Chapter 4

Signatures of mutational processes in human cancer

4.1 Introduction

In the previous chapter of this thesis, I applied a newly developed computational approach to somatic mutational data derived from breast cancer genome and exome sequences, which revealed multiple signatures with distinct patterns of somatic mutations. Comparing these mutational patterns with the scientific literature as well as statistically associating them with molecular phenotypes provided an indication for the etiology of the mutational processes responsible for these signatures. In this chapter, I expand the scope of the mutational signatures analysis and apply the developed computational framework to 30 distinct cancer types. The approach taken in this chapter is analogous to the one used for breast cancer in the previous chapter; mutational signatures are extracted from mutational catalogues based on the Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536} alphabets separately for each cancer type (with further separation for samples derived from whole-genome and exome sequencing in a single cancer type). The deciphered mutational signatures are hierarchically clustered, as demonstrated in the previous chapter, to derive the consensus mutational signatures in human cancer. In this chapter, I will focus on extracting the signatures of the operative mutational processes in 7,042 samples across 30 cancer classes, examining their patterns of somatic mutations, and discussing them in the context of the different cancer types in which they are found. It should be noted that this chapter does not discuss the potential etiologies of the identified consensus mutational signatures since these will be the focus of chapter 5.

4.2 Data generation and filtering of mutational catalogues

Similarly to breast cancer, no data were generated solely for the purposes of this thesis. Rather, I curate already identified somatic mutations from freely available previously published and (at the time) unpublished data. The curated freely available data are taken from three distinct sources:

- The data portal of The Cancer Genome Atlas (TCGA)
- The data portal of the International Cancer Genome Consortium (ICGC)
- Previously published in peer-review journals cancer genomics mutational data: (Agrawal et al., 2011; Barbieri et al., 2012; Berger et al., 2011; Biankin et al., 2012; Dulak et al., 2013; Fujimoto et al., 2012; Govindan et al., 2012; Grasso et al., 2012; Gui et al., 2011; Imielinski et al., 2012; Jiao et al., 2011; Jones et al., 2012a; Jones et al., 2010; Krauthammer et al., 2012; Le Gallo et al., 2012; Liu et al., 2012a; Liu et al., 2012b; Love et al., 2012; Morin et al., 2011; Nik-Zainal et al., 2012; Peifer et al., 2012; Puente et al., 2011; Pugh et al., 2013; Rudin et al., 2012; Sausen et al., 2013; Seo et al., 2012; Seshagiri et al., 2012; Shah et al., 2012; Stephens et al., 2012; TCGA, 2012; Wang et al., 2011; Wei et al., 2011; Wiegand et al., 2010; Wu et al., 2011; Zang et al., 2012; Zhang et al., 2013)

The unpublished data are generated internally by the Cancer Genome Project (CGP) or donated by collaborating investigators that were willing to participate in the performed large-scale pan-cancer mutational signatures analysis. The majority of exome data are taken from the ICGC data portal, TCGA data portal, or from the abovementioned published peer-reviewed publications. In contrast, the majority of whole-genomes are previously unpublished data. A summary of the number of

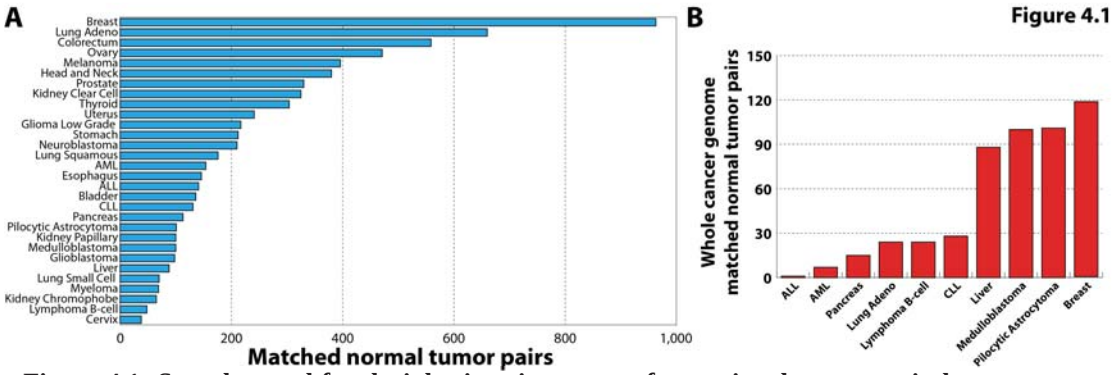

Figure 4.1: Samples used for deciphering signatures of mutational processes in human cancer. Mutational catalogues of ***(A)*** 7,042 primary cancers derived from 30 different cancer types are examined for mutational signatures, including ***(B)*** 507 whole cancer genomes with matched normal pairs.

samples based on cancer types is shown in Figure 4.1; in addition, a complete list including all samples, all examined cancer types, and their respective data sources is provided in Appendix II.

In total, I compiled the mutational catalogues of 7,042 primary cancers of 30 different classes: 507 from whole-genome and 6,535 from exome sequences (Figure 4.1). In all cases, normal DNAs from the same individuals have been sequenced to establish the somatic origin of the variants. The somatic mutations are extensively filtered to remove germline polymorphisms and sequencing artefacts as previously described for breast cancer (see chapter 3) and the final filtered dataset contains 4,938,362 somatic substitutions and small insertions/deletions (indels). The somatic mutations found in these 7,042 matched normal tumour pairs are used to decipher the mutational signatures from catalogues based on the Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536} alphabets (see below).

Examining the mutational catalogues of the 7,042 primary cancers revealed that the prevalence of somatic substitutions and indels is highly variable between and within cancer classes, ranging from about 0.001 somatic mutations per megabase to more than 400 somatic mutations per megabase (Figure 4.2). Certain childhood cancers carried fewest mutations whereas cancers related to chronic mutagenic exposures such as lung (tobacco smoking) and malignant melanoma (exposure to ultraviolet light) exhibited the highest prevalence. This variation in mutation prevalence is attributable to differences between cancers in the duration of the cellular

