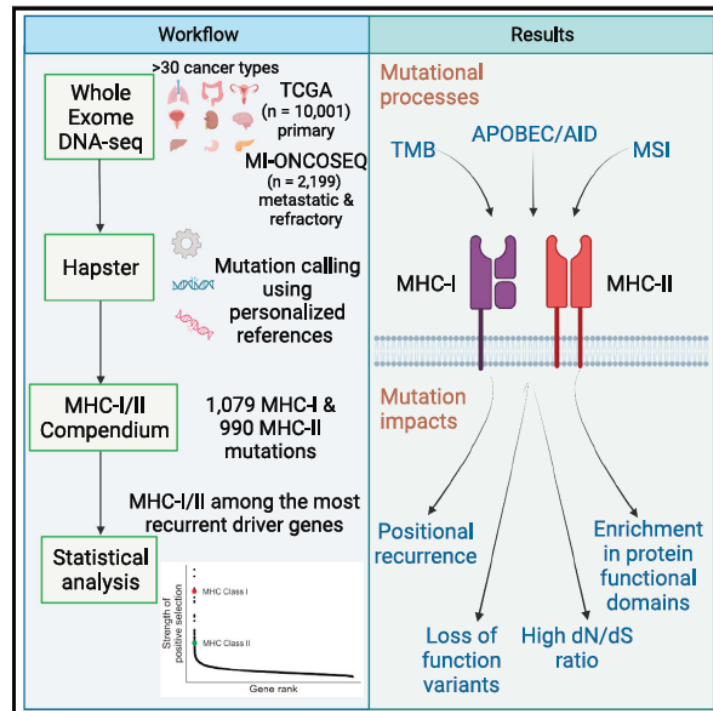


Distinct mutational processes shape selection of MHC class I and class II mutations across primary and metastatic tumors

Graphical abstract



Authors

Michael B. Mumphrey, Noshad Hosseini, Abhijit Parolia, ..., Malini Raghavan, Arul Chinnaiyan, Marcin Cieslik

Correspondence

mcieslik@med.umich.edu

In brief

Using the personalized mutation caller Hapster, Mumphrey et al. report a pan-cancer analysis of positive selection for MHC class I and class II mutations across primary and metastatic cancers. Their analysis provides evidence for the enrichment of inactivating MHC mutations in select cancers, as well as the mutational processes responsible.

Highlights

- Hapster detects MHC class I and II mutations with high sensitivity and specificity
- MHC genes are among the most recurrently mutated genes pan-cancer
- Tumor mutation burden and mutational processes shape the spectrum of MHC mutations
- MHC missense mutations are likely loss of function, disrupting B2M and antigen binding



Mumphrey et al., 2023, Cell Reports 42, 112965
August 29, 2023 © 2023 The Authors.
<https://doi.org/10.1016/j.celrep.2023.112965>



Resource

Distinct mutational processes shape selection of MHC class I and class II mutations across primary and metastatic tumors

Michael B. Humphrey,² Noshad Hosseini,² Abhijit Parolia,^{1,3} Jie Geng,⁴ Weiping Zou,^{4,6,8} Malini Raghavan,⁴ Arul Chinnaiyan,^{1,2,3,5,7,8,9} and Marcin Cieslik^{1,2,3,8,9,10,*}

¹Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA

²Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

³Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI 48109, USA

⁴Department of Microbiology & Immunology, University of Michigan, Ann Arbor, MI 48109, USA

⁵Department of Urology, University of Michigan, Ann Arbor, MI 48109, USA

⁶Center of Excellence for Cancer Immunology and Immunotherapy, University of Michigan, Ann Arbor, MI 48109, USA

⁷Howard Hughes Medical Institute, Ann Arbor, MI 48109, USA

⁸University of Michigan Rogel Cancer Center, Ann Arbor, MI 48109, USA

⁹Senior author

¹⁰Lead contact

*Correspondence: mcieslik@med.umich.edu

<https://doi.org/10.1016/j.celrep.2023.112965>

SUMMARY

Disruption of antigen presentation via loss of major histocompatibility complex (MHC) expression is a strategy whereby cancer cells escape immune surveillance and develop resistance to immunotherapy. Here, we develop the personalized genomics algorithm Hapster and accurately call somatic mutations within the MHC genes of 10,001 primary and 2,199 metastatic tumors, creating a catalog of 1,663 non-synonymous mutations that provide key insights into MHC mutagenesis. We find that MHC class I genes are among the most frequently mutated genes in both primary and metastatic tumors, while MHC class II mutations are more restricted. Recurrent deleterious mutations are found within haplotype- and cancer-type-specific hotspots associated with distinct mutational processes. Functional classification of MHC residues reveals significant positive selection for mutations disruptive to the B2M, peptide, and T cell binding interfaces, as well as to MHC chaperones.

INTRODUCTION

The immune system is capable of identifying and eliminating cancer cells via CD8⁺ T cell-mediated cytotoxicity.¹ To avoid this destruction, successful cancers often evolve strategies to disrupt T cell immunity, such as overexpression of the immunosuppressive PD-L1² or reduced expression of key proteins such as the major histocompatibility complex (MHC) class I molecules.³ Research into how tumors escape T cell surveillance has led to immunotherapies that restore cancer immunity in select patients.^{4–9} However, even the most promising immunotherapies still only provide a clinical benefit in a minority of cases.¹⁰ Understanding the mechanisms that lead to primary and acquired resistance to T cell-based immunotherapies will be critical for the continued improvement of patient outcomes.

T cells are able to identify malignant cells via the presence of mutant peptides known as neoantigens.¹ In order to detect these neoantigens, T cells require the peptides to be presented at the cell surface by the MHC. The MHC class I proteins present neoantigens to cytotoxic CD8⁺ T cells¹ and, as a result, are directly

involved in the destruction of malignant cells. Consistent with this role as a tumor suppressor, correlations between decreased MHC class I expression and poor outcomes have been repeatedly observed across many cancers.^{11–15} The MHC class II proteins present neoantigens to CD4⁺ T cells and play an important role in tumor suppression.^{16–19} Their expression can be induced in most cell types,²⁰ including cancer cells,^{21–23} and MHC class II-restricted neoantigen vaccines have been shown to trigger an anti-tumor immune response,²⁴ suggesting that cancer-cell-specific MHC class II expression may also play a tumor-suppressor role.

Identifying the mechanisms that lead to loss of the MHC will be key to understanding resistance to T cell-based immunotherapies. However, genetic studies of the MHC are hindered by the extreme polymorphism of the MHC genes.²⁵ To address this issue, we created Hapster, a generalized algorithm that constructs personalized reference sequences for alignment and mutation calling within polymorphic genes. The high sensitivity and specificity of the identified mutations enabled us to study the mutational and evolutionary processes driving somatic MHC loss in primary and metastatic tumors.



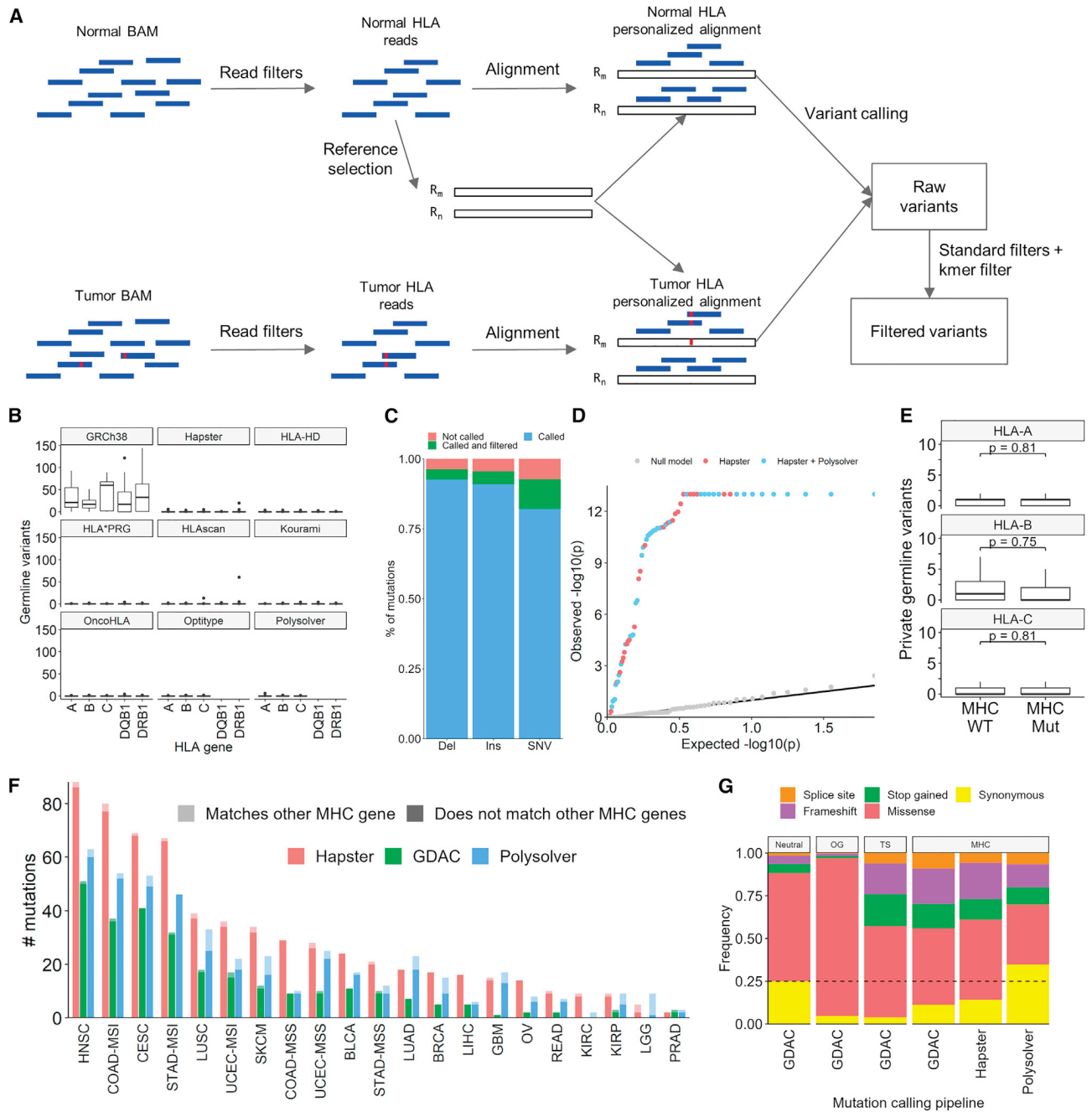


Figure 1. Hapster algorithm and validation

(A) Simplified overview of the Hapster algorithm. For each gene, Hapster infers optimal reference sequences from normal sequencing data, realigns to these personalized references, and calls mutations. For a complete description, see [STAR Methods](#).

