Computational Methods for Personalized Proteogenomic Characterizations of the HLAs

by

Michael Brodie Mumphrey

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2023

Doctoral Committee:

Assistant Professor Marcin Cieslik, Chair
Professor Alexey Nesvizhskii
Professor Gilbert Omenn
Professor Malini Raghavan
Professor Maureen Sartor

Michael Brodie Mumphrey

mumphrey@umich.edu

ORCID iD: 0000-0003-3150-8694

Dedication

This dissertation is dedicated to my family, who have always supported me in everything I do, my friends, who have always made life easier, and my cats, who are cats.

Acknowledgements

First, I would like to thank my advisor Dr. Marcin Cieslik for accepting me into his lab and providing constant support and guidance. I appreciate his encouragement to always think outside the box rather than stay inside established lanes, without which my dissertation project would probably look very different. I also am thankful for his commitment to making a fair and equitable working environment. I would also like to thank past and present members of the Cieslik lab, who are always available to bounce ideas off of, and who make the lab an enjoyable place to work.

Next I would like to thank the members of my dissertation committee Dr. Malini Raghavan, Dr. Alexey Nesvizhskii, Dr. Maureen Sartor, and Dr. Gil Omenn. They have provided valuable insight into my work, and much needed feedback as I prepare to complete my dissertation. I would particularly like to thank Dr. Malini Raghavan for her mentorship and collaborations as I delved into the nuances of HLA biology, and Dr. Alexey Nesvizhskii and members of his lab who mentored me as I learned how to develop proteomics algorithms.

I am also grateful for the funding I have received, both from the Genome Sciences Training Program under Dr. Mike Boehnke, and the Proteogenomics of Cancer Training Program under Dr. Alexey Nesvizhskii. These programs provided me not only with direct funding, but also with peer and faculty mentorship that enhanced my time at UM. I would also like to thank the Department of Computational Medicine and Bioinformatics, as well as their leadership and administrative staff, for their support throughout my time in the program.

Finally, I would like to thank all of my friends and family who have supported me through my time in graduate school. Whether it is organizing a game of Magic the Gathering, going on a kayaking trip, hanging out at a bar, or just making time to chat, it is the time spent outside of school that makes the time in school bearable. Thank you all.

Table of Contents

List of Tables

List of Figures

Abstract

Loss of HLA function is known to be a common immune escape mechanism across many cancers. Despite this, studies investigating the molecular mechanisms underpinning this loss have been hindered by the extreme genetic polymorphism of the HLA genes. Modern computational methods for high-throughput sequence-based analyses often rely on a standard reference, which is a problem for the HLA genes where most individuals will be highly divergent from the HLA sequences found in the standard reference. In this dissertation I present three methods that solve this problem by departing from the standard reference paradigm in favor of dynamic reference selection, where each individual will be analyzed relative to a personalized set of HLA reference sequences.

I first present Hapster, a genomics tool that uses DNA sequencing data to construct personalized genomic references for the HLA genes. I show that Hapster produces high quality reference sequences, and that somatic mutation calling relative to these references has higher sensitivity and specificity than existing methods. To demonstrate the utility of Hapster, I next applied it to 12,000 primary and metastatic cancers from the TCGA and MI-ONCOSEQ projects. I show that using Hapster, we were able to identify patterns of positive selection within squamous cell carcinomas, lymphomas, and cancers with microsatellite instability, and identify the likely mutational processes responsible for these mutations. I next present HLAProphet, a proteomics tool that uses known HLA types to provide personalized quantification of the HLA proteins. I show that HLAProphet's protein quantification has higher correlation to paired RNA expression data than existing methods, and that allele specific expression values reflect known

loss of function genomic events. Finally, I present MHConstruct, a graph-based genomics algorithm that produces personalized reconstructions of the entire 5 Mb MHC locus. I show that MHConstruct's personalized reference sequences have rates of germline variation below the genome-wide average of 1 snp/kb, enabling analyses of the polymorphic intergenic regions containing the promoters, enhancers, and other regions that regulate the HLA genes. In total, the tools presented here allow for a complete personalized proteogenomic characterization of the HLAs, enabling more thorough investigations of loss of HLA function as an immune escape mechanism for cancers.

Chapter 1 Introduction

1.1 Motivation

The human leukocyte antigens (HLAs) are a group of cell surface membrane bound proteins that are required for the proper function of T-cell immunity in humans. While T-cell surveillance is classically thought of as a defense against infection, in recent decades it has become clear that T-cells also play a major role in protecting against cancers. As a result, cancers are under enormous selective pressures to evolve to escape T-cell surveillance. It has become clear through immunohistochemistry (IHC) studies that loss of HLA expression is a common immune escape mechanism across many cancers. However, further studies into the molecular underpinnings of this loss are lacking due to the computational difficulties involved in studying the HLA genes. In this thesis, I will lay out three personalized computational approaches that allow us to overcome these challenges and investigate the molecular mechanisms underpinning loss of HLA function at the DNA, RNA, and protein levels.

1.2 HLA Biology

The HLAs are a family of cell surface transmembrane proteins that can be divided into four groups: classical class I HLAs, non-classical class I HLAs, class I HLA pseudogenes, and class II HLAs **(Table 1-1)**. Discussions of the HLAs often revolve around the classical class I HLAs (HLA-A, -B, -C), which are responsible for presenting peptide antigens to $CD8^+$ T-cells, and the class II HLA proteins (HLA-DP, -DQ, -DR), which are responsible for presenting peptide antigens to $CD4^+$ T-cells. The non-classical HLA proteins (HLA-E, -F, -G) each have

Table 1-1: Number of known HLA alleles

Number of known HLA alleles recorded in the IMGT/HLA[1] database version 3.52

	HLA Gene	Alleles	Proteins
Class I - Classical	A	7793	4548
	B	9274	5580
	C	7761	4311
Class I - Non-classical	E	347	140
	F	59	11
	G	117	38
Class I - Pseudogenes	H	67	0
	J	33	0
	K	6	0
	L	5	0
	N	5	0
	P	5	0
	S	7	0
	T	8	0
	U	5	0
	V	3	0
	W	11	0
	Y	3	0
Class II	DPA1	558	261
	DPB1	2332	1361
	DQA1	585	281
	DQB1	2439	1501
	DRA	46	5
	DRB	4419	2903
	DMA	58	9
	DMB	71	9
	DOA	92	14
	DOB	60	15

similar structure to the classical class I HLA proteins, but play alternative roles promoting immune tolerance, such as suppression of NK cells[2] and prevention of fetus rejection during pregnancy[3]. Similarly, the class II HLA proteins HLA-DM and HLA-DO have similar structure to the T-cell associated class II HLAs, but only play a role in MHC peptide loading[4,5]. The class I HLA pseudogenes have genetic similarity to the class I HLA genes but are not expressed. In this thesis, I will focus on the classical Class I HLAs (HLA-A, -B, -C) and the T-cell associated class II HLAs (HLA-DP, -DQ, -DR).

The class I and class II HLAs contribute to the formation of the major histocompatibility complex (MHC) in humans. The MHC is a critical component of T-cell immunity and is responsible for presenting intracellular peptide antigens to cytotoxic and helper T-cells. In this system, T-cells are tasked with identifying and eliminating any cells that express foreign proteins. However, T-cells are not able to directly detect their protein targets, as the majority of the protein content of any given cell is intracellular where T-cells do not have access. The solution to this problem is the MHC, which binds intracellular peptide fragments and presents them on the cell surface, where they can be detected by their cognate T-cells.

The class I and class II HLAs form mature MHCs with slightly different structures. MHC class I molecules consist of one membrane bound class I HLA protein in complex with the accessory protein beta-2-microglobulin (B2M). In this MHC, the binding pocket where the peptide antigen is loaded is entirely coded for by the class I HLA protein. In contrast, MHC class II molecules are composed of a dimer of class II HLA proteins. For each MHC class II there is an alpha and a beta protein that combine to make the mature MHC. In this MHC, both proteins are membrane bound, and each makes up half of the peptide binding pocket.

Loading of peptides onto the mature MHC is facilitated by a set of proteins known as the antigen processing machinery. This process begins with the normal turnover of proteins in the cell through either the ubiquitin-proteasome pathway for cytosolic proteins[6], or the lysosomal pathway for extracellular proteins[7]. MHC class I molecules are responsible for presenting intracellular peptides to $CD8^+$ T-cells and primarily present peptides originating from the ubiquitin-proteasome pathway[8]. In contrast, MHC class II molecules are responsible for presenting extracellular peptides to $CD4^+$ T-cells and primarily present peptides originating from the lysosomal pathway[9]. However, some crossover between pathways has been observed[10].

For MHC class I loading, cytosolic peptides originating from the ubiquitin-proteasome pathway are first translocated to the endoplasmic reticulum (ER) via the transporter associated with antigen processing (TAP), an ATP dependent transmembrane dimer composed of the two proteins TAP1 and TAP2[11]. TAP most efficiently transports peptide fragments of 8-12 amino acids in length[12], which is consistent with the canonical length of MHC class I peptide antigens[13]. Peptides that translocate to the ER but are larger than the canonical MHC class I length may be further trimmed by the ER peptidase ERAP1[14]. Appropriately sized peptide antigens can then be loaded onto an MHC class I molecule, which in its unloaded state is stabilized by the peptide-loading complex (PLC). The PLC consists of the general chaperone calreticulin[15], the TAP-associated protein tapasin which brings the MHC/PLC complex into proximity with TAP[15], and the thiol reductase ERp57 which plays a role in exchanging low affinity peptides for high affinity peptides within the MHC class I binding groove[16]. Once an MHC class I molecule has been loaded with a high affinity peptide, it dissociates from the PLC and is transported to the cell surface.

Loading of peptides onto MHC class II peptides follows an entirely separate pathway. Initially, MHC class II molecules are assembled in the ER and associate with the invariant chain Ii, which is cleaved by proteases into the shorter peptide fragment CLIP. CLIP fills the peptide binding groove and stabilizes the class II MHC before it is loaded[17]. Once in the endosomal compartment, displacement of the low affinity CLIP peptide is facilitated by HLA-DM and HLA-DO[4,5]. After the opening of the peptide binding groove by HLA-DM/DO, peptides with higher affinity than CLIP can be loaded into the mature MHC class II. Like the class I MHC molecules, after loading, the MHC class II is exported to the cell surface where it can present its bound antigen to the outside environment.

After export to the cell surface, MHC molecules and their bound peptide antigens can stimulate T-cells through the αβ T-cell receptor (αβTCR). αβTCRs of mature T-cells are generally specific to a single MHC:antigen complex. This high specificity is the result of stringent selection that begins with the creation of a diverse collection of αβTCRs via somatic re-arrangement of the variable (V), diversity (D), and joining (J) regions of the *TRA* and *TRB* genes in a process known as VDJ recombination[18]. This population of T-cells then undergoes both positive and negative selection in the thymus to promote clonal expansion of T-cells with high affinity to MHC-antigen complexes, while culling those T-cells that exhibit auto-reactivity[19]. Clonally expanded T-cells then translocate to the tumor microenvironment where they can come into contact with tumor cells and antigen presenting cells expressing their cognate MHC:antigen complex.

1.3 Immunogenicity of tumors

The first discovered link between the HLAs and cancers was found alongside the initial discovery of histocompatibility itself. In a series of papers published in the early 1900s

(reviewed in[20]), multiple groups simultaneously discovered that spontaneous tumors grown in mice could be transplanted successfully between inbred mice of the same stock, but when sent to partner labs could not be successfully grown in genetically distinct mouse stocks. This led to the proposal of a genetic factor responsible for mediating susceptibility or resistance to tumor transplantation[20]. In the 30s and 40s, the mouse factor responsible for this was isolated[21,22] and was named the histocompatibility factor H-2. In the 50s, the human factors were isolated and named the human leukocyte antigens, or HLAs[23]. Ultimately, it was determined that these proteins make up the MHC, and these discoveries led to the 1980 Nobel prize in medicine and physiology being awarded to Baruj Benacerraf, Jean Dausset, and George Snell.

While there was initially great interest in the immune response against transplanted tumors, this was ultimately considered a matter of tissue histocompatibility, similar to the observed rejections of other transplanted tissues. Hypotheses involving actual immune surveillance specific to malignant cells were proposed as early as 1957[24], however they failed to gain traction due to a lack of empirical evidence and clear mechanisms of action. This changed in the early 90s with direct evidence of a tumor antigen presented to T-cells via the MHC[25], as well as evidence that mice lacking an effective cytotoxic T-cell response were more likely to develop malignant lesions[26–28]. Further, meta-analyses of medical records collected over the previous decades revealed that immunocompromised individuals were more likely to develop some form of cancer over their lifetime[29].

Excitement surrounding potential anti-tumor T-cell responses was further invigorated by the discovery of the immune checkpoint inhibitors CTLA-4[30] and PD-1[31]. It was observed not only that these proteins prevent effective T-cell responses, but that tumors often overexpress them, and their blockade can directly enhance anti-tumor immunity[31,32]. These discoveries led to

the hypothesis that cancers must evolve to escape the immune system, and that there may be effective therapies that directly re-establish immune surveillance. In the United States, there are now at least eight FDA approved immune checkpoint inhibitors targeting CTLA-4[33], PD-1[34–36]/PD-L1[37–39], and LAG-3[40], with dozens more potential targets under investigation[41]. However, overexpression of immune checkpoints is only one potential immune escape mechanism, and there exist a number of other points in the cancer immunity cycle that tumors can evolve to disrupt[42].

1.4 Evasion of T-cell immunity via loss of HLA expression

In this thesis, I will focus on evasion of the T-cell response via loss of class I and class II HLA expression. Loss of the class I HLAs is the most studied phenotype and allows malignant cells to cease presentation of tumor antigens to $CD8^+$ T-cells. It has been repeatedly observed in most cancer types via immunohistochemistry (IHC) studies[43] and can occur via multiple mechanisms. Complete loss of class I HLA expression can occur via homozygous deletion of the MHC locus itself or hypermethylation of the MHC locus[44], and complete loss of cell-surface MHC class I molecules can occur through the loss of B2M or key proteins in the APM[45,46]. Haplotype loss, leading to loss of about half of the peptide presentation repertoire, has been observed to frequently occur due to loss of heterozygosity (LOH) of the MHC locus[47]. Allelic loss, resulting in the loss of one of the six overall class I HLA alleles, is the most difficult to detect via IHC, but has been observed via monomorphic monoclonal antibodies[48]. This phenotype is of particular interest, as it potentially allows for the loss of the singular HLA allele responsible for presenting an immunodominant antigen[49], while maintaining expression of the remaining HLA alleles to prevent targeting by NK cells, which act to eliminate cells that fail to express MHC molecules[50].

Class II HLA loss is not well studied as it has only recently been recognized that tumor specific MHC class II expression may also play an anti-tumor role. Traditionally, the MHC class IIs were only thought to be expressed on professional antigen presenting cells (APCs), which includes dendritic cells[51], macrophages[52], and B-cells[53]. These cells take up proteins from the extracellular environment, process them into antigenic peptides, and present them to $CD4^+$ T-cells via the MHC class II. Importantly, even though this pathway promotes an anti-tumor T-cell response and therefore could be considered a tumor suppressor, the class II HLA genes would not be expected to be under evolutionary pressure for loss of function due to the fact that their canonical function is a result of expression in healthy non-tumor cells. However, it has since been observed that during an active immune response, any cell type can be stimulated to express the MHC class II and become an APC[54]. This includes the observation that cancer cells themselves can gain MHC class II expression[55–57], and the development of MHC class II restricted neoantigen vaccines that exploit this phenomenon[58]. Given this, I hypothesize that in tumors that gain class II HLA expression, there may be selective pressures to then lose it. However, detection of this loss will be more difficult as a distinction will have to be made between tumors that lost class II HLA expression versus those that never had it.

Consistent with predictions, loss of HLA expression is associated with worse clinical outcomes across a range of cancers[59–63]. Loss of class I HLA expression is often associated with lower levels of tumor infiltrating lymphocytes (TILs)[64], which are generally considered a positive prognostic indicator[65–67]. HLA expression has also been shown to be associated with improved outcomes for T-cell based immunotherapies, including one striking melanoma case study where HLA-high metastases regressed while HLA-negative metastases progressed following combination interferon/M-Vax immunotherapy treatments[68]. However, despite the

clear evidence from IHC, PCR, and immunoblot assays linking HLA function to patient outcomes, the progression to larger high throughput sequencing based studies has been slow due to the unique genetic complexity of the HLA genes.

1.5 HLA polymorphism

The class I and class II HLA genes stand out as being among the most polymorphic genes in the human genome, with some genes having hundreds to thousands of known alleles **(Table 1-1)**. This diversity has been shown to be the result of strong balancing selection, a process that maintains a larger amount of low frequency alleles than expected due to random drift. Balancing selection of the HLA genes is thought to be caused by a number of factors, including general heterozygote advantage for a more diverse antigen presentation repertoire[69], and distinct pathogen resistance in an environment where the specific pathogens shift over time[70]. In either case, it represents a form of co-evolution between hosts and pathogens, and produces extraordinary diversity at the MHC locus.

HLA diversity is a particularly large problem for current high-throughput molecular methods. These methods frequently begin with the sequencing of small DNA, RNA, or protein molecules. These molecules are then mapped back to the genome/transcriptome/proteome by comparing their sequence to a database of known sequences known as a standard reference. Individual genes or intergenic regions can then be interrogated for mutations, differential expression, post-translational modifications, etc., based on the information obtained from the sequenced molecules. The issue that arises for the HLA genes is that current standard references often contain only one or a few HLA sequences, resulting in the failure to correctly identify the origins of sequenced HLA molecules.

In DNA or RNA sequencing studies, this can cause either misalignment or failure of alignment for many sequencing reads. One common occurrence we observe in our own data is for HLA-A alleles that have higher sequence similarity to the GRCh38 sequence for the pseudogene HLA-J than for HLA-A, causing the reads to align to the incorrect region of the genome. This can result in both failure to identify mutations in HLA-A, and in incorrect measures of gene expression. In proteomics studies, this can result in either the failure to identify tryptic peptides or the assignment of peptides to the wrong HLA protein, resulting in incorrect calculations of protein expression. Finally, given that some individuals will have HLA sequences that match standard reference sequences by random chance, these individuals will have fewer errors, causing a bias in analyses towards standard-reference matching individuals.

To address this issue, it may seem logical to simply include all known HLA alleles in the standard reference used for a given analysis. While this may help in identifying more molecules, most computational tools are not designed to handle many alternative sequences for an individual gene at once. For example, a version of GRCh38 is available that contains alternative contigs of the MHC locus to allow for better alignment of HLA sequencing reads[71]. However, while this does allow more reads to align, downstream analysis is prevented as the reads originating from one gene may be spread over more than a dozen separate alternate contigs. In this way, providing more HLA reference sequences acts as a decoy to prevent misalignment, but does not enable interrogation of the genes themselves.

A proposed solution to this issue is to discard the standard reference in favor of dynamic reference selection[72]. In this paradigm, each individual that is sequenced will be given a personalized set of reference sequences that matches their HLA type. These custom references can then be used with existing validated computational tools for DNA/RNA/proteomics analysis

to provide mutation calls or expression quantification, and integration of data resulting from different reference sequences can be handled separately. In this thesis, I will present three tools that enable the personalized analysis of the HLA genes using dynamic reference selection. In chapter 2 I will introduce Hapster which identifies HLA types and constructs personalized genomics reference sequences for use in germline and somatic mutation calling. In chapter 4 I will introduce HLAProphet, a proteomics tool which uses known HLA types to construct a personalized protein sequence database for use with FragPipe for allele-level HLA protein quantification. Finally, in chapter 5 I will introduce MHConstruct, a tool that allows for a personalized reconstruction of the entire MHC locus to allow for the study of intergenic regulatory regions near the HLA genes. In total, I show that the concept of dynamic reference selection extends to many data types, and can enable the efficient analysis of the HLA genes.

Chapter 2 Personalized Somatic Mutation Calling With Hapster

Portions of this chapter are available as a preprint[73], and are under review at *Cell Reports.*

2.1 Introduction

The current paradigm for mutation calling involves the alignment of DNA sequencing reads to a standard reference genome, followed by identification of variants relative to that reference. This approach fails for the MHC genes where high sequence divergence from the standard reference, as well as the presence of homologous pseudogenes, frequently causes reads to fail to align appropriately.

To address this problem, we have developed Hapster. Hapster models alignment of DNA sequencing reads to alternative haplotypes as a linear system that can be solved to identify the closest underlying haplotypes **(Figure 2-1)**, a method which generalizes to other polymorphic genes. We also use a curated blacklist of homologous genes and pseudogenes to remove reads that erroneously cross-map to our genes of interest. Given appropriate haplotype references and filtered reads, Hapster builds upon state-of-the-art aligners and mutation callers (BWA-MEM[74] and Mutect2[75]) to provide accurate mutation calls. As a final step, we use an alignment-free kmer search to flag variants that may have been called due to remaining erroneous alignments or sample contamination.

The Hapster reference selection algorithm, in brief, is as follows. First, reads are simulated from each sequence within a database of known MHC alleles. Reads are then aligned simultaneously to all other known sequences. An all vs all matrix A is constructed where each

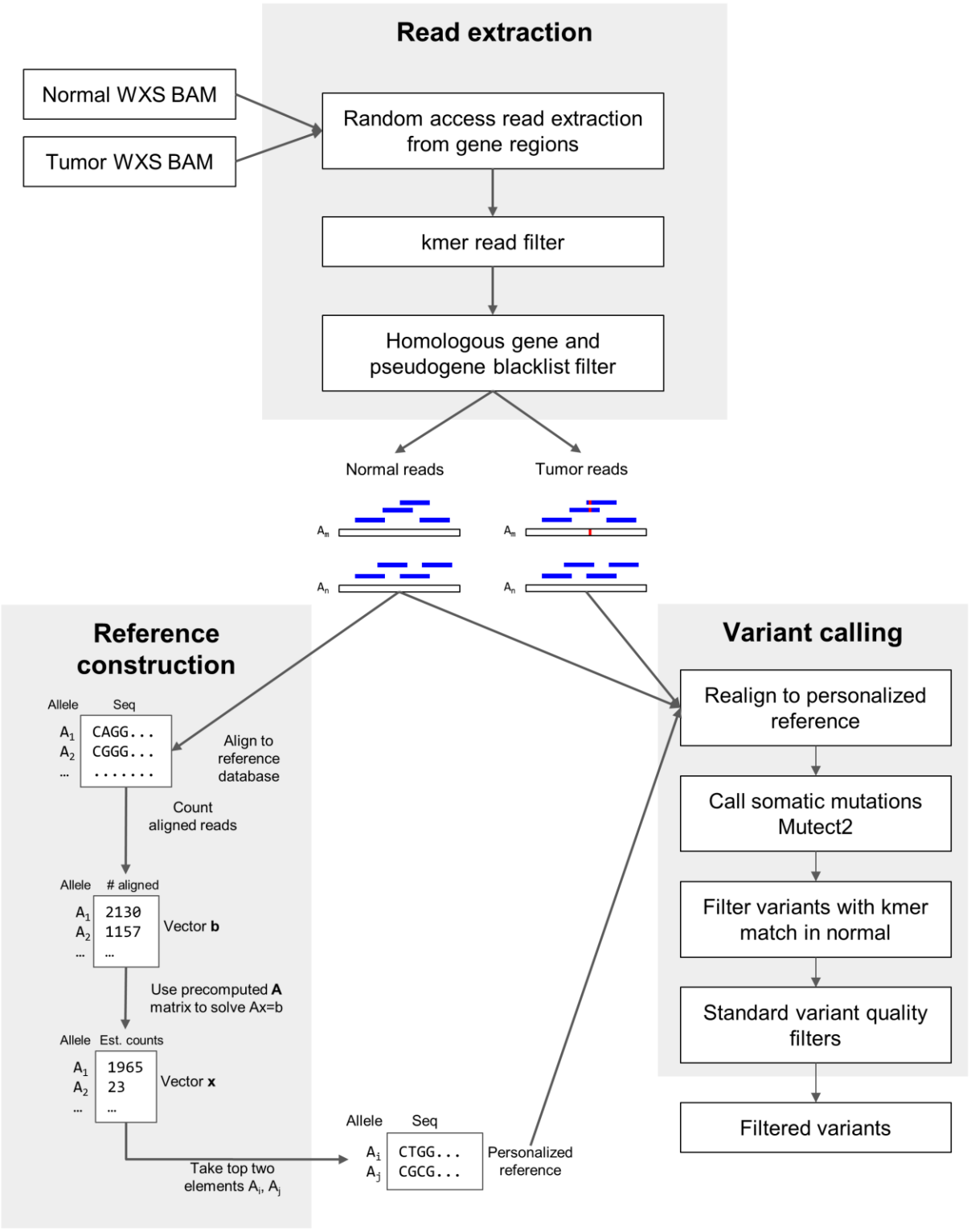

Figure 2-1: Schematic overview of Hapster's mutation calling pipeline

entry describes the percentage of reads simulated from one allele that align well to another allele. Following sequencing, MHC reads are extracted and simultaneously aligned to all alleles within the MHC sequence database. Observed counts of reads that align to each allele are put into a vector b. The system Ax = b is solved for x, where A describes alternate-sequence aware alignment, x describes the number of reads originating from each known allele, and b describes the number of reads that are aligned to each known allele. After solving, vector x should contain all values of ~0 except for the one or two (depending on whether the individual is homozygous or heterozygous at the locus) non-zero entries representing alleles that were actually present in the germline and therefore generated sequencing reads. These two non-zero entries are taken as the most likely diploid haplotype for the individual, and their reference sequences are used for downstream alignment and mutation calling. Personalized references provide improved alignments, and allow for recovery of missed mutation calls relative to the standard reference approach **(Figure 2-2)**.

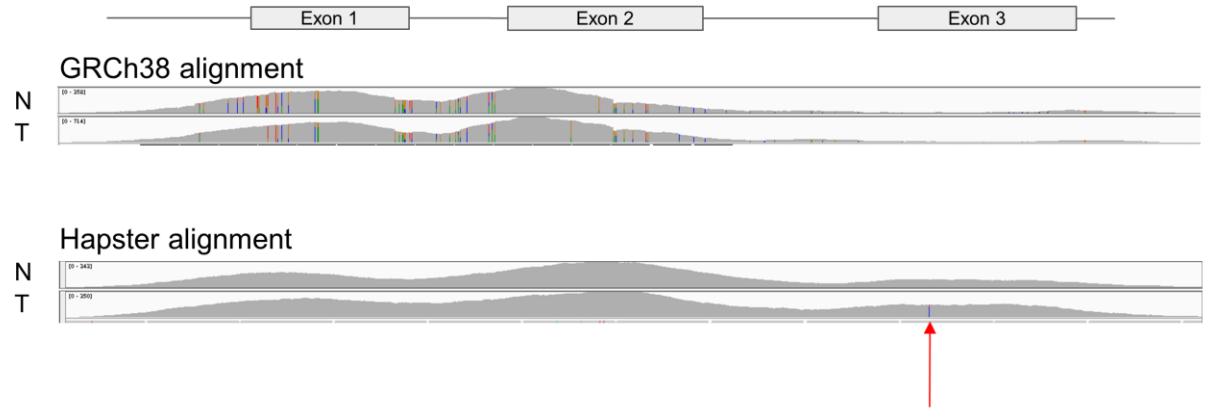

Figure 2-2: Recovery of somatic mutations missed by Grch38 reference based pipelines
IGV view of sequencing reads from paired normal and tumor whole exome sequencing data aligning to the first 3 exons of HLA-A using the standard reference GRCh38 (top) or Hapster (bottom). Arrow shows a somatic variant missed by the standard reference approach due to a failure to align reads from exon 3 to the correct location. N: Normal, T: Tumor.

2.2 Detailed overview of the Hapster algorithm

2.2.1 Alt-aware reference construction

Hapster leverages BWA mem's alt-aware alignment mode for its haplotyping, and as such needs a genomic reference that contains alternate sequences for the genes of interest. To construct an alt-aware reference for the HLA class I and class II genes we first retrieved sequences from the IMGT/HLA[1] database [IMGT/HLA release 3.51.0, https://github.com/ANHIG/IMGTHLA/tree/3510] to create a set of alternate alleles for each gene. The sequences for each alternate allele were appended to the primary assembly of Grch38 as independent contigs, and an alt index file was created by performing long read alignment of each sequence to Grch38 using minimap2[76].

2.2.2 HLA read kmer extraction

To efficiently extract true HLA reads, we created a 4 step procedure: (1) Random access retrieval of reads from all HLA regions in Grch38, all unaligned reads, and additional locations throughout Grch38 where we have found HLA reads mapping erroneously. (2) Passing reads retrieved by random access through a kmer filter, keeping any read that contains any 30-mer found in our set of alternate HLA allele sequences (3) Alt-aware alignment of kmer extracted reads to the previously created reference containing our set of alternate HLA alleles, keeping only reads that have at least one alignment to an HLA contig (4) Alignment of remaining reads to a reference containing both the sequences for our genes of interest (whitelist), as well as the sequences for any homologous genes/pseudogenes that are not of interest (blacklist), and keeping only reads that preferentially align to the whitelist.

2.2.3 HLA haplotype inference

For each HLA gene, extracted reads are aligned using BWA-MEM's alt-aware alignment mode to the constructed reference containing our set of alternate HLA sequences. The number of read pairs aligning to each allele with a total NM score of less than or equal to 1 is counted and put into vector **b**. Using vector b and the precomputed probability matrix **A** (construction of **A** matrix described below), the denoised read vector **x** is calculated by solving the equation **Ax=b**. To reduce problems caused by highly correlated alleles, a dynamic stepwise variable selection process is used to eliminate variables from **A** as follows: (1) identify alleles with pairwise correlations above a specific threshold (parameter tuning for this threshold described below) (2) For each highly correlated pair of alleles, find the magnitude difference between the alleles within vector **x** (3) Remove the most negative allele within the pair with the highest magnitude difference from both matrix **A** and vector **b** (4) Solve for **x** using the pared **A** and **b** (5) Repeat until there are no more pairs of alleles with correlations above the predetermined correlation cutoff. Once all highly correlated alleles are removed, the two highest value alleles in vector **x** are assigned to the individual's haplotype. The top two alleles are always chosen due to our observation that during alignment of reads to our inferred haplotype, a homozygous individual's reads will preferentially align to the single allele that is from their true haplotype. The presence of an extra allele's sequence in the reference file in the homozygous case therefore does not affect alignment or mutation calling of reads aligning to the true allele.

2.2.4 Construction of probability matrix A

Insert size metrics for a given sequencing protocol are calculated using the Picard tools command CollectInsertSizeMetrics. Using the read length, mean insert size, and standard deviation for the protocol, a random set of read pairs is simulated from the genomic sequence of each allele within a given HLA gene as follows: (1) Inserts of mean insert size ± 2 standard

deviations are simulated using BBMap's randomreads.sh (2) To simulate exome capture, read pairs derived from each insert are only kept if they have significant overlap with a provided set of capture probe sequences. The simulated captured reads are aligned in an alt-aware manner to our reference containing all alternate HLA alleles and are then processed and counted to create a vector b as described in the methods for haplotype inference. The vector **b** is then divided by the total number of read pairs that were simulated to obtain a vector that reflects the probability that any randomly selected read derived from the simulated allele would align to each other allele when using alt-aware alignment. This probability vector is calculated for each allele, and the probability matrix **A** is constructed using each of these probability vectors as its columns. The simulation process leads to a slightly non-symmetrical matrix, so the matrix was made to be symmetric by taking the mean of each paired off-diagonal term. The final matrix **A** is an all-vs-all matrix where each element represents the probability of reads generated from one allele aligning to another allele.

2.2.5 Parameter tuning

Parameter tuning must be performed to find the optimal correlation cutoff values for stepwise variable selection for every new sequencing experiment. Parameter tuning with HapMap samples was performed to find the optimal correlation cutoffs to minimize Levenshtein distances across all genes. For many experiments this will not be possible as ground truth haplotypes will be unavailable, making parameter tuning with Levenshtein distance impossible. However, we reasoned that every mismatch that contributes to the true Levenshtein distance could be recognizable as a germline mutation relative to the inferred reference sequences. To test this, we realigned sequences from our validation set to Hapster's inferred haplotype sequences and called germline mutations using HaplotypeCaller. We found strong correlations between the

true Levenshtein distance and the number of germline mutations called **(Figure 2-3)** suggesting that parameter tuning can be performed using the number of germline mutations called as a proxy for Levenshtein distance. For parameter tuning for TCGA and MI-ONCOSEQ samples, all correlation cutoff values between 0.70-0.99 were tested to find the optimal threshold per gene that minimizes observed germline variants.

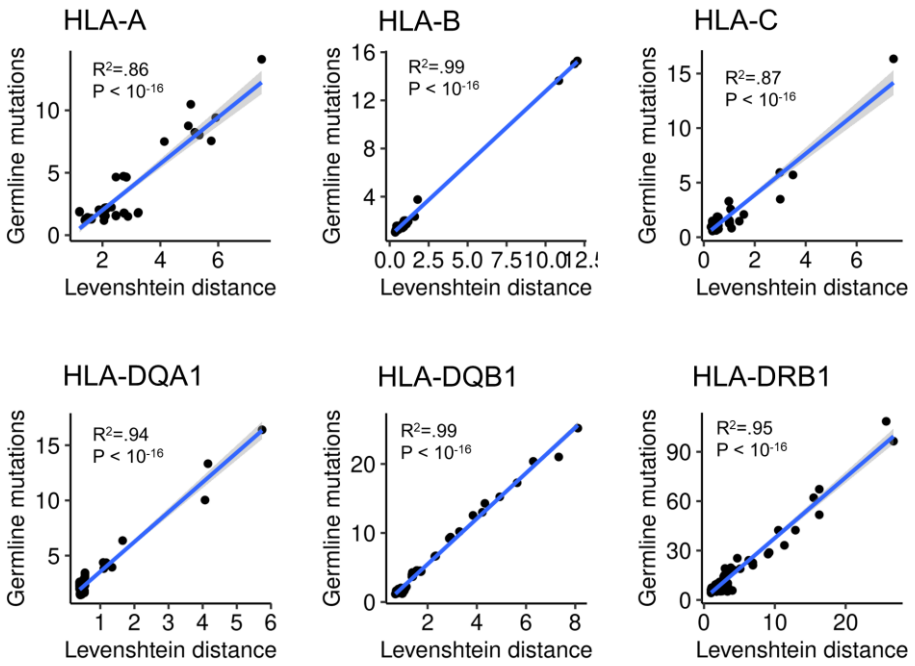

Figure 2-3: Private germline variants as a proxy for Levenshtein distance
Correlation plots showing high concordance between private germline variants and Levenshtein distance as a quantitative measure for similarity of a reference sequence to the true underlying genomic sequence.

