall plots is that they provide a perspective on the number of mutations involved in each region of hypermutation (Figure 5.1b).

At a mutation rate of ~ 1 in every 100kb to 1 in every 1Mb, most mutations in a cancer genome would therefore have an intermutation distance of ~10^5bp to ~10^6bp, approximating to where a dense cloud of mutations is situated on a rainfall plot (Figure 5.1b). Conversely, localised regions of hypermutation would present as clusters of substitutions at lower intermutation distances (Figure 5.1c).

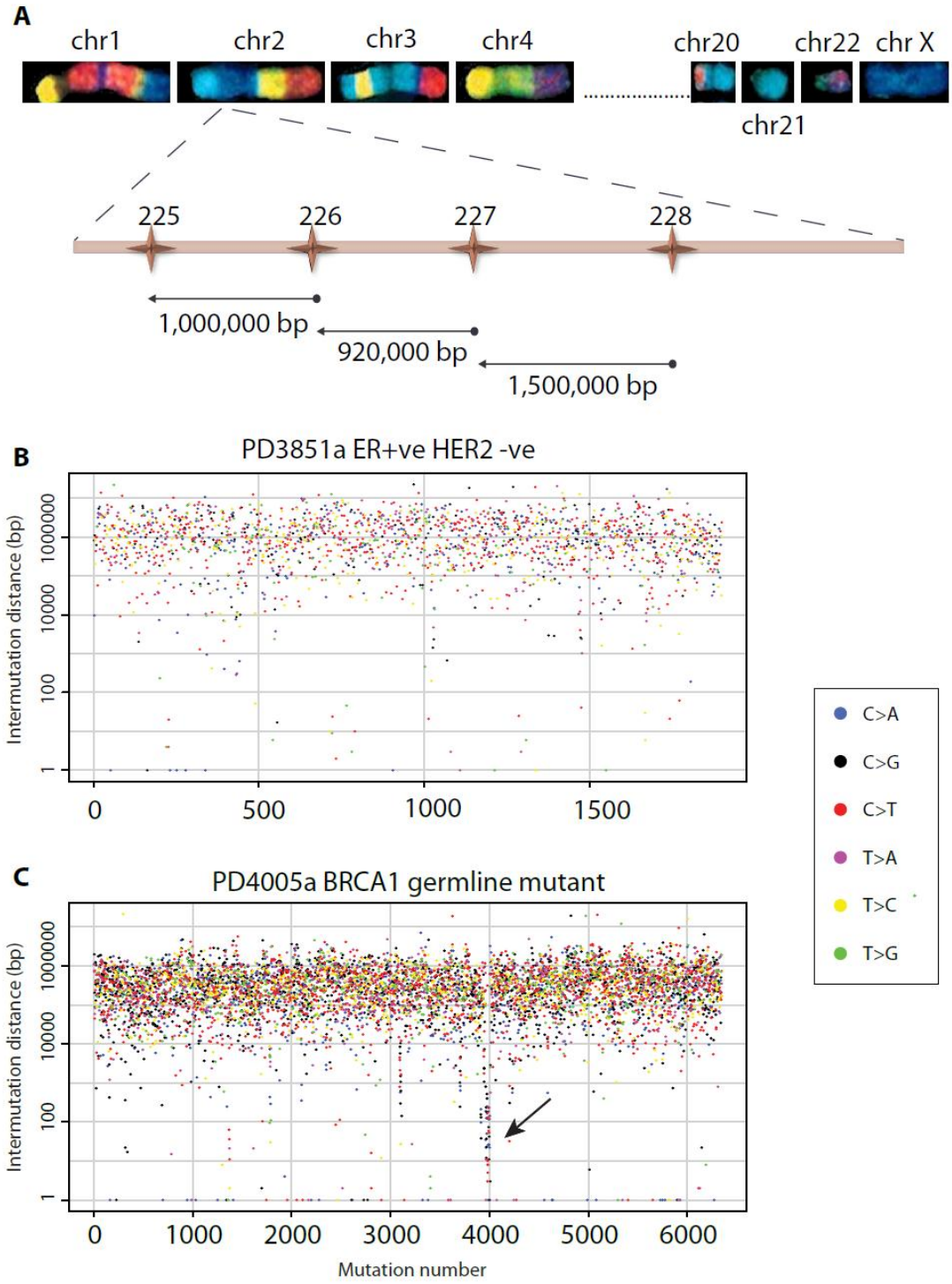


Figure 5.1: The principle behind rainfall plots. (A) Intermutation distance is the distance between each somatic substitution and the substitution immediately prior to it on the reference genome. In the example above of a region on chromosome 2p, mutation number 226 has an intermutation distance 1,000,000bp, mutation number 227 has an intermutation distance of 920,000bp and mutation number 228 has an intermutation distance of 1,500,000bp. (B) Mutations are ordered on the x axis from the first variant on the short arm of chromosome 1 to the last variant on the long arm of chromosome X and are coloured according to mutation-type. The distance between each mutation and the one prior to it (the intermutation distance) is plotted on the vertical axis on a log scale. Most mutations in this genome have an intermutation distance of ~10^5bp to ~10^6bp. (C) Mutations in a region of hypermutation present as a cluster of lower intermutation distances (example indicated by arrow).

5.3 REGIONAL HYPERMUTATION WAS OBSERVED IN THE BREAST CANCERS

Strikingly, clusters of substitution hypermutation were seen in several breast cancers and had remarkable characteristics which will be illustrated below using two cases, PDD4107a and PD4103a, as foremost examples. PD4107a, a breast cancer derived from a patient with a germline mutation in *BRCA1*, showed a markedly elevated mutation prevalence over a 14MB region on chromosome 6 (chr6:126,000,000-138,000,000) (Figure 5.2a). This accounted for 699/10291 mutations in this genome, was the largest regional cluster of mutations amongst the 21 breast cancers and exhibited several notable features which will be illustrated below.

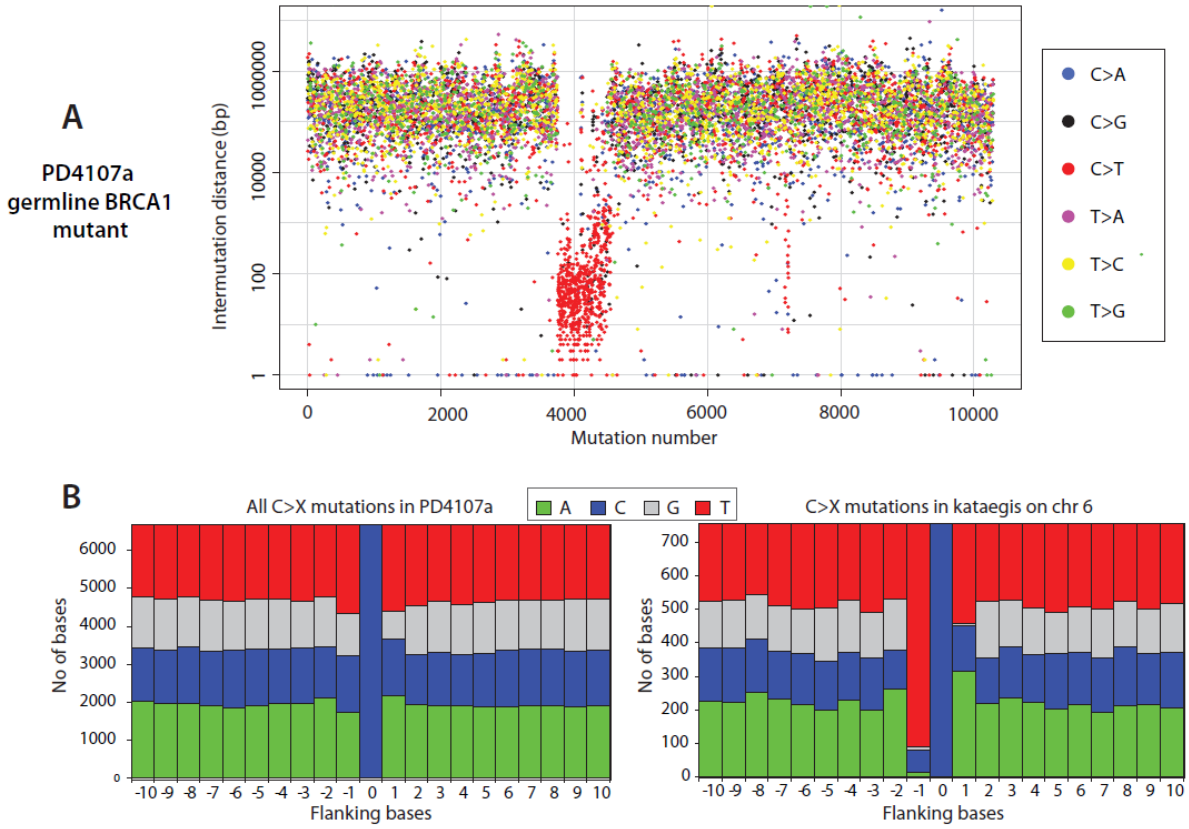


Figure 5.2: (A) Rainfall plot of PD4107a. (D) Plots of flanking sequence of all C>X mutations in PD4107a and C>X mutations within the regions of kataegis in PD4107a. Mutated base is at position 0 with ten bases of flanking sequence provided, demonstrating a strong preference for T at the -1 position in the region of kataegis.

5.3.1 "Microclusters" are present within the "macrocluster" in PD4107a

Within the hypermutated 14MB region on chromosome 6 in PD4107a, there were 699 variants. This collection of substitutions accounted for 6.79% of the total number of substitutions in this cancer and has been termed a "macrocluster". There was, however, evidence of further clustering within the macrocluster, with heavily mutated stretches of genome of a few hundred base pairs carrying anything between 6-165 mutations often separated by tens of kilobases without mutations (Figure 5.3a). These were termed "microclusters". The microcluster at chr6: 126430855-126437625 was the longest and most densely mutated cluster and contained 165 variants over a distance of ~6.7kb corresponding to a prevalence of ~2.4 mutations per 100bp. Although multiple microclusters in chromosome 6 of PD4107a were geographically macroclustered, in other breast cancers solitary microclusters were more commonly observed. Indeed, a solitary microcluster is present in another region in the same cancer, PD4107a, at chromosome 12: 10507568-10508972 (Figure 5.2a). These showers of substitution variants have been termed "kataegis", which is Greek for showers/thunderstorms/ "towards the earth".

5.3.2 *Kataegis* shows a distinctive mutational spectrum

Substitutions within this region were characterised by a distinctive mutational spectrum and sequence context (Figure 5.2b). 630 out of 699 variants (90.1%) in this region comprised C>T/G>A transitions. There was also a distinctive sequence context in which these mutations occurred. When presented in pyrimidine context, 579 out of the 630 mutations at a cytosine base (91.9%) were preceded by a 5' thymine. This was in contrast to the spectrum exhibited in the full catalogue of somatic substitutions in PD4107a where 6235/10291 (60.6%) of variants were substitutions at cytosine bases, of which 2213/6235 (35.5%) were at a TpC context. Thus C>T, and to a much lesser extent C>G and C>A mutations, at TpCpX trinucleotides were highly enriched in this region of kataegis compared to the remainder of the genome (Figure 5.2b).

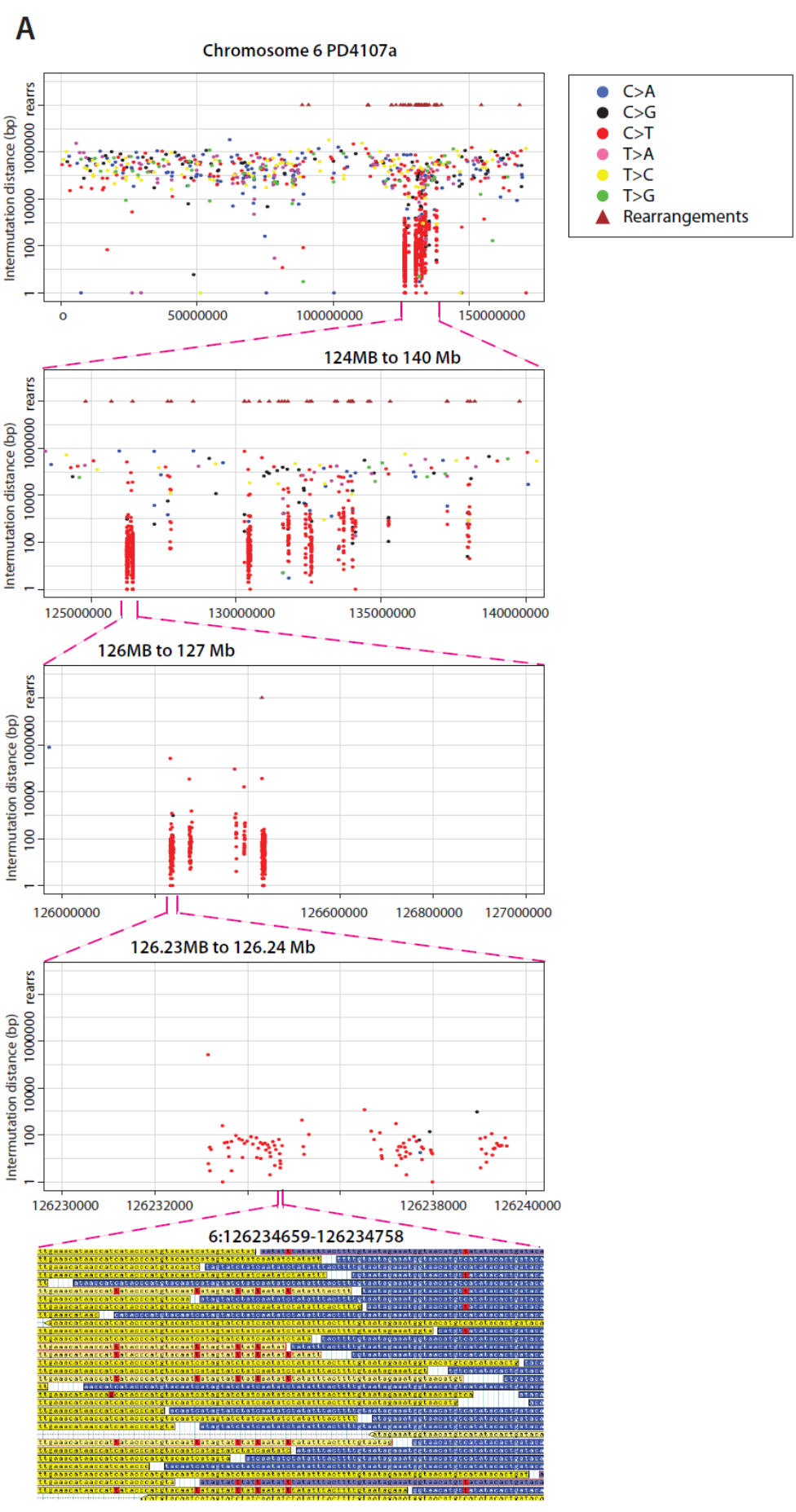
5.3.3 Mutations in microclusters of kataegis occur on the same parental chromosome.

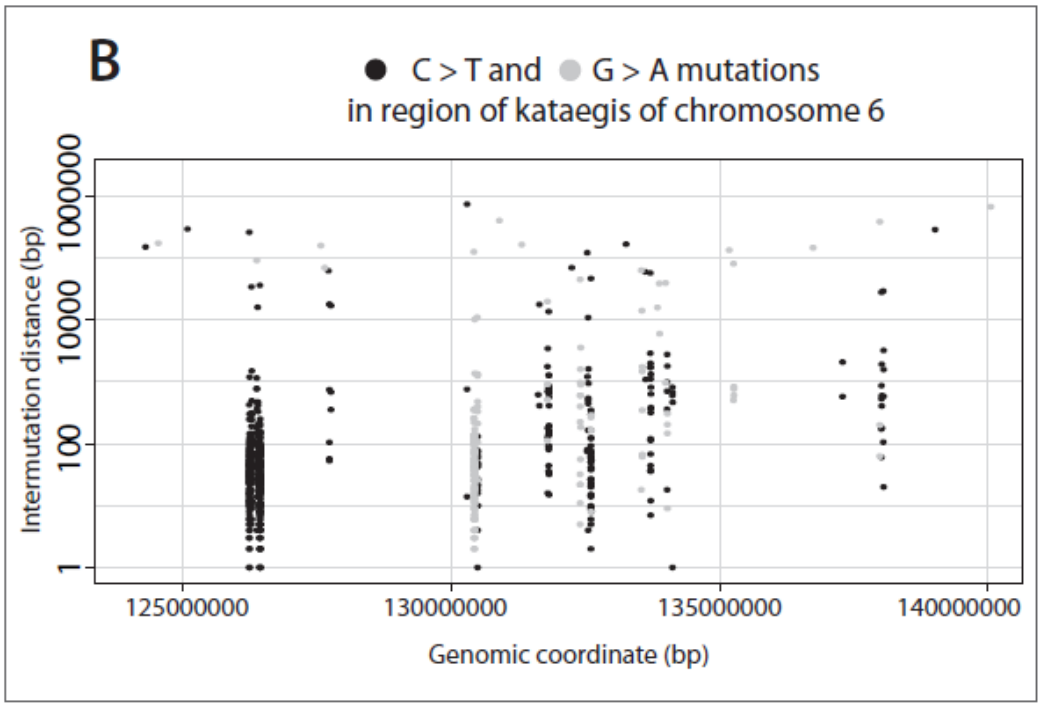
Clustering of mutations could, in principle, reflect the presence of mutations on one or alternatively on both parental alleles at particular positions. To explore these two possibilities further, individual next-generation sequencing reads which derive from individual DNA molecules were interrogated, and it was revealed that all mutations which were within one next-generation sequencing read (100bp) of another mutation within microclusters, occurred in *cis* with respect to each other (481 of

481 variants (100%), Figure 5.3a). This was subsequently verified in validation experiments where variants in kataegis were sequenced on an orthogonal platform.

5.3.4 Mutations within regions of kataegis show evidence of "processivity"

As mentioned previously, substitutions in kataegis show a predilection for C>T/G>A mutations in PD4107a. In principle, therefore, mutations in clusters of kataegis could be mixtures of C>T and G>A changes or, alternatively, runs of C>T or runs of G>A. Analysis of single sequence reads indicates, however, that mutations were generally of the same type for long genomic distances and then could switch to a different class. For example, in PD4107a, mutations in the longest microcluster, chr6: 126430855-126437625, were almost exclusively C>T (161 of 165 (97.6%) on the plus chromosomal strand). In a different cluster, chr6:130483111-130489124, ten of eleven mutations were G>A mutations in the first 4649bp and then switched to C>G and C>T mutations for the following 27 mutations in the next 1364bp (Figure 5.3b). This propensity of mutations to demonstrate this asymmetric distribution with respect to chromosomal strand has been termed processivity.


A
Chromosome 6 PD4107a
Intermutation distance (bp)
rearrs
1000000
10000
100
1
0
50000000
100000000
150000000
C>A
C>G
C>T
T>A
T>C
T>G
Rearrangements
124MB to 140 Mb
125000000
130000000
135000000
140000000
126MB to 127 Mb
126000000
126600000
126800000
127000000
126.23MB to 126.24 Mb
126230000
126232000
126238000
126240000
6:126234659-126234758



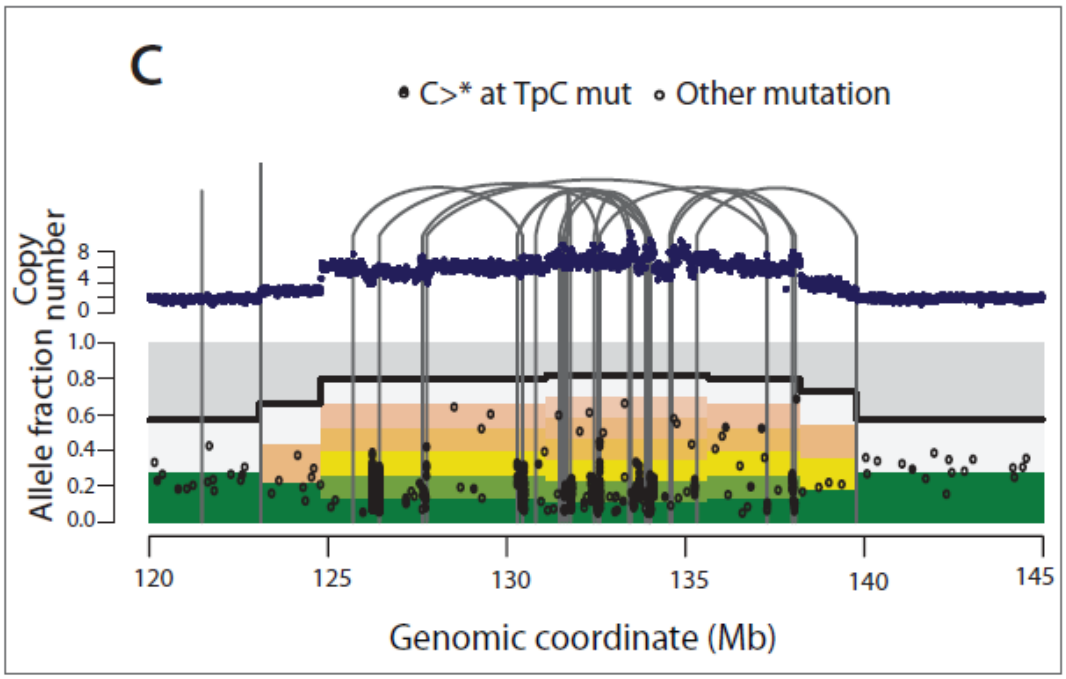


Figure 5.3: Rainfall plot for chromosome 6 of PD4107a. (A) The x axis shows the genomic coordinates of the mutations. Rearrangements are presented as brown triangles (rearrs=rearrangements). The region of kataegis is highlighted at increasing resolution to demonstrate microclusters within the macrocluster. The processive nature of C>T mutations at TpC context occurring in cis is seen in the lowest panel (G-browse image). (B) Alternating processivity of kataegis in PD4107a. Long regions of C>T mutations are interspersed with regions of G>A mutations. (C) Kataegis occurs with a variety of rearrangement architectures. Thick top line shows the copy number segments for the region of chromosome 6 of PD4107a. Point mutations are shown in lower panel as black points. X axis reflecting genomic position and y axis represents variant allele fraction. The proportions of reads derived from contaminating normal cells are depicted in grey and the fraction coming from each of the copies of that segment in the tumour cells are depicted by the multiple bars from green to yellow to pink to white. Early mutations will be found relatively higher up these bars, whereas late ones will be seen down the bottom of the variant allele fraction. Grey vertical lines represent rearrangements. Interconnecting lines indicate intrachromosomal rearrangements. On a macroscopic scale, this demonstrates how kataegis can be associated with chromothripsis (within region 130-135MB) as well as other rearrangement architectures.

5.3.5 Substitution hypermutations co-localise with rearrangements in some clusters of kataegis

To explore whether there were other characteristic features of regions of kataegis, the relationship between kataegis and other mutation classes was next examined. Surprisingly, the cluster of substitution mutations on chromosome 6 co-localised with a cluster of somatic genomic rearrangements (Figure 5.3a). Within the hypermutated region of ~14Mb, there were 18 genomic rearrangements while only eight were detected in the remaining 157Mb of chromosome 6. Most of these rearrangements were between different locations within the chromosome 6 14MB region (intrachromosomal) and only two were interchromosomal, one was involved in a rearrangement with chromosome 1 and the other with chromosome 16. Although there was clearly a positional correlation between the presence of rearrangements and substitution hypermutation, at higher resolution mutation microclusters were not usually found directly adjacent to rearrangements and were usually separated from the nearest rearrangement by many kilobases.

These regions of hypermutation coincided with a variety of different rearrangement architectures. The highly rearranged segment of chromosome 6 in PD4107a harboured a very small region of chromothripsis, nestled within a degree of low-level amplification. Hypermutated substitutions appeared to occur in conjunction with chromothripsis as well as other rearrangement architectures, and were not confined to highly rearranged regions either. For example, in PD4107a, an additional, much smaller mutation cluster, with similar mutational characteristics to the major cluster was also physically associated with a single genomic rearrangement observed on chromosome 12 (Figure 5.3c).

The rearrangement junction or the breakpoint of any structural variation in cancer genomes can provide insights into the mechanisms which have generated the rearrangements in the first place. The rearrangements that were associated with the region of kataegis appeared to show less microhomology and/or non-templated sequence at the rearrangement junction than average for the rearrangements although this did not reach statistical significance.

5.3.6 Kataegis occurs in PD4103a, a breast cancer of different histopathological subtype

An ER-positive breast cancer, PD4103a, also exhibited clusters of localised hypermutation. The pattern of mutation clustering in this cancer differed, however, in several ways from that described above for PD4107a (Figure 5.4a). The mutation clusters in PD4103a spanned shorter distances than the major cluster in PD4107a and involved many chromosomes including chromosomes 3, 4, 8, 10, 11, 12, 20 and 21. The clustered substitutions in PD4103a included C>T transitions at TpCpX dinucleotides, similar to PD4107a, but in addition, showed a greater proportion of C>G mutations which were also at TpCpX trinucleotides. In other respects, notably the mutations being in *cis* and showing a processive pattern, there were many similarities (Figure 5.4b). Moreover, in this cancer the mutation clusters were also closely associated with somatic genomic rearrangements and the characteristics of the junctional features were very similar to that of PD4107a (Table 5.2). Indeed, the regions in which mutation clusters were found were all linked together by a web of interchromosomal rearrangements (Figure 5.4c). It is notable that PD4103a is of a different histopathological subtype suggesting that this phenomenon is not restricted to a specific breast cancer subgroup.

