Chapter 6

Discussion and future explorations

6.1 Introduction

The main aim of this thesis was to improve our understanding of the mutational processes underlying cancer development by examining the molecular patterns of mutations imprinted on cancer genomes by these processes. This goal was achieved by the development of a novel theoretical model that conceptualized the idea of a *mutational signature* and mathematically connected mutational signatures with catalogues of somatic mutations identified in cancer genomes.

The developed mathematical model was used to create a computational approach to decipher the signatures of the mutational processes operative in a set of cancer genomes, based on the somatic mutations identified in the mutational catalogues of these cancers. The computational framework was extensively evaluated with a wide-range of simulated data and it was demonstrated that the framework is robust to a variety of distinct parameters and can be effectively applied to both genome and exome sequences.

The developed novel computational framework was applied to genomics data from 7,042 cancer patients to reveal the mutational processes operative across the spectrum of 30 distinct types of human cancers. This largest to date analysis of cancer genomics data has provided the first map of the signatures of the mutational processes moulding the genomes of human cancers. More than 20 distinct signatures were identified and an etiology was proposed for some of these signatures. Nevertheless, the underlying mechanisms for the majority of the mutational signatures remain mysterious and future studies will be needed to elucidate their true nature.

This chapter discusses the importance of the results presented throughout the thesis. It also provides a critical reflection on the analyses of mutational signatures and outlines potential future directions for improvement with regard to the development of novel methodologies for deciphering mutational signatures and further refining of the already identified signatures.

6.2 Implications of the identified mutational signatures

In this thesis, I report the first systematic computational analysis of large-scale cancer genomics data in order to reveal the signatures of the mutational processes underlying the development of human cancer. A brief summary of the main results of the thesis is provided in Table 6.1. The table emphasizes the characteristic mutational pattern of each mutational signature, the most common cancer types in which the signature is observed, as well as any potential etiology proposed for a mutational signature.

Signature name	**Characteristic mutational pattern**	**Most common cancer types**	**Proposed etiology**	**Etiology proposed based on**
Signature 1A	C>T at CpG	All cancer types	Deamination of 5-methylcytosine	Similarity of the mutational pattern
Signature 1B	C>T at CpG	All cancer types	Deamination of 5-methylcytosine	Similarity of the mutational pattern
Signature 2	C>T at TpC	Sixteen different cancer types	*APOBEC1*, *APOBEC3A*, or *APOBEC3B*	Similarity of the mutational pattern
Signature 3	Uniform mutational signature	Breast, ovarian, and pancreatic cancer	Defective repair of DNA double-strand breaks based on homologous recombination	Statistical association
Signature 4	C>A mutations with strong strand bias	Lung, head and neck, and liver cancer	Tobacco smoking	Similarity of the mutational pattern and statistical association
Signature 5	Mostly uniform mutational signature	Nine different cancer types	Mostly unknown but there is a weak	Some statistical association

	with some peaks of T>C mutations at ApT		association with tobacco smoking in lung cancer	
Signature 6	C>A mutations and C>T at GpC mutations	Nine different cancer types but most prevalent in colorectal and uterine cancers	Defective DNA mismatch repair	Similarity of the mutational pattern and statistical association
Signature 7	C>T at dipyrimidines	Malignant melanoma and lip cancers	Ultraviolet light	Similarity of the mutational pattern
Signature 8	C>A mutations with a moderate strand bias	Breast cancer and medulloblastoma	Higher prevalence in estrogen receptor negative breast cancers	Statistical association
Signature 9	T>G transversions at ApT and TpT	Chronic lymphocytic leukaemias and B-cell lymphomas	Polymerase η	Similarity of the mutational pattern and statistical association
Signature 10	C>A at TpCpT and C>T at TpCpG	Colorectal and uterine cancers	Polymerase ε	Statistical association
Signature 11	C>T substitutions	Malignant melanoma and glioblastoma multiforme	Treatment with temozolomide	Similarity of the mutational pattern and statistical association
Signature 12	T>C substitutions with strand bias	Liver and uterine cancer	Unknown	N/A
Signature 13	C>A and C>G at TpC	Bladder and breast cancer	*APOBEC1*, *APOBEC3A*, or *APOBEC3B* and *REV1*	Similarity of the mutational pattern
Signature 14	C>A mutations and C>T at GpC mutations	Low grade glioma and uterine cancer	Unknown	N/A
Signature 15	C>T at GpC	Stomach and lung	Defective DNA	

