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

4.1 INTRODUCTION

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

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

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

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

4.2 THE SERIES OF BREAST CANCERS USED IN THIS STUDY

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

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

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

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

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

4.2.1 Coverage

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

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

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

4.3 PUTATIVE SOMATIC DRIVER EVENTS IN TWENTY-ONE BREAST CANCERS

In the last forty years, cancer research has focused on the discovery of cancer genes which carry the "driver" mutations that confer selective clonal growth advantage and are causally implicated in oncogenesis. The search for driver mutations has led to the discovery of many cancer genes providing insights into mechanisms of tumorigenesis and targets for therapeutic intervention (Stratton et al., 2009).

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

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

Insertions and deletions

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

Substitutions

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

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

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

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

Table 4.4: Breast cancer series and total number of substitutions

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

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

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

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

4.5 EXPLORING THE SEQUENCE CONTEXT OF SOMATIC SUBSTITUTIONS IN BREAST CANCER

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

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

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

Figure 4.1

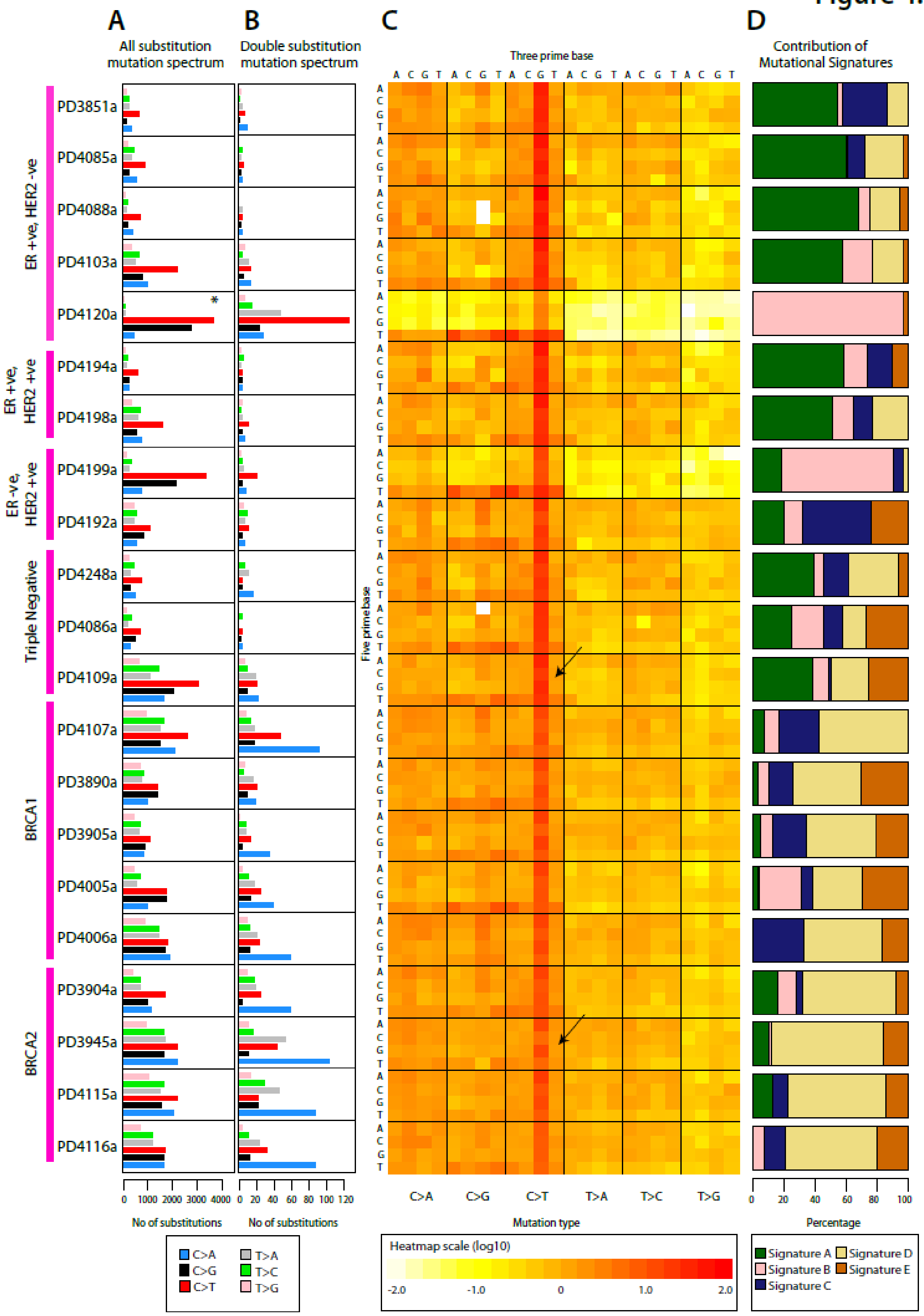

Figure 4.1: Somatic mutation profiles of 21 breast cancers. Breast cancers grouped according to subtype on the far left. (A) Base substitution mutation spectra. *Ultra-deep sequenced PD4120a has an alternative scale on the x axis (0 to 45,000). (B) Mutation spectra of double substitutions from all 21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

4.6 VISUAL IDENTIFICATION OF MUTATION PATTERNS

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

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

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

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

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

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

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

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

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

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

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

4.7 APPLICATION OF A MATHEMATICAL APPROACH TO EXTRACT MUTATION PATTERNS

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

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

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

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

In contrast, the application of other methods of extracting signal from noise produces components lacking obvious visual meaning (Berry et al., 2007). Furthermore, for individual faces, NMF is able to derive the contribution or the amount of exposure of each of those meaningful features. The desire to extract biologically meaningful mutational processes, as well as the intrinsic non-negativity of the mutation spectrum data, renders NMF an appropriate choice for decomposing the mutational spectra of the 21 cases.