Table 5.1: Junctional features of somatic structural variation in PD4107a

PD4107a	Total number of rearrangements	Rearrangements with microhomology	Rearrangements with non-templated sequence	Rearrangements with no junctional features
Whole genome	68	41	7	20
Kataegis only	18	5	1	12

Table 5.2: Junctional features of somatic structural variation in PD4103a

PD4103a	Total number of rearrangements	Rearrangements with microhomology	Rearrangements with non-templated sequence	Rearrangements with no junctional features
Whole genome	29	19	6	4
Kataegis only	188	105	32	51

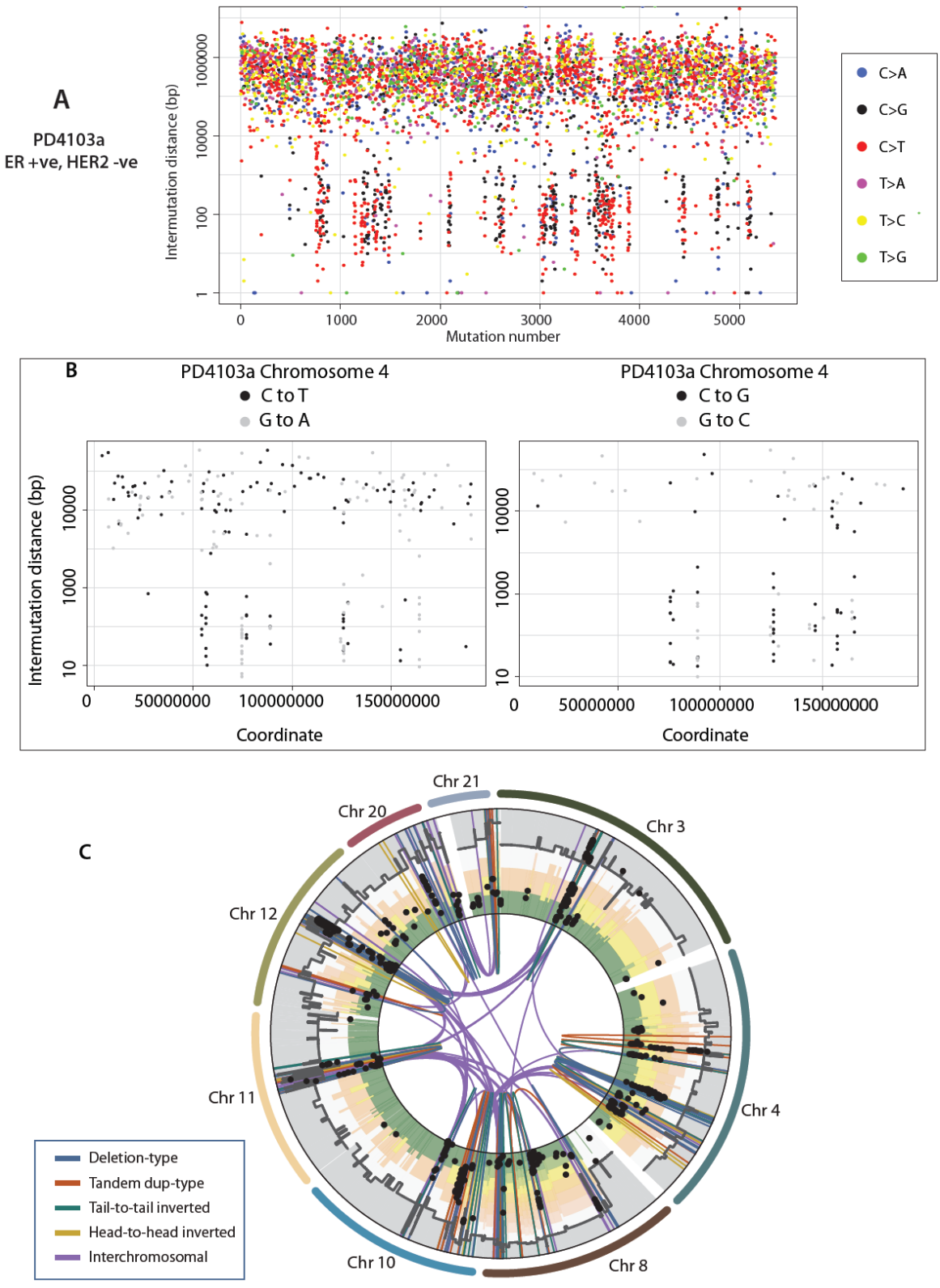


Figure 5.4: (A) Rainfall plot for PD4103a demonstrating kataegis occurring at multiple loci through the genome. (B) Stretches of C>T alternate with stretches of G>A on chromosome 4 in PD4103a. Alternating C>G and G>C mutation on the same chromosome in PD4103a. (C) The complex web of rearrangements involving 8 chromosomes in PD4103a co-localizing with kataegis. The variant allele fraction (y-axis) is represented by the coloured bars: proportion of reads derived from contaminating normal cells (grey bars) and the fraction coming from each of the copies of that segment in the tumour cells (the multiple bars from green to yellow to pink to white).

5.4 KATAEGIS IS COMMON IN THIS COHORT OF TWENTY-ONE BREAST CANCERS

In order to explore the prevalence of kataegis in this cohort of breast cancers, inspection of rainfall plots from all twenty-one breast cancers revealed variable degrees of kataegis in thirteen cases (61.9%) (PD4199a, PD4192a, PD4198a, PD4248a, PD4109a, PD4116a, PD3904a, PD3945a, PD4006a, PD4103a, and PD4107a, see Figure 5.5) encompassing all histopathological subclasses of the disease.

Regions of kataegis were defined as stretches of DNA where each of 6 consecutive mutations occurred no more than 1kb apart from its preceding neighbour mutation in the reference genome. Using this definition, 247 such stretches of kataegis were defined across the 21 genomes (Appendix 4). When ranked by the number of substitution variants involved in each stretch of kataegis, PD4107a was the most dramatic carrying eight of the most hypermutated stretches.

Table 5.3: Regions of kataegis involving the highest number of variants

Breast cancer sample	Chr	Start (coordinate)	End (coordinate)	Size of region (bp)	No of variants	Mutation rate (per kb)
PD4107a	6	126430855	126437625	6770	165	24.37
PD4107a	6	126233148	126235322	2174	48	22.08
PD4107a	6	130487760	130489124	1364	28	20.53
PD4107a	6	130436617	130438324	1707	31	18.16
PD4107a	6	126236516	126239586	3070	50	16.29
PD4005a	10	38201372	38203028	1656	25	15.10
PD4107a	6	130419337	130423519	4182	58	13.87
PD4107a	6	126274096	126277438	3342	46	13.76
PD4120a	10	37736174	37739617	3443	42	12.20
PD4107a	6	132599455	132603528	4073	47	11.54

In each case of kataegis, the features were similar to those outlined for PD4107a and PD4103a. In total, 2738 variants were involved in kataegis or 1.5% of total variants from this study. Of these, 2657 (97.0%) were mutations at cytosine, of which 274 were C>A mutations (10.3%), 770 were C>G

(29.9%) and 1613 were C>T mutations (60.7%). Of these cytosine mutations, 2388 (89.9%) were at a TpC context.

Overall, 72 rearrangements fell within 50kb of any cluster of substitutions in nine different breast cancers. Of these, 40 showed at least 1 bp microhomology at the rearrangement junction and 13 showed a degree of non-templated sequence, no different to what was observed for the all the rearrangements across all 21 breast cancers in aggregate (p=0.64). In the vast majority of cases, the rearrangements were intrachromosomal. Interchromosomal rearrangements were reported almost entirely by PD4103a bar one interchromosomal rearrangement reported in PD4088a.

5.5 KATAEGIS IS NOT HIGHLY SIGNIFICANTLY ENRICHED WITHIN ANY GENOMIC FEATURES

There were no recurrent regions of kataegis across the 21 breast cancers suggesting that the initiating events for kataegis are stochastic. However, enrichment of kataegis at specific genomic architectures, for example, genic regions, fragile sites and retro-elements, was interrogated. 1200 or 43.8% of variants in these regions of kataegis fell within a gene footprint. 477 variants from within these 247 stretches fell within 30 fragile sites (OR=0.65, CI 0.59-0.71, p=0.001), 136 variants fell within 25 LTRs (OR=0.8, CI 0.6-0.9, p=0.002) and 528 variants fell within 37 LINE elements (OR=0.98, CI 0.9-1.1, p=0.63). It should be noted that it is possibly less likely for kataegis to be found within highly repetitive features, due so systematic difficulties of mapping. Furthermore, mutations that fall within some repeat-based genomic features may be actively excluded by post-processing filters.

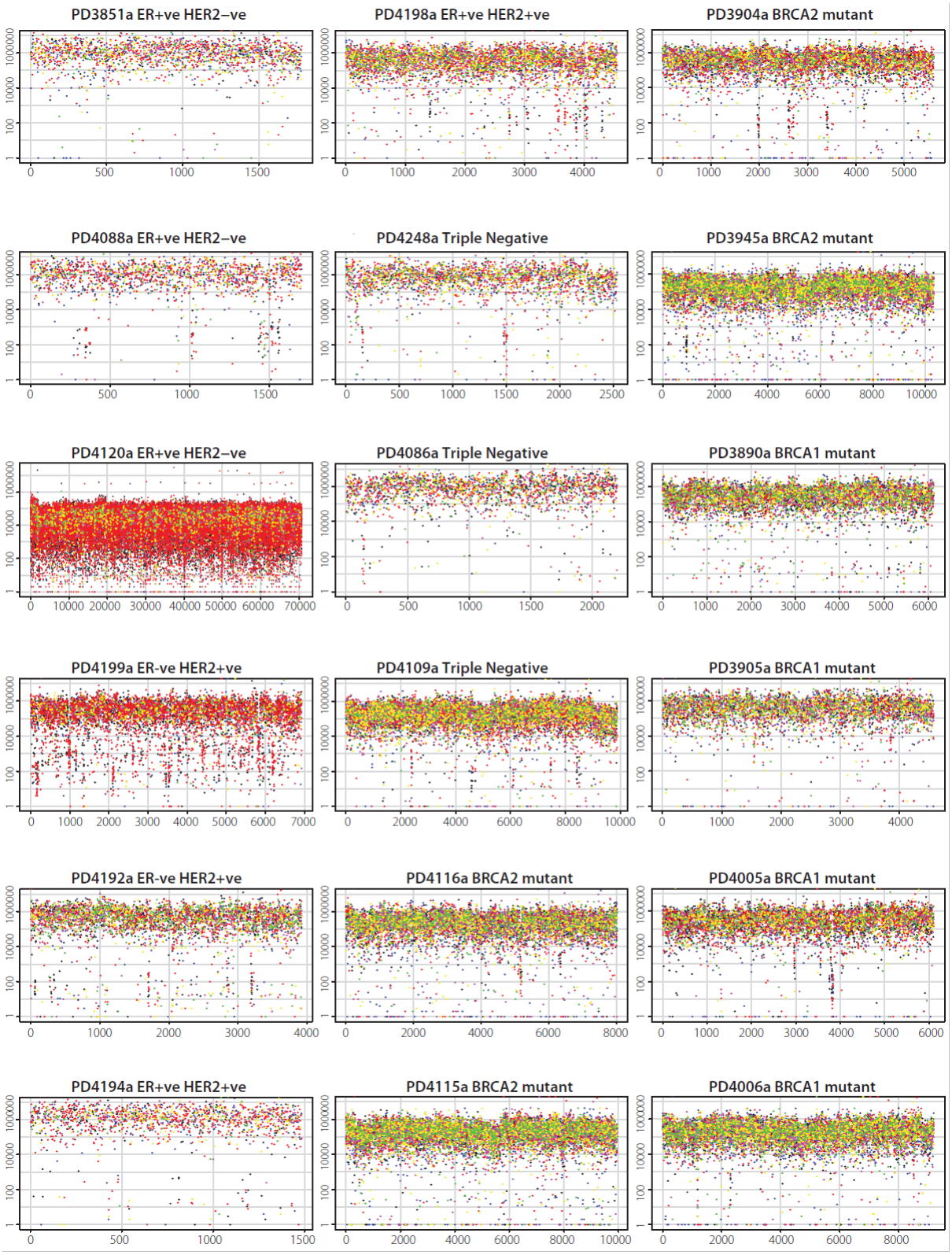


Figure 5.5: Rainfall plots for 18 genomes. Subtle regions of kataegis are present in many samples (PD4198a, PD3904a, PD4088a, PD3945a, PD4120a, PD4086a, PD4199a, PD4109a, PD4192a, PD4116a, PD4005a and PD4006a). Arrows showing some regions of kataegis.

5.6 SUBSTITUTIONS IN KATAEGIS WITHIN A MICROCLUSTER ARE LIKELY TO HAVE OCCURRED CONTEMPORANEOUSLY WHILST SUBSTITUTION IN KATAEGIS BETWEEN DIFFERENT MICROCLUSTERS MAY HAVE ARISEN AT DIFFERENT TIMES.

The localised clusters of C>T and C>G mutations occurring in a TpCpX context, showing a strong bias in strand, and closely associated with genomic rearrangements, suggest that an individual cluster of mutations may have occurred in a single event. Although the mutations within each microcluster might occur simultaneously, however, the relative timing of different clusters of kataegis remains unclear.

By studying the ploidy of kataegis mutations and the associated rearrangements, insight was gained into when they occurred. In PD4103a, there were many clusters of kataegis mutations genome-wide. Interestingly, within the amplicons involving regions of chromosomes 10, 11 and 12, these clusters occurred at several different levels of ploidy (Figure 5.6A). For example, on chromosome 12, there were several such events found at variant allele fraction of 0.8 or higher in association with rearrangements that demarcate large copy number changes. Interestingly, there was also a cluster at an allele fraction of ~0.4 and several at allele fractions <0.1. It is difficult to reconcile how mutations present at different allele fractions could have occurred in one event, although, it seems less likely that two independent hypermutation events occurring during different cell cycles could have produced regional hypermutation in the same genomic location. Rearrangements in PD4103a outside this amplicon were also associated with kataegis demonstrating that kataegis was not restricted to amplicon-generating events (Figure 5.6B). In this latter situation, it was easier to accept that clusters of mutations at different genomic sites were likely to have not all occurred in a single event in this patient.

The other patient with particularly high numbers of these clusters, PD4107a, showed a somewhat different pattern. Here, some of the kataegis substitutions were associated with a tiny chromothripsis event on chromosome 6, and were all at the same level of ploidy. Thus, it seemed very likely that these did occur in the same catastrophic cell cycle.

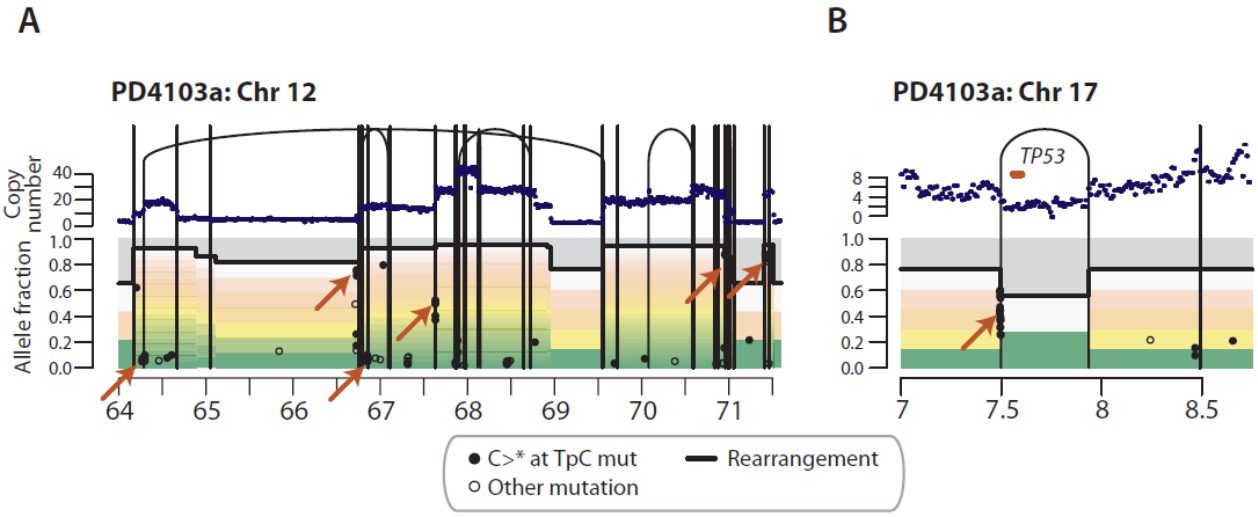


Figure 5.6: Timing kataegis events in PD4103a for the amplicon involving chromosome 12(A) and for a rearrangement resulting in a TP53 deletion (B). The top panel shows the copy number profiles with genomic rearrangements. The lower panel shows the point mutations as filled black circles for C>* mutations in a TpC context (where * is any non-reference base) and open circles for all other types of mutation. The variant allele fraction (y-axis) is represented by the coloured bars: proportion of reads derived from contaminating normal cells (grey bars) and the fraction coming from each of the copies of that segment in the tumour cells (the multiple bars from green to yellow to pink to white).

5.7 SIMILARITIES WITH MUTATIONAL PROCESSES B AND E: GLOBAL AND LOCALISED FORMS OF THE SAME MUTATIONAL PROCESS?

Kataegis is associated with a distinctive substitution mutation signature, the presence of C>T and C>G mutations at TpCpX trinucleotides. These features are similar to those of mutation Process B, and to a lesser extent, Process E described above (Figure 4.3 in chapter 4). Yet, in cancers with evidence of kataegis, mutational Processes B and E make only a small contribution to the overall mutation spectrum across the genome. Conversely, Process B overwhelmingly dominates the overall mutation spectra of PD4120a and PD4199a (Figure 4.1a, 4.1c and 4.1d in chapter 4) despite limited kataegis, and is distributed universally across the genome (Figure 5.5). Intriguingly therefore, a globally distributed and a localised form of these mutation processes may exist and the two forms may operate independently of each other.

5.8 KATAEGIS HAPPENS IN OTHER CANCER TYPES

Kataegis appears to be a relatively common occurrence in breast cancer. In order to evaluate whether this phenomenon was restricted to breast cancers, substitution mutations were sourced from published catalogues of somatic mutation. Kataegis was not seen in a malignant melanoma and a small cell lung cancer (Pleasance et al., 2010a; Pleasance et al., 2010b). Eight recent acute myeloid leukaemia genomes revealed no evidence of kataegis (Ding et al., 2012). An analysis of eight published prostate cancers (Berger et al., 2011) revealed only 1 patch of kataegis in a prostate cancer, PR1701 and this showed a mixed picture with a preponderance of TpC and CpC dinucleotides. Very recently, mutation clusters defined as "2 or more mutations in which all immediate neighbours were separated by no more than 10 kb and has a low p-value for being a clustered mutation" (Roberts et al., 2012) had been identified in an analysis involving multiple myelomas (Chapman et al., 2011), prostate cancers (Berger et al., 2011), and head and neck squamous cell carcinomas (Stransky et al., 2011). This relatively loose definition captured a lot of closely-spaced mutations including double substitutions, as well as possible kataegis.

The phenomenon of substitution hypermutation co-localising with rearrangements has been seen in cancers from other tissue-types indicating that the biological process responsible for generating kataegis is unlikely to be restricted to breast tissue. However, there are subtle and intriguing differences in the sequence context of the substitution variants involved and this may reflect the underlying DNA mutagen involved.

5.9 DISCUSSION

In this chapter, the genome-wide catalogue of substitution mutation information available was used to explore variation in mutation rate throughout cancer genomes. By first devising a metric called the intermutation distance, rainfall plots were constructed to visualise the variation in intermutation distances in each genome. Surprisingly, clusters of substitution hypermutation, termed kataegis, were found in thirteen of twenty-one breast cancers. Substitutions within these clusters exhibited striking characteristics including a predilection for cytosine mutations which were preceded by a thymine base, frequently occurring on the same parental chromosome and demonstrating marked co-localisation with rearrangements. Regional clusters of mutations in cancer have occasionally been observed in experimental models, although not at the mutation density observed here (Wang et al., 2007).

5.9.1 Kataegis in individual microclusters are likely to have arisen within a single cell cycle event

Clustered mutations in cancer could either reflect the net observable result of independent events sequentially acquired over multiple cell cycles or be due to transiently hypermutable conditions permitting the sudden accumulation of multiple mutations in short, sharp bursts.

If mutations have been cumulatively acquired over a number of different cell cycles, then a random distribution of the intermutation spacing would be expected as for mutations arising independently. Conversely, the occurrence of mutations that exhibit close proximity and processivity is more compatible with a model which postulates that these mutations have been generated non-independently in a transient moment of mutability within a single cell cycle event. These transiently permissive states may be mediated by DNA damaging mutagens, error-prone polymerases or imbalances in the nucleotide pool (El-Bayoumy et al., 2000; Weisburger et al., 1998).

The many striking characteristics of the kataegis mutations unearthed in this chapter argue against a step-wise accrual of the substitutions associated with kataegis. The propensity for cytosine mutations at a TpC dinucleotide sequence context, with multiple mutations occurring in *cis*, arising from the same parental strand over extensive genomic distances suggests an active mutagenic propensity for this motif. A hypothetical enzymatic mutagen could either latch onto one strand processively mutating its bearer or could simply have access to one parental strand. Although the mutation spectrum in kataegis may be an attribute of the replicative or translesion polymerase involved in base re-insertion or the composition of the nucleotide pool, this pattern of mutagenesis weights the argument in favour of having arisen within a single cell cycle event, at least within individual microclusters.

5.9.2 Comparison to known mutational signatures suggests that the AID/APOBEC family of enzymes may be the potential enzymatic source for kataegis

On the basis of the substitution hypermutation features described above and the similarities to mutational patterns observed in other biological contexts or in experimental systems, the characteristics of the AID/APOBEC family of cytidine deaminase proteins implicate their activity in the generation of kataegis.

The AID/APOBEC family of cytidine deaminases are characterised by their ability to deaminate cytosines to uracil. Although some knowledge of relatively defined roles have been assigned to

cytidine deaminases, their powerful intrinsic mutagenic potential raises the possibility that deamination of DNA outside the intended target could generate collateral damage, mediating substitution hypermutation of host cellular DNA when unrestricted by customary physiological constraints. Furthermore, hyperedited bases may demand correction via UNG-mediated BER which may lead to the generation of double-strand breaks and structural variation seen in these 21 breast cancer genomes.

5.9.3 Potential mechanisms for co-occurrence of kataegis with rearrangements

Even if the source for these clusters of hypermutations is verified as being due to a member of the APOBEC family, the mechanistic details behind the relationship between the substitution clusters and the rearrangements remains unclear. Kataegis is not always associated with rearrangements. Here, the detection of somatic rearrangements may be limited by the sensitivity of the structural variant calling algorithm or that there has been correct repair of double-strand break.

Likewise, there are rearrangements that do not show kataegis. Here, it may be possible that our substitution detection is limited by mapping characteristics of reads that span the rearrangement breakpoint as well as contain substitutions. Nevertheless, we would expect to see a dearth of mutations just within an insert size of any rearrangement junction, which is not what is observed. Instead, more often than not, there is a lack of substitutions for many kilobases around a rearrangement, before a mutation cluster is seen.

DNA double-strand breaks can arise from breaks induced directly in complementary strands, for example breaks induced by radiation or platinum-based compounds, with the repair of breaks resulting in hypermutation. However, an alternative model is that double strand breaks could be generated by repair of clustered damage, where the repair of these lesions in close proximity on opposing strands results in closely-opposed breaks. At present, it remains unclear which of these is the cause for the regions of kataegis that we see in the 21 breast cancers. Very recently however, data has been published suggesting that APOBECs may have an intriguing role in the repair of double-strand breaks, providing support for the former model (Nowarski et al., 2012). This hypothesis may indeed explain the loco-regional coincidence between rearrangements and substitutions.

5.9.3.1 Closely opposed lesions subjected to base excision repair can generate double-strand breaks

Base excision repair is initiated by specific DNA N-glycosylases that remove damaged bases yielding apurinic/apyrimidinic sites (AP sites). Subsequent incision of the sugar-phosphate backbone by AP endonucleases result in single-strand breaks (SSBs). Efficient SSB repair means that these are not a major threat to genome stability. However, the repair of clustered mutations could result in the formation of two closely-spaced single-strand breaks on opposing strands and might pose a risk for the secondary conversion to a double-strand break.

In a *S. cerevisiae* model, repair of clustered alkylating damage was shown to result in a double-strand break (Ma et al., 2009), in contrast to the observations from non-clustered lesions. Moreover, the delayed generation of double-strand breaks in radiation-induced clustered DNA damage following attempts to fix complex lesions or closely-opposed multifarious single-strand breaks reinforces the model that double-strand breaks can occur whilst attempting to repair closely-opposed single-strand breaks (Greinert et al., 2012).

5.9.3.2 Exposed end-resected single-stranded DNA at double-strand breaks are prone to hypermutation

There is a body of evidence that suggests that single-stranded DNA formed at double-strand breaks or at uncapped telomeres can be hypermutable. The first indication that repair of double-strand breaks can be mutagenic was seen in studies of adaptive mutagenesis in *E. coli* (Lindahl, 1993; Satoh et al., 1993). This was reiterated by yeast studies where site-specific double-strand breaks which were repaired by homologous recombination were associated with a several hundred fold increase in mutation rate (Strathern et al., 1995; Zhu et al., 1998). Hypermutability of long persistent single-stranded DNA in budding yeast was shown to occur during 5'-3' end resection (Yang et al., 2008). However, a very high mutation rate was achieved only when resection and repair was coincided with damage in the form ultraviolet radiation or MMS. The resulting strand bias and mutation spectrum led the authors to speculate that the mutations were caused on single-stranded DNA and were reliant on translesion polymerase repair of polymerase ζ. In this experimental setting, the authors observed a large number of widely-spaced mutation (6 in 4kb ORF) which is not as hypermutated as the stretches of C>T mutations observed in the breast cancers. Nevertheless, these studies marked the first observation of damage-induced localised hypermutation in transient single-stranded DNA circumstances.

5.9.3.3 A comparison of two potential models

Either of the above propositions could have generated the observations made in these 21 breast cancers. In the latter model, substitution hypermutation precedes the double-strand break which would occur as a result of BER-dependent repair. Additionally, the clusters of hypermutation would need to occur on opposing strands and be sufficiently close to each other to allow secondary conversion to a double-strand break.

In the former model, exposed single-stranded DNA would need to persist following end-resection and APOBECs will need to be recruited to these sites. Although the machinery for end-resection in mammalian cells is highly conserved and available to use, under normal physiological conditions the *capacity* for end-resection is likely to be limited. Repair mechanisms such as non-homologous end joining (NHEJ), potentially limits end-resection in mammalian cells, as it efficiently eliminates the substrate for end-resection by its actions mediating the effective joining of blunt double-stranded ends. Furthermore, other inhibitory proteins that may limit end-resection are likely to exist. For example, TP53 binding protein 1, 53BP1, is believed to limit the end-resection of long tracts in murine BRCA1 null models (Kadyrov et al., 2006) effectively encouraging error-prone repair via NHEJ instead of conservative homologous recombination. In this respect, highly rearranged regions with evidence of NHEJ, argues against a model where long end-resected tracts may have become exposed for transient hypermutability.

CHAPTER SIX: COMPLEX RELATIONSHIPS BETWEEN SUBSTITUTIONS AND TRANSCRIPTION

6.1 INTRODUCTION

Transcriptional strand bias has been described in reporter gene assays and more recently, in cancer genome sequences, and is believed to reflect the activity of nucleotide excision repair (NER). NER is a non-specific repair process activated by sensing bulky DNA distortion caused by mutagenic biochemical modifications (Nouspikel, 2009). Across the genome, DNA distortion is sensed by the XPC protein complex, which results in the opening of a denaturation bubble via the TFIIH complex. The damaged strand is incised at both the 5' and 3' ends resulting in an oligonucleotide gap which is filled in by DNA polymerase δ or DNA polymerase ε, and the nick is sealed by a DNA ligase. A particular class of NER exists that is coupled to transcription, called transcription-coupled repair (TCR). DNA lesion sensing is likely to involve stalling of RNA polymerase II (RNAPII) but otherwise repair proceeds in the same way as described for global NER (Figure 1.6 chapter 1). 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.

