Ollscoil na Gaillimhe
University of Galway

Differences in immunogenicity between cancer mutation signatures shed light on immunoediting

Noor Khehrah

A thesis presented in fulfilment of the requirements for the degree of Doctor of Philosophy

Supervisor: Cathal Seoighe

School of Mathematical and Statistical Sciences
University of Galway
Galway City, Ireland

August, 2023

Table of Contents

List of Figures

List of Tables

Declaration

I hereby declare that this thesis which I now submit for assessment as partial fulfillment of the requirements for the award of Doctor of Philosophy is all my own work and I have acknowledged any assistance or contributions and cited the published work of others where applicable. I have not obtained a degree from the University of Galway or elsewhere on the basis of any of this work.

Noor Kherreh Date

Dedication

To my mentor, Imran Khan, who has changed my way of life forever.

Acknowledgements

I sincerely want to express gratitude to my supervisor, Prof. Cathal Seoighe, for all your support and guidance over the past four years. I have learned a lot under your supervision, ranging from challenging ideas, no matter how widely accepted they are, to developing an eye for little details to always having the "Bigger Picture" in mind. I appreciate your patience in pointing out the use of inconsistent capitalization and possessive cases throughout my thesis writing. I tended to miss one or two in each draft. Thank you for your consistent and relentless supervision.

A sticky note on my workstation in ADB-1018 saying "Welcome Noor!!" on my first day, told me I would be fine in this foreign land, away from my home and family. My sincerest thanks to Barbara, Barry, Mariel, Adib, Brian, Declan, Laura, and Siobhan for making me feel at home. Thank you, Declan, for not only helping me with lugh issues but also for explaining the context of Irish jokes. Siobhan, I cannot thank you enough for all the countless things you did for me. I sincerely appreciate all your help, whether brainstorming research ideas, listening to my rants, or making mocktails for me. Sumaira!! This is no exaggeration; I would have gone crazy without you. I cannot even begin to thank you. You were always there for me. I am so lucky to call you my friend. I truly cherish the time spent with team ADB-1018. Thank you, guys. You all have been really kind.

To the people who were hundreds of miles away but still just a text message away, my friends, Saba, Nazish, Maryiam, Rabia, Mahrukh, Saad and Beeri. Thank you for always believing in me and for giving me confidence. Whenever imposter syndrome hit me, I remember how stupid you all are, and then I would immediately feel confident about my position. Thank you, Mairasiyo!!

To my elder sister, Ishwah api, all the women of my family, my puphoo, and my dearest mama who made it all easier for me. I realize I often do not have to fight many battles because they fought those grounds and set the pace for me. My siblings, Shahzil, Aniqa, Abdullah, and Muchi, who have been strong driving factors for me, thank you for always encouraging and supporting me throughout.

My partner, Zeeshan, thank you for supporting, guiding, encouraging, and pushing me, sometimes too hard but always ensuring I give my best. If it were not for you, I might have quit at some point. This could not have been possible without

you. Thank you so much. My son, Jahaan, you have been a true catalyst in this process. I love you.

Finally, my Baba, my Baba jaan. I am the luckiest girl because I am your daughter. I pray that every girl is blessed with a father like you; always there for me, enabling me, supporting me, believing in me, and making me believe in myself. This meant everything. I would have been nothing without you and your support. Thank you for everything. It's cliche, but you are all my reasons.

Abstract

Immunoediting is a process through which the immune system plays a role in shaping the mutational landscape of cancer and, consequently, in cancer progression. One critical aspect of immunoediting is the phenomenon known as neoantigen depletion. Neoantigens are mutated peptides that may arise from somatic mutations in cancer cells presented on the cell surface by the MHC molecules. Theoretically, these neoantigens can mark the cancer cells to be identified and consequently eliminated by immune cells, such as cytotoxic T-cells. Accurate neoantigen predictions allow researchers to identify which mutations generate immunogenic peptides to initiate an effective immune response against cancer cells. This has significant implications for the development of personalized immunotherapies and cancer vaccines.

Cancer immunoediting not only occurs during tumor progression but also in patients receiving anticancer immunotherapies. Neoantigen depletion during cancer progression contributes to innate resistance to immunotherapies, resulting in inconsistent results across patients and cancer types. Patients also acquire resistance to immunotherapy during treatment, leading to treatment ineffectiveness. Therefore, to effectively harness the power of the immune system against cancer and to fully understand cancer progression, a thorough understanding of cancer immunoediting is crucial. This thesis aims to gain a deeper understanding of some of the sources of variation in immunogenicity, as well as potential mechanisms to escape from immune responses. Ultimately, we are to use these findings to enhance our understanding of the impact of immunoediting on the mutational landscape of human cancers.

Two recent high profile studies have reported that recurrent driver mutations occur in the gaps in the capacity of MHC molecules to present neoantigens. This implies that the immune system selects against driver mutations that can potentially give rise to neoantigens. These findings have important implications in studying cancer progression and the role of immune system in determining how cancers develop. Interestingly, although depletion of driver mutations predicted to be immunogenic has been reported the same was not observed for passenger mutations. Therefore, in Chapter 2 we tested if the passenger mutations that are predicted to be immunogenic occur preferentially on lowly expressed or non-expressed genes which may help to

explain this observation. When we controlled for gene length and sequence context, we found no evidence to support this hypothesis. Consequently, we re-evaluated the results reported by and found that these results are based on unjustified statistical assumptions. Our analysis found no link between MHC genotype and the occurrence of driver mutations. Consistent with this, we also found no relationship between cancer risk in individuals from the UK Biobank and the coverage of common driver mutations predicted from their MHC genotypes.

In Chapter 3, we performed an analysis to predict immunogenicity of somatic mutations that arise from different cancer mutation signatures. The study found that mutated peptides resulting from specific mutation signatures were more likely to be presented by certain HLA alleles compared to peptides originating from other mutation signatures. Notably, the median activity of the mutation signatures in a given cancer could be used to predict the average number of mutations inferred to be immunogenic with high accuracy ($R^2 = 0.87$). Our results revealed that variations in the immunogenicity of mutations in tumors can be attributed to the differences in immunogenicity of mutation signatures and their activities. The limited variability in mutation signature immunogenicity and activity across different types of cancer resulted in small variation in the expected immunogenicity of various cancer types. Our findings also highlighted that the MHC-I genotype is the major determinant of the predicted immunogenicity of tumors. It was also discovered that mutation signature 20 yielded the highest proportion of immunogenic mutations, based on the HLA allele frequencies in the TCGA cohort. When comparing different types of cancer in the TCGA cohort, CESC had the highest expected number of immunogenic mutations, while PRAD had the highest observed proportion of immunogenic mutations.

Recent studies have reported that patient MHC-I genotype plays a role in determining immunotherapy responses. However, the extent of this influence appears to be inconsistent, and the underlying reasons for this inconsistency remain unclear. For instance, in the case of melanoma, the B44 HLA supertype has been linked to a better response. Interestingly, non-small cell lung cancer (NSCLC) has a similar somatic mutation burden and immunotherapy response as melanoma, but the B44 supertype has not been found to have an impact on the immunotherapy re-

sponse in NSCLC. This divergence has been attributed to underlying differences in mutational processes between melanoma and NSCLC. We performed mutation signature analysis for two ICB treated melanoma cohorts. The findings of this analysis revealed a significant enrichment of C > T mutations, which is consistent with previous studies. Furthermore, we used a combination of mutation signature activity and patient-specific HLA genotype to estimate the expected proportion of immunogenic mutations for these cohorts. A higher expected proportion of immunogenic mutations was associated with a tendency towards improved overall patient survival.

To gain insights into the role of immune selection in shaping the somatic mutation landscape and consequently the progression of cancer, we must consider the types of mutations occurring in a cancer, and the underlying mutational processes driving them. In Chapter 4, we developed a method that considers the mutational and evolutionary processes involved in tumor growth to identify and quantify the immunoediting signal. The MHC-I restricted immunoediting signal was weak and inconsistent across cancer types in the TCGA cohort. Moreover, the weak immunoediting signal persists even when we use the randomized HLA alleles. Finally, we estimated that fewer than 1% of mutations inferred to be immunogenic, were removed through immunosurveillance.

In summary, firstly, we investigated the relationship between the occurrence of driver mutations in a tumor and the MHC genotype of the patient. Then, we assessed the predicted immunogenicity of mutations arising from different somatic mutation signatures. We also examined the variation in tumor immunogenicity based on the activity of mutation signatures. We used the predicted immunogenicity of samples in the TCGA cohort to evaluate the contribution of immunoediting to the mutational landscape in cancer. We also used this method to estimate an upperbound on immunoediting signal.

1 Chapter 1: Introduction

1.1 Somatic mutations

Somatic mutations are alterations to the genome that occur in somatic cells. Because somatic cells are isolated from the germline these mutations do not get passed on to offspring [1–4]. Somatic mutations can result from errors in DNA repair mechanisms, exposure to environmental stressors such as smoking, radiation and some specific chemicals, or as a direct response to cellular stress [5, 6]. Somatic mutations can be characterised into following types:

- **Single nucleotide variants (SNVs)** are defined as alterations in the DNA sequence where a single nucleotide base is changed [7].
- **Indel mutations** also known as insertion-deletion mutations, refer to genetic alterations in which nucleotides are either inserted or deleted from a DNA sequence. These mutations can cause a shift in the reading frame, leading to changes in the amino acid sequence during protein synthesis [8].
- **Copy number variations (CNVs)** refer to the structural variation in the genome where the number of copies of certain DNA segment varies among individuals [9–12].
- **Structural variations (SVs)** refer to DNA rearrangements in the genome that involve alterations in the structure and organization of genetic material [13, 14].

These mutations can arise at any stage of life cycle of an organism and are a normal part of aging [15–21]. However, they can also give rise to oncogenesis, contributing to the development and progression of various types of cancer [22]. Somatic mutations can disrupt normal cellular processes, affect gene expression patterns, and drive uncontrolled cell growth and proliferation [20, 23–25]. Somatic mutations are important in the context of cancer as they can lead to alterations in protein function, gene regulation, and cellular processes [20, 25–27]. The impact of somatic mutations on cancer growth is twofold. Firstly, these genetic variations play a role in the development and progression of cancer by affecting crucial oncogenic pathways

and cellular signaling networks. SNVs can lead to aberrant protein structure and function, dysregulated cell growth, and evasion of tumor suppressor mechanisms, thereby promoting tumorigenesis [20, 25, 26, 28–30]. Secondly, somatic mutations play a vital role in triggering an immune response against cancer cells. Tumor-associated mutations can generate neoantigens, which are novel protein fragments derived from mutated genes [31]. The immune system targets these neoantigens, enabling the recognition and elimination of cancer cells by immune effector cells, such as cytotoxic T lymphocytes (CTLs) [31, 32].

1.2 Mechanisms of somatic mutation

Understanding the underlying mutation mechanisms is of paramount importance in comprehending the processes contributing to somatic mutations and their implications in various biological contexts. Mutations can arise through diverse mechanisms, such as errors during DNA replication and recombination, collectively referred to as endogenous mutagenic mechanisms [33–36]. Additionally, external factors, including environmental mutagenic agents [6, 37] can also induce alterations in DNA. The mutation rate is influenced by the interplay between error-producing processes and DNA repair mechanisms [37], making DNA repair a critical aspect to consider.

1.2.1 DNA replication errors

The accurate transmission of genetic information from parent cells to daughter cells is ensured by the fundamental process of DNA replication. However, despite its remarkable fidelity, DNA replication is not error-free, and mistakes can occur during this process [38]. DNA polymerase enzymes, responsible for copying the DNA template during replication, occasionally make mistakes by inserting the wrong nucleotide or by inserting too many or too few nucleotides into the growing DNA strand [38–46]. These replication errors can give rise to mutations, which are permanent alterations in the DNA sequence. While replication errors occur infrequently, the large number of DNA replication events taking place in an organism's lifetime makes them a significant source of mutations [6, 47–49].

1.2.2 DNA damage and repair

Maintaining the integrity of the genome is crucial for the survival and normal functioning of living organisms. Organisms have developed various DNA repair mechanisms to avoid the accumulation of DNA damage [45]. These repair pathways are responsible for detecting and correcting DNA lesions, ensuring the preservation of genetic information and the prevention of mutations [39–43, 50–56]. DNA repair mechanisms are known for their high efficiency and ability to fix a wide variety of DNA damage, such as single-strand and double-strand breaks, base modifications, and bulky DNA adducts [57]. They employ a sophisticated network of proteins that work together to recognize, excise, and replace damaged DNA segments with the correct nucleotides [45, 48, 58]. Importantly, these repair processes are essential for maintaining genomic stability and preventing the onset of genetic diseases, accelerated aging, and cancer [49, 59] but sometimes DNA damage is irreparable [60] or the DNA polymerases engaged in DNA repair mechanisms make mistakes and cause mutations in the DNA sequence [57].

1.2.3 Exogenous mutagenic agents

Exogenous mutational processes refer to the factors and agents external to the organism that contribute to the generation of mutations in the DNA of cells [6, 33]. These processes are distinct from endogenous mutational processes, which arise from normal cellular activities discussed in previous section. Exogenous mutational processes are influenced by various environmental and external factors. Some common examples of exogenous mutational processes include exposure to mutagens like ultraviolet (UV) light, ionizing radiation, certain chemicals, and carcinogens present in tobacco smoke or industrial pollutants. Additionally, mutational processes can be induced by therapeutic interventions, such as chemotherapy or radiation therapy [33].

1.3 Contribution of mutations to cancer development

Cancer often originates in stem cells, and the number of stem cell divisions correlates with the risk of cancer development in specific tissues [15, 20]. Somatic mutations

play a crucial role in the progression of cancer. As discussed previously these mutations are genetic changes that occur in non-germline cells during an individual's lifetime as normal part of ageing, and are not inherited from parents. However, some specific somatic mutations in DNA of a cell provide a growth advantage to the cell, leading to the clonal expansion of cells carrying these mutations. These mutations are crucial in driving the development and progression of cancer and are known as driver mutations [61]. These somatic mutations can be caused by various factors including endogenous processes, environmental exposures, genetic predispositions, and lifestyle choices that increase the likelihood of cancer occurrence. Somatic mutations incorporated through errors in DNA replication during each stem cell division contribute to cancer development. This makes age one of the most significant risk factors for cancer development [62]. Similarly, several environmental exposures have been linked to an increased risk of cancer. Exogeneous factors causing somatic mutations such as smoking, alcohol consumption, exposure to ultraviolet (UV) light, and aristolochic acid have been identified as significant risk factors in many cancers [63]. Smoking, in particular, leads to an increased mutational burden and higher lung cancer risk for smokers [64–67]. UV exposure has been linked with the development of skin cancers.

1.3.1 Role of somatic mutations in cancer development

Driver mutations provide a selective growth advantage to cancer cells, leading to uncontrolled proliferation and tumor formation [68–70]. Nine driver genes contain approximately 50% of all early clonal driver mutations, while subclonal driver mutations are found in 35 different genes, indicating a diverse set of drivers in later tumor evolution [71]. Driver genes are usually categorized into two types: oncogenes and tumor-suppressor genes. Oncogenes before they acquire mutation are known as proto-oncogenes and are involved in regulating cell division [72, 73]. When oncogenes are mutated or activated, they can drive uncontrolled cell proliferation, contributing to the development and progression of cancer [73]. On the other hand, tumor suppressor genes regulate cell growth, preventing uncontrolled division, and promoting DNA repair. In normal cells, these genes act as "brakes"to prevent the development of cancer [74]. When tumor suppressor genes are mutated or inactivated, the brakes

are released, leading to unrestrained cell growth and an increased risk of cancer [72, 74–77]. The impact of mutations in oncogenes and tumor suppressor genes on cancer development is significant.

In contrast, passenger mutations are somatic mutations that do not directly contribute to cancer development. Instead, they occur randomly and are carried along with driver mutations during tumor evolution [69]. While passenger mutations do not individually contribute to the growth of cancer, their collective presence serves as a vital baseline against which driver mutations are identified and evaluated. Through comparative analyses of tumor and normal genomes, researchers can distinguish between the stochastic background of passenger mutations and the select set of driver mutations responsible for promoting malignant transformation.

1.3.2 Role of germline mutations in cancer development

In recent studies, it has been found that germline variants not only contribute to cancer risk but also play a role in tumor progression [78]. Germline mutations are inherited genetic alterations present in the germ cells, which can be passed from one generation to the next. Patients who inherit mutations in tumor suppressor genes (TSGs) or oncogenes tend to develop cancer at a younger age compared to those without these mutations. For instance, germline mutations in BRCA1 and BRCA2 genes raise the risk of ovarian and breast cancers in women, prostate cancer in men, and pancreatic cancer in both genders [79]. Another example is Lynch syndrome, an inherited disorder that significantly increases the risk of multiple cancers, particularly colorectal cancer [80, 81]. Lynch syndrome is caused by germline mutations in genes involved in DNA mismatch repair, such as MLH1, MSH2, MSH6, or PMS2 [79–82].

1.4 Cancer progression and heterogeneity

1.4.1 Clonal expansion of cancer

The clonal expansion of cancer involves the proliferation of cancer cells with identical mutations, forming a clone within a tumor. This expansion arises from accumulated somatic mutations and genetic alterations in a subset of cancer cells, giving them a growth and survival advantage within the tumor [83–85]. This concept aligns with

the clonal theory of cancer evolution, which proposes that normal cells undergo a series of mutations, ultimately leading to the development of malignant cancerous cells [83]. Research indicates that most tumors are monoclonal, originating from a single transformed cell that proliferates into a mass of cells with a shared ancestor [86]. However, many of these accumulated somatic mutations lack growth advantages and contribute to the diversity of cells within the tumor, leading to competition for resources [87].

1.4.1.1 Clonal mutations

Occasionally a mutation occurs in a gene that enhances cell proliferation or inhibits cell death, giving the affected cell a competitive edge over others [25]. If this cell is allowed to proliferate without restraint, it leads to the expansion of a cell population with identical mutations as the original founder clone [83, 84, 88]. Such mutations are known as clonal mutations. Clonal driver mutations are crucial in initiating tumor growth and are often associated with key features of cancer progression [83, 85]. The identification of clonal mutations is essential for understanding the primary drivers of tumorigenesis and identifying potential therapeutic targets [85, 89].

1.4.1.2 Subclonal mutations

After a cell has undergone the transformation into a cancer cell, it can acquire additional somatic mutations called subclonal mutations. These mutations are found only in a fraction of the tumor cells within a tumor mass [71]. While the majority of these subclonal mutations do not offer any selective advantage [73, 74], there are cases where some of these mutations can lead to late clonal expansions, giving rise to distinct cellular populations within the tumor [90]. These mutations can be subject to random drift and may not be consistently maintained or propagated within the tumor [91, 92]. While subclonal mutations may not be the primary drivers of tumorigenesis, they can have implications for cancer treatment, as they may play a role in treatment resistance and disease progression [91, 93–96].

1.4.2 Intra-tumor heterogeneity

Intra-tumor heterogeneity (ITH) refers to the presence of diverse genetic, phenotypic, and functional characteristics within a single tumor sample [97]. It means

that different cells within the same tumor can have distinct mutations, gene expressions, and other cellular properties. Clonal evolution of cancer contributes to ITH by forming a tumor with diverse cellular populations, each driven by different mutations [98]. The presence of subclones adds complexity to the mutational landscape of tumor and response to therapy. Some subclones might be more aggressive, leading to tumor progression and treatment resistance, while others may be less harmful or responsive to treatment [83–85, 92, 95, 99]. To understand the full landscape of intra-tumor heterogeneity, it is essential to understand the role of clonal and subclonal mutations in cancer, especially in the context of the response of the immune system to cancer cells [100]. It helps in identifying critical mutations that are responsible for tumor growth and drug resistance, guiding the development of targeted therapies and personalized treatment strategies to improve patient outcomes [101, 102]. We will explore how immune selection impacts clonal and subclonal mutations in primary tumors in Chapter 4 of the thesis.

1.5 Mutation signatures

Mutation signatures characterise the patterns of somatic mutations found in the genome of cancer cells [103]. Each mutation signature reflects the underlying cause or mechanism that led to the mutations, providing insights into the molecular events driving tumor growth and evolution [103–109]. For example, mutation signatures may arise from exposure to carcinogens, such as tobacco smoke or ultraviolet radiation, which induce specific DNA damage and mutations [5, 64, 65]. Other signatures may result from defects in DNA repair mechanisms, errors during DNA replication, or the activation of specific mutagenic enzymes [110].

1.5.1 Types of mutation signatures

- **Single-Nucleotide Variants (SNVs)** mutation signatures refer to specific patterns of mutations in the genome where a single nucleotide (DNA base) is substituted for another [111]. Mutation signatures associated with SNVs can arise from various mutational processes, including exposure to environmental carcinogens, defects in DNA repair mechanisms, or errors during DNA replication [104, 105, 111]. Each mutational process leaves a distinct mark on the

genome, resulting in specific SNV patterns that can be analyzed and used to infer the underlying causes of mutations in cancer cells [108]. Understanding SNV mutation signatures is crucial for identifying the genetic alterations driving cancer development, predicting disease prognosis, and designing targeted therapies based on the specific genetic makeup of an individual's tumor.

- **Insertion and deletion signatures** Insertion and deletion signatures refer to specific patterns of mutations in the genome where nucleotides are either inserted or deleted from the DNA sequence. These mutations are collectively known as indels and can vary in length, ranging from a single nucleotide insertion or deletion (INDEL) to larger insertions or deletions of several nucleotides [112]. Like other mutation signatures, insertion and deletion signatures can result from various mutational processes, such as exposure to carcinogens, defects in DNA repair pathways, or errors during DNA replication [108].

- **Copy number signatures** refer to characteristic patterns of genomic alterations that involve changes in the number of copies of specific regions of DNA in a cancer genome. These alterations are known as copy number variations (CNVs) and can include amplifications (increased copy numbers) or deletions (decreased copy numbers) of genomic segments [113].

1.5.2 Identification and classification of mutation signatures

Non-Negative Matrix Factorization (NMF) is a widely used method for extracting mutation signatures from cancer genomic data. NMF is a linear algebra technique that decomposes a given matrix into two non-negative matrices, aiming to find a low-dimensional representation of the original data. In the context of mutational signatures, NMF is applied to a mutation count matrix, where rows represent different mutations, columns represent different samples, and the values denote the mutation counts in each sample for each mutation type. The core assumption of NMF is that the input matrix can be approximated as a product of two matrices, W and H, where both W and H are non-negative. The process of NMF involves several steps, starting with preprocessing and normalization of the mutation count matrix. The number of mutation signatures (k) to extract is an important consideration, and

it can be determined based on prior knowledge or cross-validation [106, 114–118].

Bayesian Inference is a statistical approach that can be used for extracting mutation signatures by incorporating prior knowledge and uncertainties into the analysis. In this method, the mutation count matrix representing the mutational profile of different samples is modelled as a probabilistic distribution. Prior distributions are defined for the mutation signatures to represent their expected contributions based on existing knowledge. The likelihood function describes the probability of observing the mutation count matrix given the mutation signatures and their activities in each sample. By combining the likelihood function and the prior distributions, Bayesian methods compute the posterior distribution, representing the updated beliefs about the mutation signatures [117]. To approximate the posterior distribution, Markov Chain Monte Carlo (MCMC) sampling is commonly used. The MCMC samples are then utilized to estimate the signature activities for each sample, providing insights into the contributions of different mutational processes to individual mutation profiles [119]. Bayesian Inference offers a robust and flexible approach, allowing the integration of prior knowledge and uncertainties, which is essential for analyzing complex genomic data from cancer samples [114, 117, 120–122].

Principal Component Analysis (PCA) is a dimensionality reduction technique used for mutation signature extraction from cancer genomic data. It identifies major patterns of variation in the mutation count matrix and projects the data onto a lower-dimensional space while preserving significant features. The process involves creating a covariance matrix from the normalized mutation count matrix, performing eigenvalue decomposition to identify principal components, and selecting the top components that explain most of the variance. Data is then transformed into the lower-dimensional space, representing each sample's contribution to the identified principal components. These principal components can be interpreted as mutation signatures, and their loadings reflect the association of each mutation type with the corresponding signature [123–125]. PCA is valuable for understanding dominant mutational processes in cancer samples and their potential associations with specific risk factors or DNA repair deficiencies. However, it may not capture all subtle variations, and other methods like Non-Negative Matrix Factorization (NMF) or Bayesian Inference can complement PCA in identifying a broader range of muta-

tional signatures [126].

1.5.3 Reference databases and catalogues

Mutation signatures reference databases and catalogues play a crucial role in advancing our understanding of the diverse mutational processes underlying human cancers. These databases provide categorized reference signatures, allowing researchers to analyze large cohorts of sequencing data and identify the contribution of different mutational processes to specific cancer types. The COSMIC mutational signatures database is a prominent resource in this field, curated in collaboration with Cancer Grand Challenges, the Wellcome Sanger Institute, and other institutions. It encompasses a wealth of information on mutational signatures, and it is continuously updated with the latest data from various cancer patients' genomic profiles. The signatures in COSMIC have been identified using sophisticated methods such as NMF and bayesian inference, which allow for the extraction and characterization of different mutation patterns from vast datasets. Moreover, the database provides interactive tools and visualizations for exploring and analyzing mutational signatures across different cancer types, offering a comprehensive view of their diversity and relevance in cancer research [108]. Another recently added resource is mSignatureDB, which offer valuable tools for deciphering mutational signatures in human cancers [127].

1.5.4 Aetiology of mutation signatures

Mutation signatures provide valuable biological insights into the underlying mutational processes that contribute to tumor development and progression in various cancers. While the origins of many mutational signatures remain uncertain, the analysis of mutational signatures can, in certain instances, reveal the external and internal mutational processes that have contributed to the observed genetic changes [128]. These signatures offer information about the specific mutagenic factors or biological mechanisms that lead to the accumulation of specific mutations in cancer genomes. Some of the well-studied mutation signatures are associated with smoking, ultraviolet (UV) light exposure, and APOBEC cytidine deaminases [129].

- Smoking-Associated Signatures: Smoking is a major risk factor for cancer,

and mutational signature analysis has revealed distinct patterns of mutations associated with tobacco smoke exposure (Figure 1.1). These signatures are characterized by specific base substitutions, such as C>A transversions, occurring predominantly at cytosine bases within a specific sequence context [130]. Smoking-associated signatures have been identified in various smoking-associated cancers, including lung cancer and several others, providing evidence of tobacco smoke-induced DNA damage and mutagenesis [131, 132]. The presence of these signatures in cancer genomes highlights the link between smoking and the mutational landscape of specific cancer types.

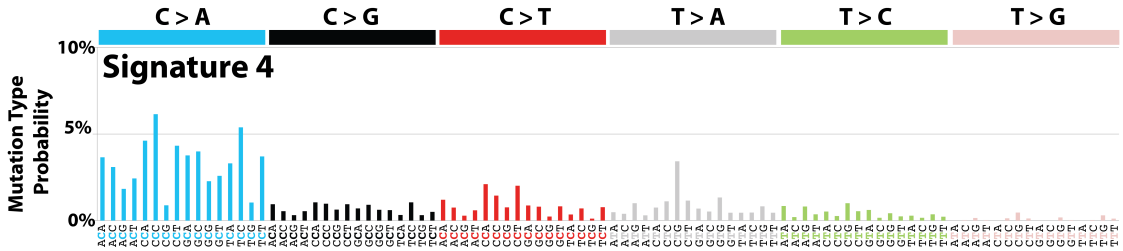

Figure 1.1: Mutation signature 4, associated with tobacco smoking. This figure is retrieved from https://cancer.sanger.ac.uk/signatures/sbs/sbs4/.

- UV Light-Induced Signatures: UV light exposure is a known risk factor for skin cancer, particularly melanoma. UV light generates specific DNA lesions, such as cyclobutane pyrimidine dimers (CPDs) and pyrimidine-pyrimidone [133–135] photoproducts, which lead to characteristic mutational patterns in cancer genomes. The mutational signatures associated with UV exposure are characterized by C>T transitions (Figure 1.2) predominantly at dipyrimidine sequences [130, 136, 137]. These signatures are prevalent in melanoma samples and are reflective of the DNA damage induced by UV light.

- APOBEC-Related Signatures: The APOBEC family of cytidine deaminases can induce mutations in cancer genomes. APOBEC enzymes target cytosines in single-stranded DNA, leading to C>G or C>T transitions, often in the context of specific trinucleotide motifs (Figure 1.3). Two major APOBEC-related signatures are commonly observed: APOBEC3A (APOBEC3A Signature) and APOBEC3B (APOBEC3B Signature) [138]. These signatures are prevalent in several cancer types and are associated with the activity of APOBEC enzymes

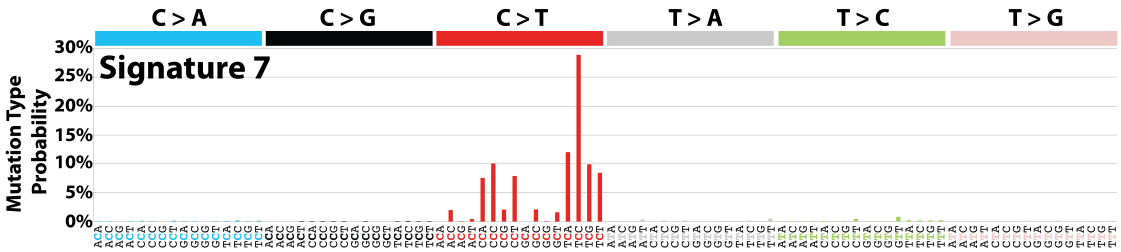

Figure 1.2: Mutation signature 7, associated with ultraviolet exposure. This figure is retrieved from https://cancer.sanger.ac.uk/signatures/sbs/sbs7/.

in cancer cells [108, 138, 139]. Additionally, recent studies have demonstrated substantial temporal and spatial variability in APOBEC-related signatures in cancer cells, providing further insights into the dynamics of mutational processes caused by APOBEC cytidine deaminases [138, 140].

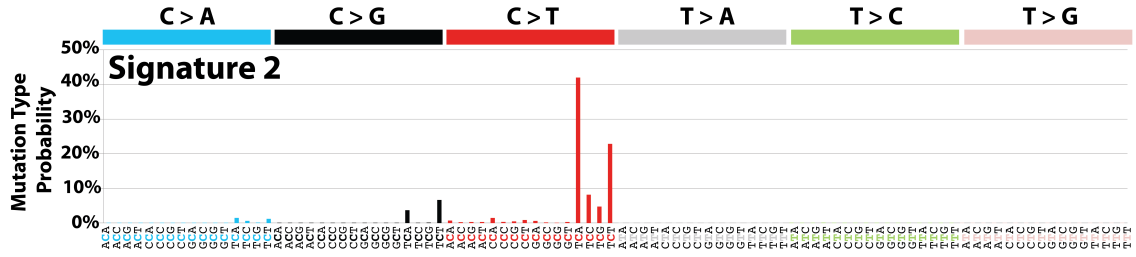

Figure 1.3: Mutation signature 2, associated with APOBEC. This figure is retrieved from https://cancer.sanger.ac.uk/signatures/.

In summary, mutational signatures associated with smoking, UV light exposure, and APOBEC activity provide critical biological insights into the DNA damage processes that contribute to the development of specific cancer types. The identification and understanding of these signatures enhance our knowledge of cancer aetiology, potential preventive measures, and personalized treatment strategies for affected individuals. We will be using COSMIC mutation signatures catalogue for our research.

1.6 Overview of the immune system

Humans have developed sophisticated innate and adaptive immune mechanisms by evolution [141, 142]. There are mainly two components of our immune system, namely innate and adaptive immune system [143]. The innate immune system acts

as the body's first line of defense, rapidly responding to any invading microorganism it encounters, even those that have not been previously encountered, in a non-specific manner [142]. On the other hand, the adaptive immune system recognizes specific threats and possesses a memory, enabling it to recall prior exposures to pathogens [144]. This memory aspect of the adaptive immune system allows it to mount quicker and more effective responses upon subsequent encounters with the same pathogen [145, 146].

When the body encounters foreign invaders, such as bacteria, viruses, or cancer cells, immune cells are activated to recognize and eliminate these threats, thus safeguarding the body against infections and diseases. Immune cell activation and effector functions are essential processes in the immune system's response to pathogens and cancerous cells once they are recognized by immune system [147–149]. Immune cell activation involves the recognition of specific antigens present on the surface of pathogens or cancer cells. Different types of immune cells, including T cells, B cells, and natural killer (NK) cells, play crucial roles in the immune response. Upon encountering antigens, these immune cells become activated and initiate a series of effector functions to eliminate the threat [148, 149].

1.6.1 Functions of the Immune System

Immune surveillance is a critical mechanism through which the immune system constantly monitors the body to identify and eliminate abnormal or transformed cells, including cancerous cells [150]. This process involves the recognition of specific antigens displayed on the surface of cancer cells, which distinguish them from normal healthy cells. When the immune system detects these antigens, it initiates an immune response to target and destroy the cancerous cells, thereby acting as a primary defense against cancer development and progression [150, 151]. Numerous studies have provided evidence supporting the concept of immune surveillance against tumors. Immunodeficient mice lacking key immune effector cells have been shown to develop spontaneous tumors at higher rates, suggesting the protective role of the immune system in suppressing tumor growth [152]. Additionally, the presence of inflammatory immune cells in human tumors raises questions about how cancer cells avoid immune attack and suggests the existence of mechanisms that mimic

peripheral immune tolerance to evade destruction [153]. T cells, a type of white blood cell, play a central role in cell-mediated immunity. They recognize antigenic peptides presented on the surface of infected cells or cancer cells through the major histocompatibility complex (MHC) molecules [154]. Activated T cells can directly kill infected or cancerous cells through the release of cytotoxic molecules, such as perforin and granzymes, or by inducing apoptosis (programmed cell death) [155]. Cytotoxic T cells (CD8+ T cells) recognize and directly attack infected cells, cancer cells, and cells presenting foreign antigens. Cytotoxic T cells play a crucial role in eliminating abnormal cells from the body [156–158], whereas, helper T cells (CD4+ T cells) do not directly kill cells but orchestrate and coordinate immune responses. They assist in activating other immune cells, such as cytotoxic T cells and B cells. Helper T cells are essential for initiating and maintaining effective immune reactions against various threats, including cancer [148]. Another important type of T cells is regulatory T cells, also known as Tregs. These cells help prevent the immune system from overreacting and causing damage to the body's own tissues. While they play a critical role in maintaining immune balance and preventing autoimmune responses, their presence can also hinder anti-cancer immune responses in certain contexts [159, 160].

B cells, another type of white blood cell, are key players in humoral immunity. When activated by specific antigens, B cells differentiate into plasma cells that produce and release antibodies. These antibodies can bind to pathogens or cancer cells, marking them for destruction by other components of the immune system or by triggering complement-mediated lysis [155]. Memory B cells are long-lived cells that "remember"previous encounters with specific antigens. They allow the immune system to mount a faster and more effective response upon re-exposure to the same antigen, contributing to immunological memory and enhancing the body's ability to fend off recurrent infections or threats [161]. Natural killer (NK) cells are part of the innate immune system and play a vital role in immunosurveillance against infected or transformed cells, including cancer cells. NK cells can recognize and directly kill target cells that lack MHC class I molecules or display stress-related ligands on their surface [162, 163]. During the initial encounter, specialized immune cells, such as T cells and B cells, are activated and differentiate into effector cells to combat the

invader. Once the infection is resolved, a subset of these immune cells transforms into memory cells [145, 146]. Immunological memory is a critical aspect of the adaptive immune system, providing the ability to recognize and respond more effectively to specific antigens upon subsequent encounters. This process plays a crucial role in both protecting against infectious agents and influencing the immune response against cancer cells [164, 165]. In the context of cancer, immunological memory also plays a crucial role. Cancer cells often express unique antigens, which can be recognized by the immune system. When cancerous cells are first detected, the immune system initiates an immune response to eliminate them. Some of the activated immune cells transform into memory cells with specificity for cancer antigens [166–168]. The formation of memory cells is characterized by three main features:

- Longevity and Independence: Memory immune cells are long-lived and persist in the body even in the absence of continuous antigen stimulation. They can be maintained through homeostatic turnover or stable maintenance, ensuring a lasting immune response [169–171].

- Antigen Specificity: Memory cells are highly specific for the antigen they encountered during the primary response. This specificity allows them to recognize and respond rapidly to the same antigen if encountered again in the future [172, 173].

- Enhanced Function: Memory cells undergo changes during the initial encounter that enhance their function. They become more effective at recognizing and eliminating the pathogen, resulting in a faster and more efficient immune response during subsequent encounters [164, 168, 171].

The immune system's ability to activate these immune cells and orchestrate their effector functions is critical in controlling infections and preventing cancer development and progression. In cancer, the activation of immune cells and their effector functions are central to the field of cancer immunotherapy [174]. Immunotherapies aim to boost the immune response against cancer by enhancing immune cell activation or overcoming the mechanisms of immune evasion employed by cancer cells [174, 175].

1.6.2 Antigen Presentation Pathway

Tumor antigens play a critical role in the interaction between the immune system and cancer cells during immune surveillance. Tumor antigens are specific molecules expressed on the surface of cancer cells by MHC-I molecules, that differentiate them from normal cells. MHC class I molecules are present on most nucleated cells and present intracellular antigens These antigens can be derived from various sources, including mutated proteins as result of SNVs, which are the primary focus of this study. When the immune system recognizes these tumor antigens as foreign or abnormal, it can mount an immune response to eliminate the cancer cells and prevent tumor development [176–179]. This recognition triggers the destruction of cancer cells through the release of cytotoxic molecules and the induction of apoptosis [144, 145, 180]. MHC-I molecules play a crucial in initiating an immune response against cancer cells. Peptide loading and presentation by MHC-I molecules are intricate processes that involve several steps (Figure 1.4):

- Antigen Processing: Antigenic peptides are generated through the degradation of intracellular proteins, such as viral proteins or cellular components, by the proteasome. These peptides are typically 8 to 11 amino acids in length [181].
- Transport into the Endoplasmic Reticulum (ER): The generated peptides are transported into the ER by the Transporter Associated with Antigen Processing (TAP) complex, which is composed of TAP1 and TAP2 subunits [182].
- Peptide Loading Complex (PLC) Formation: Inside the ER, MHC-I heavy chains associate with chaperone proteins, including tapasin, calreticulin, ERp57, and β2-microglobulin, forming the peptide loading complex (PLC) [183].
- Peptide-MHC-I Complex Formation: Tapasin plays a key role in bridging the TAP transporter and MHC-I, facilitating the loading of peptides onto MHC-I molecules. The appropriate peptides that bind with high affinity to the MHC-I groove are selected for presentation [183].
- Cell Surface Presentation: The Golgi apparatus, also known as the Golgi complex, does additional processing of the proteins received from the endoplasmic

reticulum (ER) [184]. The stable peptide-MHC-I complex is then transported to the cell surface, where it is presented for recognition by CD8+ T cells [185, 186].

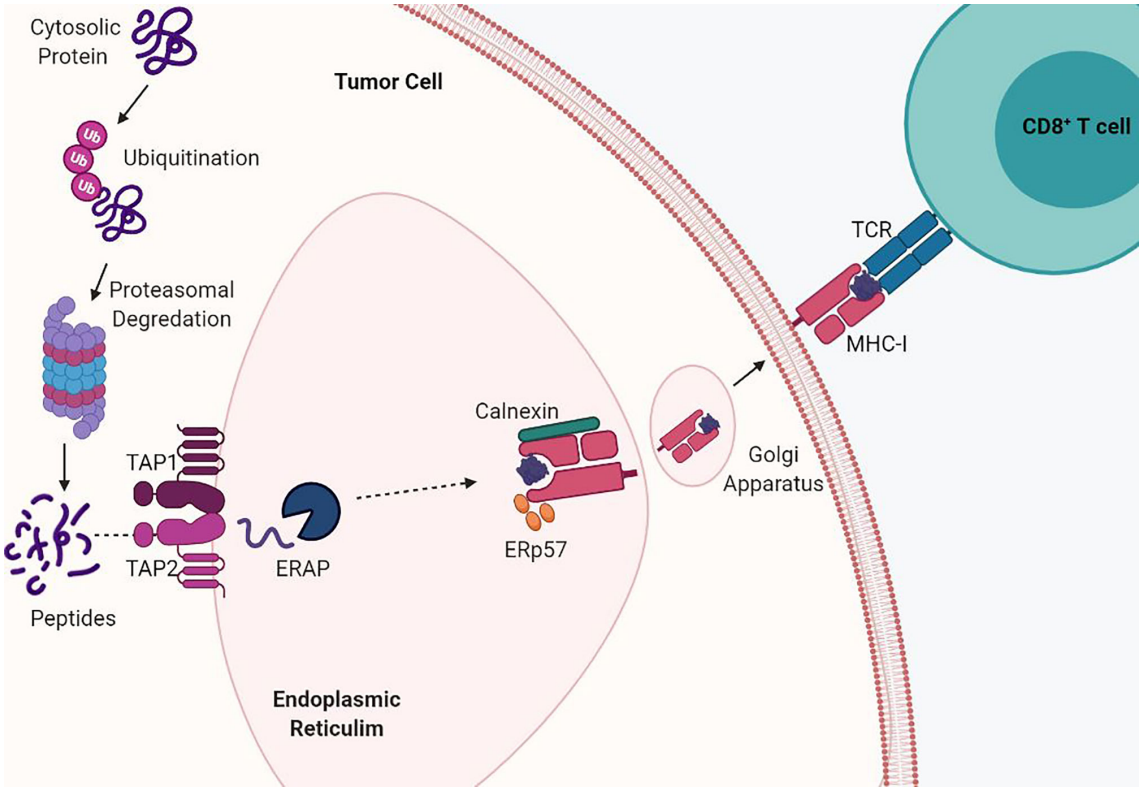

Figure 1.4: MHC-I antigen processing and presentation pathway. MHC-I presents endogenously derived peptide antigens to CD8+ T cells. The proteasome breaks down proteins in the cytosol into peptides, which are then transported into the endoplasmic reticulum by TAP transporter proteins. Afterwards, the antigen peptides are loaded onto the MHC-I α-heavy chain and beta-2-microglobulin (β2M) complex, which is then transported to the cell surface through the Golgi. Reprinted with permission from Taylor and Balko 2022 [187].