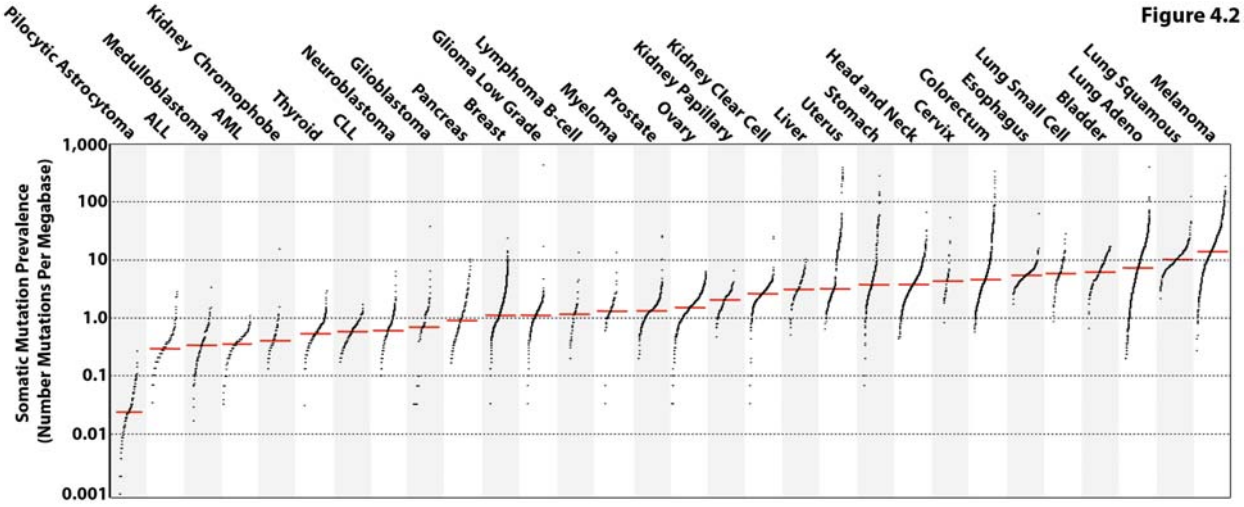

Figure 4.2: Mutational burden in human cancer. Every dot represents a sample whereas the red horizontal lines are the median numbers of mutations in the respective cancer types. The vertical axis (log scaled) shows the number of mutations per megabase whereas the different cancer types are ordered on the horizontal axis based on their median numbers of somatic mutations. ALL stands for acute lymphoblastic leukaemia; AML for acute myeloid leukaemia; CLL for chronic lymphocytic leukaemia.

lineage between the fertilized egg and the sequenced cancer cell and/or to differences in somatic mutation rates during the whole or parts of that cellular lineage (Stratton et al., 2009).

4.3 Deciphering signatures of mutational processes in 30 human cancer types

Mutational signatures are extracted using the previously defined four mutational alphabets: Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536} (Appendix I). Briefly, Ξ_{96} examines all somatic substitutions and additionally includes information on the sequence context of each substitution. This classification has 96 possible mutations since there are six classes of base substitution C>A, C>G, C>T, T>A, T>C, T>G (all substitutions are referred to by the pyrimidine of the mutated Watson-Crick base pair) and the bases immediately 5' and 3' to each mutated base are incorporated. The Ξ_{99} alphabet extends the Ξ_{96} alphabet by incorporating three additional mutation types: dinucleotide substitutions, indels at repetitive elements, and indels at microhomologies. The Ξ_{1536} alphabet examines substitutions and their immediate sequence context; however, this alphabet incorporates two bases 5' and 3' to each mutated base instead of the one base used in the Ξ_{96} alphabet. Lastly, the Ξ_{192} alphabet examines all somatic mutations in transcribed regions of the human genome. This alphabet has all the features of Ξ_{96} but it also incorporates information on whether the mutation is occurring on the transcribed or the untranscribed strand of protein-coding genes. The 96 and 1,536 substitution classifications are particularly useful for distinguishing mutational signatures which cause the same substitutions but in different sequence contexts. In contrast, the Ξ_{99} alphabet allows the evaluation of the amount of indels and dinucleotide substitutions caused by different mutational processes, while the Ξ_{192} alphabet is leveraged to evaluate the activity of repair processes operative on the transcribed regions of the human genome.

Mutational signatures are deciphered independently for each of the 30 cancer types following the same analysis procedure as the one previously used in breast cancer (chapter 3). In total, 106 mutational signatures based on the Ξ_{96} alphabet are extracted from these 30 cancer types. These mutational signatures are clustered using an unsupervised hierarchical clustering, where a cosine distance is used as a measure

for comparing mutational signatures (Figure 4.3). Any signature derived from exome sequencing data is re-normalized towards the genome trinucleotide frequency prior to applying the clustering procedure.

A threshold of 0.18 is used to separate the original 106 mutational signatures into 27 unique clusters. This threshold is conservatively selected based on visual inspection and prior biological knowledge. More specifically, annotation 1 in Figure 4.3 shows the separation of two mutational patterns overwhelmed by C>T mutations with a difference in their immediate sequence context (later referred to as Signature 7 and Signature 11, Figure 4.5). The upper branch of annotation 1 contains patterns of mutations that are consistent with exposure to ultraviolet light, while the signatures in the lower branch are exclusively found in samples that are treated with an alkylating agent (see chapter 5). Since these two sets of mutational signatures have distinct

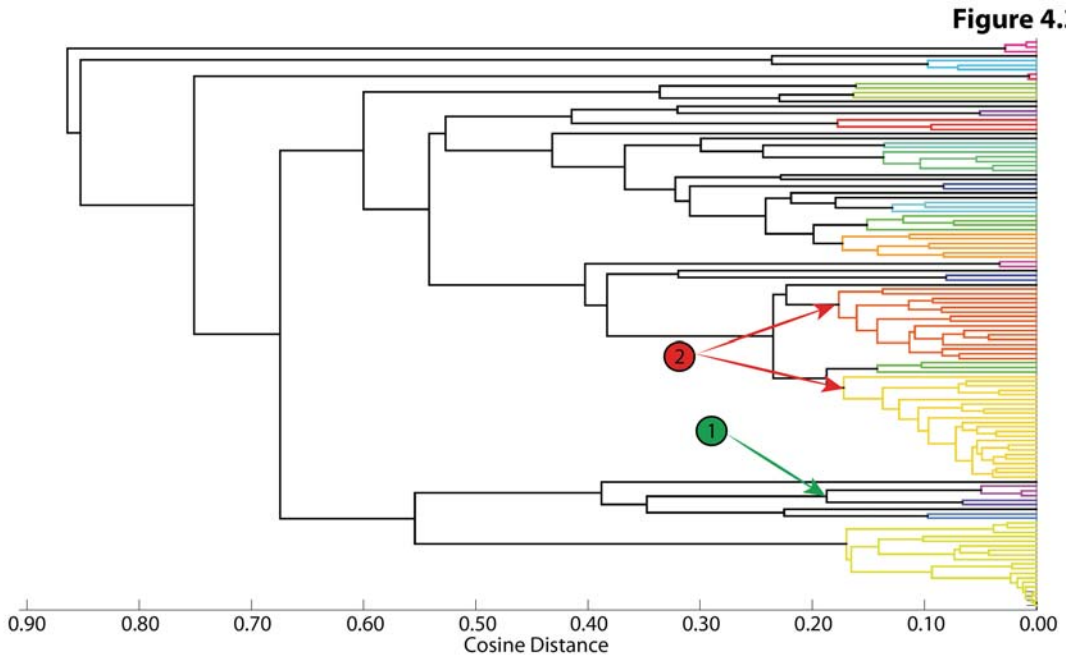

Figure 4.3: Clustering of mutational signatures. Clustering of 106 original mutational signatures deciphered from the mutational catalogues of 7,042 cancer samples. Each of the 27 unique clusters is displayed in a different colour. The cosine distance threshold for separating the signatures into clusters is set at 0.18 based on annotation 1 (green) and annotation 2 (red).