For example, DNA damage induced by short-wavelength ultraviolet light (UVB 290-320 nm and UVC < 290nm) can cause the formation of covalent modifications between two adjacent pyrimidines on the same DNA strand resulting in cyclobutane pyrimidine dimers (CPDs). These DNA distorting chemical modifications are the ideal substrate for TCR. This has been documented via reporter assays in a variety of model systems, such as mouse models (Vreeswijk et al., 1998) and more comprehensively, in a malignant melanoma cell line COLO-829. Here, a comparison of the dominant C>T/G>A transition mutation in transcribed genomic regions uncovered a strand bias, with fewer 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). It was believed that this was consistent with the past operation of transcription-coupled repair on ultraviolet-light-induced DNA damage in COLO-829. Similar analyses have been extensively documented for the by-products of tobacco-smoke. Adducts from B[a]DPE have been mapped to preferential codons in TP53 reporter assays of human bronchial epithelial cells and these have been shown to coincide with C>A/G>T transversion mutations (Denissenko et al., 1996; Pfeifer, 2000; Pfeifer et al., 2002). A strong transcriptional bias with fewer transversions on the transcribed strand has been attributed to the past activity of TCR (Hainaut and Pfeifer, 2001). Whole genome analysis of a lung cancer cell line, NCI-H209, has reiterated this verdict (Pleasance et al., 2010b).

These studies revealed how detailed analyses of comprehensive catalogues of somatic mutations in cancer could help to uncover clues regarding specific repair processes that have been operative. In this chapter, asymmetries in substitution mutation prevalence are sought, integrating the analysis with transcriptomic data where appropriate. Variation is explored between genes, between transcriptional strands in each gene and along the length of each gene, in order to gain further insight into the mutagenic exposure and repair pathways that have fashioned these twenty-one breast cancer genomes.

6.2 DEFINING STRAND BIAS OF SUBSTITUTIONS IN CANCER GENOMES

Base substitutions that fall within a genomic footprint, corresponding to ~40% of the human genome, can be classified according to transcriptional strand (Figure 6.1). The six mutation-types of substitutions can therefore be further sub-classified according to whether they are on the transcribed or non-transcribed strand.

Figure 6.1

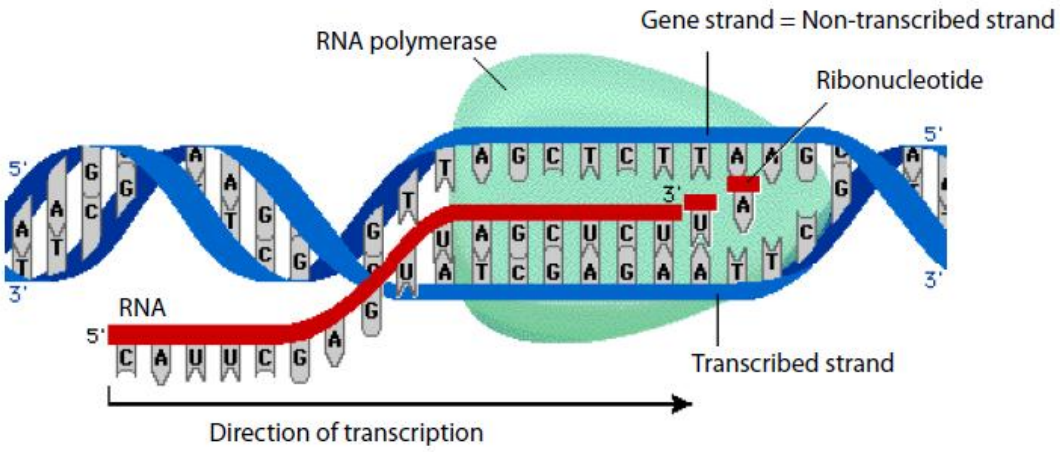


Figure 6.1: Transcriptional strands. The nucleotide sequence of transcribed RNA is identical to the sense/non-template/non-transcribed strand, except that U replaces T, and is complementary to that of the anti-sense/template/transcribed strand.

6.3 GENE EXPRESSION DATA OF THE BREAST CANCER GENOMES

Gene expression data were derived from the Illumina Human HT12 Expression BeadChip array, run in duplicate, with all seventeen samples batched together and normalised. Overall, gene expression data was available for 14,721 genes, with the genomic footprint of these genes encompassing 867,657,063 bases, corresponding to approximately 27.7% of the genome.

Standard hierarchical clustering based on expression array data showed that the seventeen samples for which expression data is available clustered according to histopathological status as would be expected (Figure 6.2).

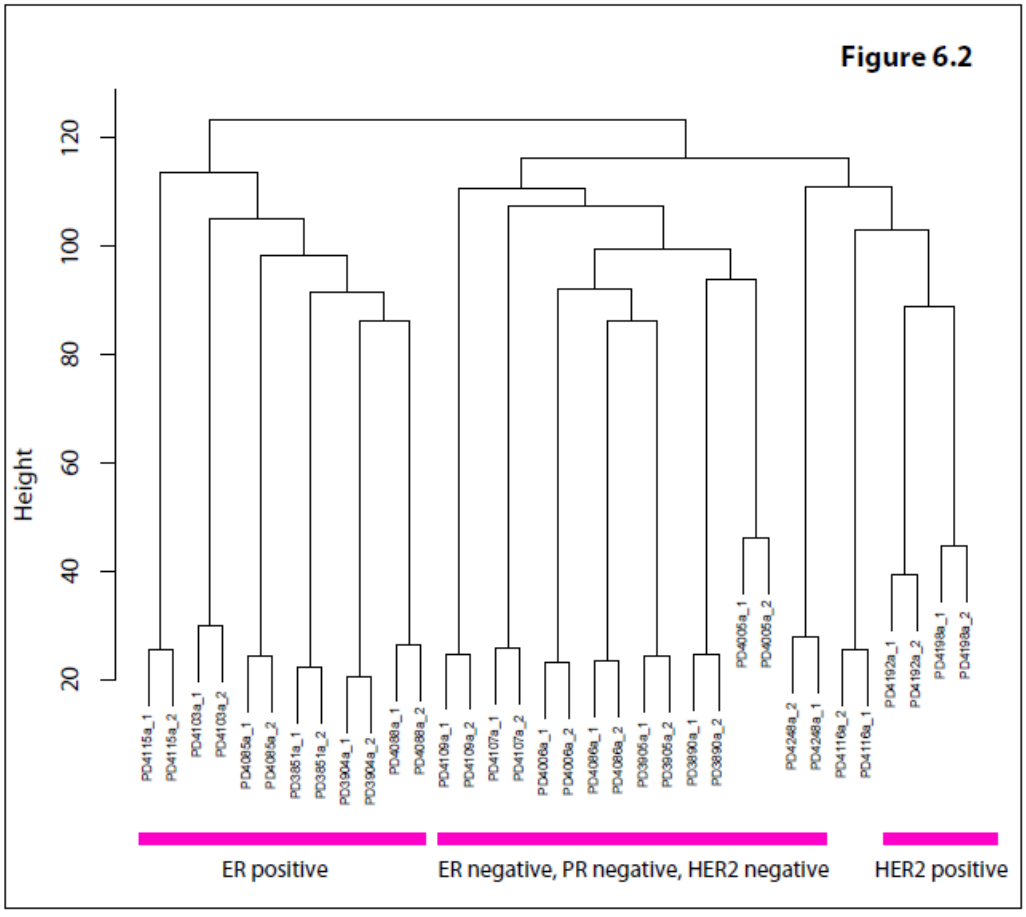


Figure 6.2: Cluster dendrogram of transcriptomic profile of seventeen breast cancer genomes.

In order to examine correlations between mutation prevalence and gene expression as well as to consider transcriptional strand biases, a Poisson regression or generalised linear mixed effects model was used for the analysis (as described in the Experimental Procedures, section 2.6.2). The overall fitted curve for each mutation-type represented the combined effects across all seventeen cases for which there was expression data. The relationships described in the following sections were based on this analysis.

6.3.1 Difference in the prevalence of mutations on the transcribed and non-transcribed strands

The relationship between transcriptional strand and the prevalence of somatic substitutions was examined first. The differences in the prevalence of mutations on the transcribed versus non-transcribed strands (transcriptional strand bias) across all protein coding genes in aggregate were sought. For a given level of gene expression, the effects of transcription-coupled repair (TCR) are revealed by the significant separation of curves for mutations on the transcribed and non-transcribed strands.

A moderate degree of transcriptional strand bias was detectable for C>A/G>T transitions across the 21 breast cancer genomes ($p=1.75 \times 10^{-15}$) and appeared to be present in almost all cases (Figure 6.3). This bias was characterised by fewer C>A mutations on the non-transcribed strand than the transcribed strand.

A strand bias was also observed for T>G/A>C mutations ($p=1.5 \times 10^{-4}$) with fewer T>G mutations on transcribed than non-transcribed strands. No evidence of a transcriptional strand bias was observed for C>G/G>C, C>T/G>A, T>A/A>T or T>C/A>G mutations.

The most widely acknowledged cause of transcriptional strand bias is TCR of nucleotide excision repair (NER) which is believed to remove nucleotides with bulky adducts from the transcribed strands of genes. Assuming that TCR was responsible for the observed strand biases, the presence of fewer C>A mutations on non-transcribed than transcribed strands would suggest that bulky adduct damage to guanine may be the cause of the observed mutations. Similarly, the presence of fewer T>G mutations on transcribed compared to non-transcribed strands would suggest that there may have been bulky adduct damage to thymine. The nature of these ubiquitous mutagenic exposures in breast cancer, and whether they are exogenous or endogenous in origin, is unknown. However, the assumption that TCR is involved is not necessarily correct and it may ultimately transpire that other DNA repair processes, or indeed DNA damage mechanisms, may differentially affect the transcribed and non-transcribed strands of genes.

6.3.2 The relationship between levels of gene expression and the prevalence of somatic mutation

It is known from studies in diverse biological lineages ranging from bacteria (Gouy and Gautier, 1982) to *Caenorhabditis elegans* and *Drosophila* (Duret and Mouchiroud, 1999), that the rate of substitution-related evolutionary change of a gene is often related to its level of expression. Relative substitution rates across genes varied widely but essentially showed a strong negative correlation with the level of gene expression across these biological extractions. Therefore, the relationship between levels of gene expression and the prevalence of somatic mutation was investigated as in a previously studied malignant melanoma and small cell lung cancer respectively (Pleasance et al., 2010a; Pleasance et al., 2010b), the authors demonstrated that levels of gene expression correlated inversely with mutation prevalence. This phenomenon was observed on both the transcribed and non-transcribed strands of genes indicating that it was independent of transcriptional strand. The mechanism underlying this phenomenon is not well-explored.

In the seventeen breast cancers for which gene expression data was available, an inverse correlation of substitution prevalence with gene was observed for C>A/G>T ($p=2.47 \times 10^{-9}$), C>T/G>A ($p=7.5 \times 10^{-3}$), T>A/A>T ($p=1.09 \times 10^{-6}$) and T>C/A>G ($p=1.83 \times 10^{-4}$) mutations for both transcribed and non-transcribed strands (Figure 6.3). This finding reinforces the observation made in a single malignant melanoma and small-cell lung cancer, but extends it to multiple cancer samples of a different tissue-type. No correlation was observed for C>G/G>C or T>G/A>C mutations.

There could be two reasons for this observation. First, for four out of the six classes of mutations, an alternative repair pathway which is related to the degree of expression but that operates on both strands and is at least as numerically important as TCR appears to be at play. Thus, significantly lower mutation prevalence, on both transcribed and non-transcribed strands, was observed in more highly expressed genes. The alternative argument would be that highly expressed genes are under enhanced selective constraint with purifying selection modulating and restricting mutagenic accumulation in highly expressed genes. However, a large proportion of mutations in the genomic footprint are within introns and it remains unclear why purifying selection would be acting on mutations in these regions.

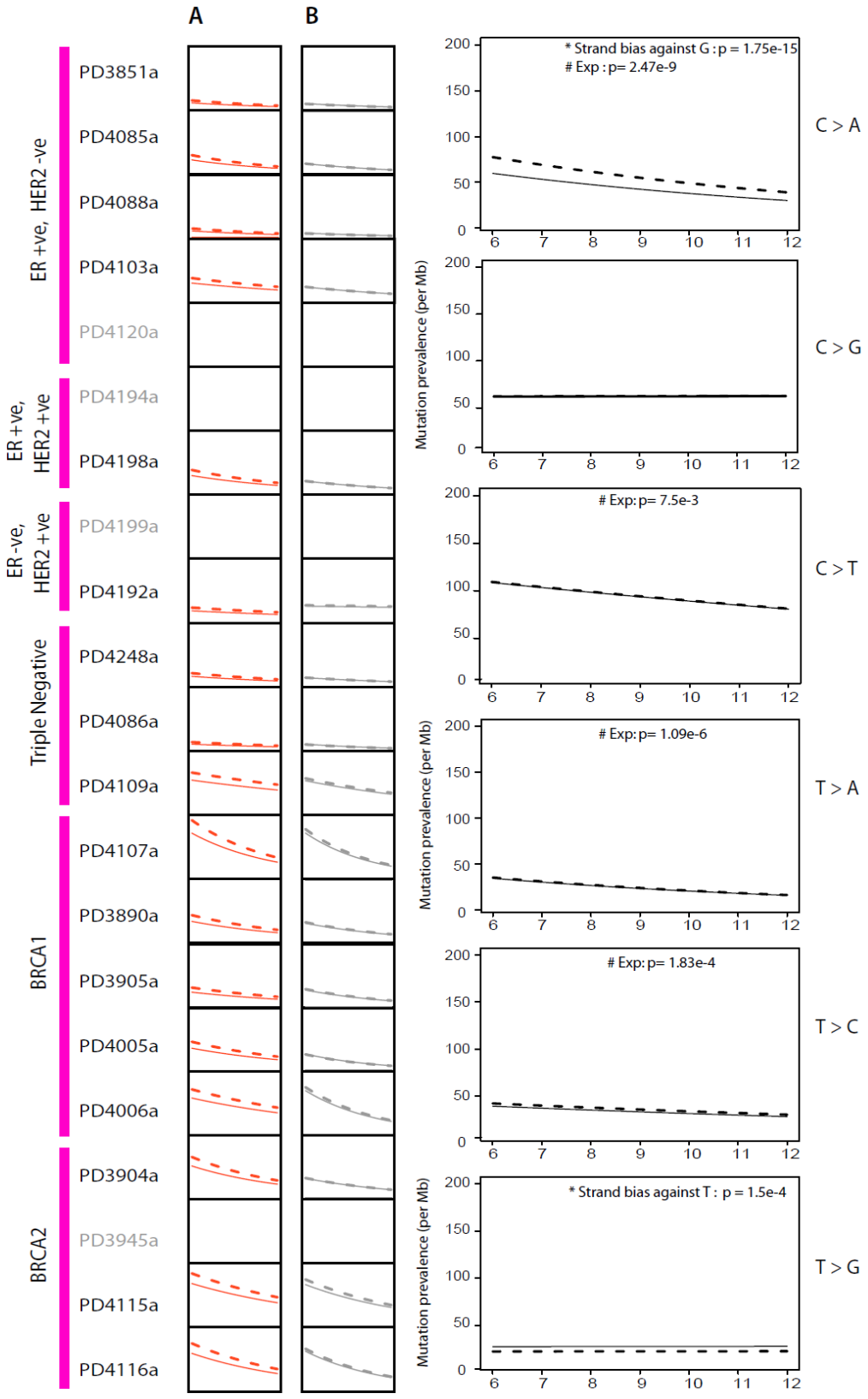


Figure 6.3: The relationship between mutation prevalence and transcription and/or expression. Mutation prevalence is expressed as the number of mutations per Mb from 0 to 2 per Mb on the vertical axis. Log 2 expression levels range from 6 to 12 on the horizontal axis. Lines are fitted curves to the data for A and B. (A) C > A mutations for each cancer genome; (B) T > A mutations for each cancer genome. Breast cancer samples without expression data are shown in gray. (C) Overall effect of transcription and gene expression on mutation prevalence by mutation type. P-values of significance are provided for each mutation-type if a strong effect was seen in either strand bias and/or relationship with expression.

6.4 THE MUTATION PREVALENCE INCREASES AT INCREASING DISTANCE FROM THE TRANSCRIPTION START SITE

Previously, a sharp decline in the mutations at methylated CpG dinucleotides was observed in germline cells in the vicinity of the 5' end of genes (Polak and Arndt, 2008). More recently, an inverse relationship between distance from transcription start site and mutation prevalence was demonstrated in a malignant melanoma cancer genome (Pleasance et al., 2010a). Therefore, the relationship between distance from the transcriptional start site and mutation prevalence in protein coding genes was next examined. Transcription start site coordinates and the genomic footprint of all genes were obtained from the Ensembl v58 API. 1 kb bins beginning from each transcription start site were defined, marching along the length of all genes. The number of genes which completely encompassed each 1kb bin was scored for each bin. The number of mutations present in each bin was also counted. The fraction of genes mutated in each bin is presented in Figure 6.4.

There was evidence of increasing mutation prevalence at increasing distance from the transcription start site (Figure 6.4a), suggesting that the influences of transcription upon mutagenesis described above wane as proximity to the transcription start site decreases. The result confirms the observation previously made on ultraviolet light induced C>T/G>A mutations in a melanoma cell line (Pleasance et al., 2010a), extending it to many more cancer samples of different classes and across many different mutation types.

The effect appears to be particularly pronounced in the first 1kb from the transcription start site (Figure 6.4b). In germline cells, a localised strand asymmetry showing an excess of C>T over G>A substitutions in the non-transcribed strand was confined to the first 1-2 kb downstream of the 5' end of genes (Polak and Arndt, 2008). The authors hypothesised that the exposed non-transcribed strand near the 5' end of genes was more susceptible to cytosine deamination of methylated CpG dinucleotides. To investigate if this was a feature of somatically acquired mutations, strand bias was interrogated in bins of 1kb from the transcription start site and then increasingly larger bins thereafter for all the C>X mutation-types, because the effect was believed to be prominent closer to the transcription start site. Here, all 21 genomes were included in the analysis. Although initially no significant difference between the transcriptional strands were seen, when C>T mutations were

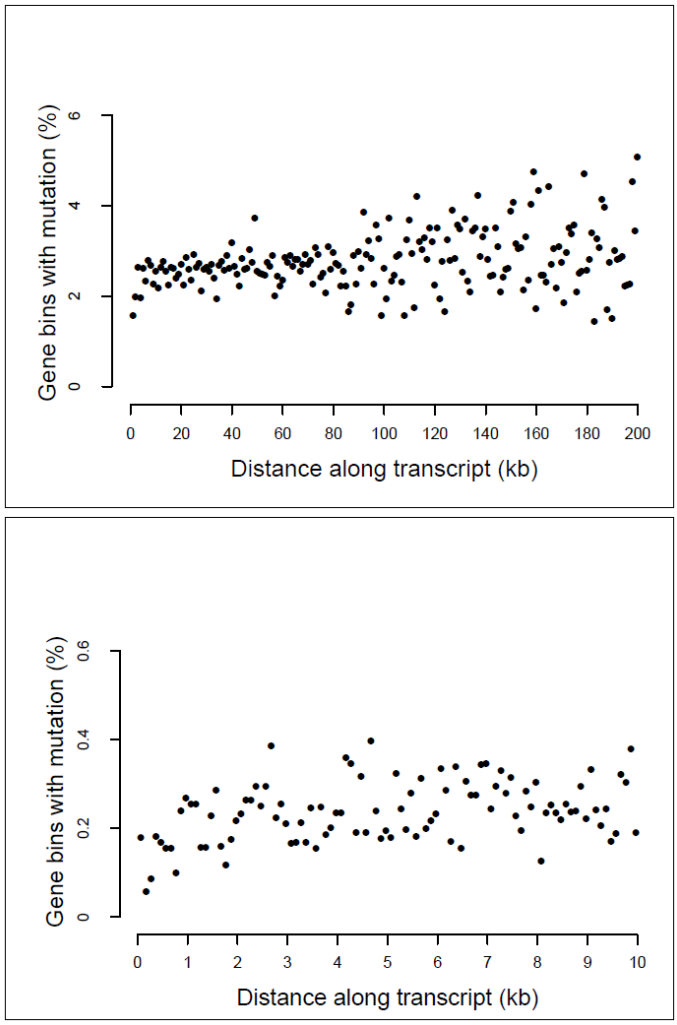


Figure 6.4: The effect of distance from transcription start site on mutation prevalence. (A) Each dot represents a 1kb bin at increasing distances from all transcription start sites (TSS) up to 200kb. The y axis shows the percentage of genes in each bin carrying a somatic mutation. The mutation prevalence increases as distance increases from the TSS. (B) This is particularly marked in the first 1kb after the TSS. Each dot represents a 100bp bin.

separated by whether they occurred at CpG dinucleotides or otherwise, an overall strand asymmetry was seen for the C>T mutations at CpGs, with more mutations seen in the non-transcribed strand. Unlike what was previously documented in the germline, this was not confined to the first 1-2kb (Figure 6.5). When C>Ts at CpGs were treated in isolation, the asymmetry extended to approximately 10kb.

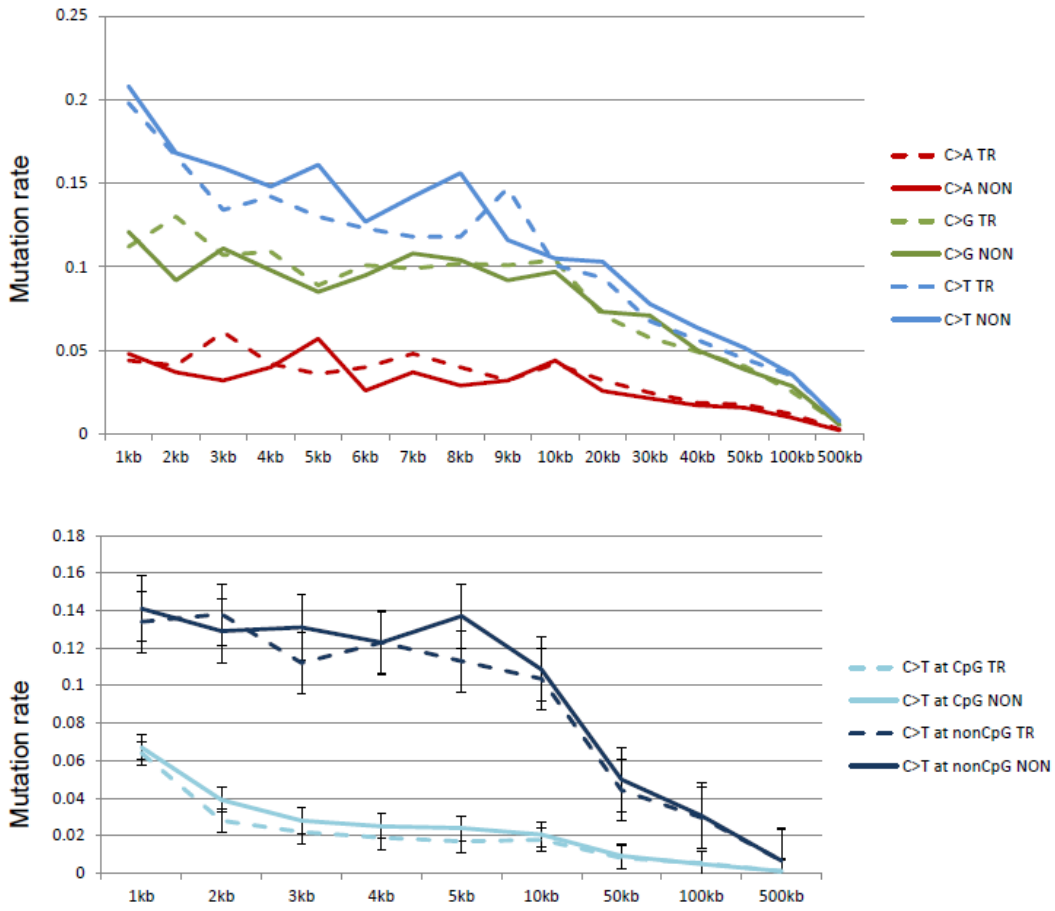


Figure 6.5: Mutation rate with distance from transcription start site, not corrected for the total number of genes involved. Note the strand bias against C>T mutations on the transcribed strand which is not limited to first 1-2kb from the transcription start site, although it is not statistically significant for each genomic bin. The horizontal axis describes the bins of genomic distance from the transcription start site. The vertical axis describes the mutation rate per Mb.

6.5 THE MUTATION PREVALENCE IS HIGHER IN INTRONS THAN IN EXONS

Previously, a reduction in mutation prevalence in exons compared to introns was reported in a single cancer genome and the possibility of the action of negative selection on mutations in the coding sequence was raised, given that preferential targeting of nucleotide excision repair has not been reported towards exons specifically. By comparing the rate of mutation in various parts of the genome, a reduction in mutation prevalence in exons (2.64 per Mb) compared to introns (2.90 per MB) was observed in the twenty-one breast cancers ($p = 1.79X10^{-5}$).

6.6 DISCUSSION

The results from these 21 genomes have yielded further insights into the complex relationship between mutagenesis and transcription in breast cancers. First, transcriptional strand bias was documented with evidence of fewer mutations on the transcribed strand of G>T mutations and for T>G mutations. Second, a marked inverse correlation between gene expression levels and mutation prevalence was seen for C>A/G>T, C>T/G>A, T>A/A>T and T>C/A>G mutations for both transcribed and non-transcribed strands. Thirdly, increasing mutation prevalence was seen with increasing distance from transcription start site. Finally, mutation prevalence was found to be higher in introns than in exons.

6.6.1 Transcriptional strand bias invokes transcription-coupled repair being operative

A transcriptional strand bias was found for C>A/G>T mutations and also for T>G/A>C mutations in most of the cancers but not for the remaining classes of mutations. This suggests that the extent of TCR differs for the various classes of mutation, possibly reflecting differences in the ability of the TCR machinery to recognize and/or repair different adduct lesions.

If TCR is responsible for these strand biases, DNA damage through covalent binding of bulky adducts may be implicated in breast cancer pathogenesis. The exposures generating such covalent modifications could, in principle, be exogenous, and indeed many carcinogens are known to cause adducts on guanine, for example the by-products of tobacco smoke. Alternatively, the exposure could be endogenous in origin, for example due to reactive oxygen species (Hori et al., 2011). Oxidised bases are, however, generally thought to be repaired by base excision repair (BER).