(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 haplotypes. A perfect reference sequence should produce 0 apparent germline variants.

(C) Fraction of simulated insertions or deletions (indels) and SNVs that were either called and passed all filters, were called and filtered by either Hapster or Mutect2, or were never called. Shown here are mutations simulated at a VAF of 0.45 and a coverage of ~100x.

(D) 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 identified by Hapster alone (red) or by both Hapster and Polysolver (blue) from WES data. A comparison is shown to randomly generated alternate bases (gray), which are only supported by noisy reads and follow the null model (diagonal black line).

(E) Boxplot showing the number of private germline variants observed per tumor in those cases with or without somatic MHC mutations. Wilcoxon rank sum test p values with BH correction.

(legend continued on next page)

RESULTS

Hapster allows for more sensitive and specific somatic mutation calls in MHC genes

Hapster is a complete mutation calling pipeline that (1) selects personalized reference haplotype sequences, (2) prunes contaminant and misaligned reads, and (3) detects^{26,27} and filters variants (Figure 1A). For the first function, in principle, any existing human leukocyte antigen (HLA) haplotyper^{28–34} could be used to identify HLA haplotype sequences. However, in practice, existing haplotypers often report HLA types for which only the sequences for exons 2 and 3 are known³⁵ (Figure S1A). This is insufficient for somatic mutation calling where a full-length sequence is needed to call variants in all exons/introns. We therefore developed a generalized haplotyping algorithm to return full-length sequences for MHC class I and class II genes. For the second and third functions, we developed MHC-specific strategies for pruning of contaminant and misaligned reads originating from other MHC genes and pseudogenes, as well as for the identification of false positive somatic mutation calls. A detailed overview of the Hapster pipeline is included (STAR Methods).

To benchmark the haplotype inference portion, we used a set of 69 whole-exome sequencing (WES) samples from the 1000 Genomes Project with reported haplotype calls.^{33,36} To compare Hapster with other methods, we called germline variants in WES data relative to each haplotyper's inferred haplotype for each individual. Using sequencing data as ground truth, a perfectly identified reference would lead to no germline variants being identified. We observe that, relative to the standard GRCh38 reference, there is a median of 17–38 germline variants observed per allele. All tested haplotypers improve upon this, with each having a median of 0 and a mean of <0.5 observed germline mutation per allele (Figure 1B), similar to the genome-wide average of ~1 variant per kilobase.³⁷ In a larger population of 10,001 normal tissue samples from TCGA, Hapster identified 0–3 germline variants per allele (Table S1).

To assess Hapster's sensitivity, we simulated 200 synthetic MHC haplotypes with a random mutation, followed by simulated WES at depths ranging from 5× to 100× and variant allele fractions (VAFs) of 0.025–0.45. Of the 200 simulated mutations, 94% (187/200) were successfully identified (Figure 1C) at 100× 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 (Table S2). As such, they are false negatives due to the capture kit design rather than to Hapster's algorithm. When looking at results over the range of coverages and VAFs, we see that as coverage and VAF decrease, sensitivity decreases as expected

due to lower read support for variants (Figure S1D). 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 S1E).

To assess specificity, we called somatic mutations in all 450 samples from the TCGA head and neck squamous cell carcinoma (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).

To assess Hapster's accuracy using an orthogonal sequencing technology, we used Hapster to call somatic mutations in WES data and then determined if these same mutations were supported by paired RNA sequencing (RNA-seq) data. Established RNA-seq validation methods are not ideal, as they rely on alignment of reads to a reference, which would be inappropriate in validating Hapster. We therefore developed an orthogonal alignment-free kmer-based approach to determine if the reads support variants in the RNA based on a beta-binomial model of sequencer error (STAR Methods), avoiding reference selection or alignment biases. Of the 80 variants 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 ($p < 0.05$ [Benjamini-Hochberg adjusted (BH-adj.)], Figure 1D), and 4 (5%) were truncating with evidence of nonsense-mediated decay (Figure S1B). This leaves only 5 variants (7%) without RNA evidence, potentially due to the limitations of our model, low tumor cellularity, loss of heterozygosity (LOH), or transcriptional silencing.³⁸

For a second orthogonal validation, we performed Sanger sequencing on tumors from MI-ONCOSEQ³⁹ with sufficient DNA or tissue samples. All 14 candidate variants called by Hapster were detected in the Sanger chromatograms from tumor specimens while being absent in traces from patient-matched normal tissues (Figure S1C). Addressing the possibility of germline variants being miscalled as somatic due to poor reference selections, we found no evidence of enrichment of somatic mutations in cases with higher numbers of private germline variants (Figure 1E).

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.^{34,40} 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 (Table S3; Figures 1F and S2A). Next we examined variant allele frequency (VAF) distributions. Hapster tended to report slightly higher VAFs (Figure S2B) due to preserving more variant reads

(F) 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.

(G) 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 (OGs): KRAS, PIK3CA, IDH1, CTNNB1, FOXA1, BRAF, AKT1, EGFR. Tumor suppressors (TSs): TP53, RB1, PTEN, APC, BRCA2, VHL. Neutral genes: all others.

See also Figures S1 and S2 and Tables S1, S2, and S3.

(Figure S2C) rather than failing to call low VAF mutations. Conversely, variants exclusive to Hapster tended to be low VAF mutations, supporting its higher sensitivity (Figure S2D). We next performed an exhaustive search for potential false positives resulting from 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-adj.) but similar to the GDAC pipeline (4%, Fisher's exact test $p = 1$, BH-adj.) (Figures 1F and S2E; Table S3). An analysis of synonymous mutation calls shows the apparently recurrent synonymous variants p.T214T and p.A269A, which are identified as somatic mutations by Polysolver (Figure S2F). These mutations are unlikely to be under extreme positive selection 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. We next 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 were under-represented when compared to neutral genes, consistent with what would be expected for a potential driver gene (Figure 1G). In contrast, we found that Polysolver had an over-representation of synonymous calls, many of which can likely be attributed to misaligned reads originating from non-classical MHC class I genes (Figure S2F; Table S3).

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

To comprehensively characterize MHC class I and class II mutation rates in human cancer, we analyzed 10,001 tumors across 35 cancer types from TCGA and 2,199 tumors across 24 cancer types from MI-ONCOSEQ³⁹ (Table S4), for a total compendium of 2,069 MHC class I and class II mutations (Figure 2A; Table S5). Samples from TCGA are mainly primary tumors, with the exception of the melanoma cohort (skin cutaneous melanoma [SKCM]), which consists only of metastatic samples. Microsatellite unstable (MSI) tumors are immunologically distinct due to their significantly higher neoantigen burden,⁴¹ and we have therefore separated them from their microsatellite stable (MSS) counterparts within the colon (colon adenocarcinoma [COAD]), stomach (stomach adenocarcinoma [STAD]), and endometrial (uterine corpus endometrial carcinoma [UCEC]) TCGA cohorts. While some other cancers also have distinct subtypes, such as breast cancer (BRCA) estrogen receptor (ER)^{+/−} and HNSC human papillomavirus (HPV)^{+/−}, no significant difference in MHC mutation rates was observed between subtypes (BH-adj. chi-squared $p = 0.27-1$). These cohorts were therefore not subdivided for further analyses. Samples from MI-ONCOSEQ are metastatic/refractory and represent a significantly more advanced form of disease compared with the corresponding TCGA cohorts, with cases generally having received multiple forms of prior systemic therapy for their primary cancer, and one or more rounds of therapy for their metastatic cancer.

Mutations were in general distributed uniformly across the gene body but occasionally concentrated within prominent hotspots (Figures 2A and S3A). We found that for MHC class I, *HLA-A* and

HLA-B contained significantly more mutations than *HLA-C*, and that for MHC class II, *HLA-DRA* contained significantly more mutations than all other MHC class II genes except for *HLA-DQB1* (Figure 2B). Within each *HLA* gene, no allele was found to bear an excess of mutations. In primary tumors, we noted substantial variation in both mutational frequency and functional consequences across tumor types and MHC gene classes (Figure 2C). We found non-synonymous MHC class I and class II mutations in 10.5% of primary tumors (ranging from 2.7% to 72.5% across cancer types) (Figure S3B), with 5.6% (range: 0.2%–62.3%) of patients harboring an MHC class I and 5.7% (range: 1.1%–21.7%) harboring an MHC class II somatic variant. Consistent with previous reports that MSI tumors are under strong pressure to lose MHC function,^{42,43} the COAD-MSI, STAD-MSI, and UCEC-MSI cohorts make up 3 of the top 4 cohorts for MHC class I mutations (Figures 2D and S3B), with the majority of variants being loss-of-function (LOF) frameshifts or stop gains. MHC class II mutations were also most prevalent in cancers with high mutation burden including MSI tumors and melanoma (Figure 2E). However, LOF mutations in the top-mutated cohorts were less frequent and the variation in mutation rates across cancer types was lower compared with MHC class I (Figures 2E and S3B). Interestingly, there was a slight pan-cancer association between both MHC class I and class II mutations and immune cell infiltration after adjusting for tumor purity (Figure S3C), but further cohort-level analysis showed that no individual cancer type had a significant association after multiple testing correction.

Prevalence of MHC class I and 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 the immunological differences between the primary and metastatic tumor microenvironment (TME).^{39,44–47} Overall, we observed non-synonymous MHC class I and class II mutations in 7.6% (range: 3.3%–20%) of metastatic/refractory patients, with substantial variation in mutational frequency and functional consequences between cancer types (Figures 2F, 2G, and S3D). 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 S6). For 15/17 pairings (88%), there were no significant changes in primary vs. metastatic MHC class I or class II mutation rates. However, for prostate and breast cancers, we observed a significant increase in MHC class I mutations in metastatic cancers compared with primary cancers (Figure S3E; prostate: $F(1, 909) = 9.35$, $p = 0.03$; breast: $F(1, 1140) = 12.8$, $p = 0.01$, BH-adj.). No significant differences were seen in MHC class II mutations.

Overall, these data provide a comprehensive look at MHC class I and 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. 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.

