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).

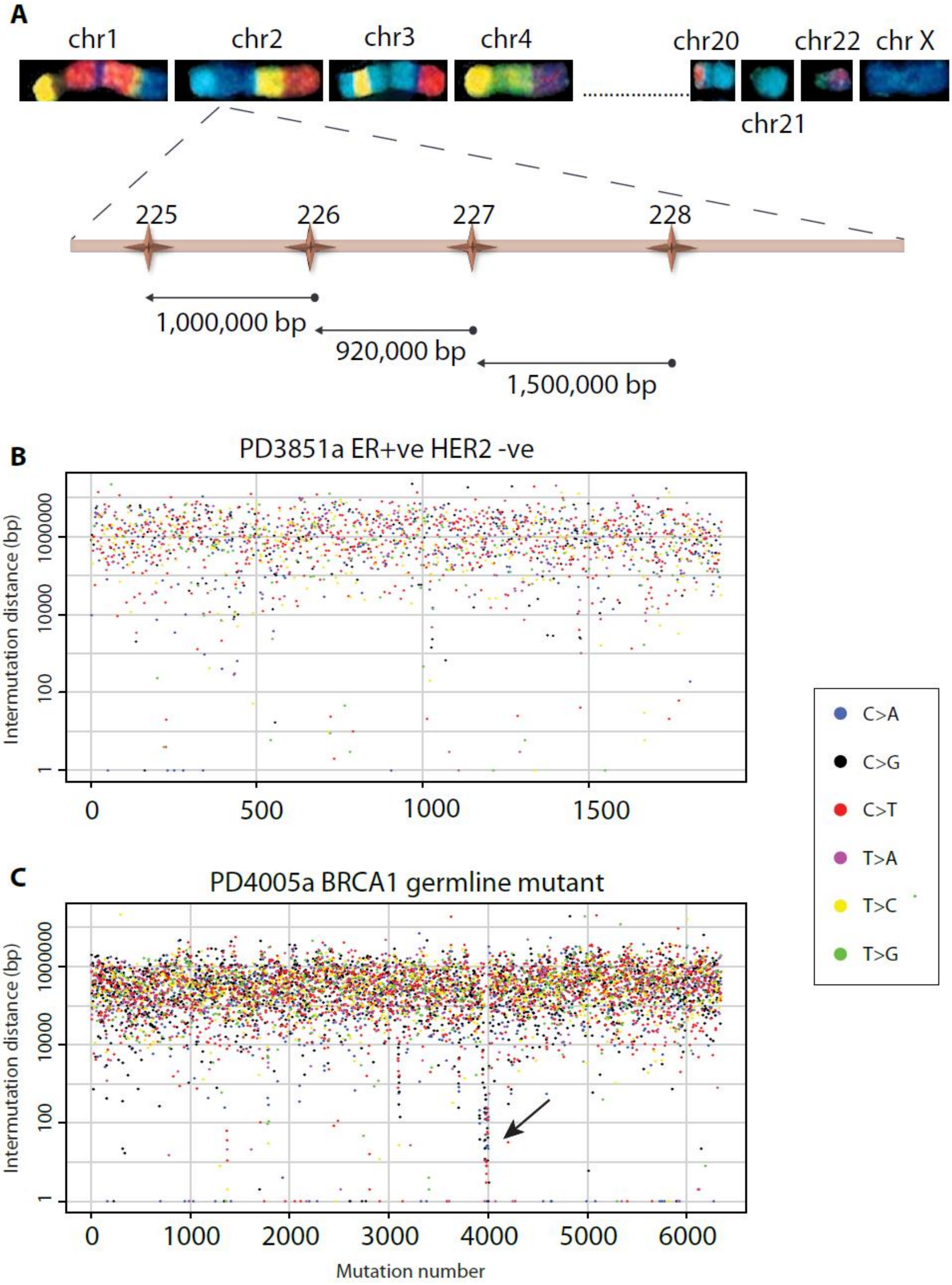


Figure 5.1: The principle behind rainfall plots. (A) Intermutation distance is the distance between each somatic substitution and the substitution immediately prior to it on the reference genome. In the example above of a region on chromosome 2p, mutation number 226 has an intermutation distance 1,000,000bp, 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.

