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).

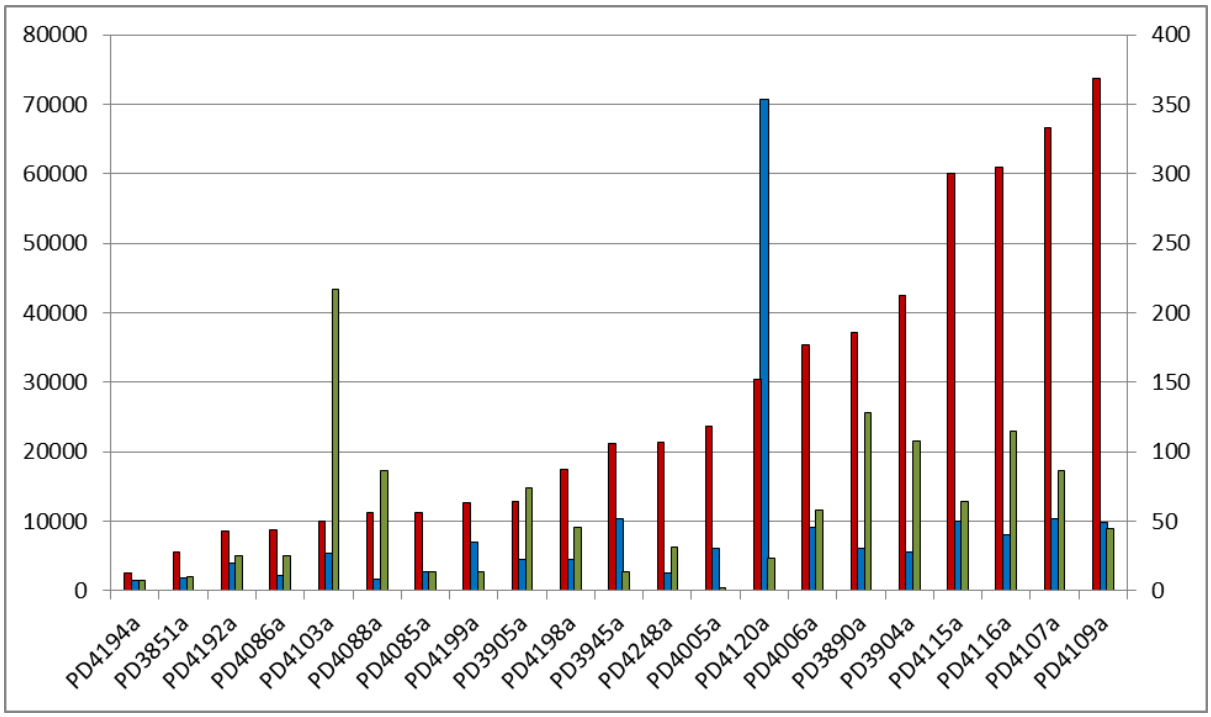


Figure 7.1: Relationship between total number of insertions/deletions and mutation burden of other classes of mutation. Indels (in red) and rearrangements (in green) are scaled to the right-hand vertical axis (total number of indels or rearrangements). Substitutions (in blue) are scaled to the left-hand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

