PLOS GENETICS

RESEARCH ARTICLE

Quantifying immune-based counterselection of somatic mutations

Fan Yang[1,2,3¤], Dae-Kyum Kim[1,2,3], Hidewaki Nakagawa[4], Shuto Hayashi[5], Seiya Imoto[5], Lincoln Stein[1,6], Frederick P. Roth[1,2,3,7,8]*

1 Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, **2** Donnelly Centre, University of Toronto, Toronto, Ontario, Canada, **3** Lunenfeld-Tanenbaum Research Institute, Mt. Sinai Hospital, Toronto, Ontario, Canada, **4** Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences, Tokyo, Japan, **5** The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, **6** Ontario Institute of Cancer Research, Toronto, Ontario, Canada, **7** Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, **8** Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America

¤ Current address: Department of Pathology, Stanford University, Palo Alto, California, United States of America
* fritz.roth@utoronto.ca

OPEN ACCESS

Citation: Yang F, Kim D-K, Nakagawa H, Hayashi S, Imoto S, Stein L, et al. (2019) Quantifying immune-based counterselection of somatic mutations. PLoS Genet 15(7): e1008227. https://doi.org/10.1371/journal.pgen.1008227

Editor: Peter McKinnon, St Jude Children's Research Hospital, UNITED STATES

Received: December 11, 2018

Accepted: June 4, 2019

Published: July 25, 2019

Data Availability Statement: PCAWG patient HLA types are provided in S3 Table. TCGA missense mutation data and TCGA RNAseq expression data are available via the Broad Institute TCGA Genome Data Analysis Center (2016) as "Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run Broad Institute of MIT and Harvard (Dataset. https://doi.org/10.7908/C11G0KM9)". Protein sequence data are publicly available via Ensembl BioMart (Release 89). TCGA patient HLA types are available via Broad Institute's Firehose (Authorization Domain: TCGA-dbGaP-Authorized).

Abstract

Somatic mutations in protein-coding regions can generate 'neoantigens' causing developing cancers to be eliminated by the immune system. Quantitative estimates of the strength of this counterselection phenomenon have been lacking. We quantified the extent to which somatic mutations are depleted in peptides that are predicted to be displayed by major histocompatibility complex (MHC) class I proteins. The extent of this depletion depended on expression level of the neoantigenic gene, and on whether the patient had one or two MHC-encoding alleles that can display the peptide, suggesting MHC-encoding alleles are incompletely dominant. This study provides an initial quantitative understanding of counter-selection of identifiable subclasses of neoantigenic somatic variation.

Author summary

Cancer immunotherapy and personalized cancer vaccines depend on clearance of cancer and pre-cancer cells by the immune system. However, little is known about the strength of this phenomenon as it acts on the cell populations which give rise to tumors. Here we provide an initial quantitative estimate of the fraction of neo-antigen-containing cells in this population that are cleared by the MHC class I-dependent immune system. The impacts of both neo-antigenic gene expression and the number of neo-antigen-displaying MHC alleles on this clearance phenomenon were examined. A more complete understanding of immune clearance of neoantigenic cells and how this phenomenon varies between patients and cancers, has the potential to guide immunotherapy and cancer vaccines.

Funding: This work was funded by the National Human Genome Research Institute of the National Institutes of Health Center of Excellence in Genomic Science grant [HG004233], the Canada Excellence Research Chairs Program, a Canadian Institutes of Health Research Foundation Grant to FR (https://www.genome.gov/10001771/centers-of-excellence-in-genomic-science/); a Project for Cancer Research and Therapeutic Evolution of Japan Agency for Medical Research and Development grant to HN (https://www.amed.go.jp/en/program/list/01/03/003.html); an Open Foundation of Shandong Provincial Key Laboratory of Network-based Intelligent Computing (SPKL2017-G001) grant to FY (http://nbic.ujn.edu.cn/index.htm); and by a Banting Postdoctoral Fellowship of Canada (http://banting.fellowships-bourses.gc.ca/en/home-accueil.html) and Basic Science Research Program through the National Research Foundation of Korea grant funded by the Ministry of Education (2017R1A6A3A03004385; http://www.nrf.re.kr/eng/main) to DKK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

In every human cell, proteins are constantly being degraded into component peptides, and a subset of this pool of peptides are displayed on MHC class I receptor proteins (encoded by human leukocyte antigen or HLA genes). As somatic mutations arise, some cause differences in MHC-displayed peptides, producing antigens that can be differentially recognized by T cells and lead to the specific destruction of tumor cells by the immune system [1]. In addition to the production and display of 'non-self' peptides that can arise directly from mutation, genetic and epigenetic alterations can cause tumor cells to express many proteins more highly [2]. Together, these changes mean that cancer cells have an altered repertoire of proteins and therefore of tumor antigens.

Tumor antigens can be classified into two categories: tumor-associated self-antigens (which may also be displayed by non-cancer cell types) and antigens derived from tumor-specific mutant proteins. The latter class of tumor-specific 'neo-antigenic' mutations are ideal targets for cancer immunotherapy, because neo-antigens that can potentially be recognized by the mature T-cell repertoire are less likely to be found in healthy cells/tissues [3]. It has been reported that neo-antigens are likely to be more immunogenic, presumably due to the T-cell maturation process in which T-cells capable of high-avidity recognition of self-antigens are eliminated [4]. Immuno-therapy approaches exploiting neo-antigenicity, however, have been hampered by the fact that every tumor possesses a unique set of mutations that must first be identified [5]. Moreover, individual patients can differ dramatically in their immune systems, based on HLA type and other allelic variation in immune genes, as well their unique repertoire of mature immune cells. Thus, personalized immuno-therapy could positively benefit the patient during cancer treatment [6–8]. After recognition, the process of tumor-cell killing by T-cells may release more tumor neo-antigens in a potentially therapeutic virtuous cycle [9].

