Chapter 3

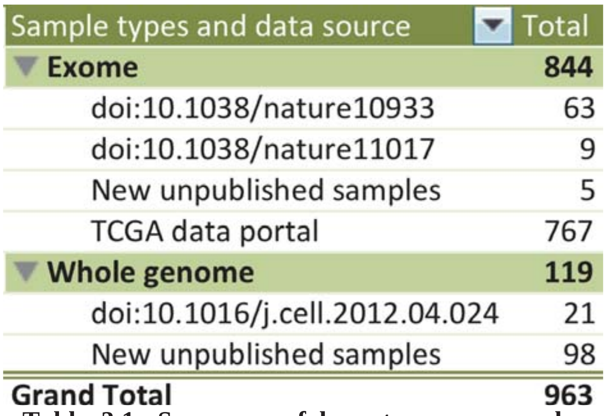
Signatures of mutational processes operative in breast cancer

3.1 Introduction

The previous chapter introduced a novel mathematical model of mutational processes operative in cancer genomes and a computational framework that allows deciphering of the signatures of these processes from a set of mutational catalogues. The newly developed computational approach was extensively evaluated with simulated data demonstrating its applicability to mutational catalogues derived from sequencing both cancer genomes and cancer exomes. Further, the performed simulations demonstrated that the method is robust to a wide range of different parameters. In this chapter, I present and discuss the application of the developed framework to experimentally generated data. The framework is used to examine the mutational catalogues derived from the sequences of 844 breast cancer exomes and 119 breast cancer whole-genomes. The aim of this chapter is to describe the signatures of the mutational processes operative in breast cancer as well as to serve as a prelude to chapter 4 in which analogous analysis will be performed for another 29 different types of human cancer.

3.2 Data generation and filtering of mutational catalogues

It should be noted that none of the examined data are generated for the purposes of this thesis. Rather, the analysis relies on previously identified somatic mutations by curating freely available published data as well as data that was unpublished at the time. Any unpublished breast cancer data were generated internally at the Cancer Genome Project (CGP) for the purposes of other projects. The majority of breast cancer exomes are taken from The Cancer Genome Atlas (TCGA) data portal as well as from peer-reviewed publications. In contrast, the majority of breast cancer whole-genomes are previously unpublished data. Summary of the numbers of samples based on their data source is provided in Table 3.1, whereas a complete list including all samples, all examined cancer types, and their respective data sources is provided in Appendix II.



Sample types and data source	Total
▼ **Exome**	**844**
doi:10.1038/nature10933	63
doi:10.1038/nature11017	9
New unpublished samples	5
TCGA data portal	767
▼ **Whole genome**	**119**
doi:10.1016/j.cell.2012.04.024	21
New unpublished samples	98
Grand Total	**963**

Table 3.1: Summary of breast cancer samples and their data sources.

The somatic mutations of the 844 breast cancer exomes and the 119 breast cancer whole-genomes are curated, filtered, and mutational catalogues are generated for each sample based on the Ξ_6, Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536} alphabets. It should be noted that there is no sample overlap between the breast cancer genomes and exomes (*i.e.*, breast cancer whole-genomes are not included twice as exomes and genomes).

As these data are retrieved from many different sources and generated using different next-generation sequencing platforms and bioinformatics approaches, quality control is performed in order to remove any germline contamination and technology specific sequencing artefacts. Germline mutations are filtered out from the list of reported mutations using the data from dbSNP (Sherry et al., 2001), 1000 genomes project (Abecasis et al., 2012), NHLBI GO Exome Sequencing Project (Fu et al., 2013), and 69 Complete Genomics panel (http://www.completegenomics.com/public-data/69-Genomes/). Any mutation at a position of a previously identified germline variant in any of these datasets is removed from the signatures analysis. Furthermore, technology specific sequencing artefacts are filtered out by using panels of (unmatched) BAM files for normal tissue containing 137 normal genomes and 532 normal exomes. Any somatic mutation present in at least three well-mapping reads in at least two normal BAM files is discarded. The remaining somatic mutations are used for the generation of mutational catalogues and the extraction of mutational signatures.

Figure 3.1

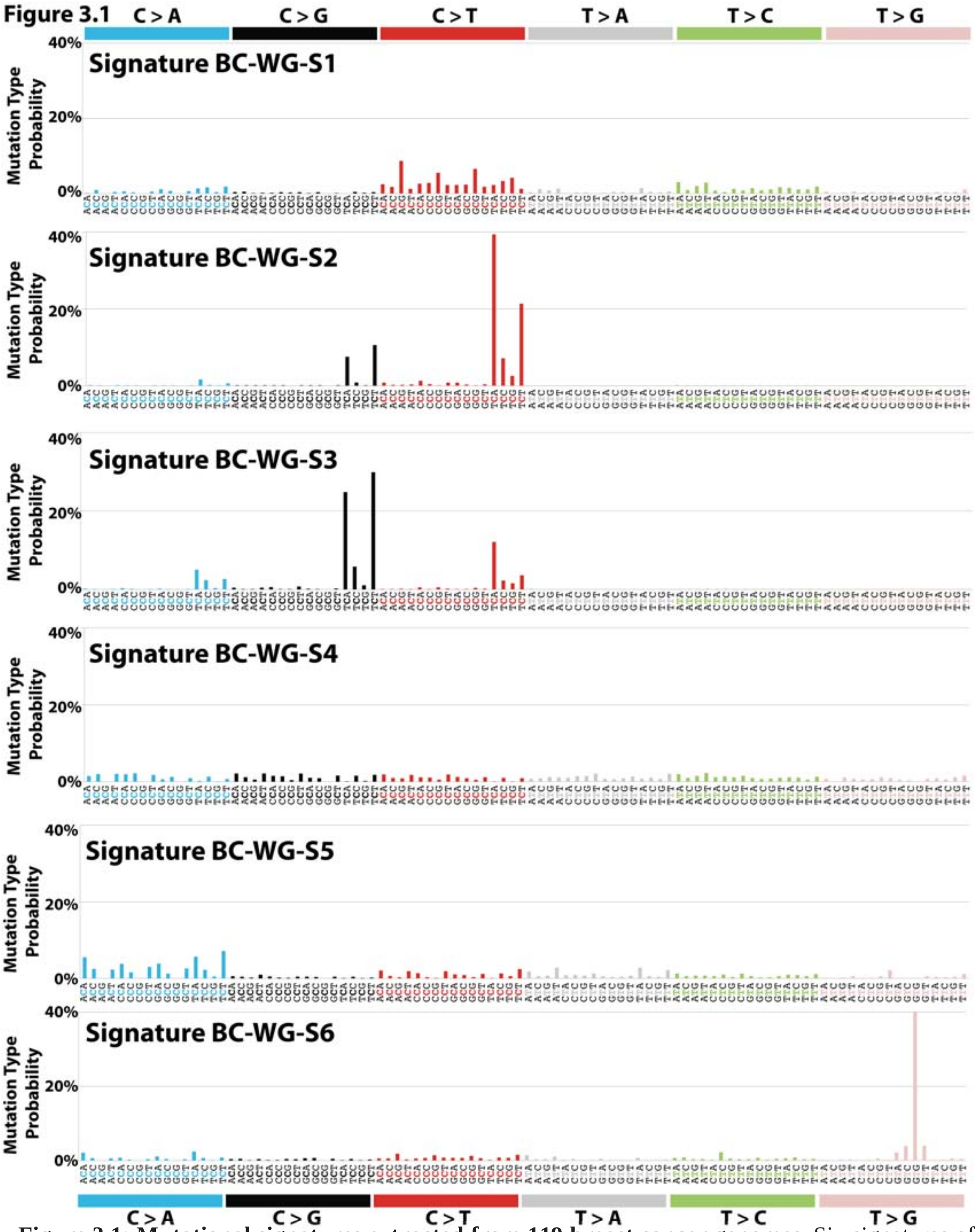


Figure 3.1: Mutational signatures extracted from 119 breast cancer genomes. Six signatures of mutational processes are deciphered from the base substitutions (including their immediate 5’ and 3’ sequence context) identified in the examined 119 breast cancer genomes. Each signature is depicted on an independent panel, where each type of substitution is displayed in a different colour. Mutational signatures are plotted based on the genome trinucleotide frequency.

