11-2015

Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes

Sachet A. Shukla
Iowa State University

Michael S. Rooney
Broad Institute of MIT and Harvard

Mohini Rajasagi
Dana-Farber Cancer Institute

Grace Tiao
Broad Institute of MIT and Harvard

Philip M. Dixon
Iowa State University, pdixon@iastate.edu

See next page for additional authors

Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes

Abstract

Detection of somatic mutations in human leukocyte antigen (HLA) genes using whole-exome sequencing (WES) is hampered by the high polymorphism of the HLA loci, which prevents alignment of sequencing reads to the human reference genome. We describe a computational pipeline that enables accurate inference of germline alleles of class I *HLA-A, B* and *C* genes and subsequent detection of mutations in these genes using the inferred alleles as a reference. Analysis of WES data from 7,930 pairs of tumor and healthy tissue from the same patient revealed 298 nonsilent HLA mutations in tumors from 266 patients. These 298 mutations are enriched for likely functional mutations, including putative loss-of-function events. Recurrence of mutations suggested that these 'hotspot' sites were positively selected. Cancers with recurrent somatic HLA mutations were associated with upregulation of signatures of cytolytic activity characteristic of tumor infiltration by effector lymphocytes, supporting immune evasion by altered HLA function as a contributory mechanism in cancer.

Disciplines

Biostatistics | Cancer Biology | Computational Biology | Genetics and Genomics

Comments

This is a manucript of an article published as Shukla, Sachet A., Michael S. Rooney, Mohini Rajasagi, Grace Tiao, Philip M. Dixon, Michael S. Lawrence, Jonathan Stevens et al. "Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes." *Nature biotechnology* 33, no. 11 (2015): 1152. doi: 10.1038/nbt.3344. Posted with permission.

Authors

Sachet A. Shukla, Michael S. Rooney, Mohini Rajasagi, Grace Tiao, Philip M. Dixon, Michael S. Lawrence, Jonathan Stevens, William J. Lane, Jamie L. Dellagatta, Scott Steelman, Carrie Sougnez, Kristian Cibulskis, Adam Kiezun, Vladimir Brusic, Catherine J. Wu, and Gad Getz

HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2016 May 01.

Published in final edited form as:

Nat Biotechnol. 2015 November ; 33(11): 1152–1158. doi:10.1038/nbt.3344.

Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes

Sachet A. Shukla[1,3,6], **Michael S. Rooney**[3,7], **Mohini Rajasagi**[1,2], **Grace Tiao**[3], **Philip M. Dixon**[6], **Michael S. Lawrence**[3], **Jonathan Stevens**[8], **William J. Lane**[8,9], **Jamie L. Dellagatta**[8], **Scott Steelman**[3], **Carrie Sougnez**[3], **Kristian Cibulskis**[3], **Adam Kiezun**[3], **Vladimir Brusic**[1,2], **Catherine J. Wu**[1,2,5,*], and **Gad Getz**[3,4,*]

[1]Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston, MA, USA

[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

[3]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[4]Massachusetts General Hospital Cancer Center and Department of Pathology, Boston, MA, USA

[5]Department of Medicine, Harvard Medical School, Boston, MA, USA

[6]Department of Statistics, Iowa State University, Ames, IA, USA

[7]Harvard/MIT Division of Health Sciences and Technology, Cambridge, MA, USA

[8]Department of Pathology, Brigham and Women's Hospital, Boston, MA

[9]Harvard Medical School, Boston, MA, USA

Abstract

Detection of somatic mutations in HLA genes using whole-exome sequencing (WES) is hampered by the high polymorphism of the HLA loci, which prevents alignment of sequencing reads to the human reference genome. We describe a computational pipeline that enables accurate inference of germline alleles of class I *HLA-A*, *-B* and *-C* genes and subsequent detection of mutations in these genes using the inferred alleles as a reference. Analysis of WES data from 7,930 pairs of tumor and healthy tissue from the same patient revealed 298 non-silent HLA mutations in tumors from 266 patients. These 298 mutations are enriched for likely functional mutations, including putative loss-of-function events. Recurrence of mutations suggested that these 'hotspot' sites were

Correspondence should be addressed to: **Catherine J. Wu, MD**, Dana-Farber Cancer Institute, 450 Brookline Ave, Rm 540B, Boston, MA- 02215, cwu@partners.org; **Gad Getz, PhD**, Broad Institute and Massachusetts General Hospital, 75 Ames Street, 4007-A, Office: (617) 714-7471, gadgetz@broadinstitute.org.

*Denotes equal contribution

AUTHOR CONTRIBUTIONS

C.J.W. proposed the initial idea of using exome data for HLA typing. S.A.S., G.G., C.J.W., P.M.D., K.C. conceived and designed POLYSOLVER and the mutation detection pipeline. G.T. and A.K. developed the ethnicity inference module. M.S.L. and S.A.S. performed the mutation significance analysis. V.B., C.J.W. and S.A.S. mapped the contact residue mutations. M.S.R. performed the gene expression analysis. J.S., W.J.L., S.S. and J.L.D. performed the experimental validation. C.S. helped with data access and management. C.J.W., S.A.S., G.G. and M.R. wrote the manuscript. C.J.W. and G.G. led the project.

COMPETING FINANCIAL INTERESTS

Work of this paper is included in patent applications.

positively selected. Cancers with recurrent somatic HLA mutations were associated with upregulation of signatures of cytolytic activity characteristic of tumor infiltration by effector lymphocytes, supporting immune evasion by altered HLA function as a contributory mechanism in cancer.

Recent large-scale WES studies have revealed the existence and relative high frequency of somatic changes in HLA class I genes in head and neck cancer, squamous cell lung cancer, stomach adenocarcinoma and diffuse large B cell lymphoma[1–5]. The HLA locus, located on chromosome 6, is among the most polymorphic regions of the human genome, with thousands of documented alleles for each gene[6]. These class I alleles are critical mediators of the cytotoxic T cell response, presenting cellular peptides on the cell surface in a form that can be recognized by the T cell receptor[7, 8]. The finding of enhanced somatic mutation rate in HLA genes has strongly implicated HLA dysfunction as a possible mechanism of immune evasion in the development and progression of certain cancers[1–5].

Each individual expresses six major MHC class I alleles, encoded by three genes (*HLA-A*, *HLA-B* and *HLA-C*) located on the two homologous copies of chromosome 6. Conventional determination of HLA type is performed using serology- and/or PCR-based methods that are labor-intensive and time-consuming[9–11]. Several protocols have recently been proposed for HLA-targeted multiplexed PCR coupled with next generation sequencing (NGS), but by design, provide information restricted to HLA alleles, and not the rest of genome[12–16]. Theoretically HLA typing information should be directly extractable from WES data, an increasingly available and cost-effective approach for the comprehensive analysis of genome-wide somatic alterations. The human reference genome, however, has a single sequence for each HLA gene and would likely misrepresent the true alleles in the individual, thereby causing suboptimal alignments. In addition, the HLA genes are GC-rich and therefore typically suffer from lower sequencing coverage due to lower efficiency in capture and amplification and increased sequencing errors that further reduce the alignment rates. Consequently, in order to accurately detect somatic mutations in the HLA genes, one needs first to accurately align all reads originating from this region in both the tumor and matched normal samples and only then apply somatic mutation detection tools. We also surmised that conventional alignment and mutation detection methods, which do not carefully treat this highly polymorphic region, would be prone to errors.

To this end, we developed the algorithm POLYSOLVER (POLYmorphic loci reSOLVER), which enables high precision HLA-typing even while using relatively low coverage WES data, and a subsequent mutation detection pipeline that uses the inferred alleles as a basis for high fidelity detection of mutations in HLA genes. By analyzing WES data from 7,930 cancer patients we demonstrate high sensitivity and specificity of our method in detecting HLA somatic mutations. Further characterization suggests a functional impact of these mutations on this biologically important and complex locus. POLYSOLVER is freely available for noncommercial use at http://www.broadinstitute.org/cancer/cga/polysolver (Supplementary Data).