T cells have receptors called T cell receptors (TCRs) that can recognize antigenic peptides when presented by MHC molecules. Helper T cells (CD4+ T cells) recognize antigens presented by MHC class II molecules, while CD8+ T cells recognize antigens presented by MHC class I molecules. When a T cell encounters a cancer cell displaying a peptide derived from a tumor antigen, TCR binds to the peptide-MHC-I complex on the cancer cell's surface. In addition to TCR binding, a co-stimulatory signal is required for full T-cell activation. This co-stimulation

is typically provided by interactions between molecules on the surface of T cells (e.g., CD28) and molecules on the cancer cell's surface (e.g., B7). Co-stimulation ensures that T-cell activation only occurs when there is a genuine threat, preventing unwarranted immune responses [188].

TCR binding and co-stimulation lead to the activation of intracellular signaling pathways within the T cell. This signaling cascade results in the production of various signaling molecules and transcription factors that promote T-cell activation and proliferation [189]. Once activated, the T cell undergoes clonal expansion, where it rapidly divides to produce a population of effector T cells specific to the tumor antigen [190]. This increases the number of T cells available to target and eliminate cancer cells. The activated T cells, now termed effector CTLs, perform their function. CTLs release granules that are cytotoxic in nature and contain perforin and granzymes. These granules induce programmed cell death, or apoptosis, in the cancer cells that they target [191–193].

1.6.2.1 Major Histocompatibility Complex Molecules (MHC-I) The MHC-I genes present on chromosome 6, are part of a highly diverse and polymorphic genetic region that plays a critical role in the adaptive immune response of vertebrates [194]. This genetic variability is particularly important at the MHC level due to its influence on various biological traits [195]. The diversity in MHC-I genes is believed to be maintained by pathogen-driven selection, either through heterozygote advantage or frequency-dependent selection [195]. Consequently, MHC-I genes are among the best candidates for studying mechanisms and the significance of molecular adaptation in vertebrates.

1.6.2.1.1 Genetic Diversity of MHC-I MHC-I genes exhibit extreme polymorphism, which refers to the presence of a large number of different alleles within a population. This high level of polymorphism is particularly evident in the peptide-binding domains of MHC-I proteins [196]. The binding platform of MHC-I proteins consists of two domains forming a slightly curved β-sheet base and two α-helices on top, allowing them to accommodate a wide range of peptide antigens [181, 183, 197]. Notably, MHC-I alleles may differ significantly in the range of antigens they bind, and this diversity has been linked to better resistance to local

parasites [198–200].

1.6.2.1.2 Impact of MHC-I expression levels on antigen presentation

MHC-I molecules play a crucial role in antigen presentation, where they present peptide antigens derived from intracellular pathogens on the cell surface for recognition by CD8+ T cells. The level of MHC-I expression on the cell surface is a critical factor in determining the efficiency of antigen presentation and subsequent T cell activation. Effective antigen presentation is essential for the immune system to recognize and eliminate transformed or infected cells expressing abnormal proteins [201, 202].

In the context of cancer and viral-mediated diseases, genetic variations in antigen-processing genes of the MHC-I pathway can influence antigen presentation and immune responses. Certain genetic variations in components of the antigen presentation machinery are risk factors for different types of cancer, highlighting their role in cancer development and progression [203].

1.6.3 Immune evasion

One of the hallmarks of cancer is immune evasion [204]. Immune evasion is the ability of cancer cells to avoid recognition and elimination by the immune system, enabling them to survive and proliferate within the host [153, 204]. Multiple mechanisms are involved in cancer immune evasion, contributing to tumor progression and resistance to immune-based therapies.

For the immune system to effectively impact tumor growth and influence the tumor genome, it requires a fully functional antigen presentation machinery and the presence of immune cells capable of recognizing and eliminating cancerous cells within the tumor microenvironment. However, cancer cells have evolved strategies to evade detection by disrupting the antigen presentation pathway pathway [204–206]. In many cancers, there are recurrent mutations observed in key players of the APM, such as MHC-1 and b2-microglobulin (B2M), resulting in their downregulation [207, 208]. B2M is essential for the formation and stabilization of MHC on the cell surface. When B2M is lost, the MHC cannot properly form, leading to a form of immune escape in cancers [209]. While mutations in the B2M gene are rare, down-

regulation of the gene is more commonly observed in cancer cells [208]. One critical mechanism is the downregulation of MHC-I molecules in general, which disable the antigen presentation on the tumor cell surface, making them invisible to cytotoxic T lymphocytes (CTLs) [210]. Furthermore, cancer cells may overexpress HLA-G, a non-classical HLA molecule known for its immunosuppressive properties, which inhibits natural killer cells and cytotoxic T lymphocytes (CTLs). HLA-G is typically expressed in immune-privileged tissues but is frequently overexpressed in tumors [207]. Loss of heterozygosity (LOH) in HLA alleles, is suggested as another way for tumors to evade the immune system. For example the loss of HLA-C08:02 which has high binding capacity for KRAS G12D neoantigen was observed in tumors that showed resistance to CD8+ T cell treatment targeting mutant KRAS [211]. Further research revealed that HLA LOH is common in lung cancer, occurring in 40% of early-stage NSCLCs [212]. Additionally, cancer cells downregulate TAP1, further contributing to their escape from immune recognition [213]. The downregulation of critical components in the antigen presentation pathway hinders the capacity of the immune system to identify antigens on the surface of cancer cells.

Another immune evasion mechanism is blocking immune activation signals. This can be achieved by interacting with immune checkpoints, which hinder the activation of the immune response [214]. These checkpoints act as barriers preventing T cells from carrying out their function of killing cancer cells. Notably, cancer cells themselves can engage these checkpoints, but they can also be activated by dendritic cells and macrophages [215]. The immune system has learned to dynamically regulate its responsiveness, and maintaining a delicate balance between activation and inhibition is crucial. This balance ensures that immune cells do not mistakenly attack healthy normal cells. Dendritic cells play a dual role in this process: they send signals to activate T cells for their anti-cancer functions, while also moderating the ability of T cells to respond effectively against a threat [159]. However, in certain cases, this balance can be disrupted, and the scale tips towards excessive inhibition. Consequently, T cells become suppressed and are unable to carry out their task of eliminating cancer cells [216]. This phenomenon contributes to the immune evasion strategies employed by cancer cells, enabling them to proliferate and evade destruction by the immune system. To compound their immune evasion

strategies, cancer cells also increase the expression of certain proteins like PDL1 and NF-kb. These proteins act as checkpoint inhibitors, effectively blocking the immune system's response [217, 218].

Another immune evasion mechanism is overly active Regulatory T cells (Tregs). Tregs typically monitor and regulate the activity of effector T cells, are found in elevated levels in various cancer types [160, 219]. Moreover, Tregs within tumors have been demonstrated to exhibit higher suppressive functionality compared to Tregs in normal tissue samples [220, 221]. Additionally, tumor cells can secrete proteins that suppress the response of effector T cells and promote the proliferation of immunosuppressive cells within the tumor microenvironment (TME), such as myeloid-derived suppressor cells (MDSCs), tumor-associated macrophages (TAMs), and regulatory T cells (Tregs) [220, 222, 223]. These factors collectively contribute to the dysfunction of T cells, hindering the immune response and facilitating tumor progression.

It has been proposed that a low mutational burden is another strategy employed by tumors to evade the immune system [223]. When a tumor has fewer mutations, the likelihood of generating neoantigens that can trigger an immune response decreases. The underlying idea is that tumor cells possessing neoantigens capable of provoking a strong immune response would have been eliminated by the immune system, leaving behind cancer cells that can evade the immune system [224, 225]. In this context, the reduction in tumor immunogenicity serves as a mechanism to escape immune surveillance. However, recent research has cast doubt on this notion, as there is limited evidence supporting the depletion of neoantigens in cancer samples [226–228].

1.6.4 Immune selection

As discussed in the previous section, immune surveillance acts as the initial line of defense, detecting and then subsequently eliminating cancer cells that present immunogenic antigens by activating various immune cells against them, thus preventing tumor development and progression [150, 151]. However, some cancer cells can evolve and develop strategies to evade the immune response, leading to immune selection and the emergence of non-immunogenic or less immunogenic tumor

variants [153].

Some tumor clones are thought to undergo "immune editing,"where the immune system selects for clones that are depleted of immunogenic antigens or neoantigens. When cancer cells display neoantigens on their surface, they become targets for immune recognition and potential elimination by CD8+ T cells. However, in the immune-editing process, tumors that lose these neoantigens can escape immune detection and evade immune responses. It has been reported that the immune system plays a significant role in shaping tumor genomes by exerting selective pressures on cancer cells based on their antigenic characteristics [227–229]. However, this remains a controversial topic, and recent studies have shown that there is a lack of evidence to support these findings [70, 230].

1.6.5 Immunoediting

Immunoediting is a dynamic process wherein the immune system exerts selective pressure on developing tumors, resulting in both tumor suppression and tumor promotion [231]. It consists of three distinct phases: elimination, equilibrium, and escape (Figure 1.5). These phases collectively shape the immunogenicity of tumors and their ability to evade immune recognition and destruction [177, 232].

Elimination phase: In this initial phase, the immune system recognizes and eliminates nascent tumor cells through a process called cancer immunosurveillance. The immune response is initiated when cells of the innate immune system detect the presence of a growing tumor, triggered in part by tissue disruption due to angiogenesis or tissue-invasive growth. The anti-tumor immune response, particularly by CD8+ T cells, targets and eliminates tumor cells expressing highly immunogenic antigens, including tumor-specific mutant neoantigens. If the immune system successfully eliminates the tumor, the immunoediting process ends at this stage without tumor progression [233].

Equilibrium phase: Some tumor cells may escape complete elimination during the initial phase and enter a state of equilibrium with the immune system. During this phase, tumor growth is balanced by ongoing immune surveillance and immune-mediated control. The immune system keeps the tumor cells in check, preventing further expansion. However, tumor cells may undergo genetic and epigenetic changes

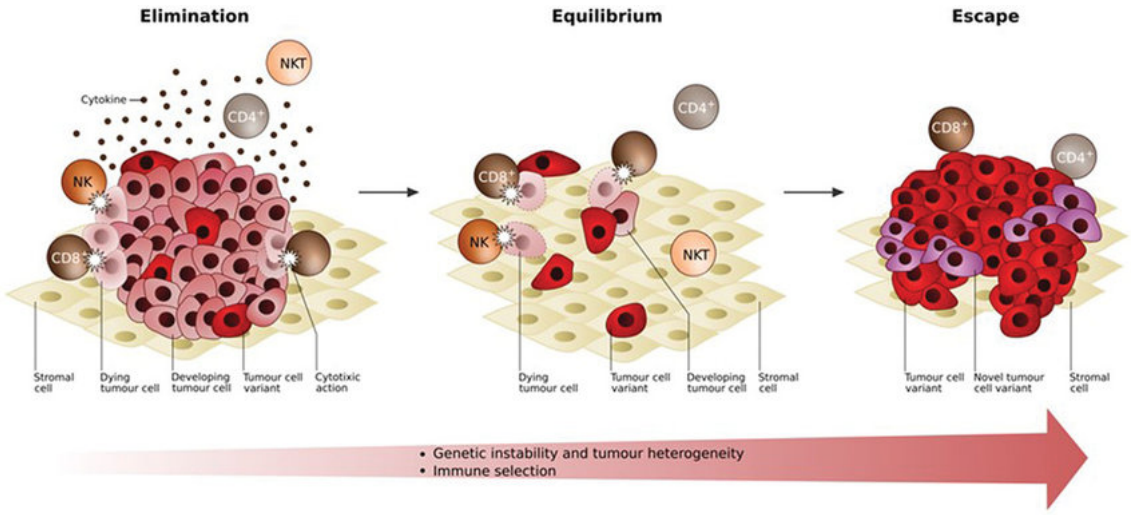

Figure 1.5: The three phases of immunoediting. (a) Elimination refers to the recognition and elimination of the tumor cells by immune system; (b) equilibrium, during this phase, the immune system selects and/or promotes the generation of immunologically resistant tumor cell variants; and (c) escape, during this phase the expansion of the tumor is now beyond the control of the immune system, and tumor cells continue to grow. Figure drawn by Wiebke Bretting, after [232].

to adapt and evade immune detection, leading to the emergence of less immunogenic variants [233].

Escape phase: In the escape phase, tumor cells that have acquired changes enabling them to evade immune recognition and destruction ultimately overcome immune control and proliferate. These tumor variants may downregulate or lose expression of strong tumor-specific antigens, becoming less visible to the immune system. This immune escape allows the tumor to grow and progress, leading to clinical manifestation and the onset of cancer [233].

MHC-I molecules play a crucial role in immunoediting [232]. MHC-I molecules are responsible for presenting antigens derived from intracellular proteins, including tumor-specific antigens, on the surface of tumor cells [228–234]. This antigen presentation allows CD8+ cytotoxic T cells to recognize and eliminate the tumor cells displaying these antigens. During the elimination phase, MHC-I presentation of tumor-specific antigens facilitates immune recognition and elimination of tumor cells [232].

In the escape phase, tumor cells may downregulate or lose the expression of

MHC-I molecules, which is a common immune evasion mechanism. This loss of MHC-I molecules prevents the presentation of tumor-specific antigens to CD8+ T cells, making the tumor cells invisible to immune surveillance. Consequently, the immune system fails to recognize and eliminate these tumor cells effectively, enabling their escape from immune control and leading to tumor progression. The downregulation of MHC-I molecules is associated with reduced responsiveness to immunotherapies, particularly immune checkpoint inhibitors that rely on MHC-I presentation to enhance anti-tumor T cell responses [187]. It has also been reported that gaps in MHC-I genotypes shape the mutational landscape of the cancer [227, 228]. However, this remains a controversial topic, and a lot of research is going on in this domain [70, 230, 234]. [234] has shown that the MHC restricted immunoediting reported by [227, 228] is caused by 13 lowly immunogenic, common hot spot mutations in 6 cancer genes. Also recent studies have highlighted the importance of considering mutational signatures [235], while estimating the impact of immunoediting on cancer.

In summary, immunoediting is a dynamic process involving three phases that collectively shape the interaction between the immune system and developing tumors. MHC-I molecules play a critical role in this process, as their expression on tumor cells allows for antigen presentation and effective immune recognition during the elimination phase, while loss or downregulation of MHC-I molecules enables immune escape and tumor progression in later phases. Understanding these mechanisms is essential for developing effective cancer immunotherapies and strategies to overcome immune evasion by tumors.

1.7 Bioinformatics methods to analyze cancer data

1.7.1 Cancer data

Several consortiums have been established with the goal to collaborate, develop and validate methods, combine resources and expertise, and generate extensive datasets to advance our knowledge and comprehension of cancer. One of the significant contributions to cancer genomics data comes from Foundation Medicine, which released genomic data for 18,004 adult cancers profiled using the FoundationOne assay [236]. The data has been collected from 162 tumor subtypes, primarily focusing on tho-

racic, gastrointestinal, breast, gynaecologic, and hepato-pancreato-biliary cancers. The American Association for Cancer Research (AACR) initiated the Genomics Evidence Neoplasia Information Exchange (GENIE) to promote data sharing among 19 different institutions. The main objective is to generate sufficient data to support clinical decision-making. GENIE comprises genomic and clinical data from an extensive cohort of 44,756 patients encompassing over 50 cancer types [237]. In the United States, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) was established in 2011 as a nationwide endeavor to expedite cancer understanding through genomic and proteomic data analysis. CPTAC focuses on 1527 samples from 9 cancer types [238]. On a global scale, the International Cancer Proteogenomic Consortium (ICPC) brings together scientists collaborating to share genomic and proteomic data from cancer samples across 12 tissue types. The overarching goal of ICPC is to utilize proteogenomic data to predict the outcomes of cancer treatments (cpc.cancer.gov).

One of the most widely used datasets in cancer research has been generated by The Cancer Genome Atlas (TCGA) [239], a collaborative initiative launched jointly by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) in December 2005. Initially beginning as a pilot program focused on three cancer types (glioblastoma, serous cystadenocarcinoma of the ovary, and lung squamous carcinoma), it has since expanded to encompass data from 33 cancer types, including primary cancer and matched normal samples. The program involves 20 collaborating institutions across Canada and the US. TCGA has produced a vast array of genomic, transcriptomic, epigenomic, and proteomic data from over 11,000 individuals, making it publicly accessible to researchers through open and controlled access types. The data is hosted on the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov).

To further aggregate and analyze whole-genome analyses, the Pan-Cancer Analysis of Whole Genomes (PCAWG) project was established. It integrates data from projects such as TCGA and the International Cancer Genome Consortium (ICGC) [240]. The PCAWG project identifies coding and non-coding variations in the cancer genomes of 2,834 individuals across 38 tumor types. The data, derived from primary tumor samples with matched normal tissues, is publicly accessible through the

ICGC database (https://dcc.icgc.org/pcawg). The International Cancer Genome Consortium (ICGC), launched in 2008, serves as a global coordinating platform for cancer genome projects [241]. It includes prominent cancer initiatives like TCGA and PCAWG, working collectively to advance cancer genomics research.

In the year 2000, the Welcome Trust Sanger Institute initiated The Cancer Genome Project with the objective of detecting genetic alterations and patterns within cancer genomes using high-throughput sequencing. The findings from this project have been made accessible through the Catalogue of Somatic Mutations in Cancer (COSMIC) database (cancer.sanger.ac.uk). This comprehensive database contains thousands of somatic mutations identified in various cancers, along with a compilation of mutational signatures discovered in human cancers [242].

1.7.2 Identification of somatic mutations

Next-generation sequencing (NGS) technologies are generally used for the identification of somatic mutations, particularly in tumor samples, revolutionizing cancer research and clinical practice [243]. In clinical settings, two common approaches are utilized for mutation detection: targeted sequencing panels and whole exome sequencing (WXS) panels. Targeted sequencing focuses on specific genes relevant to the disease, allowing for high-depth sequencing with a substantial number of reads obtained at each position [244]. On the other hand, WXS covers approximately 20,000 protein-coding genes in the human genome and typically achieves a depth of 100X across the genome [245]. Although whole genome sequencing (WGS) is an unbiased technique that covers the entire genome, its cost limitations often result in depths of only 30-50X in cancer samples [246]. This limitation poses challenges in identifying somatic mutations with low frequencies or in samples with low tumor purity [245, 246].

1.7.2.1 Sample collection Samples for bulk sequencing in cancer genomics are typically preserved using two main methods: formalin-fixed paraffin-embedded (FFPE) and fresh frozen (FF) techniques. The FFPE method involves fixing the tissue with formaldehyde solution, which halts cell metabolism, and then sealing it with paraffin to reduce oxidation rates, allowing long-term storage of the samples

[247]. On the other hand, the FF method requires freezing the sample in liquid nitrogen shortly after surgery to preserve the DNA/RNA, as thawing causes rapid degradation [248, 249].

Both methods have their advantages and limitations. FFPE is the most common method due to its lower cost, ability to store at room temperature, and longer time frame for processing after surgery [97, 250]. It also preserves tissue morphology, making it valuable for certain analyses. However, FFPE may introduce artifacts in the form of C > T mutations [251]. On the other hand, FF is advantageous in terms of better preservation of DNA/RNA compared to FFPE [248]. But it requires careful storage and handling due to its sensitivity to thawing.

To identify somatic mutations in cancer samples, matched paired tumor-normal samples are generally taken from a patient. The normal sample is usually obtained from blood, but sometimes normal tissue adjacent to the tumor is used [252]. This approach allows the identification and removal of germline variants. In some cases, a tumor sample is taken without a matched normal sample, referred to as "tumor-only,"which makes it more challenging to identify germline mutations. In such instances, databases of common variants or information from a panel of normal samples are used to remove likely germline variants [245].

1.7.2.2 Sequence alignment The initial format of the files generated by the sequencer is generally binary base call (BCL) files. However, these files are then converted to FASTQ files during the bioinformatics pipeline. FASTQ files contain the readout of the nucleotides called for each read, along with a quality score for each base. It is essential to use a tool like FASTQC to check the quality of these FASTQ files before proceeding with downstream processing. FASTQC allows researchers to identify if the reads have sufficient quality and can be used with confidence in subsequent analyses. This quality assessment is the first step of any bioinformatics pipeline. To improve the accuracy of the data, the ends of the reads are often trimmed using a tool like Trimomatic [253]. This step is necessary because the base quality tends to drop off at the 3' end during the sequencing by synthesis process of Illumina sequencing.

After the quality assessment and trimming steps, the reads need to be aligned to a reference genome to obtain positional information. There are several tools

available for read alignment, with BWA-MEM being one of the most commonly used tools for this purpose [254]. Once the reads are aligned, preprocessing steps are required before variant calling. If PCR amplification is performed during library preparation, duplicate marking is essential to identify reads generated from the same DNA molecule. However, this step should be done with caution, as it could introduce bias in the variant calling step, with some reads being over-represented in the results. Depending on the downstream analysis, an optional local realignment step using tools like GATK [255] or ABRA [256] can be performed. This step aims to limit errors caused by insertions and deletions (indels) and single nucleotide polymorphisms (SNPs). These tools utilize information from all reads at a given location to determine the best alignment of the reads, leading to more accurate variant calls. Base quality recalibration is another optional but highly recommended step. It is performed using the GATK suite of tools and helps correct inaccuracies in base quality scores assigned by the sequencer. Accurate base quality scores are crucial for reliable variant calling, and this step significantly improves the accuracy of the variant calls. However, it is computationally expensive and time-consuming. With recent improvements in sequencing technologies that have increased the accuracy of base quality scores, some researchers may choose to skip this step to reduce analysis turnaround time.

1.7.2.3 Somatic mutation calling Somatic variant calling is a crucial step in cancer research and treatment. It involves identifying genetic mutations that are present only in the tumor cells and not in the normal cells of an individual. This is typically achieved by comparing the DNA sequences of a tumor sample and a matched normal sample from the same individual. Two main types of variant callers are commonly used: position-based callers and haplotype-based callers. Position-based callers directly compare the aligned sequence of a tumor to the reference genome, while haplotype-based callers perform a local realignment step to identify regions of variation and use haplotype blocks to identify variants. Mutect2 which is one of the most commonly used caller is a haplotype caller. We use the somatic mutations identified in the TCGA data by Ellrot et al [257]. They have used a consensus based approach, employing multiple variant calling tools, which enables the identification of somatic variants with high confidence.

1.7.3 Antigen prediction methods

1.7.3.1 Somatic mutation annotation After identifying somatic mutations, the first step to predict potential neoantigens is filtering the mutations for altered protein sequences using annotation tools like VEP [258] and Annovar [259]. In this study, we focus on missense mutations, which result from SNVs encoding different amino acids at specific positions in the resulting protein and we have used VEP annotation to assess the effect of SNVs on protein sequence.

1.7.3.2 Neoantigen prediction The prediction of putative neoantigen is generally accurate because T cell-identified neoantigens have simpler structures, comprising short, linear peptides (9–15 amino acids). Peptides that bind to MHC class I are generally sized between 8 and 11 amino acids, whereas those that bind to MHC class II are longer at 12-25 amino acids and extend beyond the MHC groove, but have a minimum of 9 amino acids in the core [186]. However, some studies have shown that larger peptides can also bind to MHC but with lower immunogenic potential [185, 186, 260, 261]. These neoantigens are recognized by T-cell receptors (TCRs) when presented by MHC class I or class II. During neoantigen prediction, it is essential to consider both peptide-MHC complex and neoantigen-TCR complex bonds.

Several methods are employed to predict T-cell recognized neoantigens, including motif-based systems, matrices, SVM, empirical scoring, and molecular dynamics (MDs) methods [186]. The motif-based system was the pioneering method for neoantigen prediction. It involves predicting amino acid sequences that are likely to bind to the MHC groove, referred to as motifs. These sequences are then compared to data in a motif library, which contains previously determined binding peptide sequences and nonbinding MHC-binding motifs. The accuracy of this method may be limited due to the lack of known motifs for all HLA alleles [186].

Another approach of motif-based methods involves the development of machine learning algorithms (MLAs). Using MLAs, peptide-binding motifs can be determined based on specific classifications, such as positive values for peptide binders and negative values for nonpeptide binders. MLAs can also handle multiple classifications simultaneously. Among MLAs, artificial neural networks are widely utilized

Tool	Methods	Cite
NetMHCpan	Based on the binding propensity of peptides to different HLA alleles using artificial neural networks	[264, 265]
EpiMatrix	MHC class I and II protein binding efficiency based	[266]
IEDB	Proteasomal processing, TAP transport, and MHC class I and II binding based	[267]
NetChop	Immunoproteasome cleavage site based	[268]
NetCTL	Combination of proteasome, TAP transport and MHC subtype binding values	[157]
nHLAPred	Hybrid approach of artificial neural networks and quantitative matrices.	[269]
MHCPred	Binding value of MHC/peptide or TAP/peptideIC50	[270]
MMBPred	By determining of high-affinity MHC binding peptide that undergoes mutations	[271]
ProPred-1	Peptide binding efficiency with MHC I	[272]
SYFPEITHI	Motif binding to MHC class I and II	[273]
TAPPred	Binding affinity with TAP protein	[274]
RANKPEP	MHC I and MHC II binders using position specific scoring matrices (PSSMs)	[275]
Epijen	The immunoproteasome cleavage site and TAP binding affinity	[276]
DeepNeo	Immunogenic peptides with T-cell reactivity	[277]

Table 1: Antigen Prediction Tools

for determining motifs for peptide presentation to MHCs [181, 262]. One important resource for neoantigen prediction is the Immune Epitope Database (IEDB). IEDB offers valuable tools for predicting epitopes that B cells and T cells recognize, along with analyzing epitope characteristics to enhance prediction reliability. Researchers frequently utilize this database and its associated tools for studying epitopes in vaccine development, finding its user-friendly nature advantageous [262, 263]. However, it is essential to acknowledge that in silico studies, relying on computational approaches, have their limitations and are not 100% accurate.

In addition to motif-based systems, T cell neoantigen prediction can be achieved through molecular dynamics simulations (MDs), which calculate free binding energy for a molecular system. MDs offer insights into the individual or collective movement of atoms within a molecular system, providing a dynamic perspective. Unlike data-based methods, MDs rely on de novo predictions of all parameters that constitute the receptor-ligand complex structure, making them advantageous [260]. The tools available for predicting T cell-recognized neoantigens are summarized below in Table 1.

1.8 Immunotherapies

Immunotherapies have emerged as a revolutionary approach in the treatment of cancer, enabling the body's immune system to target and eliminate cancer cells. These

therapies have shown promising effects in various tumor types and have significantly improved the survival of patients with advanced malignancies [214]. Immunotherapies encompass a range of approaches, including immune checkpoint blockade (ICB) and cytokine-based therapies [278]. Cytokine therapies, like IL-2 and IL-15, have demonstrated immunomodulatory effects, contributing to enhanced antitumor responses [279]. There is a growing interest in ICB therapies [175, 214, 215, 280], especially in exploring the association between the neoantigen load and immunotherapies response [89, 130, 281].

The immune system employs checkpoint proteins to maintain a balanced immune response and prevent excessive damage to healthy cells. However, cancer cells can exploit these checkpoints to evade immune detection and destruction. Immune checkpoint inhibitors, such as PD-1 (programmed cell death protein 1) and CTLA-4 inhibitors, work by blocking the interactions between checkpoint proteins and their ligands, effectively releasing the brakes on the immune response. Immune checkpoint inhibitors, such as anti-PD-1, anti-PD-L1, and anti-CTLA-4 agents, have shown remarkable success by targeting regulatory mechanisms that suppress immune responses [174, 175, 214, 215, 278, 280].

For instance, PD-1 inhibitors prevent the binding of PD-1 receptors of immune cells CD8+ T cells with PD-L1 (programmed death-ligand 1) or PD-L2 (programmed death-ligand 2) of cancer cells [282]. This inhibition prevents the "off"signal that would normally restrain T cells from attacking the cancer cells, thereby allowing the immune system to target and eliminate the cancer [214, 283, 284]. Similarly, CTLA-4 immunotherapies work by blocking the inhibitory signals of CTLA-4, thereby unleashing the immune system to mount a more robust and effective attack against cancer. Monoclonal antibodies that target CTLA-4, such as ipilimumab, are administered to patients. By binding to CTLA-4 and preventing its interaction with CD80 and CD86, these antibodies allow T cells to maintain their activity and enhance their ability to recognize and eliminate cancer cells .

One important aspect of immunotherapy research is identification of biomarkers. The efficacy of immune checkpoint inhibitor (ICI) treatments varies, with only about 20-30% of patients responding positively, and responses varying among cancer types [285]. To optimize immunotherapy outcomes and manage costs, research seeks to

identify biomarkers that can predict which patients are more likely to respond well to immunotherapy. PD-L1 expression and tumor mutational burden are established biomarkers with clinical utility for predicting immunotherapy response [286]. Moreover, microsatellite instability, DNA mismatch repair, and other genomic biomarkers have been identified as predictive indicators [287]. Recent research has focused on liquid biopsy-based biomarkers for noninvasive prediction and in-treatment monitoring of immunotherapy response. Despite advancements, there remains variability in clinical response, highlighting the need for further exploration and identification of robust biomarkers.

Tumor Mutational Burden (TMB) is the number of nonsynonymous mutations per million bases (Mb) above a threshold frequency, usually 0.05 in a sample [288, 289]. TMB is associated with neoantigen load and has been proposed as a potential biomarker for ICI response [290]. Higher mutation loads in a sample indicate more potential immune response-inducing mutations. High TMB is linked to better responses to immunotherapy [290] and has been approved by the FDA as a biomarker for response to pembrolizumab [286]. Studies on predictive potential of TMB for ICI response have yielded mixed results across various cancer types [291], emphasizing the need for tumor-specific analysis.

Besides overall mutation numbers, other factors influence immunotherapy response prediction. Clonal load, representing the total mutations present in all cancer cells, affects therapy response and relapse likelihood [99, 292]. The expression level of genes containing neoantigens is crucial, as only expressed antigens can be presented to the immune system [293]. Neoantigen binding affinity to HLA alleles determines which neoantigens will be presented on the cell surface [293–297]. TMB estimates focus on SNVs, but indels can also create highly distinct neoantigens [298]. Additionally, tumor purity and intra-tumor heterogeneity impact TMB estimates and immune checkpoint blockade responses [293]. Considering all these factors is essential for developing effective biomarkers for immunotherapy.

1.9 Thesis overview and objectives

There is a growing interest in understanding how the immune system shapes the mutational landscape in cancer, as it has important implications in designing im-

munotherapies. It has been reported that antigen presentation plays a vital role in defining tumor immunogenicity [206]. In this thesis, we explore this further and assess the role of MHC-I in shaping the mutational landscape of cancer.

In two previous studies [227, 228], it was reported that common driver mutations in cancer are common due to the inability of common HLA alleles to present them to the immune system. However, the same pattern did not apply to passenger mutations. In Chapter 2, we hypothesised that the absence of a connection between clonal passenger mutations and HLA genotype might indicate other immune evasion mechanisms. This led us to re-evaluate the results reported by [227, 228].

In Chapter 3, we estimated the intrinsic immunogenicities for mutational signatures observed in cancer using most common HLA supertypes. It has been reported that there is a relation between HLA-B44 supertype and a mutational signature observed in melanoma patients, which leads to improved responses to immunotherapy [130]. Motivated by these results we performed an exhaustive characterization of the relationship between mutation signatures and common HLA supertypes. We used the activity of mutation signature and their immunogenicity to estimate the immunogenicity of a tumor type and samples.

Studies have reported that to estimate the magnitude of immunoediting in cancer accurately, it is crucial to consider the underlying mutational processes, which are characterized by mutational signatures [230, 235]. In Chapter 4, we investigated the extent to which the number of immunogenic mutations in a tumour sample can be predicted from mutation signature activities and the HLA genotype of the patient. We built on this approach to assess the evidence for MHC-I-mediated immunoediting. We also used simulations to estimate an upper bound on this immunoediting signal.

The research questions of this thesis can be summarised as follows:

1. Is downregulation of genes carrying immunogenic passenger mutations a potential immune evasion mechanism?

2. Do driver mutations occur in the gaps of MHC genotype?

3. How different are mutation signatures from each other in terms of immunogenicity?

4. Can we use the immunogenicity of mutation signatures and their median activity in cancer types to predict the immunogenicity of different cancer types?

5. If neoantigen depletion signal exist, what is the upper-bound of this signal?

2 No evidence that HLA genotype influences the driver mutations that occur in cancer patients

The results presented in this chapter motivated Kherreh N, Cleary S, Seoighe C. No evidence that HLA genotype influences the driver mutations that occur in cancer patients. Cancer Immunol Immunother. 2022 Apr;71(4):819-827. doi: 10.1007/s00262-021-03028-w. Epub 2021 Aug 21. PMID: 34417841; PMCID: PMC8921139. All results presented here are parts of the publication except for the expression analysis of genes carrying passenger mutations. I performed all data analysis except for gene expression analysis which was carried out by Siobhan Cleary and susceptibility to cancer based on HLA alleles which was carried out by Cathal Seoighe initially and then reproduced by me.

2.1 Abstract

The major histocompatibility (MHC) molecules are capable of presenting neoantigens resulting from somatic mutations on cell surfaces, potentially directing immune responses against cancer. This led to the hypothesis that cancer driver mutations may occur in gaps in the capacity to present neoantigens that are dependent on MHC genotype. If this is correct, it has important implications for understanding oncogenesis and may help to predict driver mutations based on genotype data. In support of this hypothesis, it has been reported that driver mutations that occur frequently tend to be poorly presented by common MHC alleles and that the capacity of a patient's MHC alleles to present the resulting neoantigens is predictive of the driver mutations that are observed in their tumor. Here we show that these reports of a strong relationship between driver mutation occurrence and patient MHC alleles are a consequence of unjustified statistical assumptions. Our reanalysis of the data provides no evidence of an effect of MHC genotype on the oncogenic mutation landscape.

2.2 Introduction

The adaptive immune system plays a crucial role in protecting the body against various pathogens and abnormal cells, including cancer cells [148]. This system

relies on specialized cells and processes to recognize and eliminate threats to the host organism. One essential component of the adaptive immune response is the major histocompatibility complex (MHC) molecules, which are responsible for presenting antigens to T cells and coordinating the immune response [299].

MHC molecules are a diverse group of cell surface proteins found in almost all nucleated cells of the body. They can be classified into two major classes: MHC class I and MHC class II [299]. MHC class I molecules are expressed on virtually all nucleated cells and play a crucial role in presenting antigens derived from intracellular pathogens, such as viruses, to cytotoxic CD8+ T cells. These antigens are generated within the cell through protein synthesis and subsequent processing. MHC class I molecules capture these antigenic peptides and present them on the cell surface, allowing CD8+ T cells to recognize and eliminate infected or cancerous cells [148, 299]. On the other hand, MHC class II molecules are primarily expressed on antigen-presenting cells, including macrophages, dendritic cells, and B cells. These molecules are responsible for presenting antigens derived from extracellular pathogens, such as bacteria and parasites, to CD4+ T cells. Antigen presentation by MHC class II molecules triggers the activation of CD4+ T cells, leading to the generation of an immune response against the pathogen [300].

The immune system can mount a response against cancer cells through a process called cancer immunosurveillance. This process involves the recognition of tumor-specific antigens presented by MHC molecules to T cells, triggering an immune response against the cancer cells [151, 152]. The recognition of tumor antigens by T cells is facilitated by the interaction between the T-cell receptor (TCR) and the peptide-MHC complex on the surface of cancer cells. This recognition leads to the activation of T cells and the subsequent elimination of cancer cells through various mechanisms, including the release of cytotoxic molecules and the induction of apoptosis [174, 179].

Cancer cells often develop mechanisms to evade immune recognition and destruction, allowing them to proliferate unchecked. One of those mechanisms is known as Immunoediting [177, 231, 232, 301, 302] that describes the interaction between the immune system and cancer cells, involving three distinct phases elimination, equilibrium, and escape [233]. It explains the dynamic relationship between cancer cells and

the immune system. This interaction between the cancer and immune cells is like a tug of war, as they influence behavior and survival of each other [303, 304]. In the first phase, elimination, the immune system recognizes and eliminates transformed cells that have acquired cancerous characteristics. Immune cells, such as cytotoxic T cells and natural killer cells, target and destroy these cancer cells [233, 304, 305]. This phase represents the initial response of the immune system to eliminate cancerous cells, acting as a form of immunosurveillance [233, 304, 305]. However, some tumor cells may escape elimination and progress to the next phase. During the equilibrium phase, the remaining tumor cells coexist with the immune system in a state of balance. Immune cells recognize and control the growth of these cancer cells, preventing their expansion. This phase is characterized by a dynamic interplay between the efforts of immune systems to suppress tumor growth and the ability of tumor cells to evade immune responses. The equilibrium phase can last for an extended period, during which time the immune system exerts selective pressure on the tumor, leading to the emergence of more aggressive and immunoresistant cancer cell variants [233, 304, 305]. In the final phase, escape, the tumor cells acquire the ability to evade immune recognition and elimination [233, 304, 305]. These cells develop various mechanisms to avoid immune detection, such as downregulating the expression of antigens that are recognized by immune cells or hijacking immune checkpoint pathways to suppress immune responses [306].

As a result, the tumor cells can grow and progress without effective immune control, leading to disease progression and metastasis. This tug of war between cancer cells and immune cells in immunoediting is a dynamic process influenced by multiple factors, including the genetic and phenotypic heterogeneity of cancer cells, the plasticity of immune cell populations, and the complex interplay of immunosuppressive and immune-activating signals within the tumor microenvironment [307]. Understanding the phases of immunoediting and the intricate balance between cancer cells and the immune system is crucial for developing effective cancer immunotherapies and personalized treatment strategies that can harness and enhance the immune response against tumors [302].

Cancer cells employ various mechanisms to evade the immune response, and one such mechanism involves acquiring mutations that alter antigen presentation.

Antigen presentation is a crucial step in the immune response, where antigenic peptides bind to major histocompatibility complex (MHC) molecules for recognition by immune cells. Several studies have highlighted the significance of mutations in genes associated with antigen presentation, such as the HLA genes or the B2M gene, which affects the formation of MHC class I molecules [229, 293, 308–310]. Loss or mutation of these genes has been found to correlate with increased tumor mutation burden, implying their role in immune evasion [229]. Furthermore, the absence of neoantigens capable of triggering an immune response can also contribute to cancer cells evading immune surveillance [158, 309]. Studies have reported selection against immunogenic somatic mutations in cancer, suggesting that cancers may actively deplete mutations that give rise to neoantigens [227, 228, 311]. It is worth noting that recent research has raised questions regarding the evidence for depletion of neoantigens, indicating the need for further investigation in this area [230, 234] .

Driver mutations are genetic alterations that occur in DNA of a cell, and they provide a growth advantage, leading to the clonal expansion of cells with that mutation. They are called "driver" mutations because they are responsible for driving the development and progression of cancer [68]. Driver mutations can occur in oncogenes, which are genes that promote cell growth and division, or in tumor suppressor genes, which are genes that normally inhibit cell growth and division [312]. Driver mutations are different from "passenger" mutations, which are genetic alterations that do not provide a growth advantage and are simply carried along as the cancer cells divide and grow. While passenger mutations may contribute to the overall genetic diversity of a tumor, they do not drive its growth [69]. Studies have reported that driver mutations identified in cancer patients tend to occur in regions where the MHC genotype of patient fails to present them to the immune system, hence shaping the driver mutations landscape in cancer [227, 228] .

However, it was reported that this negative selection pressure was not acting on immunogenic passenger mutations [228]. We hypothesize that it is because these passenger mutations occur in lowly expressed genes, thus not recognized by the immune system. We explore the possibility that downregulation of genes carrying immunogenic passenger mutations is another immune evasion mechanism used by cancer. When controlling for gene length and sequence context, we found no evidence

of immune evasion by these mechanisms. This led us to reanalyze the data from two high-profile studies [227, 228] that reported that the driver mutations that are found in cancer patients can be predicted from the capacity of the patient's MHC molecules to bind the resulting neoantigens. The patient harmonic mean best rank (PHBR) score was proposed in [227, 228] as a measure of whether a neoantigen resulting from a somatic mutation can be bound by MHC molecules, given the HLA genotype of a patient. The score is derived from predicted binding affinities of the patient's MHC molecules for the peptides spanning the mutation. The conclusions of both studies are based on an analysis of 1018 cancer driver mutations in patients from the cancer genome atlas (TCGA). The focus of the 2017 study is on MHC class I alleles, and the primary focus of the 2018 study is on presentation of cancer neoantigens by MHC class II molecules. The data for both comprised a binary matrix of mutation occurrences (indicating whether the driver mutation in each column has been observed in the patient in each row) and a matrix of PHBR scores corresponding to 9176 and 5942 patients for MHC class I and class II alleles, respectively. We reanalyzed these data and found that the conclusion of both papers that cancer driver mutations emerge preferentially in gaps in the patient's capacity to present neoantigens on MHC molecules is not robust. We found that there is no evidence from the data that the driver mutations seen in a patient are influenced by the patient's MHC class I or class II genotypes.

2.3 Results

If cancer evades the immune system by the downregulation of genes harboring immunogenic mutations (PHBR < 2) we would expect that the expression of these immunogenic genes would be lower than the expression of genes harboring non-immunogenic mutations (PHBR ≥ 2). To test our hypothesis, we compared the expression of genes carrying immunogenic mutations with the genes carrying non-immunogenic mutations. We used synonymous mutations as a proxy for neutrality since they do not alter the amino acid composition of the mutated peptide and thus, have relatively less selection pressure acting upon them (Figure 2.1).