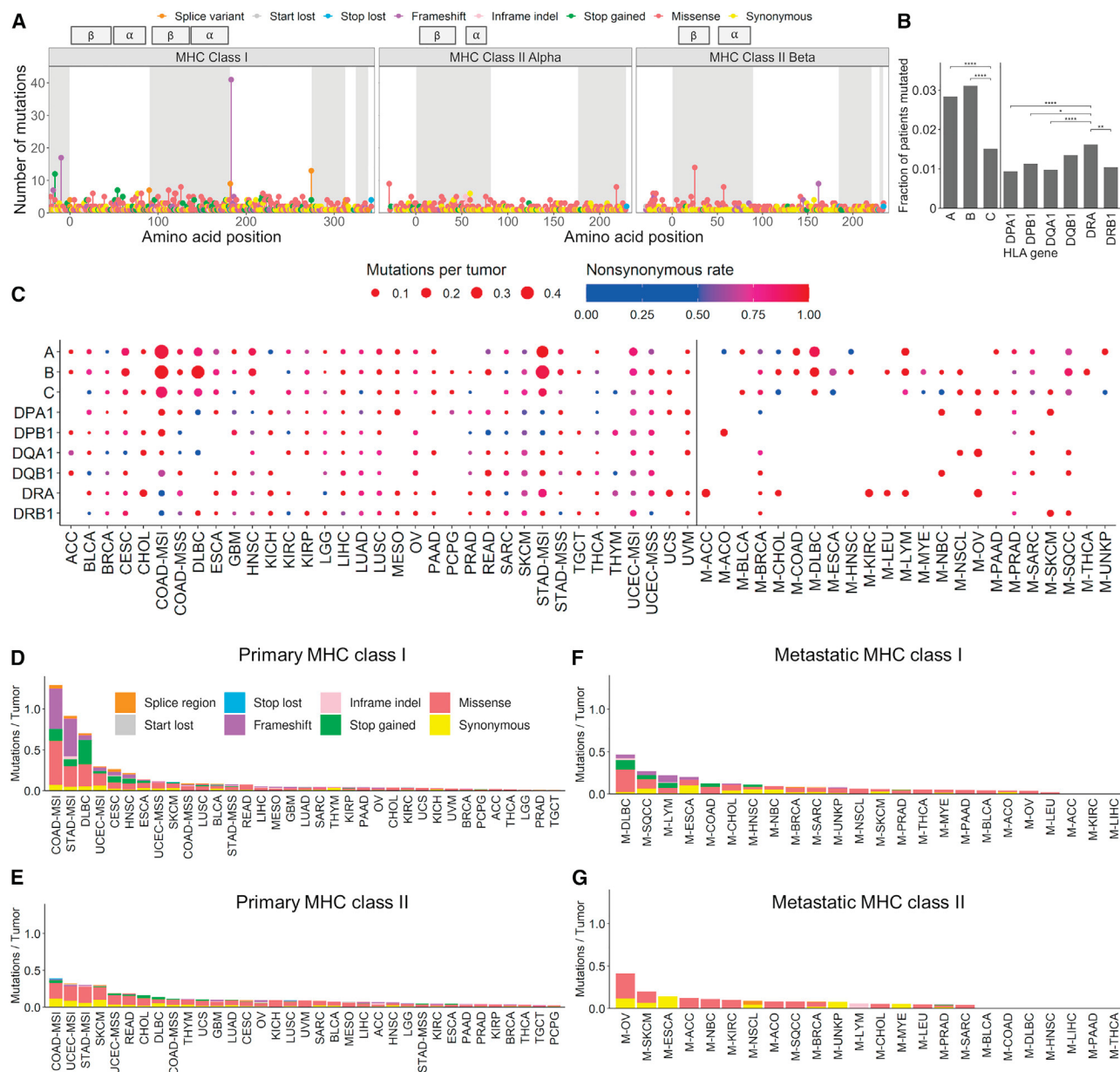


Figure 2. Compendium of MHC class I and class II mutations in primary and metastatic tumors

(A) Distribution of all observed mutations in both primary and metastatic cancers across the coding region of the MHC genes. Binding pocket secondary structures are noted above.

(B) Significant differences in the prevalence of non synonymous mutations and indels of individual MHC class I and class II genes. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$, BH adj. Fisher's exact test.

(C) Cohort specific mutation rates for MHC class I and 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 non synonymous/indel mutations.

(D) Cohort summaries of coding region mutations in MHC class I (D and F) and class II (E and G) genes in primary (D and E) and metastatic (F and G) cancers. Values are scaled by the number of individuals within each cohort.

See also Figure S3 and Tables S4 and S5.

Positive selection of non-synonymous MHC somatic mutations

To identify positive selection of functional mutations within the MHC genes, we applied CBaSE⁴⁸ to each primary and metastatic cohort from TCGA and MI-ONCOSEQ. HLA genes and haplo-

types are codominant, and each allele presents a largely unique set of neoantigens.⁴⁹ Additionally, specific T cell responses are often immunodominant and mounted against only a few of the presented neoantigens such that the mutation of a single HLA allele may result in the complete inability to present an

immunodominant neoantigen. Given this, 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,⁵⁰ taking into account the increased genomic length of this combined set of genes. In primary cancers, CBaSE identified 6 cohorts (COAD-MSI, STAD-MSI, diffuse large B cell lymphoma [DLBC], cervical squamous cell carcinoma [CESC], HNSC, LUSC) with statistically significant evidence for positive selection of non-synonymous variants in the MHC class I genes and 3 cohorts (cholangiocarcinoma [CHOL], kidney chromophobe [KICH], uveal melanoma [UVM]) for the MHC class II genes (Figure 3A). By this measure, the MHC class I genes are tied for 7th and the MHC class II genes 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 lymphoma cohorts (M-DLBC, M-LYM), making them tied for 6th most recurrent pan-cancer driver gene by number of cohorts significantly mutated (Figure 3B). 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 3C and 3D). 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 3C and 3D), and in primary cancers, the MHC class II genes scored in the top 0.5% (Figure 3C). 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.

We next 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 3E). In tumors with multiple hits, we see that in each of these cohorts, the majority (63%–100%) of cases have at least one clonal variant (Figures S4A and S4B). When there is at least one leading clonal mutation, we see that co-occurring MHC class I mutations are also primarily clonal in the DLBC (80%) and HNSC (88%) cohorts, while in the remaining cohorts, we observe a mix of clonal and subclonal mutations (Figures S4A and S4C). In cohorts showing no evidence of positive selection, the proportion of clonal mutations were significantly lower in both *HLA-A* and *HLA-B* (Figure 3E), 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 2B).

Impact of tumor mutation burden on MHC class I and 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 S4D).

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 (Figures S4E–S4G). 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.

Functional consequences of MHC class I and class II mutations

We next characterized the distributions of mutation functional consequences 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 passengers⁵¹ (Figure 3F, “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 3F, “unselected”), suggesting that these mutations 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-adj.). 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 with the expected neutral rate of ~12%. The DLBC, CESC, and 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 S4H). Notably, frameshift mutations in MHC class II were rare even in MSI tumors but were unexpectedly common in some MSS tumors including glioblastoma (GBM), ovarian cancer (OV), and liver hepatocellular carcinoma (LIHC). These cohorts were also depleted of synonymous mutations (Figure 3G).

Similar to the TCGA DLBC cohort, the refractory M-DLBC cohort showed both a high mutation rate and a bias toward truncating mutations in the MHC class I genes (35%). Other non-DLBC refractory lymphomas (M-LYM) had a lower MHC class I mutation rate but still had a bias toward truncating mutations (65%) (Figures 3F and S4I). The lymphomas alone account for 52% of stop gains observed across all MI-ONCOSEQ cohorts

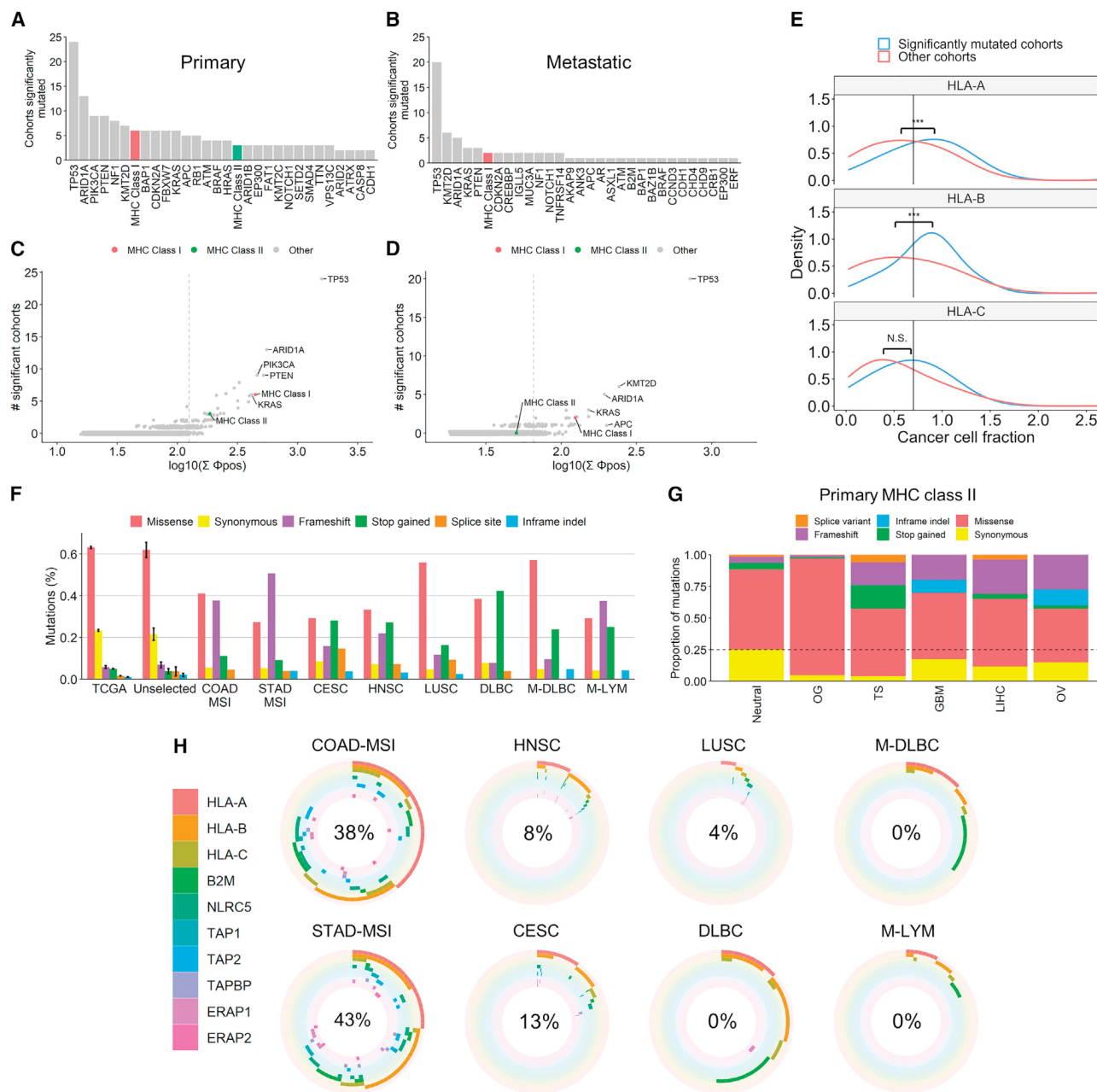


Figure 3. Evidence for strong positive selection and deleteriousness of MHC somatic mutations

(A and 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 and 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} . (E) Cancer cell fraction (CCF) of MHC class I variants in TCGA cohorts showing significant evidence of positive selection compared with all other cohorts. Vertical line shows 70% CCF, above which mutations are considered clonal. *** $p < 0.001$, Wilcoxon rank sum test after BH correction. (F) Proportion of functional consequences observed in various groups: “TCGA,” 2,600,654 pan cancer mutations from TCGA an approx. neutral model; “unselected,” cohorts showing no evidence of positive selection; others, 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. (G) Functional consequences of MHC class II mutations in select primary cohorts. Mutational consequence distribution of known OGs (KRAS, PIK3CA, IDH1, CTNNB1, FOXA1, BRAF, AKT1, EGFR) and TSs (TP53, RB1, PTEN, APC, BRCA2, VHL) are shown. (H) Sample level co occurrence of mutations in either the MHC class I or APM genes within positively selected cohorts. Percentage values show percentage of mutated samples containing a hit in both the MHC class I and APM.