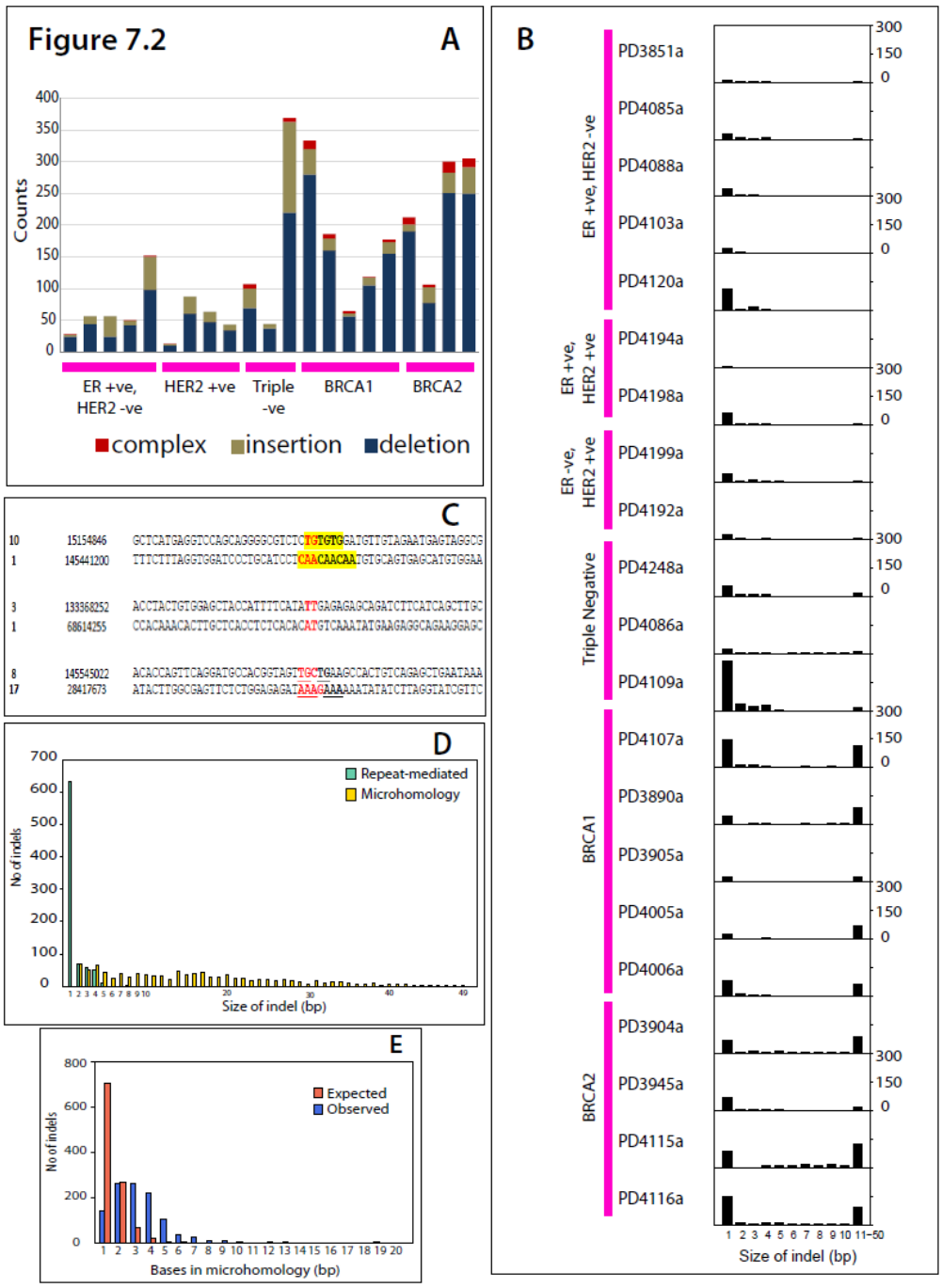


Figure 7.2: Somatic mutation profile of indels. (A) Histogram of number of indels for each breast cancer. (B) The x axis shows indel size from 1-10 and all larger indels between11-50bp in size grouped in a single bin. The y axis shows the number in each genome from 0-300bp. (C) Classification of indel 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