patterns and etiologies, the selected clustering threshold needs to separate them and as such it needs to be lower than 0.184. Visual inspection of clustering of the original mutational signatures (annotation 2, Figure 4.3) shows that all of these signatures possess similar patterns of somatic mutations (*e.g.*, C>T at CpG). However, these patterns are contaminated since, most probably, they cannot be extracted with the

same accuracy from different datasets (*e.g.*, less than 40 samples are used for signature analysis in cervical cancer versus the almost 1,000 samples used for signature analysis in breast cancer). To ensure that these visually similar mutational signatures cluster together, a threshold of 0.18 is selected. It should be noted that visual examination may be misleading and it may result in clustering mutational signatures that are different. Nevertheless, this analysis provides a conservative estimation of the mutational signatures found in human cancer and it is foreseeable that some of the reported mutational signatures are, in fact, mixtures of multiple distinct signatures. Only further samples across all types of human cancer will allow a further separation of these mutational signatures. As was previously performed for breast cancer, each consensus mutational signature is derived using a weighted average of the signatures belonging to its respective cluster and the number of somatic mutations attributed to a consensus mutational signature in a sample is set to the number of mutations of the original signature found in that sample.

In addition to deciphering mutational signatures using mutational catalogues based on the Ξ_{96} alphabet, an analysis is performed also for the Ξ_{99}, Ξ_{192}, and Ξ_{1536} alphabets. In all cases the consensus signatures results from the Ξ_{99} and Ξ_{192} catalogues are consistent with the previous observations based on the Ξ_{96} alphabet. However, deciphering mutational signatures for the Ξ_{1536} alphabet produced results only for a few of the cancer types (see below). The inability to decipher mutational signatures using the 1,536 mutation types is, most probably, due to the absence of sufficient numbers of somatic mutations in the examined mutational catalogues. This is perhaps unsurprising as ~93% of the examined mutational catalogues are derived from exome sequences, which harbour very few somatic mutations. Furthermore, the majority of whole-genome sequences are from childhood cancers and they have a low mutational burden (Figure 4.2).

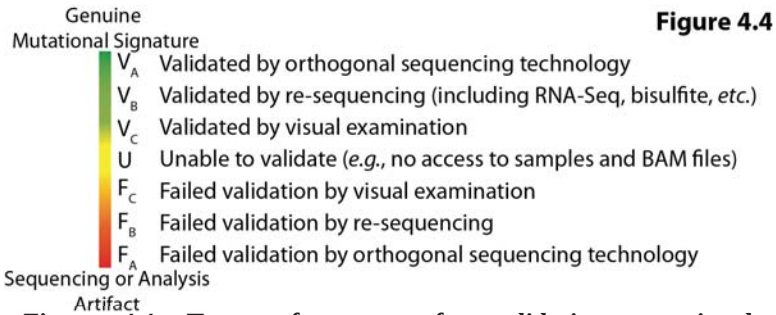

Figure 4.4: Types of statuses for validating mutational signatures.

4.4 Validating consensus mutational signatures

Validating a signature of a mutational process requires ensuring that a large set of somatic mutations, with a mutational spectrum resembling the one of the mutational signature of interest, is genuine in at least one sample in which this process is operative. As previously discussed with regard to breast cancer (chapter 3), validation is complicated as various mutational processes are found in a single cancer sample and, as such, every individual somatic mutation can be probabilistically assigned to several mutational signatures. In this analysis, I leveraged the same approach as the one used in validating mutational signatures in breast cancer: the dataset is examined 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). Since I did not have access to the biological samples, I mostly relied on previously performed validation experiments (*e.g.*, samples in TCGA sequenced by two different groups using two different next-generation sequencing technologies) as well as visual validation of BAM files by an experienced curator. Based on the data, I identified the optimal available sample for every mutational signature and attempted to validate a subset of somatic mutations attributed to this signature using one of three methods (Figure 3.3):

- Validation by re-sequencing with an orthogonal sequencing technology
- Validation by re-sequencing with the same sequencing technology (including RNA-Seq, bisulfide sequencing, *etc.*)
- Validation by visual examination of somatic mutations performed by an experienced curator using a genomic browser and BAM files for both the tumour and its matched normal

When possible, somatic mutations are validated by either re-sequencing with orthogonal technology or re-sequencing using the same sequencing technology. I resorted to visual validation only when there is no other possibility for validating a mutational signature. 22 of the 27 consensus mutational signatures were validated (Table 4.1 and Figure 4.4). Three of the mutational signatures failed validation (termed Signatures R1 to R3), while another two mutational signatures were not validated (termed Signatures U1 and U2) due to lack of access to biological samples and BAM files for the samples with sufficient numbers of somatic mutations generated by these two mutational signatures. A validation summary for all consensus

mutational signatures is provided in Table 4.1. The validated mutational signatures are depicted in Figure 4.5, while the signatures that failed validation and the signatures that remain with unknown validation status are shown respectively in Figure 4.6 and Figure 4.7.

Mutational Signature	Validation Type	Total Mutations in Sample	Total Mutations by Signature	Examined Mutations	Validated Mutations
Signature 1A	V_A	48	40	48	48 (100%)
Signature 1B	V_A	58	55	58	56 (97%)
Signature 2	V_A	76	75	76	72 (95%)
Signature 3	V_A	70	65	70	69 (99%)
Signature 4	V_A	196	192	196	182 (95%)
Signature 5	V_C	332	286	91	75 (82%)
Signature 6	V_A	598	440	598	540 (90%)
Signature 7	V_A	470	432	470	412 (88%)
Signature 8	V_A	4,514	1,558	250	227 (91%)
Signature 9	V_B	4,423	2,811	4,423	3,977 (90%)
Signature 10	V_A	12,848	10,558	12,848	9,420 (74%)
Signature 11	V_A	102	100	102	67 (66%)
Signature 12	V_C	2,808	2,327	100	93 (93%)
Signature 13	V_A	8,612	5,697	200	190 (95%)
Signature 14	V_C	12,984	12,984	100	86 (86%)
Signature 15	V_A	784	784	31	30 (97%)
Signature 16	V_A	793	678	73	69 (95%)
Signature 17	V_B	2,627	1,959	2,627	2,476 (94%)
Signature 18	V_A	158	156	158	142 (90%)
Signature 19	V_C	769	769	103	102 (99%)
Signature 20	V_A	885	488	198	198 (100%)
Signature 21	V_C	6,790	4,368	121	103(85%)
Signature U1	N/A	N/A	N/A	N/A	N/A
Signature U2	N/A	N/A	N/A	N/A	N/A
Signature R1	F_C	11,869	7,955	100	2(2%)
Signature R2	F_C	738	738	50	1(2%)
Signature R3	F_C	385	235	83	3(4%)

Table 4.1. Validating consensus mutational signatures found in human cancer. The precise validation approach is outlined in the text. The codes of validation types are explained in Figure 4.4.