6.6.1.2 Potential atypical substrates for transcription-coupled repair in breast cancers

Here, potential atypical substrates for the action of transcription-coupled repair in breast cancers are speculated. A particularly unique class of oxidative DNA lesion generated by hydroxyl radicals called cyclopurines, are characterised by a covalent bond between the purine and the sugar moiety of the sugar-phosphate backbone making them troublesome for base excision repair (BER) and ideal candidates for nucleotide excision repair (NER) (Bishop and Bell, 1985; Simon et al., 1985). Furthermore, lipid peroxidation has also been known to yield a highly reactive product, malondialdehyde, which can form bulky DNA adducts on guanine (Katzen et al., 1985) again challenging the effectiveness of base excision repair, but posing the perfect substrate for nucleotide excision repair. In fact, malondialdehyde adducts in the transcribed strand of expressed genes were shown to be strong blocks to RNA polymerase II (RNAPII) and are targets for cellular transcription-

coupled repair (TCR) (Shih et al., 1981). Furthermore, viral DNA site-specifically adducted with a malondialdehyde-analogue, exocyclic adduct propanodeoxyguanosine (PdG) and incorporated into NER-deficient and proficient strains demonstrated a 4-fold increase in the frequencies of transversions and transitions in E. coli strains deficient in NER (Johnson et al 1987) (Bishop, 1985).

Oestrogens can cause damage to DNA by generating electrophilic species that can covalently bind to DNA. This is thought to proceed through catechol oestrogen metabolites, which can be oxidised to intermediates that bind to DNA. Therefore, oestrogen could generate DNA lesions particularly to guanines which may become substrates for nucleotide excision repair. First, stable oestrogen adducts which constitute ideal candidates for nucleotide excision repair can be formed through 2,3-quinone oxidative species (Shih et al., 1981). Second, both endogenous and synthetic oestrogens have been shown to induce oxidative DNA damage in addition to specific DNA adducts (Spencer et al., 2011) (Spencer et al., 2012). However, it remains unclear whether such DNA damage would be repaired by nucleotide excision repair or base excision repair. Notwithstanding, there is ample epidemiological evidence linking this endogenously operating and widely exogenously administered mutagen to breast cancer and it should not be overlooked as a potential source for DNA damage.

Commonly occurring depurinating and base deamination events, which are not the usual substrates for transcription-coupled repair, could potentially affect the progression of the transcription complex, thereby providing an opportunity for transcription-related repair. Abasic sites on the transcribed strand were found to block transcription by mammalian RNA polymerase II (RNAPII) in in vitro transcription assays with site-specific lesions whilst not causing any such block when the abasic site was in the non-transcribed strand (Hanawalt and Spivak, 2008). The prevailing dogma is that lesions that block RNAPII will be subject to transcription-coupled repair, and these findings would suggest that an abasic site could be sufficient to initiate transcription-coupled repair (Tornaletti et al., 2006; Wang et al., 2006). This implies that bulky adduct formation is not *always* necessary to stall RNAPII and initiate transcription-coupled repair, at least in vitro.

6.6.1.3 Other forms of transcription-related DNA repair may be operative in breast cancers

Although TCR is the most widely acknowledged transcription-dependent repair pathway, it is possible that a version of base excision repair may exist which is itself coupled to transcription. Alternatively, RNAPII stalling via atypical substrates could initiate other forms of transcription-related repair. If transcription-coupled repair is not involved, the data would suggest that there exist other, currently uncharacterised, forms of transcription-related DNA repair pathways.

6.6.1.4 Strand bias could be due to transcription-related DNA *damage*, not repair

The underlying assumption that mutagenic damage is equivalent in both transcriptional strands and that strand asymmetry arises from partiality of repair to the transcribed strand may of course be incorrect. At the present time, it is hard to identify an example of a mutagen which shows strand bias but the possibility that a mutagen preferentially targets one of two strands cannot be dismissed.

6.6.2 Evidence for an alternative repair pathway associated with levels of gene expression

An inverse relationship between levels of gene expression and mutation prevalence was previously reported in a malignant melanoma and a small cell lung cancer cell line (Pleasance et al., 2010a; Pleasance et al., 2010b). This finding has been extended in this study for some mutation types to include seventeen primary breast cancers. The relationship again seems to be inverse in nature, with more somatic substitutions accumulating in poorly expressed genes. This expression-related phenomenon is independent of transcriptional strand as both strands appear to be similarly affected (Figure 6.3). T>G/A>C mutations exhibited a transcriptional strand bias but no correlation between expression and mutation prevalence. Conversely, C>T/G>A, T>A/A>T, and T>C/A>G mutations showed correlations between gene expression and mutation prevalence but no strand bias (Figure 6.3).

This relationship could be due to less efficient repair in poorly expressed genes. However, the proficient repair of the non-transcribed strand cannot be attributed to transcription-coupled repair which targets the transcribed strand of genes. One possibility is that the genome-wide form of nucleotide excision repair is recruited more effectively to highly transcribed genes, perhaps as a result of differing chromatin configuration.

There have been hints in the past of a type of global nucleotide excision repair called transcription domain-associated repair (DAR) which may account for efficient repair of both strands in expressed

genes. Whilst DAR depends upon transcription it does not depend upon RNA polymerase II stalling due to a lesion. In a series of strand-specific repair studies on HL60 and THP1 cells, repair of both strands continued to occur in parts of the gene that the polymerase never reached and continued despite blocking RNAPII activity with RNAPII inhibitors like α–amanitin (Nouspikel et al., 2006). Furthermore, using siRNA experiments, DAR has been shown to be dependent on XPC, a protein central to global genome repair and not essential for transcription-coupled repair. Conversely, transcription domain associated repair appeared to be independent of transcription coupled repair-specific proteins, CSA and CSB (Barnes et al., 1993). DAR may therefore be a subset of global nucleotide excision repair, perhaps restricted to certain genomic regions by chromatin configuration. It was proposed that genomic domains within which transcription is active are more accessible than the bulk of the genome to the recognition and repair of lesions through the global pathway (Barnes et al., 1993). Since then however, relatively little work has been seen in transcription domain associated repair. The finding here of an inverse relationship between gene expression level and mutation prevalence, acting on both transcribed and non-transcribed strands lends some support to the presence of a repair phenomenon related to expression which is independent of and different to transcription-coupled repair, and could be evidence supporting transcription domain associated repair.

These data also show a trend towards a higher prevalence of somatic substitutions at the 3′ compared to the 5′ ends of genes. This may be due to aborted transcription, such that 3′ ends are overall less transcribed than 5′ ends, with the consequence that expression-related repair processes are deployed less at 3′ ends and hence the mutation prevalence is higher.

CHAPTER SEVEN: MUTATIONAL PROCESSES REVEALED BY OTHER MUTATION CLASSES IN TWENTY-ONE BREAST CANCER GENOMES

7.1 INTRODUCTION

In the preceding chapters, the somatic single-nucleotide substitution catalogues of twenty-one breast cancers were explored in order to identify mutational signatures that have shaped the cancer genomes. However, analyses of other mutational classes can reveal underlying biological processes that have been operative in these twenty-one breast cancers.

In this chapter, further mutational signatures generating insertions and deletions, double substitutions and rearrangements will be sought. Putative cancer genes within this catalogue of somatic mutations of 21 breast cancers will also be highlighted, to complete the portraits of twenty-one breast cancer genomes.

7.2 INSERTIONS AND DELETIONS

Insertions and deletions of nucleotides in DNA, are collectively termed 'indels', and constitute common and biologically significant mutations with relevance to human disease. The biological consequence is often deleterious as an indel involving a number of bases that is not a multiple of three results in a shift in reading frame that can abolish the function of a gene. This constitutes a common mechanism of human pathology in both germline and somatic cells (Duval and Hamelin, 2002).

In 1960, shortly after the description of the structure of the DNA double helix (Watson and Crick, 1953b), models of double-helical DNA molecules containing unpaired nucleotides which formed loops were described (Fresco and Alberts, 1960) and posited to be the preliminary step towards indel formation. It was subsequently proposed that frameshift mutations resulted from strand slippage in repetitive DNA sequences, thereby creating misaligned intermediates containing unpaired bases that are eventually added or deleted (Streisinger et al., 1966; Streisinger and Owen, 1985). Furthermore, the moderation of indel formation in this classical model of mutagenesis has been shown to be critically governed by post-replicative DNA mismatch repair (Kunkel and Erie, 2005; Modrich and Lahue, 1996).

The importance of post-replicative mismatch repair as a constraint on the generation of indels during replication is emphasised by studies showing that spontaneous indel error rates in repetitive sequences increased by many orders of magnitude when mismatch repair was inactivated (Greene and Jinks-Robertson, 1997; Tran et al., 1997). Loss of mismatch repair in humans leads to 'microsatellite instability', a phenomenon characterised by variation in repeat length caused by indel errors in repetitive sequences, frequently observed in colorectal carcinomas (Ionov et al., 1993; Thibodeau et al., 1993), but not so far demonstrated to drive breast cancer carcinogenesis.

Here, the landscape of indels across the twenty-one breast cancer genomes will be described in detail. Particular attention will be paid to the junctional features immediately flanking each indel in order to identify mutational signatures which may, for example, expose deficiencies in post-replicative mismatch repair that may constitute a mutational process underlying the generation of indels in breast cancer.

7.2.1 The landscape of indels in twenty-one breast cancers

Overall, 2,869 indels were identified from the twenty-one breast cancer genomes. Of these, 2,233 were deletions, 544 insertions and 92 were complex indels. There were 21 coding indels, of which 15 were predicted to result in a translational frameshift and six were in-frame. All the indels presented have been validated by Sanger sequencing or Roche 454 pyrosequencing.

The frequency of indels did not generally associate with any histopathological subtype and did not demonstrate a clear correlation with total number of substitutions or number of rearrangements in the cancers (Figure 7.1).

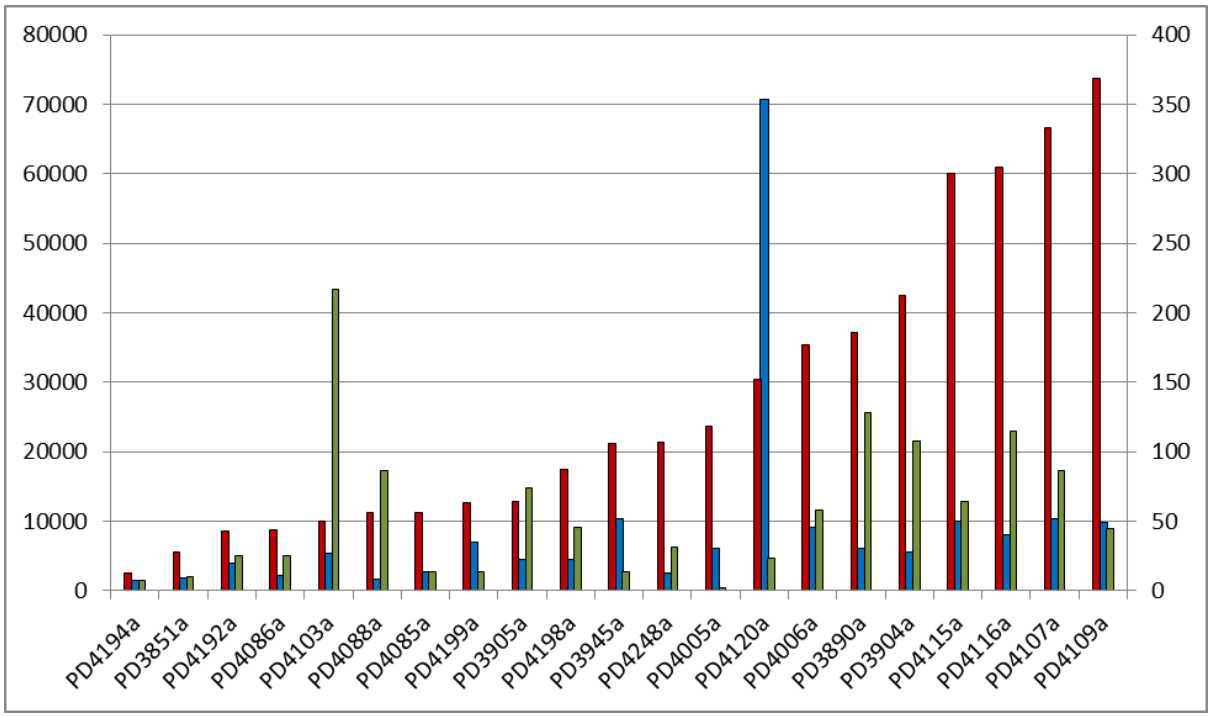


Figure 7.1: Relationship between total number of insertions/deletions and mutation burden of other classes of mutation. Indels (in red) and rearrangements (in green) are scaled to the right-hand vertical axis (total number of indels or rearrangements). Substitutions (in blue) are scaled to the left-hand vertical axis (total number of substitutions).

7.2.2 Breast cancers with defects in homologous recombination show more and larger indels

There was substantial variation in number and pattern of indel between the breast cancers. The cancer with the most number of indels was PD4109a, a triple negative breast cancer with a total of 369 indels and the cancer with the least indels was PD4194a, a lobular ER positive, PR positive and HER2 positive cancer with only 13 indels. Regardless of the wide variation in number of indels (Figure 7.2a), almost all the breast cancers showed more deletions than insertions apart from PD4088a. Furthermore, of the 2,869 validated somatic indels from the 21 breast cancers, single-base pair indels were the most common in each case. The frequency of indel by size, diminished as the size of indel increased in virtually all cases. However, in general, more indels were noted amongst the *BRCA1* and *BRCA2* germline mutant cancers. Furthermore, the distribution of indel by size of indel also demonstrated a long tail of larger-sized indels in the BRCA1 and BRCA2 mutant cancers (Figure 7.2b).

7.2.3 Analysis of flanking sequence reveals differences in processes mediating small and large indels

Given the observed difference between BRCA1/BRCA2 mutant breast cancers and sporadic breast cancers, the sequences flanking each indel were interrogated for the presence of either short tandem repeats or short stretches of identical sequence at the breakpoints (termed overlapping microhomology) (Figure 7.2C). Indels were classified according to whether they were repeat-mediated, microhomology-mediated or neither. Complex indels were excluded from the analysis given the ambiguity in classification.

Repeat-mediated indels were small (1-5bp), present in all breast cancers, and were composed of both deletions and insertions. Microhomology-mediated indels were larger (5 to 50bp), comprised mainly deletions and were considerably more common in breast cancers with mutations in *BRCA1* or *BRCA2.* The distributions of the two-groups were plotted according to indel size and a strong statistical difference was found between the two distributions, using the Kolmogorov-Smirnov test ($p = 2.2X10^{-16}$) (Figure 7.2D).

The distribution of the number of bases involved in microhomology was significantly greater than expected number of bases if microhomology were to have occurred by chance (p<1.2e-8). This signature suggests that the larger indels seen particularly in the BRCA1 and BRCA2 cancers seem to be actively mediated by microhomology-mediated repair processes. Overlapping microhomology is often considered to be a signature of non-homologous end-joining (NHEJ) DNA double-strand break repair. The segments of microhomology are likely to mediate alignment of the two DNA fragments that are joined. Since BRCA1 and BRCA2 are involved in homologous recombination based double strand break repair, the elevated frequency of microhomology-mediated indels in *BRCA1* or *BRCA2* mutant cancers presumably reflects the necessity for alternative methods of double strand break repair in these cancers (Figure 7.2E).

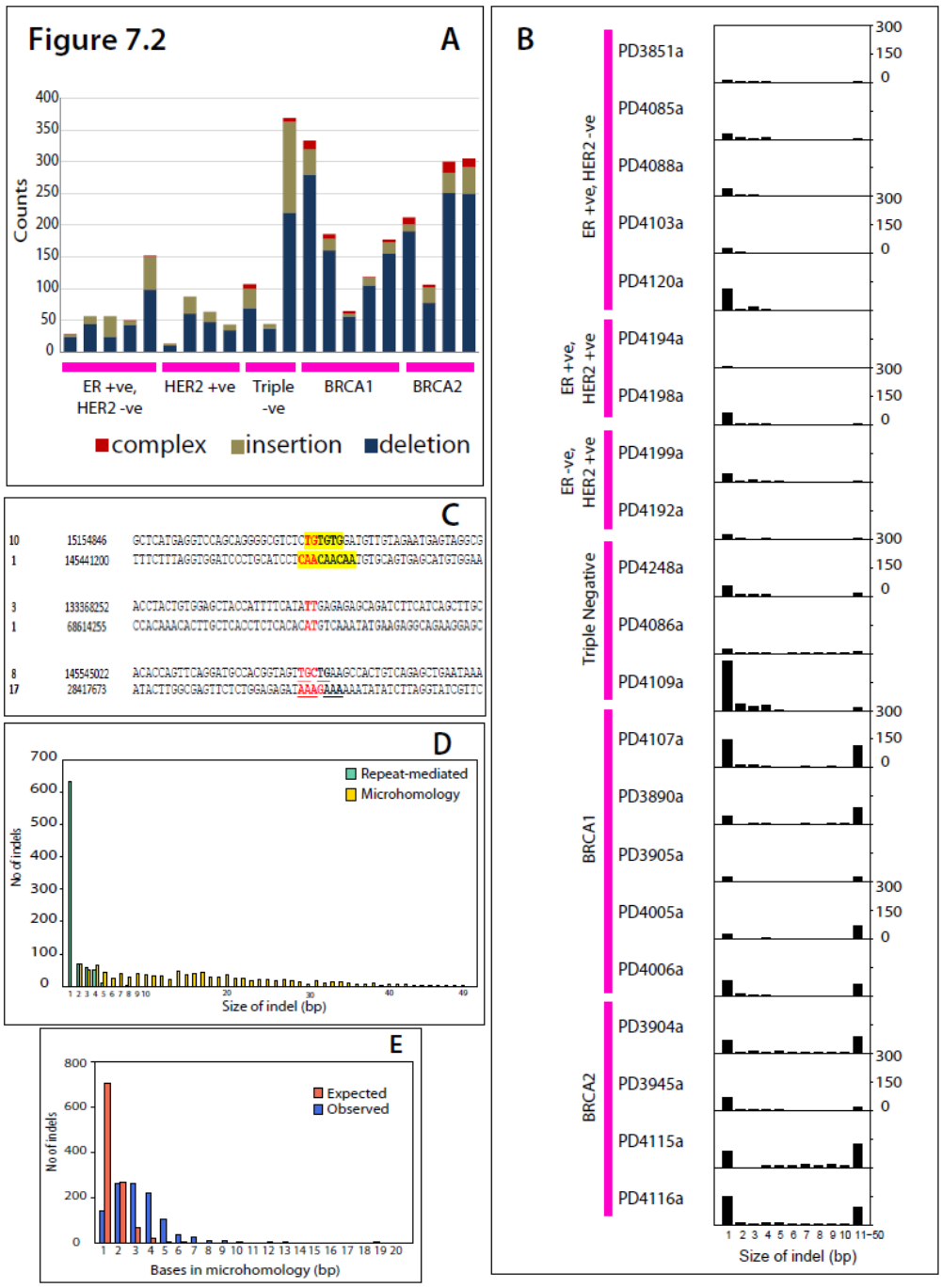


Figure 7.2: Somatic mutation profile of indels. (A) Histogram of number of indels for each breast cancer. (B) The x axis shows indel size from 1-10 and all larger indels between11-50bp in size grouped in a single bin. The y axis shows the number in each genome from 0-300bp. (C) Classification of indel by junctional characteristics. 6 examples of deletions are provided. The motif of the deletion is highlighted in red. In the first two examples, the deletion bears the same motif as the immediate tandemly repeating units and is classed as repeat-mediated. In the next two examples, there are no characteristics in common between the motif of the deletion and the flanking sequence. In the last two examples, there is some homology between the first few bases and the immediate flanking sequence. Microhomology does not involve the entire deletion motif and there are no tandem repeats and are termed microhomology-mediated indels. (D) Frequency of indels by indel size. This demonstrates how repeat-mediated indels are usually of smaller size. From a Kolmogorov-Smirnov (K-S) test, the distribution of indel lengths for repeats and microhomologies is significantly different ($p<2.2e-16$). (E) Observed number of bases involved in microhomology at junction of indels versus expected number of bases if microhomology occurred simply by chance (K-S test $p<1.2e-8$).

7.3 DOUBLE SUBSTITUTIONS

In this section, double substitutions were explored as a separate class of mutation. Double substitutions could arise due to two independent events occurring by chance at sites adjacent to each other. An alternative model would posit that mutagenic damage to one is linked to mutation at the adjacent site. This is likely to be the case, for example, for CC>TT/GG>AA mutations caused by UV–light. Apart from the documentation of this signature in *TP53* reporter gene assays and tandem BRAF mutations in malignant melanomas induced by ultraviolet damage (Thomas et al., 2004), there is very little in the literature on phenomena driving double nucleotide mutations. Some clustered mutations have been described in the immediate vicinity of radiation-induced breaks in vitro, also known as oxidatively-generated clustered DNA lesions, but these are not consistently adjacent substitutions and do not show a predilection for attacking guanines (Cadet et al., 2012).

7.3.1 Substantial enrichment of double substitutions was observed in all twenty-one breast cancers

It was observed from the construction of the rainfall plots (chapter 5), that the frequency of substitutions with an intermutation distance of 1bp, which corresponds to adjacent or double substitutions, was substantially higher in some cancers (Figure 5.5, samples PD3904a, PD3945a, PD4120a, PD3890a, PD4109a, PD4116a, PD4005a, PD4115a, PD4006a, PD4107a). Evaluating this further, double substitutions were found to comprise between ~0.5-2.5% of the total number of mutations for each cancer with no significant enrichment for any histopathological subtype (Table 7.1).

In order to test whether there was an enrichment of double substitutions compared to chance adjacency of two independent single nucleotide substitutions, 1000 Monte Carlo simulations were performed corrected for the total number of substitutions and the mutation spectrum present in each genome and the average number of double substitutions per simulation as well as the maximum number of double substitutions across the 1000 simulations were obtained (Table 7.1). The observed number of double substitutions was 75-11,000 fold higher than expected if mutations had been randomly distributed in each of the 21 cancer genomes ($p<0.001$) from the *in silico* simulations. This highly significant enrichment suggests that a mutational process must be actively driving this phenomenon. However, whether it is due to a mutagen with a propensity for damaging adjacent bases or simply a higher likelihood of base mis-incorporation adjacent to a damaged site, is uncertain.

Group	Breast cancer sample	Mean no of simulated double subs	Max no of simulated double subs	Observed number of double subs	Total no of subs	Proportion of double subs
ER +ve HER2 -ve	PD3851	0.002	2	22	1782	0.012
	PD4085	0.004	2	16	2673	0.006
	PD4088	0.000	0	12	1705	0.007
	PD4103	0.020	2	52	5360	0.010
	PD4120	3.182	14	240	70690	0.003
ER +ve HER2 +ve	PD4194	0.000	0	18	1484	0.012
	PD4198	0.018	2	28	4552	0.006
ER -ve HER2 +ve	PD4199	0.036	2	42	6932	0.006
	PD4192	0.018	2	42	3919	0.011
Triple negative	PD4248	0.004	2	40	2536	0.016
	PD4086	0.002	2	12	2199	0.005
	PD4109	0.072	4	86	9888	0.009
BRCA1	PD4107	0.076	2	192	10291	0.019
	PD3890	0.032	2	76	6124	0.012
	PD3905	0.026	2	68	4587	0.015
	PD4005	0.034	2	108	6104	0.018
	PD4006	0.070	4	134	9194	0.015
BRCA2	PD3904	0.028	2	132	5608	0.024
	PD3945	0.076	4	234	10308	0.023
	PD4115	0.070	2	216	9954	0.022
	PD4116	0.056	4	168	8026	0.021

Table 7.1: The double substitutions identified in twenty-one breast cancers are presented. Mean and maximum number of double substitutions identified from 1000 Monte Carlo simulations and observed number of double substitutions are provided.

7.3.2 Mutational spectra of double substitutions differs to that of the overall spectrum

The patterns of double nucleotide substitutions generally reflected the overall patterns of single nucleotide substitutions in each cancer. However, in most cancers there was evidence of a substantial enrichment of C>A/G>T substitutions as components of double nucleotide substitutions (Figure 4.1B) with the consequent emergence of CpC>ApA as the most common class of double nucleotide substitution (Figure 7.3) for this analysis. Mutations of the same consequence on different strands were pooled, for example, CpC>ApA is equivalent to GpG>TpT.

Oxidative lesions, such as 8-oxo-G, have been shown to generate G>T:C>A transversions. Furthermore, a site-specific GGG sequence has been associated with some oxidative damage (see section 1.3.3)(Oikawa and Kawanishi, 1999). It is possible that this mutational signature of CpC>ApA or GpG>TpT identified in double substitutions constitutes the mark of oxidative stress.

Double nucleotide substitutions were distributed throughout the genomes of the cancers in which they were found without obvious evidence for clustering, nor enrichment for particular genomic features.

		Second Mutated Base											
		A>C	A>G	A>T	C>A	C>G	C>T	G>A	G>C	G>T	T>A	T>C	T>G
First Mutated Base	A>C	6	3	6	14	3	9	14	10	25	9	3	0
	A>G	8	10	7	32	4	16	27	10	29	15	10	
	A>T	4	5	32	57	10	35	54	9	68	16		
	C>A	18	51	49	202	40	71	44	13	41			
	C>G	7	10	7	39	6	25	19	5				
	C>T	8	26	21	69	17	104	33					
	G>A	8	18	65	59	27	16						
	G>C	10	5	20	17	3							
	G>T	26	32	83	31								
	T>A	5	15	9									
	T>C	2	2										
	T>G	0											

Figure 7.3: Relationship between first and second substitution in double substitutions showing enrichment for CC>AA mutations.

7.4 REARRANGEMENTS

Structural variation is defined as differences in orientation or location of relatively large genomic segments (typically >100 bp). In cancer, the landscape of somatically acquired structural variation is extremely diverse, ranging from very few to tens or hundreds (Stephens et al., 2009) and this structural variation in cancer is sometimes referred to as 'rearrangements'. Some cancer-associated rearrangements appear to be functional, driver events and under strong selection, such as amplification of oncogenes, deletion of tumour suppressors and translocations that produce fusion genes, but many rearrangements in cancers are passenger events.

7.4.1. The landscape of somatic rearrangements in 21 primary breast cancers

In total, 1192 somatic structural variants or rearrangements were identified in the twenty-one breast cancers. There was substantial variation in the numbers of rearrangements harboured by each breast cancer ranging from 2 rearrangements in PD4005a to 217 rearrangements in PD4103a. Apart from variation in numbers, there was marked variation in distribution of rearrangements through the genome. In some cancers, rearrangements were stochastically distributed whilst in others, rearrangements appeared to cluster within and connect genomic regions associated with amplification (Figure 7.4).

7.4.2 There is marked variation in rearrangement architecture between the twenty-one breast cancers

In this thesis, a previously reported rearrangement classification system (Stephens et al., 2009) which has been derived from the orientations, copy number status and relative chromosomal locations of the two genomic segments forming each rearrangement has been employed. Rearrangement breakpoints are usually identified by comparing the structure of the cancer genome to that of the reference genome, and breakpoint positions are reported based on the coordinate system of the reference.

