INTERROGATING THE TCR-pMHC COMPLEX IN HEALTH AND DISEASE USING IMMUNOGENOMICS METHODS

by

Scott Derek Brown

B.Sc., Simon Fraser University, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in
THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

December 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Interrogating the TCR-pMHC Complex in Health and Disease Using Immunogenomics Methods

Submitted by Scott Derek Brown in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in Genome Science and Technology

Examining Committee:

Professor Robert A. Holt

Supervisor

Professor Christian Steidl

Supervisory Committee Member

Professor Scott Tebbutt

University Examiner

Professor David Granville

University Examiner

Additional Supervisory Committee Members:

Professor Brad Nelson

Supervisory Committee Member

Professor Inanç Birol

Supervisory Committee Member

Abstract

The adaptive immune system is a complex network of cells working towards a common goal: detection and elimination of foreign cells that can harm the host. In cancer, malignant cells acquire mutations which can appear foreign to the adaptive immune system. The immune cells most directly involved in destruction of cancer cells are $CD8^+$ T cells, using their T cell receptor (TCR) to recognize mutated peptides presented on cancer cells in the context of class I Major Histocompatibility Complex (MHC) molecules (pMHC). Immunogenomics methods offer ways to interrogate this TCR-pMHC interaction using genomics data.

The aim of this thesis is to adapt and apply novel and existing immunoinformatic methods to cancer datasets to identify relationships between the immune system and cancer in a pan-cancer context. This involves prediction of cancer neoantigens derived from single nucleotide variants (SNVs) from tumours, and correlation of this neoantigen burden with outcomes and markers of immune inhibition. It involves extraction of TCR sequences from RNA-seq datasets to gain value-added information from these existing datasets, with demonstrated utility in solid tumours and lymphomas. Finally, it defines and explores the size of the self-immunopeptidome to classify individuals based on their ability to present peptides on class I MHC molecules.

I show that T cell infiltration of solid tumours correlates with improved outcomes, neoantigen load, but also markers of T cell inhibition, suggesting that these individuals would benefit from checkpoint blockade therapy. In established tumours, the T cell repertoire is not clonal, and among the most abundant T cells in the tumour are viral-specific T cells also found in the normal repertoire. This information is obtained directly from existing RNA-seq datasets of tumours. When applied to RNA-seq of sorted T cell populations, clonally expanded T cells are detectable by their TCR, and alpha-beta pairing can be inferred. The self-immunopeptidome can be used to predict neoantigen load and is used to infer signatures of neoantigen immunogenicity. This thesis contributes towards a better understanding of the interaction between T cells and cancer cells, which can inform future strategies to improve immunotherapies in cancer.

Lay Summary

The human immune system is capable of detecting and responding to a wide variety of threats, recognizing when a cell does not belong to the body. Cancer cells gain mutations which can be recognized by the immune system as foreign. When this happens, the immune system will attack these cells, attempting to kill them. This interaction is directed by special receptors on the immune cells which can bind to mutated proteins on the cancer cells. In this thesis, I use computational methods to make predictions from DNA sequence datasets from thousands of cancer patients to learn about these immune cell receptors and the mutations they target. I predict which mutations will be targeted by an immune cell, and how this predicts survival. By better understanding how these immune cells interact with cancer cells, we can decide which cancer treatments will have the best chance of success for a specific individual.

Preface

The overall project was designed and conducted in collaboration with my supervisor Dr. Robert Holt. Considerations for specific chapters are as follows.

A version of chapter 2 has been published as below. I was responsible for processing the data and analysing the results. I performed the survival analysis with guidance from Dr. John Spinelli. I co-wrote the manuscript with Dr. Robert Holt. Rene Warren acquired the raw data, performed initial development of the computational pipeline to generate the peptide-MHC binding predictions, and designed the MySQL database to store the results. Gene expression quantification was performed by Rene Warren and Dr. Ewan Gibb.

> Brown, S.D., Warren, R.L., Gibb, E.A., Martin, S.D., Spinelli, J.J., Nelson, B.H., and Holt, R.A. (2014). Neo-antigens predicted by tumour genome meta-analysis correlate with increased patient survival. *Genome Research*, 24(5), 743–750.

A version of chapter 3 has been published across two manuscripts as below. In section 3.1, I was responsible for the research design and development of the methods, processing and analysis of the data, and drafting the manuscript. TCR-seq libraries were created previously by Lisa Raeburn. RNA-seq libraries were created previously by Douglas Freeman.

> Brown, S.D., Raeburn, L.A., and Holt, R.A. (2015). Profiling tissue-resident T cell repertoires by RNA sequencing. *Genome Medicine*, 7(1), 125.

In section 3.2, I was responsible for the bioinformatic analysis of the data and interpretation of the results. I co-wrote the manuscript with Dr. Greg Hapgood, who also performed sample preparation and FACS analysis.

> Brown, S.D., Hapgood, G., Steidl, C., Weng, A.P., Savage, K.J., and Holt, R.A. (2016). Defining the clonality of peripheral T cell lymphomas using RNA-seq. *Bioinformatics*, 33(8), 1111–1115.

A version of chapter 4 has been accepted as a manuscript as below. I was responsible for the research design, data generation and analysis, interpretation, and drafting of the manuscript.

> Brown, S.D. and Holt, R.A. Neoantigen characteristics in the context of the complete predicted MHC class I predicted self-immunopeptidome. *Oncoimmunology,* DOI:10.1080/2162402X.2018.1556080

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACC	Adrenocortical carcinoma
AIRE	Autoimmune regulator
AITL	Angioimmunoblastic T cell lymphoma
APC	Antigen presenting cell
BAM	Binary alignment map (file)
BCR	B cell receptor
BLCA	Bladder urothelial carcinoma
BOVIN	*Bos Taurus*
BRAFL	*Branchiostoma floridae*
BRCA	Breast invasive carcinoma
CANLF	*Canis lupus*
CAR	Chimeric antigen receptor
CDR3	Complementarity determining region 3
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHICK	*Gallus gallus*
CHLTR	*Chlamydia trachomatis*
CI	Confidence interval
CIHR	Canadian Institutes of Health Research
CIOIN	*Ciona intestinalis*
COAD	Colon adenocarcinoma
CPU	Central processing unit
CRAD	Colon and rectum adenocarcinoma (COAD and READ combined)
CTL	Cytotoxic T lymphocyte
CTLA-4	Cytotoxic T-lymphocyte-associated antigen 4
DANRE	*Danio rerio*
DC	Dendritic cell
DNA	Deoxyribonucleic acid
DRiP	Defective ribosomal product
EBV	Epstein-barr virus
ECOLI	*Escherichia coli*
EMBL	European Molecular Biology Laboratory
ENCODE	Encyclopedia of DNA Elements
ER	Endoplasmic reticulum
ERAAP	Endoplasmic reticulum aminopeptidase associated with antigen presentation
ESCA	Esophageal carcinoma

FACS	Fluorescence-activated cell sorting
FPKM	Fragments per kilobase of exon per million fragments mapped
FUSNN	*Fusobacterium nucleatum*
GBM	Glioblastoma multiform
GDC	Genomic Data Commons
GIAIC	*Giardia intestinalis*
GTEx	Genotype-Tissue Expression Project
HLA	Human leukocyte antigen
HNSC	Head and neck squamous cell carcinoma
HR	Hazard ratio
IEDB	Immune Epitope Database
Ig	Immunoglobulin
Indel	Insertion or deletion
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LEIMA	*Leishmania major*
LEPIN	*Leptospira interrogans*
LF	Leukocyte fraction
LGG	Brain lower grade glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MACMU	*Macaca mulatta*
MAF	Mutation annotation file
MHC	Major histocompatibility complex
MONDO	*Monodelphis domestica*
MOUSE	*Mus musculus*
MS	Mass spectrometry
NGS	Next-generation sequencing
NHGRI	National Human Genome Research Institute
NIH	National Institutes of Health
NK	Natural killer (cell)
NMDP	National Marrow Donor Program
OR	Odds ratio
ORNAN	*Ornithorhynchus anatinus*
OV	Ovarian serous cystadenocarcinoma

PAAD	Pancreatic adenocarcinoma
PANTR	*Pan troglodytes*
PCPG	Pheochromocytoma and paraganglioma
PCR	Polymerase chain reaction
PD-1	Programmed cell death protein 1
PD-L1	Programmed cell death protein ligand 1
PDCD1	Programmed cell death 1 (gene)
PFI	Progression-free interval
PLAF7	*Plasmodium falciparum*
pMHC	Peptide-MHC
PRAD	Prostate adenocarcinoma
PSEAE	*Pseudomonas aeruginosa*
PSSM	Position specific scoring matrix
PTCL	Peripheral T cell lymphoma
PTCL-NOS	Peripheral T cell lymphoma not otherwise specified
RACE	Rapid amplification of cDNA ends
RAG	Recombination-activating gene
RAT	*Rattus norvegicus*
READ	Rectum adenocarcinoma
RNA	Ribonucleic acid
RT-PCR	Reverse transcription polymerase chain reaction
SKCM	Skin cutaneous melanoma
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
STAD	Stomach adenocarcinoma
TAKRU	*Takifugu rubripes*
TAP	Transporter associated with antigen presentation
TCGA	The Cancer Genome Atlas
TCR	T cell receptor
TEC	Thymic epithelial cell
THCA	Thyroid carcinoma
TIL	Tumour-infiltrating lymphocyte
TPM	Transcripts per million
UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma
XENTR	*Xenopus tropicalis*

Acknowledgements

First, I would like to thank my supervisor Dr. Robert Holt for his support, guidance, and the opportunities he provided to me over the years. His repeated willingness to let me explore areas outside of my expertise and take ownership over projects allowed me to learn so much. His repeated use of “why don’t you try…,” while sometimes asking the impossible, led to numerous new avenues to explore. His encouragement to present my work and take part in collaborations ensured that I developed a well-rounded set of skills throughout my degree.

Next, I would like to thank Michael Crichton, whose novel *Jurassic Park* got me interested in science and genomics at a young age. I didn’t realize it until much later, but this work of fiction is at least partially responsible for the academic path I took that has brought me to where I am today.

Thank you to my committee members Dr. Inanc Birol, Dr. Brad Nelson, and Dr. Christian Steidl, whose scientific support and wisdom helped guide my research. Thank you also to Dr. Greg Hapgood, Dr. Kerry Savage, and Dr. Andy Weng, whose willingness to let me explore their data has led to fruitful collaborations.

I would like to thank Dr. Andy Sandford, who accepted me into his lab as a green co-op student with zero experience, and Dr. Loubna Akhabir, for teaching me how to function in a molecular biology lab and introducing me to the field of bioinformatics. Lessons learned in the Sandford lab have stuck with me throughout my scientific career. Thank you to Dr. Anthony Fejes, who guided my first foray into real bioinformatics and programming during my time in Dr. Michael Kobor’s lab. Thanks also goes to Dr. Elodie Portales-Casamar who showed me fundamentals in data analysis during my time in Dr. Paul Pavlidis’ lab.

To the members of the Holt lab, past and present, thank you for providing a warm, friendly, fun, and collaborative work environment. I’ll always remember the nights at OJs and lunchtime crosswords. Special thanks to Dr. Mauro Castellarin, Lisa Dreolini, Lisa Raeburn, Doug Freeman, and Dr. Spencer Martin who made me feel so welcome when I first joined the lab. A big thank you to Rene Warren, not only for laying the groundwork for the project that would later become the basis for my thesis, but for being

an inspiration and role model during the early stages of my scientific career. Payal Sipahimalani is also deserving of acknowledgement, helping me navigate through the complicated nuances of data usage restrictions, and dealing with the headaches I put her through while we both tried to understand ambiguous data sharing rules. None of my work would have been possible without her persistence and dedication. To my fellow graduate students, Govinda Sharma and Chris May, thank you for the support and discussions, but even more than that, thank you for the laughs. Your comradery and friendship made this whole experience so much more enjoyable. Finally, I would like to thank Daniel Woodsworth. I think back fondly on our late-night scientific discussions and attempts to come up with hairbrained schemes to solve the big problems in biology, and will always be grateful for your friendship and support.

I would like to thank those that provided the financial support that allowed me to complete this PhD; the Genome Science + Technology Program, the University of British Columbia, and the Canadian Institutes of Health Research. Thanks also goes to those that made data available for me to use; the NIH and the Cancer Genome Atlas Research Network, the National Marrow Donor Program, and ENCODE.

Finally, a special thanks to my friends and family, who have been so supportive, and only occasionally questioned my judgement for staying in school for 24 years. I would especially like to thank my parents, Tracey and Derek, for their love and support over the years. This journey has not been easy, and it has been extraordinarily reassuring to know that they were there for me without question. Likewise, thank you to my partner in life, Nafeesa Amlani. Her unending support through the ups and downs of graduate school means more to me than she knows.

To the participants of The Cancer Genome Atlas

Chapter 1: Introduction

The human adaptive immune system is a complex network of cells working towards a common goal: detection and elimination of foreign cells that can harm the host. One of the hallmarks of cancer is genomic instability, allowing these cells to have unrestricted growth. A by-product of this genomic instability is that immune cells can recognize these cancerous cells as non-self and can eliminate them. However, a game of cat-and-mouse takes place as selective pressures on the cancer cells allow them to evolve mechanisms to evade the immune system. Within this thesis, I adapt and apply novel and existing immunoinformatic methods to existing cancer datasets to identify general patterns and relationships between the class I immune system and cancer in a pan-cancer context, not dependent on tumour type. These findings contribute towards a better understanding of the interaction between T cells and cancer cells, which can inform future strategies to improve immunotherapies in cancer.

Within this introductory chapter, I review the existing literature and motivations for pursuing novel ways to interrogate the TCR-pMHC interaction computationally. Firstly, fundamentals of the adaptive immune system are reviewed, with emphasis on the class I MHC presentation pathways and T cell recognition of these presented antigens. Next, I review the interaction between cancer cells and immune cells in the context of this adaptive immunity, and how this interaction can be leveraged for cancer immunotherapies. I then review the existing immunoinformatic methods and strategies that can be applied to genomic data to uncover information regarding TCRs and pMHCs, and tools to predict immunogenicity of presented mutant peptides in cancer. Finally, the overall layout and goals of this thesis are overviewed.

1.1 The adaptive immune system

The adaptive immune system evolved to protect vertebrates from foreign pathogens or cells containing infectious agents, such as viruses. If this system identifies something on the surface of a cell as foreign, this presenting cell will be killed. These cell surface molecules can include membrane proteins which naturally exist on the surface of the cell, intracellular proteins which are degraded, processed, and presented on Major Histocompatibility Complex (MHC) class I molecules via antigen presentation pathway,

or extracellular proteins that are taken up by antigen presentation cells and shuttled into the antigen presentation pathway for display on MHC class II molecules. Proteins in their native form are recognized by B cells, whereas MHC-presented peptides derived from digested proteins are recognized by T cells (see 1.1.2 MHC antigen presentation). Of the two classes of MHC presentation, class I is present on all cell types and can result in direct cell death from interaction between the presenting target cell and an effector T cell. The work presented in this thesis focuses on this axis.

Cytotoxic T lymphocytes (CTLs) are the immune cells most directly responsible for antigen recognition and cell killing, having a cell-surface T cell receptor (TCR) which binds the peptide-MHC (pMHC) complex of other cells [1,2], and cytolytic granules which release upon pMHC recognition. Once activated (see 1.1.6 T cell priming) CTLs circulate through the body, surveying the pMHCs presented on cells. If there is sufficient engagement between TCR and pMHC, an immunological synapse is formed between the two cells, and an as-of-yet undetermined mechanism occurs resulting in initiation of the T cell signalling cascade [3]. This cascade causes the release of granzyme and perforin by the T cell into the immunological synapse, which allows granzyme to enter the target cell and ultimately cause cell death in the target cell [4].

1.1.1 T cell receptor rearrangement

The TCR determines the antigen specificity of the cell, and as such, there must exist a massive TCR sequence diversity within every individual for the immune system to be effective against a broad range of targets [5]. How is this TCR sequence diversity achieved? The TCR consists of a dimer of two subunits, or chains: alpha and beta (the most common), or gamma and delta [6]. Alpha and gamma chains are comprised of (V)ariable, (J)oining, and (C)onstant gene segments, whereas beta and delta chains have V, (D)iversity, J, and C gene segments. These gene segments are somatically recombined during T cell development in the thymus, resulting in a random combination of V, (D), J, and C genes (Figure 1.1) [7]. Additionally, at the gene junctions between the V-(D)-J genes there is random nucleotide deletion and non-templated nucleotide addition, further increasing the potential sequence diversity from what is encoded in the genome [8,9].

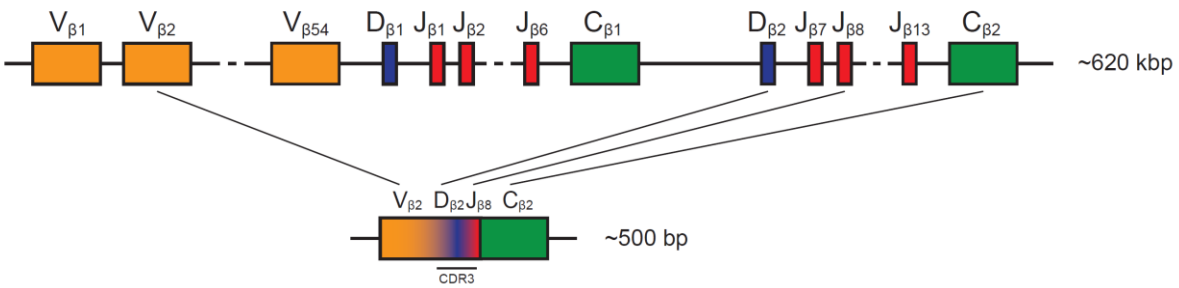

Figure 1.1: VDJ gene recombination of genes at the TCR beta locus. First, one of two D genes is joined with one of 13 J genes, followed by joining of DJ to one of 54 V genes. Nucleotides are randomly deleted and inserted at the DJ and VD junctions (depicted by the colour gradient). Finally, the recombined VDJ is joined to one of two C genes. The CDR3 region encompasses the end of the V gene to the beginning of the J gene. Adapted from [10].

The region from the end of the V gene to the beginning of the J gene is known as the complementarity determining region 3 (CDR3), and is the segment of the TCR that is most frequently studied as it encodes the hypervariable protein loop that contacts the MHC-presented peptide (Figure 1.2A). The N- and C-terminal ends of the CDR3 have low variability as they are encoded by the germline V and J gene segments, rarely affected by the random nucleotide deletion [11]. Moving towards the middle of the CDR3 results in increased deviation from the germline sequence. These conserved boundaries of the CDR3 are useful informatically as they act as landmarks to delineate the edges of the CDR3 in sequence data (see 1.3.1 TCR annotation).

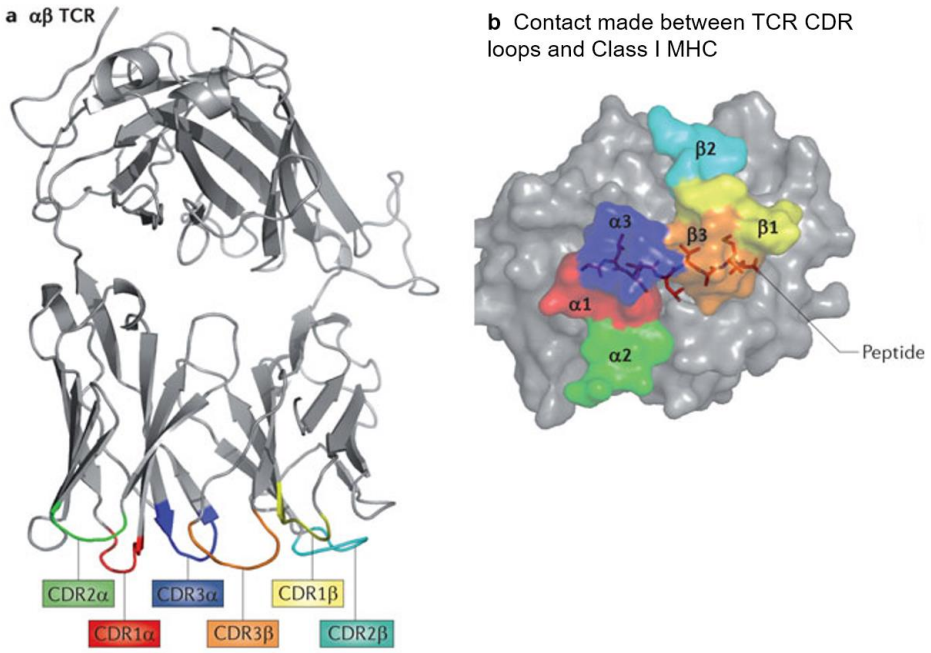

Figure 1.2: CDR loops of the T cell receptor. (A) Crystal structure of an alpha-beta TCR with coloured CDR loops. **(B)** *HLA-A*02:01* (grey) presenting peptide (stick model), and coloured regions of interaction with TCR CDR loops. The CDR3 of the alpha and beta chain contact the peptide. Adapted by permission from Springer Nature: Nature Reviews Immunology "Why must T cells be cross-reactive", Andrew K. Sewell, Copyright 2012. [12]

Within the TCR dimer, the CDR3 of the alpha and beta chains work together to scan the MHC-presented peptide (Figure 1.2B). Historically, CDR3-beta is targeted to be studied for high-throughput sequencing experiments due to the increased potential diversity (due to having a D gene) and allelic exclusion at the beta locus typically resulting in only a single TCR-beta chain being expressed per cell [13,14], however, it is important to keep in mind that the TCR recognition is dependent on both alpha and beta chains, and knowledge of the beta chain alone is not sufficient to infer binding specificity. Recent advances in methods for TCR-seq sample preparation with the goal of determining paired TCR chains combined with appropriate data analysis [15], and in single cell sequencing technologies [16], provide paired information on both the alpha and beta chains in a given T cell.

1.1.2 MHC antigen presentation

There are two classes of antigen presentation: class I and class II. Class I deals with presentation of intracellular proteins and occurs in almost all cells, whereas class II only occurs in professional antigen-presenting cells (APCs) such as dendritic cells, and presents extracellular proteins that have been endocytosed by these APCs. As class I presentation results in direct interaction between the presenting target cell with an effector CTL, the work in this thesis focuses on this TCR-pMHC axis. Within every cell of the body, proteins are continuously being degraded in the cytosol, and are replaced with newly synthesized ones. The degradation products of this process enter the class I antigen presentation pathway, resulting in the short peptide fragments being displayed on class I MHC molecules [17,18]. This degradation is primarily done by the 26S proteasome, however, upon stimulation with interferon gamma (as during an active immune response), subunits within the 26S proteasome are exchanged to form the immunoproteasome. This complex favours the creation of short peptide fragments with hydrophobic C-terminal ends, increasing the generation of peptides able to bind MHC [19]. The peptide fragments generated by these proteasomes are transported into the endoplasmic reticulum (ER) by the transporter associated with antigen presentation (TAP) [20]. Within the ER, these peptides are further digested from their N-terminal end by endoplasmic reticulum aminopeptidase associated with antigen processing (ERAAP) [21], and are loaded onto MHC molecules. These peptide-MHC complexes then migrate to the cell surface to be displayed for surveillance by T cells.

While the protein source for most of these peptides are mainly either old proteins marked for degradation via ubiquitination or defective ribosomal products (DRiPs), there are many cases where presented peptides do not originate from canonical coding sequence. These include so-called cryptic MHC I-associated peptides generated from translation of non-coding regions and translation of a non-canonical reading frames [22]. Additionally, it has been reported that peptide splicing can occur within the proteasome, generating peptides that do not derive from a continuous stretch of any coding sequence [23–25]. Due to the stochastic nature in the generation of these peptides, and our incomplete understanding of the rules governing their creation, their occurrence is not amenable to prediction by algorithms.

1.1.3 Peptide-MHC binding

The binding of a short peptide to an MHC molecule is highly dependent on the MHC molecule – itself encoded by the Human Leukocyte Antigen (HLA) genes. The HLA locus is the most polymorphic region of the human genome [26], and each HLA allele yields an MHC with distinct binding characteristics [27], resulting in only a subset of peptides having the ability to bind any given MHC molecule. This subset is potentially immunogenic: it is presented on MHC and may generate a T cell response.

Peptides bind within the binding groove of the MHC. Most of the variability within HLA sequences cause changes within this binding groove, altering which peptides are able to bind [28]. For class I MHC molecules, the ends of the binding groove are closed, restricting the length of peptides that can be presented. The typical lengths of MHC I-presented peptides are 8-11 amino acids, with the most common length (> 73 %) being 9 amino acids [29,30]. Typically, there are two positions within the presented peptide that interact with pockets within the MHC binding groove. At the N-terminal end, the anchor residue is at position 2. The C-terminal anchor occurs right at the C-terminal end (for example, position 9 for a 9mer peptide) [31]. The peptide sequence between the two anchors may bulge out from the binding groove, depending on the length of the peptide [32] (Figure 1.3). This region of the presented peptide is readily available for surveying by the TCR.

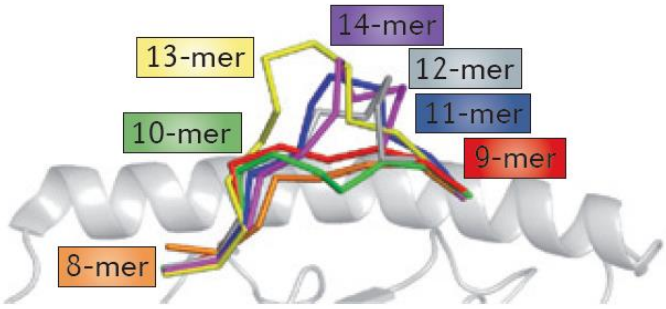

Figure 1.3: Short peptides in the MHC-I binding groove. Longer peptides exhibit greater bulging away from the MHC. Adapted by permission from Springer Nature: Nature Reviews Immunology "Why must T cells be cross-reactive", Andrew K. Sewell, Copyright 2012. [12]

1.1.4 Immunogenicity and immunodominance

Non-self peptides which are processed and presented on MHC molecules may be immunogenic – able to produce an immune response. During a viral infection there may be many peptides which are potentially immunogenic, however, it has been observed that only a small subset of these will provoke an immune response [33–35]. These peptides are immunodominant. The difference between immunogenicity and immunodominance is subtle – immunogenicity refers to the potential to generate an immune response whereas immunodominance refers to the peptide which actually drives the response. Since immunodominance is dependent on there being a T cell present which can recognize and respond to the peptide, two individuals with the same infection and same MHC type may have different immunodominant peptides [36]. Some of the factors determining which peptides will be immunodominant, reviewed by Akram & Inman [37], include prior viral infections [38], frequency of naïve T cells and the rate of their clonal expansion [39], and TCR-pMHC affinity [40]. Since prediction of immunodominance would require omniscient knowledge of all T cells in a sample as well as prior infections that have generated immune memory, immunogenicity predictions are typically used in lieu of attempting to predict immunodominance. This is true within this thesis, where the term “immunogenic” is used to describe peptides which are predicted to have the potential to generate an immune response given the available data.

1.1.5 T cell development

The sequence diversity of the TCR repertoire (the set of all TCR sequences present in a sample) is limited during T cell development in the thymus to prevent T cell recognition of self-antigens. It was first postulated by Ehrlich that there exists a mechanism to avoid self-reactivity by lymphocytes [41]. It is now understood that within the thymic cortex and medulla, thymic epithelial cells (TECs) present self-peptides in the context of MHC. Due to the tissue-specificity of gene expression within the periphery, T cells which leave the thymus must not be self-reactive against any of the potential self-peptides they will encounter. To ensure this self-tolerance will exist, the presentation of self-antigens by TECs is promiscuous, deriving from genes throughout the entire genome and not solely

genes required to be expressed in thymic cells. This promiscuous expression is driven by the autoimmune regulator (AIRE) protein [42,43], resulting in non-tissue-specific clusters of genes within the genome being expressed. As developing T cells transit the thymus, only those that can bind to pMHC complexes (positive selection), yet do not strongly recognize self-peptides (negative selection), will survive. This ensures that all T cells exiting the thymus into the periphery are effective in surveying MHC-bound peptides, and do not contain any strong, self-reactive TCRs, as this could lead to auto-immune disorders [44,45]. The resulting TCRs will have established central tolerance to the presented self-peptides present in that individual (the self-immunopeptidome). As TCR recombination is a stochastic event, and the set of self-peptides presented by TECs (the self-immunopeptidome) are dependent on a subject's HLA alleles, the set of TCRs generated in two distinct individuals would not be expected to be the same.

1.1.6 T cell priming

Naïve T cells which have exited the thymus and circulate through the body need to be primed before they can mount a cytolytic response to antigen [46]. This priming is most commonly performed by professional APCs known as dendritic cells (DCs) [47–49]. DCs acquire proteins from sites of infection and return to lymph nodes to present antigens derived from these proteins to T cells [50,51]. These proteins may be the result of direct infection of the DC (by a virus) or endocytosed exogenous proteins from the environment. Exogenous proteins are canonically presented via the class II MHC presentation pathway, but DCs perform cross-presentation of these proteins, diverting some of them to the cytosol to be processed via the class I MHC presentation pathway for surveying by $CD8^+$ T cells [52–54].

Naïve T cells require two signals to become effector cells. The first signal is recognition of pMHC on the surface of an APC, and typically takes place in the secondary lymphoid organs such as lymph nodes. The second signal is the binding of CD28 (on the T cell) by B7 molecules (CD80 or CD86, upregulated during infection) on the APC, resulting in massive clonal expansion of that T cell [55]. This expansion increases the ability of the immune system to detect that specific antigen at the time of infection. These activated cytotoxic T cells can then transit to the site of infection or

inflammation to attack the cells expressing these foreign antigens. Following clearance of the infection, most of these T cells die by apoptosis, but around 10 % remain as memory T cells, protecting against future infections by the same antigen [56]. At times when infection is not present, APCs will not express the B7 molecules. Naïve T cells which interact with self peptides presented on these APCs will become anergic, establishing peripheral tolerance to these self antigens [57,58].

1.1.7 T cell cross-reactivity

Originally, it was theorized that a single TCR would only recognize a single target [59,60], though later theoretical calculations demonstrated that this was not feasible for sufficient protection from pathogens to exist [61]. After TCR rearrangement has occurred during development, the TCR of T cells is unchanged during the cell's lifetime. Therefore, assuming a one-to-one relationship between TCR and antigen, upwards of 10^{15} T cells would be required to recognize any of the possible foreign peptides, a number of T cells which would weigh over 500 kg [12]. This is clearly not the case, as the total number of T cells present in the human repertoire is closer to 10^{12}, comfortably fitting inside the human body [62]. These observations led others to collect direct evidence of T cell cross-reactivity [63,64], generating an estimate that each T cell is capable of recognizing over one million distinct antigens [65,66]. These findings explain how the T cell repertoire present in an individual is able to offer protection against a vast space of possible foreign pathogens.

1.1.8 T cell receptor repertoire

T cells that have been primed by APCs due to an appropriate antigen being present will have undergone clonal expansion, greatly increasing the number of T cells with that specific TCR [55]. These clonal expansions alter the TCR repertoire present in the periphery, skewing the abundance of T cells with certain TCRs towards those that have recognized an antigen. By measuring the TCR repertoire, these clonally expanded T cells can be identified. Genomic approaches are used to provide a high-resolution view of the TCR repertoire. Specialized amplicon sequencing experiments (TCR-seq) are performed which first selectively amplify the CDR3 region of the TCR chain from RNA or DNA [11,67]. Subsequently, these amplicons are sequenced, and the resulting

sequence reads are annotated for the CDR3 region (see 1.3.1 TCR annotation). Studies interrogating the TCR repertoires of multiple individuals find minimal overlap between samples from different individuals [5]. Therefore, when clonally expanded TCRs are observed to be shared in multiple individuals, it is likely that they have been selected due to common antigens [68]. However, much like T cells being cross reactive against multiple antigens, it is also true that antigens can be recognized by multiple distinct T cells [69]. There is a many-to-many relationship between T cells and antigens, complicating the discovery of TCR-pMHC interactions via informatics-based approaches to analyze cancer datasets. The same antigen and TCR co-occurring in two distinct individuals, while possible, will occur at a low frequency.

1.1.9 The self-immunopeptidome

During T cell development, T cells are presented with self-peptides in the thymus. Depending on the six HLA class I alleles that a subject has, the set of self-peptides presented to T cells will vary. This set of self-peptides is known as the self-immunopeptidome and is unique to each individual. The size of this set of self-peptides can vary considerably between different HLA alleles [70,71], therefore, the self-immunopeptidome in one subject may be dramatically different than another subject with a different HLA genotype. The size and diversity of the self-immunopeptidome may influence the size and diversity of the TCR repertoire, and this may explain the ongoing hypothesis that there are “at-risk” HLA alleles. In theory, within the context of autoimmunity, a subject with a very large self-immunopeptidome will have a large number of self-pMHCs that need to be surveyed by T cells during development in the thymus, and there may be a higher risk of an auto-immune TCR escaping negative selection into the periphery simply due to this large number of self-pMHCs that need to be surveyed. In the context of cancer immunology, subjects with large self-immunopeptidomes may have a greater chance of any given mutation being presented, leading to more effective elimination of these cells, resulting in any tumour that evades this selection being highly edited and non-immunogenic (having escaped the immune system). Conversely, it is also possible that a large self-immunopeptidome will result in a greatly restricted TCR repertoire during development, limiting the ability to recognize

foreign antigens in the periphery. It is likely that there is a balancing act for the optimal amount of peptide presentation; too few and the immune system will not be able to effectively detect foreign peptides, too many and the risk of autoimmune disease, risk of foreign peptides being lost in a sea of self-peptides, and risk of extreme selection on the TCR repertoire during negative selection will be too great [72,73]. The calculation of the self-immunopeptidome size, and its relation to cancer immunology, had not yet been explored (see Chapter 4: Neoantigen characteristics in the context of the complete predicted MHC class I self-immunopeptidome).

1.2 Cancer immunology

Cancer is the result of malignant cell transformation causing unrestricted growth [74]. Early independent studies by Ehrlich and Bashford were the first to demonstrate that the immune system may be involved in tumour cell growth, able to control growth when the burden of cancer cells is low, but being less effective on established tumours [75,76]. Later studies provided evidence that the adaptive immune response was involved, demonstrating that inoculation of tumour cells can be protective against future tumour development in chemically-induced tumours [77]. This suggested that the immune system was capable of surveying and creating memory of these cancer cells. The concept of immunosurveillance [78] was investigated in studies on immunodeficient mice (RAG2$^{-/-}$), showing that these immunocompromised individuals had higher incidence of carcinogen-induced and spontaneous cancers than immunocompetent hosts, and these tumours were immunogenic when transplanted into immunocompetent hosts [79]. Further, tumours that exist in equilibrium with an immune system in an immunocompetent host will rapidly grow if that host becomes immunodeficient [80], demonstrating that the immune response is critical to prevent the development of these tumours. These studies highlighted the role the immune system has in shaping the tumour as it grows, selecting for cancer clones with reduced immunogenicity (see 1.2.4 Cancer immunoediting and immune-evasion).

Further evidence that the immune system plays an important role in cancer development can be seen in multiple case reports and studies looking at rates of cancer incidence in immunosuppressed transplant recipients [81–84], showing an increase in

the rate of non-infectious primary cancers in these individuals. Additionally, the incidence of infectious-based cancers is increased in individuals with immunodeficiencies due to HIV infection [85,86]. Finally, immunotherapies have been shown to regress tumours. The earliest recorded evidence of immunotherapies being used to treat cancers is from the Egyptian Ebers Papyrus (circa 1550 BC), which recommended treating tumours with a poultice followed by an incision [87], now understood to likely result in infection of the tumour generating an immune response and subsequent regression [88]. More modern evidence of immunotherapy having a beneficial effect on cancer was shown by Coley [89], creating a non-living bacterial vaccine for the treatment of sarcomas. Immunotherapies have since shown remarkable promise in the treatment of cancers (see 1.2.5 Immunotherapies).

1.2.1 Cancer mutational profiles

Most cancers are caused by somatic mutations (spontaneous changes to the genome that occur in a single cell). If a cell acquires one of these mutations in a key cancer-related gene, it may allow this cell to grow without restriction, developing into a tumour [74]. Some mutations are recurrent, repeatedly occurring in a specific gene or position within a gene across many individuals. Others are sporadic, occurring more randomly throughout the genomes. Recurrent mutations in cancer are typically driver mutations, either resulting in loss of function of a tumour suppressor gene (such as *TP53*) or activating an oncogene (such as *KRAS*) [90,91]. These mutations drive the progression of the cancer, enabling further mutation and development. Passenger mutations do not necessarily confer any growth advantage to the cell, and in fact may result in reduced proliferative fitness [92]. The most common types of mutations are single nucleotide variants (SNVs) and small insertions or deletions (indels), though larger indels and gene fusions caused by structural variants are also prominent. The mutational load is variable both within and across different cancer types [93], as are the frequency of specific recurring mutations [94]. In predicting neoantigens, historically SNVs have been the focus. This is because the determination of the flanking protein sequence is trivial (unlike variants which cause frameshifts or larger structural changes), and a length-matched wildtype peptide which differs only at the mutation site can be obtained.

1.2.2 T cell infiltration

Tumour-infiltrating lymphocytes (TILs) are lymphocytes that have trafficked into the tumour tissue (Figure 1.4). These are mainly comprised of various subsets of $CD3^+$ T cells [95]. These cells can help mount an immune response to the tumour (ie. $CD8^+$ CTLs or $CD4^+$ T helper cells [96]), or can be immunosuppressive (ie. $FOXP3^+$ regulatory T cells [97]). Numerous studies have shown survival benefit in individuals harbouring "hot" tumours – those with large numbers of TILs, compared to those with "cold" tumours – having few TILs [98–100]. The mechanism underlying this survival benefit likely involves T cell recognition of cancer antigens – CTLs in the tumour can directly attack the cancer cells, responding to tumour-specific antigens (neoantigens) or tumour-associated antigens (cancer-testis antigens).

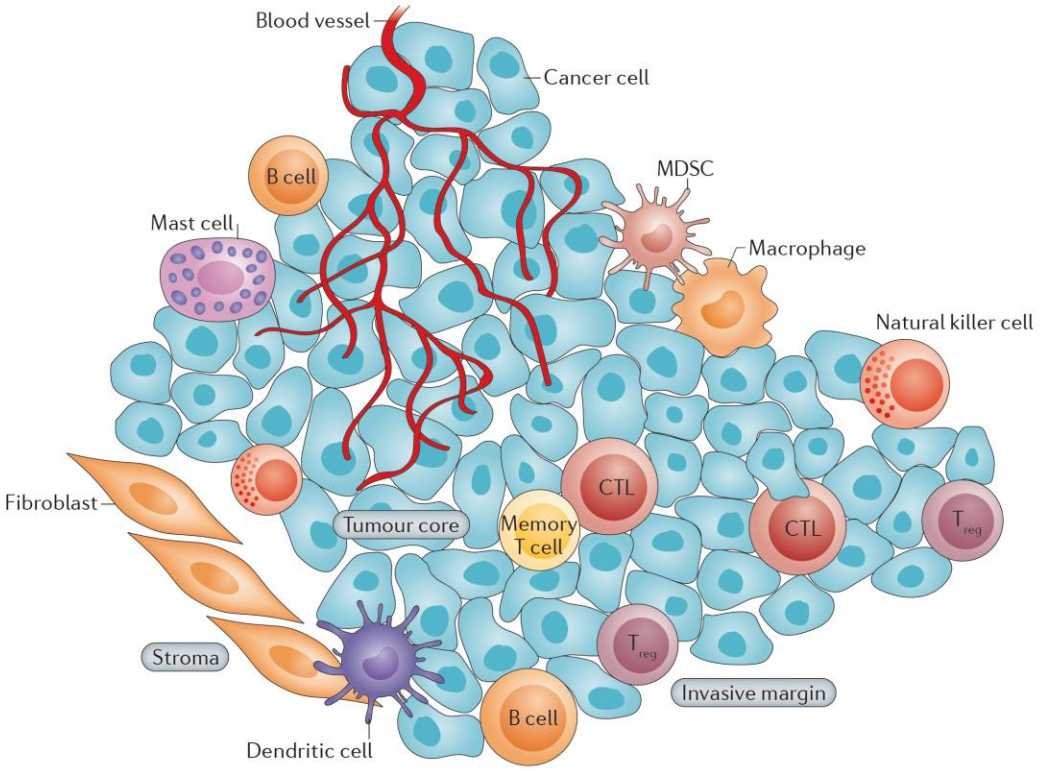

Figure 1.4: Immune infiltration within the tumour microenvironment. These infiltrating immune cells are mainly comprised of T cells (CTLs, memory T cells, helper T cells, and regulatory T cells), but can also include B cells and Natural Killer (NK) cells. Also present can be cells from the innate immune system. Adapted by permission from Springer Nature: Nature Reviews Genetics "Computational genomics tools for dissecting tumour–immune cell interactions", Hubert Hackl *et al.*, Copyright 2016. [101]

1.2.3 Cancer neoantigens and cancer-testis antigens

Cancer is a disease of genomic instability, yielding somatic changes to genomic sequence and aberrant expression of genes [74]. A non-synonymous mutation occurring in the coding region of a gene will result in an amino acid change in the protein encoded by that mutated gene. As with all proteins, this protein will be degraded through the class I antigen presentation pathway described above (see 1.1.2 MHC antigen presentation), and peptides containing the variant amino acids may be presented on the surface of the cell. There, it is potentially recognizable by T cells if it looks sufficiently different from the wild type “self” peptide, and will be a neoantigen. This type of antigen is said to be tumour-specific; an antigen that is unique to the tumour and does not occur in other tissues of the body. Alternatively, aberrant gene expression can cause genes that are typically only expressed in the human germline (and are thus immune-privileged) to be upregulated in cancer cells [102]. This cancer cell-specific aberrant expression means that MHC-presented peptides from these genes may form targets for CTLs (cancer-testis antigens) [103]. These antigens are said to be tumour-associated, resulting from aberrantly expressed unmutated self-peptides and not genomic changes specific to these cells. Thus, they are not as specific as cancer neoantigens and the possibility of off-target autoreactive CTL attack is higher.

1.2.4 Cancer immunoediting and immune-evasion

Tumour development is an evolutionary process, with tumours in constant interaction with the host immune system [78,104,105]. The cancer-immunity cycle describes the set of steps required for effective T cell recognition and attack of cancer cells [106]. First, cancer cell death results in the release of neoantigens, which are taken up by APCs (see 1.1.6 T cell priming) which transit via the efferent lymphatic vessels to the lymph nodes. There, these neoantigens are presented to naïve T cells, priming them for cancer cell recognition. Finally, these effector T cells, which have upregulated cell surface receptors allowing them to sense chemokines secreted by tumour cells, traffic into the tumour, specifically attacking the cancer cells expressing these neoantigens. Killing of these cancer cells releases more neoantigens, continuing the cycle. Nascent cancer cells enter the elimination phase of immunoediting when this immune cell attack

is uninhibited. If the cancer cells acquire alterations (genomic mutations or epigenetic changes) that confer some sort of resistance to the immune cells, immune pressure will select for these cells, killing other cells without the resistance. This is the equilibrium phase of immunoediting, where the immune system is not quite able to clear all the cancer cells. Finally, the cancer cells can acquire further alterations to evade the immune system, leading to the escape phase of immunoediting. At this point, the immune system is no longer able to control the cancer cells, and the tumour can grow uncontrollably. Due to variation in the HLA genes, different mutations will be immunogenic in different individuals. Therefore, this process results in each individual's tumour evolving its own unique set of mutations [91,107].

Even if a tumour bears mutations that make perfect T cell targets, there are ways it can evade the immune response. T cells have many co-stimulatory and inhibitory molecules on their surface, the latter of which are known as immune-checkpoint proteins, which act to modulate the T cell response to antigen [108]. These include two inhibitory receptors which are currently clinically relevant: cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4) [109] and programmed cell death protein 1 (PD-1) [110]. Upon T cell activation by TCR recognition of cognate antigen, CD28 (present on the cell membrane of T cells) amplifies TCR signalling by binding to CD80 on target cells. Shortly after, CTLA-4 is upregulated, and competitively binds CD80 to limit the T cell response. PD-1 is also upregulated on T cells post antigen-recognition, and binding by PD-L1 or PD-L2 inhibits the T cell's response. Unlike CD80, PD-L1 and PD-L2 can be upregulated on tumours [111], allowing the tumours to evade the immune system by altering the tumour microenvironment.

1.2.5 Immunotherapies

Since the seminal work on cancer immunotherapies by Coley demonstrating that inoculation of a strain of bacteria could induce an immune response to fight tumours [89], the field has expanded to include natural and engineered cell therapies, neoantigen vaccines, and targeted immune checkpoint blockade [109,110,112–115]. Despite promising results, these therapies are not yet effective for most patients [116]. Some level of personalization will be required for each individual tumour; tailoring the

therapy to unique immunogenic mutations and each patient's immune environment [117].

1.2.5.1 Checkpoint blockade

Checkpoint blockade is an antibody-based immunotherapy which blocks inhibitory receptors on T cells, releasing the brakes on the T cell response. These inhibitory receptors can be taken advantage of by cancer cells, suppressing the immune response and protecting themselves from T cell attack. Within cohorts of individuals treated with checkpoint blockade, mutational load has been shown to predict response to therapy, with better clinical outcomes typically seen for individuals having tumours with higher mutational loads [118–120]. It would appear that subsets of these mutations are eliciting T cell responses during treatment, but it is not yet known if this is due to these tumours having a greater chance of generating immunogenic mutations or is due to some other phenomena of which mutational load is a marker.

1.2.5.2 Neoantigen vaccines

Immune responses to tumours can be driven by therapeutic vaccines targeting cancer neoantigens [121,122]. Typically, this involves sequencing of the tumour to identify cancer-specific neoantigens, and creating a vaccine based on those neoantigens. Due to the high heterogeneity across cancers and subjects, a personalized approach is required, creating a novel vaccine for every individual. Recently, there have been successes in their application in melanoma, both as RNA-based and peptide based vaccination strategies [115,123], with significantly reduced rates of recurrence in patients receiving the vaccine. Work is ongoing in other cancer types, as well as using a combination therapy of neoantigen vaccines paired with checkpoint blockade [124]. Creation of cancer neoantigen vaccines would be assisted by improved bioinformatic tools to predict which neoantigens would form the most potent targets for an immune response.

1.2.6 Predicting response to immunotherapies

While it has been known for some time that patients bearing tumours containing an abundant CTL infiltrate have improved overall survival compared to those with a sparse

infiltrate [99,100], in general, the peptide targets of these CTLs remain elusive [125]. Evidence continues to build demonstrating CTL responses to neoantigens derived from point mutations within the tumour genome [112,126–130], however, predicting which mutations from the tumour genome are likely to generate *bona fide* neoantigens remains challenging as only a small fraction of mutations in a tumour can generate a T cell response [122,129–131], and further only a subset of those will be immunodominant (see 1.1.4 Immunogenicity and immunodominance). The criteria for filtering the list of mutations in a tumour down to those that are truly immunogenic and immunodominant have yet to be fully elucidated. While peptide-MHC binding predictions offer valuable information regarding what will be presented, there are other factors that can be taken into consideration.

Work has been done to characterize sequence motifs in sets of validated immunogenic peptides which may correlate with response to immunotherapy. Snyder *et al.* identified a substring signature which is shared with proteins derived from pathogens, and correlates with T cell response [119]. This is an intriguing idea that has not yet been replicated in other datasets. Conversely, there have been reports of T cell clones responding equally well to dissimilar peptides not having any shared motif [132], suggesting that, in the context of T cell recognition and activation, the primary sequence of the peptide is less important than its biophysical properties. Indeed, it seems plausible that due to the cross-reactivity of T cells, effective immune responses against tumour mutations may be at least partly the result of previous T cell activations in response to pathogenic peptides which, to the T cell receptor, appear biophysically similar. Recent studies showed that responses to checkpoint blockade depend on the presence of specific species of bacteria [133–136]. This result, when viewed in the context of cross-reactivity, suggests that the response seen may be at least partially due to cross-reactive priming of T cells to tumour antigens by bacterial antigens.

Another factor which may affect response to immunotherapies is the HLA genotype of the individual. Since MHC will present mutant peptides on the surface of cancer cells, the repertoire of presented peptides is dependent on the specific MHC molecules that individual has. This may affect response to immunotherapy in two ways: (1) restricting the TCR repertoire during T cell development, or (2) affecting the

probability that mutations are presented as neoantigens. Firstly, as TCR recombination during development is dependent on MHC presentation of self-peptides [137], TCRs that are “taught” about a large and diverse set of self-peptides may be better able to distinguish between self and non-self once in the periphery, or may result in a more diverse TCR repertoire [138,139]. Secondly, having a set of MHC molecules that can present many self-peptides should improve the chance that any given mutation will be presented as a neoantigen, compared to MHC molecules that can only present a narrow and restricted set of self-peptides. There is evidence for this in advanced melanoma and advanced non-small cell lung cancer, where individuals who were heterozygous at all HLA class I loci (and thus are able to present more peptides) showed better response to immune checkpoint blockade compared to those homozygous in at least one loci due to genetic variation or somatic loss of heterozygosity, independent of mutational load [140].

1.3 Immunoinformatics

Immunoinformatics exists at the nexus between computer science and immunology and is becoming an increasingly important tool for those working on cancer immunotherapy. What began as the development of tools predicting the binding ability of a peptide to MHC molecules [141,142] has grown into a field including, but not limited to, predicting HLA alleles from next-generation sequencing (NGS) data [143,144], presentation of mutated peptide sequences [145], peptide processing and transport prior to binding MHC [146], characterization of the T cell and B cell receptor repertoires [11,67], and analysis of immune-related gene expression [147,148]. As big data in immunology becomes even bigger and more abundant, there is a clear need for analyses that can pare down the overwhelming body of data into meaningful insights that can be applied to future studies. Additionally, one area of immunoinformatics that still has room for improvement is the prediction of peptides that elicit T cell reactivity, due to incomplete understanding of what makes a peptide immunogenic and immunodominant.

Genomic datasets are typically used for immunoinformatics due to their ability to provide information on mutational variants, gene expression, TCR recombination, and HLA variation. In the context of cancer immunology, a tremendous data resource is The

Cancer Genome Atlas (TCGA; https://cancergenome.nih.gov/). This international project worked to characterize over 10,000 individual tumours spanning 33 different tumour types, collecting tumour and matched normal tissue as well as limited clinical information. For each tumour, whole exome sequencing was performed on both normal and tumour tissue to detect somatic variants occurring in the tumour. Transcriptome sequencing (RNA-seq) was performed on most tumour samples to measure gene expression within the tumour. Despite one of the tumour specimen selection criterion being low immune infiltration, immunological insights can still be gained from these samples despite reduced immune cell content [149].