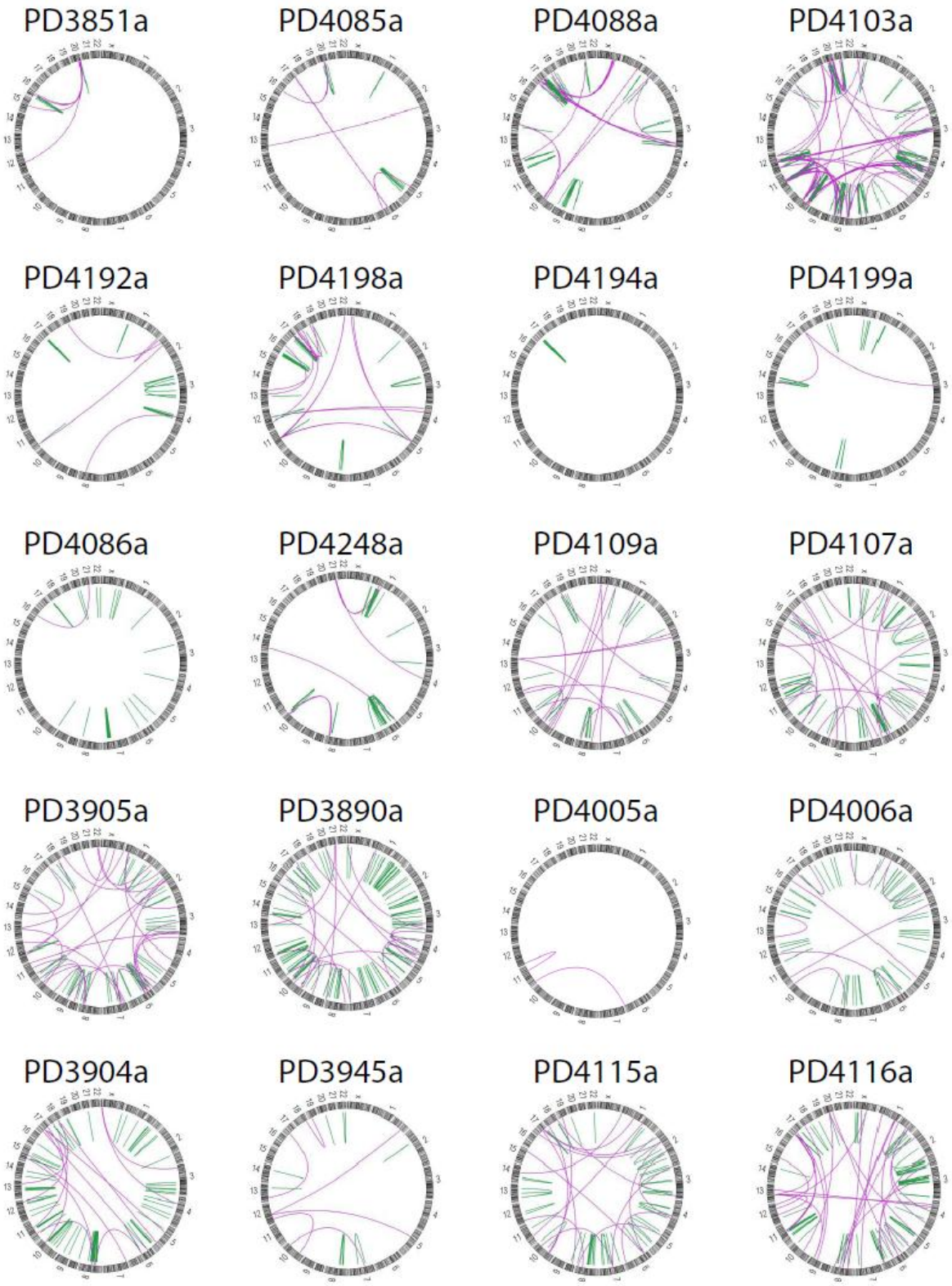


Figure 7.4: Circos plots demonstrating the rearrangements in the 20 breast cancers.