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

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

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

Figure 4.2

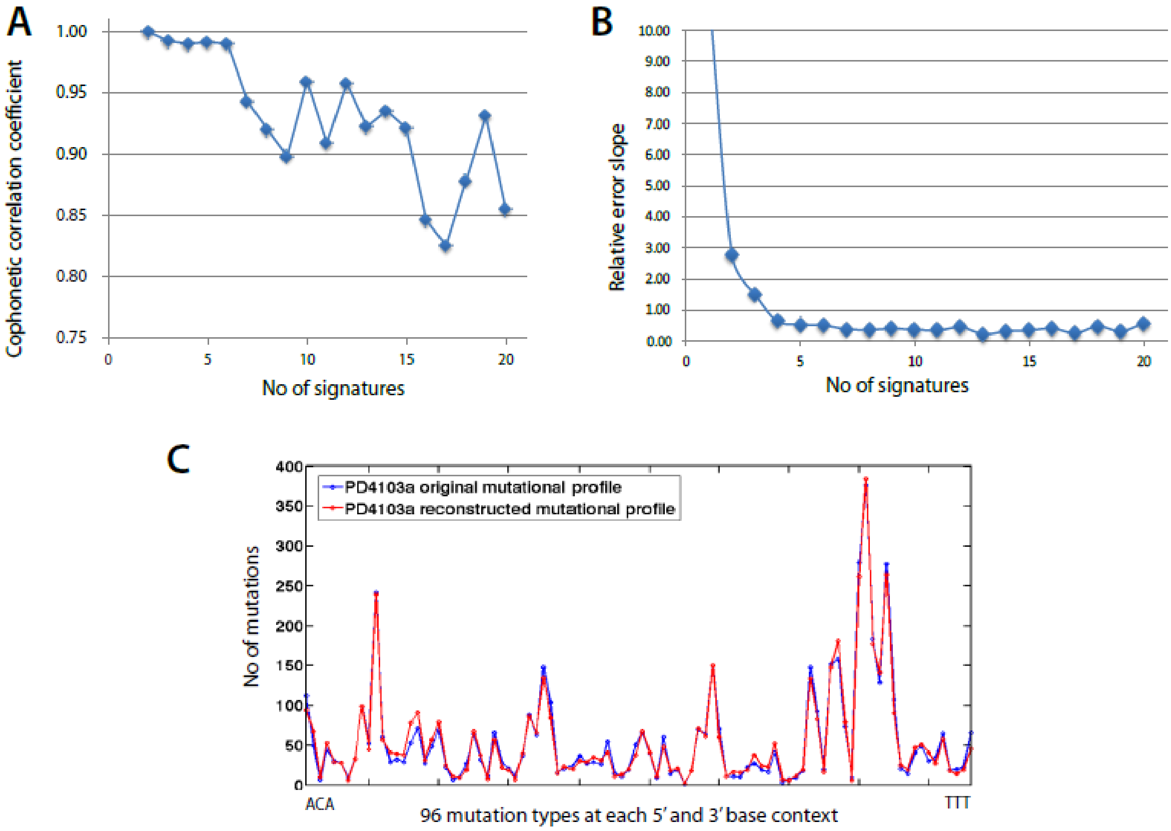

Figure 4.2: Selection of the optimal number of signatures via the NMF model selection framework. (A) The x axis depicts the number of signatures while the y axis shows the cophenetic coefficient. As an indicator of stable reproducibility, the cophenetic correlation coefficient is at its highest points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

