Chapter 5

Etiology of mutational processes operative in human cancer

5.1 Introduction

The previous chapter of this thesis presented 27 consensus mutational signatures that were extracted from the cancer genomes of 7,042 patients across 30 distinct classes of human cancer. The chapter discussed the mutational patterns of the derived consensus signatures; however, no propositions were made about potential endogenous or exogenous mutational processes associated with any of these patterns. The aim of the present chapter is to suggest etiology for the molecular and/or environmental processes underlying at least some of these mutational signatures. These suggestions will be based on either comparing the spectrum of a mutational signature with mutational patterns of known causation or by statistically associating a signature with epidemiological, biological, or molecular features specific for each of the cancer types in which the signature has been identified.

5.2 Associating cancer etiology and mutational signatures based on mutational patterns with known causation

Each mutational signature is the imprint left on a cancer genome by a mutational process that may include one or more DNA damage and/or DNA maintenance mechanisms, with the latter either functioning normally or abnormally. Here, I consider probable mechanisms or underlying causes of the identified signatures by comparing signatures with mutation patterns of known causation in the scientific literature.

Signature 1A and Signature 1B exhibit a very similar mutational pattern. This pattern is likely related to the relatively elevated rate of spontaneous deamination of 5-methylcytosine which results in C>T transitions and which predominantly occurs at NpCpG trinucleotides (Pfeifer, 2006). As discussed in chapter 4, this mutational process operates both in the germline and in somatic cells (Welch et al., 2012). Thus, Signature 1A/B is probably due to spontaneously occurring endogenous mutational processes present in most normal and neoplastic cells that are initiated by deamination of 5-methylcytosine (Pfeifer, 2006). Other signatures are likely attributable to exogenous mutagenic exposures or failure of cellular molecular mechanisms.

The mutational patterns of Signature 2 and 13 are similar as they are both composed of C>A, C>T, and C>G substitutions at TpC dinucleotides. In chapter 3, I proposed that Signature 2 could be attributed to the activity of the *AID/APOBEC* family of cytidine deaminases, while Signature 13 encompasses an interaction between *APOBEC* enzymes and the DNA repair protein *REV1*. On the basis of similarities in the sequence context of cytosine mutations caused by *APOBEC* enzymes in experimental systems, a role for *APOBEC1*, *APOBEC3A* and/or *APOBEC3B* in human cancer appears more likely than for other members of the family (Burns et al., 2013; Harris et al., 2002; Taylor et al., 2013). Furthermore, recent studies have demonstrated that there is an association between the observed patterns of somatic mutations and the expression of *APOBEC3B* (Burns et al., 2013; Taylor et al., 2013). However, the reason for extreme activation of this mutational process, such as Signatures 2 and/or 13 hypermutated samples with up to 25 somatic mutations per megabase, remains unknown. Since *APOBEC* activation constitutes part of the innate immune response to viruses and retrotransposons (Koito and Ikeda, 2013) it may be that these mutational signatures represent collateral damage on the human genome from a response originally directed at retrotransposing DNA elements or exogenous viruses. Confirmation of this hypothesis would establish an important new mechanism for initiation of human carcinogenesis. However, it is plausible that entirely different mechanisms (both endogenous and/or exogenous) are activating the *APOBEC* enzymes.

In smoking-associated lung cancer, C>A transversions are the predominant known mutational pattern induced by tobacco carcinogens (Pfeifer et al., 2002). It is

believed that this type of substitutions is due to the formation of bulky adducts on guanine. Furthermore, previous studies have shown that the tobacco carcinogenic lesions occurring on the transcribed strand are correctly identified and removed by transcription-coupled nucleotide excision repair resulting in strong transcriptional strand-bias on a genomic scale (Pfeifer et al., 2002; Pleasance et al., 2010b). In the previous chapter, I demonstrated that Signature 4 generates predominantly C>A substitutions and that it possesses a strong transcriptional strand-bias (chapter 4). Furthermore, this signature is present in cancer types with a well-known association to tobacco smoking: lung adenocarcinoma, lung squamous, small cell lung carcinomas, head and neck squamous, and liver cancers (Figure 4.9). Thus, it is reasonable to causally associate Signature 4 with tobacco smoking. This association will be further refined using statistical analysis in the next section (see below).