Figure 7.5

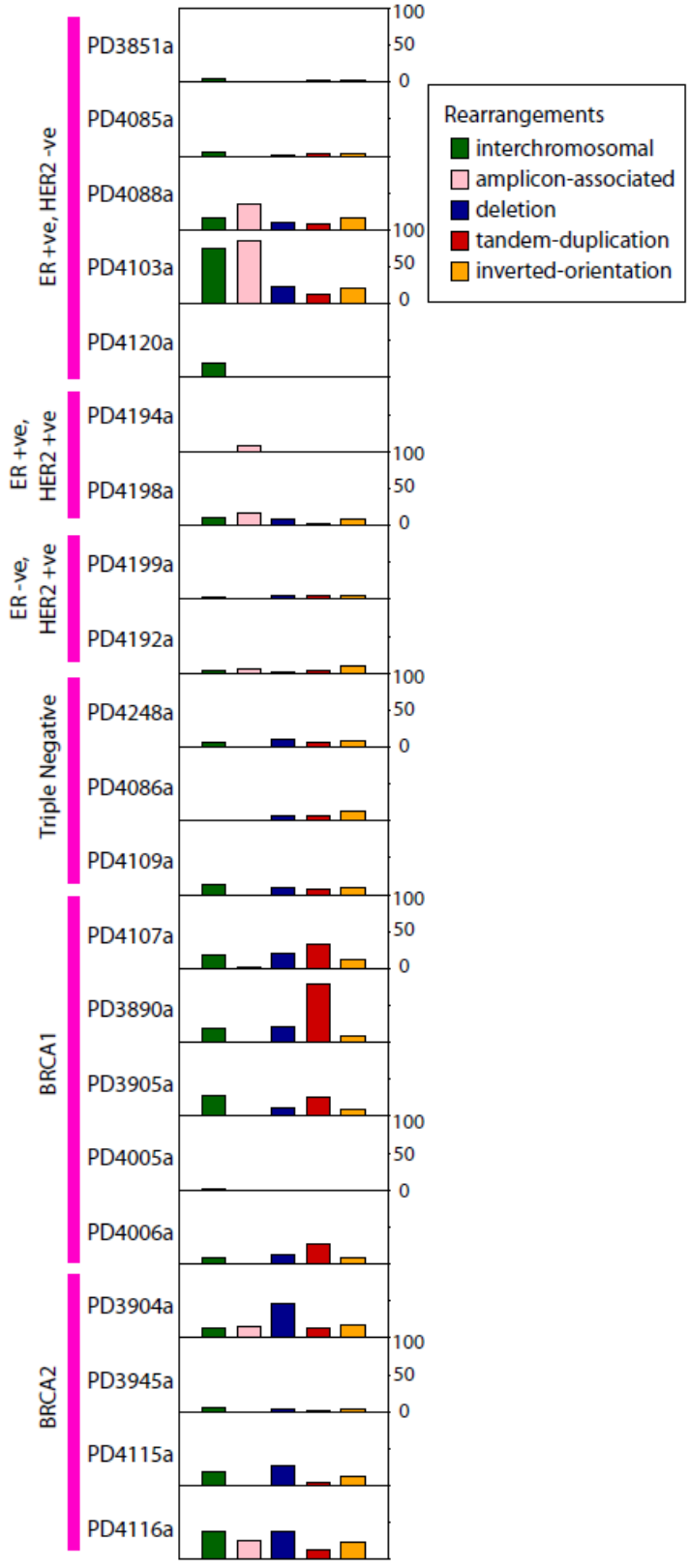


Figure 7.5: Variation in rearrangement architecture between the twenty-one breast cancers