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

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

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

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

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

Figure 4.3

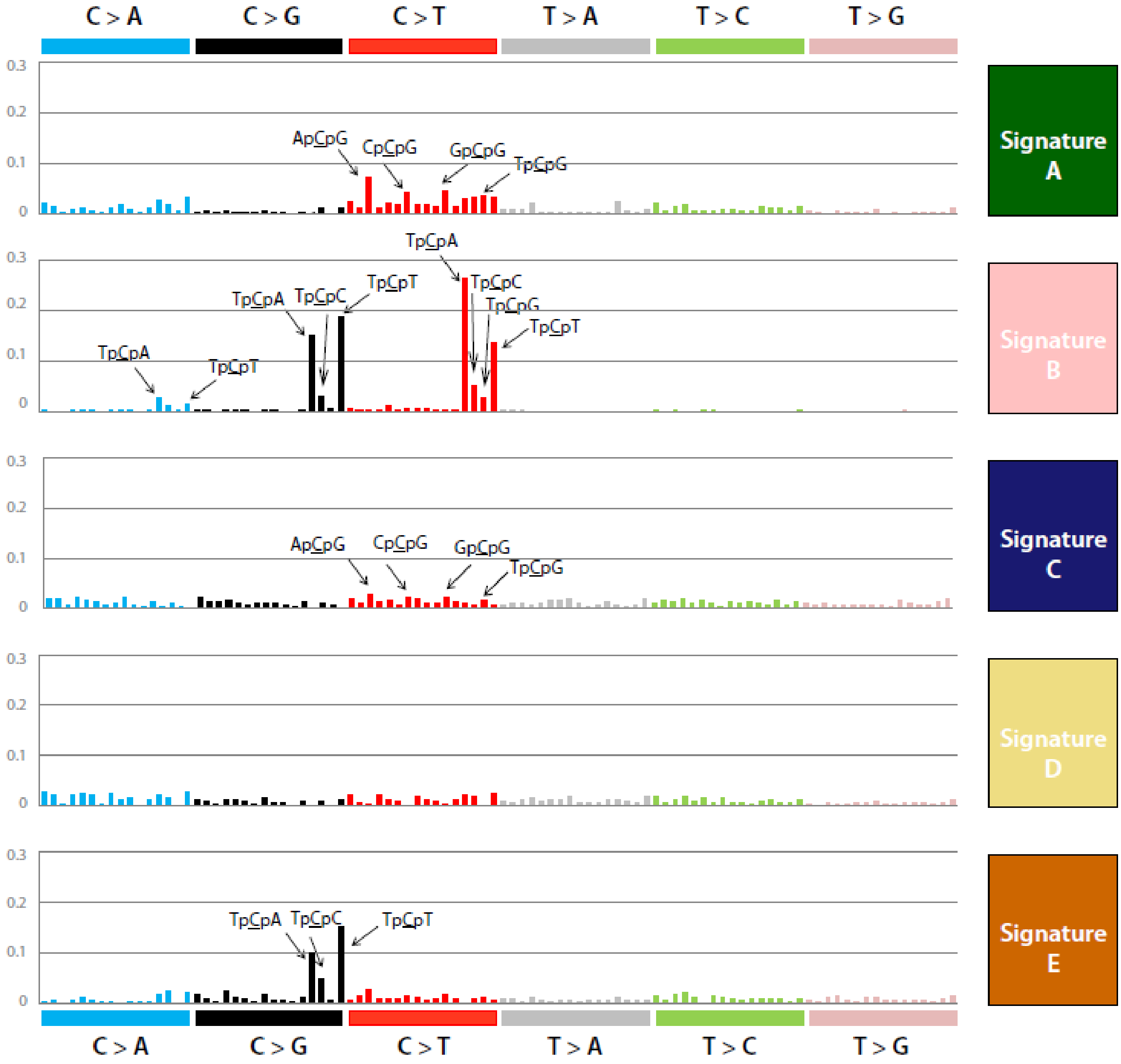

Figure 4.3: Five mutational signatures extracted by NMF in 21 breast cancers. The fraction of contribution of each mutation-type at each context for the five mutational signatures identified by NMF analysis is presented. The major components contributing to each signature are highlighted with arrows.