In principle, any coding mutation has the potential to generate a mutant peptide that can be presented by MHC class I molecules and subsequently recognized by cytotoxic T cells. However, a crucial challenge for the personalized treatment approach is determining the MHC-binding potential of non-self peptides that arise from somatic tumor mutations, and determining which among them are most likely to be potent neo-antigens in a given cancer type, and given the patients repertoire of HLA alleles that encode different MHC class I receptors.

To improve our understanding of neo-antigenicity in cancer, we conducted several analyses of somatic mutations and the ability of corresponding mutant peptides to be displayed by MHC class I receptors across different cancer types. More specifically, we quantified the impact of predicted antigenicity on the spectrum of tumor missense somatic mutations. We expected to find that somatic mutations would be less frequent in MHC-displayed peptides, presumably because the immune system is more likely to have eliminated cells bearing these mutations. Other groups have identified depletion of predicted-displayed mutations based on patient HLA-A genotypes [10], without quantifying the extent of depletion. Other work reported that predicted-MHC-displayed mutations were depleted in colorectal and clear cell renal cancer [11]. However, this phenomenon was not explored in detail, e.g., it did not consider patient genotypes at all HLA loci or consider expression levels of the displayed peptide.

Here, we quantified the extent to which somatic mutations are significantly depleted in peptides that are predicted to be displayed by MHC class I proteins (without considering patient HLA type). We further characterized the dependence of this depletion on the inferred expression level of each peptide. Next, we refined the preceding analyses by considering individual patient HLA alleles. Finally, we extended this analysis by relating depletion of somatic

mutations to the number of HLA alleles predicted to display peptides bearing that mutation. Thus, we quantitatively estimated the 'neoantigenicity' of different classes of somatic variants in individual patients.

Results

Depletion of mutations within expressed predicted MHC-binding peptides

As somatic mutations arise, we should expect that the more immunogenic mutations are more likely to be counter-selected due to clearance of the mutant cell by the immune system, and therefore depleted from observed tumor genomes. To formally test this hypothesis and to begin to quantify the expected depletion effect, we examined somatic cancer mutations in human cancer samples, beginning with data from the Pan-cancer Analysis of Whole Genomes (PCAWG) study [12].

The immunogenicity of a protein-coding mutation depends in part on whether or not it yields a mutant peptide that is displayed by a MHC class I protein receptor. MHC class I binding peptides were predicted using the NetMHC server [13, 14]. In total, we examined 121,258 missense somatic mutations from 2,834 PCAWG patients for whom HLA type had been assigned [15]. Those mutations were distributed across more than 10,700 genes. Missense somatic mutations from PCAWG were separated into two groups: either falling within or outside of predicted MHC binding peptides. For an initial analysis, we modeled all MHC class I alleles with available display predictions as being present in each patient (we revisit this issue later).

Because a mutant protein must be expressed in order to yield a displayed peptide, we also examined the dependence of missense variant depletion on gene expression levels. More specifically, we analyzed the relationship between the missense mutation density within MHC-binding peptides and the expression level of the corresponding protein in the appropriate cancer type (see Methods). Then, for mutations both within and outside of MHC binding peptides, we calculated the mutation density for five classes of peptide: those that were undetectably expressed and those in each of four gene expression quantiles (Methods).

As expected, we found that mutation density and expression level are negatively correlated, and that the average mutation density within MHC binding peptides is lower than that of MHC binding peptides for expressed peptides (Fig 1; ratio of mutation density within MHC-displayed peptides to that outside displayed peptides = 0.94; Fisher's exact test, P-value $< 2.2e^{-16}$). As a control, we further compared the mutation density within and out of MHC binding peptides in undetectably-expressed genes. Our results indicated that there was no significant depletion of missense somatic mutations within MHC binding peptides that are not detectably expressed (Fig 1; odds ratio = 1.01, P-value = 0.65). Although the odds ratio was near 1 for non-expressed proteins, as one might naively expect, we note that the sequence specificity of specific MHC class I receptor alleles can lead to HLA-allele-dependent amino acid (and therefore nucleotide-level) sequence biases in the peptides displayed, which could in turn yield sequence-dependent differences in mutation density. To account for this, we performed a correction by dividing the mutation density ratio of expressed proteins by that of non-expressed proteins. Although in this case the corrected mutational density ratio was 0.93/1.01, which is still 0.93, it did make a difference for other results below.

Thus, our analysis of PCAWG data confirmed the expected phenomenon that somatic mutations are depleted within expressed MHC-displayed peptides. Quantifying the MHC-display-dependent depletion effect in non-expressed peptides served as a crucial negative control for sequence biases of peptides displayed by particular HLA alleles.