(11/21) despite containing only 7% of patients. The HNSC, CESC, and LUSC cohorts in TCGA are all squamous cell carcinomas that 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 (squamous cell carcinoma) 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 S4J](#)). Metastatic MSI tumors are underrepresented in MI-ONCOSEQ, preventing any comparison with primary MSI tumors. 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.

Patterns of mutual exclusivity and independence of MHC mutations

We next looked at the relationships between deleterious mutations in the MHC class I genes and the APM (antigen presentation machinery)⁵² ([Figure 3H](#)). Other genes linked to MHC class I have been identified as cancer driver genes (e.g., beta-2 microglobulin [*B2M*]⁵³), and it has been shown that driver genes that fall within the same pathway frequently show mutual exclusivity. This effect is strongest in the lymphomas, where there is significant mutual exclusivity between MHC mutations and the APM ($p = 0.02$, BH-adj.), 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-adj.), 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-adj.). Finally, we looked at potential co-mutation of the MHC class I genes with known driver genes. The strongest observed co-occurrences were found in the HNSC cohort with deleterious mutations of *CASP8* ($p = 4.5e-10$), a gene that plays a role in an alternative pathway for destruction of malignant cells by the immune system,⁵⁴ and *HRAS* ($p = 9.2e-5$), a member of the RAS family of oncogenes that has been shown to be associated with increased immune activity within head and neck cancers.⁵⁵

Mutational processes shape cancer-type-specific MHC mutational patterns

For MSI cancers, mismatch repair deficiency (MMRd) is the primary mutational process leading to large numbers of insertions or deletions (indels) within microsatellites.⁵⁶ We have already observed a high rate of frameshift indels within MSI tumors ([Figure 3F](#)) and notable hotspots ([Figures 2A](#) and [S3A](#)), and upon further investigation, we found that the majority of these frameshifts (57/63, 90%) are due to single-base-pair indels at homopolymer microsatellites ([Figure 4A](#)) that occur at a rate much higher than observed in MSS cancers ($p < 1e-16$) ([Figure 4B](#)). MSI-associated indels were preferentially in longer homopolymers, while MSS-associated indels showed no relationship with homopolymer length ($p < 0.001$) ([Figure 4C](#)), which is consistent with the mutational signature for MMRd.⁵⁷

For the lymphoma and squamous cell carcinoma cohorts, we observed a striking number of positionally recurrent ($n > 2$)

stop-gain mutations ([Figure 4D](#)), all of which were C>T or C>A mutations. This is consistent with the process of cytosine deamination, which is frequently observed in lymphomas due to activation-induced cytosine deaminase (AID)⁵⁸ and in squamous cell carcinomas due to the closely related APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide) family of enzymes.⁵⁹ 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* ($W = A/T$, $R = A/G$) ([Figures 4D–4F](#)). 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* ([Figures 4D](#), [4G](#), and [4H](#)). Further, mutational signatures SBS2 and SBS13, which have been reported to be associated with APOBEC activity,⁶⁰ are significantly more active in squamous cell carcinomas with observed stop-gain mutations ([Figure 4I](#)). No further 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-adj. $p = 0.03$), were found.

Altogether, these observations suggest that truncating mutations within the MHC genes originate due to specific mutational processes active within select cancer types. The active processes are responsible for producing highly immunogenic tumors that are under pressure to select truncating mutations within the MHC class I genes and are directly responsible for creating the majority of the LOF mutations. Haplotypes harboring homopolymer repeats and AID/APOBEC templates are therefore more susceptible to this immune-escape mechanism.

Missense mutations are enriched in specific MHC functional domains

Missense mutations are often difficult to interpret as they can be LOF, neutral, gain of function, or even neomorphic.^{61,62} We hypothesized that deleterious missense mutations 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 ([Figure 5A](#); [STAR Methods](#)), which we compared to the observed mutations across MHC functional domains established through systematic expert knowledge- and crystal structure-guided annotation of individual amino acids within the MHC class I proteins ([Figures 5B](#) and [S5A–S5C](#); [Table S7](#)). The first, “simulated” null model was based on HLA mutations generated randomly taking into account HLA sequence trinucleotide contexts and observed mutational signature activities within each positively selected cohort ([STAR Methods](#)). The second, “observed” null model was based on 251 actual mutations called in cohorts that showed no evidence of positive selection. We first compared the observed null dN/dS (nonsynonymous to synonymous ratio) ratios with the simulated null ratios across all functional domains ([Figure 5C](#)). 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-adj.), showing that there are no significant differences between the

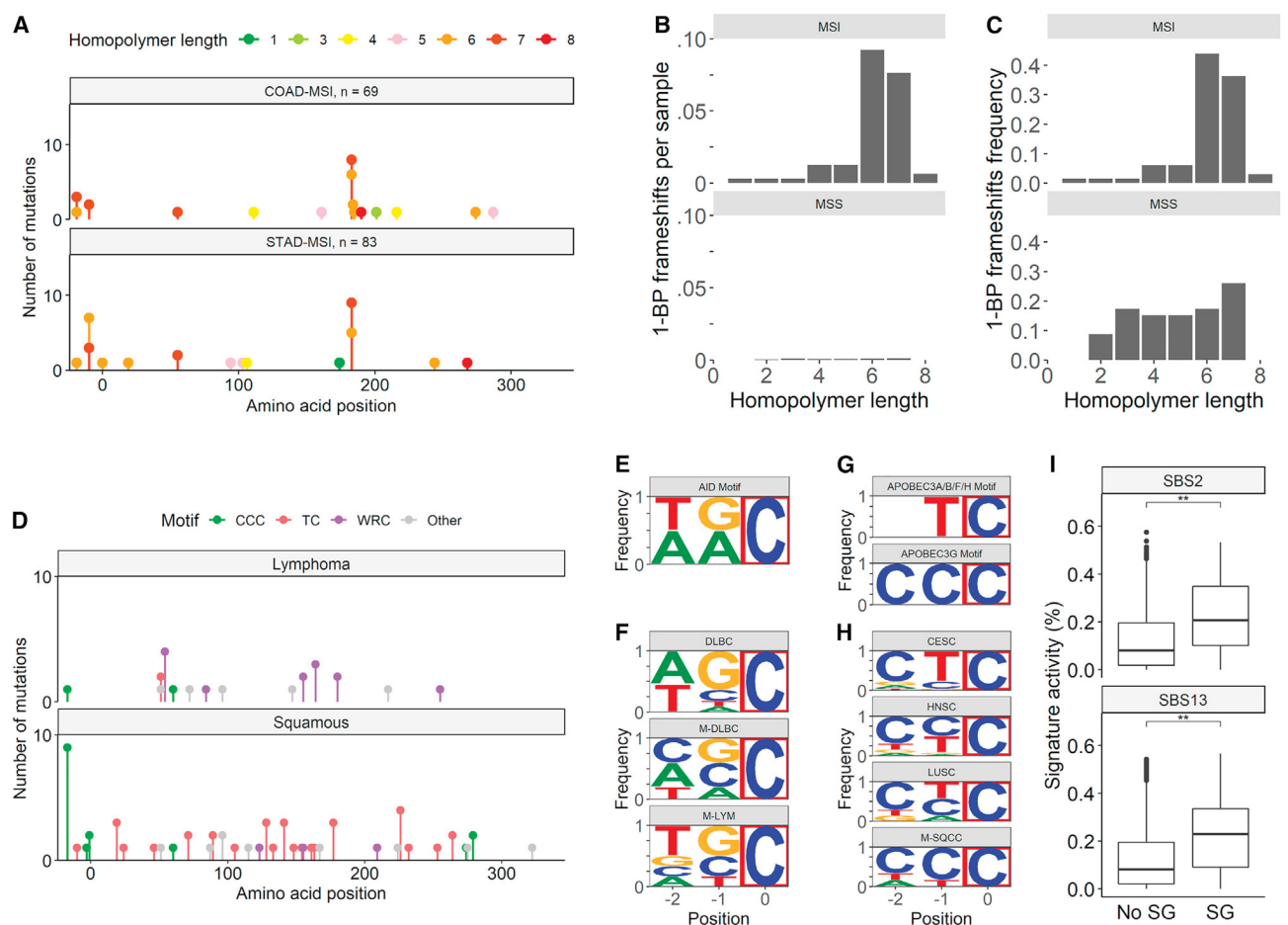


Figure 4. Distinct mutational processes shape cancer-type-specific MHC mutational patterns

(A) Position of observed frameshift mutations within the MHC class I genes of COAD MSI and STAD MSI tumors. Colors show length of homopolymer micro satellites 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.

(D) Position of observed stop gain mutations within the MHC class I genes of lymphomas and squamous cell carcinomas. Colors show motif of mutated position.

(E and F) Canonical motif for AID (E) and DNA motifs for stop gain mutations (F) observed in lymphoma cohorts. Mutated base marked with red box.

(G) Canonical motifs for APOBEC proteins.

(H) DNA motifs for stop gain mutations observed in squamous cell carcinoma cohorts. Mutated base marked with red box.

(I) Percentage of mutations within squamous cell carcinomas with (stop gain [SG]) or without (no SG) observed SGs that can be attributed to signatures SBS2 and SBS13, which have been associated with APOBEC activity. ** $p < 0.01$, BH adj. t tests.

two null models. In contrast, in cohorts showing evidence of positive selection, the local dN/dS fold change was also normally distributed (Shapiro-Wilk $p = 0.98$, $p = 0.433$) but with a mean significantly above 0 when compared with both the simulated null (mean = 0.85, Cohen's $d = 0.63$, $p = 0.03$, BH-adj.) and the observed null models (mean = 1.13, Cohen's $d = 0.81$, $p = 0.02$, BH-adj.) (Figure 5C). This shows a strong general trend toward excess non-synonymous mutations across all annotation regions.