The immediate 5’ and 3’ sequence context is extracted using the ENSEMBL Core APIs for human genome build GRCh37. Curated data originally mapped to an older version of the human genome is re-mapped using UCSC’s freely available lift genome annotations tool. Dinucleotide substitutions are identified when two substitutions are present in consecutive bases on the same chromosome (sequence context is ignored). The immediate 5’ and 3’ sequence content of all small insertions and deletions (indels) is examined and the ones present at mono/polynucleotide repeats or microhomologies are included in the analysed mutational catalogues as their respective types. Strand-bias catalogues are derived for each sample using only substitutions identified in the transcribed regions of well-annotated protein coding genes. Mutational signatures are independently derived from the mutational catalogues of breast cancer exomes and breast genomes (see below).

3.3 Deciphering the signatures of mutational processes from whole-genome sequencing of breast cancers

The developed computational approach presented in chapter 2 is applied to the mutational catalogue of 119 breast cancer whole-genomes that contain 654,308 somatic substitutions and indels. Mutational signatures are extracted based on the Ξ_{96}, Ξ_{99}, $\Xi_{192,}$ and Ξ_{1536} alphabets. The approach reveals six consistent and reproducible mutational signatures for all four alphabets – termed Signatures BC-WG-S1, BC-WG-S2, BC-WG-S3, BC-WG-S4, BC-WG-S5, and BC-WG-S6 (BC-WG-S stands here for breast cancer whole-genome signature).

The patterns of somatic substitutions for the signatures extracted using Ξ_{96} are depicted in Figure 3.1. Signature BC-WG-S1 is characterized by 50% C>T substitutions predominantly occurring at CpG dinucleotides and 25% T>C mutations with peaks at ApTpN trinucleotides. Signature BC-WG-S has predominantly (~76%) C>T mutations at TpCpN trinucleotides and (~20%) C>G mutations occurring at TpCpN trinucleotides. In contrast, Signature BC-WG-S3 is mirroring Signature BC-WG-S2 with ~65% of its substitutions being C>G at TpCpN trinucleotides, ~22% being C>T at TpCpN trinucleotides, and ~11% C>A at TpCpN trinucleotides. Signature BC-WG-S4 has a rather flat mutational pattern including all types of somatic mutations. While this mutational signature does not exhibit any strong features based on the immediate 5' or 3' sequence context, such as Signatures BC-WG-S2 or BC-WG-S3, the pattern of its substitutions is not completely uniform. Rather, the mutational pattern of Signature BC-WG-S4 has subtle trinucleotide features. Similar to BC-WG-S4, Signature BC-WG-S5 has a generally flat mutational pattern with subtle sequence context features. However, in addition, Signature BC-WG-S5 exhibits a predominance of C>A mutations (~40%) compared to the other types of substitutions. Lastly, Signature BC-WG-S6 has a very strong sequence context with ~40% of all mutations being T>G at GpTpG.

As previously demonstrated, the developed computational framework can be applied to a wider repertoire of mutation types than the 96 mutated trinucleotides. The Ξ_{96} alphabet can be extended to the Ξ_{99} alphabet by including three additional mutational subclasses: double nucleotide substitutions, indels at microhomologies, and indels at mono/polynucleotide repeats. This analysis reveals that Signature BC-WG-S4 is associated with 91% of the 8,915 indels at microhomologies found in the 119 whole breast genomes, 39% of the 12,555 indels found at mono/polynucleotide repeats, and 21% of the 3,974 dinucleotide substitutions (Figure 3.2). The activity of Signature BC-WG-S1 is associated with 52% of indels found at mono/polynucleotide repeats, whereas Signature BC-WG-S5 accounts for 65% of all dinucleotide substitutions. It should be noted that a significant proportion of the dinucleotide substitutions associated with Signature BC-WG-S5 are CC>AA. Signatures BC-WG-S2, BC-WG-S3, and BC-WG-S6 do not have a strong association with any type of indels or dinucleotide substitutions.

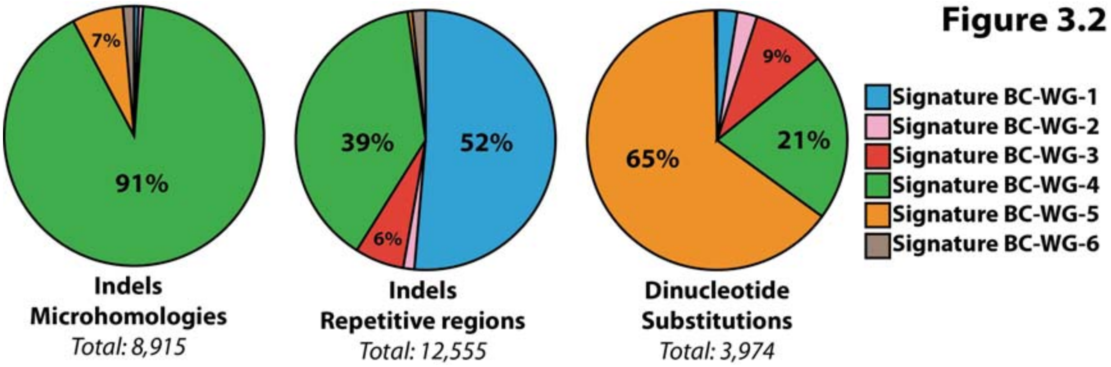


Figure 3.2: Breast cancer whole-genome mutational signatures with indels and dinucleotides. Mutational signatures analysis of the 119 breast cancer whole-genomes is extended to incorporate indels at microhomologies, indels at repetitive regions, and dinucleotide substitutions (*i.e.*, Ξ_{99} alphabet). The percentage of mutations attributed to these three additional mutation types is displayed for all signatures that contribute at least 5%. Each signature is displayed in a different colour.

Previous examination of the mutational catalogues of 21 breast cancer genome showed a weak transcriptional strand-bias for all C>A mutations (Nik-Zainal et al., 2012). This bias results in C>A mutations being more common on the transcribed than the untranscribed strands of genes (and vice versa for G>T mutations). To investigate whether a particular mutational signature is associated with this (or any other) transcriptional strand-bias, the Ξ_{96} substitution alphabet is extended to include information on whether a substitution is on the transcribed or non-transcribed strand,