2.3 Validation and benchmarking of the Hapster algorithm

2.3.1 Comparison of Hapster to existing HLA haplotypers

For the reference selection portion of the Hapster algorithm, in principle any of the many existing HLA haplotypers[72,77–82] could be used to identify HLA haplotype sequences. However, in practice, existing haplotypers often report HLA types for which only the sequence for exons 2 and 3 are known[1] **(Figure 2-4A)**. Given that we aim to identify mutations in all exons of each gene, the inability to guarantee a full-length genomic sequence for every reported HLA type

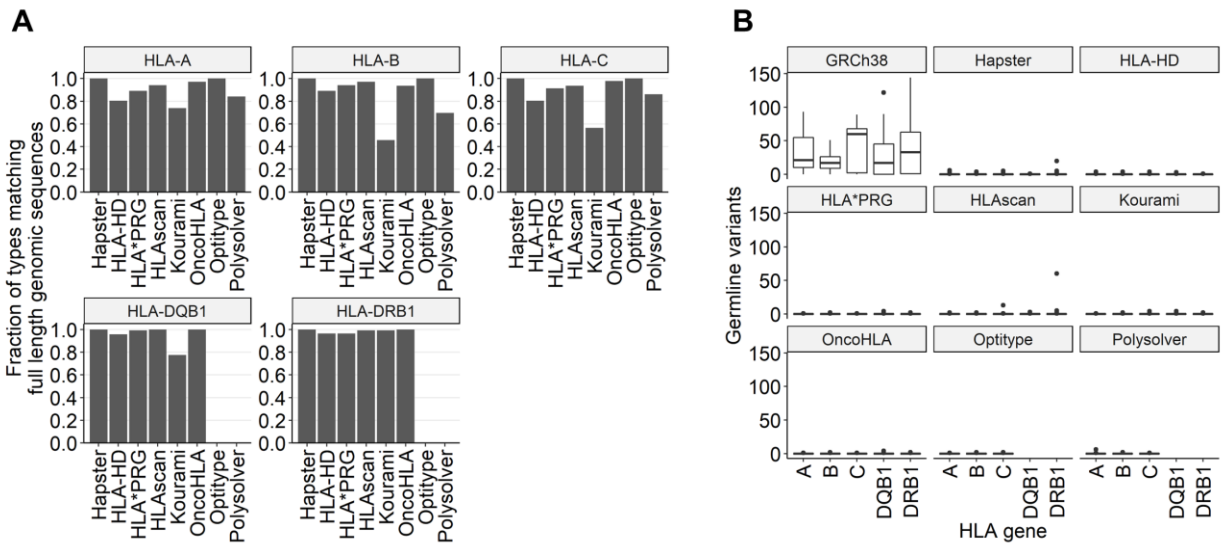

Figure 2-4: Comparison of Hapster to existing HLA haplotypers

(A) Fraction of HLA types from 69 1000 Genomes samples reported by 8 different HLA haplotypers that correspond to a full length genomic sequence found in the IMGT/HLA database. HLA haplotypers often report types that only have known sequence for the polymorphic binding pocket (exons 2 and 3), and full length genomic sequences are unavailable.**(B)** Germline variants identified from 69 WES samples from the 1000 Genomes project relative to either the standard reference GRCh38, or to dynamically selected references using 8 different haplotypers. A perfect reference sequence should produce 0 apparent germline variants.

makes existing haplotypers ineffective for this purpose. Further, some leading haplotypers such as OptiType[77] cannot type the MHC class II genes. However, we would still like to have a measure of the quality of Hapster's reference selection in comparison to existing haplotypers in those cases where they do return full length genomic sequences. To benchmark the haplotype inference portion of the Hapster pipeline we used a set of 69 whole exome sequencing (WES) samples from the 1000 genomes project that have previously reported MHC-I and MHC-II haplotype calls both via sanger sequencing[83] and seven *in silico* prediction methods[82]. MHC haplotyping algorithms are generally 'digit-optimizing', in that they attempt to maximize the number of correct digits in the names of the inferred alleles, which due to HLA nomenclature has the effect of emphasizing protein similarity over DNA similarity. However, for the purposes of mutation calling, it is most critical that the reference haplotype minimizes mismatches between an individual's germline DNA sequence and the chosen haplotypes. We therefore compared Hapster to other haplotyping methods by calling germline variants in WES

sequencing data relative to each haplotyper's inferred haplotype sequence for that individual. We consider the sequencing reads to be the ground truth, and a perfectly identified reference sequence would lead to no germline variants being identified in these reads. We see that relative to the fixed sequences of the standard reference Grch38, there are a median of 17-38 germline variants observed per gene. All tested haplotypers improve upon this, with each having a median of 0 and a mean of <0.5 observed germline mutation per gene (**Figure 2-4B**), which is on par with the ~1 variant per kilobase rate observed relative to the standard reference in other non-polymorphic regions[84].

2.3.2 Sensitivity and specificity of Hapster mutation calls

To assess somatic mutation calling sensitivity, we first simulated 200 synthetic MHC haplotypes with a random mutation, followed by simulated WES at depths ranging from 5x-100x and variant allele fractions (VAFs) of 0.025-0.45. Of the 200 simulated mutations, 94% (187/200) were successfully identified **(Figure 2-5A)** at 100x coverage and a VAF of 0.45. Following filtering, 18 of these calls were removed by either Hapster's or Mutect2's filters, giving a final sensitivity of 85% (169/200) for high coverage clonal variants. Inspection of the 13 variants that failed to be called showed that 12 were in regions of low coverage following probe capture. As such, they reflect a true loss in sensitivity when calling MHC mutations in WES data, but due to the sequencing platform and capture kit design rather than Hapster's algorithm. When looking at results over the entire range of simulated coverages and VAFs, we see that as coverage and VAF decrease, sensitivity decreases as expected due to lower read support for identified variants **(Figure 2-5A)**. A comparison of simulated vs observed VAFs for each mutation call shows that at most simulated VAFs Hapster produces calls with slightly lower observed VAFs, likely due to a slight loss of reads following read filtering **(Figure 2-5B)**. We also note an over-representation of mutation calls at half of the simulated VAF. This occurs when a mutation is called in a homologous

region between two alleles of the same gene causing the variant supporting reads to segregate between the two reference sequences, dropping the VAF by half.

To assess specificity, we called somatic mutations in 450 samples from the TCGA HNSC cohort with tumor and normal labels swapped, such that no somatic variants should be identified. In 9 cases, an apparent somatic variant was identified that passed all filters. Assuming all 9 calls are false positives gives a specificity of 98% (441/450).

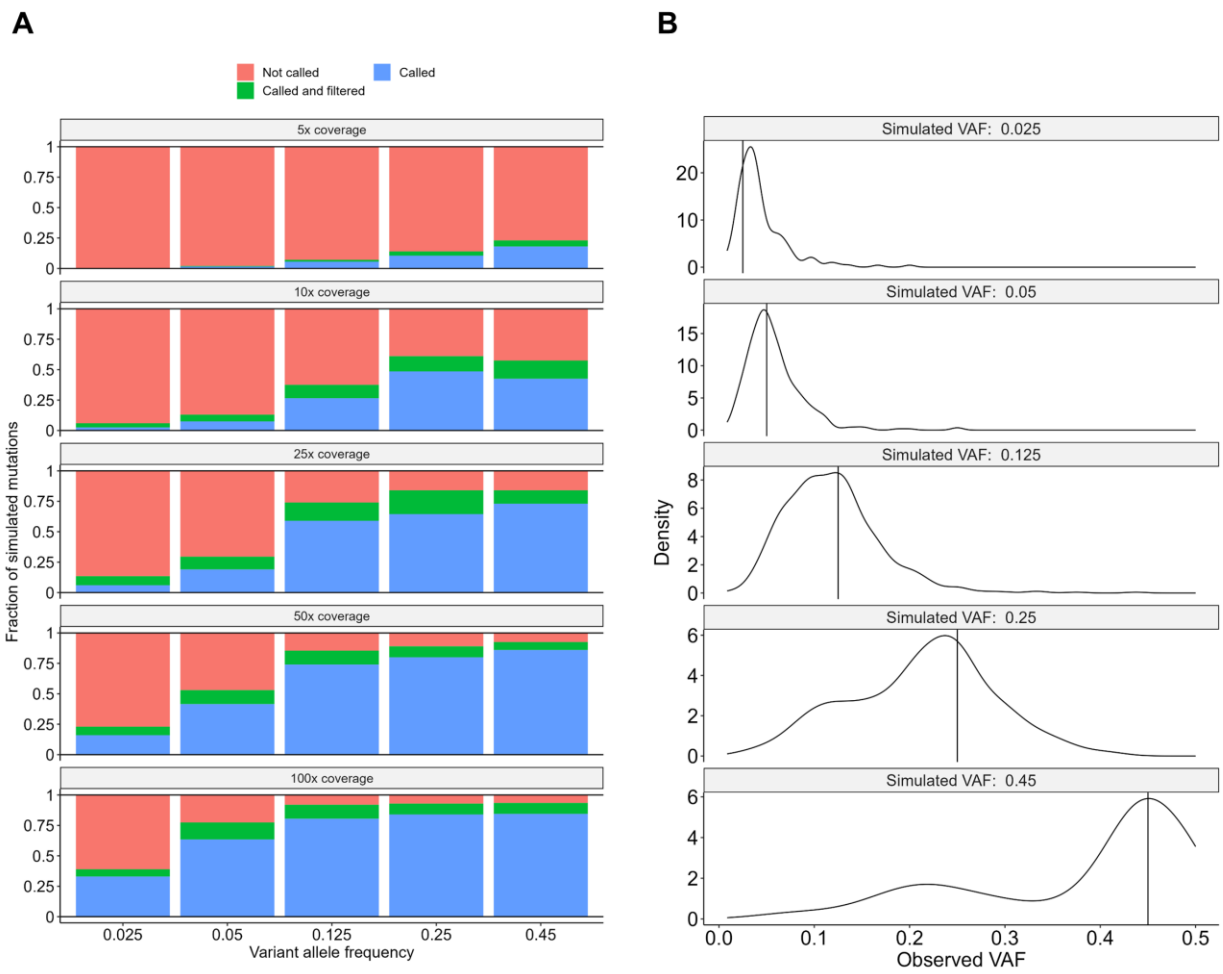

Figure 2-5: Hapster simulation benchmarking

(A) Fraction of simulated mutations either called and passing all filters, called and filtered by Hapster or Mutect2, or never called. Shown are results from whole exome simulations with coverages ranging from 5x to 100x, and mutations with VAFs ranging from 0.025 to 0.45. **(B)** Comparison of observed VAFs from successfully called mutations, compared to the VAF at which the mutation was simulated. Vertical lines show the exact VAF at which the mutation was simulated. Simulations were done probabilistically (see methods), and not with exact fractions of variant to reference supporting reads, so some variance is expected.

To further assess mutation calling accuracy using an orthogonal sequencing technology, we used Hapster to call somatic mutations from WES data for 450 TCGA HNSC samples, and then determined if these same mutations were supported by paired RNA-seq data. While established RNA-seq validation methods would be ideal, they rely on alignment of reads to a reference in order to identify mutations, which would be inappropriate in validating Hapster. We therefore developed a fully orthogonal alignment-free kmer based approach to determine if the read support for each variant in the RNA exceeds expectations based on a beta-binomial model

of sequencer error, avoiding potential reference selection or alignment biases. Of the 80 variants that were called in the WES data, 72 had high enough coverage in the RNA-seq data to undergo validation. Of these, 63 variants (88%) had read support significantly exceeding the null model of sequencing error ($p < 0.05$ (BH corrected), **Figure 2-6**), and 4 (5%) were truncating variants which may have caused the mutant RNA to undergo nonsense mediated decay. This leaves only 5 variants (7%) without RNA evidence, some of which may have failed to validate due to the

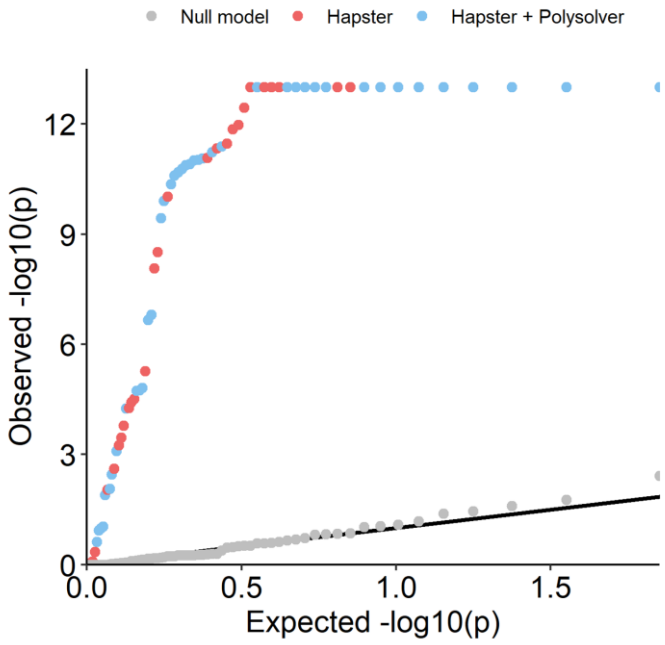

Figure 2-6: Orthogonal validation of variant calls using RNA-seq
QQ plot for observed RNA-seq read support for HLA variants, assuming read support is only due to sequencing error according to a Beta-binomial model. Variants were originally identified by Hapster alone (red), or by both Hapster and Polysolver (blue) from WES data. A comparison is shown to randomly generated alternate bases (grey) which are only supported by noisy reads and follow the null model (diagonal black line).

limitations of our statistical model, variable tumor cellularity, loss-of-heterozygosity (LOH), and ubiquitous transcriptional silencing of the MHC locus in tumors[43].

For a second orthogonal validation, we performed Sanger sequencing on tumors from MI-ONCOSEQ[85] with sufficient DNA or tissue samples. All 14 candidate variants called by Hapster were clearly detected in the Sanger chromatograms from tumor specimens, while being absent in traces from patient-matched normal tissues **(Figure 2-7**).

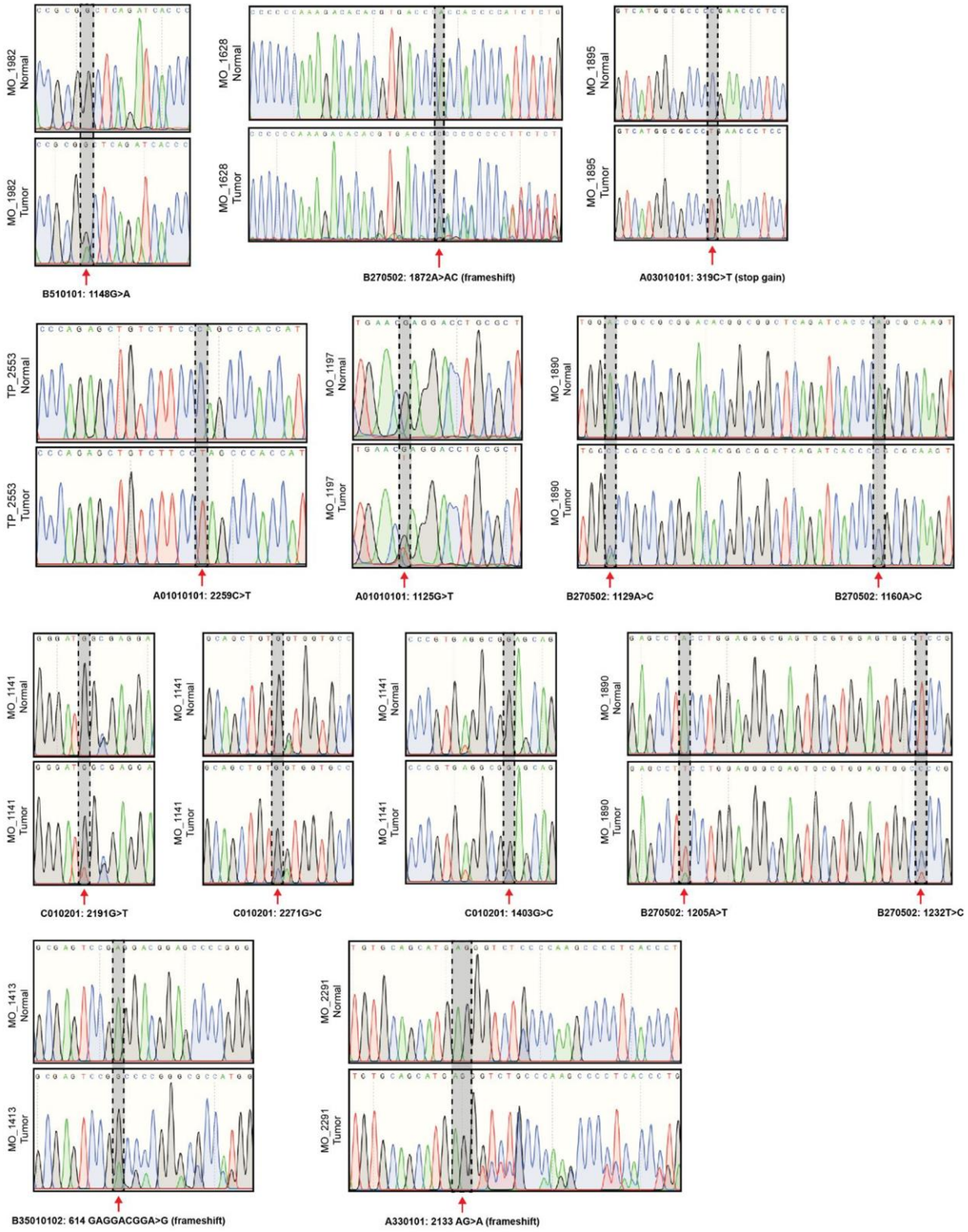

Figure 2-7: Orthogonal validation of variant calls using Sanger sequencing

2.3.3 Comparison of Hapster to the Polysolver and GDAC mutation calling pipelines

Finally, we applied Hapster to a larger set of 7,746 samples from TCGA that have previously reported mutations called by both the Broad Genomic Data Analysis Center (GDAC) standard reference based pipeline and the Polysolver personalized pipeline[72,86]. We found that when calling mutations in the MHC class I genes, Hapster detected over twice as many non-synonymous mutations as the GDAC pipeline, and 36% more than Polysolver **(Figure 2-8A)**. To assess the differences in the mutation calls between Hapster and Polysolver we first looked at the variant allele frequency (VAF) distributions. Hapster tended to identify slightly higher VAF mutations **(Figure 2-8B)**, however this is in part due to Hapster preserving more variant supporting reads than Polysolver **(Figure 2-8C)** resulting in higher VAFs for the same mutations, rather than a failure by Hapster to call low VAF mutations. Indeed, when looking at variants exclusive to Hapster, these tended to be low VAF mutations **(Figure 2-8D)** missed by Polysolver, possibly due to Hapster alone retaining enough reads to identify them.

We next performed an exhaustive search for potential alternative explanations for each variant called by each pipeline. We reasoned that given an accurate haplotype inference, the most likely cause of false positives should be misalignment of sequencing reads originating from other homologous MHC genes or pseudogenes. We found that only 6% of Hapster's non-synonymous calls matched known sequences in any other MHC gene, a rate significantly lower than that of Polysolver (15%, Fisher's exact test $p < 1e-10$, BH adjusted), but similar to the GDAC pipeline (4%, Fisher's exact test $p = 1$, BH adjusted) (**Figure 2-8A**).

Many methods to identify positive selection of mutations within a gene rely on the detection of deviations from a neutral nonsynonymous to synonymous (dN/dS) ratio, and any pipeline biases in functional consequences can confound these analyses. We therefore additionally compared the distribution of functional consequences of HLA mutations called by each of the approaches. For both Hapster and the GDAC pipeline, synonymous mutation calls

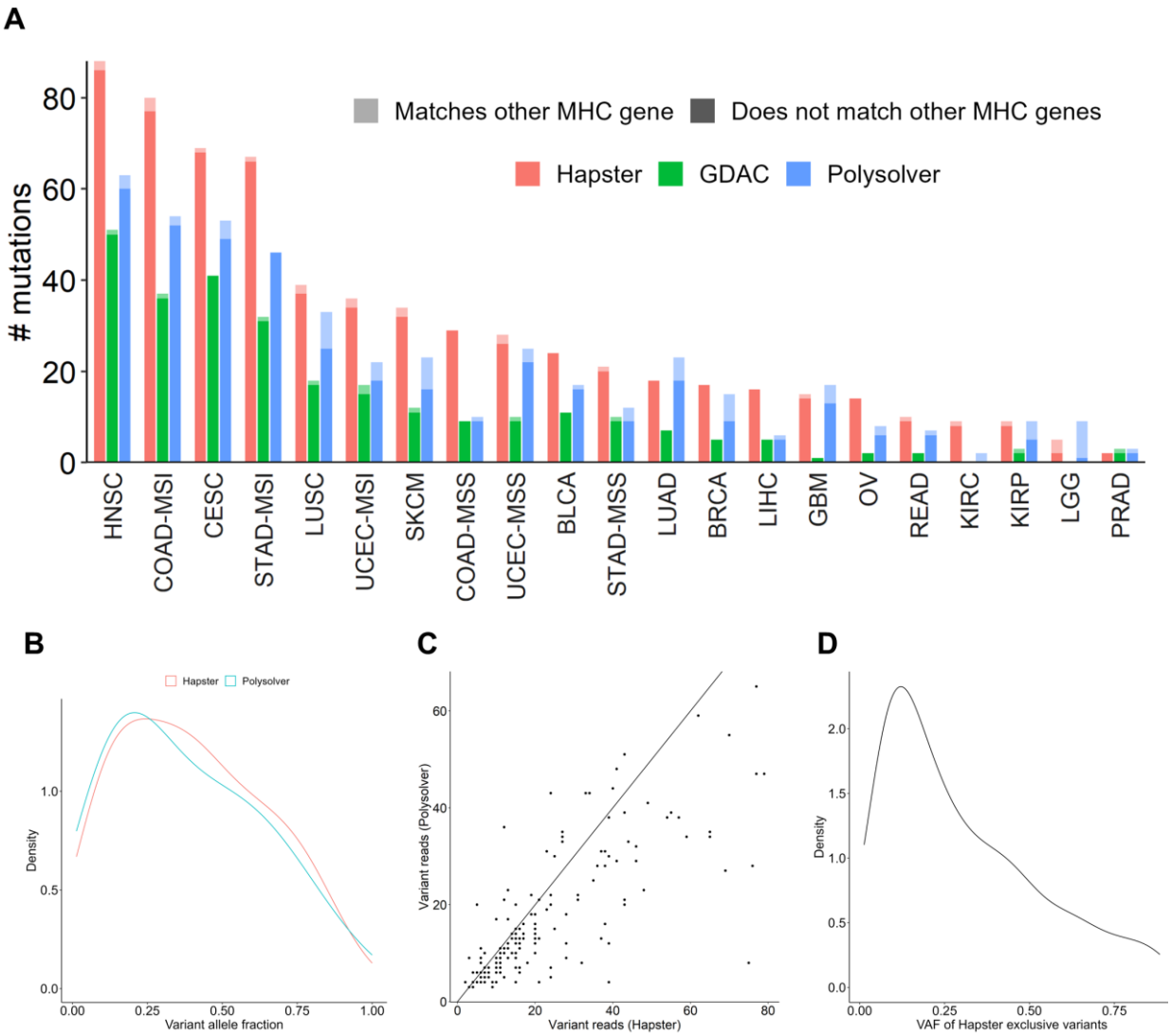

Figure 2-8: Comparison of Hapster to Polysolver and GDAC somatic variant calling pipelines
(A) Comparison of non-synonymous mutation calls for the MHC class I genes between the GDAC pipeline, Polysolver, and Hapster across various cancer types from TCGA. Lightly shaded bars represent possible false positives. **(B)** Comparison of VAF distributions for mutation calls made by either the Hapster of Polysolver pipelines. **(C)** Comparison of read support for variants that were identified by both the Hapster and Polysolver pipelines. **(D)** VAF of mutation calls unique to the Hapster pipeline, showing a relatively high proportion of low VAF variants that were not called by Polysolver.

were underrepresented when compared to neutral genes, consistent with what would be expected for a potential driver gene **(Figure 2-9A)**. In contrast, we found that Polysolver had a surprising over-representation of synonymous calls. Interestingly, an analysis of Polysolver's synonymous mutation calls shows the apparently recurrent synonymous variants p.T214T and

p.A269A that are identified as somatic variants **(Figure 2-9B)**. These mutations are unlikely to be under the extreme positive selection that would be required for such a recurrent hotspot, but have sequences exactly matching non-classical MHC class I genes *i.e.* are likely due to alignment errors from *HLA-E*, *HLA-F*, or HLA pseudogenes.

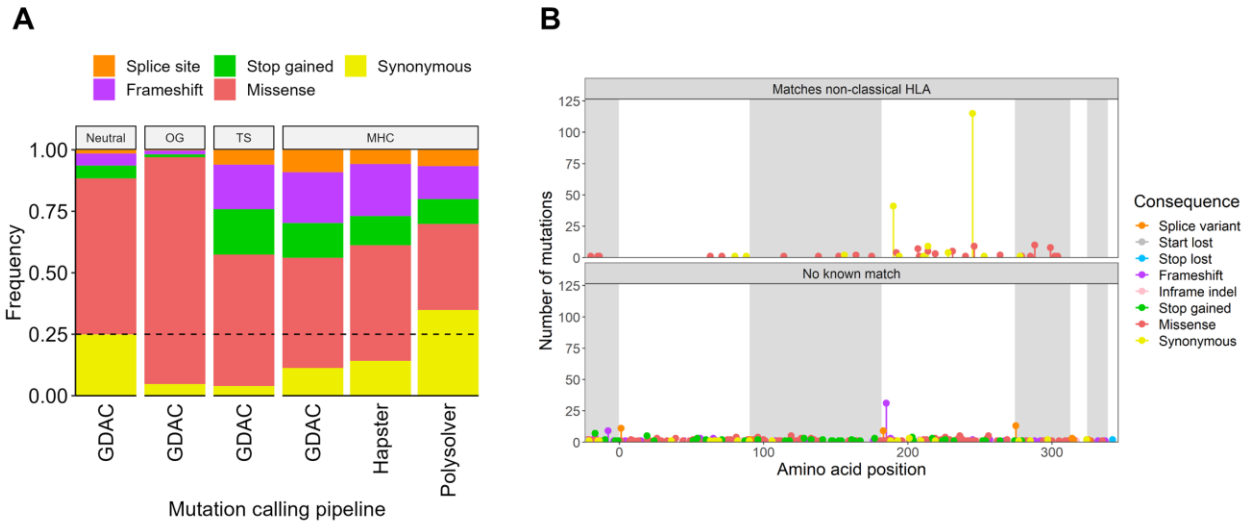

Figure 2-9: dN/dS ratio of mutation calls across Hapster, Polysolver, and GDAC mutation calling pipelines **(A)** Comparison of mutational consequences for variants called by the standard GDAC pipeline, Polysolver, or Hapster in the MHC genes vs oncogenes, tumor suppressors, and neutral gene mutations from TCGA. Oncogenes (OG): KRAS, PIK3CA, IDH1, CTNNB1, FOXA1, BRAF, AKT1, EGFR. Tumor suppressors (TS): TP53, RB1, PTEN, APC, BRCA2, VHL. Neutral genes: All others. **(B)** Positions of mutations called by the Polysolver pipeline, separated by whether or not the variant matches a known sequence in another non-classical MHC class I gene. Annotated recurrent synonymous variants are suspected false positives.

2.4 Discussion

In total, I show that by aligning DNA sequencing reads to personalized HLA reference sequences, the Hapster algorithm allows for more sensitive and specific somatic mutation calling than existing standard reference based pipelines. I also show that Hapster's novel read filters allow for the retention of more HLA reads, and that its novel mutation filters reduce false positives relative to the existing HLA mutation caller Polysolver. Importantly, Hapster also avoids the biases in mutational consequence distributions observed in Polysolver's calls, allowing for a more accurate estimation of dN/dS ratios. This improves our ability to detect

signals of positive selection within the HLA genes, and lays the groundwork for the pan-cancer analysis presented in chapter 3

2.5 Methods

2.5.1 Reference selection validation

Quantitative reference selection validation was performed using a ground truth set of 69 WES samples from the 1000 Genomes project with previously reported HLA haplotype calls[83]. For each sample and each method, an HLA reference was constructed by taking the genomic sequence in the IMGT/HLA database[1] corresponding to the called HLA type. For a comparison to GRCh38, the standard reference haplotype (A*03:01:01:01, B*07:02:01:01, C*07:02:01:01, DQB1*06:02:01:01, DRB1*15:03:01:01) was used for all samples. For samples where the reported haplotype call did not refer to a single unique sequence within the IMGT/HLA database, the alphanumerically first sequence was used. Reads from each sample were aligned to each reference using BWA-mem[74], and germline variants were called using GATK's HaplotypeCaller[87].

2.5.2 Alignment and mutation calling

HLA extracted reads were aligned to Hapster inferred reference sequences using BWA-mem[74]. During read extraction, reads from other homologous HLA genes and pseudogenes may be collected due to their high sequence similarity. For this reason, any reads aligning to blacklisted genes (HLA-E, -F, -G, -H, -J, -K, -L, -N, -P, -S, -T, -U, -V, -W, -Y, -DMA, -DMB, -DOA, -DOB, -DPB2, -DRB3, -DRB4, -DRB5) were discarded. Mutation calling was performed using GATK's Mutect2[75]. Mutations were filtered based on the following criteria: 1) Must pass the GATK filters *FilterMutectCalls* and *FilterByOrientationBias*, 2) The alternate base must not have been observed in the same position in other alleles of the same gene, 3) Read support in the tumor must be at least 3 reads, or at least 20% VAF, 4) Variant must have no more than 1

read support in the normal after a kmer search. For the kmer filter, all 25-mers covering the variant position are used to search for any matching reads in the normal in an alignment-free manner.

2.5.3 Simulation validation

200 synthetic haplotypes were constructed by taking two random allelic sequences from the IMGT/HLA database[1] for each MHC class 1 and class 2 gene (HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRA, -DRB1). For each synthetic haplotype, a mutation was inserted randomly using the following logic: 1) One allele from the haplotype was selected uniformly randomly 2) One position within the coding region of the allele was selected uniformly randomly 3) A mutational consequence was chosen with a 75% chance of creating an SNV, a 12.5% chance of creating a deletion, and a 12.5% chance of creating an insertion 4) For SNVs a random alternate base was chosen uniformly evenly, for deletions a deletion of length 1-5 nucleotides was chosen uniformly randomly, and for insertions an insertion of length 1-5 nucleotides was chosen uniformly randomly, with the insertion being uniformly random nucleotides. From each haplotype, inserts of length 125-300 were simulated from the normal reference sequences for each sample using the BBmap[88] function randomreads.sh to create a pool of inserts for the simulated normal. For the simulated tumor, a separate set of inserts was simulated from the normal reference sequences, as well as a set of inserts from the mutated reference sequences, to create both a germline and mutant pool of reads for the simulated tumor. To simulate various VAFs, inserts were mixed from the tumor germline and tumor mutant pools to create a final simulated tumor set with VAFs of 0.025, 0.05, 0.125, 0.25, and 0.45. Inserts were considered captured if they had significant overlap with MI-ONCOSEQ WES capture probes. Paired reads of length 125 were taken from the ends of each captured insert. To simulate various levels of WES coverage, a random subset of captured reads was selected to create BAM files representing 5x, 10x, 25x, 50x, and 100x

average coverage across gene exons. Mutations were then called using the Hapster pipeline and were filtered using Mutect2 and Hapster filters.

2.5.4 Label-flipping validation

To assess specificity, we performed a label-flipping experiment with the TCGA HNSC cohort. The HNSC cohort was selected as it contained a large number of samples (N = 450), and we knew it to have a high number of somatic mutation calls from previous analyses. We swapped all tumor and normal labels for each case, and called somatic mutations using the Hapster pipeline as described. Any somatic variant called after label-flipping was considered to be a false positive.

2.5.5 RNA validation of somatic mutations

All variants from the TCGA HNSC cohort with sufficiently high coverage in the RNA to detect low allelic fraction variants (>1000 reads at the variant position) were selected. MHC class II variants were not selected for RNA validation as the expression of MHC class II genes is expected to be dominated by immune cells and not cancer cells. For each somatic mutation, a set of all 25-mers containing either the called variant or the germline sequence were created. In the case that other variants were called within 25 bases of the primary variant, kmers were produced with both the primary variant alone and with other variants in combination to account for possible phasing. Kmers that contained the variant within 8 bases of the edge of the kmer were discarded to avoid confounding results when deletions or insertions in small repetitive regions lead to variant kmers that are identical to the germline sequence. All reads within the RNA seq data for a sample were searched for germline or variant supporting kmers. Due to the high depth of the RNA-seq data, we often observe reads supporting every alternate base at every position due to some reads containing sequencing errors. We therefore only consider somatic mutations to be validated if the number of variant supporting reads could not be explained by sequencing errors alone. We modeled sequencing errors as a Beta-Binomial distribution with the

probability of error as a Beta distribution determined experimentally from RNA-seq data. An experimental null sample was created by selecting random bases at random positions within sequenced samples and evaluating the support for the random variant based on the null distribution. Variants were considered validated if the probability of observing equal or more variant supporting reads was less than 5% (BH corrected) given our distribution.

2.5.6 Sanger sequencing validation of somatic mutations

14 somatic mutations called within samples from the MI-ONCOSEQ project that had tissue samples available were chosen to be validated via Sanger sequencing. For each mutation, Primer3[89] was used to create custom PCR probes for the specific HLA allele that the mutation was called within. Genomic DNA from the tumor was amplified using PCR and amplicons isolated using gel electrophoresis. Following Sanger sequencing, point mutations were considered validated if a clear peak containing the called variant was observed, and indels were considered validated if a clear peak-offset corresponding to the number of inserted or deleted bases was observed.

Chapter 3 Distinct Mutational Processes Shape Selection of MHC Class I and Class II Mutations Across Primary and Metastatic Tumors

Portions of this chapter are available as a preprint[73], and are under review at *Cell Reports*.