Signature 7 is the predominant mutational signature found in malignant melanoma. This signature bears a mutational pattern that is expected from ultraviolet light: C>T and CC>TT mutations at dipyrimidines (chapter 4). Moreover, as expected from a mutational pattern of ultraviolet light, Signature 7 exhibits a strong transcriptional strand-bias indicating that mutations occur at pyrimidines (*viz.*, by formation of pyrimidine-pyrimidine photodimers) and these mutations are being effectively repaired by transcription-coupled nucleotide excision repair. In addition to malignant melanoma, this mutational pattern is also found in two cases of squamous carcinoma of the head and neck. Further examination revealed that both these head and neck cases are the only two cancers of the lip in the dataset. Indeed, lip cancers have been previously associated with exposure to ultraviolet light (Pfeifer et al., 2002). Based on the similarity of the mutational pattern to the one observed in experimental systems exposed to ultraviolet light and the presence of Signature 7 in ultraviolet associated cancers (*viz.*, lip cancer and malignant melanoma), Signature 7 is most likely due to exposure to ultraviolet light.

Some anticancer drugs are mutagens that have specific patterns of somatic mutations (Hunter et al., 2006). Signature 11 has mutational features very similar to those previously reported in experimental studies of alkylating agents (Hunter et al., 2006). Further analysis will be performed in the next section to statistically associate Signature 11 with a specific cancer treatment.

Abnormalities in DNA maintenance may also be responsible for mutational signatures. Previous studies have demonstrated that defective DNA mismatch repair results in highly elevated numbers of somatic mutations and exhibits significant numbers of small (1bp and 2bp long) insertion and/or deletions (indels) predominantly found at repetitive elements (Tomita-Mitchell et al., 2000). Further, microsatellite unstable tumours are characteristic for colorectal, uterine, and stomach cancers. Taken together, these observations are consistent with the behaviours and patterns of three of the identified mutational signatures: Signature 6, Signature 15, and Signature 20. Thus, it is plausible that Signatures 6, 15, and 20 are due to the failure of one or more of the molecular mechanisms of DNA mismatch repair. In the next section, I will statistically demonstrate that at least one of these mutational signatures is highly elevated in microsatellite unstable samples.

Defective repair of DNA double-strand breaks based on homologous recombination has also been known to cause an elevated numbers of large indels with overlapping microhomology at breakpoint junctions (chapter 1). This pattern of mutations is consistent with the behaviour of Signature 3. Further, in chapter 3, Signature 3 is statistically associated with failure of homologous recombination in breast cancer due to mutations in *BRCA1* and/or *BRCA2*. In a latter section, I will demonstrate that this statistical association also holds for pancreatic and ovarian cancers.

Mutational signatures may also result from the abnormal function of enzymes that modify DNA or the activity of error-prone polymerases. Previous studies have demonstrated that the activity *POL η*, an error prone polymerase involved in processing *AID* induced cytidine deamination, results in an excess of T>G transversions at ApTpN and TpTpN trinucleotides in chronic lymphocytic leukaemias with mutated immunoglobulin genes (Di Noia and Neuberger, 2007; Puente et al., 2011). This pattern of mutations is consistent with Signature 9, which is found in chronic lymphocytic leukaemia and malignant B-cell lymphomas (Figure 4.9).

Similarly, previous studies have associated recurrent somatic mutations altering the functions of the error-prone polymerase *POL ε* (*POLE*) with a subset of colorectal and uterine tumours that exhibit an ultra-hypermutator phenotype. This

behaviour is consistent with Signature 10, which is found in cancers of the colorectum and uterus with an extremely high prevalence of somatic mutations.