RESULTS

Inference of class I HLA alleles using POLYSOLVER

In order to develop POLYSOLVER, we collected a training set of 8 chronic lymphocytic leukemia (CLL) patients for which WES data as well as conventional PCR-based HLA typing were available[17] (Supplementary Table 1). We first confirmed the expected poor coverage and inverse correlation between GC content and coverage in HLA genes in this set (Supplementary Fig. 1). We reasoned that coverage at these highly polymorphic regions can be substantially improved by ensuring retrieval of true HLA reads that failed to align to the canonical reference, followed by alignment to a library of all known HLA alleles. These alignments could then be used for subsequent computational inference of the individual's HLA type. Thus, POLYSOLVER consists of the following steps: (i) improved retrieval and alignment of HLA reads; (ii) inference of the HLA alleles using a two-step Bayesian classification approach (Fig. 1a and Supplementary Methods). In brief, we increased the precision of the alignment by first selecting reads from the WES data that potentially originated from the HLA region (Supplementary Fig. 2) and aligning them to a full-length genomic library of all known HLA alleles based on the IMGT database[18] (Online Methods) using a precise alignment method (Novoalign, www.novocraft.com), keeping all best-scoring alignments for each read for use in subsequent steps. Inference of the two alleles for each HLA gene was based on a Bayesian calculation that takes into account the base qualities of aligned reads, observed insert sizes, as well as the ethnicity-dependent prior probabilities of each allele[12, 19] (Supplementary Table 2).

For validation, we applied POLYSOLVER to WES data from an independent set of 253 HapMap samples with known HLA genotypes (Supplementary Tables 3 and 4). We observed that POLYSOLVER achieved an overall mean sensitivity of 97% (83% samples had all allele species correctly identified), overall mean precision of 98.8% (93.6% samples had no incorrectly identified allele species), mean overall accuracy of 97% (83% samples had all alleles correctly called), and 100% homozygosity success rate (83 of 83 homozygous cases correctly identified) in HLA typing at the protein coding level. Compared to other recently reported algorithms for inference of HLA type directly from whole-exome sequencing data, POLYSOLVER outperformed 4 of 5 other tools, and was of comparable performance to the recently described OptiType tool[20] (Fig. 1b, Supplementary Table 4 and Supplementary Methods). To accommodate future use of POLYSOLVER for samples of unknown ethnic origin, we developed a principal components (PC)-based method for exome-based ethnicity inference (Online methods and Supplementary Fig. 3), which can be used prior to analysis by POLYSOLVER to ensure maximal typing accuracy.

Detection of somatic mutations within the HLA region

A standard approach for detection of somatic mutations is to first align both tumor and normal reads to the reference genome and then scan the genome and identify mutational events observed in the tumor but not in the matched normal (e.g. as implemented in MuTect[21]). We reasoned that the accurate detection of individual native HLA type using germline data by POLYSOLVER could substantially improve alignment of reads (in both tumor and normal) and hence improve the sensitivity and specificity of somatic mutation

calling within the HLA region (Fig. 2a). In this setting, the inferred allele species for each HLA gene would serve as patient-specific reference 'chromosomes' against which pre-selected HLA reads from the tumor and germline samples are aligned separately followed by standard mutation calling. We therefore built an analysis pipeline to call somatic mutations in the HLA genes that includes the following steps: (i) ethnicity detection using the normal sample; (ii) inference of HLA type by applying POLYSOLVER on the normal sample (although other highly accurate HLA typing tools could also be used); (iii) re-alignment of the HLA reads in both tumor and normal to the inferred HLA alleles while filtering out likely erroneous alignments (Online methods); (iv) application of standard tools to detect somatic mutations (MuTect[21] and Strelka[22]) by comparing the re-aligned tumor and normal HLA reads.

To test this approach, we initially assembled a dataset of 2,545 cases of matched tumor and germline DNA spanning 12 tumor types – 10 from The Cancer Genome Atlas project (TCGA), and two separate genomic studies focusing on chronic lymphocytic leukemia and melanoma. 59 HLA gene somatic mutations were previously detected using standard methods (Supplementary Methods) and reported as part of a pan-cancer analysis effort[23] (Online Methods)[17, 24]. On re-analysis of these cases with our POLYSOLVER-based mutation detection pipeline, we detected 36 of 59 (61%) previously reported HLA mutations, as well as 37 novel somatic HLA mutations; in total, 73 mutations in 64 of 2,545 cases (Fig. 2b, Fig. 2c and Supplementary Tables 5–7). Manual review of all HLA mutation events using IGV[25] suggested that 9 of 23 mutations identified exclusively by TCGA were true events, of which 6 were just below the detection limit of our pipeline and were identified once we slightly relaxed the read filtering criteria used prior to mutation calling (Supplementary Table 8 and Supplementary Methods).

When available, we examined matched RNA-Sequencing data and sought orthogonal evidence of expression of the somatically mutated HLA allele that was detected by WES (indel calls were excluded from this analysis due to low reliability of indel alignment and detection by RNA-Seq[26]). A mutation was considered validated if there were at least two alternate allele bearing reads in the RNA-Seq data for well-powered sites (Online methods). In total, we could evaluate RNA-Seq data for 49 of 96 mutations, including 10 that were exclusively reported by TCGA, 17 detected only by our pipeline and 22 that were detected by both. We observed a high rate of RNA-Seq based validation of missense, nonsense and splice-site mutations in the set of 22 mutations found in common (8 of 8; 8 of 11; and 2 of 3 events, respectively; Fig. 2d and Supplementary Table 9). We likewise observed high rates of validation for events identified exclusively by the POLYSOLVER-based mutation detection pipeline (7 of 9; 5 of 6; and 2 of 2 events respectively). By contrast, only 2 of 10 mutations uniquely identified by TCGA were validated using RNA-Seq.

We further performed experimental validation of inferred mutation calls through direct targeted sequencing of *HLA-A* and *HLA-B* alleles of 18 TCGA samples identified as bearing HLA mutations for which DNA material was available (Online Methods)[27]. 6 of these 18 samples did not have adequate coverage at the site of mutation and were removed from the analysis due to power considerations (Online Methods). Of the remaining 12 mutations, this analysis confirmed all 11 of 11 HLA mutations that were inferred by the POLYSOLVER-

based mutation detection pipeline (5 identified by TCGA also; 6 identified exclusively by POLYSOLVER), while the sole mutation identified exclusively by TCGA did not validate (Fig. 2d and Supplementary Table 10). Altogether, these results demonstrate that the POLYSOLVER-based approach is both a sensitive and specific somatic mutation detection strategy within the highly polymorphic HLA loci.

Patterns of somatic HLA mutation across tumor types

We extended our analysis of POLYSOLVER-based mutation detection to a total of 7,930 TCGA tumor/normal pairs (including the original collection of 2,545 and 5,385 additional cases). In total, we detected 298 somatic HLA mutations in 266 of 7,930 (3.3%) individuals (Supplementary Tables 11 and 12). The median allele fraction across somatic changes was 33% (interquartile range: 16 – 58%) suggesting that most of these mutations are heterozygous (Supplementary Fig. 4a).

Amongst the cancer types, we observed differences in frequency, localization and types of somatic HLA mutations (Fig. 3). In addition to finding HLA mutations occurring significantly in head and neck (*HLA-A*, *HLA-B*), lung squamous (*HLA-A*) and stomach (*HLA-B*) cancer as previously reported, we now further identified *HLA-A* (FDR $q=2.3\times10^{-8}$) and *HLA-B* (FDR $q=3.9\times10^{-7}$) to be significantly mutated in colon adenocarcinoma. By contrast, chronic lymphocytic leukemia (n=128) and liver cancer (n=202) entirely lacked HLA mutations, and only single mutations were detected in glioblastoma (n=390) and ovarian cancer (n=432). 214 of 298 HLA mutations (71.8%) fell in 64 recurrent positions (i.e. amino acids that were mutated in at least two instances). The recurrent sites were distributed across the HLA gene (median of 2 mutated cases/recurrent site (range 2–24), Fig. 3-bottom, Supplementary Table 13 and Supplementary Fig. 4b–c).

Somatic class I HLA mutations are enriched for localization at sites affecting peptide-MHC interaction and are likely positively selected

Alterations highly likely to have a functional effect, including loss-of-function events (nonsense, frameshift indels, splice site), were significantly enriched in HLA mutations compared to non-HLA mutations (Fig. 4a, chi-squared test $P < 2.2 \times 10^{-16}$). We also observed that while potentially loss-of-function mutations occur in all functional domains of the HLA molecule, they demonstrated a strong preference for the N-terminal end in the leader peptide sequence ($P=0.0038$), which would likely result in a completely non-functional protein (Supplementary Fig. 4d). The highest frequency of mutations occurred in exon 4 (118 mutations, 39.6%) which encodes the α3 domain of the HLA protein that binds to the CD8 co-receptor of T cells[28] (Fig. 4b). Abrogation of this function could lead to a loss of T cell recognition and thereby a loss of immune reactivity. The second highest frequency of mutations occurred in exon 3 (56 mutations, 18.8%) followed by exon 2 (49 mutations, 16.4%), which encode the α1 and α2 peptide binding domains of the HLA molecule respectively which conventionally bind 9- and 10-mer peptides for antigen presentation[29].