3.1 Introduction

There is now overwhelming evidence from IHC studies that class I HLA loss is a common phenotype observed across many cancers[43]. However, due to the heterogeneous nature of cancers, it is likely that different cancers evolve this loss through different mechanisms. Identification of the specific mechanisms leading to loss of HLA expression will be key to understanding primary and acquired resistance to immunotherapies. In particular, it is necessary to determine when HLA loss has occurred through a reversible mechanism such as hypermethylation, or an irreversible mechanism such as somatic loss of one or more HLA genes. In cases with somatic loss of the HLA genes, especially an HLA responsible for presentation of an immunodominant antigen, it may indicate that T-cell based immunotherapies are no longer a viable treatment option.

Early investigations into somatic mutation of the class I HLA genes revealed significant evidence for positive selection of loss of function variants in select cancers[72]. However, reports on somatic mutation of the class I HLA genes in metastases are lacking. It is therefore unclear if class I HLA mutations are an early event in primary tumors that will already be present when metastasis takes place, or a late event that either enables metastasis or evolves at the metastatic site. The class II HLA proteins are also relatively understudied across cancers, but evidence suggests that they also play an important role in tumor suppression[90–92]. The MHC class II is

responsible for presenting neoantigens to CD4+ T cells which have both regulatory and effector functions and have been shown to play a role in tumor immunity[93]. Classically, MHC class II expression was thought to be restricted to professional APCs. However, their expression can be induced in most cell types[54], including cancer cells[55–57], and MHC class II-restricted neoantigen vaccines have been shown to transform cancer cells into APCs[58]. In this context, the MHC class II expressing cancer cells promote an anti-tumor response suggesting that loss of MHC class II function may also promote tumor survival. However, this process is not well understood, and may only be relevant to a subset of cancers. To address these gaps, we applied Hapster to 12,000 cancers from the Cancer Genome Atlas (TCGA) and Michigan Oncology Sequencing Center (MI-ONCOSEQ) cohorts to reveal the landscape of somatic mutation of the class I and class II HLA proteins in primary and metastatic cancers **(Box 3-1)**.

- MHC genes are among the most recurrently mutated genes pan-cancer
- Cancer type specific mutational processes shape the spectrum of MHC mutations
- Off-target AID/APOBEC activity likely causes stop-gain mutations in lymphomas and squamous cell carcinomas
- Microsatellite instability causes truncating frameshifts in HLA-A and HLA-B
- MHC missense mutations disrupt B2M and antigen binding

Box 3-1 – Highlights from chapter 3 pan-cancer study

3.2 Results

3.2.1 Pan-cancer compendium of MHC class I and MHC class II mutations

To comprehensively characterize MHC class I and MHC class II mutation rates in human cancer we analyzed 10,001 tumors across 35 cancer types from TCGA (**Table 3-1**) and 2,199 tumors across 24 cancer-types within MI-ONCOSEQ[85] (**Table 3-2**), for a total compendium of 2069 MHC class 1 and class 2 mutations (**Figure 3-1A**). Samples from TCGA are mainly primary tumors, with the exception of the melanoma cohort (SKCM) which consists only of metastatic

Table 3-1: Number of paired tumor/normal whole exomes used per TCGA cohort

Cohort	Short code	N
Adrenocortical carcinoma	ACC	91
Bladder Urothelial Carcinoma	BLCA	410
Breast invasive carcinoma	BRCA	1040
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	305
Cholangiocarcinoma	CHOL	50
Colon adenocarcinoma - Microsatellite instability	COAD-MSI	69
Colon adenocarcinoma - Microsatellite stable	COAD-MSS	353
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	37
Esophageal carcinoma	ESCA	180
Glioblastoma multiforme	GBM	390
Head and Neck squamous cell carcinoma	HNSC	450
Kidney Chromophobe	KICH	64
Kidney renal clear cell carcinoma	KIRC	332
Kidney renal papillary cell carcinoma	KIRP	285
Brain Lower Grade Glioma	LGG	505
Liver hepatocellular carcinoma	LIHC	362
Lung adenocarcinoma	LUAD	558
Lung squamous cell carcinoma	LUSC	494
Mesothelioma	MESO	82
Ovarian serous cystadenocarcinoma	OV	422
Pancreatic adenocarcinoma	PAAD	169
Pheochromocytoma and Paraganglioma	PCPG	179
Prostate adenocarcinoma	PRAD	438
Rectum adenocarcinoma	READ	156
Sarcoma	SARC	255
Skin Cutaneous Melanoma	SKCM	467
Stomach adenocarcinoma - Microsatellite instability	STAD-MSI	83
Stomach adenocarcinoma - Microsatellite stable	STAD-MSS	354
Testicular Germ Cell Tumors	TGCT	150
Thyroid carcinoma	THCA	491
Thymoma	THYM	118
Uterine Corpus Endometrial Carcinoma - Microsatellite instability	UCEC-MSI	162
Uterine Corpus Endometrial Carcinoma - Microsatellite stable	UCEC-MSS	364
Uterine Carcinosarcoma	UCS	57
Uveal Melanoma	UVM	79

Table 3-2: Number of paired tumor/normal samples used per MI-ONCOSEQ cohort

Cohort	Short code	WES	Capture panel	Total
Adenoid Cystic/Oral Carcinoma	M-ACO	12	39	51
Adrenocortical Carcinoma	M-ACC	8	19	27
Bladder Carcinoma	M-BLCA	15	33	48
Breast Carcinoma	M-BRCA	148	145	293
Cholangiocarcinoma	M-CHOL	19	55	74
Colorectal/Anal Carcinoma	M-COAD	12	12	24
Diffuse large B-cell lymphoma	M-DLBC	11	34	45
Gastroesophoageal Carcinoma	M-ESCA	14	16	30
Head and Neck Carcinoma	M-HNSC	6	30	36
Kidney/Renal Cell Carcinoma	M-KIRC	10	12	22
Leukemia	M-LEU	20	30	50
Liver/Gallbladder Carcinoma	M-LIHC	7	15	22
Lymphoma	M-LYM	17	92	109
Melanoma	M-SKCM	15	19	34
Myeloproliferative Neoplasms	M-MYE	19	159	178
Nervous System/Brain Carcinoma	M-NBC	18	57	75
Non-Small Cell Lung Cancer	M-NSCL	22	10	32
Ovarian/Vaginal Carcinoma	M-OV	17	10	27
Pancreatic Carcinoma	M-PAAD	12	31	43
Prostate Carcinoma	M-PRAD	493	38	531
Sarcoma	M-SARC	69	157	226
Squamous cell carcinoma	M-SQCC	36	90	126
Thyroid Carcinoma	M-THCA	3	16	19
Unknown primary	M-UNKP	26	51	77

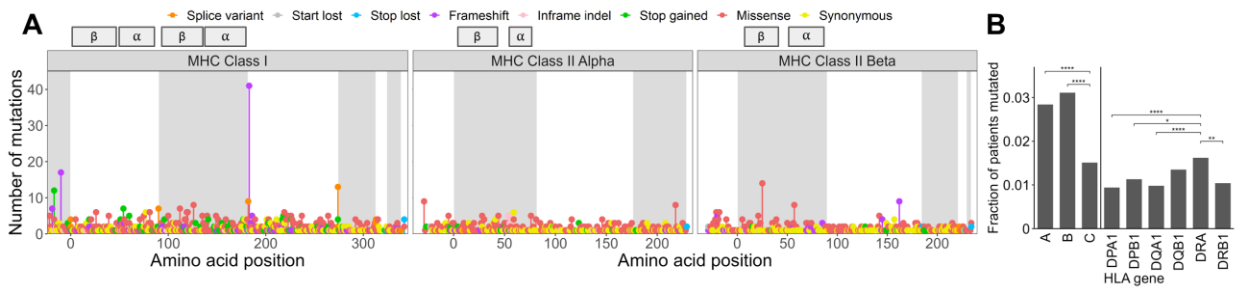

Figure 3-1: Pan-cancer of class I and class II mutations

(A) Distribution of all observed mutations in both primary and metastatic cancers across the coding region of the MHC genes. Denoted above are regions of secondary structure corresponding to either the binding pocket floor (β-sheets) or walls (α-helices). **(B)** Significant differences in the prevalence of nonsynonymous mutations and indels of individual MHC class I and MHC class II genes. *: $p < .05$; **: $p < .01$; ***: $p < .001$; ****: $p < 0.0001$, BH corrected Fisher's exact test

samples. Microsatellite unstable (MSI) tumors are immunologically distinct due to their significantly higher neoantigen burden[94] and we have therefore separated them from their microsatellite stable (MSS) counterparts within the colon (COAD), stomach (STAD), and endometrial (UCEC) TCGA cohorts. While some other cancers also have distinct subtypes, such as BRCA ER+/- and HNSC HPV+/-, no significant difference in MHC mutation rates was observed

Mutations were in general distributed uniformly across the gene body, but occasionally concentrated within prominent hotspots (**Figure 3-1A**). We found that for the MHC class I HLA-A and HLA-B contained significantly more mutations than HLA-C, and for the MHC class II HLA-DRA contained significantly more mutations than all other MHC class II genes except for HLA-DQB1 (**Figure 3-1B**). Within each HLA gene, no particular allele was found to bear an excess of mutations. In primary tumors, we noted substantial variation in both mutational frequency and their predicted consequences across tumor types and the MHC gene classes (**Figure 3-2**). We found nonsynonymous MHC class I and MHC class II mutations in 10.5% of primary tumors (ranging from 2.7% to 72.5% across cancer types) (**Figure 3-3A**), with 5.6% (range 0.2% to 62.3%) of patients harboring an MHC class I and 5.7% (range 1.1% to 21.7%) an MHC class II somatic variant. Consistent with previous reports that MSI tumors should be under strong pressure to acquire loss of MHC function[95,96], the COAD-MSI, STAD-MSI, and UCEC-MSI cohorts

make up 3 of the top 4 cohorts for MHC class I mutations with the majority being loss of function (LOF) frameshifts or stop gains (**Figure 3-3B)**. MHC class II mutations were also most prevalent in cancers with high mutation burden including MSI tumors and melanoma (**Figure 3-3B**). However, LOF mutations in the top-mutated cohorts were less frequent and the variation in mutation-rates across cancer types was lower compared to MHC class I.

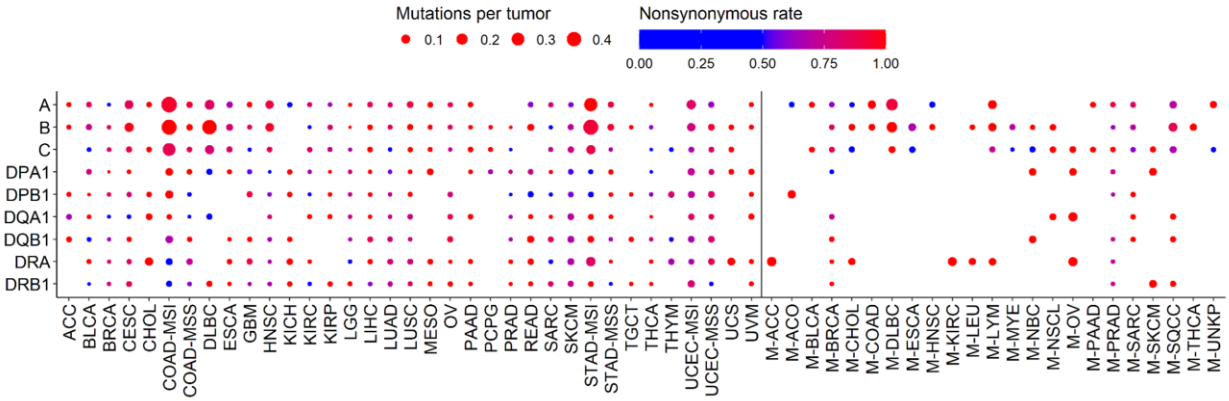

Figure 3-2: Pan-cancer overview of MHC class I and class II mutations per-gene

Cohort specific mutation rates for MHC class I and MHC class II genes across all primary and metastatic cancers. Values are scaled to the number of individuals within each cohort. Colors represent the fraction of cancers with nonsynonymous/indel mutations. At neutrality, the expected nonsynonymous rate should be approximately 0.75.

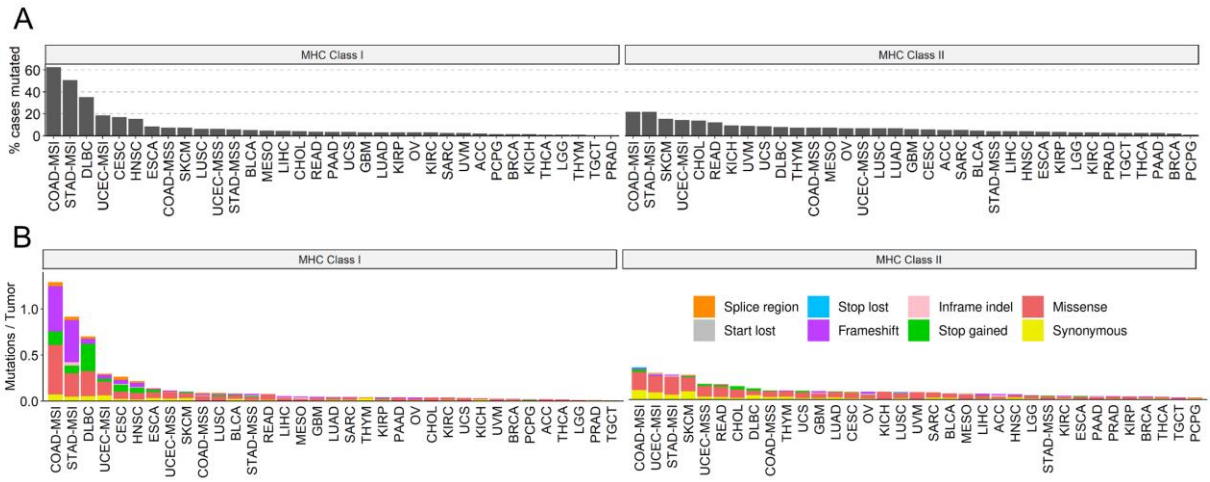

Figure 3-3: Overview of MHC class I and class II mutations within all TCGA cohorts

(A) Fraction of primary cancers from TCGA harboring mutations in any MHC class I and any MHC class II gene. **(B)** Functional consequences of coding region mutations in MHC class I and MHC class II genes in primary cancers from TCGA.

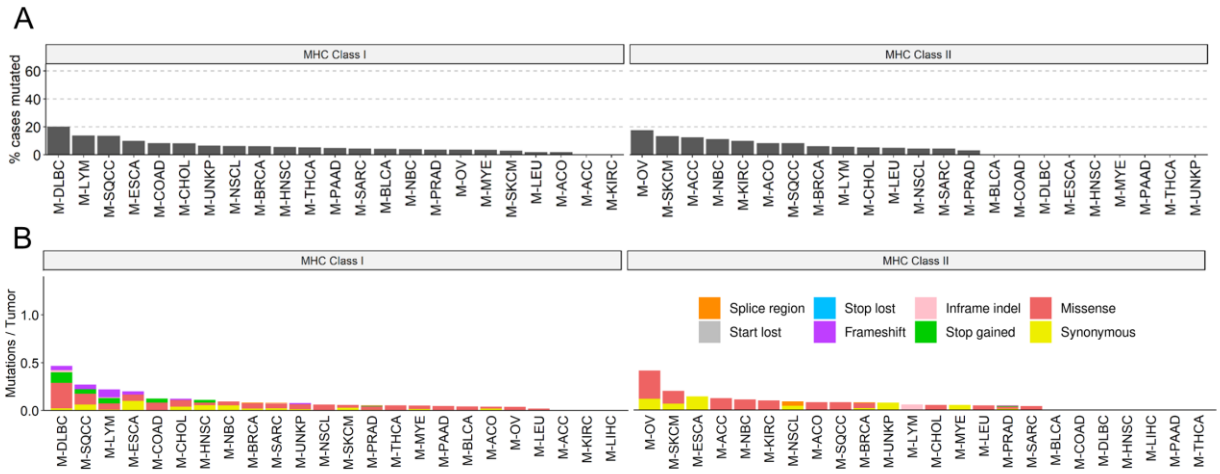

Figure 3-4: Overview of MHC class I and class II mutations within all MI-ONCOSEQ cohorts

(A) Fraction of metastatic/refractory cancers from MI-ONCOSEQ harboring mutations in any MHC class I and any MHC class II gene. **(B)** Functional consequences of coding region mutations in MHC class I and MHC class II genes in metastatic/refractory cancers from MI-ONCOSEQ.

3.2.2 Prevalence of MHC class I and MHC class II mutations in primary vs metastatic tumors

The prevalence of MHC mutations in metastatic tumors is unknown, a critical gap in knowledge considering the predominant use of immunotherapies in this setting and immunological differences between the primary and metastatic tumor micro-environment (TME)[85,97–100]. Overall, we observed nonsynonymous MHC class I and MHC class II mutations in 7.6% (range 3.3%-20.0%) of metastatic/refractory patients, with substantial variation in mutational frequency and functional consequences between cancer types (**Figures 3-3, 3-4**). To directly compare mutation rates between primary and metastatic cancers we created a set of pairings to match TCGA cohorts to MI-ONCOSEQ cohorts (**Table 3-3**). For 15/17 pairings (88%) there were no significant changes in primary vs metastatic MHC class I or MHC class II mutation rates. However, for prostate and breast cancers we observed a significant increase in MHC class I mutations in metastatic cancers compared to primary (**Figure 3-5,** prostate: $F(1, 909) = 9.35$, $p = 0.03$; breast: $F(1, 1140) = 12.8$, $p = 0.01$, BH adjusted). No significant differences were seen in MHC class II mutations.

Overall these data provide, to our knowledge, the first comprehensive look at MHC class I and MHC class II mutations pan-cancer, across both primary and metastatic tumors. We find that somatic mutations of HLA-A and HLA-B are most common while HLA-C and MHC class II genes are less frequently mutated and are likely only relevant in specific cancer types. While some significant differences in MHC mutation rate between primary and metastatic tumors are noted, the majority of MHC mutations in metastatic tumors are expected to be already present in the primary tumor.

Table 3-3: Cohorts paired by cancer/tissue type

Pairing group	Primary cohorts (TCGA)	Metastatic/Refractory cohorts (MI-ONCOSEQ)
Adenoid	ACC	M-ACC
Bladder	BLCA	M-BLCA
Breast	BRCA	M-BRCA
Cholangio	CHOL	M-CHOL
Colorectal	COAD, READ	M-COAD
Esophogus	ESCA	M-ESCA
Kidney_rcc	KIRC	M-KIRC
Liver	LIHC	M-LIHC
Lung	LUAD	M-NSCL
DLBC	DLBC	M-DLBC
Melanoma	SKCM	M-SKCM
Ovary	OV	M-OV
Pancreas	PAAD	M-PAAD
Prostate	PRAD	M-PRAD
Sarcoma	SARC	M-SARC
Squamous	CESC, HNSC, LUSC	M-SQCC
Thyroid	THCA	M-THCA

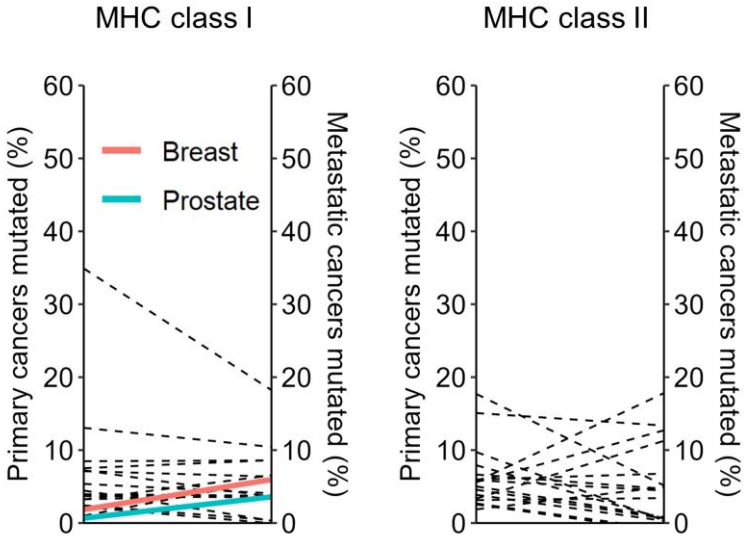

Figure 3-5: Comparison of MHC mutations between primary and metastatic/refractory cancers
Change in percent of tumors harboring MHC class I or MHC class II mutations between primary and metastatic cancers. Values are marginal means after adjusting for tumor mutation burden. Colored lines represent cohorts with significantly different numbers of mutations between primary and metastatic cases ($p < 0.05$, BH adjusted). Sample sizes are slightly reduced from full cohort sizes due to missing WES mutation calls with which to calculate the global tumor mutation burden covariate. Breast: primary n = 854, metastatic n = 289; Prostate: primary n = 434, metastatic n = 478.

3.2.3 Positive selection of non-synonymous MHC somatic mutations

Given the high proportion of deleterious mutations in cancer types with the highest frequency of MHC mutations, we asked whether there was significant evidence for positive selection of functional mutations within the MHC genes. We applied CBaSE[101], a tool that estimates the gene-specific strength of positive or negative selection for functional mutations, to each primary and metastatic cohort from TCGA and MI-ONCOSEQ, respectively. HLA genes and haplotypes are codominant and each allele presents a largely unique set of neoantigens[102]. In addition, specific T cell responses are often immunodominant and mounted against only a few of the presented neoantigens. Mutation of a single HLA allele may therefore result in the complete inability to present an immunodominant neoantigen. Accordingly, in the following analyses we treat all MHC class I genes (and separately, all MHC class II genes) as one functional unit, analogous to multiple genes of a protein complex[103], taking into account the increased genomic length of this combined set of genes. In primary cancers, CBaSE identified 6 cohorts (COAD-MSI, STAD-MSI, DLBC, CESC, HNSC, LUSC) with statistically significant evidence for positive selection of non-synonymous variants in the MHC class I genes, and 3 cohorts (CHOL, KICH, UVM) for the MHC class II genes **(Figure 3-6A)**. By this measure, the MHC class I are tied for 7th and the MHC class II are tied for the 17th most recurrent driver genes pan-cancer as determined by applying CBaSE to all protein-coding genes across primary cancers. A similar trend was identified in metastatic and refractory cancers with the MHC class I genes being mutated in two cohorts (M-DLBC, M-LYM) making them tied for 6th most recurrent pan-cancer driver gene by number of cohorts significantly mutated **(Figure 3-6B)**. As an alternative measure of positive selection, we used Fisher's method to create a combined score Φ_{pos} for the strength of selection across all cohorts **(Figures 3-6C/D)**. We found that in both primary and metastatic cancers the MHC class I genes scored in the top 0.1% of all protein-coding genes according to this metric of positive selection **(Figures 3-6C/D)**, and in primary cancers the MHC class II genes

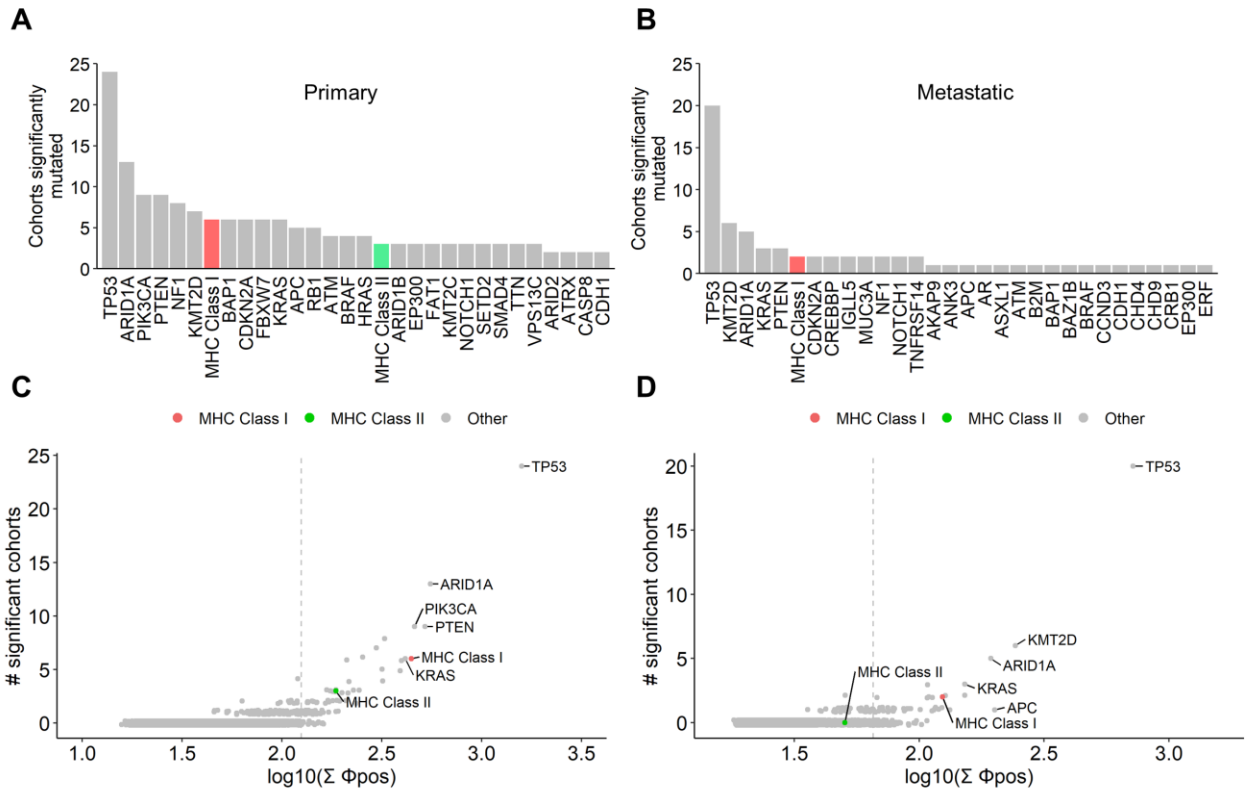

Figure 3-6: Pan-cancer strength of positive selection for all genes using CBaSE
(A, B) Top 30 genes showing evidence of positive selection in primary **(A)** or metastatic **(B)** cancers by CBaSE by number of cohorts with significant evidence. **(C, D)** Comparison of the number of cohorts significantly mutated vs pan-cancer metastatic Φ_{pos} for protein-coding genes in primary **(C)** or metastatic **(D)** cancers as measured by CBaSE. Vertical dashed lines show the cutoff for the top 0.5% of genes by Φ_{pos}.

scored in the top 0.5% **(Figure 3-6C)**. Due to the exclusion of MHC class II genes from the sequencing panel in a subset of MI-ONCOSEQ samples, we were not statistically powered to investigate selection of MHC class II genes in metastatic cohorts.

To provide further evidence for positive selection, we looked at the clonality of mutations within the 6 TCGA cohorts (COAD-MSI, STAD-MSI, DLBC, CESC, HNSC, LUSC) reported to be significantly mutated by CBaSE. We show that the majority of mutations in HLA-A (111/164, 68%) and HLA-B (132/179, 74%) within these cohorts have a cancer cell fraction (CCF) >0.7, consistent with the variants providing a survival advantage followed by a clonal sweep (**Figure 3-7**). In contrast, in cohorts showing no evidence of positive selection the proportion of clonal mutations were significantly lower in both HLA-A and HLA-B, consistent with these being mostly subclonal

passenger mutations. In both groups HLA-C is primarily subclonal, indicating that mutations in this gene may not provide as much survival benefit, consistent with our earlier finding that HLA-C is less frequently mutated than HLA-A and HLA-B **(Figure 3-1B)**.

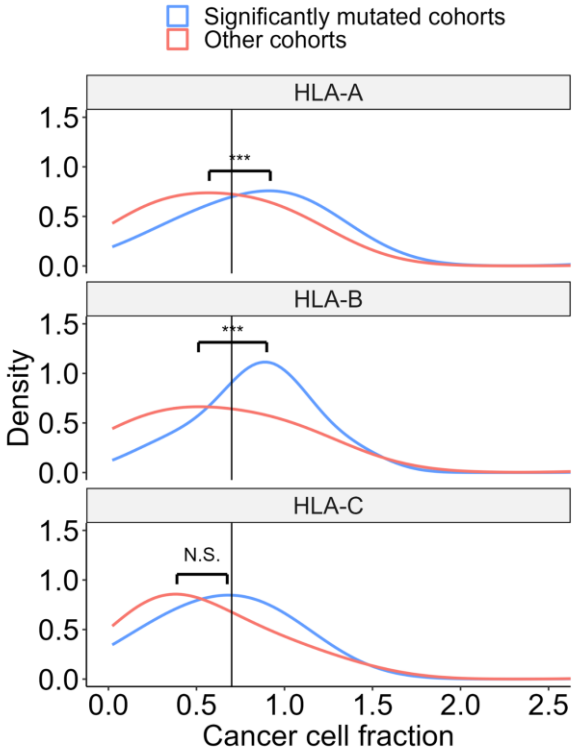

Figure 3-7: Clonality of MHC mutations within cohorts showing evidence of positive selection
Cancer cell fraction (CCF) of MHC class I variants in TCGA cohorts showing significant evidence of positive selection compared to all other cohorts. Vertical line shows 70% CCF, above which mutations are considered clonal. ***: $p < 0.001$, Wilcoxon rank-sum test after BH correction.

3.2.4 Impact of tumor mutation burden on MHC class I and MHC class II mutation frequency

To investigate the association between tumor mutational burden (TMB) and MHC mutations we compared the local TMB within the MHC genes to the genome-wide TMB for each cancer cohort. As TMB increases, we expect the number of passenger mutations in a gene to increase stochastically. However, as TMB increases, neoantigen burden also increases, and we would expect increased selective pressures for LOF MHC mutations. We therefore expect all cancer types to show a positive association between TMB and MHC mutations, but in cohorts

with significant evidence of positive selection this increase should be elevated due to the added effect of both TMB and neoantigen induced selective pressures. We show this to be the case, with significantly mutated cohorts having a higher local TMB within the MHC genes than other cohorts of similar global TMB **(Figure 3-8)**.

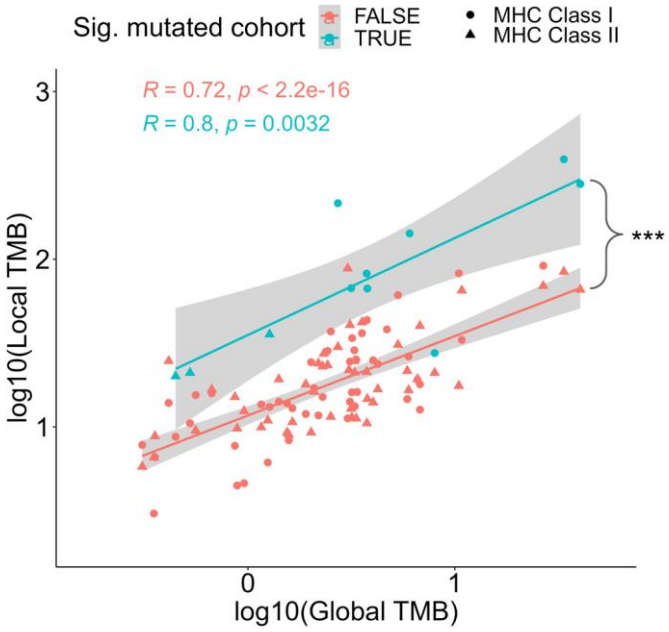

Figure 3-8: Association between global tumor mutational burden and MHC mutations
Comparison between global TMB and local TMB within MHC class I and MHC class II genes in all TCGA and MI-ONCOSEQ cohorts. Average global TMB is calculated based on non-synonymous mutations in all protein-coding genes, average MHC local TMB is the number of mutations in MHC class I or class II genes divided by their length. Significant differences between regression lines was measured using the Chow test, ***: p < 0.001.

We originally hypothesized that somatic loss of MHC class II should mirror that of MHC class I given that both have been shown to promote anti-tumor immune responses. However, there was no association at the cohort level between MHC class I mutations and MHC class II mutations after controlling for TMB **(Figure 3-9)**. Strikingly, while MHC class I mutations appeared to be most prevalent in cancer types with high TMB, MHC class II mutations were frequently increased in low TMB cancers with few MHC class I mutations.

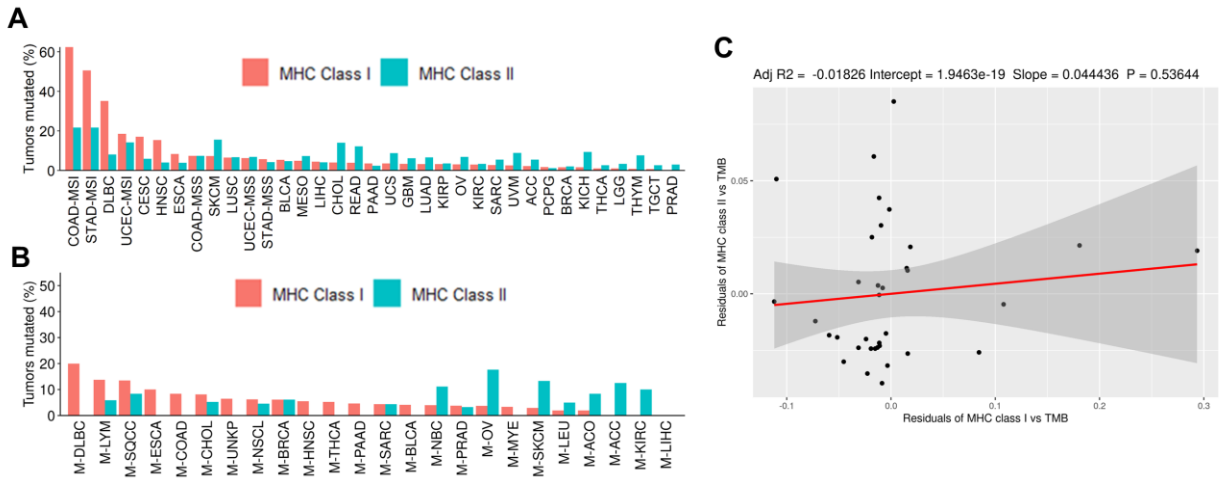

Figure 3-9: Association between MHC class I and class II mutations pan-cancer
(A, B) Proportion of tumors with mutations in MHC class I or MHC class II in **(A)** primary and **(B)** metastatic tumors. **(C)** Residual correlation between the numbers of MHC class I and class II mutations in primary tumors after adjusting for tumor mutational burden. Points correspond to each of the 35 cancer types from TCGA.