Many of the validated mutational signatures do not, however, have an established or proposed underlying mutational process or etiology. Some, for example Signatures 8, 12 and 16, show strong transcriptional strand-bias (Figure 4.8) and possibly reflect the involvement of transcription-coupled nucleotide excision repair acting on bulky DNA adducts due to exogenous carcinogens. Others, for example Signatures 14 and 21, show an overwhelming activity in a small number of cancer cases and are perhaps due to currently uncharacterized defects in DNA maintenance or abnormal activity of DNA polymerases.

In addition to the 22 validated consensus mutational signatures, there are another 5 consensus signatures identified through extraction of mutational signatures. The mutational patterns of Signature U1 and Signature U2 (Figure 4.7) are too uniform and unspecific to unambiguously associate them with any previously published patterns of somatic mutations. In contrast, the mutational patterns of Signatures R1, R2, and R3 are extremely specific and sequence-context dependent (Figure 4.6). Further, as discussed in chapter 4, these artifactual mutational signatures seem to be confined to data generated within specific sequencing centres. Signature R1 is associated with the next generation sequencing protocol used at the Wellcome Trust Sanger Institute. This protocol has been optimized to avoid the generation of this signature. Signature R2 is present in data from the Broad Institute and in-depth investigation revealed its pattern of mutations is due to the generation of 8-oxoguanine during DNA shearing (Costello et al., 2013). Lastly, Signature R3 is confined to colorectal data generated by the Baylor College of Medicine. After investigation, this pattern is attributed to the settings of the used bioinformatics pipelines, which are set to call somatic mutations from only a few reads in genes previously associated with colorectal cancer.

5.3 Associating cancer etiology and mutational signatures based on statistical analysis

In the previous section, a mutational signature is causally associated with a potential etiology based on the similarity of its pattern to mutational patterns of