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

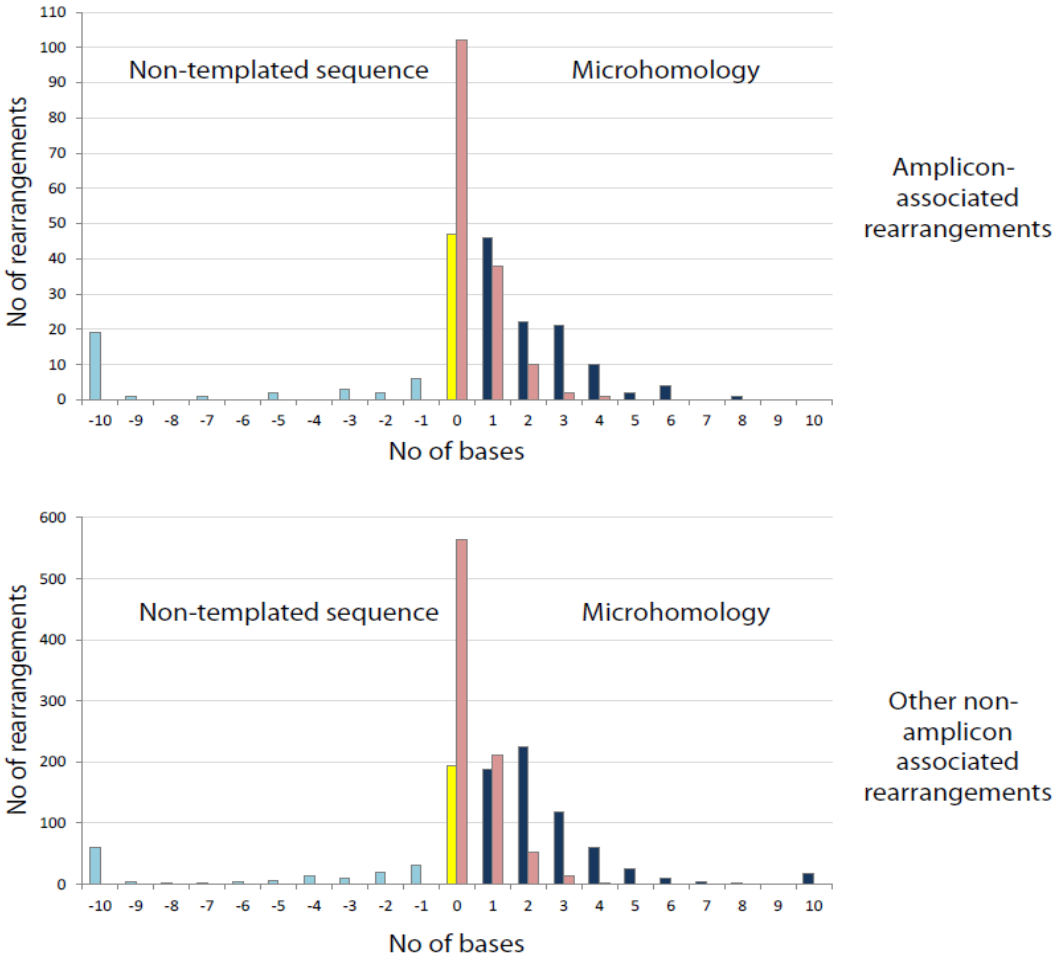


Figure 7.6: Patterns of microhomology (dark blue) and non-templated sequence (light blue) at rearrangement breakpoints of twenty-one breast cancers. The occurrence of microhomology by chance