3.2.5 Functional consequences of MHC class I and MHC class II mutations

To better understand the impact of positive selection of non-synonymous mutations in MHC genes, we characterized their functional consequences and compared their distributions in cohorts with and without evidence of positive selection. We constructed an approximately neutral model by looking at the distribution of functional consequences across 2.6 million mutations called from the entirety of the TCGA, the overwhelming majority of which are known to be passengers[104] (**Figure 3-10A,** "TCGA"). MHC class I mutations within cohorts showing no evidence of positive selection showed a consequence distribution nearly identical to that of the neutral model (**Figure 3-10A,** "Unselected"), supporting the notion that mutations observed in these cohorts are primarily passengers. However, in each of the 8 cancer types that did show positive selection, there was a significant difference in consequence distributions when compared to the TCGA derived neutral model (Chi-squared tests, p < 1e-3 to 1e-16, BH adjusted). Consistent with the MHC's role as a tumor suppressor, this deviation was caused by an increase in truncating mutations which accounted for more than 40% of mutations in most cohorts, as compared to the expected neutral

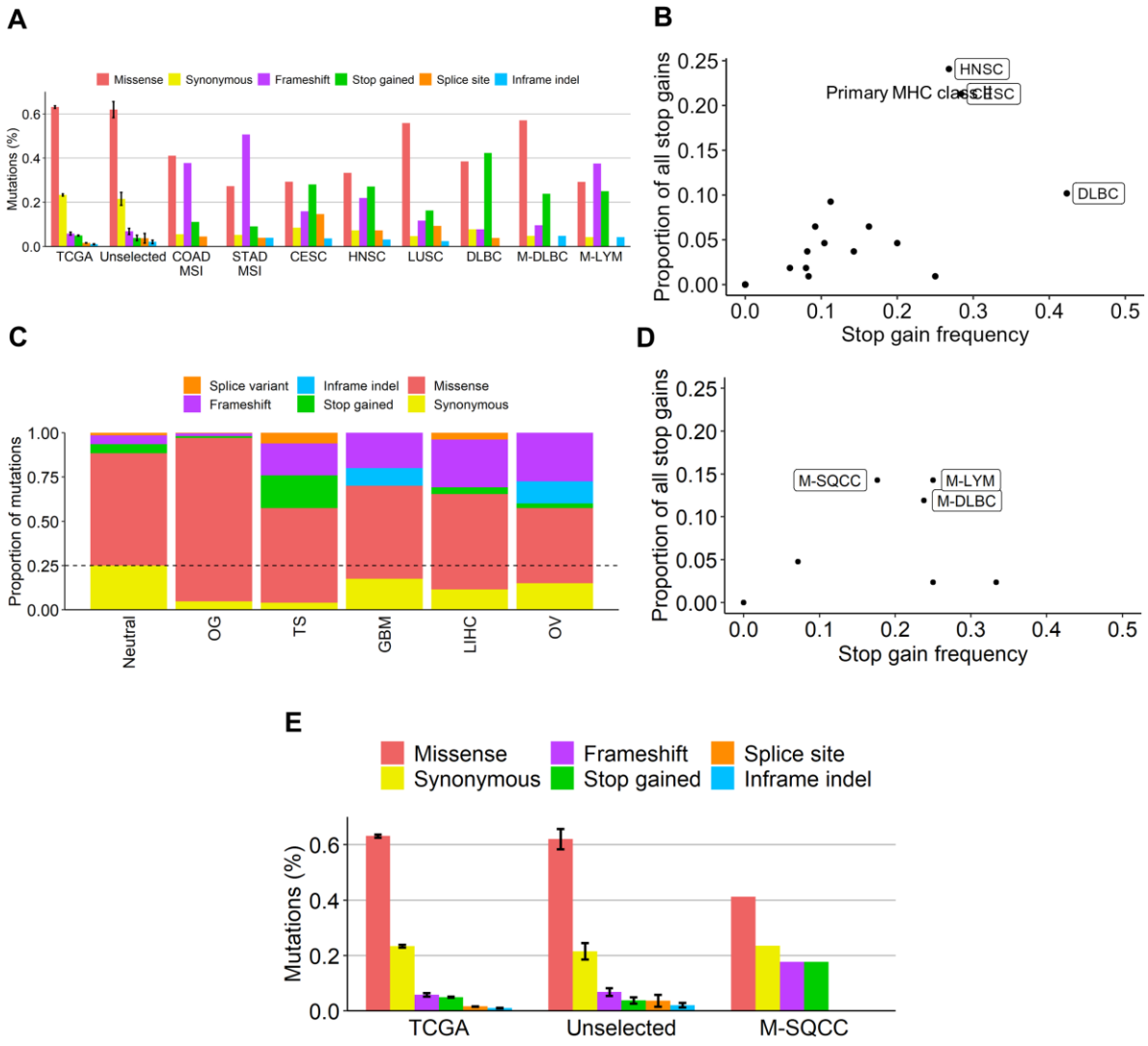

Figure 3-10: Distribution of functional consequences for non-synonymous mutations within cohorts showing evidence of positive selection

(A) Proportion of functional consequences observed in various groups: "TCGA" 2,600,654 pan-cancer mutations from TCGA an approx. neutral model; "Unselected" - MHC class I mutations from all primary and metastatic cohorts showing no evidence of positive selection; others - MHC class I mutations from cohorts showing evidence of positive selection (n = 21-96 mutations within positively selected cohorts). "TCGA" and "Unselected" are average frequencies across cohorts, with error bars showing SEM. (B) Cohort specific MHC class I stop gain frequency compared to overall proportion of MHC class I stop gains contributed by each TCGA cohort. Enriched cohorts are labeled. (C) Functional consequences of MHC class II mutations in select primary cohorts. For comparison, mutational consequence distribution of known oncogenes (OG: KRAS, PIK3CA, IDH1, CTNNB1, FOXA1, BRAF, AKT1, EGFR) and tumor suppressors (TS: TP53, RB1, PTEN, APC, BRCA2, VHL) are shown. (D) Cohort specific MHC class I stop gain frequency compared to overall proportion of MHC class I stop gains contributed by each MI-ONCOSEQ cohort. Enriched cohorts are labeled. (E) Proportion of functional consequences observed in various groups: "TCGA" 2,600,654 pan-cancer mutations from TCGA, an approx. neutral model; "Unselected" - MHC class I mutations from all primary and metastatic cohorts showing no evidence of positive selection; "M-SQCC" - MHC class I mutations from the M-SQCC cohort. "TCGA" and "Unselected" are average frequencies across cohorts, with error bars showing SEM.

rate of ~12%. The B-cell lymphoma (DLBC), cervical (CESC), and head and neck (HNSC) cohorts all have a high proportion of stop gains (46%, 32%, and 28%, respectively) within the MHC class I genes, accounting for 56% (60/108) of all observed stop gains despite only comprising 8% (792/10,001) of TCGA patients (**Figure 3-10B**). Notably, frameshift mutations in MHC class II were rare even in MSI tumors, but unexpectedly common in some MSS tumors including GBM, OV, and LIHC. These cohorts were also depleted of synonymous mutations (**Figure 3-10C**).

Similar to the DLBC cohort from TCGA, the refractory M-DLBC cohort showed both a high mutation rate and a strong bias towards truncating mutations in the MHC class I genes (35%). Other non-DLBC refractory lymphomas (M-LYM) had a lower overall MHC class I mutation rate, but still had a large bias towards truncating mutations (65%) **(Figure 3-10A)**. The lymphomas alone account for 52% of stop gains observed across all MI-ONCOSEQ cohorts (11/21) despite containing only 7% of patients **(Figure 3-10D)**. The HNSC, CESC, and LUSC cohorts in TCGA are all types of squamous cell carcinomas which correspond to a single cohort M-SQCC within MI-ONCOSEQ. Similar to what was observed across the primary squamous cancers, the pan-squamous M-SQCC cohort showed an overall elevated mutation rate and a high rate of LOF mutations when considering frameshifts, stop gains, and splice region variants (35%, **Figure 3-10E**). Metastatic MSI tumors are underrepresented in MI-ONCOSEQ preventing any comparison to primary MSI. Altogether, these data reveal striking differences in mutation frequency and deleteriousness not only across cancer types but also between MHC class I and class II genes.

3.2.6 Patterns of mutual exclusivity and independence of MHC mutations

We next sought to determine whether mutations in the MHC are independent of other mutational drivers. Since the ability to present antigens can be restricted by mutations of other genes that make up the antigen processing machinery (APM), we next looked at the relationships between deleterious mutations in the MHC class I genes and the APM[105] (**Figure 3-11**). Other genes linked to the MHC class 1 have been identified as cancer driver genes (e.g. *B2M*[106]), and it has been

shown that driver genes that fall within the same pathway frequently show mutual exclusivity. This effect is most clear in the lymphomas where there is significant mutual exclusivity between MHC mutations and the APM ($p = 0.02$, BH adjusted) with no observed tumors having hits in both gene sets. However, there is no mutual exclusivity in the squamous cell carcinomas ($p = 0.99$, BH adjusted) with the LUSC, HNSC, and CESC cohorts having 4%, 8%, and 13%, respectively, of mutated tumors with simultaneous hits in the MHC genes and the APM. MSI tumors had even higher overlap with 38% of COAD-MSI and 43% of STAD-MSI tumors containing simultaneous hits, which does not support mutual exclusivity ($p = 0.99$, BH adjusted). These data suggest that allelic loss of HLA does not significantly reduce (or increase) the pressure to select for additional mutations limiting antigen presentation in solid tumors, but appears dominant in lymphomas. However, it is also possible that low mutual exclusivity is the result of high tumor heterogeneity, with multiple subclones having independent loss of APM function that only appear to co-occur due to bulk sequencing.

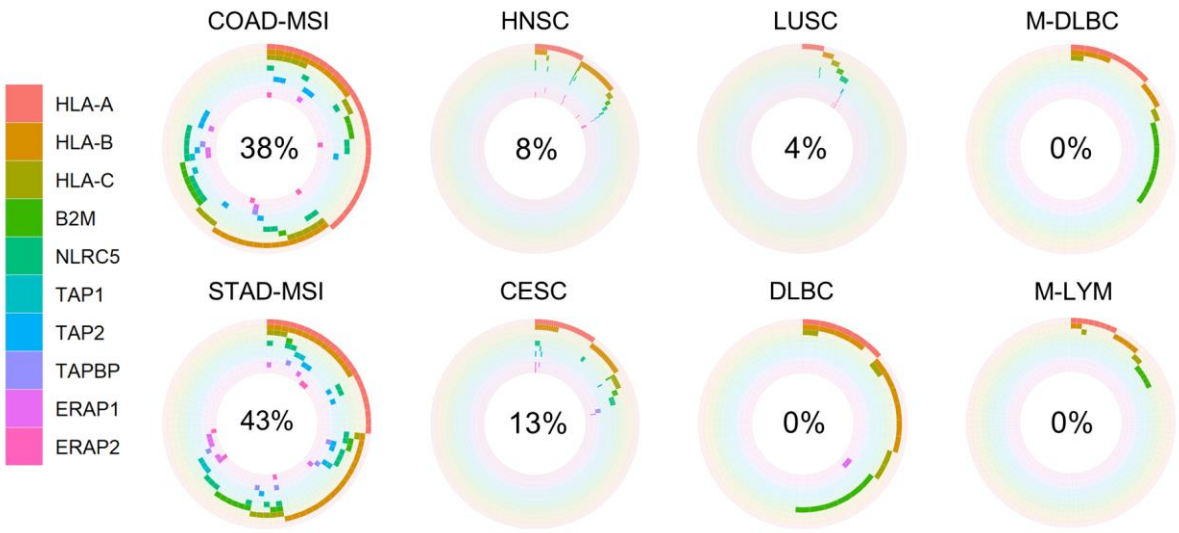

Figure 3-11: Mutual exclusivity of MHC and APM mutations
Association of HLA mutations and LOH, numbers indicate portion of mutated tumors (mutated | total) **(I)** Sample level co-occurrence of mutations in either the MHC class I or APM genes within positively selected cohorts. Percentage values show percent of mutated samples containing a hit in both the MHC class I and APM, with lower percentages suggesting mutual exclusivity.

Finally, we also looked at potential co-mutation of the MHC class I genes with known driver genes (**Figure 3-12A**). The strongest observed co-occurrence was found in the HNSC cohort with deleterious mutations of *CASP8*, a gene that plays a key role in an alternative pathway for destruction of malignant cells by the immune system[107]. This observation has been confirmed in an independent cohort of primary HNSCC (**Figure 3-12B**). Also observed in the HNSC cohort was co-mutation with HRAS, a member of the RAS family of oncogenes that has been shown to be associated with increased immune activity within head and neck cancers[108] **(Figure 3-12A)**.

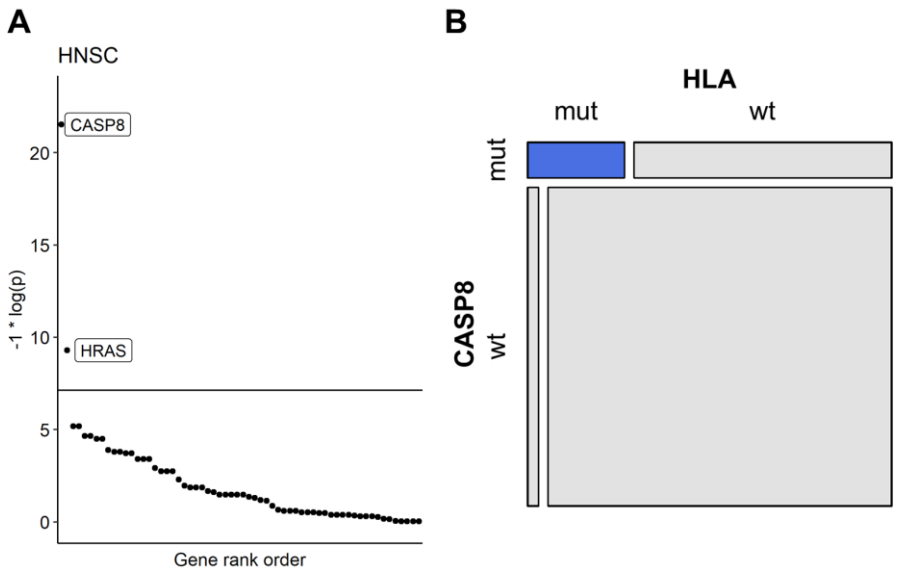

Figure 3-12: Co-occurrence of MHC mutations with CASP8 and HRAS in HNSC cancers
(A) Co-mutation analysis for functional mutations in all cancer gene census tier 1 genes vs MHC class I mutations in the HNSC cohort. Horizontal lines show significance cutoff after Bonferroni correction. Significantly co-mutated genes are labeled. **(B)** Pearson residual plot showing the enrichment of tumors with both CASP8 and MHC class I mutations in an independent cohort of 109 HNSCC tumors.

3.2.7 Mutational processes shape cancer type specific MHC mutational patterns

We next sought to determine which mutational processes may contribute to the generation of the non-synonymous mutations within cohorts showing evidence of positive selection. For MSI cancers, mismatch repair deficiency (MMRd) is the primary mutational process leading to a large number of indels within microsatellites[109]. We have already observed a high rate of frameshift indels within MSI tumors (**Figure 3-3B**) and notable hotspots (**Figure 3-1A**), and upon further

investigation the majority of these frameshifts (57/63, 90%) are due to single base pair insertions or deletions at homopolymer microsatellites (**Figure 3-13A**) which occur at a rate much higher than observed in MSS cancers ($p < 1e-16$) (**Figure 3-13B**). We also observed that MSI-associated indels were preferentially in longer homopolymers, while MSS-associated indels showed no relationship with homopolymer length ($p < 0.001$) (**Figure 3-13C**), which is consistent with the mutational signature for MMRd[110].

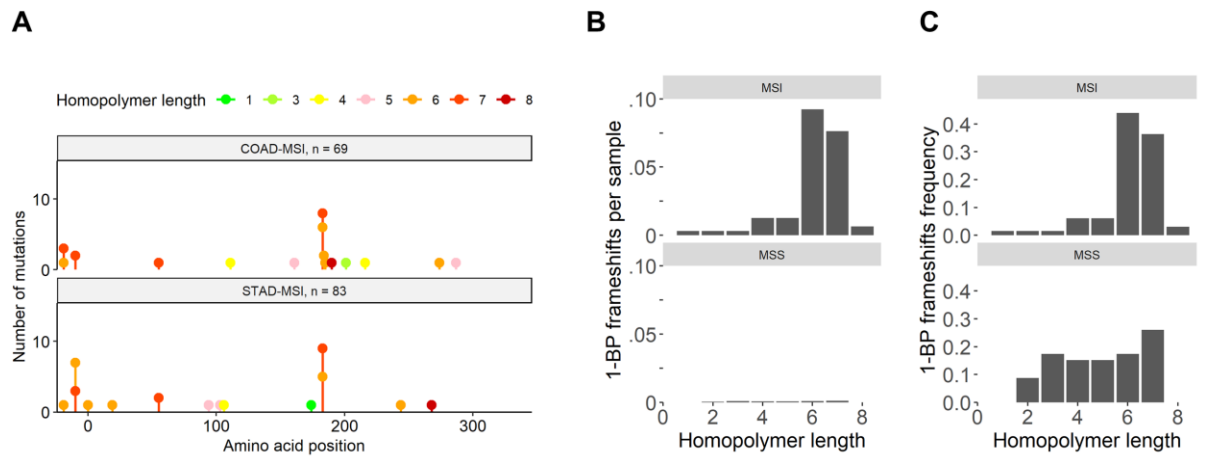

Figure 3-13: Frameshift hotspots in MSI cancers fall within coding region microsatellites
(A) Position of observed frameshift mutations within the MHC class I genes of COAD-MSI and STAD-MSI tumors. Colors show length of homopolymer microsatellites at each observed frameshift. **(B)** Total number of 1-BP frameshifts observed per tumor at homopolymers of varying length in MSI and MSS tumors. **(C)** Frequency of 1-BP frameshifts observed at homopolymers of varying length in MSI and MSS tumors.

For the lymphoma and squamous cell carcinoma cohorts, we observe a striking number of stop gain mutations, including multiple recurrent ($n > 2$) hotspot positions (**Figure 3-14A**). Interestingly, 100% of the stop gains that we observed in these cohorts are caused by C>T or C>A mutations. This is consistent with the well characterized process of cytosine deamination which is frequently observed in lymphomas due to activation-induced cytosine deaminase (AID)[111] and in squamous cell carcinomas due to the closely related APOBEC family of enzymes[112]. Both AID and APOBEC have distinct sequence preferences for their deaminase activity that should be visible in the sequence motifs surrounding each mutation. In the lymphomas we find that 13/22 (59%) of the observed stop gains match the canonical AID motif WRC (W = A/T, R = A/G) (**Figure**

-14A, B, C). Similarly, across the squamous cell carcinoma cohorts we find that 49/62 (79%) of observed stop gains match either the APOBEC3A/B/H/F motif TC or the APOBEC3G motif CCC (**Figure 3-14A,D, E**). Further analysis showed that mutational signatures SBS2 and SBS13, which have been reported to be associated with APOBEC activity[113], are significantly more active in squamous cell carcinomas with observed stop gain mutations (**Figure 3-14F**). A pan-cancer mutational signature analysis found no further significant associations between specific signatures and MHC mutations, with the exception of a small reduction in signature SBS5 in melanomas with MHC class II mutations (Wilcoxon rank sum test, BH-adjusted p = 0.03).

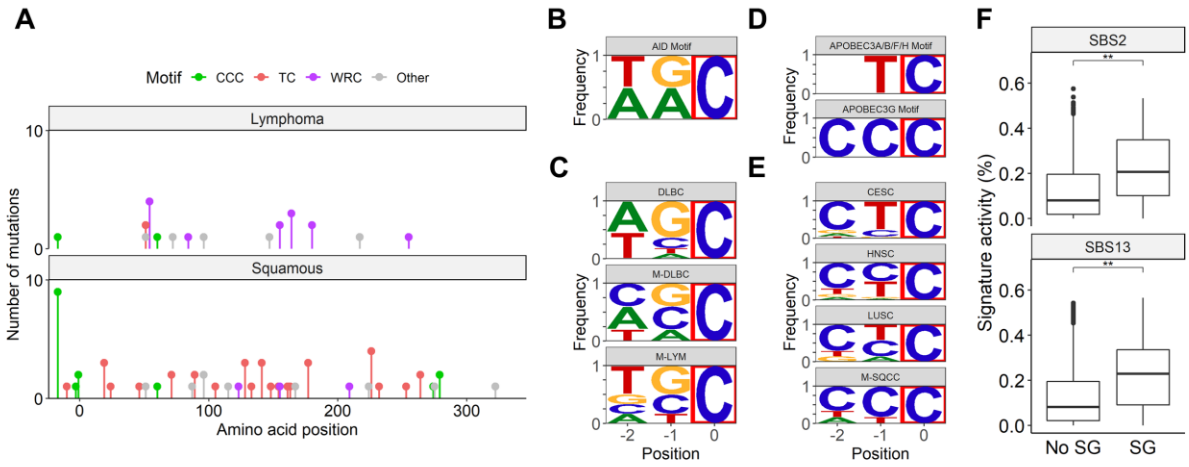

Figure 3-14: Stop gain hotspots match AID/APOBEC motifs
(A) Position of observed stop gain mutations within the MHC class I genes of lymphomas and squamous cell carcinomas. Colors show motif of mutated position. **(B)** Canonical motif for AID **(C)** DNA motifs for stop gain mutations observed in lymphoma cohorts. Mutated base marked with red box**. (D)** Canonical motifs for APOBEC proteins. **(E)** DNA motifs for stop gain mutations observed in squamous cell carcinoma cohorts. Mutated base marked with red box. **(F)** % of mutations within Squamous cell carcinomas with (SG) or without (No SG) observed stop gains that can be attributed to signatures SBS2 and SBS13, which have been associated with APOBEC activity. **: $p < 0.01$, BH corrected t-tests

Altogether, these observations strongly suggest that truncating mutations within the MHC genes originate due to specific mutational processes active within select cancer types. The active mutational processes are responsible not only for producing highly immunogenic tumors that are under pressure to select truncating mutations within the MHC class I genes, but are also directly responsible for creating the majority of the LOF mutations in the first place. Haplotypes harboring

homopolymer repeats and AID/APOBEC templates are therefore potentially more susceptible to this immune-escape mechanism.

3.2.8 Missense mutations are enriched in specific MHC functional domains

While frameshift and stop gain mutations are easy to classify as LOF, missense mutations are more difficult to interpret as they can be LOF, neutral, gain of function, or even neomorphic[114,115]. We hypothesized that deleterious missense mutations within positively selected cohorts should accumulate predominantly within the functional domains that provide the most immune escape potential for a tumor. To detect this enrichment we constructed two null models which were compared to the observed mutations across MHC functional domains which we established through systematic expert-knowledge and crystal structure guided annotation of individual amino acids within the MHC class I proteins (**Figure 3-15**). The first 'simulated' null model was based on a large number of HLA mutations generated randomly taking into account HLA sequence trinucleotide contexts and observed mutational signature activities within each positively selected cohort **(Figure 3-16A)**. The second 'observed' null was based on 251 actual mutations called in cohorts that showed no evidence of positive selection. We first compared the observed null dN/dS ratios to the simulated null ratios across all functional domains (**Figure 3-16B**). The log fold change of the observed *vs* simulated local dN/dS ratios were normally distributed (Shapiro-Wilk p = 0.99) with a mean not significantly different than 0 (p = 0.38, BH corrected), showing appropriately that there are no significant differences between the two null models. In contrast, in cohorts showing evidence of positive selection, the local dN/dS fold change was also normally distributed (Shapiro-Wilk p = 0.98, p = .433) but with a mean significantly above 0 when compared to both the simulated null (mean = 0.85, Cohen's d = 0.63, p = 0.03, BH corrected) and the observed null (mean = 1.13, Cohen's d = 0.81, p = 0.02, BH corrected) (**Figure 3-16B**). This shows a strong general trend towards excess nonsynonymous mutations across all annotation regions.

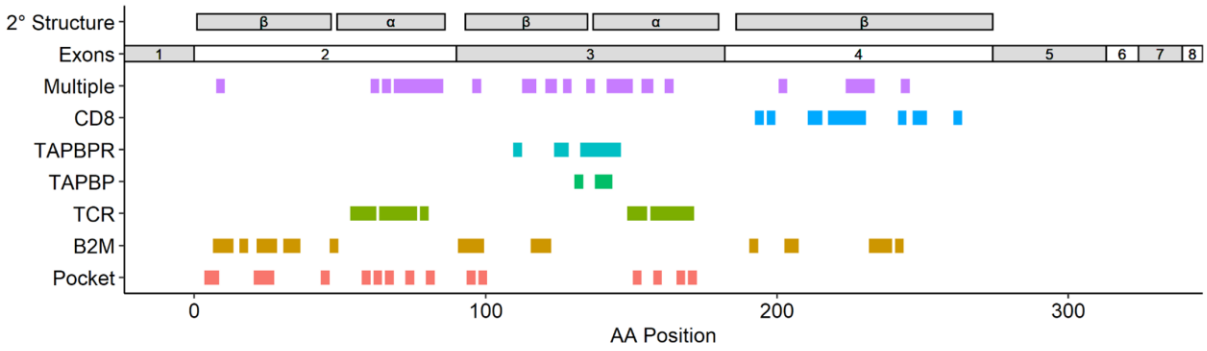

Figure 3-15: Annotation of class 1 HLA amino acids and their interacting protein partners
Schematic overview of HLA proteins showing secondary structure, exon boundaries, and amino acid interactions with various binding partners.

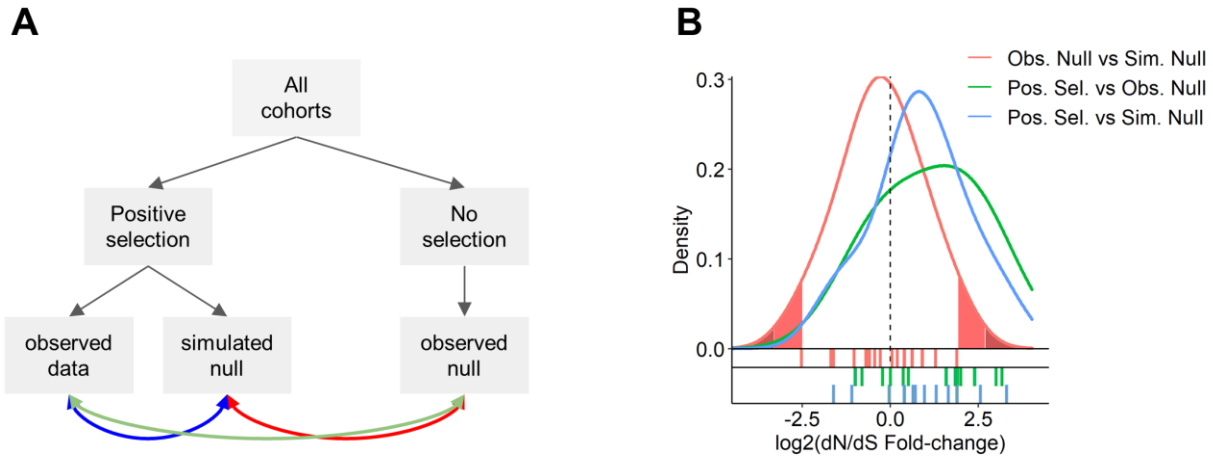

Figure 3-16: dN/dS ratio analysis within specific annotation regions
(A) Schematic representation of the constructed null models and their comparisons (arrows, colors match comparisons in panel C). **(B)** Distribution of observed vs simulated dN/dS ratio fold-change for amino acids that are predicted to interact with various MHC interacting partners. Observed dN/dS ratios are compared to dN/dS ratios from simulations taking into account the mutational signature activities within each of the cohorts showing evidence of positive selection. Rug plots show individual data points. Filled regions show tails of empirical null distribution. (light-red: top and bottom 5%; dark-red: top and bottom 1%)

Analysis of missense mutations in the context of functional domains acts not only as a signal for positive selection, but also helps in the interpretation of the mutations' likely functional consequences. Therefore, using our amino acid annotations, we examined which functional domains had the highest enrichment of non-synonymous mutations in both primary and metastatic tumors (**Figure 5D-E**). Compared to both the null models, 5 of the 7 annotated

functional domains showed a two-fold or higher enrichment of non-synonymous mutations. Some differences were noted between primary and metastatic tumors. Specifically, based on the 'simulated' null model, multi-functional residues are under a stronger positive selection in metastatic compared to primary tumors (**Figure 5D-E**), while on the other hand residues involved specifically in the B2M, TCR and CD8 binding interfaces show strong enrichment only in the primary tumors (**Figure 5D-E**).

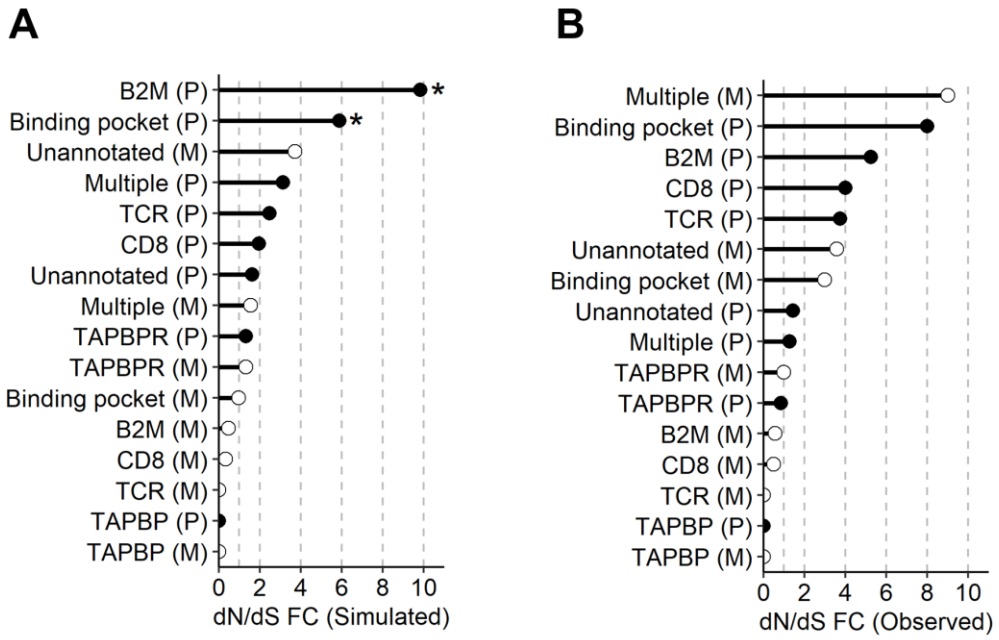

Figure 3-17: Significant increase in dN/dS ratios within specific amino acid annotation regions
(A,B) dN/dS ratio fold-changes vs simulated / theoretical **(A)** and observed / empirical **(B)** null models for individual annotation regions within cohorts showing evidence of positive selection (P: primary cancers; M: Metastatic/refractory cancers). Stars denote observations above the 95th percentile based on the observed null distribution.

3.2.9 Mutations at the B2M interface are predicted to disrupt MHC-B2M complex formation

In primary tumors, particularly striking are the B2M interacting and binding pocket residues displaying dN/dS ratios of 5 to 10 fold with respect to both null models **(Figure 3-17A,B)**. This suggests that within cohorts showing evidence of positive selection, missense mutations may be LOF by disrupting the ability of the mature HLA proteins to interact with either B2M or their cognate

neoantigen peptides. To examine this, we overlaid all missense mutations from the positively selected cohorts on the crystal-structure of the MHC class 1 - B2M complex[116] (**Figure 3-18A**), which revealed a clustering of recurrently mutated positions in 3D space at both the interface between the MHC and B2M proteins and at the anchor points of the peptide binding pocket (**Figure 3-18A**), strongly suggesting that mutations disrupt this interface leading to loss of MHC function. This is consistent with previous studies identifying B2M itself as a driver gene in all cancer types implicated here[101,117,118]. To determine whether missense mutations at the MHC class 1 - B2M interface are potentially deleterious we used SSIPe[119] to predict the change in binding energy resulting from each observed mutation in comparison to that of our previously 'simulated' null mutations (**Figure 3-18B**). Observed mutations had a significantly higher predicted ΔΔG than appropriately simulated mutations (median ΔΔG 1.37 vs 0.15, $p < 1e\text{-}4$, **Figure 3-19D**). Additionally, 42% of observed mutations had a ΔΔG > 1.5 kcal, the threshold suggested by SSIPe as evidence for significant disruption of a protein-protein interface. Overall, this suggests that the observed somatic mutations in residues at the MHC-B2M interface are more disruptive than expected by chance. Interestingly, even though *B2M* itself was enriched for loss of function mutations **(Figure 3-20A,B)**, there were no observed missense mutations in *B2M* in residues at the MHC class 1 interface within positively selected cohorts **(Figure 3-20C)**. This is consistent with previous reports that disruption of B2M protein interfaces are predominantly due to mutations within its interacting partners rather than B2M itself[120]. Altogether, these findings demonstrate that observed missense MHC class I mutations are strongly enriched at residues enabling MHC antigen binding complex formation, and have a likely deleterious function.

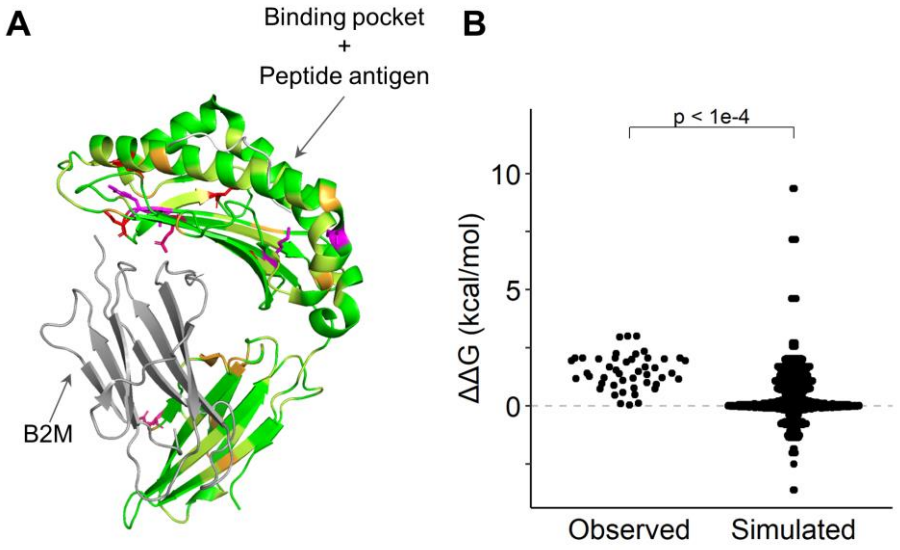

Figure 3-18 :Missense mutations disrupt the HLA:B2M binding interface
(A) Structure of the MHC/B2M complex showing 3D clustering of recurrently mutated amino acids. Positions within the MHC protein are colored based on the number of observed mutations (0: green, 1: yellow, 2: orange, 3: pink, 4: magenta, 5: red). Positions mutated 3 or more times are shown with side chains visible. B2M and bound peptide in grey. **(B)** Change in binding energy due to mutations in amino acids at the HLA:B2M interface as predicted by the software SSIPe.