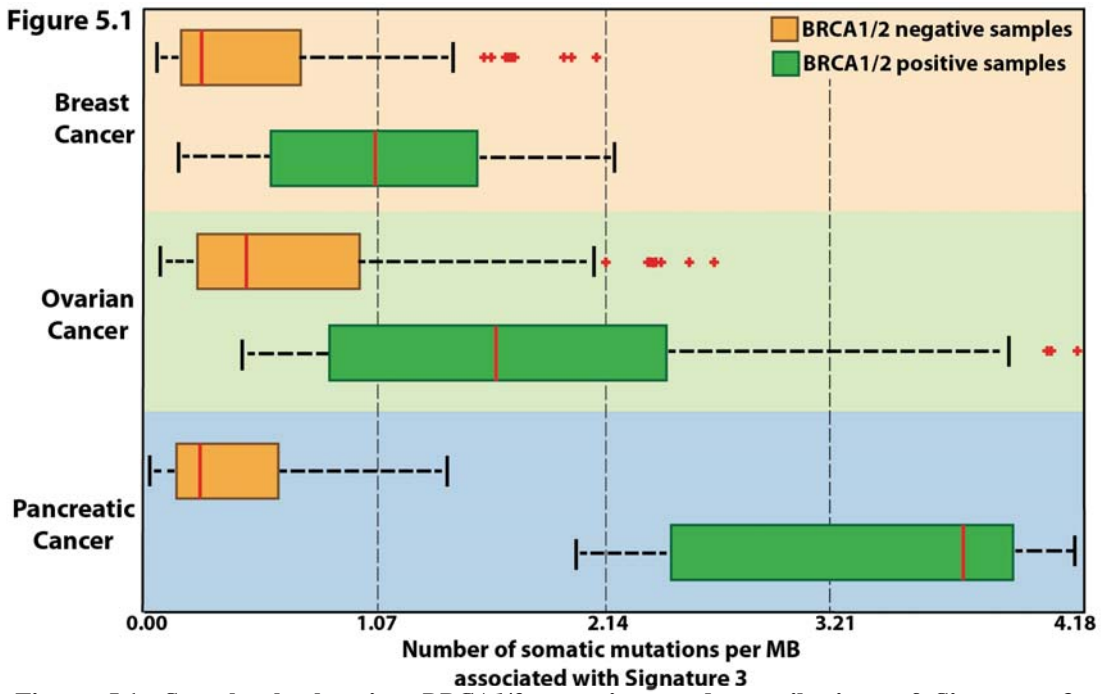

Figure 5.1: Samples harbouring *BRCA1/2* mutations and contributions of Signature 3. Signature 3 is examined in breast, ovarian, and pancreatic cancers. In each cancer type samples are separated into two sets: *BRCA1/2* positive samples (green) and *BRCA1/2* negative samples (orange). A box plot of the mutations contributed by Signature 3 in each cancer type is displayed for each of the two sets. Outliers with more than 4.18 mutations per megabase are not shown but they are included in the statistical analysis. All consensus mutational signatures are evaluated for statistical association with *BRCA1/2* in their respective cancer types. The only statistically significant difference in signatures' contributions between the *BRCA1/2* positive and negative sets is the one due to Signature 3 ($Q = 1.6 \times 10^{-8}$ for breast cancer; $Q = 2.3 \times 10^{-7}$ for ovarian cancer; $Q = 0.02$ for pancreatic cancer).

known causation in the scientific literature. This section will focus on re-confirming (or identifying new) associations via statistical analysis. Briefly, a cancer type is split based on a feature of interest (*e.g.*, smoking status separating lung adenocarcinomas in smokers and non-smokers) and statistical analysis is performed for all signatures found in that cancer type. The analysis checks whether mutations attributed to the signature in question are statistically different between the set of samples possessing the feature (*e.g.*, smokers) and the set of samples without the feature (*e.g.*, non-smokers). Any samples with missing information about a selected feature (*e.g.*, when the smoking status is unknown) are ignored. In all cases, q-values are reported for all statistically significant associations between a signature and a feature of interest. In most cases, only a single mutational signature associates with a particular feature. Features of interest are selected based on prior biological knowledge or based on advice from collaborators who are experts in a specific cancer type.

Previous analysis of breast cancer data demonstrated that samples harbouring *BRCA1* and/or *BRCA2* mutations have an elevated numbers of somatic mutations attributable to Signature 3 (Figure 3.12). Mutations associated to other mutational

signatures found in breast cancer are not statistically different between *BRCA1/2* wild type samples and *BRCA1/2* mutants (Figure 3.12). Analogous analysis is performed for the two additional cancer types in which Signature 3 is found: ovarian and pancreatic cancer (Figure 4.9). The subset of cases from these three cancer classes, known to be due to inactivating mutations in *BRCA1* and *BRCA2*, is strongly associated with the presence of Signature 3 ($Q = 1.6 \times 10^{-8}$ for breast cancer; in all cases Q refers to a q-value, see chapter 7; $Q = 2.3 \times 10^{-7}$ for ovarian cancer; $Q = 0.02$ for pancreatic cancer; Figure 5.1 and Figure 5.3). Similarly to breast cancer, no other mutational signature associated with the BRCA1/2 status in pancreatic and ovarian cancers. Interestingly, every single pancreatic cancer that harboured *BRCA1/2* mutations exhibited an extremely elevated mutational burden for Signature 3. Indeed, almost all cases with *BRCA1* and *BRCA2* mutations in breast and ovarian cancers also showed a large contribution from Signature 3. However, some ovarian and breast cancers with a substantial contribution from Signature 3 do not have *BRCA1/2* mutations, which suggests that other mechanisms of *BRCA1/2* inactivation or abnormalities of other genes may also generate the mutational pattern.