We hypothesized that if immunogenic mutations occur preferentially in lowly expressed genes, there should be a higher proportion of genes with an immuno-

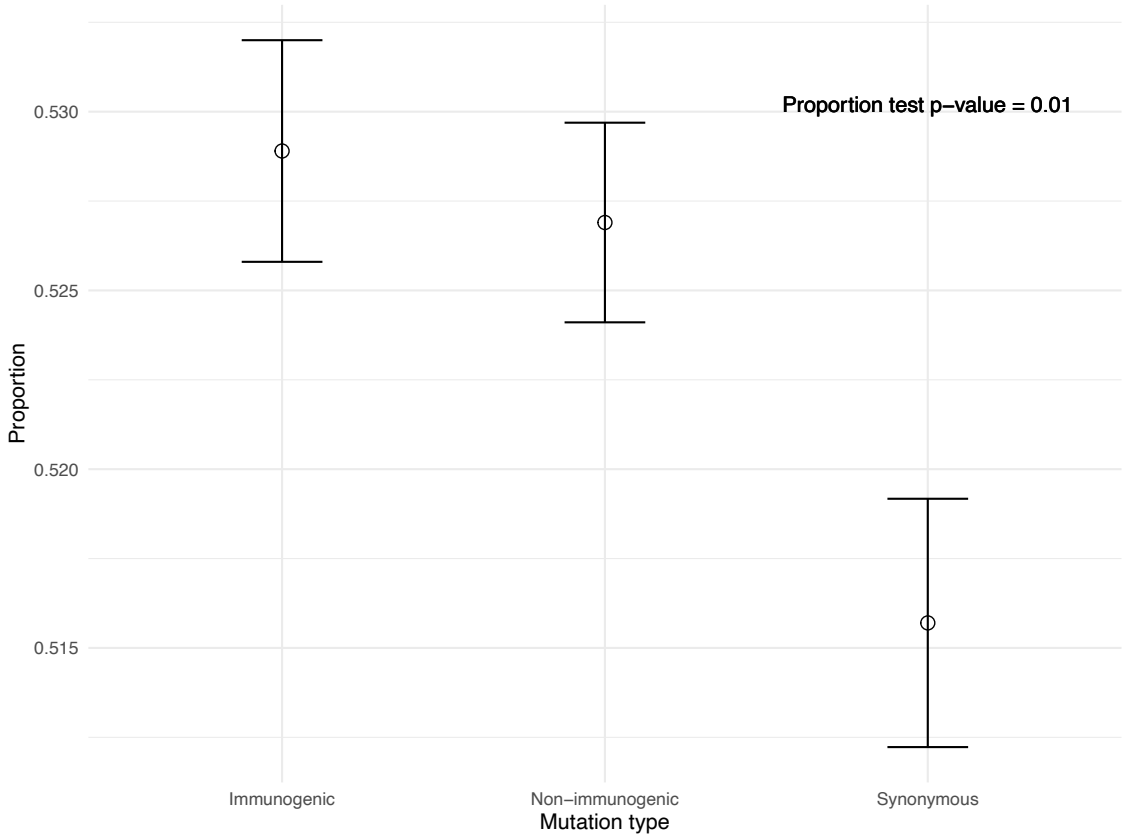

Figure 2.1: Points in this plot represent the proportions of genes containing mutations of the type shown on the x-axis for which the gene expression value is lower than the median expression of all genes in the corresponding TCGA sample.

genic mutation with expression values lower than the median expression of all genes within that sample (Figure 2.1). We compared the proportion of genes with immunogenic mutations with expression lower than the median expression of all genes within the sample to the same proportion for non-immunogenic and synonymous mutations (Figure 2.1). Our analysis revealed a significant difference (proportion test p-value = 0.01) between the groups, with immunogenic mutations tending to occur in genes with lower expression levels than non-immunogenic or synonymous mutations. However, the difference was small with proportions of mutations on genes expressed below the median level of 0.5289, 0.5269, and 0.5157 for genes with immunogenic, non-immunogenic, and synonymous mutations, respectively. When we controlled for the gene length and sequence context of mutations (See figures 6.1 - 6.4 in Appendix), we found no evidence that the immunogenic mutations are preferentially occurring in lowly expressed as an immune evasion mechanism. This

led us to conclude that the observed tendency for immunogenic mutations to occur preferentially in lowly expressed genes is not caused by the immunoediting.

2.3.1 Relationship between immunogenicity and driver mutation occurrence across patients

Due to our inability to confirm our hypothesis originating from the Marty *et al.* papers [227, 228], we conducted a re-analysis of their data. Using the predicted immunogenicities of driver mutations derived by [227, 228] we re-investigated the relationship between immunogenicity and driver mutation occurrence across patients. In both [227, 228], the predicted capacity of the MHC to present cancer driver mutations was compared between patients with and without the mutation. Higher values of the PHBR score (corresponding to low predicted capacity to bind neoantigens resulting from the mutation) in the patients in which the driver mutations occur were presented as evidence that driver mutations preferentially arise in patients who lack the MHC alleles that are capable of presenting them to T cells. In these comparisons of groups of PHBR scores, one group consists of the scores of driver mutations in patients in which the mutation is present (the Mutation group) and the other group (the No Mutation group) consists of PHBR scores of the driver mutations in the patients without the mutation. A given driver mutation can appear many times in the Mutation group in these comparisons—once for each patient in which it occurs. This is problematic, because the PHBR scores of mutations are highly correlated (Figure 2.2A, 2.4D) and, thus, the data points are not independent. For example, a driver mutation that occurs in 500 patients will contribute 500 PHBR scores to the Mutation group and N – 500 scores to the No Mutation group, where N is the total number of patients. If the PHBR score of the mutation is generally high or generally low across patients, it will clearly have a disproportionate impact on the distribution of PHBR scores in the Mutation group.

The correlation in PHBR scores between patients is not solely due to sharing of HLA alleles. Even the PHBR scored using HLA alleles from different allele groups is significantly correlated (Figure 2.4D), but the scores of driver mutations were effectively treated as independent observations by the studies that reported an effect of HLA alleles on driver mutations. Marty Pyke *et al.* [227] used a statistical

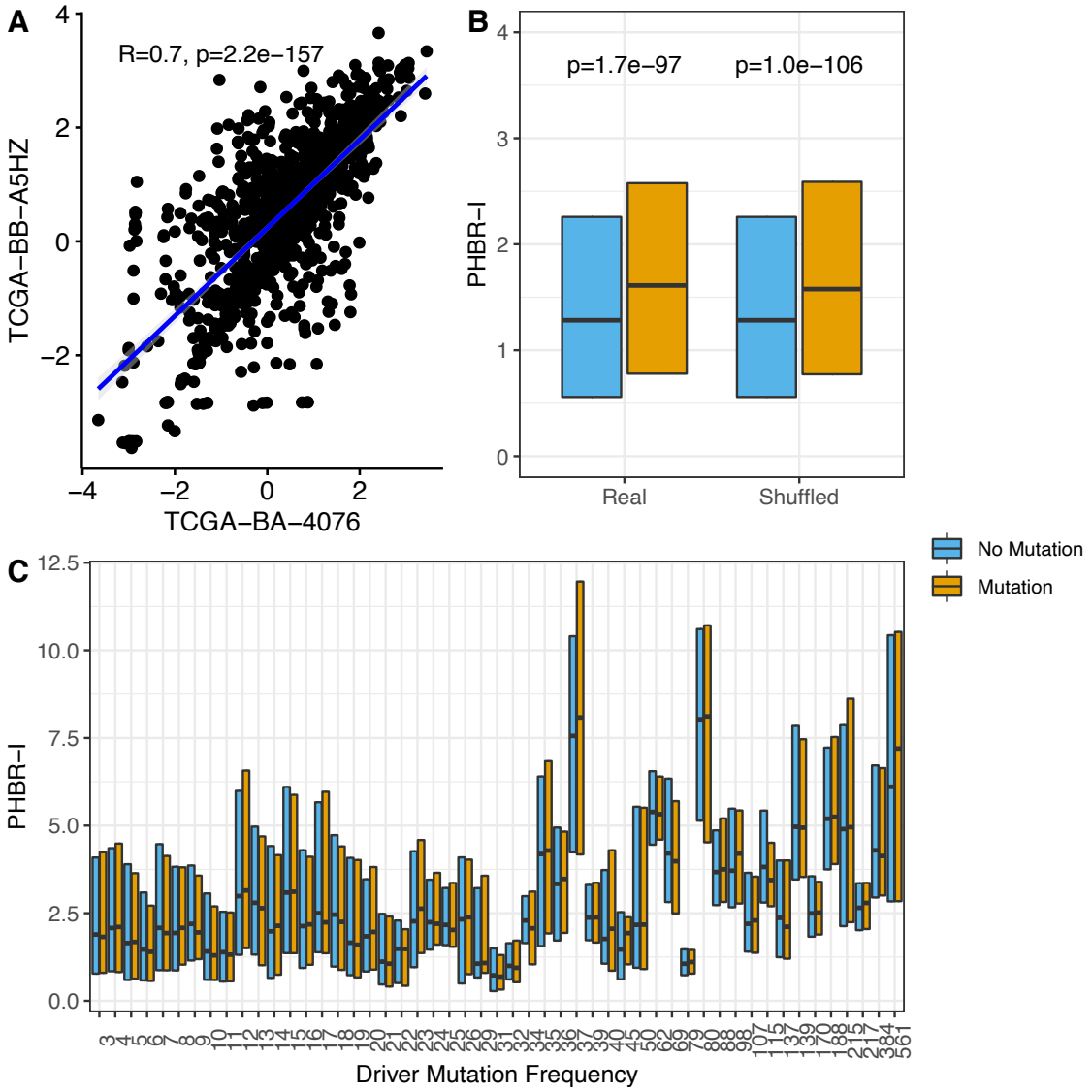

Figure 2.2: (A) Scatterplot of log PHBR-I scores of all driver mutations, calculated using the HLA genotypes of two randomly selected patients from TCGA. (B) Median and interquartile range of PHBR-I score in the No Mutation (blue) and Mutation (orange) groups for the real data and for data in which the MHC genotypes have been randomized between patients. (C) Median and interquartile range of PHBR-I scores in the No Mutation (blue) and Mutation (orange) groups in bins of mutation recurrence. The number of observations corresponding to each bin is provided in Table 2.

test (the Mann–Whitney U test) to compare the median PHBR-II score between the Mutation and No Mutation groups and reported a higher median score in the Mutation group with a p-value $< 2.2 \times 10^{-16}$. This was interpreted as evidence that

the HLA genotype of patient influences the driver mutations that occur in cancer patients. However, the fundamental assumption of the test is that the observations in each group are independent, and this assumption is clearly violated. We found that the differences between the Mutation and No Mutation groups are, in fact, just as large when the MHC genotypes are randomized between patients, indicating that this difference is not driven by patient genotype (Figure 2.2A). Moreover, when we compared PHBR scores, grouped by driver mutation frequency (so that each driver mutation contributes the same number of observations to the Mutation group in each comparison), we saw no consistent differences (Figure 2.2C).

In 100 randomizations of the HLA class I genotypes, the median PHBR-I scores of the Mutation group in the randomized data in fact exceeded the median of the Mutation group in the real data 94 times (the difference was not statistically significant; p-value = 0.12 for the two-sided randomization-based test for a difference in PHBR-I scores between the groups). Similarly, when we shuffled the HLA class II genotypes, the median PHBR-II score of the Mutation group in the shuffled data exceeded that of the real data 36 times; again, there was no significant difference in median score between groups (p-value = 0.72). Thus, comparison of PHBR scores between the Mutation and No Mutation group does not provide any support for the hypothesis that driver mutations occur preferentially in patients with MHC molecules that are not capable of binding the resulting neoantigens. In [227, 228], PHBR scores of driver mutation occurrences were also compared against scores of occurrences for different mutation classes (e.g., germline mutations and passenger mutations). Because they contribute many times to the Mutation group, the existence of a small number of highly recurrent cancer driver mutations with high PHBR scores (i.e., low binding affinity) may be sufficient to skew all of these comparisons. This problem is compounded by the fact that the 1,018 driver mutations that are the basis of this study occur on just 168 different genes and PHBR scores are statistically significantly correlated between mutations in the same gene, particularly for class II alleles (Figure 2.4B, C). The number of distinct genes among the most highly recurrent cancer driver mutations is smaller still (Figure 2.3A).

2.3.2 Regression models relating log-PHBR score to mutation probability

In addition to comparing PHBR scores between the Mutation and No Mutation groups, [228] proposed two mixed effects logistic regression models to relate the log odds that a driver mutation is found in a patient to the log of the PHBR-I score for the mutation, given patient MHC genotype. In one model (referred to as the within-mutation model), a random effect is used to correct for differences in the frequency with which different driver mutations occur. In the other model (referred to as the within-patient model), the random effect models differences in the abundance of driver mutations between patients, but there is no correction for differences in the frequency of different driver mutations. Mathematical descriptions of both models are reproduced in the Methods. In [228], there was no significant effect of log PHBR-I on the log odds of driver mutations using the within-mutation model. Although the results of the within-mutation model are not reported in [227], log PHBR-II is not significantly associated with driver mutation occurrence with this model either. The failure of the within-mutation model to detect an effect of log PHBR-I on the probability of a driver mutation was explained in [227] as resulting from the fact that the impact of immune presentation on the probability of a mutation was captured by the random effect. In other words, the tendency for a driver mutation not to be recognized by common HLA alleles resulted in a high driver mutation frequency, and this was captured by the random effect in the model. This is not a strong argument, however, because the median PHBR score does not explain much, if any, of the variance in driver mutation frequency in cancer patients (Figure 2.3B, C). Even if the variation in driver mutation frequency was entirely driven by the MHC class I genotype, it should not fully capture the relationship between driver mutation occurrence and MHC genotype. That is, the rare driver mutations should still be found associated with the rare MHC genotypes that are not capable of presenting them and the common driver mutations should be found associated with the relatively more common MHC genotypes that cannot present them. This should be detectable with the within-mutation model, even after accounting for differences in driver mutation frequencies.

In contrast to the lack of a signal from the model that accounted for differences

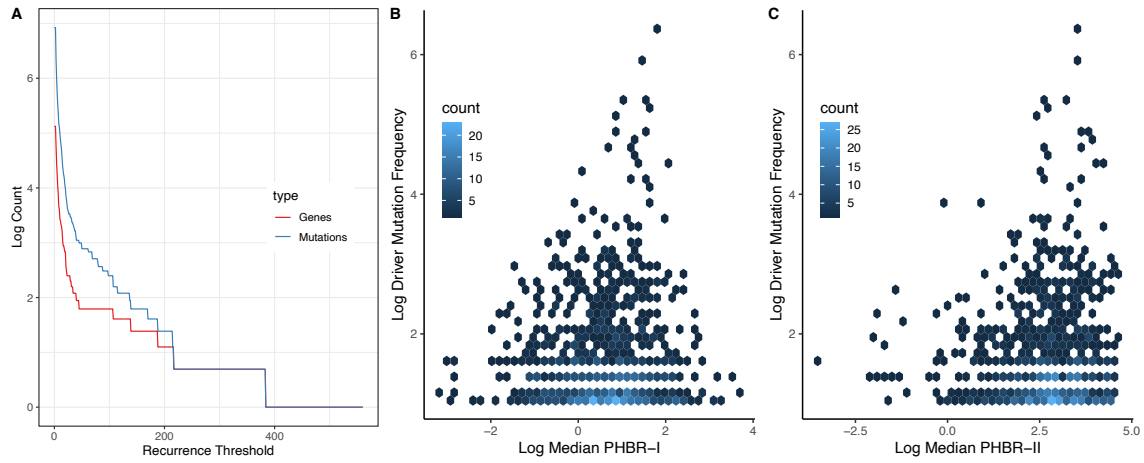

Figure 2.3: (A) The blue line shows the logarithm of the number of driver mutations that recur across patients at least as often as the recurrence threshold on the x-axis. The red line shows the logarithm of the number of distinct genes in which these mutations occur. (B), (C) Hexbin plots illustrating the relationship between the logarithm of median PHBR-I (B) and PHBR-II (C) scores and driver mutation frequency (across patients)

in frequencies between driver mutations, Marty and colleagues [228] reported a very strong effect of log PHBR-I on the log odds of driver mutations using the within-patient model (which accounts for differences in driver mutation burden between patients). Quoting a p-value of $< 2.2 \times 10^{-16}$, the authors estimate an increase of 28% in the log odds of occurrence of a mutation with each unit increase in log PHBR-I (95% CI: [25%, 31%]). However, this result is affected by the same failure to take account of the non-independence of observations of the same driver mutation that led to the spurious between group comparisons of PHBR scores discussed above. This can be seen from the fact that the results are not affected by randomization of the patient genotypes. We randomly shuffled the patient genotypes for the real data so that, for each patient, driver mutations were scored with the HLA genotypes of a randomly selected patient. We then fitted the within-patient model to the shuffled data. When we did this, we found that the increase in the log odds of a driver mutation occurrence per unit increase in log PHBR-I was 25.1% (standard error 1%), slightly higher than we obtained using the real data (we obtained an estimate of 24.7% when we implemented the within-patient model on the PHBR-I data, a little below the 28% reported by [228]). The difference between the real and shuffled

data was not statistically significant (p-value = 0.69). Similarly, the relationship between PHBR-II was just as strong using the shuffled and unshuffled data (27.0% and 26.9% increase in the log odds of mutation occurrence per unit log PHBR-II for the shuffled and unshuffled data, respectively). Again, these results provide no indication of a relationship between the patient HLA genotypes and driver mutation occurrence. We performed a simple simulation to demonstrate how the spurious results obtained with the within-patient model can come about. We simulated the case of a single driver mutation that occurs at high frequency and has a high PHBR score across patients. The remaining mutations occurred at lower frequency and had a lower PHBR score distribution (details of the simulation are provided in Methods). Because the within-patient model of [227, 228] treats PHBR scores of a given mutation as though they were independent observations (despite the strong correlation in the scores of different mutations between patients seen in Figure 2.2A), this single common driver mutation with a high PHBR score was sufficient to give a highly significant association between PHBR score and driver mutation occurrence (p-value = 2×10^{-52}). This trivial example illustrates how failure to account for the high degree of correlation in the immunogenicities of driver mutations across patients can give highly misleading results.

2.3.3 No evidence that driver mutations in cancer patients are adapted to patient MHC genotypes

Under a null model of no effect of MHC genotype on driver mutation occurrence, the probability that the patient can present a given driver mutation can be estimated from the proportion of all patients that can present that mutation. This provides a straightforward means to compare the observed to the expected total number of driver mutations with PHBR scores below the threshold for presentation. If the driver mutation landscape is shaped by patient-specific MHC binding capacity and if this is captured by PHBR scores, then the observed number of driver mutations that can be presented in the patients in which they occur should be smaller than the expected number. For MHC-I, the observed number of driver mutations with PHBR-I scores below the threshold for presentation of 2 applied in [228] was, in fact, slightly (but not statistically significantly) larger than the expected number

(3,669 compared to 3,657.5 ± 68.8; p-value = 0.73 from the cumulative distribution function of the Poisson-binomial distribution). For MHC-II, the observed number of driver mutations with PHBR-II scores below the threshold of 10 applied in [227] was slightly (and again not statistically significantly) below the expected number (1,119 compared to 1,142.3 ± 36.4; p-value = 0.21). Similar results were obtained when the thresholds that were used to define strong binding (0.5 and 2 for MHC-I and MHC-II, respectively) were applied (p-value = 0.92 and p-value = 0.71, respectively). These results provide no evidence that driver mutations occur significantly less often in patients with MHC alleles that are capable of binding them.

2.3.4 Prediction of driver mutation occurrence from MHC genotype

The study of [228] includes the claim that the PHBR scores derived from patient MHC-I genotype could be used to predict the driver mutations that are observed in cancer patients; however, this claim is never tested directly. For each driver mutation, we fitted a logistic regression model to relate the log odds of a driver mutation occurring to the patient specific log PHBR-I score. For example, the most common driver mutation in the dataset, V600E in BRAF, occurs in 561 individuals. When we fitted a logistic regression model, treating the log odds of occurrence of this mutation as the response variable and with log PHBR-I for V600E, cancer type and population of origin of the patient as predictor variables, there was no significant effect of log PHBR-I on the occurrence of this mutation (P = 0.67). It could be argued that common mutations are common because they cannot be presented by common HLA alleles (i.e., they have generally high PHBR scores across patients). While it is the case that V600E in BRAF has a high mean PHBR-I score, there were still 704 patients whose MHC-I alleles were predicted to be capable of presenting this mutation (PHBR-I < 2, the threshold used in [228] to indicate MHC class I binding). Of these patients, 5.5% actually carried the V600E mutation in BRAF, almost identical to the frequency of the mutation in the patients with PHBR-I ≥ 2 (6.2%; p-value = 0.57 from Fisher's exact test). We fitted logistic regression models for each driver mutation and found that no driver mutation was significantly predicted by log PHBR-I, after correction for multiple testing (minimum p-value = 0.003; adjusted p-value = 1, using the Holm method). We repeated this procedure using PHBR-II

scores and again found no significant association with driver mutation occurrence following correction for multiple testing (minimum p-value = 0.004; adjusted p-value = 1). There is, therefore, no evidence that patient HLA alleles are predictive of the driver mutations that occur in the patient.

2.3.5 The association between driver mutation frequency and PHBR scores

The strong associations previously reported between driver mutations and immune presentation scores could be explained by a small number of driver mutations with high frequencies that have high PHBR scores (and therefore are not well presented by HLA alleles). The authors in [228] implies that the high frequency of some driver mutations is caused by the fact that these mutations are not well presented by common HLA alleles, thus enabling them to occur in many individuals. This is illustrated by a significant correlation between the frequency of driver mutation occurrence (within bins of driver mutation frequency) and median PHBR-I scores in the bin (this relationship can be seen in the upward trend of the median values from left to right in Figure 2.2C). Although 1018 driver mutations were included in the studies of [227, 228], they are associated with just 168 different genes. Based on an analysis of 1000 randomly sampled pairs of germline mutations from the same genes, we found that the PHBR scores of mutations in the same gene are positively correlated (Figure 2.4B, C), likely reflecting amino acid or domain content of the proteins. For example, peptides of proteins with a large proportion of hydrophobic residues may be more likely to be presented on MHC molecules [230, 313, 314]. The driver mutations with the highest frequencies across patients are dominated by a relatively small number of genes (Figure 2.3A). If a subset of these genes tend to have relatively high PHBR scores this could induce a correlation between driver mutation frequency across patients and median PHBR score. Indeed, when we restricted to only the highest frequency driver mutation for each driver gene, the relationship between PHBR-I score and driver mutation frequency was no longer significant (Spearman $\rho = 0.24$; p-value = 0.28). Thus, the reported association between driver mutation frequency and median PHBR-I score is not robust.

.

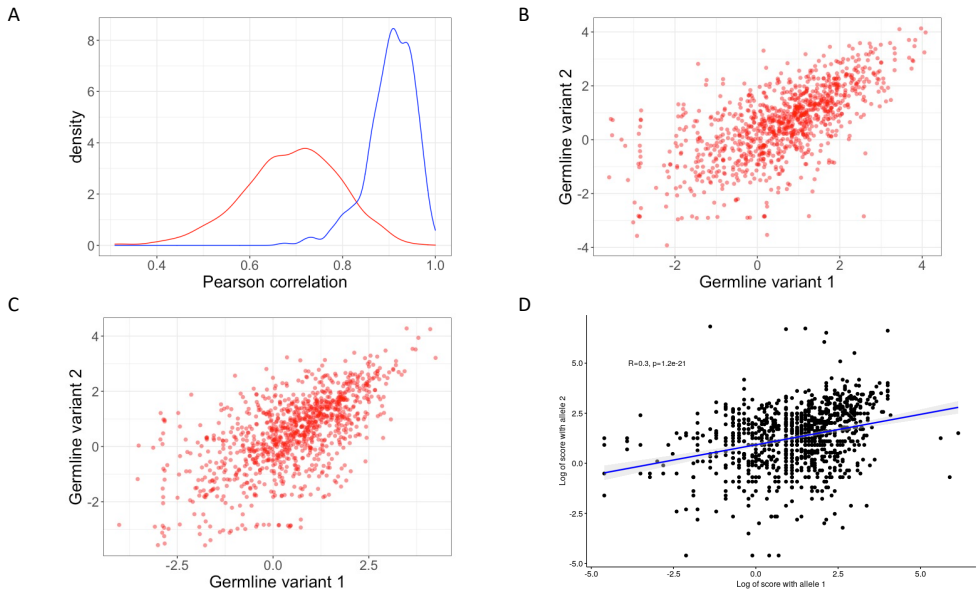

Figure 2.4: (A) Distribution of pairwise correlations between PHBR-I (red) and PHBR-II (blue) scores of 1,000 randomly sampled pairs of patients. (B) Scatterplot of log PHBR-I scores for pairs of germline mutations from the same genes.(C) Scatterplot of log PHBR-II scores for pairs of germline mutations from the same genes. (D) Correlations of PHBR-I scores from different HLA alleles.

2.3.6 No evidence that driver mutation coverage predicts cancer risk

If the frequency of driver mutations across cancer patients was determined to a large extent by the binding affinities of common HLA alleles, we would expect the number of recurrent cancer driver mutations that can be bound by a patient's MHC molecules to be associated with cancer risk. In [227], the driver mutation coverage is defined as the number of driver mutations that can be presented by the patient's MHC molecules. This can be calculated for MHC-I (for which a threshold of PHBR-I < 2 was used to indicate binding) and for MHC-II (for which the threshold was PHBR-II < 10). MHC-I (but not MHC-II) coverage was found to be correlated with age of diagnosis for TCGA patients [227, 228]. Interestingly, the strongest correlations between PHBRI coverage and age at diagnosis are for cervical and liver cancers, two cancers that are strongly associated with viral infections [315–317], suggesting that the relationship between coverage and age at diagnosis may reflect HLA-dependent differences in susceptibility to these viral infections. To test, more

generally, whether there is any relationship between PHBR-I coverage and cancer risk we fitted a logistic regression model to the log odds of cancer status (a binary variable to indicate whether the individual has self-reported a diagnosis of cancer of any type) to PHBR-I coverage for 377,790 participants from the UK Biobank. Treating age and sex as covariates, we found no significant association between PHBR-I coverage and cancer risk (p-value = 0.15). The lack of an association between cancer risk and driver mutation coverage does not support a model in which cancer driver mutations occur in gaps in the capacity of the individual's MHC molecules to bind the associated neoantigens.

2.4 Discussion

The relationship between MHC genotype and the driver mutations that are found in cancer patients, reported by [227, 228], is unchanged when the MHC genotypes of patients are shuffled. This includes the effect of log PHBR score on the occurrence of a driver mutation, as inferred from the within patient model, as well as the difference in median PHBR scores between the Mutation and No Mutation groups. It is therefore clear that any effect of PHBR scores on the driver mutation landscape is not dependent on individual level MHC genotypes.

It is still conceivable that MHC genotype affects the driver mutation landscape at the population level, such that poorly presented driver mutations are relatively common; however, it is implausible that the population level effect could arise in the absence of any association between PHBR score and driver mutation occurrence within individual patients. If immune responses cause driver mutations that can be recognized by common MHC alleles to be rare, we would expect these driver mutations to be more frequent among individuals with MHC alleles that are incapable of presenting them. No such effect of MHC genotype on driver mutation occurrence within individuals was apparent from the data.

Furthermore, the relationship that was reported between driver mutation frequency and median PHBR score might be accomplished without affecting cancer cell proliferation, the requirement of the cancer cells for continued expression of genes carrying driver mutations may prevent downregulation of these genes. One difficulty in attempting to reconcile these findings in this way is that the effect of

MHC genotype on the driver mutation landscape was reported for both oncogenes and tumor suppressor genes (and was stronger for the latter group in [227]). It is not clear that the requirement for expression of the gene that carries the driver mutation should apply to driver mutations in tumor suppressor genes, where loss of function is the expected mode of action.

Our reanalysis of cancer driver mutations from the TCGA indicates that there is no evidence that selection exerted by the immune response influences the driver mutations observed in cancer. This result complements the recently reported lack of overall depletion of neoantigens among somatic mutations observed in cancer [230, 234]. It remains possible, however, that the capacity of the MHC to present neoantigens at the cell surface does have an appreciable influence on the driver mutations observed in cancer, but that this capacity is not sufficiently well captured by the PHBR score. Given the experimental evidence for the capacity of PHBR-I and PHBR-II scores to predict MHC-I and MHC-II binding affinity [227, 228], this seems unlikely. Alternatively, it is possible that the availability of immunogenic non-synonymous mutations is not what limits the capacity of the immune response to prevent cancer development. The wide range of mutation burdens in human cancers [318] and the relationship between mutation burden and the efficacy of immune checkpoint inhibitors [88, 319, 320] argue against this suggestion, unless the immune response to the developing cancer is distinct to the response following immune checkpoint inhibitor therapy. The lack of a relationship between MHC genotype and driver mutation content suggests that if the immune system plays a major role in cancer prevention, this does not involve the prevention of specific driver mutations in a way that depends strongly on MHC genotype. of the mutation is weak and no longer significant when we restricted to a single driver mutation per driver gene. This restriction is necessary, given the correlation we observed between PHBR scores derived from the same gene, even for germline mutations.

If, as [228] suggests, cancer arises in gaps in an individual's capacity to present driver mutations, then we would expect the number of such gaps that an individual has for cancer driver mutations to be a strong risk factor for cancer development. Indeed, [228] reports effect MHC-I driver mutation coverage on age at cancer diagnosis, where coverage was defined as the number of driver mutations in the study that

were predicted to be bound by the patient's MHC class I molecules. We tested this using data from the UK Biobank. Given the size of the data set (377,790 individuals, including 32,802 with a self-reported cancer diagnosis) even a weak relationship between MHC-I coverage and cancer risk should be detectable; however, we found no significant effect of coverage on cancer status when we fitted a logistic regression model that included sex and age as covariates. If the reported effect of MHC genotype on driver mutation landscape was robust, this would be an important negative result, as it addresses the proposal by [228] that PHBR-I scores of driver mutations may prove useful for assessing risk of development of certain cancers. This negative result has not previously been reported to the best of our knowledge.

The reported depletion of cancer neoantigens [158, 309, 321] applies to all non-synonymous immunogenic mutations and not specifically to driver mutations. However, [227, 228] reported no evidence of an influence of patient MHC on passenger mutations. This finding is surprising, given that both driver and passenger mutations (particularly clonal, non-synonymous, immunogenic passenger mutations) should have the capacity to elicit immune responses. In principle, this could be explained by downregulation of genes carrying immunogenic mutations. Indeed, a recent study [322] suggested that the extent of depletion of neoantigens depends on the expression level of the gene. While for neoantigens resulting from passenger mutations, this downregulation might be accomplished without affecting cancer cell proliferation, the requirement of the cancer cells for continued expression of genes carrying driver mutations may prevent downregulation of these genes. But when we investigated this possibility, we found there was lack of evidence of downregulation of genes carrying immunogenic passenger mutations. Another difficulty in attempting to reconcile these findings in this way is that the effect of MHC genotype on the driver mutation landscape was reported for both oncogenes and tumor suppressor genes (and was stronger for the latter group in [227]). It is not clear that the requirement for expression of the gene that carries the driver mutation should apply to driver mutations in tumor suppressor genes, where loss of function is the expected mode of action.

In conclusion, our reanalysis of cancer driver mutations from the TCGA indicates that there is no evidence that selection exerted by the immune response

influences the driver mutations observed in cancer. This result complements the recently reported lack of overall depletion of neoantigens among somatic mutations observed in cancer [230, 234]. It remains possible, however, that the capacity of the MHC to present neoantigens at the cell surface does have an appreciable influence on the driver mutations observed in cancer, but that this capacity is not sufficiently well captured by the PHBR score. Given the experimental evidence for the capacity of PHBR-I and PHBR-II scores to predict MHC-I and MHC-II binding affinity [227, 228], this seems unlikely. Alternatively, it is possible that the availability of immunogenic non-synonymous mutations is not what limits the capacity of the immune response to prevent cancer development. The wide range of mutation burdens in human cancers [318] and the relationship between mutation burden and the efficacy of immune checkpoint inhibitors [319, 320, 323] argue against this suggestion, unless the immune response to the developing cancer is distinct to the response following immune checkpoint inhibitor therapy. The lack of a relationship between MHC genotype and driver mutation content suggests that if the immune system plays a major role in cancer prevention, this does not involve the prevention of specific driver mutations in a way that depends strongly on MHC genotype.

Studies [230, 235, 324, 325] have shown that to understand how natural selection operates in cancer cells, we need to consider the types of mutations that are occurring and the processes that are driving them. It is currently unclear whether and how mutational processes, which are characterized by mutational signatures and their sequence context preferences (the specific DNA sequences where mutations are more likely to occur) affect signals of neoantigen depletion. [230] has focused on mutations arising from different mutational processes. The study used HLA affinity predictions to annotate the human genome for its translatability to HLA binding peptides. They reported that the apparent neoantigen depletion signals became negligible when considering the background mutational processes [230]. However, their annotation of the HLA binding region in human genome is very conservative and may lead to biased results. In the next chapter we estimate intrinsic immunogenicities of different mutational signatures and their impact on the immunogenicity of different cancer types.

2.5 Methods

2.5.1 Data

We performed a reanalysis of cancer driver mutations in TCGA and their predicted immunogenicities, reported in [227, 228]. Both papers calculate a score that is used to predict the extent to which neoantigens are presented on MHC-I or MHC-II molecules, given the patient genotype. The score is calculated by considering all peptides of a specific length or range of lengths that contain the mutation. A rank-based presentation score was obtained for each peptide using NetMHCpan3.0 [265], and for each of the patient's HLA alleles the best rank value was retained. The PHBR score is then the harmonic mean (across the patient's HLA alleles) of these best-rank scores (see [227, 228] for details). This score was calculated for class I MHC alleles in [228] where it was based on peptides with lengths ranging from 8 to 11 amino acids and for class II alleles in [227], where it was based on peptides of length 15 amino acids. We applied the methodology as described to the TCGA data to obtain a binary matrix of driver mutation occurrences across patients and matrices of PHBR-I and PHBR-II scores across patients for each driver mutation. In order to ensure our results were precisely comparable to the published results, we also requested the data matrices that were the basis of the original studies and these were kindly provided by the authors (following confirmation of the appropriate data access permissions).

2.5.2 Logistic regression models relating mutation occurrences to PHBR scores

Following the notation of [228], consider a mutation matrix, with entries $y_{ij} \in \{0, 1\}$, indicating the presence/absence of driver mutation j in patient i and a matrix of PHBR-I or PHBR-II scores with real-valued entries, x_{ij}, corresponding to the score of mutation$j$, given the MHC alleles of individual i. Two mixed effects logistic regression models were used in [228] to relate the log-odds of $y_{ij} = 1$ to the log of x_{ij}. The first model, referred to as the *within-mutation* model, has a normally distributed random effect,β_j, that models differences in the frequencies of different

driver mutations:

$$\text{logit}(P(y_{ij} = 1|x_{ij})) = \beta_j + \gamma \log(x_{ij}) \quad (1)$$

The second model, referred to as the *within-patient* model, uses a random effect, η_i, to model differences in the abundance of driver mutations between patients, but does not model differences in the frequencies with which different driver mutations occur:

$$\text{logit}(P(y_{ij} = 1|x_{ij})) = \eta_i + \gamma \log(x_{ij}) \quad (2)$$

2.5.3 Simulation

We designed a simple simulation scenario to illustrate how spurious results can be obtained from the within-patient model due to a failure to account for non-independence of the PHBR scores across patients (some driver mutations tend to have higher scores across patients, while others have lower scores, leading to the high degree of correlation in the scores of driver mutations between patients seen in Figure 2.1A). The simulation consisted of 100 driver mutations, one of which had a high frequency (20% of 500 patients) and a relatively high PHBR score (normally distributed across patients with mean 10 and standard deviation 2). The remaining mutations occurred at low frequency (1%) and had normally distributed PHBR scores with mean 5 and standard deviation 2. We then fitted the within-patient model to this simulated dataset.

2.5.4 Relationship between MHC-I coverage and cancer risk in UK Biobank

We retrieved HLA class I alleles from participants in the UK Biobank. These alleles were inferred using HLA*IMP:02 [326]. Only alleles that were called with imputation posterior probability greater than 0.5 and only participants with six HLA class I alleles called were retained. This left a total of 377,790 individuals. For each individual, we determined the driver mutation coverage as the number of driver mutations with PHBR-I scores < 2, given the individual's HLA genotype. We retrieved the self-reported cancer status (data field 20001) for these individuals. Treating the self-report of any cancer type as a case, we fitted a logistic regression model to case status as a function of age, sex and PHBR-I coverage.

3 Chapter 3: Variation in the predicted immunogenicity of mutation types

3.1 Abstract

The presentation of intracellular antigens on the cell surface by Major Histocompatibility Complex class-I (MHC-I) molecules is one of the major determinants for CD8+ T cell activation. Research has shown that patient MHC-I genotype influences immunotherapy responses; however, this influence appears to be inconsistent, and it is not clear why this is the case. For example, the B44 HLA supertype is associated with a better response in melanoma. Non-small cell lung cancer (NSCLC) has a similar somatic mutation burden and immunotherapy response to melanoma; however, the B44 supertype has not been found to influence immunotherapy response in NSCLC. This difference has been attributed to underlying differences in mutational processes active in melanoma compared to NSCLC.

To generalize these findings, we performed an exhaustive characterization of the predicted immunogenicity of mutations arising from all cancer mutation signatures for the common HLA supertypes. We observed that mutations resulting from some mutation signatures were more likely to be presented by specific HLA alleles than mutations from other signatures. The average number of mutations inferred to be immunogenic in a cancer type could be predicted with high accuracy ($R^2 = 0.87$) using the median activity of the mutation signatures in that cancer. Mutation signature 20 resulted in the highest proportion of immunogenic mutations, given the HLA allele frequencies in the TCGA cohort. The highest proportion of somatic mutations and immunogenic somatic mutations in the TCGA cohort was contributed by mutation signature 5. When comparing different types of cancer in the TCGA cohort, CESC had the highest expected number of immunogenic mutations, while PRAD had the highest observed proportion of immunogenic mutations. We used our method to predict expected immunogenicity in two ICB-treated cohorts and observed that in both cohorts, higher expected immunogenicity was associated with improved immunotherapy response and overall survival.

3.2 Introduction

Cancer frequently develops through an evolutionary process, in which cancer genes accumulate somatic mutations that confer a fitness advantage to the affected cells, resulting in positive selection on these somatic mutations [25]. Recent studies have reported that neoantigens resulting from somatic mutations play a crucial role in shaping the immune response to cancer [158, 311, 327, 328]. Neoantigens are immunogenic peptides, resulting from somatic mutations, that are presented on cell surface by MHC-I molecules. MHC-I molecules play a key role in initiating an immune response against cancer cells. These molecules are responsible for presenting peptide fragments derived from intracellular proteins on the surface of cells [329].

Somatic mutations arise through distinct mutational processes and the activities of these processes play a critical role in shaping how cancer develops and progresses. The distribution and characteristics of somatic mutations in cancer vary considerably and are influenced by various factors [330]. These different mutational processes are characterized by mutational signatures that have been identified using mathematical methods [106, 126, 331, 332]. Mutational signatures are specific patterns of mutations that provide insights into the mutational processes that have occurred throughout the development of the disease. By characterizing mutational signatures in the genome of a cancer patient we can gain valuable information about these underlying mechanisms. Alexandrov *et al.* have made available a curated census of signatures, known as the COSMIC mutation signatures and provide the mutational profile, proposed aetiology and tissue distribution of each signature [108]. COSMIC mutation signatures were identified using 96 triplet mutation contexts, consisting of the mutated nucleotide and the adjacent nucleotides (5' and 3'). and the adjacent nucleotides (5' and 3').

MHC-I alleles are polymorphic, meaning they have multiple variants within the population. This polymorphism results in differences in the amino acid sequences of the peptide-binding cleft of MHC-I molecules [196]. These differences contribute to variations in the binding capability and preferences of MHC-I alleles for different peptides. The polymorphic nature of MHC-I alleles enables the presentation of a diverse range of antigens and plays a crucial role in immune recognition and response. The difference in MHC-I genotype can lead to differences in neoantigen

load among patients, and neoantigen load has been reported to be associated with the clinical benefits of immunotherapy treatment in melanoma and NSCLC cohorts [319, 320, 333].

Chowell *el al.* [137] showed that germline HLA-I genotypes influence ICB responses, and heterozygosity at HLA-I loci was associated with better survival than homozygosity for one or more HLA-I genes. In two independent melanoma cohorts, patients carrying the HLA-B44 supertype exhibited extended survival, while patients with the HLA-B62 supertype or somatic loss of heterozygosity at HLA-I were associated with poor outcomes [137]. However, a recently published study reported that this association is not present in patients of European ancestry [334]. Cumming *et al.* showed that this beneficial effect of the B44 supertype in melanoma patients is caused by the mutational pattern of this cancer type [130]. Melanoma patients having mutated peptides that include radical glutamic acid substitutions in the anchor position have enhanced B44 binding [130, 137]. Therefore, mutational processes that enrich these mutations, such as sun exposure in melanoma, benefit patients with the B44 allele. In contrast, cancers with less favourable mutation patterns may lack or show an opposite association with the B44 supertype, as seen in NSCLC [130, 281].

Understanding the association between mutational signatures and MHC molecules can provide insights into the mechanisms of immune evasion by cancer cells and also inform the development of personalized cancer therapies. By considering mutational signatures and MHC molecules, we can better understand the interplay between the mutational landscape of cancer cells and the immune response and ultimately develop more effective strategies for diagnosing and treating cancer. Although previous studies [130, 137, 281] have reported a relationship between the particular background mutational processes and HLA-I supertypes and their importance in predicting ICB efficacy, the general relationship between the two factors is understudied. To generalize these findings, we set out to perform an exhaustive characterization of the predicted immunogenicity of mutations arising from different mutation processes for all the major HLA supertypes.

We used COSMIC mutation signatures [108] in our study and predicted their immunogenicity for the major HLA supertypes. HLA Supertypes are groups of HLA

alleles that bind common sets of peptides due to sharing specific residues at the anchor position [335]. We also predicted the expected immunogenicity of various cancer types based on the median activity of the mutational signatures, i.e. the median number of mutations contributed by each mutation signature in that cancer type. Further, we predict expected immunogenicity for the TCGA cohort [318] by analyzing the activity of various mutation signatures in individual samples together with HLA genotype. We validated our results by assessing the correlations of expected immunogenicity (immunogenicity estimated based on mutational signature activities) with empirical immunogenicity (immunogenicity estimated directly from the observed mutations).