presented in pink. Difference in distribution of number of bases involved in microhomology between observed and 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

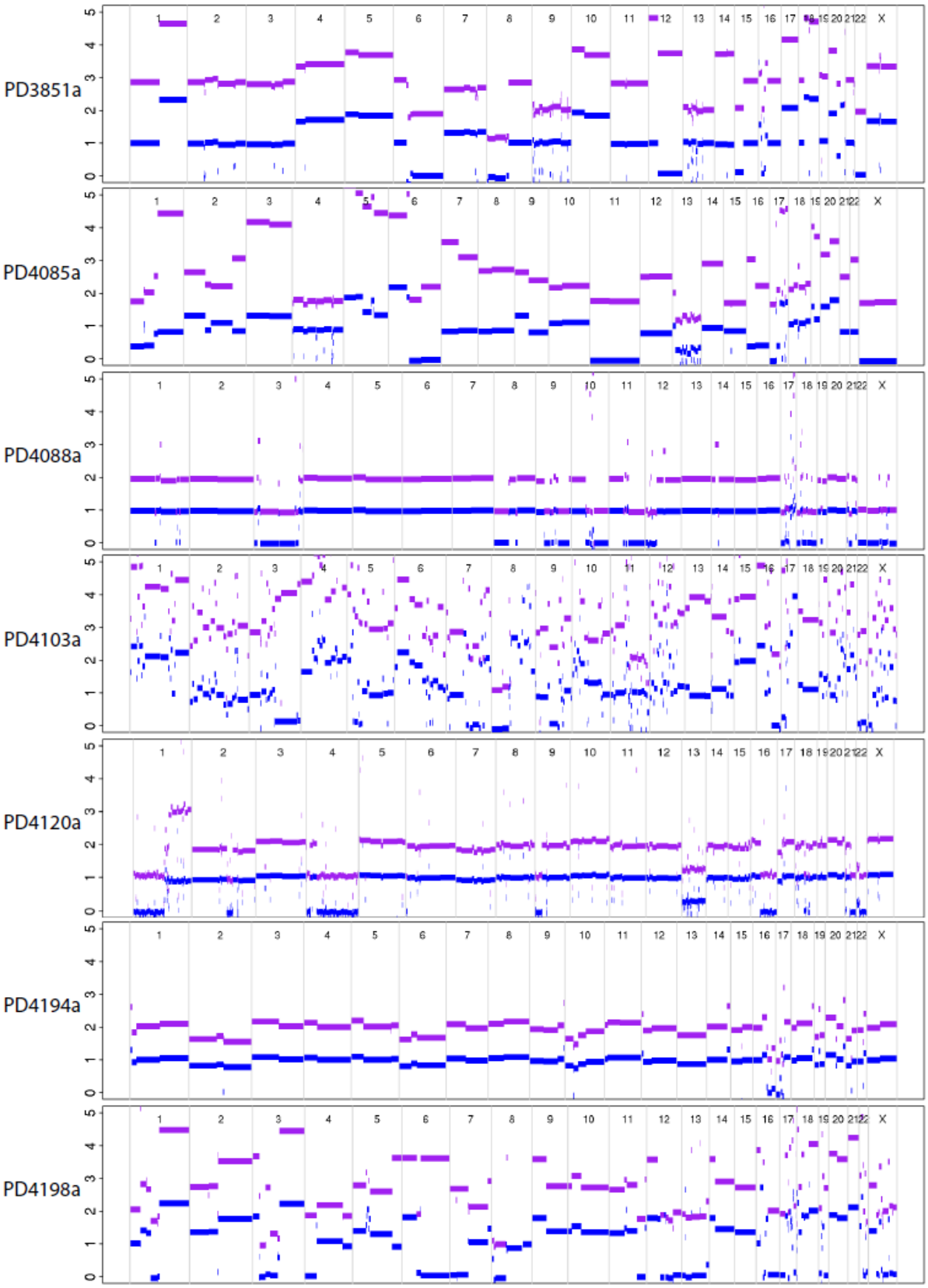
7.5.1 The observed variation in gross copy number changes between breast cancers