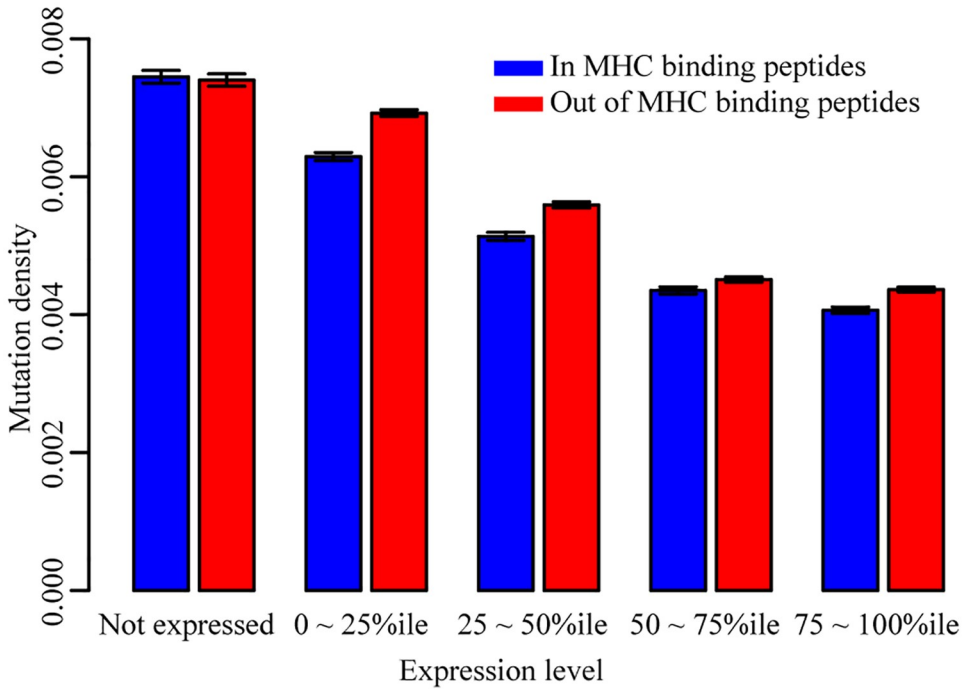

Fig 1. MHC-display-dependent mutation densities for genes with different expression levels. Blue bars are the mutation density within the predicted MHC binding peptides. Red bars are the mutation density out of the predicted MHC binding peptides. Mutations were separated into five categories based on the expression levels of their genes.

https://doi.org/10.1371/journal.pgen.1008227.g001

Depletion of mutations within predicted patient-displayed MHC-binding peptides

For a mutant protein to yield a peptide that is displayed by a given allele of the MHC class I receptor, that allele must of course be present in the cells of that patient. Because the analyses above were based on a hypothetical (and unrealistic) patient who bears all 12 of the common HLA alleles for which display predictions are available, the depletion effect sizes estimated above are likely to be conservatively small. Indeed, individual patients can differ dramatically in their immune systems, in part due to allelic variation in HLA genes. Therefore, we sought to characterize the mutation depletion phenomenon using, for each somatic variant, only peptide display predictions for the subset of HLA alleles carried by the patient in which that somatic variant was detected.

Re-examining the PCAWG data, there were 12,552 genes in which at least one variant was predicted to be neo-antigenic, e.g., presented by the MHC class I protein of the patient carrying this mutated gene. For these genes, we again examined the tendency for depletion of mutations within MHC binding peptides relative to non-MHC binding peptides, now taking patient HLA type into account. Within expressed proteins, the ratio of mutation density within predicted-displayed MHC binding peptides to that outside predicted-displayed peptides was 0.82 (Fisher's exact test, P-value $< 2.2e^{-16}$). Within non-expressed proteins, the corresponding ratio was 0.98 (Fisher's exact test, P-value = 0.19), yielding a corrected mutational density ratio for expressed proteins of 0.83 (0.82/0.98).

Our analysis showed that missense mutations tend to be counter-selected within MHC binding peptides, both in an idealized patient with unknown HLA type, and when accounting for HLA type in each specific patient sample. In each case, the phenomenon depended on

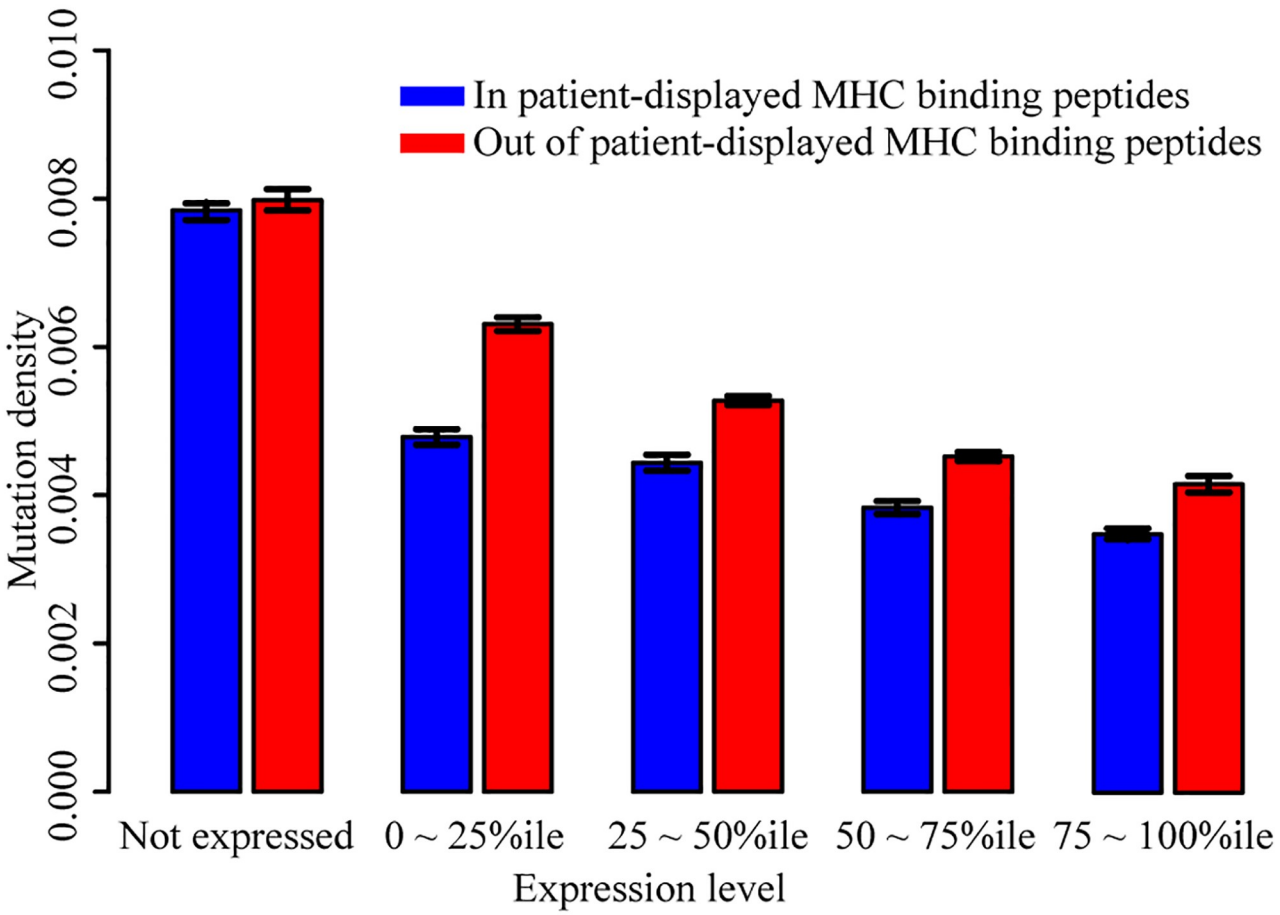

Fig 2. MHC-display-dependent mutation densities for genes with different expression levels, considering each patient's HLA type. Blue bars are the mutation density within the predicted patient-displayed MHC binding peptides. Red bars are the mutation density out of the patient-displayed predicted MHC binding peptides. Mutations were separated into five categories based on the expression levels of their genes.