Analysis of the position of the mutated residues within exons 2 and 3 in relationship to their predicted interaction with binding peptide[29] further strongly suggests alteration of immune function by these somatic HLA mutations (Supplementary Table 14). The two major anchor

grooves in the HLA molecule bind to positions 2 and 9 respectively of the peptide, and mutation in either groove would be expected to profoundly affect the biochemical stability of the MHC-peptide complex[29]. A secondary anchor groove that interacts primarily with the sixth amino acid of the peptide lies between the two primary anchor grooves[30]. Overall, 28.6% of mutations (30 of 105) in the peptide binding domains were in residues that come in contact with the peptide and 80% (24 of 30) of these were in positions that comprised one of the two primary anchor grooves (Fig. 4c).

We hypothesized that loss-of-function HLA mutations would more likely arise in the presence of selective pressure imposed by the host immune response against the tumor. A growing body of studies has shown that higher mutational burdens in cancers give rise to a higher load of mutation-derived immunogenic epitopes and that immune responses against these are associated with clinical benefit[31]. These immune responses are presumably driven by the presentation of tumor-derived epitopes by antigen presenting cells to stimulate effector lymphocyte responses. Consistent with the idea that a tumor would evolve in a manner to escape recognition and destruction by tumor-directed T or NK cells, we detected an association between the presence of HLA somatic mutations and tumor expression signatures consistent with infiltration by effector lymphocytes, as recently defined[32] (Supplementary Table 15 and Fig. 4d). Although putative loss-of-function somatic mutations in tumor HLA genes could lead to a decrease in the presentation of immunogenic epitopes by the tumor cell and evasion of immunologic targeting, these same mutations would not impact the ability of non-tumor host antigen-presenting cells to ingest and present tumor antigens to T cells, thereby stimulating immune infiltration. For these analyses, we examined the expression of 18,000 genes in matched RNA-Seq data from 4,512 samples across 11 tumor types and found the strongest associations in 6 of 11 cancer types (stomach, endometrial, cervical, head and neck, colorectal, and glioma), suggesting that reduced MHC class I activity may be particularly important for driving immune escape in these tumor types. From this unbiased analysis, the most significantly enriched genes were interferon gamma (IFNG), T cell attractive chemokines (CXCL9, CXCL10, CXCL11), lytic molecules (GZMA, GZMH, PRF1, GNLY), as well as the “Cytolytic Activity” metagene (analyzed previously as a measure of anti-tumor T/NK cell activity[32]). These results suggest that acquisition of HLA mutations without abrogation of expression may provide a complementary immunosurveillance escape mechanism in which potential destruction of the tumor by T cells and natural killer (NK) cells is precluded.

DISCUSSION

Immune evasion is a critical process in tumor biology and is enabled by several mechanisms including immune-editing[33], down-regulation of HLA expression[34], secretion of immunosuppressive mediators[35], and expression of proteins that modulate immune checkpoints[36]. Most recently, somatic mutation of HLA genes was revealed to be a significantly frequent process in some tumor types[4]. Improved sensitivity and accuracy of somatic HLA mutation detection could better characterize this already strongly implicated mechanism of immune evasion across cancers. We therefore created POLYSOLVER, a model-based algorithm for accurate inference of HLA typing information from germline exome-capture data which enables more sensitive and specific detection of somatic HLA

mutations compared to standard techniques reliant on alignment to the canonical reference genome.

We have demonstrated that POLYSOLVER infers HLA-type information with 97% sensitivity and 98% precision from exome-capture sequencing data and is among the best-performing tools for the analysis of HLA loci from WES data. Indeed, different typing tools, or a combination thereof, may be used for optimizing different aspects of HLA mutation detection performance, e.g. a consensus approach that only uses allele species commonly identified by multiple tools as basis for mutation detection would favor increased specificity at the cost of sensitivity. The improved performance of HLA mutation detection was assessed to be primarily due to use of inferred alleles as reference and employment of stringent criteria for filtering aligned reads prior to mutation calling. We estimate an increase in sensitivity from 58.8% to 94.1% and specificity from 20% to 53.3% over standard methods based on validation of point mutations in RNA-Seq data, a performance similar to that of the recently published OptiType tool[20]. An expected limitation of POLYSOLVER is its restriction to identification of known alleles, but future versions may be augmented by an assembly-driven module which would enable discovery of novel HLA alleles, and by representing a wider range of ethnic groups. POLYSOLVER, as well as other available HLA typing tools that can be used with WES, are also not yet suitable for clinical use where much higher accuracy (>99.9%) is required. However, the POLYSOLVER-based mutation detection pipeline can still be used effectively for detecting somatic changes in HLA genes once experimentally determined HLA typing information is available.

In this study, we performed a comprehensive characterization of HLA mutations in 7,930 samples across 20 different tumor types. We have shown that, in comparison to previous studies, the HLA mutational spectrum elucidated by our analysis has significantly reduced false positives and detects additional somatic mutations. Several biologic insights emerged from our analysis. First, we identified colon adenocarcinoma to be significantly affected by somatic mutation in class I HLA genes in addition to head and neck, lung squamous and stomach cancer, thus further supporting HLA mutation as a common oncogenic mechanism. In contrast, other cancers such as glioblastoma, ovarian cancer and chronic lymphocytic leukemia largely lacked mutations in HLA genes. Second, several characteristics of the identified nonsynonymous mutations suggest that they functionally impact antigen presentation. We identified 29 sites across the HLA genes that were recurrently mutated in at least 3 cases, and 35 sites by 2 cases suggesting positive selection at these positions. We further noted a significant enrichment in loss-of-function events in the HLA genes such as frameshifting indels, nonsense and splice site mutations. These events would be expected to abrogate HLA class I surface expression on tumors[37–39], thereby impacting antigen presentation to immune cells. We determined that the majority of the detected mutations map to regions critical for antigen presentation. More than a third of the mutations (39.6%) were in exon 4 that encodes the MHC class I allele α3 domain, which binds to the CD8 co-receptor on T cells[28]. Mutations in this domain have been previously shown to abrogate binding to CD8[40]. Exons 2 and 3 harbored 35.2% of the mutations – these exons encode the surfaces that present peptides to immune cells. We found evidence that exon 2 and 3 HLA mutations preferentially localized to residues critical for anchoring peptide to the MHC

binding grooves, and would be expected to interfere with the fundamental process of antigen presentation[29, 30].

Finally, we observed a strong association between effector lymphocyte gene expression signatures and HLA mutations, which is consistent with the hypothesis that somatic changes in these genes are a plausible immune escape mechanism that arise in response to increased cytolytic activity in several tumor types. However, additional experiments are required to better understand this mechanism.

Improvements in massively-parallel sequencing technologies are now enabling increased coverage and longer read lengths, which should further help POLYSOLVER in resolving somatic changes in HLA regions. Further efforts will be focused on extending the methodology to other data modalities including RNA-Seq and whole genome sequencing. In addition to enabling better detection of HLA mutations, accurate HLA typing by POLYSOLVER can also be used to study germline associations of HLA alleles in diseases such as autoimmune diseases and cancer. It could be used prospectively for preliminary screening for matches for allogeneic organ transplantation. Finally, as described in the current report, POLYSOLVER can be potentially extended to extract sequence and mutation information from other polymorphic regions in the genome such as MHC class II, non-classical MHC alleles, TAP1 and TAP2 genes, and MIC-A and MIC-B ligands, and hence is a generally applicable analysis framework to address these otherwise challenging loci.

Online Methods

Whole-exome sequencing data

All samples were obtained under Institutional Review Board approval and with documented informed consent. A complete list of TCGA samples is given in Supplementary Table 11. Mutational spectra of chronic lymphocytic leukemia (CLL)[17, 45] and melanoma[24] have previously been reported, while mutations lists for lung squamous carcinoma (LUSC), lung adenocarcinoma (LUAD), bladder (BLCA), head and neck (HNSC), colon (COAD) and rectum (READ), glioblastoma (GBM), ovarian (OV), uterine corpus endometrial carcinoma (UCEC), and breast (BRCA) were obtained from the Sage Bionetworks' Synapse resource (http://www.synapse.org/#!SYNAPSE:syn1729383). For a subset of CLL patients (N=8), HLA typing was performed by molecular typing (Tissue Typing Laboratory, Brigham and Women's Hospital, Boston, MA) and these cases were used as a training set for the POLYSOLVER algorithm (Supplementary Table 1). The validation set comprised 253 samples from 183 distinct individuals (47 Caucasians, 50 Blacks, 41 Chinese and 45 Japanese individuals) that had both exome data and experimentally determined HLA type information[12] (http://www.1000genomes.org).