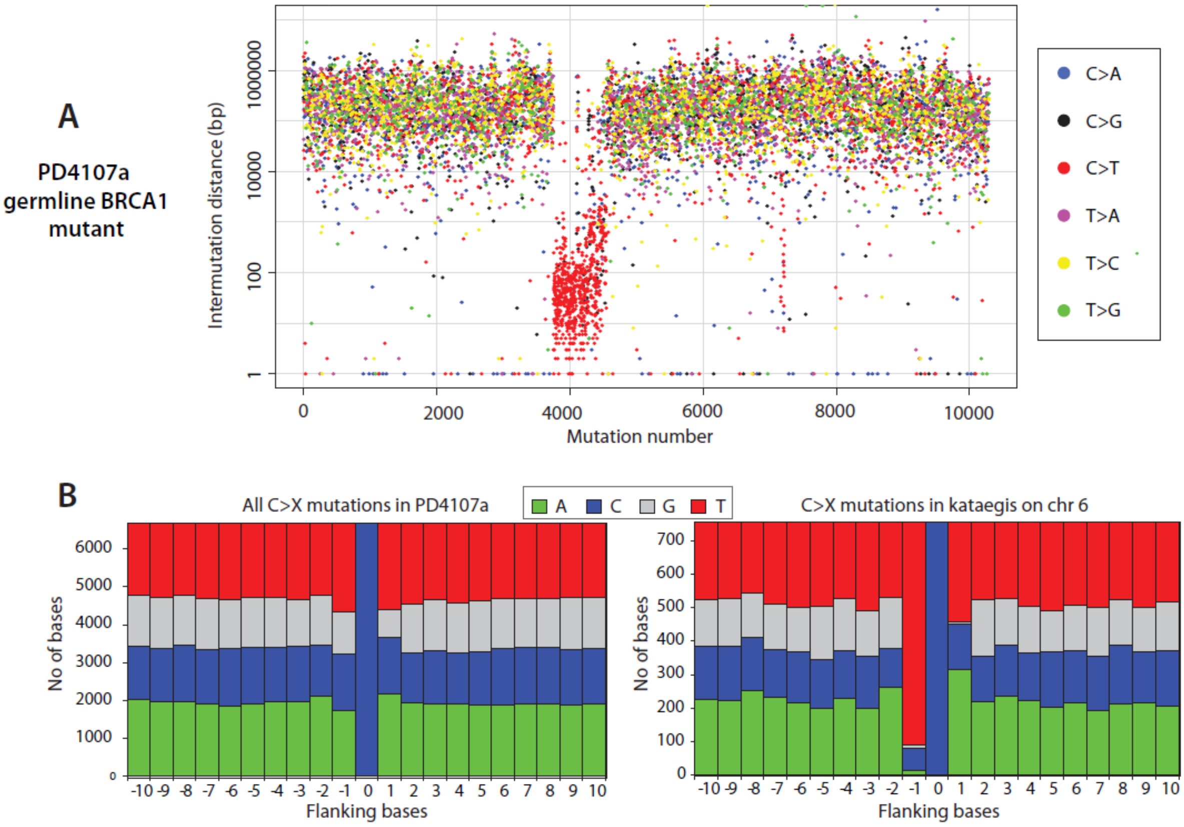


Figure 5.2: (A) Rainfall plot of PD4107a. (D) Plots of flanking sequence of all C>X mutations in PD4107a and C>X mutations within the regions of kataegis in PD4107a. Mutated base is at position 0 with ten bases of flanking sequence provided, demonstrating a strong preference for T at the -1 position 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 Tp$\underline{\text{C}}$ context. Thus C>T, and to a much lesser extent C>G and C>A mutations, at Tp$\underline{\text{C}}$pX trinucleotides were highly enriched in this region of kataegis compared to the remainder of the genome (Figure 5.2b).