We next examined which functional domains had the highest enrichment of non-synonymous mutations in both primary and metastatic tumors (Figures 5D and 5E). Compared to both the null models, 5 of the 7 annotated functional domains showed a 2-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 with primary tumors (Figures 5D and 5E), while on the other hand, residues involved specifically in the B2M, T cell receptor (TCR), and CD8 binding interfaces show strong enrichment only in the primary tumors (Figures 5D and 5E).

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 (Figures 5B, 5D, and

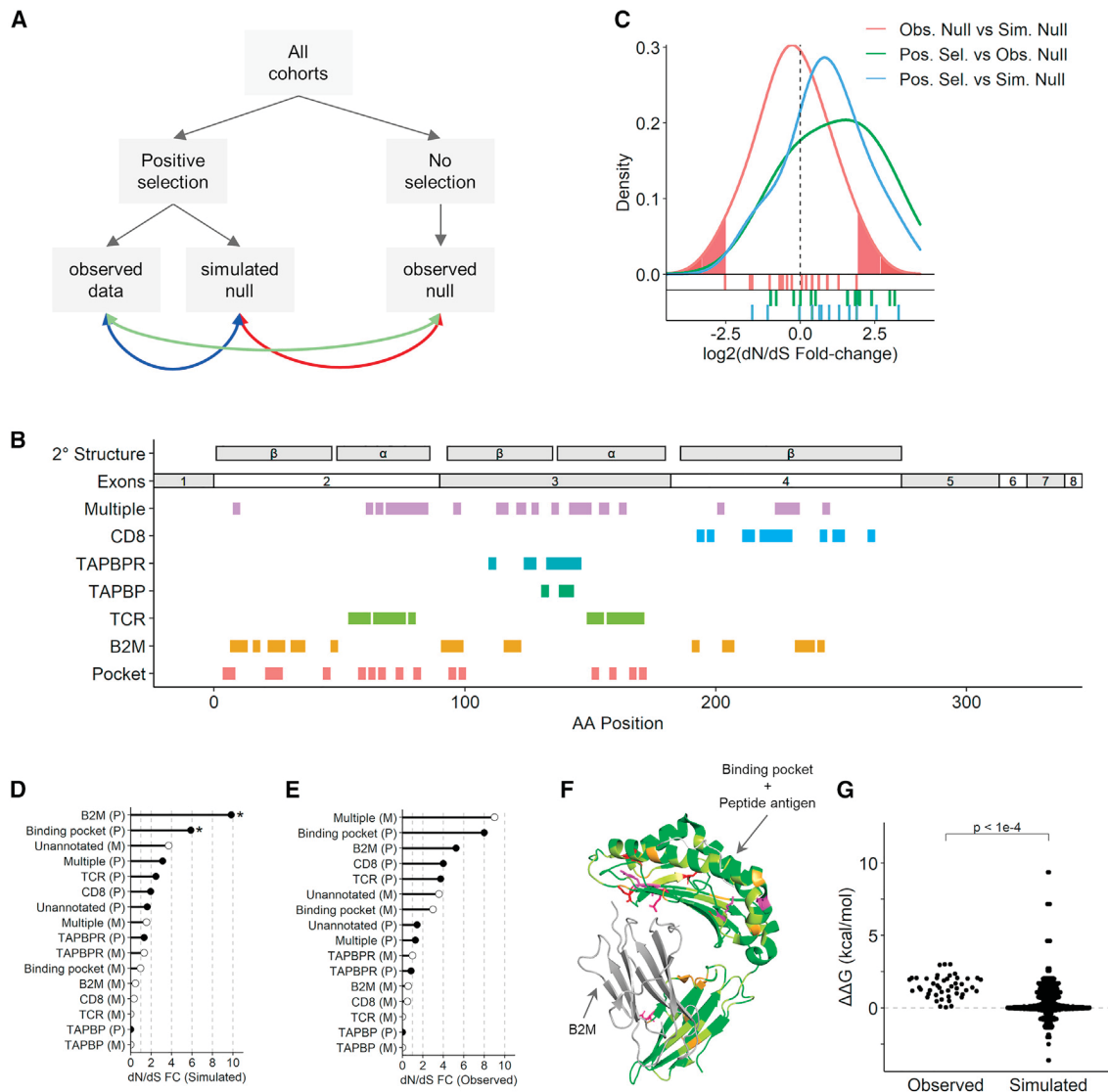


Figure 5. MHC class I missense mutations are enriched within specific protein functional domains

(A) Schematic representation of the constructed null models and their comparisons (arrows, colors match comparisons in C).

(B) Schematic overview of HLA proteins showing secondary structure, exon boundaries, and amino acid interactions with various binding partners.

(C) Distribution of observed vs. simulated dN/dS ratio fold change for amino acids that are predicted to interact with various MHC interacting partners. Filled regions show tails of empirical null distribution. (light red: top and bottom 5%; dark red: top and bottom 1%).

(D and E) dN/dS ratio fold changes vs. simulated/theoretical (D) and observed/empirical (E) 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.

(F) 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 are in gray.

(G) Change in binding energy due to mutations in amino acids at the HLA:B2M interface as predicted by the software SSIPe. Significance determined by Monte Carlo simulation of 10,000 runs.

See also [Figure S5](#) and [Table S7](#).

5E). 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 I-B2M complex⁶³ ([Figure 5F](#)), 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 5F](#)), 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.^{13,48,64} To determine whether missense mutations at the MHC class I:B2M interface are potentially deleterious, we used SSiPe⁶⁵ to predict the change in binding energy resulting from each observed mutation in comparison with that of our previously simulated null mutations (Figure 5G). Observed mutations had a significantly higher predicted $\Delta\Delta G$ than simulated mutations (median $\Delta\Delta G$ 1.37 vs. 0.15, $p < 1e-4$; Figure S5D). Additionally, 42% of observed mutations had a $\Delta\Delta G > 1.5$ kcal, the threshold suggested by SSiPe as evidence for significant disruption of a protein-protein interface.

Overall, our data suggest that the observed somatic mutations in residues at the MHC class I:B2M interface are more disruptive than expected by chance. Interestingly, even though *B2M* itself was enriched for LOF mutations (Figures S5E and S5F), there were no observed missense mutations in *B2M* in residues at the MHC class I interface within positively selected cohorts (Figure S5G). 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.⁶⁶ Altogether, these findings demonstrate that observed missense MHC class I mutations are strongly enriched at residues enabling MHC antigen-binding complex formation and that they have a likely deleterious function.

DISCUSSION

Despite overwhelming evidence from immunohistochemistry (IHC) studies demonstrating frequent loss of MHC expression pan-cancer,³⁸ there is still much unknown about the underlying molecular mechanisms. To address this issue, we developed Hapster, an algorithm that can detect germline and somatic mutations in the MHC locus with high sensitivity and specificity. Using Hapster, we identified the cancer types most affected by somatic mutation of the MHC, identified the mutational processes that may generate these mutations, quantified positive selection of MHC mutations pan-cancer, and characterized the deleterious consequences of truncating and missense mutations at the level of functional residues and protein domains.

We have identified six cancer types that are significantly enriched for somatic non-synonymous mutations of MHC class I. Notably, all of these cancers display above-average levels of tumor-immune infiltration,⁶⁷ and at the pan-cancer level, MHC mutant tumors are significantly more likely to have approved immunotherapies ($p < 2.2e-16$, odds ratio [OR]: 2.29). We consider it likely that immunologically “hot” tumors are both more likely to respond to immunotherapies and to be under higher selective pressure to escape T cell surveillance. Given that nearly 10% of patients harbor a functional mutation in the MHC class I genes that could render them invisible to T cells,³⁸ it will be vital to characterize MHC loss in the context of immunotherapy to better understand its role in primary and acquired resistance to immune checkpoint inhibitors (ICIs).

Metastatic cancers typically have a higher TMB and therefore may be more immunogenic than primary tumors. However, metastases may also originate from less immunogenic subclones and from seed locations that are immunosuppressive.^{68,69} IHC

studies have also observed both MHC⁺ primary cancers that produce MHC[−] metastases and MHC[−] primary tumors that produce metastases that regain MHC function.⁷⁰ 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 those of primary tumors.

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 responses¹⁹ and can act as a therapeutic target for CD4⁺ T cell-based cancer vaccines.^{24,71} 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 genes. 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. What was not expected, however, was the lack of a relationship between mutations of MHC class I and class II given that both sets of proteins act to present neoantigens 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 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 or by 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 and identified MMRd, APOBEC activity, and AID activity as major drivers of MHC mutations in specific cancer types. In MSI cancers, squamous cell carcinomas, and lymphomas, the most likely mechanism to acquire MHC mutations is specific to the mutational processes active in these cancers.

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 focused on immunotherapy-naïve tumors, future studies examining tumors post-immunotherapy may reveal MHC mutations in acquired immunotherapy resistance. While we have focused on MHC mutations, in other cancer types, MHC function may be lost via structural variants in the locus, LOH,¹⁵ transcriptional repression, or post-translational inactivation. A multi-faceted examination of MHC abnormalities will provide a deeper understanding of mechanisms that can lead to loss of MHC expression, immune evasion, and immunotherapy resistance.

Limitations of the study

We note a number of limitations to our study. First, while our total number of cases is high at 12,200, select cohorts have relatively low sizes, which reduces our power to detect positive selection. Second, our investigation of the MHC class II genes is based on the hypothesis that MHC class II expression is induced in some tumors to promote an anti-tumor response. Given the pan-cancer

nature of our analysis, we may have reduced power to detect positive selection of MHC class II mutations relative to an experiment that is restricted to only those tumors that have gained MHC class II expression. Finally, we note that our study only investigates SNVs and indels within the MHC genes, and a study investigating large-scale genomic events such as LOH or homozygous deletion will be necessary to complete the picture for somatic loss of the MHC.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
 - Lead contact
 - Materials availability
 - Data and code availability
- **METHOD DETAILS**
 - Hapster algorithm
 - Reference selection validation
 - Alignment and mutation calling
 - Simulation validation
 - Label-flipping validation
 - RNA validation of somatic mutations
 - Sanger sequencing validation of somatic mutations
 - Tumor mutational burden calculations
 - Positive selection of somatic mutations
 - Cancer cell fraction calculations
 - Mutual exclusivity and co-mutation analyses
 - AID/APOBEC mutational signature activity
 - dN/dS null model simulations
 - MHC amino acid annotations
 - PPI binding energy predictions
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.celrep.2023.112965>.

ACKNOWLEDGMENTS

The results published here are in part based upon data generated by the TCGA Research Network (<https://www.cancer.gov/tcga>). This manuscript was supported in part by funding by the Institute for Cancer Informatics (ICI398672) and the V Cancer Research Foundation to M.C. M.B.M. was supported by the Genome Sciences Training Program (T32 HG00040) from the National Genome Research Institute of the National Institutes of Health and the Advanced Proteogenomics of Cancer Training Program (T32CA140044).