POLYSOLVER allele database creation

To maximally retrieve true HLA reads, we constructed a full length genomic reference library of known HLA alleles (6597 unique entries) based on the Multiple Sequence Alignment (MSA) files provided in the IMGT database (v3.10; http://www.ebi.ac.uk/ipd/imgt/hla/), similar to the approach described in Erlich *et al*[12]. We first used the cDNA file to impute exons in an incompletely sequenced allele by using a reference allele that had

protein-level identity with the allele in question, as was evident by concordance of 4-digit nomenclature. If no such reference allele was available, we set as reference an allele that derived from the same allele group, as was evident by concordance of 2-digit nomenclature. In cases where there were multiple such possibilities for choosing the reference allele, we chose the first listed allele in the MSA. A similar approach was used to impute the missing components of the sequences listed in genomic (gDNA) MSA file. Finally the full length genomic sequence of each allele was imputed by assembling exons from the cDNA imputation step and introns from the gDNA imputation.

Ethnicity inference and prior probability estimation

4-digit allele frequencies for different ethnicities were calculated by taking a sample-size weighted average of all relevant population studies in the Allele Frequency Net Database (http://www.allelefrequencies.net/).

A rapid principal components analysis (PCA) based method was developed to infer ethnicity for samples of unknown racial origin (Kiezun *et al*, manuscript in preparation). Exome data for samples of known (self-described) ethnicity from the 1000 Genomes and HapMap projects (n=1,398, with 911 Caucasians, 375 Blacks, 54 Asians, and 58 South Asians) was genotyped at a predefined set of 5,845 loci chosen based on considerations related to known linkage disequilibrium between different loci, representation on population genotyping platforms and consistency between genome releases[46]. A PCA revealed distinct segregation of Caucasian, Black, Asian, and South Asian samples in the 2-dimensional space defined by the first two principal components. Any new sample of unknown ethnicity can now be projected in this space and its Euclidean distance from the clusters centroids can be computed. Ethnicity is inferred based on the cluster of minimal distance from the sample projection.

Allele inference

The posterior probability calculations for alleles corresponding to each HLA gene (A, B or C) are performed separately as described below:

Let

$N_A \equiv$ # alleles corresponding to the HLA gene

$N \equiv$ # reads aligning to at least one allele

$N_m \equiv$ # reads aligning to allele a_m

$N_T \equiv$ # reads in the sequencing run

$f_m \equiv$ population based prior probability of allele m

$r_{k1} \equiv$ first read of read pair r_k

$r_{k2} \equiv$ second read of read pair r_k

$d_k \equiv$ insert length of read pair r_k

$l_{k1} \equiv$ length of first read of read pair r_k

$l_{k2} \equiv$ length of second read of read pair r_k

$q_i \equiv$ Phred-like quality of sequenced base *i*

$e_i \equiv$ probability that the sequenced base *i* is an error

$$e_i = 10^{-\frac{q_i}{10}}$$

The quality scores of the alignment were used to build a model for the sequencing process. Let us say that a given read pair r_k does in fact derive from an allele a_m and their sequence relationship allowing for miscalls in the sequencing process is accurately captured in the alignment. Let Y_{Ai}, Y_{Ci}, Y_{Gi} and Y_{Ti} denote random variables corresponding to observing bases A, C, G and T respectively at position *i* in read pair r_k in it's alignment to allele a_m. Then

$$Y_{Ai},\ Y_{Ci},\ Y_{Gi},\ Y_{Ti} \sim Multinomial\,(n{=}1;\ \alpha_{Ai},\ \alpha_{Ci},\ \alpha_{Gi},\ \alpha_{Ti})$$

where

$\alpha_{Bi} = 1 - e_i$ if reference base at position *i* in a_m is B

$= e_i/3$ otherwise

Let D denote a random variable for the observed insert length of a paired read in the sequencing run based on alignment to the complete genome. For a given read pair r_k, the empirical insert size distribution can be used to estimate the probability of observing the insert length d_k as

$$P\,(D{=}d_k) = \frac{\sum_{l=1}^{N_T} I(d_l{=}d_k)}{N_T}$$

Assuming positional independence of quality scores, and independence of generated reads and their insert sizes, the probability of observing r_k given allele a_m is then

$$P\,(r_k|a_m) = \begin{cases} \prod_{i=1}^{l_{k1}} \alpha_i \prod_{j=1}^{l_{k2}} \alpha_j \cdot P\,(D{=}d_k) & if\ r_k\ aligns\ to\ a_m \\ s_k & otherwise \end{cases}$$

where s_k corresponds to the lowest theoretical probability achievable for read pair r'$_k$ with perfect base qualities and segment lengths equal to those of r_k. Since 93 is the maximum achievable base quality under Illumina 1.8+ format, s_k is computed as

$$s_k = (l_{k1}{+}l_{k2}) \cdot log\frac{10^{-9.3}}{3} \approx -23 \cdot (l_{k1}{+}l_{k2})$$

The posterior probability of allele a_m using all reads that align to it is given by

$$P\left(a_m|r_1, r_2, \ldots, r_N\right) = \frac{\prod_{k=1}^{N} P\left(r_k|a_m\right) \cdot f_m}{\prod_{k=1}^{N} P(r_k)}$$

Log transformation of the above equation yields

$$L_m = \sum_{k=1}^{N_m} \sum_{i=1}^{l_{k1}} log\ \alpha_i + \sum_{k=1}^{N_m} \sum_{j=1}^{l_{k2}} log\ \alpha_j + \sum_{k=1}^{N_m} log\ P\left(D=d_k\right) + \left(N - N_m\right) s_k + log\ f_m - \sum_{k=1}^{N} log\ P(r_k)$$

Note that the terms $N \cdot s_k$ and $\sum_{k=1}^{N} log\ P(r_k)$ are constants for all alleles and can be ignored. The first allele is inferred as the one that maximizes the posterior probability.

$$a_w = \underset{a_m}{\operatorname{argmax}}\ L_m$$

To infer the second allele we had to handle the fact that different alleles are very similar to each other, including the winning allele. Therefore, we weight reads aligning to multiple alleles according to the likelihood of reads with respect to the different alleles by applying a heuristic strategy. For a given allele a_m, the likelihood l_m^k of a read r_k that also mapped to the winning allele a_w with likelihood l_w^k was weighted by a factor equal to $l_m^k/(l_m^k + l_w^k)$. Consequently, reads mapping exclusively to a_m with respect to a_w were assigned a weight of 1. The read insert size and allele prior probability components were preserved from the first allele inference step. The second winner at each locus was identified as the allele with the maximal reevaluated score.

Pre- and post-processing steps for HLA mutation detection

Prior to detection of somatic changes using Mutect and Strelka by comparison of tumor and normal HLA reads aligned to POLYSOLVER-inferred HLA alleles, the following changes and filters were implemented: (i) *NotPrimaryAlignment* bit flag was turned off from all alignments since several reads mapped to multiple alleles; (ii) mapping quality was changed to a non-zero value (=70) for all reads; (iii) alignments where both mates did not align to the same reference allele were discarded; and (iv) alignments where at least one mate had more than one mutation, insertion or deletion event compared to the reference allele were discarded. Soft-clipping of the reads was not allowed during the alignment. Alleles with multiple detected somatic changes were removed from the analysis. In cases where both inferred alleles were identical in the region of detected somatic mutation, the mutation was assigned to the more common allele in the population. All somatic events were visualized using IGV (Mutect: ‘KEEP’ entries in call_stats file, Strelka: All entries in all.somatic.indels.vcf file) and the ones that passed manual review were further annotated for the gene compartment (intron, exon, splice site) and protein change. Splice sites were defined as the set of splice consensus sequence positions that had a bit score of at least 1 in

either the human major/U2 or human minor/U12 introns at the exon/intron boundaries (9 positions at the 5′ splice donor end of the intron including the ultimate base in the upstream exon, and 2 positions at the 3′ splice acceptor end of the intron)[47].