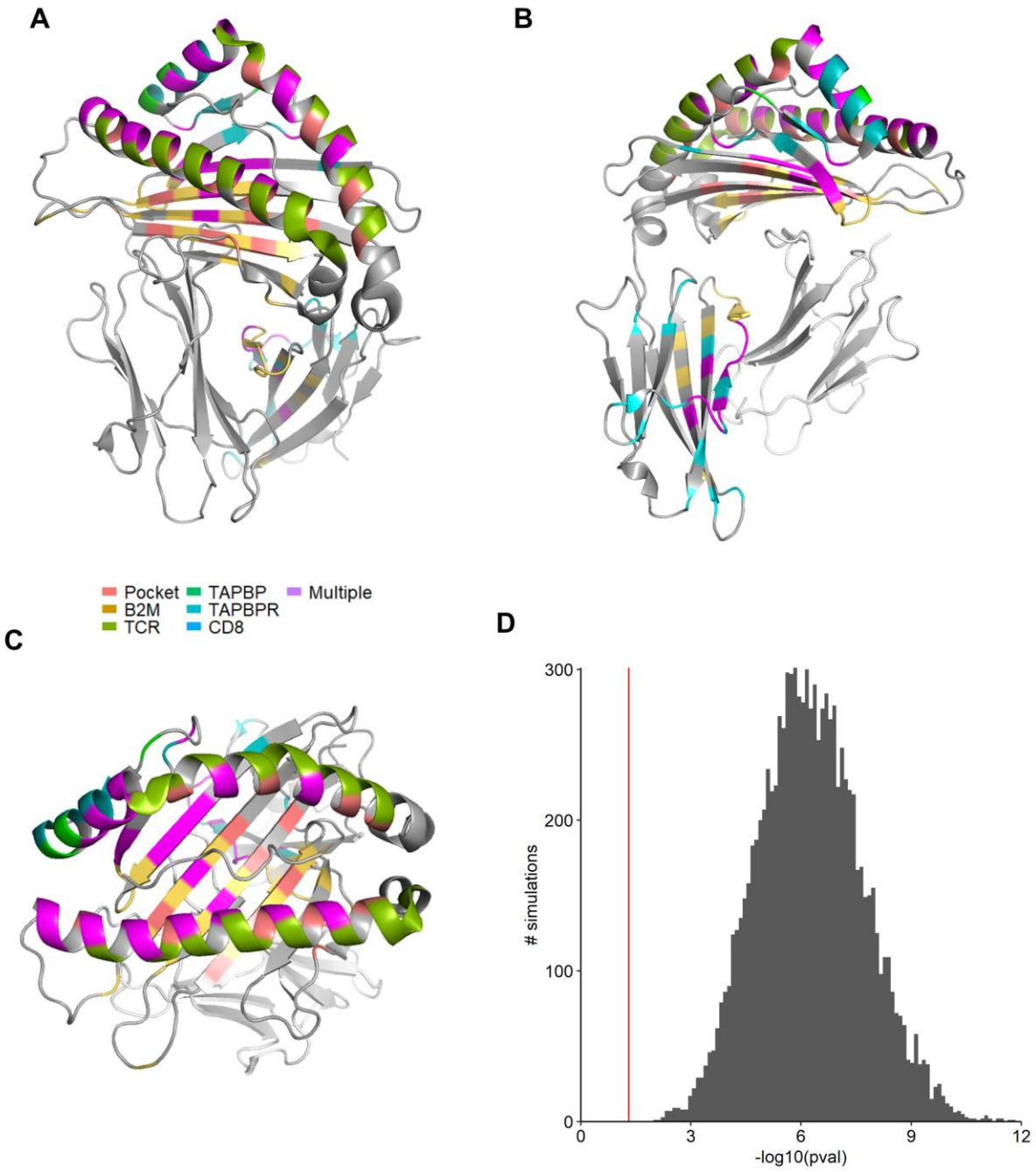

Figure 3-19: Mutations at the MHC1:B2M interface

(A-C) Multiple views of the MHC class I crystal structure. Residues are colored based on interactions with MHC associated proteins. **(D)** P-values (t-test) comparing observed mutations at the β2m interface to a set of random samples of the same size from a pool of simulated mutations. 10,000 random samples were taken and tested, all of which showed a significant difference (red line, $p < 0.05$) *i.e.* all random sets of mutations had significantly smaller $\Delta\Delta G$ compared to the observed ones.

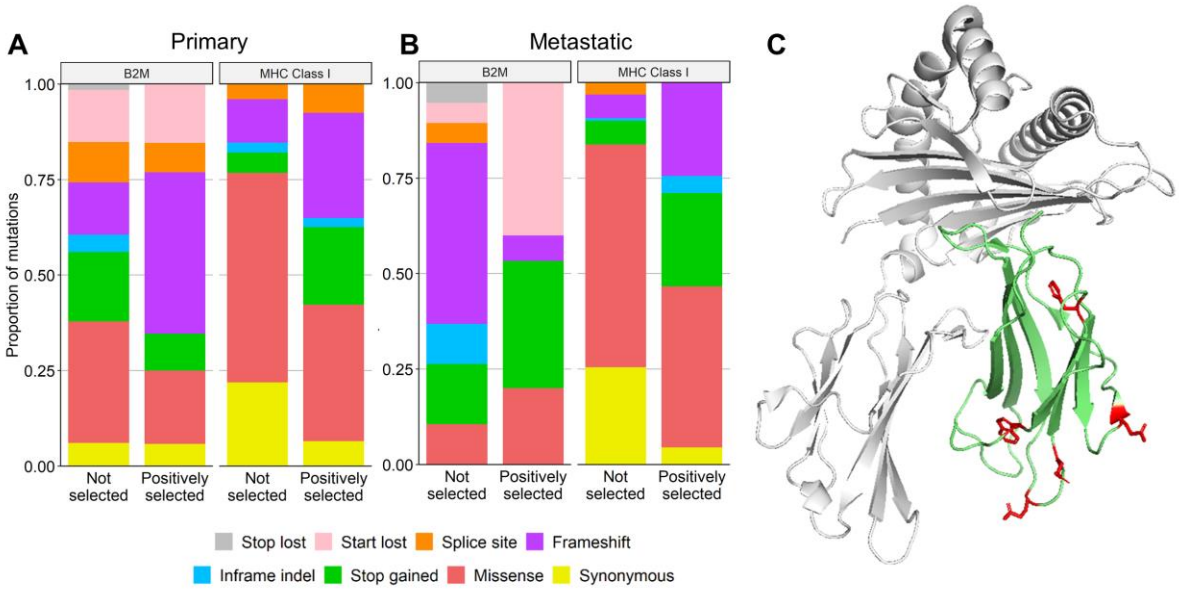

Figure 3-20: Mutations within B2M

(A, B) B2M shows a high rate of loss of function mutations in both **(E)** primary and **(F)** metastatic/refractory cancers that do or do not show evidence of positive selection for MHC class I mutations. Mutational consequence distributions of mutations within the MHC class I proteins are provided for comparison. **(C)** Crystal structure showing missense mutations (red) within the B2M protein in significantly mutated cohorts. Unlike mutations within the class 1 HLA proteins, which are enriched at the HLA:B2M interface, missense mutations within B2M were absent from this interface and only found in external residues. Green: B2M protein; White: Class 1 HLA protein.

3.3 Discussion

Despite overwhelming evidence from IHC studies demonstrating frequent loss of MHC expression pan-cancer[43], there is still much that is unknown about the molecular mechanisms that drive it. Using Hapster we identified the cancer types most affected by somatic mutation of the MHC and the mutational processes that promote and generate these mutations. We quantified positive selection and patterns of mutual exclusivity and independence of MHC mutations. We also characterized the deleterious consequences of truncating and missense mutations at the level of functional residues and protein domains.

We have identified six cancer types (colon and stomach adenocarcinomas with microsatellite instability; head and neck, cervical, and lung squamous cell carcinomas; lymphomas) that are significantly enriched for somatic non-synonymous mutations of the MHC class I. Notably, all of these cancers display above-average levels of tumor-immune infiltration[121], and at the pan-cancer level MHC mutant tumors are significantly more likely to have approved immunotherapies ($p < 2.2e-16$, OR: 2.29). A logical interpretation of these results is that at the cohort-level immunologically 'hot' tumors tend to both respond to immune checkpoint inhibitors (ICI) and mutate the MHC as an immune-escape mechanism. However, individual patients with impaired MHC function may be partially or completely invisible to T cells[43], making them less likely to respond to ICI. Since somatic loss of one or more MHC genes is essentially irreversible[122] it may preclude T cell based immunotherapies as viable treatments of MHC-mutated tumors. Within the above 6 cancer types, nearly 10% of patients harbor a functional mutation within the MHC class I genes, highlighting the necessity of further studies of MHC function in the context of both primary and acquired resistance to ICI.

We also provide a first look at MHC mutations in metastatic and refractory cancers obtained using personalized genomics. Metastatic cancers typically have a higher TMB, and therefore neoantigen load, and may be expected to be more immunogenic than primary tumors.

However, metastases may also originate from less immunogenic sub-clones in the first place, and seed locations that are immunosuppressive such as bone marrow[123] or liver[124], making them less visible targets for the immune system. IHC studies have also observed both MHC^+ primary cancers that produce MHC^- metastases, and MHC^- primary tumors that produce metastases that regain MHC function[125]. It is therefore unclear how pan-cancer mutational patterns in the MHC locus should compare to primary tumors. We show here that, broadly, MHC mutations in metastatic cancers mirror that of primary tumors. In refractory lymphomas there is a downward trend in MHC class I mutations when compared to primary, however this trend did not reach significance and evidence for positive selection still remains very high. Similarly, in the metastatic squamous cell carcinomas we see that there is no longer evidence for positive selection of MHC class I mutations as was observed in the primary cases. In contrast, we note a significant increase in MHC class I mutations in metastatic breast and prostate cancers. However, due to the limited metastatic sample size uncertainty remains If this reflects selection of biologically distinct molecular subtypes with higher MHC class I mutation burden, or MHC class I mutations themselves.

We also provide novel insights into MHC class II mutations pan-cancer. While MHC class II function is not strictly required for identification of neoantigens by $CD8^+$ T cells, it has been shown to regulate anti-tumor T cell responses[93] and can act as a therapeutic target for $CD4^+$ T cell based cancer vaccines[58,126]. While select cohorts (CHOL, KICH, UVM) did show evidence for positive selection of functional MHC class II mutations, there were overall fewer mutations in the MHC class II genes than in the MHC class I. This is expected as MHC class II expression is not constitutive and must first be induced before a tumor can be under pressure to select for a LOF mutation, which is in contrast to the MHC class I where selective pressures are present from the onset of tumor formation. What was not expected, however, was the lack of a relationship between cancer types that lose MHC class I and those that lose MHC class II function given that both sets of proteins act to present neo-antigens to T cells to promote an anti-tumor response. We observed

no overlap in cohorts showing evidence of positive selection for LOF mutations in the MHC class I and MHC class II genes, with MHC class I loss being more prevalent in high TMB cancer types and MHC class II loss being more prevalent in low TMB cancer types. It will take future studies to determine if this is driven by differences in MHC class II induction in different tissue types, differences in APC or $CD4^+$ T cell infiltration in the TME of different cancers, or if this is an actual relationship between low TMB tumors and MHC class II loss.

We have investigated the role of TMB overall as well as specific mutational processes in MHC mutagenesis. We identified three mutational signatures associated with mismatch repair deficiency, APOBEC activity, and AID activity as major drivers of MHC mutations in specific cancer types. We posit that for any given cancer under pressure to lose MHC function, it will be lost via "the path of least resistance" given the processes active in each different cancer type. In MSI cancers, squamous cell carcinomas, and lymphomas, the path of least resistance may be specific to the mutational processes active in these cancers. In contrast, there is melanoma which is known to have high TMB and to respond to immunotherapy, but is not observed to have many MHC mutations. However, transcriptional downregulation of MHC class I, as well as *B2M* mutations, are known acquired resistance mechanisms in this cancer type. Other cancers, such as prostate adenocarcinoma, are shown in IHC studies to have extremely high rates of loss of MHC expression [39], yet show no apparent somatic mutations in the MHC genes.

Altogether, our study demonstrates the high prevalence, positive selection, and deleterious nature of MHC mutations, and suggests that immune-escape through MHC mutagenesis is a common and early step in the progression of several common cancer types. While this study primarily focused on immunotherapy naive tumors, future studies looking at tumors post-immunotherapy may reveal MHC mutations to be drivers of acquired immunotherapy resistance. In this work we have focused on MHC mutations, however, in other cancer types, MHC function may be more easily lost via structural loss of the MHC locus, LOH[47], transcriptional

repression, or even post-translational inactivation. Given Hapster's ability to create personalized haplotype references, we aim in future studies to apply Hapster to other sequence based analytical methods to identify patterns of loss at the RNA and protein level. This will provide a deeper understanding of the diversity of mechanisms that can lead to loss of MHC expression, and as a result immune evasion and immunotherapy resistance.

3.4 Methods

3.4.1 Tumor mutational burden calculations

Global mutation burden was calculated using mutations obtained from MAF files provided by the Broad Institute's GDAC Firehose (https://gdac.broadinstitute.org/). Mutations observed within the capture kit region for each WES sample was divided by the total area covered by the capture kit. Capture kit information for each sample was obtained from official TCGA sample level metadata. Local mutation burden for each MHC gene was reported as the number of coding region mutations divided by the total exon length of each gene.

3.4.2 Positive selection of somatic mutations

Positive selection was evaluated by CBaSE[101], a tool that provides gene-specific measures of the strength of positive or negative selection for functional mutations based on the distribution of synonymous and non-synonymous variants after accounting for sequence contexts and cohort-specific mutational signature activities. CBaSE was run on each TCGA and MI-ONCOSEQ cohort separately. For MHC genes, CBaSE was run on mutations called by Hapster. For all other genes, CBaSE was run on mutations reported via MAF files obtained from the Broad Institute's GDAC Firehose (https://gdac.broadinstitute.org/). To calculate the pan-cancer metastatistic ϕ_{pos}, the cohort-level ϕ_{pos} reported by CBaSE was summed across all cohorts.

3.4.3 Cancer cell fraction calculations

The cancer cell fraction (CCF) of a variant was calculated using the following formula, where *f* is the VAF of the variant, *p* is the purity of the tumor sample, *m* is the multiplicity, and n_T is the total estimated copy number for the tumor:

$$CCF = \frac{f}{pm}((p * n_T) + 2 * (1 - p))$$

Mutations were considered clonal if CCF > 0.7.

3.4.4 Mutual exclusivity and co-mutation analyses

Both mutual exclusivity and co-mutation analyses were performed at the cohort level. Mutual exclusivity of MHC class I and APM mutations were evaluated using CoMEt[127]. Exact tests for mutual exclusivity were calculated by comparing any mutation with the MHC class I genes to any mutation within an APM gene. Co-mutation analyses were performed using a SNP-seq kernel association test as implemented in the R package SKAT[128]. Only driver genes listed in the COSMIC Tier 1 Cancer Gene Census[129] were considered. Within each cohort, only genes mutated in more than 2% of patients were analyzed.

3.4.5 AID/APOBEC mutational signature activity

Mutational signature activity within specific TCGA samples was obtained from the ICGC Pan Cancer Analysis Mutational Signatures Working Group (https://doi.org/10.7303/syn11726601). For each tumor within the CESC, HNSC, LUSC, and DLBC cohorts, the relative activity of AID/APOBEC associated mutational signatures SBS2 and SBS13 were identified. Cases were split into those either containing or not containing stop gain mutations within the MHC genes, and differences in SBS2/SBS13 activity were tested using t-tests.

3.4.6 dN/dS null model simulations

A simulated null model for the dN/dS ratio was created by simulating mutations using cancer-type specific background mutation rates. Mutations were simulated for each significantly mutated cohort by taking into account trinucleotide mutational signatures that are active in each cancer type. Mutational signature activity was obtained from the ICGC Pan Cancer Analysis Mutational Signatures Working Group (https://doi.org/10.7303/syn11726601). 10,000 mutations for each cohort were simulated as follows: 1) Each base along the length of HLA-A*01:01:01:01 was weighted based on its trinucleotide context, and the probability of that trinucleotide being mutated given known mutational signature activity. 2) A random position was picked based on the weighted probabilities of each base being mutated. 3) A random alternate base was picked for the selected position, with each potential alternate base weighted based on trinucleotide context and mutational signature activity.

3.4.7 MHC amino acid annotations

Protein interaction analysis was performed using annotations derived from structures of the relevant MHC class I complexes (PDB IDs provided below). The distance cutoff of the contacts was ≤4 Å. Structures used for MHC class I : peptide interaction and the interaction between MHC class I heavy chain and B2M were 4NQV, 6IEX, 3MGO, 3KPL, 3RL2, 3DX7, 4F7M, 1E28, 4HX1, 2RFX, 5EO0, 5IM7, 1XR8, 5W6A, 1JGE, 5VGE, 3LKR and 6JTP. Structure used for MHC class I : CD8 interaction was 3DMM. Structures used for MHC class I : TCR interaction were 5WKF, 4G8G, 5WKH, 4G9F, 5NQK, 5XOT, 6EQA, 4PRP, 3VXM, 6BJ2, 3W0W, 3MV7, 6AVF, 4JRX, 6AVG, 4JRY, 4QRP, 4PRI, 1MI5, 3DXA, 3FFC, 3KPR, 3SJV, 3KPS and 4QRP. Structures used for MHC class I : TAPBPR interaction were 5OPI and 5WER. Structure used for MHC class I : TAPBP interaction was 6ENY.

3.4.8 PPI binding energy predictions

Predicted changes to the MHC-B2M interface binding energy were generated using SSIPe[119]. All predictions were performed using the 3D crystal structure at https://www.rcsb.org/structure/4U6X. While the MHC class I proteins are highly polymorphic, the residues that make up the B2M interface are highly conserved, allowing us to predict binding energy changes using a single 3D structure regardless of which MHC protein the mutation was called in.

3.4.9 Quantification and statistical analysis

Total sample sizes for each cohort are provided in supplementary tables 5 and 6. Where sample sizes change due to data availability, N values are noted directly in figure legends. All statistics were performed in R. Tests used are specified directly in figure legends, alongside multiple-testing corrections used. Briefly, for contingency tables, Fisher's exact test was used unless sample sizes were too large, in which case Chi-squared tests were performed. For comparisons of two group means, t-tests were used unless the normality distribution assumption was not met, in which case Wilcoxon rank-sum tests were performed. For comparison of regression line coefficients, the Chow test was performed. For multiple testing corrections the Benjamini-Hochberg procedure was performed with a FDR of 5%, unless otherwise noted.

Chapter 4 Personalized Quantification of the HLA Proteins With HLAProphet

Portions of this chapter are available as a preprint[130], and are being compiled for submission for peer review.

4.1 Introduction

In chapter 2 I demonstrated a personalized method for alignment and mutation calling using DNA sequencing data, and in chapter 3 I showed the application of this method to two large pan-cancer cohorts to characterize the landscape of HLA somatic mutations. I also noted that due to the similarities between DNA sequencing and RNA sequencing, the Hapster algorithm naturally extends to RNA alignment and quantification. However, when attempting to quantify the HLA proteins, the nature of MS/MS data prevents the direct application of Hapster. Therefore, a personalized method for quantification of HLA protein expression is still needed.

A number of methods have been developed to allow identification and/or quantification of proteins containing variant sequences[131–136]. The general approach involves identification of germline or somatic variants using paired DNA sequencing data, *in silico* translation of these variant proteins, and then concatenation of these variant sequences with a standard protein reference to create an augmented search database. Unfortunately, these methods often restrict their analyses to single amino acid variants (SAAVs), which is not useful for the HLA proteins where we frequently observe multiple variants per tryptic peptide. Haplosaurus[137] improved upon SAAV approaches by allowing quantification of proteins with an arbitrary number of amino acid variants by imputing complete phased genotypes into all protein coding genes before *in silico* translation. However, this generalized approach requires high quality genotypes. This is a problem for the HLA genes where their polymorphism precludes traditional genotyping, and

where germline sequencing instead requires the use of specialized HLA typing software. This issue can be seen when looking at the 709 personalized HLA-A protein haplotypes inferred by Haplosaurus from 2504 samples from 1000 Genomes. Even though all 1000 Genomes samples have HLA types matching entries in the IMGT/HLA database[138], none of the personalized HLA-A proteins reported by Haplosaurus can be found in the database [IMGT/HLA release 3.51.0]. This is likely caused by the presence of at least one error in genotyping or phasing in each sample, leading to the *in silico* translation of non-existent proteins. This demonstrates that solutions for personalized HLA quantification require a specialized approach that works from full HLA types, not individual variant calls. To address this we have developed HLAProphet, an algorithm that leverages the FragPipe computational platform to provide personalized quantification of the HLA proteins from multiplexed tandem mass tag (TMT) based quantitative mass spectrometry data using known HLA types **(Figure 4-1)**.

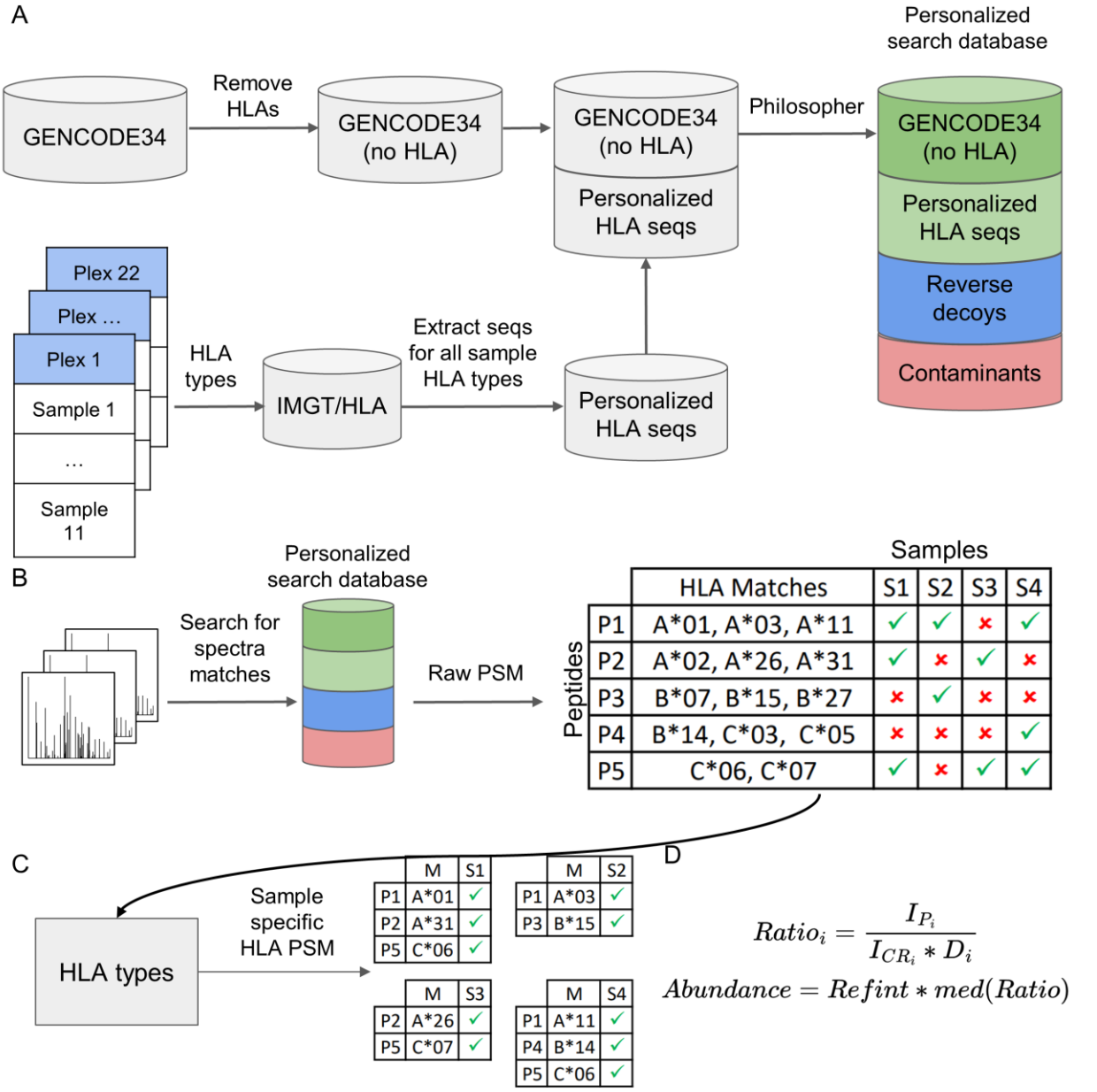

Figure 4-1: HLAProphet schematic overview

(A) HLA types for all samples across all plexes of a multiplex TMT MS/MS experiment are retrieved from the IMGT/HLA database and combined with GENCODE34, after removing GENCODE HLA sequences. This combined database is then run through philosopher to generate a final search database with reverse decoys and common contaminant sequences. **(B)** Searches for peptide matches for each MS/MS spectrum are performed using the personalized HLAProphet database to produce a PSM table with rows equal to the number of identified peptides, and columns equal to the number of samples in a plex. Notably, this table contains entries for all combinations of peptide to sample, even though only a subset of peptides are expected to be coded for given each sample's HLA type (green check) while the rest are not (red x). **(C)** The raw PSM table is broken down into individual sample tables, only retaining peptides predicted to be coded for in that sample's genome based on known HLA types, and with peptide-to-protein assignments restricted to only the HLA proteins in the HLA type (column M). **(D)** Peptide MS2 intensities (IP) are divided by common reference MS2 intensities (ICR) adjusted by a dilution factor (D, see methods) to generate a peptide intensity ratio. Protein abundances are calculated by multiplying the median peptide ratio across all assigned peptides to the reference intensity of that protein.

4.2 Results

4.2.1 Issues faced with standard reference based HLA proteomics

To demonstrate the issues faced by traditional proteomics methods when quantifying the HLAs, we used FragPipe and the GENCODE34[139] protein database to quantify all proteins in 108 tumors and 100 normal adjacent tissues from the CPTAC lung squamous cell carcinoma cohort[140], for a total of 208 samples. First, due to the polymorphic nature of the HLA genes, most HLA proteins will have tryptic peptides not found in a standard reference database. This is evidenced by the HLA proteins having a significantly lower fraction of predicted tryptic peptides identified than other proteins of similar size and abundance **(Figure 4-2)**.

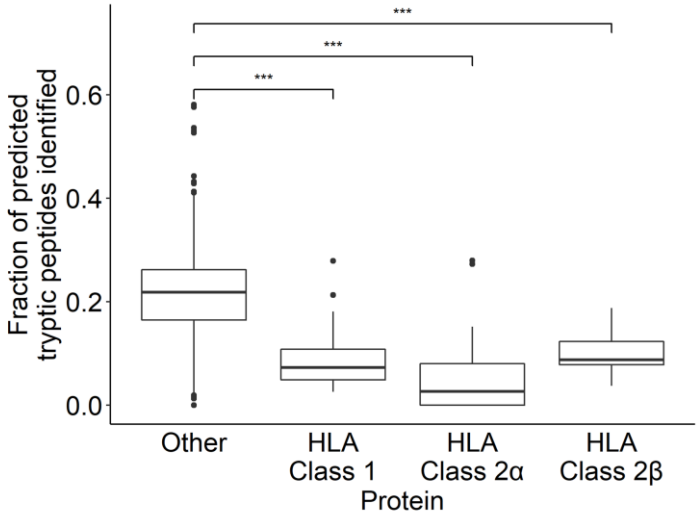

Figure 4-2: Fraction of tryptic peptides identified with standard reference based searches
Fraction of predicted tryptic peptides identified when performing standard reference (GENCODE v34) based searches to identify peptides in the CPTAC LSCC cohort (n = 208) using TMT mass spectrometry. Class 1, class 2α, and class 2β HLA proteins all have significantly reduced peptide identifications when compared to all other proteins of similar length and abundance. ***: Wilcoxon rank-sum test, $p < 0.001$

Second, when peptides are identified, abundance calculations are often done under the assumption that all uniquely mapping peptides are coded for twice due to the diploid nature of the human genome. To investigate the validity of this assumption for HLA tryptic peptides we use a measure called the allele count, which is the total number of alleles across all HLA genes

that code for the peptide **(Figure 4-3A)**. When we take all HLA tryptic peptides that appear to be uniquely mapping and diploid (allele count 2) based on the GENCODE34 reference and use HLA type information to calculate the true number of alleles coding for each peptide, we see that their true allele count is often different **(Figure 4-3B)**. We also show that incorrect identification of the allele count of a peptide within an HLA type is problematic given that the allele count directly relates to peptide abundance **(Figure 4-3C).** The most common issue (41% of identified peptides) is the identification of peptides that are not actually coded for in an individual's genome based on their HLA type (true allele count 0). This occurs in multiplexed samples when some samples code for a peptide while others do not. The peptide will be captured in MS1, and then a non-zero TMT MS2 reporter ion intensity value will be assigned to all samples in a plex, even if the abundance is actually zero, due to peptide co-isolation within the mass spectrometer[141]. The assumption that all samples are able to express peptides captured in MS1 does not apply here, and using these peptides would lead to an underestimation of HLA abundance. The next most common issue (35% of identified peptides) is when an individual is heterozygous for two different alleles of an HLA gene, with many unique peptides being coded for by only a single chromosome (true allele count 1). This clashes with the standard assumption that an identified peptide is coded for twice in a diploid human genome. When calculating gene-level abundances, these peptides will lead to underestimates since they only provide information about expression for one of the two alleles for that gene.

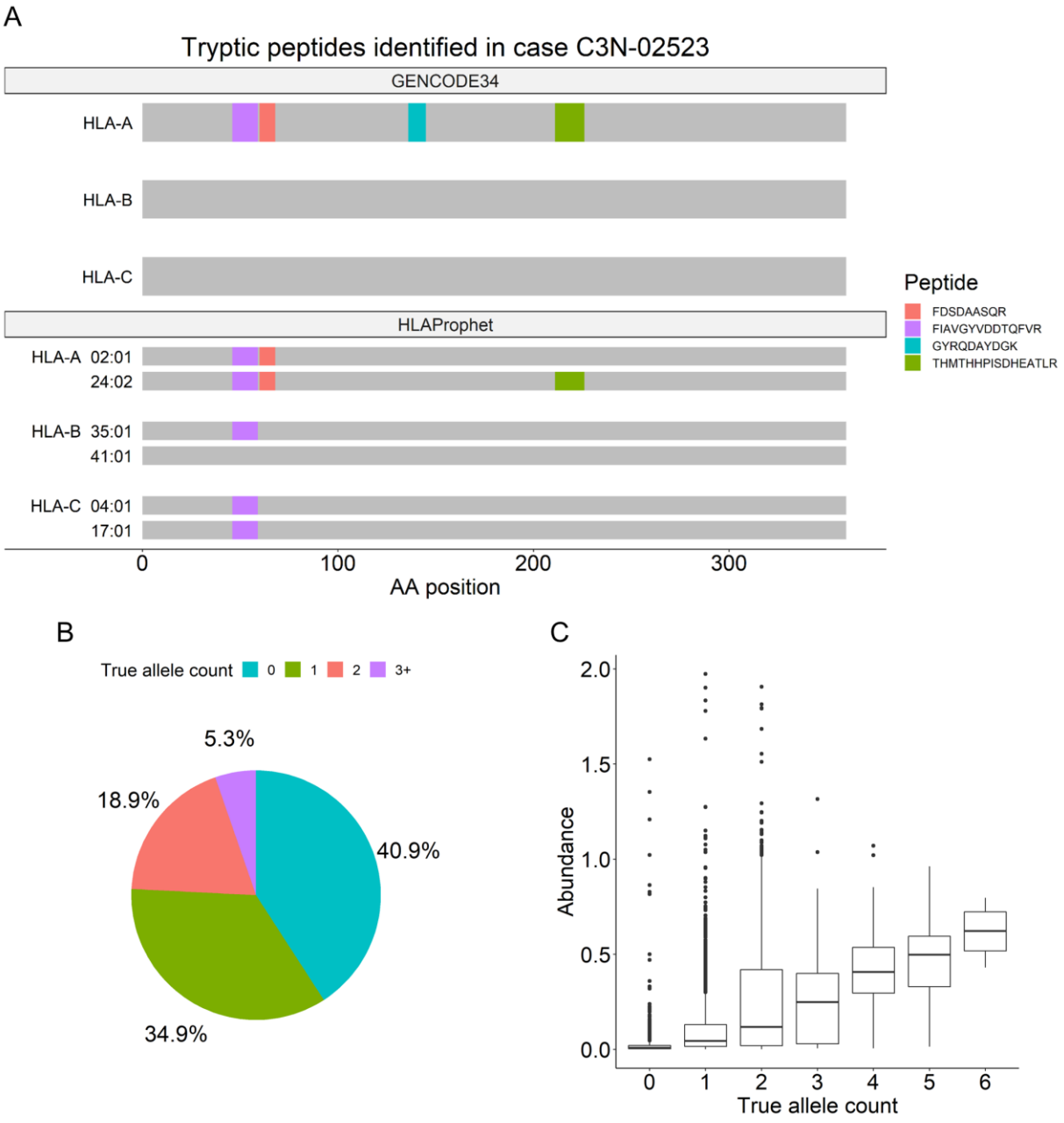

Figure 4-3: Allele count errors when identifying peptides using GENCODE34

Allele count errors when using standard reference (GENCODE v34) based proteomics to quantify HLA proteins in the CPTAC LSCC cohort (n = 208) using TMT mass spectrometry **(A)** Visualization of four peptides identified in a single CPTAC case, with peptide position within the class 1 HLA proteins colored. Positions are shown relative to either GENCODE34 (top) or HLAProphet personalized (bottom) reference sequences. All 4 peptides appear to be uniquely mapping to HLA-A using GENCODE34 and are assumed to be diploid (allele count of 2). The true number of alleles containing each peptide based on HLAProphet references are shown below. **(B)** True allele counts of all class 1 HLA tryptic peptides that are expected to be uniquely mapping and diploid (allele count 2) based on the GENCODE v34 protein database. **(C)** Association between abundance and true allele counts for HLA tryptic peptides that are expected to be uniquely mapping and diploid (allele count 2) using the GENCODE v34 database.