3.3 Results

The extent of variation in immunogenicity of each mutation type was estimated for the most common HLA supertypes (see Table 1). We observed that trinucleotide contexts of mutations influence their immunogenicity, but the differences in the proportion of immunogenic mutations for each mutation type were smaller compared to HLA supertypes. Specifically, the analysis showed that HLA-C supertypes exhibited the highest proportion of mutations that were predicted to be immunogenic, followed by HLA-B supertypes (Figure 3.1). On the other hand, HLA-A and HLA-B alleles displayed similar proportions of immunogenic mutations. This finding aligns with previous studies that have reported similar functional diversity between HLA-A and HLA-B alleles [320] while highlighting the distinct binding pattern of HLA-C [336].

3.3.1 Mutation signatures MHC-I affinities

The relationship between MHC-I supertype allele affinities and COSMIC mutation signatures was investigated to gain insights into how mutational signatures vary from each other in terms of immunogenicities. The results demonstrated that the variation in expected immunogenicity was greater across the HLA supertypes than between the mutation signatures within a specific HLA supertype (see Figure 3.2a, b), consistent with the findings observed for different mutation types (Figure 3.1).

Two different binding thresholds were employed to determine the proportion of immunogenic mutations, as described in the Methods section. The stricter threshold

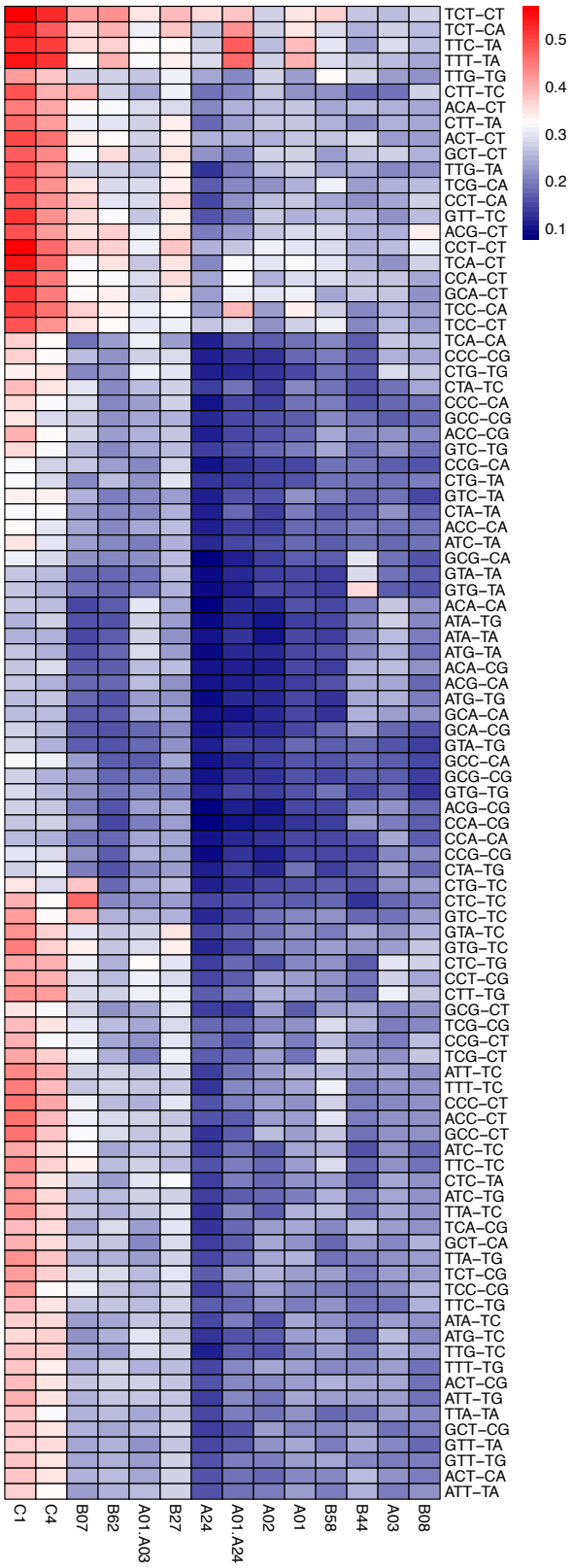

Figure 3.1: Heatmap showing the expected proportion of immunogenic mutations of each mutation type and HLA supertype, using the rank < 2 threshold

of percentage rank < 0.5 estimated by using NetMHCPAN 4.0 [264](see Methods for details) allowed for the consideration of only strong binders (see Figure 3.2b) whereas percentage rank < 2 includes weak binders too (Figure 3.2a). Notably, the overall patterns observed in each analysis mirrored those observed for different mutation types. Specifically, HLA-C supertypes exhibited the highest binding proportions, followed by HLA-B supertypes, reaffirming the distinctive binding characteristics associated with HLA-C supertypes.

It is interesting to note that although mutational signatures have the highest immunogenicity for HLA-C supertypes, HLA-C is usually expressed at much lower levels than HLA-A and HLA-B [337]. This suggests that another immune evasion mechanism is the downregulation of HLA molecules with a higher capacity of neoantigen presentation. Downregulation of HLA molecules in general for immune evasion is a well-documented mechanism, but the observed higher presentation capacity of HLA-C could explain why it has lower expression levels than HLA-A and HLA-B [338].

3.3.2 Estimating MHC-I affinities across different cancers using their mutational landscape

In this study, we defined intrinsic immunogenicity of a tumor type as the proportion of somatic mutations within a tumor that generate putative neoantigens. To estimate the median number of mutations contributed by each mutation signature in the TCGA cohort, we utilized the attribution matrix of mutation signatures reported by [108]. By combining the median activity of mutation signatures and the expected immunogenicity from the previous section, we estimated the intrinsic immunogenicity of different tumor types for various HLA supertypes. This approach provides a comprehensive understanding of the potential of a tumor to induce an immune response through the presentation of neoantigens.

Consistent with previous studies [130, 137], we observed that certain HLA supertypes, such as B44, have a higher proportion of immunogenic mutations in specific tumor types, such as SKCM (skin cutaneous melanoma), than in other tumor types, such as LUAD (lung adenocarcinoma)() and LUSC (lung squamous cell carcinoma) (Figure 3.3,3.4) (0.236, 0.215 and 0.216 proportion of immunogenic mutations re-

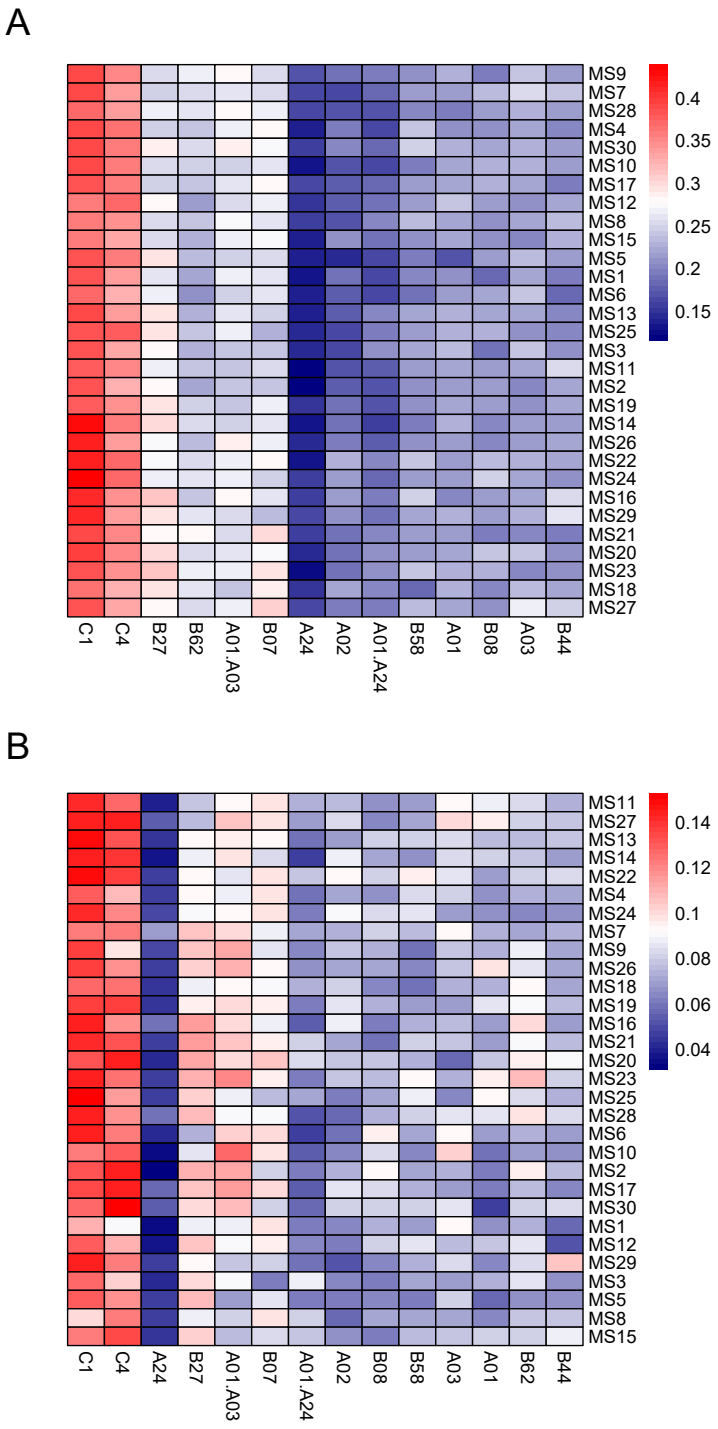

Figure 3.2: (a) Heatmap showing the expected proportion of immunogenic mutations of each mutation signature for common HLA supertypes, using the percentage rank < 2 threshold. (b) Heatmap showing the expected proportion of immunogenic mutations of each mutation signature for common HLA supertypes, using the percentage rank < 0.5 threshold.

spectively). Interestingly, we also found that when only strong binders are considered, B44 had a higher proportion of immunogenic mutations in CESC (cervical squamous cell carcinoma and endocervical adenocarcinoma) and TCC (transitional cell carcinoma) (0.0699 and 0.0720 respectively) (Figure 3.4a), which preferentially exhibit transition mutations, particularly C>T, similar to melanoma [108, 339], which showed 0.0723 proportion of immunogenic mutations. This suggests that the underlying mutational processes and genetic variation in MHC-I alleles can play a crucial role in determining its immunogenicity.

We compared the expected immunogenic proportions with the empirical proportions obtained from the TCGA cohort (Methods). We found a strong correlation between the two estimates of immunogenicity of cancer types ($r = 0.96$ and $r = 0.87$ for percentage rank less than 2 and 0.5, respectively; p-value $< 2.2 * 10^{-16}$ for both analyses; Figure 3.3b, 3.4b). These results suggest that our method accurately predicts the intrinsic immunogenicity of various tumor types for different HLA supertypes. However, it is important to note that this correlation is influenced by the variations in HLA alleles. As shown in Figure 3.3a,3.4a; the differences between HLA alleles had a much greater effect on the variation in predicted immunogenicity than the differences in mutational patterns within cancers.

Next, we used patient level mutation signature activity and patient-specific HLA genotypes within the TCGA cohort to determine the expected immunogenicity of a cancer type (Methods). We compared the median expected proportion of immunogenic mutations among samples within a specific tumor type with the corresponding median observed proportion of such mutations across the same set of samples (Figure 3.5). We did not find correlation between the two estimates using patient-specific HLA genotype and mutation signatures. This can be explained by the lack of variation in mutation signature activity across cancer types in the TCGA cohort (Figure 3.6a) and also by the low variation in the expected immunogenicity of mutation signatures (Figure 3.6b) which results in marginal variation in the expected immunogenicities of cancer types (Figure 3.6c).

To further test the robustness of the relationship between the expected and empirical immunogenicity, we controlled the confounding effect of HLA genotype variation. We randomly selected two HLA alleles from the set of unique HLA alleles

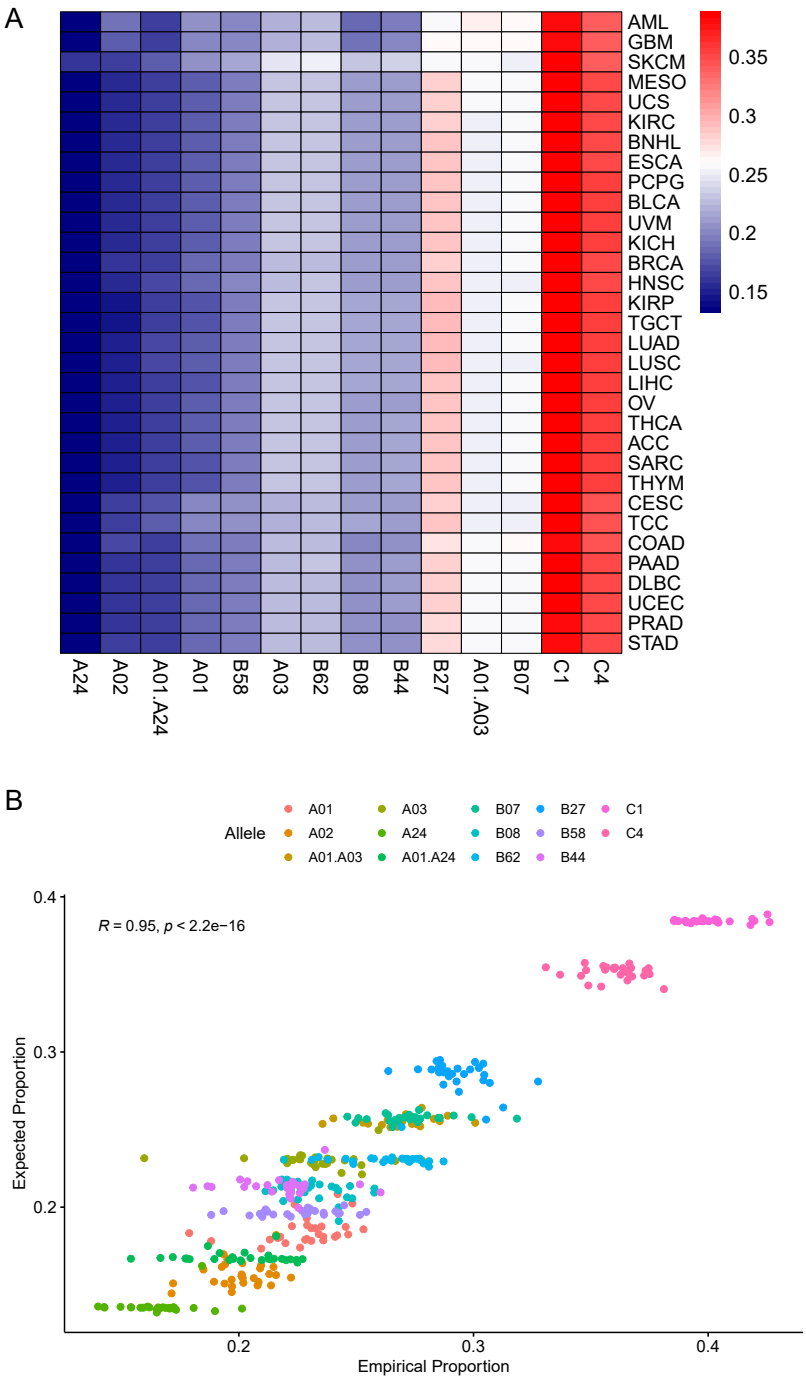

Figure 3.3: (a) Heatmap showing the expected proportion of immunogenic mutations of each cancer type for most common HLA supertypes, using the rank < 2 threshold. (b) Scatterplot showing the relationship between the expected and empirical proportion of immunogenic mutations in a cancer type, using the percentage rank < 2. Each point represents the proportion of immunogenic mutations for a cancer type, and they are grouped by HLA supertypes.

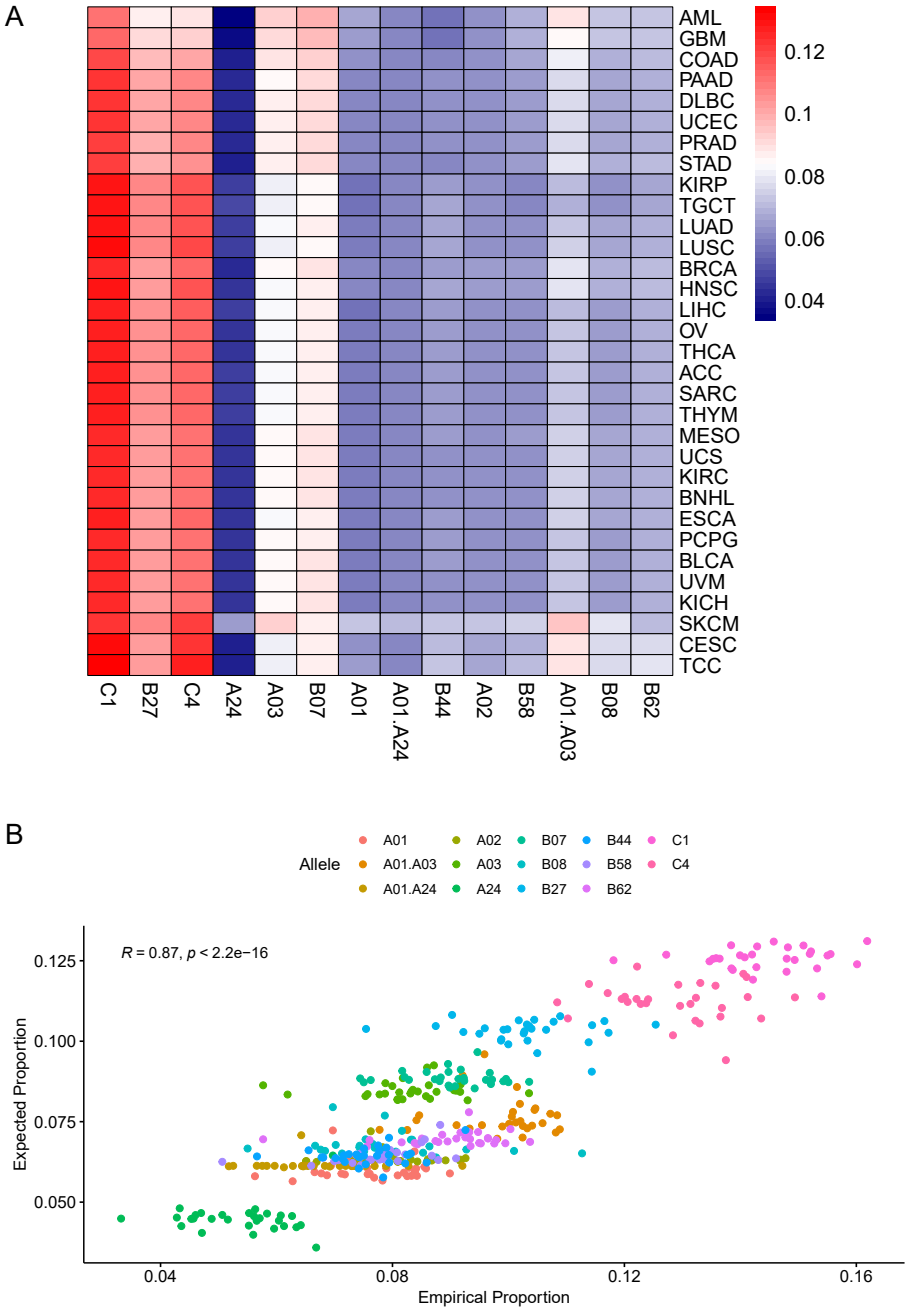

Figure 3.4: (a) Heatmap showing the expected proportion of immunogenic mutations of each cancer type for most common HLA supertypes, using the rank < 0.5 threshold. (b) Scatterplot showing the relationship between the expected and empirical proportion of immunogenic mutations in a cancer type, using the percentage rank < 0.5. Each point represents the proportion of immunogenic mutations for a cancer type, and they are grouped by HLA supertypes.

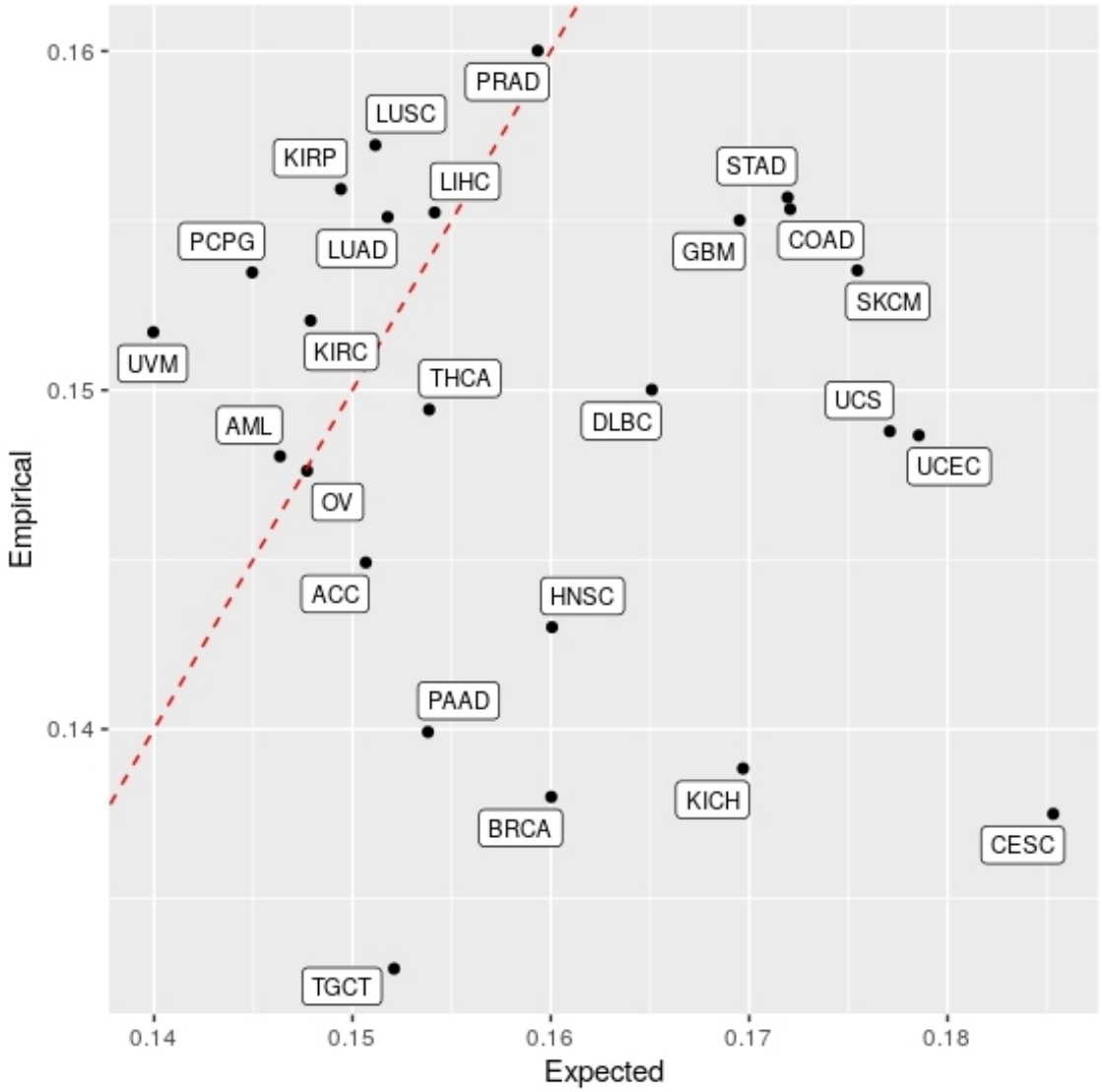

Figure 3.5: Scatterplot showing the relationship between the median expected and empirical immunogenicity of cancer types across samples in the TCGA cohort using patient-specific HLA genotype

found in the TCGA samples. We estimated the expected and empirical proportion of immunogenic mutations in various cancer types using the median mutation signature activity and randomly sampled HLA alleles (Figure 3.7a, b). Our results showed there is a positive correlation between the expected and empirical proportion of immunogenic mutations for both alleles. This indicates we can predict the variation in immunogenicity of various cancer types to some extent using specific HLA alleles and mutation signature activity. Figure 3.7a, b).

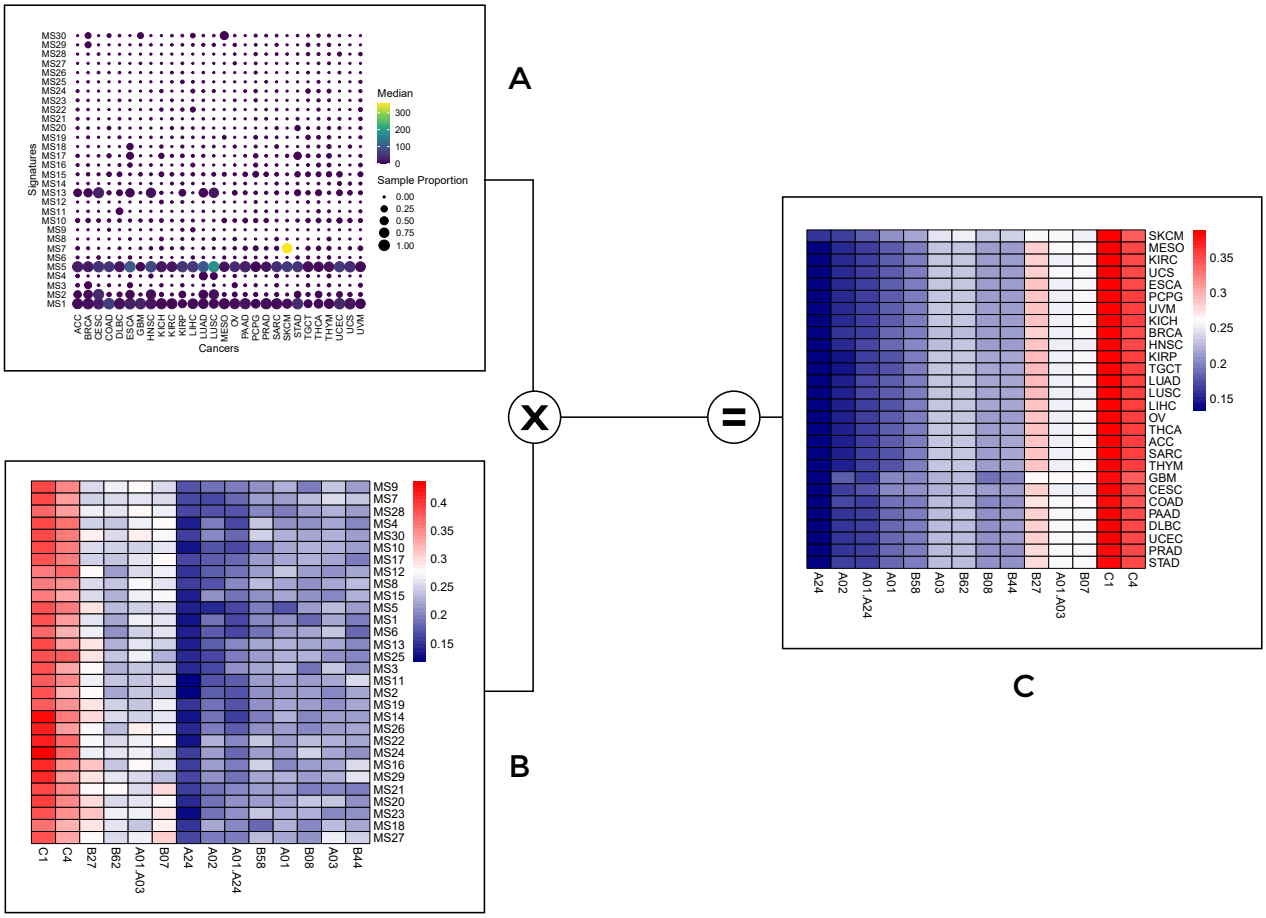

Figure 3.6: (a) Dot heatmap showing the median activity of mutation signatures in the TCGA cohort. The color of each dot represents the median value, whereas the size of the dots represents the proportion of samples in which a given mutation signature is present. (b) The expected proportion of immunogenic mutations for each mutation signature (row) for each HLA supertype (column). (c) The product of matrix multiplication of (a) and (b), showing the expected proportion of immunogenic mutations in each cancer type

3.3.3 Estimating MHC-I affinities of TCGA cohort using their mutational landscape

We calculated the expected immunogenicity of each mutation signature in the TCGA cohort using patient-specific HLA genotypes (Methods). We observed that the mutation signatures differ from each other in immunogenicity but the differences were small (Figure 3.8a), as we observed in (Figure 3.2). Overall, mutation signature MS20 had the highest predicted immunogenicity in the TCGA cohort, while MS1

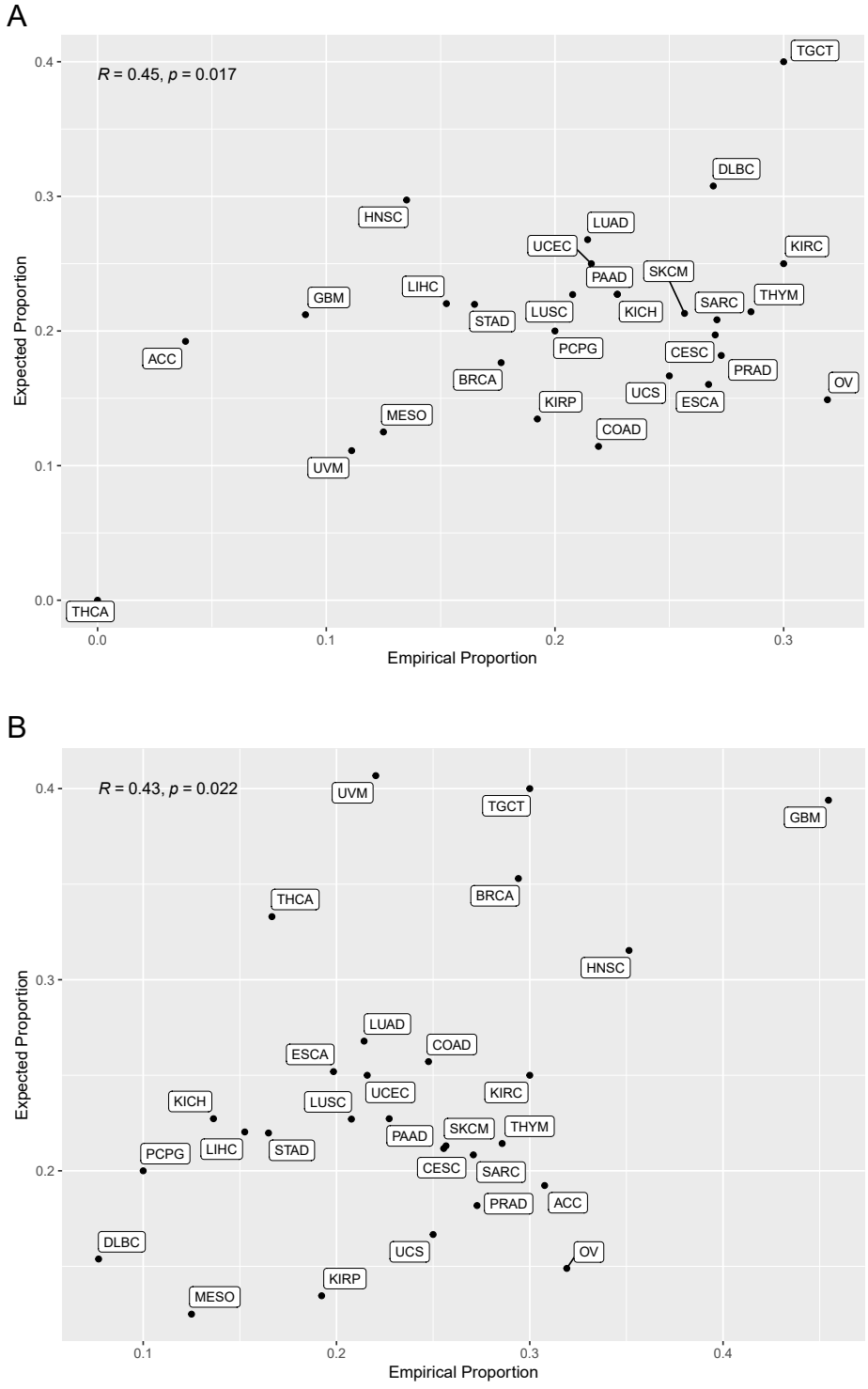

Figure 3.7: Scatterplots showing the relationship between the expected and empirical proportion of immunogenic mutations using the randomly sampled HLA alleles.

had the lowest predicted immunogenicity. When we restricted this analysis to the activity of each signature in TCGA samples, we observed that the number of im-

munogenic mutations contributed by each signature was directly proportional to the number of somatic mutations contributed by that signature (Figure 3.8b). This is consistent with previous reports [340, 341] suggesting that the higher the TMB the higher the neoantigen load and the better the response to immunotherapy. Interestingly, mutation signature 5 had the highest activity and second lowest immunogenicity. Mutation signature 1, with the lowest immunogenicity, was the third most active signature.

We extended our method to estimate the expected immunogenicity of each TCGA sample, using the six HLA alleles of each patient and the mutation burden contributed by each mutation signature in their tumour (Methods). To evaluate the potential of our results, we compared the expected immunogenicity of each sample with its empirical immunogenicity, which was calculated by using the PHBR score method [227, 228]. We compared the expected and empirical proportion of immunogenic mutations; in this case, the correlation dropped drastically, with a rho value of 0.09 and a p-value of $1x10^{-4}$. This drop was expected as the expected immunogenicity at the individual patient level was highly stochastic relative to the expected immunogenicities for cancer types. This difference is because, in the case of cancer types, we sampled 1000 mutations for each cancer type reducing the stochasticity, whereas individual patients had much smaller numbers of mutations.

3.3.4 Can mutational landscapes shape immunotherapy outcomes?

Studies have shown that neoantigen load can play a role in predicting the efficacy of immunotherapy. Higher neoantigen load has been associated with improved responses to immunotherapy in some cancer types (Hutchinson 2016). We tested whether the expected proportion of immunogenic mutations in a tumor sample, has the same association with immunotherapy outcomes. We carried out mutational signature analysis for the Dana Farber (DF) [333] and the Memorial Sloan (MSKCC) [319] melanoma cohorts and observed that both cohorts were enriched for C>T mutations (58.32% and 82% respectively) (Figure 8), similar to the UCLA melanoma cohort as reported by Cumming et al [130].

We used these mutation signatures and predicted the expected proportion of immunogenic mutations for each sample using their HLA genotype and the activity

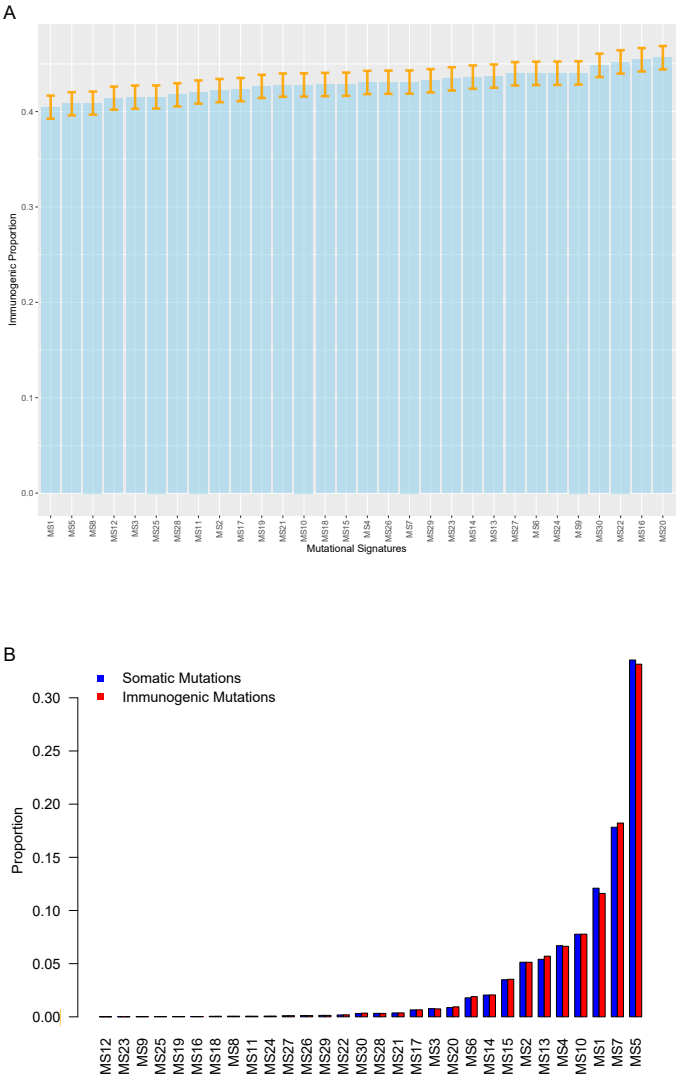

Figure 3.8: (a) Expected proportions of immunogenic mutations for each mutation signature across the TCGA cohort. (b) Grouped barplots, showing the proportion of somatic mutations and immunogenic mutations contributed by each signature in the TCGA cohort.

of these signatures. We then tested if it can be used to predict the efficacy of immunotherapy treatment. Our goal was to establish whether the relationships found between different mutation signatures and HLA supertypes are useful and reflective of observed data.

In the present study, we analyzed two separate cohorts, namely the DFCI cohort consisting of 110 patients and the MSKCC cohort comprising 60 patients. In the DFCI cohort, 32% of patients exhibited at least one B44 supertype allele, while in the

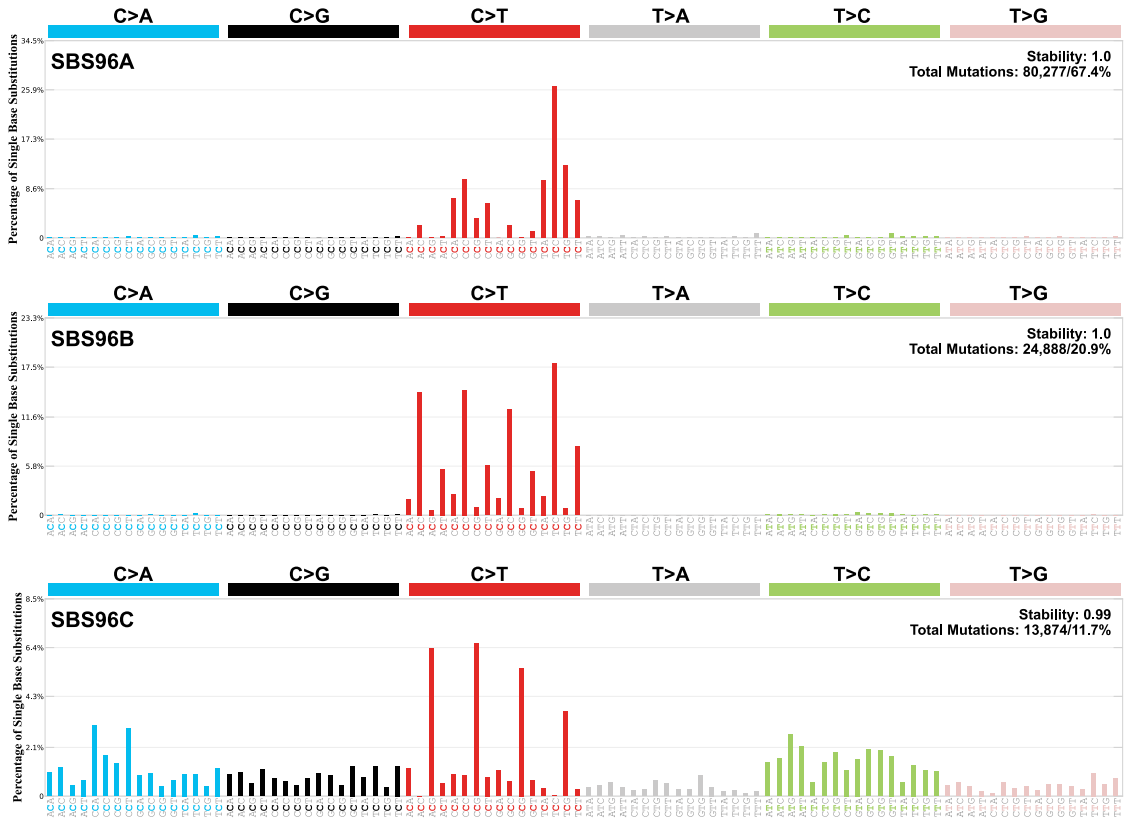

Figure 3.9: Extracted mutational signatures for two melanoma cohorts. These signatures showed that these melanoma cohorts are enriched with C>T mutations.

MSKCC cohort, this proportion was 28%. To further investigate the impact of HLA alleles, we also examined the presence of B27 supertype alleles due to their similar binding pocket characteristics to B44. Within the DFCI cohort, approximately 23% of patients carried at least one B27 supertype allele, whereas in the MSKCC cohort, the proportion was approximately 18%. Notably, there were 7 overlapping patients between the B44 and B27 groups in the DFCI cohort and 4 patients in the MSKCC cohort. We categorized our data based on the estimated immunogenicity of the patients, into low and high expected immunogenicity groups after estimating the optimal cut points for survival analysis (Methods).