1.3.1 TCR annotation

Random TCR recombination events lead to unique CDR3 sequences in T cells. The CDR3 sequence of rearranged TCRs is unique to each cell, and can therefore be used as a barcode to track each T cell. Surveying the CDR3 sequences in a quantifiable manner allows the measurement of clonal expansion of T cells. The random variation present in the CDR3 region does not make it amenable to classical alignment-based algorithms for annotation since these approaches require a reference sequence for alignment. Since each CDR3 sequence is flanked by one of a limited set of V and J genes, local alignments are performed for these regions, and the sequence between these regions is inferred to be the CDR3 sequence. MiTCR [150] is one such algorithm, relying on the conserved cysteine and phenylalanine/tryptophan codons which flank the CDR3 region.

The ability to annotate TCR sequences is dependent on sequence read length. Due to highly variable CDR3 sequence, assembly of short reads (which do not completely span the CDR3 region) into larger contigs may introduce false chimeric receptor sequences due to overlapping sequence within the CDR3 region. Truly distinct CDR3 sequences may only differ by a single nucleotide, therefore, partial sequence of the CDR3 is not sufficient to uniquely identify a TCR. The most accurate algorithms rely on single reads to detect CDR3s, however, this requires that single reads can span the entire CDR3 region, creating a technical upper bound on the length of CDR3s that can be recovered. CDR3-beta is typically 45 nts long [11], so 50 bp reads (standard for

many existing datasets) have limited ability to detect this CDR3 length with sufficient V and J gene sequence left for alignment. Longer CDR3s are undetectable using 50 bp read datasets. Typically, TCR-seq experiments are designed to maximize the coverage of the CDR3 region to obtain the maximal amount of information regarding the TCR repertoire [11,67]. The utility of bulk sequencing datasets (RNA-seq) for TCR repertoire characterization was previously unknown (but was addressed by my work in Chapter 3: Exploring the TCR repertoire of solid and liquid tumours by bulk RNA-seq).

1.3.2 HLA allele nomenclature

The HLA locus is the most polymorphic region of the human genome, with over 10,000 known class I HLA alleles in the human population [26]. As such, a standardized nomenclature was developed to name and distinguish HLA alleles [151]. First, the HLA gene is given (*HLA-A*), followed by two digits describing the allele family (*HLA-A*02*). Four-digit resolution describes protein coding changes to the allele (*HLA-A*02:01*). Six- and eight-digit resolution describe synonymous nucleotide changes within coding regions and intronic variants, respectively (*HLA-A*02:01:01, HLA-A*02:01:01:01*). Many of the allelic variants differ only in non-coding sequence, but for peptide-MHC predictions, only knowledge of protein coding variants (4-digit HLA allele resolution) is required.

1.3.3 HLA genotype predictions

Knowledge of which MHC molecules are present in an individual is a requirement to perform personalized peptide-MHC binding predictions. Motivated by this, numerous HLA calling algorithms have been developed to extract HLA allele information from NGS data. While most of the algorithms are quite accurate at determining 4-digit HLA alleles from NGS reads 75 bps or longer, they all suffer from ambiguity in discriminating alleles from shorter 50 bp reads due to insufficient alignment with reference sequences. The simplest solution to this problem is to sequence longer reads, however, many existing datasets only contain 50 bp reads. One solution for these existing datasets may be the integration of multiple data types (RNA-seq and exome-seq) to increase the amount of sequence that can be analyzed, or may include new methods which infer HLA alleles using observed population HLA haplotype frequencies to resolve ambiguous calls

[152,153]. Additionally, new algorithms utilizing a combination of alignment and expectation maximization methods, which select the genotype which explains the greatest number of observed sequence reads, have demonstrated improved performance on shorter read datasets [154].

1.3.4 Peptide-MHC binding predictions

Peptide-MHC binding prediction algorithms, initially designed to find the optimal epitopes from viral genomes for vaccine development, have more recently begun to be applied to tumour genomes in the hopes of determining which mutant peptides are immunogenic [145,155–157]. These prediction algorithms require empirically determined training data of pMHC interactions to predict binding, with more training data leading to better predictions. Due to the highly polymorphic HLA locus in the human genome, it is not feasible to obtain *in vitro* training data for peptide binding to all alleles. Additionally, early attempts to construct basic peptide-MHC binding predictors quickly revealed that purely position-specific scoring matrices (PSSMs) or sequence motif-based predictors were insufficient to explain the peptide-binding preferences of MHC [158]. To date, the most successful approach has been neural-network-based, using training data from known alleles and inferring predictions for new alleles. The NetMHCpan algorithm is one such algorithm to take this approach, and when tested against other algorithms on experimentally-derived data not included in the training data for any algorithm, has been shown to be the best pan-specific class I pMHC binding prediction algorithm [159].

Training data for peptide-MHC interactions generated by different labs is deposited at the Immune Epitope Database (IEDB) [30], and contains information on the binding affinity of these interactions. Typically, binding affinity is inferred from the half maximal inhibitory concentration (IC_{50}), or the concentration of peptide required to competitively bind half of the available MHC. As such, the output of NetMHCpan provides a predicted IC_{50}, as well as a percentile rank measurement comparing the strength of that peptide-MHC binding interaction compared to the binding strength for a set of random peptides to the same MHC. Classically, a predicted IC_{50} threshold of < 50 nM is used to classify strong binders, and < 500 nM is used to classify weak

binders [160]. More recently, it has been suggested that a percentile rank threshold may control for variable median IC_{50} thresholds for different MHC molecules [161], though this threshold assumes that all MHCs present peptide repertoires of equal size. Currently, there is insufficient evidence to support this assumption [70], and within the literature both thresholds have been used to gain meaningful information. In the context of creating peptide vaccines against infectious agents, where the goal is to determine the most immunogenic peptide(s) from an entire foreign proteome, the percentile rank threshold may be optimal [162,163]. Conversely, to predict which mutated self-peptides (if any) present in a tumour will yield cancer neoantigens, where T cell tolerance exists, the IC_{50} threshold may perform better [164]. Thus, the best choice for binding threshold metric is likely application-dependent, and the optimal threshold for predicting the subset of self-peptides presented in an individual is unknown.

1.3.5 Methods for predicting immunogenicity of presented peptides

While peptide-MHC binding is necessary for immunogenicity, it is not sufficient. To be truly immunogenic, the peptide must also interact with the TCR to elicit a T cell response. Traditional approaches for predicting protein-protein interactions, such as molecular dynamics simulations, are currently intractable for TCR-pMHC [165]. This is due to the limited number of experimentally derived crystal structures for pMHCs and TCRs, and the enormous level of complexity that occurs at their interface (a variable MHC molecule binds a variable peptide sequence, which contacts a variable T cell receptor). Therefore, a more inferential approach is to reduce the problem to co-occurrence, identifying putative pMHCs and TCRs that are found together. The main assumptions to this approach are that TCRs and pMHCs that co-occur also interact, and that sequence-based information is sufficient to identify patterns in TCR and pMHC interactions. The former is only able to be confirmed by *in vitro* validation, as it may be that co-occurring TCR-pMHCs are simply due to both parts being common in the population. The latter can be relaxed by considering the similarity of biophysical properties of an amino acid sequence rather than exact sequence, looking for TCRs of a certain “category” co-occurring with pMHCs of a certain “category.” This could lead to a better understanding of the biophysical laws governing cross-reactivity, and would

allow for more complex comparisons between mutanomes and TCR repertoires in the context of T cell recognition.

More direct ways of inferring the immunogenicity of presented peptides do not require knowledge of the TCR repertoire, and instead focus on specific characteristics of the presented peptides. This may involve incorporating binding predictions for the wildtype as well as the mutant peptides, and identifying mutant peptide sequence similarity to peptides derived from infectious agents, to determine how “non-self” the neoantigen will appear to the immune system [119,166]. Mutant peptides that have sequence similarity to known infectious agents may be better able to elicit a T cell response due to pre-existing priming of T cells towards these peptides, or due to as-of-yet undiscovered sequence “motifs” present in pathogen-derived sequences. Likewise, mutant peptides where the corresponding wildtype peptide also binds may be less immunogenic due to existing tolerance to the wildtype peptide by the TCR repertoire, though there is evidence that existing tolerance does not preclude T cell reactivity towards mutated peptides [167]. Other methods of predicting immunogenicity of peptides involve determining the hydrophobicity of amino acids at the residues that typically contact the TCR. Immunogenic peptides have been observed to have increased hydrophobicity at these residues [168]. It is likely that future predictors will use a concert of these and other undiscovered metrics to predict immunogenicity of peptides, aided by increased availability of data for validated positive and negative TCR-pMHC interactions.

1.4 Thesis overview

The aim of this thesis is to adapt and apply novel and existing immunoinformatic methods to cancer datasets to identify general patterns and relationships between the immune system and cancer in a pan-cancer context, not dependent on the cancer type. This involves the creation of a strategy to predict cancer neoantigens derived from SNVs from tumours, and correlation of this neoantigen burden with survival and markers of immune inhibition, amenable to checkpoint blockade therapies. It involves extraction of TCR sequences from RNA-seq datasets to gain additional information from these existing datasets, beyond the scope of their original design, with demonstrated

utility in data from solid tumours and lymphomas. Finally, it develops a novel metric (self-immunopeptidome size) to classify individuals based on their ability to present peptides on class I MHC molecules. This thesis contributes towards a better understanding of the interaction between T cells and cancer cells, which can inform future strategies to improve immunotherapies in cancer.

Chapter 2 details a novel approach to predicting neoantigens from SNV mutations in the TCGA dataset and associates the neoantigen load with clinical outcomes. Here, we surveyed all six tumour types which were available from TCGA at the time and found patterns that were not tumour-type specific. Neoantigen load was associated with increased patient survival. Moreover, the corresponding tumours had higher CTL content, inferred from *CD8A* gene expression, and elevated expression of the CTL exhaustion markers *PDCD1* and *CTLA4*. Neoantigens were very scarce in tumours without evidence of CTL infiltration. These findings suggest that the abundance of predicted immunogenic mutations may be useful for identifying patients likely to benefit from checkpoint blockade and related immunotherapies.

Chapter 3 describes the extraction of TCR sequence information directly from RNA-seq data derived from 6,738 tumour and 604 control tissues. This method circumvents the need for PCR amplification of TCR template, reducing cost by avoiding dedicated sequencing of just the TCR locus and allowing analysis of existing RNA-seq datasets. TCR information is provided in the context of global gene expression, allowing integrated analysis of extensive RNA-seq data resources. It provides evidence that abundant, shared TCR sequences found in multiple distinct tumours recognize common viral antigens. Further, this chapter describes the application of this method to RNA-seq from sorted cell subsets from peripheral T cell lymphomas, demonstrating increased sensitivity and diagnostic ability of RNA-seq over conventional flow cytometry.

Chapter 4 uses comprehensive peptide-MHC binding predictions for the entire human proteome to the set of all class I MHC to generate the human immunopeptidome, a novel resource allowing individual HLA genotypes to be numerically quantified based on the number of self-peptides they are predicted to present (the self-immunopeptidome). Self-immunopeptidome sizes correlate with survival, with larger self-immunopeptidome sizes having improved survival outcomes in

a pan-cancer population. The size of the self-immunopeptidome can be used to predict neoantigen load in tumours from their mutational load, suggesting a potential clinical utility of HLA genotypes as a biomarker for predicting response to immunotherapy. Further, we identified evidence of immune editing in the TCGA data – containing fewer presented mutations than would be expected by chance (based on that tumour's HLA genotype), and identified specific positions within presented peptides which, when mutated, are predicted to influence immunogenicity of the presented peptide.

Finally, chapter 5 gives an overall analysis of the research and conclusions presented in this thesis within the context of current work being done in the field. This chapter also comments on strengths and limitations of the thesis research and presents possible future research directions in the field drawing on the work of this thesis.

Contemporaneously to the work described in this thesis, I contributed towards the TCGA PanCancer Atlas [169] efforts to comprehensively characterize the immune landscape across the entire TCGA cohort [149]. I provided neoantigen predictions for SNV mutations built upon the work presented in chapter 2. I also built upon the TCR extraction work presented in chapter 3 and provided the analysis pipeline to complete these extractions on the entire cohort. As the PanCancer Atlas project was a collaborative consortium effort, the specific methods and results pertaining to those analyses are not included in this thesis, which instead focuses on my primary work only.

Chapter 2: Neoantigens predicted by tumour genome meta-analysis correlate with increased patient survival

2.1 Introduction

The accumulation of somatic mutations underlies the initiation and progression of most cancers, by conferring upon tumour cells unrestricted proliferative capacity [74]. The analysis of cancer genomes has revealed that tumour mutational landscapes [91] are extremely variable among patients, among different tumours from the same patient, and even among the different regions of a single tumour [107]. There is a need for personalized strategies for cancer therapy that are compatible with mutational heterogeneity, and in this regard immune interventions that aim to initiate or enhance anti-tumour immune responses hold much promise. Therapeutic antibodies and chimeric antigen receptor (CAR) technologies have shown anti-cancer efficacy [170], but such antibody-based approaches are limited to cell surface target antigens [171–175]. By contrast, most tumour mutations are point mutations in genes encoding intracellular proteins. Short peptide fragments of these proteins, after intracellular processing and presentation at the cell surface as MHC ligands, can elicit T-cell immunoreactivity. Further, the presence of TIL, in particular $CD8^+$ T cells, has been associated with increased survival [98–100,176–178] suggesting that the adaptive immune system can mount protective anti-tumour responses in many cancer patients [170,179]. The antigen specificities of tumour infiltrating T cells remain almost completely undefined [125] but there are numerous examples of cytotoxic T cells recognizing single amino acid coding changes originating from somatic tumour mutations [112,126–130,180]. Thus, the notion that tumour mutations are reservoirs of exploitable neoantigens remains compelling [180].

For a mutation to be recognized by $CD8^+$ T cells, the mutant peptide must be presented by MHC I molecules on the surface of the tumour cell. The ability of a peptide to bind a given MHC I molecule with sufficient affinity for the peptide-MHC complex to be stabilized at the cell surface is the single most limiting step in antigen presentation and T cell activation [181]. Recently, several algorithms have been developed that can

predict which peptides will bind to given MHC molecules [30,158,182–184], thereby providing guidance into which mutations are immunogenic.

TCGA (http://cancergenome.nih.gov/) is an initiative of the National Institutes of Health that has created a comprehensive catalogue of somatic tumour mutations identified using deep sequencing. As a member of The Cancer Genome Atlas Research Network our centre has generated extensive tumour RNA-seq data. Here, we have used public TCGA RNA-seq data to explore the T-cell immunoreactivity of somatic missense mutations across six tumour sites. This type of analysis is challenged not only by large numbers of mutations unique to individual patients, but also by the complexity of personalized antigen presentation by MHC arising from the extreme HLA allelic diversity in the outbred human population. Previous studies have explored the potential immunogenicity of tumour mutations [145,185,186], but these have been hampered by small sample size and the inability to specify autologous HLA restriction. Recently we described a method of HLA calling from RNA-seq data that shows high sensitivity and specificity [143]. Here, we have obtained matched tumour mutational profiles and *HLA-A* genotypes from TCGA subjects and used these data to predict patient-specific mutational epitope profiles. The evaluation of these data together with RNA-seq derived markers of T-cell infiltration and overall patient survival provides the first comprehensive view of the landscape of potentially immunogenic mutations in solid tumours.

2.2 Methods

2.2.1 TCGA mutation annotation files

Mutation Annotation Files (MAF) for unrestricted TCGA cancer sites were downloaded from https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumour/. We parsed every available MAF file regardless of level (https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+%28MAF%29+Specification) however only listed variants predicted to yield non-synonymous missense coding mutations and associated with a predicted RefSeq identifier at the specified genomic location were ultimately tracked. The MAF format specification enabled the selection of putative whole-genome shotgun screen variants that had been verified by orthogonal methods. The screen identified a total of 74,535 verified missense SNVs

from 1,069 TCGA patients and seven cancer sites, including BRCA (breast invasive carcinoma), COAD (colon adenocarcinoma), READ (rectum adenocarcinoma), GBM (glioblastoma multiform), KIRC (kidney renal clear cell carcinoma), LUSC (lung squamous cell carcinoma), and OV (ovarian serous cystadenocarcinoma). Parsing scripts, written in Perl, tallied corresponding RNA-seq BAM filenames for each of the 1,069 TCGA patients for use in conjunction with HLA prediction and gene expression profiling.

2.2.2 HLA predictions

RNA-seq BAM files for each of the 1,069 subjects were downloaded from CGhub and used directly as input for HLAminer [143]. HLAminer was run with default values, in parallel on a computer cluster. The two highest-scoring 4-digit HLA predictions for the *HLA-A* locus were retained (highest score at ranks 1 and 2). Patients with 4-digit HLA predictions that were ambiguous, that is, with two or more 4-digit HLA alleles scoring equally, were excluded from analysis. RNA-seq read length strongly influences the performance of HLA calling, and ambiguous HLA calls from tumour types where only short reads (50 nt) were available (lung, breast and kidney) represented that largest source of attrition of TCGA subjects from the meta-analysis. HLAminer predictions, including the genes, rank, group allele, coding allele, score, expect value, confidence and number of predictions were stored in a MySQL relational database. A custom script was developed to integrate the automated HLA predictions with SNV-specific information, and used as input for HLA epitope predictions.

2.2.3 HLA ligand binding predictions

A tab-separated file that listed all 74,535 filtered SNVs along with the predicted amino acid coding mutation and protein sequence was split by cancer type and each used as input for PERL scripts designed to query IEDB (http://www.iedb.org/) offline (http://tools.immuneepitope.org/analyze/html_mhcibinding20090901B/download_mhc_I_binding.html) as previously described [145]. Briefly, entire protein sequences were submitted sequentially using default values in their unchanged wildtype and mutated form based on mutant position predictions. When supported, 8 - 11mer peptide prediction were selected, each with a specific HLA allele determined computationally

from RNA-seq data for the patient under scrutiny. The output epitope prediction was captured, parsed and all peptides encompassing the amino acid of interest were tracked, including binding prediction rank and score when applicable.

2.2.4 Gene expression from RNA-seq data

Raw sequence reads were extracted from the 1,069 BAM files using bam2fastq v.1.1.0. Extracted reads were subsequently aligned to the human reference genome and transcriptome (hg19, Ensembl v70) using the ultra-fast aligner STAR v.2.3.0e [187] with the following parameters: minimum / maximum intron size set to 30 and 500000 respectively, non-canonical, unannotated junctions were removed, maximum tolerated mismatches was set to 10, and the outSAMstrandField intronMotif option was enabled. The Cuffdiff command included with Cufflinks v.2.0.2 [188] was used to calculate the Fragments Per Kilobase of exon per Million fragments mapped [188] (FPKM) with upper quartile normalization, fragment bias correction, and multi read correction enabled. All other options were set to default.

2.2.5 Clinical data sets

TCGA clinical datasets were downloaded from https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumour/DISEASE_CODE/bcr/biotab/clin/. For each cancer site, we obtained clinical_drug_XXX.txt, clinical_follow_up_vX.X_XXX.txt, clinical_patient_XXX.txt and clinical_radiation_XXX.txt. The files were parsed, and pertinent clinical information extracted and saved into a MySQL relational database.

2.2.6 Data analysis

Pertinent data was extracted from the MySQL database using custom queries, and the results were saved to tab delimited text files. These files were read into R v3.0.1 [189] for further statistical analysis. Colon and rectum cancers were combined for all analyses as colorectal cancer. A single colorectal patient with a total mutation count 20.3 standard deviations away from the mean mutation count of all patients was removed from all analysis.

To count the overall number of putatively immunogenic mutations for each patient, we summed the total number of point mutations that contained a peptide predicted to be presented by the MHC molecules encoded by the *HLA-A* alleles identified, unambiguously, for that patient. The requirement of unambiguous *HLA-A* prediction resulted in a sample size of 515. To count the number of putatively immunogenic mutations, we first took the "best" peptide for each point mutation which were those with the highest predicted binding affinity (lowest IC_{50}) to its respective autologous MHC variant. We filtered these peptides by keeping those that had an IC_{50} value below 500 nM. We then filtered these peptides to those that were expressed at a level higher than the median expression for their given gene. We further filtered these peptides to those where the *HLA-A* gene expression was higher than the median of all *HLA-A* gene expression values. These cut offs were selected to maximize the probability that a given peptide was able to be seen by a T cell receptor, in which case it should be highly expressed and bind to an MHC variant that is also highly expressed. The number of peptides which passed these criteria was used as the number of predicted immunogenic mutations for each patient.

2.2.7 Statistical analysis

We modified a random re-assignment method, described previously [190], to test the significance of associations with TIL gene expression markers. First, the percent of mutations that belonged to tumours with above median *CD8A* expression was calculated. Next, counts of mutations were randomly re-assigned to tumours 1,000,000 times using the boot package [191] in R. The percent of total mutations belonging to tumours with above median *CD8A* expression was calculated after each random re-assignment, and the bootstrap *P* value was equal to the proportion of randomizations where the number of mutations belonging to tumours with above median *CD8A* expression was equal to or greater than the number of mutations belonging to tumours with above median *CD8A* expression in the original, non-randomized data. This same method was used to test the significance of associations between the presence of predicted immunogenic mutations and elevated expression of all three genes, *PDCD1, CTLA4* and *CD8A*.

Survival times were calculated as the number of days from initial pathological diagnosis to death or the number of days from initial pathological diagnosis to the last time the patient was known to be alive was recorded. These times were used in the construction of the Kaplan-Meier survival curves and Cox proportional hazard models. Potential confounders age, gender, cancer and tumour stage were examined. The R survival package [192] was used to construct Kaplan-Meier curves and fit the univariate and multivariate Cox proportional hazard models. 512 patients were used in the survival analysis investigating *CD8A* and *HLA-A* after removing 3 patients without survival information. The 16 brain tumour patients were excluded from the analysis as they were missing tumour stage information. The 24 breast patients were also excluded from analysis as the low mortality rate (1/24) was not informative. Additionally, 7 patients were not used in the survival analysis as their prognostic information was incomplete. This resulted in a sample size of 468 for the multivariate survival analysis.

2.2.8 Hive plots

An R script was designed to create hive plot input files from the original data, converting from a table format to the graph format, DOT. These input files were imported into jhive v0.0.18 (http://hiveplot.com/distro/jhive-0.0.18.zip) to create the hive plots [193].

2.3 Results

2.3.1 Summary of available data

Raw TCGA RNA-seq data plus clinical metadata and complete profiles of sequence verified missense mutations were obtained with permission from the Cancer Genomics Hub (https://cghub.ucsc.edu). Our analysis covers six tumour sites, including colon and rectum (combined as colorectal), ovary, breast, brain, kidney and lung. These were the only tumour sites with complete and non-embargoed data at the time of this study. The RNA-seq data were first processed using HLAminer [143] to predict, at 4-digit resolution, the two *HLA-A* alleles carried by each subject. Data from 515 patients with unambiguous *HLA-A* calls were processed further. The distribution of missense mutation counts across patients with different tumour types is shown in Figure 2.1. For each of the 22,758 total missense mutations, we evaluated binding of all possible 8-

11mer mutant and wild-type peptide variants to autologous *HLA-A* encoded MHC proteins using the IEDB T cell epitope-MHC Binding Prediction Tool [30] (http://www.iedb.org/). We focused our analysis on *HLA-A* alleles because (1) MHC I proteins (encoded by *HLA-A, -B* and *-C* genes) present antigens to $CD8^+$ cytotoxic T cells, which are the subset of T cells most strongly linked to patient survival, and (2) *HLA-A* alleles of MHC I yield the most accurate peptide binding affinity predictions by IEDB and most other algorithms due to the abundance of *HLA-A* specific training data [194]. All mutational data, RNA-seq derived *HLA-A* calls, IEDB epitope predictions, RNA-seq derived gene expression values, and clinical metadata were compiled in a MySQL database for analysis.

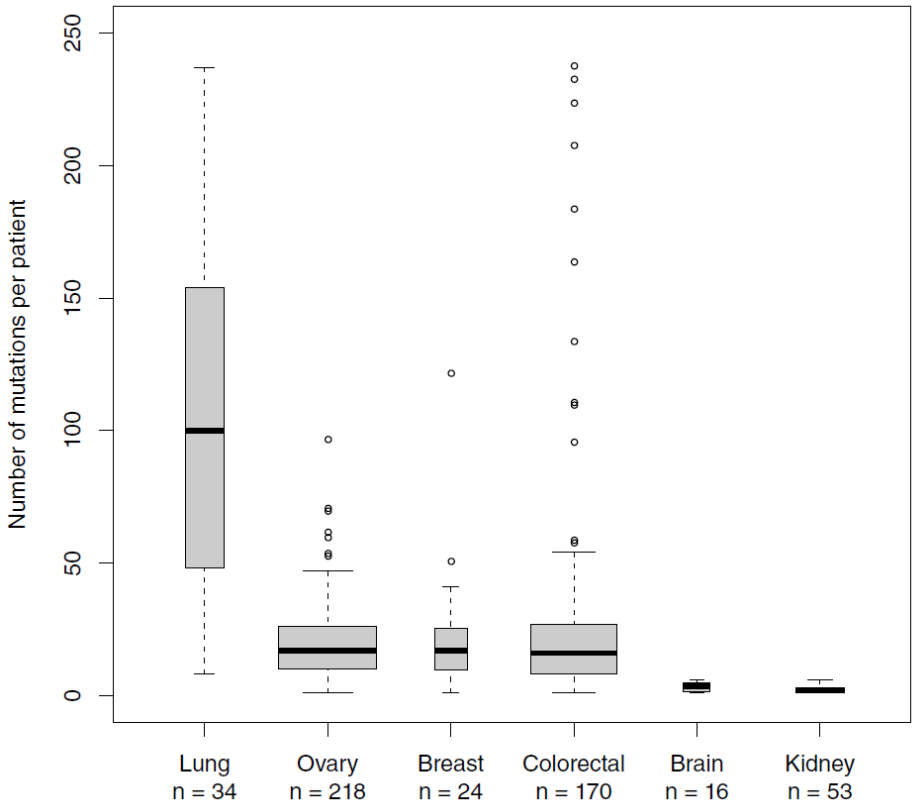

Figure 2.1: Boxplots for the number of mutations per patient for each cancer type. The Y-axis is cut off at 250 mutations for better visualization of the majority of the data. The dark horizontal bar shows the median, whereas the box encompasses the interquartile range (middle 50 % of the data). Whiskers reach the farthest data point that is within 1.5x the interquartile range from the nearest box edge (quartile). Box width is proportional to the sample size (Lung: 34, ovary: 218, breast: 24, colorectal: 170, brain: 16, kidney: 53).

2.3.2 *CD8A* expression is associated with survival

We first asked if we could reproduce the known association between increased numbers of tumour-infiltrating $CD8^+$ T cells and increased overall survival [98–100,176–178]. $CD8^+$ TIL levels are usually measured by immunohistological staining. To interrogate RNA-seq data, we used the expression of *CD8A* (one component of the CD8 dimer), as a surrogate for $CD8^+$ TIL levels. We observed significantly higher overall survival for patients with high *CD8A* expression than for those patients with low *CD8A* expression (HR = 0.71, 95 % CI = 0.53 to 0.94, $P = 1.7 \times 10^{-2}$) (Figure 2.2A). Likewise, the data recapitulated the known association between high *HLA-A* expression and improved overall survival [195–199] (HR = 0.59, 95 % CI = 0.44 to 0.81, $P = 8.6 \times 10^{-4}$) (Figure 2.2B). Based on these positive findings with established T cell and MHC markers, we proceeded to evaluate candidate peptide epitopes, which represent the third molecular component required for T cell recognition and destruction of target cells.

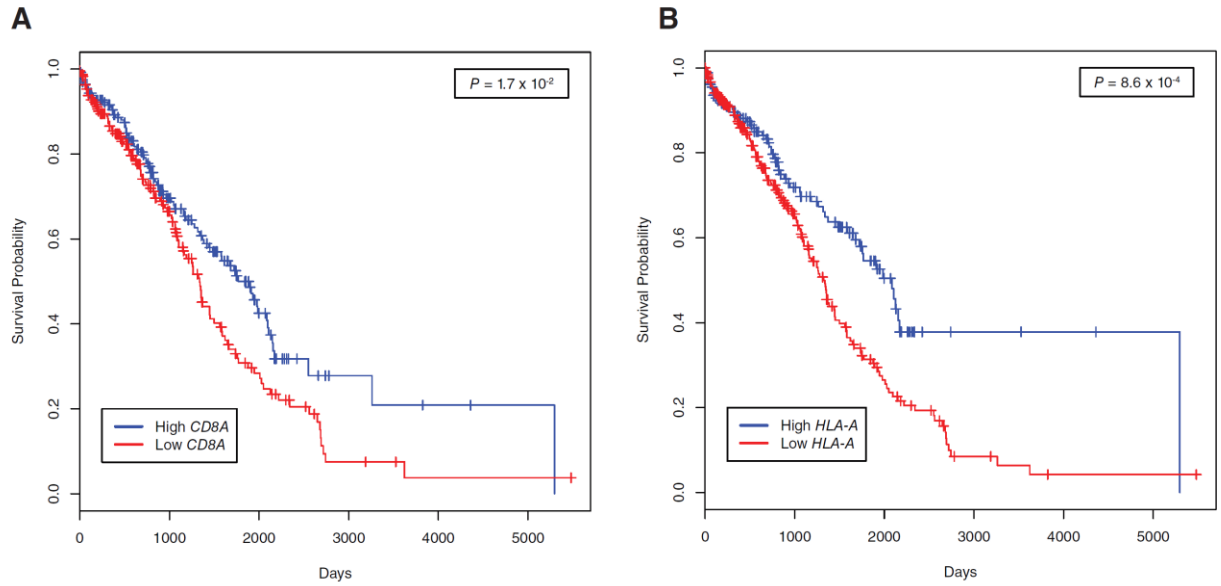

Figure 2.2: Overall survival for patients based on *CD8A* or *HLA-A* expression. Kaplan-Meier curves were constructed to look at the difference in survival of patients (n = 512) with low and high expression levels of **(A)** *CD8A* or **(B)** *HLA-A*. Patients were split into two groups based on the median expression value. Patients with high expression showed increased survival compared to those with low expression of either **(A)** *CD8A* (HR = 0.71, 95 % CI = 0.53 to 0.94, $P = 1.7 \times 10^{-2}$) or **(B)** *HLA-A* (HR = 0.59, 95 % CI = 0.44 to 0.81, $P = 8.6 \times 10^{-4}$). Tick marks on the graph denote the last time when survival status was known for living patients.

2.3.3 Tumours with high numbers of missense mutations have more tumour infiltrating lymphocytes

Initially, we asked if there is a relationship between overall mutation count and $CD8^+$ TIL. Ranking patients by decreasing *CD8A* expression and displaying the mutation count for each patient's tumour revealed a skewed distribution whereby tumours with low *CD8A* expression had sparse mutations and tumours with high mutation counts were among those with elevated *CD8A* expression (Figure 2.3A). Tumours with above median *CD8A* expression contained 73.6 % of the total mutations ($P = 2.0 \times 10^{-6}$ by iterative randomization and resampling as described in Methods). However, there was no association between total mutation count and overall survival (Figure 2.3B) (HR = 0.91, 95 % CI = 0.68 to 1.23, $P = 5.5 \times 10^{-1}$).

2.3.4 Tumour missense mutations that have predicted immunoreactivity are associated with increased survival

We reasoned that missense mutations yielding peptides with poor MHC I binding would be immunologically silent and hence likely to obscure any association between missense mutations, anti-tumour immunoreactivity, and survival. To address this, we repeated the above analysis focusing on those mutations that were most likely to be immunogenic by several criteria, including (1) the expression of the gene in the tumour bearing the mutation was above the median expression level of that same gene in all tumours, (2) *HLA-A* expression was above the median expression in all tumours, and (3) the predicted autologous HLA-A binding affinity of the best scoring peptide containing a given mutation had an IC_{50} value of 500 nM or less. This value has been estimated, experimentally, to be the affinity necessary for an epitope to elicit an immune response [160]. Applying these filters, the predicted immunogenic mutation count was zero in 334 patients. The remaining 181 patients had predicted immunogenic mutation counts ranging from 1 to 147, with a median of 3. The predicted immunogenic mutation count showed a strong relationship with tumour *CD8A* expression, where tumours with higher numbers of such mutations had higher *CD8A* expression (Figure 2.3C). Of all predicted immunogenic mutations, 84.7 % were in tumours with above median *CD8A* expression ($P = 1.0 \times 10^{-6}$). We did not see any relationship between predicted

immunogenic mutation count and *CD4* expression by tumours (P = 6.9 x 10^{-1}) (Figure A.1), consistent with the fact that we had assessed epitopes presented by MHC class I, which is recognized exclusively by $CD8^+$ T cells. Interestingly, patients with tumours containing at least one predicted immunogenic mutation showed markedly increased overall survival compared to those without predicted immunogenic mutations (HR = 0.53, 95 % CI = 0.36 to 0.80, P = 2.1 x 10^{-3}) (Figure 2.3D).

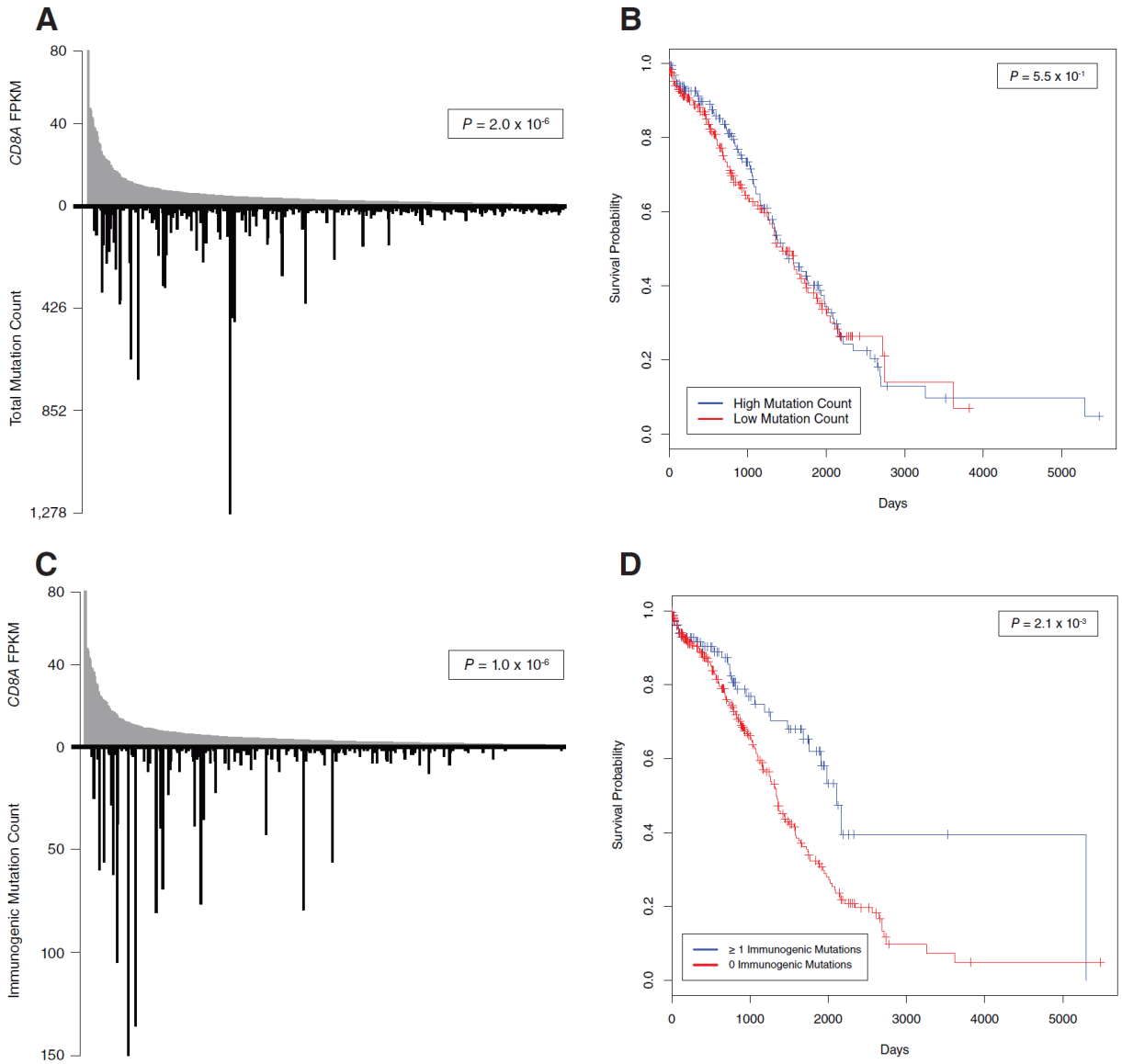

Figure 2.3: The total number of mutations in tumours is not associated with survival, while the number of predicted immunogenic mutations is associated with survival. A "skew plot" was made for all patients (n = 515), ordering patients along the x-axis according to their *CD8A* expression. Each patient's *CD8A* expression was plotted above the x-axis, and **(A)** total mutation count or **(C)** predicted

immunogenic mutation count was plotted below the x-axis. 73. 6% of the total mutation count belonged to patients with above median *CD8A* expression (P = 2.0 x 10^{-6}), and 84.7 % of the total predicted immunogenic mutation count belonged to patients with above median *CD8A* expression (P = 1.0 x 10^{-6}). Kaplan-Meier curves were constructed to look at the difference in survival between patients with low versus high numbers of mutations. Patients (n = 468) were split into two groups based on the median mutation count. There was no difference in survival between the two groups when stratifying on total mutation count **(B)** (HR = 0.91, 95 % CI = 0.68 to 1.23, P = 5.5 x 10^{-1}), but there was a statistically significant difference between the two groups when stratifying on predicted immunogenic mutation count **(D)** (HR = 0.53, 95 % CI = 0.36 to 0.80, P = 2.1 x 10^{-3}). Tick marks on the Kaplan-Meier graphs denote the last time when survival status was known for living patients.

To further examine this association, we fit a model including all available prognostic factors (age, gender, cancer type, and tumour stage), as well as predicted immunogenic mutations. This model also showed significantly improved overall survival for patients with predicted immunogenic mutations relative to those without (HR = 0.50, 95 % CI = 0.31 to 0.80, P = 3.9 x 10^{-3}), indicating that the effect of predicted immunogenic mutations was independent of the other prognostic factors. Fitting a model which contained an interaction between cancer type and predicted immunogenic mutations did not yield a significant result (P = 9.2 x 10^{-1}), indicating that the prognostic effect is not limited to a specific cancer diagnosis. Given that tumour *HLA-A* expression alone is a known indicator of favourable patient survival (Figure 2.2B), we asked if the number of predicted immunogenic mutations provides additional predictive value independent of *HLA-A* expression. After removing the *HLA-A* expression requirement from the definition of a predicted immunogenic mutation, we fit a model including all prognostic factors to the subset of patients with high (above median) tumour *HLA-A* expression. Within this subset of patients, we observed that patients with at least one predicted immunogenic mutation had a significantly lower relative risk of death than those without (HR = 0.44, 95 % CI = 0.22 to 0.88, P = 2.0 x 10^{-2}). Evaluating the reciprocal group of patients with low (below median) *HLA-A* expression, where the potential of immunogenic mutations to elicit *bona fide* anti-tumour responses is expected to be curtailed, there was no significant association between the presence of predicted immunogenic mutations and survival (HR = 1.30, 95 % CI = 0.83 to 2.04, P = 2.6 x 10^{-1}). The results from all survival analyses are summarized in Table 2.1.

Table 2.1: Summary of survival analysis.

Predictor	HR	95 % CI	P-value	
CD8A expression	0.71	0.53 – 0.94	1.7×10^{-2}	*
HLA-A expression	0.59	0.44 – 0.81	8.6×10^{-4}	**
Total mutations	0.91	0.68 – 1.23	5.5×10^{-1}	
[&]Predicted immunogenic mutations	0.50	0.31 – 0.80	3.9×10^{-3}	**
[&]Predicted immunogenic mutations, Low *HLA-A*	1.30	0.83 – 2.04	2.6×10^{-1}	
[&]Predicted immunogenic, High *HLA-A*	0.44	0.22 – 0.88	2.0×10^{-2}	*

(*) denotes P-values < 0.05. (**) denotes P-values < 0.005. (&) denotes analysis that accounted for variation from known prognostic factors.

2.3.5 Predicted immunogenic mutation counts correlate with the expression of T cell exhaustion markers

PDCD1 and CTLA4 are T cell surface molecules that can inhibit anti-tumour T cell responses [200,201]. Blockade of these inhibitory receptors by targeted monoclonal antibodies can disinhibit anti-tumour immunity and improve clinical outcomes [109,110,202–206]. Given that many patients in the current study had clinically significant cancer despite having predicted immunogenic mutations and $CD8^+$ TIL, we asked if there was an association between immunogenic mutation load and expression of *PDCD1* or *CTLA4*. We found that patients with higher numbers of predicted immunogenic mutations had increased expression of not only *CD8A* but also *PDCD1* and *CTLA4.* Displaying these values in a 3-way hive plot [193] highlights the association between these T cell markers and immunogenic mutation load (Figure 2.4). Significance was assessed by iterative randomization and resampling (as described in Methods). Of all tumours with predicted immunogenic mutations, 45.9 % had above median expression of all three of *PDCD1, CTLA4,* and *CD8A* ($P = 1.0 \times 10^{-6}$).

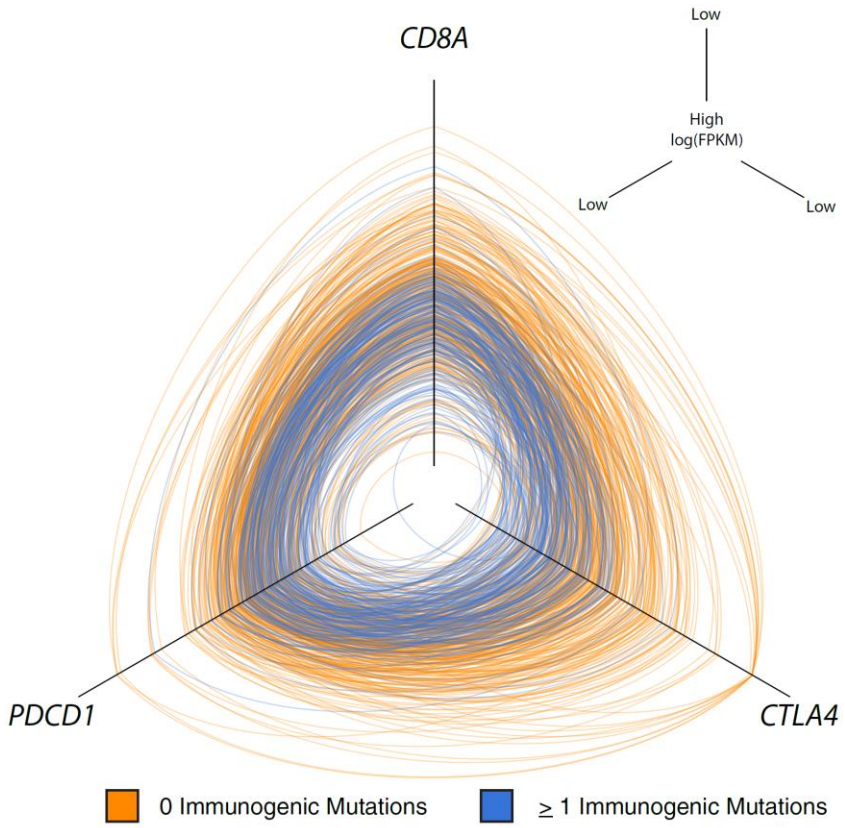

Figure 2.4: Hive plot showing tumours with high immunogenic mutation counts have higher expression of *CD8A, PDCD1,* and *CTLA4*. On each axis is the log expression value (log(FPKM)) for *CD8A* (top), *PDCD1* (left), and *CTLA4* (right). Values go from small to large moving towards the center of the plot. Each ring represents one patient, and the intersection with the axis represents that patient's value for that axis. Patients with 0 predicted immunogenic mutations are coloured orange, and patients with at least 1 predicted immunogenic mutation are coloured blue. Blue rings tend to cluster around the center of the plot indicating concordance between increased predicted immunogenic mutation count and elevated *CD8A, PDCD1,* and *CTLA4* expression ($P = 1.0 \times 10^{-6}$).

2.4 Discussion

The adaptive immune system opposes tumour development, and the elicitation of immunogenic cell death is a key component of both targeted immunotherapies and conventional treatment modalities including radiation and chemotherapy [207]. There is a robust association between T cell infiltration of solid tumours and favourable patient outcomes. Missense variants are the most frequent type of oncogenic mutation, which raises the question of whether missense mutations also underlie tumour

immunoreactivity. Exome analysis in mice has revealed specific missense mutations that encode MHC class I presented mutational epitopes that are capable of eliciting T cell mediated tumour rejection [122,127]. Moreover, human tumour exome sequencing studies have identified mutational epitopes recognized by autologous $CD8^+$ TIL [112,129,130,180]. However, from these investigations it appears that missense mutations with demonstrable endogenous immunoreactivity are relatively rare. They are a small minority of total missense mutations. It is likely the case that only one or a few mutations per tumour are immunodominant, and tumours with a higher mutational burden simply have an increased likelihood of bearing a highly immunogenic mutation. This is consistent with our results, where total mutations (Figure 2.3A) greatly outnumber mutations that are predicted to be immunogenic (Figure 2.3C), but the distributions are similar. Looking at cancers individually (Figure A.2) it is interesting that colorectal tumours, many of which had very high mutational loads, showed the strongest association between predicted immunogenic mutation counts and *CD8A* expression. Unfortunately, however, in the current meta-analysis the number of subjects varied widely among cancer types. A comprehensive evaluation of immunogenic mutations specific to individual cancer types remains an important topic for future study.

Our meta-analysis focussed exclusively on missense mutations because in addition to these being most abundant, they were sequence verified and therefore of high confidence. Moreover, they were amenable to evaluation using existing computational epitope prediction tools. We observed that nearly all patient tumours with high missense mutation counts also had elevated $CD8^+$ TIL inferred by *CD8A* expression, and elevated counts of predicted immunogenic mutations. However, the association was directional, with many tumours having high $CD8^+$ TIL but few or no predicted immunogenic mutations. This suggests that while the expression of immunogenic missense mutations may induce $CD8^+$ TIL responses in some tumours, in other tumours $CD8^+$ TIL may be attracted by other classes of mutation or other factors altogether. In patients with hereditary nonpolyposis colorectal cancer, microsatellite instability is the major determinant of dense tumour infiltration by activated $CD8^+$ T cells [208], thus, a mutator tumour phenotype may in general enhance immunoreactivity. Other classes of potentially immunogenic mutations require exploration, such as gene

fusions resulting from genomic rearrangements. Instances of tumours with high $CD8^+$ TIL but few immunogenic mutations may also be due to immune editing [127,209]. Specifically, tumour cells bearing highly immunogenic mutations may have been selectively eliminated by T cells, resulting in accumulation of $CD8^+$ TIL but fewer immunogenic mutations remaining to be detected.

The results of the present study have clinical implications. We have shown that patients with tumours bearing missense mutations predicted to be immunogenic have a survival advantage (Figure 2.3D). These tumours also show evidence of higher $CD8^+$ TIL, which suggests that a number of these mutations might be immunoreactive. The existence of these mutations is encouraging because in principle they could be leveraged by personalized therapeutic vaccination strategies or adoptive transfer protocols to enhance anti-tumour immunoreactivity. Likewise, patients with tumours showing naturally immunogenic mutations and associated TIL are potential candidates for treatment with immune modulators such as CTLA4 or PDCD1 targeted antibodies. There is evidence that such therapies are most effective against tumours infiltrated by T cells [210,211]. Our results indicate that tumours bearing predicted immunogenic mutations have not only elevated *CD8A* expression (Figure 2.3C) but also elevated expression of *CTLA4* and *PDCD1* (Figure 2.4), reinforcing the notion that these patients may be optimal candidates for immune modulation. Importantly, we observed that tumours with low levels of $CD8^+$ TIL invariably have far fewer immunogenic mutations. Such patients would be better suited to conventional therapy, or to immunotherapies such as chimeric antigen receptor modified T cells that target non-mutated antigens.

Chapter 3: Exploring the TCR repertoire of solid and liquid tumours by bulk RNA-seq

Primary sequence analysis of the highly variable CDR3 of rearranged TCR genes provides insight into the adaptive immune response. T cells recognize peptide epitopes presented on the surface of cells on MHC (major histocompatibility complex) molecules. CDR3 is the TCR motif that directly binds MHC-presented peptide epitopes and this binding interaction is the main factor conferring T cell antigen specificity. Typically, CDR3 sequence information is acquired by performing TCR-seq experiments on peripheral T cells isolated from blood [10,11]; amplifying the CDR3 region with a conserved C gene primer followed by 5'RACE [11], or a multiplexed set of V and J gene primers [67]. TCR-seq applied to tissue specimens can provide insight into tumour-infiltrating lymphocytes [177,212], T cells associated with autoimmune pathology [213–215] and infection [216], and the properties of normal primary and secondary lymphatic tissues [217,218]. Further, TCR-seq applied to lymphoblastic leukemias and lymphomas can identify the malignant clonally expanded cell and track minimal residual disease [219].

Conventional TCR-seq methods provide a detailed view of TCR diversity [5,11,67]. However, because they rely on targeted amplicon sequencing, they do not evaluate TCR variation in the context of the overall genetic diversity of the specimen from which the data are derived. NGS technology has made whole genome and transcriptome sequencing routine, and provided opportunities for extraction of immunological data, such as HLA types, using specialized software tools [143,144]. Here, we describe an optimized approach for T cell receptor CDR3 extraction from RNA-seq datasets from solid tumours, for the purpose of characterizing T cell populations present in the tumour environment (3.1 Profiling tissue-resident T cell repertoires by RNA sequencing). We then apply this approach to RNA-seq of sorted cell subsets from peripheral T cell lymphoma specimens and identify aberrant cells with increased sensitivity compared to conventional methods (3.2 Defining the clonality of peripheral T cell lymphomas using RNA-seq).

3.1 Profiling tissue-resident T cell repertoires by RNA sequencing

3.1.1 Introduction

Compared to TCR-seq, the main challenge in CDR3 extraction from tumour RNA-seq data is the disproportionally large number of non-TCR transcripts (Figure 3.1). For a pure lymphocyte population, only one in approximately 2,000 transcripts are TCR transcripts (see 3.1.2.4 Approximation of TCR transcript abundance from percent T cell infiltration) and in tissues T cells represent a minor cell type, further decreasing TCR transcript representation. This necessitates an analytical approach that is both fast and accurate for TCR extraction from tissue-derived RNA-seq datasets. Here, we describe the application and optimization of existing TCR clonotype annotation tools designed for TCR-seq datasets to RNA-seq datasets from solid tumours, and describe the extracted TCR repertoires from these samples.