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

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

Figure 4.4

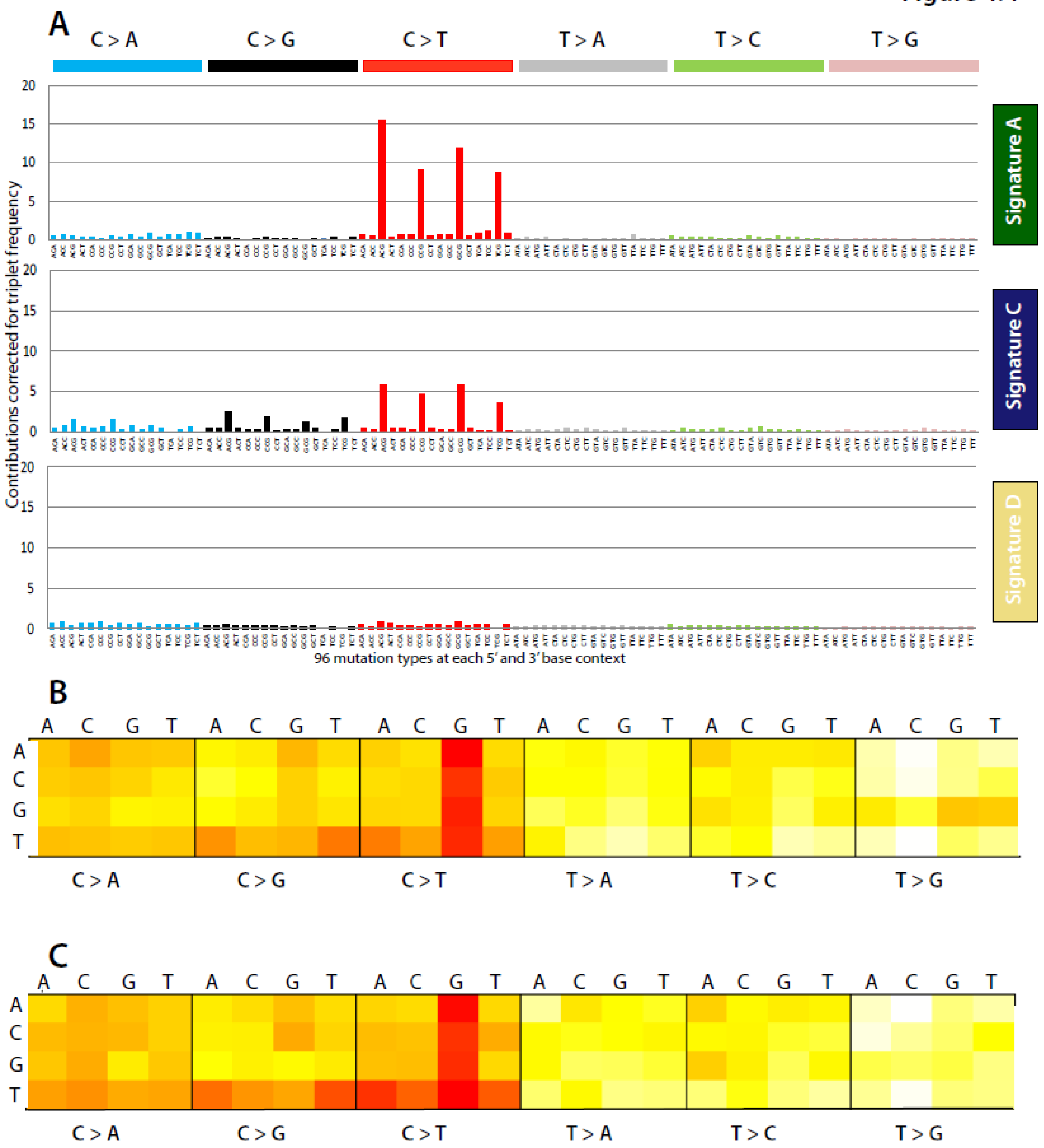

Figure 4.4: Contrasting and validating signatures. (A) Signatures A,C and D with contributions from each of the 96 trinucleotides corrected for the frequency of trinucleotides in the genome. This form of representation highlights the contrast between Signature A and C, as well as demonstrates the differences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

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