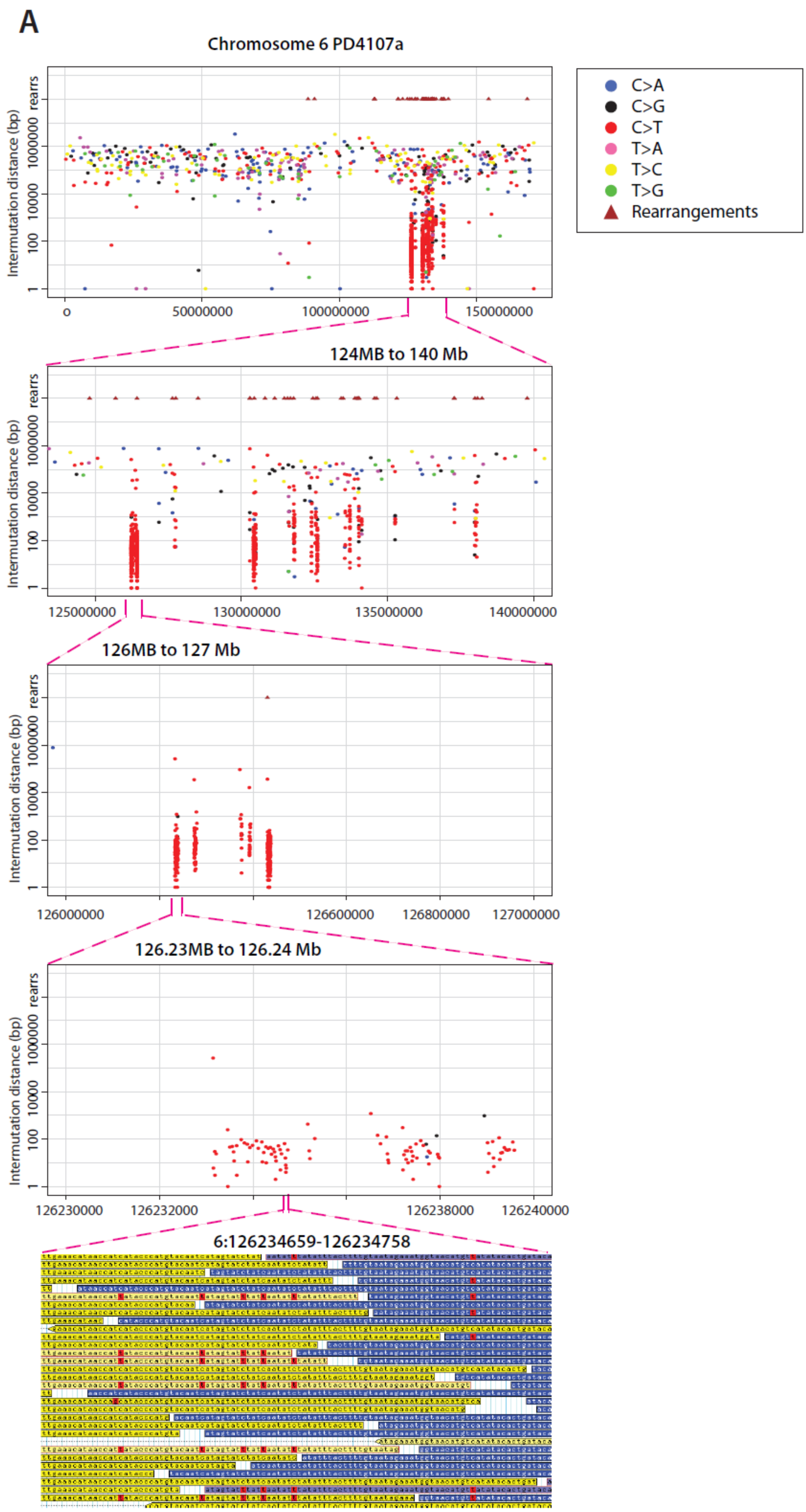
5.3.3 Mutations in microclusters of kataegis occur on the same parental chromosome.