	mutations	cancer	mismatch repair	Similarity of the mutational pattern
Signature 16	T>C mutations at ApT with extremely strong strand-bias	Liver cancer	Unknown	N/A
Signature 17	T>G at TpT and T>C at CpT	Oesophagus cancer, liver cancer, stomach cancer, and B-cell lymphoma	Unknown	N/A
Signature 18	C>A mutations	Neuroblastoma	Amplification of *N-Myc*	Statistical association
Signature 19	C>T mutations	Pilocytic astrocytoma	Unknown	N/A
Signature 20	C>A and C>T mutations	Stomach cancer	Defective DNA mismatch repair	Similarity of the mutational pattern
Signature 21	T>C mutations	Stomach cancer	Unknown	N/A
Signature R1	T>G at GpTpG	Breast cancers generated by the Sanger Institute	Sequencing artifact	Fine-tuning a sequencing protocol
Signature R2	C>A mutations	Lung and kidney cancers generated by the Broad Institute	Sequencing artifact	Fine-tuning a sequencing protocol
Signature R3	T>C mutations	Colorectal cancers generated by the Baylor College of Medicine	Bioinformatics analysis artifact	Fine-tuning a bioinformatics analysis

Signature U1	Uniform mutational signature	Glioblastoma and prostate cancer	Unknown	N/A
Signature U2	Uniform mutational signature	Liver and kidney cancer	Unknown	N/A

Table 6.1: Summary of the deciphered signatures of mutational processes in human cancer.

This thesis has three potential implications for cancer research and cancer treatment. First, from a basic science perspective, the thesis provides the first roadmap of the mutational signatures underlying human cancer and it reveals that these signatures have a complex landscape both in an individual cancer type and across multiple cancer types.

Second, from a targeted therapeutics perspective, many of the described mutational signatures are believed to reflect failure of DNA repair mechanisms and, as such, they might be better predictors of clinical outcome when compared to mutations in genes. For example, Signature 3 is associated with mutations in *BRCA1* and *BRCA2* and it is believed to reflect failure of repair of DNA double-strand breaks based on homologous recombination (Table 6.1). This mutational signature is observed in many breast and ovarian samples lacking any *BRCA1/2* mutations and it could potentially be used for targeted treatment especially for cancers such as triple negative breast cancer. A similar logic may be applied to some of the other mutational signatures reflecting failure of DNA repair mechanisms; however, future studies will be required to reveal the applicability of mutational signatures in the clinic.

Third, some of the identified mutational signatures reflect exposures to exogenous mutagens. These signatures might be useful for the development of cancer prevention strategies. For example, Signature 4 is due to tobacco smoking while Signature 7 is associated with exposure to ultraviolet light (Table 6.1). It is foreseeable that some of the other deciphered mutational signatures might be due to or triggered by environmental exposures. For example, Signature 2 is found in 16 cancer types and it is believed that this signature is due to the activity of the APOBEC family of enzymes, which could get activated by viral infection. In support of this claim, Signature 2 is found overwhelmingly in cervical cancer, which is by far the most common HPV-related cancer. It is highly plausible that Signature 2 is indeed triggered by viral infection in cervical cancer and it is foreseeable that this might be the case in one or more of the other fifteen cancer types in which Signature 2 is observed. While future analysis will be required to evaluate the validity of this hypothesis, confirming it will establish an important new mechanism for initiation of human carcinogenesis with significant potential for cancer prevention.

6.3 Limitations of the performed analyses of mutational signatures

The mutational signatures analyses have a number of shortcomings pertaining to the developed computational approach and the examined mutational data.

With regard to the data limitations, the majority of the work is restricted to certain classes of mutations, namely substitutions and small insertions/deletions (indels), with no attention to rearrangements and copy number changes. Further, the examined data are taken from a range of different sources (*e.g.*, publications, data portals, collaborators, *etc*.) in which the quality of DNA sequencing and mutation identification is highly variable. This is especially true for indels where the quality of the data allowed only limited exploration of indel-based mutational signatures.

Most of the analysed cancer cases are derived from exome sequencing data. Power calculations (chapter 2) and empirical observations indicate that, in general, a small number of whole-genome sequences are more powerful than a large number of exome sequences in extracting substitution and indel signatures. Indeed, in some cancer types the number of substitutions and indels available from exome sequences is so limited that only a very crude assessment of the landscape of mutational signatures is possible (*e.g.,* ovarian and thyroid cancers). Moreover, some cancer types with known patterns of mutations are not included at all in the analyses as data are either not freely available or non-existent (*e.g.*, cancer types due to exposure to aristolochic acid or aflatoxin).

