Chapter 7

Materials and methods

7.1 Introduction

This chapter provides further details about the materials and methods used. As one of the main results of this thesis is the development of a novel method for analysing patterns of somatic mutations, the majority of materials and methods have already been presented in chapter 2. Thus, to avoid repetition, this chapter only discusses additional methods that were used through this thesis. It should be noted that I did not personally perform any DNA sequencing or mutation identification but I rather relied on somatic mutations previously identified by others. Thus, this chapter will not cover any experimental procedures for DNA sequencing or bioinformatics algorithms for identifying somatic mutations from next-generation sequencing data.

7.2 Deciphering signatures of mutational processes

Mutational signatures are deciphered independently for each of the 30 cancer types using the previously developed computational framework. The algorithm deciphers the minimal set of mutational signatures that optimally explains the proportion of each mutation type found in each catalogue and then estimates the contribution of each signature to each catalogue. Mutational signatures are also extracted separately for genomes and exomes. Mutational signatures extracted from exomes are normalized using the observed trinucleotide frequency in the human exome to the trinucleotide frequency of the human genome. All mutational signatures are clustered using unsupervised agglomerative hierarchical clustering and a threshold is selected to identify the set of consensus mutational signatures. Misclustering is avoided by manual examination and, whenever necessary, re-assignment of all

signatures in all clusters. 27 consensus mutational signatures are identified across the 30 cancer types. The computational framework for deciphering mutational signatures as well as all the data used in this study are freely available and can be downloaded from:

http://www.mathworks.com/matlabcentral/fileexchange/38724

7.3 Displaying mutational signatures

Mutational signatures are displayed using a 96 substitution classification defined by the substitution class and the sequence context immediately 5' and 3' to the mutated base. Mutational signatures are displayed in the main text (unless otherwise specified) based on the observed trinucleotide frequency of the human genome, *i.e.*, representing the relative proportions of mutations generated in each signature based on the actual trinucleotide frequencies of the reference human genome.

7.4 Filtering and generating mutational catalogues

In all examined samples, normal DNAs from the same individuals are sequenced to establish the somatic origin of variants. Extensive filtering is performed to remove any residual germline mutations and technology specific sequencing artefacts prior to analysing the data. Germline mutations are filtered out from the lists of reported mutations using the complete list of germline mutations 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/). Technology specific sequencing artefacts are filtered out by using panels of BAM files of (unmatched) normal tissues containing more than 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 are discarded. The remaining somatic mutations are used for generating a mutational catalogue for every sample.

The immediate 5′ and 3′ sequence context is extracted using the ENSEMBL Core APIs for human genome build GRCh37. Curated somatic mutations that originally mapped to an older version of the human genome are re-mapped using UCSC's freely available lift genome annotations tool (any somatic mutations with ambiguous or missing mappings are discarded). 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 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. Genomic regions of bidirectional transcription are excluded from the strand bias analysis.

7.5 Statistical evaluation of associations

Generalized linear models (GLMs) are used to fit signatures' exposures (*i.e.*, number of mutations assigned to a signature) and the age of cancer diagnoses. For each cancer type, all mutational signatures operative in it are evaluated using GLMs. The Benjamini–Hochberg false discovery rate (FDR) procedure is used to adjust for multiple hypothesis testing and in all cases q-values are reported.

Associations between all other etiologies and signature exposures are performed using two-sample Kolmogorov-Smirnov tests between two sets of samples. The first set encompasses the signature exposures of the samples with the "desired feature" (*e.g.*, samples that contain immunoglobulin gene hypermutation) and the second set encompasses the signature exposures of the samples without the "desired feature" (*e.g.*, samples that do NOT contain immunoglobulin gene hypermutation). Samples with unknown features status (*e.g.*, not knowing the hypermutation status of the immunoglobulin gene) are ignored. Kolmogorov-Smirnov tests are performed for all signatures and all examined "features" in a cancer type. Similarly, the Benjamini–Hochberg false discovery rate (FDR) procedure is used to adjust for multiple hypothesis testing in a particular cancer class and in all cases q-values are reported.