4.5 The landscape of consensus mutational signatures in human cancer

Applying the developed computational approach to the 7,042 samples derived from 30 cancer types revealed 22 distinct and validated mutational signatures (Figure 4.5; an individual figure for each signature can be found in Appendix III). These 22 mutational signatures show substantial diversity in their patterns of somatic mutations. There are signatures characterized by the prominence of only one or two of the 96 possible substitution mutations, indicating a remarkable specificity of mutation type and sequence context. One such example is Signature 10, which is

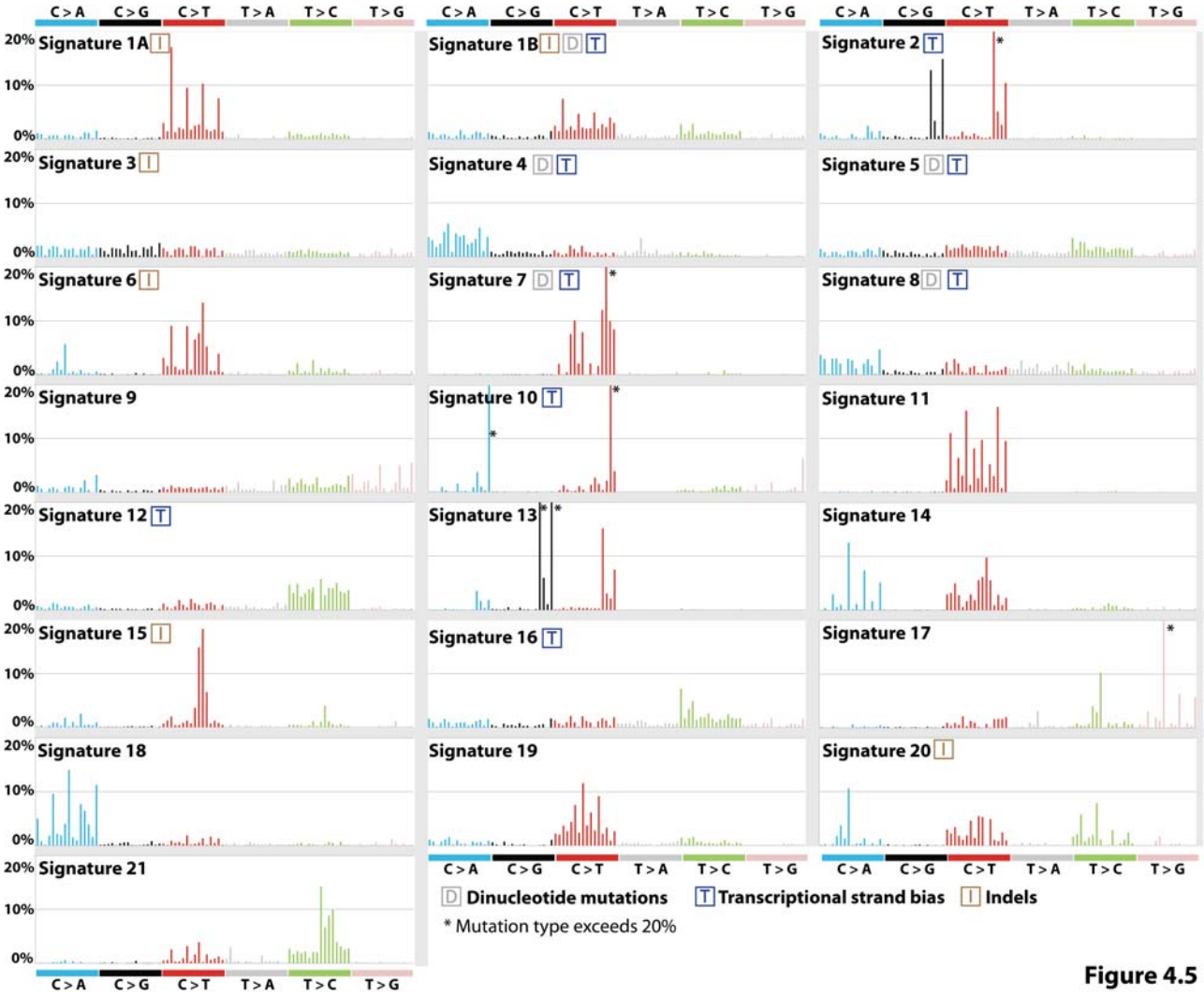

Figure 4.5

Figure 4.5: Consensus validated mutational signatures in human cancer. Each signature is displayed according to the 96 substitution classification defined by the substitution class and sequence context immediately 3' and 5' to the mutated base. The probability bars for the six types of substitutions are displayed in different colours. The mutation types are on the horizontal axes, whereas vertical axes depict the percentage of mutations attributed to a specific mutation type. All mutational signatures are displayed on the basis of the trinucleotide frequency of the human genome. A higher resolution of each panel is found Appendix III. Asterisk indicates mutation type exceeding 20%.

predominantly characterized by C>A mutations at TpCpT and C>T mutations at TpCpG. At the other extreme, some mutational signatures exhibit a more-or-less equal representation of all 96 mutations. Examples of such mutational signatures are Signatures 3 and 8. A large proportion of the validated consensus mutational signatures are characterized predominantly by C>T substitutions at different trinucleotide sequence contexts: Signatures 1A, 1B, 6, 7, 11, 15, and 19. Signatures 4, 8, and 18 have a prevalence for C>A mutations, while Signatures 5, 12, 16, and 21 exhibit a preference for T>C substitutions. Signatures 9 and 17 exhibit a preference of T>G mutations at specific sequence contexts. Lastly, no mutational signatures in this series are dominated by T>A substitutions.

Signatures 1A and 1B are observed in 25 of the 30 cancer classes (Figure 4.9). Both are characterized by a prominence of C>T substitutions at NpCpG trinucleotides. Since they are almost mutually exclusive among tumour types (Figure 4.9) they probably represent the same underlying process, with Signature 1B representing a less efficient separation from other signatures in some cancer types. Signature 1A/B is most likely related to the relatively elevated rate of spontaneous deamination of 5-methyl-cytosine which results in C>T transitions and which predominantly occurs at NpCpG trinucleotides (Pfeifer, 2006). This mutational process operates in the germline, where it has resulted in substantial depletion of NpCpG sequences, as well as in normal somatic cells (Welch et al., 2012).

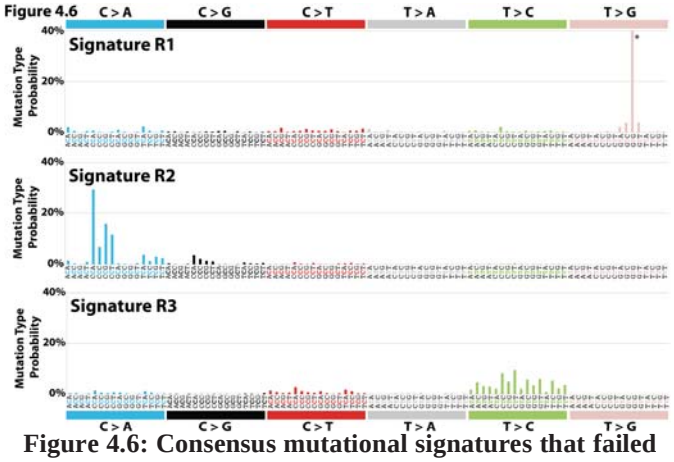

Figure 4.6: Consensus mutational signatures that failed validation. Each signature is displayed according to the 96 substitution classification defined by the substitution class and sequence context immediately 3' and 5' to the mutated base. The probability bars for the six types of substitutions are displayed in different colours. The mutation types are on the horizontal axes, whereas vertical axes depict the percentage of mutations attributed to a specific mutation type. All mutational signatures are displayed on the basis of the trinucleotide frequency of the human genome. A higher resolution of each panel is found Appendix III. Asterisk indicates mutation type exceeding 40%.