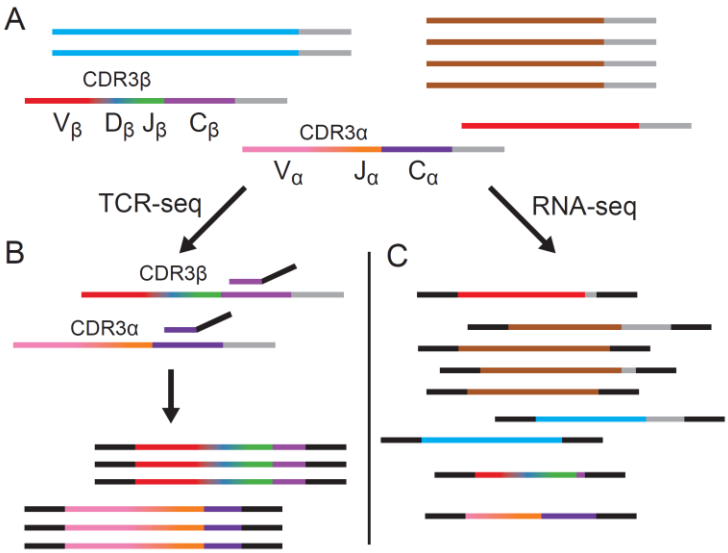

Figure 3.1: Schematic representation of TCR-seq versus RNA-seq. Horizontal lines represent mRNA transcripts with grey poly-A tails. Each colour represents a unique gene sequence. **(A)** A pool of all mRNA in a sample is depicted, which contains irrelevant transcripts (blue, brown, and red) as well as recombined TCR transcripts (multi-coloured). **(B)** TCR-seq involves selective amplification of the CDR3 region of TCR transcripts (displayed as a colour gradient) by RT-PCR, shown using a conserved C-gene primer (purple with black sequencing adapter tails) for the initial reverse transcription step and resulting, after PCR (not shown), in an enriched set of recombined TCR sequences. **(C)** RNA-seq employs shotgun sequencing, generating fragments from all transcripts present in the sample, which then have sequencing adapters ligated (black). The resulting sequencing library will contain fragments which, by chance, contain

CDR3 encoding sequence. Additionally, these libraries may contain fragments which share sequence similarity to recombined TCR sequences (ex. the red transcript), potentially leading to false-positives.

3.1.2 Methods

3.1.2.1 Ethics

The research described herein conformed to the Helsinki Declaration. All clinical specimens not part of The Cancer Genome Atlas were obtained previously [190] with informed consent by the BC Cancer Agency Tumour Tissue Repository (BCCA-TTR), which operates as a dedicated biobank with approval from the University of British Columbia-British Columbia Cancer Agency Research Ethics Board (BCCA REB; certificate #H09-01268).

3.1.2.2 Extraction of T cell receptor CDR3 sequences from RNA-seq data

We deployed MiTCR [150] v1.0.3, which is well suited for annotation of CDR3 sequences from sequencing reads. However, upon initial application of MiTCR to tumour RNA-seq data using the default parameters we identified hundreds of non-specific and out-of-frame CDR3 sequences per sample, which prompted us to explore alternative parameters. Closer inspection of the bogus CDR3s identified low similarity between these sequences and the putative flanking TCR V and J gene segments, suggesting that the false positives were spurious, non-TCR hits to TCR-like sequences elsewhere in the transcriptome. Therefore, we optimized settings using positive and negative control RNA-seq data. The positive control data set was comprised of TCR sequences generated *in silico*, as follows. V, (D), J, and C gene reference sequences for human T cell receptor alpha and beta chains were downloaded from the ImMunoGeneTics information system [220]. To generate each transcript, V, (D), J, and C genes were chosen randomly. Non-templated nucleotide addition and deletion frequencies at the V-(D)-J gene junctions were modelled from observed frequencies in normal TCR beta repertoires [11]. Due to the absence of D genes in the alpha chain, the number of bases added between the V and J genes was selected by averaging the number to add to both the V-D and D-J junctions in a beta chain. Out of frame transcripts and those that contained stop codons were removed. Full length recombined TCR transcript sequences were run through MiTCR using stringent alignment

parameters (minimum V and J alignment length both set to 20 in the XML parameter file; default value is 12) to annotate the CDR3 region in the transcript, and to ensure the *in silico* recombination created a CDR3 sequence able to be detected by MiTCR. 10,000 transcripts each of alpha and beta were generated, with 8,573 alpha and 8,804 beta sequences successfully being identified by MiTCR and being used as the source for the positive control dataset. The distribution of CDR3 lengths for the *in silico* generated alpha and beta chains are displayed in Figure B.1.

For negative control RNA-seq data, we used paired-end 101 nt RNA-seq data from seven TCR-negative cell lines, downloaded from ENCODE [221] and pooled for use as a negative control (Table B.1). To create negative datasets for shorter read lengths, reads from the 101 nt datasets were truncated to 76 nt and 50 nt reads. For positive control data sets error-free reads (101, 76, and 50 nts) were created for each in silico generated CDR3, with the center of the CDR3 region positioned at the center of the read.

An unbiased parameter space exploration was performed across all pairwise combinations of V gene minimum alignments and J gene minimum alignments (values 8 to 26 explored, all other parameters set as default) to determine optimal parameters. For each of the 361 parameter pairs, MiTCR was run on the negative and positive control datasets. For negative control datasets, the number of detected bogus CDR3s was tracked, and for positive control datasets, the number of correctly annotated CDR3s was tracked. Optimal parameters were assessed for each TCR chain - read length combination. Sensitivity was calculated for each parameter pair by dividing the number of recovered CDR3s by the maximum number of recovered CDR3s for all parameter pairs of that TCR chain – read length combination, giving a relative sensitivity value. We used a binary categorization to bin the false discovery rates as acceptable or not. For a set of false discovery rates, we selected the best parameter pair which had an acceptable false discovery rate and highest sensitivity. In the case of multiple parameter pairs being equally acceptable, the pair which minimized the V and J alignment parameters was selected. These optimal parameters are summarized in Table B.2.

3.1.2.3 Benchmarking CDR3 extraction efficiency using simulated data

The Flux Simulator [222] v1.2.1 is a computational tool which generates RNA-seq datasets by simulating a transcriptome expression profile, library construction, and sequencing errors. We simulated a range of sequencing depths (10^4 – 10^8 reads) and read lengths (50, 76, and 101 nt) to determine the importance of different factors on characterizing the ability to detect a given CDR3 sequence. Full-length *in silico* recombined TCR sequences were annotated as single exon genes in a reference synthetic chromosome sequence file, which we added to the human genome (GRCh38) to be used as the reference genome for Flux Simulator. Flux Simulator was run with the following command line flags: –t simulator, –x (to simulate expression), –l (to simulate library construction), –s (to simulate sequencing), and –p parameterFile.par. The parameter file contained the following parameters: REF_FILE_NAME: path to .gtf file; GEN_DIR: directory with genome reference files; FASTA: true; ERR_FILE: 76; READ_LENGTH: one of 50, 76, 101; PAIRED_END: true, UNIQUE_IDS: true; READ_NUMBER: one of 10000, 50000, 100000, 500000, 1000000, 5000000, 10000000, 50000000, 100000000; and TMP_DIR: path to temporary directory.

Ten RNA-seq datasets were simulated for each read length and sequencing depth combination to minimize the risk of any stochastic effects on transcript abundance in any one simulation confounding the variables which explain CDR3 recovery. Each simulated dataset was run through MiTCR using the optimized parameter sets for a theoretical 0 % false discovery rate, and results from all 270 simulations were pooled for analysis (Figure B.2). There are two requirements for detection of a CDR3: (1) the TCR transcript must be expressed, and (2) the sequence read length must be longer than the CDR3 length (Figure B.3). Before modelling, we filtered the data to cases that met these two criteria (n = 362,233). A multivariate logistic regression model was fit using half of the data (n = 181,116), leaving half the data for a validation set (n = 181,117). The fit of the logit function is shown in Equation 3.1. The performance of the model is summarized in Table 3.1, and resulted in 63.8 % sensitivity and 93.1 % specificity.

Equation 3.1:

$$\begin{aligned} logit(CDR3\ detected) \\ &= -5.38 \\ &+ 1.98(\log_{10}(transcripts\ per\ million)) + 0.51\left(\frac{sequencing\ depth}{10,000,000}\right) \\ &+ 0.04(read\ length) - 0.04(CDR3\ length) \end{aligned}$$

Table 3.1: Observed and predicted detection of CDR3s in the validation set by logistic regression with cut-off of 0.50.

	Predicted		
Observed	**Detected**	**Not Detected**	**% Correct**
Detected	31163	17698	63.8
Not Detected	9162	123094	93.1
Overall			85.2

3.1.2.4 Approximation of TCR transcript abundance from percent T cell infiltration

We queried Illumina BodyMap (http://www.ebi.ac.uk/gxa/experiments/E-MTAB-513) and GTEx (http://www.gtexportal.org/home/) to find the expression of *TRAC* and *TRBC1/2* in healthy whole blood. An average expression of approximately 150 FPKM was observed for these genes. As these genes are roughly 1 kb in length, this level of expression translates to a transcript fraction on the order of 1.5×10^{-4}. As lymphocytes are generally between 20 – 40 % of white blood cells, a pure lymphocyte population would have a TCR transcript fraction of approximately 5×10^{-4}. Assuming a similar cellular composition between peripheral blood lymphocytes and TIL, a tumour with 2 % TIL (the median value for TCGA tumours, parsed from TCGA biospecimen slide data) would have a TCR transcript fraction of 1×10^{-5}. This value can be inserted into the logistic regression model in order to predict the minimum sequence depth required to have a 50 % chance of detecting a CDR3 sequence with this abundance and other known properties.

3.1.2.5 TCGA RNA-seq data analysis

All available RNA-seq fastq files for solid tumours and matched normal tissues were downloaded with permission from Cancer Genomics Hub (https://cghub.ucsc.edu/). This includes 8,655 samples from 24 tumour sites. The majority (79 %) of this TCGA RNA-seq data are 50 nt reads, while 21 % are 76 nt, with sequencing depths ranging from 5.27×10^6 to 4.51×10^8 reads. To be able to directly compare the extracted CDR3s across all samples, we performed a pre-normalization step by truncating all reads to 50 nt, and randomly sub-sampling 1.00×10^8 reads from every sample. This resulted in removal of 41.5 % of all sequence data (5.20×10^{11} reads) and 15.2 % of samples (1,313) due to insufficient depth, leaving 7,342 (6,738 tumour and 604 normal) samples and 7.34×10^{11} total reads for analysis. Prior to running MiTCR, the fastq files were cleaned by only retaining reads longer than 40 nt and reads containing standard (ACTGN) bases. For every sample, MiTCR was run with the optimized V and J alignment parameters for a theoretical 0 % false discovery rate with 50 nt reads for both alpha and beta chains, keeping all other parameters default. Extracted CDR3 sequences that contained stop codons or frame shifts were removed prior to all further analysis.

3.1.2.6 TCGA gene expression datasets

In order to correlate extracted TCR diversity with expression of immune-related genes, all available RNASeqV2 data from the TCGA Data Portal was downloaded. This data provides gene expression information generated using MapSplice [223] for alignment and RSEM [224] to quantify gene expression. The reported *scaled_estimate* value was multiplied by 10^6 to obtain transcripts per million (TPM). To obtain consensus gene expression values, we summed the TPM values within each of the following groups: HLA class I (*HLA-A, HLA-B, HLA-C, HLA-E, HLA-F, HLA-G*), Class II (*HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRA, HLA-DRB1, HLA-DRB5*), CD8 (*CD8A*, *CD8B*) or CD3 (*CD3D*, *CD3E*, *CD3G*). Pearson correlations were calculated between these genes and the number of distinct CDR3 sequences in each subject (Figure B.5, Figure B.6).

3.1.2.7 Inferred pairing of TCR alpha and beta subunits

For each tumour sample, all possible pairwise combinations of TCR alpha and beta subunit CDR3 sequences derived from that sample were specified (n = 1,286,810). We then looked for recurrent alpha-beta pairs among all TCGA tumour samples, and identified 188 distinct alpha-beta pairs that were found in at least two subjects. To test if this was a stronger alpha-beta co-occurrence than would be expected by chance, we randomized the relationship between sample identifiers and their corresponding TCR alpha-beta subunit pairs. We then determined the frequency of recurrent alpha-beta pairs in this randomized dataset. Randomization was repeated for 100 iterations, and the proportion of trials which had a degree of sharing greater than or equal to the original, non-randomized data was taken as the P value.

3.1.2.8 Shared peptide-MHC and CDR3 sequences

We predicted HLA Class I alleles and MHC-presented point mutations for 1,361 TCGA subjects as previously described [156], with an IC_{50} value of 500 nM for MHC predicted binding affinity taken as the maximum threshold for potential immunogenicity. We counted the frequency of each pMHC (n = 305,438), and found 393 to be recurrent. In order to determine if any of the subjects sharing a pMHC also shared a similar CDR3 sequence, we took all CDR3 sequences from all 1,361 subjects, and clustered them to allow for inexact CDR3 sequence matches. We clustered at 95 % identity using CD-HIT [225,226] v4.6 with the following parameters: -c 0.95; -n 5; -l 5. We then checked each cluster to see if it contained sequences from at least two subjects, and if those subjects shared a pMHC. We found one cluster of two subjects that met this criteria. To test how frequently this would be expected to happen by chance, we randomly selected two subjects from the set, and checked if they shared a pMHC. We measured the fraction of successes from 1,000,000 trials, then multiplied by the number of clusters that contain two subjects (n = 1,457) to correct for multiple testing to give the adjusted P value.

3.1.2.9 Data analysis

All data analysis was performed in R [189] v3.1.2 or Python (http://www.python.org) v2.4.3 and v3.2.2.

3.1.2.10 Code availability

Custom code is available at https://github.com/scottdbrown/TCR-from-RNAseq2015.

3.1.3 Results

3.1.3.1 Somatically rearranged T cell receptor sequences can be effectively recovered from RNA-seq data.

We optimized the extraction of TCR alpha and beta chain sequences from RNA-seq datasets by evaluating negative and positive control datasets and adjusting search parameters. We identified optimal V and J alignment parameters that yielded an average of 94 % sensitivity for 100 % specificity using the shortest (50 nt) reads (Table B.2, see 3.1.2 Methods). Sensitivity is limited, ultimately, by the inability to detect the small proportion of CDR3s that are longer than a sequencing read (Figure B.3).

To estimate the yield of TCR transcripts that could be expected from typical RNA-seq experiments, we used Flux Simulator [222] to generate simulated RNA-seq data spiked with *in silico* recombined TCR transcripts, and processed these data as described in Methods, above. As expected, the most abundant TCR transcripts are the most readily detected and the sensitivity of the method increases with increasing sequencing depths and read lengths (Figure B.2). We fit a multivariate logistic regression model to explain the odds of detecting a CDR3 sequence from the RNA-seq data using the explanatory variables $\log_{10}$(transcripts per million) (OR = 7.242, 95 % CI = 7.079-7.411, $P < 2 \times 10^{-16}$), sequencing depth in tens of millions of reads (OR = 1.667, 95 % CI = 1.658-1.677, $P < 2 \times 10^{-16}$), sequence read length (OR = 1.0358, 95 % CI = 1.035-1.037, $P < 2 \times 10^{-16}$), and CDR3 nucleotide sequence length (OR = 0.958, 95 % CI = 0.956-0.960, $P < 2 \times 10^{-16}$), and observe that initial TCR transcript abundance is the most important factor in predicting whether a CDR3 would be detected.

For tumour tissue, where the degree of T cell infiltration is typically about 2 % (see 3.1.2.4 Approximation of TCR transcript abundance from percent T cell infiltration), we would expect 0.001 % of the total transcripts to be recombined TCR transcripts. Assuming a monoclonal infiltrate, an RNA-seq depth of 70 million 50 nt reads is required to have a greater than 50 % chance of detecting a 45 nt long CDR3 (the most frequent CDR3-beta length [11]). Additional probabilities are presented in Table B.3. To

further evaluate the yield of TCR sequences from RNA-seq data, we generated TCR-seq for TCR-beta (1,411,056 reads; 2,823 unique CDR3-beta sequences) and RNA-seq for total cDNA (56,067,687 reads; 9 unique CDR3-beta sequences) data from the same colorectal tumour tissue sample, using previously described methods [5,227] and observed that all high confidence CDR3-betas identified by RNA-seq (n = 9) fall within the top 2.1 % (n = 60) of CDR3-betas detected by TCR-seq, ranked by abundance (Figure B.4), confirming that at a modest depth of sequencing, RNA-seq can identify the most abundant CDR3 encoding transcripts.

3.1.3.2 TCR sequence diversity in tumour-associated T cell repertoires.

We extracted TCR alpha and beta chain CDR3 sequences from all available RNA-seq datasets from the TCGA project. This included 7,342 total data sets derived from 6,738 solid tumour and 604 matched normal tissues, from 24 different tumour sites. In tumours, the yield per subject ranged from 0 – 702 (median of 9) reads containing a full CDR3 sequence (Figure 3.2), and this translated to a range of 0 – 538 (median of 7) distinct CDR3 amino acid sequences per subject. Kidney renal clear cell carcinoma (KIRC) produced the greatest yield of CDR3s, whereas brain lower grade glioma (LGG) produced the least. As expected, there is a strong correlation between number of distinct CDR3 amino acid sequences and *CD3* expression (Figure B.5). Comparing the gene expression of HLA class I and class II genes with the number of distinct CDR3 amino acid sequences per subject, we observed a positive correlation, with markedly stronger correlations seen for class II genes (Figure B.6; $P = 9.3 \times 10^{-10}$, paired t test). This is consistent with recent reports highlighting the immunoreactivity of T cells with specificity for MHC class II presented tumour antigens [114,228,229].

Next, we evaluated the differential abundance of CDR3s between tumours and matched normal control tissues for all subjects where the RNA-seq data was available for both (n = 462, Figure 3.3, Table B.4). Of 6,611 total alpha chains in this set, 3,560 (53.8 %) were unique to tumour samples and 2,826 (42.7 %) were unique to matched control samples. Likewise, of the 7,664 beta chains, 4,279 (55.8 %) were unique to tumour and 3,277 (42.8 %) were unique to matched control samples. A total of 225 unique CDR3-alpha and 108 unique CDR3-beta sequences were present in both

tumour and control tissues. Notably, almost all CDR3s that are unique to tumour or normal tissues had only a single supporting read (Figure 3.3), and do not show evidence of clonal expansion. Thus, while there is evidence for a larger and more diverse T cell infiltrate in tumour compared to control tissues ($P < 2.2 \times 10^{-16}$, chi-squared test), these results suggest that a large proportion of tumour associated T cells are bystanders, not readily distinguishable from the normal population of tissue resident T cells. A single KIRC subject was a notable outlier in this analysis. The tumour sample from this subject yielded the three most abundant tumour specific CDR3-alphas and the two most abundant tumour specific CDR3-betas in the entire cohort, suggesting the possibility of an acute anti-tumour T cell response in this individual.

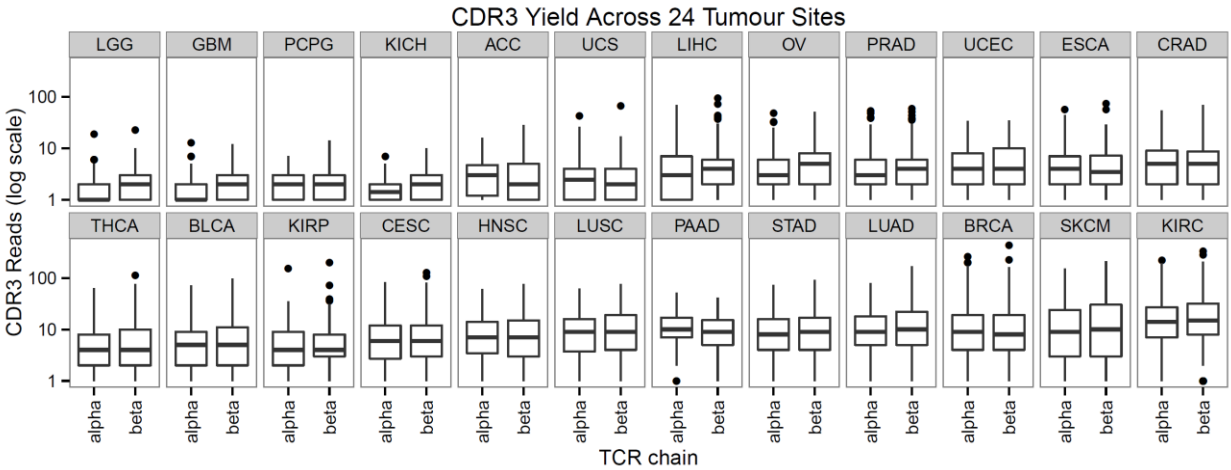

Figure 3.2: The number of reads containing CDR3 sequences varies across tumour sites. LGG: brain lower grade glioma; GBM: glioblastoma multiforme; PCPG: pheochromocytoma and paraganglioma; KICH: kidney chromophobe; ACC: adrenocortical carcinoma; UCS: uterine carcinosarcoma; LIHC: liver hepatocellular carcinoma; OV: ovarian serous cystadenocarcinoma; PRAD: prostate adenocarcinoma; UCEC: uterine corpus endometrial carcinoma; ESCA: esophageal carcinoma; CRAD: colon and rectum adenocarcinoma (combination of COAD and READ); THCA: thyroid carcinoma; BLCA: bladder urothelial carcinoma; KIRP: kidney renal papillary cell carcinoma; CESC: cervical squamous cell carcinoma and endocervical adenocarcinoma; HNSC: head and neck squamous cell carcinoma; LUSC: lung squamous cell carcinoma; PAAD: pancreatic adenocarcinoma; STAD: stomach adenocarcinoma; LUAD: lung adenocarcinoma; BRCA: breast invasive carcinoma; SKCM: skin cutaneous melanoma; KIRC: kidney renal clear cell carcinoma.

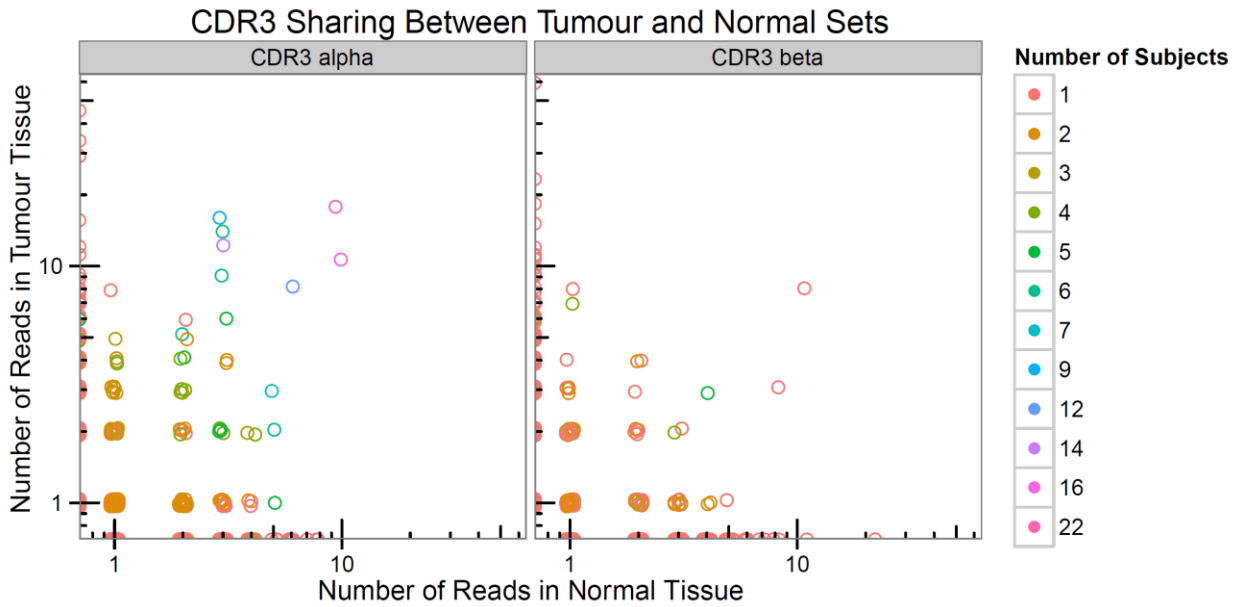

Figure 3.3: The majority of CDR3s recovered from tumour/normal control tissue pairs are unique to tumour or normal. For every CDR3, the number of reads in the set of tumours is plotted on the y-axis, with the number of reads in the set of normal samples on the x-axis. Points are coloured by the number of subjects in which that CDR3 sequence is detected.

3.1.3.3 Public T cells are common in the tumour environment

To explore the recurrence of TCRs mined from the tumour environment, we compared the CDR3-beta sequences extracted from the complete set of analyzed TCGA samples to the approximately 1.1 million distinct CDR3-beta sequences we previously identified by deep TCR-seq analysis of peripheral blood from a healthy subject [5]. Of all 49,672 distinct TCGA CDR3-beta sequences we observed, 22.8 % were found in the peripheral repertoire of the healthy subject (Figure B.7). We found the level of overlap with this healthy individual for those CDR3-betas that are seen in multiple TCGA subjects (public TCRs; 76.5 % of 2,197 shared TCGA CDR3-betas found in the healthy repertoire) was substantially greater than for those that are unique to a single TCGA subject (private TCRs; 20.3 % of 47,475 unique TCGA CDR3-betas found in the healthy repertoire), suggesting these shared, tumour associated CDR3-betas are derived, predominantly, from public T cells. Indeed, when we queried tumour-associated CDR3-beta sequences for matches to CDR3-beta sequences in the literature with defined antigen specificity, we found numerous matches to known viral-specific TCRs [230] (Figure 3.4). TCGA

CDR3-betas with viral specificity were much more common within the set of shared TCGA CDR3-betas than within the set of unique TCGA CDR3-betas. Specifically, nine of 2,197 CDR3-betas (0.41 %) that were shared among TCGA subjects were identifiable as being viral-specific, whereas only three of the 47,475 CDR3-betas (0.0063 %) that were unique to a single TCGA subject were similarly identifiable.

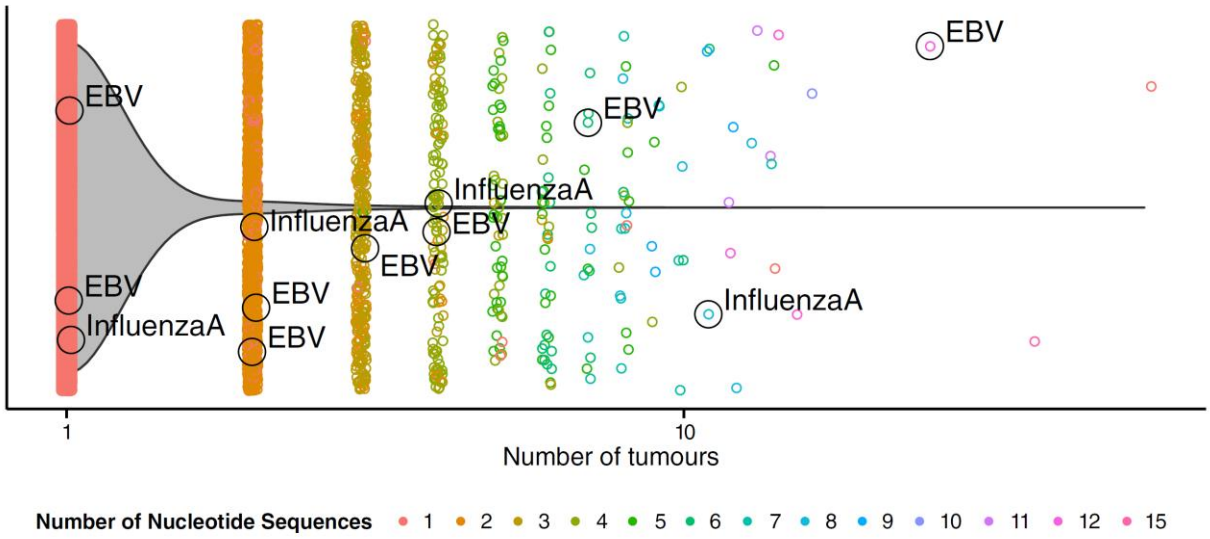

Figure 3.4: Sharing of CDR3-beta sequences. All 49,672 CDR3-beta sequences derived from tumours are plotted along the x-axis according to the number of tumours they are found in. Colour defines the number of nucleotide sequences that were found to generate the same CDR3-beta amino acid sequence. Violin plot overlay shows that most recovered CDR3-beta sequences are unique to an individual, though there is notable sharing (4.4 %) between subjects. Known, public, viral-specific CDR3-beta sequences [230] are labelled with their antigen specificity.

Given the observation of substantial sharing of CDR3 sequences among subjects, we asked if some of the shared alpha and beta CDR3 sequences may represent shared dimeric TCRs. To test this, we generated all possible alpha-beta pairs within each subject's alpha and beta repertoires and looked for sharing of any pairs between two or more subjects. We observed 188 distinct alpha-beta pairs that were found in at least 2 subjects, which was not significantly more than would be expected by chance (P = 0.42, random resampling, see 3.1.2.7 Inferred pairing of TCR alpha and beta subunits). We also asked if subjects with shared mutations and shared HLA alleles may also share TCR sequences. Previously, for a subset of TCGA subjects, we identified tumour point mutations predicted to yield class I pMHCs [156]. We clustered

the CDR3 amino acid sequences from all subjects with predicted mutant pMHCs at 95 % amino acid sequence identity. Of the 17,092 total CDR3 sequence clusters (including singletons), only one cluster contained sequences from two subjects with matching mutant pMHCs (Table 3.2). Although this is not statistically significant (P = 0.35, random resampling, see 3.1.2.8 Shared peptide-MHC and CDR3 sequences), with deeper sequencing data from large numbers of subjects, this approach may prove useful for matching TCR sequences to the neoantigens they recognize.

Table 3.2: Summary of a CDR3-beta sequence cluster that shares pMHC.

Cluster Number	Subjects	Sequences	Mutant Gene	Peptide	HLA
6473	TCGA-HU-A4G8	CASSRDSSYEQYF	PGM5	GRLIIGQNGV	B*27:05P
6473	TCGA-BR-8081	CASSLRDSSYEQYF	PGM5	GRLIIGQNGV	B*27:05P
6473	TCGA-HU-A4G8	CASSRDSSYEQYF	PGM5	GRLIIGQNGVL	B*27:05P
6473	TCGA-BR-8081	CASSLRDSSYEQYF	PGM5	GRLIIGQNGVL	B*27:05P

3.1.4 Discussion

In future, as sequence costs continue to decline, there will be increasing opportunities to derive immune signatures from unbiased data types. Here, we have optimized an analytical strategy for extracting T cell repertoire information from RNA-seq datasets, and used it to characterize tumour associated T cell repertoires. We have provided optimal parameters for mining RNA-seq datasets of varying read lengths, and provided a range of parameters for varying levels of acceptable false positive rates. This procedure was validated on simulated RNA-seq datasets with known recombined T cell receptor transcripts and was also compared to classical TCR-seq data, showing that the subset of TCRs we detect using RNA-seq are the most abundant T cells in the sample. The most abundant T cells may or may not be the most biologically relevant T cells. In solid tumours, the relationship between clonal abundance of T cells and anti-tumour immunity is not yet clear as it is obscured by the presence of bystander T cells in the tumour environment. Likewise, in cases such as acute infection, the most abundant T cells are likely those that are most biologically relevant. TCR transcripts from rare T

cells will become more accessible in future because continually declining sequencing costs will allow deeper and deeper transcriptome sampling by RNA-seq.

The expected yield of TCR reads from RNA-seq data is ultimately dependent on the level of T cell infiltration in the sample and the clonality of the infiltrate. Assuming a similar cellular composition to TCGA tumours, one can expect on the order of 1 TCR reads from 10 million sequence reads. Our analysis has highlighted a strong and novel correlation between tumour TCR diversity and tumour MHC class II expression and high prevalence of public T cells in the tumour environment. Further, within the limitations of the available data, we have explored the association between alpha-beta TCR pairs, and linked TCR sequences to specific pMHC complexes. Analyses of this nature may inform future cancer immunotherapy strategies, and we expect that this same approach will have value in exploring other immune related pathologies, where large RNA-seq datasets already exist or can be obtained. In cases such as T cell lymphoma, where the T cell is the cancerous cell and is highly expanded, the most abundant CDR3 sequences found by RNA-seq should generally be adequate to identify the tumour clone and monitor disease progression [219]. We next test this by applying our analysis to RNA-seq datasets from peripheral T cell lymphomas (PTCLs).

3.2 Defining the clonality of peripheral T cell lymphomas using RNA-seq

3.2.1 Introduction

Peripheral T cell lymphomas (PTCLs) represent 10-15 % of non-Hodgkin lymphomas [231]. PTCL not otherwise specified (PTCL-NOS) and angioimmunoblastic T cell lymphoma (AITL) are the most common PTCL subtypes [231,232]. Both are aggressive lymphomas with cure rates of 20-30 % by chemotherapy [233].

Lymphoid cancers are believed to arise from a single lymphocyte that acquires somatic mutations sufficient for malignant transformation. Progeny tumour cells obtain an increasingly diverse mutational landscape through tumour evolution, but all progeny share identical (clonal) rearrangements in TCR or, in the case of B cell lymphomas, immunoglobulin (Ig) genes. Clonal TCR and Ig rearrangements are useful for distinguishing malignant cells from the polyclonal background of normal lymphocytes. The use of flow cytometry to identify aberrance in cell surface markers is an important

diagnostic tool in PTCLs [234]; however, it is not directly informative regarding T cell clonality. Multiplex PCR methodologies are the current standard for inferring clonal TCR rearrangements in clinical practice. The PCR products of TCR genes can be analysed for putative clonality using either heteroduplex analysis [235] or GeneScanning [236], but only TCR sequence analysis can identify T cell clones unequivocally.

Deep sequencing of TCR amplicons (TCR-seq) is a powerful and sensitive method for characterizing the T cell repertoire [10,11,67,219,237]. However, transcriptome sequencing (RNA-seq) is more informative, providing data from all transcribed genes present in the sample, and has proven utility in personalized oncology [238,239]. Obtaining TCR sequences directly from RNA-seq data can, in some settings, provide sufficient information on T cell clonal abundance to obviate the need for dedicated TCR-seq assays [240] which are associated with considerable added time and cost.

Here, we used flow cytometry/fluorescence-activated cell sorting (FACS) to identify and purify malignant T cell populations from non-malignant cells. We aimed to determine the utility of RNA-seq in establishing TCR clonality in samples with either an aberrant or normal (non-aberrant) T cell immunophenotype without the need for amplicon-based TCR-seq.

3.2.2 Methods

3.2.2.1 Clinical specimens and cell sorting

This study was approved by the University of British Columbia/British Columbia Cancer Agency (BCCA) Research Ethics Board (H14-01235). Sixty diagnostic lymph node cell suspensions were obtained from the BCCA Lymphoid Cancer Tumour Bank (32 PTCL-NOS, 28 AITL), collected from 1990 to 2014. Excess aliquots from each diagnostic specimen were placed in DMSO and stored at -80°C. One additional lymph node cell suspension from a pre-diagnostic time point was included for one subject (PTCL-NOS), as were five tonsil biopsies from healthy subjects to be used as controls.

All specimens were stained with an 11-antibody panel to identify aberrant T and T follicular helper (T_{FH}) cells. The panel (Table B.5) consisted of CD45 (common leukocyte marker), lineage-specific T cell antibodies (CD3, CD4, CD8), pan-T cell

antibodies (CD2, CD5, CD7), T_{FH} cell antibodies (CXCR5, PD1), CD10 (for the detection of aberrant T cells in AITL), and CD19 (for the detection of B cells). Data was acquired on a Becton Dickinson FACSAria3 instrument as part of a sorting experiment to isolate tumour cell subpopulations. Data was analysed by conventional gating and bivariate plot display using FlowJo software (version 10.0.8).

FACS was used to identify and purify specimens with an aberrant immunophenotype (Figure 3.5A). T cell surface marker aberrance was defined as the loss of one or more lineage-specific (CD3, CD4 or CD8) or pan-T cell marker (CD2, CD5 or CD7) or the gain of CD10. Skewing of the CD4:CD8 ratio (*e.g.,* > 10 or < 0.5) was not a criterion for aberrancy; however, we acknowledge that marked skewing is suggestive of clonal dominance. A population was defined as ≥ 1 % of viable lymphocytes. For specimens without an aberrant immunophenotype, and for tonsil biopsies from healthy controls, T_H, T_{FH}, and CTL populations were sorted (Figure 3.5C).

3.2.2.2 Sequencing

RNA was extracted from FACS-sorted cells using Qiagen Allprep DNA/RNA column-based extraction kits as per the manufacturer's instructions. RNA-seq was performed on DNase-treated samples using an RNA-seq lite plate-based protocol with SMART cDNA amplification. Non-aberrant cell subsets with the highest RNA quality for each subject were selected for RNA-seq library construction and sequencing (> 10 ng RNA and RNA quality score ≥ 6.4), as these were most likely to produce informative sequencing data to identify a dominant clone among immunophenotypically normal T cells. All samples were initially subjected to shallow sequencing (average 4.1 million total reads per sample, range 2.4 – 6.0 million) to facilitate the identification of the malignant T-cell clone in the sorted populations. To evaluate the importance of sequence depth for this application, all samples were re-sequenced deeply (average 92 million total reads per sample, range 63 – 156 million). All sequencing was performed using 125 nucleotide paired-end reads on an Illumina HiSeq 2500 instrument at the Genome Sciences Centre in Vancouver, Canada.

3.2.2.3 Analysis of clonality

MiTCR software [150] with modified settings (TRAV minAlignmentMatches: 12, TRAJ minAlignmentMatches: 19, TRBV minAlignmentMatched: 14, TRBJ minAlignmentMatches: 16) [240] was used to identify TCR alpha and beta CDR3-containing reads present in the sequencing data, generating a list of CDR3s (clonotypes) and their relative abundances. Non-productive CDR3 sequences (containing a frame-shift or stop codon) were removed from the analysis as they are most likely the result of incomplete allelic exclusion [241]. CDR3 sequences initially classified, inappropriately, as both alpha and beta were subsequently resolved by assigning the chain that had greater read support. Low-abundance CDR3s that had equal numbers of supporting reads for both chains were marked as ambiguous. TCR gamma or delta chain sequences were not interrogated because extensive optimization and validation of TCR extraction from transcriptome data is required [240] and this has only been done for alpha-beta TCR analysis. A dominance metric *D* for each recovered clonotype *c* in sample *s* was calculated as shown in Equation 3.2, where R_{sc} denotes the number of reads in sample *s* supporting clonotype *c*, R_{sHc} denotes the number of reads in sample *s* supporting clonotypes of chain *H* that matches the chain of clonotype *c*, and R_s denotes the total number of reads in sample *s*. This metric is the product of two proportions: the proportion of chain-specific TCR reads and the proportion of total sequence reads. Together, these provide a measure of the clonotype abundance relative to all clonotypes identified, as well as relative to the size of the sequence dataset.

Equation 3.2:

$$D_{cs} = \frac{{R_{sc}}^2}{R_{sHc} \times R_s}$$

To determine which CDR3 sequences were dominant (above background), chain-specific thresholds were set as the maximum D_{cs} derived from control samples for each chain; clonotypes above the threshold are dominant (5.38×10^{-7} for shallow alpha, 2.28×10^{-7} for shallow beta, 2.27×10^{-8} for deep alpha, 9.85×10^{-8} for deep beta).

These thresholds are experiment-specific, and controls containing the expected normal polyclonal background of T cells for each future experiment would be required.

TCR gamma gene rearrangements are observed in most alpha-beta T cells, and thus is used as the target for heteroduplex TCR analysis [235]. This assay was performed on 54 of the 60 PTCL samples at the time of diagnosis (from 1990 to 2014). GeneScanning was not available at our cancer center during the period of sample collection. The results from the heteroduplex analysis were compared to the MiTCR analysis to determine if there was improved sensitivity from the sequence-based approach.

3.2.2.4 Estimating tumour content

As all samples are comprised of sorted T cell populations, the tumour content of a sample can be estimated using the relative abundance of the dominant T cell. This approach is valid assuming all cells sequenced are alpha-beta T cells, all cells are expressing the TCR at a roughly equal level, and all TCR transcripts have an equal probability of being captured and sequenced. Due to allelic exclusion, each T cell clone should only express one beta chain [13], therefore the most abundant beta clonotype (if present, otherwise alpha) was used to define each clonal T cell and estimate tumour content. As an additional metric for assessing tumour content, the Shannon Entropy [242] was calculated for each sample using the entropy package [243] for R (v3.1.1). The Shannon Entropy quantifies the information content of a set of entities with associated abundances by measuring the uncertainty associated with predicting the identity of a randomly chosen entity. A high value corresponds to high uncertainty, and thus high diversity, whereas a low value corresponds to a set with low diversity.

3.2.2.5 Gene expression

RNA-seq files were aligned to the hg38 reference transcriptome using Bowtie2 [244] (v2.0.2) and gene expression was quantified using RSEM [224] (v1.2.29). TPM (transcripts per million) values for TCR constant genes were centered and scaled, and used to determine if samples showed evidence of alpha-beta or gamma-delta TCR expression.

3.2.2.6 Code availability

Custom code is available at https://www.github.com/scottdbrown/RNAseq-TcellClonality.

3.2.3 Results

3.2.3.1 Identification of dominant TCRs

For 60 cases, diagnostic lymph node cell suspensions from T cell lymphomas were sorted by FACS into either aberrant (loss of one or more lineage-specific (CD3, CD4 or CD8) or pan-T cell marker (CD2, CD5 or CD7) or the gain of CD10) or non-aberrant populations (no evidence of aberrance; they are immunophenotypically normal). In 45 of the 60 total cases, at least one aberrant population (Figure 3.5A) of varying abundance (median 27.1 % of lymphocytes; Figure 3.5B) was isolated by FACS. The remaining 15 cases and 5 controls had no observable aberrant population, and were sorted into at least one of CTL, T_H, and T_{FH} populations (Figure 3.5C). Two cases had two distinct aberrant populations each, which were sorted. In total, 82 sorted cell populations (samples) were isolated from the 65 specimens (60 cases and 5 controls). These were subjected to RNA-seq and bioinformatic extraction of TCR alpha- and beta-chain sequences.

The threshold for determining the dominance of a TCR clonotype was computed by setting the background as the highest dominance metric observed from the 15 control samples (Figure B.8). Using this threshold, evidence of a dominant TCR clonotype was obtained in 96 % of samples (45/47) that were aberrant by flow cytometry (Figure B.9), but also in 80 % of samples (16/20) that had appeared non-aberrant by flow cytometry (Figure B.10). The samples that were aberrant by flow cytometry and also appeared aberrant by TCR sequence analysis typically showed a highly abundant dominant clonotype with a minimal background of low-abundance clonotypes (Figure 3.6). Samples that were non-aberrant by flow cytometry, but aberrant by TCR sequence analysis, generally had a dominant clone, but also a larger background repertoire, similar to the diverse repertoire of normal T cells seen in healthy controls (Figure 3.6). Thus, although the specimens that were non-aberrant by flow cytometry appeared to be immunophenotypically normal, they clearly contained an

expanded malignant clone. This suggests that malignant lymphocytes can retain a normal immunophenotype despite underlying clonality, and demonstrates the increased sensitivity and diagnostic ability of RNA-seq over FACS. In general, deeper sequencing results did not provide additional utility, as results were consistent with those obtained by shallow sequencing (Figure B.11).

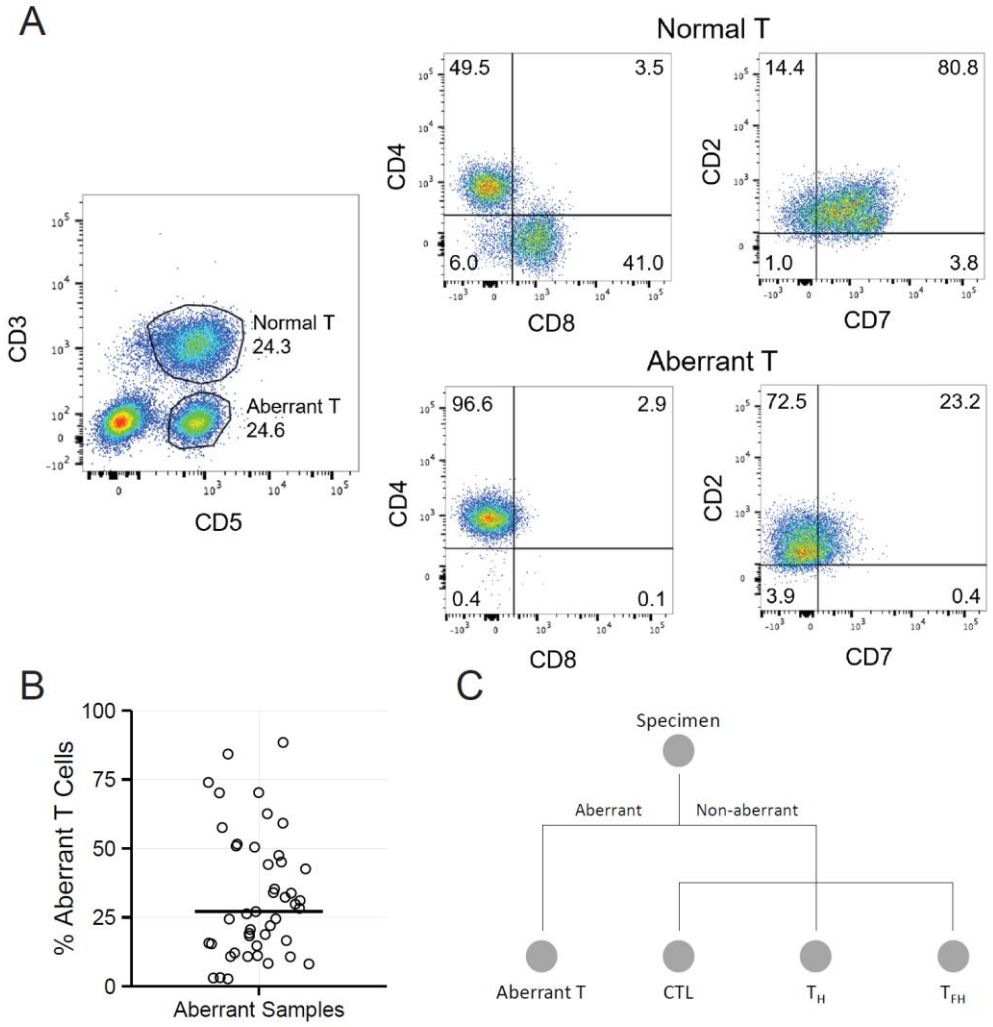

Figure 3.5: Specimen processing and cell sorting. (A) After gating on CD45$^+$ cells, normal (CD3$^+$, 24.3 %) and aberrant (CD3$^-$, 24.6 %) T cell populations are identified. The normal population is composed of a mixture of CD4$^+$ and CD8$^+$ cells, and has no aberrant loss of CD7. The aberrant population is composed exclusively of CD4$^+$ cells, and demonstrates aberrant loss of CD3 and CD7. **(B)** Aberrant T cell populations were present at a range of frequencies (2.7 - 88.5 %) in lymph node cell suspensions (horizontal bar marks median). **(C)** Overview of the FACS sorting strategy. All specimens were stained identically as described in Methods. If an aberrant population was identified, it was sorted to purity. Non-aberrant and control cases were sorted into CTL, T_H, and T_{FH} subpopulations.

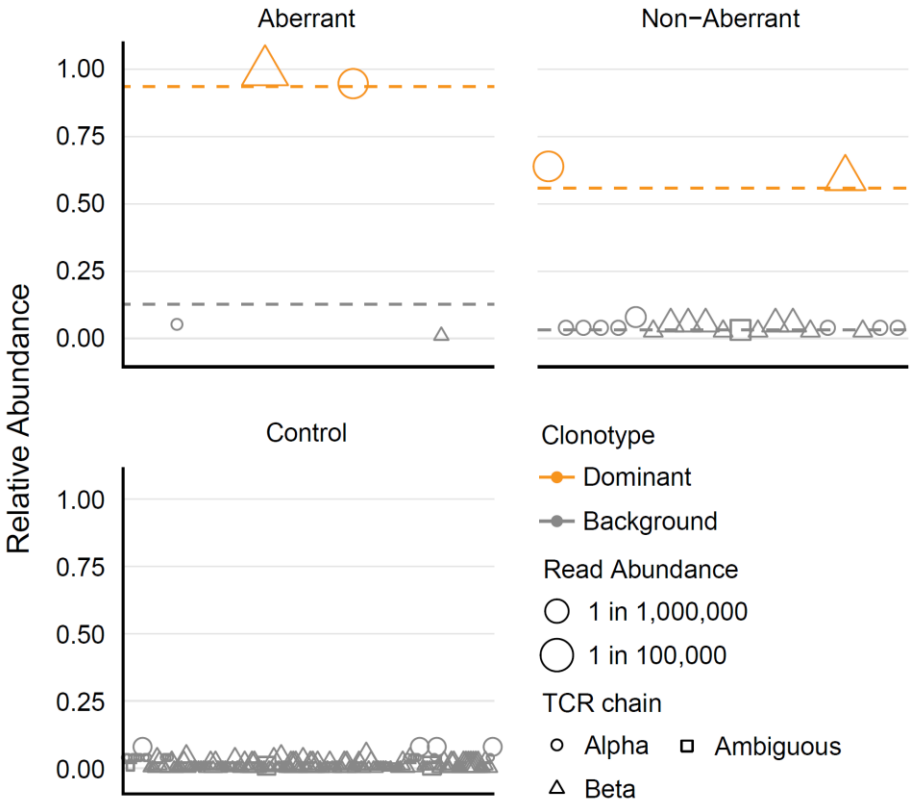

Figure 3.6: Identification of dominant clonotypes. Clonotypes from one representative sample for each category are displayed. The relative abundance of each clonotype is shown on the y-axis, calculated as the abundance of each clonotype relative to the total abundance of all clonotypes of the same chain in that sample. Clonotypes are plotted along the x-axis in lexicographical order. Clonotypes determined to be dominant are coloured orange. The read abundance of each clonotype is represented by its size. Mean relative abundance of dominant or background clonotypes for all samples of each category are shown as dashed horizontal lines.

3.2.3.2 Comparison to existing clinical assay

Of the 45 diagnostic specimens which were immunophenotypically aberrant, 39 had clinical heteroduplex testing performed [235]. This test is based on PCR of the TCR locus using a multiplexed set of V and C gene primers, followed by denaturation and cooling of the amplicons to induce duplex formation. Samples which contain monoclonal cells will re-anneal with their complementary DNA strand, resulting in homoduplexes. Duplex formation from polyclonal samples will mainly result in heteroduplexes (sequences from two different TCRs annealing), These heteroduplexes are larger due to imperfect base pairing. Duplexes are loaded onto a non-denaturing polyacrylamide

gel and run to visualize differences in sizes of the duplexes. Samples with smaller bands are positive by this test, demonstrating evidence of a monoclonal cell population. Of the 39 specimens that had this testing done, 31 tested positive, and all 31 also showed evidence of a dominant clone by RNA-seq analysis. Of the eight that tested negative, seven had a dominant clone by RNA-seq analysis. Of the 20 diagnostic specimens that were immunophenotypically normal, 15 had clinical heteroduplex testing performed. Of these, 13 tested positive, and 11 of these showed a dominant clone by RNA-seq analysis. The two heteroduplex-positive and RNA-seq-negative specimens likely reflect missing sequence data; these specimens had non-sequenced sorted cell populations due to poor RNA quality, and one of these non-sequenced cell populations may contain the malignant clone. Of the two that tested negative by clinical heteroduplex testing, one was positive by RNA-seq analysis.