Validation of somatic HLA mutations by RNA-Seq evaluation

The MutationValidator tool (manuscript in preparation) was used for orthogonal confirmation of mutations in RNA-Seq data. A mutation was considered validated in RNA-Seq if there were at least 2 reads supporting the mutation. In brief, to determine the power, we first model the distribution of allelic fraction of the mutation based on the exome data as a Beta(a+1, r+1) distribution, where a is the number of reads bearing the alternate allele and r is the number of reads bearing the reference allele at the site of mutation. Then, given the total number of reads aligning at the position in the RNA-Seq data (N), power was calculated as the probability that we would detect at least two reads bearing the alternate allele in the RNA-Seq data (assuming it has the same underlying allele fraction as the DNA) using the Beta-binomial distribution *Beta-Binom*(N,a+1, r+1) i.e.

$$Power=\sum_{k=2}^{N}\left(\begin{array}{c} N \\ k \end{array}\right)\frac{B(k+a+1,n-k+r+1)}{B(a+1,r+1)}\,where$$
$$B\,(x,y)=\frac{(x-1)!(y-1)!}{(x+y-1)!}$$

A threshold of 80% power was used to consider a site to be powered to detect the mutation in the RNA-Seq data. Sites that had less than 80% power were removed from the analysis.

Standard HLA typing

Standard HLA typing was performed at the Brigham and Women's Hospital Tissue Typing Laboratory using a combination of sequence-specific oligonucleotide probe (SSO) and sequence specific primer (SSP) techniques. Genomic DNA samples were initially typed using locus specific LabType® SSO kits (One Lambda Inc.) and analyzed using a Luminex 200. Loci for which there were more than one common well documented (CWD) allele were subsequently resolved by PCR-SSP kits (One Lambda Inc. and Life Technologies) and analyzed using gel electrophoresis.

Validation of inferred somatic HLA mutations by targeted long sequencing of *HLA-A* and –*B*

***HLA-A* and *HLA-B* amplification of TCGA samples—**HLA locus-specific amplification for *HLA-A* and *HLA-B* sequences were performed separately using HGSgo-AmpX kits from GenDX (Utrecht, Netherlands). Briefly, for each sample, 100ng of genomic DNA was mixed with 1ul of AmpX primer (GenDX), 1.25uL dNTP mix (Qiagen), 2.5uL LongRange PCR Buffer (Qiagen), 0.4uL LongRange PCR Enzyme (Qiagen) and nuclease free water was added to a final volume of 25uL per reaction. Samples were then placed in a thermal cycler and PCR was performed using the following conditions: initial denaturation at 95°C for 3 min, followed by 35 cycles of 95°C for 15 sec., 65°C for 30 sec. and 68°C for 6 min, followed by a final incubation at 68°C for 10 min. All PCR reactions were then purified using Agencourt AMPureXP beads according to the manufacturer's protocol

(Beckman Coulter). Following AMPureXP purification, the concentrations of the amplification products (approximately 3.1–3.4 kb) were confirmed by Quant-iT (Life Technologies) and the sizes were confirmed using an Agilent Bioanalyzer DNA 7500 kit.

Library construction and long sequencing—SMRTbell DNA template libraries were prepared from the *HLA-A* and *HLA-B* amplicons according to the manufacturer's suggested protocol (5kb Template Preparation and Sequencing, Pacific Biosciences). Briefly, equimolar pools of *HLA-A* and *HLA-B* amplicons were prepared for each sample. Pooled amplicons were then end repaired and ligated to barcoded SMRTbell adapters. Following the addition of barcoded SMRTbell adapters, all samples were pooled and exonuclease treated according to the manufacturer's suggested protocol. Pooled, barcoded libraries were then purified using AMPure PB beads (Pacific Biosciences) and quantified using an Agilent Bioanalyzer DNA 7500 kit. Pooled samples were sequenced in SMRTCells using a Pacific Biosciences RSII instrument using the P6 DNA/Polymerase Binding Kit in conjunction with the DNA Sequencing Reagent 4.0. Barcoded subreads were analyzed using the SMRT Analysis (version 2.3.0) Long Amplicon Analysis (LAA) protocol.

Analysis—We confirmed the accuracy of the Pacific Biosciences-based long sequencing approach through testing 6 samples from normal volunteers with known HLA typing (performed at BWH Tissue Typing laboratory based on a combination of sequence-specific SSO and SSP techniques, see above), wherein we observed 100% concordance between the two approaches. The LAA phased consensus fastq sequences and HLA typing for each sample were derived using a set of publicly available analysis tools (https://github.com/bnbowman/HlaTools). In total, data was generated from 28 samples corresponding to 18 different mutations (10 tumor/normal pairs and 8 tumor only cases). The median number of subreads generated per sample was 20,120 (range: 7,464 – 40,990). For validation of POLYSOLVER-predicted mutations, the subreads from the corresponding samples were split into contiguous 76-mers, aligned to alleles comprising the inferred HLA type for the individual using Novoalign and visualized using IGV. Only reads that had no more than one somatic event of the same type (mismatch, insertion, deletion) as the mutation being assessed were retained. After filtering, the median number of 76-mer reads mapping to the allele predicted to have the mutation was 1,046 (range: 9 –3,860). Power was calculated using the MutationValidator tool as described above, and a threshold of 80% power was used in evaluating the mutations.

Identifying changes in gene expression associated with non-silent MHC Class I mutation

Gene expression data were obtained and processed as described[32]. In short, "Level_3" gene-level data were obtained from GDAC Firehose (http://gdac.broadinstitute.org). Read counts were tallied per gene symbol and divided by the gene symbol's maximum transcript length (as defined by UCSC Genome Browser's table "knownIsoforms" (hg19 version)). For each sample, these values were rescaled to sum to a total of one million, such that expression estimates may be interpreted as Transcripts Per Million transcripts (TPM).

For each gene (of ~18,000 quantified pan-cancer), a one-sided Wilcoxon rank-sum test was applied to determine whether the mutants (those samples non-silently mutated in any of the

six HLA alleles) demonstrated significantly higher expression than the non-mutants. In performing this rank-based test, random tie breaks were applied when two samples exhibited identical gene expression. Note that in addition to the 18,000 genes tested, "cytolytic activity" (defined previously as the geometric mean of GZMA and PRF1 expression[32]) was also included. This process was executed separately per tumor type and excluded tumor types for which the count of mutated samples with available expression data was fewer than three (which excluded glioblastoma, chronic lymphocytic leukemia, kidney clear cell cancer, liver cancer, ovarian cancer, prostate cancer, melanoma, and thyroid cancer). This resulted in a matrix of p-values (11 tumor types by 18,000 genes). Fisher's method was applied to each gene to assess its overall significance across the 11 tumor types. Per-cancer and pan-cancer p-values are presented (Supplementary Table 15). Effect sizes (estimated by taking the ratio of median expression in the mutants to median expression in the non-mutants) for top genes (defined as those with unadjusted $P < 10^{-10}$) are depicted in the form of a heatmap (Fig. 4d). For this heatmap, row and column orderings reflect hierarchical clustering (on the basis of the effect size variable), though dendrograms are not shown.

This entire process was repeated but reversing the directionality of the one-sided Wilcoxon rank-sum tests in order to identify genes with lower expression in HLA mutants. Per-cancer and pan-cancer p-values for this analysis are presented in Supplementary Table 16, and the effect size heatmap appears as Supplementary Figure 5.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

C.J.W. acknowledges support from the Blavatnik Family Foundation, AACR (SU2C Innovative Research Grant), NHLBI (1RO1HL103532-01), and NCI (1R01CA155010-01A1). This work has made extensive use of data generated by TCGA, a project of the National Cancer Institute and National Human Genome Research Institute. We thank Eran Hodis for providing access to the melanoma data. We would also like to thank Caryn McCowan (Broad Technology Labs), Terrance Shea (Broad Technology Labs), Sarah Young (Broad Technology Labs) and Mike Weiand (Pacific Biosciences) for their help in setting up, performing and analyzing data using Pacific Biosciences RSII instruments. We are grateful to Edward Fritsch for critical reading of the manuscript and providing valuable feedback.