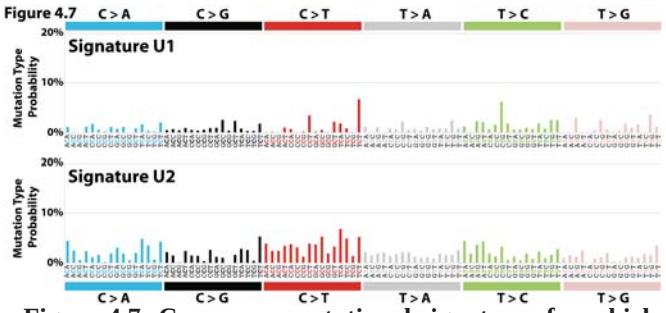

Figure 4.7: Consensus mutational signatures for which it is not possible to perform validation. Each signature is displayed according to the 96 substitution classification defined by the substitution class and sequence context immediately 3' and 5' to the mutated base. The probability bars for the six types of substitutions are displayed in different colours. The mutation types are on the horizontal axes, whereas vertical axes depict the percentage of mutations attributed to a specific mutation type. All mutational signatures are displayed on the basis of the trinucleotide frequency of the human genome. A higher resolution of each panel is found Appendix III.

In addition to the 22 consensus mutational signatures that validated (Table 4.1), three signatures failed validation, and thus most likely reflect technology specific sequencing artefacts (Figure 4.6). Signature R1 is previously described in chapter 3 and is predominantly characterized by T>G mutations at GpGpTpGpG. Signature R2 exhibits a C>A pattern of

mutations with a preference for CpC and TpC dinucleotides. Finally, Signature R3 is predominantly composed of T>C mutations with a specific trinucleotide pattern (Figure 4.6). Interestingly, these mutational signatures are confined to samples from specific sequencing centres. Signature R1 is found in samples analysed by the Sanger Institute, Signature R2 in samples sequenced at the Broad Institute, and Signature R3 is found only in data generated by the Baylor College of Medicine. This observation further confirms the suspicion that these three mutational processes reflect technical/analysis artefacts rather than real biological processes.

For three of the 27 consensus mutational signatures, I was unable to identify available samples that could be used to validate these signatures (Figure 4.7). Both Signatures U1 and U2 exhibit a rather uniform pattern of mutations across the six types of substitutions without any mutation type exceeding 10%. It should be noted that the patterns of these two mutational signatures are different from the previously identified and validated uniform mutational signatures: Signature 3 and Signature 8 (Figure 4.5).

Lastly, all of the previously identified breast cancer mutational signatures are found by this pan-cancer analysis. Breast cancer Signature BC-1 (chapter 3) has the same pattern of mutations as the global consensus Signature 1B, Signature BC-2 corresponds to Signature 2, Signature BC-3 corresponds to Signature 13, Signature BC-4 corresponds to Signature 3, Signature BC-5 corresponds Signature 8, and Signature BC-6 corresponds to Signature R1.

4.5.1 Consensus mutational signatures with transcriptional strand-bias

The efficiency of DNA damage and DNA maintenance processes can differ between the transcribed and untranscribed strands of genes. The most celebrated cause of this phenomenon is transcription-coupled nucleotide excision repair (NER) that operates exclusively on the transcribed strand of genes and is recruited by RNA polymerase II when it encounters bulky DNA helix-distorting lesions (Hanawalt and Spivak, 2008). Evaluation of the efficiency of transcription-coupled DNA repair is done analogously to the analysis performed for breast cancer (chapter 3). Briefly,

mutational signatures are re-extracted incorporating the transcriptional strand on which each mutation has taken place.