AUTHOR CONTRIBUTIONS

Conceptualization, M.B.M. and M.C.; methodology, M.B.M., N.H., M.R., and M.C.; software, M.B.M.; validation, M.B.M. and A.P.; formal analysis, M.B.M. and N.H.; investigation, M.B.M. and N.H.; resources, A.C. and M.C.; data curation, M.B.M. and J.G.; writing – original draft, M.B.M., N.H., J.G., and M.C.; writing – review & editing, M.B.M., N.H., A.P., W.Z., M.R., A.C., and M.C.; visualization, M.B.M., N.H., and A.P.; supervision, A.C. and M.C.; funding acquisition, A.C. and M.C.

DECLARATION OF INTERESTS

An innovation disclosure related to this manuscript has been submitted to the University of Michigan Innovation Partnerships office.

Received: January 10, 2023

Revised: May 15, 2023

Accepted: July 26, 2023

Published: August 21, 2023

REFERENCES

1. Boon, T., Cerottini, J.C., Van den Eynde, B., van der Bruggen, P., and Van Pel, A. (1994). Tumor antigens recognized by T lymphocytes. *Annu. Rev. Immunol.* 12, 337–365. <https://doi.org/10.1146/annurev.iy.12.040194.002005>.
2. Latchman, Y., Wood, C.R., Chernova, T., Chaudhary, D., Borde, M., Chernova, I., Iwai, Y., Long, A.J., Brown, J.A., Nunes, R., et al. (2001). PD L2 is a second ligand for PD 1 and inhibits T cell activation. *Nat. Immunol.* 2, 261–268. <https://doi.org/10.1038/85330>.
3. Lopez Nevot, M.A., Garcia, E., Pareja, E., Bonal, F.J., Martin, J., Ruiz Cabello, F., Serrano, S., and Garrido, F. (1986). Differential expression of HLA class I and II antigens in primary and metastatic melanomas. *J. Immunogenet.* 13, 219–227.
4. Hodi, F.S., O'Day, S.J., McDermott, D.F., Weber, R.W., Sosman, J.A., Haanen, J.B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J.C., et al. (2010). Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* 363, 711–723. <https://doi.org/10.1056/NEJMoa1003466>.
5. Wolchok, J.D., Kluger, H., Callahan, M.K., Postow, M.A., Rizvi, N.A., Lesokhin, A.M., Segal, N.H., Ariyan, C.E., Gordon, R. A., Reed, K., et al. (2013). Nivolumab plus ipilimumab in advanced melanoma. *N. Engl. J. Med.* 369, 122–133. <https://doi.org/10.1056/NEJMoa1302369>.
6. Garon, E.B., Rizvi, N.A., Hui, R., Leighl, N., Balmanoukian, A.S., Eder, J.P., Patnaik, A., Aggarwal, C., Gubens, M., Horn, L., et al. (2015). Pembrolizumab for the Treatment of Non Small Cell Lung Cancer. *N. Engl. J. Med.* 372, 2018–2028. <https://doi.org/10.1056/NEJMoa1501824>.
7. Rittmeyer, A., Barlesi, F., Waterkamp, D., Park, K., Ciardiello, F., von Pawel, J., Gadgeel, S.M., Hida, T., Kowalski, D.M., Dols, M.C., et al. (2017). Atezolizumab versus docetaxel in patients with previously treated non small cell lung cancer (OAK): a phase 3, open label, multicentre randomised controlled trial. *Lancet* 389, 255–265. [https://doi.org/10.1016/S0140-6736\(16\)32517-X](https://doi.org/10.1016/S0140-6736(16)32517-X).
8. Davila, M.L., Riviere, I., Wang, X., Bartido, S., Park, J., Curran, K., Chung, S.S., Stefanski, J., Borquez Ojeda, O., Olszewska, M., et al. (2014). Efficacy and toxicity management of 19 28z CART T cell therapy in B cell acute lymphoblastic leukemia. *Sci. Transl. Med.* 6, 224ra25. <https://doi.org/10.1126/scitranslmed.3008226>.
9. Fyfe, G., Fisher, R.I., Rosenberg, S.A., Sznol, M., Parkinson, D.R., and Louie, A.C. (1995). Results of treatment of 255 patients with metastatic renal cell carcinoma who received high dose recombinant interleukin 2 therapy. *J. Clin. Oncol.* 13, 688–696. <https://doi.org/10.1200/JCO.1995.13.3.688>.
10. Haslam, A., and Prasad, V. (2019). Estimation of the Percentage of US Patients With Cancer Who Are Eligible for and Respond to Checkpoint Inhibitor Immunotherapy Drugs. *JAMA Netw. Open* 2, e192535. <https://doi.org/10.1001/jamanetworkopen.2019.2535>.
11. Cabrera, T., Maleno, I., Lopez Nevot, M.A., Redondo, M., Fernandez, M.A., Collado, A., and Garrido, F. (2003). High frequency of HLA B44 allelic losses in human solid tumors. *Hum. Immunol.* 64, 941–950. [https://doi.org/10.1016/s0198-8859\(03\)00164-2](https://doi.org/10.1016/s0198-8859(03)00164-2).
12. Polprasert, C., Takeuchi, Y., Makishima, H., Wudhikarn, K., Kakiuchi, N., Tangnuntachai, N., Assanasen, T., Sitthi, W., Muhamad, H., Lawasut, P., et al. (2021). Frequent mutations in HLA and related genes in extranodal

- NK/T cell lymphomas. *Leuk. Lymphoma* 62, 95–103. <https://doi.org/10.1080/10428194.2020.1821011>.
13. Bicknell, D.C., Kakkamanis, L., Hampson, R., Bodmer, W.F., and Karan, P. (1996). Selection for $\beta 2$ microglobulin mutation in mismatch repair defective colorectal carcinomas. *Curr. Biol.* 6, 1695–1697. [https://doi.org/10.1016/S0960-9822\(02\)70795-1](https://doi.org/10.1016/S0960-9822(02)70795-1).
14. Carretero, F.J., Del Campo, A.B., Flores Martín, J.F., Méndez, R., García López, C., Cozar, J.M., Adams, V., Ward, S., Cabrera, T., Ruiz Cabello, F., et al. (2016). Frequent HLA class I alterations in human prostate cancer: molecular mechanisms and clinical relevance. *Cancer Immunol. Immunother.* 65, 47–59. <https://doi.org/10.1007/s00262-015-1774-5>.
15. McGranahan, N., Rosenthal, R., Hiley, C.T., Rowan, A.J., Watkins, T.B.K., Wilson, G.A., Birkbak, N.J., Veeriah, S., Van Loo, P., Herrero, J., et al. (2017). Allele Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* 171, 1259–1271.e11. <https://doi.org/10.1016/j.cell.2017.10.001>.
16. Mortara, L., Castellani, P., Meazza, R., Tosi, G., De Lerma Barbaro, A., Procopio, F.A., Comes, A., Zardi, L., Ferrini, S., and Accolla, R.S. (2006). CLITA induced MHC class II expression in mammary adenocarcinoma leads to a Th1 polarization of the tumor microenvironment, tumor rejection, and specific antitumor memory. *Clin. Cancer Res.* 12, 3435–3443. <https://doi.org/10.1158/1078-0432.CCR-06-0165>.
17. Johnson, D.B., Estrada, M.V., Salgado, R., Sanchez, V., Doxie, D.B., Opatlenik, S.R., Vilgelm, A.E., Feld, E., Johnson, A.S., Greenplate, A.R., et al. (2016). Melanoma specific MHC II expression represents a tumour autonomous phenotype and predicts response to anti PD 1/PD L1 therapy. *Nat. Commun.* 7, 10582. <https://doi.org/10.1038/ncomms10582>.
18. Forero, A., Li, Y., Chen, D., Grizzle, W.E., Updike, K.L., Merz, N.D., Downs Kelly, E., Burwell, T.C., Vaklavas, C., Buchsbaum, D.J., et al. (2016). Expression of the MHC Class II Pathway in Triple Negative Breast Cancer Tumor Cells Is Associated with a Good Prognosis and Infiltrating Lymphocytes. *Cancer Immunol. Res.* 4, 390–399. <https://doi.org/10.1158/2326-6066.CIR-15-0243>.
19. Haabeth, O.A.W., Tveita, A.A., Fauskanger, M., Schjesvold, F., Lørvik, K.B., Hofgaard, P.O., Omholt, H., Munthe, L.A., Dembic, Z., Corthay, A., and Bogen, B. (2014). How Do CD4(+) T Cells Detect and Eliminate Tumor Cells That Either Lack or Express MHC Class II Molecules? *Front. Immunol.* 5, 174. <https://doi.org/10.3389/fimmu.2014.00174>.
20. Steimle, V., Siegrist, C.A., Mottet, A., Lisowska-Grospierre, B., and Mach, B. (1994). Regulation of MHC class II expression by interferon gamma mediated by the transactivator gene CIITA. *Science* 265, 106–109. <https://doi.org/10.1126/science.8016643>.
21. Dunne, M.R., Phelan, J.J., Michielsen, A.J., Maguire, A.A., Dunne, C., Martin, P., Noonan, S., Tosetto, M., Geraghty, R., Fennelly, D., et al. (2020). Characterising the Prognostic Potential of HLA DR during Colorectal Cancer Development. *Cancer Immunol. Immunother.* 69, 1577–1588.
22. He, Y., Rozeboom, L., Rivard, C.J., Ellison, K., Dziadziuszko, R., Yu, H., Zhou, C., and Hirsch, F.R. (2017). MHC class II expression in lung cancer. *Lung Cancer* 112, 75–80. <https://doi.org/10.1016/j.lungcan.2017.07.030>.
23. Park, I.A., Hwang, S. H., Song, I.H., Heo, S. H., Kim, Y. A., Bang, W.S., Park, H.S., Lee, M., Gong, G., and Lee, H.J. (2017). Expression of the MHC class II in triple negative breast cancer is associated with tumor infiltrating lymphocytes and interferon signaling. *PLoS One* 12, e0182786. <https://doi.org/10.1371/journal.pone.0182786>.
24. Accolla, R.S., Ramia, E., Tedeschi, A., and Forlani, G. (2019). CLITA Driven MHC Class II Expressing Tumor Cells as Antigen Presenting Cell Performers: Toward the Construction of an Optimal Anti tumor Vaccine. *Front. Immunol.* 10, 1806. <https://doi.org/10.3389/fimmu.2019.01806>.
25. Buhler, S., and Sanchez Mazas, A. (2011). HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. *PLoS One* 6, e14643. <https://doi.org/10.1371/journal.pone.0014643>.
26. Li, H. (2013). Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA MEM. Preprint at arXiv. arXiv:1303.3997..
27. Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. Preprint at bioRxiv. <https://doi.org/10.1101/861054>.
28. Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., and Kohlbacher, O. (2014). OptiType: precision HLA typing from next generation sequencing data. *Bioinformatics* 30, 3310–3316. <https://doi.org/10.1093/bioinformatics/btu548>.
29. Dilthey, A.T., Gourraud, P. A., Mentzer, A.J., Cereb, N., Iqbal, Z., and McVean, G. (2016). High Accuracy HLA Type Inference from Whole Genome Sequencing Data Using Population Reference Graphs. *PLoS Comput. Biol.* 12, e1005151. <https://doi.org/10.1371/journal.pcbi.1005151>.
30. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R., and Matsuda, F. (2017). HLA HD: An accurate HLA typing algorithm for next generation sequencing data. *Hum. Mutat.* 38, 788–797. <https://doi.org/10.1002/humu.23230>.
31. Ka, S., Lee, S., Hong, J., Cho, Y., Sung, J., Kim, H. N., Kim, H. L., and Jung, J. (2017). HLAScan: genotyping of the HLA region using next generation sequencing data. *BMC Bioinf.* 18, 258. <https://doi.org/10.1186/s12859-017-1671-3>.
32. Lee, H., and Kingsford, C. (2018). Accurate Assembly and Typing of HLA using a Graph Guided Assembler Kourami. *Methods Mol. Biol.* 1802, 235–247. https://doi.org/10.1007/978-1-4939-8546-3_17.
33. Sverchkova, A., Anzar, I., Stratford, R., and Clancy, T. (2019). Improved HLA typing of Class I and Class II alleles from next generation sequencing data. *Hladnikia* 94, 504–513. <https://doi.org/10.1111/tan.13685>.
34. Shukla, S.A., Rooney, M.S., Rajasagi, M., Tiao, G., Dixon, P.M., Lawrence, M.S., Stevens, J., Lane, W.J., Dellagatta, J.L., Steelman, S., et al. (2015). Comprehensive analysis of cancer associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* 33, 1152–1158. <https://doi.org/10.1038/nbt.3344>.
35. Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P., and Marsh, S.G.E. (2015). The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* 43, D423–D431. <https://doi.org/10.1093/nar/gku1161>.
36. Gourraud, P. A., Khankhanian, P., Cereb, N., Yang, S.Y., Feolo, M., Maier, M., Rioux, J.D., Hauser, S., and Oksenberg, J. (2014). HLA diversity in the 1000 genomes dataset. *PLoS One* 9, e97282. <https://doi.org/10.1371/journal.pone.0097282>.
37. Shen, H., Li, J., Zhang, J., Xu, C., Jiang, Y., Wu, Z., Zhao, F., Liao, L., Chen, J., Lin, Y., et al. (2013). Comprehensive characterization of human genome variation by high coverage whole genome sequencing of forty four Caucasians. *PLoS One* 8, e59494. <https://doi.org/10.1371/journal.pone.0059494>.
38. Dhatchinamoorthy, K., Colbert, J.D., and Rock, K.L. (2021). Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation. *Front. Immunol.* 12, 636568. <https://doi.org/10.3389/fimmu.2021.636568>.
39. Robinson, D.R., Wu, Y. M., Lonigro, R.J., Vats, P., Cobain, E., Everett, J., Cao, X., Rabban, E., Kumar Sinha, C., Raymond, V., et al. (2017). Integrative clinical genomics of metastatic cancer. *Nature* 548, 297–303. <https://doi.org/10.1038/nature23306>.
40. Castro, A., Ozturk, K., Pyke, R.M., Xian, S., Zanetti, M., and Carter, H. (2019). Elevated neoantigen levels in tumors with somatic mutations in the HLA A, HLA B, HLA C and B2M genes. *BMC Med. Genomics* 12, 107. <https://doi.org/10.1186/s12920-019-0544-1>.
41. Howitt, B.E., Shukla, S.A., Sholl, L.M., Ritterhouse, L.L., Watkins, J.C., Rodig, S., Stover, E., Strickland, K.C., D'Andrea, A.D., Wu, C.J., et al. (2015). Association of polymerase ϵ mutated and microsatellite unstable endometrial cancers with neoantigen load, number of tumor infiltrating lymphocytes, and expression of PD 1 and PD L1. *JAMA Oncol.* 1, 1319–1323.
42. Kloor, M., Becker, C., Benner, A., Woerner, S.M., Gebert, J., Ferrone, S., and von Knebel Doeberitz, M. (2005). Immunoselective pressure and human leukocyte antigen class I antigen machinery defects in microsatellite