In essence, each rearrangement was classified according to:

- whether it is within an amplicon,
- if not in an amplicon, whether it is interchromosomal or intrachromosomal,

- if intrachromosomal, whether if results in a deletion, tandem duplication or rearrangement with inverted orientation

There were 839 intrachromosomal and 353 interchromosomal rearrangements in aggregate across the twenty-one breast cancers, with 56.9% being within 2MB of each other. Therefore, intrachromosomal rearrangements outnumbered interchromosomal rearrangements by this analysis, presumably reflecting the greater sensitivity of detection of small intrachromosomal events by second-generation sequencing techniques when compared to historic methods of detecting structural variation in cancer.

The most commonly observed rearrangement architecture in each cancer varied from one cancer to another, but showed some correlation with histopathological subtype. Deletions were commonest in BRCA2 germline mutant cancers and frequent in BRCA1 cancers, although the most common rearrangement architecture in the latter group was tandem duplications. Two ER positive breast cancers, PD4103a and PD4088a were characterised by an excess of amplicon-associated and interchromosomal rearrangements.

Apart from these more common rearrangement architectures, three loci in the 21 genomes reveal evidence of 'chromothripsis' (in PD4248a chr6:6.3-9.9MB ; PD4107a chr6: 130-135MB and PD4120a chr21:16.9-32.6MB) characterised by extraordinarily complex intrachromosomal and/or interchromosomal rearrangements, clustered in a highly non-random manner and associated with defined copy number states (usually two).

Figure 7.4

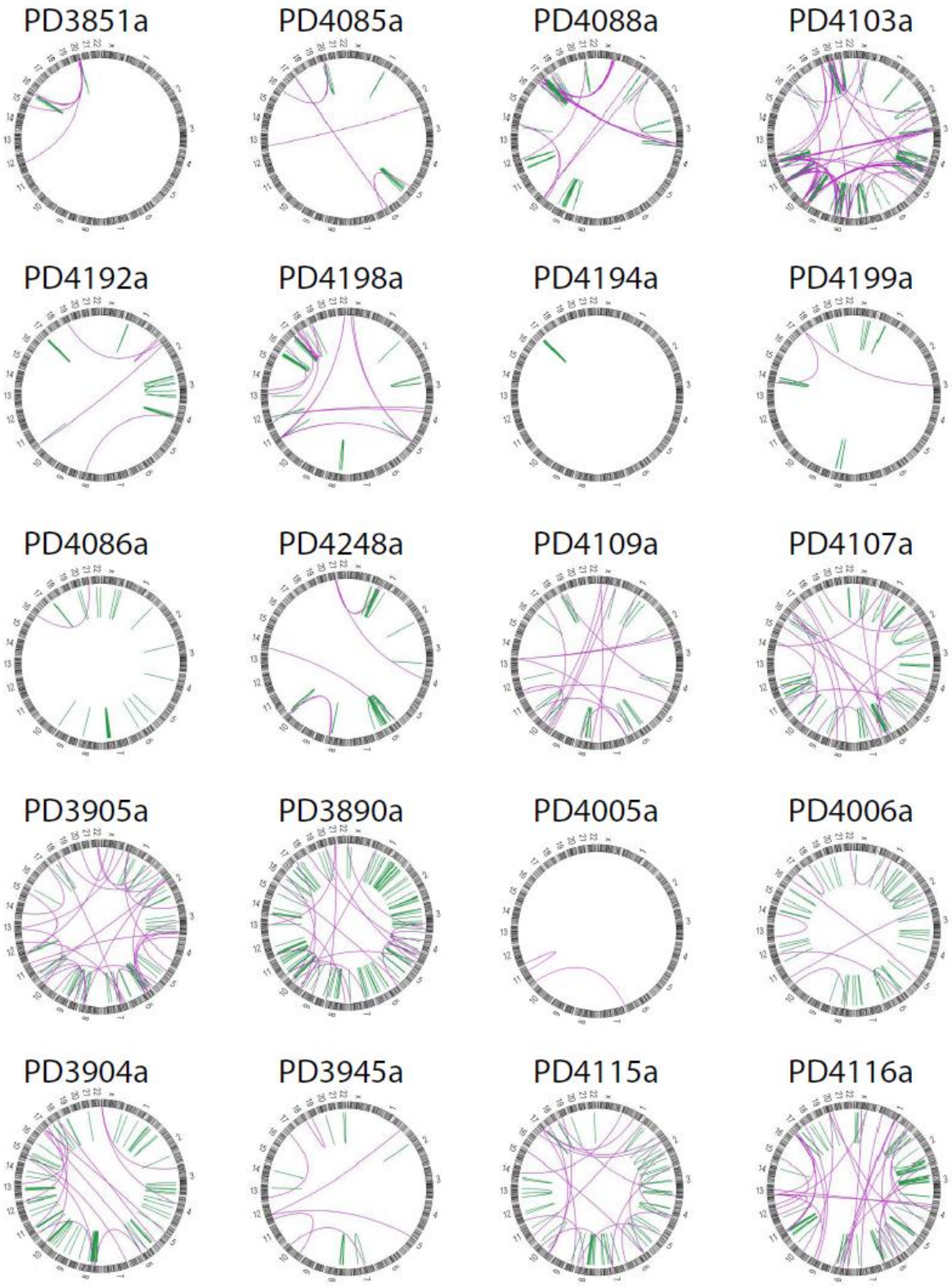


Figure 7.4: Circos plots demonstrating the rearrangements in the 20 breast cancers.

Figure 7.5

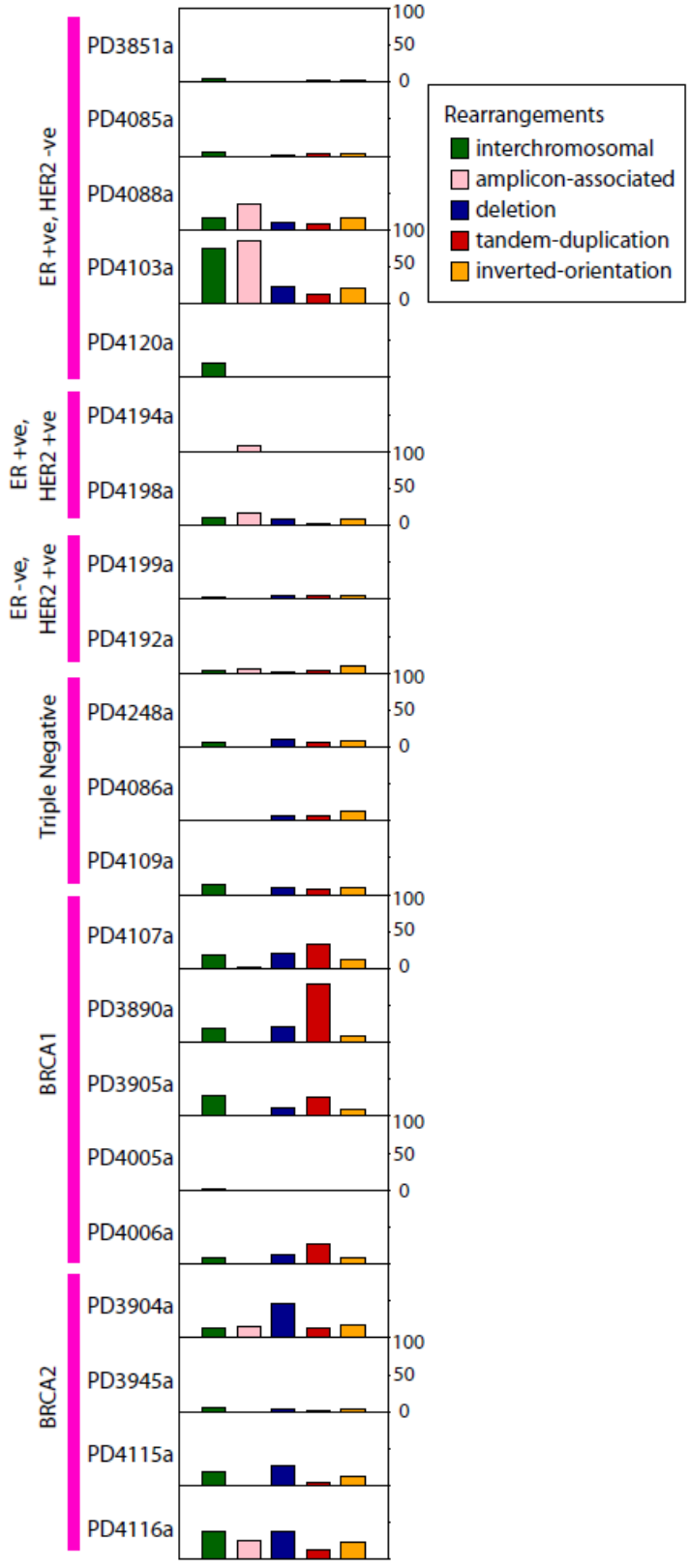


Figure 7.5: Variation in rearrangement architecture between the twenty-one breast cancers

7.4.3 Junctional features at rearrangement breakpoints demonstrate increased microhomology-mediated rearrangements

The sequences either side of each rearrangement junction can reveal insights into the underlying mechanisms involved in generating these rearrangements. Previously, it was shown in low-coverage rearrangement screens of cancers, that in the majority of cases, the two contributing DNA segments either side of a rearrangement junction showed a short stretch of identical sequence, known as an overlapping microhomology, immediately adjacent to the rearrangement junction (Campbell et al., 2008; Stephens et al., 2009). A smaller proportion (~15% in the breast cancer rearrangement screen) showed non-templated sequence at the rearrangement junction.

In this study, 757 of 1192 rearrangements demonstrated at least 1bp of microhomology (63.5%) with 167 rearrangements (14%) showing non-templated sequence of up to 50bp. A further 26 rearrangements (2.2%) had lengths greater than 50bp from elsewhere in the genome interposed between the rearrangement breakpoints identified by paired-end sequencing. These have previously been termed 'genomic shards' (Bignell et al., 2007; Campbell et al., 2008) and the longest segment was 256bp.

Overlapping microhomologies and non-templated sequences at rearrangement junctions are often considered to be signatures of a non-homologous end-joining (NHEJ) DNA double-strand break repair process (Hastings et al., 2009; Hefferin and Tomkinson, 2005; van Gent and van der Burg, 2007; Weterings and Chen, 2008). The segments of overlapping microhomology are believed to facilitate alignment of the two DNA fragments that are combined. It has also been proposed that complex germline rearrangements with genomic shards and overlapping microhomology might be due to replicative mechanisms (Hastings et al., 2009).

It was demonstrated (Stephens et al., 2009) that in some breast cancers, rearrangements with zero base pairs of microhomology were most frequent, whereas in others rearrangements with two or more base pairs were the commonest class. In these twenty-one breast cancers, rearrangements with zero base pairs of microhomology were most common for amplicon-associated rearrangements. In other classes of rearrangement, although zero base pairs of microhomology was still very high the modal class of microhomology was 2 bp (Figure 7.7). These differences suggest two distinct classes of NHEJ repair may be are operative to different extents in different somatic rearrangement architectures. This difference relative to chance occurrence was highly significant (KS-test, $P<0.0001$ for both).

Figure 7.6

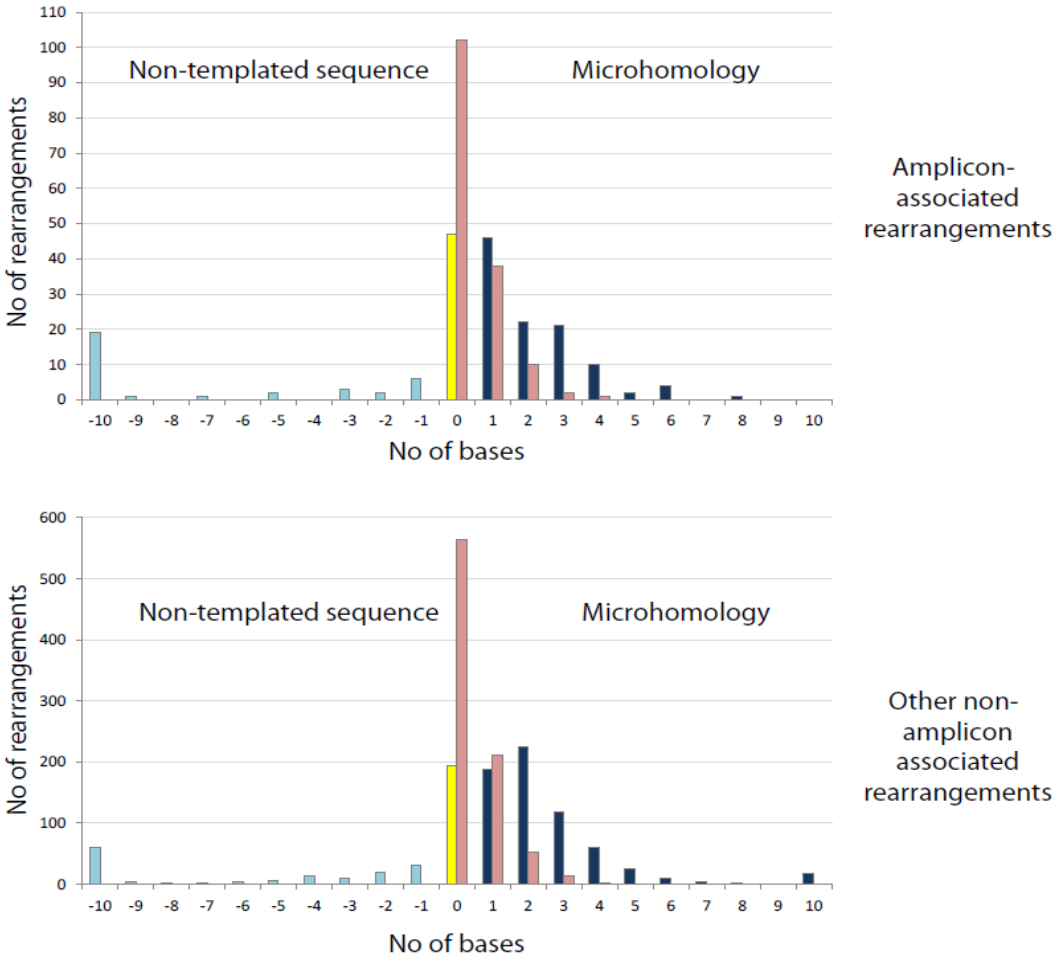


Figure 7.6: Patterns of microhomology (dark blue) and non-templated sequence (light blue) at rearrangement breakpoints of twenty-one breast cancers. The occurrence of microhomology by chance presented in pink. Difference in distribution of number of bases involved in microhomology between observed and chance were highly significant (KS-test $p < 0.0001$) for both amplicon-associated and non-amplicon associated rearrangements.

7.4.4 Rearrangements involving protein-coding genes

61% of rearrangements had breakpoints falling within the footprint of a protein coding gene compared to chance (p=2.8e-5). This observation was made previously in a rearrangement screen of 24 breast cancers (Stephens et al., 2009). The reason for this enrichment of rearrangements in genic regions is not clear. It is conceivable that some of this effect may be due to selection for rearrangements which are located in genes that confer selective advantage on a cancer clone and therefore that a subset of rearrangements is implicated in cancer development. However, it is also likely that there are structural properties of genic regions that increase the likelihood of a DNA double-strand break occurring, perhaps through chromatin configuration or active transcription.

130 rearrangements were predicted to generate in-frame rearrangements, of which 88 were in-frame internally rearranged genes. 42 rearrangements were predicted to generate in-frame gene fusions. In-frame fusion genes are potentially of biological interest as candidates for new cancer genes. However, fusion genes implicated in cancer development are likely to be recurrent. None of the novel fusion genes identified in this analysis was present in more than one out of the 21 cancers screened. In a previous low-coverage rearrangement screen of 24 breast cancers, three expressed, in-frame fusion genes were examined by FISH (*ETV6–ITPR2*, *NFIA–EHF* and *SLC26A6–PRKAR2A*) and twenty by RT–PCR across the rearranged exon–exon junction in 288 additional breast cancer cases. No examples of recurrence were found, indicating that they are either passenger events or that they contribute infrequently to breast cancer development. None of these three were found in the twenty-one breast cancer genomes.

Thirty-two genes were rearranged in multiple cancers. One gene, ADAM2 was rearranged in three different cancers. Some of these recurrently hit genes were in known targets of genomic amplification in breast cancer. It is likely that these are recurrently rearranged because of the high density of rearrangements associated with these regions of recurrent genomic amplification. Others, however, generally had large genomic footprints and may simply represent bigger targets for randomly positioned rearrangements. For some, however, an elevated local rate of DNA double strand breakage (‘fragility’) may also contribute to the clustering of rearrangements.

7.5 COPY NUMBER CHANGES

Gross chromosomal anomalies were amongst the earliest genetic aberrations identified as being characteristic of cancer. Genomic DNA copy number aberrations in cancer may take the form of copy number gains or losses and may contribute to alterations in the expression of tumour-suppressor genes and oncogenes, respectively. In the last 15 years, cancer genomes have been extensively charted by modern platforms of gene dosage analysis including array-comparative genomic hybridization (Bergamaschi et al., 2006) and SNP6.0 arrays (Bignell et al., 2010).

The importance of the identification of copy number aberrations is seen in how hemizygous and homozygous deletions achieve functional inactivation (e.g. p53, PTEN, CDKN2A), in contrast to genomic amplification which contributes to uncontrolled positive growth signaling (e.g. ERBB2). The copy number status of cancer genes can also serve as prognostic markers in various cancer types and, as in the case of ERBB2, and can constitute an effective target for therapy. Furthermore, the increasing resolution of gene dosage analyses have allowed highly accurate localization of specific genetic alterations and revealed associations with tumour progression and response to treatment [reviewed in (Kallioniemi, 2008).

Modern platforms, such as the affymetrix genome-wide SNP6.0 platform, offer gene dosage analyses and perform genotyping experiments across millions of single nucleotide polymorphisms (SNPs) simultaneously, which produce copy number information in addition to SNP genotypes. Additional non-polymorphic probes are present and designed to give greater genomic resolution of copy number in regions of lower SNP density. These methods are however restricted to detecting non-reciprocal or unbalanced structural changes where there is a physical change in copy number of a region of the genome.

An algorithm called “ASCAT” or allele-specific copy number analysis of tumors was used to estimate the fraction of aberrant cells and the tumor ploidy, as well as whole-genome allele-specific copy number profiles. ASCAT is an algorithm (Van Loo et al., 2010) that has considered and modeled the following two properties in cancer; that tumours often deviate from a diploid state (Holland and Cleveland, 2009; Rajagopalan and Lengauer, 2004) and that cancers are likely to comprise multiple populations of both tumour and non-tumour cells (Witz and Levy-Nissenbaum, 2006). ASCAT is therefore able to provide these estimates (Table 7.2) in the twenty-one breast cancers.

Table 7.2: Estimates of aberrant cell fraction and ploidy are made by ASCAT. Normal DNA content is derived by the following: 2 X (1-aberrant cell fraction). Tumour DNA content is obtained from the product of aberrant cell fraction and ploidy. Total DNA content is obtained from the addition of normal and tumour DNA together. Normal contamination is the fraction of normal DNA from the total DNA content.

Sample	Aberrant cell fraction	Ploidy	Normal DNA content	Tumour DNA content	Total DNA content	Normal contamination
PD3851a	0.63	3.20	0.74	2.01	2.754	0.269
PD3890a	0.49	1.78	1.02	0.87	1.890	0.540
PD3904a	0.79	1.97	0.42	1.56	1.976	0.213
PD3905a	0.8	3.72	0.4	2.97	3.373	0.119
PD3945a	0.44	3.94	1.12	1.74	2.855	0.392
PD4005a	0.45	1.84	1.1	0.83	1.927	0.571
PD4006a	0.59	2.93	0.82	1.73	2.548	0.322
PD4085a	0.68	2.81	0.64	1.91	2.549	0.251
PD4086a	0.37	3.06	1.26	1.13	2.392	0.527
PD4088a	0.63	1.81	0.74	1.14	1.880	0.394
PD4103a	0.56	3.89	0.88	2.18	3.061	0.287
PD4107a	0.57	2.86	0.86	1.63	2.492	0.345
PD4109a	0.5	3.32	1	1.66	2.660	0.376
PD4115a	0.69	3.92	0.62	2.71	3.327	0.186
PD4116a	0.67	3.18	0.66	2.13	2.790	0.237
PD4192a	0.22	4.68	1.56	1.03	2.590	0.602
PD4194a	0.57	1.98	0.86	1.13	1.990	0.432
PD4198a	0.32	3.05	1.36	0.97	2.335	0.583
PD4199a	0.56	1.69	0.88	0.94	1.825	0.482
PD4248a	0.29	3.09	1.42	0.90	2.316	0.613

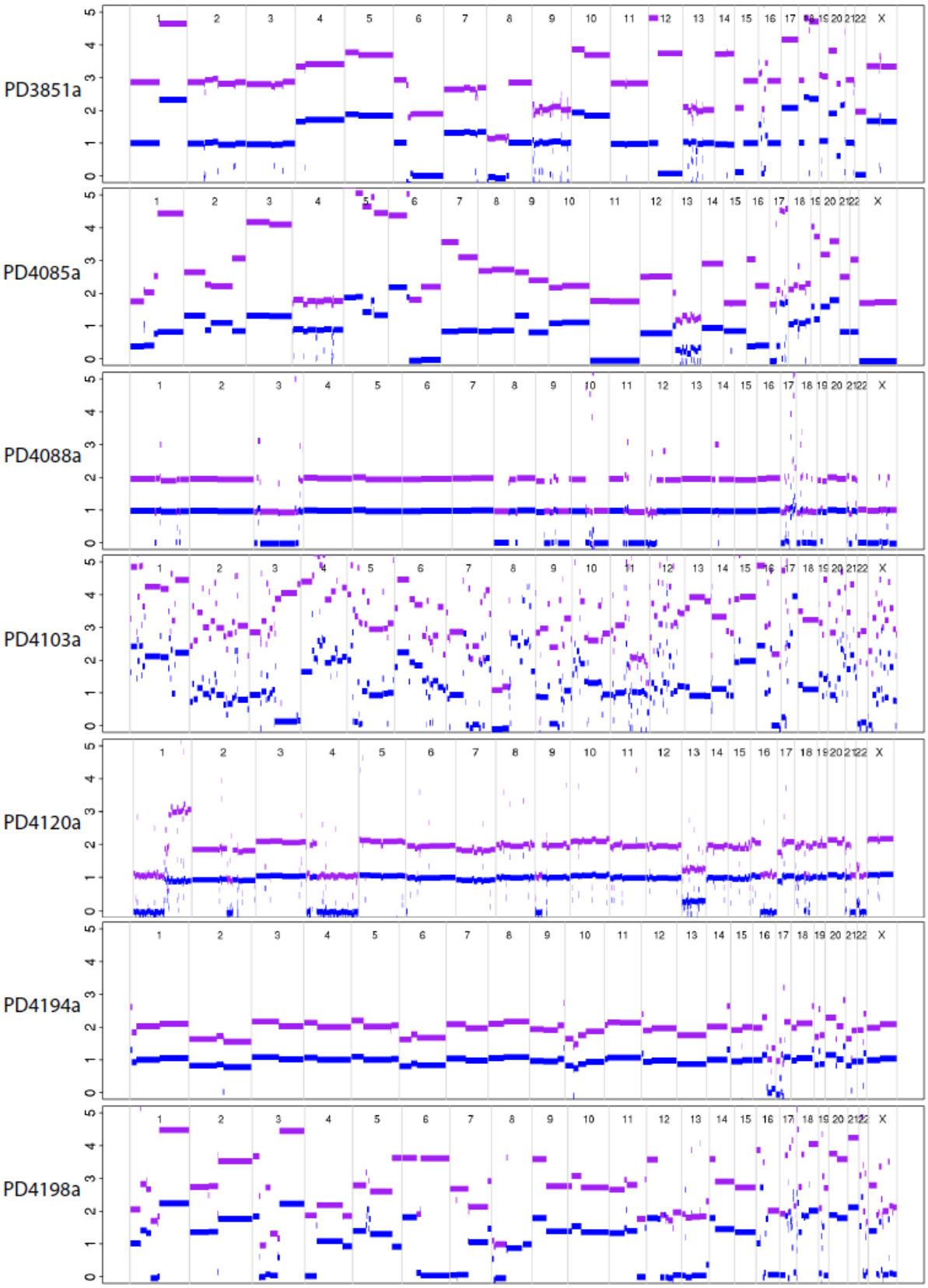
7.5.1 The observed variation in gross copy number changes between breast cancers

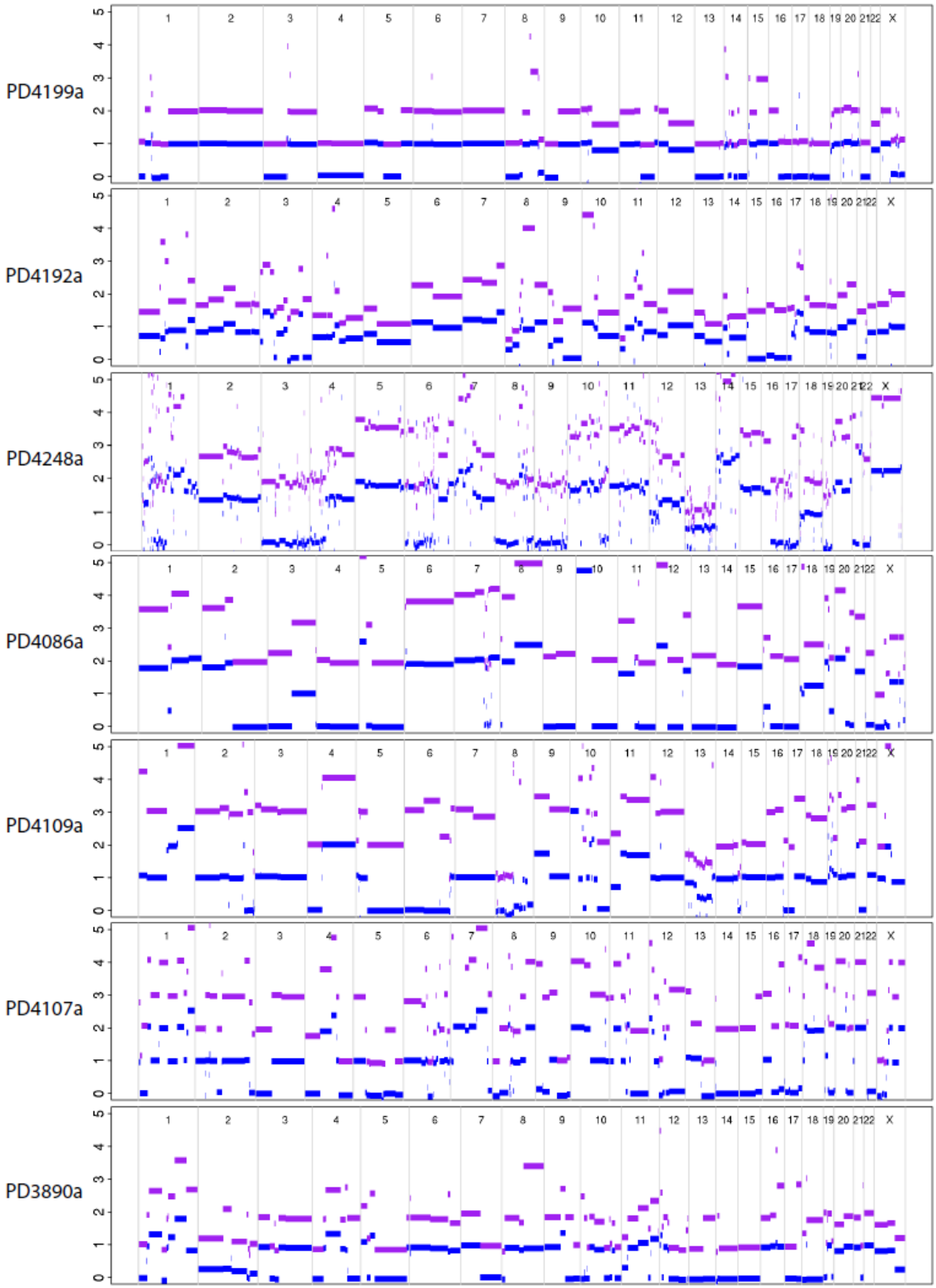
There was distinct copy number variation between breast cancers with copy number changes typical of previous descriptions of breast cancers. Samples PD4194a and PD4088a showed relatively quiescent copy number profiles compared to the rest of the cancers. Frequently observed copy number aberrations included gain of chromosomal regions 1q (PD4109a, PD4198a, PD3851a, PD3890a, PD3945a, PD4005a, PD4006a, PD4085a, PD4120a), 8q (all breast cancers apart from PD4085a, PD4088a and PD4194a) and 17q (PD4005a, PD4086a, PD4194a and PD4199a) and loss of 1p (PD3890a, PD3904a, PD4006a, PD4107a, PD4115a, PD4199a, PD4120a), 8p (PD3851a, PD390a, PD3945a, PD4088a, PD4103a, PD4107a, PD4109a, PD4192a, PD4198a, PD4199a), 13q (PD3945a, PD4006a, PD4107a, PD4085a, PD4120a) and 17p (all bar PD3851a), in-keeping with previous reports of common gains and losses in breast cancer (Knuutila et al., 2000).