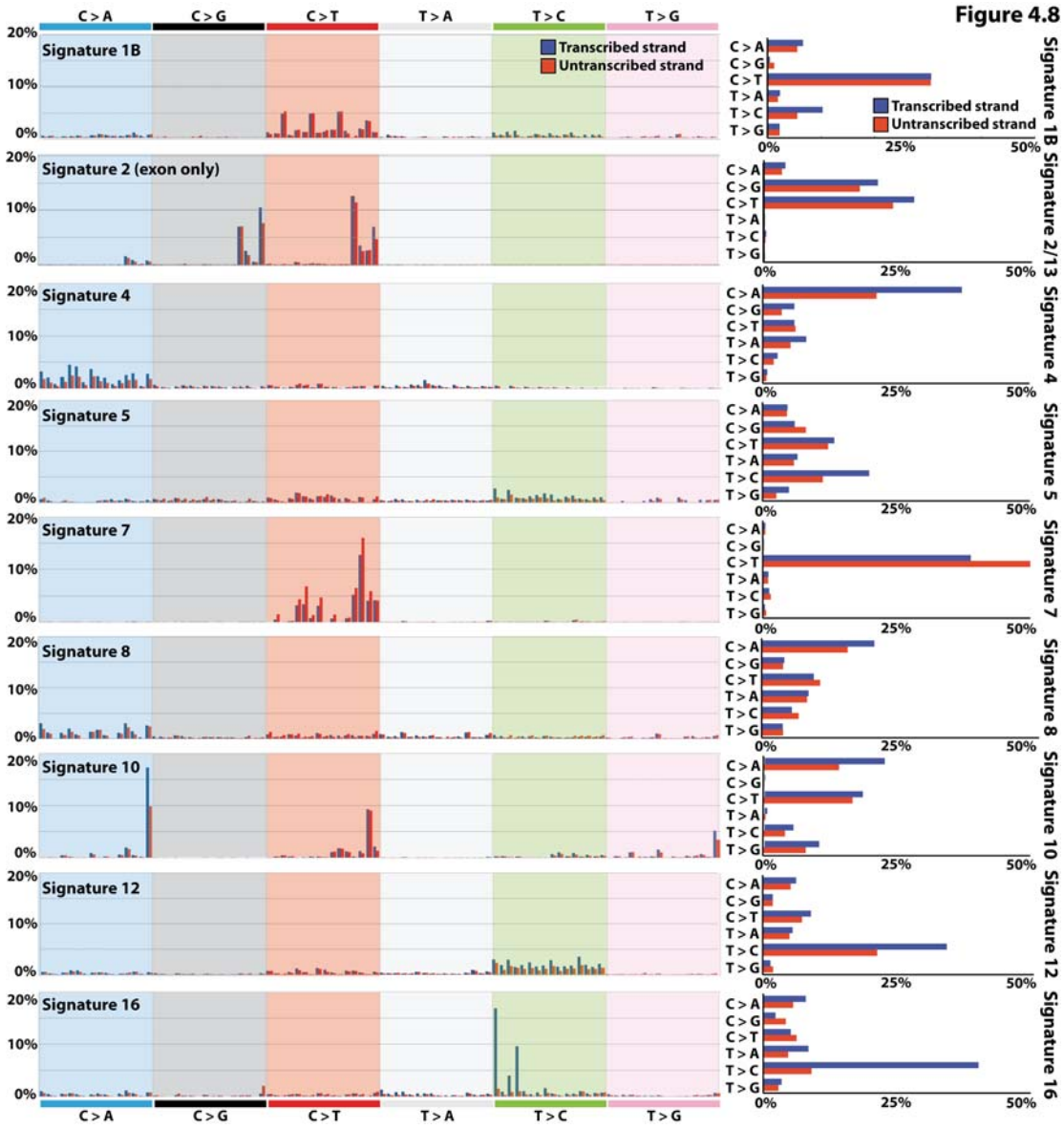

Figure 4.8: Consensus mutational signatures with strand-bias. Mutations are shown according to the 192 mutation classification incorporating the substitution type, the sequence context immediately 5' and 3' to the mutated base and whether the mutated pyrimidine is on the transcribed or untranscribed strand. The mutation types are displayed on the horizontal axis, whereas the vertical axis depicts the percentage of mutations attributed to a specific mutation type. A higher resolution version of all mutational signatures with transcriptional strand-bias is found in Appendix IV.

Nine consensus signatures showed substantial differences in mutation prevalence between transcribed and untranscribed strands, known as transcriptional strand-bias (Figure 4.8; an individual figure for each signature can be found in Appendix IV). This strand-bias is observed only for validated mutational signatures (Figure 4.5) and it is absent in the signatures that failed validation (Figure 4.6) or for

which validation is not possible (Figure 4.7). In eight of these nine signatures the strand-bias is observed across the complete footprints of transcribed protein coding genes. In contrast, the strand-bias in Signature 2 is observed only in exons and it is lacking in intronic regions.

Two of the nine mutational signatures likely implicate activity of transcription-coupled nucleotide excision repair. Signature 4 shows transcriptional strand-bias for C>A mutations (Figure 4.8). Signature 4 is observed in lung adenocarcinoma, squamous and small cell carcinomas, head and neck squamous, and liver cancers (Figure 4.9), most of which are caused by tobacco smoking. Therefore, Signature 4 is probably an imprint of the bulky DNA adducts generated by polycyclic hydrocarbons found in tobacco smoke and their removal by transcription-coupled NER (Pfeifer et al., 2002). The higher prevalence of C>A mutations on transcribed compared to untranscribed strands is consistent with the propensity of many tobacco carcinogens to form adducts on guanine.

Similarly, Signature 7, mainly found in malignant melanoma, shows a higher prevalence of C>T mutations on the untranscribed compared to the transcribed strands consistent with the formation, through ultraviolet light exposure, of pyrimidine dimers and other lesions which are known to be repaired by transcription-coupled NER (Pfeifer et al., 2005).

Beyond these known examples of DNA damage processed by transcription-coupled NER, other signatures show strong transcriptional strand-bias: Signatures 1B, 2, 5, 8, 10, 12, and 16. Notably, Signature 16, which is characterized by T>C mutations at ApTpA, ApTpG, and ApTpT trinucleotides and is observed in hepatocellular carcinomas, shows the strongest transcriptional strand-bias of any signature, with T>C mutations occurring almost exclusively on the transcribed strand (Figure 4.8). Similarly, Signature 12, which features T>C mutations at NpTpN trinucleotides, also found in hepatocellular carcinomas, shows strong transcriptional strand-bias with more T>C mutations on the transcribed than untranscribed strands (Figure 4.8). Based on the assumption that the transcriptional strand-biases in Signatures 12 and 16 are introduced by transcription-coupled NER, these currently unexplained signatures might be the result of bulky DNA helix distorting adducts on adenine. However, there is no prior basis for invoking transcription-coupled NER in

the genesis of these signatures (or any of the other mutational signatures) and other causes of transcriptional strand-bias may exist.

4.5.2 Mutational signatures with dinucleotide substitutions and indels

Mutational signatures are re-extracted including, in addition to the 96 substitution types, three further classes of mutation: dinucleotide substitutions, indels at short nucleotide repeats, and indels with overlapping microhomology at breakpoint junctions. This analysis also revealed 27 consensus mutational signatures (annotated on Figure 4.5). No indels or dinucleotide substitutions are found in the signatures that are not validated. Six of the validated mutational signatures are associated with indels, while five of the validated mutational signatures are associated with double nucleotide substitutions.

Signature 1A and Signature 1B both associate with indels at repetitive elements. Interestingly, these mutational signatures do not contribute large amounts of indels (or substitutions) in any given sample but, rather, these mutational signatures are present at low background levels in almost all samples in which they are found.

Four of the 22 base substitution signatures associated with large numbers of indels. Signature 6, which is characterized predominantly by C>T at NpCpG mutations, but is distinct from Signature 1A/B, contributes very large numbers of substitutions and small indels (mostly of 1bp) at nucleotide repeats to subsets of colorectal, uterine, liver, kidney, prostate, oesophageal and pancreatic cancers.

Signature 15 and Signature 20 also contribute very large numbers of substitutions and small indels at nucleotide repeats but, compared to Signature 6, Signature 15 exhibits greater prominence of C>T at GpCpN trinucleotides, whereas Signature 20 contains C>A and T>C mutations. Signature 15 is found in several samples of lung and stomach cancer, whereas Signature 20 is found only in few gastric carcinomas (Figure 4.9). The origin of both mutational signatures is currently unknown.

By contrast, substantial numbers of larger deletions (up to 50 bp) with overlapping microhomology at breakpoint junctions are found in breast, ovarian and

pancreatic cancer cases with major contributions from Signature 3. In the chapter 3, I associated this particular mutational signature with inactivating mutations in *BRCA1* and/or *BRCA2* in breast cancer. This association will be further elaborated upon in the next chapter for ovarian and pancreatic cancers.

Signatures 1B, 5, 4, 7, and 8 are associated with double nucleotide substitutions. Samples with Signature 1B, 5, or 8 have low numbers of dinucleotide substitutions. In contrast, overwhelming numbers of dinucleotide substitutions are present in samples in which Signature 4 or Signature 7 is found. CC>AA/GG>TT or TC>AA/GA>TT are the predominant types of dinucleotide substitutions caused by Signatures 1B and 5. Signature 4 and 8 generate mostly CC>AA/GG>TT mutations, while Signature 7 is characterized by CC>TT/GG>AA mutations occurring predominantly at dipyrimidines.