Finally, for the class 1 HLA proteins, some peptides within the relatively conserved backbone can be shared across genes, but in a way that is not captured in the standard reference database sequences. This causes the peptide to appear to be unique to an HLA gene, when in fact the allele count is higher than expected (true allele count 3+). For these peptides (5% of identified peptides), HLA abundance will be overestimated. Only 19% of peptides that appear to be diploid and unique (true allele count 2) based on GENCODE34 are correctly identified as such. The final issue with standard reference based methods is that they report a single abundance for each gene. However, for each HLA gene, heterozygous haplotypes will generally code for two unique proteins, each with a unique peptide presentation repertoire. Therefore, it is important for a personalized HLA quantification algorithm to report allele level protein abundances.

4.2.2 Improvements to HLA protein quantification with HLAProphet

To demonstrate HLAProphet's improvements in peptide identification and protein quantification, we applied it to the same set of 208 samples from the CPTAC LSCC cohort. HLAProphet uses HLA types derived from paired DNA sequencing data to generate a personalized HLA protein reference sequence database via *in silico* translation. We use here HLA types called by Hapster, but any source of HLA types can be used. The personalized HLA database is then concatenated with an existing reference protein database, here GENCODE34[139] with HLA sequences removed, and is run through the FragPipe quantification workflow. The use of HLAProphet's personalized HLA reference increases peptide identification by ~3-fold over GENCODE34 **(Figure 4-4A)**. These peptides cover the majority of each HLA protein **(Figure 4-4B)**, with the exception of the leader peptide which is absent from mature HLA proteins, and the transmembrane domain which is devoid of tryptic sites and produces a single long hydrophobic peptide that may be difficult to detect using tandem mass spectrometry[142]. We also find that within heterozygous samples, even though allele-specific peptides are a smaller

subset of all identified peptides, they still make up a large enough fraction of identified peptides to allow us to quantify individual allele abundances **(Figure 4-4C)**. Following peptide identifications, peptide spectrum matches (PSMs) are then used to quantify protein abundance using a personalized, HLA-aware quantification approach, with three major modifications relative to conventional algorithms such as TMT-integrator[143].

First, conventional algorithms for peptide to gene/protein roll-up of quantitative proteomics data, including TMT-Integrator, assume that all reference peptides are coded for in the genomes of all samples. This is not true for the HLA proteins. We therefore calculate HLA protein abundance on an individual basis, only using intensity values from peptides predicted to be present in a sample based on its HLA haplotype **(Figure 4-1B, C)**. At this step we also are able to filter out peptides with poor signal to noise ratios. Within a plex, we consider the MS2 intensity of a peptide in samples that do not code for the peptide to be purely the result of noise in the measurement. We expect this noise intensity to be lower than the MS2 intensity in samples coding for the peptide. This is often the case **(Figure 4-5A, C)**, but some peptides have poor signal to noise ratios **(Figure 4-5B)**. To filter out peptides with poor signal, we apply the Kolmogorov-Smirnov test and reject all peptides where the signal distribution is not significantly different than the noise distribution **(Figure 4-5D, E)**.

Second, a conventional strategy for TMT-based multiplexed quantitative proteome profiling is to normalize peptides by taking the ratio of a sample's MS2 intensity to a common reference (CR) created by pooling all samples from the experiment together. The CR acts as a physical average for tryptic peptide abundance, given that all samples have the peptide coded for in their genome. However, for the HLA proteins, some peptides will only be coded for in a handful of samples, and therefore their average abundance in a pool of the entire cohort will be diluted by the samples with HLA types that do not code for the peptide. These diluted CR MS2 intensities shrink the denominator of the ratio calculation, causing inflated ratios and overestimation of abundance for rare peptides **(Figure 4-6A)**. To address this issue, we multiply

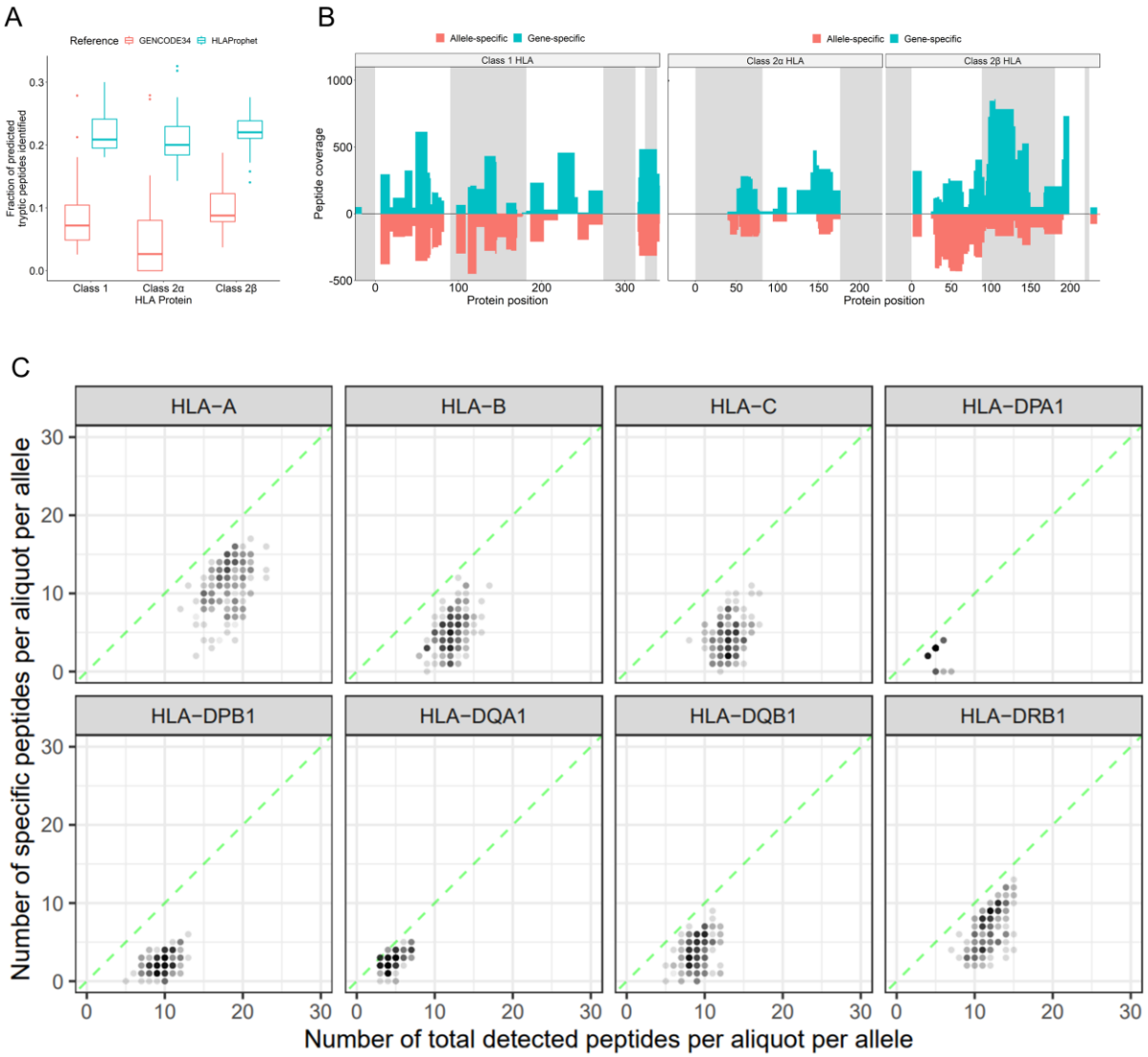

Figure 4-4: HLA Peptide identifications with HLAProphet

(A) Fraction of predicted tryptic peptides for HLA proteins identified using either the GENCODE34 protein reference or HLAProphet's personalized HLA reference. HLAProphet produces ~3x as many peptide identifications. **(B)** Peptide coverage for each position within the class 1, class 2α, and class 2β HLA proteins across all heterozygous samples. Blue bars show counts of gene-specific peptides (present in both alleles of one gene), while red bars show counts of allele-specific peptides. Alternating grey and white boxes show exon boundaries. Negative positions denote leader peptides, which are absent from the mature protein. **(C)** Fraction of peptides identified within heterozygous samples that are allele specific for each HLA gene. For each heterozygous HLA protein, a variable number of total peptides will be identified (shown on the x-axis). However, many of these will be found in both alleles of the gene, and cannot be used to differentiate between the two allele's abundances. To calculate allele-specific abundances, only those peptides specific to one allele can be used (shown on the y-axis).

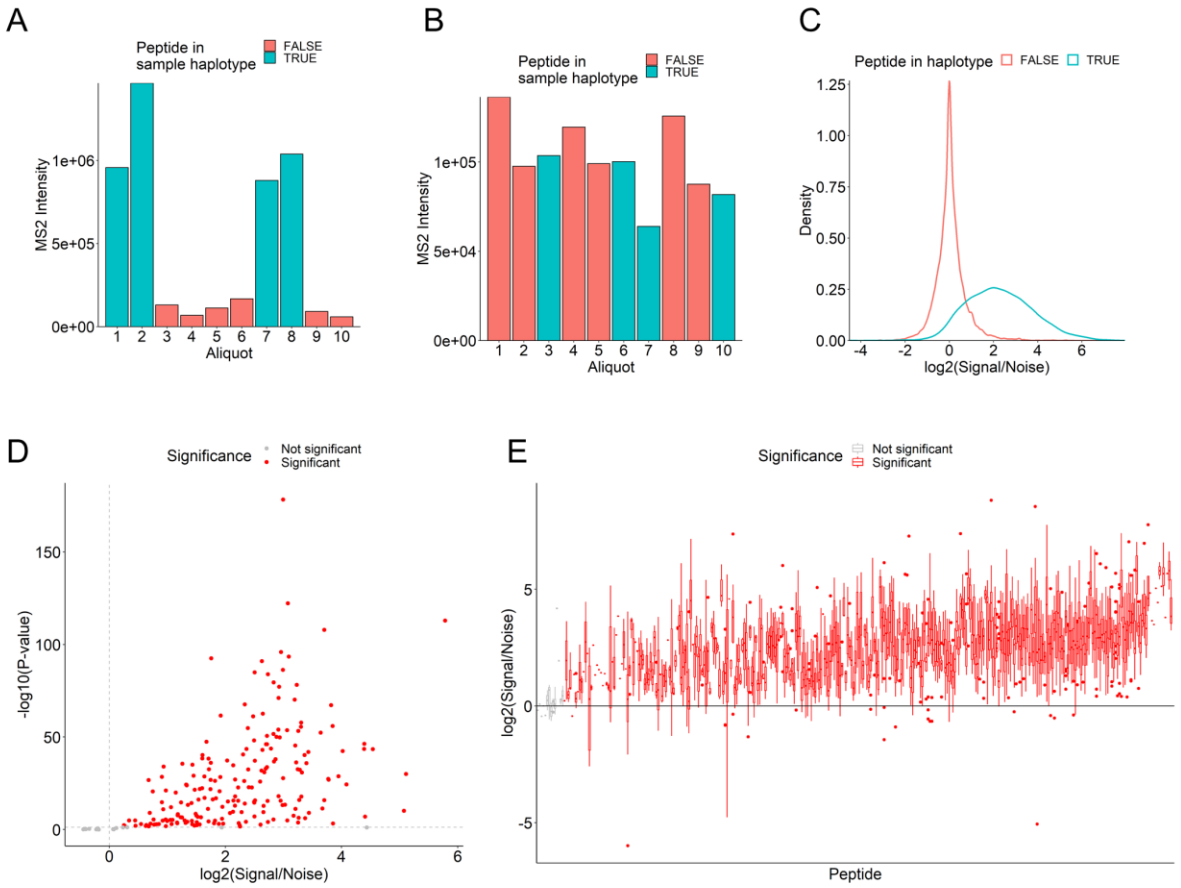

Figure 4-5: Detection of peptides with poor signal to noise ratios

(A, B) Example MS2 intensities for two different peptides measured in a 10-plex. Each aliquot is colored based on whether or not the peptide is predicted to be coded for in that individual's genome based on their HLA type. Shown are (A) an example peptide with good signal to noise ratios, and (B) an example peptide with poor signal to noise ratios. **(C)** Comparison of signal to noise ratios for 1,845 peptide identifications across the entire LSCC cohort. Noise is calculated as median MS2 intensity of aliquots that are not predicted to express a given peptide (red bars in E). **(D)** Results of applying the Kolmogorov-Smirnov (K-S) test to all HLA tryptic peptides to determine if the signal MS2 intensities come from a different distribution than the noise MS2 intensities, where noise is defined as the MS2 intensity within cases that do not code for the peptide. K-S test P-values are shown relative to the Signal/Noise ratios for the peptides, showing that only a small number of peptides fail this filter and are rejected (grey dots). **(E)** Signal to noise ratios for all peptides ordered by their K-S test P-values, from least to most significant. Only those cases with the most consistently low Signal/Noise ratio are rejected (grey boxplots).

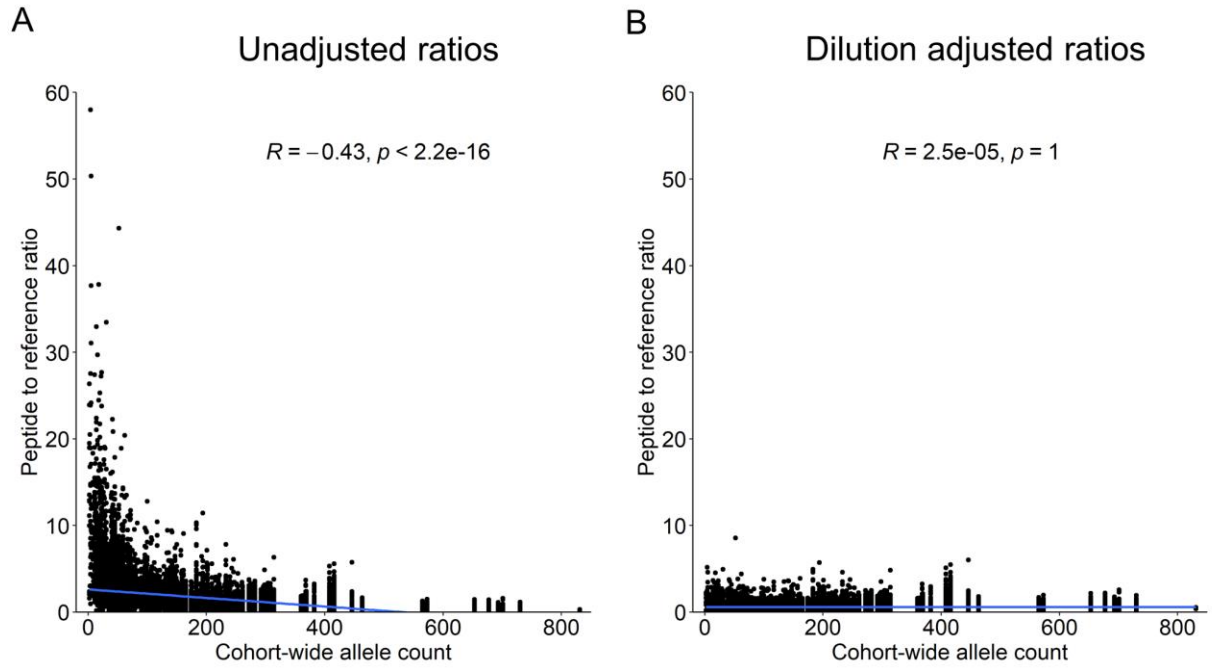

Figure 4-6: Dilution adjustment for rare peptides

(A) Ratio of peptide intensities to reference channel intensities for all identified HLA tryptic peptides compared to cohort-wide peptide allele counts for the CPTAC LSCC cohort (n = 208). A standard uniquely mapping diploid peptide would be found twice per sample for a total allele count of 416 across the cohort. HLA peptides have variable cohort representation, and rare peptides show a bias towards elevated ratios due to dilution in the reference channel which shrinks the denominator of the ratio calculation. **(B)** Ratio of peptide intensities to reference channel intensities, adjusted for reference dilution (see methods). The bias caused by cohort allele count is completely removed.

our CR ratios by a dilution factor **(see methods)**, removing the association between cohort population frequency and peptide CR ratio **(Figure 4-6B)**.

Third, when calculating gene level abundances for heterozygous HLAs, we will generally have two sets of peptides: those peptides that are coded for in only a single allele, and those that are coded for in both alleles. This is an issue given that the allele count of a peptide directly correlates to that peptide's abundance **(Figure 4-7A)**, with 1-allele peptides having lower abundance than 2-allele peptides. When calculating gene-level abundances, using 1-allele peptides directly is not appropriate, as it would cause an underestimation of gene abundance. However, expression differences due to variable allele counts are predictable and can be adjusted for. We see that 2-allele peptides do not quite reach double the intensity of 1-allele peptides, but consistently show an ~80% increase **(Figure 4-7B)**. Rather than excluding 1-allele peptides when calculating gene-level abundances, we multiply their intensity by the corresponding scaling factor (1.8) to get an estimate of what their intensity would be if they were present in both alleles (**Figure 4-7C-D**). This allows us to include 1-allele peptides in our total gene-level abundance calculations, providing a larger total peptide set and reducing the variance of the final values. Peptides with an allele count higher than 2 (~6% of all peptides) are excluded.

4.2.3 Benchmarking HLAProphet's HLA protein quantifications

To evaluate HLAProphet's HLA protein quantification, we compared protein abundance to paired personalized RNA expression for all samples. We see that when using the GENCODE34 reference for peptide searches and the traditional TMT-integrator algorithm for quantification, the class 1 HLAs show very low R^2 values of .017-.062 **(Figure 4-8A)**. For the class 2 HLAs, the correlation is slightly higher with R^2 values of .01-.59. HLAProphet's protein abundances show improved correlations to RNA in all cases, with the class 1 HLA R^2 values rising to .27-.50 and the class 2 HLA R^2 values improving to .58-.78 **(Figure 4-8B)**. We also see

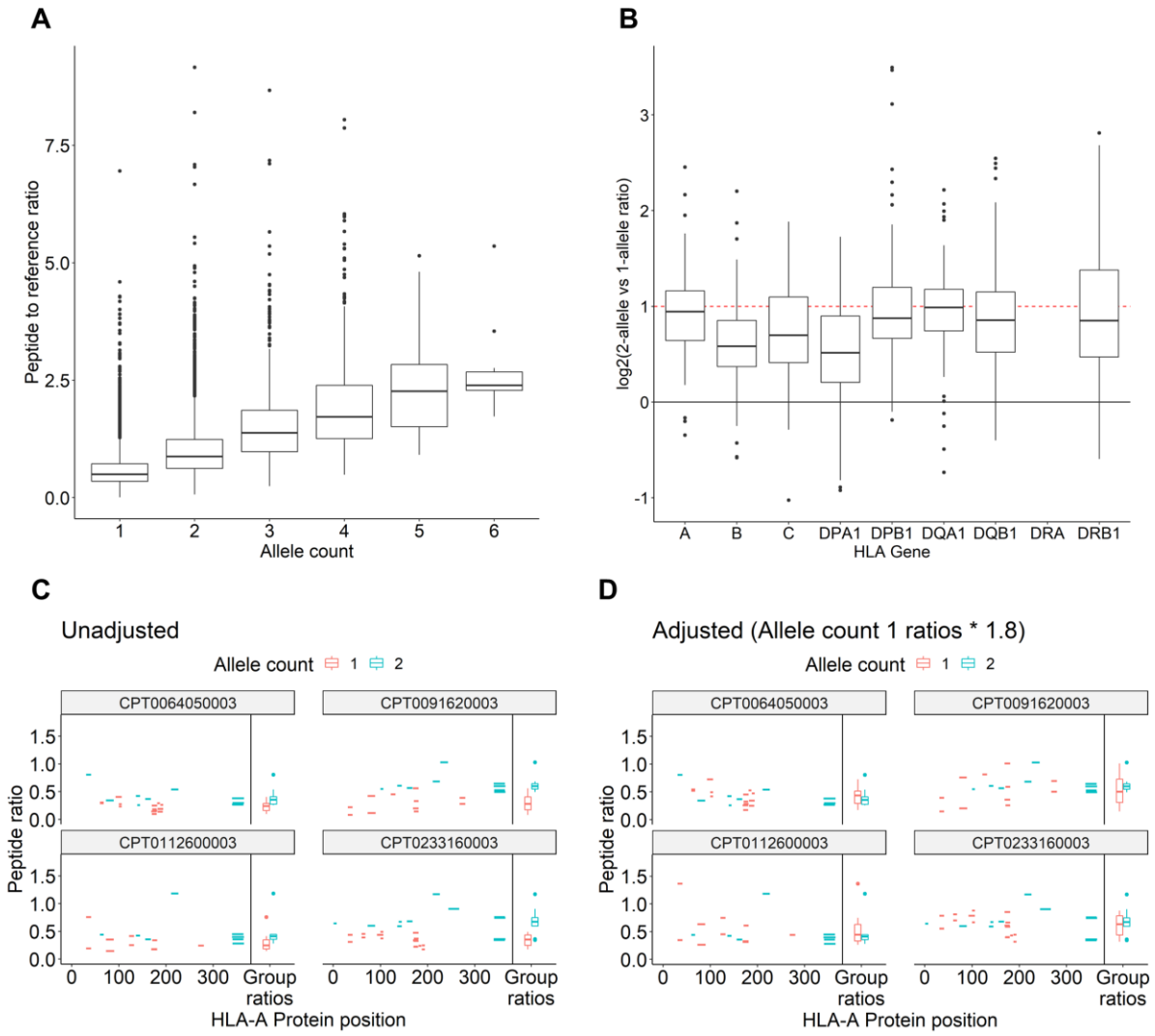

Figure 4-7: Ratio adjustment for allele specific peptides

(A) Peptide to reference ratios compared to allele counts for all class 1 HLA tryptic peptides identified using the HLAProphet search database. **(B)** log2 abundance ratio of peptides coded for twice to those coded for once within the same gene of a given sample. Red dotted line shows the threshold for 2-fold increase in abundance. **(C, D)** Four aliquot examples demonstrating the effect of allele count adjustment for gene-level abundance calculations. **(C)** Ratios of individual peptides across the length of HLA-A within four aliquots. Blue peptides are coded for in both HLA-A alleles within a given aliquot, while red bars are only coded for in 1 allele. Box plots on the right show the median ratio across all peptides in each group, with 1-allele peptides showing lower expression. **(D)** Same as in (C), but with 1-allele peptide ratios (red bars) increased by 80% (the experimental offset identified in (B)). This provides an estimate of what that peptide's abundance would be if present in both alleles, allowing them to be combined with real 2-allele peptides for final gene-level abundance calculations.

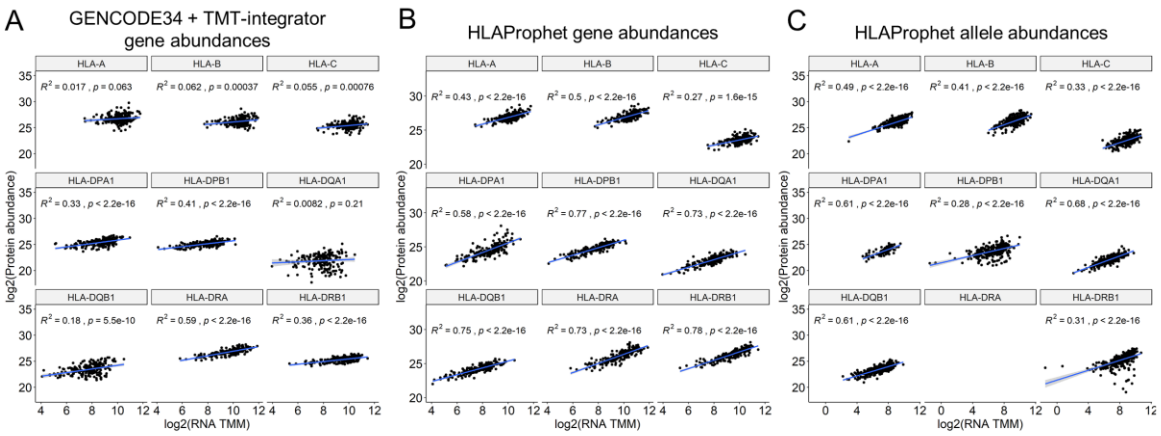

Figure 4-9: Correlation between HLA RNA and protein expression

Correlation of RNA expression to protein expression using GENCODE34 + TMT-integrator **(A)**, or HLAProphet's personalized quantification of gene level **(B)** and allele level **(D)** abundances for the CPTAC LSCC cohort. HLA-DRA is excluded from the allele level quantifications due to its low polymorphism, and low availability of allele-specific peptides.

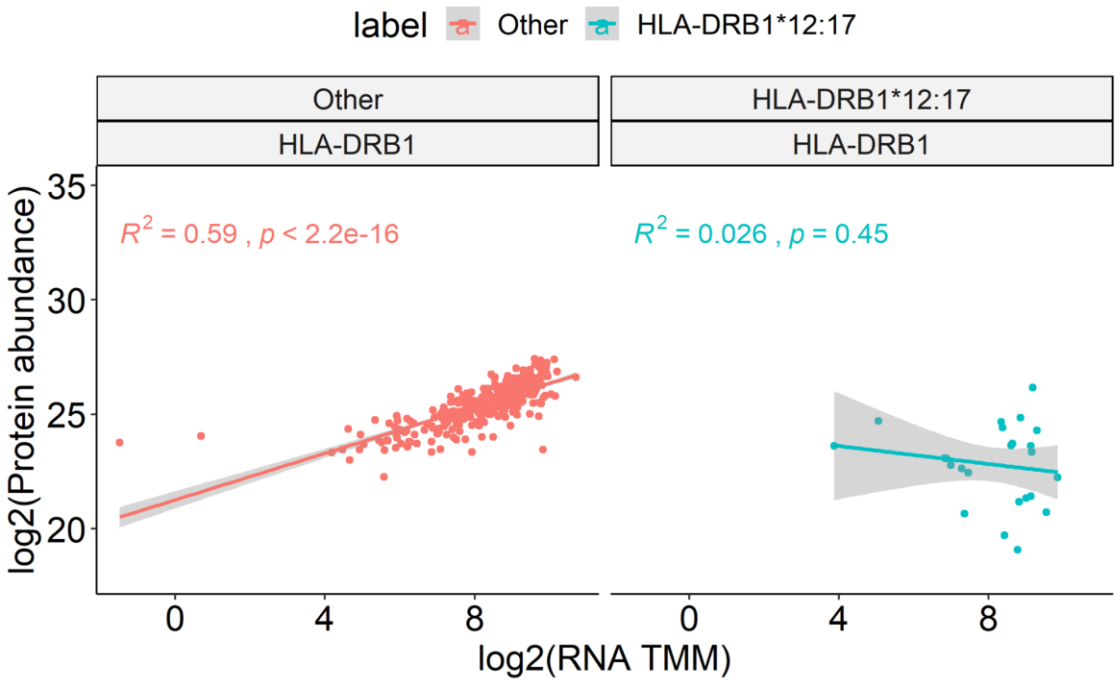

Figure 4-8: Lack of correlation between RNA and protein by HLA-DRB1*12:17

Correlation of allele specific RNA and protein expression for HLA-DRB1 (Data from Figure 2F) with allele HLA-DRB1*12:17 separated out. Most DRB1 alleles show strong correlation between RNA and protein (left, red), with the exception of HLA-DRB1*12:17 (right, blue) showing no correlation.

that at the allele level, protein expression remains highly correlated with allele-specific personalized RNA expression for all genes, demonstrating HLAProphet's ability to report allele level HLA protein abundances **(Figure 4-8C)**. We do, however, see a drop in R^2 for HLA-DPB1 and HLA-DRB1. Interestingly, for HLA-DRB1, this loss of correlation is almost completely due to a single allele HLA-DRB*12:17 **(Figure 4-9)**, which shows no correlation between RNA and protein expression. It will take further experiments to determine if this is an artifact or a real allele-specific effect. HLA-DRA has low polymorphism and does not contain sufficient allele-specific peptides to reliably quantify at the allele level, and so it was excluded from subsequent analysis.

To further demonstrate that HLAProphet can calculate allele-level abundances with minimal loss of information, we compare gene-level abundances calculated in two distinct ways. First, we calculate gene-level abundances as described. Then, for each heterozygous HLA type we calculate allele-level abundances and sum the abundance of the two alleles to get an alternate estimate of total gene abundance. We show that both methods produce a gene-level abundance with nearly perfect correlation (R^2 = 0.97, slope = 1, intercept not significantly different from 0, **Figure 4-10**), suggesting minimal information is lost when calculating allele level abundances, despite there being less peptides available for this measurement. Where measurements do show large differences, this is often due to only a single peptide being available for one of the calculations, which increases variance **(Figure 4-10, blue dots)**.

To provide further evidence for HLAProphet's ability to produce reliable allele level expression values, we examined allelic HLA protein abundances in tumor samples with loss of heterozygosity (LOH) events indicated by paired DNA sequencing data **(Figure 4-11)**. We see that for the class 1 HLA proteins, in tumors with LOH the lost allele has significantly lower expression than the retained allele, with an effect size that increases as tumor purity increases. In tumors with no observed LOH, there is no clear difference between the two alleles. For the

class 2 HLA proteins, which are not expected to be expressed in most tumor cells, we see no effect of LOH.

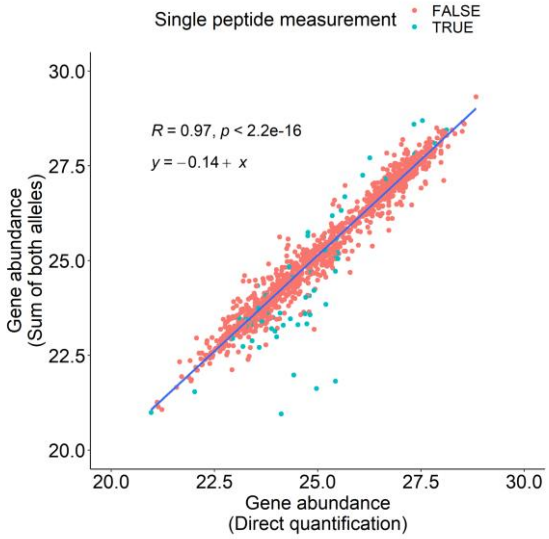

Figure 4-10: Correlation between allele-level and gene-level HLA protein abundances
Correlation between gene-level abundances calculated two ways. First, gene abundances are calculated directly as described in the methods. Second, gene abundances are calculated by summing allele-level abundances for all heterozygous HLA types. Blue dots show measurements based on a single tryptic peptide, which can cause high variance.

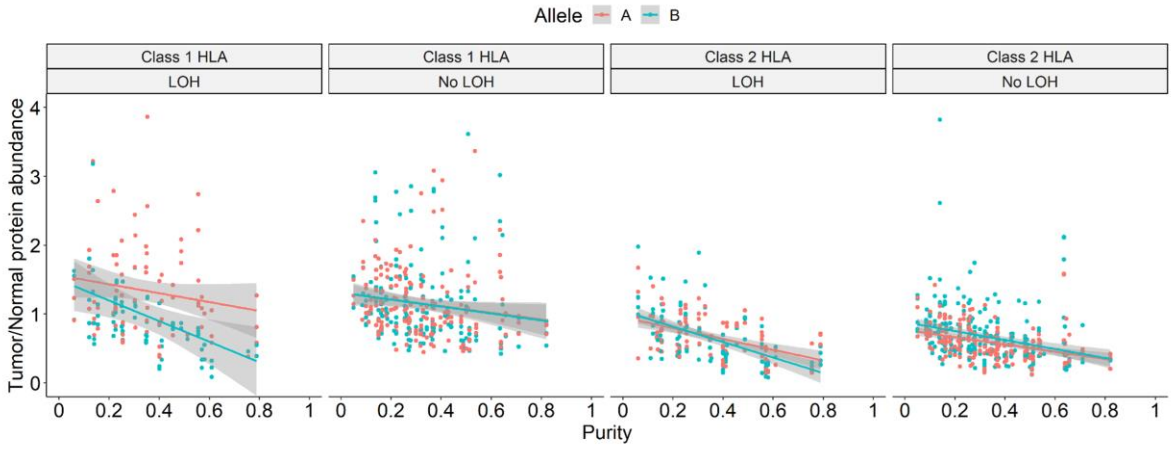

Figure 4-11: Allele specific loss of protein expression in cases of LOH
HLA protein abundance in tumors as a fraction of normal tissue abundance. The "B" allele is the allele that is lost following an LOH event, while the "A" allele is the retained allele. Allele labels are randomly assigned for cases with no observed LOH events. Significant differences between regression coefficients were tested using the Chow test.

To provide an initial look at HLA protein expression across multiple cancer types, we additionally applied HLAProphet to the CCRCC, HNSCC, LUAD, PDAC, and UCEC CPTAC cohorts. We show that in almost all cases HLAProphet's HLA quantification shows higher correlation with known antigen processing machinery (APM) proteins that should be under co-regulation **(Figure 4-12)**, suggesting that HLAProphet is improving HLA quantification across all cancer types. To provide a picture of allelic imbalance across cancers we identified cases where the difference in the tumor vs normal log2 expression fold change (Δlog2FC) between two alleles of an HLA gene was more than 3 standard deviations away from the mean Δlog2FC across all samples **(Figure 4-13)**. While overall HLA expression changes are confounded by changes in tumor purity and infiltrating immune cells, we expect allelic imbalance to be a tumor-cell specific phenomenon. We show that this is most evident in the HNSC (Class I: 6%, Class II: 8%) and PDAC (Class I: 6%, Class II: 18%) cohorts. However, we note that while tumor purity does not confound allelic imbalance, it does mute the signal as the observed Δlog2FC will be lower in a low purity sample, giving this thresholding approach relatively low sensitivity. Additionally, no distinction is made here between allelic imbalance as a result of allelic gains vs losses. More nuanced analyses will be required moving forward to provide a complete picture of HLA allelic imbalance at the protein level.