https://doi.org/10.1371/journal.pgen.1008227.g002

expression level of the gene encoding that peptide (Fig 2). In all subsequent analyses, we considered only peptides expressed according to RNA-Seq analysis of the appropriately-matched cancer type.

Dependence of depletion on the number of mutation-displaying alleles

In the above analysis, we only considered for each peptide whether or not the patient carried an HLA allele predicted to display that peptide but did not consider how many copies of the displaying allele were present in that patient. However, peptides for which two copies of the displaying HLA alleles were present could be more efficiently displayed. (This could be due either to increased expression of the displaying allele by increased gene dosage, or a decreased chance that the displaying allele would be silenced where the phenomenon of mono-allelic expression occurs [16]). We assessed this hypothesis further by testing, for patient samples where 'likely-displayed' mutations were found, whether the number of alleles that can display the MHC binding peptides was associated with the extent of mutation depletion.

Missense variants from the 2,834 PCAWG patient samples were separated into three types (Fig 3). "D0," where the patient has zero HLA class I alleles that are predicted to display the mutant peptide; "D1", where only one HLA class I allele type can display the peptide, i.e., the

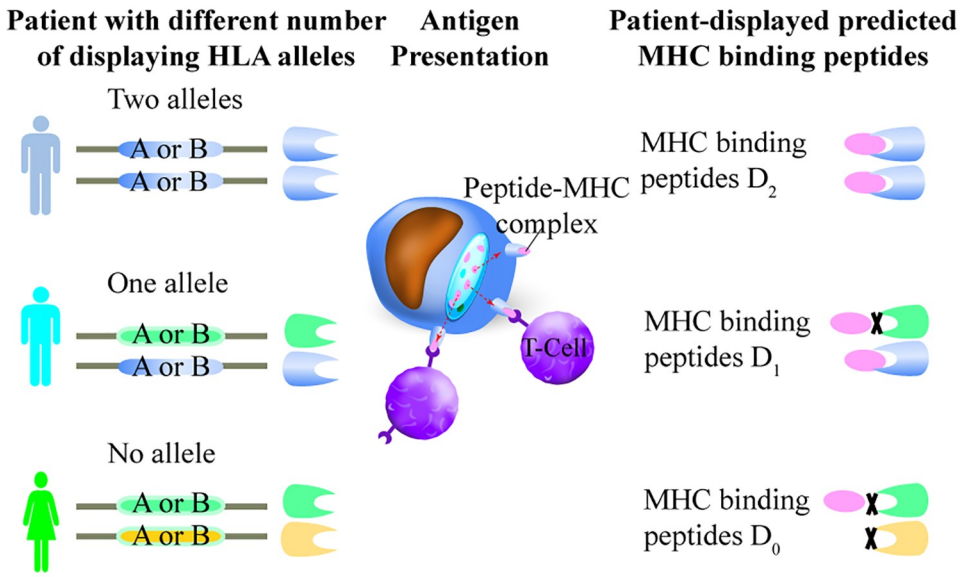

Fig 3. Three types of MHC binding peptides based on patient HLA allele types.

https://doi.org/10.1371/journal.pgen.1008227.g003

patient is heterozygous at the relevant HLA locus and the patient has only one HLA allele that can display the peptide; and "D2", where two HLA class I alleles are predicted to display the peptide. These two alleles can either be two copies of the same MHC allele (i.e., the patient is homozygous for a displaying allele) or be two different alleles (i.e., the patient is heterozygous with alleles that are both predicted to display the peptide).

For both D1 and D2 mutations, we found that the mutation density within patient-displayed MHC binding peptides is lower than that observed outside of MHC binding peptides of the same protein. For expressed MHC binding peptides of type D1, the ratio of mutation density within displayed peptides to that outside of displayed peptides was 0.91 (Fisher's exact test, *P*-value = $1.96e^{-7}$). This ratio for non-expressed peptides was 0.99 (Fisher's exact test, *P*-value = 0.39), yielding a corrected mutational density ratio of 0.92 (0.91/0.99) for expressed D1 peptides.

For expressed displayed peptides of type D2, the ratio was 0.79 (Fisher's exact test, *P*-value = $9.73e^{-9}$). The corresponding ratio in non-expressed displayed peptides D2 that can be displayed by two distinct HLA alleles is 1.02 (Fisher's exact test, *P*-value = 0.64). Thus, a corrected mutational density ratio 0.77 (0.79/1.02) was observed for expressed D2 peptides displayed by two HLA alleles. Thus, we find that the depletion for mutations in MHC-displayed peptides is stronger if the patient has more alleles predicted to display a mutant peptide (Fig 4), and therefore that HLA alleles are incompletely dominant.