3.2.3.3 Using diversity to identify malignant clones that do not express alpha-beta TCR

Shannon entropy [242] was calculated as a measure of TCR diversity for each sample. The relative abundance of the dominant T cell clone was used as a surrogate for tumour content, and showed an expected negative correlation with Shannon entropy (Pearson r = -0.90; Figure 3.7). There were two aberrant outliers with low entropy and no dominant clonotype identified, suggesting these aberrant cells do not express TCR alpha or beta. These samples had low expression of TCR alpha and beta constant genes (*TRAC*, *TRBC1*/*2*) and high expression of TCR gamma and delta constant genes (*TRGC1*/*2*, *TRDC*; $P \leq 0.032$, Mann-Whitney *U* tests), suggesting these may be gamma-delta expressing PTCLs.

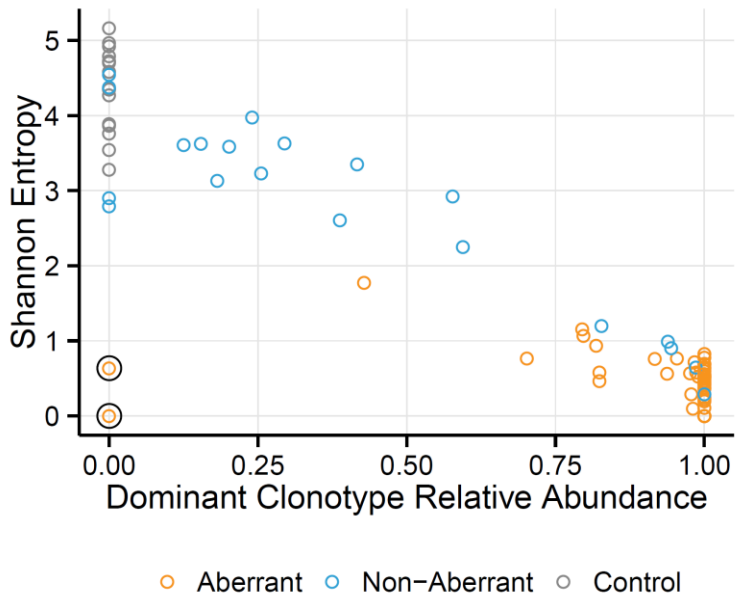

Figure 3.7: Characterization of sample diversity. Relationship between dominant clone relative abundance and Shannon entropy. Increasing dominant clone abundance shows decreasing Shannon entropies. The two aberrant samples with low Shannon entropy and absence of a dominant clone (circled) may represent cases where the malignant clone did not express an alpha-beta TCR.

3.2.3.4 Recurrent TCR sequences

There was no dominant TCR shared across subjects. There were four examples of dominant TCR sharing between different samples from the same subject: two of these occurred in subjects with two immunophenotypically distinct aberrant populations, possibly due to cell surface marker diversification post-malignant transformation; and two occurred in separate non-aberrant subsets, likely reflecting impurity in the sorting from FACS as the dominant clone showed unequal abundance between samples.

3.2.4 Discussion

These data demonstrate the utility of mining TCR sequences from RNA-seq data obtained from diagnostic lymph nodes to define malignant T cells. This is feasible using only light-coverage RNA-seq data. Deeper sequencing, while producing more robust data, was largely unnecessary as it did not improve the ability to detect dominant clonotypes in almost all cases, highlighting the usefulness and cost-effectiveness of shallow RNA-seq (4 million reads per sample) to detect clonality in PTCLs. For the current analysis, data was generated using sorted cell populations, but we expect that

malignant clones would be similarly recognizable from their dominant TCRs in unsorted samples or from whole blood. Thus, in the future, analysis of TCR profiles extracted from RNA-seq data from unsorted PTCL populations should be feasible and could serve as a useful assay in the diagnosis of T cell lymphoproliferative disorders. Further, this method yields a unique sequence identifier in clinical samples that could find utility as a personalized marker to monitor response to treatment, assess minimal residual disease, identify the onset of recurrence, and track tumour evolution.

Chapter 4: Neoantigen characteristics in the context of the complete predicted MHC class I self-immunopeptidome

4.1 Introduction

Tumour neoantigens are mutated self-peptides presented by tumour cell MHC molecules, and are capable of eliciting anti-cancer T cell responses [112,126–130,180]. The self-immunopeptidome is the collection of all self-peptides presented by the MHC molecules present in an individual. In principle, individuals with large self-immunopeptidomes should be more able to present a diversity of neoantigens (due to a general increased ability to present peptides) and these individuals may, therefore, be better able to mount natural immune responses to control malignant cell growth. Indeed, there is evidence for improved response to cancer immunotherapies for individuals having higher diversity of their class I HLA loci [140]. It is possible that individuals with smaller self-immunopeptidomes would be more vulnerable to immune threats such as cancer and/or infectious disease. In the context of immune surveillance of cancer, it has been observed that in individuals with cancer, mutations that are poorly presented across a range of MHC occur at higher frequencies than mutations that are readily presented by many MHC [245], suggesting that tumour cells can exploit gaps in the self-immunopeptidome, and that individuals with smaller self-immunopeptidomes will have greater cancer risk.

Here, we measured the range of self-immunopeptidome sizes present in human cohorts by predicting, computationally, the fraction of the human proteome able to be presented by each class I MHC molecule. By performing this exhaustive computation up-front, we were then able to query the results for any given individual with a known HLA class I genotype to predict the overall size of their self-immunopeptidome. Our analysis of TCGA data revealed a small but significant decrease in size of self-immunopeptidomes for cancer subjects compared to non-cancer subjects. We also explored the phenomenon of immune-editing [246] by predicting the immunogenicity of mutations. Here, immunogenicity of a mutation is defined by the number of mutant peptide-MHC pairs (neoantigens) containing the mutation that are predicted to bind MHC-I with IC_{50} < 500 nM. We compared the immunogenicity of mutations found in

cancer subjects to sets of matched *in silico* generated mutations. By this approach we identified the amino acid positions in peptide epitopes that have the strongest influence on immunogenicity.

4.2 Methods

4.2.1 Reference proteome

The human reference proteome was downloaded from EMBL-EBI (http://ftp.ebi.ac.uk/pub/databases/reference_proteomes/QfO/Eukaryota/UP000005640_9606.fasta.gz and http://ftp.ebi.ac.uk/pub/databases/reference_proteomes/QfO/Eukaryota/UP000005640_9606_additional.fasta.gz), using the April 2016 Qf0 release. This contained references for 21,006 protein sequences from the canonical set of proteins, plus 71,173 additional isoform sequences. These were combined for the complete analysis to ensure all unique peptides that exist in the human reference proteome were captured.

4.2.2 Condensing the proteome

Within the reference proteome, a specific 8-11mer peptide sequence may occur multiple times (non-unique peptides). To reduce the amount of computation required, we will only compute peptide-MHC binding for the set of unique 8-11mer peptides. All unique 8-11mer peptides were extracted from the complete proteome and written to a file. However, we determined that the compute time per peptide using NetMHCpan 3.0 [161] is significantly sped up by providing NetMHCpan with longer protein sequences and having it automatically extract all *n*-mers by sliding window rather than providing each *n*-mer individually (two orders of magnitude, data not shown). To take advantage of this, we desired a set of amino acid sequences which, when parsed with a sliding window, only contain peptides from these unique sets, and only contain each peptide exactly once. To achieve this, we re-assembled all unique *n*-mers into sets of artificial protein sequences using the following greedy algorithm (in pseudocode):

```
for n in [8,9,10,11]:
  start a set of artificial proteins Sn for peptides of length n;
  for each unique peptide of length n:
    if the first or last n-1 amino acids of the peptide matches the
    last or first n-1 amino acids of any artificial protein in Sn:
      extend that artificial protein with the additional N- or C-
      terminal amino acid;
    else:
      start a new artificial protein in Sn with the peptide
      sequence;
  end for;
end for;
```

This resulted in four sets of artificial protein sequences (one for each peptide length) containing each unique 8-11mer peptide exactly once when parsing with a sliding window. This shrunk the amino acid space required to be explored from the 36,688,307 amino acids of the reference × 4 peptide lengths = 146,753,228 total amino acids to 12,600,566 (for 8mers) + 12,635,023 (for 9mers) + 12,734,064 (for 10mers) + 12,835,955 (for 11mers) = 50,805,608 total amino acids (34 % of the reference).

4.2.3 Selecting the most suitable binding prediction threshold

Within the literature, multiple thresholds are used to classify pMHC binders using NetMHCpan algorithms [70,160,161,247]. The two most common thresholds to classify binders are a binding affinity threshold (IC_{50} < 500 nM), and a rank-based threshold (Percentile Rank < 2 %). While the most correct threshold is as-of-yet undiscovered, and likely will depend on the source of the data being used (self vs. mutated vs. infectious agent peptides), these two thresholds have been demonstrated to provide useful and informative results [155,156,245]. To determine which of these two thresholds would perform best for our purposes of estimating the size of the set of self-peptides presented by each class I MHC, we used publicly available mass-spectrometry (MS) data from SysteMHC Atlas [248]. All datasets containing Human peptide data in the context of MHC class I molecules were downloaded from systemhcatlas.org on

February 7, 2018 (n = 194). These data comprised 66 MHC molecules (15 *HLA-A*, 34 *HLA-B*, 17 *HLA-C*) and 135,092 total pMHC interactions.

We performed binding predictions using NetMHCpan 3.0 [161] for all unique 8-11mer peptides obtained from a 10 % random subsample of the human proteome (proteins were randomly selected from the human proteome until 10 % of the total human proteome length was achieved, results from this depth correlate strongly with those from the full dataset, data not shown) to all class I MHC molecules available for prediction. For each MHC, we tallied the number of pMHC by either the IC_{50} < 500 nM, or the Rank < 2 % threshold. For 66 MHC alleles with Human peptide data available from SysteMHC, we compared the number of unique peptides predicted to bind using the two thresholds to the number of peptides observed to bind these MHC using MS data.

By Spearman's rank correlation, thresholding by IC_{50} yields a better correlation to the observed peptide data than thresholding by rank (IC_{50} $\rho = 0.558$, $p = 1.1 \times 10^{-6}$, Figure C.1, and Rank $\rho = 0.314$, $p = 0.010$). To control for any effect that sample size might have in the SysteMHC data, we generated a linear model using either IC_{50}- or Rank-based threshold counts to explain the observed peptide counts, with the number of samples in SysteMHC data for each MHC as a covariate. Using this model, we found the counts using the IC_{50}-based threshold to be a better predictor of the number of *in vivo* presented peptides (IC_{50} adjusted $R^2 = 0.686$; Rank adjusted $R^2 = 0.591$). Therefore, we used the IC_{50} < 500 nM threshold to predict pMHC binding.

4.2.4 Running peptide-MHC binding predictions

Calls to NetMHCpan were batched into sets of approximately 1,000 artificial proteins and a single HLA and split into 1,676,125 separate jobs to run on a compute cluster. As the size of each artificial protein varied, proteins were sorted by length and then distributed across all jobs to ensure that on average each job had a similar number of total peptides to be predicted (and thus took a similar length of time to complete). On average, each job took 35 minutes, totaling over 110 CPU years to complete all predictions.

4.2.5 Classifying TCGA tumours as hot or cold

Within the TCGA The Immune Landscape of Cancer project [149], DNA methylation information was acquired. The top differentially methylated probes between pure leukocyte cells and normal tissue were identified and used to predict the leukocyte fraction (LF) in the TCGA tumour samples. For this analysis, the TCGA tumours in the top third of LF values were classified as “hot”, whereas the bottom third was classified as “cold”. The number of tumours of each type for each cancer site are shown in Figure C.2.

4.2.6 Comparing distributions of self-immunopeptidome sizes

For a cancer dataset, we used HLA data from TCGA [149]. Class I HLA calling was performed using OptiType [154] for 9,957 samples. For a non-cancer dataset, we obtained HLA genotypes from the National Marrow Donor Program (NMDP) [249]. Class I HLA typing was performed using a mix of DNA and sequence-based techniques for 13,996 participants. For each individual, we calculated their self-immunopeptidome size as the number of distinct peptides predicted to bind to their set of class I MHCs. To control for any effect from differing ethnicities skewing the frequencies of certain HLA alleles between the two datasets, we first restricted our data to the individuals with Caucasian ethnicity (TCGA n = 6,415, NMDP n = 7,867). We then tested if the two distributions were statistically different from each other by performing a T test.

4.2.7 Tallying mutations and neoantigens in TCGA

TCGA mutation information was tallied by Ellrott *et al.* [93] and downloaded from the Genomic Data Commons (GDC) (https://gdc.cancer.gov/about-data/publications/mc3-2017; id: 8b851024-2915-4d66-8a84-d03199b616fd; filename: mc3.v0.2.8.CONTROLLED.maf.gz). Class I HLA genotypes were performed by Thorsson *et al.* [149] and downloaded from the GDC (https://gdc.cancer.gov/about-data/publications/panimmune; id: cf05dd5-9653-497a-8c7e-45ba0d1d237a; filename: OptiTypeCallsHLA_20171207.tsv). pMHC predictions were performed as described in Thorsson *et al.* [149].

4.2.8 Survival analysis in TCGA

TCGA clinical data was obtained from Liu *et al.* [250] Supplemental Table S1. Cox proportional hazard models were built using the survival package in R, using the progression free interval variable “PFI.1” from Liu *et al*. [250]. Covariates in survival models were “age at initial pathologic diagnosis”, “gender”, “race”, and “cancer type”. Tumour stage was only available for a subset of cancer sites, so was excluded as a covariate from these pan-cancer survival models.

4.2.9 Comparing presentation of TCGA mutations (*in vivo*) to simulated mutations (*in silico*)

All non-synonymous SNVs from the TCGA mutation file were used, and the frequency of every possible amino acid change was tallied (Figure C.3). These amino acid change frequencies were then used to generate a pool of 50,000 random amino acid changes across the reference human proteome. First, 50,000 positions were randomly selected across the proteome. For each position, the reference amino acid was randomly mutated to a different amino acid using the measured amino acid change frequencies from the TCGA data. All peptides containing these mutations had pMHC predictions generated for all available HLA alleles (5,456,375,705 unique combinations), and results were stored in a database for querying.

For each TCGA subject, a random sample of the above simulated mutations was selected to match the size of the number of non-synonymous SNVs from that subject. Of these selected mutations, all pMHCs corresponding to this subset of random mutations and that TCGA subject’s specific HLA genotype were selected and tallied, acting as a matched, simulated pMHC repertoire.

4.2.10 Identification of expressed SNVs in TCGA

To determine if a TCGA SNV is expressed, we used the Samtools [251] v0.1.8 mpileup command to obtain all bases seen at the genomic coordinate of the SNV from the RNA-seq bam file of that subject. An SNV was classified as expressed if the mutated base was observed at least three times.

4.2.11 Measuring differences in variant position usage from TCGA pMHCs compared to *in silico* pMHCs

This analysis was performed for each peptide length separately. For each subject, we first enumerate the variant position usage within the peptides from TCGA pMHCs, and repeat this for the in silico pMHCs. For each subject, we then filter the data to retain positions that have at least one pMHC with the variant at that position from both the TCGA and in silico sets. We then calculate the frequency that each position is used by dividing the count of each position by the number of peptides of that length in the subject. We then calculate delta, the difference in these frequency values for the TCGA pMHCs compared to the in silico pMHCs. For each position, we perform a T test on these delta values to see if there is evidence that they are significantly different from zero. For visualization, we plot the mean of these delta values for each position, with bars showing the 95 % confidence interval on the mean as reported by the T test. We use the Bonferroni correction to adjust the p-values for multiple testing.

4.2.12 Code and data availability

All custom code relating to the prediction of the self-immunopeptidome is available at https://github.com/scottdbrown/self-immunopeptidome_cancer/. The human immunopeptidome dataset is available at http://doi.org/10.5281/zenodo.1453418.

4.3 Results

4.3.1 Exhaustive binding prediction of all self-peptides to MHC-I

The complete human reference proteome was downloaded from EMBL-EBI, containing 21,006 protein and 71,173 additional isoform sequences. Typically, class I MHC peptides are restricted to 8-11mer peptides due to the closed ends of the MHC binding groove [252]. As such, all possible 8-11mer peptides were extracted from this reference sequence using a sliding window, yielding over 146,000,000 peptides, of which 46,029,730 were unique. For each of the 46,029,730 unique 8-11mer peptides we predicted binding to each of 2,915 HLA class I alleles available in NetMHCpan v3.0. Executing these 134,176,662,950 binding predictions required over 110 CPU years of compute (see 4.2 Methods) and provides a new human immunopeptidome resource.

We tested whether an IC_{50}- or Rank-based threshold would better represent the number of observed MHC-eluted peptides (see 4.2.3 Selecting the most suitable binding prediction threshold). For 66 MHC alleles with Human peptide data available from SysteMHC Atlas [248], we compared the number of unique peptides predicted to bind using the two thresholds to the number of peptides observed to bind these MHC using MS data. By Spearman's rank correlation, thresholding by IC_{50} yielded a better correlation to the observed peptide data than thresholding by rank (IC_{50} $\rho = 0.558$, $p = 1.1 \times 10^{-6}$, Figure C.1, and Rank $\rho = 0.314$, $p = 0.010$). Importantly, this demonstrates that the number of predicted self-peptides as defined by $IC_{50} < 500$ nM correlates with observed experimental data. We filtered the output to include the 987,968,036 pMHCs (0.7 % of all combinations tested) that had predicted $IC_{50} < 500$ nM and this set was used to calculate self-immunopeptidome sizes. The results of this compute are now made available for researchers, obviating the need for these computational predictions to be repeated (http://doi.org/10.5281/zenodo.1453418).

4.3.2 MHC frequency in NetMHCpan training data correlates weakly with peptide presentation properties

As HLA alleles with greater representation in the NetMHCpan training data likely have more reliable binding predictions, we computed the correlation between the fraction of all unique human peptides presented by an MHC, and the number of datapoints for the HLA allele encoding that MHC variant in the NetMHCpan training data (http://tools.immuneepitope.org/static/main/binding_data_2013.zip). We observed a weak correlation between the fraction of all unique human peptides predicted to be presented by an MHC and number of datapoints in the training data (Spearman $\rho = 0.388$, $p = 5.5 \times 10^{-5}$). We saw no difference in fraction of all unique human peptides presented by an MHC for MHC included in the training data vs. those with no training data ($p = 0.1185$, T test). As a separate test, we checked for any effect that HLA population frequency may have on the size of predicted self-immunopeptidomes. This was done using the 330 HLA alleles with non-zero population frequencies in the USA NMDP Caucasian dataset from http://www.allelefrequencies.net [253], selected as most ethnicities within the TCGA dataset are Caucasian. We observed no significant

correlation between population frequency and fraction of unique peptides presented (Spearman ρ = -0.096, p = 0.082). For single MHC molecules, fractions of unique peptides ranged from 0.0 % (*HLA-B*15:137*) to 4.5 % (*HLA-A*02:229*) of all 8-11mer self-peptides (Figure 4.1A). Within the set of all peptides comprising the human immunopeptidome, most peptides are able to be presented by relatively few (< 250) MHC, while some can be presented by upwards of 1,500 different MHC (Figure 4.1B). Taking all 2,915 MHC together, 29.7 % of all 8-11mer self-peptides are predicted to be presented, showing that there is significant overlap between the repertoire of peptides presented by different MHC. Additionally, this suggests that over 70 % of the human peptidome is unable to be presented by MHC and is not surveyed (nor naturally tolerized) by T cells.

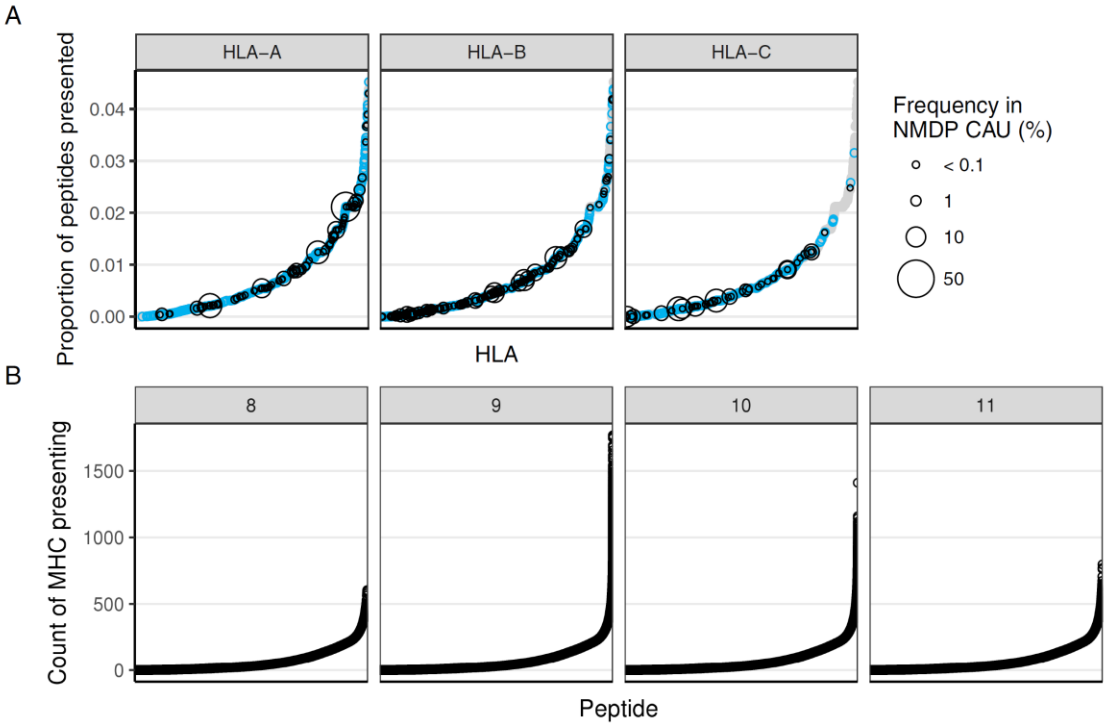

Figure 4.1: Characterization of the self-immunopeptidome. (A) Fraction of the human peptidome presented by each of 2,915 class I MHC. MHC are plotted along the x-axis in increasing fraction (y-axis). Gray points are all MHC combined, and the overlaid blue points identify those from the specific gene (*HLA-A, -B,* or *-C*) in each panel. Black circles show the population frequency, when available, of that allele in the USA NMDP Caucasian dataset. **(B)** Number of MHC able to present each presented peptide. Peptides are plotted along the x-axis in increasing numbers of MHC (y-axis), with each peptide length in separate panels (8mer: *n* = 1,522,052, 9mer: *n* = 4,667,489, 10mer: *n* = 4,704,530, 11mer: *n* = 2,790,531).

4.3.3 The distribution of self-immunopeptidome sizes are similar between cancer and non-cancer datasets

The self-immunopeptidome presented by any individual is dependent on the up to six different class I HLA alleles encoded by their genome. We define the self-immunopeptidome size for an individual to be the size of the (possibly) overlapping sets of distinct 8-11mer peptides predicted to be presented by each of their MHC-I. To compare the distribution of self-immunopeptidome sizes for individuals with cancer compared to those without, we used data from TCGA [149] and NMDP [249].

We obtained class I HLA types for the TCGA dataset, predicted using OptiType [154] as part of "The Immune Landscape of Cancer" [149], and for the NMDP dataset [249] where typing was done by PCR- and amplicon sequencing-based techniques. As most ethnicities in the TCGA dataset are Caucasian, and to control for potential confounding effects of varying allele usage in different ethnicities, we filtered the TCGA and NMDP data to exclude non-Caucasian subjects for this analysis only. The resulting distributions of self-immunopeptidome sizes for both TCGA and NMDP datasets are shown in Figure 4.2 (TCGA 1,767,986 ± 561,474 (mean ± SD), n = 6,415; NMDP 1,797,092 ± 553,010 (mean ± SD), n = 7,867). Self-immunopeptidome sizes for TCGA are slightly smaller than the NMDP distribution ($p = 1.9 \times 10^{-3}$, two sample T test), though the distributions are not distinct enough to have any practical utility in predicting if an unknown subject would belong to either group.

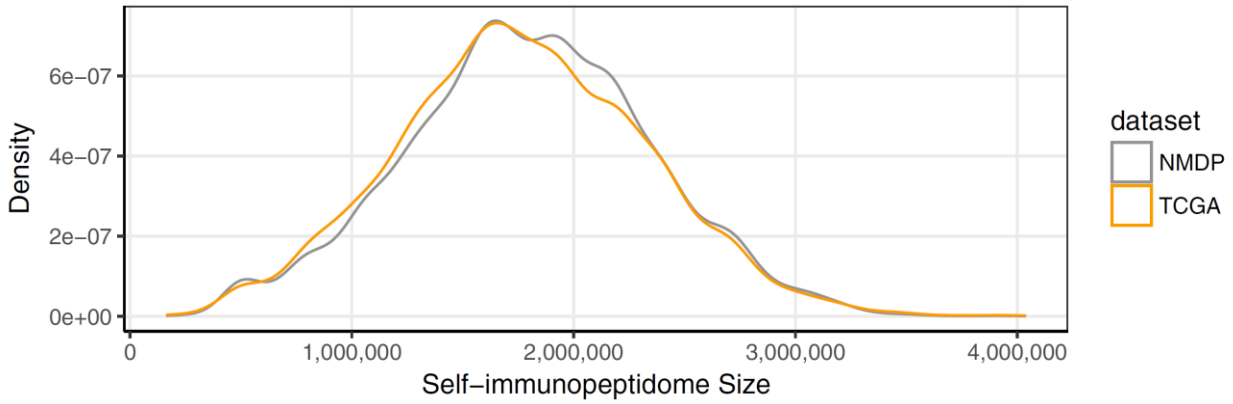

Figure 4.2: Self-immunopeptidome sizes for TCGA and NMDP subjects. Density plot showing the relative frequency (y-axis) of self-immunopeptidomes of varying sizes (x-axis). TCGA (orange) subjects, on average, have smaller self-immunopeptidomes than NMDP (gray) donors.

We tested whether tumours that have a large immune infiltrate (hot) had different self-immunopeptidome sizes than tumours with low levels of immune infiltrate (cold). We classified each tumour as being hot or cold based on the computed leukocyte fraction [149]. We did not see any difference in self-immunopeptidome size distributions when stratifying TCGA tumours by immune infiltration levels (Figure 4.3; ANOVA p = 0.506).

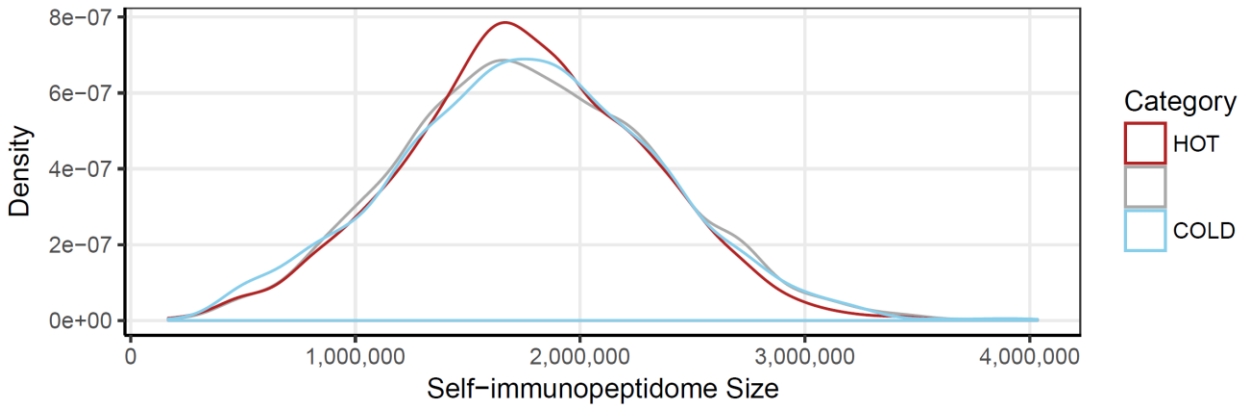

Figure 4.3: Self-immunopeptidome sizes for hot and cold TCGA tumours. Density plot showing the relative frequency (y-axis) of self-immunopeptidomes of varying sizes (x-axis). Hot (red; n = 2,144), cold (blue; n = 1,994), and mid-level (gray; n = 2,115) tumours have overlapping self-immunopeptidome sizes.

4.3.4 In cancer, self-immunopeptidome size correlates with predicted neoantigen load and progression free interval

Despite there being only a small difference in self-immunopeptidome size between cancer and non-cancer datasets, self-immunopeptidome size may have a clinically relevant effect within the cancer dataset. We hypothesized that in TCGA, individuals with larger self-immunopeptidomes would have improved outcomes due to there being a higher probability of mutations in these tumours generating neoantigens.

As SNV neoantigen data from TCGA was calculated as the number of pMHC containing a mutated amino acid [149], we can combine the coding SNV mutational load and self-immunopeptidome size in these subjects to approximate their SNV neoantigen load. If we express the self-immunopeptidome size as a fraction of all unique peptides that are presented by that genotype, and multiply this by the coding SNV mutational load to get an approximated SNV neoantigen load, we observe a strong positive correlation between approximated SNV neoantigen load and TCGA SNV neoantigen

load (Pearson r = 0.987, $p < 2.2 \times 10^{-16}$). This result suggests that using a combination of coding mutational load and self-immunopeptidome size together, as approximated SNV neoantigen load, could be very useful as an indicator of tumour immunogenicity because it does not require exhaustive neoantigen predictions to be performed.

To test the utility of self-immunopeptidome size as a measure of tumour immunogenicity in a pan-cancer context, we performed Cox proportional hazard survival analysis on the TCGA data using progression free intervals as the endpoint [250]. In a multivariate Cox-PH model containing race, age, gender, cancer type, and HLA diversity as covariates, increases in self-immunopeptidome size alone did not significantly decrease the hazard rate, (HR = 0.921 for an increase in self-immunopeptidome size of 1 million peptides, p = 0.054, 95 % CI: 0.847 – 1.002; details of full model in Table C.1). Creating an interaction term between self-immunopeptidome size and cancer type did not show statistical significance in any cancer type, likely due to decreased sample size within each cancer type and demonstrating any survival effect is not restricted to any one cancer type. Further, an ANOVA test comparing the two models with and without the interaction term showed that the model with the interaction does not explain any more of the observed variance (p = 0.2325), and one model does not fit the data better than the other (*AIC* = 25,373.41 for the model with the interaction term *vs.* 25,359.40 without). When switching the predictor in the model without the interaction term from self-immunopeptidome size to the approximated SNV neoantigen load described above, a significant protective effect is observed (HR = 0.995, p = 0.003, 95 % CI: 0.991 – 0.998; details of full model in Table C.2). We obtain comparable results when using the comprehensive TCGA SNV neoantigen load (Table C.3) and observe that the two models fit the data equally well (*AIC* = 25,350.55 using comprehensive TCGA SNV neoantigen load *vs*. 25,348.77 using approximated SNV neoantigen load), demonstrating that in the context of outcomes, self-immunopeptidome size and mutational load combine to provide the same clinical information obtained by comprehensive neoantigen predictions.

4.3.5 Differential patterns of peptide presentation derived from *in vivo* and *in silico* mutations are consistent with immunoediting

We hypothesized that evidence of immune surveillance and immune evasion would be detectable by comparing pMHCs derived from TCGA SNVs to pMHCs originating from *in silico* generated random mutations which have not undergone immunoediting. For every TCGA subject, we used the predicted SNV neoantigens from above [149]. Then, we generated a matched set of *in silico* coding SNVs from random positions throughout the proteome (with amino acid change frequencies modelled after those observed in the TCGA SNVs (Figure C.3)) and we predicted pMHCs from these *in silico* SNVs. As expected, there is a high correlation between the number of TCGA and *in silico* mutant pMHCs per subject (Pearson $r = 0.999$, $p < 2.2 \times 10^{-16}$), as these were derived from the same number of starting mutations and the same set of HLA alleles. We further stratified the TCGA predicted pMHCs by the expression of the source mutation. For each of 5,748 TCGA subjects that we have RNA-seq data for, we classified each of the 1,181,367 coding SNVs as expressed if there were at least 3 sequence reads containing the variant base. We identified evidence of expression for 417,335 (35 %) of these coding SNVs.

To investigate the effect of immune editing in the TCGA subjects, we compared the predicted immunogenicity of expressed SNVs, non-expressed SNVs, and random *in silico* generated SNVs. Predicted immunogenicity was calculated as the number of neoantigens per SNV. Within each subject, potential neoantigens are defined as the subset of the up to 38 peptides (all 8-11mers containing the variant) × up to 6 HLA alleles = up to 228 peptide-MHC pairs that are predicted to bind. Importantly, every SNV may generate zero or a few neoantigens. We hypothesized that there would be fewer neoantigens per expressed TCGA SNV (lower predicted immunogenicity) because cells carrying SNVs generating many neoantigens would have been depleted by neoantigen-reactive T cells. Indeed, we observed fewer neoantigens per expressed SNV compared to both *in silico* and non-expressed SNVs (Figure 4.4; $p < 2.2 \times 10^{-16}$, paired T tests). Interestingly, we see more neoantigens per non-expressed SNV than per random SNV (Figure 4.4; $p = 1.9 \times 10^{-14}$, paired T test), suggesting that the cancer cells that downregulate mutated genes can avoid detection by T cells, resulting in the

accumulation of inconsequential, non-expressed potentially neoantigenic mutations in these non-expressed alleles. When looking at each cancer site individually, the trend of more neoantigens per non-expressed SNV and fewer neoantigens per expressed SNV was maintained for all cancer sites except LIHC, CESC, BLCA, and SKCM (Figure C.4). It should be noted that unlike other TCGA cancer sites, the majority of SKCM samples are from lymph node metastasis [250], and as such they represent tumours at a different stage of development and with biased immune cell content compared to the rest of the dataset.

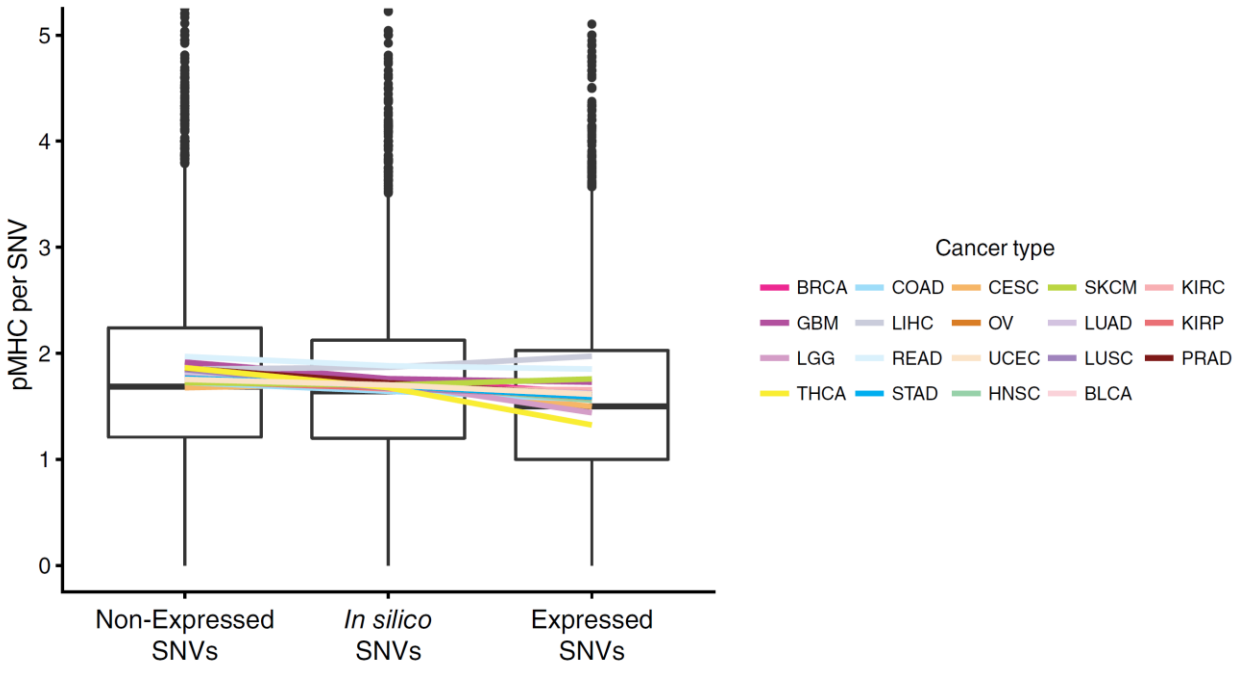

Figure 4.4: Evidence of immunosurveillance. Boxplots for the average immunogenicity (neoantigens per SNV; y-axis) per subject for non-expressed, *in silico*, and expressed SNVs (x-axis). Coloured lines showing the average number of neoantigens per SNV for each cancer type are overlaid. Outliers above a threshold of 5 neoantigens per SNV are omitted from plot to simplify the display.

Directly comparing numbers of neoantigens per SNV from expressed and non-expressed SNVs in each TCGA subject, we observe that the general trend of more neoantigens per non-expressed SNV (as shown in Figure 4.4) appears to reverse for samples that have greater than two neoantigens per expressed SNV (Figure 4.5A). One interpretation of this observation is that tumours with higher numbers of neoantigens per expressed SNV have been able to retain more expressed neoantigens because they

have suppressed the immune response by mechanisms such as an immunosuppressive microenvironment or downregulation of MHC. To explore this further, we classified samples as having a suppressed immune response if they met the following criteria: (1) they have more neoantigens per expressed SNV than neoantigens per non-expressed SNV, and (2) they have greater than two neoantigens per expressed SNV. Survival analysis comparing these two groups supports this notion (Figure 4.5B), with the samples having a putatively suppressed immune response showing decreased Progression Free Intervals (PFI) (HR = 1.138, p = 0.027, 95 % CI = 1.015 – 1.275; multivariate Cox-PH model with cancer type, age, race, and gender as covariates).

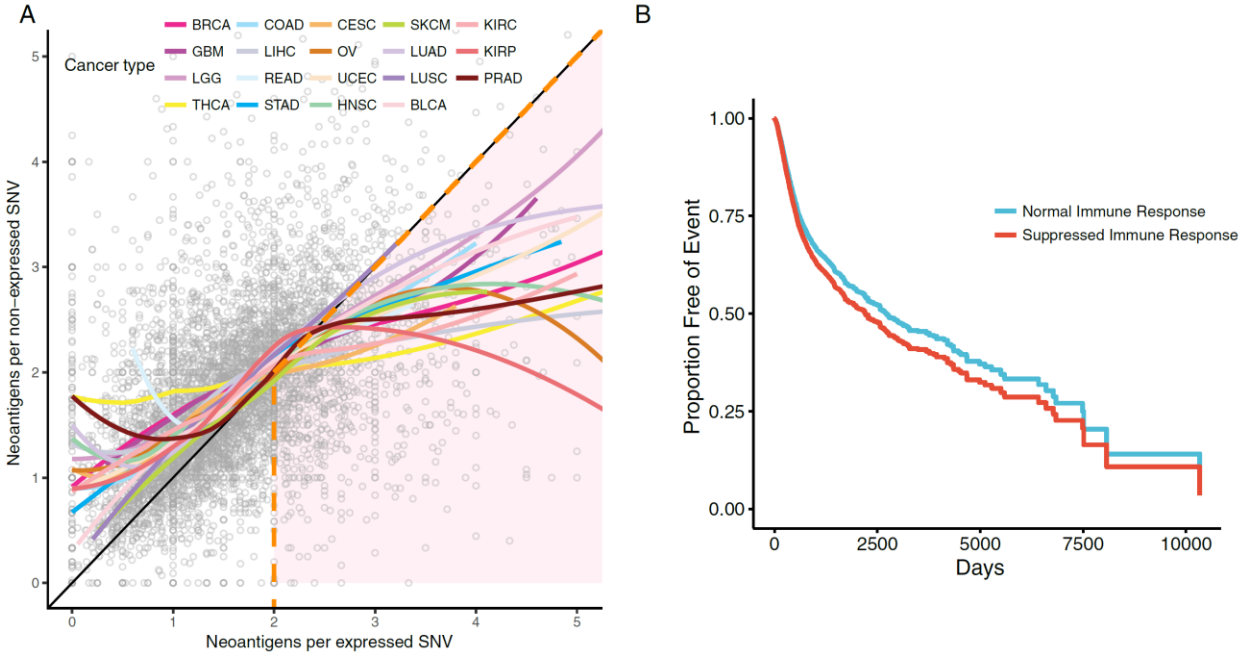

Figure 4.5: Evidence of immune-evasion. (A) Scatterplot showing direct comparison of the number of neoantigens per expressed (x-axis) or non-expressed (y-axis) SNVs. Coloured lines show the locally weighted average neoantigens per non-expressed SNV (LOESS) for each cancer type across x-axis values. Orange dashed line separates subjects predicted to have suppressed immune response (right of line, pink shade, *n* = 956) from those with normal immune response (left of line, *n* = 4,792). Plot zoomed in to show bulk of the data – 55 outliers of 5,748 points fall outside of this window. **(B)** Survival curves showing the effect of evidence of normal immune response (blue) compared to suppressed immune response (red), adjusted for the effect of covariates from the Cox proportional hazards multivariate model.

4.3.6 Relative depletion of variants in MHC-binding anchor positions of peptide epitopes identify potentially immunogenic positions

We were interested in whether different amino acid positions within neoantigens show different signatures of immunogenicity. Within the set of samples with evidence of a normal immune response (Figure 4.5), expressed mutations from TCGA exist within established tumours and coexist with the host immune system. Therefore, neoantigens originating from these mutations were not immunogenic enough to result in tumour eradication within these subjects. One factor that may influence immunogenicity is the position within the presented peptide that the variant resides [254]. We investigated whether there was a bias in the usage of positions within the peptide for the variant amino acid in neoantigens from expressed TCGA variants compared to random, *in silico* generated mutations. We noted that an important factor to consider is whether, for each neoantigenic peptide, the corresponding wildtype peptide is also predicted to bind to the same MHC. If the wildtype peptide is also presented, this would be expected to result in T cell tolerance to the wildtype peptide and may highlight certain amino acid positions as being relevant to breaking or taking advantage of this pre-existing tolerance when mutated. To investigate this, we looked up wildtype pMHC binding scores for all expressed TCGA and *in silico* mutations in our human immunopeptidome dataset. All analysis was performed on the set of neoantigens where both the mutant and matched wildtype peptide was also predicted to bind with an $IC_{50} < 500$ nM.

We limited our analysis to 9mers, which are the most common peptide length and have the most well-defined MHC-binding interactions [161]. Looking at neoantigens derived from random, *in silico* generated mutations, we observe a depletion of variants at positions 2 and 9 of these neoantigens compared to the other positions when the corresponding wildtype peptide also binds the MHC (Figure 4.6A; top panel). This is expected, as these positions are the most influential on peptide-MHC binding and most likely to confer loss of MHC binding when mutated [255]; these are the two canonical MHC-binding anchor positions. This trend was recapitulated in the neoantigens derived from expressed TCGA mutations (Figure 4.6A; bottom panel), confirming that this effect is intrinsic to peptide-MHC binding and not an artefact of *in silico* mutagenesis. Ignoring

wildtype binding status yielded a uniform distribution across the possible variant positions (Figure C.5).

To control for this non-uniform distribution of variant position usage, we directly compared the frequency that each position was used in the TCGA and *in silico* datasets. Positions which are relatively depleted of mutations in the expressed TCGA dataset relative to the random *in silico* dataset may be the result of immune editing during tumour development, deleting cells which carry mutations at these positions. Similarly, positions that are relatively enriched in the TCGA dataset may be non-immunogenic, being able to persist while co-existing with the host immune system. To identify positions that have an enrichment or depletion of mutations in the TCGA dataset relative to the random *in silico* mutation dataset, we first converted the raw count of neoantigens with variants at each position into the frequency that each variant position is used within each subject's neoantigen repertoire. For each subject, we then calculated the difference in frequencies at each position between the TCGA and random *in silico* derived mutations, and tested, over the entire dataset, whether these differences were statistically significant.

We observed a significant depletion of position 2, 3, and 9 variants in TCGA neoantigen data compared to the *in silico*-derived neoantigens (Figure 4.6B). Note that positions 2 and 9 held the lowest number of mutations in both datasets because we limited the data to cases where the wildtype and mutant versions are both predicted to bind, and mutations at these positions will typically impair peptide-MHC binding. However, mutations at positions 2 and 9 do not always result in loss of binding, and in some cases can even increase binding affinity [256]. Further, while these positions would be expected to remain buried in the MHC groove and thus hidden from the immune system, they may enhance the immunogenicity through conformational changes elsewhere in the peptide. Therefore, the relative depletion of TCGA position 2, 3 and 9 variants relative to *in silico* variants may be due to enhancement of immunogenicity conferred by these mutations as a result of improved peptide-MHC binding stability [257,258], or from changes to the peptide conformation within the MHC binding groove. It is well established that subtle changes to peptide conformation can have large effects on T cell reactivity [259,260], exposing regions of the peptide to the

TCR that were previously concealed. Conversely, we observed a relative enrichment of TCGA position 8 (so called anchor-adjacent) variants, suggesting changes here are tolerated because they do not increase immunogenicity (Figure 4.6B). This general trend of depletion at the anchors and enrichment just interior to the anchors is seen across all peptide lengths (Figure C.6). In the context of overcoming existing T cell tolerance to wildtype peptides, these data suggest that, counterintuitively, variants at anchor positions are the most immunogenic and are selected against during tumour development.

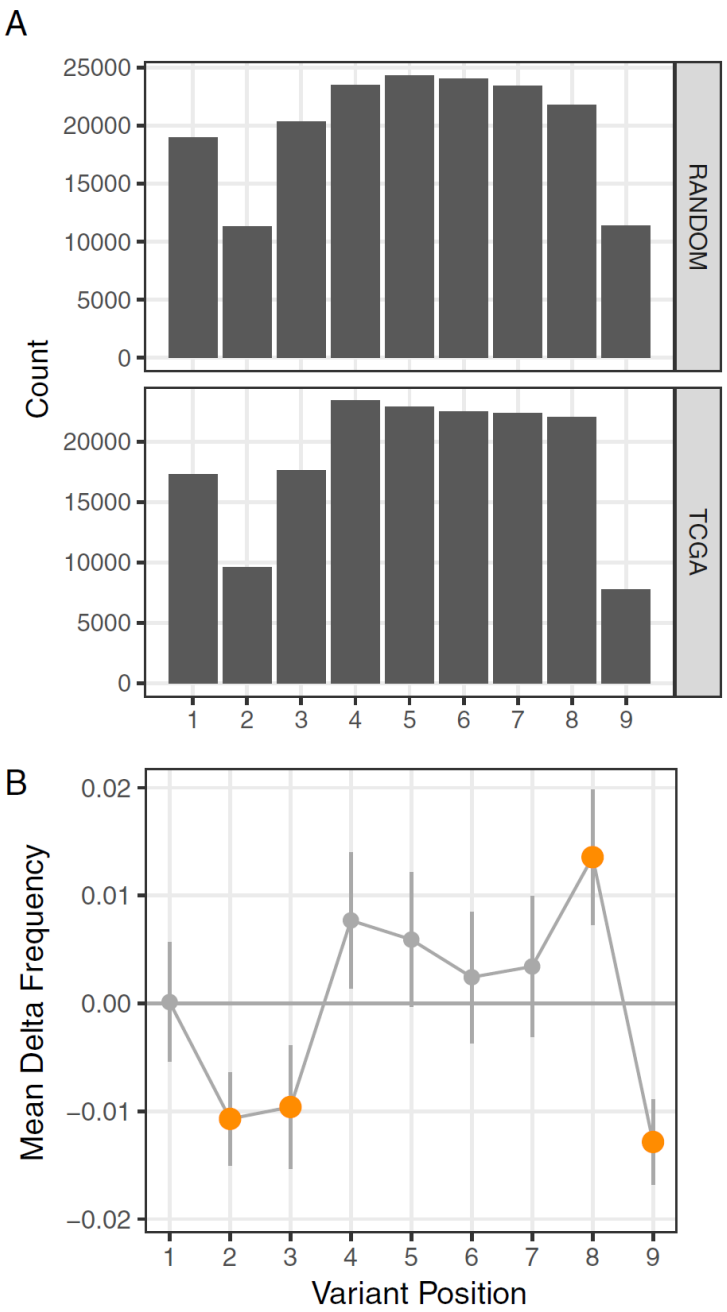

Figure 4.6: Usage of positions for variants within presented peptides. (A) Counts of neoantigens containing the variant at each position within the peptide. Variants occur in positions 2 and 9 at the lowest frequency. This trend is consistent for both random *in silico* derived mutations (top panels) and TCGA-derived mutations (bottom panels). **(B)** Differences in frequency (y-axis) of the variant amino acid being in each position (x-axis) of a presented peptide for TCGA mutations compared to random mutations. Mean values are shown (points), with lines showing 95 % confidence intervals of the means. Positions with significant enrichment or depletion ($p_{adj} < 0.05$, T test) are displayed larger and coloured orange. Only data for 9mers shown; 165,248 TCGA neoantigens and 179,002 *in silico* neoantigens for 4,533 subjects.

4.4 Discussion

We performed exhaustive binding predictions between every unique 8-11mer peptide that exists in the reference human proteome to nearly 3,000 MHC molecules available for prediction, generating, to our knowledge, the largest set of peptide-MHC binding predictions to date, and which can now serve as a community resource. This resource supported the fast and efficient characterization of thousands of individuals from TCGA and NMDP for their predicted ability to present self-peptides. It is important to note that these predictions ignore protein expression and abundance in different cell types, antigen processing requirements, epitope destruction, and issues such as proteasome-generated spliced peptides that might represent a substantial component of the self-immunopeptidome [261,262]. Currently, it is not possible to perform our predicted self-immunopeptidome analysis for these spliced peptides as there are no algorithms to predict their occurrence nor are there well annotated databases of those that exist.

Individuals with cancer from TCGA have marginally smaller self-immunopeptidomes compared to non-cancer NMDP individuals. More importantly, within the TCGA dataset having larger self-immunopeptidomes correlates with better outcomes, independent of HLA diversity. This supports previous findings of an HLA-effect on survival [140,195,263]. We did not see any correlation between hot or cold tumours and self-immunopeptidome size. This is not entirely unexpected, as self-immunopeptidome size is a measure at the level of the subject, whereas each subject can have both hot and cold tumours [264]. Additionally, self-immunopeptidome size and mutational load combine to approximate the neoantigen load, and this strongly correlates with the actual predicted neoantigen load (based on personalized binding predictions of all mutant peptides to MHC), supporting the potential utility of approximated neoantigen load as a clinical metric in assessing the immunogenicity of tumours without the need to perform more exhaustive neoantigen predictions. This approach may also facilitate the calculation of approximated neoantigen loads in subjects from all mutation types, not limited to SNVs, without the additional neoantigen prediction processing required for more complex mutation types.