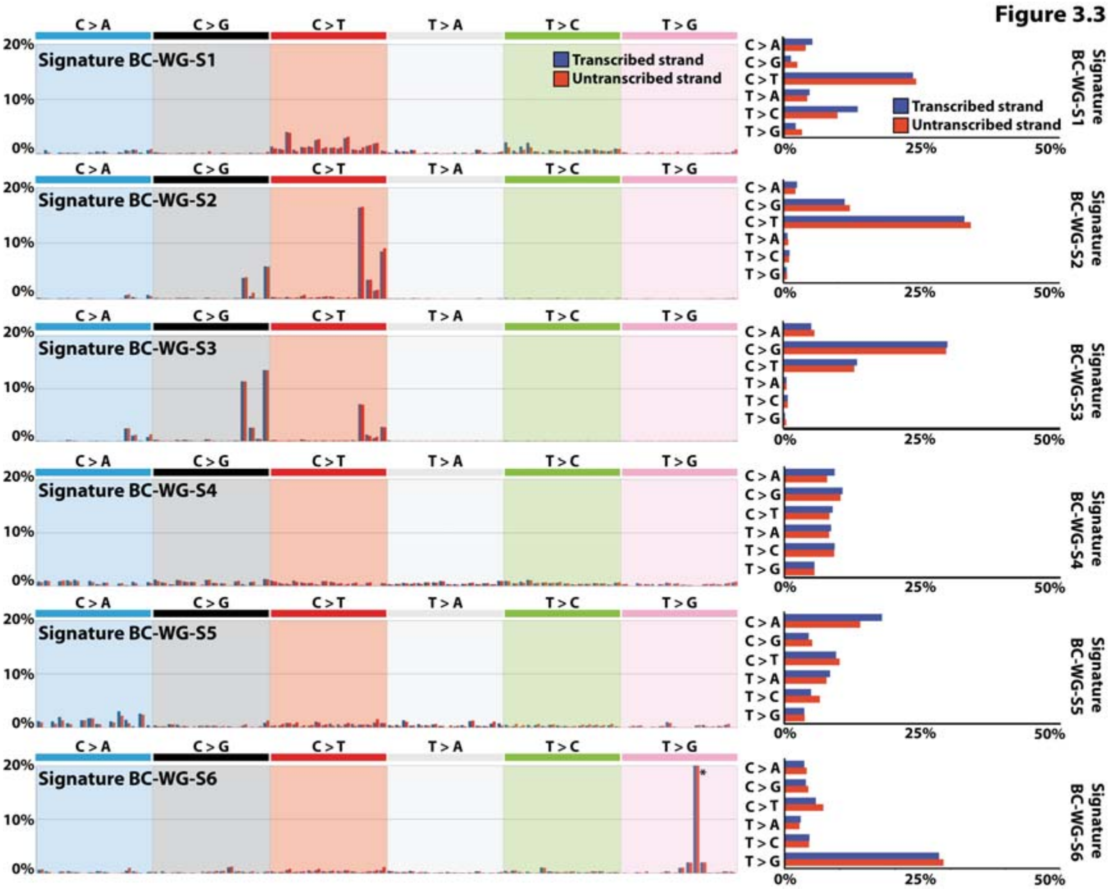


Figure 3.3: Breast cancer whole-genome mutational signatures with strand-bias. Signatures of mutational processes with strand-bias are extracted from the mutational catalogues of 119 breast cancer genomes. Six mutational signatures deciphered from the base substitutions (including their immediate 5' and 3' sequence context) identified in the transcribed regions of 119 breast cancer genomes. Each signature is depicted on an independent panel, where each type of substitution is highlighted in a different colour. The probability of a mutation to occur on a transcribed strand is depicted in blue, while red is used to display the probability of a mutation to occur on the untranscribed strand. Mutational signatures are plotted based on the genome trinucleotide frequency. Asterisk indicates mutation type exceeding 20%.

thus increasing the 96 trinucleotide substitutions to 192. The developed model selection approach again reveals the signature of six reproducible mutational processes (Figure 3.3) with patterns resembling the ones based on the Ξ_{96} alphabet (Figure 3.1). Examining the mutational signatures based on the Ξ_{192} alphabet reveals that Signature BC-WG-S2, Signature BC-WG-S3, Signature BC-WG-S4, and Signature BC-WG-S6 do not have statistically significant strand-bias (Figure 3.3). In contrast, Signature BC-WG-S1 exhibits a weak T>C strand-bias ($Q = 1.4 \times 10^{-3}$; in all cases Q refers to a q-value, see chapter 7), while Signature BC-WG-S5 is associated with a C>A strand-bias ($Q = 5.2 \times 10^{-7}$). The nature of the mutational process(es) underlying these transcription strand-biases is currently unknown, but it could be due to past activity of transcription-coupled nucleotide excision repair.

The previous assessment of the impact of sequence context on classification of mutational processes is limited to the bases immediately 5’ and 3’ to each mutated base. However, other sequence motifs close to or distant from the mutant base could be important in defining a mutational process. Here, I extend the sequence context to include the two bases 5’ and 3’ to each mutated base, which results in 1,536 possible mutated pentanucleotides (*i.e.*, mutational signatures are examined based on the Ξ_{1536} alphabet). The model selection approach is able to find six reproducible mutational signatures based on the 1,536 mutation types. New sequence context dependencies are found in several of the previously identified mutational signatures. Signature BC-WG-2 substitutions at TpCpN trinucleotides are dependent on the next base 5’, which is predominantly a pyrimidine (Figure 3.4A and 3.4B). Of all C>X at TpCpN mutations caused by Signature 2, 40% are at CpTpCpN, 33% at TpTpCpN and the remaining 27% are either G or A 5’ to the TpCpN trinucleotide (Figure 3.4C). Such a tetranucleotide distribution is highly unlikely to happen purely by chance in the human genome (Q = 7.1×10^{-14}). Exactly the same set of observations can be made for Signature BC-WG-3 when additional sequence context is included (data not shown).

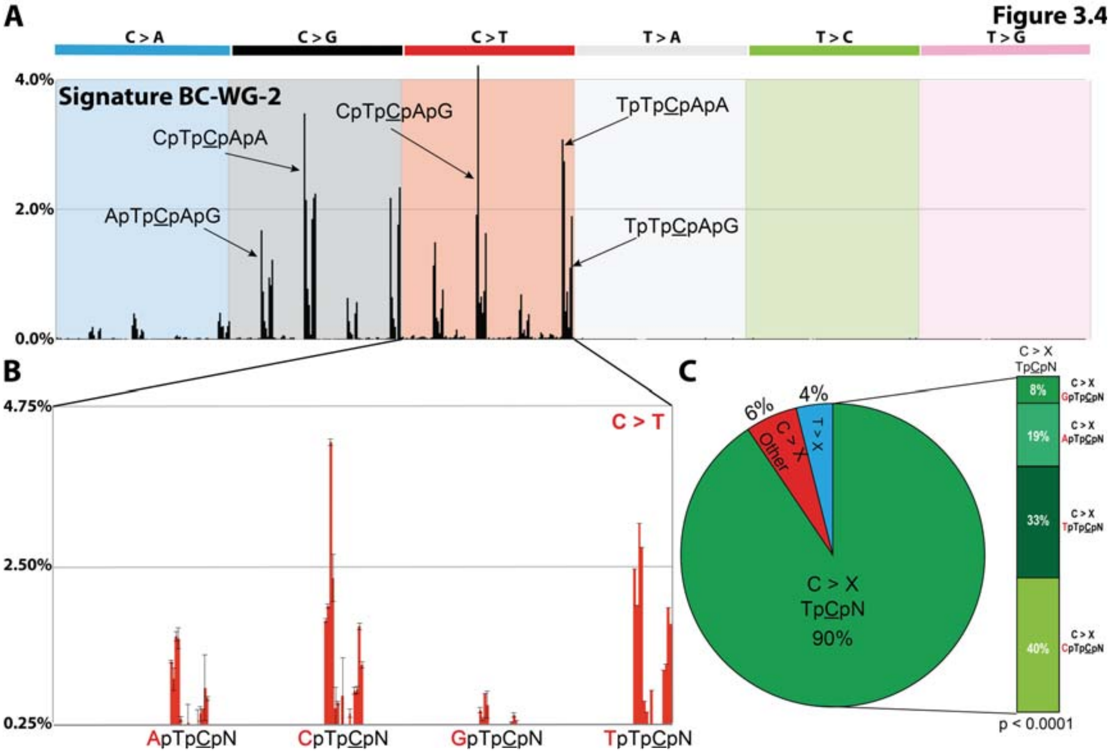


Figure 3.4: Signature BC-WG-2 with additional sequence context. (A) Signature BC-WG-2 is deciphered from the base substitutions (including the two bases 5’ and 3’ to each mutated base resulting in 1,536 possible mutated pentanucleotides) identified in 119 breast cancer genomes. (B) Detailed view of C>T mutation types in Signature BC-WG-2. (C) Summary of all mutation types caused by Signature BC-WG-2.