Validation of mutation depletion phenomena in an independent dataset

We repeated the above analyses using the missense somatic mutations detected from 5,213 patient samples provided by the TCGA project [17], examining the distribution pattern of 676,171 missense mutations detected in more than 10,800 genes. Analysis of this TCGA data confirmed the tendency of depletion of mutations within MHC binding peptides relative to non-MHC binding peptides, both with and without considering patient HLA types (S1 Fig). Considering only patient-displayed MHC binding peptides, the corrected mutational density ratio was 0.54 (with 95% confidence interval of 0.539–0.545 estimated by bootstrap

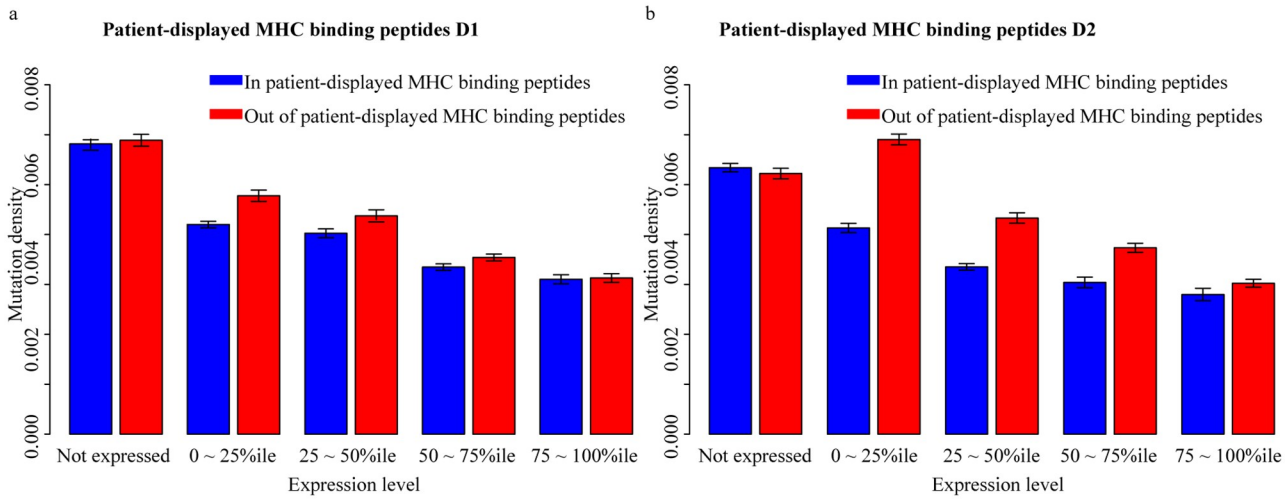

Fig 4. MHC-display-dependent mutation densities for genes with different expression levels, considering the number of displaying HLA alleles. Average mutation density in peptides predicted to be displayed by one or two of the 12 common HLA-A or HLA-B allele types. A. Mutation density in peptides predicted to be displayed in patients by only one HLA allele. B. Mutation density in peptides predicted to be displayed in patients with two displaying HLA alleles.

https://doi.org/10.1371/journal.pgen.1008227.g004

resampling; S2 Fig). Our analysis of the TCGA data confirmed that mutations displayed by two display-enabling HLA alleles (mutations of type "D2") were more strongly depleted than mutations displayed by a single display-enabling allele (S2 Fig), further supporting the conclusion that HLA alleles are incompletely dominant.

To address concerns that the depletion phenomenon stems from a bias in the spectrum or rate of mutation for expressed genes, we also analyzed 1,048,575 synonymous mutations in 5,134 samples. We did not find depletion of synonymous mutations within patient-displayed MHC binding peptides (S3 Fig). Within expressed proteins, the ratio of synonymous mutation density within predicted-displayed MHC binding peptides to that outside predicted-displayed peptides was 1.05 (Fisher's exact test, *P*-value = 0.99). Within non-expressed proteins, the corresponding ratio was 1.03 (Fisher's exact test, *P*-value = 0.78), yielding a corrected mutational density ratio for expressed proteins of 1.01 (1.05/1.03). The 95% confidence interval of the corrected mutational density ratio for synonymous variants was 1.00 to 1.01 (based on bootstrap resampling 500 times; S2 Fig). That we observed no depletion of synonymous mutations in patient displayed MHC binding peptides is consistent with the hypothesis that the depletion phenomenon arises from a selection that depends on expression of the mutant protein.

We next repeated our analysis by considering different cancer types separately. Here, we chose the six different types for which the most samples were available: breast cancer (BRCA, 973 samples), thyroid cancer (THCA, 386 samples), skin cutaneous melanoma (SKCM, 341 samples), prostate adenocarcinoma (PRAD, 329 samples), gastric adenocarcinoma (STAD, 275 samples) and uterine corpus endometrial carcinoma (UCEC, 240 samples (see S2 Table). Significant depletion of predicted-displayed mutations had (without considering patient HLA type or peptide expression) been found previously for BRCA and STAD [11]. We also included adenomatous colorectal cancer (COAD, 60 samples), because Rooney *et al*. noted significant depletion for this cancer type. Considering patient HLA genotypes and proteins in the 75–100%ile of expression level, we could confirm the trend of depletion of mutations in MHC-binding peptides for BRCA, STAD and COAD. We also found depletion for THCA, which had not been previously reported. Although we could not confirm depletion of mutations in

UCEC, SKCM and PRAD for genes at 75%-100% expression percentile, depletion was seen for UCEC and SKCM at other expression percentiles (S4 Fig).

As a negative control, we performed the same analysis for synonymous mutation density within predicted-displayed MHC binding peptides relative to that outside predicted-displayed peptides. The depletion ratios, which did not vary greatly from unity for any of the seven cancer types, were as follows: COAD, 0.97; BRCA, 0.96; THCA, 0.94; STAD, 0.96; UCEC, 0.99; SKCM, 0.95; and PRAD, 1.00. For non-expressed genes, the corresponding results were: COAD, 1.04; BRCA, 1.01; THCA, 1.03; STAD, 0.98; UCEC, 1.02; SKCM, 0.97; and PRAD, 1.00. Only for BRCA were there enough samples to separate mutations into the three categories, D0, D1 and D2, although even for BRCA only 8–10 mutations fell into the D2 category. Still, the increased depletion that had been seen for D2 vs D1 and D0 when considering all cancer types together could be confirmed for BRCA (S5 Fig).