References

1. Stransky N, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011; 333:1157–1160. [PubMed: 21798893]
2. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–525. [PubMed: 22960745]
3. Lohr JG, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc Natl Acad Sci U S A. 2012; 109:3879–3884. [PubMed: 22343534]
4. Lawrence MS, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014; 505:495–501. [PubMed: 24390350]
5. Cancer Genome Atlas Research, N. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014; 513:202–209. [PubMed: 25079317]
6. Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. Nature. 1999; 401:921–923. [PubMed: 10553908]

7. Townsend A, Bodmer H. Antigen recognition by class I-restricted T lymphocytes. Annu Rev Immunol. 1989; 7:601–624. [PubMed: 2469442]
8. Bjorkman PJ, Parham P. Structure, function, and diversity of class I major histocompatibility complex molecules. Annu Rev Biochem. 1990; 59:253–288. [PubMed: 2115762]
9. Welsh K, Bunce M. Molecular typing for the MHC with PCR-SSP. Rev Immunogenet. 1999; 1:157–176. [PubMed: 11253945]
10. Fernandez-Vina MA, Falco M, Sun Y, Stastny P. DNA typing for HLA class I alleles: I. Subsets of HLA-A2 and of -A28. Hum Immunol. 1992; 33:163–173. [PubMed: 1618656]
11. Tiercy JM, et al. Oligotyping of HLA-A2, -A3, and -B44 subtypes. Detection of subtype incompatibilities between patients and their serologically matched unrelated bone marrow donors. Hum Immunol. 1994; 41:207–215. [PubMed: 7868376]
12. Erlich RL, et al. Next-generation sequencing for HLA typing of class I loci. BMC Genomics. 2011; 12:42. [PubMed: 21244689]
13. Wang C, et al. High-throughput, high-fidelity HLA genotyping with deep sequencing. Proc Natl Acad Sci U S A. 2012; 109:8676–8681. [PubMed: 22589303]
14. Lank SM, et al. Ultra-high resolution HLA genotyping and allele discovery by highly multiplexed cDNA amplicon pyrosequencing. BMC Genomics. 2012; 13:378. [PubMed: 22866951]
15. Danzer M, et al. Rapid, scalable and highly automated HLA genotyping using next-generation sequencing: a transition from research to diagnostics. BMC Genomics. 2013; 14:221. [PubMed: 23557197]
16. Cao H, et al. An integrated tool to study MHC region: accurate SNV detection and HLA genes typing in human MHC region using targeted high-throughput sequencing. PLoS One. 2013; 8:e69388. [PubMed: 23894464]
17. Wang L, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med. 2011; 365:2497–2506. [PubMed: 22150006]
18. Robinson J, et al. The IMGT/HLA database. Nucleic Acids Res. 2013; 41:D1222–1227. [PubMed: 23080122]
19. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. Nucleic Acids Res. 2011; 39:D913–919. [PubMed: 21062830]
20. Szolek A, et al. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014; 30:3310–3316. [PubMed: 25143287]
21. Cibulskis K, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013; 31:213–219. [PubMed: 23396013]
22. Saunders CT, et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012; 28:1811–1817. [PubMed: 22581179]
23. Omberg L, et al. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. Nat Genet. 2013; 45:1121–1126. [PubMed: 24071850]
24. Hodis E, et al. A landscape of driver mutations in melanoma. Cell. 2012; 150:251–263. [PubMed: 22817889]
25. Robinson JT, et al. Integrative genomics viewer. Nat Biotechnol. 2011; 29:24–26. [PubMed: 21221095]
26. Engstrom PG, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods. 2013; 10:1185–1191. [PubMed: 24185836]
27. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. Genome Biol. 2013; 14:405. [PubMed: 23822731]
28. Fayen J, et al. Class I MHC alpha 3 domain can function as an independent structural unit to bind CD8 alpha. Mol Immunol. 1995; 32:267–275. [PubMed: 7723772]
29. Brusic V, Petrovsky N, Zhang G, Bajic VB. Prediction of promiscuous peptides that bind HLA class I molecules. Immunol Cell Biol. 2002; 80:280–285. [PubMed: 12067415]
30. Ruppert J, et al. Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. Cell. 1993; 74:929–937. [PubMed: 8104103]

31. Brown SD, et al. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. Genome Res. 2014; 24:743–750. [PubMed: 24782321]
32. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015; 160:48–61. [PubMed: 25594174]
33. Schreiber RD, Old LJ, Smyth MJ. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. Science. 2011; 331:1565–1570. [PubMed: 21436444]
34. Bubenik J. MHC class I down-regulation: tumour escape from immune surveillance? (review). Int J Oncol. 2004; 25:487–491. [PubMed: 15254748]
35. Zou W. Regulatory T cells, tumour immunity and immunotherapy. Nat Rev Immunol. 2006; 6:295–307. [PubMed: 16557261]
36. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer. 2012; 12:252–264. [PubMed: 22437870]
37. Norgaard L, Fugger L, Madsen HO, Svejgaard A. Identification of 4 different alternatively spliced HLA-A transcripts. Tissue Antigens. 1999; 54:370–378. [PubMed: 10551420]
38. Brady CS, et al. Multiple mechanisms underlie HLA dysregulation in cervical cancer. Tissue Antigens. 2000; 55:401–411. [PubMed: 10885560]
39. Jimenez P, et al. A nucleotide insertion in exon 4 is responsible for the absence of expression of an HLA-A*0301 allele in a prostate carcinoma cell line. Immunogenetics. 2001; 53:606–610. [PubMed: 11685475]
40. Pittet MJ, et al. Alpha 3 domain mutants of peptide/MHC class I multimers allow the selective isolation of high avidity tumor-reactive CD8 T cells. J Immunol. 2003; 171:1844–1849. [PubMed: 12902485]
41. Boegel S, et al. HLA typing from RNA-Seq sequence reads. Genome Med. 2012; 4:102. [PubMed: 23259685]
42. Kim HJ, Pourmand N. HLA haplotyping from RNA-seq data using hierarchical read weighting. PLoS One. 2013; 8:e67885. [PubMed: 23840783]
43. Bai Y, Ni M, Cooper B, Wei Y, Fury W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genomics. 2014; 15
44. Warren RL, et al. Derivation of HLA types from shotgun sequence datasets. Genome Med. 2012; 4:95. [PubMed: 23228053]
45. Landau DA, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013; 152:714–726. [PubMed: 23415222]
46. Purcell SM, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014; 506:185–190. [PubMed: 24463508]
47. Irimia M, Roy SW. Origin of spliceosomal introns and alternative splicing. Cold Spring Harb Perspect Biol. 2014; 6