In addition to Signatures BC-WG-2 and BC-WG-3, Signature BC-WG-6 also exhibits a strong context dependency when it is examined based on the Ξ_{1536} alphabet (Figure 3.5). Approximately 20% of all somatic mutations due to this mutational signature are T>G at GpGpTpGpG pentanucleotides ($Q = 2.7 \times 10^{-31}$). It should be noted that, when extracted based on the Ξ_{1536} alphabet, Signatures BC-WG-1, BC-WG-4, and BC-WG-5 do not show any specific pentanucleotide patterns.

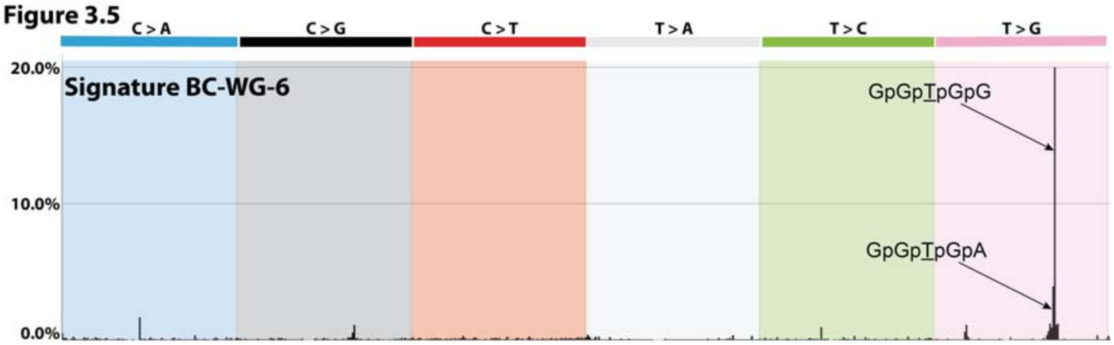


Figure 3.5: Signature BC-WG-6 with additional sequence context. Signature BC-WG-5 is deciphered from the base substitutions (including the two bases 5' and 3' to each mutated base resulting in 1,536 possible mutated pentanucleotides) identified in 119 breast cancer genomes.

3.4 Deciphering the signatures of mutational processes from exome sequencing of breast cancers

The developed computational approach presented in chapter 2 is applied to the mutational catalogues of 884 breast cancer exomes that contain 39,480 somatic substitutions and indels. Mutational signatures are extracted based on the Ξ_{96}, Ξ_{99}, and Ξ_{192} alphabets. The approach reveals three reproducible mutational signatures for all alphabets – termed Signatures BC-EX-S-1, BC-EX-S-2, and BC-EX-S-3 (BC-EX-S stands here for breast cancer exome signature). The numbers of somatic mutations in these exome data (average ~45 somatic mutations per sample) are found to be too low to perform signature analysis using 1,536 mutation types and, as such, no mutational signatures are derived based on the Ξ_{1536} alphabet.

The patterns of somatic substitutions for the signatures extracted from the breast cancer exomes using Ξ_{96} are depicted in Figure 3.6. Signature BC-EX-S-1 is characterized by 60% C>T substitutions predominantly occurring at CpG dinucleotides and 17% T>C mutations with peaks at ApTpN trinucleotides. The pattern of mutations of Signature BC-EX-S-1 (Figure 3.6) closely resembles the one of Signature BC-WG-S1 (Figure 3.1). In fact, these two mutational signatures have a

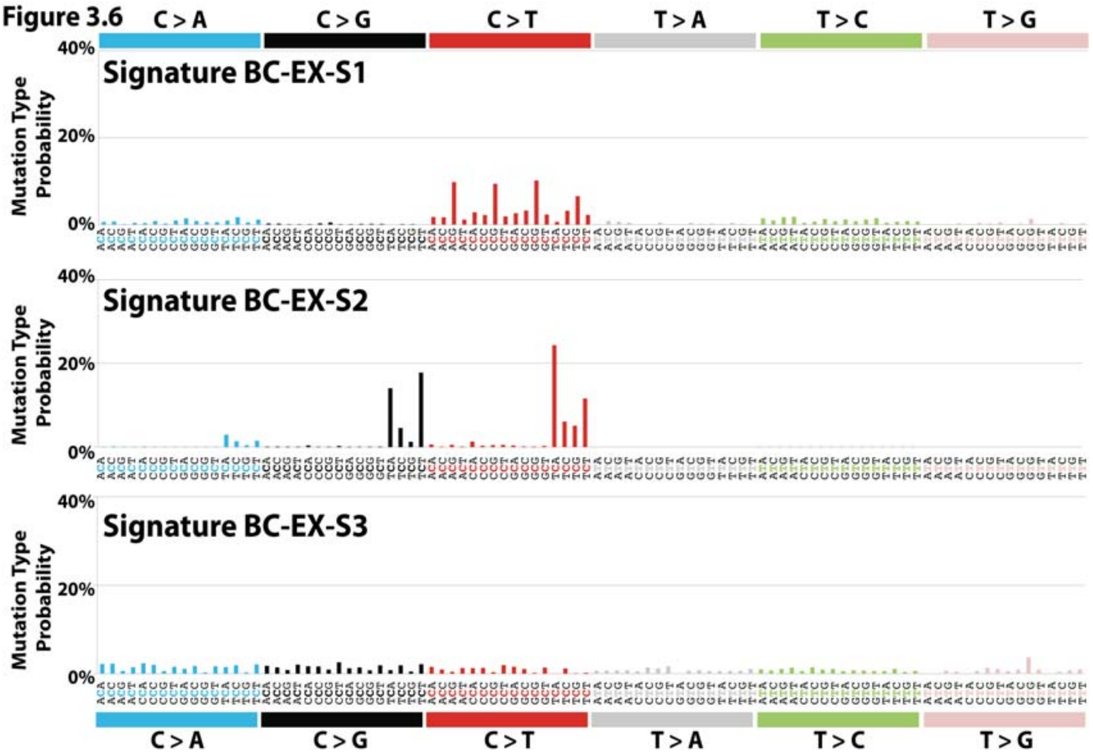


Figure 3.6: Mutational signatures extracted from 884 breast cancer exomes. Signatures of mutational processes are extracted from the mutational catalogues of 884 breast cancer exomes. Three mutational signatures deciphered from the base substitutions (including their immediate 5' and 3' sequence context) identified in the 884 breast cancer exomes. Each signature is depicted on an independent panel, where each type of substitution is displayed in a different colour. Mutational signatures are plotted based on the exome trinucleotide frequency.

Pearson correlation of 0.91. It should be noted that Signature BC-EX-S-1 is extracted from exome sequencing data while Signature BC-WG-S1 is extracted from whole-genome sequencing data. As exome sequencing samples only ~1.5% of the human genome, the examined trinucleotide frequencies in exomes is different than the one found in whole-genome sequencing. Correcting for the trinucleotide frequencies in the exome derived mutational signatures improves the correlation between Signatures BC-WG-S1 and BC-EX-S1 to 0.95.