The median survival for the high expected immunogenicity group was 10.6 months and 6.74 months for the low expected immunogenicity group in the DFCI cohort. Whereas in the MSKCC cohort, the median survival for the group with high expected immunogenicity was 94.6 months and 15 months for the low immunogenicity group. In both cohorts, we found that the expected proportion of immunogenic

mutations was predictive of immunotherapy efficacy (p-value 0.042 and 0.022 for the DFCI and the MSKCC cohorts, respectively) (Figure 3.10a, b). It is important to note that 69% of the patients with higher expected immunogenicity had B44 or B27 alleles in the MSKCC cohort, whereas in the DFCI cohort this measure was 57%. Previously it had been shown that these two supertypes have beneficial effects in melanoma patients [130, 137, 281]

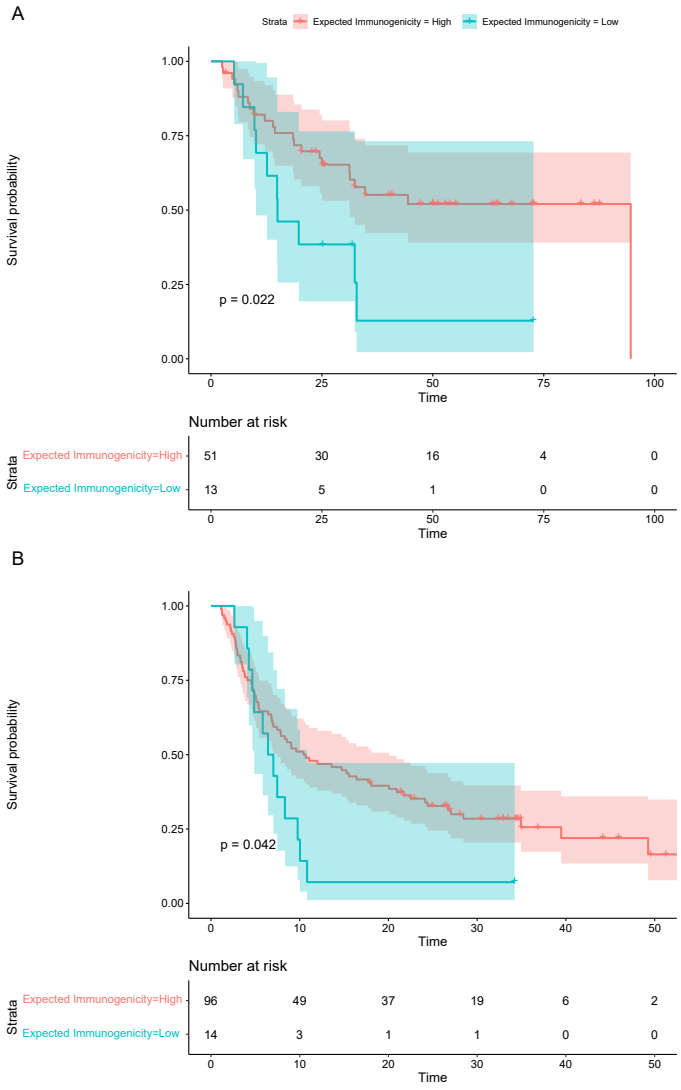

Figure 3.10: Survival estimated using the Kaplan–Meier method. (a) In the MSKCC cohort patients with high expected immunogenicity had a median OS of 94.6 months versus 15 months in the low expected immunogenicity group, P= 0.022. (b) In the DFCI cohort patients with high expected immunogenicity had a median OS of 10.6 months versus 6.74 months in the low expected immunogenicity group, P=0.042.

3.4 Discussion

In this chapter, we investigated the relationship between different HLA supertypes and mutational signatures in the context of immunogenicity. Our analysis revealed variations in the immunogenicity of mutational signatures, although these differences were relatively smaller compared to the divergences observed for different HLA supertypes. We found that HLA-C supertypes have the highest immunogenicity for mutational signatures, suggesting that the reported downregulation of HLA-C [338] as compared to HLA-A and HLA-B in cancer patients could be another mechanism of immune evasion.

The beneficial effect of HLA B44 supertype in melanoma patients is associated with an enrichment of C>T mutations due to underlying mutational processes [130]. To investigate this further, we performed signature analysis on two ICB treated melanoma cohorts, namely the Dana Farber (DF) and MSKCC cohorts. Both cohorts exhibited an enrichment of C>T mutations, consistent with previous reports in melanoma patients. We then explored whether the expected proportion of immunogenic mutations could predict immunotherapy efficacy, similar to the observed neoantigen load. Our analysis revealed a positive relationship between the expected proportion of immunogenic mutations and immunotherapy efficacy. This finding has important implications for predicting the efficacy of immunotherapy.

Recent studies have shown that tumor mutation burden (TMB) may not be a reliable universal biomarker for predicting immunotherapy response across all cancer types [291, 342, 343]. It has been previously proposed that for samples with low tumor purity, where it is difficult to accurately measure TMB, mutation signature analysis can be used as a proxy for TMB [344]. We propose that one potential proxy for TMB is the assessment of intrinsic immunogenicities of mutational signatures or mutational patterns in the tumor. Our results suggest that this may have potential for the prediction of immunotherapy efficacy. However, it is important to note that we employed the maximally selected rank statistic method to determine an optimal cutoff point for dividing our data into high and low expected immunogenicity groups. This method provides a value of a cutoff point that corresponds to the most significant relationship with the outcome, survival in this case. The statistical tests are biased due to the same data being used to define the threshold

and to assess the significance of the difference between the samples above and below the threshold. Nonetheless, our results suggest a positive trend, with patients with higher expected proportions of immunogenic mutations exhibiting better survival. Notably, we established a correlation between expected immunogenicities of various tumor types, which were estimated using mutational signature activity, and observed immunogenicities. This correlation supports the existence of a relationship between the probability of mutation occurrence in specific nucleotide sequences and the predicted HLA affinities for corresponding peptides. Our findings are consistent with a previous study by [230], who observed a similar association; however, their study employed a more conservative approach that encompassed only six HLA alleles. In contrast, our analysis was exhaustive, considering a broader range of HLA alleles, which enhances the robustness of our results.

The observed correlation between the expected and empirical proportion of immunogenic mutations, indicates that underlying mutational processes can play a role in defining the immunogenicity of a cancer type. This observation has implications for investigations of neoantigen depletion, as we would expect that in the presence of negative selection, the expected immunogenicity will be greater than the empirical immunogenicity of tumor (i.e., there will be fewer immunogenic mutations than you would expect, given the mutation signature activity profile).

Negative selection is thought to eliminate cells carrying mutations that elicit an immune reaction [177, 232, 233, 301, 304, 307, 321], but its effects on cancer genomes are not fully understood. Studies have provided clear evidence that mutational signatures - patterns of mutations in DNA - need to be considered when detecting selection signals in cancer [230, 234, 235], as they can bias metrics used for the detection of immunoediting signals. [230] showed that the apparent neoantigen depletion signal disappears when mutation signatures are considered. Another study reported that in melanoma cancer ultraviolet light dimerization gives rise to C>T mutations resulting in an increased rate of synonymous mutations in hydrophobic amino acid codons [235]. This increased rate of synonymous mutations creates a bias in the dN/dS metric, incorrectly suggesting that negative selection is acting on somatic mutations in cancer.

The same mutational pattern (C>T) is associated with increased immunother-

apy efficacy in patients with HLA-B44 supertype, suggesting mutation signatures also influence immunotherapy responses [136, 137]. This effect is associated with the presence of radical glutamic acid substitutions at the anchor position, resulting in neoepitopes for B44 [130]. Another analysis revealed that the expression of genes encoding non-motif neoepitopes was higher compared to the expression of genes encoding motif neoepitopes [281]. This observation suggests the possibility of tumors having an evolutionary advantage in evading immunosurveillance through a decreased availability of motif neoepitopes. Nonetheless, the precise impact of mutational signatures and their sequence context preferences on signals of neoantigen depletion and immunotherapies remains uncertain.

Overall, the study demonstrated that the affinities of MHC-I supertype alleles and the mutational landscape of a tumor can play a crucial role in determining its immunogenicity. We estimated the intrinsic immunogenicity of mutation signatures and tumor types. This method can be applied to detect immunoediting signals by comparing the estimated intrinsic immunogenicity and observed immunogenicity of a sample. This study laid the foundation for the analysis we performed in the next chapter, where we test and quantify the immunoediting signal.

3.5 Methods

3.5.1 Data acquisition

We used version 2 of COSMIC mutational signatures [108] to predict the intrinsic immunogenicity for HLA supertypes. The classification of HLA supertypes was gathered from [345]. To acquire the variant annotation of the TCGA cohort, we accessed the TCGA portal and utilized the variant annotated files generated by Multi-Center Mutation Calling in Multiple Cancers project [346]. To ensure high quality variant calls, we further refined the annotations by using only those variants that passed all filters and were called by at least two variant callers. We used the HLA typing of the TCGA cohort provided by [228] in our analysis. The data of cohorts used for the prediction of ICB treatment efficacy was taken from previously published studies [319, 320].

3.5.2 Peptide binding

Most of all known MHC-I ligands are of length nine, so we only considered peptides of length 9 (ninemers) containing the residue (mutated amino acid) for binding prediction. Peptide binding affinity predictions for ninemers were obtained for various HLA alleles using the NetMHCPan-4.0 tool. NetMHCPan-4.0 returns IC 50 scores and corresponding allele-based ranks. Peptides with rank < 2 and < 0.5 were considered weak and strong binders respectively [264].

3.5.3 Empirical Immunogenicity

We defined empirical immunogenicity as the proportion of immunogenic mutations among observed missense mutations in a sample. We calculated the empirical immunogenicity of a cancer type for an HLA supertype by first randomly sampling mutations observed in samples of that cancer type and estimating their binding affinity for the HLA supertype. To determine the empirical immunogenicity of a patient, we used the PHBR scoring method described in [227, 228]. This method gives an aggregated binding score to each mutation observed in that patient, considering all six HLA-I alleles of the patient.

3.5.4 Expected Immunogenicity of Mutation Signatures

To estimate the immunogenicity of a mutation signature, we first randomly sampled genomic positions having the same trinucleotide context as mutations in the probability vector of the mutation signature, which consists of probabilities of a mutation occurring in these trinucleotide contexts. We then simulated single nucleotide substitutions corresponding to the mutation types in the probability vector of a mutation signature in these sampled genomic positions. In the next step, we calculated the binding affinity for these simulated mutations with HLA supertypes reported in [345] using the method described above. Finally, an immunogenicity scores i.e., proportion of immunogenic mutations, were assigned to each mutation signature and HLA supertype pair, which was calculated by dividing the number of immunogenic mutations by the total number of missense mutations. We also estimated the expected variation in the immunogenicity of various mutation signatures in the TCGA cohort using patient specific HLA genotype. For this purpose, we

simulated one mutation for each mutation signature for all the TCGA samples, and then used PHBR method to check if simulated mutation is immunogenic using HLA genotype of a given patient. Then, we calculated the proportion of immunogenic mutations for each mutation signature across all TCGA samples.

3.5.5 Expected immunogenicity of a cancer type

We predicted the expected immunogenicity of a cancer type for an HLA supertype by using the expected immunogenicity of mutation signatures with that HLA supertype, and the median activity of that mutation signature in cancers of that type in the TCGA cohort. We performed a dot product calculation between two vectors: one representing the immunogenicity of mutation signatures, and the other representing the median number of mutations attributed to each mutation signature in a specific cancer type. We used the attribution matrix reported by [108] for the TCGA cohort to estimate the medians. This attribution matrix was obtained using the (SigProfilerAttribution) function of the SigProfiler package.

3.5.6 Expected Immunogenicity of TCGA cohort

To calculate the expected immunogenicity of a TCGA sample, we used the attribution matrix of mutation signatures in the TCGA cohort reported by [108]. This attribution matrix presents the number of mutations contributed by each signature in a sample. We sampled mutations corresponding to each signature based on the number of mutations attributed to that signature in each sample. This resulted in a simulated mutation dataset with the same size as the observed mutations dataset in that sample. We again used the PHBR method and threshold of PHBR < 2 to predict whether a mutation will be presented on the cell surface in the sample. Finally, using the number of presented mutations in a sample, we predicted the proportion of immunogenic mutations in a sample. To estimate the expected immunogenicity of mutation signatures across the TCGA cohort we sampled one mutation for each signature in each TCGA sample. We determined if that mutation was a putative neoantigen given the HLA genotype of the patients. We used a PHBR < 2 threshold for a mutation to be considered immunogenic.

3.5.7 Mutational Signature and Survival Analysis

We used the SigProfiler R package [108] to extract mutational signatures of the DF and MSKCC cohorts. We first extracted trinucleotide contexts of mutations observed in the samples. The matrix consisting of trinucleotide contexts of mutations is passed to sigprofilerextractor function, which extracts mutation signatures from the given set of samples. For survival analysis we used R packages survminer and survival to perform survival analysis for both cohorts. We first divided our data based on the expected proportion of immunogenic mutations into two groups (high and low) using the surv_cutpoint function to find the optimal cutpoint for survival analysis.

4 Chapter 4: Assessment and quantification of immunoediting in human cancers

4.1 Abstract

Neoantigens arising from somatic mutations potentially initiate immune responses against tumors. Neoantigens are mutated peptides presented on the cancer cell surface. It has been reported that these putative immunogenic mutations are removed by selection, in a process known as immunoediting. Because patient MHC-I genotype plays a vital role in initiating the immune response, it has been studied widely in the immunoediting context, and studies have claimed that it restricts the mutational landscape of tumors. However, this remains controversial, and rigorous research is ongoing to determine the strength of this immunoediting signal. Here, we present a method incorporating the mutational and evolutionary processes active during tumor development to detect and quantify the immunoediting signal. We estimate that fewer than 1% of mutations are removed through immune surveillance, and the immunoediting signal is weak to absent in most tumor types. These results could have significant implications in predicting immunotherapy responses and studying the role of immune surveillance in cancer prevention.

4.2 Introduction

Understanding the effects of the immune system on cancer development has been challenging. The idea that the immune system can influence cancer growth, especially that it can prevent cancer growth, has been debated for more than a century [347, 348]. Ehrlich *et al.* argued that cancer would be much more common in long-lived organisms if there was no protective system, such as the immune system [349]. This argument was further strengthened by Burnet *et al.* in 1957, and the term "immune surveillance" was coined by them in their paper [350]. The immune surveillance theory postulated that adaptive immunity plays a role in protecting the immunocompetent host from cancer . However, this hypothesis was abandoned when studies published in 1974 and 1975 by Stutman *et al.* did not support it and showed by experiments that tumor susceptibility was similar in both immunocom-

petent and immunodeficient mice models [351, 352]. This debate continued until the 1990s, when it was settled with the help of improved mice models and advanced technologies that the immune system does act as a tumor suppressor in immunocompetent mice [353, 354].

In the early 2000s, studies showed that the immune system not only controls cancer growth, i.e., tumor quantity but also plays a role in defining tumor immunogenicity, i.e., tumor quality [355]. Shankaran *et al.* showed that tumors in an immune-deficient mice model were more immunogenic than similar tumors grown in an immunocompetent mouse [356]. This observation that the tumor was unedited in a host lacking a competent immune system led to the formation of the immunoediting hypothesis [204], which proposed that negative selection pressure acts to remove somatic mutations that have the potential to initiate an immune response in a tumor [177]. This immune response is initiated when the MHC-I molecule presents neoantigens resulting from somatic mutations on the surface of the cancer cells [299]. CD8+ killer T cells then recognize these neoantigen-presenting cancer cells, and, if sufficient co-stimulatory signals are present, these cancer cells are eliminated by the immune system [156]. We call such somatic mutations putative immunogenic mutations, and the negative selection acting on them is called immunoediting.

Although immunoediting is considered to be confirmed in mice models [357], it still lacks convincing evidence in humans. Several studies have explored immunoediting and its extent in humans in the past few years. Hannah Carter and colleagues at UCSD reported that MHC-I genotypes restrict the oncogenic mutational landscape in humans[228]; however, the findings of this study were not supported by [230]. Their study did not find a strongly detectable neoantigen depletion signal. Two other studies argued the reported signals of immunoediting were the result of low immunogenicity of common driver mutations [70, 234]. We have already discussed [70] in Chapter 2 of this thesis.

Various mutational processes can cause somatic mutations in cancer during tumor evolution [86, 106, 108]. The properties of mutational processes in cancer have been characterized by trinucleotide-based mutation signatures [86, 106, 108]. This method assumes that the occurrence probability of a single nucleotide mutation at a locus depends on the upstream and downstream nucleotide and on the active mu-

tational processes. Van Den Eyden *et al.* reported that the immunoediting signal becomes negligible when mutation signatures are taken into consideration, but their method had some limitations [230], such as a very conservative but general estimate of the HLA binding region [358].

More recent studies, such as Marta *et al.*, have suggested that neoantigen quality rather than quantity is a better predictor of immunoediting signal [294]. Marta, L. *et al.* showed that an immunoediting signal exists by studying the evolution of a cohort of pancreatic cancers over ten years [294]. They found that long-term survivors with a more robust immune response in the primary tumor have fewer immunogenic mutations in the recurrent tumors, despite having more time to accumulate mutations. They use the mutation quality to quantify the immunoediting signal in this cohort. Another study used K-S statistics to estimate the difference between the cumulative distributions of cancer cell fraction of antigenic mutations and non-antigenic mutations in each sample of the TCGA cohort to quantify the immunoediting signal. They reported a strong immunoediting signal in many cancer types [358] and argued that the method of immunoediting signal detection used in the paper [230] was problematic. They did not find the neoantigen depletion signal because the actual neoantigens did not exist in the region they defined as the HLA-binding region [230, 358] .

Immunotherapies have shown very promising results for specific cancer types; however, they do not work for all patients because of the heterogeneous nature of cancer, its evolutionary potential, and the diversity of the human immune system. Cancer grows by a repeating process of cell multiplication, genetic changes, and selection, happening in the body's natural environment [359]. Studies have shown that immune system intervention may deplete cancer clones and erode their habitats, but this may also result in selection pressure for the elimination of immunogenic mutations or expansion of resistant variants [226]. Given the importance of immunotherapy in cancer treatment and the need to predict clinical responses accurately [175] the extent to which immunoediting shapes the landscape of somatic mutations in cancer is a critical question.

Many previous studies have tested for the effect of immunoediting on the somatic mutations that are observed in cancer, with conflicting results, as described

above. Here, we present a novel method to quantify the extent of immunoediting using patient MHC-I profiles and mutation signature activities. This method allows us to place an upper bound on the number of mutations removed by immune surveillance during cancer development. We observed a weak immunoediting signal in a pan-cancer analysis, but this immunoediting signal was lacking in many individual cancer types. Furthermore this apparent purifying selection does not disappear when we use random HLA alleles rather than patient-specific ones, indicating that HLA binding affinity mainly depends on protein sequence composition, which is in line with previous reports [230, 235]. Our analysis provides no substantial evidence to support the notion that the mutational landscape of a tumor is influenced by patient HLA genotype.

4.3 Results

In this study, we developed a patient-specific method to estimate the expected and empirical immunogenicity scores of tumors in the TCGA cohort. Expected immunogenicity refers to the hypothetical proportion of immunogenic mutations that would be present in the absence of negative selection pressure. To estimate expected immunogenicity, we simulated a random mutation with the same mutational signature for each mutation observed in the patient's tumor sample. This generated a simulated tumor sample with a similar mutational background and the same number of somatic mutations as the original sample. The empirical proportion of immunogenic mutations in the patient's original tumor sample reflects the proportion of immunogenic mutations that were able to escape negative selection pressure and are present in the tumor. We calculated the empirical immunogenicity using a similar approach as the expected immunogenicity, with the difference that it considers the actual mutations found in the tumor sample.

Our hypothesis was that if negative selection pressure is acting on immunogenic mutations in a tumor sample, then the expected proportion of immunogenic mutations in the simulated tumor sample would be greater than the empirical proportion of immunogenic mutations. Our results showed that the mean expected empirical proportion of immunogenic mutations in the TCGA cohort were 0.158 and 0.148, respectively. We interpreted the difference in mean immunogenicity as evidence of

immunoediting, as the expected immunogenicity was significantly higher than the empirical immunogenicity (mean difference = 0.010, pairwise t-test, p = 2.22 x 10^{-12} ; see Figure 4.1).

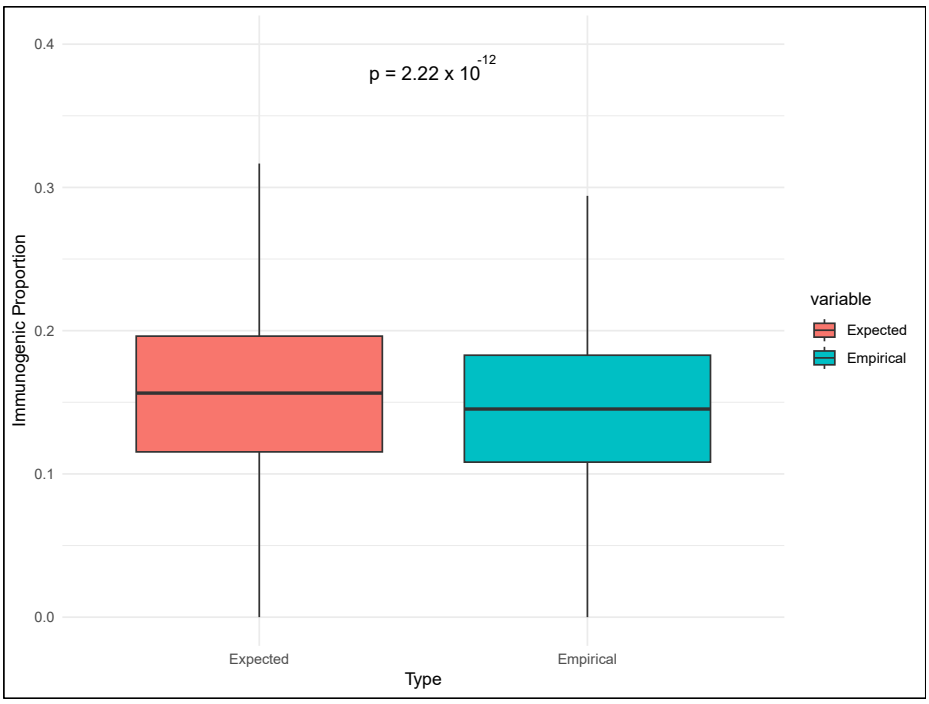

Figure 4.1: Boxplots comparing expected and empirical proportions of immunogenic mutations.

We next inspected the strength of the immunoediting signal in individual cancer types. Interestingly, we observed that most cancer types did not show a significant immunoediting signal when comparing the expected and empirical immunogenicity of tumor samples within each cancer type, contrary to what was observed in the pan-cancer analysis (16 out of 27 cancer types lacked a significant signal of immunoediting; see Figure 4.2). Among the cancer types that did show a significant signal, Transitional Cell Carcinoma (TCC) exhibited the strongest immunoediting signal with a mean difference of 0.02 (p-value = 1.15 x 10^{-8}). On the other hand, we observed a negative mean difference of -0.01 (p-value = 0.04) in Lung Squamous cell carcinoma (LUSC). A previous study [230] reported a negative signal of immunoediting in 8 out of the 11 cancers where we found evidence of immunoediting, specifically in HNSC, LUAD, LUSC, BRCA, SKCM, BLCA, CESC, and UCEC.

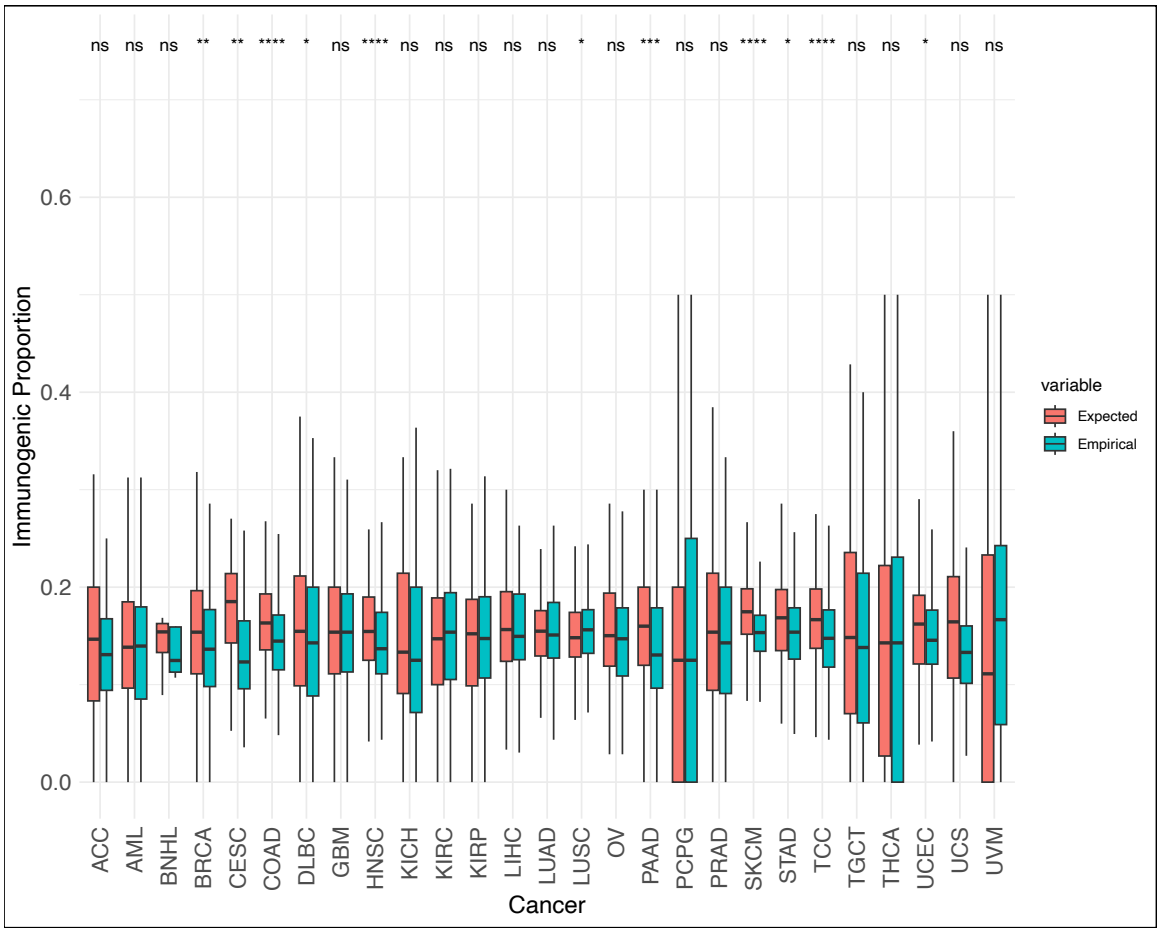

Figure 4.2: Boxplots comparing expected and empirical proportions of immunogenic mutations for individual cancer types.

4.3.1 Immunogenicity of clonal and subclonal mutations

To incorporate diverse evolutionary processes and selection pressures evident during different stages of cancer development into our analysis, we categorized the somatic mutations observed in tumor samples as clonal or subclonal mutations (Methods). As tumors progress, certain clones may acquire the ability to evade immune surveillance by downregulating or eliminating the expression of immunogenic markers, thereby creating a selective pressure favoring the expansion of less immunogenic or immune-resistant tumor clones [92, 99, 307].

In the early stages of cancer, we anticipate a stronger negative selection pressure by the adaptive immune system on immunogenic clonal mutations. This can be attributed to the limited development of immune evasion mechanisms during this phase, making these mutations more vulnerable to elimination. A recent study

conducted by Rosenthal *et al.* has also reported the presence of stronger negative selection pressure acting upon clonal mutations [293]. Consequently, we hypothesized that, under immunoediting pressure, lower proportion of clonal mutations would result in putative neoantigens as compared to subclonal mutations. However, our pan-cancer analysis revealed a minimal mean difference in immunogenicity between clonal and subclonal mutations (mean difference = 0.006, paired t-test p-value = 3.33 x 10^{-2}, Figure 4.3). Intriguingly, the mean clonal immunogenicity across the TCGA cohort was slightly higher (0.149) than the mean subclonal immunogenicity (0.143), contradicting this hypothesis.

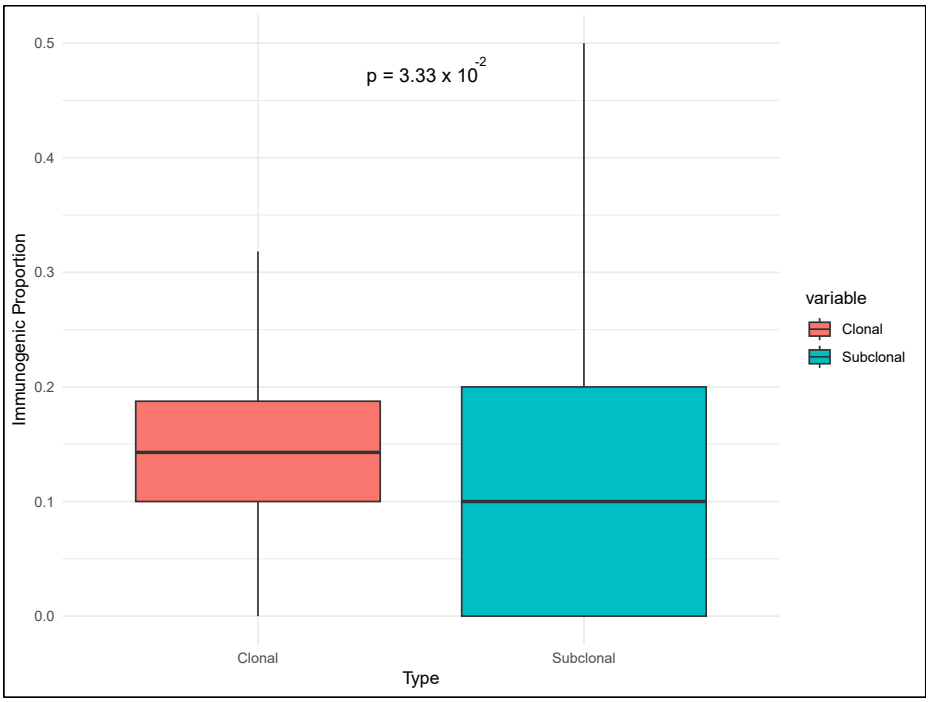

Figure 4.3: Boxplots comparing the proportion of immunogenic mutations between clonal and subclonal mutations.

To further investigate this phenomenon, we explored the variation in this signal among individual cancer types (Figure 4.4). The significance level of boxplots shows if there is a significant difference between the mean of the groups. Among the 27 cancer types analyzed, only four—AML, LUAD, SKCM, and UVM exhibited a significant difference in mean immunogenicity between clonal and subclonal mutations. In these cancer types, the proportion of immunogenic mutations was statistically significantly higher among clonal mutations, except for AML where the mean difference

was -0.052 (paired t-test p-value 0.055). Overall our findings was contrary to the hypothesis that in the case of immunoediting, the proportion of immunogenic mutation would be lower in clonal mutations because in that phase advanced immune evasion mechanisms are absent but in line with our previous results for pan-cancer analysis.

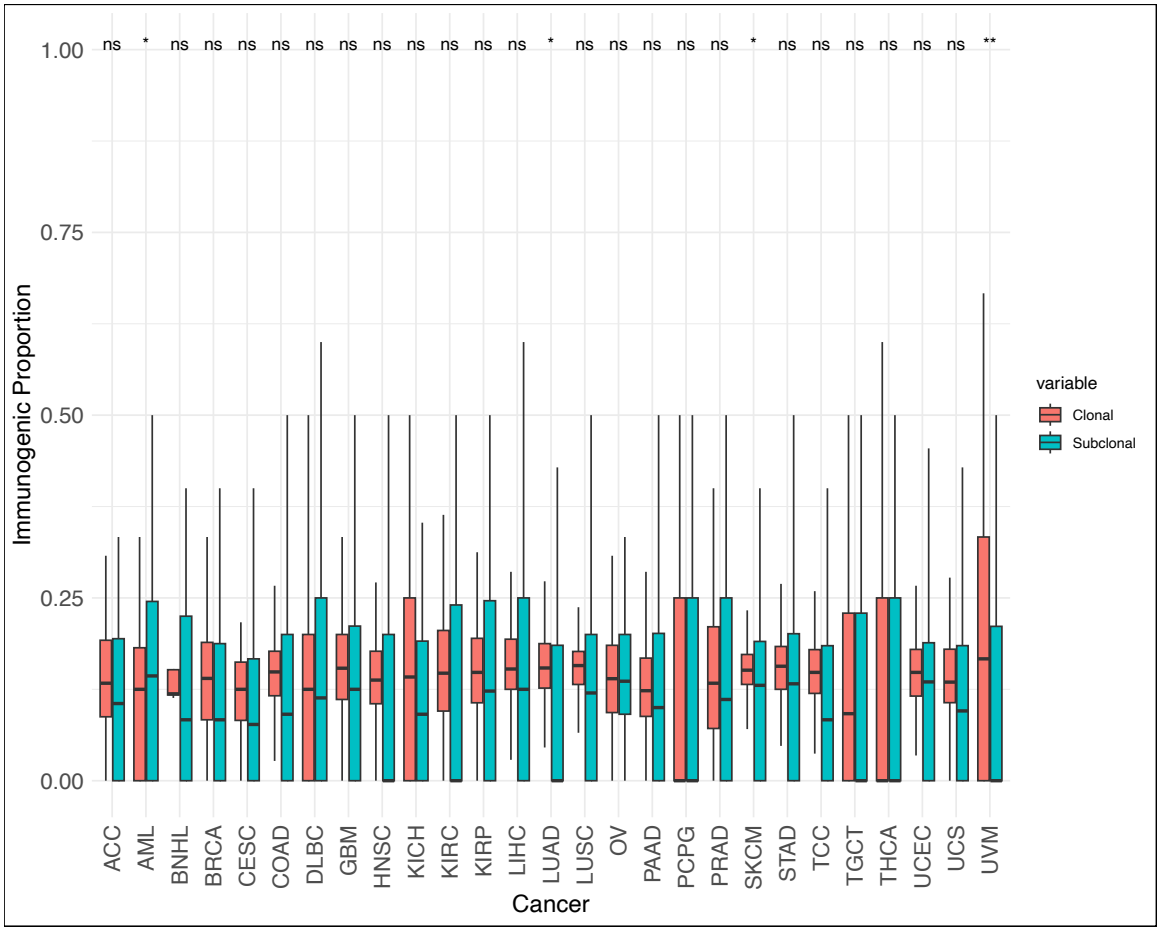

Figure 4.4: Boxplots comparing empirical proportions of immunogenic mutations in clonal and subclonal mutations for individual cancer types.

To further test our hypothesis, we compared the expected proportion of immunogenic mutations with the empirical proportion for both clonal and subclonal mutations (Figure 4.5). Our analysis revealed that both clonal and subclonal mutations exhibited higher expected immunogenicity compared to their empirical immunogenicity, indicating the presence of immunoediting in both mutation subsets. The mean difference between expected and empirical proportions of immunogenic mutations for clonal mutations was 0.01 (paired t-test p-value = 9.19×10^{-7}), while

for subclonal mutations, it was 0.017 (paired t-test p-value = 4.52 x 10^{-6}). These results are consistent with our previous findings suggesting that neoantigens have been eliminated by the immune system, resulting in lower empirical immunogenicity compared to expected immunogenicity for both clonal and subclonal mutations.

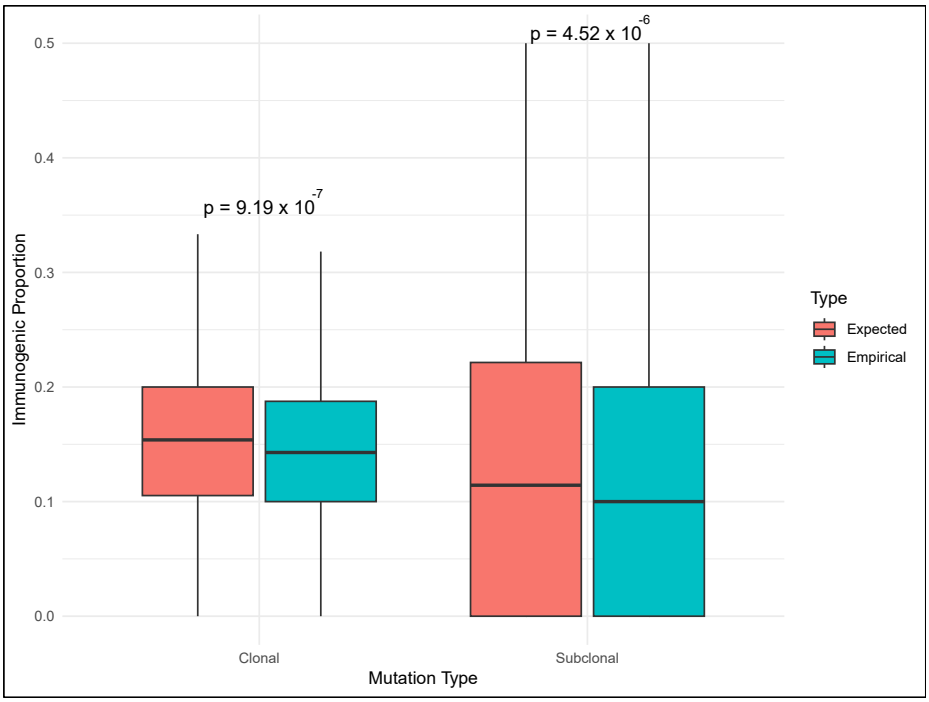

Figure 4.5: Boxplots comparing expected and empirical proportions of immunogenic mutations among clonal and subclonal mutations in the TCGA cohort.

We anticipated that if immunoediting were occurring, the difference between expected and empirical immunogenic proportions would be higher for clonal mutations compared to subclonal mutations. This expectation is based on the notion that negative selection is likely stronger on clonal mutations due to the higher expression levels of genes carrying these mutations, their essentiality for tumor growth, and the absence of advanced immune evasion mechanisms. To ascertain this, we compared the expected and empirical proportion of immunogenic mutations between both these subsets of somatic mutations. We observed that the difference between the expected and empirical immunogenic proportions of subclonal mutations was slightly higher than that of clonal mutations, although the magnitude was very small and statistically non-significant (0.007, paired t-test p-value = 0.12) (Figure 4.5). These findings suggest a lack of a neoantigen depletion signal. While our re-

sults indicate a weak immunoediting signal, we note that our analysis was limited by the use of the immunogenicity score as a proxy for immune recognition. Other factors, such as expression levels of genes carrying these mutations, the presence of antigen specific TCRs or the activity of immune cells in the tumor microenvironment, may also contribute to the immune response to cancer cells. Furthermore, our results may be influenced by sample size limitations, as the number of subclonal mutations in our dataset varied widely between cancer types.

We extended this comparison to both clonal and subclonal mutations separately for individual cancer types too. We observed a trend of higher expected proportions of immunogenic mutations in both clonal (Figure 4.6) and subclonal (Figure 4.7) subset of mutations. Although this difference was significant in more cancer types in clonal subset as compared to subclonal.

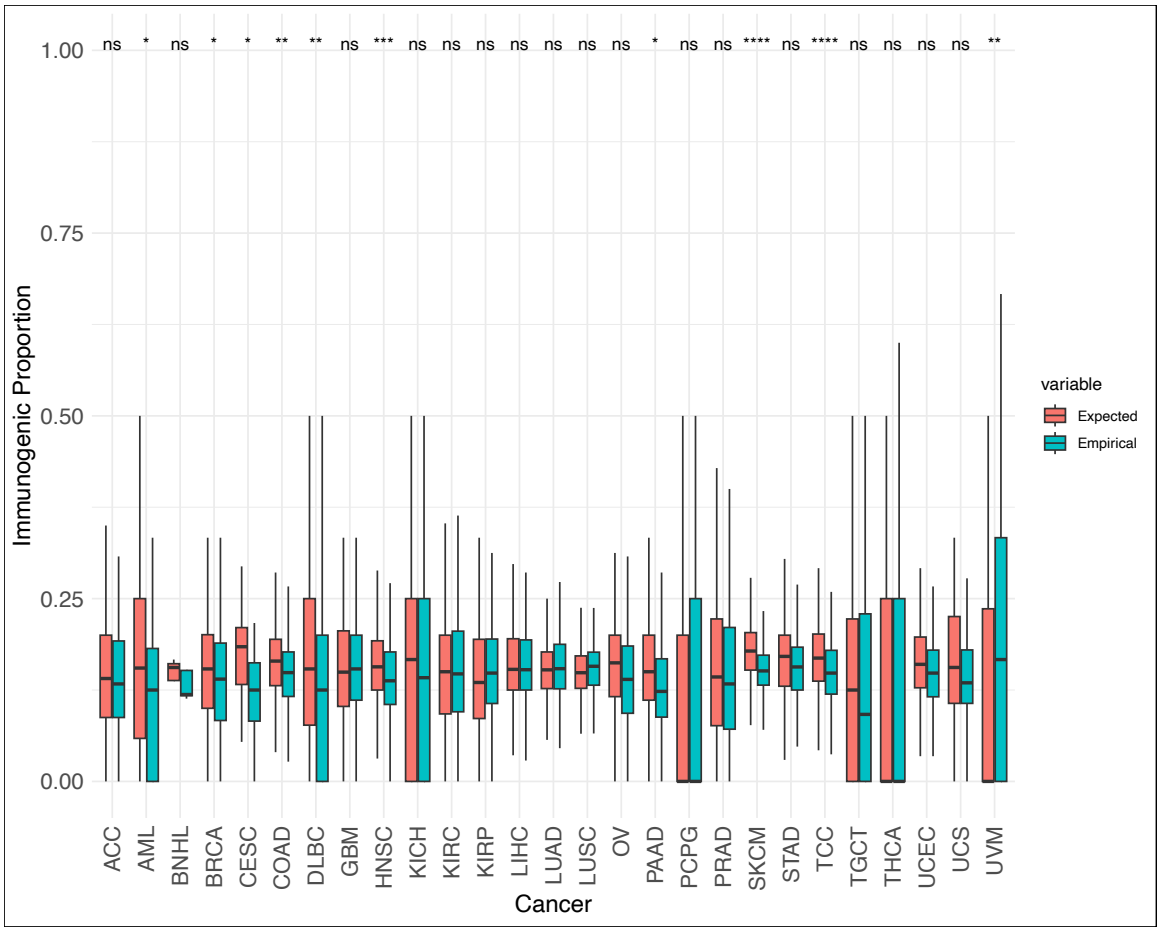

Figure 4.6: Boxplots comparing expected and empirical proportions of clonal immunogenic mutations in individual cancer types.

Overall, our results indicate at most weak evidence of immunoediting in both clonal and subclonal mutations. Although certain cancer types may exhibit stronger immunoediting signals, the overall impact of immunoediting on the genetic landscape of tumors appears to be relatively weak. This observation aligns with previous reports highlighting the overall limited signals of negative selection in cancer [[70, 230, 234, 235].