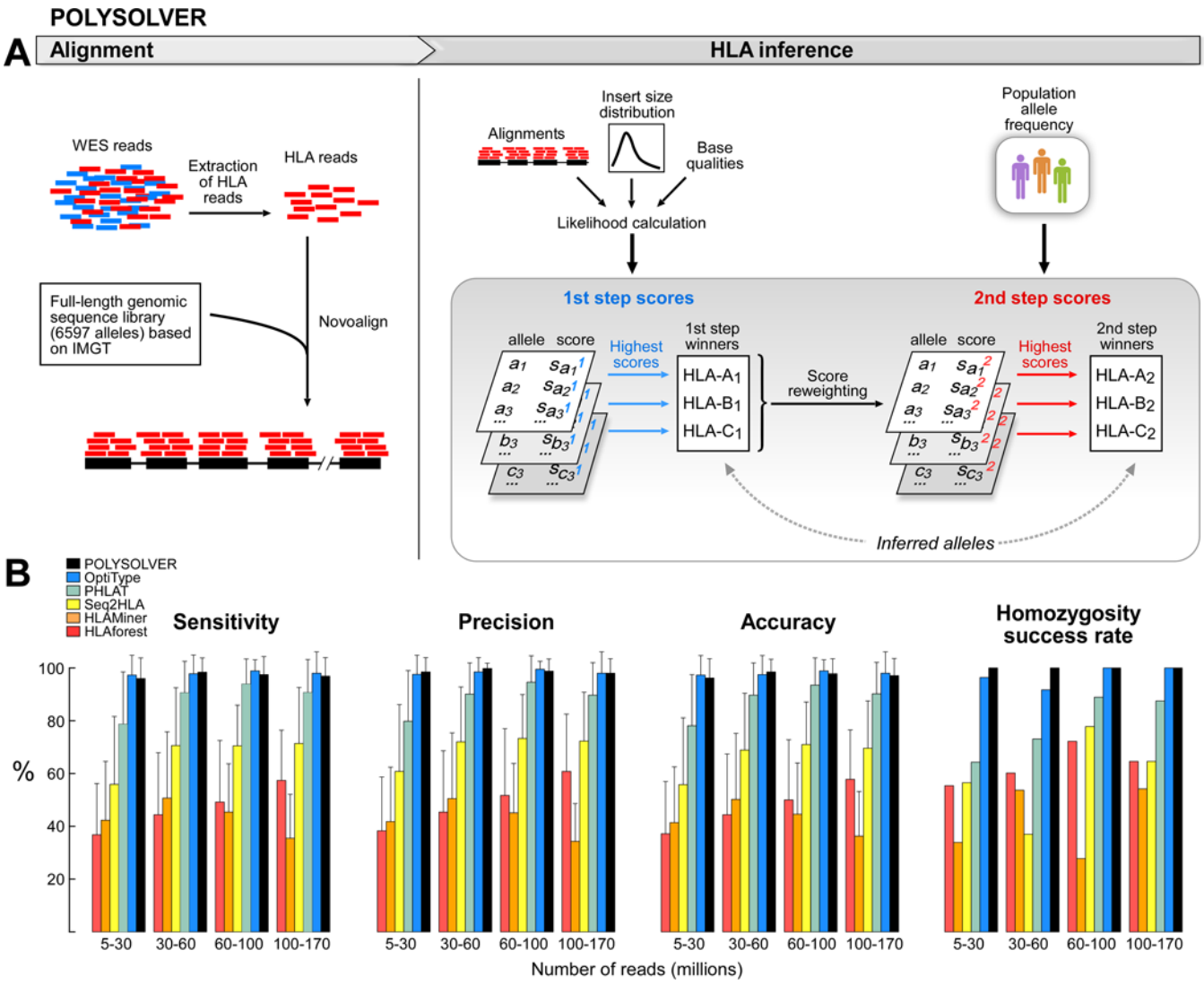

Figure 1. Development and validation of POLYSOLVER for inference of MHC class I type

(a) Schematic of the POLYSOLVER algorithm. **(b)** Comparative performance of POLYSOLVER (black bars) and other previously reported algorithms[20, 41–44] by library size (error bars correspond to s.d.) using the following performance criteria: (i) *sensitivity* – the proportion of all true allele species that are correctly identified by the algorithm; (ii) *precision* – the probability that an inferred allele species is correct; (iii) a*ccuracy* – the fraction of total number of alleles that are correctly called; and (iv) *homozygosity success rate* – the fraction of all homozygous cases that are correctly inferred.

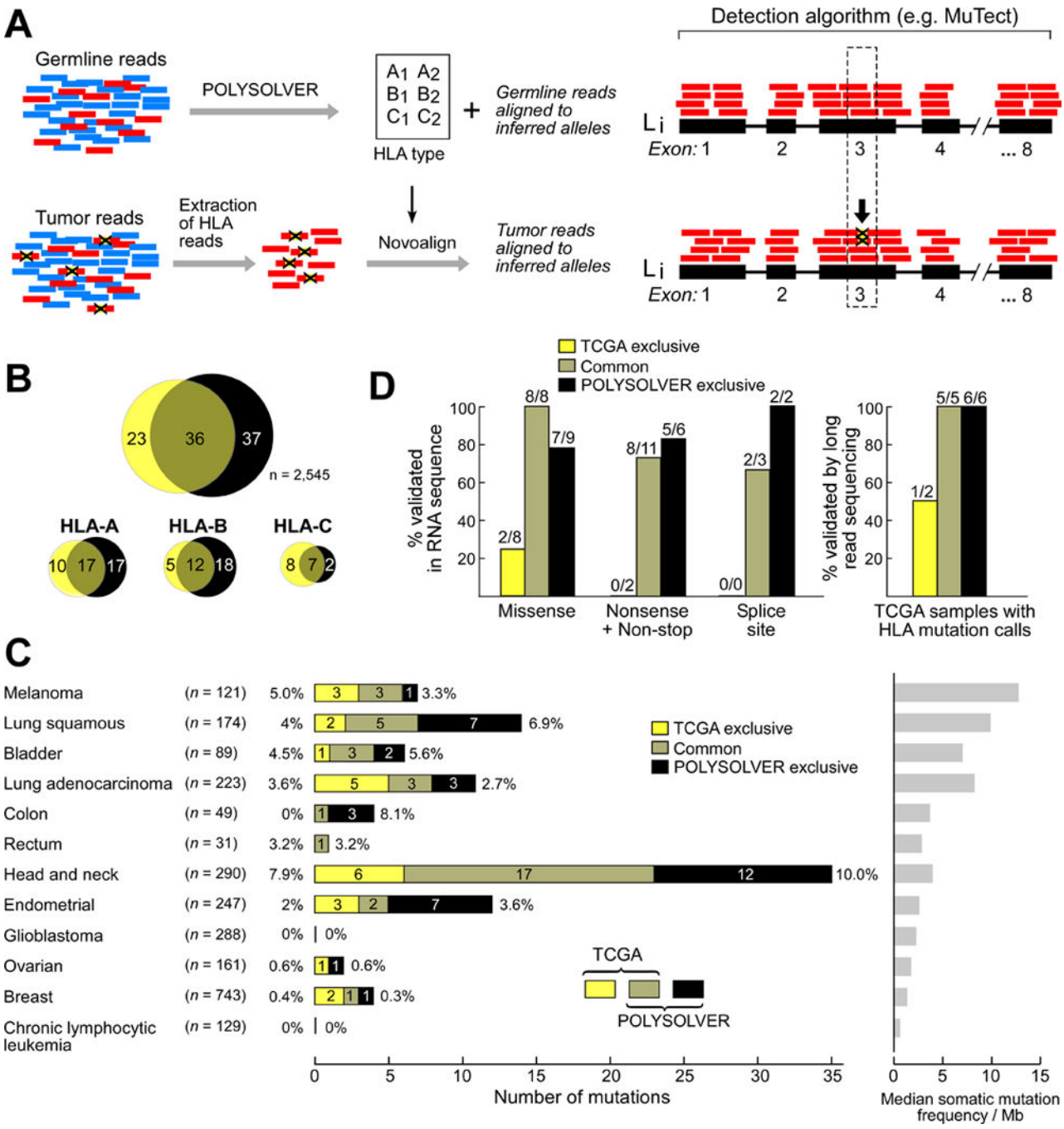

Figure 2. POLYSOLVER for the detection of somatic mutations in MHC class I alleles across cancers

(a) Schema for detection of somatic changes in HLA genes using POLYSOLVER. Mutation detection algorithms Mutect[21] and Strelka[22] were incorporated for calling point mutations and indels respectively, following MHC class I typing of the germline by POLYSOLVER. **(b)** Comparison of somatic HLA mutations identified by TCGA (yellow) across cancers using standard approaches to those identified by POLYSOLVER (black) (n=2,545). Green – mutations found in common between the two datasets. **(c)** Number of HLA mutations and

the percentage of samples bearing HLA mutations per cancer type identified by standard methods (yellow) and POLYSOLVER (black). **(d)** Validation of mutations using RNA-Seq and long read sequencing. RNA-Seq based validation was restricted to 49 samples with HLA point mutations (missense, nonsense, non-stop, splice site) identified by exome analysis and with available RNA-Seq data. Long read sequencing using Pacific Biosciences' SMRT® technology was performed on HLA alleles from 18 samples with available DNA material (Online Methods)[27].