The pattern of somatic substitutions of Signature BC-EX-S-2 is predominantly C>T, C>G, and C>A mutations at Tp<u>C</u>pN trinucleotides. This exome-extracted signature resembles Signature BC-WG-S2, which is extracted from the mutational catalogues of whole-genomes. Nevertheless, Signature BC-EX-S-2 exhibits a strong preference of C>G mutations at Tp<u>C</u>pN trinucleotides which is not as pronounced as the one in Signature BC-WG-S2. Thus, Signature BC-EX-S-2 is most likely a linear combination between Signatures BC-WG-S2 and BC-WG-S3 (Figure 3.1 and 3.6).

Signature BC-EX-S-3 is characterized by a flat mutational pattern with only subtle features based on the immediate sequence context. This subtle pattern of

Figure 3.7

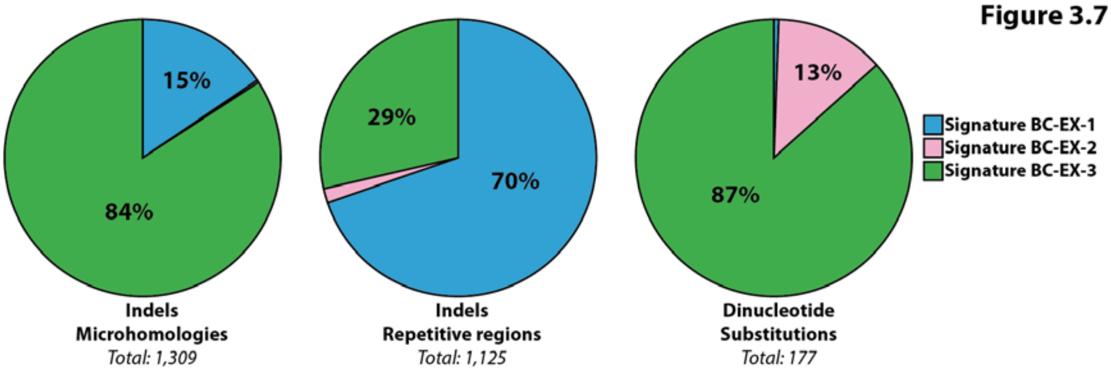


Figure 3.7: Breast cancer exome mutational signatures with indels and dinucleotides. The mutational signatures analysis is extended to incorporate indels at microhomologies, indels at repetitive regions, and dinucleotide substitutions (*i.e.*, Ξ_{99} alphabet). The percentage of mutations attributed to these three additional mutation types is displayed for all signatures that contribute at least 5%. Each signature is displayed in a different colour.

mutations resembles to some degree two of the signatures extracted from whole-genome sequencing data: Signature BC-WG-S-4 (Pearson correlation 0.65, after correcting for trinucleotide context) and Signature BC-WG-S-5 (Pearson correlation 0.49, after correcting for trinucleotide context). Signature BC-EX-S-3 has almost no correlation with any of the other mutational signatures extracted from whole-genome sequencing data. Thus, Signature BC-EX-S-3 is likely a combination of at least two previously identified signatures: Signature BC-WG-S-4 and Signature BC-WG-S-5.

The whole-genome signatures analysis is based on 654,308 somatic mutations and it reveals 6 distinct mutational signatures. In contrast, the exome signatures analysis is based on only 39,480 somatic substitutions and indels (~6% of the whole-genome data) and it reveals only 3 mutational signatures. The performed analyses demonstrate that mutational catalogues from exomes can be used to extract signatures of mutational processes. Furthermore, regardless of the fact that the DNA sequencing and initial bioinformatics analysis of these data were performed by different sequencing centres, the mutational signatures deciphered using exome sequencing are very similar to the ones extracted from whole-genome sequencing data. This illustrates the overall reproducibility of the results together with some vulnerability, particularly when the amount of data are limited or some of the mutational signatures are similar to each other. While using whole-genome sequencing data provides a great resolution for examining common mutational signatures, analysis of smaller, exome derived mutational catalogues (or catalogues from other subcomponents of the genome) may be beneficial as thousands of samples will allow sampling for the activity of mutational processes that are present only in rare cancer cases.

The mutational signatures analysis of breast cancer exomes is extended to evaluate double nucleotide substitutions, indels at microhomologies, and indels at mono/polynucleotide repeats. The results from this analysis are consistent with the indel/dinuc mutational signatures analysis of whole breast cancer genomes (Figure 3.2). Signature BC-EX-3 (which appears to be a mixture of Signatures BC-WG-S-4 and BC-WG-S-5) associated with the majority (>80%) of indels at microhomologies and dinucleotide substitutions as well as with some (~29%) indels at repetitive elements (Figure 3.7). Furthermore, Signature BC-EX-1 accounted for ~70% of indels at repetitive elements (Figure 3.7).

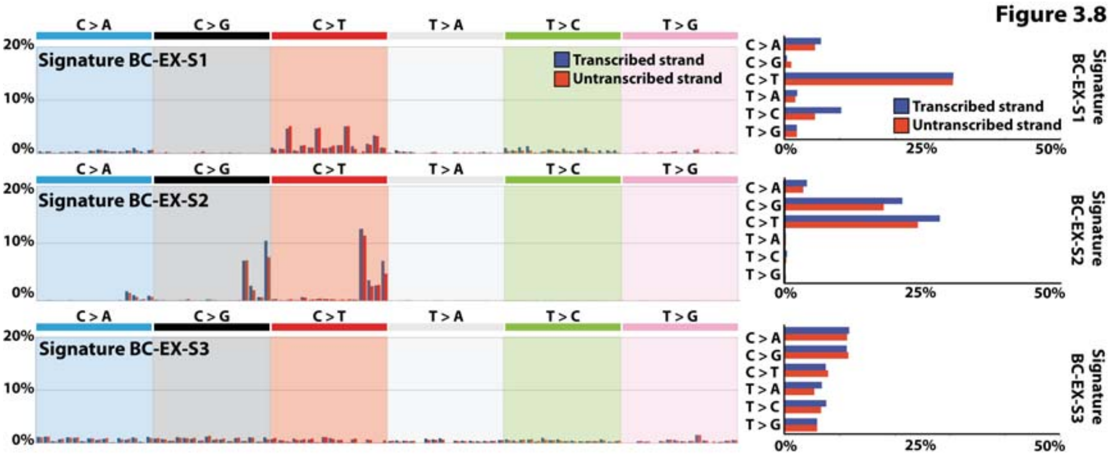


Figure 3.8: Breast cancer exome mutational signatures with strand-bias. Signatures of mutational processes with strand-bias are extracted from the mutational catalogues of 884 breast cancer exomes. Three mutational signatures are deciphered from the base substitutions (including their immediate 5’ and 3’ sequence context). Each signature is depicted on an independent panel, where each type of substitution is highlighted in a different colour. The probability of a mutation to occur on the transcribed strand is depicted in blue, while red is used to display the probability of a mutation to occur on the untranscribed strand. Mutational signatures are plotted based on the exome trinucleotide frequency.

Analysis of smaller, exome derived mutational catalogues (or catalogues from other subcomponents of the genome) may also be useful in detecting biologically revealing features of mutational processes that are particular to coding, transcribed, non-transcribed, or other functionally distinct regions. Consistent with the strand-bias analysis of whole-genome cancer samples, Signature BC-EX-S1 exhibited a weak T>C strand-bias ($Q = 7.2 \times 10^{-4}$). In contrast, no C>A strand-bias is observed in any of the mutational signatures derived from exome sequences (Figure 3.8). This could be due to the lack of somatic mutations to definitively separate Signature BC-EX-S-3 into two distinct mutational signatures. Further, incorporating transcriptional strand in the analysis of the 884 breast cancer exomes reveals strand-bias in BC-EX-S-2 for C>T and C>G mutations with a preference for specific trinucleotide context, *i.e.*, TpCpT (Figure 3.8). However, this strand-bias is not observed in the versions of Signature BC-EX-S-3 (*i.e.*, Signatures BC-WG-S-2 and BC-WG-S-3) extracted from whole cancer genome sequences, which include complete footprints (including introns and untranslated exons) of protein coding genes, suggesting that the underlying mechanism generating strand-bias is restricted to exons (Figures 3.8 and 3.3). Examining only the exon compartments of the whole cancer genome sequences reveals the presence of this strand-bias in samples with substantial exposure to Signature BC-WG-S-2 and/or Signature BC-WG-S-3, supporting this conclusion. This result is biologically surprising and the mechanism underlying this difference in strand-bias between exons and introns is currently unknown.