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
C>A
C>G
C>T
T>A
T>C
T>G
Rearrangements
Intermutation distance (bp)
rearrs
1000000
10000
100
1
0
50000000
100000000
150000000
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

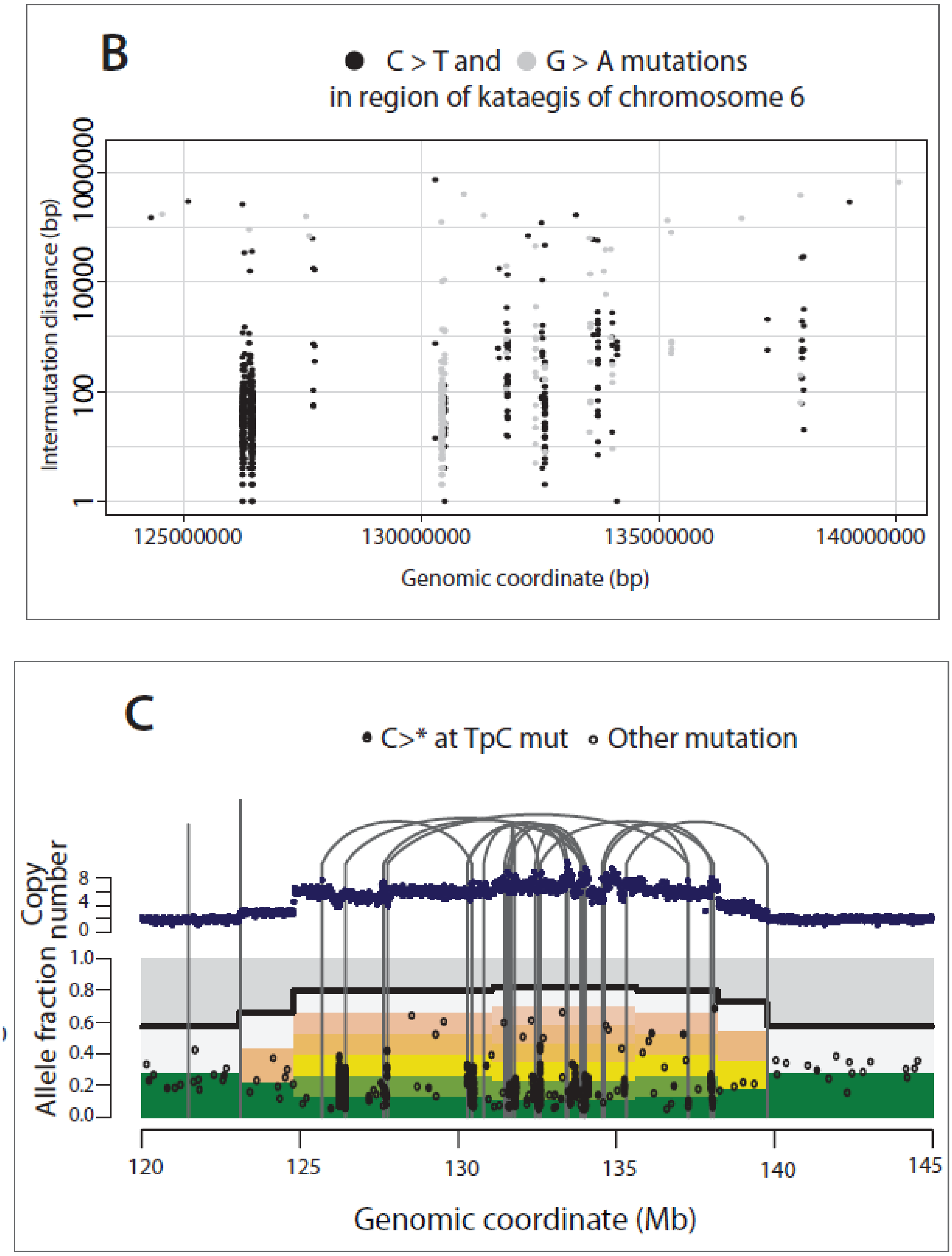


Figure 5.3: Rainfall plot for chromosome 6 of PD4107a. (A) The x axis shows the genomic coordinates of the mutations. Rearrangements are presented as brown triangles (rearrs=rearrangements). The region of kataegis is highlighted at increasing resolution to demonstrate microclusters within the macrocluster. 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