4.3 Discussion

In summary, we present here HLAProphet, a tool that enables personalized allele level quantification of the HLA proteins from TMT-MS/MS data. We show that HLAProphet improves upon the existing state-of-the-art standard reference based approaches by significantly increasing peptide identifications, and by providing protein expression values that show higher concordance with RNA expression and known genomic events. Moving forward, HLAProphet

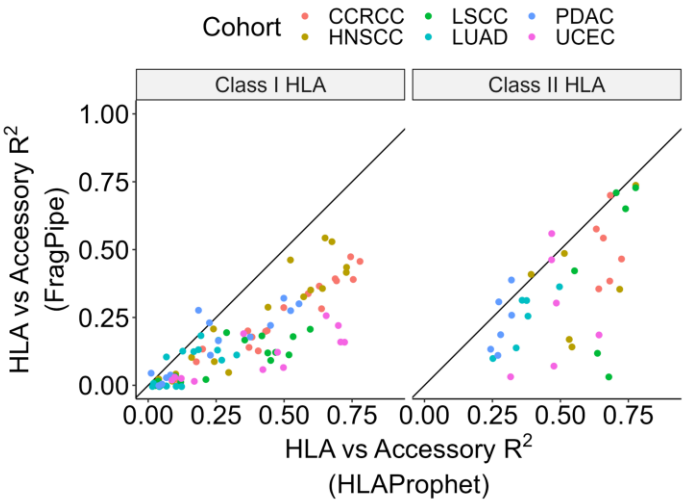

Figure 4-12: Correlation of HLA expression to accessory proteins

R^2 values for correlation between class I and class II HLA proteins as quantified by either FragPipe or HLAProphet, compared to APM proteins which should be under co-regulation. Data are shown for 6 CPTAC cohorts (N = 100-212 per cohort) Class I APM proteins: CANX, CALR, TAP1, TAP2, TAPBP, PDIA3. Class II APM protein: CD74

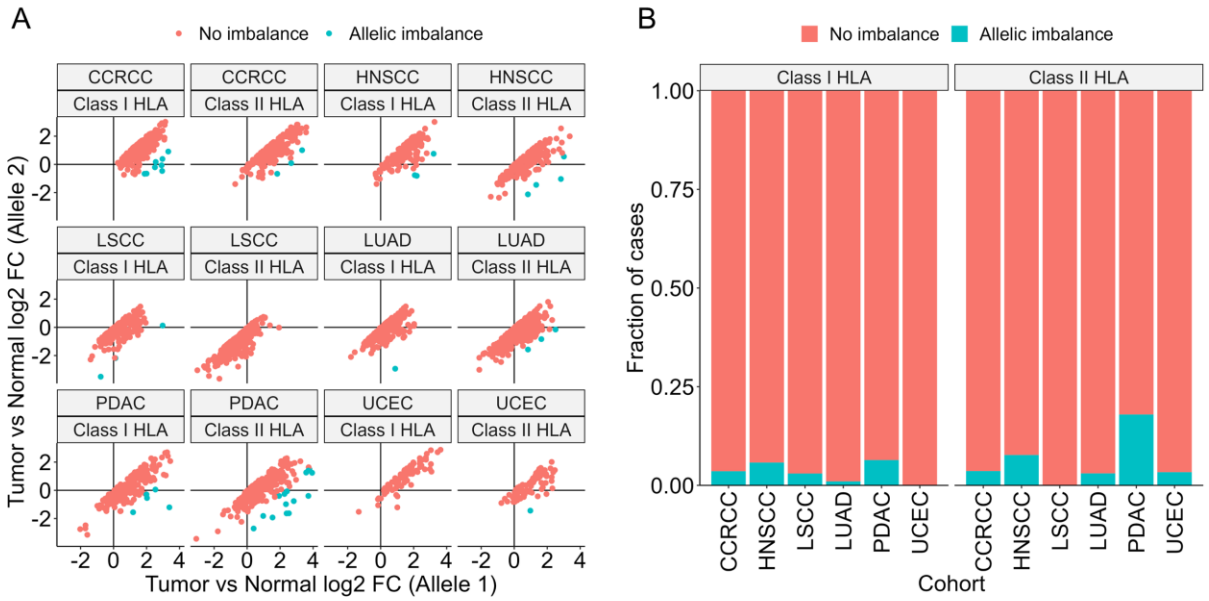

Figure 4-13: Pan-cancer HLA allelic imbalance

Allelic imbalance across 6 CPTAC cohorts. Samples are considered imbalanced if the Δlog2FC are more than 3 standard deviations from the mean Δlog2FC across all genes and all samples. (N = 100-212 per cohort). (A) log2FC values for the A allele (Allele 1) and the B allele (Allele 2) for all proteins across all samples. (B) Fraction of cases from each cohort showing at least one allele with an imbalance.

will enable the study of HLA loss at the protein level in tumors, providing greater insights into how cancers evade T-cell surveillance.

In addition to the use cases presented here, HLAProphet could also be used for studies of post-translational modifications (PTMs) of the HLAs. The HLA proteins are known glycoproteins, with sequential modification of the main PTM being thought to be required for the efficient folding, loading, and export of the mature MHC[144]. HLAProphet's personalized protein database should naturally extend to PTM quantification methods, allowing for insight into potential disruptions of this glycosylation. Further, as a generalized algorithm HLAProphet is not restricted to cancer studies, and could empower investigations into germline associations between specific HLA alleles and infection or autoimmunity.

4.4 Methods

4.4.1 Code availability

HLAProphet code is available at https://github.com/mctp/HLAProphet

4.4.2 Data acquisition

Molecular data for the CPTAC lung squamous cell carcinoma cohort were generated as described in Satpathy et al.[140]. DNA and RNA sequencing reads can be downloaded from the NIH GDC portal at https://portal.gdc.cancer.gov/repository. For proteomics searches, raw mzML files were downloaded from the CPTAC Data portal at https://proteomic.datacommons.cancer.gov/pdc/study/PDC000234. In total, paired DNA, RNA, and proteomics data were available for 108 tumors. From normal adjacent tissue, DNA sequencing data were available for all cases, RNA sequencing data were available for 95 cases, and TMT proteomics data were available for 100 cases.

4.4.3 Fixed reference proteomics searches

Protein extraction and tryptic digestion, as well as common reference pool construction, TMT-11 labeling and LC-MS/MS workflow were described in Satpathy et al[140]. Briefly, a total of 108 LSCC tumor samples, 100 paired normal adjacent tissue (NAT) samples, 22 aliquots from a common reference pool, and 8 other tumor samples were assigned to 22 TMT 11-plex sets. All of the tumors were in the C channels, all of the normals were in the N channels, and CRs were in the 11th channel of each plex.

Raw mass spectrometry files were converted into open mzML format using the msconvert utility of the Proteowizard software suite, and analyzed using FragPipe computational platform (fragpipe.nesvilab.org) using the TMT11-bridge workflow. MS/MS spectra were searched using the database search tool MSFragger v3.7[145] against a harmonized *Homo sapiens* GENCODE34 protein sequence database appended with an equal number of reverse decoy sequences. Whole cell lysate MS/MS spectra were searched using a precursor-ion mass tolerance of 20 ppm, and allowing C12/C13 isotope errors −1/0/1/2/3. Mass calibration and parameter optimization were enabled. Cysteine carbamidomethylation (+57.0215) and lysine TMT labeling (+229.1629) were specified as fixed modifications, and methionine oxidation (+15.9949), N-terminal protein acetylation (+42.0106), and TMT labeling of peptide N terminus and serine residues were specified as variable modifications. The search was restricted to tryptic peptides, allowing up to two missed cleavage sites. Peptide to spectrum matches (PSMs) were further processed using Percolator[146] to compute the posterior error probability, which was then converted to posterior probability of correct identification for each PSM. The resulting files from Percolator were converted to pep.xml format, and then processed together to assemble peptides into proteins (protein inference) using ProteinProphet[147] run via the Philosopher toolkit v4.8.1[148] to create a combined set of high confidence protein groups. The combined prot.xml file and the individual PSM lists for each TMT experiment were further processed using the Philosopher filter command as follows.

Each peptide was assigned either as a unique peptide to a particular protein group or assigned as a razor peptide to a single protein group that had the most peptide evidence. The protein groups assembled by Percolator were filtered to 1% protein-level False Discovery Rate (FDR) using the target-decoy strategy and the best peptide approach (allowing both unique and razor peptides). The PSM lists were filtered using a sequential FDR strategy, keeping only those PSMs that passed 1% PSM-level FDR filter and mapped to proteins that also passed the global 1% protein-level FDR filter. In addition, for all PSMs corresponding to a TMT-labeled peptide, reporter ion intensities were extracted from the MS/MS scans (using 0.002 Da window) using Philosopher and the precursor ion purity scores were calculated using the intensity of the sequenced precursor ion and that of other interfering ions observed in MS1 data (within a 0.7 Da isolation window).

4.4.4 Fixed reference based abundance quantification

The PSM output files were further processed using TMT-Integrator v2.1.5 to generate summary reports at the gene level and modification site level. TMT-Integrator[143] (https://github.com/Nesvilab/TMT-Integrator) used as input the PSM tables generated by the Philosopher pipeline as described above and created integrated reports with quantification across all samples. First, PSMs were filtered to remove all entries that did not pass at least one of the quality filters, such as PSMs with (a) no TMT label; (b) precursor-ion purity less than 50%; (c) summed reporter ion intensity (across all channels) in the lower 5% percentile of all PSMs in the corresponding PSM.tsv file. In the case of redundant PSMs (i.e., multiple PSMs in the same MS run sample corresponding to the same peptide ion), only the single PSM with the highest summed TMT intensity was retained for subsequent analysis. Both unique and razor peptides were used for quantification, while PSMs mapping to common external contaminant proteins (that were included in the searched protein sequence database) were excluded. Next, for each PSM the intensity in each TMT channel was converted into a log2-based ratio to the reference channel.

The PSMs were grouped to various summarizing levels (i.e., peptide, protein, gene), and summarizing group ratios were computed as the median of the corresponding PSM ratios after outlier removal. Ratios were then converted back to absolute intensity in each sample by using the reference intensity estimated by the median of weighted sum of the MS1 intensities of the top 3 most intense peptide ions for each plex.

4.4.5 HLA typing

HLA haplotypes were inferred from paired WES DNA sequencing data using the Hapster software as described in Mumphrey et al[73].

4.4.6 HLAProphet personalized protein reference construction

Protein sequences for HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRA, and -DRB1 were downloaded from the IMGT/HLA database (Release 3.51.0, https://github.com/ANHIG/IMGTHLA). For each sample, protein sequences for each allele of each gene in the predicted HLA type were taken from the IMGT/HLA database and added to a common fasta file. For HLA types with different IDs but identical protein products, a single sequence is included in the final reference with a harmonized type ID to avoid duplicate protein sequences in the final database. A relationship database is provided with a list of all original HLA types and their harmonized IDs, so that abundances can be tied back to their original HLA type during post-processing. The HLAProphet reference fasta was then combined with GENCODE34 (HLA sequences removed) using Philosopher, which adds reverse decoy and common contaminant sequences.

4.4.7 Proteomics searches using personalized HLAProphet databases

Proteomics searches and PSM filtering were performed as described above, using the HLAProphet personalized protein reference in place of the original GENCODE34 protein database.

4.4.8 Personalized protein abundance calculations

Personalized HLA abundances were calculated from PSM tables using the TMT-integrator filters and workflow as described above, with the following changes: 1) Peptides were only used in the final abundance calculation if they were predicted to be coded for in the genome of a sample based on that sample's HLA type. 2) As an additional filter, peptides were discarded if they had a poor signal to noise ratio (Supplementary figure 4). This was done by calculating a noise level within each plex as the mean MS2 intensity of all samples not predicted to code for the peptide, as their true expression should be zero. For all samples, the log2-ratio was taken of sample intensity to noise intensity. Peptides were discarded if log2-ratios for samples predicted to code for the peptide did not appear to come from a different distribution than samples not predicted to code for the peptide based on the Kolmogorov-Smirnov test. 3) Razor peptides were not used. 4) Peptide to reference ratios were multiplied by a dilution factor (described below) to adjust for ratio inflation in rare peptides. 5) For gene-level abundance calculations, CR ratios for peptides with an allele count of 1 were multiplied by the experimentally determined allele count adjustment factor of 1.8 (Supplementary figure 6B) to produce an abundance value reflecting what their expression would be if they were present in 2 copies.

4.4.9 Sample specific peptide to haplotype assignments

Tryptic peptide predictions for all HLA proteins within the HLAProphet database were performed using the cleave function from the python package Pyteomics. Predicted cleavages were performed for the trypsin protease, with minimum length 7 and maximum length 50, with

up to 1 missed cleavage. For each HLA tryptic peptide, true allele counts within a sample were determined by counting the number of times a tryptic peptide was predicted to be coded for in the HLA type of each individual sample

4.4.10 Peptide ratio dilution factor calculation

Physical differences between peptides often cause large changes in the efficiency with which the peptides are measured within a mass spectrometer, and therefore make it difficult to directly integrate multiple peptides into a final protein abundance using intensity values directly. For this reason, a ratio of peptide MS2 intensity in a sample relative to the MS2 intensity in a pooled common reference (CR) is often used to create a value that can be compared between peptides. In this case, the CR acts as a physical average of peptide expression across all samples in an experiment. For an experiment with N samples, the CR abundance A_{ref} can be modeled as the mean abundance across all samples:

$$A_{ref} = \frac{\sum_{i=1}^{N} A_i}{N}$$

For the HLA proteins, however, many tryptic peptides will be coded for in only a subset of the N sample genomes. The set of all sample abundances A can therefore be thought of as the union of two sets, A^+ containing abundances from N^+ samples coding for the peptide and A^- containing abundances from N^- samples that do not code for the peptide, with all A^- values necessarily being zero:

$$A = A^+ \cup A^- \qquad N = N^+ + N^-$$
$$A^+ = \{A_1^+, \ldots, A_{N^+}^+\}$$
$$A^- = \{A_1^-, \ldots, A_{N^-}^-\} = \{0, \ldots, 0\}$$

However, the pooled CR still physically averages all samples in the experiment, diluting the expression of a given peptide by combining it with other samples that cannot express it and therefore contribute no information about its expression:

$$A_{ref} = \frac{\sum_{i=1}^{N^+} A_i^+ + \sum_{j=1}^{N^-} A_j^-}{N^+ + N^-} = \frac{\sum_{i=1}^{N^+} A_i^+ + 0}{N^+ + N^-}$$

We can adjust for this dilution by multiplying the CR abundance by a dilution factor D that corrects the denominator from N to N^+. In practice, we do not have a direct measurement of CR abundance, but instead apply this dilution factor to the MS2 intensity value for the CR. Due to noise within MS/MS measurements, MS2 intensity values will not truly go to zero as N^+ shrinks, so we add a noise term *c* to prevent over-correcting intensity values for rare peptides:

$$A_{ref_{adj}} = \frac{\sum_{i=1}^{N^+} A_i^+}{N^+ + N^-} * D$$

$$D = \frac{N^+ + N^- + c}{N^+ + c}$$

The optimal value of c is determined by testing all integers from 0 to 10000 to find the value that creates a set of adjusted ratios that have the lowest association with cohort-wide peptide allele count as measured by linear regression.

4.4.11 HLA RNA expression quantification

Personalized genomic HLA reference sequences were produced for all samples using paired DNA sequencing data and the Hapster software as described previously[149]. Genomic reference sequences were then spliced *in silico* to produce personalized transcript reference sequences. HLA specific RNA reads were extracted from each sample's BAM file using the Hapster command extract_reads. Transcript counts were then generated from HLA specific

RNA reads and personalized transcript reference sequences using the Kallisto[25] command quant.

4.4.12 HLA loss of heterozygosity

Loss of heterozygosity for HLA genes was identified using the tool MHCnvex[150] . Briefly, MHCnvex incorporates personalized germline references for the MHC locus to find an accurate coverage, log2-ratio, and B-allele frequency for the region. It then integrates coverage values from the MHC region with coverage of MHC-flanking regions to perform genomic-segmentation. Finally, MHCnvex incorporates integrated segmentation results into its likelihood based model to find the most likely copy number of each HLA gene.

4.4.13 Statistics

For the comparison of the predicted fraction of tryptic peptides identified, pairwise t-tests with Bonferroni corrections were performed. For correlations between peptide ratios vs cohort-wide allele counts, protein expression vs RNA expression, and tumor abundance vs tumor purity, linear regression was performed. To compare the regression line coefficients between A and B alleles for tumor abundance vs tumor purity, the Chow test was performed.

Chapter 5 Personalized Reconstruction of the MHC Locus With MHConstruct

5.1 Introduction

In our previous chapters we focused on analysis of the class I and class II HLA genes using Hapster and HLAProphet. However, given that these tools focus on only the HLA gene regions themselves, neither approach can identify disruption of intergenic regulatory regions that could potentially affect HLA expression. It has been shown that the class I HLA genes are under direct control of the NLRC5 complex which associates with an upstream SXY motif[151]. Similarly, the class II HLA genes are under strict control of the CIITA transcription factor, which again associates with an upstream SXY motif[152]. Further, each gene has also been shown to be controlled by more distant interferon signaling response elements (ISREs)[153], enhancers under the control of NF-kb[154], and potentially other as-yet undiscovered regulators. Given that disruption of these regulatory elements would likely impact HLA expression, it will be critical to investigate these intergenic regions when determining the molecular mechanisms underpinning HLA loss across various cancers.

To investigate HLA regulatory elements we need to focus on a region of chromosome 6 known as the MHC locus which contains all HLA genes, as well as many other genes related to antigen processing[155]. The general explanation for the extreme polymorphism of the HLA genes is that the proteins need to maintain a diverse peptide presentation repertoire and are under evolutionary pressure for balancing selection[156]. It is therefore reasonable to think that while the HLA genes themselves have extreme diversity, the intergenic regions in the MHC locus may not be as variable and would be amenable to traditional standard reference based methods.

Unfortunately, this is not the case, and it has been shown that while variation is still highest in the HLA coding regions, intergenic regions in the MHC locus are still highly polymorphic, with extreme levels of both SNPs and structural variants[157]. To address this issue of intergenic polymorphism we have developed MHConstruct, a genome-graph based algorithm that allows for the reconstruction of a personalized, linear, diploid pair of reference sequences covering the MHC locus that improves alignments and enables downstream analyses.

5.2 Results

5.2.1 The MHConstruct algorithm

MHConstruct begins with the construction of a variation graph. To easily create and manipulate variation graphs, MHConstruct relies on the vg toolkit[158]. For this example we will construct the graph from a simple multiple sequence alignment (MSA) of 5 short sequences containing 9 variants using the command *vg construct* **(Figure 5-1A)**. While MSAs are the most effective way to construct a variation graph, optimal MSAs are difficult to construct from very long sequences with large amounts of variation, and as such larger variation graphs are generally constructed using heuristics to allow for more complicated alignments[159].

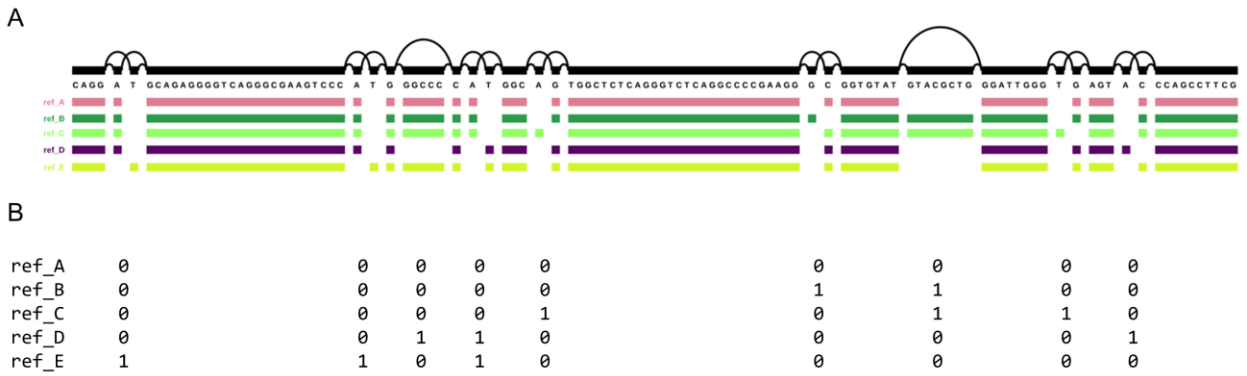

Figure 5-1: Graph representation of genomic variation
(A) Graph representation of 5 genomic sequences containing 9 variant positions. Represented are single nucleotide polymorphisms, insertions, and deletions. **(B)** Genotype calls for each reference path through the graph. Genotypes are coded as 0 if they match the call for path ref_A, and 1 if they represent the alternate allele.

In this graph, we can see that both single nucleotide and indel variants are represented in a single data structure. Similar to how we would annotate variants in a linear reference, we can construct a panel of variants observed in the graph. Here we can arbitrarily choose sequence A to be the reference sequence, and other sequences will have genotype calls relative to that sequence **(Figure 5-1B)**.

With a reference panel, we can now take in sequencing data from a new sample and genotype it relative to the panel. For this example we construct a synthetic diploid genome based on sequences from the original MSA **(Figure 5-2A)**. This synthetic individual has one chromosome containing sequence from references B and C, and one chromosome containing sequences from references D and E. Short 20 bp reads were simulated from this synthetic genotype and were aligned to the variation graph using the command *vg map* **(Figure 5-2B)**. Each variant position in the graph was then genotyped using the command *vg call*, producing a diploid set of genotype calls for every position in the genotype panel **(Figure 5-2C)**.

Once every variant position has been genotyped, we can identify the original haplotypes by phasing the variants. Here, the original panel was used to phase our variants into two

haplotypes with the Eagle[160] phasing tool **(Figure 5-2D)**. With phased variants, the original underlying reference sequences can then be reconstructed by representing the reference genotype panel as a hidden Markov model. In this model, the reference haplotypes represent the hidden states, and the variants in the panel represent a series of emissions, with each emission being either 0 or 1 depending on the genotype. The probability of changing states between each emission represents the probability of a chromosomal crossover event, and the emission probabilities reflect the probability of observing an alternate base at a position due to novel germline variation. Each of the two phased haplotype calls represents a set of observed emissions from the Markov model, and the Viterbi algorithm[161,162] can be used to reconstruct the most likely set of states leading to those emissions **(Figure 5-2E)**. The original haplotype sequences can then be constructed by extracting the path sequences from the reference graph that correspond to the haplotype states reported by the Viterbi algorithm.

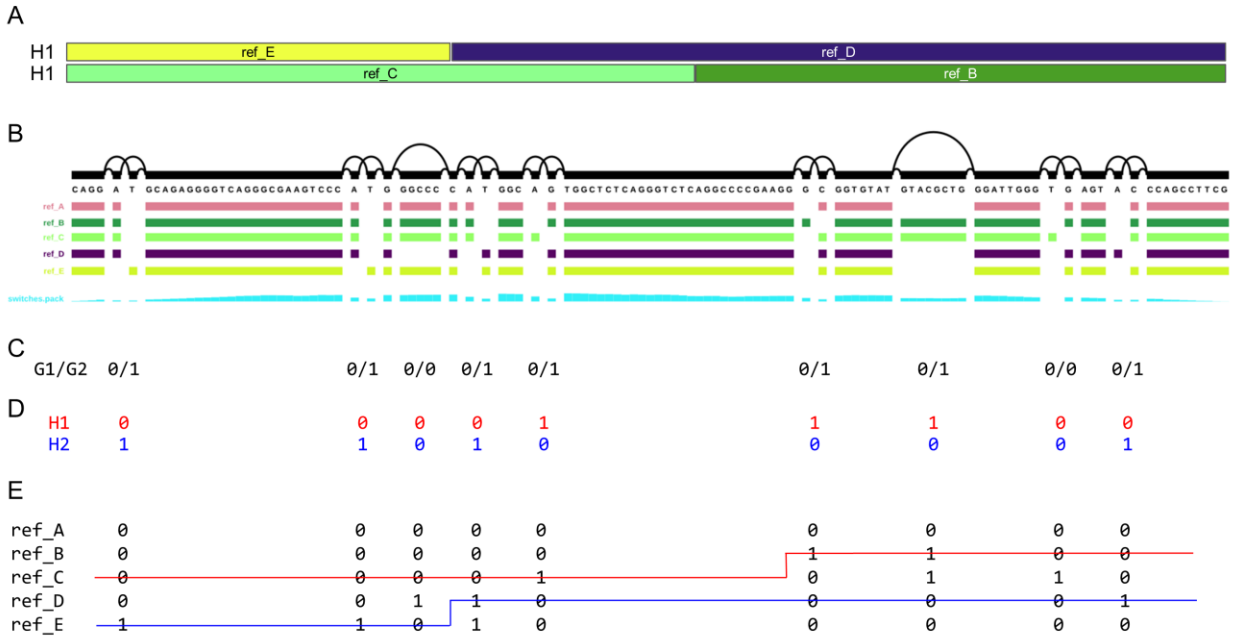

Figure 5-2: Reconstruction of a synthetic diploid haplotype

(A) Synthetic diploid haplotype constructed from partial sequences from reference paths used to construct the variation graph. Haplotype 1 (H1) is constructed from sequences originating from ref_E (yellow) and ref_D (blue). Haplotype 2 (H2) is constructed from sequences originating from ref_C (light green) and ref_B (dark green). **(B)** Coverage map of reads simulated from H1 and H2 after alignment to the variation graph. Coverage for each node is shown below the graph in light blue. Nodes show full coverage in positions where the synthetic haplotype is homozygous, half coverage in positions where the synthetic haplotype is heterozygous, and no coverage for nodes containing variation not found in the synthetic haplotype. **(C)** Genotypes for each position called from the simulated read alignments. **(D)** Phased haplotypes H1 and H2 output by the Eagle phasing algorithm. **(E)** Genotype reference panel for the variation graph, with Viterbi paths shown as colored lines. Red lines show the path for haplotype H1, blue lines show the path for haplotype H2. Note that the two Viterbi paths correspond to the original states in the synthetic haplotypes in A.

5.2.2 MHConstruct produces reference sequences with low germline variation

To reconstruct the full MHC locus for an individual, MHConstruct requires a high quality genome graph to operate on. I use here the first draft reference pangenome provided by the Human Pangenome Reference Consortium (HPRC)[159]. This draft pangenome consists of phased diploid genome assemblies from 47 individuals representing diverse ancestries, and therefore captures a large amount of variation at the MHC locus. The HPRC makes available three separate versions of the draft genome, each constructed using one of either the Minigraph[163], Minigraph-Cactus[164], or PanGenome Graph Builder[165] construction algorithms. We use here the

HPRC reference graph constructed by Minigraph-Cactus as its publicly released representation in GFA format natively encodes haplotype paths through the graph.

As a demonstration of the MHConstruct algorithm, it was next applied to a random 40x WGS sample from the CPTAC melanoma cohort, and results were compared to standard alignment pipelines. Reads from the MHC locus were first aligned to the standard reference GRCh38 region chr6:28510120-33532223, which covers the entire region of chromosome 6p21 from *GPX5* to *ZBTB9*[157]. To evaluate the quality of the GRCh38 MHC locus sequence as a reference for this random individual, germline variants were called using this alignment. If alignment was performed using a perfect reference sequence, we would expect no apparent germline variants to be called, as the reference sequence would exactly match the sequencing reads being aligned. A poor reference sequence would have a large number of differences relative to the sequencing reads, resulting in many germline variant calls. For this individual, when partitioning the MHC locus into 10kb buckets, 74% of buckets have a rate of variation higher than the genome wide average of 1 variant per kb[84], 25% of buckets have a rate higher than 5 variants per kb, and many regions have a rate higher than 20 variants per kb relative to the GRCh38 reference sequence **(Figure 5-3)**. Of note, in regions of extremely high variance this may even be an underestimate, as failures to align will result in missed germline variant calls. As expected, we see that the regions with many germline variants surround the HLA genes **(Figure 5-3**, vertical grey lines**)**, suggesting that both the HLA coding regions and their neighboring intergenic regions have a high amount of germline polymorphism.

Next, MHConstruct was used to reconstruct a personalized diploid representation of the MHC locus for this same individual, using the same set of MHC locus reads. The reads were then aligned to the MHConstruct reference sequences, and germline variants were again called.

In this alignment, only 3.5% of 10kb buckets have a rate of germline variation higher than 1 variant per kb, and less than 0.1% of buckets are above 5 variants per kb. 99.7% of reads successfully aligned, showing that the drop in germline variant calls was not due to reduced coverage or poor alignments, but rather due to a true increase in the concordance between the reference sequences and the sequencing reads.

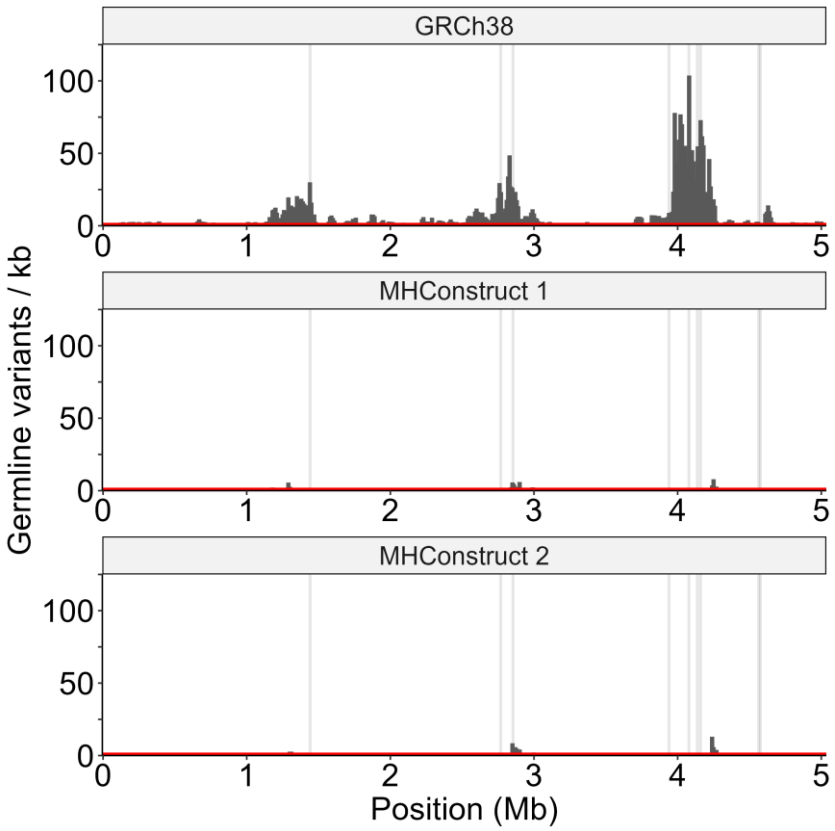

Figure 5-3: Comparison of linear vs MHConstruct references
Histogram of germline SNPs across the MHC locus observed in a random normal tissue sample from the CPTAC melanoma cohort. Germline SNPs were called relative to the representative sequence for the MHC locus found in the standard reference GRCh38 region chr6:28510120-33532223 (top), or relative to the two reference sequences created by MHConstruct (middle, bottom). Horizontal red line represents the genome-wide average rate of variation of 1 variant per kilobase. Vertical light grey bars represent coding regions corresponding to the class I and class II HLA genes that were investigated in previous chapters.

5.3 Discussion

We show here that MHConstruct can be used to create large, personalized, diploid reconstructions of the MHC locus. In this case study, the reconstructed reference sequences

allow for improved alignments which enable downstream analyses of intergenic regulatory regions that may control HLA or other antigen processing related genes. In the future we will provide additional validation studies to demonstrate the efficacy of MHConstruct in reducing germline SNPs across diverse HLA types. Additionally, these validation studies will investigate the ability of MHConstruct to properly identify large structural variants, such as those at the C3/C4 locus and the DRB3/4/5 locus.

We also aim to perform thorough benchmarking of MHConstruct's ability to improve the identification of somatic variation in tumors relative to normal samples. While improved references for the most polymorphic sequences, such as the coding regions of the HLA genes, nearly doubles the number of somatic variants called in tumors (Chapter 2), it remains to be seen how much improvement MHConstruct will provide in the surrounding intergenic regions. The application of MHConstruct to tumors will be of particular clinical interest when studying cancers such as prostate adenocarcinomas which are frequently completely HLA negative[166], but show no clear genomic alterations to the coding regions of the HLA genes (Chapter 3). Further studies using personalized MHConstruct reference sequences may be able to identify the specific regulatory mechanisms leading to suppression of HLA expression in these cases, providing a better understanding of primary and acquired resistance to T-cell based immunotherapies.

We also note that MHConstruct is not restricted to the MHC locus. With the release of the HPRC reference pangenome, any region of interest could be reconstructed for personalized analysis. In addition to the MHC locus, other polymorphic regions such as those surrounding the KIR[167] genes could also be constructed for use with downstream analyses. Further, even in relatively homogenous regions, it has been shown that graph based representations can improve

linear alignments[159]. This may ultimately allow for reduced bias towards standard reference sequence homology in any sequenced based analysis.

5.4 Methods

5.4.1 Code and availability

Custom code is available at https://github.com/MBMumphrey/MHConstruct3

5.4.2 MHConstruct example workflow

Full code for the MSA example provided in figures 5-1 and 5-2 is available at https://github.com/MBMumphrey/MHConstruct3/tree/main/example

5.4.3 MHConstruct example graph

A multiple sequence alignment (MSA) of 5 short sequences with 9 variants was manually constructed as a text fasta file. A variation graph representation of the MSA was constructed using the command *vg construct*. The variant panel was constructed from the variation graph using the command *vg deconstruct*. Alignment indexes for the variation graph were constructed using the command *vg autoindex.*

5.4.4 MHConstruct example synthetic haplotype

A synthetic haplotype was manually constructed by stitching together sequences from the original MSA. For the first synthetic chromosome, sequence from ref_E was taken covering the first 3 variant positions, followed by sequence from ref_D covering the last 6 variant positions. For the second synthetic chromosome, sequence from ref_C was taken covering the first 5 variant positions, followed by sequence from ref_B covering the last 4 variant positions. A variation graph was constructed from these two synthetic contigs using the command *vg*

construct, and reads of length 20 were simulated from each haplotype using the command *vg sim*.

5.4.5 MHConstruct example genotyping

Simulated reads were aligned to the original variation graph using the command *vg map*. Node read coverages were calculated using the command *vg pack*. Genotypes were called using the command *vg call*. Genotypes were phased using Eagle[160] v2.4.1.

5.4.6 MHConstruct Viterbi algorithm

Optimal state paths through the genotype reference panel were identified using the Viterbi algorithm[161,162]. Briefly, the genotype reference panel was modeled as a hidden Markov model with haplotypes as hidden states, genotypes as state emissions, and variant sets as a Markov chain. Transition probabilities were calculated as the probability of a crossover event occurring between two variants, with an estimated rate of 1% chance of crossover per 1Mb. The probability of emitting a genotype different from the observed state emission was estimated to be 1/3000. This estimate was derived from the genome wide SNP rate of 1 variant per kilobase, or 1 in 1000. Then, given that each SNP in our panel has an exact alternate base observed, the probability of hitting that exact alternate base is 1 in 3, which results in a combined probability of 1 in 3000.