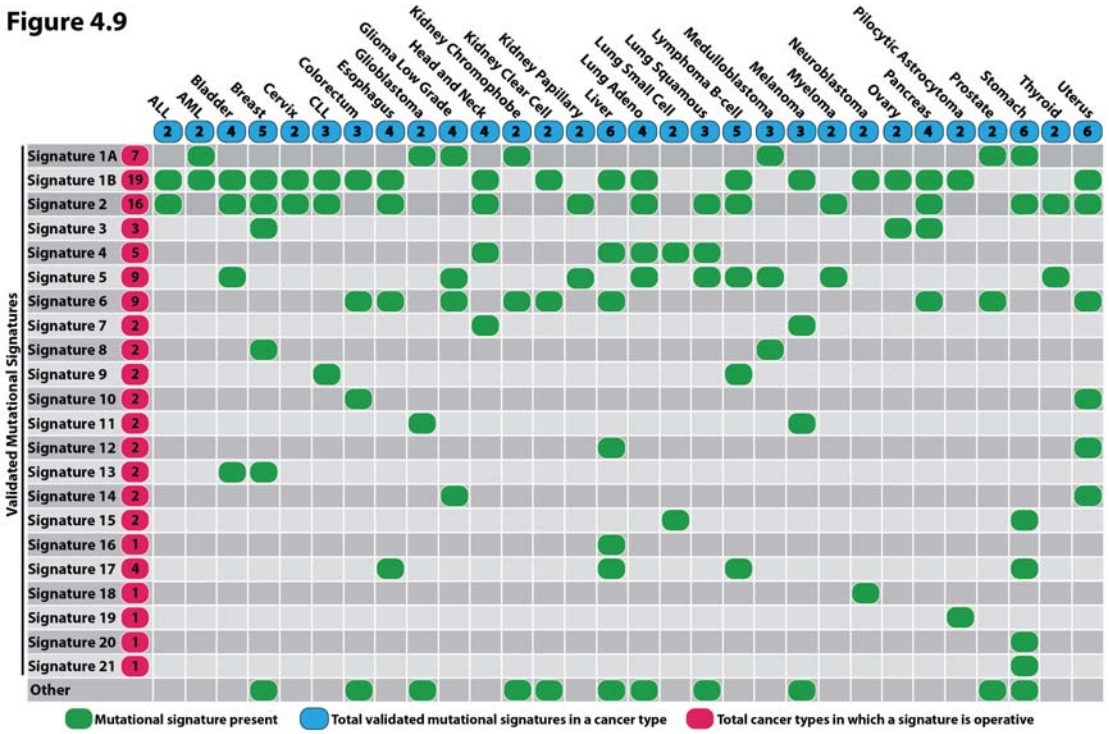

Figure 4.9: Signatures of mutational processes and the cancer types in which they are found. Cancer types are ordered alphabetically as columns, whereas mutational signatures are displayed numerically as rows. 'Other' indicates mutational signatures for which validation was not performed or for which validation failed.

4.5.3 Mutational signatures with additional sequence context

Mutational signatures are further extracted using mutational catalogues based on the Ξ_{1536} alphabet. Unfortunately, the majority of the examined cancer types are

derived from exome sequencing data and, as such, they harbour too few somatic mutations for this analysis. The examination of low numbers of somatic mutations based on a classification system that contains 1,536 types of mutations resulted in predominantly binary matrix data and, for the majority of cancer types, the analysis either fails or it does not reveal any further elaborations of the consensus mutational signatures.

Nevertheless, there are four mutational signatures that are refined by this analysis. As previously demonstrated for breast cancer, Signature 2 and Signature 13 exhibit a preference for a pyrimidine prior to the mutated TpC dinucleotide while the majority of Signature R1's T>G substitutions occur at T>G at GpGpTpGpG pentanucleotides (chapter 3). Further, this analysis demonstrated that the T>X peaks at CpT dinucleotides characteristic for Signature 17 are, in fact, dependent on the presence of an adenine located 5' prior to the dinucleotide; thus these peaks occur at ApCpTpN tetranucleotides ($Q = 1.3 \times 10^{-11}$; in all cases Q refers to a q-value, see chapter 7). Lastly, Signature 10 also displays a pentanucleotide pattern different than the one expected purely by chance ($Q = 4.5 \times 10^{-42}$). The three large peaks of Signature 10 are highly dependent on either an adenine or thymine two bases 5' to the somatic mutation.

4.6 Prevalence of consensus mutational signatures in human cancer

The previous sections of this chapter discussed the identified consensus mutational signatures. In this section, I will examine and summarize their prevalence across the analysed 30 human cancer types. In most cancer classes at least two mutational signatures are observed, with a maximum of six in cancers of the liver,

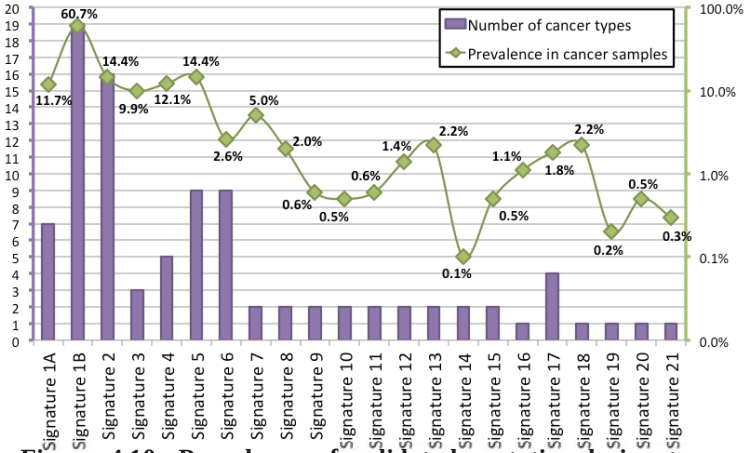

Figure 4.10: Prevalence of validated mutational signatures across all cancer types. The X-axis depicts the mutational signatures. The right Y-axis reflects the number of cancer types in which the validated consensus signature has been identified, while the left Y-axis indicates the percentage of samples from the data set of 7,042 cancers in which the signature contributed a significant number of somatic mutations.

uterus, and stomach (Figure 4.9 and Figure 4.10). Although these differences may, in part, be due to the available data in each cancer type, it seems likely that some cancers have a more complex repertoire of mutational processes than others. Signature 1A/B is found in the majority of the samples (Figure 4.10), while Signatures 2, 3, 4, 5, and 7 are present in at least ~5% of the samples. Notably Signature 2 is found in 16 of the 30 cancer types and in ~14% of all samples.

Most individual cancer genomes exhibit more than one mutational signature and many different combinations of signatures are observed (Figure 4.11). An individual figure for each cancer type depicting the contributions of the mutational signatures in each sample of that cancer type can be found in Appendix V. Further, an individual figure for each cancer type depicting the summary of the signatures' contributions in that cancer type can be found in Appendix VI. Liver cancers have the richest mutational landscape since the average liver cancer sample has at least 5 signatures imprinted by different mutational processes (Appendix V).