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

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

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

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

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

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

Figure 4.5

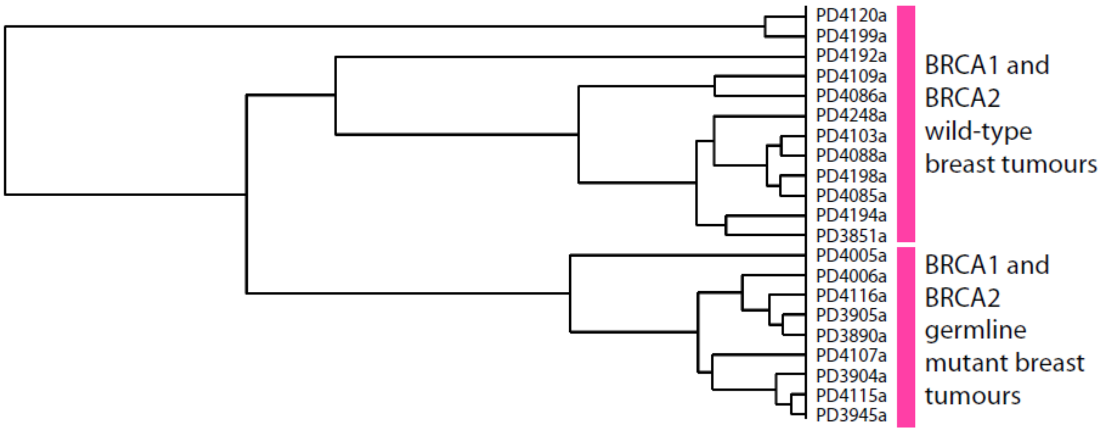

Figure 4.5: Cluster dendrogram generated by unsupervised hierarchical clustering based on contributions of the five mutational signatures identified by NMF for the 21 breast cancer genomes.