- unstable colorectal cancers. *Cancer Res.* 65, 6418–6424. <https://doi.org/10.1158/0008.5472.CAN.05.0044>.
43. Ozcan, M., Janikovits, J., von Knebel Doeberitz, M., and Kloor, M. (2018). Complex pattern of immune evasion in MSI colorectal cancer. *Oncology* 7, e1445453. <https://doi.org/10.1080/2162402X.2018.1445453>.
44. Giraldo, N.A., Becht, E., Remark, R., Damotte, D., Sautès-Fridman, C., and Fridman, W.H. (2014). The immune contexture of primary and metastatic human tumours. *Curr. Opin. Immunol.* 27, 8–15. <https://doi.org/10.1016/j.coi.2014.01.001>.
45. Szekely, B., Bossuyt, V., Li, X., Wali, V.B., Patwardhan, G.A., Frederick, C., Silber, A., Park, T., Harigopal, M., Pelekanou, V., et al. (2018). Immunological differences between primary and metastatic breast cancer. *Ann. Oncol.* 29, 2232–2239. <https://doi.org/10.1093/annonc/mdy399>.
46. Remark, R., Alifano, M., Cremer, I., Lupo, A., Dieu-Nosjean, M. C., Riquet, M., Crozet, L., Ouakrim, H., Goc, J., Cazes, A., et al. (2013). Characteristics and clinical impacts of the immune environments in colorectal and renal cell carcinoma lung metastases: influence of tumor origin. *Clin. Cancer Res.* 19, 4079–4091. <https://doi.org/10.1158/1078.0432.CCR.12.3847>.
47. Halama, N., Michel, S., Kloor, M., Zoernig, I., Benner, A., Spille, A., Pommerenke, T., von Knebel, D.M., Folprecht, G., Lubber, B., et al. (2011). Localization and density of immune cells in the invasive margin of human colorectal cancer liver metastases are prognostic for response to chemotherapy. *Cancer Res.* 71, 5670–5677. <https://doi.org/10.1158/0008.5472.CAN.11.0268>.
48. Weghorn, D., and Sunyaev, S. (2017). Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.* 49, 1785–1788. <https://doi.org/10.1038/ng.3987>.
49. Chowell, D., Krishna, C., Pierini, F., Makarov, V., Rizvi, N.A., Kuo, F., Morris, L.G.T., Riaz, N., Lenz, T.L., and Chan, T.A. (2019). Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Nat. Med.* 25, 1715–1720. <https://doi.org/10.1038/s41591.019.0639.4>.
50. Leiserson, M.D.M., Vandin, F., Wu, H. T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2015). Pan cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. <https://doi.org/10.1038/ng.3168>.
51. Iranzo, J., Martincorena, I., and Koonin, E.V. (2018). Cancer mutation network and the number and specificity of driver mutations. *Proc. Natl. Acad. Sci. USA* 115, E6010–E6019. <https://doi.org/10.1073/pnas.1803155115>.
52. Leone, P., Shin, E. C., Perosa, F., Vacca, A., Dammacco, F., and Raccanelli, V. (2013). MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells. *J. Natl. Cancer Inst.* 105, 1172–1187. <https://doi.org/10.1093/jnci/djt184>.
53. Challa Malladi, M., Lieu, Y.K., Califano, O., Holmes, A.B., Bhagat, G., Murty, V.V., Dominguez Sola, D., Pasqualucci, L., and Dalla Favera, R. (2011). Combined genetic inactivation of β 2 Microglobulin and CD58 reveals frequent escape from immune recognition in diffuse large B cell lymphoma. *Cancer Cell* 20, 728–740. <https://doi.org/10.1016/j.ccr.2011.11.006>.
54. Muzio, M., Chinnaiyan, A.M., Kischkel, F.C., O'Rourke, K., Shevchenko, A., Ni, J., Scaffidi, C., Bretz, J.D., Zhang, M., Gentz, R., et al. (1996). 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. [https://doi.org/10.1016/S0092.8674\(00\)81266.0](https://doi.org/10.1016/S0092.8674(00)81266.0).
55. Lyu, H., Li, M., Jiang, Z., Liu, Z., and Wang, X. (2019). Correlate the TP53 Mutation and the HRAS Mutation with Immune Signatures in Head and Neck Squamous Cell Cancer. *Comput. Struct. Biotechnol. J.* 17, 1020–1030. <https://doi.org/10.1016/j.csbj.2019.07.009>.
56. Ionov, Y., Peinado, M.A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* 363, 558–561. <https://doi.org/10.1038/363558a0>.
57. Lang, G.I., Parsons, L., and Gammie, A.E. (2013). Mutation rates, spectra, and genome wide distribution of spontaneous mutations in mismatch repair deficient yeast. *G3* 3, 1453–1465. <https://doi.org/10.1534/g3.113.006429>.
58. Okazaki, I. M., Kotani, A., and Honjo, T. (2007). Role of AID in tumorigenesis. *Adv. Immunol.* 94, 245–273. [https://doi.org/10.1016/S0065.2776\(06\)94008.5](https://doi.org/10.1016/S0065.2776(06)94008.5).
59. Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S.A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G.V., Carter, S.L., Sakseena, G., et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is wide spread in human cancers. *Nat. Genet.* 45, 970–976. <https://doi.org/10.1038/ng.2702>.
60. Chan, K., Roberts, S.A., Klimczak, L.J., Sterling, J.F., Saini, N., Malc, E.P., Kim, J., Kwiatkowski, D.J., Fargo, D.C., Mieczkowski, P.A., et al. (2015). An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* 47, 1067–1072. <https://doi.org/10.1038/ng.3378>.
61. Parolia, A., Cieslik, M., Chu, S. C., Xiao, L., Ouchi, T., Zhang, Y., Wang, X., Vats, P., Cao, X., Pitchiaya, S., et al. (2019). Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. *Nature* 571, 413–418. <https://doi.org/10.1038/s41586.019.1347.4>.
62. Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., Golub, T.R., Lander, E.S., and Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. USA* 112, E5486–E5495. <https://doi.org/10.1073/pnas.1516373112>.
63. Gao, G.F., Tormo, J., Gerth, U.C., Wyer, J.R., McMichael, A.J., Stuart, D.I., Bell, J.I., Jones, E.Y., and Jakobsen, B.K. (1997). Crystal structure of the complex between human CD8 α (α) and HLA A2. *Nature* 387, 630–634. <https://doi.org/10.1038/42523>.
64. Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202–209. <https://doi.org/10.1038/nature13480>.
65. Huang, X., Zheng, W., Pearce, R., and Zhang, Y. (2020). SSIPe: accurately estimating protein binding affinity change upon mutations using evolutionary profiles in combination with an optimized physical energy function. *Bioinformatics* 36, 2429–2437. <https://doi.org/10.1093/bioinformatics/btz926>.
66. Engin, H.B., Kreisberg, J.F., and Carter, H. (2016). Structure Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLoS One* 11, e0152929. <https://doi.org/10.1371/journal.pone.0152929>.
67. Gentles, A.J., Newman, A.M., Liu, C.L., Bratman, S.V., Feng, W., Kim, D., Nair, V.S., Xu, Y., Khuong, A., Hoang, C.D., et al. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* 21, 938–945. <https://doi.org/10.1038/nm.3909>.
68. Landi, L., D'Incà, F., Gelibter, A., Chiari, R., Grossi, F., Delmonte, A., Pasaro, A., Signorelli, D., Gelsomino, F., Galetta, D., et al. (2019). Bone metastases and immunotherapy in patients with advanced non small cell lung cancer. *J. Immunother. Cancer* 7, 316. <https://doi.org/10.1186/s40425.019.0793.8>.
69. Yu, J., Green, M.D., Li, S., Sun, Y., Journey, S.N., Choi, J.E., Rizvi, S.M., Qin, A., Waninger, J.J., Lang, X., et al. (2021). Liver metastasis restrains immunotherapy efficacy via macrophage mediated T cell elimination. *Nat. Med.* 27, 152–164. <https://doi.org/10.1038/s41591.020.1131.x>.
70. López Nevot, M.A., Esteban, F., Ferrón, A., Gutiérrez, J., Oliva, M.R., Romero, C., Huelin, C., Ruiz Cabello, F., and Garrido, F. (1989). 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. <https://doi.org/10.1038/bjc.1989.45>.
71. Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., et al. (2017). An immunogenic