Immunoediting is a well accepted phenomenon that occurs during cancer development [105]. By comparing observed mutations from immune-exposed TCGA

tumours to *in silico* generated mutations, we were able to detect signals of immune-evasion within the TCGA data. It is important to note that our *in silico* mutations are random and do not necessarily confer the same cancer growth advantage, or in fact any biological relevance, that is likely found in the set of TCGA mutations and thus are used as a measure of baseline pMHC generation. Compared to *in silico* mutations, we observed a general trend of decreased immunogenicity for expressed TCGA mutations, and an increase for non-expressed mutations. This supports the view that the majority of TCGA tumours are immune-edited, as we see higher immunogenicity from non-expressed SNVs than would be expected by chance. This is likely the result of immune-editing over time shaping the mutational profile of these cancer cells and resulting in the relative accumulation of immunogenic mutations in non-expressed genes in contrast to the deletion of cancer cell clones containing immunogenic mutations in expressed genes. Under this framework, we also identified samples that showed evidence of a suppressed immune response, permitting relatively more immunogenic mutations in expressed genes. These subjects demonstrate decreased progression free survival, supporting the concept that these individuals harbour tumours which have suppressed the natural immune response to the tumour.

Given that highly immunogenic mutations could be rapidly recognized by the immune system and cancer cells containing these mutations would not survive, we assume that the mutations we see have decreased immunogenicity. By comparing the variant positions within the presented peptides to the positions containing the *in silico* mutations, we were able to identify positions that were depleted (more immunogenic) and enriched (less immunogenic). Immunogenic positions were canonical MHC binding positions, likely resulting in significant changes to the topography of the presented peptide and a greater likelihood of breaking T cell tolerance, or increasing the stability of the peptide-MHC complex [254]. Non-immunogenic positions were anchor-adjacent. Changes in anchor-adjacent positions may represent an optimization between effects on MHC-binding and visibility of the variant to the T cell receptor. This is supported by work describing positions important for MHC binding and T cell interaction [254,265–267], summarized in Figure 4.7. These observations on the relative importance of

certain positions in influencing the immunogenicity of peptides may help refine epitope immunogenicity prediction algorithms.

Figure 4.7: Theoretical importance of positions in presented peptides. Interactions with MHC (gray) and TCR (green) based on existing knowledge of immunogenicity in the literature. Two positions representing the minimization of effects on interactions with both MHC and TCR are denoted with arrows.

Our findings were generalizable across cancer types, with the identification of immune-edited tumours not being restricted to any one cancer type. Similar efforts to characterize different immune subtypes within TCGA have recently been undertaken [149]. Based on gene expression data, six immune subtypes were identified in TCGA tumours, described according to different immune pathways that are most active in each subtype. Across these TCGA immune subtypes, our measure of average mutation immunogenicity (neoantigens per expressed SNV) was smaller for subjects in the immunologically quiet (C5) cluster, though this may be due to a higher number of zero counts in this cluster due to a relative dearth of mutations.

Given the TCGA tumour samples were obtained pre-treatment, our measure of immunogenicity is relevant within the context of the natural immune response. These predictions and trends may not extend to cases where immunotherapies are used to modulate the immune response – pMHCs present in these immune-exposed samples which are assumed to be not naturally immunogenic may still form potent immune targets for immunotherapies. Future studies applying similar comparative approaches to the mutational landscapes of tumours before and after immunotherapy would be useful in identifying predictive measures of immunogenicity in these contexts.

This immunopeptidome analysis can readily be extended to proteomes from different sources and may yield novel insights into MHC-presentation of peptides from infectious agents. Human MHC variability has been shaped by hundreds of thousands of years of evolution in the presence of pathogens, exerting selective pressure on the human genome to maintain a high level of MHC variability. Preliminary data exploring this phenomenon using the techniques in this chapter to create immunopeptidomes for a variety of species is presented in Appendix D (Evolutionary analysis of immunopeptidomes).

Chapter 5: Discussion, conclusions, and future directions

Cancer immunotherapies have shown tremendous potential for treatment of a variety of cancers [110,113,205,206,229], leveraging the existing immune capabilities of the individual to recognize and respond to the cancer. Further, we now understand that the natural immune response is capable of recognizing and suppressing tumour development, effectively targeting and killing cells that have acquired mutations [268,269]. However, we still do not fully understand the specifics defining the immunogenicity of mutations, the relevance of the T cell receptor repertoire, or if genetic markers exist that can inform the effectiveness of class I MHC-presentation of antigens in an individual.

This thesis summarizes my work on these open research questions relating to the TCR-pMHC axis of immune cell recognition, focussing on immune recognition of cancer cells. I apply novel computational methods to existing genomics datasets to obtain a pan-cancer view of tumour-immune cell interactions. My analysis takes advantage of the diverse TCGA dataset to obtain a pan-cancer view of these interactions, highlighting general trends of the interaction between T cells and tumours that are not specific to certain tumour types.

In chapter 2, I performed the first meta-analysis of solid tumour neoantigen landscapes, identifying the neoantigen burden in each subject by performing exhaustive pMHC binding predictions and strict filtering to identify those most likely to be immunogenic. This advance was enabled by the development of methods to perform HLA typing from NGS data [143]. Using this data, I correlated neoantigen load with levels of T cell infiltration, survival, and markers of immune inhibition. It was known that T cell infiltration conveyed a survival advantage, and my work provided evidence that it was neoantigens which were driving the T cell infiltration. Further, it demonstrated that while there were tumours containing a dense T cell infiltrate and numerous neoantigen targets for these T cells, the T cells showed evidence of being inhibited. This suggested that these tumours may be prime candidates for checkpoint blockade therapies, supporting the use of neoantigen load as a marker for checkpoint blockade success. An important implication of these results was that personalized computational predictions of neoantigens could be used to yield clinically relevant information, able to filter the set of

all mutant peptide-MHC combinations down to a set likely to be immunogenic, which correlates with T cell infiltration and overall survival.

Having assessed the personalized neoantigen burden in the TCGA tumours from genome sequencing data, in chapter 3 I assessed the utility of the same bulk sequencing datasets for TCR repertoire characterization. I performed extensive testing and optimization of tools for TCR sequence annotation, ensuring sensitivity was maintained while specificity was maximized when applying these tools to bulk RNA-seq from solid tumours or sorted T cell populations. This demonstrated that RNA-seq data contains detectable TCR information, the utility of which is dependent on the immune cell content of the sample (ie. solid tumours with infiltrating T cells vs. sorted pure T cells). Within solid tumours, where the number of T cells is low, I showed that the TCRs detected via RNA-seq are from most abundant T cells in the sample and are not just a random sampling from all T cells present, highlighting the utility of this approach. Importantly, there was significant overlap between the tumour and matched normal TCR repertoires, indicating that most TIL are bystanders and are not necessarily enriched in tumour-specific T cells, a result which is gaining acceptance [270]. Within sorted T cell populations, the extracted TCRs can readily identify clonally expanded T cells and can identify the presence of malignant clones with greater sensitivity than flow cytometry using cell surface markers.

In chapter 4, I performed exhaustive computational predictions to generate the human immunopeptidome and used this to calculate the self-immunopeptidome size for thousands of individuals with or without cancer. This showed that HLA genotype defines the diversity of peptides able to be presented, but does not offer any predictive information on the risk of developing cancer. Self-immunopeptidome size taken together with coding mutational load was able to approximate the classically predicted neoantigen load, and correlated with progression free survival. By assessing the neoantigen landscape from TCGA mutations as well as *in silico*-generated mutations, I was able to gain information on the immunogenicity of neoantigens, identifying certain positions that have a greater influence on immunogenicity of presented mutant peptides.

Due to the heavy use of TCGA data in this thesis, it is important to note that characterizing the immune composition in tumours was not a goal of TCGA, and as such, tumour specimens used for sequencing were selected to maximize tumour cell content, consequently limiting the immune cell content. Despite this, immune cell information is present in TCGA data, and broad trends can be uncovered with this limited information, as demonstrated by the TCGA PanCancer Atlas “Immune Landscape of Cancer” project [149]. In this paper, TCGA expression data was used to identify six immune subtypes which spanned cancer tissue types. These subtypes had varying immune cell composition, mutational loads, and outcomes, and recapitulated known immune trends which may be responsible for these differing outcomes.

Another limitation of this data is the restricted nature of TCGA sample access; results generated from this data are hypothesis generating, unable to be directly validated. I have undertaken orthogonal validation measures, correlating results with observed clinical measures, generating simulated datasets, and utilizing separate distinct datasets, where applicable.

In the remainder of this chapter I address some of the limitations of this work, relevant recent advances, outstanding problems, and future directions for the field.

5.1 Predicting cancer neoantigens from tumour genome data

It has become well recognized that T cells can respond to tumour neoantigens, and these neoantigens can predict response to immunotherapies [269,271,272]. Identification of these neoantigens, however, is still a challenging endeavour. For my analysis, I focused on class I MHC-presented neoantigens for two reasons. Biologically, these are responsible for the direct cytolytic attack by $CD8^+$ T cells, and computationally, the peptide-MHC prediction algorithms and HLA prediction algorithms are more accurate for class I than class II. It has been demonstrated that class II MHC-presented neoantigens also play an important role in cancer [114,228,229,273], so as prediction algorithms improve it will be beneficial to perform comprehensive neoantigen predictions for class II antigens in addition to class I.

The general framework developed in chapter 2 for using mutation data to generate mutant peptide sequences from the reference proteome, and predicting

binding of these mutant peptides to MHC, has been incorporated into multiple modern neoantigen prediction tools [164,274–277]. However, this framework does not consider (1) phasing of proximal mutations, (2) germline variation, or (3) silencing of the gene. Firstly, under the framework used in chapter 2, every mutation is processed individually. If there are two proximal mutations that occur on the same allele, and are within 8 – 11 amino acids from each other, the mutated peptides generated, computationally, would not contain both of these mutations, and thus would not be an accurate representation of the mutant peptides truly present in the sample. Secondly, since the flanking peptide sequence around each mutation is obtained from the reference proteome, germline variation present in the sample, such as SNPs, were ignored. Both of these differences in peptide sequence between what is generated computationally and what is truly present in the sample can have a significant impact on the peptide-MHC binding predictions, resulting in incorrect neoantigen predictions. Thirdly, while the work presented in chapter 2 required expression of the parent gene, it did not require allele-specific expression of the mutation. It is possible that expression of a single copy of a gene is sufficient for cell survival, and a cancer cell containing an immunogenic mutation in such a gene may silence the mutant copy of the gene, effectively negating the immunogenicity of the mutation [127]. Recent work by Rubinsteyn and colleagues describes a computational tool (Isovar) which offers an elegant solution to these three problems by using RNA-seq reads instead of the reference proteome to generate mutant peptide sequences, ensuring that all generated peptide sequences are derived from nucleotide sequences observed to be present in the tumour [278].

An additional benefit of Isovar is its ability to generate mutant peptide sequences for indel variants, whereas my work in chapter 2 is limited to SNVs. Indels, as well as gene fusion variants, represent a greater change from normal sequence compared to SNVs, and thus should offer greater potential for immunogenicity. Indeed, analyses predicting neoantigens from indel mutations found that there were a greater number of predicted neoantigens per indel variant than per SNV variant [149,279]. Importantly though, predictions of immunogenicity on indel-derived pMHCs is made difficult as the definition of the corresponding wildtype peptide for these variants is unclear. My work in chapter 4 offers insights into the immunogenicity of variants at specific positions within

the presented peptide that is dependent on the corresponding wildtype peptide also being presented. Additionally, many state-of-the-art methods in immunogenicity prediction require a matched wildtype peptide-MHC binding prediction for their measures of immunogenicity [280–282]. This is an open issue that will have to be solved in order to facilitate improved neoantigen predictions from a variety of mutation types, integrating peptide-MHC binding as well as other measures to predict the immunogenicity of a variant [168,254,281,283].

The neoantigen predictions performed in chapter 2 are primarily based on binding affinity predictions between peptide and MHC. While peptide-MHC binding is required for antigen presentation, it alone is not sufficient. Much of the training data for NetMHCpan (≤ v3.0) is derived from low-throughput measures of peptide-MHC affinity via *in vitro* peptide-binding assays. As such, these predictions do not encompass all aspects of the antigen presentation pathway. Data derived from tandem MS of acid-eluted MHC-peptides provides a more direct view of the immunopeptidome present on a cell [284,285]. Peptides identified using MS can be assigned to specific MHC molecules either by performing peptide-MHC binding predictions [286], or by using cells modified to express a single HLA allele [287]. Recent work to create a central repository for MS data from MHC-eluted peptides hopes to improve these peptide-MHC predictions [248,288]. Indeed, the creators of various peptide-MHC prediction algorithms have begun to improve their algorithms by incorporating this MS-derived MHC-eluted peptide information to generate predictions on whether a peptide will be found presented by an MHC, rather than predicting binding affinity [289–292]. These predictions indirectly encompass the upstream processing of peptides that end up on the cell surface, offering improved predictions.

Other methods of prediction are also being developed. There is renewed interest in PSSM-based peptide-MHC binding predictors, with new tools showing improved performance [293]. Structure-based approaches and molecular docking methods are also being actively pursued [294,295], and as more structure-based training data is created and computational power increases, these approaches will have the potential to surpass sequence-based prediction algorithms.

It is important to note limitations of the TCGA tumour data used for the analyses presented in this thesis. The genomic datasets represent a single view of the tumour, ignoring both spatial and temporal heterogeneity. Within a single tumour biopsy, multiple cancer cell clones may exist with different mutational profiles [104,296,297]. Perhaps unsurprisingly, there is evidence that neoantigens derived from clonal mutations (those shared among all cancer cells in the tumour) are better targets, and predict response to immunotherapy better than sub-clonal mutations (those found in a subset of cancer clones) [166,272]. Targeting these clonal neoantigens will ensure that all cancer cells are attacked, whereas targeting sub-clonal neoantigens will only delete those mutation-carrying cells, selecting for cells not carrying that mutation and leading to resistance to the therapy. Computational methods exist that can utilize the variant allele frequencies to infer the underlying population structure of cancer cells [298–301], however, this still does not give information on spatially distinct tumours nor the temporal dynamics that occur as immune cells interact with cancer cells. This type of information is only attainable by performing multiple biopsies, unavailable in the TCGA data. In other cohorts, this strategy has yielded high resolution views of the dynamics of tumour heterogeneity over time [264,271,302–305], and as this type of data becomes more available, the spatiotemporal dynamics of the interactions between immune cells and cancer cells will become attainable.

The work presented in chapter 2 was performed on a subset of TCGA data that was available at the time, and was further strictly filtered to samples for which I was able to obtain unambiguous *HLA-A* calls (n = 515). Subsequent to this analysis, the general approach and findings were replicated by Rooney et al., performing an analogous analysis on an expanded set of TCGA data (n = 4,486) [155]. There, they used an alternative measure of CTL infiltration which is an expression marker of cytolytic activity rather than bulk *CD8* expression, and did not include expression filtering on the neoantigen predictions. Their results also demonstrate the association between neoantigen load and cytolytic activity of $CD8^+$ T cells, and show that the presence of viruses correlates with this cytolytic activity marker. Recently, I contributed the neoantigen prediction pipeline from chapter 2, modified to only require IC_{50} < 500 nM, to the TCGA PanCancer Atlas Immune Landscape of Cancer project [149]. These

neoantigen predictions were performed on the entire solid tumour dataset (*n* = 8,546). Here, when tested within each cancer type, the correlation between neoantigen load and survival was limited to a small subset of cancer types. The difference between this result and those reported above may be because neoantigen predictions were performed for all HLA genes (*HLA-A*, *-B*, and *-C*) instead of *HLA-A* only, a more accurate HLA caller was used, and more up-to-date survival data was used. Neoantigen load was predictive of survival in five of six immune subtypes, suggesting that the prognostic effect of neoantigen load is dependent on the overall immune signalling present in the tumour rather than the anatomical tumour type.

While immune checkpoint blockade has shown excellent responses in a variety of cancers, it is not effective in 100 % of patients (Table 5.1). Therefore, there is a need for biomarkers that can predict response. Studies have investigated the clinical utility of a variety of markers including T cell infiltration [304,306], T cell clonality [307], and *PD-L1* expression [308]. Intuitively, for this kind of immunotherapy to be effective, there must exist sufficient T cell targets for the T cells to respond to. For this reason, mutational load and neoantigen load have also been investigated for their use as markers of response [118–120,129,272,309,310]. In chapter 4 I describe the self-immunopeptidome as defined by an individual's *HLA* genotype. I show that a metric based on combining the mutational load of coding mutations with the relative self-immunopeptidome size (the approximated neoantigen load) can yield comparable information to comprehensive neoantigen predictions. Thus, I propose that this approximated neoantigen load could be used as a metric to predict response to immunotherapies in the clinic, negating the need for more computationally intensive neoantigen predictions. It is likely that a combination of this metric and other biomarkers would be required to effectively stratify patients by probability of response. As more data becomes available on immunotherapy response, more accurate biomarkers will become attainable by interrogating how these markers differ between responders and non-responders.

Table 5.1: Summary of approved indications and their Observed Response Rates (ORR) for checkpoint blockade.

Drug	Target	Approved Indication	ORR
Pembrolizumab (KEYTRUDA) [311]	PD-1	Melanoma	33 %
		Non-small cell lung cancer	45 %
		Head and neck	16 %
		Classical Hodgkin lymphoma	69 %
		Urothelial carcinoma	29 %
Nivolumab (OPDIVO) [312]	PD-1	Melanoma	32 %
		Non-small cell lung cancer	20 %
		Head and neck	13.3 %
		Classical Hodgkin lymphoma	65 %
		Urothelial carcinoma	23.4 %
		Renal cell carcinoma	21.5 %
Avelumab (BAVENCIO) [313]	PD-L1	Urothelial carcinoma	13.3 %
		Metastatic Merkel cell carcinoma	33 %
Durvalumab (IMFINZI) [314]	PD-L1	Urothelial carcinoma	17 %
Atezolizumab (TECENTRIQ) [315]	PD-L1	Urothelial carcinoma	14.8 %
Ipilimumab (YERVOY) [316]	CTLA-4	Melanoma	5.7 %
		Renal cell carcinoma *	41.6 %
		Colorectal carcinoma *	49 %

*Ipilimumab administered in combination with nivolumab.

5.2 Extracting TCR repertoire information from RNA-seq

While I have demonstrated that TCR repertoire information can be obtained from RNA-seq datasets, the information obtained can be quite limited. As discussed earlier, TCGA tumour sample accrual standards were set to maximize the tumour cell content, disqualifying tumour samples with a strong immune infiltrate, limiting the number of T cells present to be surveyed. Additionally, the majority of the TCGA RNA-seq datasets contain reads 50 bp in length, which restricts the length of CDR3s that can be detected (the median length for beta chains is 45 bp; Figure B.3). Finally, due to variability in sequencing read length and depth across the TCGA dataset, I subsampled all of the data to the lowest common denominator (100,000,000 × 50 bp sequence reads). This acted as an upfront normalization method, ensuring that observed changes in diversity were not due to technical differences in the samples due to read length or sequencing depth. In this regard, it was successful, however, at the cost of reduced yield. Indeed, in my analysis, the yield of TCRs was very low for TCGA tumours, with approximately 1 in 10,000,000 reads being annotated as a TCR CDR3. In an analysis where the goal is

discovery of TCR CDR3 sequences, this upfront normalization would not be the optimal strategy. For the TCGA PanCancer Atlas Immune Landscape of Cancer project [149], I contributed my TCR extraction pipeline without this upfront normalization. In the entire TCGA dataset, the findings were comparable to my results presented in chapter 3. Additionally, it was observed that TCR diversity varied by immune subtype, irrespective of tumour type. This larger dataset yielded greater statistical power, with nearly 3,000 TCR alpha-beta pairs and 400 TCR-pMHC pairs observed to co-occur at a statistically significant level in at least two subjects. The most highly reoccurring of these could be targets for follow up *in vivo* studies to determine if these represent true interactions.

Despite low yields in the TCGA data, I was able to identify shared TCR sequences found in tumours from multiple individuals. Given the enormous potential diversity of TCRs, identifying shared, abundant TCRs in multiple individuals suggests a common antigen is present in these individuals. In my attempts to determine what the shared antigen in these cases may be, I discovered that there was not a central repository for TCRs and their antigens, nor were there standardized ways of report TCR sequences in the literature. Recently there has been a push from the community to standardize adaptive immune receptor reporting [317], and databases have begun to emerge which aim to collect all known TCR-pMHC interactions [318,319]. These databases will offer enormous potential, both for identifying the antigen targets of T cells that have been previously observed, and in generating training datasets for TCR-pMHC interaction prediction algorithms. Ultimately, these types of algorithms may be able to predict targets of TCRs *de novo*. In addition to these interaction databases, TCR structure databases are being developed which will provide more useful information in the development of new prediction algorithms [320].

With this goal, two methods were recently published which attempt to identify recognition motifs in TCRs that recognize a shared antigen [321,322]. These approaches rely on co-occurrence of TCR and pMHC, conceptually similar to the approach I took in chapter 3, however they begin with enriched T cell populations resulting from either vaccination or pMHC-tetramer sorting. By analyzing large sets of TCR sequences that all share a common antigen, they were able to identify conserved motifs that direct antigen recognition and extend these results to predict if novel TCRs

would recognize the same antigens. A similar method was described which strives to identify TCRs that recognize public antigens among distinct repertoires, without the need for artificially enriching the dataset for reactive T cells by vaccination or tetramer-sorting [323]. As the TCR-pMHC databases mentioned above acquire more data, these methods can be further developed to learn more about the specific motifs that are important for antigen recognition [324].

Subsequent to my work presented in chapter 3, the creators of MiTCR released an updated version of their tool called MiXCR [325] which is able to identify both TCRs and B cell receptors (BCRs). Further, an update to MiXCR introduced a parameter to allow it to extract TCR information directly from RNA-seq data [326]. This updated tool can assemble multiple non-ambiguous partial CDR3 reads into longer contigs, improving sensitivity. Others in the field have also seen the benefit of extracting TCR information from RNA-seq datasets, and more tools have been created to perform the extraction and annotations [327,328]. Additionally, with single cell genomics increasing in prevalence, tools designed specifically to extract TCR information from RNA-seq of single cells have been developed [329,330].

In the context of RNA-seq from PTCL samples, the utility of direct TCR extraction is clear. Aberrant samples that are expected to have a clonally expanded T cell show clear evidence of clonal expansion, typically with one CDR3 alpha and beta having far greater abundance than other T cells in the sample. In this way, the paired alpha-beta chains of the TCR can be inferred. Subsequent to the analysis presented in chapter 3, Gong and colleagues published their work which replicates my findings [331], using the initial release of MiXCR followed by filtering of false-positives. It is unclear if their RNA-seq data is derived from sorted T cell populations or bulk lymphocytes, so it is difficult to compare the sensitivity between the two methods. Based on my findings directly comparing MiTCR to MiXCR, the sensitivity is comparable. The ability of MiXCR to annotate BCRs should allow for the natural extension of this analysis to identify the malignant clones in B cell malignancies.

5.3 Utility of the predicted self-immunopeptidome

The human immunopeptidome dataset generated in chapter 4 is the result of exhaustive peptide-MHC binding predictions for all unique peptides in the human proteome to all class I MHC molecules. Class I was selected for this initial proof of concept for a number of reasons: class I presented peptides are more restricted in length (8-11mers) compared to class II (13-25mers with a 9aa core [332,333]), the class I MHC binding groove is encoded by a single gene whereas class II MHC binding grooves are the product of an alpha and beta subunit coming together in a combinatorial manner, and peptide binding predictions for class I are much more accurate than for class II [334]. Given recent work demonstrating the importance of class II epitopes in cancer immunology [114,228,229,273], and the results I present in chapter 4, it would be of value to predict the class II human immunopeptidome using analogous methods as improvements to the prediction algorithms are made.

The class I human immunopeptidome dataset, while useful in ranking MHC molecules according to their ability to present a repertoire of human peptides, does not necessarily accurately represent the set of self-peptides presented by each MHC molecule. The fundamental basis for this dataset is the predicted binding affinity as measured by IC_{50} by NetMHCpan v3.0. I demonstrated that classifying peptide-MHC combinations as binders using this metric correlated best with the limited observed MHC-bound peptide data. While this metric performed better than a percentile rank threshold, it still did not explain all the variability in observed self-immunopeptidome size. For my purposes, I am less concerned with the exact set of peptides predicted to be presented, and more concerned with ordering the MHC molecules by how diverse a set of peptides they can present. The calculation of self-immunopeptidome size takes the overlapping set of peptides presented by all the MHC in the individual, so peptide identity does play a role, however it is likely that any bias due to errors in pMHC predictions will be systematic and thus have a minimal effect on self-immunopeptidome size prediction. NetMHCpan uses the sequence of the MHC binding pocket to make predictions, so similar MHC molecules should have similar repertoires of peptides predicted to be bound, and the overlap will not be significantly affected by a small number of erroneous predictions.

Another factor to consider is that binding predictions by NetMHCpan are most accurate for MHC that have the most training data available [247]. I checked for any effect from this by comparing HLA frequency in the NetMHCpan v3.0 training data to the number of presented peptides and saw a weak positive correlation. This likely represents a source of noise in the dataset, however, it is also feasible that some HLAs have less data in the training set due to decreased numbers of peptides being presented.

HLA alleles have different population frequencies in different ethnic groups. As such, the self-immunopeptidome size for different ethnic populations can vary due to biases in HLA allele usage. I controlled for this in chapter 4 by filtering out the non-Caucasian subjects from the datasets when comparing the distributions of self-immunopeptidome sizes, leaving the majority of the data for analysis. Ethnicity information in these samples is self-reported and is therefore subject to errors. Principal component analysis can be applied to SNP data to reveal genetic ancestry groups, which would provide a genetically determined ethnicity [335], removing the reliance on self-reported data. This technique would be possible for the TCGA data [149], but no SNP information was available for the NMDP data.

In my work investigating immunoediting and immunosurveillance in tumours, there are two key assumptions. Firstly, the set of mutations found in the TCGA data were not strongly immunogenic and have been depleted of immunodominant epitopes. The rationale for this assumption is that the cancer cells have been able to co-exist with the host immune system and develop into a clinically relevant tumour. Based on the theory of immunoediting [105], cells containing immunogenic mutations would be deleted by the immune system. My analysis supports this, showing decreased immunogenicity of expressed mutations compared to non-expressed mutations for over 80 % of the samples. However, for the remaining samples, some immunogenic mutations likely remain due to immune evasion. This may take the form of downregulation of the antigen-presentation machinery by the tumour [336], inhibition of the T cells [337], or other immune suppression by the tumour microenvironment [338]. Despite this, in general the majority of mutations that are able to co-exist with a host immune system are not expected to be immunogenic, and “holes” in variant position

usage for presented peptides (when compared to a set of mutations that have not experienced immune pressure) can be informative.

Secondly, the set simulated mutations generated *in silico* represents mutations that have not experienced immune pressure. The mutations were generated at random positions throughout the proteome, but followed the same amino acid transition frequencies that were observed in the TCGA data. This is a very simplistic model for generating amino acid changes, and the resulting mutations may or may not affect the biological activity of the source protein. Because of this, these mutations do not necessarily accurately reflect the natural set of cancer mutations that would be acquired in a tumour free from immune pressure. They do, however, provide a baseline immunogenicity measure for random amino acid changes. Mutation datasets from tumours allowed to grow without immune pressure, for example in a RAG-1 deficient mouse, would be useful as comparators, though the mutations in such a dataset may contain biases due to development in a non-human host.

5.4 Future directions

The ultimate goal of immunoinformatics applied to cancer immunology is to accurately predict the biological activity of the immune system given tumour genomic data. Relevant to this thesis, the main areas of prediction that have the greatest opportunities for improvement are neoantigen immunogenicity and response to immunotherapy.

Currently, neoantigen predictions depend on the contextual placement of somatic mutations within flanking reference protein sequence, without regard for germline variation or proximal mutations. They also typically focus only on SNVs. Above, I described a recent tool which uses RNA-seq reads to generate the mutant peptides, which is a step in the right direction, however, I believe this can be further improved. Rather than calling somatic mutations and fishing for RNA-seq reads that support these mutations, tumour RNA-seq reads could be directly aligned (or assembled and aligned) to the matched normal exome or genome data, and all cases of sequence that does not match the normal genome could be identified. These sequences would represent expressed non-self sequence that are the result of mutations within the tumour, not limited to SNVs. This would result in a mutation-type-agnostic view of the mutant

peptide repertoire, outputting all peptides that would not be expected to be present in that individual's healthy cells. This approach would require guidance by the reference transcriptome to infer the reading frame of the sequences, so would not be a completely *de novo* approach, but would better reflect the peptides present in the tumour.

Once we are able to accurately capture the repertoire of mutant peptides present in a tumour, we need to better predict the immunogenicity and immunodominance of these peptides. Immunogenicity predictions are mainly based on peptide-MHC binding and amino acid composition, but these still suffer from poor sensitivity and specificity when the predictions are validated. Immunogenicity trends can be found in existing data, but purpose-built datasets will be more informative. Genomics from tumours paired with MS on the MHC-eluted peptides would greatly improve the ability to determine the correlation between predictable features and absence or presence of the peptide on the surface of the cell. These types of datasets would also be informative for improving immunogenicity predictions for class II MHC-presented peptides, providing orthogonal information to peptide-MHC binding data. Following these studies up with new methods that can determine the targets of T cells [339,340], it will be possible to then determine which of these presented peptides are immunogenic and immunodominant. As TCR-pMHC databases continue to grow, new insights on the interaction network between TCR and pMHC will aid in predicting which features can influence recognition by the TCR.

While improved immunogenicity predictions will result in a highly detailed view of the neoantigens displayed on a tumour and will aid in the development of cancer vaccines or target selection for engineered T cell therapies, predicting response to checkpoint blockade immunotherapies will likely benefit from integration of a broader range of data. Predictions for immunotherapy success will likely be derived from comprehensive genomic profiling and deep learning, taking into consideration metrics such as the self-immunopeptidome size, mutational load, characteristics of the neoantigens present, TIL stratification, gene expression, tumour heterogeneity, and TCR diversity. Historically this type of comprehensive data has been limited by the resources available, but as more all-inclusive data is being collected and generated as part of new clinical cancer genomics programs [239,341,342], the clinical utility of these

datasets will grow. Additionally, as the number of studies tracking response to immunotherapy increases, the combined results will be able to train future algorithms to enable accurate predictions on the optimal immunotherapy strategy for individual cases of cancer.

Looking farther into the future, it seems plausible that nearly perfect predictions of TCR-pMHC interactions will become possible. After characterization of the neoantigen and TCR repertoire present in a tumour, this, along with the advancements in predicting response to immunotherapies suggested above, would allow for the optimal therapy to be selected for that individual. If one or more TCRs able to respond to clonal neoantigens present on the tumour are found, and these TCRs appear to be inhibited, some combination of checkpoint blockade may be desired. If no appropriate TCRs are found, then an engineering approach would be possible, creating an engineered T cell with a TCR capable of recognizing the appropriate antigen [343]. This would also allow for a combination approach if there is no appropriate clonal neoantigen, designing a set of T cells to be able to effectively target sub-clonal neoantigens and attack all the cancer clones present.

Bibliography

[1] Dembić Z, Haas W, Weiss S, Mccubrey J, Kiefer H, Von Boehmer H, Steinmetz M. Transfer of specificity by murine α and β T-cell receptor genes. *Nature*. 1986;320(6059):232–238. doi:10.1038/320232a0

[2] Saito T, Germain RN. Predictable acquisition of a new MHC recognition specificity following expression of a transfected T-cell receptor β-chain gene. *Nature*. 1987;329(6136):256–259. doi:10.1038/329256a0

[3] van der Merwe PA, Dushek O. Mechanisms for T cell receptor triggering. *Nat. Rev. Immunol.* 2011;11(1):47–55. doi:10.1038/nri2887

[4] Darmon AJ, Nicholson DW, Bleackley RC. Activation of the apoptotic protease CPP32 by cytotoxic T-cell-derived granzyme B. *Nature*. 1995;377(6548):446–448. doi:10.1038/377446a0

[5] Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* 2011;21(5):790–7. doi:10.1101/gr.115428.110

[6] Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature*. 1988;334(6181):395–402. doi:10.1038/334395a0

[7] van Gent DC, Ramsden DA, Gellert M. The RAG1 and RAG2 proteins establish the 12/23 rule in V(D)J recombination. *Cell*. 1996;85(1):107–13. doi:10.1016/S0092-8674(00)81086-7

[8] Gauss GH, Lieber MR. Mechanistic constraints on diversity in human V(D)J recombination. *Mol. Cell. Biol.* 1996;16(1):258–69.

[9] Lu H, Schwarz K, Lieber MR. Extent to which hairpin opening by the Artemis:DNA-PKcs complex can contribute to junctional diversity in V(D)J recombination. *Nucleic Acids Res.* 2007;35(20):6917–23. doi:10.1093/nar/gkm823

[10] Woodsworth DJ, Castellarin M, Holt RA. Sequence analysis of T-cell repertoires in health and disease. *Genome Med.* 2013;5(10):98. doi:10.1186/gm502

[11] Freeman JD, Warren LR, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* 2009;19(10):1817–24. doi:10.1101/gr.092924.109

[12] Sewell AK. Why must T cells be cross-reactive? *Nat. Rev. Immunol.* 2012;12(9):669–677. doi:10.1038/nri3279
[13] Khor B, Sleckman BP. Allelic exclusion at the TCRβ locus. *Curr. Opin. Immunol.* 2002;14(2):230–234. doi:10.1016/S0952-7915(02)00326-6
[14] Balomenos D, Balderas RS, Mulvany KP, Kaye J, Kono DH, Theofilopoulos AN. Incomplete T cell receptor V beta allelic exclusion and dual V beta-expressing cells. *J. Immunol.* 1995;155(7):3308–12.
[15] Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, Kirsch I, Vignali M, Rieder MJ, Carlson CS, et al. High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.* 2015;7(301):301ra131-301ra131. doi:10.1126/scitranslmed.aac5624
[16] Afik S, Yates KB, Bi K, Darko S, Godec J, Gerdemann U, Swadling L, Douek DC, Klenerman P, Barnes EJ, et al. Targeted reconstruction of T cell receptor sequence from single cell RNA-seq links CDR3 length to T cell differentiation state. *Nucleic Acids Res.* 2017;45(16):e148–e148. doi:10.1093/nar/gkx615
[17] Townsend A, Öhlén C, Bastin J, Ljunggren H-G, Foster L, Kärre K. Association of class I major histocompatibility heavy and light chains induced by viral peptides. *Nature*. 1989;340(6233):443–448. doi:10.1038/340443a0
[18] Schubert U, Antón LC, Gibbs J, Norbury CC, Yewdell JW, Bennink JR. Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature*. 2000;404(6779):770–774. doi:10.1038/35008096
[19] Cascio P, Hilton C, Kisselev AF, Rock KL, Goldberg AL. 26S proteasomes and immunoproteasomes produce mainly N-extended versions of an antigenic peptide. *EMBO J.* 2001;20(10):2357–66. doi:10.1093/emboj/20.10.2357
[20] Grandea AG, Androlewicz MJ, Athwal RS, Geraghty DE, Spies T. Dependence of peptide binding by MHC class I molecules on their interaction with TAP. *Science*. 1995;270(5233):105–8. doi:10.1126/SCIENCE.270.5233.105
[21] Serwold T, Gonzalez F, Kim J, Jacob R, Shastri N. ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature*. 2002;419(6906):480–483. doi:10.1038/nature01074
[22] Laumont CM, Daouda T, Laverdure J-P, Bonneil É, Caron-Lizotte O, Hardy M-P,

Granados DP, Durette C, Lemieux S, Thibault P, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* 2016;7:10238. doi:10.1038/ncomms10238
[23] Vigneron N, Stroobant V, Chapiro J, Ooms A, Degiovanni G, Morel S, van der Bruggen P, Boon T, Van den Eynde BJ. An antigenic peptide produced by peptide splicing in the proteasome. *Science*. 2004;304(5670):587–90. doi:10.1126/science.1095522
[24] Hanada K, Yewdell JW, Yang JC. Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature*. 2004;427(6971):252–256. doi:10.1038/nature02240
[25] Dalet A, Vigneron N, Stroobant V, Hanada K-I, Van den Eynde BJ. Splicing of distant peptide fragments occurs in the proteasome by transpeptidation and produces the spliced antigenic peptide derived from fibroblast growth factor-5. *J. Immunol.* 2010;184(6):3016–24. doi:10.4049/jimmunol.0901277
[26] Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA database. *Nucleic Acids Res.* 2012;41(D1):D1222–D1227. doi:10.1093/nar/gks949
[27] Murphy K. Janeway's Immunobiology. 8th ed. New York, New York, USA: Garland Science; 2014.
[28] van Deutekom HWM, Keşmir C. Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most? *Immunogenetics*. 2015;67(8):425–36. doi:10.1007/s00251-015-0849-y
[29] Trolle T, McMurtrey CP, Sidney J, Bardet W, Osborn SC, Kaever T, Sette A, Hildebrand WH, Nielsen M, Peters B. The Length Distribution of Class I-Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference. *J. Immunol.* 2016;196. doi:10.4049/jimmunol.1501721
[30] Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B. The immune epitope database 2.0. *Nucleic Acids Res.* 2010;38(Database issue):D854-62. doi:10.1093/nar/gkp1004
[31] Johansen TE, McCullough K, Catipovic B, Su XM, Amzel M, Schneck JP. Peptide binding to MHC class I is determined by individual pockets in the binding groove.

Scand. J. Immunol. 1997;46(2):137–46.
[32] Guo H-CC, Jardetzky TS, Garrettt TPJJ, Lane WS, Strominger JL, Wiley DC. Different length peptides bind to HLA-Aw68 similarly at their ends but bulge out in the middle. *Nature*. 1992;360(6402):364–366. doi:10.1038/360364a0
[33] Zinkernagel RM, Doherty PC. MHC-Restricted Cytotoxic T Cells: Studies on the Biological Role of Polymorphic Major Transplantation Antigens Determining T-Cell Restriction-Specificity, Function, and Responsiveness. *Adv. Immunol.* 1979;27:51–177. doi:10.1016/S0065-2776(08)60262-X
[34] Gotch F, McMichael A, Smith G, Moss B. Identification of viral molecules recognized by influenza-specific human cytotoxic T lymphocytes. *J. Exp. Med.* 1987;165(2):408–16. doi:10.1084/JEM.165.2.408
[35] Townsend ARM, Rothbard J, Gotch FM, Bahadur G, Wraith D, McMichael AJ. The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides. *Cell*. 1986;44(6):959–968. doi:10.1016/0092-8674(86)90019-X
[36] Betts MR, Casazza JP, Patterson BA, Waldrop S, Trigona W, Fu TM, Kern F, Picker LJ, Koup RA. Putative immunodominant human immunodeficiency virus-specific CD8(+) T-cell responses cannot be predicted by major histocompatibility complex class I haplotype. *J. Virol.* 2000;74(19):9144–51.
[37] Akram A, Inman RD. Immunodominance: A pivotal principle in host response to viral infections. *Clin. Immunol.* 2012;143(2):99–115. doi:10.1016/j.clim.2012.01.015
[38] Bodewes R, Kreijtz JHCM, Hillaire MLB, Geelhoed-Mieras MM, Fouchier RAM, Osterhaus ADME, Rimmelzwaan GF. Vaccination with whole inactivated virus vaccine affects the induction of heterosubtypic immunity against influenza virus A/H5N1 and immunodominance of virus-specific CD8+ T-cell responses in mice. *J. Gen. Virol.* 2010;91(7):1743–1753. doi:10.1099/vir.0.020784-0
[39] La Gruta NL, Rothwell WT, Cukalac T, Swan NG, Valkenburg SA, Kedzierska K, Thomas PG, Doherty PC, Turner SJ. Primary CTL response magnitude in mice is determined by the extent of naive T cell recruitment and subsequent clonal expansion. *J. Clin. Invest.* 2010;120(6):1885–94. doi:10.1172/JCI41538
[40] Meijers R, Lai C-C, Yang Y, Liu J, Zhong W, Wang J, Reinherz EL. Crystal

Structures of Murine MHC Class I H-2 Db and Kb Molecules in Complex with CTL Epitopes from Influenza A Virus: Implications for TCR Repertoire Selection and Immunodominance. *J. Mol. Biol.* 2005;345(5):1099–1110. doi:10.1016/J.JMB.2004.11.023
[41] Silverstein AM. Autoimmunity versus horror autotoxicus: The struggle for recognition. *Nat. Immunol.* 2001;2(4):279–281. doi:10.1038/86280
[42] Derbinski J, Gäbler J, Brors B, Tierling S, Jonnakuty S, Hergenhahn M, Peltonen L, Walter J, Kyewski B. Promiscuous gene expression in thymic epithelial cells is regulated at multiple levels. *J. Exp. Med.* 2005;202(1):33–45. doi:10.1084/jem.20050471
[43] Heino M, Peterson P, Kudoh J, Nagamine K, Lagerstedt A, Ovod V, Ranki A, Rantala I, Nieminen M, Tuukkanen J, et al. Autoimmune Regulator Is Expressed in the Cells Regulating Immune Tolerance in Thymus Medulla. *Biochem. Biophys. Res. Commun.* 1999;257(3):821–825. doi:10.1006/bbrc.1999.0308
[44] Liblau RS, Wong FS, Mars LT, Santamaria P. Autoreactive CD8 T Cells in Organ-Specific Autoimmunity. *Immunity*. 2002;17(1):1–6. doi:10.1016/S1074-7613(02)00338-2
[45] Yu W, Jiang N, Ebert PJR, Kidd BA, Müller S, Lund PJ, Juang J, Adachi K, Tse T, Birnbaum ME, et al. Clonal Deletion Prunes but Does Not Eliminate Self-Specific αβ CD8(+) T Lymphocytes. *Immunity*. 2015;42(5):929–41. doi:10.1016/j.immuni.2015.05.001
[46] Zinkernagel RM, Doherty PC. Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature*. 1974;248(5450):701–702. doi:10.1038/248701a0
[47] Miller MJ, Wei SH, Cahalan MD, Parker I. Autonomous T cell trafficking examined in vivo with intravital two-photon microscopy. *Proc. Natl. Acad. Sci. U. S. A.* 2003;100(5):2604–9. doi:10.1073/pnas.2628040100
[48] Itano AA, Jenkins MK. Antigen presentation to naive CD4 T cells in the lymph node. *Nat. Immunol.* 2003;4(8):733–739. doi:10.1038/ni957
[49] Bousso P, Robey E. Dynamics of CD8+ T cell priming by dendritic cells in intact lymph nodes. *Nat. Immunol.* 2003;4(6):579–585. doi:10.1038/ni928
[50] Albert ML, Sauter B, Bhardwaj N. Dendritic cells acquire antigen from apoptotic cells and induce class I-restricted CTLs. *Nature*. 1998;392(6671):86–89.

doi:10.1038/32183
[51] Bell D, Chomarat P, Broyles D, Netto G, Harb GM, Lebecque S, Valladeau J, Davoust J, Palucka KA, Banchereau J. In breast carcinoma tissue, immature dendritic cells reside within the tumor, whereas mature dendritic cells are located in peritumoral areas. *J. Exp. Med.* 1999;190(10):1417–26.
[52] Bevan MJ. Cross-priming for a secondary cytotoxic response to minor H antigens with H-2 congenic cells which do not cross-react in the cytotoxic assay. *J. Exp. Med.* 1976;143(5):1283–8. doi:10.1084/JEM.143.5.1283
[53] Huang AY, Golumbek P, Ahmadzadeh M, Jaffee E, Pardoll D, Levitsky H. Role of bone marrow-derived cells in presenting MHC class I-restricted tumor antigens. *Science*. 1994;264(5161):961–5. doi:10.1126/SCIENCE.7513904
[54] Ackerman AL, Kyritsis C, Tampé R, Cresswell P. Early phagosomes in dendritic cells form a cellular compartment sufficient for cross presentation of exogenous antigens. *Proc. Natl. Acad. Sci. U. S. A.* 2003;100(22):12889–94. doi:10.1073/pnas.1735556100
[55] Burnet FM. A Modification of Jerne's Theory of Antibody Production using the Concept of Clonal Selection. *Aust. J. Sci.* 1957;20(3):67–9.
[56] Murali-Krishna K, Altman JD, Suresh M, Sourdive DJD, Zajac AJ, Miller JD, Slansky J, Ahmed R. Counting antigen-specific CD8 T cells: A reevaluation of bystander activation during viral infection. *Immunity*. 1998;8(2):177–187. doi:10.1016/S1074-7613(00)80470-7
[57] Bretscher P, Cohn M. A theory of self-nonself discrimination. *Science*. 1970;169(3950):1042–9. doi:10.1126/SCIENCE.169.3950.1042
[58] Steinman RM, Nussenzweig MC. Avoiding horror autotoxicus: the importance of dendritic cells in peripheral T cell tolerance. *Proc. Natl. Acad. Sci. U. S. A.* 2002;99(1):351–8. doi:10.1073/pnas.231606698
[59] Jerne NK. The Natural-Selection Theory of Antibody Formation. *Proc. Natl. Acad. Sci.* 1955;41(11):849–857. doi:10.1073/pnas.41.11.849
[60] Jerne NK. The somatic generation of immune recognition. *Eur. J. Immunol.* 1971;1(1):1–9. doi:10.1002/eji.200425132
[61] Mason D. A very high level of crossreactivity is an essential feature of the T- cell

receptor. *Immunol. Today*. 1998;19(9):395–404. doi:10.1016/S0167-5699(98)01299-7
[62] Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alphabeta T cell receptor diversity. *Science*. 1999;286(5441):958–61. doi:10.1126/SCIENCE.286.5441.958
[63] Ignatowicz L, Kappler J, Marrack P. The repertoire of T cells shaped by a single MHC/peptide ligand. *Cell*. 1996;84(4):521–529. doi:10.1016/S0092-8674(00)81028-4
[64] Ignatowicz L, Rees W, Pacholczyk R, Ignatowicz H, Kushnir E, Kappler J, Marrack P. T cells can be activated by peptides that are unrelated in sequence to their selecting peptide. *Immunity*. 1997;7(2):179–186. doi:10.1016/S1074-7613(00)80521-X
[65] Hiemstra HS, van Veelen PA, Willemen SJM, Benckhuijsen WE, Geluk A, de Vries RRP, Roep BO, Drijfhout JW. Quantitative determination of TCR cross-reactivity using peptide libraries and protein databases. *Eur. J. Immunol.* 1999;29(8):2385–2391. doi:10.1002/(SICI)1521-4141(199908)29:08<2385::AID-IMMU2385>3.0.CO;2-B
[66] Wooldridge L, Ekeruche-Makinde J, van den Berg HA, Skowera A, Miles JJ, Tan MP, Dolton G, Clement M, Llewellyn-Lacey S, Price DA, et al. A single autoimmune T cell receptor recognizes more than a million different peptides. *J. Biol. Chem.* 2012;287(2):1168–77. doi:10.1074/jbc.M111.289488
[67] Robins HS, Campregher P V, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS. Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood*. 2009;114(19):4099–107. doi:10.1182/blood-2009-04-217604
[68] Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, Davenport MP. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc. Natl. Acad. Sci. U. S. A.* 2006;103(49):18691–6. doi:10.1073/pnas.0608907103
[69] Cole DJ, Weil DP, Shamamian P, Rivoltini L, Kawakami Y, Topalian S, Jennings C, Eliyahu S, Rosenberg SA, Nishimura MI, et al. Identification of MART-1-specific T-Cell Receptors: T Cells Utilizing Distinct T-Cell Receptor Variable and Joining Regions Recognize the Same Tumor Epitope. *Cancer Res.* 1994;54(20):5265–5268.
[70] Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and

immunogenicity. *J. Immunol.* 2013;191(12):5831–9. doi:10.4049/jimmunol.1302101
[71] Rao X, Fontaine Costa AICA, van Baarle D, Kesmir C, Isabel A, Costa CAF, Baarle D Van, Kes C. A Comparative Study of HLA Binding Affinity and Ligand Diversity: Implications for Generating Immunodominant CD8+ T Cell Responses. *J. Immunol.* 2009;182(3):1526–1532. doi:10.4049/jimmunol.182.3.1526
[72] Nowak MA, Tarczy-Hornoch K, Austyn JM. The optimal number of major histocompatibility complex molecules in an individual. *Proc. Natl. Acad. Sci. U. S. A.* 1992;89(22):10896–9.
[73] Woelfing B, Traulsen A, Milinski M, Boehm T. Does intra-individual major histocompatibility complex diversity keep a golden mean? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 2009;364(1513):117–28. doi:10.1098/rstb.2008.0174
[74] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646–674.
[75] Ehrlich P. Collected studies on immunity. J. Wiley & sons; 1906.
[76] Bashford E, Murray JA, Halland M, Bowen WH. General results of propagation of malignant new growths. *Third Sci. Rep. Investig. Imp. Cancer Res. Fund*. 1908;3:262–283.
[77] Prehn RT, Main JM. Immunity to Methylcholanthrene-Induced Sarcomas. *JNCI J. Natl. Cancer Inst.* 1957;18(6):769–778. doi:10.1093/jnci/18.6.769
[78] Burnet FM. The Concept of Immunological Surveillance. Vol. 13. Karger Publishers; 1970. pp. 1–27.
[79] Shankaran V, Ikeda H, Bruce AT, White JM, Swanson PE, Old LJ, Schreiber RD. IFNγ and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature*. 2001;410(6832):1107–1111. doi:10.1038/35074122
[80] Koebel CM, Vermi W, Swann JB, Zerafa N, Rodig SJ, Old LJ, Smyth MJ, Schreiber RD. Adaptive immunity maintains occult cancer in an equilibrium state. *Nature*. 2007;450(7171):903–907. doi:10.1038/nature06309
[81] Penn I. Malignant melanoma in organ allograft recipients. *Transplantation*. 1996;61(2):274–8.
[82] Suranyi MG, Hogan PG, Falkk MC, Axelsen RA, Rigby R, Hawley C, Petrie J. Advanced donor origin melanoma in a renal transplant recipient: Immunotherapy, cure,

and retransplantation. *Transplantation*. 1998;66(5):655–661. doi:10.1097/00007890-199809150-00020
[83] MacKie RM, Reid R, Junor B. Fatal Melanoma Transferred in a Donated Kidney 16 Years after Melanoma Surgery. *N. Engl. J. Med.* 2003;348(6):567–568. doi:10.1056/NEJM200302063480620
[84] Birkeland SA, Storm HH, Lamm LU, Barlow L, Blohmé I, Forsberg B, Eklund B, Fjeldborg O, Friedberg M, Frödin L, et al. Cancer risk after renal transplantation in the nordic countries, 1964-1986. *Int. J. Cancer*. 1995;60(2):183–189. doi:10.1002/ijc.2910600209
[85] Silverberg MJ, Chao C, Leyden WA, Xu L, Tang B, Horberg MA, Klein D, Quesenberry CP, Towner WJ, Abrams DI, et al. HIV infection and the risk of cancers with and without a known infectious cause. *AIDS*. 2009;23(17):2337–45. doi:10.1097/QAD.0b013e3283319184
[86] Simard EP, Engels EA. Cancer as a Cause of Death among People with AIDS in the United States. *Clin. Infect. Dis.* 2010;51(8):957–962. doi:10.1086/656416
[87] Ebbell B. The papyrus Ebers. The greatest Egyptian medical document. *London Oxford Univ. Press*. 1937:135.
[88] Jessy T. Immunity over inability: The spontaneous regression of cancer. *J. Nat. Sci. Biol. Med.* 2011;2(1):43. doi:10.4103/0976-9668.82318
[89] Coley WB. the Treatment of Malignat Tumors By Repeated Inoculations of Erysipelas. *Am. J. Med. Sci.* 1893;105(5):487–510. doi:10.1097/00000441-189305000-00001
[90] Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*. 2018;173(2):371–385.e18. doi:10.1016/j.cell.2018.02.060
[91] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. *Science*. 2013;339(6127):1546–58. doi:10.1126/science.1235122
[92] McFarland CD, Yaglom JA, Wojtkowiak JW, Scott JG, Morse DL, Sherman MY, Mirny LA. The Damaging Effect of Passenger Mutations on Cancer Progression. *Cancer*