The patterns of contribution to individual cancer samples vary markedly between signatures. Signature 1A/B contributes relatively similar numbers of mutations to most cancer cases whereas most other signatures contribute overwhelming numbers of mutations to some cancer samples but very few to others of the same cancer class. Examples of such mutational signatures are Signatures 2, 3, 4, 6, 7, 9, 10, 11, and 13 (Figure 4.11).

Some mutational signatures are found in significant proportions of samples in some cancer types, while contributing only to a subset of samples in other cancer types. Notably, Signature 2 is identified in the majority of cervical (79%), thyroid (52%), and bladder (51%) samples but it is found only in a limited set of multiple myelomas (6%), B-cell lymphomas (11%), and breast cancers (18%) (Appendix V). Other examples include: Signature 6, identified in 20% of colorectal samples but only present in 0.6% of prostate cancer samples; Signature 13, identified in 67% of bladder samples but only present in 7% of breast cancer samples; Signature 17, found in 44% of oesophageal cancers but only in 14% of stomach cancers; Signature 3, found in 30% of breast cancers but only in 12% of pancreatic cancers (Appendix V). The reasoning behind the tissue specificity of the identified mutational signatures remains elusive. However, it is possible that some (or even most) mutational signatures are not as variable by cancer type as currently appears to be the case and examination of more

genomics data will reveal the presence of these mutational signatures in the majority of cancer types (albeit with low prevalence). Nevertheless, there are at least some mutational signatures that are most likely specific to a set of cancer types. For

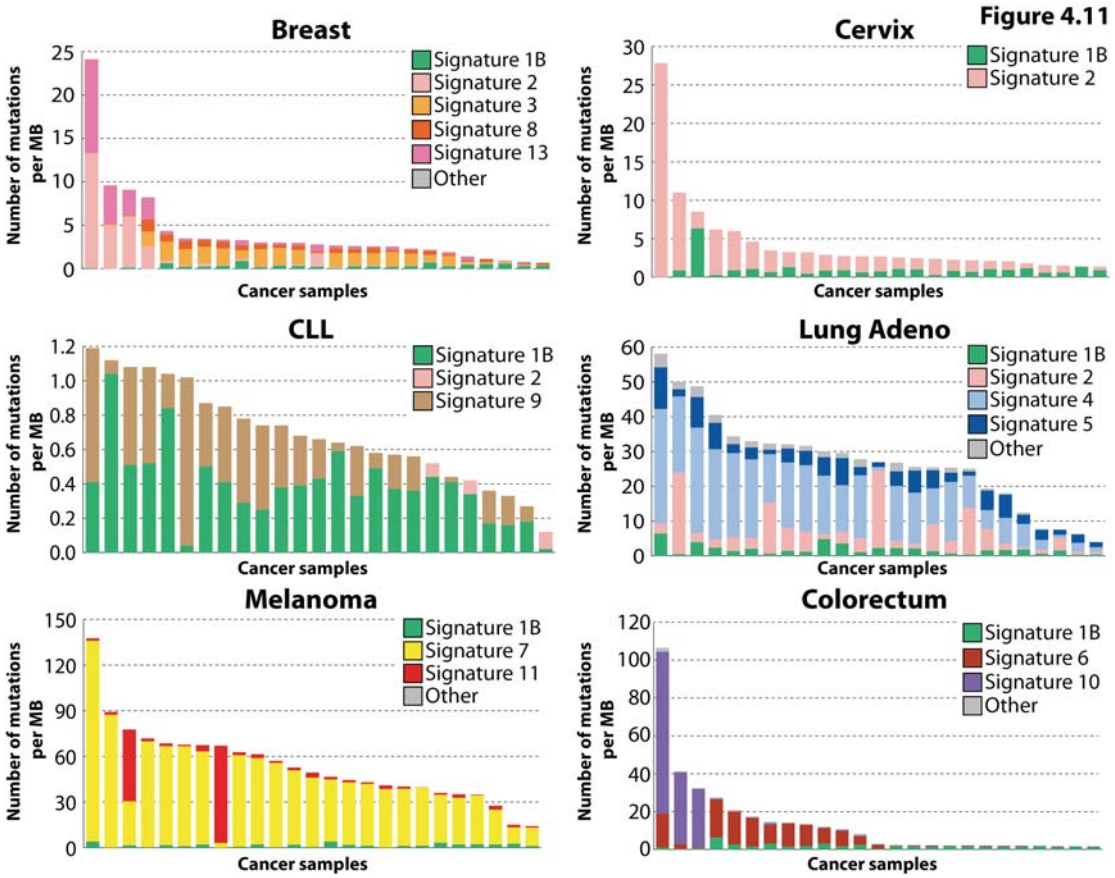

Figure 4.11: Contributions of mutational signatures in a selected set of cancer types. 25 samples are displayed for each cancer type. Each sample is displayed as a column with a height corresponding to the number of somatic mutations per megabase found in this sample. Every column is proportionately coloured to reflect the percentage of mutations attributed to different mutational signatures. 'Other' indicates mutational signatures for which validation is not performed or for which validation failed.

example, one would not expect to find the mutational signature of ultraviolet light in a primary colorectal cancer.

4.7 Discussion

In this chapter, I presented and discussed mutational signatures analysis encompassing 7,042 samples derived from 30 human cancers. The results revealed more than 20 consensus mutational signatures with a complex landscape across the different cancer types and, even, across individual cancer samples.

It should be noted that as in any computational analysis, the extraction of mutational signatures is not a perfect process. In chapter 2, I described in detail the

factors that influence the extraction of mutational signatures. These included the number of available samples, the mutation prevalence in samples, the number of mutations contributed by different mutational signatures, the similarity between the signatures of mutational processes operative in cancer samples, as well as the limitations of the developed computational approach.

In this chapter, I examined datasets with varying sizes from 30 different cancer types and great care has been taken to report only validated mutational signatures. However, the developed approach identified two similar patterns most likely representing the same biological process, viz., Signature 1A and 1B. The reasons for this is, for some cancer types, sufficient numbers of samples and/or mutations are available (i.e., statistical power) to decipher the cleaner version (i.e., Signature 1A) while for other cancer types there are not sufficient data and the approach extracts a version of the signature which is more contaminated by other, likely partially correlated, signatures present in that cancer type (i.e., Signature 1B). Nevertheless, the two signatures are visually very similar and they have been named 1A and 1B. Being almost mutually exclusive amongst cancer types (i.e., finding either Signature 1A or Signature 1B in each cancer type but not usually both) is supportive of the notion that they represent the same underlying process as is the fact that Signatures 1A and 1B have the same overall pattern of contributions to individual cancer genomes. Indeed, it is likely that if there were sufficient data, Signature 1B would disappear and the algorithm would extract only Signature 1A.

In summary, through examination of the mutational patterns buried within cancer genomes, this analysis revealed the diversity and complexity of somatic mutational processes underlying carcinogenesis in human beings. It is likely that more mutational signatures will be extracted, together with more precise definition of their features, as the number of whole-genome sequenced cancers increases and analytic methods are further refined.