Loss of heterozygosity (LOH) or loss of one parental allele with or without duplication of the remaining allele was seen consistently and involved a total of 1182 regions (Appendix 5). LOH with reduplication occurred in 774 regions and were informative for the analysis of the timing of mutational events described in Chapter 4 and 5. LOH was most frequent on chromosome arms 8p, 11q, 16q, and 17p. A higher frequency of LOH specifically in the triple negative (basal-like subtype) of breast cancers was apparent ($P = 1.0 \times 10^{-7}$ by a *t* test looking for differences between triple negative breast carcinomas and other carcinomas).

All tumours derived from *BRCA1* or *BRCA2* germline mutation carriers showed loss of the wild type haplotypes at 17q21 or 13q12 respectively, as expected of recessive cancer genes. All the breast cancers apart from PD3851a showed loss of a wild-type haplotype at 17p13 (*TP53*).

Figure 7.7




PD4199a
PD4192a
PD4248a
PD4086a
PD4109a
PD4107a
PD3890a

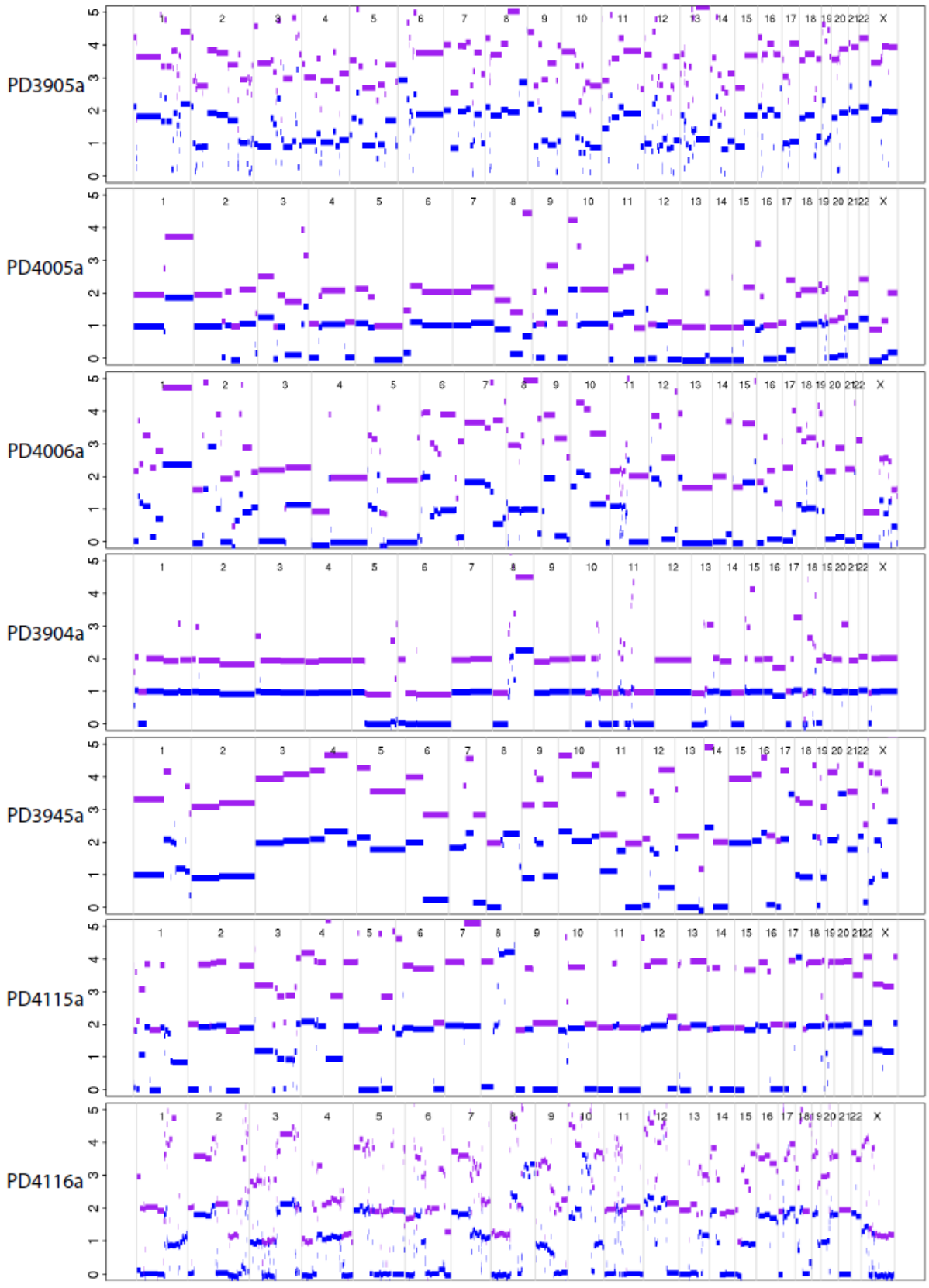


Figure 7.7: Copy number plots for all twenty-one breast cancers. Chromosomes provided along the horizontal axis and copy number values on the vertical axis for each cancer. Purple lines denote total copy number whilst blue denotes minor copy number values.

7.5.2 Fourteen regions of amplification involving putative target genes were identified

Previous efforts at characterization of genomic copy number profiles in breast cancers had identified sites of localised high-level DNA amplification harbouring oncogenes. These include 7p12 (*EGFR*), 8q24 (*MYC*), 11q13 (*CCND1*), 12q14 (*MDM2*), 17q12 (*ERBB2*), 20q12 (*AIB1*), and 20q13 (*ZNF217*) [reviewed ((Al-Kuraya et al., 2004) and references therein)]. In order to identify regions of amplification in the twenty-one breast cancers, a total copy number threshold was set as follows. Genomic segments in breast cancers which were estimated as overall diploid (copy number less than 2.5) by ASCAT, had to exceed a total copy number of more than or equal to 5 in any particular segment in order to be considered a region of amplification. Genomic segments in breast cancers which had a higher overall ploidy (copy number more than or equal to 2.5) had to exceed a total copy number threshold of more than or equal to 9 in any segment to qualify as a region of amplification. Altogether, 180 segments were identified as amplifications across the twenty-one breast cancers encompassing 583Mb of genome in total.

In order to identify putative amplification target genes, the segments identified by the criteria described in the paragraph above were mapped to the amplified cancer gene census in COSMIC and fourteen putative target gene regions of amplification were identified in nine of the twenty-one breast cancer genomes. The highest levels of amplification were seen at the ERBB2 locus of the four HER positive breast cancers in this cohort. Two breast cancers showed two independent target gene loci of amplification and one breast cancer, PD4103a, an ER positive PR positive HER2 negative breast cancer showed four regions of amplification with putative target genes, which were involved in an interconnected web of rearrangements. The list of potential targets of amplification is provided along with the genomic loci of the region of amplification in Table 7.3 below.

Table 7.3: Amplifications identified in twenty-one breast cancer genomes

Amplifications					
Sample	**Chr**	**Start position (bp)**	**End position(bp)**	**Putative Target Gene**	**Copy number**
PD4192a	17	37833600	38018803	ERBB2	51
PD4194a	17	37833600	38018803	ERBB2	18
PD4198a	17	37833600	38018803	ERBB2	14
PD4199a	17	37833600	38018803	ERBB2	29
PD4103a	11	69224506	69556470	CCND1	22
PD4116a	11	69224506	69556470	CCND1	18
PD4198a	11	69224506	69556470	CCND1	12
PD3904a	8	37353781	37489508	FGFR1/ZNF703	9
PD4005a	8	128504497	1.29E+08	MYC	5
PD4103a	8	128504497	1.29E+08	MYC	10
PD4115a	8	128504497	1.29E+08	MYC	13
PD4116a	8	128504497	1.29E+08	MYC	10
PD4103a	20	52065876	52723895	ZNF217	19
PD4103a	12	69038072	70197123	MDM2	28

As expected, HER2 positive (or ERBB2-subtype) tumours, characterised by overexpression of *ERBB2* and its neighbors exhibited consistent amplification at 17q12-q21 which harbours the *HER2*/*ERBB2* gene.

7.5.3 Sixteen homozygous deletions were identified in ten breast cancers

Homozygous deletions were identified as regions of the genome where the total copy number state was zero. Sixteen homozygous deletions were identified in ten of the twenty-one breast cancers. Putative cancer genes were sought via the Cancer Gene Census and only *MAP2K4* was identified as a significant tumour suppressor candidate (Table 7.4).

Table 7.4: Homozygous deletions identified by ASCAT

Homozygous deletions				
Sample	**Chr**	**Start position (bp)**	**End position (bp)**	**Annotation**
PD4116a	1	37855871	37885161	
PD4006a	2	136849419	145317330	
PD4006a	5	59519068	60420580	
PD4006a	6	144982326	145195890	UTRN
PD3851a	7	11727099	11773517	THSD7A
PD4006a	7	117929448	118035293	
PD4088a	10	82497699	83204305	
PD4116a	11	85834374	85877316	
PD4248a	13	28685062	32338443	
PD4248a	13	39239090	44876701	
PD4088a	17	11645786	12337597	MAP2K4
PD4198a	17	58802859	58811328	BCAS3
PD4199a	17	70116518	70516791	
PD3904a	18	5872382	8002993	
PD3904a	18	41698511	41710757	
PD4006a	X	31971839	33354165	DMD

7.6 DISCUSSION

So far, the derivation of mutational signatures has been focused on those which are discernible within somatic substitutions. In this chapter, mutational signatures from other mutation classes namely double substitutions, insertions/deletions and rearrangements were sought. Double substitutions were enriched in all 21 breast cancers, and showed a preponderance for C>A mutations. Furthermore, CC>AA mutations were the most common double substitution. The mechanism underlying this pattern is unknown, although it is possible that these are remnants of oxidative DNA lesions. Two mutational signatures were appreciable in insertions/deletions. Within indels, a signature was observable in small indels (<5bp) flanked by small tandem repeats, evidence of an accumulation of oversights of post-replicative mismatch repair. A second signature was identifiable, enriched from amongst the breast cancers with *BRCA1* and *BRCA2* germline mutation carriers, and comprising larger indels (>=5bp) sharing a degree of microhomology with flanking sequence. This is postulated to be the mark of microhomology-mediated repair of non-homologous end-joining. Microhomology-mediated repair of breakpoints was not restricted to insertions/deletions and were also seen in somatic rearrangements invoking the activity of similar microhomology-mediated repair mechanisms in the generation of large-scale variation in cancers. This chapter demonstrates how other biological processes that shape the mutation landscape in cancers are not confined to somatic substitutions but may leave traces of activity in other mutation classes.

7.6.1 The mutational process generating double substitutions is unknown

The best described double nucleotide substitutions in human cancer are the CpC>TpT mutations found in skin tumours, generally attributed to the presence of pyrimidine dimers that arise as a consequence of ultraviolet light exposure. This highly specific mutational signature is unlikely to be the source of CpC>ApA mutations in breast cancer. Clustered substitutions which culminate as double substitutions generated near sites of damage by ionizing radiation are not known to generate any particular signature. However, secondary oxidative DNA lesions, or those from reactive oxygen species are believed to have a predilection for guanines (Cadet et al., 2012), so may underlie the excess of these mutations in breast cancers.

7.6.2 Two mutational processes are present generating insertions/deletions

Two insertion/deletion signatures were instantly appreciable from an analysis of the indels in these breast cancers and were compared to the table in the introductory chapter (Table 1.1). Firstly, the architecture of small indels (< 5bp) occurring at tandemly-repeating sequences is a feature of errors accumulated by post-replicative mismatch repair. It is thought that insertion-deletion loops form around sites of simple sequences such as repeat tracts during replication. Indels accumulate at such regions producing a signature of small indels (1-3 bp) forming predominantly around simple repeat tracts. Although post-replicative mismatch repair improves the error rate in replication significantly, an error rate still exists.

This signature was universally present in twenty-one breast cancers without exception. Unlike the observation in some colorectal cancers, however, the breast cancers were not overwhelmed by insertions and deletions at microsatellite repeat tracts, and did not have mutations in genes associated with post-replicative mismatch repair. Therefore, given the ubiquitous nature of this indel mutational signature, it is postulated that this mutational process is simply one that is occurring in all tissues. It may represent the usual rate of error of post-replicative mismatch repair but perhaps seen at a higher prevalence because of the increased number of mitoses in each cancer, with some variation between cancers resulting in the variation in the total number of small indels.

In contrast, the enrichment of microhomology-mediated indels in breast cancers derived from women with *BRCA1* and *BRCA2* mutations plausibly suggests a microhomology-mediated repair process compensating for the defective repair by homologous recombination of double-strand breaks. This alternative mutational process was restricted to germline mutated breast cancers, and was clearly distinct from the mutational process generating small indels. This analysis demonstrates how multiple mutational processes may be discernible even within one class of mutation that is indels.

7.6.3 Multiple mutational processes are at play generating large-scale rearrangements in cancer

Amplicon-associated rearrangements have zero base pairs of microhomology as a modal feature of flanking bases at the rearrangement junction implying that the double-strand repair involved is likely to be mediated by blunt end-to-end fusion. In-contrast, non-amplicon-associated rearrangements demonstrated dependence on microhomology-mediated processes of repair suggesting that at least two different repair processes are at play in generating somatic rearrangements. However, it should be emphasised that the numbers in this study are small and perhaps limited by the sensitivity of the rearrangement –calling algorithm.

CHAPTER EIGHT: DISCUSSION

8.1 INTRODUCTION

In the course of this thesis, catalogues of all classes of somatic mutation from twenty-one whole genome sequenced breast cancers have been curated and archived. Detailed analyses of these catalogues have yielded several insights into underlying mutational processes which were defined in a previous chapter as comprising some combination of DNA damaging and DNA reparative mechanisms.

Different mutational processes have been highlighted by the different chapters in this thesis using a variety of methods including mathematical methods and integration of different mutation-types. The characteristic features of each mutational process have been sought and compared to the collection of known mutational signatures reviewed in the introduction. Possible biological candidates for each mutational signature discovered have been discussed.

In this chapter, the wealth of biological information that is revealed by the detailed analysis of the data is highlighted. Potential future directions are also discussed.

8.2 THE EXTRACTION OF COMPONENTS OF MUTAGENESIS AND REPAIR FROM MUTATIONAL SIGNATURES

8.2.1 At least eleven different mutational signatures were identified in this study

In the introductory chapter, a variety of known DNA mutagens and DNA repair pathways were described and mutational signatures related to these mutagenic/repair pathways were sought from the literature. Subsequently, using mathematical methods, five independent single nucleotide substitution processes were extracted from cancer genome datasets, generating the observed variation in mutation numbers and patterns between cancers as described in chapter 4. Analysis of variation in mutation density revealed another mutational process characterised by localised hypermutation, termed kataegis, in chapter 5. Integration of substitution data with transcriptomics revealed evidence of transcription-coupled and expression-related repair being operative in chapter 6. Finally, analyses of other mutation types revealed a double substitution mutational process, two

different insertion/deletion signatures and rearrangement phenotypes in chapter 7. In total, eleven clear mutational signatures were identified in this study.

The individual features of each mutational process have been characterised and compared to the collection of known mutational signatures reviewed in the introduction. Plausible biological interpretations for some mutational signatures are discussed in the next section.

8.2.2 Unravelling the components of mutagenesis and repair from mutational signatures

Because a mutational signature is the imprint left by a mutational process governed by any combination of mutagenic and repair mechanisms, these signatures can be compared and contrasted to one another in order to tease apart the components of each mutational process. For example, Signature E also exhibits mutations at TpCpX trinucleotides, but is characterised by a much lower fraction of C>T mutations than Signature B. It is possible that both Signature B and E result from cytosine to uracil deamination by an APOBEC family member, but that the different signatures are sequelae of different repair mechanisms following the deamination step. C>T transitions may simply result from DNA replication across uracil. However, if uracil is excised by uracil-DNA glycosylase (UNG) as part of base excision repair (BER), an abasic site is generated (Wilson and Bohr, 2007). The partiality for C>G transversions in Signature E may reflect preferential insertion of cytosine opposite such an UNG-mediated abasic site. The propensity to introduce cytosine opposite an abasic site is characteristic of REV1 translesion polymerase (Jansen et al., 2006; Ross and Sale, 2006). Thus, Signature B may be caused by a combination of replicative polymerases, while Signature E may be the imprint of the almost exclusive activity of REV1 translesion polymerase. Contrast this with the results from chapter 6, where transcriptional strand bias was identified in two mutation-types, C>A/G>T and T>G/A>C mutations. Although both showed evidence of strand bias, a proxy for the activity of transcription-coupled repair, it is plausible that given the different nature of the mutated base, disparate mutagenic assaults have been resolved by the same repair pathway.

In essence, the mutational signatures identified are the remnants of the processes that have been operative for which the mutagenic and repair components can be teased apart. In the two examples described above, the first describes a situation where the same mutagenic damage may be repaired or resolved by different mechanisms, and the second demonstrates different mutagenic effects repaired by the same pathway. The processes appear to have been acting in combination, either contemporaneously or during different phases of evolution of the cancer clone. Additional subtle processes may exist, and sharper definition of currently characterised processes may follow refinements of NMF and inclusion of other mutational features in the models.

8.2.3 Potential future directions: Exploring biological processes underlying mutational signatures identified in cancer genomes

It is anticipated that whole genome sequencing of many hundreds of breast cancers as well as other types of cancers will reveal further mutational signatures. The biological mechanisms underlying these mutational signatures will, in large part, be uncertain. A potential future direction would be to explore the biological basis of these and other mutational signatures that emerge from sequencing cancer genomes. Using engineered model systems, components of repair/replication pathways could be systematically manipulated for targeted over-expression or knock-down experiments and second generation sequencing technologies can be used to obtain genomic readouts. Ultimately, the aim would be to compare cryptic signatures extracted from cancer genomes to this archive of controlled signatures in order to elucidate their pathogenesis, an extension of the over-arching method used in this thesis, where cancer-detected signatures were compared to the limited collection of well-described signatures obtained from the literature.

8.3 MODELS FOR THE MECHANISMS UNDERLYING SIGNATURE B AND KATAEGIS

The detailed analyses performed in this thesis have revealed subtle variations in mutational signatures which have important biological connotations. In this section, using similarities and differences between Signature B and kataegis, as an example, hypothetical models describing the genesis of these two patterns are discussed, reiterating the potential weight of the biological message that is hidden in these large datasets.

8.3.1 Signature B and kataegis share startling similarities in their mutational features but also have locoregional differences

In chapter 4, Signature B, characterised by C>T, C>G, and C>A substitutions at TpCpX trinucleotides, was found to be responsible for the overwhelming majority of mutations in two cancer samples, PD4120a and PD4199a. It is believed that this signature is present in this dominant form in approximately 10% of ER positive breast cancers (Stephens et al., 2012). In chapter 5, a remarkable process generating regional hypermutation called kataegis, was found to be frequently operative in breast cancer. Mutations within regions of kataegis bear similarities to those in Signature B, notably the preponderance of C>T and C>G substitutions at TpCpX trinucleotides. Additionally, they are closely associated with regions of rearrangement and occur on the same chromosome and chromosomal strand over long genomic distances, suggesting that they occur simultaneously or in a processive manner over a short time span (Chen et al., 2012a). Despite sharing common mutational features, however, the mutational process generating signature B appears to be unleashed globally, mutating the whole genome with little regard for the presence of rearrangements as opposed to being regionally targeted in the vicinity of rearrangements in kataegis.

8.3.2 The APOBEC family of cytidine deaminases are implicated in Signature B and kataegis

The APOBEC family of proteins has been implicated in kataegis and/or in Signature B because of the similarities to mutational patterns observed in other biological contexts or in experimental systems. Further studies are, however, required to explore whether and how APOBEC family members contribute to these two forms of mutagenesis in cancer.

8.3.3 A pre-requisite for APOBEC activity is single-stranded DNA

APOBECs possess an intriguing requirement for single-stranded DNA in order to accomplish the task of cytosine deamination. It remains unclear how and when an APOBEC would gain access to single-stranded DNA in these cancers. However, based on this requirement we can posit two models for the generation of kataegis and Signature B respectively.

8.3.4 A model for localised bursts of activity of APOBEC resulting in kataegis

One potential model is that resection of one strand at the broken ends of double-strand breaks exposes single-stranded DNA for APOBEC deamination. Recently, a study on lymphoma cells expressing high APOBEC3G levels displayed efficient repair of genomic double strand breaks induced by ionizing radiation, with transient localization of APOBEC3G to damage foci (Nowarski et al., 2012). APOBEC3G knockdown resulted in deficient repair whilst reconstitution reinstated efficient repair, suggesting a role for APOBEC3G in processing of DNA flanking a double-strand break, providing support for this hypothesis. This model would explain the stochastic nature of the topographical occurrences of kataegis, explain the clustering of kataegis with rearrangements and inform the temporal relationship between kataegis and rearrangements. Furthermore, it may explain a further observation made in these breast cancer genomes. Rearrangements which do not appear to have any associated kataegis may simply not have been exposed to APOBECs. The converse could also be true. Kataegis may be the only trace of what was an exposed section of single-stranded DNA from a double strand break which has been repaired correctly.

Other mechanisms and enzymatic activities may, however, be responsible for kataegis. If so, the question of which constitutes the primary set of lesions, the rearrangements or the substitutions observed in kataegis, remains to be addressed. If a stochastic event in a cell nucleus results in a DNA DSB and repair of this break is associated with accumulation of substitutions in the vicinity of the consequent rearrangement, this could provide an explanation for the regional targeting of kataegis. Indeed, such mechanisms have been reported in yeast (Deem et al., 2011; Hicks et al., 2010; Roberts et al., 2012).

8.3.5 A model for APOBECs generating globally mutated cancers

In contrast to the localised hypermutation observed in kataegis, a globally hypermutated phenotype is observed in up to 10% of ER positive breast cancers (Stephens et al., 2012). If APOBECs were involved in generating this phenotype, then the availability of long stretches of single-stranded DNA would be required at some point during the cell cycle.

The unwinding of DNA by topoisomerases and helicases during replication in S phase could provide such an opportunity, transiently exposing a stretch of persistent single-stranded DNA as a substrate for APOBECs, perhaps through uncoupling between the leading and lagging strands of the replication fork. In 1979, the Lindahl laboratory revealed the presence of single-stranded DNA in nuclei of cancerous human Molt-4 acute lymphoblastic leukaemia cell line and Raji Burkitt lymphoma cell lines (Bjursell et al., 1979). Through fractionation, identification of sedimentation coefficients, efficient removal of isolated material through treatment with pancreatic DNase but not pancreatic RNase as well as repeated electron microscopy observations, they confirmed that the isolated material comprised long single-stranded DNA of 11-35 microns in length corresponding to 25000-80000 unstretched nucleotides (assuming unstretched ssDNA single base size of 0.43nm) (Tinland, B.; Pluen, A.; Sturm, J.; Weill, G. Macromolecules 1997, 30 (19), 5763–5765). Moreover, they showed that the isolation of this fraction of material was confined to the S phase, supporting the above notion that long stretches of single-stranded DNA can become available for APOBEC deamination activity during the synthesis phase of replication, providing the opportunity for globally hypermutated sequences in a cancer genome.

In support of this model, it is observed that processive C>T and C>G mutations at a TpC context given by Signature B shared the same variant allele fraction over a region of equivalent ploidy. This argues that the individual processive stretches of Signature B are occurring during at the same instant within a single cell cycle. However, different processive stretches can occur at different variant allele fractions, with some occurring below the variant allele fraction expected for that level of ploidy and normal contamination (Table 8.1) indicative of the activity of Signature B occurring in subclonal populations. This does suggest that APOBEC activity occurred early in the evolution of the cancer, such that processive patches are present in all the cells in the cancer, but that APOBEC deamination also occurred later in the phylogenetic evolution of the cancer, hence its subclonal imprints. This is an important biological insight indicating that transient hypermutability conferred by a deaminating enzyme can occur multiple times over the evolutionary lifetime of the cancer. It is postulated that in the 10% of hypermutated breast cancers projected to exist (Stephens et al., 2012), APOBECs are somehow permitted to strike the genome recurrently over the evolution of these cancers.