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

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

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

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

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

4.10 THE CONTRIBUTIONS OF MUTATIONAL SIGNATURES CHANGE OVER EVOLUTIONARY TIME

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

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

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

4.10.2 The rationale for interrogating timing of mutational processes

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

4.10.3 The temporality of mutational processes

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

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

Figure 4.6

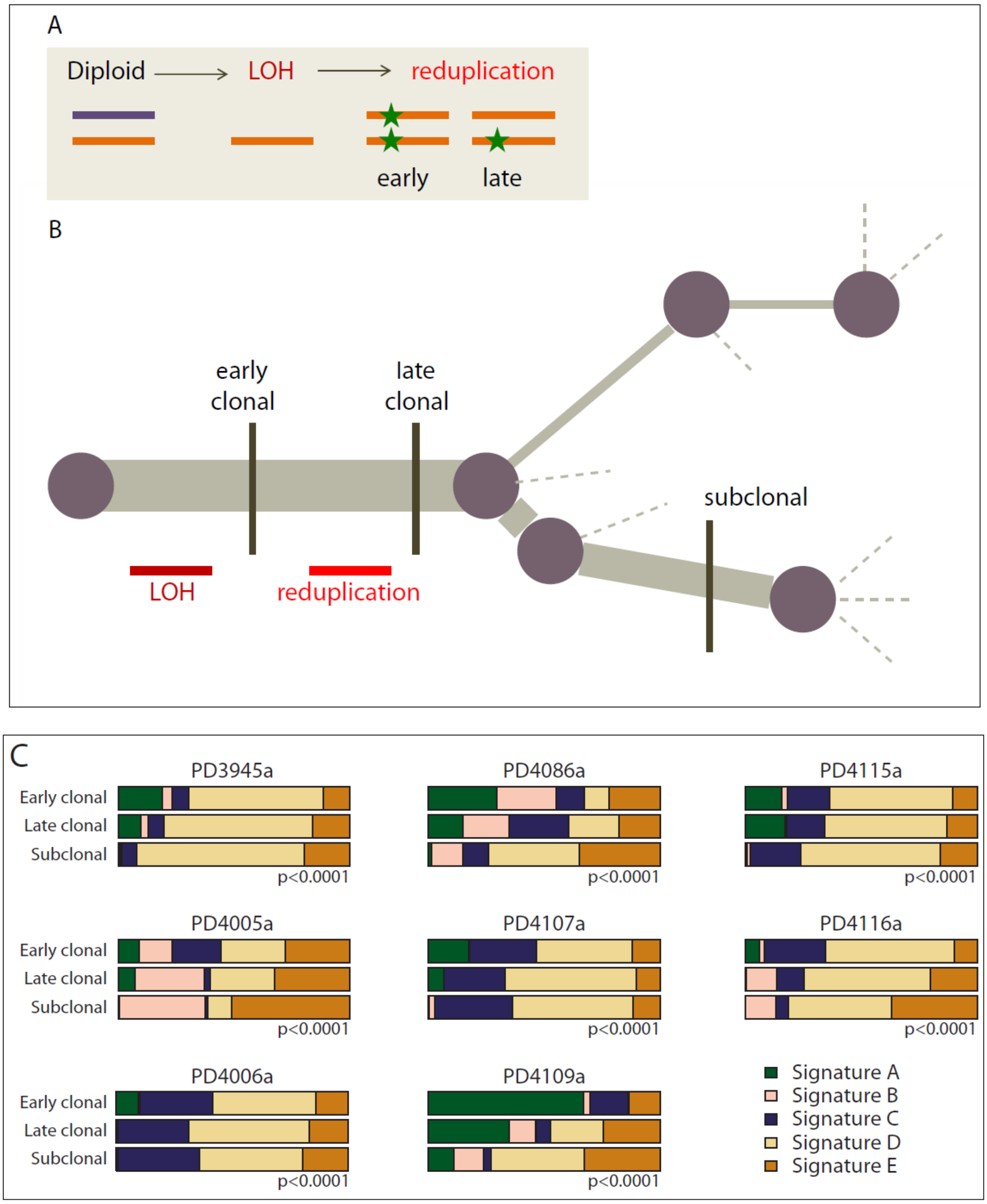

Figure 4.6: Temporality of mutational processes (A) Starting from a diploid state, loss of one parental allele will lead to a state of loss of heterozygosity and a ploidy of 1. However, reduplication of this single allele can occur. If a mutation occurs on the single allele early in the evolution of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

4.11 DISCUSSION

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

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

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

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

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

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

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

The mechanisms underlying the remainder are currently unknown. The most distinctive of these signatures, Signature B, is characterised by C>T, C>G, and to a lesser extent, C>A substitutions at Tp$\underline{C}$pX trinucleotides, is responsible for the overwhelming majority of mutations in two cancer samples, PD4120a and PD4199a. These two cancers are most similar to each other and most dissimilar to the other breast cancers, despite having an order of magnitude difference in mutation burden (70690 versus 6932 total substitutions). These two cancers are also of divergent histopathological subtypes; PD4120a is an ER positive, PR positive and HER2 negative breast cancer, whilst PD4199a is an ER negative, PR negative and HER2 positive cancer suggesting that the underlying mutational process generating this striking signature is independent of and unrelated to expression-based profiling. Signature B has similarities with the mutational signature produced by the endogenous deaminating enzyme superfamily described in the introductory chapter, the APOBEC family.

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

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

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

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

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

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

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

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

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