Discussion

In this study, we examined signatures of immune selection pressure on the distribution of somatic mutations, quantifying the extent to which somatic mutations are significantly depleted in peptides that are predicted to be displayed by MHC class I proteins, and characterizing the dependence of this depletion on expression level. We also examined whether the extent of immune selection pressure on somatic mutations depends on whether there are one or two HLA alleles that can display the peptide. It is important to note that peptides, whether displayed by class I MHC receptors or not, are subject to other forms of purifying selection. This could be due to essentiality of encoded functions or immunogenicity arising by mechanisms other than MHC class I display (e.g., MHC class II display [18]). Forms of purifying selection that are independent of class I MHC display should tend to lower mutational density in both displayed and non-displayed peptides. Although this phenomenon is expected to shrink the observed absolute difference in mutational density between displayed and non-displayed peptides, it should not affect the relative difference.

Only expressed MHC binding peptides that can be displayed by at least one patient HLA allele are immunogenic in terms of class I MHC display. In our analysis using the PCAWG dataset, we found mutation densities to be similar for mutations within or out of the predicted MHC binding peptides when the gene was not expressed (Fig 1). That proteins must be expressed to be antigenic is one explanation for the fact that many "likely-displayed" mutations were nevertheless observed in a tumor. We also note that, although expression levels were obtained from tumors of matched type, they were generally not taken from precisely the same tumors for which we had somatic missense variant data. Thus, an explanation for presence of a likely-displayed mutation in an apparently-expressed gene is that this gene is not actually expressed in the specific tumor sample in which it appears. This could be due to differences in environment, germline or somatic genetic background, or epigenetic escape by silencing.

More refined estimates of the depletion effect in future studies might come from using expression data from a specific patient tumor sample. We also noticed that at the 75-100th percentile of gene expression level, there are only weak or even no differences between nonsynonymous and synonymous mutation density for several comparisons. It has been reported that dN/dS diminishes in more highly expressed genes, presumably due to a tendency towards heightened purifying selection for the function of highly expressed proteins in cancers [19, 20]. Our data is consistent with this phenomenon, considering only peptides that are not predicted to be displayed by MHC. Although it stands to reason that MHC display would provide additional purifying selection, and indeed we see this for several comparisons, our statistical

power to detect significant differences must necessarily decrease where there is reduced mutational density in non-displayed peptides.

We note that the terms "in MHC binding peptides" and "out of MHC binding peptides" were applied based on whether or not peptides were predicted to be displayed by at least one of the 12 common HLA-A or HLA-B allele types. We expect to observe depletion of somatic mutations *out* of MHC binding peptides if patients do not have a common displaying allele type. This is because failure to display by any of the common alleles increases the chance that there is display for another allele, e.g., one of the HLA-C alleles or less common HLA-A or HLA-B allele types.

We expect that this information will be useful in building a model that predicts the antigenicity of any given missense mutation detected by whole genome or whole exome sequencing. Although scores for observed mutations based on counter-selection of similar mutations may over-estimate neoantigenicity (if a somatic mutation has been observed, it has obviously not yet been cleared by the immune system), such scores could point to 'cryptic immunogenicity' of a somatic variant. In cases of cryptic immunogenicity, some therapies might enable immune clearance of cancer cells by revealing this immunogenicity, e.g. by relieving tumor-derived suppression of immune cells. The ability to score each observed somatic mutation in a specific tumor for its potential to stimulate an immune response would therefore be potentially useful in scoring tumors with greatest potential to benefit from immunotherapy. Similarly, improved ability to predict which somatically mutated peptides are more likely to be neo-antigens could potentially help in choosing peptides as personalized cancer vaccines to specifically stimulate immune cells to recognize and specifically clear the patient's tumor cells.

Our results also supported the idea that having two copies of the display-enabling allele is more effective for peptide display than having just one copy. This could result from a gene-dosage effect (i.e., incomplete dominance as suggested earlier). Alternatively, it could result from monoallelic expression (MAE). MAE, the phenomenon that only one allele of a given gene is expressed, is a frequent genomic event in normal tissues. MAE-derived silencing of one or more HLA-encoded alleles could potentially cause failure to express MHC binding-peptide-encoding genes, which may, in turn, alter the immunogenicity of somatic mutations. A previous study showed that the genome-wide rate of MAE was higher in tumor cells than in normal tissues, and the MAE rate was increased with specific tumor grade. Oncogenes exhibited significantly higher MAE in high-grade compared with low-grade tumors [16, 21, 22]. The role of MAE in immunogenicity of cancerous cells is entirely unclear. Because HLA alleles are known to be subject to MAE [16], it may be interesting in future studies to assess the impact of MAE by comparing the mutation rates between homozygous (same alleles) and heterozygous (two different alleles) samples at HLA class I loci A and B respectively using the allele-specific expression data. One example of a potential therapy that might emerge from this study is that de-silencing (either global or targeted) could lead to the display of otherwise-cryptic neo-antigens and therefore to immune clearance of cancerous cells, especially when used in combination with current immunotherapy strategies. If we can better understand the interplay between individual immune systems and the likelihood that cancer cells bearing specific somatic mutations are cleared, we will gain insight into the therapeutic potential of MAE modulation. For example, if MAE can indeed limit peptide display efficiency, then therapies reducing MAE could potentially increase the efficiency of immune clearance of tumor cells.

With the analysis conducted here, we can begin to quantify the efficiency of immune clearance of somatically mutated cells. For example, for somatic mutations in proteins expressed in a given cancer type, the depletion ratios we observed were as low as 0.77 in the PCAWG data and as low as 0.54 for TCGA data (in each case this was for expressed peptides predicted to be displayed by an MHC receptor encoded by two copies of the same HLA allele). This result

allows us to predict that cells bearing somatic mutations falling within DNA segments encoding such peptides are cleared roughly 23–46% of the time by the immune system at tumor stages that are earlier than those examined in PCAWG sequencing studies. Because any inaccuracy in estimating protein expression levels or peptide display would be expected to diminish our ability to detect the depletion phenomenon, this estimate of immune clearance rate is likely conservatively low.