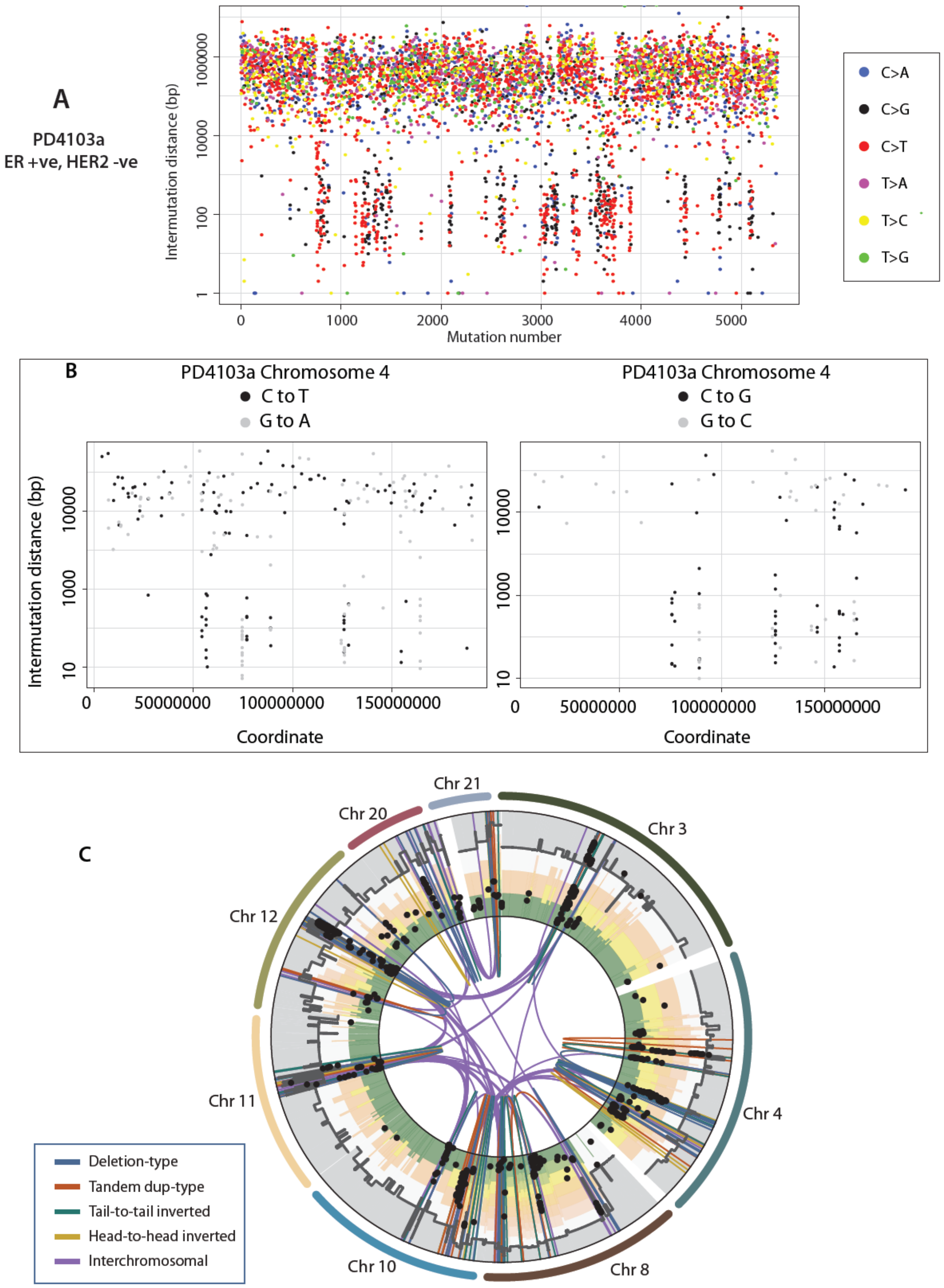


Figure 5.4: (A) Rainfall plot for PD4103a demonstrating kataegis occurring at multiple loci through the genome. (B) Stretches of C>T alternate with stretches of G>A on chromosome 4 in PD4103a. Alternating C>G and G>C mutation on the same chromosome in PD4103a. (C) The complex web of rearrangements involving 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.