BRCA1 and *BRCA2* are implicated in homologous recombination-based DNA double-strand break repair (Thompson, 2012). The abrogation of their functions results in non-homologous end-joining mechanisms, which can utilize microhomology at rearrangement junctions to re-join double-strand breaks, taking over DNA double-strand break repair. The results show that, in addition to the genomic structural instability conferred by defective double-strand break repair, a base substitution mutational signature is associated with *BRCA1/2* deficiency in three distinct cancer types.

The statistical analysis performed in chapter 3 associated Signature 8 with estrogen receptor negative breast cancer samples. Signature 8 is also found in medulloblastoma (Figure 4.9); however, the mutations attributed to this mutational signature do not associate with any molecular subtype of medulloblastoma.

In the previous section, a causal association is proposed between tobacco smoking and Signature 4 based on the similarity between the mutational pattern of the signature and the mutational pattern observed in experimental systems exposed to tobacco carcinogens. This relationship is supported by a strong elevation of the

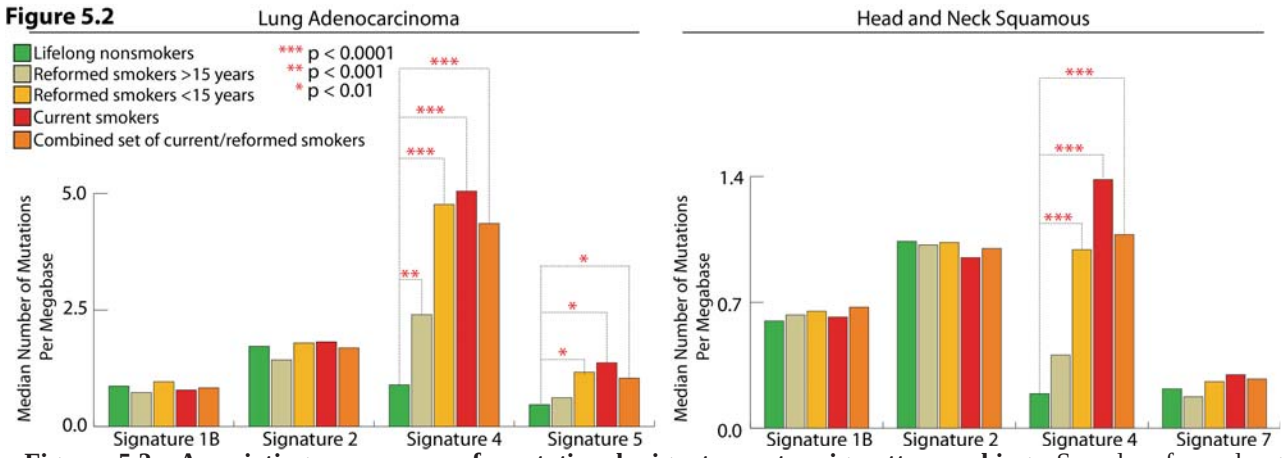

Figure 5.2: Associating exposures of mutational signatures to cigarette smoking. Samples from lung adenocarcinomas and head and neck squamous are examined. Each of the two cancer types is separated in 5 categories: lifelong non-smokers (dark green); reformed smokers for more than 15 years (light green); reformed smokers for less than 15 years (yellow); current smokers (red); a combined set containing all current and reformed smokers (orange). Statistical analysis is performed for every mutational signature by comparing the set of non-smokers with the other four sets. All reported p-values have been adjusted for multiple hypothesis testing.
The X-axis depicts the mutational signatures operative in the respective cancer types, while the Y-axis reflects the median numbers of somatic mutations attributed to each signature in each of the five categories. Note that the two Y-axes have a different scale.

mutations attributed to Signature 4 in current smokers when compared to non-smokers ($Q = 1.1 \times 10^{-7}$ for lung adenocarcinomas; $Q = 2.4 \times 10^{-5}$ for head and neck squamous; Figure 5.2). Further, there is even a statistically significant difference between the numbers of mutations attributed to Signature 4 in lung adenocarcinomas from non-smokers when compared to the mutations found in adenocarcinomas from people who stopped smoking more than fifteen years prior to their tumour diagnosis (Figure 5.2). This association is not found in head and neck cancers; however, that might be partly explained by the low number of head and neck squamous cancers from patients that stopped smoking more than 15 years prior to their diagnosis. At the very least, this result confirms that tobacco smoking leaves a strong and long lasting mutational imprint on the genome of a lung cancer.