3.5 Deriving and validating consensus mutational signatures in breast cancer

Figure 3.9

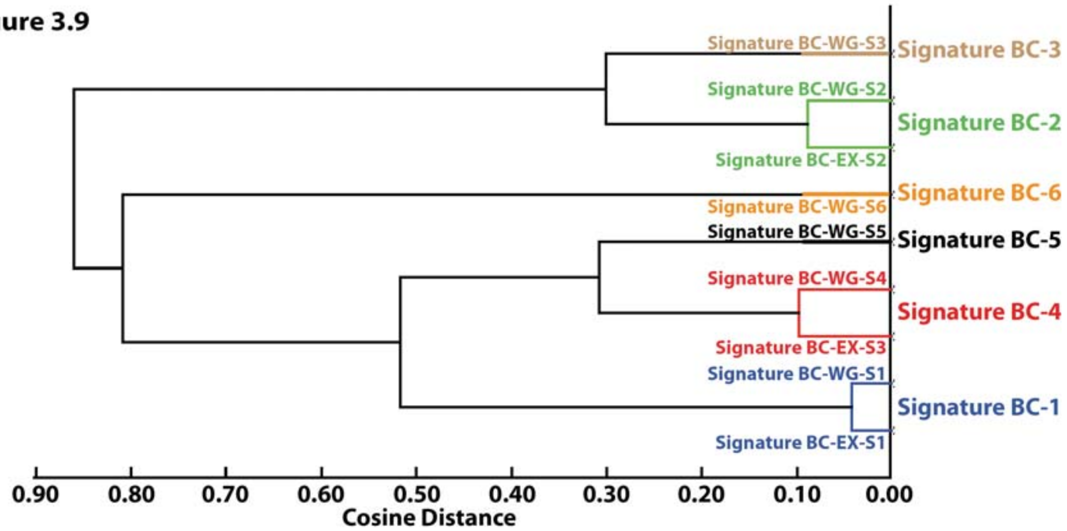


Figure 3.9: Clustering of breast cancer signatures derived from whole-genome and exome data. The originally deciphered mutational signatures are displayed inside the dendrogram near their respective branches. The consensus mutational signatures are displayed on the right-hand side of the dendrogram. Each of the six unique clusters is displayed in a distinct colour. Cosine distance threshold for separating the signatures into clusters is set at 0.09. Note that any threshold between 0.09 and 0.29 results in exactly the same clusters.

In the previous two sections, the signatures of the operative mutational processes in breast cancer are extracted by performing two independent analyses. One encompasses 654,308 somatic substitutions and indels derived from the mutational catalogues of 119 whole breast cancer genomes and reveals the existence of 6 mutational signatures. The second analysis examines only 39,480 somatic mutations from the mutational catalogues of 884 breast cancer exomes and it reveals the existence of 3 mutational signatures. While the patterns of somatic mutations between the signatures extracted from genomes and exomes are very similar, in this section, I use the previous two analyses and leverage unsupervised hierarchical clustering to derive consensus mutational signatures that are operative in breast cancer. The previously extracted 9 mutational signatures (3 from exome mutational catalogues and 6 from genome mutational catalogues) are clustered using a cosine distance (Figure 3.9). The exome derived mutational signatures are re-normalized towards the genome trinucleotide frequency prior to clustering and a threshold of 0.09 is used to separate the original 9 mutational signatures into 6 unique consensus clusters (Figure 3.9).

The value of 0.09 is selected as a conservative measure for the different mutational signatures operative in breast cancer. This threshold is low enough to not cluster mutational signatures with different characteristics (*e.g.*, Signature BC-WG-S5 which exhibits C>A strand-bias and Signature BC-WG-S4 which is associated with indels at microhomologies) and it is high enough to cluster together extremely similar mutational signatures (*e.g.*, Signature BC-EX-S1 and Signature BC-WG-S1, which have a Pearson correlation of 0.95). Nevertheless, this threshold may result in a conservative estimate of the consensus mutational signatures as it may be clustering and mixing together distinct mutational patterns.

Each consensus mutational signature is derived using a weighted average of the signatures belonging to its respective cluster. For example, Signature BC-2 is constructed as a weighted average of genome Signature BC-WG-S2, which accounts for 152,762 somatic mutations, and exome Signature BC-EX-S2, which accounts only for 19,922 somatic mutations (Figure 3.9). As the majority of somatic mutations are found in the whole-genome sequencing data, in this case, the patterns of somatic mutations in the consensus mutational signatures are visually indistinguishable from the ones derived from whole-genome sequencing data (Figure 3.1). Thus, the pattern of mutations of the consensus Signature BC-2 is very similar to the one of Signature BC-WG-S2. It should be noted that the number of mutations attributed to a consensus mutational signature in a sample is set to the number of mutations of the original mutational signature identified in this sample and belonging to the cluster used to derive the consensus mutational signature. For example, Signature BC-2 contributes 69 somatic mutations in exome sample PD6042a as this is the number of somatic mutations attributed to Signature BC-EX-S2 in this sample. In total, Signature BC-2 accounts for 172,684 somatic mutations in the exome and genome breast cancer data (~24.9% of all mutations used in this breast cancer analysis).

In addition to deriving the consensus mutational signatures, in this section, I validate these signatures to check whether any of them might be due to sequencing artefacts or bioinformatics analysis. Validating a mutational signature requires ensuring that a large set of somatic mutations attributed to its pattern is genuine in at least one sample. Validation is complicated as multiple mutational processes are usually operative in most cancer samples, and thus every individual somatic mutation can be probabilistically assigned to several mutational signatures. To overcome this limitation, when possible, I examine the curated dataset for samples that are predominantly generated by one mutational signature (*i.e.*, more than 50% of the somatic mutations in the sample belong to an individual mutational signature) and for which validation data were available. The optimal sample for validating each of the six mutational signatures is identified and a subset of somatic mutations characteristic for this signature (*e.g.*, C>T and C>G substitutions at TpC dinucleotides for Signature BC-2) are chosen for validation through re-sequencing with an orthogonal sequencing technology.

Mutational Signature	Validation Status	Total Mutations in Sample	Total Mutations by Signature	Examined Mutations	Validated Mutations
Signature BC-1	PASS	58	55	58	56 (97%)
Signature BC-2	PASS	76	75	76	72 (95%)
Signature BC-3	PASS	8,612	5,697	200	190(95%)
Signature BC-4	PASS	70	65	70	69 (99%)
Signature BC-5	PASS	4,514	1,558	250	227 (91%)
Signature BC-6	FAIL	11,869	7,955	100	2(2%)

Table 3.2: Validating consensus mutational signatures found in breast cancer. Validation is performed with an orthogonal sequencing approach. The precise validation approach is outlined in the text.

The results reveal that Signatures BC-1, BC-2, BC-3, BC-4 and BC-5 are most likely genuine biological patterns of somatic mutations as they have validation rates of more than 90% (Table 3.2). In contrast, Signature BC-6 is probably due to a sequencing artifact as 98% of the mutations characteristic for this signature (*i.e.*, T>G at GpTpG trinucleotides) failed to validate using an alternative orthogonal sequencing approach. Further investigation into this signature reveals that it is an artifact specific to the configuration of some Illumina sequencing machines at the Wellcome Trust Sanger Institute.