Res. 2017;77(18):4763–4772. doi:10.1158/0008-5472.CAN-15-3283-T
[93] Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan MD, et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* 2018;6(3):271–281.e7. doi:10.1016/j.cels.2018.03.002
[94] Kandoth C, Mclellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, Mcmichael JF, Wyczalkowski M a, et al. Mutational landscape and significance across 12 major cancer types. *Nature*. 2013;502(7471):333–9. doi:10.1038/nature12634
[95] Whiteside TL, Jost LM, Herberman RB. Tumor-infiltrating lymphocytes. Potential and limitations to their use for cancer therapy. *Crit. Rev. Oncol. Hematol.* 1992;12(1):25–47. doi:10.1016/1040-8428(92)90063-V
[96] Mosmann TR, Cherwinski H, Bond MW, Giedlin MA, Coffman RL. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *J. Immunol.* 1986;136(7):2348–57. doi:10.1111/j.1442-9071.2011.02672.x
[97] Curiel TJ, Coukos G, Zou L, Alvarez X, Cheng P, Mottram P, Evdemon-Hogan M, Conejo-Garcia JR, Zhang L, Burow M, et al. Specific recruitment of regulatory T cells in ovarian carcinoma fosters immune privilege and predicts reduced survival. *Nat. Med.* 2004;10(9):942–949. doi:10.1038/nm1093
[98] Oble DA, Loewe R, Yu P, Mihm MC. Focus on TILs: prognostic significance of tumor infiltrating lymphocytes in human melanoma. *Cancer Immun.* 2009;9:3.
[99] Hwang W-T, Adams SF, Tahirovic E, Hagemann IS, Coukos G. Prognostic significance of tumor-infiltrating T cells in ovarian cancer: a meta-analysis. *Gynecol. Oncol.* 2012;124(2):192–8. doi:10.1016/j.ygyno.2011.09.039
[100] Gooden MJM, de Bock GH, Leffers N, Daemen T, Nijman HW. The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis. *Br. J. Cancer*. 2011;105(1):93–103. doi:10.1038/bjc.2011.189
[101] Hackl H, Charoentong P, Finotello F, Trajanoski Z. Computational genomics tools for dissecting tumour-immune cell interactions. *Nat. Rev. Genet.* 2016;17(8):441–458. doi:10.1038/nrg.2016.67
[102] Chen YT, Scanlan MJ, Sahin U, Türeci O, Gure AO, Tsang S, Williamson B,

Stockert E, Pfreundschuh M, Old LJ. A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening. *Proc. Natl. Acad. Sci. U. S. A.* 1997;94(5):1914–8. doi:10.1073/PNAS.94.5.1914
[103] van der Bruggen P, Traversari C, Chomez P, Lurquin C, De Plaen E, Van den Eynde B, Knuth A, Boon T. A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science*. 1991;254(5038):1643–1647. doi:10.1126/science.1840703
[104] Nowell P. The clonal evolution of tumor cell populations. *Science*. 1976;194(4260):23–28. doi:10.1126/science.959840
[105] Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat. Immunol.* 2002;3(11):991–998. doi:10.1038/ni1102-991
[106] Chen DS, Mellman I. Oncology Meets Immunology: The Cancer-Immunity Cycle. *Immunity*. 2013;39(1):1–10. doi:10.1016/J.IMMUNI.2013.07.012
[107] Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 2012;366(10):883–92. doi:10.1056/NEJMoa1113205
[108] Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer*. 2012;12(4):252–64. doi:10.1038/nrc3239
[109] Hodi FS, Mihm MC, Soiffer RJ, Haluska FG, Butler M, Seiden M V, Davis T, Henry-Spires R, MacRae S, Willman A, et al. Biologic activity of cytotoxic T lymphocyte-associated antigen 4 antibody blockade in previously vaccinated metastatic melanoma and ovarian carcinoma patients. *Proc. Natl. Acad. Sci. U. S. A.* 2003;100(8):4712–4717.
[110] Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, Powderly JD, Carvajal RD, Sosman JA, Atkins MB, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* 2012;366(26):2443–54. doi:10.1056/NEJMoa1200690
[111] Dong H, Strome SE, Salomao DR, Tamura H, Hirano F, Flies DB, Roche PC, Lu J, Zhu G, Tamada K, et al. Tumor-associated B7-H1 promotes T-cell apoptosis: a potential mechanism of immune evasion. *Nat. Med.* 2002;8(8):793–800.

doi:10.1038/nm730
[112] Robbins PF, Lu Y-C, El-Gamil M, Li YF, Gross C, Gartner J, Lin JC, Teer JK, Cliften P, Tycksen E, et al. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* 2013;19(6):747–52. doi:10.1038/nm.3161
[113] Grupp SA, Kalos M, Barrett D, Aplenc R, Porter DL, Rheingold SR, Teachey DT, Chew A, Hauck B, Wright JF, et al. Chimeric antigen receptor-modified T cells for acute lymphoid leukemia. *N. Engl. J. Med.* 2013;368(16):1509–18. doi:10.1056/NEJMoa1215134
[114] Schumacher T, Bunse L, Pusch S, Sahm F, Wiestler B, Quandt J, Menn O, Osswald M, Oezen I, Ott M, et al. A vaccine targeting mutant IDH1 induces antitumour immunity. *Nature*. 2014;512:324–327. doi:10.1038/nature13387
[115] Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, Löwer M, Bukur V, Tadmor AD, Luxemburger U, Schrörs B, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*. 2017;547(7662):222–226. doi:10.1038/nature23003
[116] Agur Z, Vuk-Pavlović S. Personalizing immunotherapy: Balancing predictability and precision. *Oncoimmunology*. 2012;1(7):1169–1171. doi:10.4161/onci.20955
[117] Wayteck L, Breckpot K, Demeester J, De Smedt SC, Raemdonck K. A personalized view on cancer immunotherapy. *Cancer Lett.* 2013;352(1):113–125. doi:10.1016/j.canlet.2013.09.016
[118] Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, Sucker A, Hillen U, Geukes Foppen MH, Goldinger SM, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*. 2015;350(6257):207–211. doi:10.1126/science.aad0095
[119] Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, Walsh LA, Postow MA, Wong P, Ho TS, et al. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* 2014;371(23):141119140020009. doi:10.1056/NEJMoa1406498
[120] Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, et al. Mutational landscape determines sensitivity to PD-1

blockade in non – small cell lung cancer. *Science*. 2016;348(6230):124. doi:10.1126/science.aaa1348
[121] Dubey P, Hendrickson RC, Meredith SC, Siegel CT, Shabanowitz J, Skipper JC, Engelhard VH, Hunt DF, Schreiber H. The immunodominant antigen of an ultraviolet-induced regressor tumor is generated by a somatic point mutation in the DEAD box helicase p68. *J. Exp. Med.* 1997;185(4):695–705. doi:10.1084/JEM.185.4.695
[122] Castle JC, Kreiter S, Diekmann J, Löwer M, van de Roemer N, de Graaf J, Selmi A, Diken M, Boegel S, Paret C, et al. Exploiting the mutanome for tumor vaccination. *Cancer Res.* 2012;72(5):1081–91. doi:10.1158/0008-5472.CAN-11-3722
[123] Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, Zhang W, Luoma A, Giobbie-Hurder A, Peter L, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*. 2017;547(7662):217–221. doi:10.1038/nature22991
[124] Li L, Goedegebuure SP, Gillanders WE. Preclinical and clinical development of neoantigen vaccines. *Ann. Oncol.* 2017;28(suppl_12):xii11-xii17.
[125] Andersen RS, Thrue CA, Junker N, Lyngaa R, Donia M, Ellebæk E, Svane IM, Schumacher TN, Thor Straten P, Hadrup SR. Dissection of T-cell antigen specificity in human melanoma. *Cancer Res.* 2012;72(7):1642–50. doi:10.1158/0008-5472.CAN-11-2614
[126] Lennerz V, Fatho M, Gentilini C, Frye RA, Lifke A, Ferel D, Wölfel C, Huber C, Wölfel T. The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. *Proc. Natl. Acad. Sci. U. S. A.* 2005;102(44):16013–8. doi:10.1073/pnas.0500090102
[127] Matsushita H, Vesely MD, Koboldt DC, Rickert CG, Uppaluri R, Magrini VJ, Arthur CD, White JM, Chen Y-S, Shea LK, et al. Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature*. 2012;482(7385):400–4. doi:10.1038/nature10755
[128] Lu Y-C, Yao X, Li YF, El-Gamil M, Dudley ME, Yang JC, Almeida JR, Douek DC, Samuels Y, Rosenberg SA, et al. Mutated PPP1R3B is recognized by T cells used to treat a melanoma patient who experienced a durable complete tumor regression. *J. Immunol.* 2013;190(12):6034–42. doi:10.4049/jimmunol.1202830
[129] van Rooij N, van Buuren MM, Philips D, Velds A, Toebes M, Heemskerk B, van

Dijk LJA, Behjati S, Hilkmann H, El Atmioui D, et al. Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.* 2013;31(32):e439-42. doi:10.1200/JCO.2012.47.7521

[130] Wick DA, Webb JR, Nielsen JS, Martin SD, Kroeger DR, Milne K, Castellarin M, Twumasi-Boateng K, Watson PH, Holt RA, et al. Surveillance of the tumor mutanome by T cells during progression from primary to recurrent ovarian cancer. *Clin. Cancer Res.* 2014;20(5):1125–1134. doi:10.1158/1078-0432.CCR-13-2147

[131] van Buuren MM, Calis JJ, Schumacher TN, Directed TI. High sensitivity of cancer exome-based CD8 T cell neo-antigen identification. *Oncoimmunology*. 2014;3(May):e28836. doi:10.4161/onci.28836

[132] Stone JD, Kranz DM. Role of T cell receptor affinity in the efficacy and specificity of adoptive T cell therapies. *Front. Immunol.* 2013;4(August):244. doi:10.3389/fimmu.2013.00244

[133] Sivan A, Corrales L, Hubert N, Williams JB, Aquino-Michaels K, Earley ZM, Benyamin FW, Man Lei Y, Jabri B, Alegre M-L, et al. Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science*. 2015;350(6264):1084–1089. doi:10.1126/science.aac4255

[134] Matson V, Fessler J, Bao R, Chongsuwat T, Zha Y, Alegre M-L, Luke JJ, Gajewski TF. The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. *Science*. 2018;359(6371):104–108. doi:10.1126/science.aao3290

[135] Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews MC, Karpinets T V., Prieto PA, Vicente D, Hoffman K, Wei SC, et al. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science*. 2018;359(6371):97–103. doi:10.1126/science.aan4236

[136] Routy B, Le Chatelier E, Derosa L, Duong CPM, Alou MT, Daillère R, Fluckiger A, Messaoudene M, Rauber C, Roberti MP, et al. Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. *Science*. 2018;359(6371):91–97. doi:10.1126/science.aan3706

[137] Mizushima N, Kohsaka H, Nanki T, Ollier WE, Carson DA, Miyasaka N. HLA-dependent peripheral T cell receptor (TCR) repertoire formation and its modification by rheumatoid arthritis (RA). *Clin. Exp. Immunol.* 1997;110(3):428–33. doi:10.1046/J.1365-

2249.1997.4331451.X
[138] Ureta-Vidal A, Firat H, Pérarnau B, Lemonnier FA. Phenotypical and functional characterization of the CD8+ T cell repertoire of HLA-A2.1 transgenic, H-2KbnullDbnull double knockout mice. *J. Immunol.* 1999;163(5):2555–2560. doi:ji_v163n5p2555 [pii]
[139] Dyall R, Messaoudi I, Janetzki S, Nikolic-Zugić J. MHC polymorphism can enrich the T cell repertoire of the species by shifts in intrathymic selection. *J. Immunol.* 2000;164(4):1695–8. doi:10.4049/JIMMUNOL.164.4.1695
[140] Chowell D, Morris LGT, Grigg CM, Weber JK, Samstein RM, Makarov V, Kuo F, Kendall SM, Requena D, Riaz N, et al. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*. 2017 Dec 7:eaao4572. doi:10.1126/science.aao4572
[141] Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* 1994;152(1):163–75.
[142] Rammensee H-G, Bachmann J, Emmerich NPN, Bachor OA, Stevanović S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*. 1999;50(3–4):213–219. doi:10.1007/s002510050595
[143] Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, Holt RA. Derivation of HLA types from shotgun sequence datasets. *Genome Med.* 2012;4(12):95. doi:10.1186/gm396
[144] Boegel S, Löwer M, Schäfer M, Bukur T, Graaf J De, Boisguérin V, Türeci Ö, Diken M, Castle JC, Sahin U, et al. HLA typing from RNA-Seq sequence reads. *Genome Med.* 2012;4(12):102. doi:10.1186/gm403
[145] Warren RL, Holt RA. A census of predicted mutational epitopes suitable for immunologic cancer control. *Hum. Immunol.* 2010;71(3):245–54. doi:10.1016/j.humimm.2009.12.007
[146] Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, Schatz MM, Kloetzel P-M, Rammensee H-G, Schild H, Holzhütter H-G. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage,TAP transport and MHC class I binding. *C. Cell. Mol. Life Sci.* 2005;62(9):1025–1037. doi:10.1007/s00018-005-4528-2
[147] Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W,

Treviño V, Shen H, Laird PW, Levine DA, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 2013;4:2612. doi:10.1038/ncomms3612
[148] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*. 2015;12(5):453–457. doi:10.1038/nmeth.3337
[149] Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, et al. The Immune Landscape of Cancer. *Immunity*. 2018;48(4):812–830.e14. doi:10.1016/j.immuni.2018.03.023
[150] Bolotin DA, Shugay M, Mamedov IZ, Putintseva E V, Turchaninova M a, Zvyagin I V, Britanova O V, Chudakov DM. MiTCR: software for T-cell receptor sequencing data analysis. *Nat. Methods*. 2013;10(9):813–4. doi:10.1038/nmeth.2555
[151] Marsh SGE, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Fernández-Viña M, Geraghty DE, Holdsworth R, Hurley CK, et al. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*. 2010;75(4):291–455. doi:10.1111/j.1399-0039.2010.01466.x
[152] Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, Stevens J, Lane WJ, Dellagatta JL, Steelman S, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* 2015;33:1152–1158. doi:10.1038/nbt.3344
[153] Listgarten J, Brumme Z, Kadie C, Xiaojiang G, Walker B, Carrington M, Goulder P, Heckerman D. Statistical resolution of ambiguous HLA typing data. *PLoS Comput. Biol.* 2008;4(2):e1000016. doi:10.1371/journal.pcbi.1000016
[154] Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*. 2014;30(23):1–7. doi:10.1093/bioinformatics/btu548
[155] Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell*. 2015;160(1–2):48–61. doi:10.1016/j.cell.2014.12.033
[156] Brown SD, Warren RL, Gibb EA, Martin SD, Spinelli JJ, Nelson BH, Holt RA. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient

survival. *Genome Res.* 2014;24(5):743–750. doi:10.1101/gr.165985.113
[157] Gubin MM, Artyomov MN, Mardis ER, Schreiber RD. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J. Clin. Invest.* 2015;125(9):1–9. doi:10.1172/JCI80008
[158] Lundegaard C, Lund O, Nielsen M. Prediction of epitopes using neural network based methods. *J. Immunol. Methods*. 2011;374(1–2):26–34. doi:10.1016/j.jim.2010.10.011
[159] Zhang L, Udaka K, Mamitsuka H, Zhu S. Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. *Brief. Bioinform.* 2012;13(3):350–64. doi:10.1093/bib/bbr060
[160] Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, Melief CJ, Oseroff C, Yuan L, Ruppert J, et al. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *J. Immunol.* 1994;153(12):5586–92.
[161] 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 Med.* 2016;8(1):33. doi:10.1186/s13073-016-0288-x
[162] Pradhan D, Yadav M, Verma R, Khan NS, Jena L, Jain AK. Discovery of T-cell Driven Subunit Vaccines from Zika Virus Genome: An Immunoinformatics Approach. *Interdiscip. Sci. Comput. Life Sci.* 2017;9(4):468–477. doi:10.1007/s12539-017-0238-3
[163] dos Santos Franco L, Oliveira Vidal P, Amorim JH. In silico design of a Zika virus non-structural protein 5 aiming vaccine protection against zika and dengue in different human populations. *J. Biomed. Sci.* 2017;24(1):88. doi:10.1186/s12929-017-0395-z
[164] Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, Griffith M. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* 2016;8(1):11. doi:10.1186/s13073-016-0264-5
[165] Kass I, Buckle AM, Borg NA. Understanding the structural dynamics of TCR-pMHC interactions. *Trends Immunol.* 2014;35(12):604–612. doi:10.1016/j.it.2014.10.005
[166] Balachandran VP, Luksza M, Zhao JN, Makarov V, Moral JA, Remark R, Herbst B, Askan G, Bhanot U, Senbabaoglu Y, et al. Identification of unique neoantigen

qualities in long-term survivors of pancreatic cancer. *Nature*. 2017;551(7681):S12–S16. doi:10.1038/nature24462
[167] Kessels HWHG, de Visser KE, Tirion FH, Coccoris M, Kruisbeek AM, Schumacher TNM. The Impact of Self-Tolerance on the Polyclonal CD8+ T Cell Repertoire. *J. Immunol.* 2004;172(4):2324–2331. doi:10.4049/jimmunol.172.4.2324
[168] Chowell D, Krishna S, Becker PD, Cocita C, Shu J, Tan X, Greenberg PD, Klavinskis LS, Blattman JN, Anderson KS. TCR contact residue hydrophobicity is a hallmark of immunogenic CD8 $^{+}$ T cell epitopes. *Proc. Natl. Acad. Sci.* 2015;112(14):E1754–E1762. doi:10.1073/pnas.1500973112
[169] Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*. 2018;173(2):283–285.
[170] Fox BA, Schendel DJ, Butterfield LH, Aamdal S, Allison JP, Ascierto PA, Atkins MB, Bartunkova J, Bergmann L, Berinstein N, et al. Defining the critical hurdles in cancer immunotherapy. *J. Transl. Med.* 2011;9(1):214. doi:10.1186/1479-5876-9-214
[171] Kalos M, Levine BL, Porter DL, Katz S, Grupp SA, Bagg A, June CH. T cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia. *Sci. Transl. Med.* 2011;3(95):95ra73. doi:10.1126/scitranslmed.3002842
[172] Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* 2001;344(11):783–92. doi:10.1056/NEJM200103153441101
[173] Yang JC, Haworth L, Sherry RM, Hwu P, Schwartzentruber DJ, Topalian SL, Steinberg SM, Chen HX, Rosenberg SA. A randomized trial of bevacizumab, an anti-vascular endothelial growth factor antibody, for metastatic renal cancer. *N. Engl. J. Med.* 2003;349(5):427–34. doi:10.1056/NEJMoa021491
[174] Coiffier B, Lepage E, Briere J, Herbrecht R, Tilly H, Bouabdallah R, Morel P, Van Den Neste E, Salles G, Gaulard P, et al. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *N. Engl. J. Med.* 2002;346(4):235–42. doi:10.1056/NEJMoa011795
[175] Cunningham D, Humblet Y, Siena S, Khayat D, Bleiberg H, Santoro A, Bets D,

Mueser M, Harstrick A, Verslype C, et al. Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *N. Engl. J. Med.* 2004;351(4):337–45. doi:10.1056/NEJMoa033025
[176] Yamada N, Oizumi S, Kikuchi E, Shinagawa N, Konishi-Sakakibara J, Ishimine A, Aoe K, Gemba K, Kishimoto T, Torigoe T, et al. CD8+ tumor-infiltrating lymphocytes predict favorable prognosis in malignant pleural mesothelioma after resection. *Cancer Immunol. Immunother.* 2010;59(10):1543–9. doi:10.1007/s00262-010-0881-6
[177] Sato E, Olson SH, Ahn J, Bundy B, Nishikawa H, Qian F, Jungbluth AA, Frosina D, Gnjatic S, Ambrosone C, et al. Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. *Proc. Natl. Acad. Sci. U. S. A.* 2005;102(51):18538–43. doi:10.1073/pnas.0509182102
[178] Nelson BH. The impact of T-cell immunity on ovarian cancer outcomes. *Immunol. Rev.* 2008;222:101–16. doi:10.1111/j.1600-065X.2008.00614.x
[179] Kim R, Emi M, Tanabe K. Cancer immunoediting from immune surveillance to immune escape. *Immunology*. 2007;121(1):1–14. doi:10.1111/j.1365-2567.2007.02587.x
[180] Heemskerk B, Kvistborg P, Schumacher TNM. The cancer antigenome. *EMBO J.* 2013;32(2):194–203. doi:10.1038/emboj.2012.333
[181] Yewdell JW, Bennink JR. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu. Rev. Immunol.* 1999;17:51–88. doi:10.1146/annurev.immunol.17.1.51
[182] Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, Brunak S, Lund O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* 2003;12(5):1007–17. doi:10.1110/ps.0239403
[183] Bui H-H, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton K-A, Mothé BR, Chisari F V, Watkins DI, Sette A. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*. 2005;57(5):304–14. doi:10.1007/s00251-005-0798-y
[184] Peters B, Sette A. Generating quantitative models describing the sequence

specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*. 2005;6:132. doi:10.1186/1471-2105-6-132
[185] Segal NH, Parsons DW, Peggs KS, Velculescu V, Kinzler KW, Vogelstein B, Allison JP. Epitope landscape in breast and colorectal cancer. *Cancer Res.* 2008;68(3):889–92. doi:10.1158/0008-5472.CAN-07-3095
[186] Khalili JS, Hanson RW, Szallasi Z. In silico prediction of tumor antigens derived from functional missense mutations of the cancer gene census. *Oncoimmunology*. 2012;1(8):1281–1289. doi:10.4161/onci.21511
[187] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21. doi:10.1093/bioinformatics/bts635
[188] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 2010;28(5):511–5. doi:10.1038/nbt.1621
[189] R Core Team. R: A Language and Environment for Statistical Computing Team RDC, editor. *R Found. Stat. Comput.* 2013;1(2.11.1). (R Foundation for Statistical Computing).
[190] Warren RL, Freeman DJ, Pleasance S, Watson P, Moore RA, Cochrane K, Allen-Vercoe E, Holt RA. Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome*. 2013;1(1):16. doi:10.1186/2049-2618-1-16
[191] Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-7. 2012.
[192] Therneau T. A package for survival analysis in S. R package version 2.37-4. 2013.
[193] Krzywinski M, Birol I, Jones SJM, Marra MA. Hive plots--rational approach to visualizing networks. *Brief. Bioinform.* 2012;13(5):627–44. doi:10.1093/bib/bbr069
[194] Hoof I, Peters B, Sidney J, Pedersen LE, Sette a, Lund O, Buus S, Nielsen M. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*. 2009;61(1):1–13. doi:10.1007/s00251-008-0341-z
[195] Bijen CBM, Bantema-Joppe EJ, de Jong RA, Leffers N, Mourits MJE, Eggink HF, van der Zee AGJ, Hollema H, de Bock GH, Nijman HW. The prognostic role of classical

and nonclassical MHC class I expression in endometrial cancer. *Int. J. Cancer*. 2010;126(6):1417–27. doi:10.1002/ijc.24852
[196] Concha A, Cabrera T, Ruiz-Cabello F, Garrido F. Can the HLA phenotype be used as a prognostic factor in breast carcinomas? *Int. J. Cancer*. 1991;47(S6):146–154. doi:10.1002/ijc.2910470726
[197] Kitamura H, Honma I, Torigoe T, Asanuma H, Sato N, Tsukamoto T. Down-regulation of HLA class I antigen is an independent prognostic factor for clear cell renal cell carcinoma. *J. Urol.* 2007;177(4):1269–72; discussion 1272. doi:10.1016/j.juro.2006.11.082
[198] Han LY, Fletcher MS, Urbauer DL, Mueller P, Landen CN, Kamat AA, Lin YG, Merritt WM, Spannuth WA, Deavers MT, et al. HLA class I antigen processing machinery component expression and intratumoral T-Cell infiltrate as independent prognostic markers in ovarian carcinoma. *Clin. Cancer Res.* 2008;14(11):3372–9. doi:10.1158/1078-0432.CCR-07-4433
[199] Ogino T, Shigyo H, Ishii H, Katayama A, Miyokawa N, Harabuchi Y, Ferrone S. HLA class I antigen down-regulation in primary laryngeal squamous cell carcinoma lesions as a poor prognostic marker. *Cancer Res.* 2006;66(18):9281–9. doi:10.1158/0008-5472.CAN-06-0488
[200] Blank C, Mackensen A. Contribution of the PD-L1/PD-1 pathway to T-cell exhaustion: an update on implications for chronic infections and tumor evasion. *Cancer Immunol. Immunother.* 2007;56(5):739–45. doi:10.1007/s00262-006-0272-1
[201] Schneider H, Downey J, Smith A, Zinselmeyer BH, Rush C, Brewer JM, Wei B, Hogg N, Garside P, Rudd CE. Reversal of the TCR stop signal by CTLA-4. *Science*. 2006;313(5795):1972–5. doi:10.1126/science.1131078
[202] Hodi FS, Butler M, Oble D a, Seiden M V, Haluska FG, Kruse A, MacRae S, Nelson M, Canning C, Lowy I, et al. Immunologic and clinical effects of antibody blockade of cytotoxic T lymphocyte-associated antigen 4 in previously vaccinated cancer patients. *Proc. Natl. Acad. Sci. U. S. A.* 2008;105(8):3005–3010.
[203] Hamanishi J, Mandai M, Iwasaki M, Okazaki T, Tanaka Y, Yamaguchi K, Higuchi T, Yagi H, Takakura K, Minato N, et al. Programmed cell death 1 ligand 1 and tumor-infiltrating CD8+ T lymphocytes are prognostic factors of human ovarian cancer. *Proc.*

Natl. Acad. Sci. U. S. A. 2007;104(9):3360–3365.
[204] Mansh M. Ipilimumab and cancer immunotherapy: a new hope for advanced stage melanoma. *Yale J. Biol. Med.* 2011;84(4):381–9.
[205] Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, Gonzalez R, Robert C, Schadendorf D, Hassel JC, et al. Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* 2010;363(8):711–23. doi:10.1056/NEJMoa1003466
[206] Brahmer JR, Tykodi SS, Chow LQM, Hwu W-J, Topalian SL, Hwu P, Drake CG, Camacho LH, Kauh J, Odunsi K, et al. Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *N. Engl. J. Med.* 2012;366(26):2455–65. doi:10.1056/NEJMoa1200694
[207] Kroemer G, Galluzzi L, Kepp O, Zitvogel L. Immunogenic cell death in cancer therapy. *Annu. Rev. Immunol.* 2013;31:51–72. doi:10.1146/annurev-immunol-032712-100008
[208] Dolcetti R, Viel A, Doglioni C, Russo A, Guidoboni M, Capozzi E, Vecchiato N, Macrì E, Fornasarig M, Boiocchi M. High prevalence of activated intraepithelial cytotoxic T lymphocytes and increased neoplastic cell apoptosis in colorectal carcinomas with microsatellite instability. *Am. J. Pathol.* 1999;154(6):1805–13. doi:10.1016/S0002-9440(10)65436-3
[209] Vesely MD, Schreiber RD. Cancer immunoediting: antigens, mechanisms, and implications to cancer immunotherapy. *Ann. N. Y. Acad. Sci.* 2013;1284:1–5. doi:10.1111/nyas.12105
[210] Moschos SJ, Edington HD, Land SR, Rao UN, Jukic D, Shipe-Spotloe J, Kirkwood JM. Neoadjuvant treatment of regional stage IIIB melanoma with high-dose interferon alfa-2b induces objective tumor regression in association with modulation of tumor infiltrating host cellular immune responses. *J. Clin. Oncol.* 2006;24(19):3164–71. doi:10.1200/JCO.2005.05.2498
[211] Hamid O, Chasalow SD, Tsuchihashi Z, Alaparthy S, Galbraith S, Berman D. Association of baseline and on-study tumor biopsy markers with clinical activity in patients with advanced melanoma treated with ipilimumab. *J. Clin. Oncol.* 2009;27:A9008.

[212] Clemente CG, Mihm MC, Bufalino R, Zurrida S, Collini P, Cascinelli N. Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma. *Cancer*. 1996;77(7):1303–10. doi:10.1002/(SICI)1097-0142(19960401)77:7<1303::AID-CNCR12>3.0.CO;2-5
[213] Cantaert T, Brouard S, Thurlings RM, Pallier A, Salinas GF, Braud C, Klarenbeek PL, de Vries N, Zhang Y, Soulillou J-P, et al. Alterations of the synovial T cell repertoire in anti-citrullinated protein antibody-positive rheumatoid arthritis. *Arthritis Rheum.* 2009;60(7):1944–56. doi:10.1002/art.24635
[214] Jonasson L, Holm J, Skalli O, Bondjers G, Hansson GK. Regional accumulations of T cells, macrophages, and smooth muscle cells in the human atherosclerotic plaque. *Arterioscler. Thromb. Vasc. Biol.* 1986;6(2):131–138. doi:10.1161/01.ATV.6.2.131
[215] Traugott U, Reinherz E, Raine C. Multiple sclerosis: distribution of T cell subsets within active chronic lesions. *Science*. 1983;219(4582):308–310. doi:10.1126/science.6217550
[216] Matloubian M, Concepcion RJ, Ahmed R. CD4+ T cells are required to sustain CD8+ cytotoxic T-cell responses during chronic viral infection. *J. Virol.* 1994;68(12):8056–63.
[217] Moon JJ, Chu HH, Pepper M, McSorley SJ, Jameson SC, Kedl RM, Jenkins MK. Naive CD4(+) T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity*. 2007;27(2):203–13. doi:10.1016/j.immuni.2007.07.007
[218] Parrott DM V., Tait C, MacKenzie S, Mowat AM, Davies MDJ, Micklem HS. Analysis of the effector functions of different populations of mucosal lymphocytes. *Ann. N. Y. Acad. Sci.* 1983;409(1 The Secretory):307–320. doi:10.1111/j.1749-6632.1983.tb26879.x
[219] Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, Greisman HA, Sabath DE, Wood BL, Robins H. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci. Transl. Med.* 2012;4(134):134ra63. doi:10.1126/scitranslmed.3003656
[220] Lefranc M-P. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb. Protoc.* 2011;2011(6):595–603. doi:10.1101/pdb.top115

[221] Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57–74. doi:10.1038/nature11247
[222] Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, Sammeth M. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* 2012;40(20):10073–83. doi:10.1093/nar/gks666
[223] Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* 2010;38(18):e178. doi:10.1093/nar/gkq622
[224] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323. doi:10.1186/1471-2105-12-323
[225] Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*. 2001;17(3):282–3.
[226] Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*. 2002;18(1):77–82.
[227] Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.* 2012;22(2):299–306. doi:10.1101/gr.126516.111
[228] Kreiter S, Vormehr M, van de Roemer N, Diken M, Löwer M, Diekmann J, Boegel S, Schrörs B, Vascotto F, Castle JC, et al. Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature*. 2015;520:692–696. doi:10.1038/nature14426
[229] Tran E, Turcotte S, Gros A, Robbins PF, Lu Y-C, Dudley ME, Wunderlich JR, Somerville RP, Hogan K, Hinrichs CS, et al. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science*. 2014;344(6184):641–5. doi:10.1126/science.1251102
[230] Lossius A, Johansen JN, Vartdal F, Robins H, Jūratė Šaltytė B, Holmøy T, Olweus J. High-throughput sequencing of TCR repertoires in multiple sclerosis reveals intrathecal enrichment of EBV-reactive CD8+ T cells. *Eur. J. Immunol.*

2014;44(11):3439–52. doi:10.1002/eji.201444662
[231] Swerdlow S, Campo E, Harris N, Jaffe E, Pileri S, Stein H, Thiele J, Vardiman J. WHO classification of tumours of haematopoietic and lymphoid tissues. 4th Editio. IARC; 2008.
[232] Savage KJ, Chhanabhai M, Gascoyne RD, Connors JM. Characterization of peripheral T-cell lymphomas in a single North American institution by the WHO classification. *Ann. Oncol.* 2004;15(10):1467–75. doi:10.1093/annonc/mdh392
[233] Vose J, Armitage J, Weisenburger D. International peripheral T-cell and natural killer/T-cell lymphoma study: pathology findings and clinical outcomes. *J. Clin. Oncol.* 2008;26(25):4124–30. doi:10.1200/JCO.2008.16.4558
[234] Chen W, Kesler M V, Karandikar NJ, McKenna RW, Kroft SH. Flow cytometric features of angioimmunoblastic T-cell lymphoma. *Cytometry B. Clin. Cytom.* 2006;70(3):142–8. doi:10.1002/cyto.b.20107
[235] Langerak AW, Szczepański T, van der Burg M, Wolvers-Tettero IL, van Dongen JJ. Heteroduplex PCR analysis of rearranged T cell receptor genes for clonality assessment in suspect T cell proliferations. *Leukemia*. 1997;11(12):2192–9.
[236] van Dongen JJM, Langerak AW, Brüggemann M, Evans PAS, Hummel M, Lavender FL, Delabesse E, Davi F, Schuuring E, García-Sanz R, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia*. 2003;17(12):2257–317. doi:10.1038/sj.leu.2403202
[237] Kirsch IR, Watanabe R, O'Malley JT, Williamson DW, Scott L-LL, Elco CP, Teague JE, Gehad A, Lowry EL, Leboeuf NR, et al. TCR sequencing facilitates diagnosis and identifies mature T cells as the cell of origin in CTCL. *Sci. Transl. Med.* 2015;7(308):308ra158-308ra158. doi:10.1126/scitranslmed.aaa9122
[238] Jones SJM, Laskin J, Li YY, Griffith OL, An J, Bilenky M, Butterfield YS, Cezard T, Chuah E, Corbett R, et al. Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biol.* 2010;11(8):R82. doi:10.1186/gb-2010-11-8-r82
[239] Laskin J, Jones S, Aparicio S, Chia S, Ch'ng C, Deyell R, Eirew P, Fok A, Gelmon K, Ho C, et al. Lessons learned from the application of whole-genome analysis to the

treatment of patients with advanced cancers. *Mol. Case Stud.* 2015;1(1):a000570. doi:10.1101/mcs.a000570
[240] Brown SD, Raeburn LA, Holt RA. Profiling tissue-resident T cell repertoires by RNA sequencing. *Genome Med.* 2015;7(1):125. doi:10.1186/s13073-015-0248-x
[241] Shulin L, Wilkinson MF, Li S, Wilkinson MF. Nonsense surveillance in lymphocytes? Minireview. *Immunity*. 1998;8(2):135–141. doi:10.1016/S1074-7613(00)80466-5
[242] Shannon CE. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 1948;5(3):3. doi:10.1002/j.1538-7305.1948.tb01338.x
[243] Hausser J, Strimmer K. entropy: Estimation of Entropy, Mutual Information and Related Quantities. 2014.
[244] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 2012;9(4):357–9. doi:10.1038/nmeth.1923
[245] Marty R, Kaabinejadian S, Rossell D, Slifker MJ, van de Haar J, Engin HB, de Prisco N, Ideker T, Hildebrand WH, Font-Burgada J, et al. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*. 2017;171(6):1272–1283.e15. doi:10.1016/j.cell.2017.09.050
[246] Schreiber RD, Old LJ, Smyth MJ. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science*. 2011;331(6024):1565–70. doi:10.1126/science.1203486
[247] Kim Y, Sidney J, Buus SS, Sette A, Nielsen M, Peters B. Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. *BMC Bioinformatics*. 2014;15(1):241. doi:10.1186/1471-2105-15-241
[248] Shao W, Pedrioli PGA, Wolski W, Scurtescu C, Schmid E, Vizcaíno JA, Courcelles M, Schuster H, Kowalewski D, Marino F, et al. The SysteMHC Atlas project. *Nucleic Acids Res.* 2018;46(D1):D1237–D1247. doi:10.1093/nar/gkx664
[249] Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum. Immunol.* 2013;74(10):1313–1320. doi:10.1016/J.HUMIMM.2013.06.025
[250] Liu J, Lichtenberg TM, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee A V., et al. An Integrated TCGA Pan-Cancer

Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*. 2018;173(2):400–416.e11. doi:10.1016/j.cell.2018.02.052
[251] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. doi:10.1093/bioinformatics/btp352
[252] Eichmann M, de Ru A, van Veelen PA, Peakman M, Kronenberg-Versteeg D. Identification and characterisation of peptide binding motifs of six autoimmune disease-associated human leukocyte antigen-class I molecules including *HLA-B*39:06*. *Tissue Antigens*. 2014;84(4):378–388. doi:10.1111/tan.12413
[253] González-Galarza FF, Takeshita LYC, Santos EJM, Kempson F, Maia MHT, Da Silva ALS, Teles E Silva AL, Ghattaoraya GS, Alfirevic A, Jones AR, et al. Allele frequency net 2015 update: New features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res.* 2015;43(D1):D784–D788. doi:10.1093/nar/gku1166
[254] Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, Keşmir C, Peters B. Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* 2013;9(10):e1003266. doi:10.1371/journal.pcbi.1003266
[255] Lund O, Nielsen M, Kesmir C, Petersen AG, Lundegaard C, Worning P, Sylvester-Hvid C, Lamberth K, Roder G, Justesen S, et al. Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics*. 2004;55(12):797–810. doi:10.1007/s00251-004-0647-4
[256] Bjerregaard A-M, Nielsen M, Jurtz V, Barra CM, Hadrup SR, Szallasi Z, Eklund AC. An Analysis of Natural T Cell Responses to Predicted Tumor Neoepitopes. *Front. Immunol.* 2017;8:1566. doi:10.3389/fimmu.2017.01566
[257] Borbulevych OY, Baxter TK, Yu Z, Restifo NP, Baker BM. Increased immunogenicity of an anchor-modified tumor-associated antigen is due to the enhanced stability of the peptide/MHC complex: implications for vaccine design. *J. Immunol.* 2005;174(8):4812–20. doi:10.4049/JIMMUNOL.174.8.4812
[258] Harndahl M, Rasmussen M, Roder G, Dalgaard Pedersen I, Sørensen M, Nielsen M, Buus S. Peptide-MHC class I stability is a better predictor than peptide affinity of CTL

immunogenicity. *Eur. J. Immunol.* 2012;42(6):1405–1416. doi:10.1002/eji.201141774
[259] Degano M, Garcia KC, Apostolopoulos V, Rudolph MG, Teyton L, Wilson IA. A functional hot spot for antigen recognition in a superagonist TCR/MHC complex. *Immunity*. 2000;12(3):251–61.
[260] Sharma AK, Kuhns JJ, Yan S, Friedline RH, Long B, Tisch R, Collins EJ. Class I major histocompatibility complex anchor substitutions alter the conformation of T cell receptor contacts. *J. Biol. Chem.* 2001;276(24):21443–9. doi:10.1074/jbc.M010791200
[261] Liepe J, Marino F, Sidney J, Jeko A, Bunting DE, Sette A, Kloetzel PM, Stumpf MPH, Heck AJR, Mishto M. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science*. 2016;354(6310):354–358. doi:10.1126/science.aaf4384
[262] Granados DP, Laumont CM, Thibault P, Perreault C. The nature of self for T cells-a systems-level perspective. *Curr. Opin. Immunol.* 2015;34:1–8. doi:10.1016/j.coi.2014.10.012
[263] Tsukahara T, Kawaguchi S, Torigoe T, Asanuma H, Nakazawa E, Shimozawa K, Nabeta Y, Kimura S, Kaya M, Nagoya S, et al. Prognostic significance of HLA class I expression in osteosarcoma defined by anti-pan HLA class I monoclonal antibody, EMR8-5. *Cancer Sci.* 2006;97(12):1374–80. doi:10.1111/j.1349-7006.2006.00317.x
[264] Zhang AW, McPherson A, Milne K, Kroeger DR, Hamilton PT, Miranda A, Funnell T, Little N, de Souza CPE, Laan S, et al. Interfaces of Malignant and Immunologic Clonal Dynamics in Ovarian Cancer. *Cell*. 2018;0(0):1–15. doi:10.1016/j.cell.2018.03.073
[265] Frankild S, de Boer RJ, Lund O, Nielsen M, Kesmir C. Amino Acid Similarity Accounts for T Cell Cross-Reactivity and for “Holes” in the T Cell Repertoire Zhang L, editor. *PLoS One*. 2008;3(3):e1831. doi:10.1371/journal.pone.0001831
[266] Yadav M, Jhunjhunwala S, Phung QT, Lupardus P, Tanguay J, Bumbaca S, Franci C, Cheung TK, Fritsche J, Weinschenk T, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*. 2014;515(7528):572–576. doi:10.1038/nature14001
[267] Bristol JA, Schlom J, Abrams SI. Development of a Murine Mutant Ras CD8+ CTL Peptide Epitope Variant That Possesses Enhanced MHC Class I Binding and

Immunogenic Properties. *J. Immunol.* 1998;160:2433–2441.
[268] Vesely MD, Kershaw MH, Schreiber RD, Smyth MJ. Natural innate and adaptive immunity to cancer. *Annu. Rev. Immunol.* 2011;29:235–71. doi:10.1146/annurev-immunol-031210-101324
[269] Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science*. 2015;348(6230):69–74. doi:10.1126/science.aaa4971
[270] Simoni Y, Becht E, Fehlings M, Loh CY, Koo S-L, Teng KWW, Yeong JPS, Nahar R, Zhang T, Kared H, et al. Bystander CD8+ T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature*. 2018;557(7706):575–579. doi:10.1038/s41586-018-0130-2
[271] Anagnostou V, Smith KN, Forde PM, Niknafs N, Bhattacharya R, White J, Zhang T, Adleff V, Phallen J, Wali N, et al. Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. *Cancer Discov.* 2017;7(3):264–276. doi:10.1158/2159-8290.CD-16-0828
[272] McGranahan N, Furness AJS, Rosenthal R, Ramskov S, Lyngaa R, Saini K, Jamal-Hanjani M, Wilson GA, Birkbak NJ, Hiley CT, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*. 2016;490(March):1–11. doi:10.1126/science.aaf1490
[273] Cai W, Zhou D, Wu W, Tan WL, Wang J, Zhou C, Lou Y. MHC class II restricted neoantigen peptides predicted by clonal mutation analysis in lung adenocarcinoma patients: implications on prognostic immunological biomarker and vaccine design. *BMC Genomics*. 2018;19(1):582. doi:10.1186/s12864-018-4958-5
[274] Bjerregaard A-M, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC. MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol. Immunother.* 2017 Apr 20:1–8. doi:10.1007/s00262-017-2001-3
[275] Zhou Z, Lyu X, Wu J, Yang X, Wu S, Zhou J, Gu X, Su Z, Chen S. TSNAD: an integrated software for cancer somatic mutation and tumour-specific neoantigen detection. *R. Soc. Open Sci.* 2017;4(4):170050. doi:10.1098/rsos.170050
[276] Bais P, Namburi S, Gatti DM, Zhang X, Chuang JH. CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. *Bioinformatics*. 2017;33(19):3110–3112. doi:10.1093/bioinformatics/btx375

[277] Schubert B, Walzer M, Brachvogel H-P, Szolek A, Mohr C, Kohlbacher O. FRED 2: an immunoinformatics framework for Python. *Bioinformatics*. 2016;32(13):2044–6. doi:10.1093/bioinformatics/btw113
[278] Rubinsteyn A, Kodysh J, Hodes I, Mondet S, Aksoy BA, Finnigan JP, Bhardwaj N, Hammerbacher J. Computational Pipeline for the PGV-001 Neoantigen Vaccine Trial. *Front. Immunol.* 2018;8:1807. doi:10.3389/fimmu.2017.01807
[279] Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, Wong YNS, Rowan A, Kanu N, Al Bakir M, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* 2017;18(8):1009–1021. doi:10.1016/S1470-2045(17)30516-8
[280] Duan F, Duitama J, Al Seesi S, Ayres CM, Corcelli SA, Pawashe AP, Blanchard T, McMahon D, Sidney J, Sette A, et al. Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J. Exp. Med.* 2014;211(11):jem.20141308-. doi:10.1084/jem.20141308
[281] Kim S, Kim HS, Kim E, Lee MG, Shin E, Paik S, Kim S. Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann. Oncol.* 2018;29(4):1030–1036. doi:10.1093/annonc/mdy022
[282] Ghorani E, Rosenthal R, McGranahan N, Reading JL, Lynch M, Peggs KS, Swanton C, Quezada SA. Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma. *Ann. Oncol.* 2018;29(1):271–279. doi:10.1093/annonc/mdx687
[283] Trolle T, Nielsen M. NetTepi: an integrated method for the prediction of T cell epitopes. *Immunogenetics*. 2014 May 27. doi:10.1007/s00251-014-0779-0
[284] Rötzschke O, Falk K, Deres K, Schild H, Norda M, Metzger J, Jung G, Rammensee H-G. Isolation and analysis of naturally processed viral peptides as recognized by cytotoxic T cells. *Nature*. 1990;348(6298):252–254. doi:10.1038/348252a0
[285] Hunt DF, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, Cox AL, Appella E, Engelhard VH. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science*. 1992;255(5049):1261–3.

doi:10.1126/SCIENCE.1546328
[286] Hassan C, Kester MGD, de Ru AH, Hombrink P, Drijfhout JW, Nijveen H, Leunissen JAM, Heemskerk MHM, Falkenburg JHF, van Veelen PA. The Human Leukocyte Antigen–presented Ligandome of B Lymphocytes. *Mol. Cell. Proteomics*. 2013;12(7):1829–1843. doi:10.1074/mcp.M112.024810
[287] Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, Stevens J, Lane W, Zhang GL, Eisenhaure TM, et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity*. 2017;46(2):315–326. doi:10.1016/J.IMMUNI.2017.02.007
[288] Bassani-Sternberg M, Bräunlein E, Klar R, Engleitner T, Sinitcyn P, Audehm S, Straub M, Weber J, Slotta-Huspenina J, Specht K, et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* 2016;7:13404. doi:10.1038/ncomms13404
[289] Jurtz VI, 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. *J. Immunol.* 2017;199(9):3360–3368. doi:10.4049/jimmunol.1700893
[290] O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst.* 2018;7(1):129–132.e4. doi:10.1016/j.cels.2018.05.014
[291] Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, Kandalaft LE, Coukos G, Gfeller D. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity Hertz T, editor. *PLOS Comput. Biol.* 2017;13(8):e1005725. doi:10.1371/journal.pcbi.1005725
[292] Gfeller D, Bassani-Sternberg M. Predicting Antigen Presentation—What Could We Learn From a Million Peptides? *Front. Immunol.* 2018;9:1716. doi:10.3389/fimmu.2018.01716
[293] Liu G, Li D, Li Z, Qiu S, Li W, Chao C, Yang N, Li H, Cheng Z, Song X, et al. PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity. *Gigascience*. 2017;6(5):1–11. doi:10.1093/gigascience/gix017

[294] Mukherjee S, Bhattacharyya C, Chandra N. HLaffy: estimating peptide affinities for Class-1 HLA molecules by learning position-specific pair potentials. *Bioinformatics*. 2016;32(15):2297–2305. doi:10.1093/bioinformatics/btw156
[295] Antunes DA, Devaurs D, Moll M, Lizée G, Kavraki LE. General Prediction of Peptide-MHC Binding Modes Using Incremental Docking: A Proof of Concept. *Sci. Rep.* 2018;8(1):4327. doi:10.1038/s41598-018-22173-4
[296] Salk JJ, Horwitz MS. Passenger mutations as a marker of clonal cell lineages in emerging neoplasia. *Semin. Cancer Biol.* 2010;20(5):294–303. doi:10.1016/j.semcancer.2010.10.008
[297] Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*. 2012;486(7403):395–9. doi:10.1038/nature10933
[298] Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Côté A, Shah SP. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*. 2014;11(4):396–398. doi:10.1038/nmeth.2883
[299] Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, Vij R, Tomasson MH, Graubert TA, Walter MJ, et al. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution Beerenwinkel N, editor. *PLoS Comput. Biol.* 2014;10(8):e1003665. doi:10.1371/journal.pcbi.1003665
[300] Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, Melnyk N, McPherson A, Bashashati A, Laks E, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* 2014;24(11):1881–93. doi:10.1101/gr.180281.114
[301] Deshwar AG, Vembu S, Yung CK, Jang G, Stein L, Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* 2015;16(1):35. doi:10.1186/s13059-015-0602-8
[302] Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, Gelmon K, Chia S, Mar C, Wan A, et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*. 2015;518(7539):422–426. doi:10.1038/nature13952
[303] Jamal-Hanjani M, Hackshaw A, Ngai Y, Shaw J, Dive C, Quezada S, Middleton G, de Bruin E, Le Quesne J, Shafi S, et al. Tracking Genomic Cancer Evolution for

Precision Medicine: The Lung TRACERx Study. *PLoS Biol.* 2014;12(7):e1001906. doi:10.1371/journal.pbio.1001906
[304] Chen P-LL, Roh W, Reuben A, Cooper ZA, Spencer CN, Prieto PA, Miller JP, Bassett RL, Gopalakrishnan V, Wani K, et al. Analysis of Immune Signatures in Longitudinal Tumor Samples Yields Insight into Biomarkers of Response and Mechanisms of Resistance to Immune Checkpoint Blockade. *Cancer Discov.* 2016;6(8). doi:10.1158/2159-8290.CD-15-1545
[305] Roh W, Chen P-L, Reuben A, Spencer CN, Prieto PA, Miller JP, Gopalakrishnan V, Wang F, Cooper ZA, Reddy SM, et al. Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci. Transl. Med.* 2017;9(379):eaah3560. doi:10.1126/scitranslmed.aah3560
[306] Teng MWL, Ngiow SF, Ribas A, Smyth MJ. Classifying Cancers Based on T-cell Infiltration and PD-L1. *Cancer Res.* 2015;75(11):2139–45. doi:10.1158/0008-5472.CAN-15-0255
[307] Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJM, Robert L, Chmielowski B, Spasic M, Henry G, Ciobanu V, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*. 2014;515(7528):568–571. doi:10.1038/nature13954
[308] Patel SP, Kurzrock R. PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. *Mol. Cancer Ther.* 2015;14(4):847–56. doi:10.1158/1535-7163.MCT-14-0983
[309] Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, Lu S, Kemberling H, Wilt C, Luber BS, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science*. 2017;357(6349):409–413. doi:10.1126/science.aan6733
[310] Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, Stephens PJ, Daniels GA, Kurzrock R. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol. Cancer Ther.* 2017;16(11):2598–2608. doi:10.1158/1535-7163.MCT-17-0386
[311] US-FDA. KEYTRUDA. 2017. https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/125514s017s018lbl.pdf
[312] US-FDA. OPDIVO. 2017.