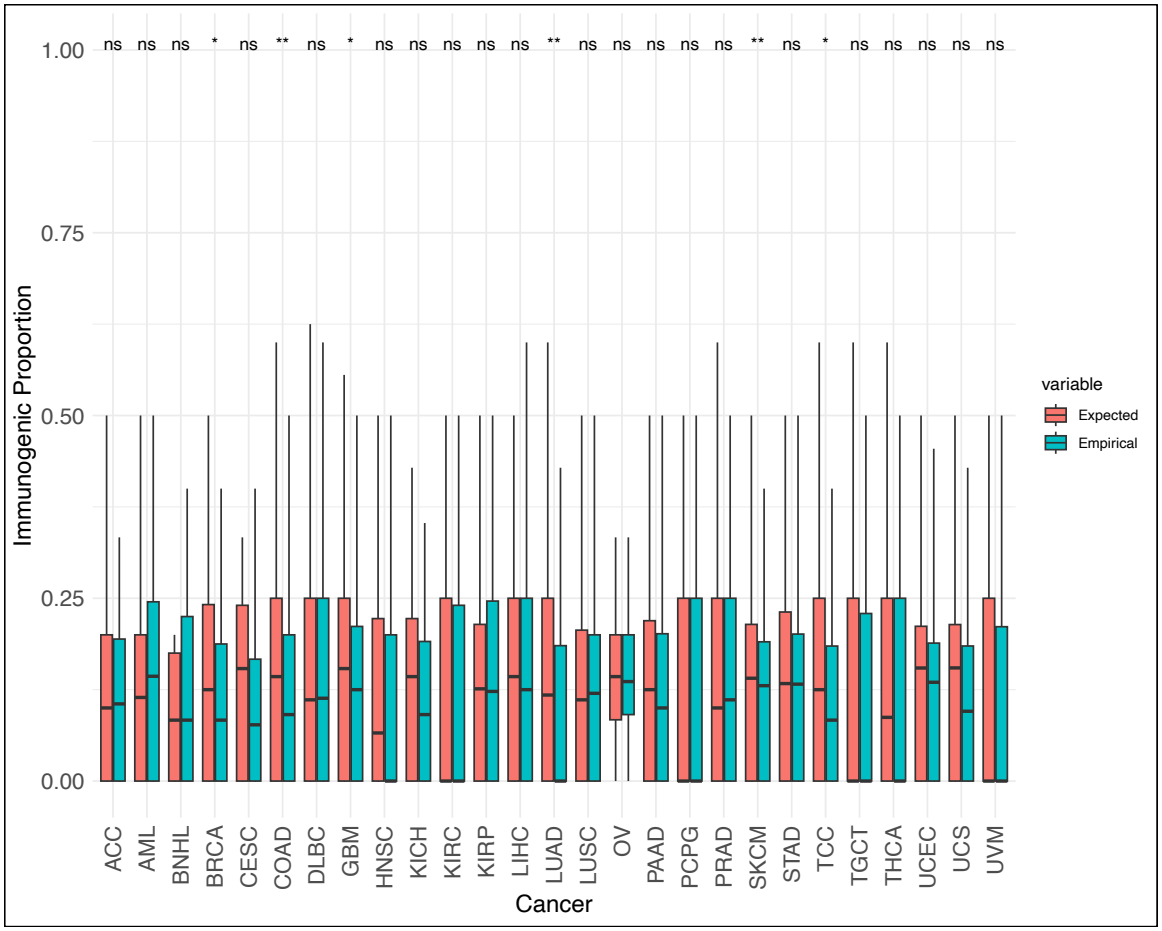

Figure 4.7: Boxplots comparing expected and empirical proportions of subclonal immunogenic mutations in individual cancer types.

4.3.2 Immunoediting signal persists even with random HLA alleles

In this section, we investigated the possibility that the weak immunoediting signal observed in our analysis may be due to differences in sequence composition on longer length scales between the observed mutations and random mutations with the same triplet context. MHC-I binding is dependent on peptides of length 8-11 amino

acids. When we select a mutation at random with the same nucleotide triplet context as an observed mutation the peptides in which it occurs would typically be different and systematic differences between the amino acid sequence contexts in which the somatic mutations and the randomly sampled mutations are found could cause differences in the immunogenicity between the two groups. Factors such as the presence of specific epitopes, the abundance of rare codons, and the overall structure of the sequence can all impact the immunogenicity of a given mutation. Therefore, it is crucial to consider the sequence composition when assessing the immunogenic potential of somatic mutations. We repeated our analysis using random MHC-I alleles instead of patient-specific alleles. The primary objective of this analysis was to determine whether negative selection on immunogenic mutations inferred using a patient's HLA alleles would still be present or significantly reduced when using random HLA alleles sampled from the population to assess the immunogenicity of the mutation.

Firstly, we evaluated the empirical proportions of immunogenic clonal and subclonal mutations using random HLA alleles (Figure 4.8). We expected that if the immunoediting signal observed in Figure 4.3 is driven by patient MHC-I genotypes, then this will diminish in the case of the randomized HLA alleles. But our results showed that the magnitude of difference between the clonal and subclonal empirical immunogenicity was similar to what we observed using the patient's own HLA alleles 4.3. We inferred from this finding that the observed neoantigen depletion signal is not driven by patient MHC-I genotype (Figure 4.8).

Next, we compared the expected and empirical immunogenicity of clonal and sub-clonal mutations using randomized HLA alleles (Figure 4.9). Again we expected that the differences between expected and empirical proportions of immunogenic mutations, which we interpret as the signal of negative selection, will be smaller or insignificant in case of randomized HLA alleles. But similar to our previous findings, the magnitude of the differences was very close to the mean immunogenicity differences we observed in the case of patients specific HLA alleles (Figure 4.5). The fact that immunogenicity inferred using the HLA alleles of the patient did not yield stronger immunoediting signals than randomly sampled alleles casts doubt on the existence of immunoediting signal. These results again are consistent with a lack

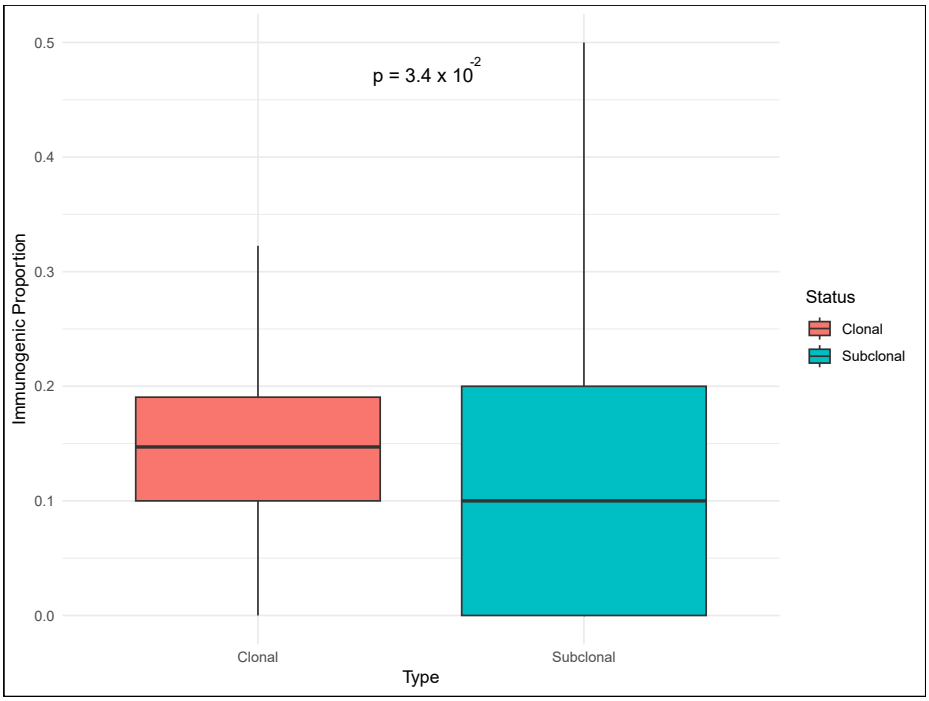

Figure 4.8: Boxplots comparing the proportion of immunogenic mutations between clonal and subclonal mutations using randomized HLA alleles.

of MHC-I restricted immunoediting signal in human, in keeping with our previous findings discussed in Chapter 2 of a lack of evidence for a role of MHC-I in shaping mutational landscape in cancers [70]. This could also imply that very few mutations are removed by immune surveillance.

We compared the empirical immunogenicity of clonal and sub-clonal mutations estimated using patients' MHC-I genotype with that estimated using random MHC-I alleles, sampled at random from the population (Figure 4.10). Our results revealed no significant difference between the two, indicating that immunogenicity is not primarily dependent on MHC-I genotype (Figure 4.11,4.12). It also indicates the important role of sequence composition of the mutated peptide in potential immunogenicity. Our results also highlight the limitation of triplet context of a mutation in fully covering the sequence composition of resultant mutated peptide.

Our finding is consistent with previous research studies that have also reported a strong association between sequence composition and the immunogenicity of somatic mutations, including the studies by [327] and [230]. These studies emphasized the significance of assessing sequence composition to evaluate the immunogenic potential

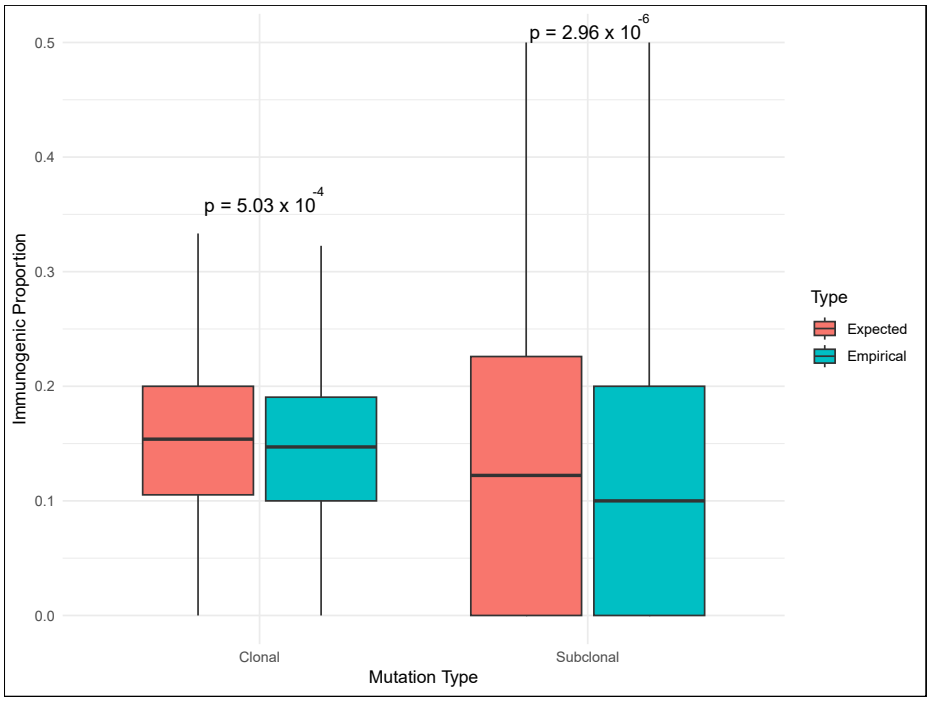

Figure 4.9: Boxplots comparing expected and empirical proportions of immunogenic mutations among clonal and subclonal mutations in the TCGA cohort using randomized HLA alleles.

of somatic mutations.

4.3.3 Determination of an upper bound on the contribution of immunoediting

The magnitude of the differences in predicted immunogenicity between observed somatic mutations and randomly sampled mutations was very small. This is consistent with the presence of a weak immunoediting signal; however, the fact that this signal persists when immunogenicity was predicted using random HLA alleles rather than the patient's own HLA alleles suggests that the signal may result from failure of the mutation signatures to fully capture the sequence context-dependence of the observed somatic mutations. Even if we assume that the weak signal does indeed reflect immunoediting, the weakness of the signal is not consistent with a major role for immunoediting in shaping the observed somatic mutations. To investigate this further we designed a simulation to infer an upper bound for the number of mutations that are removed by immunosurveillance during cancer development. The

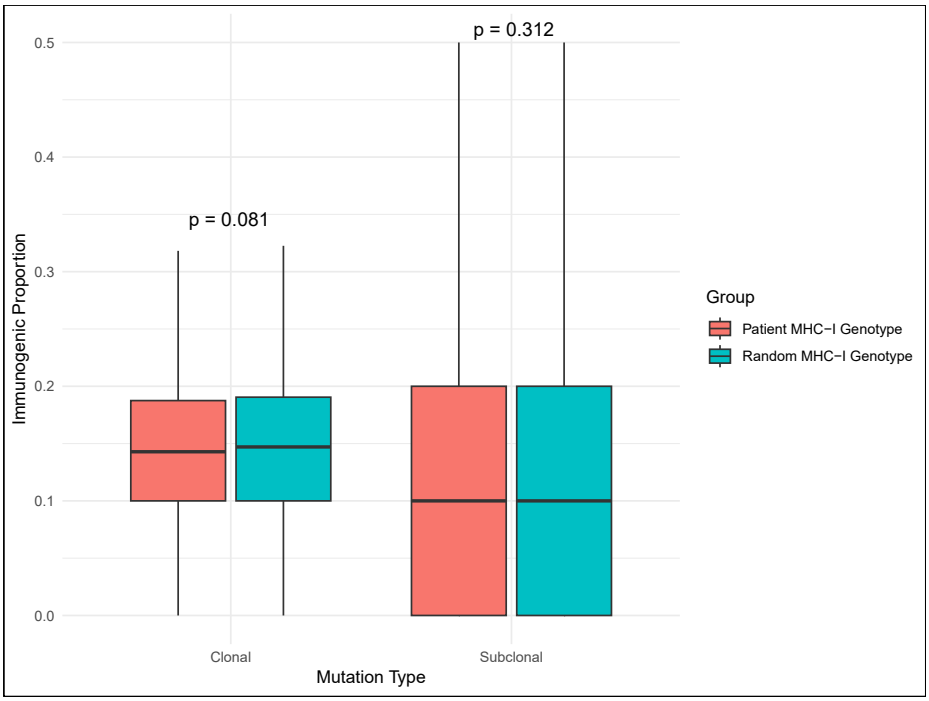

Figure 4.10: Boxplots comparing proportions of immunogenic mutations estimated using Patient MHC-I genotype and randomized MCH-I genotype.

method involved determining how many immunogenic mutations must be removed until a difference is observed between the real and shuffled data (see Methods). Our results suggest that fewer than 1% of the immunogenic mutations have been removed by immunosurveillance during cancer development.

4.4 Discussion

In this study we compared the proportion of somatic mutations inferred to be immunogenic in TCGA patients to the expected proportion, based on random sampling of somatic mutations from the patient-specific mutation signature activity profile. There was a slightly lower proportion of immunogenic mutations among the observed compared to the randomly sampled somatic mutations, suggesting a weak immunoediting signal. This result aligned with previous studies that have reported similar observations of a little or no impact of negative selection on the mutational landscape in cancer [70, 230, 234, 235]. Van Den Eynden *et al.* demonstrated that accounting for mutational patterns substantially reduces the immunoediting signal, although their method had certain limitations, also highlighted by [120]. The au-

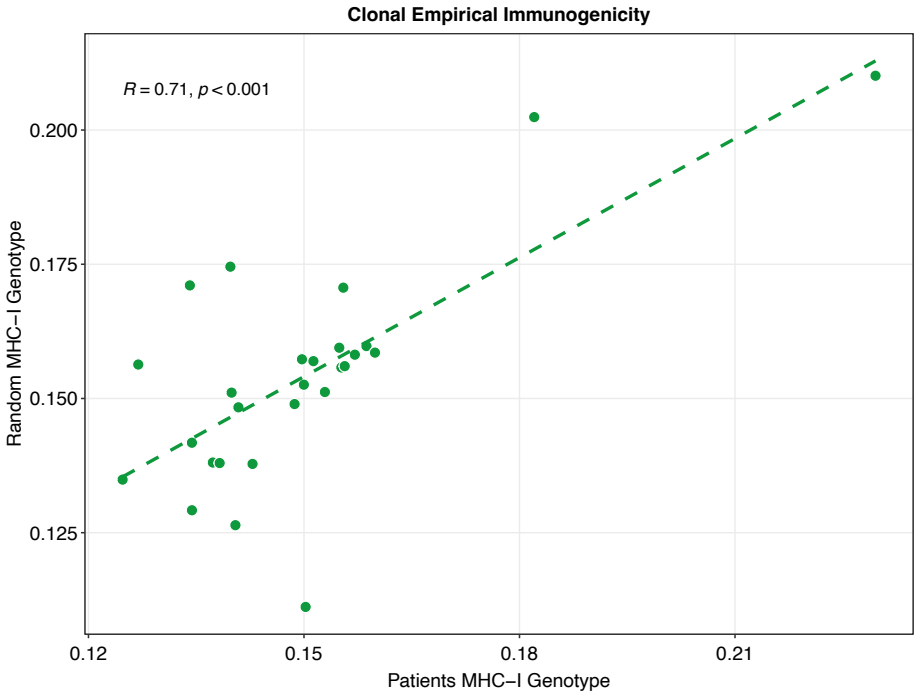

Figure 4.11: Correlation between proportions of immunogenic mutations estimated using Patient MHC-I genotype and randomized MCH-I genotype in clonal and subclonal mutations.

thors in [120] argued that although [230] annotated the whole coding genome, they only used six HLA alleles for the annotation of the HLA binding region, leading to a conservative yet broad estimate of the HLA binding region [120]. The presence of a weak immunoediting signal implies that immune surveillance may have a limited role in eliminating immunogenic mutations during cancer development. We extended this analysis to individual cancer types, and we observed a similar pattern in individual cancer types, but this signal was not statistically significant in most cancer types. The cancer types with statistically significant neoantigen depletion signal are reported to have higher mutational burden relatively, and studies have provided evidence that higher mutational burden is associated with increased neoantigen production, supporting the argument that higher mutational burden contributes to the apparent signal of immunoediting [212, 319, 323, 327]

Previous studies have highlighted the crucial role of clonal architecture in determining immunogenicity, emphasizing that neoantigens originating from clonal mutations are more likely to elicit an immune response compared to those arising

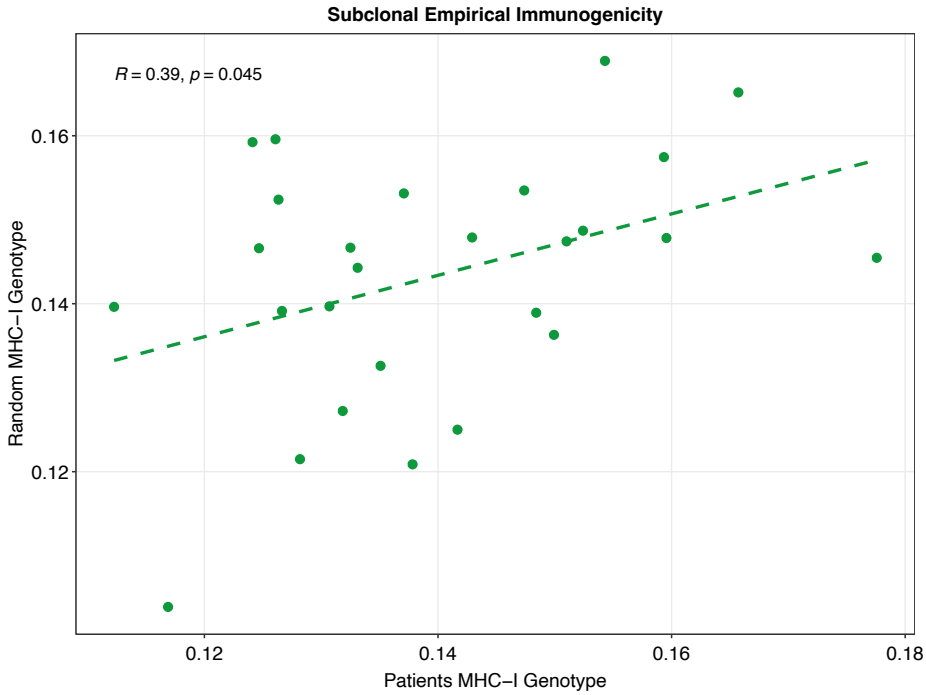

Figure 4.12: Correlation between proportions of immunogenic mutations estimated using Patient MHC-I genotype and randomized MCH-I genotype in clonal and subclonal mutations.

from subclonal mutations [212, 293, 327, 360]. This led us to hypothesize that negative selection pressure would be stronger on clonal mutations and result into lower proportion of immunogenic mutations in clonal mutations as compared to subclonal mutations. We observed an opposite trend in pan-cancer analysis, although the difference between the proportion of immunogenic mutations was minimal. A similar pattern was observed by [311], which suggests that negative selection mechanisms in the immune system act to prevent subclonal neoantigens from becoming highly prevalent within a tumor. This effect is more prominent in tumors with higher mutation rates, where the immune system can effectively eliminate or suppress cells displaying subclonal neoantigens. The lack of statistical significance in the observed results may be attributed to the lesser statistical power resulting from the smaller dataset size of each cancer type. The absence of an immunoediting signal may indicate effective immune escape mechanisms acquired by the tumor over time, and this finding is in line with the previously reported studies [212, 293, 327, 360].

To evaluate the proportion of immunogenic somatic mutations in both clonal

and subclonal populations, we compared them to the expected proportion. The expected proportion was determined by randomly sampling somatic mutations from the patient-specific mutation signature activity profile at both clonal and subclonal levels. Our analysis yielded results consistent with our previous findings, indicating slightly lower proportions of immunogenic mutations among the observed somatic mutations compared to the randomly sampled mutations in both clonal and subclonal subsets. These results suggest a weak immunoediting signal. We hypothesized that there would be a greater difference between clonal and subclonal mutations, with stronger negative selection acting on clonal mutations. However, contrary to our expectations in the pan-cancer analysis, we found that the difference in expected and observed immunogenic proportions was slightly larger for subclonal mutations. Nonetheless, this difference was not significantly larger than the difference observed in clonal mutations. Furthermore, when we performed the same analysis for individual cancer types, this signal was not significantly present in most cancer types for either the clonal or subclonal subsets of mutations. These results suggest that immunoediting signal is weak and inconsistent across different cancer types.

We further evaluated the strength of this apparent immunoediting signal by redoing our analysis using random HLA alleles rather than patient-specific HLA alleles. Our results showed that the immunoediting signal persists even with random HLA alleles. This suggests that the immunoediting signal observed in the study is not entirely dependent on patient-specific HLA alleles. The differences we observed in the proportion of immunogenic mutations and expected proportion could be because of the differences in the sequence compositions of the observed and randomly sampled mutations. This may lead to systematic differences between the amino acid sequences of the peptides within which the observed and randomly sampled mutations occur. Indeed, HLA binding is influenced not only by the amino acid sequence of the peptide itself but also by the surrounding residues [230, 361]. These amino acid sequence context effects are, of course, not captured by the triplet mutation context. This results in the observed differences in expected and empirical proportion of immunogenic mutations in both cases when we use MHC-I genotype of patients and randomized MHC-I genotype. This indicates that these differences are not mainly driven by MHC-I genotype. This can be implied from this observa-

tion that the mutations with the same trinucleotide context may occur in regions with different sequence composition resulting in, on average, different effects on the peptide sequence or, more importantly, in the anchor position, which heavily influences the HLA binding affinity. Thus leading to the observed differences between the immunogenicity of the two groups.

The observation that similar proportions of immunogenic mutations are predicted in cancer samples when we use patient-specific HLA data and shuffled HLA data suggests that either immunoediting does not make a substantial contribution to the mutation landscape in cancer or that the immunogenicity of somatic mutations cannot be predicted accurately from HLA data. These findings have important implications for cancer immunotherapy, particularly the development of personalized cancer vaccines. Personalized vaccines are designed to target patient-specific neoantigens and HLA typing is a crucial step in identifying these neoantigens. However, the results of this study question whether the HLA type data is sufficient to predict neoantigen immunogenicity. If immunoediting does indeed make a substantial contribution to the somatic mutation landscape in cancer, then personalized vaccine design may require factors beyond just HLA typing for the accurate prediction immunogenic neoantigens.

The lack of significant immunoediting signal in most cancer types observed in our study is consistent with previous studies reporting the small overall impact of negative selection on cancer evolution [70, 230, 234]. However, a comprehensive method for quantifying the strength of this phenomenon is lacking in the literature. Recently some studies have presented methods to quantify the immunoediting signal [120, 294]. In this study, we aimed to contribute to the field by proposing a novel approach to estimating an upper bound on the immunoediting signal. We found that the immune system removes fewer than 1% of mutations. This finding is consistent with previous studies that have reported a high level of tolerance for somatic mutations in cancer cells [85, 235, 362, 363]. One possible explanation for this tolerance is that the immune system has difficulty distinguishing between cancerous and normal cells [364]. This also suggests that most tumors have developed efficient immune escape mechanisms, which enable them to evade the immune system and grow unchecked. This is a cause for concern, as it highlights the challenges

that must be overcome to develop effective immunotherapeutic strategies for cancer treatment.

It is important to note that the study has limitations, such as the use of computationally predicted, rather than experimentally measured, HLA binding affinities. Many factors can influence the accuracy of neoantigen predictions, including the quality of the sequencing data and the accuracy of the HLA typing. As a result, the predicted neoantigens may not always reflect the true immunogenicity of the tumor. One recent study reported that only 6% of predicted neoantigens are actually immunogenic [365]. Beyond the presence of neoantigens, tumor immunogenicity is influenced by multiple factors, such as the quality of neoantigens, accessibility to T cells, and potential for T cell recognition. The tumor microenvironment also plays a crucial role in elimination of cells carrying immunogenic mutations. Although MHC-I binding has been widely used in studies as a proxy for the immunogenicity of a tumor [130, 230, 234, 313, 366] , it is also important to consider the T cell reactivity of predicted neoantigens while estimating a signal of neoantigen depletion. One recently published paper has addressed this issue, and has introduced a tool, DeepNeo which considers both neoantigen presentation and T cell reactivity [277].

In conclusion, to determine the prevalence and strength of the immunoediting signal in tumors, we developed a method that takes account of mutation signature activities. Our analysis indicates that the immunoediting signal is weak to absent in most tumor types. This implies that either only a small portion of the predicted neoantigens are actually recognized by the immune system, or developing tumors have developed effective ways to evade immune system, such as through HLA loss or PDL1 amplification, such that the presence of immunogenic neoantigens is not a key determinant of whether a cell clone can develop into a tumor. This study also proposed a novel approach to estimating an upper bound on the immunoediting signal and found that the immune system removes at most 1% of mutations as a result of their inferred immunogenicity, suggesting the presence of effective immune evasion mechanisms. We also emphasize the limitations of neoantigen prediction methods and encourage further validation studies. By providing insight into the prevalence and strength of the immunoediting signal in various tumor types, our study advances the understanding of the relationship between tumor evolution and the immune sys-

tem and lays the groundwork for future research in cancer immunotherapy.

4.5 Methods

4.5.1 Data Acquisition

To acquire the variant annotation of TCGA cohort, we accessed the TCGA portal and utilized the variant annotated files generated by [257]. To ensure high quality variant calls, we further refined the annotations by using only those variants that passed all filters and were called by at least two variant callers. This filtering step was crucial to ensure accuracy and reduce false positives in our subsequent analysis. We used the HLA typing provided by [366] in our analysis.

4.5.2 Mutation Signature Analysis

We used the R package SigProfiler to analyze mutation signatures in the TCGA cohort. First, we obtained mutation data from the TCGA portal, using the variant annotated files generated by MC3 [257]. We then refined the data by selecting only variants that passed all filters and were called by at least two variant callers. To identify mutation signatures, we applied the non-negative matrix factorization (NMF) algorithm implemented in SigProfiler. NMF is a computational method that identifies the underlying mutational processes contributing to a set of mutations. We used the recommended settings for NMF analysis in SigProfiler, including a range of signatures from 1 to 30, and a minimum cosine similarity of 0.75 for signature robustness. The output of the analysis included a matrix of signature activities for each sample, representing the contribution of each signature to the mutational burden of that sample. This allowed us to perform a comprehensive analysis of mutational signatures in the TCGA cohort and to identify the contribution of each mutational signature to the clonal and subclonal mutations.

4.5.3 Clonal & Subclonal Mutation Calling

The clonality of variants was determined by Siobhan Cleary based on the methods defined by [367]. Variant frequency was adjusted to take account of tumor purity and ploidy, and the cancer cell fraction (CCF) was calculated (the cancer cell fraction is

defined as the proportion of cells in the tumor that have the variant). Clopper and Pearson's method was used to calculate 99% confidence intervals for the CFF [368]. After calculating the CCF confidence intervals, variants were classified as clonal or subclonal based on their upper and lower CCF confidence intervals. Variants were considered clonal if the upper confidence interval for the CFF greater than 0.8, and the lower confidence interval was greater than 0.4. These thresholds were found to be optimal for defining clonal variants in unpublished work by Siobhan Cleary. Variants were considered subclonal if the upper confidence interval on the CCF was less than 0.5.

4.5.4 Peptide Binding

In our study, we focused on predicting the immunogenicity of mutated peptides using computational tools. It is important to note that the majority of known major histocompatibility complex class I (MHC-I) ligands are of length nine. Therefore, we only considered peptides containing the mutated amino acid of length nine, also known as ninemers, for binding prediction. We used NetMHCPan-4.0 [264], a widely used computational tool for peptide binding affinity predictions, to predict the binding affinity of ninemers for various HLA alleles NetMHCPan-4.0 returns IC50 scores and corresponding allele-based ranks for each peptide. Peptides with a rank less than 2 are considered weak binders, while peptides with a rank less than 0.5 are considered strong binders. To identify potential immunogenic mutations, we only considered mutations with a predicted PHBR less than 0.5. The PHBR method, defined by [366], is used to calculate the likelihood of a peptide being presented on the cell surface in a patient sample. By applying this stringent threshold, we ensured that only peptides with a high likelihood of being presented on the cell surface were considered as potential immunogenic mutations.

4.5.5 Empirical and expected proportions of Immunogenic mutations

We defined the empirical proportion of immunogenic mutations as the proportion of the missense mutations in a given patient sample for which the inferred immunogenicity passes a defined threshold. In order to calculate the expected immunogenicity of a TCGA sample, we utilized a simulation-based approach. We divide human

genome GRch38 into lists of loci for each triplet context, then we sampled genomic positions from these lists based on the triplet contexts of each mutation type and its occurrence probability in a given mutation signature. These occurrence probabilities of different mutation types were defined by [108]. This enabled us to simulate a set of mutations for each mutation signature. To determine the expected proportion of immunogenic mutations, we utilize the estimates provided by Alexandrov *et al.*, which indicate the contributions of each mutation signature to individual samples using the SigProfilerAttribution function of their SigProfiler package [108]. To perform this calculation, we randomly select the number of mutations contributed by each signature from the simulated mutations specific to that signature, for each TCGA sample. This allowed us to create a set of random mutations with similar mutational profile and same mutation numbers as of observed mutations in a sample. Because of the reduced data size of clonal and subclonal mutations subsets which limits an accurate signature analysis, we used a slightly different approach. We simulated mutations with same triplet context as of observed mutations in these subsets. Once we had the simulated mutation dataset, we used the PHBR method to predict whether a mutation would be presented on the cell surface. Finally, by counting the number of mutations that were predicted to be immunogenic in the simulated dataset, we were able to predict the proportion of immunogenic mutations in the sample. We refer to this proportion as the expected immunogenicity of the sample.

4.5.6 Determination of an upper bound on the contribution of immunoediting

To estimate an upper bound on the immunoediting signal, we assigned a random probability to each observed immunogenic mutation in the observed mutations dataset using the runif function in R. Then, we applied a series of thresholds, starting from 0.01 (indicating 1% chance of removal by immune system), to remove mutations with a probability lower than the threshold. The aim was to keep increasing the threshold until we see a statistically significant difference between the mean of the real data and the randomized data. We used paired t-test to determine this difference between the two groups.

5 Conclusions

5.1 Overview

The main focus of this thesis was to study the role of the immune system in shaping the mutational landscape of cancer. Somatic mutations occurring within genes responsible for encoding self-proteins can lead to alterations in the amino acid sequence, consequently giving rise to neoantigens. These neoantigens hold the potential to trigger an immune response when they are presented to T cells by the Major Histocompatibility Complex (MHC) [181]. The immune system, plays a pivotal role in the immunosurveillance of cancers [150], and also has the power to shape the genetic makeup of cancer by targeting cells carrying immunogenic neoantigens, thus contributing to the evolution of the cancer genome. Recent high-profile papers have indicated that driver mutations prevalent among cancer patients tend to arise in regions that the patient fails to effectively present to the immune system [227, 228]. Interestingly, this negative selection pressure was not observed for immunogenic passenger mutations [228]. We hypothesised that this disparity might be explained by passenger mutations that are predicted to be immunogenic occurring on lowly expressed or non-expressed genes. This could arise from the occurrence of these passenger mutations in genes with low expression levels (due to cancers that contain immunogenic mutations on highly expressed genes having been eliminated by the immune response) or through escape from immune recognition through the downregulation of genes carrying immunogenic mutations. Our analysis, rigorously accounting for factors such as gene length and sequence context, showed no substantiated evidence of immune editing or immune evasion through these mechanisms.

In light of these results, in Chapter 2, our focus shifted towards re-evaluating two earlier studies that had claimed that driver mutations occur preferentially in HLA genotype-dependent gaps in the capacity of the patient to present them [227, 228]. These studies reported a connection between the prevalence of specific driver mutations in cancer and the gaps within the MHC genotype of patients. Our reanalysis showed that these findings resulted from unjustified statistical assumptions. Deconstructing the observed relationship between MHC genotype and the occurrence of driver mutations, we found that it originated from the coexistence of numerous

high-frequency mutations, each possessing closely correlated MHC binding affinity scores. Upon controlling for these factors, we found no evidence to support the presence of the signal initially highlighted by [227, 228].

Moreover, if the prevalence of driver mutations among individuals with cancer were substantially influenced by the binding affinities exhibited by prevalent HLA alleles, a logical expectation would be a correlation between the number of recurring driver mutations, capable of binding to a patient's MHC molecules, and the level of associated cancer risk. However, the absence of any detectable link between the extent of cancer risk and the coverage of driver mutations offers a counter argument against the hypothesis suggesting that cancer driver mutations primarily arise within the gaps of MHC-I genotypes of patient.

In Chapter 3, our focus was on assessing the intrinsic immunogenicities of the mutational signatures found in cancer. This was achieved by using the common HLA supertypes. Our results revealed that mutated peptides resulting from specific mutation signatures had a greater probability of being presented by particular HLA alleles compared to peptides originating from other mutation signatures. Furthermore, we demonstrated that differences in the immunogenicity of mutations arising from different mutation signatures and variation in mutation signature activities can explain some of the variation in immunogenicity between tumor types and across individual samples. Notably, our findings indicated that due to the limited variation in mutation signature activity across different cancer types and the predicted immunogenicity of these mutation signatures, the overall variance in expected immunogenicity across cancer types remained relatively small. These results also demonstrated that patient MHC-I genotype is the most important determinant of the predicted immunogenicity of tumors. This is consistent with the reported loss of MHC molecules and the downregulation of MHC molecules as immune evasion mechanisms (Dhatchinamoorthy, Colbert, and Rock 2021). Moreover, our results showed, HLA-C which usually has lower expression level than HLA-A and HLA-B [338] has higher predicted immunogenicity, suggesting down regulation of specific MHC molecules capable of presenting neoantigens as another immune evasion mechanism.

We also performed a mutation signature analysis of two melanoma cohorts and

showed that they are enriched with C >T mutations, as previously reported [130]. We estimated the expected proportion of immunogenic mutations using the mutation signature activity and patient-specific HLA genotype for these two cohorts and performed a survival analysis. Our results indicated a positive trend, where patients with higher expected proportions of immunogenic mutations exhibit better survival.

Recent studies emphasized the need to analyze the types of mutations occurring and the underlying processes driving them to fully understand the impact of negative selection pressure acting on immunogenic mutations [230, 235]. In chapter 4, we aimed to estimate the expected proportion of immunogenic mutations in TCGA samples, comparing it against the actual empirical proportion. This analysis revealed a marginal difference between these two proportions, indicating a subtle neoantigen depletion signal. Notably, this weak neoantigen signal persisted even when we used randomized HLA alleles. It was deduced that the observed differences were not caused by negative selection, but instead resulted from inherent differences in sequence composition. This discrepancy is attributed to the fact that a MHC-I binding peptide typically consists of nine amino acids, leading to dissimilar sequence compositions for mutations sharing the same triplet context.

We also developed a randomization approach to estimate an upper bound for the proportion of immunogenic mutations that have been eliminated through immunoediting. The method involved iteratively removing immunogenic mutations until observable differences between the actual and shuffled data emerged. Our findings indicate that immunosurveillance likely eliminates at most 1% of immunogenic mutations. This outcome aligns with prior investigations demonstrating a significant tolerance for somatic mutations in cancer cells and indication negative selection is generally weak in cancers [85, 235, 362, 363, 369]. Additionally, our results are in line with the absence of strong immunoediting-driven loss in the cancer mutation landscape, or the potential effect of immunoediting being too subtle to be detected [230, 234].

5.2 Future perspectives

The scope of this thesis has been limited to neoantigens originating from SNVs. However, growing attention is being directed towards alternative sources of neoanti-

gens, including frame-shift and splicing mutations, as well as those arising from non-coding regions. Encompassing a comprehensive analysis of the immunogenicity of tumors by considering all potential neoantigen sources could significantly enhance our understanding of the role of the immune system in eliminating cancer. Such an approach might have implications for the development of immunotherapies and the identification of biomarkers.

While we found no evidence supporting the substantial elimination of cancerous cells carrying immunogenic mutations by the immune system, it is noteworthy that instances of natural regression in certain cancer cases have been reported [370]. Although rarely reported, these observations hint at the presence of effective immunosurveillance. A compelling question for further exploration could be an assessment of the prevalence of successful cancer elimination through immunosurveillance if, indeed, the immune system plays a role in spontaneous cancer removal. In particular, a key research goal should be to quantify the contribution of immune surveillance to cancer prevention. Investigating the association of different cancer types with various immunodeficiencies in large cohorts should provide one approach to tackling this question. The theory of immunosurveillance implies a higher incidence of cancer in immunodeficient than in immunocompetent individuals, but the magnitude of the increased prevalence of specific cancers associated with specific classes of immunodeficiencies should prove informative.

Bibliography

[1] Beerman, I.; Maloney, W. J.; Weissmann, I. L.; Rossi, D. J. Stem cells and the aging hematopoietic system. *Current opinion in immunology* **2010**, *22*, 500–506.

[2] Sándor, S.; Kubinyi, E. Genetic Pathways of Aging and Their Relevance in the Dog as a Natural Model of Human Aging. *Frontiers in genetics* **2019**, *10*, 948.

[3] Ema, H.; Kobayashi, T.; Nakauchi, H. Principles of Hematopoietic Stem Cell Biology. *Hematopoietic Stem Cell Biology* **2010**, 1–36.

[4] Kondo, M. *Hematopoietic Stem Cell Biology*; Springer, 2009.

[5] Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **2020**, *578*, 266–272.

[6] Helleday, T.; Eshtad, S.; Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature reviews genetics* **2014**, *15*, 585–598.

[7] Acuna-Hidalgo, R.; Bo, T.; Kwint, M. P.; Vorst, M. V. D.; Pinelli, M.; Veltman, J. A.; Hoischen, A.; Vissers, L. E.; Gilissen, C. Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *American journal of human genetics* **2015**, *97*, 67–74.

[8] Sehn, J. K. Insertions and Deletions (Indels). *Clinical Genomics* **2015**, 129–150.

[9] Salem, R. M.; Rodriguez-Murillo, L. Copy Number Variant (CNV). *Encyclopedia of Behavioral Medicine* **2013**, 500–501.

[10] Palacios, R.; Jáuregui, C. G.; Flores, M.; Palacios-Flores, K. Copy Number Variation. *Reference Module in Life Sciences* **2022**,

[11] Pös, O.; Radvanszky, J.; Buglyó, G.; Pös, Z.; Rusnakova, D.; Nagy, B.; Szemes, T. DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomedical Journal* **2021**, *44*, 548–559.

[12] Zhang, F.; Gu, W.; Hurles, M. E.; Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics* **2009**, *10*, 451–481.

[13] Hollox, E. J.; Zuccherato, L. W.; Tucci, S. Genome structural variation in human evolution. *Trends in Genetics* **2022**, *38*, 45–58.

[14] Cook, G. W. *et al.* Structural variation and its potential impact on genome instability: Novel discoveries in the EGFR landscape by long-read sequencing. *PLoS ONE* **2020**, *15*.

[15] Oh, J.; Lee, Y. D.; Wagers, A. J. Stem cell aging: mechanisms, regulators and therapeutic opportunities. *Nature medicine* **2014**, *20*, 870–880.

[16] Rossi, D. J.; Bryder, D.; Zahn, J. M.; Ahlenius, H.; Sonu, R.; Wagers, A. J.; Weissman, I. L. Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proceedings of the National Academy of Sciences* **2005**, *102*, 9194–9199.

[17] Warren, L. A.; Rossi, D. J. Stem cells and aging in the hematopoietic system. *Mechanisms of ageing and development* **2009**, *130*, 46–53.

[18] Rossi, D. J.; Jamieson, C. H.; Weissman, I. L. Stems Cells and the Pathways to Aging and Cancer. *Cell* **2008**, *132*, 681–696.

[19] Bailey, K.; Maslov, A.; Pruitt, S. Accumulation of mutations and somatic selection in aging neural stem/progenitor cells. *Aging Cell* **2004**, *3*, 391–397.

[20] Kennedy, S. R.; Loeb, L. A.; Herr, A. J. Somatic mutations in aging, cancer and neurodegeneration. *Mech. Ageing Dev.* **2012**, *133*, 118–126.

[21] Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **2018**, *359*, 555–559.

[22] Shen, H. B.; Li, J.; Yao, Y. S.; Yang, Z. H.; Zhou, Y. J.; Chen, W.; Hu, T. J. Impact of Somatic Mutations in Non-Small-Cell Lung Cancer: A Retrospective Study of a Chinese Cohort. *Cancer Management and Research* **2020**, *12*, 7427.