- personal neoantigen vaccine for patients with melanoma. *Nature* 547, 217–221. <https://doi.org/10.1038/nature22991>.
72. Bushnell, B. (2014). BBMap: A Fast, Accurate, Splice Aware Aligner (Lawrence Berkeley National Lab. (LBNL)).
 73. Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012). Primer3 – new capabilities and interfaces. *Nucleic Acids Res.* 40, e115. <https://doi.org/10.1093/nar/gks596>.
 74. Leiserson, M.D.M., Wu, H. T., Vandin, F., and Raphael, B.J. (2015). CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* 16, 160. <https://doi.org/10.1186/s13059-015-0700-7>.
 75. Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L.G., and Lee, S. (2020). UK Biobank Whole Exome Sequence Binary Phenome Analysis with Robust Region Based Rare Variant Test. *Am. J. Hum. Genet.* 106, 3–12. <https://doi.org/10.1016/j.ajhg.2019.11.012>.
 76. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
 77. Poplin, R., Ruano Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy Moonshine, A., Roazen, D., et al. (2018). Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples. Preprint at bioRxiv. <https://doi.org/10.1101/201178>.
 78. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* 18, 696–705. <https://doi.org/10.1038/s41568-018-0060-1>.



STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA WES DNA	TCGA	https://portal.gdc.cancer.gov
Broad GDAC Firehose	Broad Institute	https://gdac.broadinstitute.org
ICGC TCGA SBS signatures	ICGC	https://www.synapse.org , ID: syn11801497
1000 genomes HLA types	Sverchkova et al. ³⁴	https://doi.org/10.1111/tan.13685
Polysolver TCGA HLA variants	Shukla et al., ³⁵ Castro et al. ⁴¹	https://doi.org/10.1038/nbt.3344 , https://doi.org/10.1186/s12920-019-0544-1
Hapster HLA variants	This paper	Table S7
HLA sequence database	IMGT	https://github.com/ANHIG/IMGTHLA
COSMIC Tier 1 cancer gene Census	Sondka et al. ⁷²	https://cancer.sanger.ac.uk/census
HLA crystal structures	RCSB PDB	PDB IDs: 4NQV, 6IEX, 3MGO, 3KPL, 3RL2, 3DX7, 4F7M, 1E28, 4HX1, 2RFX, 5EO0, 5IM7, 1XR8, 5W6A, 1JGE, 5VGE, 3LKR, 6JTP, 5WKF, 4G8G, 5WKH, 4G9F, 5NQK, 5XOT, 6EQA, 4PRP, 3VXM, 6BJ2, 3W0W, 3MV7, 6AVF, 4JRX, 6AVG, 4JRY, 4QRP, 4PRI, 1MI5, 3DXA, 3FFC, 3KPR, 3SJV, 3KPS, 4QRP, 5OPI, 5WER, 6ENY, 4U6X, 3OXR
Software and algorithms		
Hapster	This paper	https://github.com/mctp/hapster
BWA mem	Heng Li ²⁷	https://github.com/lh3/bwa
GATK4	Broad Institute	https://github.com/broadinstitute/gatk
Primer3	Untergasser et al. ⁷³	https://github.com/primer3-org/primer3
CBaSE	Weghorn et al. ⁴⁹	http://genetics.bwh.harvard.edu/cbase/
CoMEt	Leiserson et al. ⁷⁴	https://github.com/raphael-group/comet
SKAT	Zhao et al. ⁷⁵	https://github.com/leelabsg/SKAT
SSIPE	Huang et al. ⁶⁶	https://zhanggroup.org/SSIPE/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Marcin Cieslik (mcieslik@med.umich.edu)

Materials availability

This study did not produce new unique materials or reagents.

Data and code availability

- All HLA mutation calls made by Hapster in TCGA and MI-ONCOSEQ samples are reported in this manuscript (Table S5). All raw data will be shared by the lead contact upon request.
- All source code for Hapster is available at <https://github.com/mctp/hapster>.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

METHOD DETAILS

Hapster algorithm

Algorithm overview

Hapster (See Figure S6 for overview) models alignment of sequencing reads to alternative haplotypes as a linear system that can be solved to identify the closest underlying haplotypes, a method which generalizes to any polymorphic genes. Hapster also uses 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-MEM²⁶ and Mutect2²⁷) 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 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 $\mathbf{Ax} = \mathbf{b}$ 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.

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 database³⁵ 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

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.

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 $\mathbf{Ax} = \mathbf{b}$. 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.

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 are 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 BMap'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.

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, 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 and 0.99 were tested to find the optimal threshold per gene that minimizes observed germline variants.

Mutation calling and filtering

For each gene, the reads extracted from the normal sample are re-aligned with BWA-mem²⁶ to the new genomic reference file dynamically constructed to only contain the genomic sequence of each allele in the inferred HLA haplotype. In this new genomic reference file, each HLA allele is present as an individual contig. Germline variants are first called using the GATK4 function HaplotypeCaller. Any variant with a VAF >0.9 is assumed to be a private germline variant and is imputed into the genomic reference file to create a finalized set of personal reference sequences. Extracted reads from both the tumor and normal are next re-aligned to the germline-imputed reference using BWA-mem. Mutations are called in the re-aligned BAM using the GATK4 function Mutect2,²⁷ and variants are filtered using the function FilterMutectCalls. The GATK4 filter FilterByOrientationBias was applied to remove 8-Oxog artifacts that were observed within TCGA cohorts. We additionally removed any low coverage variants that had 100% orientation bias even if there were not enough reads to fail the bias test. We next identify duplicates that arise for mutations called within homologous regions between two alleles of the same gene where reads randomly segregate between the two alleles. For each duplicate pair, the mutation is randomly assigned to one of the two alleles. We next constructed a custom kmer filter that searches for the presence of any reads supporting a called variant in an alignment free manner, removing any variants that find kmer support in the normal. 25-mers covering the variant were used, excluding those kmers where the variant is within 8 bases from the edge of the kmer. As a final filter, we removed any somatic variants that matched a known polymorphism for an alternative allele of the same gene within the HLA database. Known polymorphisms were identified by looking at a variant position within the global multiple sequence alignment of all alleles of a given gene. Only single nucleotide variants were identified this way.

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.³³ For each sample and each method, an HLA reference was constructed by taking the genomic sequence in the IMGT/HLA database (<https://www.ebi.ac.uk/ipd/imgt/hla>) 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,²⁶ and germline variants were called using GATK's HaplotypeCaller.⁷⁷

Alignment and mutation calling

HLA extracted reads were aligned to Hapster inferred reference sequences using BWA-mem.²⁶ 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 Mutect2.²⁷ 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.

Simulation validation

200 synthetic haplotypes were constructed by taking two random allelic sequences from the IMGT/HLA database (<https://www.ebi.ac.uk/ipd/imgt/hla>) for each MHC-I and class II 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 BBmap⁷² 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.

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.

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, resulting in a set of 75 variants. MHC-II variants were not selected for RNA validation as the expression of MHC-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-adj.) given our distribution.

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, Primer3⁷³ 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.

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.

Positive selection of somatic mutations

Positive selection was evaluated by CBaSE,⁴⁸ 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 metastatic ϕ_{pos} , the cohort-level ϕ_{pos} reported by CBaSE was summed across all cohorts.



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:

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

Mutations were considered clonal if $CCF > 0.7$.

Mutual exclusivity and co-mutation analyses

Both mutual exclusivity and co-mutation analyses were performed at the cohort level. Mutual exclusivity of MHC-I and APM mutations were evaluated using CoMet.⁷⁴ Exact tests for mutual exclusivity were calculated by comparing any mutation with the MHC-I genes to any mutation within an APM gene. Co-mutation analyses were performed using an SNP-seq kernel association test as implemented in the R package SKAT.⁷⁵ Only driver genes listed in the COSMIC Tier 1 Cancer Gene Census⁷⁸ were considered. Within each cohort, only genes mutated in more than 2% of patients were analyzed.

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.

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.

MHC amino acid annotations

Protein interaction analysis was performed using annotations derived from structures of the relevant MHC-I complexes (PDB IDs provided below). The distance cutoff of the contacts was $\leq 4 \text{ \AA}$. Structures used for MHC-I: peptide interaction and the interaction between MHC-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-I: CD8 interaction was 3DMM. Structures used for MHC-I: TCR interaction were 5WK6, 4G8G, 5WKH, 4G9F, 5NQG, 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-I: TAPBP interaction were 5OPI and 5WER. Structure used for MHC-I: TAPBP interaction was 6ENY.

PPI binding energy predictions

Predicted changes to the MHC-B2M interface binding energy were generated using SSIPe.⁶⁵ All predictions were performed using the 3D crystal structure at <https://www.rcsb.org/structure/4U6X>. While the MHC-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.

QUANTIFICATION AND STATISTICAL ANALYSIS

Total sample sizes for each cohort are provided (Table S4). 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 an FDR of 5%, unless otherwise noted.