Cigarette smoke contains over 60 carcinogens (Pfeifer et al., 2002) and it is possible that this complex mixture may initiate other mutational processes. Signature 1B, 2, and 7 are identified in head and neck squamous but they do not associate with the smoking statuses of the examined patients (Figure 5.2). However, Signature 5, but not Signatures 1A/B and 2, also showed a positive correlation between smoking history and mutation contribution in lung adenocarcinomas ($Q = 8.0 \times 10^{-3}$, Figure 5.2). Thus, in lung cancer, Signature 5 may also be generated by tobacco carcinogens.

From the carcinogens present in tobacco smoke, vinyl chloride and ethyl carbamate have been reported to generate the T>C mutations characteristic of Signature 5 (Pfeifer et al., 2002). However, Signature 5 is also present in nine other cancer types, most of which are not strongly associated with tobacco consumption, and therefore its overall etiology remains unclear (Figure 4.9).

The mutational pattern of Signature 6's indels, often termed "microsatellite instability", is characteristic of cancers with defective DNA mismatch repair (Boland and Goel, 2010). Consistent with this explanation, the presence of Signature 6 is strongly associated with the inactivation of DNA mismatch repair genes in colorectal cancer ($Q = 3.3 \times 10^{-5}$ for colorectal cancers; Figure 5.3).

Signature 9 is observed in chronic lymphocytic leukaemia and malignant B-cell lymphomas. This signature is characterized by a pattern of mutations that has been attributed to polymerase η, which is implicated with the activity of *AID* during somatic hypermutation (Puente et al., 2011). Examining chronic lymphocytic leukaemias that possess immunoglobulin gene hypermutation (IGHV-mutated) reveals a statistically significant elevation of Signature 9 ($Q = 2.5 \times 10^{-4}$; Figure 5.3). This analysis is not performed for B-cell lymphomas due to the lack of sufficient number of IGHV-mutated samples. Nevertheless, only one of the B-cell lymphomas is IGHV-mutated and this sample exhibits an extremely high level of Signature 9 (Appendix V).

Signature 10 generates huge numbers of mutations in subsets of colorectal and uterine cancers. It has been proposed that the mutational process underlying this signature is due to the altered activity of the error-prone polymerase *POLE*. To support this hypothesis, a high number of recurrent function modifying somatic mutations, *viz.*, Pro286Arg and Val411Leu, have been observed in *POLE* in colorectal and uterine samples with high mutational burden (Kandoth et al., 2013; TCGA, 2012). Statistical analysis reveals an extremely strong association between these recurrent somatic mutations and the contributions of Signature 10 ($Q = 3.1 \times 10^{-22}$ for colorectal cancer; $Q = 8.8 \times 10^{-9}$ for uterine cancer; Figure 5.3).

Signature 11 exhibits a mutational pattern resembling the one of an alkylating agent and this signature is identified in malignant melanoma and glioblastoma

multiforme. Examining information from the patients' histories revealed a statistical association between treatments with the alkylating agent temozolomide in both cancer