[23] Jia, P.; Zhao, Z. Impacts of somatic mutations on gene expression: an association perspective. *Briefings in Bioinformatics* **2017**, *18*, 413.

[24] Brunner, S. F.; Roberts, N. D.; Wylie, L. A.; Moore, L.; Aitken, S. J.; Davies, S. E.; Sanders, M. A.; Ellis, P.; Alder, C.; Hooks, Y.; Abascal, F.; Stratton, M. R.; Martincorena, I.; Hoare, M.; Campbell, P. J. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **2019**, *574*, 538–542.

[25] Vogelstein, B.; Papadopoulos, N.; Velculescu, V. E.; Zhou, S.; Diaz, L. A.; Kinzler, K. W. Cancer genome landscapes. *Science* **2013**, *339*, 1546–1558.

[26] Stratton, M. R.; Campbell, P. J.; Futreal, P. A. The cancer genome. *Nature* **2009**, *458*, 719–724.

[27] Stoler, D. L.; Chen, N.; Basik, M.; Kahlenberg, M. S.; Rodriguez-Bigas, M. A.; Petrelli, N. J.; Anderson, G. R. The onset and extent of genomic instability in sporadic colorectal tumor progression. *Proceedings of the National Academy of Sciences* **1999**, *96*, 15121–15126.

[28] Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **2015**, *348*, 880–886.

[29] Risques, R. A.; Kennedy, S. R. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet.* **2018**, *14*.

[30] Martincorena, I.; Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **2015**, *349*, 1483–1489.

[31] Xie, N.; Shen, G.; Gao, W.; Huang, Z.; Huang, C.; Fu, L. Neoantigens: promising targets for cancer therapy. *Signal Transduction and Targeted Therapy* **2023**, *8*, 9.

[32] Wang, S.; Xie, K.; Liu, T. Cancer immunotherapies: from efficacy to resistance mechanisms–not only checkpoint matters. *Frontiers in immunology* **2021**, *12*, 690112.

[33] Cannataro, V. L.; Mandell, J. D.; Townsend, J. P. Attribution of cancer origins to endogenous, exogenous, and preventable mutational processes. *Molecular Biology and Evolution* **2022**, *39*, msac084.

[34] Marnett, L. J.; Plastaras, J. P. Endogenous DNA damage and mutation. *Trends in Genetics* **2001**, *17*, 214–221.

[35] Bont, R. D.; van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **2004**, *19*, 169–185.

[36] Davidson, J. F.; Guo, H. H.; Loeb, L. A. Endogenous mutagenesis and cancer. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **2002**, *509*, 17–21.

[37] Strauss, B. S. Mechanisms of Mutation. *Genetic Diagnosis of Endocrine Disorders: Second Edition* **2016**, 3–18.

[38] Marian, A. J. Errors in DNA replication and genetic diseases. *Current opinion in cardiology* **2013**, *28*, 269–271.

[39] Kunkel, T. A.; Bebenek, K. DNA replication fidelity. *Annual review of biochemistry* **2000**, *69*, 497–529.

[40] Kunkel, T. A. Evolving Views of DNA Replication (In)Fidelity. *Cold Spring Harbor symposia on quantitative biology* **2009**, *74*, 91.

[41] McCulloch, S. D.; Kunkel, T. A. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell research* **2008**, *18*, 148.

[42] Shevelev, I. V.; Hübscher, U. The 3–5 exonucleases. *Nature Reviews Molecular Cell Biology 2002 3:5* **2002**, *3*, 364–376.

[43] Reha-Krantz, L. J. DNA polymerase proofreading: Multiple roles maintain genome stability. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2010**, *1804*, 1049–1063.

[44] Kolodner, R. D.; Marsischky, G. T. Eukaryotic DNA mismatch repair. *Current Opinion in Genetics Development* **1999**, *9*, 89–96.

[45] Li, G. M. Mechanisms and functions of DNA mismatch repair. *Cell Research 2008 18:1* **2007**, *18*, 85–98.

[46] Pray, L. DNA replication and causes of mutation. *Nature education* **2008**, *1*, 214.

[47] Drost, J.; Boxtel, R. V.; Blokzijl, F.; Mizutani, T.; Sasaki, N.; Sasselli, V.; Ligt, J. D.; Behjati, S.; Grolleman, J. E.; Wezel, T. V.; Nik-Zainal, S.; Kuiper, R. P.; Cuppen, E.; Clevers, H. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science (New York, N.Y.)* **2017**, *358*, 234–238.

[48] Barnes, D. E.; Lindahl, T. Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu. Rev. Genet.* **2004**, *38*, 445–476.

[49] Hsieh, P.; Yamane, K. DNA mismatch repair: Molecular mechanism, cancer, and ageing. *Mechanisms of ageing and development* **2008**, *129*, 391.

[50] Jiricny, J. The multifaceted mismatch-repair system. *Nature Reviews Molecular Cell Biology 2006 7:5* **2006**, *7*, 335–346.

[51] Lyer, R. R.; Pluciennik, A.; Burdett, V.; Modrich, P. L. DNA mismatch repair: Functions and mechanisms. *Chemical Reviews* **2006**, *106*, 302–323.

[52] Kunkel, T. A.; Erie, D. A. DNA mismatch repair. *Annu. Rev. Biochem.* **2005**, *74*, 681–710.

[53] Johnson, K. A. The kinetic and chemical mechanism of high-fidelity DNA polymerases. *Biochimica et biophysica acta* **2010**, *1804*, 1041.

[54] Yang, W. Structure and function of mismatch repair proteins. *Mutation Research - DNA Repair* **2000**, *460*, 245–256.

[55] Schaaper, R. M. Base selection, proofreading, and mismatch repair during DNA replication in Escherichia coli. *Journal of Biological Chemistry* **1993**, *268*, 23762–23765.

[56] Cooper, G. M. *DNA Repair*; Sinauer Associates, 2000.

[57] Chatterjee, N.; Walker, G. C. Mechanisms of DNA damage, repair and mutagenesis. *Environmental and molecular mutagenesis* **2017**, *58*, 235.

[58] Meira, L. B.; Burgis, N. E.; Samson, L. D. Base excision repair. *Advances in Experimental Medicine and Biology* **2005**, *570*, 125–173.

[59] Jeggo, P. A.; Pearl, L. H.; Carr, A. M. DNA repair, genome stability and cancer: a historical perspective. *Nature Reviews Cancer 2015 16:1* **2015**, *16*, 35–42.

[60] Fumagalli, M.; Rossiello, F.; Clerici, M.; Barozzi, S.; Cittaro, D.; Kaplunov, J. M.; Bucci, G.; Dobreva, M.; Matti, V.; Beausejour, C. M.; Herbig, U.; Longhese, M. P.; Fagagna, F. D. D. Telomeric DNA damage is irreparable and causes persistent DNA damage response activation. *Nature cell biology* **2012**, *14*, 355.

[61] Piraino, S. W.; Thomas, V.; O'Donovan, P.; Furney, S. J. Mutations: Driver Versus Passenger. *Encyclopedia of Cancer* **2019**, 551–562.

[62] Berben, L.; Floris, G.; Wildiers, H.; Hatse, S. Cancer and Aging: Two Tightly Interconnected Biological Processes. *Cancers* **2021**, *13*, 1–20.

[63] Mons, U.; Gredner, T.; Behrens, G.; Stock, C.; Brenner, H. Cancers Due to Smoking and High Alcohol Consumption: Estimation of the Attributable Cancer Burden in Germany. *Deutsches Ärzteblatt International* **2018**, *115*, 571.

[64] Li, X.; Yan, H.; Wu, J.; Zhang, L. Tobacco smoking associates with NF1 mutations exacerbating survival outcomes in gliomas. *Biomarker Research* **2022**, *10*, 1–3.

[65] Ernst, S. M.; Mankor, J. M.; Riet, J. V.; Thüsen, J. H. V. D.; Dubbink, H. J.; Aerts, J. G. J. V.; Langen, A. J. D.; Smit, E. F.; Dingemans, A.-M. C.; Monkhorst, K. Tobacco Smoking-Related Mutational Signatures in Classifying Smoking-Associated and Nonsmoking-Associated NSCLC. *Journal of Thoracic Oncology* **2023**, *18*, 487–498.

[66] Sun, L. Y.; Cen, W. J.; Tang, W. T.; Long, Y. K.; Yang, X. H.; Ji, X. M.; Yang, J. J.; Zhang, R. J.; Wang, F.; Shao, J. Y.; Du, Z. M. Smoking status

combined with tumor mutational burden as a prognosis predictor for combination immune checkpoint inhibitor therapy in non-small cell lung cancer. *Cancer Medicine* **2021**, *10*, 6610.

[67] Wang, X.; Ricciuti, B.; Nguyen, T.; Li, X.; Rabin, M. S.; Awad, M. M.; Lin, X.; Johnson, B. E.; Christiani, D. C. Association between Smoking History and Tumor Mutation Burden in Advanced Non-Small Cell Lung Cancer. *Cancer research* **2021**, *81*, 2566–2573.

[68] Ostroverkhova, D.; Przytycka, T. M.; Panchenko, A. R. Cancer driver mutations: predictions and reality. *Trends in Molecular Medicine* **2023**, *29*, 554–566.

[69] Pon, J. R.; Marra, M. A. Driver and Passenger Mutations in Cancer. *Annual Review of Pathology: Mechanisms of Disease* **2015**, *10*, 25–50.

[70] Kherreh, N.; Cleary, S.; Seoighe, C. No evidence that HLA genotype influences the driver mutations that occur in cancer patients. *Cancer Immunology, Immunotherapy* **2022**, *71*, 819.

[71] Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **2020**, *578*, 122–128.

[72] Lee, E. Y.; Muller, W. J. Oncogenes and Tumor Suppressor Genes. *Cold Spring Harbor Perspectives in Biology* **2010**, *2*.

[73] Botezatu, A.; Iancu, I. V.; Popa, O.; Plesa, A.; Manda, D.; Huica, I.; Vladoiu, S.; Anton, G.; Badiu, C. Mechanisms of oncogene activation. *New aspects in molecular and cellular mechanisms of human carcinogenesis* **2016**, *9*.

[74] Velez, A. M. A.; Howard, M. S. Tumor-suppressor Genes, Cell Cycle Regulatory Checkpoints, and the Skin. *North American Journal of Medical Sciences* **2015**, *7*, 176.

[75] Caldas, C.; Venkitaraman, A. Tumor Suppressor Genes. *Encyclopedia of Genetics* **2001**, 2081–2088.

[76] Levine, A. J.; Hu, W.; Feng, Z. Tumor suppressor genes. *The Molecular Basis of Cancer* **2008**, 31–38.

[77] Caldas, C.; Venkitaraman, A. R. Tumor Suppressor Genes. *Brenner's Encyclopedia of Genetics: Second Edition* **2013**, 232–237.

[78] Chatrath, A.; Ratan, A.; Dutta, A. Germline Variants That Affect Tumor Progression. *Trends in Genetics* **2021**, *37*, 433–443.

[79] Liede, A.; Karlan, B. Y.; Narod, S. A. Cancer risks for male carriers of germline mutations in BRCA1 or BRCA2: A review of the literature. *Journal of Clinical Oncology* **2004**, *22*, 735–742.

[80] Bhattacharya, P.; McHugh, T. W. Lynch Syndrome. *Encyclopedia of Gastroenterology, Second Edition* **2023**, 490–494.

[81] Barrow, E.; Hill, J.; Evans, D. G. Cancer risk in Lynch Syndrome. *Familial Cancer* **2013**, *12*, 229–240.

[82] Lynch, H. T.; De la Chapelle, A. Hereditary colorectal cancer. *New England Journal of Medicine* **2003**, *348*, 919–932.

[83] Greaves, M.; Maley, C. C. Clonal evolution in cancer. *Nature* **2012**, *481*, 306–313.

[84] Nowell, P. C. The Clonal Evolution of Tumor Cell Populations. *Science* **1976**, *194*, 23–28.

[85] Martincorena, I. Somatic mutation and clonal expansions in human tissues. *Genome medicine* **2019**, *11*, 1–3.

[86] Alberts, B. Molecular biology of the cell 4th edition. *(No Title)* **2002**,

[87] Hall, M. W.; Shorthouse, D.; Alcraft, R.; Jones, P. H.; Hall, B. A. Mutations observed in somatic evolution reveal underlying gene mechanisms. *Communications Biology* **2023**, *6*, 753.

[88] Waanders, E. *et al.* Mutational Landscape and Patterns of Clonal Evolution in Relapsed Pediatric Acute Lymphoblastic Leukemia. *Blood Cancer Discovery* **2020**, *1*, 96–111.

[89] Merhi, M. *et al.* Identification of Clonal Neoantigens Derived From Driver Mutations in an EGFR-Mutated Lung Cancer Patient Benefitting From Anti-PD-1. *Frontiers in Immunology | www.frontiersin.org* **2020**, *1*, 1366.

[90] Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **2021**, *184*, 2239–2254.e39.

[91] Ben-David, U.; Beroukhim, R.; Golub, T. R. Genomic Evolution of Cancer Models: Perils and Opportunities. *Nature reviews. Cancer* **2019**, *19*, 97.

[92] McGranahan, N.; Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **2017**, *168*, 613–628.

[93] Errico, A. Clonal and subclonal events in cancer evolution—optimizing cancer therapy. *Nature Reviews Clinical Oncology* **2015**, *12*, 372–372.

[94] Craig, D. J.; Bailey, M. M.; Noe, O. B.; Williams, K. K.; Stanbery, L.; Hamouda, D. M.; Nemunaitis, J. J. NC-ND license Subclonal landscape of cancer drives resistance to immune therapy. *Cancer Treatment and Research Communications* **2022**, *30*, 100507.

[95] Bhang, H. E. C. *et al.* Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nature Medicine* **2015**, *21*, 440–448.

[96] Schmitt, M. W.; Loeb, L. A.; Salk, J. J. The influence of subclonal resistance mutations on targeted cancer therapy. *Nature reviews. Clinical oncology* **2016**, *13*, 335.

[97] Cassidy, J. W.; Bruna, A. Tumor Heterogeneity. *Patient Derived Tumor Xenograft Models: Promise, Potential and Practice* **2017**, 37–55.

[98] Prasetyanti, P. R.; Medema, J. P. Intra-tumor heterogeneity from a cancer stem cell perspective. *Molecular cancer* **2017**, *16*, 1–9.

[99] McGranahan, N.; Furness, A. J.; Rosenthal, R.; Ramskov, S.; Lyngaa, R.; Saini, S. K.; Jamal-Hanjani, M.; Wilson, G. A.; Birkbak, N. J.; Hiley, C. T., *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **2016**, *351*, 1463–1469.

[100] Marusyk, A.; Almendro, V.; Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nature reviews cancer* **2012**, *12*, 323–334.

[101] Ramón y Cajal, S.; Sesé, M.; Capdevila, C.; Aasen, T.; De Mattos-Arruda, L.; Diaz-Cano, S. J.; Hernández-Losa, J.; Castellví, J. Clinical implications of intratumor heterogeneity: challenges and opportunities. *Journal of Molecular Medicine* **2020**, *98*, 161–177.

[102] Rosa, S. L.; Notohara, K.; Tibiletti, M. G.; Stanta, G.; Bonin, S. Article 85 1 Citation: Stanta G and Bonin S (2018) Overview on Clinical Relevance of Intra-Tumor Heterogeneity. *Front. Med* **2018**, *5*, 85.

[103] Koh, G.; Degasperi, A.; Zou, X.; Momen, S.; Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nature Reviews Cancer 2021 21:10* **2021**, *21*, 619–637.

[104] Koh, G.; Zou, X.; Nik-Zainal, S. Mutational signatures: experimental design and analytical framework. *Genome Biol.* **2020**, *21*.

[105] Hoeck, A. V.; Tjoonk, N. H.; van Boxtel, R.; Cuppen, E. Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer* **2019**, *19*, 457.

[106] Alexandrov, L. B.; Nik-Zainal, S.; Wedge, D. C.; Campbell, P. J.; Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **2013**, *3*, 246.

[107] Zhuravleva, E.; O'Rourke, C. J.; Andersen, J. B. Mutational signatures and processes in hepatobiliary cancers. *Nature Reviews Gastroenterology Hepatology* **2022**, *19*, 367–382.

[108] Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **2020**, *578*, 94–101.

[109] van den Heuvel, G. R.; Kroeze, L. I.; Ligtenberg, M. J.; Grünberg, K.; Jansen, E. A.; von Rhein, D.; de Voer, R. M.; van den Heuvel, M. M. Mutational signature analysis in non-small cell lung cancer patients with a high tumor mutational burden. *Respiratory Research* **2021**, *22*.

[110] Zou, X. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nature cancer* **2021**, *2*, 643.

[111] Gehring, J. S.; Fischer, B.; Lawrence, M.; Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **2015**, *31*, 3673–3675.

[112] Kucab, J. E.; Zou, X.; Morganella, S.; Joel, M.; Nanda, A. S.; Nagy, E.; Gomez, C.; Degasperi, A.; Harris, R.; Jackson, S. P.; Arlt, V. M.; Phillips, D. H.; Nik-Zainal, S. A Compendium of Mutational Signatures of Environmental Agents. *Cell* **2019**, *177*, 821.

[113] Steele, C. D.; Abbasi, A.; Islam, S. M. A.; Bowes, A. L.; Khandekar, A.; Haase, K.; Hames-Fathi, S.; Ajayi, D.; Verfaillie, A.; Dhami, P.; Mclatchie, A.; Lechner, M.; Light, N.; Shlien, A. Signatures of copy number alterations in human cancer. *Adrienne M. Flanagan* **2022**, *606*.

[114] Islam, S. A.; Díaz-Gay, M.; Wu, Y.; Barnes, M.; Vangara, R.; Bergstrom, E. N.; He, Y.; Vella, M.; Wang, J.; Teague, J. W., *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* **2022**, *2*.

[115] Kičiatovas, D.; Guo, Q.; Kailas, M.; Pesonen, H.; Corander, J.; Kaski, S.; Pitkänen, E.; Mustonen, V. Identification of multiplicatively acting modulatory mutational signatures in cancer. *BMC bioinformatics* **2022**, *23*, 522.

[116] Lyu, X.; Garret, J.; Rätsch, G.; Lehmann, K. V. Mutational signature learning with supervised negative binomial non-negative matrix factorization. *Bioinformatics* **2020**, *36*, i154–i160.

[117] Rosales, R. A.; Drummond, R. D.; Valieris, R.; Dias-Neto, E.; Silva, I. T. D. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **2017**, *33*, 8–16.

[118] Pelizzola, M.; Laursen, R.; Hobolth, A. Model selection and robust inference

of mutational signatures using Negative Binomial non-negative matrix factorization. *BMC bioinformatics* **2023**, *24*, 1–24.

[119] Sadegh, M.; Vrugt, J. A. Approximate Bayesian computation using markov chain Monte Carlo simulation: DREAM(ABC). *Water Resources Research* **2014**, *50*, 6767–6787.

[120] Wu, Y.; Chua, E. H. Z.; Ng, A. W. T.; Boot, A.; Rozen, S. G. Accuracy of mutational signature software on correlated signatures. *Scientific Reports* **2022**, *12*, 390.

[121] Pandey, P.; Arora, S.; Rosen, G. L. MetaMutationalSigs: comparison of mutational signature refitting results made easy. *Bioinformatics* **2022**, *38*, 2344–2347.

[122] Omichessan, H.; Severi, G.; Perduca, V. Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. *PLoS ONE* **2019**, *14*.

[123] Goldberg, M. E.; Harris, K. Mutational Signatures of Replication Timing and Epigenetic Modification Persist through the Global Divergence of Mutation Spectra across the Great Ape Phylogeny GBE. *Genome Biol. Evol* **2021**, *14*.

[124] Levatić, J.; Salvadores, M.; Fuster-Tormo, F.; Supek, F. Mutational signatures are markers of drug sensitivity of cancer cells. *Nature Communications 2022 13:1* **2022**, *13*, 1–19.

[125] Berglund, A. E.; Welsh, E. A.; Eschrich, S. A. Characteristics and Validation Techniques for PCA-Based Gene-Expression Signatures. *International Journal of Genomics* **2017**, *2017*.

[126] Baez-Ortega, A.; Gori, K. Computational approaches for discovery of mutational signatures in cancer. *Brief. Bioinforma.* **2019**, *20*, 77–88.

[127] Huang, P. J.; Chiu, L. Y.; Lee, C. C.; Yeh, Y. M.; Huang, K. Y.; Chiu, C. H.; Tang, P. mSignatureDB: a database for deciphering mutational signatures in human cancers. *Nucleic Acids Research* **2018**, *46*, D964.

[128] Hu, X.; Xu, Z.; De, S. Characteristics of mutational signatures of unknown etiology. *NAR Cancer* **2020**, *2*.

[129] Phillips, D. H. Mutational spectra and mutational signatures: Insights into cancer aetiology and mechanisms of DNA damage and repair. *DNA Repair* **2018**, *71*, 6.

[130] Cummings, A. L. *et al.* Mutational landscape influences immunotherapy outcomes among patients with non-small-cell lung cancer with human leukocyte antigen supertype B44. *Nature Cancer* **2020**, *1*, 1167–1175.

[131] Chen, Z.; Wen, W.; Cai, Q.; Long, J.; Wang, Y.; Lin, W.; Shu, X.-o.; Zheng, W.; Guo, X. From tobacco smoking to cancer mutational signature: a mediation analysis strategy to explore the role of epigenetic changes. *BMC cancer* **2020**, *20*, 1–11.

[132] Alexandrov, L. B.; Ju, Y. S.; Haase, K.; Loo, P. V.; Martincorena, I.; Nik-Zainal, S.; Totoki, Y.; Fujimoto, A.; Nakagawa, H.; Shibata, T.; Campbell, P. J.; Vineis, P.; Phillips, D. H.; Stratton, M. R. Mutational signatures associated with tobacco smoking in human cancer. *Science* **2016**, *354*, 618–622.

[133] Hegedűs, C.; Juhász, T.; Fidrus, E.; Janka, E. A.; Juhász, G.; Boros, G.; Paragh, G.; Uray, K.; Emri, G.; Éva Remenyik,; Bai, P. Cyclobutane pyrimidine dimers from UVB exposure induce a hypermetabolic state in keratinocytes via mitochondrial oxidative stress. *Redox Biology* **2021**, *38*.

[134] Kuzminov, A. Pyrimidine Dimers. *Brenner's Encyclopedia of Genetics: Second Edition* **2013**, 538–539.

[135] Rastogi, R. P.; Kumar, A.; Tyagi, M. B.; Sinha, R. P. Access to. *Research Journal of Nucleic Acids* **2010**, *2010*, 32.

[136] Chowell, D.; Krishna, C.; Pierini, F.; Makarov, V.; Rizvi, N. A.; Kuo, F.; Morris, L. G.; Riaz, N.; Lenz, T. L.; Chan, T. A. Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Nature Medicine* **2019**, *25*, 1715–1720.

[137] Chowell, D. *et al.* Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science (New York, N.Y.)* **2018**, *359*, 582.

[138] Petljak, M.; Maciejowski, J. Molecular Origins of APOBEC-Associated Mutations in Cancer. *DNA repair* **2020**, *94*, 102905.

[139] Harris, R. S. Molecular mechanism and clinical impact of APOBEC3B-catalyzed mutagenesis in breast cancer. *Breast Cancer Research : BCR* **2015**, *17*.

[140] Seplyarskiy, V. B.; Soldatov, R. A.; Popadin, K. Y.; Antonarakis, S. E.; Bazykin, G. A.; Nikolaev, S. I. APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Research* **2016**, *26*, 174.

[141] Kaufman, J. Evolution and immunity. *Immunology* **2010**, *130*, 459.

[142] Simon, A. K.; Hollander, G. A.; McMichael, A. Evolution of the immune system in humans from infancy to old age. *Proceedings of the Royal Society B: Biological Sciences* **2015**, *282*, 20143085.

[143] Flajnik, M. F.; Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics 2010 11:1* **2009**, *11*, 47–59.

[144] Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *The Adaptive Immune System*; Garland Science, 2002.

[145] Natoli, G.; Ostuni, R. Adaptation and memory in immune responses. *Nature Immunology 2019 20:7* **2019**, *20*, 783–792.

[146] Abolhassani, H. Immunologic Memory. *Encyclopedia of Infection and Immunity* **2022**, *1*, 221–229.

[147] Lu, L. L.; Suscovich, T. J.; Fortune, S. M.; Alter, G. Beyond binding: antibody effector functions in infectious diseases. *Nature Reviews Immunology* **2018**, *18*, 46–61.

[148] Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Helper T Cells and Lymphocyte Activation*; Garland Science, 2002.

[149] Pearce, E. L.; Pearce, E. J. Metabolic Pathways In Immune Cell Activation And Quiescence. *Immunity* **2013**, *38*, 633.

[150] Fleuren, G. J.; Gorter, A.; Kuppen, P. J. *Immune Surveillance*; Elsevier, 1998; pp 1243–1247.

[151] Swann, J. B.; Smyth, M. J. Immune surveillance of tumors. *The Journal of Clinical Investigation* **2007**, *117*, 1137–1146.

[152] Ribatti, D.; Ribatti,; Domenico, The concept of immune surveillance against tumors: The first theories. *Oncotarget* **2016**, *8*, 7175–7180.

[153] Mergener, S.; Peña-Llopis, S. A new perspective on immune evasion: escaping immune surveillance by inactivating tumor suppressors. *Signal Transduction and Targeted Therapy* **2022**, *7*, 1–2.

[154] Hewitt, E. W. The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology* **2003**, *110*, 163–169.

[155] Cano, R. L. E.; Lopera, H. D. E. *Autoimmunity: From Bench to Bedside [Internet]*; El Rosario University Press, 2013.

[156] Halle, S.; Halle, O.; Förster, R. Mechanisms and Dynamics of T Cell-Mediated Cytotoxicity In Vivo. *Trends in Immunology* **2017**, *38*, 432–443.

[157] Larsen, M. V.; Lundegaard, C.; Lamberth, K.; Buus, S.; Lund, O.; Nielsen, M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC bioinformatics* **2007**, *8*, 1–12.

[158] Rooney, M. S.; Shukla, S. A.; Wu, C. J.; Getz, G.; Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **2015**, *160*, 48–61.

[159] Maldonado, R. A.; von Andrian, U. H. How tolerogenic dendritic cells induce regulatory T cells. *Advances in immunology* **2010**, *108*, 111.

[160] Bayati, F.; Mohammadi, M.; Valadi, M.; Jamshidi, S.; Foma, A. M.; Sharif-Paghaleh, E. The therapeutic potential of regulatory T cells: challenges and opportunities. *Frontiers in immunology* **2021**, *11*, 585819.

[161] Kurosaki, T.; Kometani, K.; Ise, W. Memory B cells. *Nature Reviews Immunology* **2015**, *15*, 149–159.

[162] Rahman, M.; Bordoni, B. *StatPearls [Internet]*; StatPearls Publishing, 2022.

[163] Sconocchia, G.; Fabbri, M.; Pardo, J.; Paul, S.; Lal, G. The Molecular Mechanism of Natural Killer Cells Function and its importance in Cancer immunotherapy. **2017**, *8*, 1.

[164] Janeway Jr, C. A.; Travers, P.; Walport, M.; Shlomchik, M. J. *Immunobiology: The Immune System in Health and Disease. 5th edition*; Garland Science, 2001.

[165] Ratajczak, W.; Niedźwiedzka-Rystwej, P.; Tokarz-Deptuła, B.; DeptuŁa, W. Immunological memory cells. *Central-European Journal of Immunology* **2018**, *43*, 194.

[166] Pagès, F. *et al.* Effector Memory T Cells, Early Metastasis, and Survival in Colorectal Cancer. *New England Journal of Medicine* **2005**, *353*, 2654–2666.

[167] Barnaba, V. T Cell Memory in Infection, Cancer, and Autoimmunity. *Frontiers in Immunology* **2022**, *12*.

[168] Palm, A.-K. E.; Henry, C. Remembrance of things past: long-term B cell memory after infection and vaccination. *Frontiers in immunology* **2019**, *10*, 1787.

[169] Sun, J. C.; Beilke, J. N.; Lanier, L. L. Immune memory redefined: characterizing the longevity of natural killer cells. *Immunological reviews* **2010**, *236*, 83.

[170] Gotuzzo, E.; Yactayo, S.; Córdova, E. Review article: Efficacy and duration of immunity after yellow fever vaccination: Systematic review on the need for

a booster every 10 years. *American Journal of Tropical Medicine and Hygiene* **2013**, *89*, 434–444.

[171] Macallan, D. C.; Borghans, J. A.; Asquith, B. Human T Cell Memory: A Dynamic View. *Vaccines* **2017**, *5*.

[172] Ellebedy, A. H.; Jackson, K. J.; Kissick, H. T.; Nakaya, H. I.; Davis, C. W.; Roskin, K. M.; McElroy, A. K.; Oshansky, C. M.; Elbein, R.; Thomas, S., *et al.* Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nature immunology* **2016**, *17*, 1226–1234.

[173] Shah, H. B.; Smith, K.; Wren, J. D.; Webb, C. F.; Ballard, J. D.; Bourn, R. L.; James, J. A.; Lang, M. L. Insights from analysis of human antigen-specific memory B cell repertoires. *Frontiers in Immunology* **2019**, *9*, 3064.

[174] Waldman, A. D.; Fritz, J. M.; Lenardo, M. J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nature Reviews Immunology* **2020**, *20*, 651–668.

[175] Zhang, Y.; Zhang, Z. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. *Cellular Molecular Immunology* **2020**, *17*, 807–821.

[176] Hansen, T. H.; Bouvier, M. MHC class I antigen presentation: learning from viral evasion strategies. *Nature Reviews Immunology* **2009**, *9*, 503–513.

[177] Vesely, M. D.; Schreiber, R. D. Cancer Immunoediting: antigens, mechanisms and implications to cancer immunotherapy. *Annals of the New York Academy of Sciences* **2013**, *1284*, 1.

[178] Garcia-Lora, A.; Algarra, I.; Garrido, F. MHC class I antigens, immune surveillance, and tumor immune escape. *Journal of Cellular Physiology* **2003**, *195*, 346–355.

[179] Pishesha, N.; Harmand, T. J.; Ploegh, H. L. A guide to antigen processing and presentation. *Nature Reviews Immunology* **2022**, *22*, 751–764.

[180] May Jr, K. F.; Jinushi, M.; Dranoff, G. *Cancer Immunotherapy*; Elsevier, 2013; pp 101–113.

[181] Lafuente, E.; Reche, P. Prediction of MHC-peptide binding: a systematic and comprehensive overview. *Current pharmaceutical design* **2009**, *15*, 3209–3220.

[182] Agrawal, S.; Reemtsma, K.; Bagiella, E.; Oluwole, S. F.; Braunstein, N. S. Role of TAP-1 and/or TAP-2 antigen presentation defects in tumorigenicity of mouse melanoma. *Cellular Immunology* **2004**, *228*, 130–137.

[183] Blees, A.; Januliene, D.; Hofmann, T.; Koller, N.; Schmidt, C.; Trowitzsch, S.; Moeller, A.; Tampé, R. Structure of the human MHC-I peptide-loading complex. *Nature 2017 551:7681* **2017**, *551*, 525–528.

[184] Hua, Z.; Graham, T. R. *Madame Curie Bioscience Database [Internet]*; Landes Bioscience, 2013.

[185] Peters, B.; Nielsen, M.; Sette, A. T Cell Epitope Predictions. *Annual review of immunology* **2020**, *38*, 123–145.

[186] Tsurui, H.; Takahashi, T. Prediction of T-cell epitope. *Journal of pharmacological sciences* **2007**, *105*, 299–316.

[187] Taylor, B. C.; Balko, J. M. Mechanisms of MHC-I downregulation and role in immunotherapy response. *Frontiers in immunology* **2022**, *13*, 844866.

[188] Magee, C. N.; Boenisch, O.; Najafian, N. The role of costimulatory molecules in directing the functional differentiation of alloreactive T helper cells. *American Journal of Transplantation* **2012**, *12*, 2588–2600.

[189] Ruíz-Argüelles, A. The cascade of the immune response. *Current Therapeutic Research* **1996**, *57*, 8–13.

[190] Denizot, F.; Wilson, A.; Battye, F.; Berke, G.; Shortman, K. Clonal expansion of T cells: a cytotoxic T-cell response in vivo that involves precursor cell proliferation. *Proceedings of the National Academy of Sciences of the United States of America* **1986**, *83*, 6089.

[191] Peters, P. J.; Borst, J.; Oorschot, V.; Fukuda, M.; Krahenbuhl, O.; Tschopp, J.; Slot, J. W.; Geuze, H. J. Cytotoxic T lymphocyte granules are secretory lysosomes, containing both perforin and granzymes. *The Journal of Experimental Medicine* **1991**, *173*, 1099.

[192] Kikuchi, Y.; Tokita, S.; Hirama, T.; Kochin, V.; Nakatsugawa, M.; Shinkawa, T.; Hirohashi, Y.; Tsukahara, T.; Hata, F.; Takemasa, I.; Sato, N.; Kanaseki, T.; Torigoe, T. CD8+ T–cell immune surveillance against a tumor antigen encoded by the oncogenic long noncoding RNA PVT1. *Cancer Immunology Research* **2021**, *9*, 1342–1353.

[193] Prokhnevska, N. *et al.* CD8+ T cell activation in cancer comprises an initial activation phase in lymph nodes followed by effector differentiation within the tumor. *Immunity* **2023**, *56*, 107.

[194] Choo, S. Y. The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. *Yonsei Medical Journal* **2007**, *48*, 11–23.

[195] Kulski, J. K.; Shiina, T.; Dijkstra, J. M. Genomic diversity of the major histocompatibility complex in health and disease. 2019.

[196] Williams, T. M. Human Leukocyte Antigen Gene Polymorphism and the Histocompatibility Laboratory. *The Journal of molecular diagnostics : JMD* **2001**, *3*, 98–104.

[197] Truong, H. V.; Sgourakis, N. G. Dynamics of MHC-I molecules in the antigen processing and presentation pathway. *Current Opinion in Immunology* **2021**, *70*, 122–128.

[198] Radwan, J.; Babik, W.; Kaufman, J.; Lenz, T. L.; Winternitz, J. Advances in the Evolutionary Understanding of MHC Polymorphism. *Trends in Genetics* **2020**, *36*.

[199] Zhang, M.; He, H. Parasite-mediated selection of major histocompatibility complex variability in wild brandt's voles (Lasiopodomys brandtii) from Inner Mongolia, China. *BMC Evolutionary Biology* **2013**, *13*, 1–15.

[200] Radwan, J.; Kuduk, K.; Levy, E.; LeBas, N.; Babik, W. Parasite load and MHC diversity in undisturbed and agriculturally modified habitats of the ornate dragon lizard. *Molecular ecology* **2014**, *23*, 5966–5978.

[201] Manning, J.; Indrova, M.; Lubyova, B.; Pribylova, H.; Bieblova, J.; Hejnar, J.; Simova, J.; Jandlova, T.; Bubenik, J.; Reinis, M. Induction of MHC class I molecule cell surface expression and epigenetic activation of antigen-processing machinery components in a murine model for human papilloma virus 16-associated tumours. *Immunology* **2008**, *123*, 218.

[202] Rock, K. L.; Reits, E.; Neefjes, J. Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends in immunology* **2016**, *37*, 724.

[203] Mehta, A. M.; Jordanova, E. S.; Wezel, T. V.; Uh, H. W.; Corver, W. E.; Kwappenberg, K. M.; Verduijn, W.; Kenter, G. G.; Burg, S. H. V. D.; Fleuren, G. J. Genetic variation of antigen processing machinery components and association with cervical carcinoma. *Genes, Chromosomes and Cancer* **2007**, *46*, 577–586.

[204] Hanahan, D.; Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **2011**, *144*, 646–674.

[205] Dhatchinamoorthy, K.; Colbert, J. D.; Rock, K. L. Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation. *Frontiers in Immunology* **2021**, *12*, 636568.

[206] Jhunjhunwala, S.; Hammer, C.; Delamarre, L. Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion. *Nature Reviews Cancer* **2021**, *21*, 298–312.

[207] Lee, J. H. *et al.* Transcriptional downregulation of MHC class I and melanoma de- differentiation in resistance to PD-1 inhibition. *NatCo* **2020**, *11*, 1897.

[208] Zhao, Y.; Cao, Y.; Chen, Y.; Wu, L.; Hang, H.; Jiang, C.; Zhou, X. B2M gene expression shapes the immune landscape of lung adenocarcinoma and determines the response to immunotherapy. *Immunology* **2021**, *164*, 507–523.

[209] Campo, A. B. D.; Kyte, J. A.; Carretero, J.; Zinchencko, S.; Méndez, R.; González-Aseguinolaza, G.; Ruiz-Cabello, F.; Aamdal, S.; Gaudernack, G.; Garrido, F.; Aptsiauri, N. Immune escape of cancer cells with beta2-microglobulin loss over the course of metastatic melanoma. *International Journal of Cancer* **2014**, *134*, 102–113.

[210] Rouas-Freiss, N.; Gonçalves, R. M. B.; Menier, C.; Dausset, J.; Carosella, E. D. Direct evidence to support the role of HLA-G in protecting the fetus from maternal uterine natural killercytolysis. *Proceedings of the National Academy of Sciences of the United States of America* **1997**, *94*, 11520.

[211] Tran, E.; Robbins, P. F.; Lu, Y.-C.; Prickett, T. D.; Gartner, J. J.; Jia, L.; Pasetto, A.; Zheng, Z.; Ray, S.; Groh, E. M.; Kriley, I. R.; Rosenberg, S. A. T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *The New England journal of medicine* **2016**, *375*, 2255.

[212] McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **2017**, *171*, 1259.

[213] Ling, A.; Löfgren-Burström, A.; Larsson, P.; Li, X.; Wikberg, M. L.; Öberg, Å.; Stenling, R.; Edin, S.; Palmqvist, R. TAP1 down-regulation elicits immune escape and poor prognosis in colorectal cancer. *Oncoimmunology* **2017**, *6*, e1356143.

[214] Sharpe, A. H. Introduction to checkpoint inhibitors and cancer immunotherapy. *Immunological Reviews* **2017**, *276*, 5–8.

[215] He, X.; Xu, C. Immune checkpoint signaling and cancer immunotherapy. *Cell research* **2020**, *30*, 660–669.

[216] Seliger, B.; Massa, C.; Hargadon, K. M.; Mohamadzadeh, M.; De, L.; Cruz-Merino, L. The dark side of dendritic cells: development and exploitation of tolerogenic activity that favor tumor outgrowth and immune escape. **2013**,

[217] Maeda, T.; Hiraki, M.; Jin, C.; Rajabi, H.; Tagde, A.; Alam, M.; Bouillez, A.; Hu, X.; Suzuki, Y.; Miyo, M.; Hata, T.; Hinohara, K.; Kufe, D. MUC1-

C INDUCES PD-L1 AND IMMUNE EVASION IN TRIPLE-NEGATIVE BREAST CANCER. *Cancer research* **2018**, *78*, 205.

[218] Gonzalez, H.; Hagerling, C.; Werb, Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes Development* **2018**, *32*, 1267–1284.

[219] Ohue, Y.; Nishikawa, H. Regulatory T (Treg) cells in cancer: Can Treg cells be a new therapeutic target? *Cancer Science* **2019**, *110*, 2080.

[220] Gasparoto, T. H.; Malaspina, T. S. D. S.; Benevides, L.; Melo, E. J. F. D.; Costa, M. R. S. N.; Damante, J. H.; Ikoma, M. R. V.; Garlet, G. P.; Cavassani, K. A.; Silva, J. S. D.; Campanelli, A. P. Patients with oral squamous cell carcinoma are characterized by increased frequency of suppressive regulatory T cells in the blood and tumor microenvironment. *Cancer Immunology, Immunotherapy* **2010**, *59*, 819–828.

[221] Yokokawa, J.; Cereda, V.; Remondo, C.; Gulley, J. L.; Arlen, P. M.; Schlom, J.; Tsang, K. Y. Enhanced Functionality of CD4+CD25highFoxP3+ Regulatory T Cells in the Peripheral Blood of Patients with Prostate Cancer. *Clinical Cancer Research* **2008**, *14*, 1032–1040.

[222] Mattei, F.; Chouaib, S.; Roussy, G.; De, F. L.; Cruz-Merino, L.; Labani-Motlagh, A.; Ashja-Mahdavi, M.; Loskog, A. The Tumor Microenvironment: A Milieu Hindering and Obstructing Antitumor Immune Responses. *Frontiers in Immunology | www.frontiersin.org* **2020**, *1*, 940.

[223] Li, K.; Shi, H.; Zhang, B.; Ou, X.; Ma, Q.; Chen, Y.; Shu, P.; Li, D.; Wang, Y. Myeloid-derived suppressor cells as immunosuppressive regulators and therapeutic targets in cancer. *Signal Transduction and Targeted Therapy* **2021**, *6*, 362.

[224] Viola, J. P. B.; Flores-Borja, F.; Yang, Y.; Li, C.; Liu, T.; Dai, X.; Bazhin, A. V. Myeloid-Derived Suppressor Cells in Tumors: From Mechanisms to Antigen Specificity and Microenvironmental Regulation. *Frontiers in Immunology | www.frontiersin.org* **2020**, *1*, 1371.

[225] Tie, Y.; Tang, F.; quan Wei, Y.; wei Wei, X. Immunosuppressive cells in cancer: mechanisms and potential therapeutic targets. *Journal of Hematology Oncology 2022 15:1* **2022**, *15*, 1–33.

[226] Beatty, G. L.; Gladney, W. L. Immune escape mechanisms as a guide for cancer immunotherapy. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **2015**, *21*, 687–692.

[227] Pyke, R. M.; Thompson, W. K.; Salem, R. M.; Font-Burgada, J.; Zanetti, M.; Carter, H. Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell* **2018**, *175*, 416–428.e13.

[228] Marty, R.; Kaabinejadian, S.; Rossell, D.; Slifker, M. J.; van de Haar, J.; Engin, H. B.; de Prisco, N.; Ideker, T.; Hildebrand, W. H.; Font-Burgada, J.; Carter, H. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell* **2017**, *171*, 1272–1283.e15.