3.6 Prevalence of mutational processes in breast cancer samples

In the previous sections of this chapter, I extract mutational signatures separately from exome and genome sequencing data, and identified the consensus mutational signatures operative in breast cancer. However, the developed computational approach (chapter 2) also allows quantifying the number of somatic mutations attributed to each mutational signature in each cancer sample.

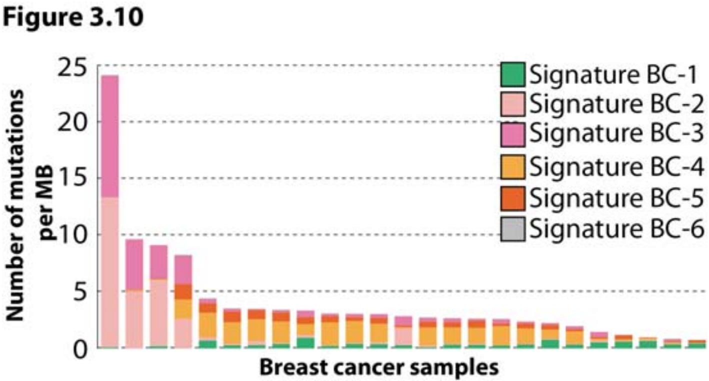


Figure 3.10: Contributions of mutational signatures in a selected set of 25 breast cancer samples. Each sample is displayed as a column with a height corresponding to the number of somatic mutations per megabase found in this sample. Every column is proportionately coloured to reflect the percentage of mutations attributed to different mutational signatures.

An example of a selected set of 25 cancer samples is displayed in Figure 3.10 (note that the contributions of all mutational signatures in all examined cancer samples is provided in Appendix V). This plot reveals the diversity of the activity of the mutational processes underlying the signatures identified in these breast cancer samples. For example, a small minority of samples exhibit a hypermutator phenotype with somatic mutational patterns best explained by Signatures BC-2 and BC-3 (Figure 3.10). A further subset of samples seems to be overwhelmed by the activity of the mutational process underlying Signature BC-4. In contrast, Signature BC-1 is

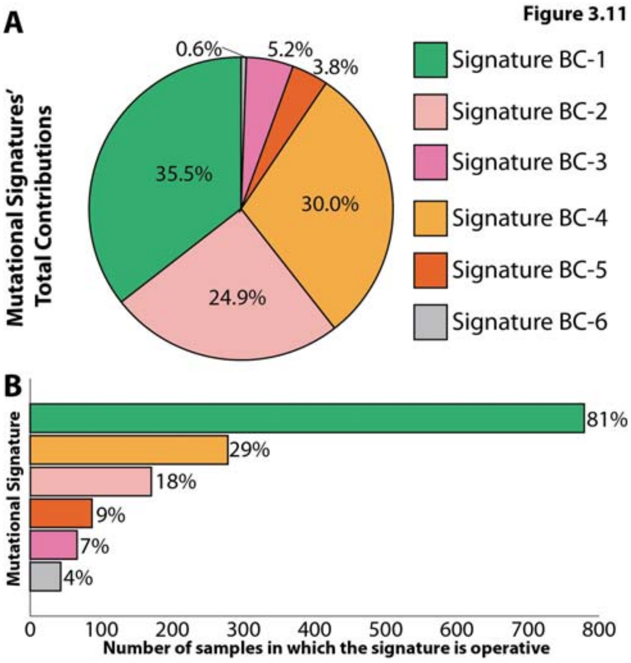


Figure 3.11: Summary of the contributions of the mutational signatures in breast cancer. ***(A)*** Percentage of total mutations contributed by each of the operative mutational signatures. ***(B)*** Percentage and number of samples in which each mutational signature contributes significant number of somatic mutations. For most signatures, significant number of mutations in a sample is defined as more than 100 substitutions or more than 25% of all mutations in that sample. Mutational signatures are displayed in distinct colours.

ubiquitously found at low levels in almost every examined sample (Figure 3.10).

In addition to examining the contributions of mutational signatures at the level of individual samples, one can evaluate the contributions of these signatures across all breast cancer samples and thus provide a mutational signature summary (Figure 3.11). Such an evaluation reveals that while Signature BC-1 accounts for only ~35% of all somatic mutations, it is the most prevalent mutational signature in breast cancer as it is found in 81% of all examined samples (Figure 3.11). In contrast, the next most prevalent signature is Signature BC-4, which is found in only 29% of the samples. Examining the prevalence of mutational signatures across breast cancer samples provides the means to propose etiologies underlying these mutational signatures based on statistical associations.

3.7 Etiology of the consensus mutational signatures in breast cancer

The analysis of breast cancer samples reveals the signatures of 6 distinct mutational processes. However, no molecular mechanisms or etiologies are proposed here for the identified mutational signatures. In principle, several approaches can be leveraged to make propositions for the mechanisms of the underlying mutational mechanisms. In this section, I consider potential mechanisms or underlying causes by comparing signatures with mutation patterns of known causation in the scientific literature or by associating contributions of mutational signatures with epidemiological and biological features specific for breast cancer.

The mutational pattern of Signature BC-1 is predominantly C>T mutations occurring at CpG dinucleotides. This signature is likely due to deamination of 5-methylcytosine, a relatively well-characterized endogenous mutational process present in most normal and neoplastic cells (chapter 1).

Signature BC-2 exhibits predominantly C>T mutations occurring at TpC dinucleotides, while Signature BC-3 generates mostly C>G substitutions occurring at TpC dinucleotides. On the basis of similarities in the sequence context of cytosine mutations caused by *APOBEC* deaminases in experimental systems, these two mutational signatures may be attributable to the activity of *APOBEC1*, *APOBEC3A* and/or *APOBEC3B* (chapter 1). Previous experimental studies have demonstrated that the activity of these proteins results in enzymatic deamination of cytosine to thymine at TpC dinucleotides and it has been speculated that these C>T mutations arise through replication across the uracil. Furthermore, it has been shown that these

Figure 3.12

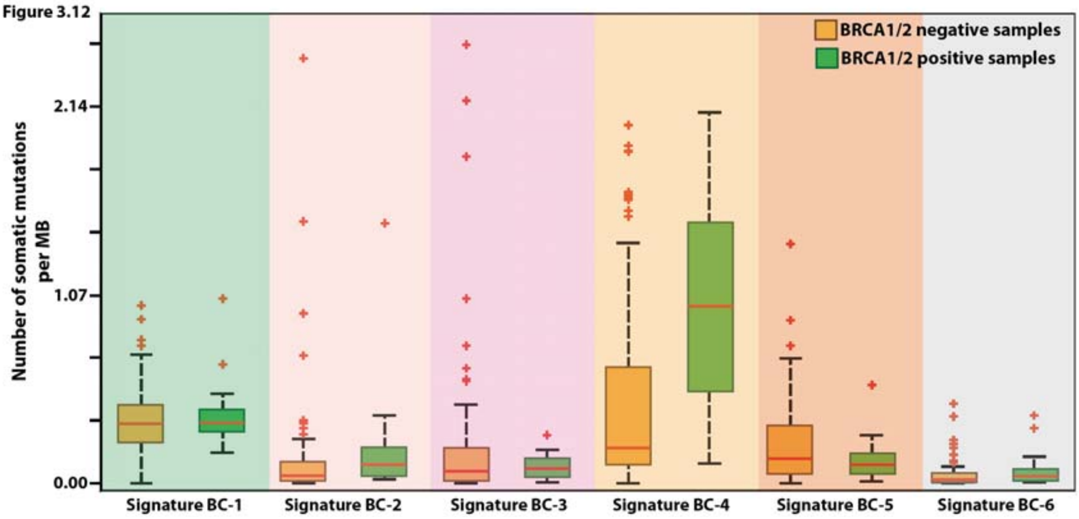


Figure 3.12: Samples harbouring *BRCA1/2* mutations and contributions of mutational signatures. Samples are separated into two sets: *BRCA1/2* positive samples (*i.e.*, with BRCA1/2 mutations, green) and *BRCA1/2* negative samples (*i.e.*, without BRCA1/2 mutations orange). A box plot of the mutations contributed by each mutational signature is displayed for each of the two sets. Outliers with more than 2.5 mutations per megabase are not shown but they are included in the statistical analysis. The only statistically significant difference in signature's contributions between the *BRCA1/2* positive and negative sets is the one due to Signature BC-4 (Q = 1.6 × 10^{-8}).