Chr	Coordinate	Wildtype base	Mutant base	a_count	c_count	g_count	t_count	Read Depth	Variant Allele Fraction
10	16803895	G	A	15	0	164	0	179	0.084
10	16819013	G	A	20	0	218	0	238	0.084
10	16857153	G	C	0	16	181	0	197	0.081
10	17273005	G	A	66	0	123	0	189	0.349
10	17314552	G	C	0	77	104	0	181	0.425
10	17389545	G	A	70	0	121	1	192	0.365

Table 8.1: The lower three variants are examples of processive heterozygous mutations occurring at the expected variant allele fraction (~35%) for a clonal population with a diploid chromosome in a sample with ~30% normal contamination. The processive heterozygous mutations occurring at a much lower variant allele fraction (~8%) suggests that the mechanism generating different groups of processive mutations are continuously occurring throughout the evolution of the cancer and has occurred in the ancestral clone of the cancer as well as occurred in a subclonal population.

Both of the globally mutated cancers, PD4199a and PD4120a harboured driver somatically acquired *TP53* mutations (http://www.sanger.ac.uk/genetics/CGP/cosmic/). *TP53* mutations are however common in breast cancer and it is impossible to draw any conclusions on such a small number of samples in this thesis. However, it is interesting to consider that a potentially permissive state, such as down-regulation of checkpoint control may be necessary in order to generate a globally hypermutated phenotype.

8.3.6 The absence of apparent mutations in the APOBEC gene family

So far, no recurrent substitutions, indels or rearrangements have been identified in any of the APOBECS in order to explain the apparent mutational signatures seen. The possibility of up-regulation through gene fusion which has not been detected by the current rearrangement-calling algorithm cannot be dismissed. The APOBEC3 gene family comprises a family of seven highly homologous genes residing in tandem on chromosome 22 having arisen as a gene expansion in placental mammals. This region shows problematic mapping of short read sequences and may curb the detection of mutations.

The lack of any detectable relationship between APOBEC expression and hypermutable phenotype may in part be due to the lack of expression data in two key samples, PD4120a and PD4199a, but may also simply reflect the transcriptional state of the cancer at the time of expression analysis.

Notwithstanding, persistently elevated expression of an APOBEC gene would not explain why some cancers are globally hypermutated and others show localised hypermutation.

It is plausible that there is no aberrant APOBEC activity. If APOBECs are in fact somehow involved in the conduct of normal repair or replication, then what we see as localised hypermutation or global hypermutation may simply reflect the normal effects of APOBECs under abnormal circumstances. In most normal cells, this hyper-editing activity may be poorly tolerated and may lead to cell death. However, under circumstances which allow cancer cells to survive (a permissive state), this phenomena becomes apparent and reflects the abrogation of controlled checkpoint activation and cell cycle arrest.

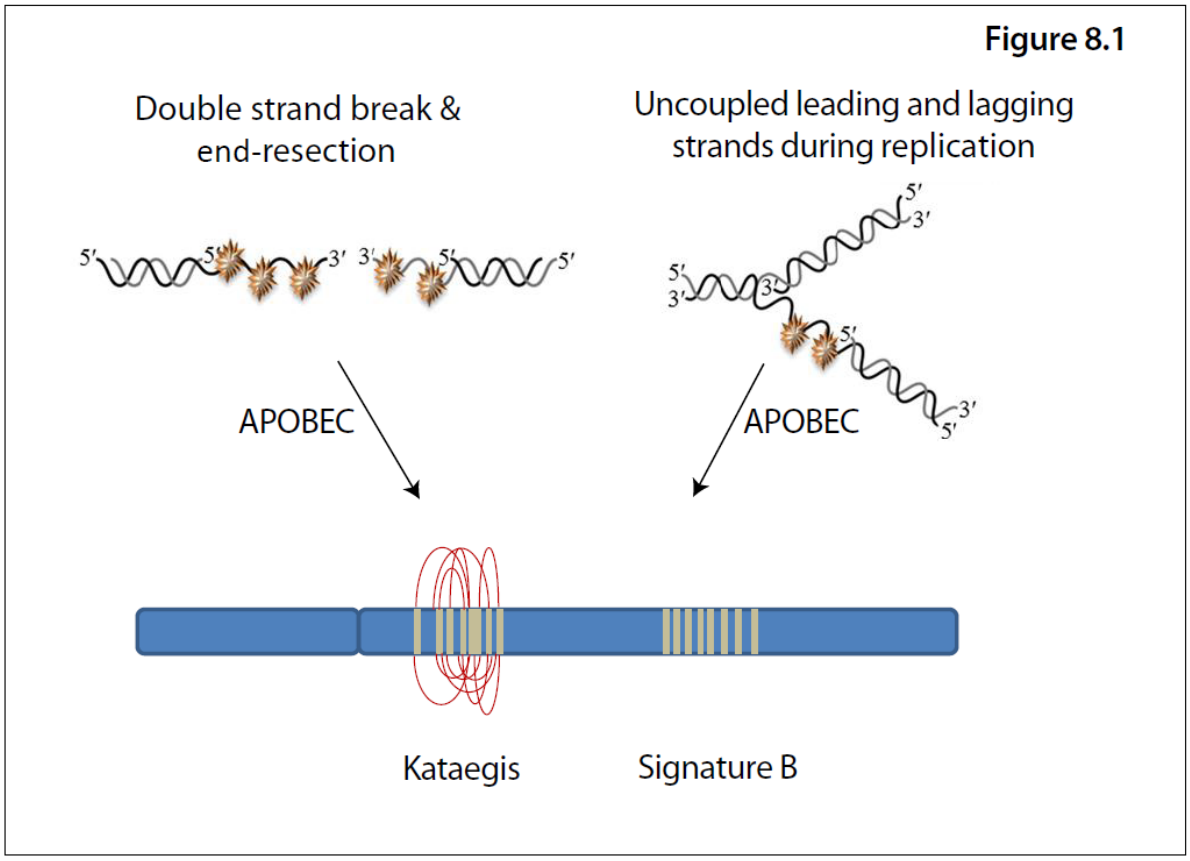


Figure 8.1: Models for APOBEC activity in the genesis of kataegis and Signature B

8.3.7 Future directions for delineating the role of APOBECs in cancer

In the first instance, it would be interesting to attempt to recapitulate Signature B and kataegis by enforced over-expression of cytidine deaminases and identify the most plausible APOBEC candidate responsible in an experimental system. Subsequent exploration could include how different experimental backgrounds affect the mutational signatures. For example, cytosine deamination to uracil invokes uracil-N-glycosylase (Ung) activity of the base excision repair pathway (BER). Induced over-expression of APOBECs on an Ung -/- background may generate more mutations, given the lack of repair via BER, and may change the overall mutation signature. Additionally, given the marked co-localisation of base substitution hypermutation with rearrangements in the kataegis observed in the primary breast cancers, but seemingly stochastic nature of the occurrences, the mechanistic relationship between APOBECs and structural variation can be explored with experiments involving targeted double-strand break induction.

8.4 FINAL SUMMARY

The set of somatic mutations observed in a cancer genome is the aggregate outcome of the activity of one or more biological processes that have been operative over the lifetime of a patient. Each of these biological processes can be characterised by the pattern of mutations that it leaves on the cancer genome. The pattern of mutations or mutational signature characterising each process will be determined both by the underlying mechanisms of DNA damage and of DNA repair that constitute the biological process. The final catalogue of somatic mutations observed in a cancer genome will thus be determined by the strength and duration of exposure to each of the biological processes that have been operative in that cancer.

In this thesis, the aim was to extract the mutational signatures characterising the biological processes that have been operative in the 21 breast cancers studied. Catalogues of somatic mutation of all classes of mutation from twenty-one whole-genome sequenced breast cancers were generated using an integrated suite of bioinformatic algorithms. Mathematical methods were applied in order to extract features of the underlying mutational signatures. Multiple distinct single-nucleotide substitution and their relative contribution to each cancer genome, double-nucleotide substitution and insertion/deletion signatures, were discernible. Integration of copy number information with substitution data revealed how temporal variation in mutational processes can be determined through the development of a cancer. Integration of substitution and expression data revealed transcription-related mutational processes. All these different signatures were compared to other known, curated mutational signatures and the potential biological sources of these processes were postulated. In addition, other distinctive phenomena such as localised hypermutation have been unearthed by analyses of breast cancer genomes at this scale. Furthermore, profound biological insights can be gleaned from the detailed and integrated analyses that have been performed here.

This study harnesses the full scale of whole-genome sequencing technology providing insights into hitherto unrecognised mutational signatures present in breast cancer genomes. It is the first of its kind and demonstrates the wealth of biological information that is hidden within these large datasets.

BIBLIOGRAPHY

Abnizova, I., Leonard, S., Skelly, T., Brown, A., Jackson, D., Gourtovaia, M., Qi, G., Te Boekhorst, R., Faruque, N., Lewis, K.*, et al.* (2012). Analysis of context-dependent errors for illumina sequencing. J Bioinform Comput Biol *10*, 1241005.

Aboussekhra, A., Biggerstaff, M., Shivji, M.K., Vilpo, J.A., Moncollin, V., Podust, V.N., Protic, M., Hubscher, U., Egly, J.M., and Wood, R.D. (1995). Mammalian DNA nucleotide excision repair reconstituted with purified protein components. Cell *80*, 859-868.

Al-Kuraya, K., Schraml, P., Torhorst, J., Tapia, C., Zaharieva, B., Novotny, H., Spichtin, H., Maurer, R., Mirlacher, M., Kochli, O.*, et al.* (2004). Prognostic relevance of gene amplifications and coamplifications in breast cancer. Cancer research *64*, 8534-8540.

Antoniou, A., Pharoah, P.D., Narod, S., Risch, H.A., Eyfjord, J.E., Hopper, J.L., Loman, N., Olsson, H., Johannsson, O., Borg, A.*, et al.* (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet *72*, 1117-1130.

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

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

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M., Lawrence, M.S., Sivachenko, A.Y., Sougnez, C., Zou, L.*, et al.* (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. Nature *486*, 405-409.

Barnes, D.E., Lindahl, T., and Sedgwick, B. (1993). DNA repair. Curr Opin Cell Biol *5*, 424-433.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R.*, et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53-59.

Bergamaschi, A., Kim, Y.H., Wang, P., Sorlie, T., Hernandez-Boussard, T., Lonning, P.E., Tibshirani, R., Borresen-Dale, A.L., and Pollack, J.R. (2006). Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. Genes Chromosomes Cancer *45*, 1033-1040.

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

Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., and Plemmons, R.J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. Comput Stat Data An *52*, 155-173.

Bignell, G.R., Barfoot, R., Seal, S., Collins, N., Warren, W., and Stratton, M.R. (1998). Low frequency of somatic mutations in the LKB1/Peutz-Jeghers syndrome gene in sporadic breast cancer. Cancer research *58*, 1384-1386.

Bignell, G.R., Greenman, C.D., Davies, H., Butler, A.P., Edkins, S., Andrews, J.M., Buck, G., Chen, L., Beare, D., Latimer, C.*, et al.* (2010). Signatures of mutation and selection in the cancer genome. Nature *463*, 893-898.

Bignell, G.R., Santarius, T., Pole, J.C., Butler, A.P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S.*, et al.* (2007). Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome research *17*, 1296-1303.

Birch, J.M., Alston, R.D., McNally, R.J., Evans, D.G., Kelsey, A.M., Harris, M., Eden, O.B., and Varley, J.M. (2001). Relative frequency and morphology of cancers in carriers of germline TP53 mutations. Oncogene *20*, 4621-4628.

Bishop, J.M. (1985). Viral oncogenes. Cell *42*, 23-38.

Bishop, W.R., and Bell, R.M. (1985). Assembly of the endoplasmic reticulum phospholipid bilayer: the phosphatidylcholine transporter. Cell *42*, 51-60.

Bjursell, G., Gussander, E., and Lindahl, T. (1979). Long regions of single-stranded DNA in human cells. Nature *280*, 420-423.
Bonner, W.M., Redon, C.E., Dickey, J.S., Nakamura, A.J., Sedelnikova, O.A., Solier, S., and Pommier, Y. (2008). GammaH2AX and cancer. Nat Rev Cancer *8*, 957-967.
Brown, J.R., and Thornton, J.L. (1957). Percivall Pott (1714-1788) and chimney sweepers' cancer of the scrotum. Br J Ind Med *14*, 68-70.
Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences of the United States of America *101*, 4164-4169.
Butlin, H.T. (1892). Three Lectures on Cancer of the Scrotum in Chimney-Sweeps and Others: Delivered at the Royal College of Surgeons of England. Br Med J *2*, 66-71.
Cadet, J., Ravanat, J.L., Tavernaporro, M., Menoni, H., and Angelov, D. (2012). Oxidatively generated complex DNA damage: Tandem and clustered lesions. Cancer Lett.
Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C*., et al.* (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nature genetics *40*, 722-729.
Carpten, J.D., Faber, A.L., Horn, C., Donoho, G.P., Briggs, S.L., Robbins, C.M., Hostetter, G., Boguslawski, S., Moses, T.Y., Savage, S*., et al.* (2007). A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. Nature *448*, 439-444.
Chaires, J.B. (1990). Biophysical chemistry of the daunomycin-DNA interaction. Biophys Chem *35*, 191-202.
Chaires, J.B. (1998). Drug--DNA interactions. Curr Opin Struct Biol *8*, 314-320.
Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.P., Ahmann, G.J., Adli, M*., et al.* (2011). Initial genome sequencing and analysis of multiple myeloma. Nature *471*, 467-472.
Chen, J., Lindblom, P., and Lindblom, A. (1998). A study of the PTEN/MMAC1 gene in 136 breast cancer families. Hum Genet *102*, 124-125.
Chen, J.M., Ferec, C., and Cooper, D.N. (2012a). Transient hypermutability, chromothripsis and replication-based mechanisms in the generation of concurrent clustered mutations. Mutation research *750*, 52-59.
Chen, S., Qiu, J., Chen, C., Liu, C., Liu, Y., An, L., Jia, J., Tang, J., Wu, L., and Hang, H. (2012b). Affinity maturation of anti-TNF-alpha scFv with somatic hypermutation in non-B cells. Protein Cell.
Chin, K., DeVries, S., Fridlyand, J., Spellman, P.T., Roydasgupta, R., Kuo, W.L., Lapuk, A., Neve, R.M., Qian, Z., Ryder, T*., et al.* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell *10*, 529-541.
Chin, S.F., Teschendorff, A.E., Marioni, J.C., Wang, Y., Barbosa-Morais, N.L., Thorne, N.P., Costa, J.L., Pinder, S.E., van de Wiel, M.A., Green, A.R*., et al.* (2007). High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol *8*, R215.
Ching, H.C., Naidu, R., Seong, M.K., Har, Y.C., and Taib, N.A. (2011). Integrated analysis of copy number and loss of heterozygosity in primary breast carcinomas using high-density SNP array. Int J Oncol *39*, 621-633.
Chiou, C.C., and Yang, J.L. (1995). Mutagenicity and specific mutation spectrum induced by 8-methoxypsoralen plus a low dose of UVA in the hprt gene in diploid human fibroblasts. Carcinogenesis *16*, 1357-1362.
Conticello, S.G. (2008). The AID/APOBEC family of nucleic acid mutators. Genome Biol *9*, 229.
Cox, A., Dunning, A.M., Garcia-Closas, M., Balasubramanian, S., Reed, M.W., Pooley, K.A., Scollen, S., Baynes, C., Ponder, B.A., Chanock, S*., et al.* (2007). A common coding variant in CASP8 is associated with breast cancer risk. Nature genetics *39*, 352-358.
Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y*., et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346-352.
Daley, J.M., and Wilson, T.E. (2005). Rejoining of DNA double-strand breaks as a function of overhang length. Molecular and cellular biology *25*, 896-906.

Deem, A., Keszthelyi, A., Blackgrove, T., Vayl, A., Coffey, B., Mathur, R., Chabes, A., and Malkova, A. (2011). Break-induced replication is highly inaccurate. PLoS biology *9*, e1000594.
DeMarini, D.M., Landi, S., Tian, D., Hanley, N.M., Li, X., Hu, F., Roop, B.C., Mass, M.J., Keohavong, P., Gao, W.*, et al.* (2001). Lung tumor KRAS and TP53 mutations in nonsmokers reflect exposure to PAH-rich coal combustion emissions. Cancer research *61*, 6679-6681.
Demple, B., and DeMott, M.S. (2002). Dynamics and diversions in base excision DNA repair of oxidized abasic lesions. Oncogene *21*, 8926-8934.
Denissenko, M.F., Pao, A., Tang, M., and Pfeifer, G.P. (1996). Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. Science *274*, 430-432.
Denny, W.A. (2001). DNA minor groove alkylating agents. Curr Med Chem *8*, 533-544.
Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D.*, et al.* (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506-510.
Doane, A.S., Danso, M., Lal, P., Donaton, M., Zhang, L., Hudis, C., and Gerald, W.L. (2006). An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. Oncogene *25*, 3994-4008.
Duret, L., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America *96*, 4482-4487.
Duval, A., and Hamelin, R. (2002). Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer research *62*, 2447-2454.
Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R.*, et al.* (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature *447*, 1087-1093.
Eckert, K.A., and Hile, S.E. (2009). Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. Mol Carcinog *48*, 379-388.
Eisenbrand, G., Muller, N., Denkel, E., and Sterzel, W. (1986). DNA adducts and DNA damage by antineoplastic and carcinogenic N-nitrosocompounds. J Cancer Res Clin Oncol *112*, 196-204.
El-Bayoumy, K., Chae, Y.H., Rosa, J.G., Williams, L.K., Desai, D., Amin, S., and Fiala, E. (2000). The effects of 1-nitropyrene, 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine and 7,12-dimethylbenz[a]anthracene on 8-hydroxy-2'-deoxyguanosine levels in the rat mammary gland and modulation by dietary 1,4-phenylenebis(methylene) selenocyanate. Cancer Lett *151*, 7-13.
Ellis, M.J., Ding, L., Shen, D., Luo, J., Suman, V.J., Wallis, J.W., Van Tine, B.A., Hoog, J., Goiffon, R.J., Goldstein, T.C.*, et al.* (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature *486*, 353-360.
Eot-Houllier, G., Eon-Marchais, S., Gasparutto, D., and Sage, E. (2005). Processing of a complex multiply damaged DNA site by human cell extracts and purified repair proteins. Nucleic Acids Res *33*, 260-271.
Evans, M.D., Dizdaroglu, M., and Cooke, M.S. (2004). Oxidative DNA damage and disease: induction, repair and significance. Mutation research *567*, 1-61.
Fang, M., Toher, J., Morgan, M., Davison, J., Tannenbaum, S., and Claffey, K. (2011). Genomic differences between estrogen receptor (ER)-positive and ER-negative human breast carcinoma identified by single nucleotide polymorphism array comparative genome hybridization analysis. Cancer *117*, 2024-2034.
Farber, E., and Cameron, R. (1980). The sequential analysis of cancer development. Adv Cancer Res *31*, 125-226.
Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., Macgrogan, G., Bergh, J., Cameron, D., Goldstein, D.*, et al.* (2005). Identification of molecular apocrine breast tumours by microarray analysis. Oncogene *24*, 4660-4671.
Fearon, E.R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*, 759-767.

Feldmeyer, N., Schmeiser, H.H., Muehlbauer, K.R., Belharazem, D., Knyazev, Y., Nedelko, T., and Hollstein, M. (2006). Further studies with a cell immortalization assay to investigate the mutation signature of aristolochic acid in human p53 sequences. Mutation research *608*, 163-168.
Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., and Parkin, D.M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International journal of cancer Journal international du cancer *127*, 2893-2917.
Fong, P.C., Boss, D.S., Yap, T.A., Tutt, A., Wu, P., Mergui-Roelvink, M., Mortimer, P., Swaisland, H., Lau, A., O'Connor, M.J.*, et al.* (2009). Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. The New England journal of medicine *361*, 123-134.
Foulds, L. (1958). The natural history of cancer. J Chronic Dis *8*, 2-37.
Fresco, J.R., and Alberts, B.M. (1960). The Accommodation of Noncomplementary Bases in Helical Polyribonucleotides and Deoxyribonucleic Acids. Proceedings of the National Academy of Sciences of the United States of America *46*, 311-321.
Frosina, G., Fortini, P., Rossi, O., Carrozzino, F., Raspaglio, G., Cox, L.S., Lane, D.P., Abbondandolo, A., and Dogliotti, E. (1996). Two pathways for base excision repair in mammalian cells. J Biol Chem *271*, 9573-9578.
Giglia-Mari, G., and Sarasin, A. (2003). TP53 mutations in human skin cancers. Hum Mutat *21*, 217-228.
Gouy, M., and Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res *10*, 7055-7074.
Greene, C.N., and Jinks-Robertson, S. (1997). Frameshift intermediates in homopolymer runs are removed efficiently by yeast mismatch repair proteins. Molecular and cellular biology *17*, 2844-2850.
Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C.*, et al.* (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153-158.
Greinert, R., Volkmer, B., Henning, S., Breitbart, E.W., Greulich, K.O., Cardoso, M.C., and Rapp, A. (2012). UVA-induced DNA double-strand breaks result from the repair of clustered oxidative DNA damages. Nucleic Acids Res *40*, 10263-10273.
Hahn, W.C., and Weinberg, R.A. (2002). Rules for making human tumor cells. The New England journal of medicine *347*, 1593-1603.
Hainaut, P., and Pfeifer, G.P. (2001). Patterns of p53 G-->T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. Carcinogenesis *22*, 367-374.
Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57-70.
Hanawalt, P.C., and Spivak, G. (2008). Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev Mol Cell Biol *9*, 958-970.
Harris, R.S., Petersen-Mahrt, S.K., and Neuberger, M.S. (2002). RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. Molecular cell *10*, 1247-1253.
Hartley, J.A., Lown, J.W., Mattes, W.B., and Kohn, K.W. (1988). DNA sequence specificity of antitumor agents. Oncogenes as possible targets for cancer therapy. Acta Oncol *27*, 503-510.
Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nat Rev Genet *10*, 551-564.
Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P.*, et al.* (2001). Gene-expression profiles in hereditary breast cancer. The New England journal of medicine *344*, 539-548.
Hefferin, M.L., and Tomkinson, A.E. (2005). Mechanism of DNA double-strand break repair by non-homologous end joining. DNA Repair (Amst) *4*, 639-648.
Heikkinen, K., Rapakko, K., Karppinen, S.M., Erkko, H., Knuutila, S., Lundan, T., Mannermaa, A., Borresen-Dale, A.L., Borg, A., Barkardottir, R.B.*, et al.* (2006). RAD50 and NBS1 are breast cancer susceptibility genes associated with genomic instability. Carcinogenesis *27*, 1593-1599.
Hicks, J., Krasnitz, A., Lakshmi, B., Navin, N.E., Riggs, M., Leibu, E., Esposito, D., Alexander, J., Troge, J., Grubor, V.*, et al.* (2006). Novel patterns of genome rearrangement and their association with survival in breast cancer. Genome research *16*, 1465-1479.

Hicks, J., Muthuswamy, L., Krasnitz, A., Navin, N., Riggs, M., Grubor, V., Esposito, D., Alexander, J., Troge, J., Wigler, M.*, et al.* (2005). High-resolution ROMA CGH and FISH analysis of aneuploid and diploid breast tumors. Cold Spring Harb Symp Quant Biol *70*, 51-63.
Hicks, W.M., Kim, M., and Haber, J.E. (2010). Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. Science *329*, 82-85.
Hofr, C., Farrell, N., and Brabec, V. (2001). Thermodynamic properties of duplex DNA containing a site-specific d(GpG) intrastrand crosslink formed by an antitumor dinuclear platinum complex. Nucleic Acids Res *29*, 2034-2040.
Hori, M., Suzuki, T., Minakawa, N., Matsuda, A., Harashima, H., and Kamiya, H. (2011). Mutagenicity of secondary oxidation products of 8-oxo-7,8-dihydro-2'-deoxyguanosine 5'-triphosphate (8-hydroxy-2'- deoxyguanosine 5'-triphosphate). Mutation research *714*, 11-16.
Hultquist, J.F., Lengyel, J.A., Refsland, E.W., LaRue, R.S., Lackey, L., Brown, W.L., and Harris, R.S. (2011). Human and rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H demonstrate a conserved capacity to restrict Vif-deficient HIV-1. Journal of virology *85*, 11220-11234.
Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A.*, et al.* (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature genetics *39*, 870-874.
Hwang, D.G., and Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proceedings of the National Academy of Sciences of the United States of America *101*, 13994-14001.
Ionov, Y., Peinado, M.A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature *363*, 558-561.
Jiricny, J. (2006). The multifaceted mismatch-repair system. Nat Rev Mol Cell Biol *7*, 335-346.
Jones, S., Zhang, X., Parsons, D.W., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A.*, et al.* (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science (New York, NY *321*, 1801-1806.
Jovanic, T., Roche, B., Attal-Bonnefoy, G., Leclercq, O., and Rougeon, F. (2008). Ectopic expression of AID in a non-B cell line triggers A:T and G:C point mutations in non-replicating episomal vectors. PloS one *3*, e1480.
Kadyrov, F.A., Dzantiev, L., Constantin, N., and Modrich, P. (2006). Endonucleolytic function of MutLalpha in human mismatch repair. Cell *126*, 297-308.
Kallioniemi, A. (2008). CGH microarrays and cancer. Curr Opin Biotechnol *19*, 36-40.
Karran, P., and Lindahl, T. (1980). Hypoxanthine in deoxyribonucleic acid: generation by heat-induced hydrolysis of adenine residues and release in free form by a deoxyribonucleic acid glycosylase from calf thymus. Biochemistry *19*, 6005-6011.
Katzen, A.L., Kornberg, T.B., and Bishop, J.M. (1985). Isolation of the proto-oncogene c-myb from D. melanogaster. Cell *41*, 449-456.
Key, T.J., Verkasalo, P.K., and Banks, E. (2001). Epidemiology of breast cancer. Lancet Oncol *2*, 133-140.
King, C.R., Kraus, M.H., and Aaronson, S.A. (1985). Amplification of a novel v-erbB-related gene in a human mammary carcinoma. Science *229*, 974-976.
Kinzler, K.W., and Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. Cell *87*, 159-170.
Klarer, A.C., and McGregor, W. (2011). Replication of damaged genomes. Crit Rev Eukaryot Gene Expr *21*, 323-336.
Knobel, P.A., and Marti, T.M. (2011). Translesion DNA synthesis in the context of cancer research. Cancer Cell Int *11*, 39.
Knox, R.J., Lydall, D.A., Friedlos, F., Basham, C., and Roberts, J.J. (1987). The effect of monofunctional or difunctional platinum adducts and of various other associated DNA damage on the expression of transfected DNA in mammalian cell lines sensitive or resistant to difunctional agents. Biochim Biophys Acta *908*, 214-223.
Knuutila, S., Autio, K., and Aalto, Y. (2000). Online access to CGH data of DNA sequence copy number changes. Am J Pathol *157*, 689.