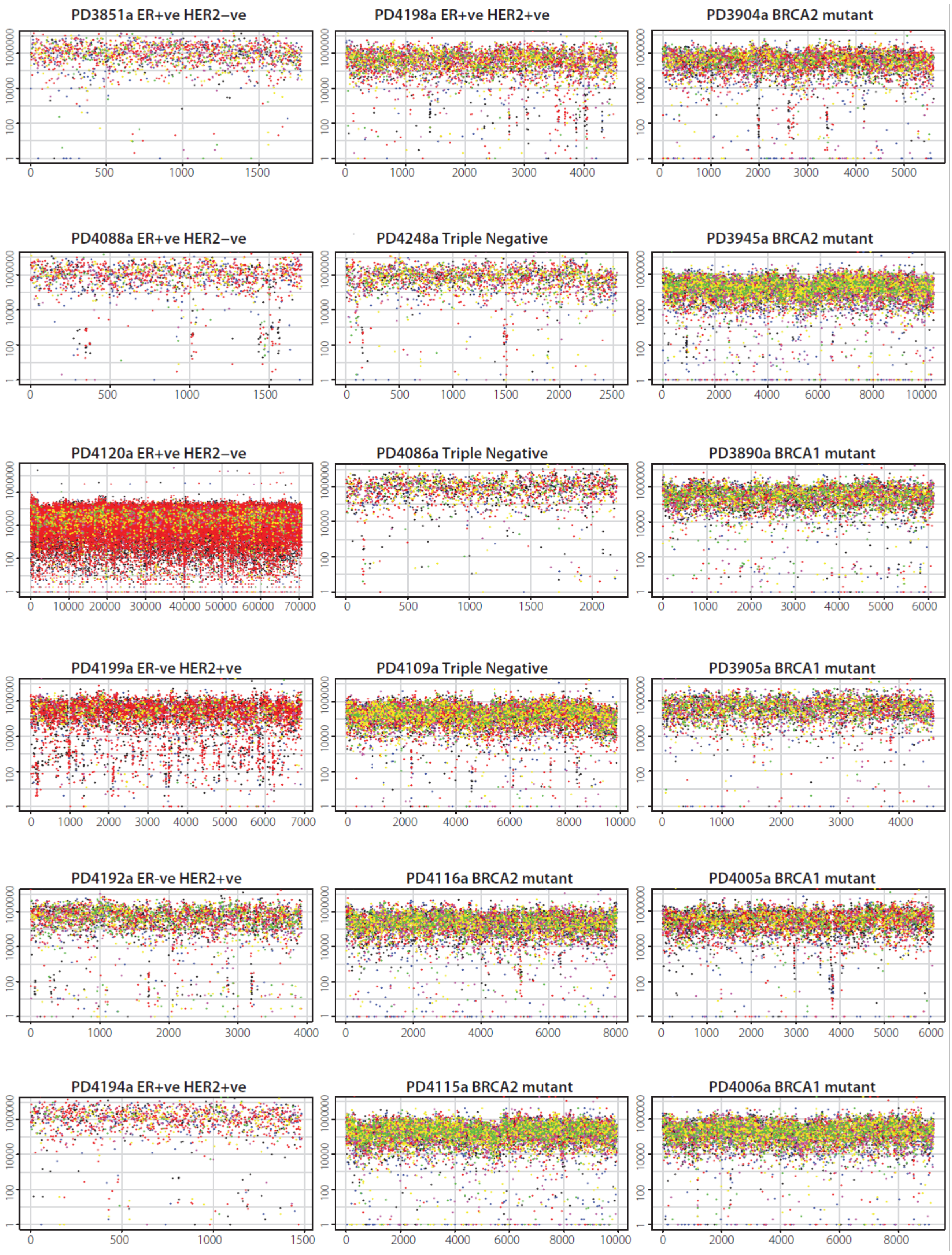


Figure 5.5: Rainfall plots for 18 genomes. Subtle regions of kataegis are present in many samples (PD4198a, PD3904a, PD4088a, PD3945a, PD4120a, PD4086a, PD4199a, PD4109a, PD4192a, PD4116a, PD4005a and PD4006a). Arrows showing some regions of kataegis.

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.

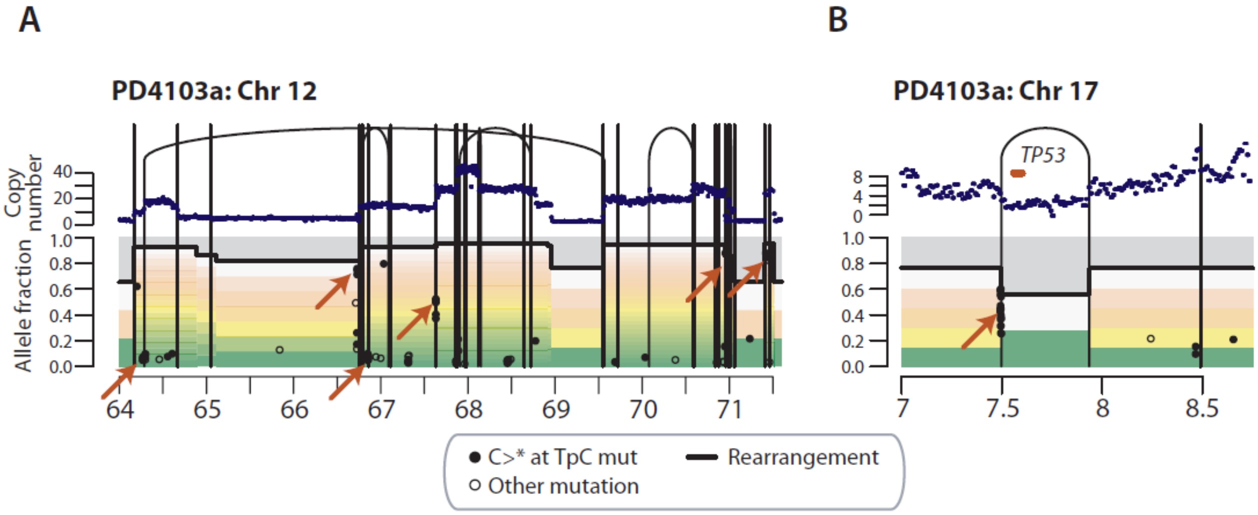


Figure 5.6: Timing kataegis events in PD4103a for the amplicon involving chromosome 12(A) and for a rearrangement resulting in a TP53 deletion (B). The top panel shows the copy number profiles with genomic rearrangements. The lower panel shows the point mutations as filled black circles for C>* mutations 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.