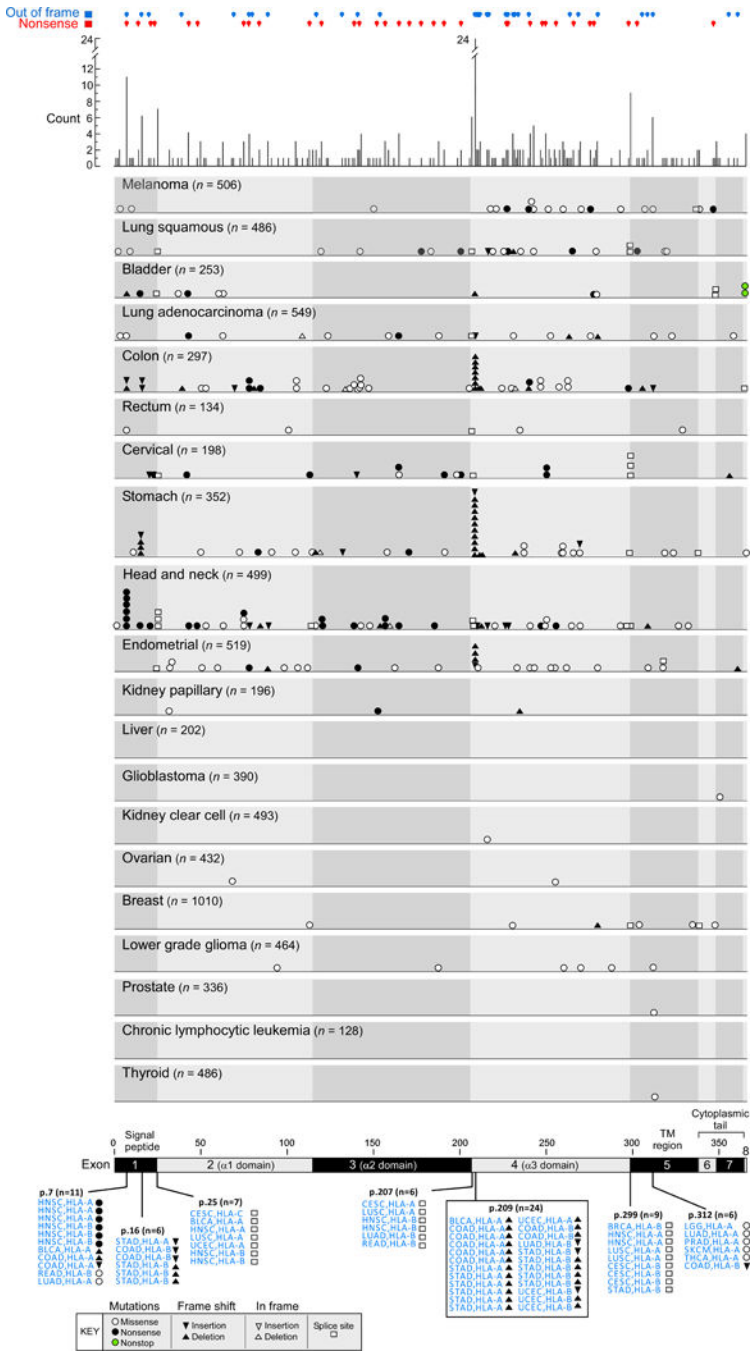

Figure 3. Distribution of HLA mutations across cancers

Distribution of HLA mutations across functional domains and tumor types. *Top* – Distribution of potential loss-of-function events; out of frame (blue) and nonsense mutations (red). The histogram summarizes the number of events identified at each position. *Central panel* – Pattern of mutations detected in each tumor type. *Bottom* – Recurrent events; recurrent positions (with disease, allele group) with frequency >= 5 cases/recurrent site are shown.

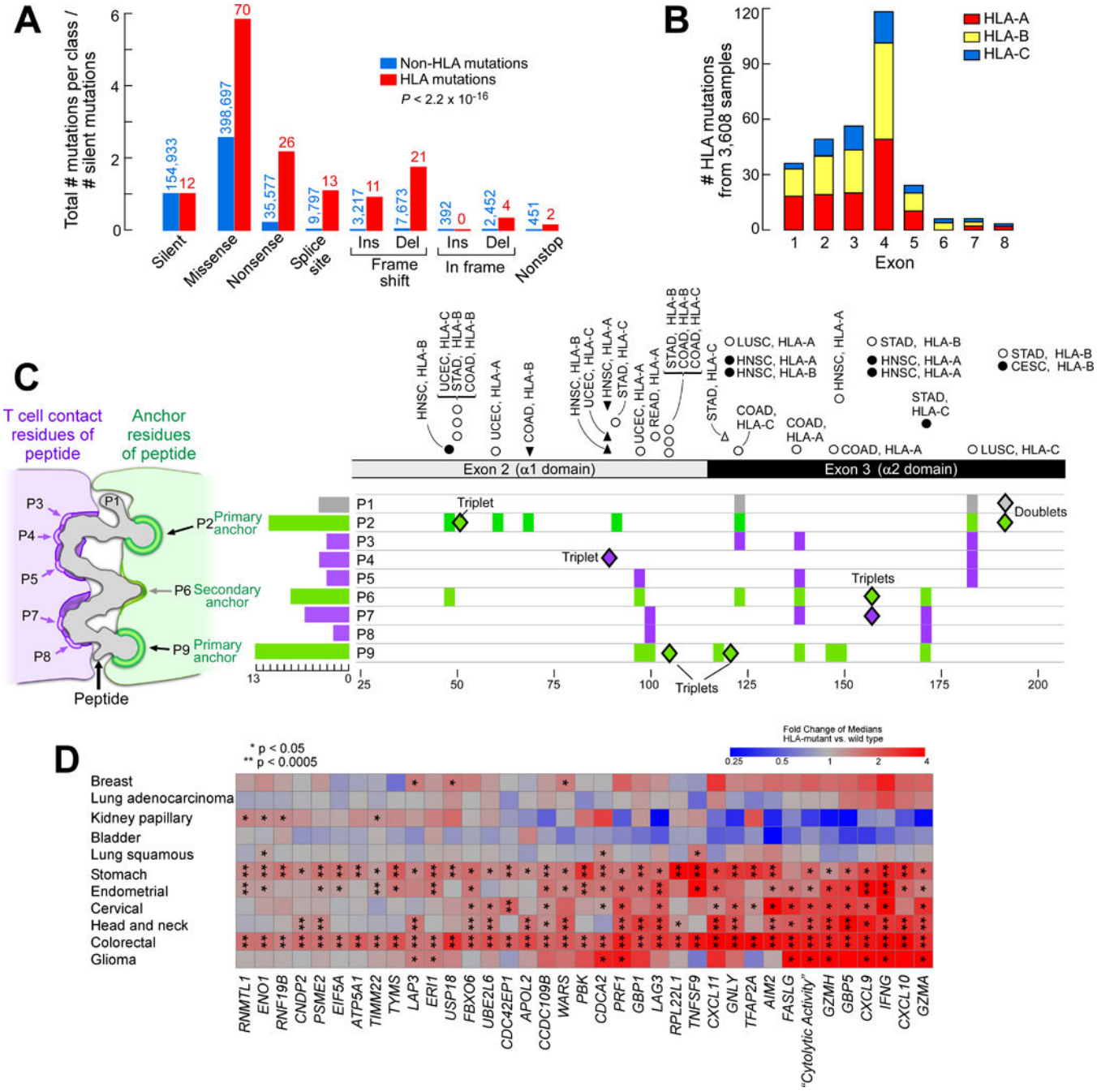

Figure 4. Distribution of MHC class I mutations and evidence of positive functional selection

(a) Comparison of spectrum of mutations in non-HLA genes and HLA genes. The ratio of number of mutations of a particular type to the number of silent mutations is compared between the non-HLA and HLA genes for all mutation types (chi-square test, $P < 2.2 \times 10^{-16}$). **(b)** Distribution of HLA mutations across exons. **(c)** Mutations in HLA positions that are in actual physical contact with the peptide (contact residues). *Left panel* – The relative orientation of a 9-mer peptide with respect to the HLA and T cell molecules. Positions 2 and 9 constitute the primary anchors while position 6 forms the secondary anchor with HLA.

The remaining position interacts with the T cell molecule. *Right panel* – The 9 amino acids of the peptide and their corresponding HLA contact residues are indicated along the rows (orange – HLA interacting anchor positions, blue – T cell interacting positions). The histogram depicts the frequency of observed HLA mutations in contact residues corresponding to each peptide position[29]. **(d)** Killer lymphocyte effector genes are more highly expressed in tumors exhibiting MHC Class I mutation. Unbiased statistical analysis was employed to find genes more highly expressed in tumors harboring a mutation in an MHC class I allele. Heatmap displays color-coded expression ratio of medians (HLA-mutant vs. non-mutant samples) for genes (columns) in each cancer type (rows), excluding cancer types with fewer than 3 instances of HLA mutation in the cohort. Asterisks (* or **, see key) indicate the significance of the association for the given gene in the given cancer type according to one-sided Wilcoxon rank-sum test (null hypothesis: expression is not greater in the mutants). Cytolytic activity (geometric mean of GZMA and PRF1 expression) is included as though a gene. The depicted genes are those for which expression in MHC Class I-mutated tumors was most significantly elevated pan-cancer (unadjusted $P < 10^{-10}$ combined by Fisher's method, Supplementary Table 15). Corresponding analysis for genes with reduced expression in MHC Class I mutants was also performed (Supplementary Fig. 5 and Supplementary Table 16).