Kunkel, T.A., and Erie, D.A. (2005). DNA mismatch repair. Annu Rev Biochem *74*, 681-710.
Kunz, B.A., Straffon, A.F., and Vonarx, E.J. (2000). DNA damage-induced mutation: tolerance via translesion synthesis. Mutation research *451*, 169-185.
Kuraguchi, M., Edelmann, W., Yang, K., Lipkin, M., Kucherlapati, R., and Brown, A.M. (2000). Tumor-associated Apc mutations in Mlh1-/- Apc1638N mice reveal a mutational signature of Mlh1 deficiency. Oncogene *19*, 5755-5763.
Landry, S., Narvaiza, I., Linfesty, D.C., and Weitzman, M.D. (2011). APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. EMBO reports *12*, 444-450.
Laquerbe, A., Guillouf, C., Moustacchi, E., and Papadopoulo, D. (1995). The mutagenic processing of psoralen photolesions leaves a highly specific signature at an endogenous human locus. J Mol Biol *254*, 38-49.
Leary, R.J., Lin, J.C., Cummins, J., Boca, S., Wood, L.D., Parsons, D.W., Jones, S., Sjoblom, T., Park, B.H., Parsons, R.*, et al.* (2008). Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. Proceedings of the National Academy of Sciences of the United States of America *105*, 16224-16229.
Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788-791.
Lee, J.A., Carvalho, C.M., and Lupski, J.R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell *131*, 1235-1247.
Lee, W., Jiang, Z., Liu, J., Haverty, P.M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K.P., Bhatt, D.*, et al.* (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature *465*, 473-477.
Liao, W., Hong, S.H., Chan, B.H., Rudolph, F.B., Clark, S.C., and Chan, L. (1999). APOBEC-2, a cardiac- and skeletal muscle-specific member of the cytidine deaminase supergene family. Biochem Biophys Res Commun *260*, 398-404.
Lindahl, T. (1993). Instability and decay of the primary structure of DNA. Nature *362*, 709-715.
Litman, R., Peng, M., Jin, Z., Zhang, F., Zhang, J., Powell, S., Andreassen, P.R., and Cantor, S.B. (2005). BACH1 is critical for homologous recombination and appears to be the Fanconi anemia gene product FANCJ. Cancer Cell *8*, 255-265.
Longerich, S., Basu, U., Alt, F., and Storb, U. (2006). AID in somatic hypermutation and class switch recombination. Current opinion in immunology *18*, 164-174.
Lutsenko, E., and Bhagwat, A.S. (1999). Principal causes of hot spots for cytosine to thymine mutations at sites of cytosine methylation in growing cells. A model, its experimental support and implications. Mutation research *437*, 11-20.
Lynch, H.T., and de la Chapelle, A. (1999). Genetic susceptibility to non-polyposis colorectal cancer. J Med Genet *36*, 801-818.
Ma, W., Panduri, V., Sterling, J.F., Van Houten, B., Gordenin, D.A., and Resnick, M.A. (2009). The transition of closely opposed lesions to double-strand breaks during long-patch base excision repair is prevented by the coordinated action of DNA polymerase delta and Rad27/Fen1. Molecular and cellular biology *29*, 1212-1221.
Mace, K., Aguilar, F., Wang, J.S., Vautravers, P., Gomez-Lechon, M., Gonzalez, F.J., Groopman, J., Harris, C.C., and Pfeifer, A.M. (1997). Aflatoxin B1-induced DNA adduct formation and p53 mutations in CYP450-expressing human liver cell lines. Carcinogenesis *18*, 1291-1297.
Masciari, S., Larsson, N., Senz, J., Boyd, N., Kaurah, P., Kandel, M.J., Harris, L.N., Pinheiro, H.C., Troussard, A., Miron, P.*, et al.* (2007). Germline E-cadherin mutations in familial lobular breast cancer. J Med Genet *44*, 726-731.
McCulloch, S.D., and Kunkel, T.A. (2008). The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell Res *18*, 148-161.
Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., Hollestelle, A., Houben, M., Crepin, E., van Veghel-Plandsoen, M.*, et al.* (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nature genetics *31*, 55-59.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W.*, et al.* (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science *266*, 66-71.
Modrich, P., and Lahue, R. (1996). Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annu Rev Biochem *65*, 101-133.
Nassif, N., Penney, J., Pal, S., Engels, W.R., and Gloor, G.B. (1994). Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. Molecular and cellular biology *14*, 1613-1625.
Nedelko, T., Arlt, V.M., Phillips, D.H., and Hollstein, M. (2009). TP53 mutation signature supports involvement of aristolochic acid in the aetiology of endemic nephropathy-associated tumours. International journal of cancer Journal international du cancer *124*, 987-990.
Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E.*, et al.* (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272-276.
Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A.*, et al.* (2012a). Mutational Processes Molding the Genomes of 21 Breast Cancers. Cell *149*, 979-993.
Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M.*, et al.* (2012b). The life history of 21 breast cancers. Cell *149*, 994-1007.
Nouspikel, T. (2009). DNA repair in mammalian cells : Nucleotide excision repair: variations on versatility. Cell Mol Life Sci *66*, 994-1009.
Nouspikel, T., and Hanawalt, P.C. (2000). Terminally differentiated human neurons repair transcribed genes but display attenuated global DNA repair and modulation of repair gene expression. Molecular and cellular biology *20*, 1562-1570.
Nouspikel, T.P., Hyka-Nouspikel, N., and Hanawalt, P.C. (2006). Transcription domain-associated repair in human cells. Molecular and cellular biology *26*, 8722-8730.
Nowarski, R., Wilner, O.I., Cheshin, O., Shahar, O.D., Kenig, E., Baraz, L., Britan-Rosich, E., Nagler, A., Harris, R.S., Goldberg, M.*, et al.* (2012). APOBEC3G enhances lymphoma cell radioresistance by promoting cytidine deaminase-dependent DNA repair. Blood *120*, 366-375.
Nussenzweig, A., and Nussenzweig, M.C. (2010). Origin of chromosomal translocations in lymphoid cancer. Cell *141*, 27-38.
Oikawa, S., and Kawanishi, S. (1999). Site-specific DNA damage at GGG sequence by oxidative stress may accelerate telomere shortening. FEBS Lett *453*, 365-368.
Oikawa, S., Tada-Oikawa, S., and Kawanishi, S. (2001). Site-specific DNA damage at the GGG sequence by UVA involves acceleration of telomere shortening. Biochemistry *40*, 4763-4768.
Okuyama, S., Marusawa, H., Matsumoto, T., Ueda, Y., Matsumoto, Y., Endo, Y., Takai, A., and Chiba, T. (2012). Excessive activity of apolipoprotein B mRNA editing enzyme catalytic polypeptide 2 (APOBEC2) contributes to liver and lung tumorigenesis. International journal of cancer Journal international du cancer *130*, 1294-1301.
Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T.*, et al.* (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England journal of medicine *351*, 2817-2826.
Palacios, J., Robles-Frias, M.J., Castilla, M.A., Lopez-Garcia, M.A., and Benitez, J. (2008). The molecular pathology of hereditary breast cancer. Pathobiology : journal of immunopathology, molecular and cellular biology *75*, 85-94.
Papadopoulo, D., Laquerbe, A., Guillouf, C., and Moustacchi, E. (1993). Molecular spectrum of mutations induced at the HPRT locus by a cross-linking agent in human cell lines with different repair capacities. Mutation research *294*, 167-177.
Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J.S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C.*, et al.* (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. The New England journal of medicine *365*, 1384-1395.

Parker, R.C., Varmus, H.E., and Bishop, J.M. (1984). Expression of v-src and chicken c-src in rat cells demonstrates qualitative differences between pp60v-src and pp60c-src. Cell *37*, 131-139.
Pathania, S., Nguyen, J., Hill, S.J., Scully, R., Adelmant, G.O., Marto, J.A., Feunteun, J., and Livingston, D.M. (2011). BRCA1 is required for postreplication repair after UV-induced DNA damage. Molecular cell *44*, 235-251.
Pena-Diaz, J., and Jiricny, J. (2012). Mammalian mismatch repair: error-free or error-prone? Trends Biochem Sci *37*, 206-214.
Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A*., et al.* (2000). Molecular portraits of human breast tumours. Nature *406*, 747-752.
Perucho, M., Goldfarb, M., Shimizu, K., Lama, C., Fogh, J., and Wigler, M. (1981). Human-tumor-derived cell lines contain common and different transforming genes. Cell *27*, 467-476.
Petit, V., Guetard, D., Renard, M., Keriel, A., Sitbon, M., Wain-Hobson, S., and Vartanian, J.P. (2009). Murine APOBEC1 is a powerful mutator of retroviral and cellular RNA in vitro and in vivo. J Mol Biol *385*, 65-78.
Pfeifer, G.P. (2000). p53 mutational spectra and the role of methylated CpG sequences. Mutation research *450*, 155-166.
Pfeifer, G.P., Denissenko, M.F., Olivier, M., Tretyakova, N., Hecht, S.S., and Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene *21*, 7435-7451.
Pfeifer, G.P., You, Y.H., and Besaratinia, A. (2005). Mutations induced by ultraviolet light. Mutation research *571*, 19-31.
Pham, P., Bransteitter, R., Petruska, J., and Goodman, M.F. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature *424*, 103-107.
Pickrell, J.K., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics *27*, 2144-2146.
Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordonez, G.R., Bignell, G.R*., et al.* (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191-196.
Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C*., et al.* (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature *463*, 184-190.
Polak, P., and Arndt, P.F. (2008). Transcription induces strand-specific mutations at the 5 ' end of human genes. Genome research *18*, 1216-1223.
Prat, A., Parker, J.S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J.I., He, X., and Perou, C.M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast cancer research : BCR *12*, R68.
Puente, X.S., Pinyol, M., Quesada, V., Conde, L., Ordonez, G.R., Villamor, N., Escaramis, G., Jares, P., Bea, S., Gonzalez-Diaz, M*., et al.* (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature *475*, 101-105.
Pulciani, S., Santos, E., Lauver, A.V., Long, L.K., Robbins, K.C., and Barbacid, M. (1982). Oncogenes in human tumor cell lines: molecular cloning of a transforming gene from human bladder carcinoma cells. Proceedings of the National Academy of Sciences of the United States of America *79*, 2845-2849.
Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T*., et al.* (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nature genetics *39*, 165-167.
Reddy, E.P., Reynolds, R.K., Santos, E., and Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. Nature *300*, 149-152.

Reid, S., Schindler, D., Hanenberg, H., Barker, K., Hanks, S., Kalb, R., Neveling, K., Kelly, P., Seal, S., Freund, M.*, et al.* (2007). Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. Nature genetics *39*, 162-164.
Reis-Filho, J.S., and Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet *378*, 1812-1823.
Renan, M.J. (1993). How many mutations are required for tumorigenesis? Implications from human cancer data. Mol Carcinog *7*, 139-146.
Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Van Loo, P., Van Den Bossche, M., Catthoor, K., Sabbe, B.*, et al.* (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nat Biotechnol *30*, 61-68.
Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M.*, et al.* (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science *328*, 636-639.
Roberts, J.D., and Kunkel, T.A. (1988). Fidelity of a human cell DNA replication complex. Proceedings of the National Academy of Sciences of the United States of America *85*, 7064-7068.
Roberts, S.A., Sterling, J., Thompson, C., Harris, S., Mav, D., Shah, R., Klimczak, L.J., Kryukov, G.V., Malc, E., Mieczkowski, P.A.*, et al.* (2012). Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. Molecular cell *46*, 424-435.
Robertson, A.B., Klungland, A., Rognes, T., and Leiros, I. (2009). DNA repair in mammalian cells: Base excision repair: the long and short of it. Cell Mol Life Sci *66*, 981-993.
Rogozin, I.B., Basu, M.K., Jordan, I.K., Pavlov, Y.I., and Koonin, E.V. (2005). APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. Cell Cycle *4*, 1281-1285.
Rowley, J.D. (1973). Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature *243*, 290-293.
Sale, J.E., Lehmann, A.R., and Woodgate, R. (2012). Y-family DNA polymerases and their role in tolerance of cellular DNA damage. Nat Rev Mol Cell Biol *13*, 141-152.
Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., Szabo, S., Yan, H., Gazdar, A., Powell, S.M., Riggins, G.J.*, et al.* (2004). High frequency of mutations of the PIK3CA gene in human cancers. Science *304*, 554.
Satoh, M.S., Jones, C.J., Wood, R.D., and Lindahl, T. (1993). DNA excision-repair defect of xeroderma pigmentosum prevents removal of a class of oxygen free radical-induced base lesions. Proceedings of the National Academy of Sciences of the United States of America *90*, 6335-6339.
Seal, S., Thompson, D., Renwick, A., Elliott, A., Kelly, P., Barfoot, R., Chagtai, T., Jayatilake, H., Ahmed, M., Spanova, K.*, et al.* (2006). Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nature genetics *38*, 1239-1241.
Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G.*, et al.* (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature *486*, 395-399.
Sheehy, A.M., Gaddis, N.C., Choi, J.D., and Malim, M.H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature *418*, 646-650.
Shen, J.C., Rideout, W.M., 3rd, and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res *22*, 972-976.
Shibata, D., Peinado, M.A., Ionov, Y., Malkhosyan, S., and Perucho, M. (1994). Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation. Nature genetics *6*, 273-281.
Shih, C., Padhy, L.C., Murray, M., and Weinberg, R.A. (1981). Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. Nature *290*, 261-264.
Simon, M.A., Drees, B., Kornberg, T., and Bishop, J.M. (1985). The nucleotide sequence and the tissue-specific expression of Drosophila c-src. Cell *42*, 831-840.
Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N.*, et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268-274.

Slupphaug, G., Kavli, B., and Krokan, H.E. (2003). The interacting pathways for prevention and repair of oxidative DNA damage. Mutation research *531*, 231-251.
Smith, C.E., Llorente, B., and Symington, L.S. (2007). Template switching during break-induced replication. Nature *447*, 102-105.
Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S*., et al.* (2001a). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America *98*, 10869-10874.
Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S*., et al.* (2001b). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America *98*, 10869-10874.
Spencer, W.A., Vadhanam, M.V., Jeyabalan, J., and Gupta, R.C. (2011). Oxidative DNA damage following microsome/Cu(II)-mediated activation of the estrogens, 17beta-estradiol, equilenin and equilin: Role of reactive oxygen species. Chemical research in toxicology.
Spencer, W.A., Vadhanam, M.V., Jeyabalan, J., and Gupta, R.C. (2012). Oxidative DNA damage following microsome/Cu(II)-mediated activation of the estrogens, 17beta-estradiol, equilenin, and equilin: role of reactive oxygen species. Chemical research in toxicology *25*, 305-314.
Stacey, S.N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S.A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A*., et al.* (2007). Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nature genetics *39*, 865-869.
Stacey, S.N., Manolescu, A., Sulem, P., Thorlacius, S., Gudjonsson, S.A., Jonsson, G.F., Jakobsdottir, M., Bergthorsson, J.T., Gudmundsson, J., Aben, K.K*., et al.* (2008). Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nature genetics *40*, 703-706.
Staszewski, O., Baker, R.E., Ucher, A.J., Martier, R., Stavnezer, J., and Guikema, J.E. (2011). Activation-induced cytidine deaminase induces reproducible DNA breaks at many non-Ig Loci in activated B cells. Molecular cell *41*, 232-242.
Stenglein, M.D., Burns, M.B., Li, M., Lengyel, J., and Harris, R.S. (2010). APOBEC3 proteins mediate the clearance of foreign DNA from human cells. Nature structural & molecular biology *17*, 222-229.
Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., Bignell, G., Teague, J., Smith, R., Stevens, C*., et al.* (2005). A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nature genetics *37*, 590-592.
Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A*., et al.* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27-40.
Stephens, P.J., McBride, D.J., Lin, M.L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J*., et al.* (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature *462*, 1005-1010.
Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R*., et al.* (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400-404.
Stojic, L., Brun, R., and Jiricny, J. (2004). Mismatch repair and DNA damage signalling. DNA Repair (Amst) *3*, 1091-1101.
Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A*., et al.* (2011). The mutational landscape of head and neck squamous cell carcinoma. Science *333*, 1157-1160.
Strathern, J.N., Shafer, B.K., and McGill, C.B. (1995). DNA synthesis errors associated with double-strand-break repair. Genetics *140*, 965-972.
Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719-724.
Strauss, B.S. (2002). The "A" rule revisited: polymerases as determinants of mutational specificity. DNA Repair (Amst) *1*, 125-135.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., and Inouye, M. (1966). Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday. Cold Spring Harb Symp Quant Biol *31*, 77-84.
Streisinger, G., and Owen, J. (1985). Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. Genetics *109*, 633-659.
Suspene, R., Aynaud, M.M., Guetard, D., Henry, M., Eckhoff, G., Marchio, A., Pineau, P., Dejean, A., Vartanian, J.P., and Wain-Hobson, S. (2011). Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. Proceedings of the National Academy of Sciences of the United States of America *108*, 4858-4863.
Teng, B., Burant, C.F., and Davidson, N.O. (1993). Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science *260*, 1816-1819.
Thibodeau, S.N., Bren, G., and Schaid, D. (1993). Microsatellite instability in cancer of the proximal colon. Science *260*, 816-819.
Thomas, N.E., Alexander, A., Edmiston, S.N., Parrish, E., Millikan, R.C., Berwick, M., Groben, P., Ollila, D.W., Mattingly, D., and Conway, K. (2004). Tandem BRAF mutations in primary invasive melanomas. J Invest Dermatol *122*, 1245-1250.
Thompson, D., Duedal, S., Kirner, J., McGuffog, L., Last, J., Reiman, A., Byrd, P., Taylor, M., and Easton, D.F. (2005). Cancer risks and mortality in heterozygous ATM mutation carriers. J Natl Cancer Inst *97*, 813-822.
Thompson, W.D. (1994). Genetic epidemiology of breast cancer. Cancer *74*, 279-287.
Tiraby, G., Fox, M.S., and Bernheimer, H. (1975). Marker discrimination in deoxyribonucleic acid-mediated transformation of various Pneumococcus strains. J Bacteriol *121*, 608-618.
Tornaletti, S., Maeda, L.S., and Hanawalt, P.C. (2006). Transcription arrest at an abasic site in the transcribed strand of template DNA. Chemical research in toxicology *19*, 1215-1220.
Tran, H.T., Keen, J.D., Kricker, M., Resnick, M.A., and Gordenin, D.A. (1997). Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. Molecular and cellular biology *17*, 2859-2865.
Turnbull, C., and Rahman, N. (2008). Genetic predisposition to breast cancer: past, present, and future. Annu Rev Genomics Hum Genet *9*, 321-345.
Usary, J., Llaca, V., Karaca, G., Presswala, S., Karaca, M., He, X., Langerod, A., Karesen, R., Oh, D.S., Dressler, L.G.*, et al.* (2004). Mutation of GATA3 in human breast tumors. Oncogene *23*, 7669-7678.
van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T.*, et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530-536.
van Gent, D.C., and van der Burg, M. (2007). Non-homologous end-joining, a sticky affair. Oncogene *26*, 7731-7740.
Van Loo, P., Nordgard, S.H., Lingjaerde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B.*, et al.* (2010). Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences of the United States of America *107*, 16910-16915.
Vincent-Salomon, A., Lucchesi, C., Gruel, N., Raynal, V., Pierron, G., Goudefroye, R., Reyal, F., Radvanyi, F., Salmon, R., Thiery, J.P.*, et al.* (2008). Integrated genomic and transcriptomic analysis of ductal carcinoma in situ of the breast. Clinical cancer research : an official journal of the American Association for Cancer Research *14*, 1956-1965.
Voulgaridou, G.P., Anestopoulos, I., Franco, R., Panayiotidis, M.I., and Pappa, A. (2011). DNA damage induced by endogenous aldehydes: current state of knowledge. Mutation research *711*, 13-27.
Vreeswijk, M.P., Overkamp, M.W., Westland, B.E., van Hees-Stuivenberg, S., Vrieling, H., Zdzienicka, M.Z., van Zeeland, A.A., and Mullenders, L.H. (1998). Enhanced UV-induced mutagenesis in the UV61 cell line, the Chinese hamster homologue of Cockayne's syndrome B, is associated with defective transcription coupled repair of cyclobutane pyrimidine dimers. Mutation research *409*, 49-56.
Walker, B., Jr., and Gerber, A. (1981). Occupational exposure to aromatic amines: benzidine and benzidine-based dyes. Natl Cancer Inst Monogr, 11-13.
Wang, D., Kreutzer, D.A., and Essigmann, J.M. (1998). Mutagenicity and repair of oxidative DNA damage: insights from studies using defined lesions. Mutation research *400*, 99-115.

Wang, J., Gonzalez, K.D., Scaringe, W.A., Tsai, K., Liu, N., Gu, D., Li, W., Hill, K.A., and Sommer, S.S. (2007). Evidence for mutation showers. Proceedings of the National Academy of Sciences of the United States of America *104*, 8403-8408.
Wang, Y., Sheppard, T.L., Tornaletti, S., Maeda, L.S., and Hanawalt, P.C. (2006). Transcriptional inhibition by an oxidized abasic site in DNA. Chemical research in toxicology *19*, 234-241.
Waters, T.R., and Swann, P.F. (2000). Thymine-DNA glycosylase and G to A transition mutations at CpG sites. Mutation research *462*, 137-147.
Watson, J.D., and Crick, F.H. (1953a). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.
Watson, J.D., and Crick, F.H. (1953b). The structure of DNA. Cold Spring Harb Symp Quant Biol *18*, 123-131.
Weigelt, B., Geyer, F.C., and Reis-Filho, J.S. (2010). Histological types of breast cancer: how special are they? Mol Oncol *4*, 192-208.
Weinberg, R.A. (1989). Oncogenes, antioncogenes, and the molecular bases of multistep carcinogenesis. Cancer research *49*, 3713-3721.
Weisburger, J.H., Dolan, L., and Pittman, B. (1998). Inhibition of PhIP mutagenicity by caffeine, lycopene, daidzein, and genistein. Mutation research *416*, 125-128.
Weterings, E., and Chen, D.J. (2008). The endless tale of non-homologous end-joining. Cell Res *18*, 114-124.
Witz, I.P., and Levy-Nissenbaum, O. (2006). The tumor microenvironment in the post-PAGET era. Cancer Lett *242*, 1-10.
Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J.*, et al.* (2007). The genomic landscapes of human breast and colorectal cancers. Science *318*, 1108-1113.
Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., and Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. Nature *378*, 789-792.
Wu, L., and Hickson, I.D. (2003). The Bloom's syndrome helicase suppresses crossing over during homologous recombination. Nature *426*, 870-874.
Yamanaka, S., Balestra, M.E., Ferrell, L.D., Fan, J., Arnold, K.S., Taylor, S., Taylor, J.M., and Innerarity, T.L. (1995). Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. Proceedings of the National Academy of Sciences of the United States of America *92*, 8483-8487.
Yang, M. (2011). A current global view of environmental and occupational cancers. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev *29*, 223-249.
Yang, S.C., Lin, J.G., Chiou, C.C., Chen, L.Y., and Yang, J.L. (1994). Mutation specificity of 8-methoxypsoralen plus two doses of UVA irradiation in the hprt gene in diploid human fibroblasts. Carcinogenesis *15*, 201-207.
Yang, Y., Sterling, J., Storici, F., Resnick, M.A., and Gordenin, D.A. (2008). Hypermutability of damaged single-strand DNA formed at double-strand breaks and uncapped telomeres in yeast Saccharomyces cerevisiae. PLoS Genet *4*, e1000264.
Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865-2871.
Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research *18*, 821-829.
Zhu, J., Yue, Z., and Wang, J. (1998). [Oxidative DNA damage of cartilage in osteoarthrosis]. Hunan Yi Ke Da Xue Xue Bao *23*, 438-440.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

(6-4)PPs	6-4 pyrimidine-pyrimidone photoproducts
8-oxo-dG	8-oxo-2'-deoxyguanosine
AID	activation-induced cytidine deaminase
APC	adenomatous polyposis coli
AP site	apurinic/apyrimidinic site
APE1	AP-endonuclease
APOBEC	apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like
ASCAT	allele-specific copy number analysis of tumors
ATM	ataxia-telangiectasia mutated gene
B[a]DPE	benzo[a]pyrene diolepoxide
BIR	break-induced replication
BWA	Burrows-Wheeler Alignment
CaVEMan	Cancer Variants Through Expectation Maximization
CDH1	cadherin1
CGH	comparative genomic hybridisation
CNA	copy number aberration
CPD	cyclobutane pyrimidine dimers
DAR	domain-associated repair
DCIS	ductal carcinoma in-situ
DNA	deoxyribose nucleic acid
DNA-PK	DNA-dependent protein kinase
DSBR	double-strand break repair
ENU	ethyl nitrosurea
ER	oestrogen receptor
FEN1	flap structure-specific endonuclease 1
GGR	global genome repair

GWAS	genome-wide association study
HPRT	hypoxanthine guanine phosphoribosyltransferase 1
HR	homologous recombination
HRT	hormone replacement therapy
ICGC	International Cancer Genome Consortium
IHC	immunohistochemisty
K–S	Kolmogorov-Smirnov test
LCIS	lobular carcinoma in situ
LOH	loss of heterozygosity
MMR	mismatch repair
MMS	methyl methane sulfonate
MNNG	methylnitronitrosoguanidine
MNU	methyl nitrosurea
mRNA	messenger RNA
NER	nucleotide excision repair
NGS	next-generation sequencing technology
NHEJ	non-homologous end-joining
NBS1	Nijmegen breakage syndrome
NMF	non-negative matrix factorization
OCDL	oxidative clustered DNA lesion
OCP	oral contraceptive pill
PAH	polyaromatic hydrocarbon
PALB2	partner and localizer of BRCA2
PCNA	proliferating cell nuclear antigen
PCR	polymerase chain reaction
PhIP	amino-1-methyl-6-phenylimidazo [4,5-b] pyridine
Pol	polymerase
PPV	positive predictive value

PR	progesterone receptor
RFC	replication factor C
RNA	ribonucleic acid
RPM	revolutions per minute
SDSA	synthesis-dependent strand annealing
SNP	single-nucleotide polymorphism
SSA	single-strand annealing
TCR	transcription-coupled repair
TFIIH	transcription-factor IIH
TP53	tumour protein p53
UNG	uracil-N-glycosylase
UV	ultraviolet
XP	xeroderma pigmentosa
XRCC1	X-ray repair, complementing defective, in chinese hamster

APPENDICES (INCLUDED CD)

Appendix 1: Table of substitutions obtained from twenty-one breast cancers. A subset of these mutations have been validated and the status shown in the "current_conf_status" column.

Appendix 2: Table of insertions and deletions from twenty-one breast cancer genomes. All indels presented here have been validated and confirmed as somatic by sequencing on an orthogonal sequencing platform.

Appendix 3: Table of rearrangements. All rearrangements presented here have been validated by capillary sequencing or local reassembly.

Appendix 4: 247 regions of kataegis defined as 6 successive mutations with an intermutation distance less than 1kb.

Appendix 5: Table of copy number segments.

Please see the README.txt file which contains a description of the abbreviations used in the tables.

Image of the cover of the journal (Cell, May 25th 2012) as well as the associated publications, have been included at the end of this thesis.