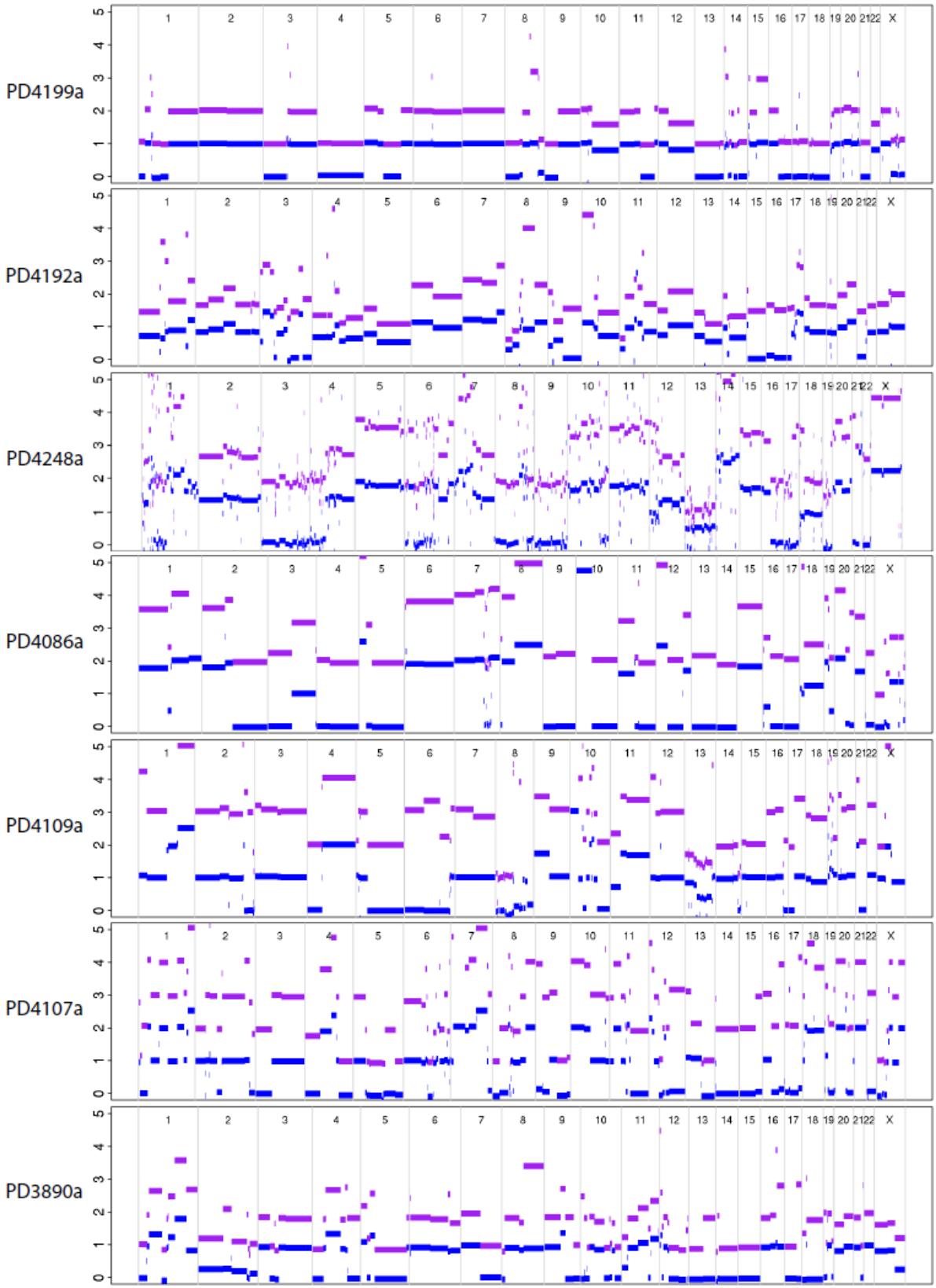
There was distinct copy number variation between breast cancers with copy number changes typical of previous descriptions of breast cancers. Samples PD4194a and PD4088a showed relatively quiescent copy number profiles compared to the rest of the cancers. Frequently observed copy number aberrations included 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