5.4.7 MHConstruct reference sequence construction

Reference sequences were reconstructed using the optimal state paths through the genotype reference panel as determined by the Viterbi algorithm. Starting with the first state in the path, sequence is taken from all nodes in the variation graph along the path for the corresponding state. Whenever a state switch occurs, we continue through the variation graph

taking sequences from nodes corresponding to the new state. Sequences from each node are appended to each other to create a single linear reference string.

5.4.8 Reference pangenome graph

MHC locus reconstruction was performed based on the first draft human pangenome reference graph[159]. Reference graph and index files were downloaded from the HPRC repository at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=pangenomes/freeze/freeze1/minigraph-cactus/.

5.4.9 Germline variant calling

Reference sequence quality was determined by identifying the number of germline variants called relative to either GRCh38 or MHConstruct references for the MHC locus. First a random 40x WGS normal tissue sample was selected from the CPTAC melanoma cohort. This sample was originally aligned to a version of GRCh38 containing alt contigs for the entire MHC locus, as well as alt contigs for many different HLA alleles. MHC locus specific reads were extracted by taking all reads from GRCh38 region GRCh38 region chr6:28510120-33532223, chr6 alt contigs (GL000250v2, GL000251v2, GL000252v2, GL000253v2, GL000255v2, GL000256v2), and all alt HLA contigs. An MHC locus specific GRCh38 reference was created by taking only GRCh38 region chr6:28510120-33532223 and constructing a new set of reference indexes. MHConstruct references were constructed as outlined in the methods. The extracted read set was then aligned to both sets of references using BWA-mem[74]. Germline variants were called using GATK's HaplotypeCaller[87].

Chapter 6 Concluding Remarks

Despite its widespread prevalence and potential impact on immunotherapy outcomes, loss of HLA expression remains an understudied phenomenon due to the extreme polymorphism of the HLA genes. In this thesis I presented three new computational techniques that solve this problem by departing from the standard-reference paradigm of molecular sequencing analysis in favor of dynamic reference selection, allowing for a personalized analysis of each individual's HLA genes.

In chapter 2 I first presented Hapster, a genomics tool that takes as input DNA sequencing data and produces as output personalized genomic reference sequences, alignments, and mutation calls. In chapter 3 I showed the utility of Hapster in uncovering new biology by applying it to 12,000 tumors sequenced by the TCGA and MI-ONCOSEQ projects. Broadly, I show that pan-cancer positive selection for functional mutations in the HLA genes is stronger than previously known, making the HLA genes among the most commonly mutated tumor suppressor genes. With a more fine-grained analysis, I showed that in squamous cell carcinomas and lymphomas there is significant evidence for positive selection of stop gain mutations in the class I HLA genes, potentially as a result of increased off target APOBEC and AID activity. I also showed that in colorectal and stomach adenocarcinomas with microsatellite instability there is significant evidence for positive selection of truncating frameshifts that occur specifically within coding region microsatellites in the class I HLA genes. Finally, I showed that missense

mutations are enriched at the HLA:B2M interface and within the peptide binding pocket, supporting the idea that these are also loss of function variants.

Hapster's use also extends far beyond identification of somatic mutations in the HLA genes. As shown by the tool MHCnvex[150], the personalized alignments produced by Hapster can be used to more accurately identify copy number variation in the MHC locus. The ability to identify both small somatic variants and complete copy loss of individual HLA alleles will be required when attempting to provide a complete picture of somatic loss of the HLAs. Hapster is also not restricted to DNA sequencing analyses, but can additionally provide personalized alignments of paired RNA-sequencing data. Due to the diploid nature of Hapster's constructed reference sequences, this not only improves expression quantifications by personalizing the analysis, but also enables allele-specific expression quantification of each individual HLA allele. The ability to analyze HLA RNA expression data will be key in determining the molecular mechanisms leading to transcriptional loss in cancers such as prostate adenocarcinoma where tumors are often HLA negative but show no apparent mutation of the HLA genes themselves. Finally, the generalized Hapster algorithm is not restricted to the HLA genes, but could in principle be expanded to any set of polymorphic genes, enabling analysis of families such as the KIRs[167].

In chapter 4 I presented HLAProphet, a tool that allows for the personalized quantification of HLA proteins from multiplexed TMT MS/MS data. I showed that HLAProphet's personalized protein databases improve HLA tryptic peptide identifications by nearly 3-fold relative to searches performed using the standard GENCODE34 protein database. I also demonstrated how HLAProphet utilizes HLA haplotype information to adjust for variable peptide allele counts both within samples and within the pooled reference channel, preventing

both over- and under-estimation of protein abundance. Finally, I showed that HLAProphet is capable of reporting both total protein expression for each gene, as well as allele-specific protein expression. In all cases, HLAProphet's expression values showed higher concordance to paired RNA-sequencing data than standard approaches, suggesting that overall quantification was improved. Additionally, allele-specific expression loss was confirmed in cases with known genomic LOH events, supporting HLAProphet's ability to report allele-specific protein expression values.

Similar to Hapster, HLAProphet's abilities extend beyond what is reported here. Of particular interest is its potential use in identifying alterations in post-translational modifications (PTMs) of the HLA proteins. The HLA proteins are glycoproteins, and the passage of an HLA protein through the folding, loading, and export process occurs alongside a complex series of modifications to a conserved PTM[144]. HLAProphet's personalized databases should naturally extend to the quantification of HLA PTMs, and allow for identification of potential PTM disruptions that could help explain cases with reduced cell surface HLA expression in the absence of apparent alterations at the DNA or RNA level.

In chapter 5 I presented MHConstruct, a tool that allows for a personalized diploid reconstruction of the entire MHC locus. While Hapster and HLAProphet enable personalized analyses of the HLA genes themselves, MHConstruct enables personalized analysis of the surrounding intergenic regions. This will enable more accurate investigations into the disruption of HLA-specific promoters and enhancers, as well as improved identification of QTLs. The version of MHConstruct demonstrated here utilizes the first draft of the human pangenome reference release by the HPRC and early research versions of the VG toolkit. However, as graph

based methods improve over the coming years, the human pangenome reference will begin to capture more known human variation, enabling more accurate reconstructions of the MHC locus.

In the future, we plan to apply the methods presented here to provide a complete picture of HLA loss pan-cancer. For any given tumor that evolves to disrupt the HLAs, it should be possible via Hapster, HLAProphet, and MHConstruct to identify if this loss is the result of somatic loss of function, transcriptional repression, or post-translational loss of protein expression. With the public availability of large cohorts such as CPTAC that provide paired DNA, RNA, and proteomics data for their samples, we aim to identify the true rates of HLA loss across cancer types, as well as characterize the relative frequencies of complete vs allelic loss.

We also aim to apply the tools presented here to better characterize tumor responses to immunotherapy using sequencing data from large ICI cohorts. Within these samples we will identify those cases that were HLA negative before treatment to determine if this has significant negative impacts on patient outcomes, as has been previously proposed. Further, we aim to identify those tumors that were HLA positive before ICI, but lose HLA expression after treatment. These cases will help to determine if HLA loss is a common mechanism of acquired resistance to immunotherapy. Finally, the specific mechanisms identified for any observed HLA loss will provide insights into how often this acquired resistance is reversible, and can potentially be treated with combination therapies that recover lost expression.

Finally, we aim to more robustly validate the methods presented here so that they may make their way to the clinic. At the individual patient level, identification of HLA loss is vital when making decisions both for potential treatment and enrollment in clinical trials. In cases with irreversible loss of HLA function, it may save patients valuable time by recognizing that they may not be good candidates for T-cell based immunotherapies, indicating other potential

treatments instead. For companies enrolling patients in new immunotherapy clinical trials, it will also be vital to identify patients who are HLA negative and may therefore be poor responders. This improvement in patient selection could both save research money, and make sure that valid immunotherapies don't fail at the trial stage.

References

1. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research* **43**, D423–D431 (2015).

2. Braud, V. M. *et al.* HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature* **391**, 795–799 (1998).

3. Hunt, J. S., Petroff, M. G., McIntire, R. H. & Ober, C. HLA-G and immune tolerance in pregnancy. *The FASEB Journal* **19**, 681–693 (2005).

4. Denzin, L. K. & Cresswell, P. HLA-DM induces clip dissociation from MHC class II αβ dimers and facilitates peptide loading. *Cell* **82**, 155–165 (1995).

5. Liljedahl, M. *et al.* HLA-DO is a lysosomal resident which requires association with HLA-DM for efficient intracellular transport. *The EMBO Journal* **15**, 4817–4824 (1996).

6. Glickman, M. H. & Ciechanover, A. The Ubiquitin-Proteasome Proteolytic Pathway: Destruction for the Sake of Construction. *Physiological Reviews* **82**, 373–428 (2002).

7. Watts, C. The endosome–lysosome pathway and information generation in the immune system. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1824**, 14–21 (2012).

8. Tanoka, K. & Kasahara, M. The MHC class I ligand-generating system: roles of immunoproteasomes and the interferon-4gMY-inducible proteasome activator PA28. *Immunological Reviews* **163**, 161–176 (1998).

9. Nakagawa, T. Y. & Rudensky, A. Y. The role of lysosomal proteinases in MHC class Il-mediated antigen processing and presentation. *Immunological Reviews* **172**, 121–129 (1999).

10. Joffre, O. P., Segura, E., Savina, A. & Amigorena, S. Cross-presentation by dendritic cells. *Nat Rev Immunol* **12**, 557–569 (2012).

11. Momburg, F. & Hämmerling, G. J. Generation and TAP-Mediated Transport of Peptides for Major Histocompatibility Complex Class I Molecules. in *Advances in Immunology* (ed. Dixon, F. J.) vol. 68 191–256 (Academic Press, 1998).

12. Momburg, F., Roelse, J., Hämmerling, G. J. & Neefjes, J. J. Peptide size selection by the major histocompatibility complex-encoded peptide transporter. *Journal of Experimental Medicine* **179**, 1613–1623 (1994).

13. Rammensee, H. G., Falk, K. & Rötzschke, O. Peptides Naturally Presented by MHC Class I Molecules. *Annual Review of Immunology* **11**, 213–244 (1993).

14. Chang, S.-C., Momburg, F., Bhutani, N. & Goldberg, A. L. The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a “molecular ruler” mechanism. *Proceedings of the National Academy of Sciences* **102**, 17107–17112 (2005).

15. Sadasivan, B., Lehner, P. J., Ortmann, B., Spies, T. & Cresswell, P. Roles for Calreticulin and a Novel Glycoprotein, Tapasin, in the Interaction of MHC Class I Molecules with TAP. *Immunity* **5**, 103–114 (1996).

16. Wearsch, P. A. & Cresswell, P. Selective loading of high-affinity peptides onto major histocompatibility complex class I molecules by the tapasin-ERp57 heterodimer. *Nat Immunol* **8**, 873–881 (2007).

17. Ghosh, P., Amaya, M., Mellins, E. & Wiley, D. C. The structure of an intermediate in class II MHC maturation: CLIP bound to HLA-DR3. *Nature* **378**, 457–462 (1995).

18. Bassing, C. H., Swat, W. & Alt, F. W. The Mechanism and Regulation of Chromosomal V(D)J Recombination. *Cell* **109**, S45–S55 (2002).

19. Jameson, S. C. & Bevan, M. J. T-cell selection. *Current Opinion in Immunology* **10**, 214–219 (1998).

20. Little, C. C. & Tyzzer, E. E. Further experimental studies on the inheritance of susceptibility to a Transplantable tumor, Carcinoma (J. W. A.) of the Japanese waltzing Mouse. *J Med Res* **33**, 393–453 (1916).

21. Gorer, P. A. The genetic and antigenic basis of tumour transplantation. *The Journal of Pathology and Bacteriology* **44**, 691–697 (1937).

22. Gorer, P. A., Lyman, S. & Snell, G. D. Studies on the Genetic and Antigenic Basis of Tumour Transplantation. Linkage between a Histocompatibility Gene and ‘Fused’ in Mice. *Proceedings of the Royal Society of London. Series B, Biological Sciences* **135**, 499–505 (1948).

23. Dausset, J. Iso-Leuko-Antibodies*. *Vox Sanguinis* **3**, 40–41 (1958).

24. Burnet, M. Cancer—A Biological Approach. *Br Med J* **1**, 841–847 (1957).

25. van der Bruggen, P. *et al.* A Gene Encoding an Antigen Recognized by Cytolytic T Lymphocytes on a Human Melanoma. *Science* **254**, 1643–1647 (1991).

26. Dighe, A. S., Richards, E., Old, L. J. & Schreiber, R. D. Enhanced in vivo growth and resistance to rejection of tumor cells expressing dominant negative IFNγ receptors. *Immunity* **1**, 447–456 (1994).

27. Shankaran, V. *et al.* IFNγ and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature* **410**, 1107–1111 (2001).

28. Russell, J. H. & Ley, T. J. Lymphocyte-mediated cytotoxicity. *Annu Rev Immunol* **20**, 323–370 (2002).

29. Grulich, A. E., van Leeuwen, M. T., Falster, M. O. & Vajdic, C. M. Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. *The Lancet* **370**, 59–67 (2007).

30. Krummel, M. F. & Allison, J. P. CD28 and CTLA-4 have opposing effects on the response of T cells to stimulation. *Journal of Experimental Medicine* **182**, 459–465 (1995).

31. Ishida, Y., Agata, Y., Shibahara, K. & Honjo, T. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *The EMBO Journal* **11**, 3887–3895 (1992).

32. Leach, D. R., Krummel, M. F. & Allison, J. P. Enhancement of Antitumor Immunity by CTLA-4 Blockade. *Science* **271**, 1734–1736 (1996).

33. Hodi, F. S. *et al.* Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. *New England Journal of Medicine* **363**, 711–723 (2010).

34. Topalian, S. L. *et al.* Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer. *New England Journal of Medicine* **366**, 2443–2454 (2012).

35. Hamid, O. *et al.* Safety and Tumor Responses with Lambrolizumab (Anti–PD-1) in Melanoma. *New England Journal of Medicine* **369**, 134–144 (2013).

36. Burova, E. *et al.* Characterization of the Anti–PD-1 Antibody REGN2810 and Its Antitumor Activity in Human PD-1 Knock-In Mice. *Molecular Cancer Therapeutics* **16**, 861–870 (2017).

37. Kaufman, H. L. *et al.* Avelumab in patients with chemotherapy-refractory metastatic Merkel cell carcinoma: a multicentre, single-group, open-label, phase 2 trial. *The Lancet Oncology* **17**, 1374–1385 (2016).

38. Necchi, A. *et al.* Atezolizumab in platinum-treated locally advanced or metastatic urothelial carcinoma: post-progression outcomes from the phase II IMvigor210 study. *Annals of Oncology* **28**, 3044–3050 (2017).

39. Massard, C. *et al.* Safety and Efficacy of Durvalumab (MEDI4736), an Anti–Programmed Cell Death Ligand-1 Immune Checkpoint Inhibitor, in Patients With Advanced Urothelial Bladder Cancer. *J Clin Oncol* **34**, 3119–3125 (2016).

40. Tawbi, H. A. *et al.* Relatlimab and Nivolumab versus Nivolumab in Untreated Advanced Melanoma. *New England Journal of Medicine* **386**, 24–34 (2022).

41. Marin-Acevedo, J. A., Kimbrough, E. O. & Lou, Y. Next generation of immune checkpoint inhibitors and beyond. *Journal of Hematology & Oncology* **14**, 45 (2021).

42. Chen, D. S. & Mellman, I. Oncology Meets Immunology: The Cancer-Immunity Cycle. *Immunity* **39**, 1–10 (2013).

43. Dhatchinamoorthy, K., Colbert, J. D. & Rock, K. L. Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation. *Frontiers in Immunology* **12**, (2021).

44. Ye, Q. *et al.* Hypermethylation of HLA class I gene is associated with HLA class I down-regulation in human gastric cancer. *Tissue Antigens* **75**, 30–39 (2010).

45. Kaklamanis, L. *et al.* Loss of major histocompatibility complex-encoded transporter associated with antigen presentation (TAP) in colorectal cancer. *Am J Pathol* **145**, 505–509 (1994).

46. Romero, J. M. *et al.* Coordinated downregulation of the antigen presentation machinery and HLA class I/β2-microglobulin complex is responsible for HLA-ABC loss in bladder cancer. *International Journal of Cancer* **113**, 605–610 (2005).

47. McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171**, 1259-1271.e11 (2017).

48. Cabrera, T. *et al.* High frequency of HLA-B44 allelic losses in human solid tumors. *Human Immunology* **64**, 941–950 (2003).

49. Zinkernagel, Rol. M. & Doherty, Pete. C. MHC-Restricted Cytotoxic T Cells: Studies on the Biological Role of Polymorphic Major Transplantation Antigens Determining T-Cell Restriction-Specificity, Function, and Responsiveness. in *Advances in Immunology* (eds. Kunkel, H. G. & Dixon, F. J.) vol. 27 51–177 (Academic Press, 1979).

50. Ljunggren, H. G. & Kärre, K. Host resistance directed selectively against H-2-deficient lymphoma variants. Analysis of the mechanism. *Journal of Experimental Medicine* **162**, 1745–1759 (1985).

51. Sunshine, G. H., Katz, D. R. & Feldmann, M. Dendritic cells induce T cell proliferation to synthetic antigens under Ir gene control. *J Exp Med* **152**, 1817–1822 (1980).

52. Erb, P. *et al.* Characterization of accessory cells required for helper T cell induction in vitro: evidence for a phagocytic, Fc-receptor, and Ia-bearing cell type. *The Journal of Immunology* **125**, 2504–2507 (1980).

53. Janeway, C. A., Jr, Ron, J. & Katz, M. E. The B cell is the initiating antigen-presenting cell in peripheral lymph nodes. *The Journal of Immunology* **138**, 1051–1055 (1987).

54. Steimle, V., Siegrist, C.-A., Mottet, A., Lisowska-Grospierre, B. & Mach, B. Regulation of MHC Class II Expression by Interferon-γ Mediated by the Transactivator Gene CIITA. *Science* **265**, 106–109 (1994).

55. Park, I. A. *et al.* Expression of the MHC class II in triple-negative breast cancer is associated with tumor-infiltrating lymphocytes and interferon signaling. *PLOS ONE* **12**, e0182786 (2017).

56. He, Y. *et al.* MHC class II expression in lung cancer. *Lung Cancer* **112**, 75–80 (2017).

57. Dunne, M. R. *et al.* Characterising the prognostic potential of HLA-DR during colorectal cancer development. *Cancer Immunol Immunother* **69**, 1577–1588 (2020).

58. Accolla, R. S., Ramia, E., Tedeschi, A. & Forlani, G. CIITA-Driven MHC Class II Expressing Tumor Cells as Antigen Presenting Cell Performers: Toward the Construction of an Optimal Anti-tumor Vaccine. *Frontiers in Immunology* **10**, (2019).

59. Mizukami, Y. *et al.* Downregulation of HLA Class I molecules in the tumour is associated with a poor prognosis in patients with oesophageal squamous cell carcinoma. *Br J Cancer* **99**, 1462–1467 (2008).

60. Bandoh, N. *et al.* HLA class I antigen and transporter associated with antigen processing downregulation in metastatic lesions of head and neck squamous cell carcinoma as a marker of poor prognosis. *Oncology Reports* **23**, 933–939 (2010).

61. Kageshita, T., Hirai, S., Ono, T., Hicklin, D. J. & Ferrone, S. Down-Regulation of HLA Class I Antigen-Processing Molecules in Malignant Melanoma: Association with Disease Progression. *The American Journal of Pathology* **154**, 745–754 (1999).

62. Ling, A. *et al.* TAP1 down-regulation elicits immune escape and poor prognosis in colorectal cancer. *OncoImmunology* **6**, e1356143 (2017).

63. Pedersen, M. H. *et al.* Downregulation of antigen presentation-associated pathway proteins is linked to poor outcome in triple-negative breast cancer patient tumors. *OncoImmunology* **6**, e1305531 (2017).

64. Perea, F. *et al.* The absence of HLA class I expression in non-small cell lung cancer correlates with the tumor tissue structure and the pattern of T cell infiltration. *International Journal of Cancer* **140**, 888–899 (2017).

65. Mihm, M. C., Clemente, C. G. & Cascinelli, N. Tumor infiltrating lymphocytes in lymph node melanoma metastases: a histopathologic prognostic indicator and an expression of local immune response. *Lab Invest* **74**, 43–47 (1996).

66. Ménard, S. *et al.* Lymphoid infiltration as a prognostic variable for early-onset breast carcinomas. *Clinical Cancer Research* **3**, 817–819 (1997).

67. Ropponen, K. M., Eskelinen, M. J., Lipponen, P. K., Alhava, E. & Kosma, V.-M. Prognostic value of tumour-infiltrating lymphocytes (TILs) in colorectal cancer. *The Journal of Pathology* **182**, 318–324 (1997).

68. Carretero, R. *et al.* Analysis of HLA class I expression in progressing and regressing metastatic melanoma lesions after immunotherapy. *Immunogenetics* **60**, 439–447 (2008).

69. Doherty, P. C. & Zinkernagel, R. M. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature* **256**, 50–52 (1975).

70. Hedrick, P. W. Pathogen Resistance and Genetic Variation at Mhc Loci. *Evolution* **56**, 1902–1908 (2002).

71. Zheng-Bradley, X. *et al.* Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *Gigascience* **6**, 1–8 (2017).

72. Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol* **33**, 1152–1158 (2015).

73. Mumphrey, M. B. *et al.* Distinct mutational processes shape selection of MHC class I and class II mutations across primary and metastatic tumors. 2023.01.22.523447 Preprint at https://doi.org/10.1101/2023.01.22.523447 (2023).

74. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).

75. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. 861054 Preprint at https://doi.org/10.1101/861054 (2019).

76. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

77. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).

78. Dilthey, A. T. *et al.* High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. *PLOS Computational Biology* **12**, e1005151 (2016).

79. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Human Mutation* **38**, 788–797 (2017).

80. Ka, S. *et al.* HLAscan: genotyping of the HLA region using next-generation sequencing data. *BMC Bioinformatics* **18**, 258 (2017).

81. Lee, H. & Kingsford, C. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome Biology* **19**, 16 (2018).

82. Sverchkova, A., Anzar, I., Stratford, R. & Clancy, T. Improved HLA typing of Class I and Class II alleles from next-generation sequencing data. *HLA* **94**, 504–513 (2019).

83. Gourraud, P.-A. *et al.* HLA Diversity in the 1000 Genomes Dataset. *PLOS ONE* **9**, e97282 (2014).

84. Shen, H. *et al.* Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. *PLOS ONE* **8**, e59494 (2013).

85. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).

86. Castro, A. *et al.* Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes. *BMC Med Genomics* **12**, 107 (2019).

87. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. 201178 Preprint at https://doi.org/10.1101/201178 (2018).

88. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*. https://www.osti.gov/biblio/1241166 (2014).

89. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Research* **40**, e115 (2012).

90. Mortara, L. *et al.* CIITA-Induced MHC Class II Expression in Mammary Adenocarcinoma Leads to a Th1 Polarization of the Tumor Microenvironment, Tumor Rejection, and Specific Antitumor Memory. *Clinical Cancer Research* **12**, 3435–3443 (2006).

91. Johnson, D. B. *et al.* Melanoma-specific MHC-II expression represents a tumour-autonomous phenotype and predicts response to anti-PD-1/PD-L1 therapy. *Nat Commun* **7**, 10582 (2016).

92. Forero, A. *et al.* Expression of the MHC Class II Pathway in Triple-Negative Breast Cancer Tumor Cells Is Associated with a Good Prognosis and Infiltrating Lymphocytes. *Cancer Immunology Research* **4**, 390–399 (2016).

93. Haabeth, O. A. W. *et al.* How Do CD4+ T Cells Detect and Eliminate Tumor Cells That Either Lack or Express MHC Class II Molecules? *Frontiers in Immunology* **5**, (2014).

94. Howitt, B. E. *et al.* Association of Polymerase e–Mutated and Microsatellite-Instable Endometrial Cancers With Neoantigen Load, Number of Tumor-Infiltrating Lymphocytes, and Expression of PD-1 and PD-L1. *JAMA Oncology* **1**, 1319–1323 (2015).

95. Kloor, M. *et al.* Immunoselective Pressure and Human Leukocyte Antigen Class I Antigen Machinery Defects in Microsatellite Unstable Colorectal Cancers. *Cancer Research* **65**, 6418–6424 (2005).

96. Ozcan, M., Janikovits, J., von Knebel Doeberitz, M. & Kloor, M. Complex pattern of immune evasion in MSI colorectal cancer. *OncoImmunology* **7**, e1445453 (2018).

97. Giraldo, N. A. *et al.* The immune contexture of primary and metastatic human tumours. *Current Opinion in Immunology* **27**, 8–15 (2014).

98. Szekely, B. *et al.* Immunological differences between primary and metastatic breast cancer. *Annals of Oncology* **29**, 2232–2239 (2018).

99. Remark, R. *et al.* Characteristics and Clinical Impacts of the Immune Environments in Colorectal and Renal Cell Carcinoma Lung Metastases: Influence of Tumor Origin. *Clinical Cancer Research* **19**, 4079–4091 (2013).

100. Halama, N. *et al.* Localization and Density of Immune Cells in the Invasive Margin of Human Colorectal Cancer Liver Metastases Are Prognostic for Response to Chemotherapy. *Cancer Research* **71**, 5670–5677 (2011).

101. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat Genet* **49**, 1785–1788 (2017).

102. Chowell, D. *et al.* Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Nat Med* **25**, 1715–1720 (2019).

103. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**, 106–114 (2015).

104. Iranzo, J., Martincorena, I. & Koonin, E. V. Cancer-mutation network and the number and specificity of driver mutations. *Proceedings of the National Academy of Sciences* **115**, E6010–E6019 (2018).

105. Leone, P. *et al.* MHC Class I Antigen Processing and Presenting Machinery: Organization, Function, and Defects in Tumor Cells. *JNCI: Journal of the National Cancer Institute* **105**, 1172–1187 (2013).

106. Challa-Malladi, M. *et al.* Combined Genetic Inactivation of β2-Microglobulin and CD58 Reveals Frequent Escape from Immune Recognition in Diffuse Large B Cell Lymphoma. *Cancer Cell* **20**, 728–740 (2011).

107. Muzio, M. *et al.* FLICE, A Novel FADD-Homologous ICE/CED-3–like Protease, Is Recruited to the CD95 (Fas/APO-1) Death-Inducing Signaling Complex. *Cell* **85**, 817–827 (1996).

108. Lyu, H., Li, M., Jiang, Z., Liu, Z. & Wang, X. Correlate the TP53 Mutation and the HRAS Mutation with Immune Signatures in Head and Neck Squamous Cell Cancer. *Computational and Structural Biotechnology Journal* **17**, 1020–1030 (2019).

109. Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D. & Perucho, M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* **363**, 558–561 (1993).

110. Lang, G. I., Parsons, L. & Gammie, A. E. Mutation Rates, Spectra, and Genome-Wide Distribution of Spontaneous Mutations in Mismatch Repair Deficient Yeast. *G3 Genes|Genomes|Genetics* **3**, 1453–1465 (2013).

111. Okazaki, I., Kotani, A. & Honjo, T. Role of AID in Tumorigenesis. in *Advances in Immunology* vol. 94 245–273 (Academic Press, 2007).

112. Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45**, 970–976 (2013).

113. Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet* **47**, 1067–1072 (2015).

114. Parolia, A. *et al.* Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. *Nature* **571**, 413–418 (2019).

115. Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences* **112**, E5486–E5495 (2015).

116. Gao, G. F. *et al.* Crystal structure of the complex between human CD8αα and HLA-A2. *Nature* **387**, 630–634 (1997).

117. Bicknell, D. C., Kaklamanis, L., Hampson, R., Bodmer, W. F. & Karran, P. Selection for β2-microglobulin mutation in mismatch repair-defective colorectal carcinomas. *Current Biology* **6**, 1695–1697 (1996).

118. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).

119. Huang, X., Zheng, W., Pearce, R. & Zhang, Y. SSIPe: accurately estimating protein–protein binding affinity change upon mutations using evolutionary profiles in combination with an optimized physical energy function. *Bioinformatics* **36**, 2429–2437 (2020).

120. Engin, H. B., Kreisberg, J. F. & Carter, H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLOS ONE* **11**, e0152929 (2016).

121. Gentles, A. J. *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med* **21**, 938–945 (2015).

122. Garrido, F., Algarra, I. & García-Lora, A. M. The escape of cancer from T lymphocytes: immunoselection of MHC class I loss variants harboring structural-irreversible "hard" lesions. *Cancer Immunol Immunother* **59**, 1601–1606 (2010).

123. Landi, L. *et al.* Bone metastases and immunotherapy in patients with advanced non-small-cell lung cancer. *j. immunotherapy cancer* **7**, 316 (2019).

124. Yu, J. *et al.* Liver metastasis restrains immunotherapy efficacy via macrophage-mediated T cell elimination. *Nat Med* **27**, 152–164 (2021).

125. López-Nevot, M. A. *et al.* HLA class I gene expression on human primary tumours and autologous metastases: demonstration of selective losses of HLA antigens on colorectal, gastric and laryngeal carcinomas. *Br J Cancer* **59**, 221–226 (1989).

126. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).

127. Leiserson, M. D., Wu, H.-T., Vandin, F. & Raphael, B. J. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biology* **16**, 160 (2015).

128. Zhao, Z. *et al.* UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test. *The American Journal of Human Genetics* **106**, 3–12 (2020).

129. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696–705 (2018).

130. Mumphrey, M. B., Li, G. X., Hosseini, N., Nesvizhskii, A. & Cieslik, M. HLAProphet: Personalized allele-level quantification of the HLA proteins. 2023.01.29.526142 Preprint at https://doi.org/10.1101/2023.01.29.526142 (2023).

131. Sheynkman, G. M., Shortreed, M. R., Frey, B. L., Scalf, M. & Smith, L. M. Large-Scale Mass Spectrometric Detection of Variant Peptides Resulting from Nonsynonymous Nucleotide Differences. *J. Proteome Res.* **13**, 228–240 (2014).

132. Ruggles, K. V. *et al.* An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer*. *Molecular & Cellular Proteomics* **15**, 1060–1071 (2016).

133. Yeom, J. *et al.* A proteogenomic approach for protein-level evidence of genomic variants in cancer cells. *Sci Rep* **6**, 35305 (2016).

134. Wingo, T. S. *et al.* Integrating Next-Generation Genomic Sequencing and Mass Spectrometry To Estimate Allele-Specific Protein Abundance in Human Brain. *J. Proteome Res.* **16**, 3336–3347 (2017).

135. Alfaro, J. A. *et al.* Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines. *Genome Med* **9**, 62 (2017).

136. Shi, J. *et al.* Determining Allele-Specific Protein Expression (ASPE) Using a Novel Quantitative Concatamer Based Proteomics Method. *J. Proteome Res.* **17**, 3606–3612 (2018).

137. Spooner, W. *et al.* Haplosaurus computes protein haplotypes for use in precision drug design. *Nat Commun* **9**, 4128 (2018).

138. Abi-Rached, L. *et al.* Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLOS ONE* **13**, e0206512 (2018).

139. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**, D766–D773 (2019).

140. Satpathy, S. *et al.* A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**, 4348-4371.e40 (2021).

141. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat Methods* **8**, 937–940 (2011).

142. Eichacker, L. A. *et al.* Hiding behind Hydrophobicity: TRANSMEMBRANE SEGMENTS IN MASS SPECTROMETRY*. *Journal of Biological Chemistry* **279**, 50915–50922 (2004).

143. Djomehri, S. I. *et al.* Quantitative proteomic landscape of metaplastic breast carcinoma pathological subtypes and their relationship to triple-negative tumors. *Nat Commun* **11**, 1723 (2020).

144. Hoek, M., Demmers, L. C., Wu, W. & Heck, A. J. R. Allotype-Specific Glycosylation and Cellular Localization of Human Leukocyte Antigen Class I Proteins. *J. Proteome Res.* **20**, 4518–4528 (2021).

145. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat Methods* **14**, 513–520 (2017).

146. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4**, 923–925 (2007).

147. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).

148. da Veiga Leprevost, F. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat Methods* **17**, 869–870 (2020).

149. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527 (2016).

150. Hosseini, N., Mumphrey, M. B. & Cieslik, M. MHCnvex: Likelihood-based model for calling the copy number variations and loss of heterozygosity in MHC class I and II locus. 2023.01.29.526131 Preprint at https://doi.org/10.1101/2023.01.29.526131 (2023).

151. Meissner, T. B., Li, A. & Kobayashi, K. S. NLRC5: a newly discovered MHC class I transactivator (CITA). *Microbes and Infection* **14**, 477–484 (2012).

152. Steimle, V., Otten, L. A., Zufferey, M. & Mach, B. Complementation cloning of an MHC class II transactivator mutated in hereditary MHC class II deficiency (or bare lymphocyte syndrome). *Cell* **75**, 135–146 (1993).

153. Gobin, S. J. P., Zutphen, M. van, Woltman, A. M. & Elsen, P. J. van den. Transactivation of Classical and Nonclassical HLA Class I Genes Through the IFN-Stimulated Response Element1. *The Journal of Immunology* **163**, 1428–1434 (1999).

154. Gobin, S. J. P., Keijsers, V., van Zutphen, M. & van den Elsen, P. J. The Role of Enhancer A in the Locus-Specific Transactivation of Classical and Nonclassical HLA Class I Genes by Nuclear Factor κB1. *The Journal of Immunology* **161**, 2276–2283 (1998).

155. Horton, R. *et al.* Gene map of the extended human MHC. *Nat Rev Genet* **5**, 889–899 (2004).

156. Solberg, O. D. *et al.* Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. *Human Immunology* **69**, 443–464 (2008).

157. Norman, P. J. *et al.* Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. *Genome Res.* **27**, 813–823 (2017).

158. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* **36**, 875–879 (2018).

159. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

160. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443–1448 (2016).

161. Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* **13**, 260–269 (1967).

162. Forney, G. D. The viterbi algorithm. *Proceedings of the IEEE* **61**, 268–278 (1973).

163. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol* **21**, 265 (2020).

164. Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol* 1–11 (2023) doi:10.1038/s41587-023-01793-w.

165. Garrison, E. *et al.* Building pangenome graphs. 2023.04.05.535718 Preprint at https://doi.org/10.1101/2023.04.05.535718 (2023).

166. Carretero, F. J. *et al.* Frequent HLA class I alterations in human prostate cancer: molecular mechanisms and clinical relevance. *Cancer Immunol Immunother* **65**, 47–59 (2016).

167. Martin, A. M., Freitas, E. M., Witt, C. S. & Christiansen, F. T. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. *Immunogenetics* **51**, 268–280 (2000).