Here, we did not consider finer-grained cancer subtypes (e.g., triple-negative BRCA). Although such an analysis would be very interesting and could help identify immune-isolated tumor types, it would require more samples with the requisite HLA type information to be well-powered.

Methods

Obtaining catalogs of somatic variants in cancer samples

This study used two different collections of cancer-cell-derived somatic variants. First, we used data from the Pan-cancer Analysis of Whole Genomes (PCAWG, May 2016 version 1.1) project [23, 24], including 121,258 missense somatic cancer mutations in 10,745 genes detected from 2,834 patient samples. The number of patient samples for each cancer type is shown in S1 Table.

Second, we examined data downloaded from The Cancer Genome Atlas (TCGA) project, obtaining 676,171 missense somatic cancer mutations in 18,106 genes detected from 5,213 patient samples (S2 Table). For TCGA data, we restricted ourselves to cancer types with more than five samples, a known expression level for each gene in a tumor sample of broadly-matched type, and HLA type information for each patient. We also examined 1,048,575 synonymous mutations in 5134 samples as a control. Data were downloaded from Broad Institute TCGA Genome Data Analysis Center (2016-01-28).

Mapping somatic variants to proteins

Protein sequences were downloaded using BioMart R package [25] based on the Ensemble Protein IDs provided in PCAWG and TCGA datasets. Each missense mutation was mapped to the corresponding protein based on the position of the mutation with respect to a given protein (Fig 5). Also, we validated that the wild type residue given for the mutation was found at the corresponding position within the downloaded protein sequence.

Predicting peptides bound by class I MHC receptors

We used the NetMHC server, version 3.4 (13, 14) to predict MHC binding peptides associated with 12 common HLA class I alleles: HLA-A*0101, HLA-A*0201, HLA-A*0301, HLA-A*2402, HLA-A*2601, HLA-B*0702, HLA-B*0801, HLA-B*1501, HLA-B*2705, HLA-B*3901, HLA-B*4001, and HLA-B*5801. For this study, NetMHC scores were obtained for MHC binding peptides of length nine (Although it is possible for peptides with 10 or 11 residues to bind, this is less common and such cases are more difficult to predict). Also, only strong MHC class I binding peptides with NetMHC affinity score of 50 or less were selected (smaller NetMHC scores correspond to higher affinity).

Calculating the depletion of mutations within MHC class I binding peptides

For each class of proteins and variants examined, we determined the total number of mutations falling within and outside of predicted MHC binding peptide regions for each protein.

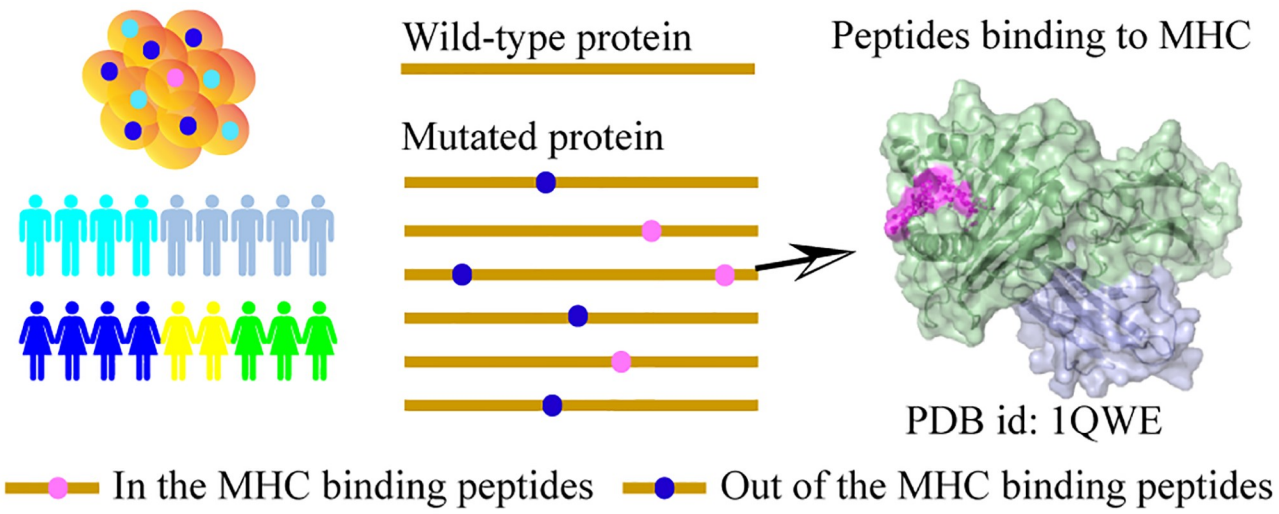

Fig 5. Predicting MHC-binding peptides and calculating mutation densities. Mutations within the MHC binding peptides are shown in blue dots, and mutations out of the MHC binding peptides are shown in pink dots. Protein sequence are shown as yellow line.

https://doi.org/10.1371/journal.pgen.1008227.g005

To test for significant differences in proportions of counts in different groups of peptides, we performed Fisher's exact test using the "stats" package in R.

Estimating transcript expression levels

We estimated gene expression levels for TCGA patient samples using TCGA RNA-Seq data [26]. Data were downloaded from Broad Institute TCGA Genome Data Analysis Center (2016-01-28). The expression level of each gene for each cancer type was estimated using the median expression level of that gene across all TCGA samples of that cancer type. Genes were classified as detectably expressed (if the RNA-Seq by Expectation Maximization or RSEM normalized expression value was greater than 0). Detectably expressed genes were grouped into four expression quantiles according to the RSEM normalized expression value.

Classifying human leukocyte antigen (HLA) types

For PCAWG samples, the four-digit HLA type for 2834 patients was determined by a Bayesian method ALPHLARD (BioRxiv; https://doi.org/10.1101/323766) and all HLA types are shown in S3 Table. For TCGA samples, the four-digit HLA type of the 5213 TCGA patients was predicted using PolySolver [17].

Ethics statement

This study has been approved by the Research Ethics Committee of University of Toronto and the NCBI dbGaP (the Database of Genotypes and Phenotypes) Authorized Access system, project # 15046).