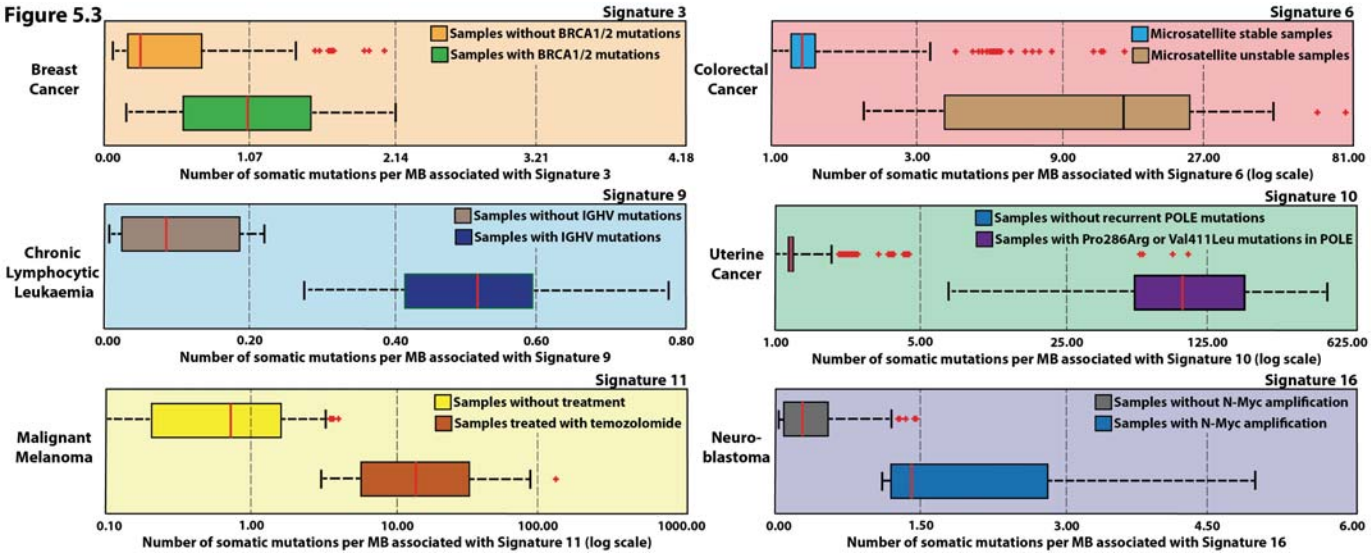

Figure 5.3: **Associating molecular or clinical features with the activity of mutational signatures**. In each case, all signatures found in a given cancer type are evaluated for a potential association with a selected feature. Only a single mutational signature associates with a selected feature in each of the examined cases. Each boxplot represents numbers of somatic mutations for a given signature for samples possessing or lacking a specific feature in a specific cancer type. The examined cancer type is annotated on the left of each panel, while the evaluated mutational signature is displayed in the upper right corner of each panel. In all cases, the X-axis depicts the number of mutations per megabase attributable to a give signature. Note that a logarithmic scale is used for the X-axes of Signatures 6, 10, and 11. For clarity, some outliers are not displayed but all data are included in the statistical analysis.

types ($Q = 4.0 \times 10^{-3}$ for malignant melanomas; $Q = 2.1 \times 10^{-3}$ for glioblastoma multiforme; Figure 5.3).

Signature 18 has a very specific mutational pattern of C>A transversions which is observed only in neuroblastomas. *N-Myc* amplification is a common feature of neuroblastomas (Brodeur et al., 1984) and statistical analysis reveals that samples with *N-Myc* amplification exhibit a significantly higher numbers of somatic mutations attributed to Signature 18 when compared to samples without this amplification ($Q = 1.2 \times 10^{-7}$; Figure 5.3).

5.4 Activity of mutational signatures and association with age of diagnosis

The origin of a cancer cell can be traced back to the zygote and, hence, the accumulation of somatic mutations identified by cancer genome sequencing can be roughly separated into mutations occurring prior to neoplastic development and mutations occurring after tumour initiation. The mutations occurring prior to

neoplastic development can be further separated as spontaneous somatic mutations occurring due to the activity of normal cellular processes and sporadic somatic mutations triggered by environmental exposures or lifestyle choices. Assuming that the accumulation of spontaneous mutations is (on average) the same across different people and that spontaneous pre-neoplastic mutations can be separated from all other somatic mutations found in a cancer, one would expect to see a strong correlation between the numbers of spontaneous pre-neoplastic somatic mutations and the age of cancer diagnosis in a large cohort of people.

