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relies on previously identified somatic mutations by curating freely available published data as well as data that was unpublished at the time. Any unpublished breast cancer data were generated internally at the Cancer Genome Project (CGP) for the purposes of other projects.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 In contrast, the majority of breast cancer whole-genomes are previously unpublished data.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New unpublished samples	5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New unpublished samples	98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 , , , , and 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 (/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.

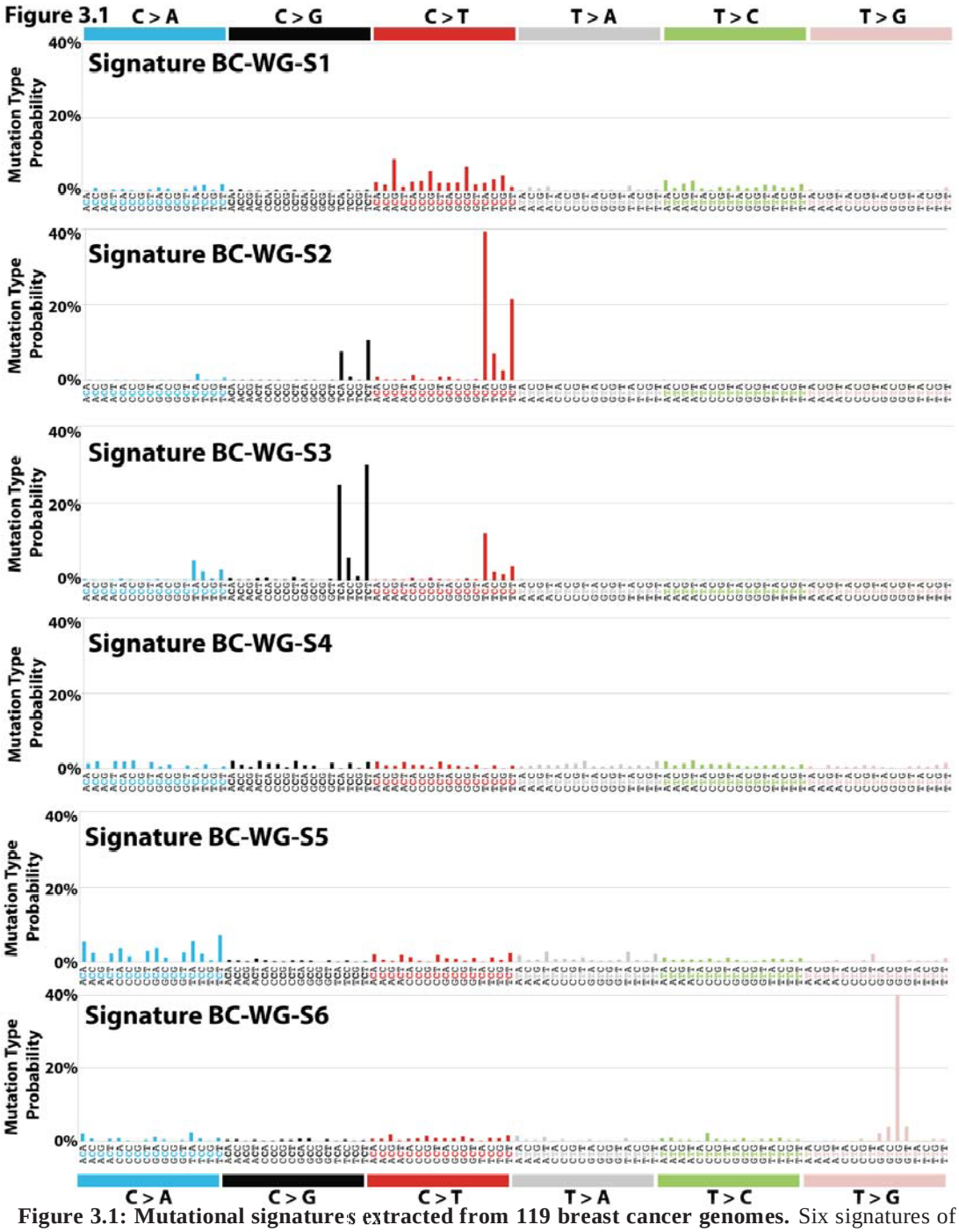

Figure 3.1: Mutational signatures extracted from 119 breast cancer genomes. Six signatures of mutational processes are deciphered from the base substitutions (including their immediate 5' and 3' sequence context) identified in the examined 119 breast cancer genomes. Each signature is depicted on an independent 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 , , and 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 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