While the developed computational approach is extensively evaluated with simulated data, this evaluation did not foresee the extreme variability of the numbers of somatic mutations found in cancer genomes. For example, an average cancer genome of a pilocytic astrocytoma has ~100 somatic mutations while a representative malignant melanoma harbours about 40,000 somatic mutations in its cancer genome (Figure 4.2). Extracting mutational signatures from a set containing equal numbers of mutational catalogues from melanomas and pilocytic astrocytomas will only result in finding the signatures of the mutational processes that are operative in malignant melanoma. In this example, pilocytic astrocytomas account for only 0.25% of all mutations in the dataset, well-below the 5% threshold used for optimizing and testing the computational framework (chapter 2). These differences in mutational burdens across cancer types required performing independent mutational signatures analyses for each of the 30 cancer types as, otherwise, the highly mutated cancers would overwhelm the extraction of mutational signatures. Further, for each of the individual cancer types, great care is taken to perform the analyses with and without hypermutated samples that may be skewing the extracted mutational signatures. Improving the developed method to allow analysis of all mutational catalogues together would be extremely beneficial, for example, to decipher common mutational signatures that contribute only very few mutations to a large set of samples belonging to different cancer types. Such mutational signatures would be most likely associated with underlying spontaneous endogenous mutational processes.

Remarkably, despite all the listed obscuring factors, the analyses allowed identification and validation of more than 20 distinct mutational signatures. Nevertheless, future studies will be required to both improve and extend this compendium of mutational signatures.

6.4 Future explorations

The developed roadmap of mutational signatures is in no way final or exclusive, and future work will be required to further refine it. This will include both improvement of the computational approach as well as generation of more whole-genome sequences across the complete spectrum of human cancer types.

Briefly, the computational method will need to allow analysis, in a single run, of thousands of mutational catalogues (including hypermutators and ultra-hypermutators) from multiple distinct classes of human cancer rather than artificially separating samples by cancer types. This will most probably require extending the developed framework to a hierarchical nonnegative matrix approach, where the current method would be applied multiple consecutive times and well-explained samples would be removed from further analysis after each of the performed iterations. Moreover, minimizing the Frobenius norm between original and reconstructed samples (chapter 2) might not be optimal as outliers can affect this measure. A more robust measure (*i.e.*, average Spearman correlation) may prove to give better results with this highly variable dataset. No matter what improvements are made to the developed computational framework, extensive validation with simulated data will be require to confirm its ability to better decipher mutational signatures.

In the previous analysis, the majority of examined data are derived from cancer exomes and I heavily rely on somatic mutations of variable quality as these mutations are identified using different mutation bioinformatics algorithms. Using the same mutational-calling algorithm will provide consistent results and allow exploring indels in greater detail and including previously neglected mutation types (*viz.*, rearrangements and copy number changes). Using cancer exomes limited the extent to which the genome landscape is introduced into signature characterization. In principle, there could be many features of the landscape that can be used to distinguish between signatures (*e.g.*, origins of replications, regions of open or closed chromatin, *etc.*) and hence provide further insights into the etiology and mechanisms underlying each signature. Further studies using whole-genome sequencing would be required to perform this analysis.

It is highly likely that a future large-scale mutational signatures analysis will become a reality in the next year as part of the forthcoming International Cancer Genome Consortium's pan-cancer initiative. This analysis will encompass 2,000 to 3,000 whole-genome sequences and ~10,000 exome sequences across the complete spectrum of human cancer. The somatic mutations of these cancer samples will be identified by a predefined set of optimized mutation-calling algorithm and include all types of somatic mutations. I am currently working on improving the developed computational framework to address its current limitations and apply it to this set of cancer genomics data. This large dataset will allow substantial improvements to the biological insights into mutational signatures.

6.5 Thesis summary

In this thesis, I introduced and mathematically connected the concepts of mutational processes and mutational signatures. A *mutational process* was defined as a mixture of DNA damage and repair mechanisms that act together and have the ability to cause mutations in somatic cells. A *mutational signature* was described as a characteristic pattern of somatic mutations exhibited by an operative mutational process in a genome of a cell. The mutational catalogue of a cancer represents the aggregated outcome of the activity of all mutational processes that have been operative since the very first division of the fertilized egg. Thus, a mutational catalogue of a cancer genome is a linear mixture of mutational signatures and this catalogue can be used as an archaeological record to identify the patterns of mutations exhibited by the mutational processes that have been operative in the cancer.

In this thesis, I developed a novel computational framework that allows extracting mutational signatures from a set of mutational catalogues, then exhaustively evaluated the developed method with simulated data, and applied it to 7,042 samples across 30 distinct classes of human cancer. This revealed more than 20 distinct signatures of mutational processes, for some of which I was able to propose an underlying mechanism.

In summary, this study examined a large scale of whole-genome and whole-exome sequencing data and provided insights into hitherto unrecognized mutational signatures present across the spectrum of human cancer. This study is the first of its kind and demonstrates the wealth of biological information that is hidden within the genomes of cancer cells.