A first order of approximation of this logic entails using cancer genomics data

Cancer Type	Samples with age information	Mutational Signature	P-value (FDR corrected)
ALL	**106**	**Signature 1B**	2.13E-04
AML	**151**	**Signature 1B**	6.81E-06
Breast	**879**	**Signature 1B**	7.23E-04
Colorectum	**488**	**Signature 1B**	2.89E-02
Glioma Low Grade	**154**	**Signature 1A**	1.50E-07
Head and Neck	**299**	**Signature 1B**	4.54E-03
Kidney Chromophobe	**21**	**Signature 1A**	3.53E-02
Kidney Clear Cell	**294**	**Signature 1B**	7.34E-12
Kidney Papillary	**95**	**Signature 5**	3.10E-03
Lymphoma B-cell	**24**	**Signature 1B**	1.06E-02
Medulloblastoma	**100**	**Signature 1A**	2.83E-10
Melanoma	**216**	**Signature 1B**	1.33E-03
Melanoma	**216**	**Signature 7**	2.00E-03
Neuroblastoma	**192**	**Signature 1B**	2.84E-05
Ovary	**425**	**Signature 1B**	7.18E-09
Pilocytic Astrocytoma	**63**	**Signature 1B**	4.76E-02
Stomach	**148**	**Signature 1A**	3.43E-02
Thyroid	**157**	**Signature 5**	2.95E-03

Table 5.1: Mutational signatures and age of diagnosis. All statistically significant correlations between exposures of consensus mutational signatures and age of cancer diagnosis are shown.

and attempting to correlate the age of cancer diagnosis with the mutational burden of the previously identified mutational signatures. Thus, examination is performed in

each cancer type for correlations between the age of diagnosis and the number of mutations attributable to each signature in each sample.

Signature 1A/B exhibits strong positive correlations with the age of diagnosis in the majority of cancer types of both childhood and adulthood (Table 5.1). No other mutational signature shows a consistent correlation with the age of diagnosis. Exposure to Signature 5 also correlates with the age of diagnosis in kidney papillary and thyroid cancers. However, in both cancer types, Signature 1A/B is not detected/extracted due to low number of mutations in their samples and it is likely that Signatures 1A/B and Signature 5 are mixed together. Further studies involving whole-genome sequences will be needed to validate this hypothesis. Interestingly, in melanoma, the age of diagnosis also correlates with exposure to Signature 7, which has been associated with exposure to ultraviolet light. Presumably this exposure is due to the relatively uniform chronic exposure to ultraviolet light throughout a person's lifetime.

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 1A/B 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. The absence of consistent correlation of all other signatures with age of diagnosis suggests that mutations associated with these signatures have been generated at different rates in different people, possibly as a consequence of different mutagenic exposures or after neoplastic change has been initiated.

5.5 Summary

In this chapter, I examined the mechanistic basis of the signatures of the mutational processes operative in 30 distinct types of human cancer. An etiology is proposed either by performing a statistical comparison between sets of samples with and without specific characteristics or by comparing the observed mutational patterns with the ones in the scientific literature.

This chapter provides an indication of the processes underlying the observed patterns of somatic mutations for at least some of the mutational signatures. However, for many of the processes their etiology remains speculative or unknown.

Further elucidating the underlying mutational processes will depend upon two major streams of investigation. First, compilation of mutational signatures from model systems exposed to known mutagens or perturbations of the DNA maintenance machinery and comparing those to the ones found in human cancers. Second, correlating the contributions of mutational signatures with other biological characteristics of each cancer through diverse approaches ranging from molecular profiling to epidemiology. Collectively, these studies will advance understanding of cancer etiology with potential implications for prevention and treatment.