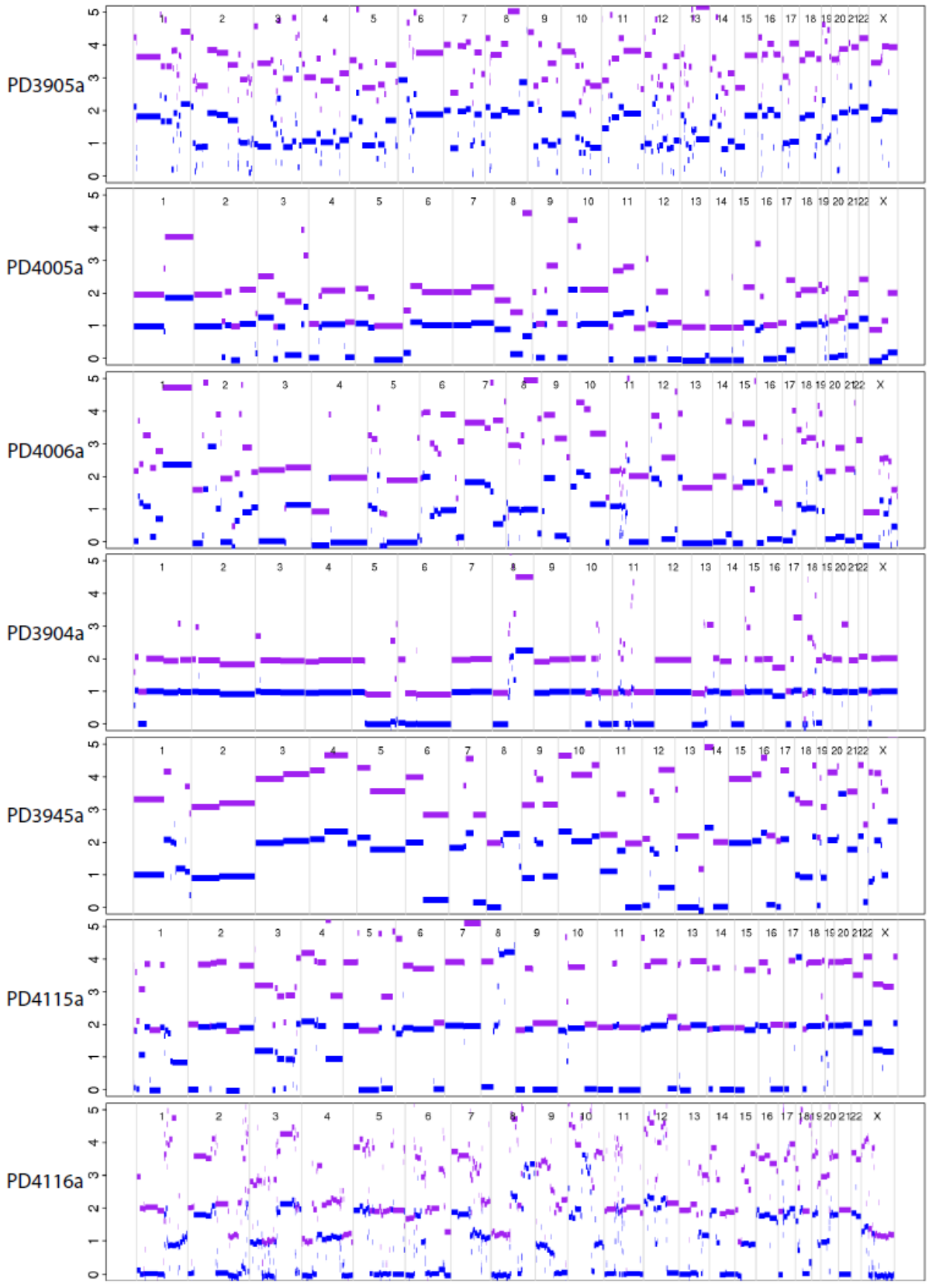


Figure 7.7: Copy number plots for all twenty-one breast cancers. Chromosomes provided along the horizontal axis and copy number values on the vertical axis for each cancer. Purple lines denote total copy number whilst blue denotes minor copy number values.

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.

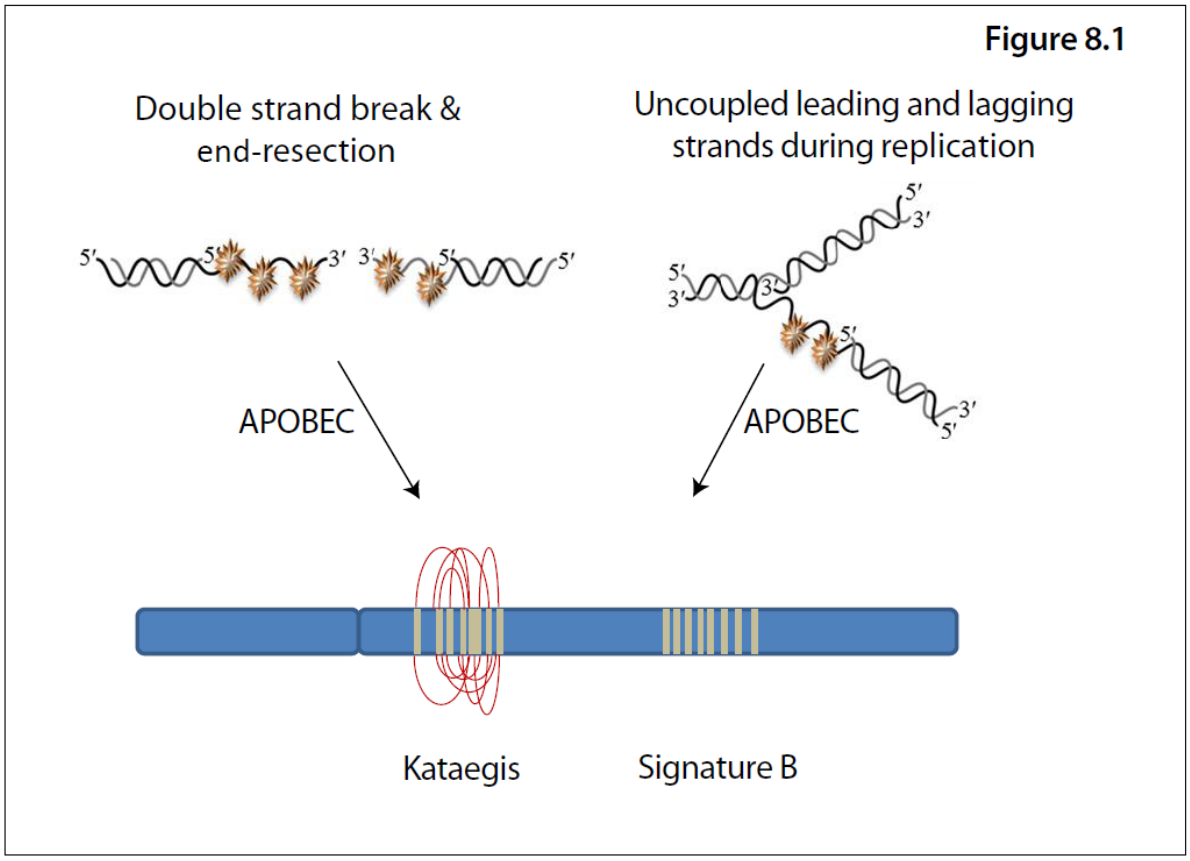


Figure 8.1: Models for APOBEC activity in the genesis of kataegis and Signature B

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.