https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/125554s024lbl.pdf
[313] US-FDA. BAVENCIO. 2017.
https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/761078s000lbl.pdf
[314] US-FDA. IMFINZI. 2017.
https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/761069s000lbl.pdf
[315] US-FDA. TECENTRIQ. 2016.
https://www.accessdata.fda.gov/drugsatfda_docs/label/2016/761034s000lbl.pdf
[316] US-FDA. YERVOY. 2018.
https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/125377s096lbl.pdf
[317] Rubelt F, Busse CE, Bukhari SAC, Bürckert JP, Mariotti-Ferrandiz E, Cowell LG, Watson CT, Marthandan N, Faison WJ, Hershberg U, et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat. Immunol.* 2017;18(12):1274–1278.
[318] Olsen LR, Tongchusak S, Lin H, Reinherz EL, Brusic V, Zhang GL. TANTIGEN: a comprehensive database of tumor T cell antigens. *Cancer Immunol. Immunother.* 2017 Mar 9:1–5. doi:10.1007/s00262-017-1978-y
[319] Shugay M, Bagaev D V, Zvyagin I V, Vroomans RM, Crawford JC, Dolton G, Komech EA, Sycheva AL, Koneva AE, Egorov ES, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* 2018;46(D1):D419–D427. doi:10.1093/nar/gkx760
[320] Leem J, de Oliveira SHP, Krawczyk K, Deane CM. STCRDab: the structural T-cell receptor database. *Nucleic Acids Res.* 2018;46(D1):D406–D412. doi:10.1093/nar/gkx971
[321] Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*. 2017;547(7661):89–93. doi:10.1038/nature22383
[322] Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature*. 2017;547(7661):94–98. doi:10.1038/nature22976
[323] Pogorelyy M V, Minervina AA, Chudakov DM, Mamedov IZ, Lebedev YB, Mora T,

Walczak AM. Method for identification of condition-associated public antigen receptor sequences. *Elife*. 2018;7. doi:10.7554/eLife.33050
[324] Bentzen AK, Such L, Jensen KK, Marquard AM, Jessen LE, Miller NJ, Church CD, Lyngaa R, Koelle DM, Becker JC, et al. T cell receptor fingerprinting enables in-depth characterization of the interactions governing recognition of peptide–MHC complexes. *Nat. Biotechnol.* 2018 Nov 19. doi:10.1038/nbt.4303
[325] Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva E V, Chudakov DM. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods*. 2015;12(5):380–381. doi:10.1038/nmeth.3364
[326] Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, Hemmers S, Putintseva E V, Obraztsova AS, Shugay M, et al. Antigen receptor repertoire profiling from RNA-seq data. *Nat. Biotechnol.* 2017;35(10):908–911. doi:10.1038/nbt.3979
[327] Li B, Li T, Pignon J-C, Wang B, Wang J, Shukla SA, Dou R, Chen Q, Hodi FS, Choueiri TK, et al. Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.* 2016 May 30. doi:10.1038/ng.3581
[328] Li B, Li T, Wang B, Dou R, Zhang J, Liu JS, Liu XS. Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA–seq data. *Nat. Genet.* 2017;49(4):482–483. doi:10.1038/ng.3820
[329] Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, Teichmann SA. T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods*. 2016;13(4):329–332. doi:10.1038/nmeth.3800
[330] Redmond D, Poran A, Elemento O. Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. *Genome Med.* 2016;8(1):80. doi:10.1186/s13073-016-0335-7
[331] Gong Q, Wang C, Zhang W, Iqbal J, Hu Y, Greiner TC, Cornish A, Kim J-H, Rabadan R, Abate F, et al. Assessment of T-cell receptor repertoire and clonal expansion in peripheral T-cell lymphoma using RNA-seq data. *Sci. Rep.* 2017;7(1):11301. doi:10.1038/s41598-017-11310-0
[332] Chicz RM, Urban RG, Lane WS, Gorga JC, Stern LJ, Vignali DAA, Strominger JL. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-

related molecules and are heterogeneous in size. *Nature*. 1992;358(6389):764–768. doi:10.1038/358764a0
[333] Holland CJ, Cole DK, Godkin A. Re-Directing CD4+ T Cell Responses with the Flanking Residues of MHC Class II-Bound Peptides: The Core is Not Enough. *Front. Immunol.* 2013;4:172. doi:10.3389/fimmu.2013.00172
[334] Nielsen M, Lund O, Buus S, Lundegaard C. MHC class II epitope predictive algorithms. *Immunology*. 2010;130(3):319–28. doi:10.1111/j.1365-2567.2010.03268.x
[335] Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, Drineas P. PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations. *PLoS Genet.* 2007;3(9):e160. doi:10.1371/journal.pgen.0030160
[336] Khong HT, Restifo NP. Natural selection of tumor variants in the generation of “tumor escape” phenotypes. *Nat. Immunol.* 2002;3(11):999–1005. doi:10.1038/ni1102-999
[337] Zou W, Chen L. Inhibitory B7-family molecules in the tumour microenvironment. *Nat. Rev. Immunol.* 2008;8(6):467–477. doi:10.1038/nri2326
[338] Zou W. Immunosuppressive networks in the tumour environment and their therapeutic relevance. *Nat. Rev. Cancer*. 2005;5(4):263–274. doi:10.1038/nrc1586
[339] Sharma G, Holt RA. T-cell epitope discovery technologies. *Hum. Immunol.* 2014;75(6):514–519. doi:10.1016/J.HUMIMM.2014.03.003
[340] Holt RA, Sharma G. T-cell epitope identification. US Patent Application US20180052176A1. 2015 Mar 25.
[341] Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu Y-M, Cao X, Kalyana-Sundaram S, Sam L, Balbin OA, Quist MJ, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci. Transl. Med.* 2011;3(111):111ra121. doi:10.1126/scitranslmed.3003161
[342] Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich DC, Kryukov G, Carter SL, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat. Med.* 2014;20(6):682–688. doi:10.1038/nm.3559
[343] Sadelain M, Rivière I, Riddell S. Therapeutic T cell engineering. *Nature*. 2017;545(7655):423–431. doi:10.1038/nature22395

[344] Eizaguirre C, Lenz TL, Kalbe M, Milinski M. Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nat. Commun.* 2012;3:621. doi:10.1038/ncomms1632
[345] Ibana JA, Schust DJ, Sugimoto J, Nagamatsu T, Greene SJ, Quayle AJ. Chlamydia trachomatis immune evasion via downregulation of MHC class I surface expression involves direct and indirect mechanisms. *Infect. Dis. Obstet. Gynecol.* 2011;2011:420905. doi:10.1155/2011/420905
[346] Antoniou AN, Powis SJ. Pathogen evasion strategies for the major histocompatibility complex class I assembly pathway. *Immunology*. 2008;124(1):1–12. doi:10.1111/j.1365-2567.2008.02804.x
[347] Schofield L, Villaquiran J, Ferreira A, Schellekens H, Nussenzweig R, Nussenzweig V. γ Interferon, CD8+ T cells and antibodies required for immunity to malaria sporozoites. *Nature*. 1987;330(6149):664–666. doi:10.1038/330664a0
[348] Khusmith S, Sedegah M, Hoffman SL. Complete protection against Plasmodium yoelii by adoptive transfer of a CD8+ cytotoxic T-cell clone recognizing sporozoite surface protein 2. *Infect. Immun.* 1994;62(7):2979–83.
[349] Vinetz JM, Kumar S, Good MF, Fowlkes BJ, Berzofsky JA, Miller LH. Adoptive transfer of CD8+ T cells from immune animals does not transfer immunity to blood stage Plasmodium yoelii malaria. *J. Immunol.* 1990;144(3):1069–74.
[350] Boyd MF, editor. Malariology ; a comprehensive survey of all aspects of this group of disease from a global standpoint , by 65 contributors . v . 1 Public Domain , Google-digitized. London: W. B. Saunders Company; 1949.
[351] Blümel J, Burger R, Drosten C, Gröner A, Gürtler L, Heiden M, Jansen B, Klamm H, Ludwig WD, Montag-Lessing T, et al. Malaria. *Transfus. Med. Hemotherapy*. 2009;36(1):48–60.

Appendices

Appendix A Chapter 2: Supplementary Material

A.1 Supplementary Figures

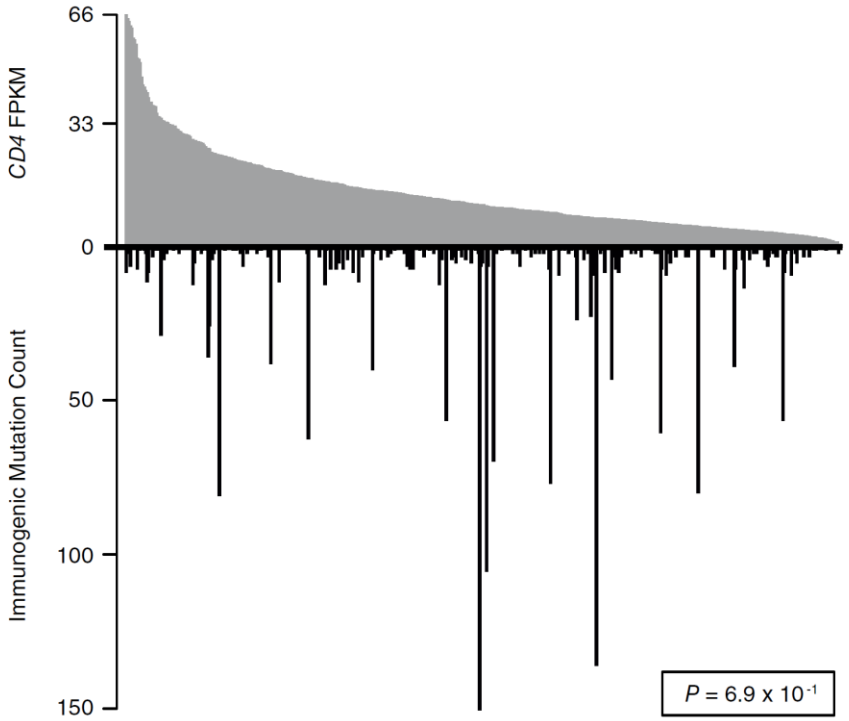

Figure A.1: Skew plot of *CD4* expression and predicted immunogenic mutations. Patients were ordered along the x-axis according to their *CD4* expression. Each patient's *CD4* expression was plotted above the x-axis, and count of predicted immunogenic mutations below. No significant skew was observed ($P = 6.9 \times 10^{-1}$).

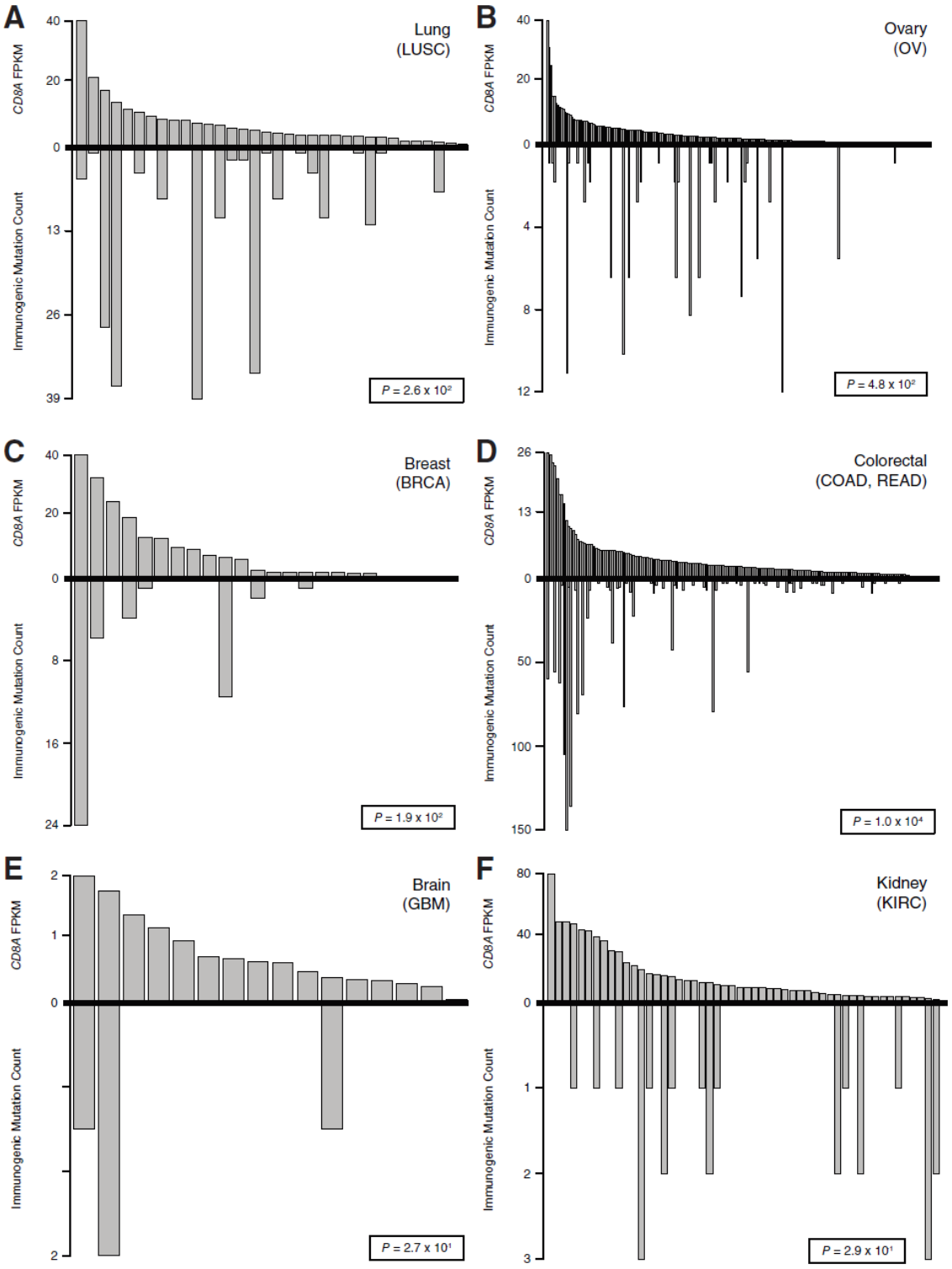

Figure A.2: Skew plots for each cancer type individually. Patients were ordered along the x-axis according to their *CD8A* expression. Each patient's *CD8A* expression was plotted along the x-axis, and count of predicted immunogenic mutations below for **(A)** lung (LUSC), **(B)** ovary (OV), **(C)** breast (BRCA), **(D)** colorectal (COAD/READ), **(E)** brain (GBM) and **(F)** kidney (KIRC).

Appendix B Chapter 3: Supplementary Material

B.1 Supplementary Figures

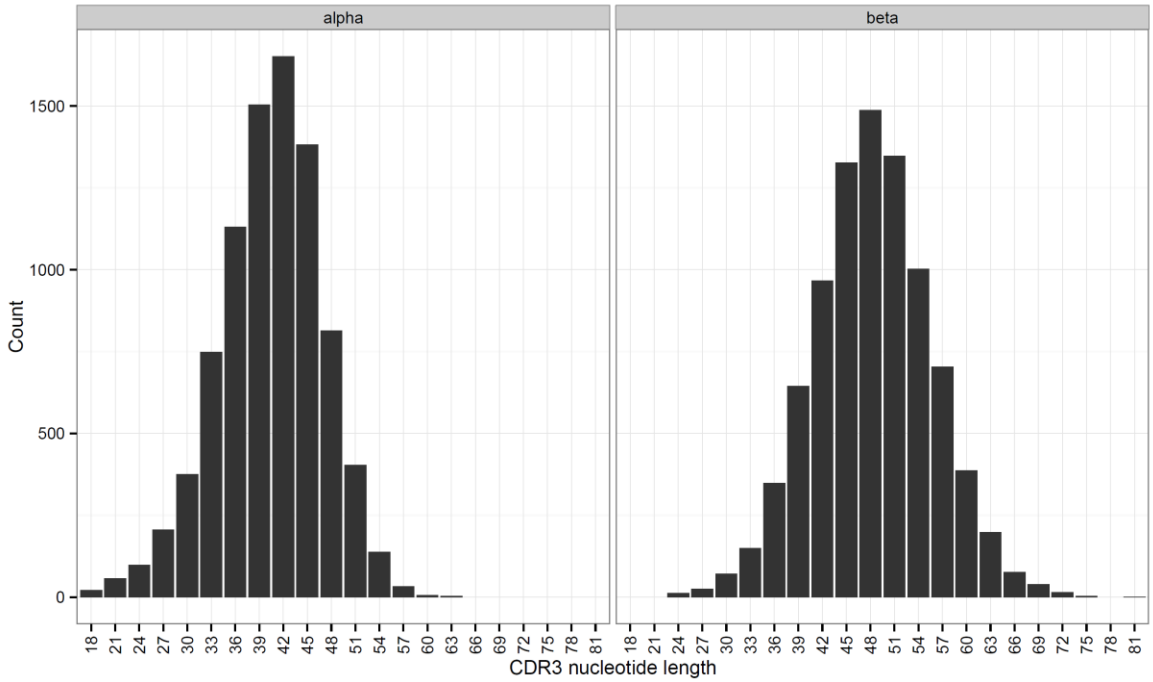

Figure B.1: Length distributions of *in silico* generated CDR3 sequences. CDR3-alpha and CDR3-beta sequence lengths plotted separately. CDR3-alpha has mean 40.35 (standard deviation 6.54), and CDR3-beta has mean 48.29 (standard deviation 7.41).

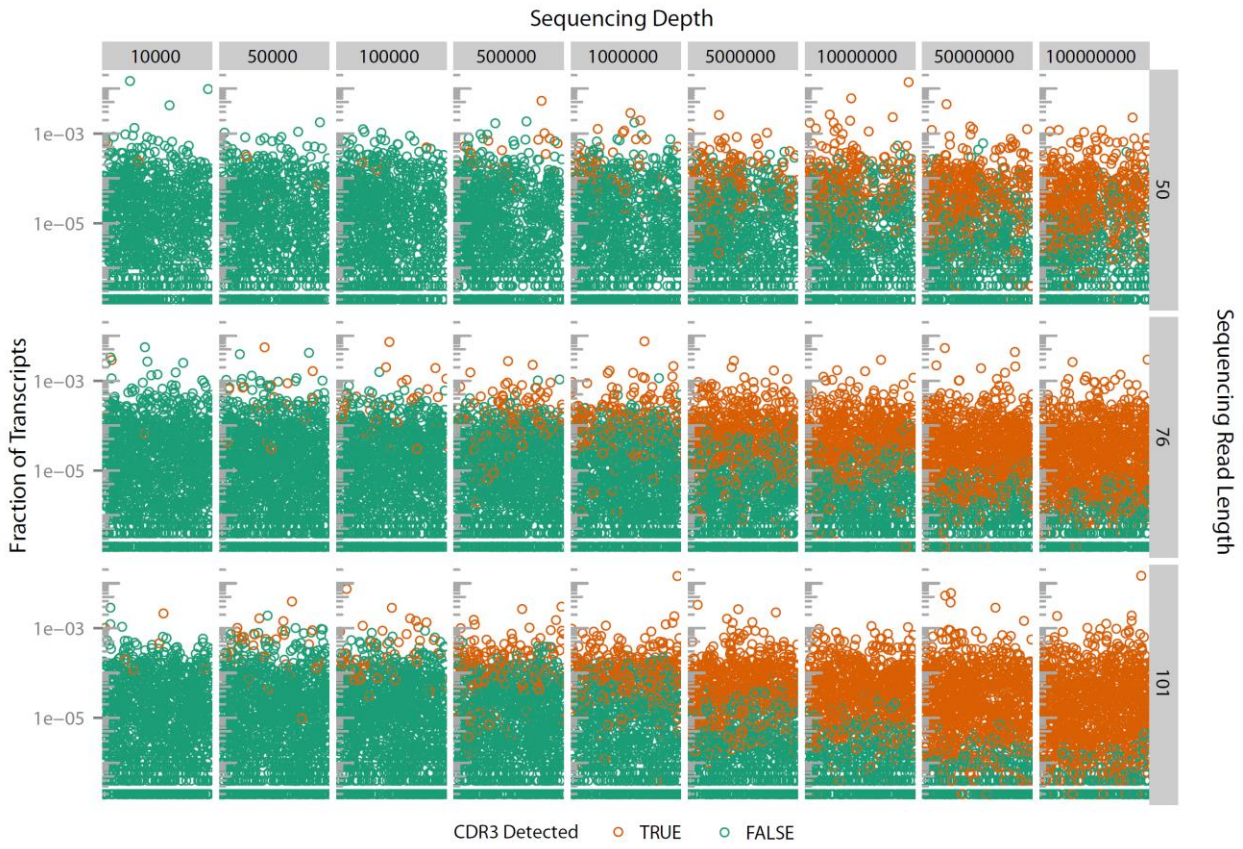

Figure B.2: TCR transcript abundance vs. sequencing depth. For the *in silico* data, the simulated TCR transcript abundance was tracked. Simulated RNA libraries that were sequenced deeper allowed lower abundance TCR transcripts to be detected.

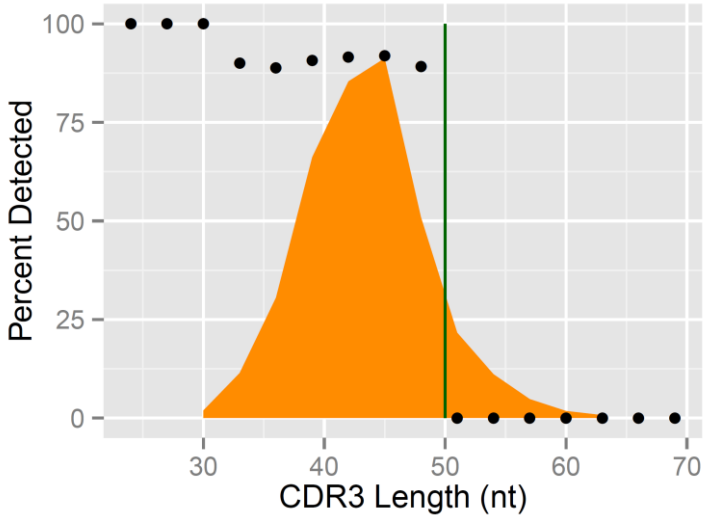

Figure B.3: Detection probability of CDR3-betas with varying lengths using error-free 50 nt reads centered on the CDR3 region. Orange density plot shows the distribution of CDR3-beta lengths in the normal population [11]. CDR3s that are longer than the read length (50 nts, green line) are not detected.

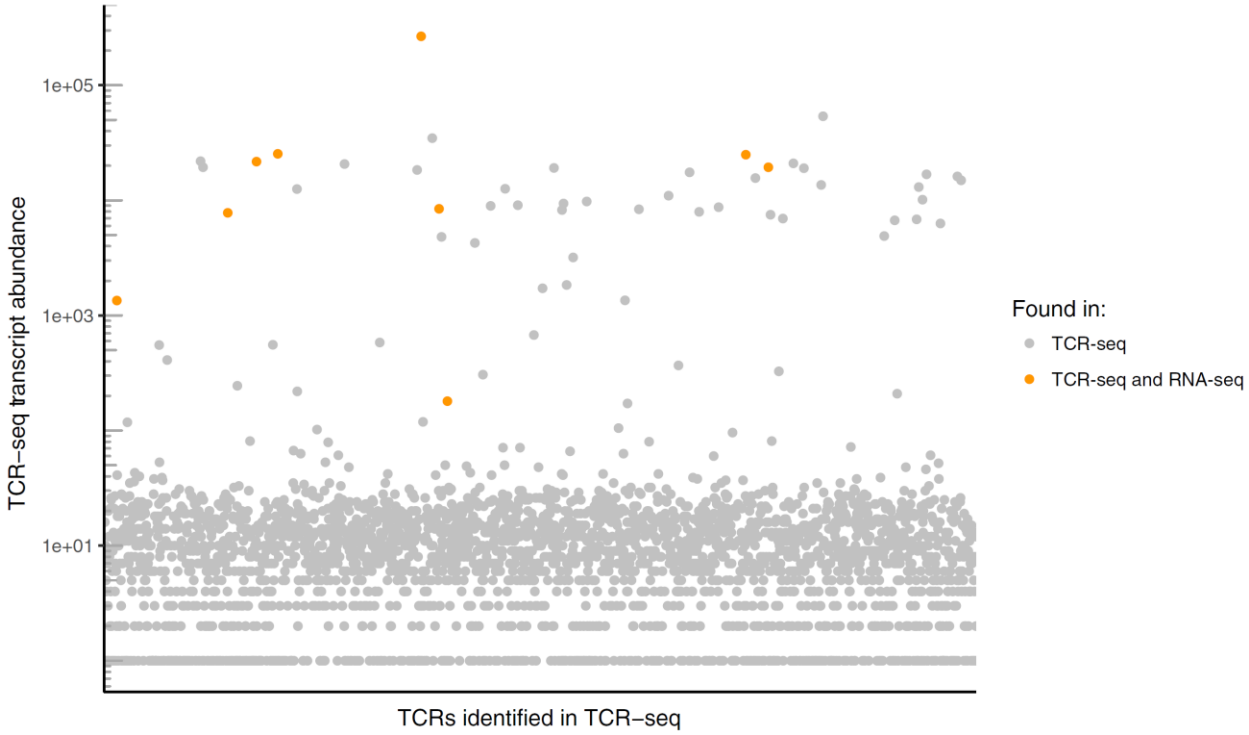

Figure B.4: Comparison of TCRs detected by conventional TCR-seq and RNA-seq. All TCRs detected by TCR-seq are displayed in lexicographical order along the x-axis, with the number of sequence reads assigned to each clonotype on the y-axis. TCRs which were also found in the RNA-seq dataset are coloured orange.

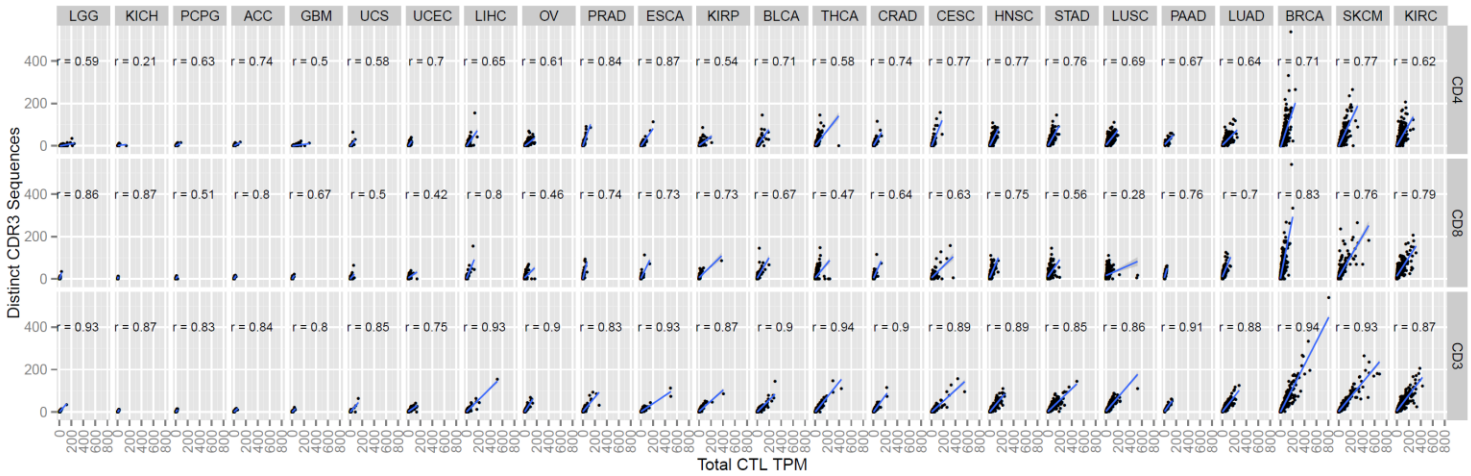

Figure B.5: Relationship between number of CDR3 amino acid sequences extracted and *CD4*, *CD8*, and *CD3* expression in tumour samples. Pearson correlation coefficients are displayed.

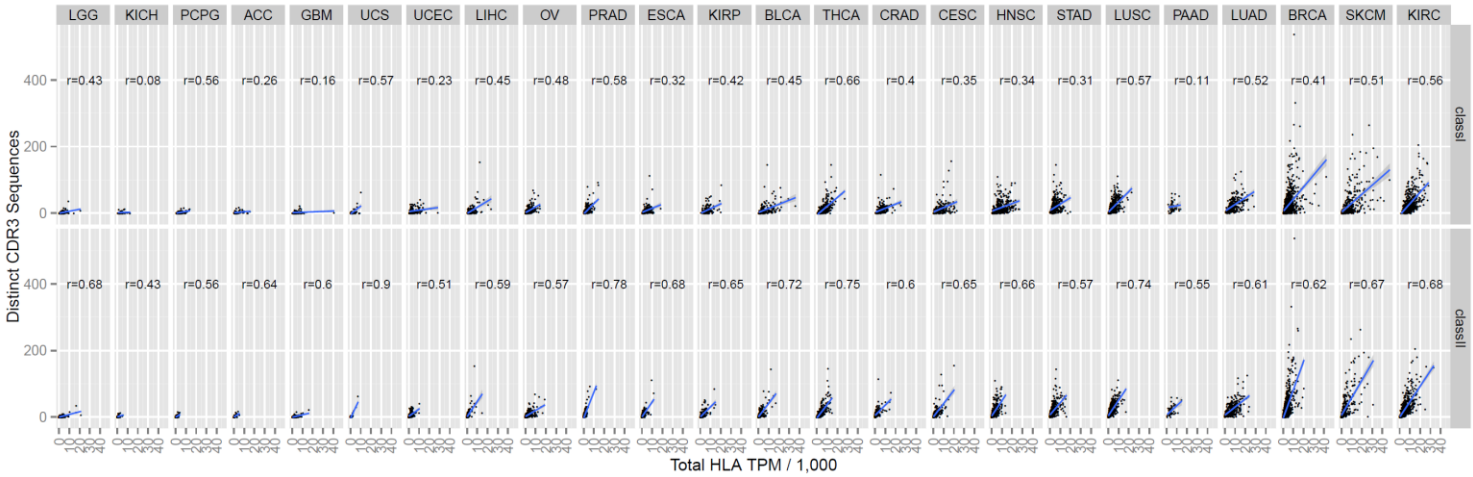

Figure B.6: Relationship between number of CDR3 amino acid sequences extracted and HLA class I and class II expression in tumour samples. Pearson correlation coefficients are displayed.

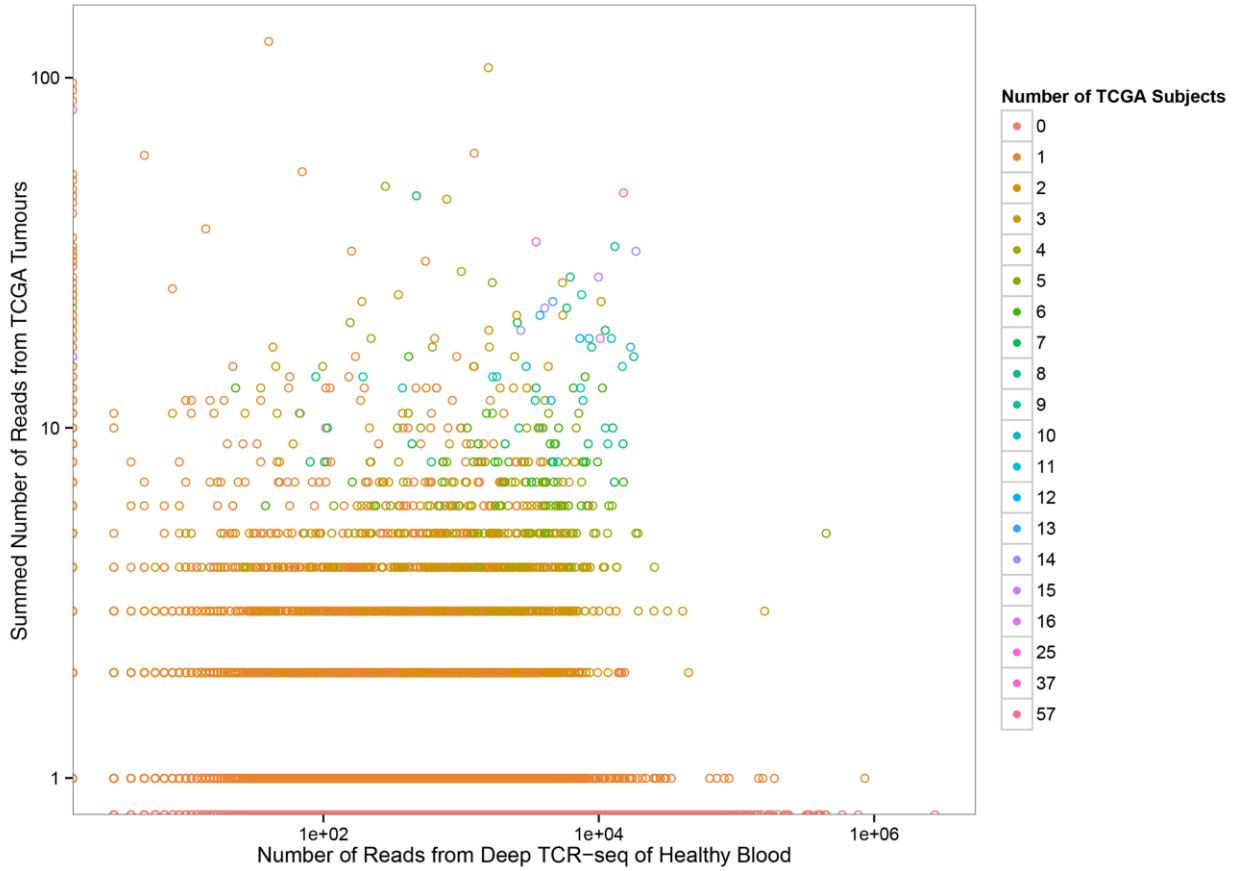

Figure B.7: Overlap between CDR3-beta extracted from TCGA tumours and one individual's deeply sequenced healthy blood sample. CDR3s that are found in multiple TCGA subjects are more likely to be found in the healthy individual's TCR repertoire

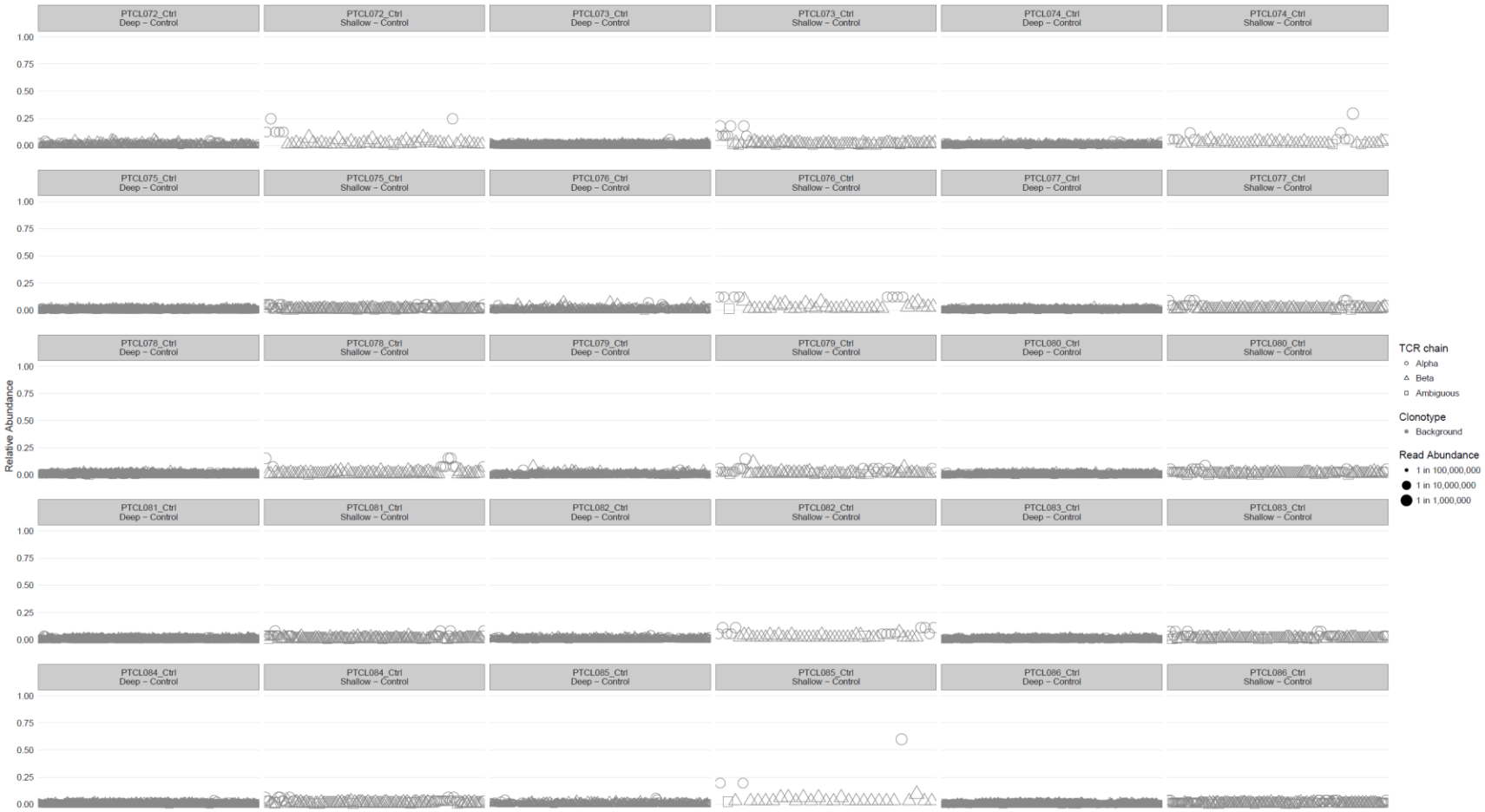

Figure B.8: Relative abundance of clonotypes in control samples. Deep and shallow sequencing for each subject shown. Y-axis showing the abundance of each clonotype relative to the total abundance of all clonotypes of the same chain in that sample. Clonotypes are plotted along the x-axis in lexicographical order. The read abundance of each clonotype is represented by its size.

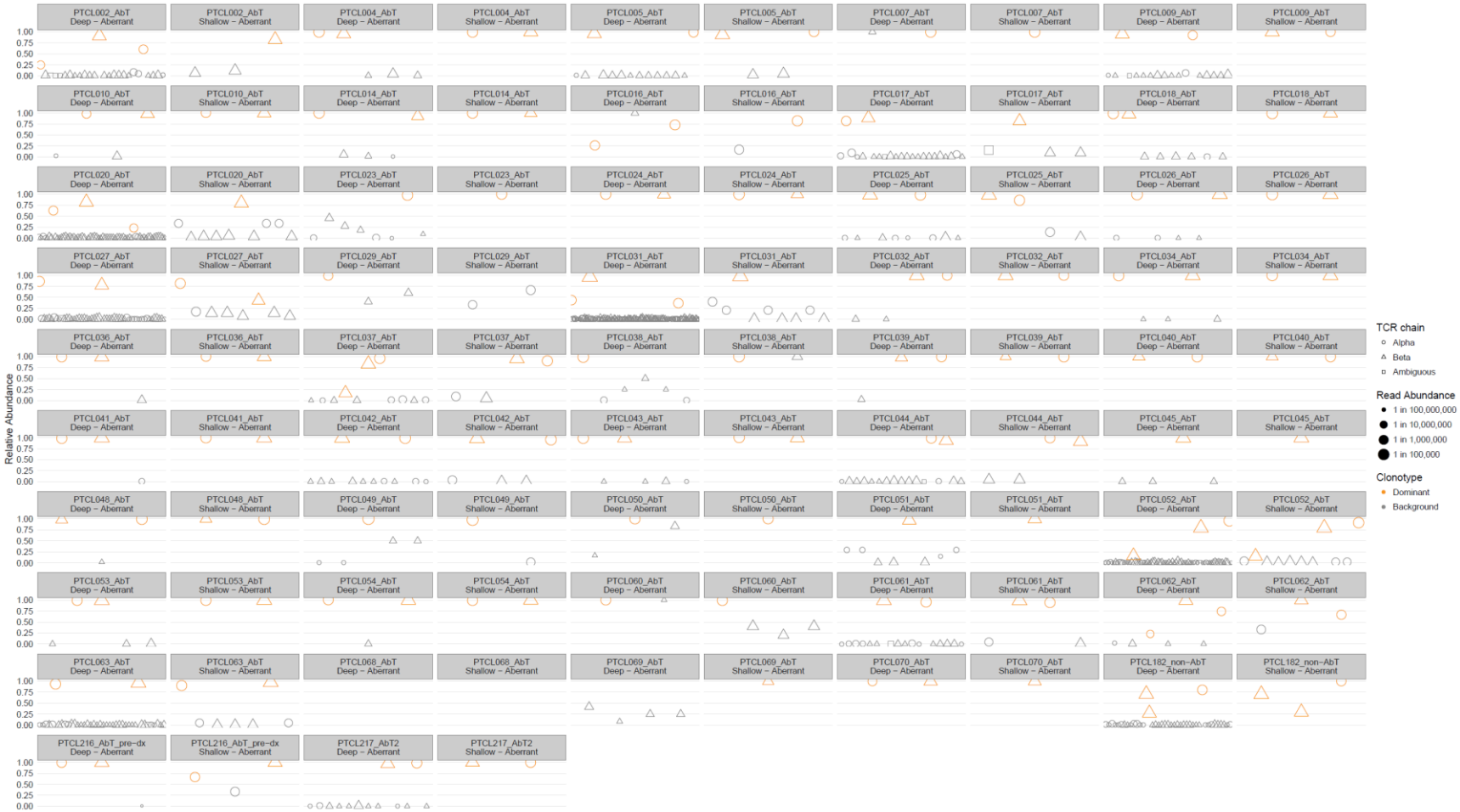

Figure B.9: Relative abundance of clonotypes in samples that were aberrant by flow cytometry. Deep and shallow sequencing for each subject shown. Y-axis showing the abundance of each clonotype relative to the total abundance of all clonotypes of the same chain in that sample. Clonotypes are plotted along the x-axis in lexicographical order. Clonotypes determined to be dominant are coloured orange. The read abundance of each clonotype is represented by its size.

Figure B.10: Relative abundance of clonotypes in samples that were not aberrant by flow cytometry. Deep and shallow sequencing for each subject shown. Y-axis showing the abundance of each clonotype relative to the total abundance of all clonotypes of the same chain in that sample. Clonotypes are plotted along the x-axis in lexicographical order. Clonotypes determined to be dominant are coloured orange. The read abundance of each clonotype is represented by its size.

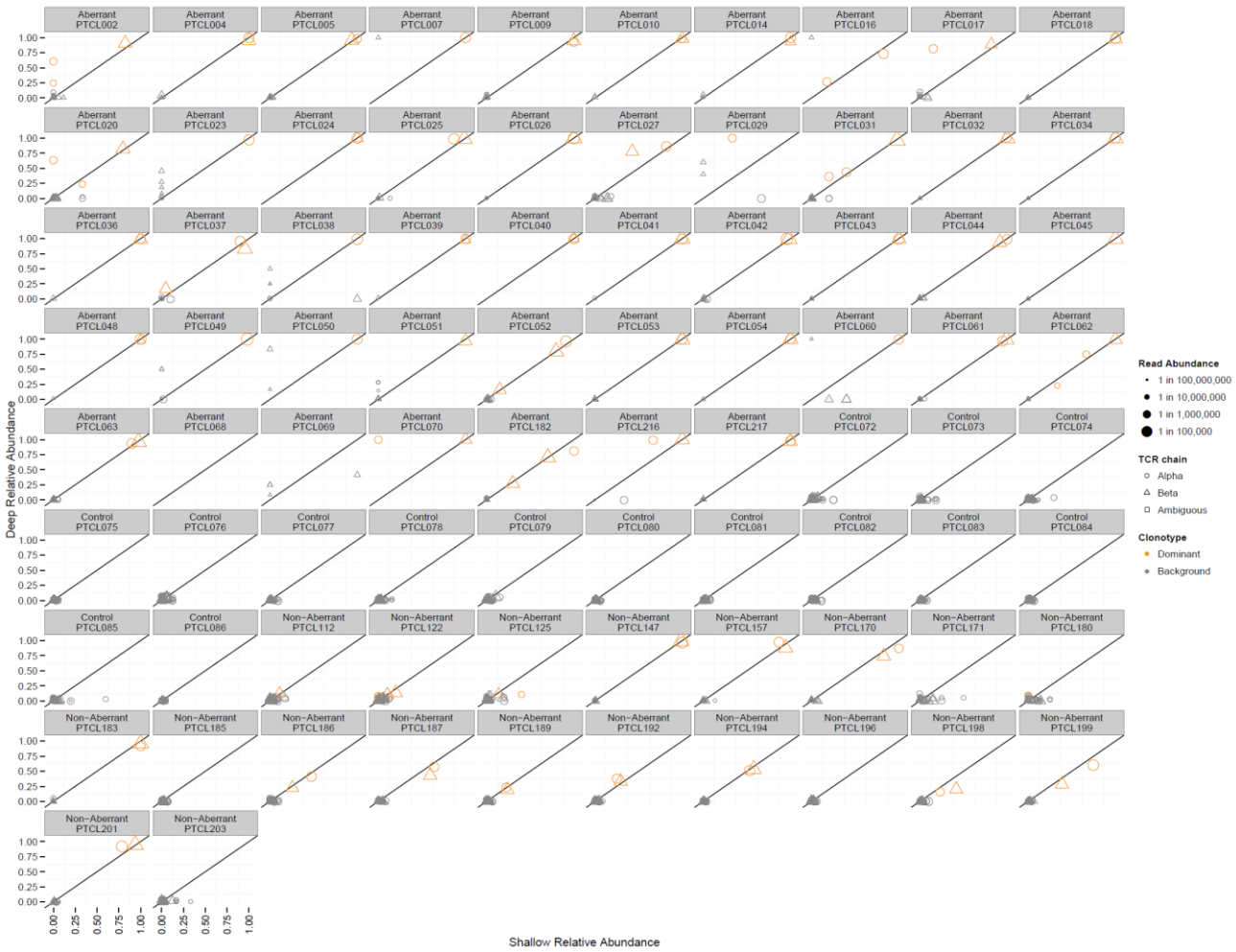

Figure B.11: Comparison of clonotype relative abundance in shallow versus deep sequencing for all 82 samples. X-axis shows the relative abundance of each clonotype in the shallow sequence data, and y-axis shows the relative abundance of each clonotype in the deep sequence data. Clonotypes falling along the diagonal have equal abundances in both shallow and deep datasets. Clonotypes determined to be dominant are coloured orange. The read abundance of each clonotype is represented by its size. Dominance and read abundance are based on deep sequence data unless only found in shallow data.

B.2 Supplementary Tables

Table B.1: Table of ENCODE datasets merged for use as a negative control.

Library	Biosample	File Accessions	Total number of reads
ENCLB113QPT	ENCBS780PCJ - smooth muscle cell	ENCFF548JWS ENCFF004IRQ	208,920,126
ENCLB615ALX	ENCBS077RUJ – hepatocyte	ENCFF245VTB ENCFF369QXD	200,577,302
ENCLB160QNF	ENCBS018TPT - neural progenitor cell	ENCFF939FVE ENCFF201WLO	202,290,896
ENCLB714MUL	ENCBS514GVM - SK-N-DZ	ENCFF482SFO ENCFF691TRA	156,710,634
ENCLB534MTC	ENCBS234AAA - LHCN-M2	ENCFF119TIN ENCFF494PBN	173,859,996
ENCLB011AUM	ENCBS367AAA - fibroblast of arm	ENCFF002DMN ENCFF002DMO	182,860,148
ENCLB059TNM	ENCBS518AAA - SK-MEL-5	ENCFF002DLD ENCFF002DLF	196,017,866
All Combined			**1,317,236,968**

Table B.2: Table of optimized parameters for a range of false discovery rates.

Allowed false positives per 100M reads	Sensitivity (%)[a]	TCR Chain	Read length	minAlignmentMatches Parameter V	minAlignmentMatches Parameter J
0	98.15	Alpha	50	10	20
0	90.76	Beta	50	12	16
0	100.00	Alpha	76	18	11
0	99.98	Beta	76	12	18
0	100.00	Alpha	101	12	19
0	100.00	Beta	101	14	16
1	98.55	Alpha	50	10	17
1	94.73	Beta	50	13	14
1	100.00	Alpha	76	17	11
1	99.99	Beta	76	8	18
1	100.00	Alpha	101	19	9
1	100.00	Beta	101	14	14
5	98.70	Alpha	50	8	17
5	97.00	Beta	50	12	13
5	100.00	Alpha	76	12	15
5	99.99	Beta	76	12	14
5	100.00	Alpha	101	14	14
5	100.00	Beta	101	8	17
10	98.89	Alpha	50	10	15
10	97.85	Beta	50	12	12
10	100.00	Alpha	76	10	16
10	99.99	Beta	76	11	14
10	100.00	Alpha	101	12	15
10	100.00	Beta	101	9	16

[a] Calculated as the count of CDR3s recovered with that parameter pair divided by the maximum count of CDR3s recovered in all parameter pairs tested.

Table B.3: Table of predictions from model for some relevant explanatory variable values. <u>Bold underlined**</u>** values vary within each group.

Transcript Fraction	Sequencing Depth	Read Length	CDR3 Length	Probability of detection (95 % CI)
1×10^{-5}	70,000,000	50	45	0.503 (0.495 – 0.512)
1×10^{-6}	50,000,000	76	48	0.100 (0.097 – 0.102)
5×10^{-6}	50,000,000	76	48	0.306 (0.302 – 0.311)
1×10^{-5}	50,000,000	76	48	0.445 (0.440 – 0.450)
2.5×10^{-5}	50,000,000	76	48	0.638 (0.633 – 0.643)
1×10^{-5}	**10,000,000**	76	48	0.094 (0.092 – 0.096)
1×10^{-5}	**25,000,000**	76	48	0.183 (0.180 – 0.186)
1×10^{-5}	**50,000,000**	76	48	0.445 (0.440 – 0.450)
1×10^{-5}	**100,000,000**	76	48	0.912 (0.908 – 0.915)
1×10^{-5}	50,000,000	**50**	48	0.243 (0.237 – 0.249)
1×10^{-5}	50,000,000	**76**	48	0.445 (0.440 – 0.450)
1×10^{-5}	50,000,000	**101**	48	0.659 (0.653 – 0.665)
1×10^{-5}	50,000,000	50	**41**	0.302 (0.296 – 0.308)
1×10^{-5}	50,000,000	50	**45**	0.267 (0.261 – 0.273)
1×10^{-5}	50,000,000	50	**48**	0.243 (0.237 – 0.249)
1×10^{-5}	50,000,000	76	**39**	0.541 (0.535 – 0.546)
1×10^{-5}	50,000,000	76	**45**	0.477 (0.472 – 0.482)
1×10^{-5}	50,000,000	76	**51**	0.414 (0.408 – 0.420)

Table B.4: Sample numbers for tumour-normal pairs in each tumour site.