[229] Castro, A.; Ozturk, K.; Pyke, R. M.; Xian, S.; Zanetti, M.; Carter, H. Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes. *BMC medical genomics* **2019**, *12*, 107.

[230] den Eynden, J. V.; Jiménez-Sánchez, A.; Miller, M. L.; Larsson, E. Lack of detectable neoantigen depletion signals in the untreated cancer genome. *Nature Genetics* **2019**, *51*, 1741–1748.

[231] Gubin, M. M.; Vesely, M. D. Cancer Immunoediting in the Era of Immuno-oncology. *Clinical Cancer Research* **2022**, *28*, 3917–3928.

[232] Dunn, G. P.; Bruce, A. T.; Ikeda, H.; Old, L. J.; Schreiber, R. D. Cancer immunoediting: from immunosurveillance to tumor escape. *Nature Immunology* **2002**, *3*, 991–998.

[233] Dunn, G. P.; Old, L. J.; Schreiber, R. D. The three Es of cancer immunoediting. *Annual Review of Immunology* **2004**, *22*, 329–360.

[234] Claeys, A.; Luijts, T.; Marchal, K.; van den Eynden, J. Low immunogenicity of common cancer hot spot mutations resulting in false immunogenic selection signals. *PLoS Genetics* **2021**, *17*.

[235] den Eynden, J. V.; Larsson, E. Mutational Signatures Are Critical for Proper Estimation of Purifying Selection Pressures in Cancer Somatic Mutation Data When Using the dN/dS Metric. *Frontiers in Genetics* **2017**, *8*, 74.

[236] Hartmaier, R. J.; Albacker, L. A.; Chmielecki, J.; Bailey, M.; He, J.; Goldberg, M. E.; Ramkissoon, S.; Suh, J.; Elvin, J. A.; Chiacchia, S.; Frampton, G. M.; Ross, J. S.; Miller, V.; Stephens, P. J.; Lipson, D. High-throughput genomic profiling of adult solid tumors reveals novel insights into cancer pathogenesis. *Cancer Research* **2017**, *77*, 2464–2475.

[237] Micheel, C. M.; Sweeney, S. M.; LeNoue-Newton, M. L.; André, F.; Bedard, P. L.; Guinney, J.; Meijer, G. A.; Rollins, B. J.; Sawyers, C. L.; Schultz, N.; Shaw, K. R. M.; Velculescu, V. E.; Levy, M. A.; on behalf of the AACR Project GENIE Consortium, American Association for Cancer Research Project Genomics Evidence Neoplasia Information Exchange: From Inception to First Data Release and Beyond—Lessons Learned and Member Institutions' Perspectives. *JCO Clinical Cancer Informatics* **2018**, *2*, 1–14.

[238] Ellis, M. J.; Gillette, M.; Carr, S. A.; Paulovich, A. G.; Smith, R. D.; Rodland, K. K.; Townsend, R. R.; Kinsinger, C.; Mesri, M.; Rodriguez, H.; Liebler, D. C. Connecting Genomic Alterations to Cancer Biology with Proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer discovery* **2013**, *3*, 1108.

[239] Collins, A.; Project Team, T. C. G. A. The cancer genome atlas (TCGA) pilot project. *Cancer Research* **2007**, *67*, LB–247.

[240] Pan-cancer analysis of whole genomes. *Nature* **2020**, *578*, 82–93.

[241] Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **2010**, *464*, 993.

[242] Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **2019**, *47*, D941.

[243] Huang, A. Y.; Lee, E. A. Identification of Somatic Mutations From Bulk and Single-Cell Sequencing Data. *Frontiers in Aging* **2021**, *2*.

[244] Singh, R. R. Target Enrichment Approaches for Next-Generation Sequencing Applications in Oncology. *Diagnostics* **2022**, *12*.

[245] Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Medicine 2020 12:1* **2020**, *12*, 1–13.

[246] Franco, I.; Helgadottir, H. T.; Moggio, A.; Larsson, M.; Vrtačnik, P.; Johansson, A.; Norgren, N.; Lundin, P.; Mas-Ponte, D.; Nordström, J.; Lundgren, T.; Stenvinkel, P.; Wennberg, L.; Supek, F.; Eriksson, M. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. *Genome Biology* **2019**, *20*, 1–22.

[247] Lou, J. J.; Mirsadraei, L.; Sanchez, D. E.; Wilson, R. W.; Shabihkhani, M.; Lucey, G. M.; Wei, B.; Singer, E. J.; Mareninov, S.; Yong, W. H. A review of room temperature storage of biospecimen tissue and nucleic acids for anatomic pathology laboratories and biorepositories. *Clinical biochemistry* **2014**, *47*, 267.

[248] Auer, H. *et al.* The effects of frozen tissue storage conditions on the integrity of RNA and protein. *Biotechnic histochemistry : official publication of the Biological Stain Commission* **2014**, *89*, 518.

[249] Mager, S. R.; Oomen, M. H.; Morente, M. M.; Ratcliffe, C.; Knox, K.; Kerr, D. J.; Pezzella, F.; Riegman, P. H. Standard operating procedure for the collection of fresh frozen tissue samples. *European Journal of Cancer* **2007**, *43*, 828–834.

[250] Gao, X. H.; Li, J.; Gong, H. F.; Yu, G. Y.; Liu, P.; Hao, L. Q.; Liu, L. J.; Bai, C. G.; Zhang, W. Comparison of Fresh Frozen Tissue With Formalin-Fixed Paraffin-Embedded Tissue for Mutation Analysis Using a Multi-Gene Panel in Patients With Colorectal Cancer. *Frontiers in Oncology* **2020**, *10*.

[251] Robbe, P. *et al.* Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. *Genetics in medicine : official journal of the American College of Medical Genetics* **2018**, *20*, 1196.

[252] Gross, A. M.; Kreisberg, J. F.; Ideker, T. Analysis of Matched Tumor and Normal Profiles Reveals Common Transcriptional and Epigenetic Signals Shared across Cancer Types. *PLOS ONE* **2015**, *10*, e0142618.

[253] Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120.

[254] Jung, Y.; Han, D. BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics* **2022**, *38*, 2404–2413.

[255] der Auwera, G. V.; O'Connor, B.; an O'Reilly Media Company. Safari, Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. *Genomics in the Cloud* **2020**, 300.

[256] Mose, L. E.; Wilkerson, M. D.; Hayes, D. N.; Perou, C. M.; Parker, J. S. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics* **2014**, *30*, 2813–2815.

[257] Ellrott, K.; Bailey, M. H.; Saksena, G.; Covington, K. R.; Kandoth, C.; Stewart, C.; Hess, J.; Ma, S.; Chiotti, K. E.; McLellan, M., *et al.* Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell systems* **2018**, *6*, 271–281.

[258] McLaren, W.; Gil, L.; Hunt, S. E.; Riat, H. S.; Ritchie, G. R.; Thormann, A.; Flicek, P.; Cunningham, F. The Ensembl Variant Effect Predictor. *Genome Biology* **2016**, *17*.

[259] Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **2010**, *38*, e164.

[260] Flower, D. R.; Phadwal, K.; Macdonald, I. K.; Coveney, P. V.; Davies, M. N.; Wan, S. T-cell epitope prediction and immune complex simulation using molecular dynamics: state of the art and persisting challenges. *Immunome Research* **2010**, *6*, 1–18.

[261] Burrows, S. R.; Rossjohn, J.; McCluskey, J. Have we cut ourselves too short in mapping CTL epitopes? *Trends in Immunology* **2006**, *27*, 11–16.

[262] Jørgensen, K. W.; Rasmussen, M.; Buus, S.; Nielsen, M. NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology* **2014**, *141*, 18–26.

[263] Jespersen, M. C.; Peters, B.; Nielsen, M.; Marcatili, P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research* **2017**, *45*, W24.

[264] Jurtz, V.; Paul, S.; Andreatta, M.; Marcatili, P.; Peters, B.; Nielsen, M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of Immunology (Baltimore, Md.: 1950)* **2017**, *199*, 3360–3368.

[265] Nielsen, M.; Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Medicine* **2016**, *8*, 33.

[266] Ardito, M.; De Groot, A.; McMurry, J.; Moise, L.; Yang, W.; Martin, B. EpiMatrix: tool for accelerated epitope selection and vaccine design. VACCINE. 2008.

[267] Vita, R.; Mahajan, S.; Overton, J. A.; Dhanda, S. K.; Martini, S.; Cantrell, J. R.; Wheeler, D. K.; Sette, A.; Peters, B. The immune epitope database (IEDB): 2018 update. *Nucleic acids research* **2019**, *47*, D339–D343.

[268] Saxová, P.; Buus, S.; Brunak, S.; Keşmir, C. Predicting proteasomal cleavage sites: a comparison of available methods. *International Immunology* **2003**, *15*, 781–787.

[269] Lata, S.; Bhasin, M.; Raghava, G. P. Application of machine learning techniques in predicting MHC binders. *Methods in molecular biology (Clifton, N.J.)* **2007**, *409*, 201–215.

[270] Guan, P.; Hattotuwagama, C. K.; Doytchinova, I. A.; Flower, D. R. MHCPred 2.0: an updated quantitative T-cell epitope prediction server. *Applied bioinformatics* **2006**, *5*, 55–61.

[271] Bhasin, M.; Raghava, G. P. Prediction of promiscuous and high-affinity mutated MHC binders. *Hybridoma and hybridomics* **2003**, *22*, 229–234.

[272] Singh, H.; Raghava, G. P. ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics* **2003**, *19*, 1009–1014.

[273] Rammensee, H. G.; Bachmann, J.; Emmerich, N. P. N.; Bachor, O. A.; Stevanović, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **1999**, *50*, 213–219.

[274] Bhasin, M.; Lata, S.; Raghava, G. P. TAPPred prediction of TAP-binding peptides in antigens. *Methods in molecular biology (Clifton, N.J.)* **2007**, *409*, 381–386.

[275] Reche, P. A.; Glutting, J. P.; Zhang, H.; Reinherz, E. L. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics* **2004**, *56*, 405–419.

[276] Doytchinova, I. A.; Guan, P.; Flower, D. R. EpiJen: a server for multistep T cell epitope prediction. *BMC bioinformatics* **2006**, *7*.

[277] Kim, J. Y.; Bang, H.; Noh, S.-J.; Choi, J. K. DeepNeo: a webserver for predicting immunogenic neoantigens. *Nucleic Acids Research* **2023**, gkad275.

[278] Wang, S.; Xie, K.; Liu, T. Cancer Immunotherapies: From Efficacy to Resistance Mechanisms – Not Only Checkpoint Matters. *Frontiers in Immunology* **2021**, *12*.

[279] Conlon, K. C.; Miljkovic, M. D.; Waldmann, T. A. Cytokines in the Treatment of Cancer. *Journal of Interferon Cytokine Research* **2019**, *39*, 6.

[280] Ribas, A.; Wolchok, J. D. Cancer Immunotherapy Using Checkpoint Blockade. *Science (New York, N.Y.)* **2018**, *359*, 1350.

[281] Castro, A.; Carter, H. Mutagenic exposures shape immunotherapy responses. *Nature Cancer* **2020**, *1*, 1132–1133.

[282] Akhtar, M.; Rashid, S.; Al-Bozom, I. A. PDL1 immunostaining: what pathologists need to know. *Diagnostic Pathology* **2021**, *16*.

[283] Yang, F.; Wang, J. F.; Wang, Y.; Liu, B.; Molina, J. R. Comparative Analysis of Predictive Biomarkers for PD-1/PD-L1 Inhibitors in Cancers: Developments and Challenges. *Cancers* **2022**, *14*, 109.

[284] Wang, Y.; Tong, Z.; Zhang, W.; Zhang, W.; Buzdin, A.; Mu, X.; Yan, Q.; Zhao, X.; Chang, H. H.; Duhon, M.; Zhou, X.; Zhao, G.; Chen, H.; Li, X. FDA-Approved and Emerging Next Generation Predictive Biomarkers for Immune Checkpoint Inhibitors in Cancer Patients. *Frontiers in Oncology* **2021**, *11*.

[285] Gascón, M.; Isla, D.; Cruellas, M.; Gálvez, E. M.; Lastra, R.; Ocáriz, M.; Paño, J. R.; Ramírez, A.; Sesma, A.; Torres-Ramón, I., *et al.* Intratumoral versus circulating lymphoid cells as predictive biomarkers in lung cancer patients treated with immune checkpoint inhibitors: Is the easiest path the best one? *Cells* **2020**, *9*, 1525.

[286] Wang, Y.; Tong, Z.; Zhang, W.; Zhang, W.; Buzdin, A.; Mu, X.; Yan, Q.; Zhao, X.; Chang, H.-H.; Duhon, M., *et al.* FDA-approved and emerging next generation predictive biomarkers for immune checkpoint inhibitors in cancer patients. *Frontiers in oncology* **2021**, *11*, 683419.

[287] Roberts, S. A.; Gordenin, D. A. Hypermutation in human cancer genomes: footprints and mechanisms. *Nature reviews. Cancer* **2014**, *14*, 786–800.

[288] Meléndez, B.; Campenhout, C. V.; Rorive, S.; Remmelink, M.; Salmon, I.; D'Haene, N. Methods of measurement for tumor mutational burden in tumor tissue. *Translational Lung Cancer Research* **2018**, *7*, 661.

[289] Song, X.; Zheng, Q.; Zhang, R.; Wang, M.; Deng, W.; Wang, Q.; Guo, W.; Li, T.; Ma, X. Potential Biomarkers for Predicting Depression in Diabetes Mellitus. *Frontiers in Psychiatry* **2021**, *12*, 731220.

[290] Chabanon, R. M.; Pedrero, M.; Lefebvre, C.; Marabelle, A.; Soria, J. C.; Postel-Vinay, S. Mutational landscape and sensitivity to immune checkpoint blockers. *Clinical Cancer Research* **2016**, *22*, 4309–4321.

[291] Klempner, S. J.; Fabrizio, D.; Bane, S.; Reinhart, M.; Peoples, T.; Ali, S. M.; Sokol, E. S.; Frampton, G.; Schrock, A. B.; Anhorn, R.; Reddy, P. Tumor Mu-

tational Burden as a Predictive Biomarker for Response to Immune Checkpoint Inhibitors: A Review of Current Evidence. *The Oncologist* **2020**, *25*, e147.

[292] Gejman, R. S.; Chang, A. Y.; Jones, H. F.; Dikun, K.; Hakimi, A. A.; Schietinger, A.; Scheinberg, D. A. Rejection of immunogenic tumor clones is limited by clonal fraction. *eLife* **2018**, *7*.

[293] Rosenthal, R. *et al.* Neoantigen directed immune escape in lung cancer evolution. *Nature* **2019**, *567*, 479.

[294] Łuksza, M. *et al.* Neoantigen quality predicts immunoediting in survivors of pancreatic cancer. *Nature* **2022**, *606*, 389–395.

[295] Levink, I. J.; Brosens, L. A.; Rensen, S. S.; Aberle, M. R.; Damink, S. S. O.; Cahen, D. L.; Buschow, S. I.; Fuhler, G. M.; Peppelenbosch, M. P.; Bruno, M. J. Neoantigen Quantity and Quality in Relation to Pancreatic Cancer Survival. *Frontiers in Medicine* **2022**, *8*.

[296] McGranahan, N.; Swanton, C. Neoantigen quality, not quantity. *Science Translational Medicine* **2019**, *11*.

[297] Wolf, Y.; Sameuls, Y. Neoantigens in cancer immunotherapy: quantity vs. quality. *Molecular Oncology* **2023**, *17*, 1457–1459.

[298] Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *The Lancet Oncology* **2017**, *18*, 1009–1021.

[299] Neefjes, J.; Jongsma, M. L. M.; Paul, P.; Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature Reviews Immunology* **2011**, *11*, 823–836.

[300] Roche, P. A.; Furuta, K. The ins and outs of MHC class II-mediated antigen processing and presentation. *Nature Reviews Immunology* **2015**, *15*, 203–216.

[301] Borroni, E. M.; Grizzi, F. Cancer Immunoediting and beyond in 2021. *International Journal of Molecular Sciences* **2021**, *22*, 13275.

[302] Kim, R.; Emi, M.; Tanabe, K. Cancer immunoediting from immune surveillance to immune escape. *Immunology* **2007**, *121*, 1–14.

[303] Garrido, F.; Aptsiauri, N. Cancer immune escape: MHC expression in primary tumours versus metastases. *Immunology* **2019**, *158*, 255–266.

[304] Tavakoli, F.; Sartakhti, J. S.; Manshaei, M. H.; Basanta, D. Cancer immunoediting: A game theoretical approach. *In Silico Biology* **2023**, *14*, 1–12.

[305] Mittal, D.; Gubin, M. M.; Schreiber, R. D.; Smyth, M. J. New insights into cancer immunoediting and its three component phases — elimination, equilibrium and escape. *Current opinion in immunology* **2014**, *27*, 16–25.

[306] O'Donnell, J. S.; Teng, M. W. L.; Smyth, M. J. Cancer immunoediting and resistance to T cell-based immunotherapy. *Nature Reviews Clinical Oncology* **2019**, *16*, 151–167.

[307] Schreiber, R. D.; Old, L. J.; Smyth, M. J. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science (New York, N.Y.)* **2011**, *331*, 1565–1570.

[308] Brown, S. D.; Warren, R. L.; Gibb, E. A.; Martin, S. D.; Spinelli, J. J.; Nelson, B. H.; Holt, R. A. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Research* **2014**, *24*, 743–750.

[309] Davoli, T.; Uno, H.; Wooten, E. C.; Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science (New York, N.Y.)* **2017**, *355*, eaaf8399.

[310] Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nature biotechnology* **2015**, *33*, 1152.

[311] Lakatos, E.; Williams, M. J.; Schenck, R. O.; Cross, W. C.; Househam, J.; Zapata, L.; Werner, B.; Gatenbee, C.; Robertson-Tessi, M.; Barnes, C. P.; Anderson, A. R.; Sottoriva, A.; Graham, T. A. Evolutionary dynamics of neoantigens in growing tumors. *Nature genetics* **2020**, *52*, 1057.

[312] Waks, Z.; Weissbrod, O.; Carmeli, B.; Norel, R.; Utro, F.; Goldschmidt, Y. Driver gene classification reveals a substantial overrepresentation of tumor suppressors among very large chromatin-regulating proteins. *Scientific Reports* **2016**, *6*, 38988.

[313] Chowell, D.; Krishna, S.; Becker, P. D.; Cocita, C.; Shu, J.; Tan, X.; Greenberg, P. D.; Klavinskis, L. S.; Blattman, J. N.; Anderson, K. S. TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*, E1754–1762.

[314] Huang, L.; Kuhls, M. C.; Eisenlohr, L. C. Hydrophobicity as a driver of MHC class I antigen processing. *The EMBO journal* **2011**, *30*, 1634–1644.

[315] Beasley, R. P. Hepatitis B virus. The major etiology of hepatocellular carcinoma. *Cancer* **1988**, *61*, 1942–1956.

[316] Wallin, K. L.; Wiklund, F.; Angström, T.; Bergman, F.; Stendahl, U.; Wadell, G.; Hallmans, G.; Dillner, J. Type-specific persistence of human papillomavirus DNA before the development of invasive cervical cancer. *The New England Journal of Medicine* **1999**, *341*, 1633–1638.

[317] zur Hausen, H. Viruses in human cancers. *Science (New York, N.Y.)* **1991**, *254*, 1167–1173.

[318] Network, C. G. A. R.; Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R. M.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J. M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **2013**, *45*, 1113–1120.

[319] Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *The New England Journal of Medicine* **2014**, *371*, 2189–2199.

[320] Allen, E. M. V. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science (New York, N.Y.)* **2015**, *350*, 207–211.

[321] Zapata, L.; Pich, O.; Serrano, L.; Kondrashov, F. A.; Ossowski, S.; Schaefer, M. H. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biology* **2018**, *19*, 67.

[322] Yang, F.; Kim, D.-K.; Nakagawa, H.; Hayashi, S.; Imoto, S.; Stein, L.; Roth, F. P. Quantifying immune-based counterselection of somatic mutations. *PLoS genetics* **2019**, *15*, e1008227.

[323] Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science (New York, N.Y.)* **2015**, *348*, 124–128.

[324] Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **2013**, *499*, 214–218.

[325] Martincorena, I.; Raine, K. M.; Gerstung, M.; Dawson, K. J.; Haase, K.; Loo, P. V.; Davies, H.; Stratton, M. R.; Campbell, P. J. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **2017**, *171*, 1029–1041.e21.

[326] Dilthey, A. T.; Moutsianas, L.; Leslie, S.; McVean, G. HLA*IMP–an integrated framework for imputing classical HLA alleles from SNP genotypes. *Bioinformatics (Oxford, England)* **2011**, *27*, 968–972.

[327] McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **2016**, *351*, 1463–1469.

[328] Rajasagi, M. *et al.* Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* **2014**, *124*, 453–462.

[329] Robinson, J.; Halliwell, J. A.; Hayhurst, J. D.; Flicek, P.; Parham, P.; Marsh, S. G. E. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research* **2015**, *43*, 423–431.

[330] Blokzijl, F.; Janssen, R.; van Boxtel, R.; Cuppen, E. MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. *Genome Medicine* **2018**, *10*.

[331] Baez-Ortega, A.; Gori, K. Computational approaches for discovery of mutational signatures in cancer. *Briefings in bioinformatics* **2019**, *20*, 77–88.

[332] Koh, G.; Zou, X.; Nik-Zainal, S. Mutational signatures: experimental design and analytical framework. *Genome biology* **2020**, *21*, 1–13.

[333] Miao, D. *et al.* Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nature Genetics* **2018**, *50*, 1271–1281.

[334] Chhibber, A.; Huang, L.; Zhang, H.; Shaw, P. M.; Hellmann, M. D.; Snyder, A. Germline HLA landscape does not predict efficacy of pembrolizumab monotherapy across solid tumor types. *Immunity* **2022**, *55*, 56–64.e4.

[335] Wang, M.; Claesson, M. H. Classification of human leukocyte antigen (HLA) supertypes. *Methods in Molecular Biology (Clifton, N.J.)* **2014**, *1184*, 309–317.

[336] Buhler, S.; Nunes, J. M.; Sanchez-Mazas, A. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics* **2016**, *68*, 401–416.

[337] Zemmour, J.; Parham, P. HLA Class I nucleotide sequences, 1992. *Tissue Antigens* **1992**, *40*, 221–228.

[338] González-Galarza, F. F.; Takeshita, L. Y.; Santos, E. J.; Kempson, F.; Maia, M. H. T.; Silva, A. L. S. D.; Silva, A. L. T. E.; Ghattaoraya, G. S.; Alfirevic, A.; Jones, A. R.; Middleton, D. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Research* **2015**, *43*, D784.

[339] Gui, Y.; Guo, G.; Huang, Y.; Hu, X.; Tang, A.; Gao, S.; Wu, R.; Chen, C.; Li, X.; Zhou, L., *et al.* Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nature genetics* **2011**, *43*, 875–878.

[340] Ward, J. P.; Gubin, M. M.; Schreiber, R. D. The Role of Neoantigens in Naturally Occurring and Therapeutically Induced Immune Responses to Cancer. *Advances in immunology* **2016**, *130*, 25.

[341] Wang, P.; Chen, Y.; Wang, C. Beyond tumor mutation burden: tumor neoantigen burden as a biomarker for immunotherapy and other types of therapy. *Frontiers in oncology* **2021**, *11*, 672677.

[342] Jardim, D. L.; Goodman, A.; de Melo Gagliato, D.; Kurzrock, R. The Challenges of Tumor Mutational Burden as an Immunotherapy Biomarker. *Cancer cell* **2021**, *39*, 154.

[343] McGrail, D. J. *et al.* High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *Annals of Oncology* **2021**, *32*, 661–672.

[344] Anagnostou, V. *et al.* Multimodal genomic features predict outcome of immune checkpoint blockade in non-small-cell lung cancer. *Nature cancer* **2020**, *1*, 99–111.

[345] Sidney, J.; Peters, B.; Frahm, N.; Brander, C.; Sette, A. HLA class I supertypes: A revised and updated classification. *BMC Immunology* **2008**, *9*, 1–15.

[346] Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell systems* **2018**, *6*, 271.

[347] Ribatti, D. The concept of immune surveillance against tumors: The first theories. *Oncotarget* **2017**, *8*, 7175.

[348] Swann, J. B.; Smyth, M. J., *et al.* Immune surveillance of tumors. *The Journal of clinical investigation* **2007**, *117*, 1137–1146.

[349] Ehrlich, P. Partial cell functions. *Nobel Lecture* **1908**, *11*.

[350] Burnet, M. Cancer—A Biological Approach. *British Medical Journal* **1957**, *1*, 779–786.

[351] Stutman, O. Tumor development after 3-methylcholanthrene in immunologically deficient athymic-nude mice. *Science (New York, N.Y.)* **1974**, *183*, 534–536.

[352] Stutman, O. Delayed tumour appearance and absence of regression in nude mice infected with murine sarcoma virus. *Nature* **1975**, *253*, 142–144.

[353] Old, L. J.; Boyse, E. A. Immunology of experimental tumors. *Annual review of medicine* **1964**, *15*, 167–186.

[354] Klein, G. Tumor antigens. *Annual Reviews in Microbiology* **1966**, *20*, 223–252.

[355] Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nature medicine* **2018**, *24*, 541.

[356] Shankaran, V.; Ikeda, H.; Bruce, A. T.; White, J. M.; Swanson, P. E.; Old, L. J.; Schreiber, R. D. IFNgamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature* **2001**, *410*, 1107–1111.

[357] Teng, M. W.; Galon, J.; Fridman, W. H.; Smyth, M. J. From mice to humans: developments in cancer immunoediting. *The Journal of Clinical Investigation* **2015**, *125*, 3338.

[358] Wu, T.; Wang, G.; Wang, X.; Wang, S.; Zhao, X.; Wu, C.; Ning, W.; Tao, Z.; Chen, F.; Liu, X.-S. Pan-cancer quantification of neoantigen-mediated immunoediting in cancer evolution. *bioRxiv* **2022**, 2022.04.08.487711.

[359] Greaves, M.; Maley, C. C. CLONAL EVOLUTION IN CANCER. *Nature* **2012**, *481*, 306.

[360] Jamal-Hanjani, M.; Wilson, G. A.; McGranahan, N.; Birkbak, N. J.; Watkins, T. B.; Veeriah, S.; Shafi, S.; Johnson, D. H.; Mitter, R.; Rosenthal, R., *et al.* Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine* **2017**, *376*, 2109–2121.

[361] Chujoh, Y.; Sobao, Y.; Miwa, K.; Kaneko, Y.; Takiguchi, M. The role of anchor residues in the binding of peptides to HLA-A*1101 molecules. *Tissue Antigens* **1998**, *52*, 501–509.

[362] den Eynden, J. V.; Basu, S.; Larsson, E. Somatic Mutation Patterns in Hemizygous Genomic Regions Unveil Purifying Selection during Tumor Evolution. *PLOS Genetics* **2016**, *12*, e1006506.

[363] Weghorn, D.; Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nature Genetics* **2017**, *49*, 1785–1788.

[364] Houghton, A. N.; Guevara-Patiño, J. A., *et al.* Immune recognition of self in immunity against cancer. *The Journal of clinical investigation* **2004**, *114*, 468–471.

[365] Wells, D. K.; van Buuren, M. M.; Dang, K. K.; Hubbard-Lucey, V. M.; Sheehan, K. C.; Campbell, K. M.; Lamb, A.; Ward, J. P.; Sidney, J.; Blazquez, A. B., *et al.* Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* **2020**, *183*, 818–834.

[366] Marty, R.; Kaabinejadian, S.; Rossell, D.; Slifker, M. J.; van de Haar, J.; Engin, H. B.; de Prisco, N.; Ideker, T.; Hildebrand, W. H.; Font-Burgada, J.; Carter, H. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell* **2017**, *171*, 1272.

[367] Dentro, S. C.; Wedge, D. C.; Van Loo, P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harbor perspectives in medicine* **2017**, *7*.

[368] Clopper, C. J.; Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **1934**, *26*, 404–413.

[369] Alfieri, F.; Caravagna, G.; Schaefer, M. H. Cancer genomes tolerate deleterious coding mutations through somatic copy number amplifications of wild-type regions. *Nature Communications* **2023**, *14*, 3594.

[370] Papac, R. J. Spontaneous regression of cancer. *Cancer Treatment Reviews* **1996**, *22*, 395–423.

6 Appendix

Bin	N (No Mutation)	N (Mutation)
3	4008601	1311
4	1550068	676
5	907929	495
6	541030	354
7	357591	273
8	348384	304
9	100837	99
10	174154	190
11	146640	176
12	100804	132
13	100793	143
14	137430	210
15	119093	195
16	64120	112
17	45795	85
18	36632	72
19	64099	133
20	36624	80
21	64085	147
22	27462	66
23	36612	92
24	27456	72
25	9151	25
26	18300	52
29	18294	58
31	9145	31
32	18288	64
34	9142	34

35	9141	35
36	9140	36
37	9139	37
39	18274	78
40	18272	80
45	9131	45
50	18252	100
62	9114	62
69	18214	138
79	9097	79
80	9096	80
88	9088	88
98	9078	98
107	18138	214
115	9061	115
137	9039	137
139	9037	139
170	9006	170
188	8988	188
215	8961	215
217	8959	217
384	8972	384
561	8615	561

Table 2: Number of mutations corresponding to each bin in Figure 2.2C

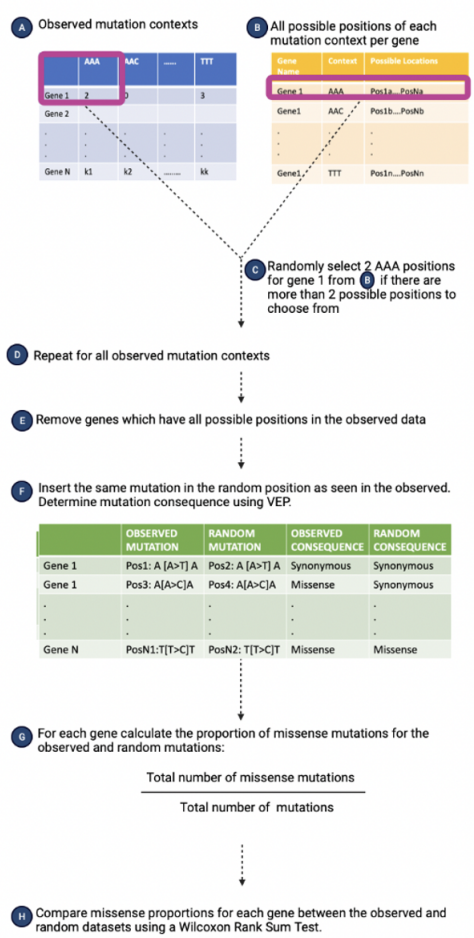

Figure 6.1: Workflow describing the creation of a random dataset of mutations with the same mutational context as observed mutations. (A) Mutation contexts are assigned to each observed mutation, and the total number of observed mutations of each context type is counted for each gene. (B) A list of all possible positions that could be mutated for each context for each gene. (C and D) The exact number of positions as observed in the real dataset were randomly sampled from the list of all possible positions for that context in that gene. (E) The gene was removed from the analysis if all possible positions were present in the observed data. (F) The random position was mutated to the same allele as in the observed data, and the variant consequence was annotated using VEP online tool. (G) The proportion of missense mutations for each gene was calculated for the observed and random datasets. (H) A paired Wilcoxon rank sum test was performed to assess whether the two datasets varied. This simulation and corresponding figures 6.2,6.3,6.4 were contributed by Siobhan Cleary.

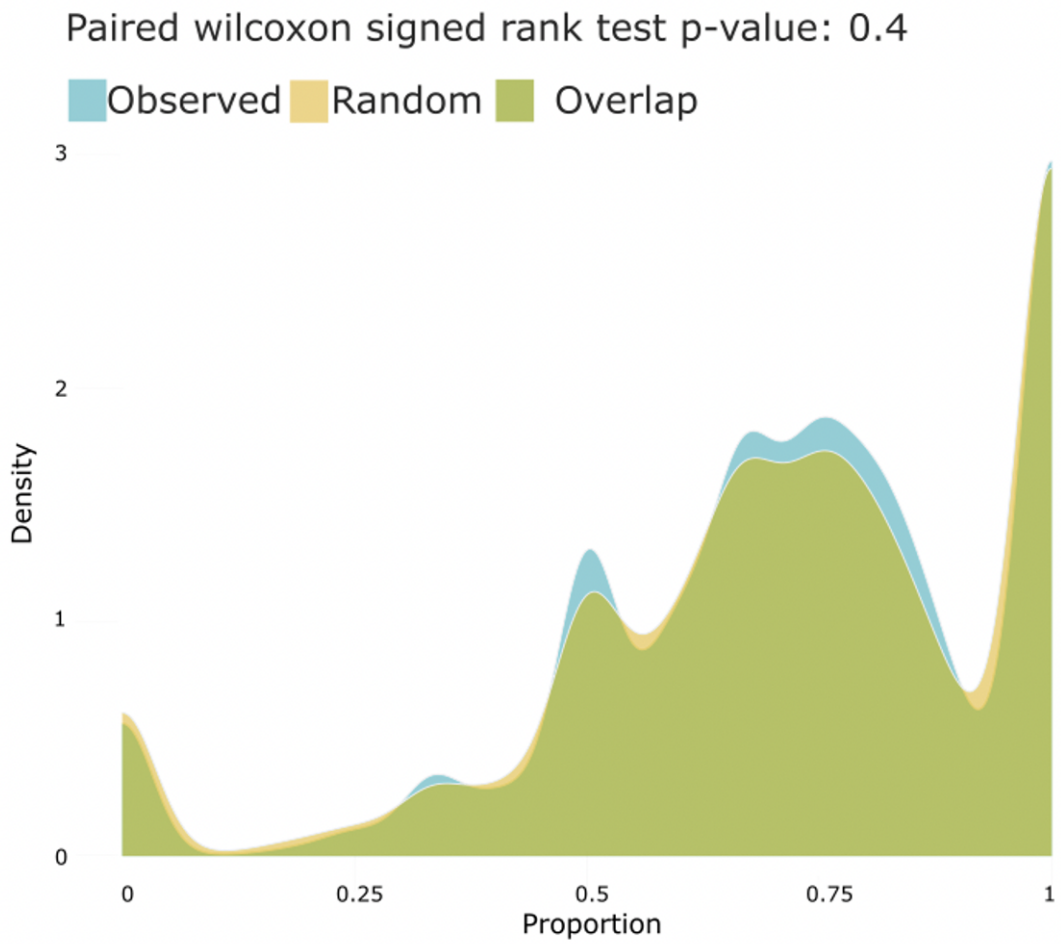

Figure 6.2: Comparison of the proportion of missense mutations per gene for the observed versus random dataset. Overlapping density plots showing the proportion of mutations classified as missense for each gene in the observed data (blue) and in the randomly assigned mutations for the same mutational context (yellow).

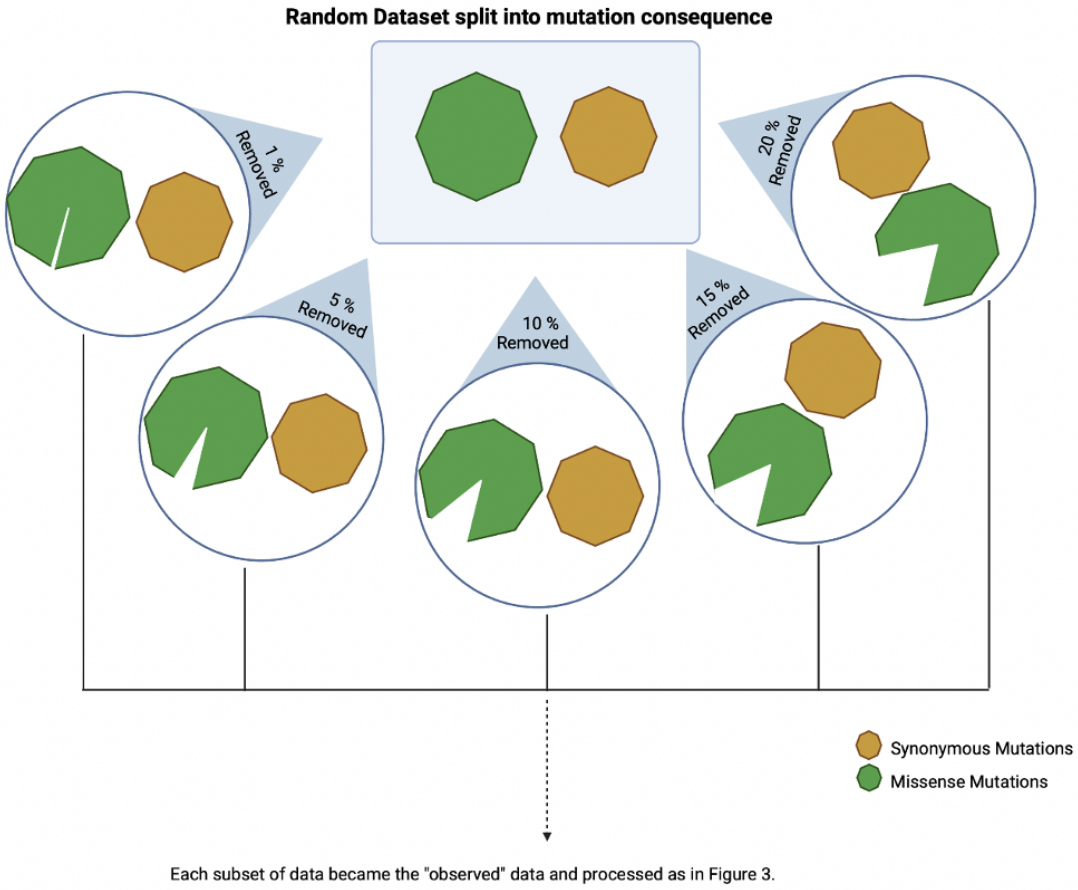

Figure 6.3: Schematic to illustrate the process of randomly removing different proportions of missense mutations from the data. Randomly removed 1, 5, 10, 15, and 20% of the missense mutations from the random dataset created in Figure 6.1

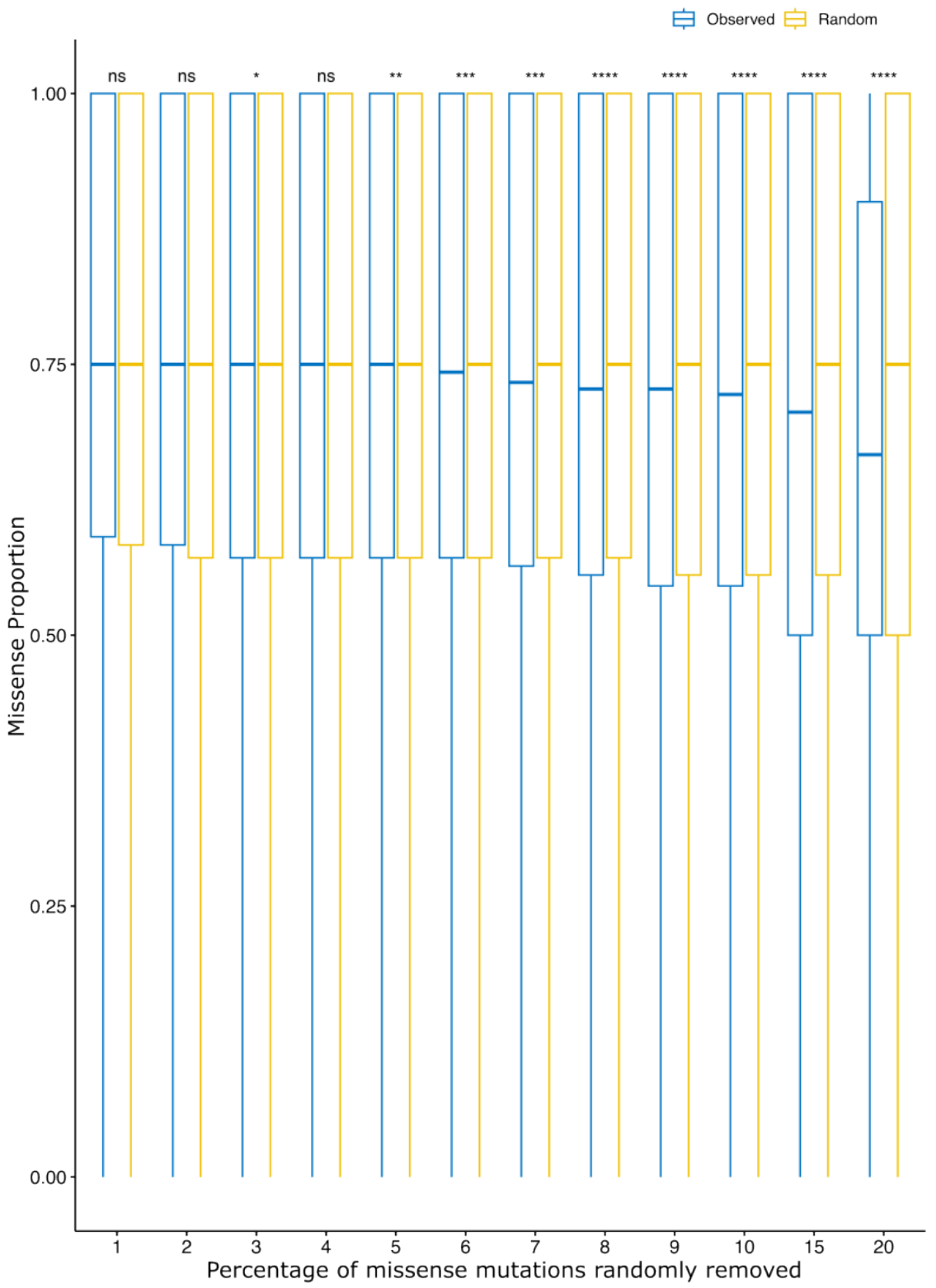

Figure 6.4: Comparison of the proportion of missense mutations per gene for the simulated datasets with a proportion of missense mutations re- moved versus the corresponding random dataset. Boxplot P-values are from paired Wilcoxon rank sum tests.