deaminases can also generate C>G substitutions at Tp<u>C</u> dinucleotides and it has been suggested that this mutational pattern is generated when an *APOBEC* deaminated cytosine is excised by uracil-DNA glycosylase with subsequent non-templated DNA replication across the abasic site by *REV1* (Taylor et al., 2013). Thus, Signature BC-2 is likely due to the activity of the *APOBEC* family of deaminases, while Signature BC-3 encompasses an interaction between *APOBEC* enzymes and *REV1*.

Substantial numbers of larger deletions (up to 50 bp) with overlapping microhomology at breakpoint junctions are found in some breast cancer samples with major contributions from Signature BC-4 (Figure 3.2). A subset of breast cancer cases is known to be due to inactivating mutations in *BRCA1* and *BRCA2*, and the presence of Signature BC-4 is strongly associated (Q = 1.6 × 10^{-8}) with *BRCA1* and *BRCA2* mutations (Figure 3.12). No other mutational signature associated with the numbers of mutations in samples harbouring *BRCA1* and/or *BRCA2* mutations (Figure 3.12). *BRCA1* and *BRCA2* are implicated in homologous-recombination-based DNA double-strand break repair. Abrogation of their functions results in recruitment of non-homologous end-joining mechanisms, which can use microhomology at rearrangement junctions to re-join double-strand breaks, to take over DNA double-strand break repair. Indeed, almost all cases with *BRCA1* and *BRCA2* mutations showed a large contribution from Signature BC-4. However, some cases with a substantial contribution from Signature BC-4 do not have *BRCA1* and *BRCA2* mutations, suggesting that other mechanisms of *BRCA1* and *BRCA2* inactivation or abnormalities of other genes may also generate this mutational signature.

Evaluating the enrichment of mutational signatures based on the molecular subtypes of breast cancer reveals that estrogen receptor negative breast cancer samples have significantly higher numbers of mutations due to Signature BC-4, Q = 7.9×10^{-5}, and Signature BC-5, Q = 1.6×10^{-6} (Figure 3.13). No other molecular subtype associated with the numbers of somatic mutations attributed to any other mutational signature. Estrogen receptor negative breast cancer samples are enriched for *BRCA1* and *BRCA2* mutations. To evaluate whether the differences of contributions of mutational signatures are due to *BRCA1/2* mutations, these samples are re-examined after stratification. *BRCA1/2* wild-type samples do not show statistically significant differences based on their estrogen receptor status for Signature BC-4 (Q = 0.09). However, estrogen receptor negative *BRCA1/2* wild-type samples have significantly higher numbers of mutations attributable to Signature BC-5 when compared to estrogen receptor positive *BRCA1/2* wild-type breast cancers (Q = 3.8×10^{-3}).

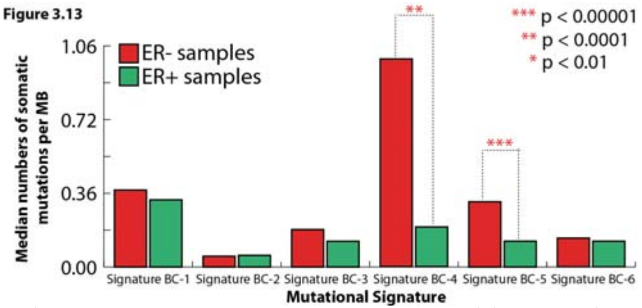


Figure 3.13: Estrogen receptor positive/negative samples and contributions of mutational signatures. Samples are separated into two sets: estrogen receptor negative samples (red) and estrogen receptor negative samples (green). The distributions of somatic mutations between the two sets are compared for each of the mutational signatures.

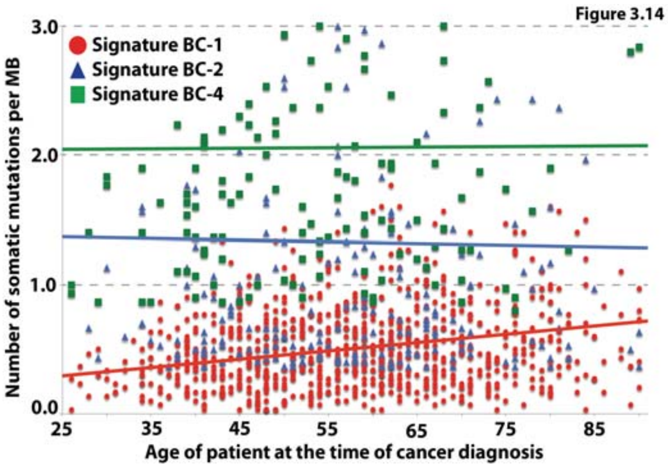


Figure 3.14: Age of diagnosis and mutations due to different mutational signatures. Each sign corresponds to a contribution of a given mutational signature for a patient at a given age. From the six mutational signatures identified in breast cancer, only Signature BC-1 (shown in red) correlates with age of diagnosis. Signatures BC-2 (blue) and BC-4 (green) are shown to illustrate the lack of correlation of other mutational signatures.

The performed validation experiments (Table 3.2) indicate that Signature BC-6 is most likely a centre specific sequencing artifact.

Lastly, I evaluate the correlations between age of diagnosis and the number of mutations attributable to each signature in each sample. Only Signature BC-1 exhibited a strong positive correlation with age of diagnosis, Q = 1.5×10^{-8} (Figure 3.14). The mutations in a cancer genome may be acquired at any stage in the cellular lineage from the fertilized egg to the sequenced cancer cell. The correlation with age of diagnosis is consistent with the hypothesis that a substantial proportion of Signature BC-1 mutations in cancer genomes have been acquired over the lifetime of the cancer patient, at a relatively constant rate that is similar in different people, probably in normal somatic tissues.

3.8 Discussion

In this chapter of the thesis, I examine the mutational catalogues of 119 breast cancer genomes as well as 884 breast cancer exomes. Mutational signatures are deciphered separately from genome and exome sequencing data. The signatures analysis incorporated somatic single base substitutions and their immediate sequencing context as well as indels at mono/polynucleotide repeats, indels at microhomologies, and dinucleotide substitutions.

The identified genome-based and exome-based mutational signatures are used to derive the 6 consensus breast cancer signatures. Validation using an orthogonal sequencing technology reveals that one of these mutational signatures is most likely due to a sequencing artifact, while the remaining five are most likely genuine. An etiology is proposed for each of these five mutational signatures based on similarities of the mutational patterns with experimental data previously reported in the literature or a statistical association with a specific molecular phenotype.

Lastly, it should be noted that one of the objectives of this chapter is to serve as an exemplar for performing mutational signatures analysis in a cancer type. The next chapter presents analogous analyses performed for another 29 types of human cancer.