Tumour Site	Number of Samples
BRCA	96
KIRC	56
THCA	47
LUSC	42
PRAD	40
HNSC	34
STAD	30
LIHC	27
CRAD	21
KIRP	15
KICH	14
LUAD	13
ESCA	10
BLCA	9
CESC	3
UCEC	3
PCPG	2
Total	**462**

Table B.5: Primary antibodies used for flow cytometry/FACS experiments.

Antibody	Fluorochrome	Clone	Supplier
CD2	PE-Cy7	S5.2	Becton Dickinson
CD3	BV510	UCHT1	Becton Dickinson
CD4	PE	RPA-T4	Becton Dickinson
CD5	BV711	UCHT2	Becton Dickinson
CD7	PE-CF594	M-T701	Becton Dickinson
CD8	APC-H7	SK1	Becton Dickinson
CD10	BV605	HI10a	Becton Dickinson
CD19	PE-Cy5	HIB19	Becton Dickinson
CD45	Alexa-700	HI30	Becton Dickinson
CD279	APC	MIH4	Becton Dickinson
CXCR5	BV421	RF8B2	Becton Dickinson

Appendix C Chapter 4: Supplementary Material

C.1 Supplementary Figures

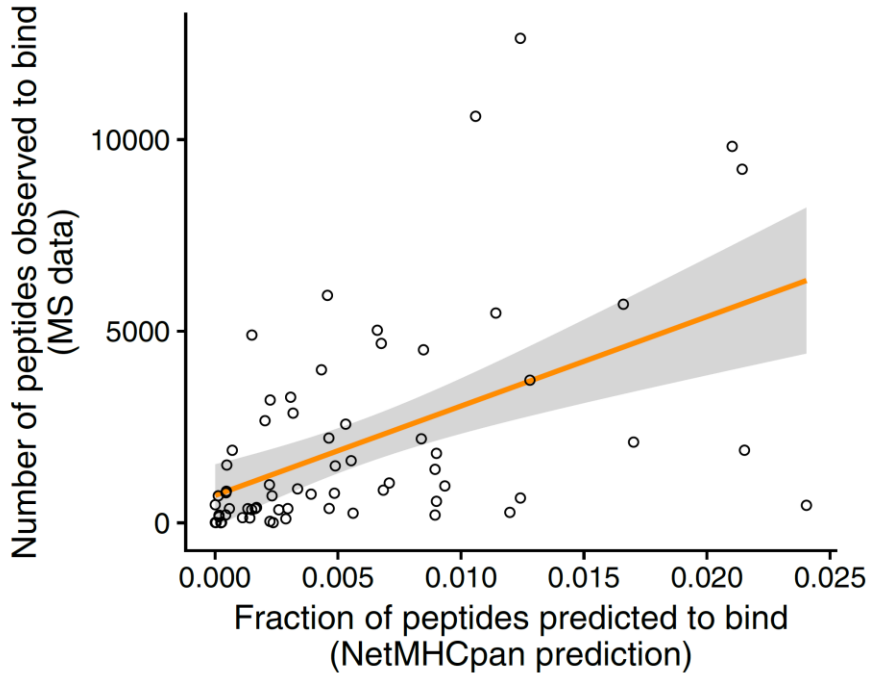

Figure C.1: Comparison of number of distinct peptides bound to 66 MHC. The fraction of self-peptides predicted to bind to each MHC is displayed on the x-axis, with the number of distinct peptides observed to bind in the SysteMHC MS data on the y-axis. The linear regression is displayed by the orange line with grey region showing the standard error.

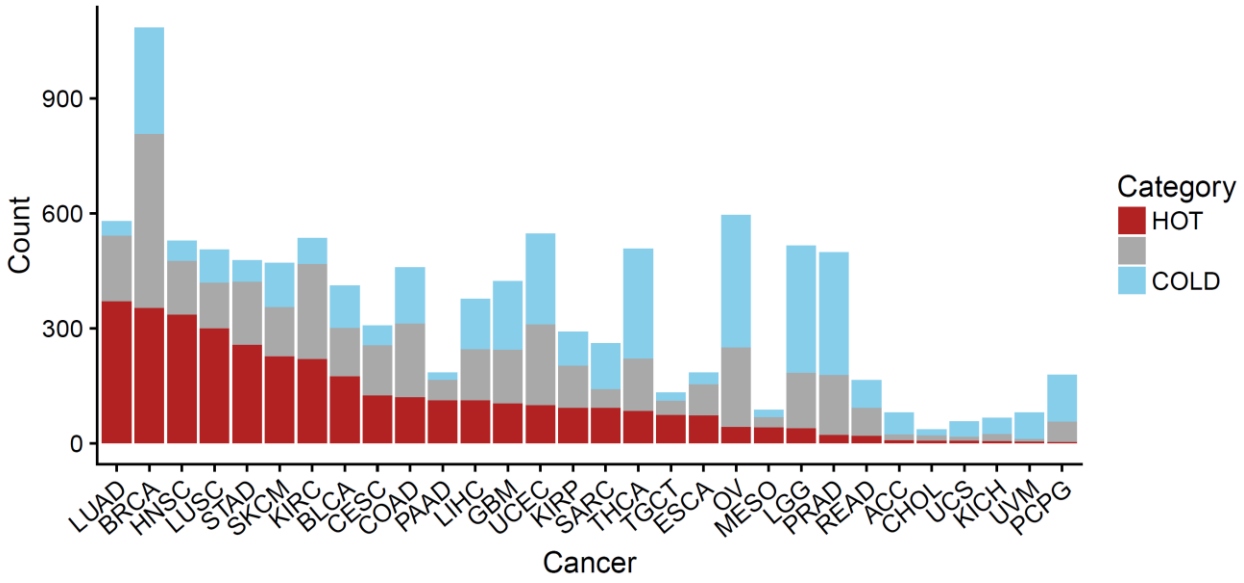

Figure C.2: Distribution of hot and cold tumours across TCGA. Within each tumour type, the number of tumours classified as hot (red), cold (blue), or neither (gray) are displayed.

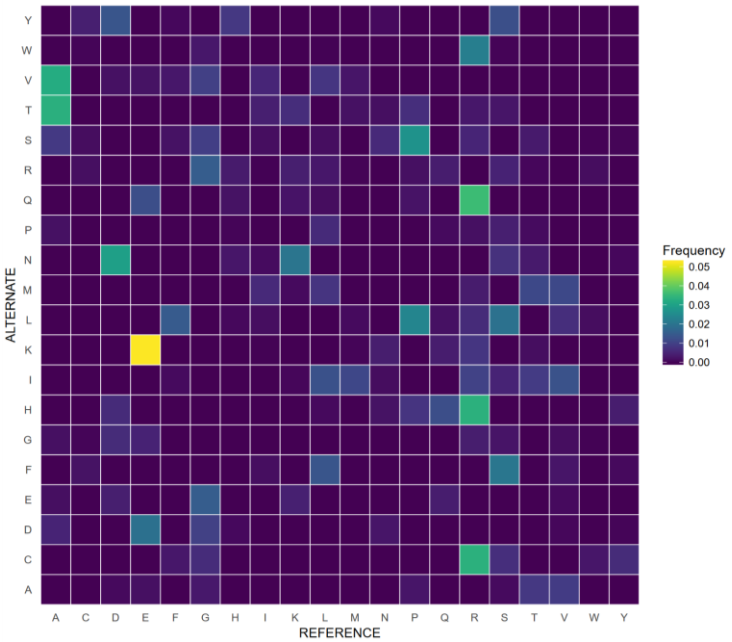

Figure C.3: Heatmap showing the observed frequencies of amino acid changes within the set of TCGA coding SNV mutations. The most common change is glutamic acid (E; reference amino acids on x-axis) to lysine (K, alternate amino acids on y-axis).

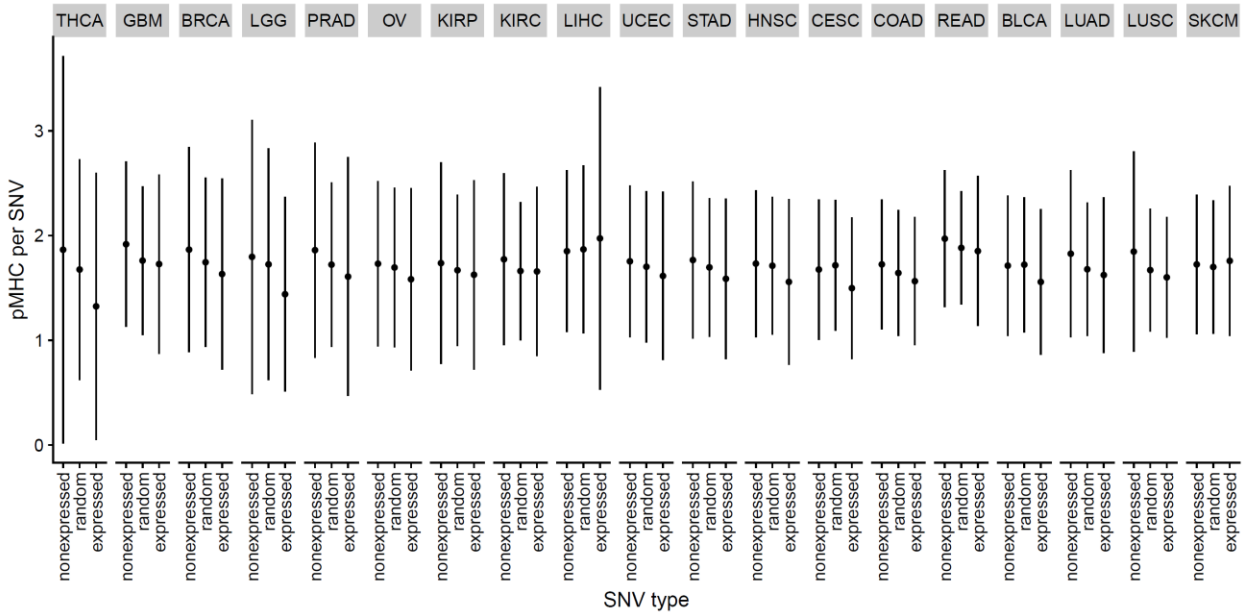

Figure C.4: Average immunogenicity of mutations for each tumour type. Average immunogenicity (pMHC per SNV; y-axis) per subject for non-expressed, random, and expressed SNVs (x-axis), split by cancer type. Points denote mean values and lines show ± one standard deviation.

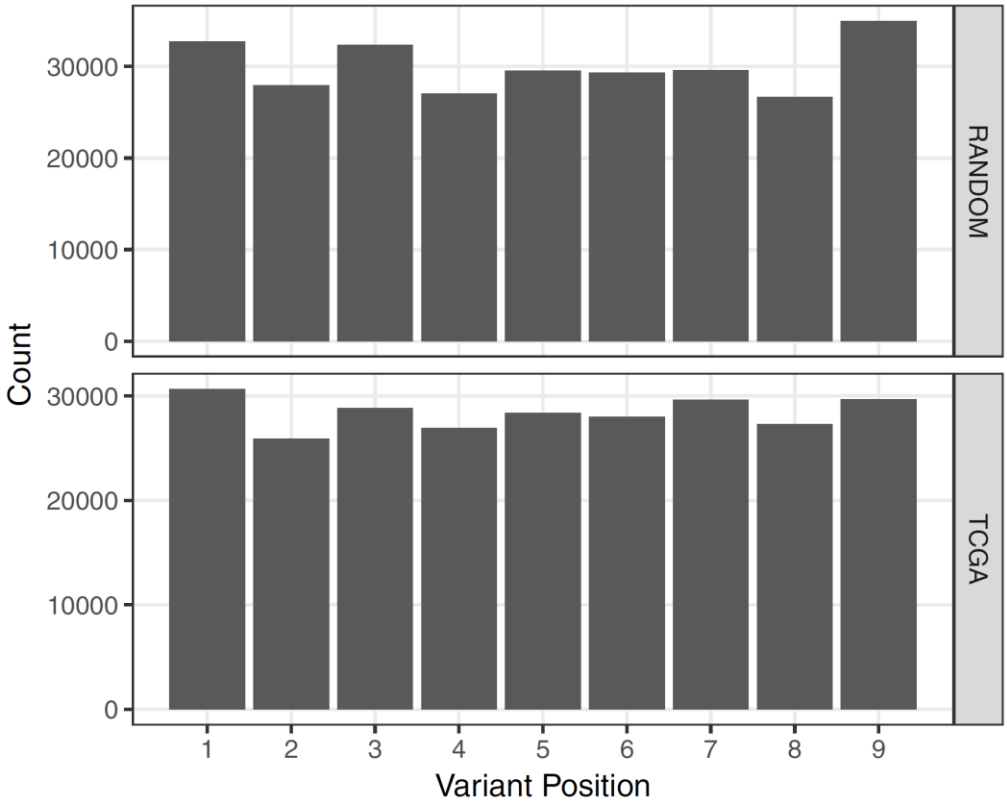

Figure C.5: Count of 9mer neoantigens containing the variant at each position, ignoring corresponding wildtype peptide binding status.

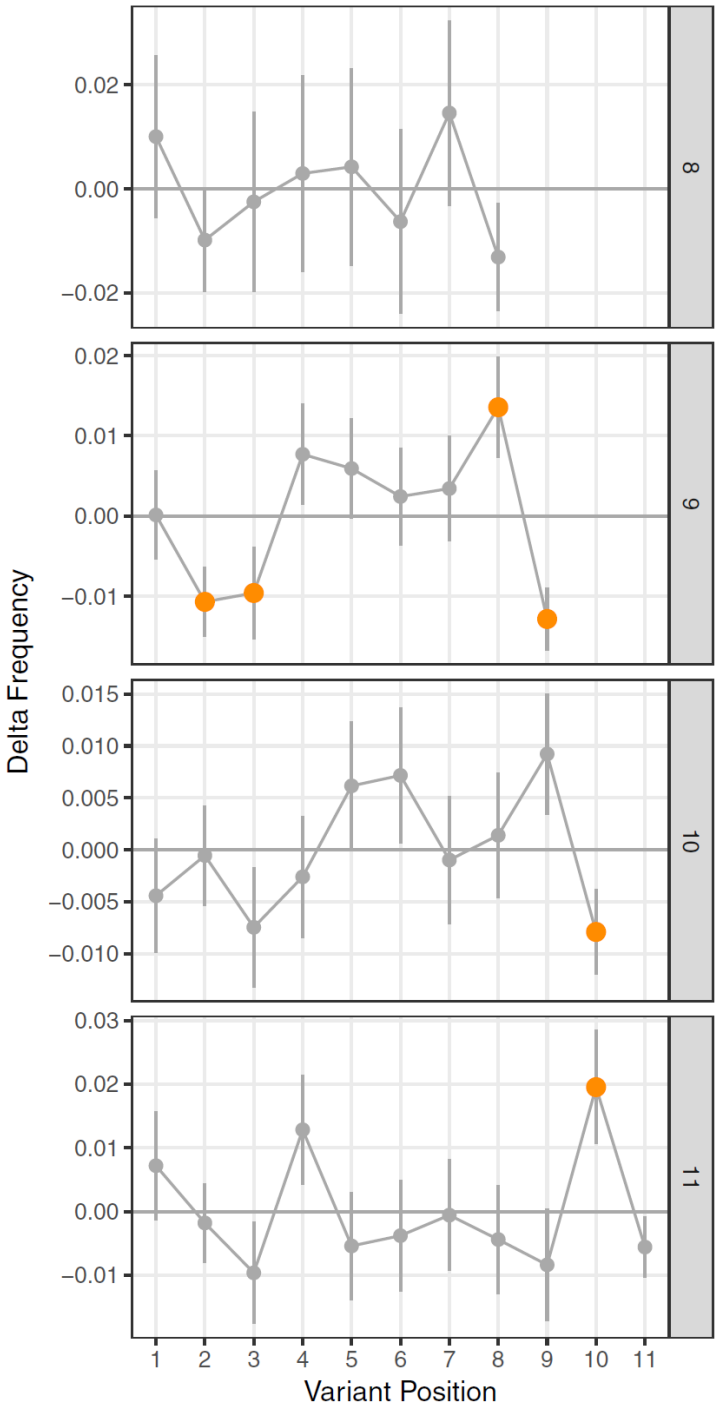

Figure C.6: Summary of variant position usage in presented peptides. Differences in frequency (y-axis) of the variant amino acid being in each position (x-axis) of a presented peptide for TCGA mutations compared to random mutations, for 8-, 9-, 10-, and 11-mers (panels from top to bottom). Mean values are shown (dots), with lines showing 95 % confidence intervals of the means. Positions with significant enrichment or depletion ($p_{adj} < 0.05$, T test) are coloured orange.

C.2 Supplementary Tables

Table C.1: Details of multivariate Cox-PH model predicting progression free intervals. Variable of interest: Self-immunopeptidome size. Bolded p-values < 0.05.

Variable	Units	Missing	Hazard Ratio	95 % CI	p-value
Race		22			
	White		1.00 (reference)		
	None		0.93	[0.76;1.14]	0.5083
	Black or African America		1.31	[1.10;1.56]	**0.00204**
	Asian		1.15	[0.90;1.47]	0.25138
	Native Hawaiian or other Pacific Islander		1.29	[0.41;4.07]	0.66446
Age at Dx	Year (*continuous)*	32	1.01	[1.00;1.01]	**< 0.001**
HLA Diversity		0			
	Heterozygous (6 distinct alleles)		1.00 (reference)		
	Homozygous		0.93	[0.82;1.04]	0.21412
Gender		0			
	Female		1.00 (reference)		
	Male		1.05	[0.93;1.18]	0.45971
Cancer type		0			
	BLCA		1.00 (reference)		
	BRCA		0.2	[0.15;0.27]	**< 0.001**
	CESC		0.36	[0.23;0.55]	**< 0.001**
	COAD		0.56	[0.41;0.77]	**< 0.001**
	GBM		4.86	[3.69;6.40]	**< 0.001**
	HNSC		0.76	[0.59;0.98]	**0.03706**
	KIRC		0.4	[0.30;0.54]	**< 0.001**
	KIRP		0.28	[0.18;0.43]	**< 0.001**
	LGG		0.95	[0.72;1.26]	0.71812
	LIHC		1.53	[1.14;2.05]	**0.00451**
	LUAD		0.86	[0.67;1.11]	0.2532
	LUSC		0.5	[0.37;0.67]	**< 0.001**
	OV		1.76	[1.33;2.33]	**< 0.001**
	PRAD		0.24	[0.17;0.36]	**< 0.001**
	READ		0.36	[0.21;0.62]	**< 0.001**
	SKCM		1.35	[0.92;2.00]	0.12881
	STAD		0.93	[0.48;1.78]	0.81954
	THCA		0.16	[0.11;0.23]	**< 0.001**
	UCEC		0.34	[0.26;0.46]	**< 0.001**
Self-immunopeptidome size	1,000,000 peptides (*continuous*)	0	0.92	[0.85;1.00]	0.05415

Table C.2: Details of multivariate Cox-PH model predicting progression free intervals. Variable of interest: Approximated SNV neoantigen load. Bolded p-values < 0.05.

Variable	Units	Missing	Hazard Ratio	95 % CI	p-value
Race		22			
	White		1.00 (reference)		
	None		0.94	[0.77;1.14]	0.51534
	Black or African America		1.28	[1.08;1.53]	**0.00442**
	Asian		1.16	[0.91;1.47]	0.23829
	Native Hawaiian or other Pacific Islander		1.37	[0.43;4.33]	0.59
Age at Dx	Year (*continuous)*	32	1.01	[1.00;1.01]	**< 0.001**
HLA Diversity		0			
	Heterozygous (6 distinct alleles)		1.00 (reference)		
	Homozygous		0.95	[0.84;1.06]	0.34986
Gender		0			
	Female		1.00 (reference)		
	Male		1.05	[0.93;1.18]	0.43945
Cancer type		0			
	BLCA		1.00 (reference)		
	BRCA		0.19	[0.14;0.25]	**< 0.001**
	CESC		0.35	[0.23;0.54]	**< 0.001**
	COAD		0.57	[0.42;0.79]	**< 0.001**
	GBM		4.63	[3.52;6.11]	**< 0.001**
	HNSC		0.74	[0.58;0.96]	**0.02234**
	KIRC		0.38	[0.28;0.52]	**< 0.001**
	KIRP		0.27	[0.18;0.41]	**< 0.001**
	LGG		0.91	[0.68;1.20]	0.48525
	LIHC		1.46	[1.09;1.96]	**0.01101**
	LUAD		0.85	[0.66;1.10]	0.22715
	LUSC		0.5	[0.37;0.66]	**< 0.001**
	OV		1.71	[1.29;2.26]	**< 0.001**
	PRAD		0.23	[0.16;0.34]	**< 0.001**
	READ		0.36	[0.21;0.62]	**< 0.001**
	SKCM		1.38	[0.93;2.04]	0.10784
	STAD		0.91	[0.47;1.74]	0.76608
	THCA		0.15	[0.10;0.22]	**< 0.001**
	UCEC		0.37	[0.28;0.50]	**< 0.001**
Approximated SNV neoantigen load	(SNV count × self-immunopeptidome size) / total unique peptides (*continuous*)	0	0.99	[0.99;1.00]	**0.00318**

Table C.3: Details of multivariate Cox-PH model predicting progression free intervals. Variable of interest: SNV neoantigen load. Bolded p-values < 0.05.

Variable	Units	Missing	Hazard Ratio	95 % CI	p-value
Race		22			
	White		1.00 (reference)		
	None		0.94	[0.77;1.14]	0.51291
	Black or African America		1.29	[1.08;1.53]	**0.00426**
	Asian		1.16	[0.91;1.48]	0.23209
	Native Hawaiian or other Pacific Islander		1.37	[0.43;4.32]	0.59223
Age at Dx	Year (*continuous)*	32	1.01	[1.00;1.01]	**< 0.001**
HLA Diversity		0			
	Heterozygous (6 distinct alleles)		1.00 (reference)		
	Homozygous		0.94	[0.84;1.06]	0.34054
Gender		0			
	Female		1.00 (reference)		
	Male		1.05	[0.93;1.18]	0.43756
Cancer type		0			
	BLCA		1.00 (reference)		
	BRCA		0.19	[0.14;0.26]	**< 0.001**
	CESC		0.35	[0.23;0.54]	**< 0.001**
	COAD		0.57	[0.42;0.79]	**< 0.001**
	GBM		4.67	[3.54;6.15]	**< 0.001**
	HNSC		0.75	[0.58;0.96]	**0.02458**
	KIRC		0.39	[0.29;0.52]	**< 0.001**
	KIRP		0.27	[0.18;0.42]	**< 0.001**
	LGG		0.91	[0.69;1.21]	0.52127
	LIHC		1.47	[1.10;1.97]	**0.00971**
	LUAD		0.86	[0.66;1.11]	0.23952
	LUSC		0.5	[0.37;0.67]	**< 0.001**
	OV		1.72	[1.30;2.28]	**< 0.001**
	PRAD		0.23	[0.16;0.35]	**< 0.001**
	READ		0.36	[0.21;0.62]	**< 0.001**
	SKCM		1.38	[0.94;2.04]	0.10383
	STAD		0.91	[0.47;1.75]	0.77538
	THCA		0.15	[0.10;0.22]	**< 0.001**
	UCEC		0.37	[0.28;0.50]	**< 0.001**
TCGA SNV neoantigen load	Count of mutant pMHC (*continuous*)	0	1	[1.00;1.00]	**0.0052**

Appendix D Evolutionary analysis of immunopeptidomes

Class I antigen presentation evolved to enable detection of intracellular pathogenic organisms. Within humans, there exists a selective pressure for MHC variants that can present pathogenic peptides [344], maintaining the high level of variability of MHC genes. At the same time, pathogens are under selective pressure to evade MHC-presentation, typically through dysregulation of the antigen presentation pathway [345,346]. I tested if evidence of this co-evolution could be detected by performing human MHC immunopeptidome predictions for different proteomes from a variety of sources.

D.1 Subsampling proteomes

Based on the computational and time resources required for the full human immunopeptidome predictive analysis (Chapter 4: Neoantigen characteristics in the context of the complete predicted MHC class I self-immunopeptidome), I determined that performing predictions for the complete proteomes of a set of different species would not be feasible, so tested if randomly subsampling the proteins in the proteome would yield sufficient information. I randomly selected proteins from the human reference proteome until 0.01, 0.05, 0.1, 0.5, 1, 2, 3, 5, 10, 25 and 50 % of the total amino acid count was reached. Five replicate subsamples were created at each depth, and the fraction of peptides presented by each allele was compared to the data from the non-subsampled dataset.

The fraction of input peptides predicted to be presented was calculated for each dataset, with each MHC having a measurement for the fraction of peptides presented in each dataset. For each subsampled dataset, the slope and correlation for fraction of peptides presented in the full dataset compared to the fraction of peptides presented in the subsampled dataset was calculated (Figure D.1). Based on this data, subsampling the proteome to 10 % of the total amino acid length was determined to be sufficient to reproduce the relationships between the number of peptides presented by different MHC molecules (mean slope of 1.00075, Pearson r = 0.9997), while significantly reducing the amount of computational time required to perform the predictions (approximately 1/10th). This subsampling depth was independent of proteome size

(subsampling tests were repeated on *Mycobacterium tuberculosis*, proteome size approximately 10 % that of human, slope = 1.00543 and r = 0.9997).

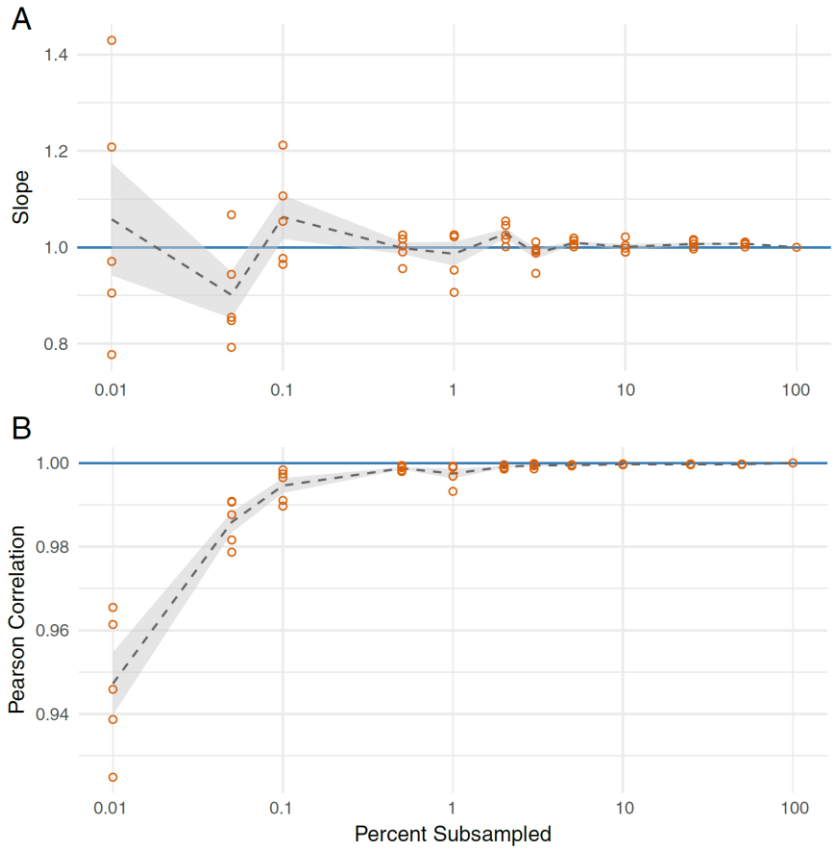

Figure D.1: Effect of subsampling the proteome on the fraction of peptides presented by each MHC. (A) The slope between the fraction of peptides presented for the full dataset and subsampled dataset was plotted for each of five replicates (orange circles) for each subsampling depth (x-axis). The black dotted line traces the average for each subsampling depth, with the grey shaded region marking ± 1 standard error on the mean. **(B)** Same as in (A) but showing the Pearson correlation coefficient instead of the slope.

D.2 Synthetic proteomes

As an initial comparison point, I wanted to determine how the fraction of peptides presented changed as the peptide sequences diverged from those found in the human proteome. To test this, I generated three synthetic proteomes: (1) reversed, (2) weighted, and (3) random. These were based on a 10 % subsample of the human proteome. For each protein in this human subsample (the source protein), (1) a protein was created by reversing the source protein (reversed), (2) a protein comprised of random amino acids selected based on their observed frequencies in the total canonical

human reference proteome was created to be the same length as the source protein (weighted), and (3) a protein comprised of random amino acids selected from a uniform distribution was created to be the same length as the source protein (random). These three datasets provided increasing divergence from human sequence. Binding of all peptides from these datasets to all class I MHC was predicted as in chapter 4.

I quantified the difference in the ability of two proteomes to be presented by comparing the fraction of all peptides presented by each MHC molecule for the two proteomes. By calculating the $\log_{10}$(ratio) between the fraction of a proteome's peptides that are presented by an MHC molecule to the fraction of the human proteome's peptides, we get a measure of how much better or worse the proteome is presented, with a $\log_{10}$(ratio) of zero having equal presentation, and a positive $\log_{10}$(ratio) meaning peptides from the proteome are, in general, presented better than the human proteome. When I performed this analysis to compare the human proteome with the reversed human proteome, I see a mean $\log_{10}$(ratio) of 0.00715 ± 0.037 (Figure D.2). Compared to the weighted proteome, the mean $\log_{10}$(ratio) is 0.0117 ± 0.058 (Figure D.2). Compared to the completely random proteome, the mean $\log_{10}$(ratio) is 0.161 ± 0.18 (Figure D.2). For these three synthetic proteomes, as the sequence becomes more and more diverged from the natural human proteome sequence, there is increasing levels of presentation by MHC molecules. This suggests that human MHC preferentially presents non-human peptide sequence – a useful property to be able to present non-self peptides from a nearly infinite space of possibilities.

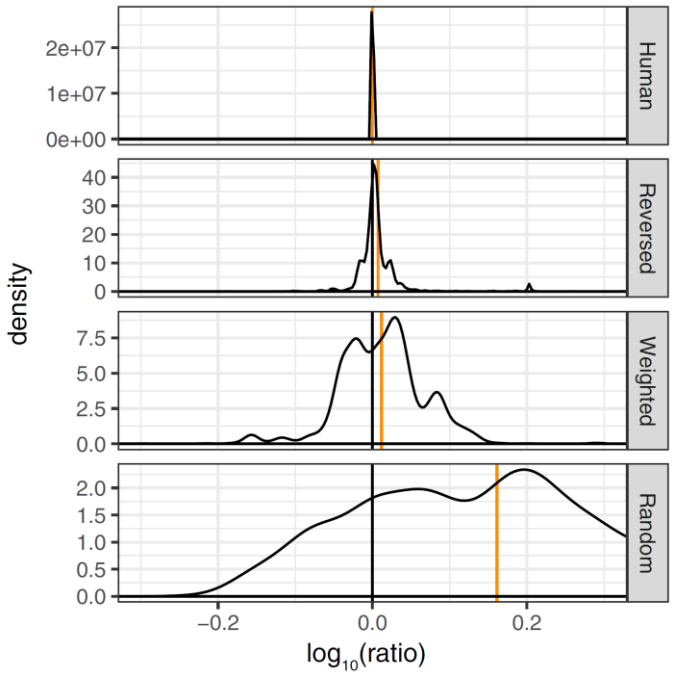

Figure D.2: Distribution of the $\log_{10}$(ratio) for synthetic human-like proteomes. For each MHC, the ratio between the fraction of peptides which are predicted to be presented from the indicated proteome compared to the fraction of peptides which are predicted to be presented from the human proteome was calculated, and the distribution of the $\log_{10}$(ratio) (x-axis) for all MHC molecules was plotted in density plots, with the y-axis density values set so the area under the curve is equal to one. A vertical orange line was drawn at the mean for each proteome. With increasing divergence from human sequence, the mean $\log_{10}$(ratio) becomes more positive.

D.3 Immunopeptidomes from vertebrate species evolutionarily diverged from humans

With the observation that protein sequences artificially diverged from humans had increased presentation, I next performed immunopeptidome predictions for a set of vertebrate species across the tree of life (Table D.1). The reference proteomes were downloaded from EMBL (ftp.ebi.ac.uk/pub/databases/reference_proteomes/QfO/), using the April 2016 Qf0 release. In general, presentation of proteomes from non-human vertebrate species was increased relative to human (Figure D.3), though there was no significant correlation between increase in presentation and evolutionary distance from humans (correlation between average $\log_{10}$(ratio) and millions of years to most recent common ancestor, Spearman's $\rho = 0.165$, $p = 0.557$). This provided evidence that

human MHC has a preference to present non-human peptides, consistent with the role of MHC in presenting foreign peptides to the immune system.

Table D.1: Vertebrate species used for testing evolutionary effect on immunopeptidome size.

OSCODE	Species	Name	Millions of Years Diverged† from Humans
HUMAN	*Homo sapiens*	Human	0
PANTR	*Pan troglodytes*	Chimpanzee	6.65
MACMU	*Macaca mulatta*	Rhesus macaque	29.44
RAT	*Rattus norvegicus*	Rat	90
MOUSE	*Mus musculus*	Mouse	90
CANLF	*Canis lupus*	Dog	96
BOVIN	*Bos taurus*	Bovine	96
MONDO	*Monodelphis domestica*	Gray short-tailed opossum	159
ORNAN	*Ornithorhynchus anatinus*	Duckbill platypus	177
CHICK	*Gallus gallus*	Chicken	312
XENTR	*Xenopus tropicalis*	Western clawed frog	352
TAKRU	*Takifugu rubripes*	Japanese pufferfish	435
DANRE	*Danio rerio*	Zebrafish	435
CIOIN	*Ciona intestinalis*	Transparent sea squirt	676
BRAFL	*Branchiostoma floridae*	Florida lancelet	684

† Divergence time estimates from http://www.timetree.org/

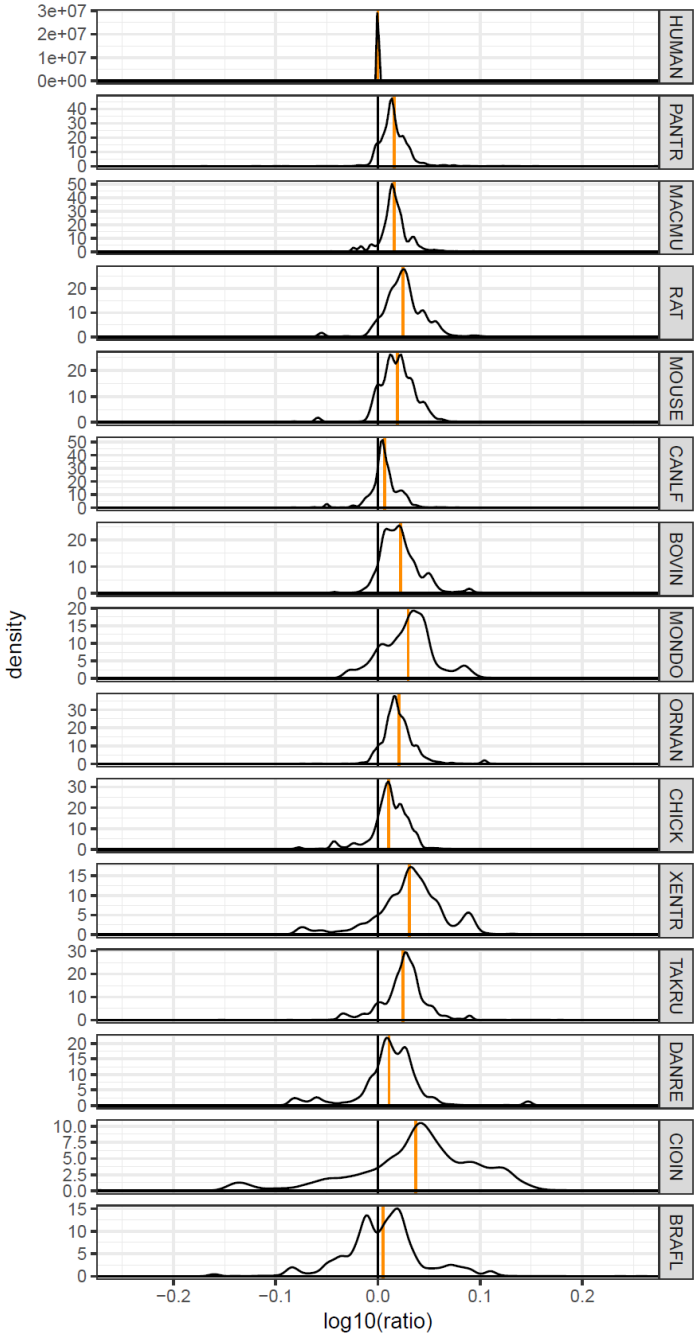

Figure D.3: Distribution of the $\log_{10}$(ratio) for different vertebrate proteomes. For each MHC, the ratio between the fraction of peptides which are predicted to be presented from the indicated proteome compared to the fraction of peptides which are predicted to be presented from the human proteome was calculated, and the distribution of the $\log_{10}$(ratio) (x-axis) for all MHC molecules was plotted in density plots, with the y-axis density values set so the area under the curve was equal to one. A vertical orange line was drawn at the mean for each proteome. With increasing divergence from human sequence, the mean $\log_{10}$(ratio) becomes more positive. Vertebrate species are ordered by increasing values of millions of years diverged from humans moving down the figure.

D.4 Immunopeptidomes from intra- and extra-cellular pathogens

Given that human MHC showed increased presentation of peptides for non-human vertebrate species, I next wanted to test if there was a difference in presentation of peptides derived from intra- or extra-cellular pathogens. Intra-cellular pathogens would exert a strong selective pressure on MHC to have strong presentation, whereas extra-cellular pathogens should not be exerting any such selective pressure.

Proteomes for eight pathogenic species were used from EMBL, four intra-cellular (*Leishmania major* (LEIMA), *Plasmodium falciparum* (PLAF7), *Chlamydia trachomatis* (CHLTR), and *Fusobacterium nucleatum* (FUSNN)), and four extra-cellular (*Giardia intestinalis* (GIAIC), *Leptospira interrogans* (LEPIN), *Escherichia coli* (ECOLI), and *Pseudomonas aeruginosa* (PSEAE)). In general, there was increased presentation of these pathogenic peptides compared to human (Figure D.4). Exceptions were for PLASM (intra-cellular; reduced presentation relative to human), and ECOLI (extra-cellular, no change relative to human). The three other extra-cellular pathogens, which are not expected to be under selective pressure, show increased presentation of peptides compared to human. The other three intra-cellular pathogens also show increased presentation of peptides compared to human, but don’t appear to be presented stronger than the extra-cellular pathogens. CHLTR has been reported to downregulate HLA expression during infection [345], which is a potential mechanism for it to evade immune detection despite having increased presentation of peptides. It is possible that similar mechanisms exist in LEIMA and FUSNN.

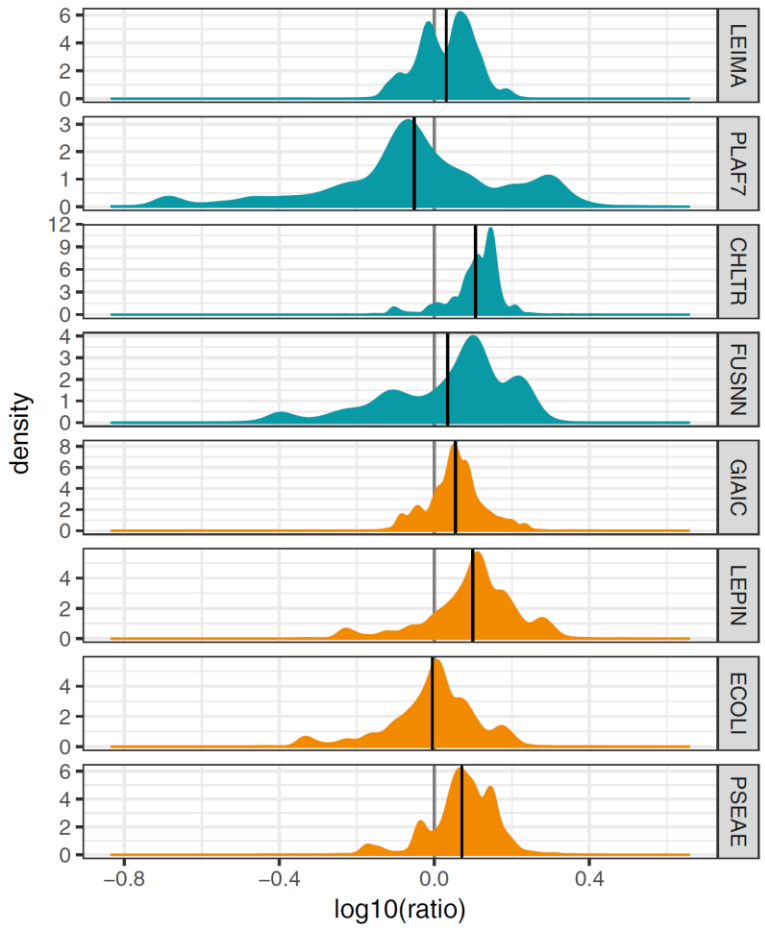

Figure D.4: Distribution of the $\log_{10}$(ratio) for intra- and extra-cellular pathogens. For each MHC, the ratio between the fraction of peptides which are predicted to be presented from the indicated proteome compared to the fraction of peptides which are predicted to be presented from the human proteome was calculated, and the distribution of the $\log_{10}$(ratio) (x-axis) for all MHC molecules was plotted in density plots, with the y-axis density values set so the area under the curve was equal to one. A vertical black line was drawn at the mean for each proteome. Intra-cellular pathogens are coloured blue (top four panels), while extra-cellular pathogens are coloured orange (bottom four panels). LEIMA; *Leishmania major*, PLAF7; *Plasmodium falciparum*, CHLTR; *Chlamydia trachomatis*, FUSNN; *Fusobacterium nucleatum*, GIAIC; *Giardia intestinalis*, LEPIN; *Leptospira interrogans*, ECOLI; *Escherichia coli*, PSEAE; *Pseudomonas aeruginosa*.

D.5 Immunopeptidomes from different species of *Plasmodium*

The comparison between intra- and extra-cellular pathogens may have limited ability to detect changes in presentation of peptides due to the heterogeneity of the different species. An alternative approach to identify signatures of co-evolution between pathogen and host is to compare different species of the same genus, where some

species infect humans and others do not. I performed this test on different species of *Plasmodium*, as there are numerous reference sequences, and information regarding the targets of each species. The first stage of the *Plasmodium* lifecycle which occurs in the vertebrate host, sporozoite invasion of hepatocytes in the liver, is the only time it is susceptible to T cell attack [347,348]. Once the *Plasmodium* merozites have infected the erythrocytes, which lack MHC, they are no longer visible to T cells [349]. Consequently, there is a limited window for the host immune system to clear *Plasmodium* infection, potentially resulting in selection of MHC which are able to efficiently present *Plasmodium* peptides. *Plasmodium* reference proteomes were downloaded from UniProt (https://www.uniprot.org/proteomes/?query=plasmodium).

In general, there is decreased presentation of *Plasmodium* peptides relative to human (Figure D.5), a trend which was not observed in any of the other species analysed. It is possible that this phenomenon is due to a long history of *Plasmodium* needing to evade immune-detection, sculpting the proteome to avoid peptides which can readily be presented. Of all the *Plasmodium* species analyzed, *P. falciparum* had the highest relative presentation of peptides by human MHC. As this is the deadliest species of *Plasmodium* in humans [350], this may be an effect of strong selection on human MHC for variants that are able to present peptides from this species.

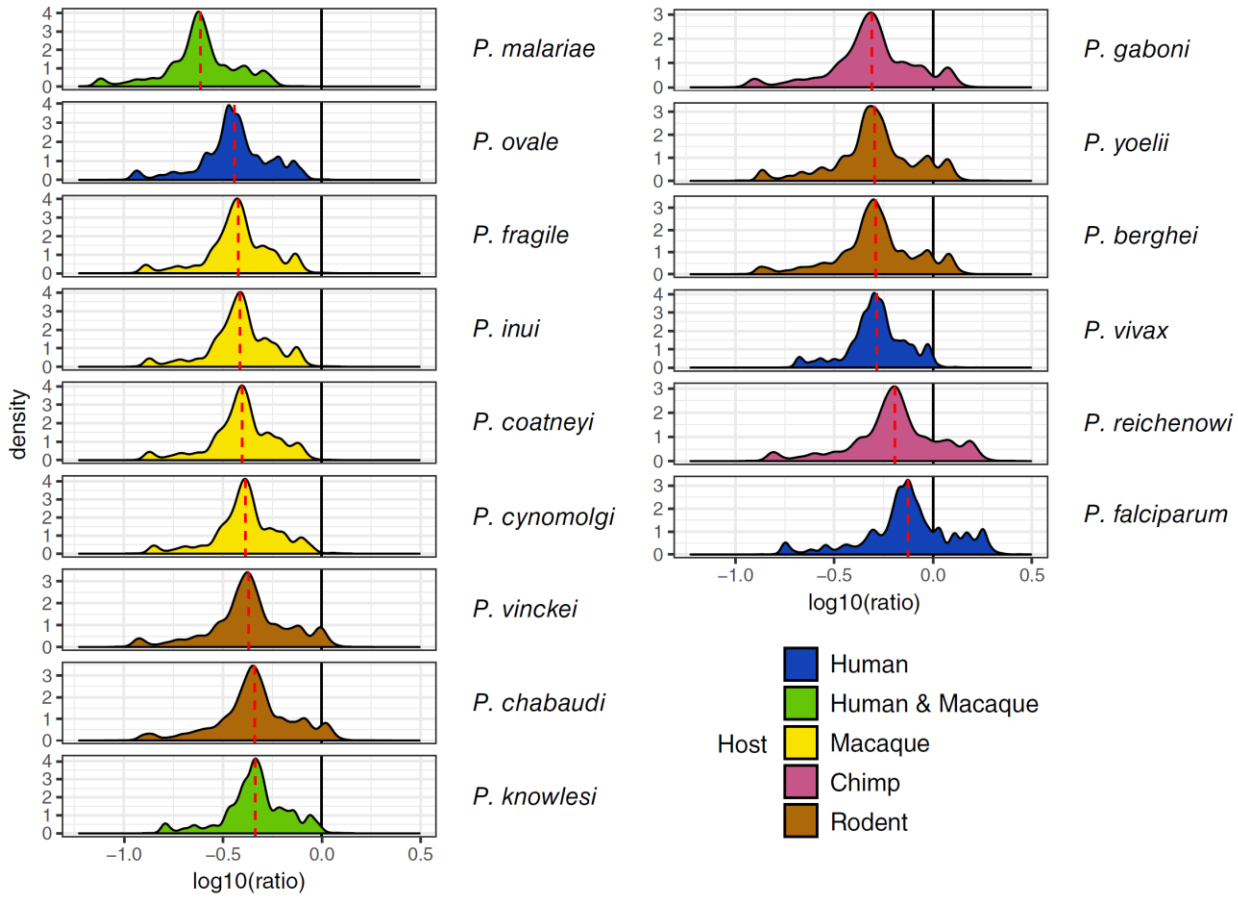

Figure D.5: Distribution of the $\log_{10}$(ratio) for species of *Plasmodium*. For each MHC, the ratio between the fraction of peptides which are predicted to be presented from the indicated proteome compared to the fraction of peptides which are predicted to be presented from the human proteome was calculated, and the distribution of the $\log_{10}$(ratio) (x-axis) for all MHC molecules is plotted in density plots, with the y-axis density values set so the area under the curve was equal to one. A vertical dotted red line was drawn at the mean for each proteome. Distributions are coloured according to their host organism. Panels are labelled with the species of *Plasmodium*. Species are ordered from lowest to highest relative presentation moving down the columns from left to right.

D.6 Discussion

In addition to the human immunopeptidome data presented in chapter 4, I have created immunopeptidomes for three human-like synthetic proteomes, fourteen vertebrate species, eight pathogens, and fifteen species of *Plasmodium*. This was made computationally tractable by determining that subsampling the proteome to 10 % of the total amino acid length is sufficient to measure the trends in MHC presentation.

Using the human immunopeptidome as the reference point, I showed that all three non-human synthetic proteomes have larger fractions of their peptides presented by human MHC than the human peptides. The reversed human proteome was the closest in sequence to the human proteome, consisting of the same amino acids in the opposite order. The weighted human proteome was the next closest, having amino acids randomly selected based on the observed frequency within the human reference proteome. Both of these synthetic proteomes had slightly better presentation compared to human. The random proteome was the most different in amino acid composition, and had the highest presentation. This suggests that human MHC are evolved to favour presentation of non-human peptides, consistent with their role in adaptive immunity.

Computing the presentation of proteomes from different vertebrate species, they all show increased presentation compared to human. The increase is not dependent on millions of years of divergence from humans. Since there would be no evolutionary selection for human MHC to present peptides from other vertebrates, this increase in presentation provides more evidence that human MHC has evolved to present non-human peptides.

There is no clear trend in the presentation of intra-cellular pathogens compared to extra-cellular pathogens. Given the strong selective pressure that pathogens exert on MHC variability, it would be expected that intra-cellular pathogens would have increased presentation compared to extra-cellular pathogens, where no such selective pressure exists. I do not see any evidence of that in this data, and in fact, the only organism to have decreased presentation relative to human, *Plasmodium*, is an intra-cellular pathogen. There are a couple of possible explanations for this. First, the intra- and extra-cellular pathogens selected for this test were pathogens with proteomes available from EMBL. A more careful selection of pathogens based on whether or not they are obligate intracellular pathogens may be informative. The lifecycle of obligate intracellular pathogens, like *Plasmodium*, are dependent on surviving within the host organism, and thus they will have a stronger selective pressure to co-exist with the host immune system. Second, stratifying pathogens based on the severity of the disease they cause may also provide a clearer distinction on presentation of pathogen-derived peptides.

Upon analyzing a variety of species of *Plasmodium* with different hosts, I show that generally, *Plasmodium* is not well presented by human MHC. For different species that infect the same host, presentation is similar. There are some clear outliers to this trend: species infecting humans range from the highest presentation observed for all *Plasmodium* species, to the lowest. Still, within this data, we see evidence of the pathogen-derived pressure on humans to increased presentation by MHC. The two most lethal forms of *Plasmodium*, *P. falciparum* and *P. vivax* [351], have relatively good presentation, whereas the less dangerous *P. malariae* (so-called “benign malaria”) and *P. ovale* [351] have lower presentation.