Supporting information

S1 Fig. a. MHC-display-dependent mutation densities for genes with different expression levels using TCGA dataset. Blue bars indicate mutation density within the predicted MHC-binding peptides. Red bars are the mutation density out of the predicted MHC-binding peptides. Mutations were separated into five categories based on the expression levels of their

genes. **b. MHC-display-dependent mutation densities for genes with different expression levels, considering each TCGA patient's HLA type. Blue bars are the mutation density within the predicted patient-displayed MHC binding peptides**. Red bars are the mutation density out of the patient-displayed predicted MHC binding peptides. Mutations were separated into five categories based on the expression levels of their genes.
(PDF)

S2 Fig. a. Exploring uncertainty in corrected mutation density ratio for TCGA mutations in patient-displayed MHC binding peptides. Bootstrap resampling was used for both missense variants (left panel) and synonymous variants (right panel) Observed values are indicated with a vertical dashed line. **b. MHC-display-dependent mutation densities for genes with different expression levels, considering the number of displaying HLA alleles. Average mutation density in peptides predicted to be displayed by one or two of the 12 common HLA-A or HLA-B allele types**. A. Mutation density in peptides predicted to be displayed in patients by only one HLA allele. B. Mutation density in peptides predicted to be displayed in patients with two displaying HLA alleles.
(PDF)

S3 Fig. MHC-display-dependent synonymous mutation densities for genes with different expression levels, considering each TCGA patient's HLA type. Blue bars are the synonymous mutation density within the predicted patient-displayed MHC binding peptides. Red bars are the synonymous mutation density out of the patient-displayed predicted MHC binding peptides. Synonymous mutations were separated into five categories based on the expression levels of their genes.
(PDF)

S4 Fig. MHC-display-dependent mutation densities for genes with different expression levels in different cancer types. Blue bars are the mutation density within the predicted patient-displayed MHC binding peptides. Red bars are the mutation density out of the patient-displayed predicted MHC binding peptides. Mutations were separated into five categories based on the expression levels of their genes.
(PDF)

S5 Fig. In breast cancer, MHC-display-dependent mutation densities for genes with different expression levels, considering the number of displaying HLA alleles. Average mutation density in peptides predicted to be displayed by one or two of the 12 common HLA-A or HLA-B allele types. A. Mutation density in peptides predicted to be displayed in patients by only one HLA allele. B. Mutation density in peptides predicted to be displayed in patients with two displaying HLA alleles.
(PDF)

S1 Table. List of 37 different PCAWG cancer types with number of samples and mutated genes of each cancer type.
(PDF)

S2 Table. List of 31 different TCGA cancer types with number of samples and mutated genes of each cancer type.
(PDF)

S3 Table. HLA genotyping for PCAWG patients.
(TSV)

Acknowledgments

We thank Dr. X. Shirley Liu from the Dana-Farber Cancer Institute and Dr. Sachet Shukla from the Broad Institute for assistance with TCGA sample HLA types.

Author Contributions

Conceptualization: Fan Yang, Lincoln Stein, Frederick P. Roth.

Data curation: Fan Yang.

Formal analysis: Fan Yang.

Methodology: Fan Yang, Seiya Imoto, Lincoln Stein, Frederick P. Roth.

Resources: Dae-Kyum Kim, Hidewaki Nakagawa, Shuto Hayashi, Seiya Imoto.

Supervision: Lincoln Stein, Frederick P. Roth.

Validation: Fan Yang.

Writing – original draft: Fan Yang.

Writing – review & editing: Dae-Kyum Kim, Hidewaki Nakagawa, Shuto Hayashi, Lincoln Stein, Frederick P. Roth.

References

1. Vesely MD, Schreiber RD. Cancer immunoediting: antigens, mechanisms, and implications to cancer immunotherapy. Ann N Y Acad Sci. 2013; 1284:1–5.
2. Gubin MM, Artyomov MN, Mardis ER, Schreiber RD. Tumor neoantigens: building a framework for personalized cancer immunotherapy. J Clin Invest. 2015; 125(9):3413–21. https://doi.org/10.1172/JCI80008 PMID: 26258412
3. Kelderman S, Kvistborg P. Tumor antigens in human cancer control. Biochim Biophys Acta. 2016; 1865 (1):83–9. https://doi.org/10.1016/j.bbcan.2015.10.004 PMID: 26542849
4. Yadav M, Jhunjhunwala S, Phung QT, Lupardus P, Tanguay J, Bumbaca S, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature. 2014; 515 (7528):572–6. https://doi.org/10.1038/nature14001 PMID: 25428506
5. Heemskerk B, Kvistborg P, Schumacher TN. The cancer antigenome. EMBO J. 2013; 32(2):194–203. https://doi.org/10.1038/emboj.2012.333 PMID: 23258224
6. Hutchinson E. Tumour immunology: Differing roles for MYD88 in carcinogenesis. Nat Rev Immunol. 2012; 12(10):681. https://doi.org/10.1038/nri3304 PMID: 22955844
7. Prehn RT. The relationship of immunology to carcinogenesis. Ann N Y Acad Sci. 1969; 164(2):449–57.
8. Haughton G, Amos DB. Immunology of carcinogenesis. Cancer Res. 1968; 28(9):1839–40. PMID: 4877743
9. Chen DS, Mellman I. Oncology meets immunology: the cancer-immunity cycle. Immunity. 2013; 39 (1):1–10. https://doi.org/10.1016/j.immuni.2013.07.012 PMID: 23890059
10. Brown SD, Warren RL, Gibb EA, Martin SD, Spinelli JJ, Nelson BH, et al. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. Genome Res. 2014; 24(5):743–50. https://doi.org/10.1101/gr.165985.113 PMID: 24782321
11. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015; 160(1–2):48–61. https://doi.org/10.1016/j.cell.2014.12.033 PMID: 25594174
12. Campbell PJ, Getz G, Stuart JM, Korbel JO, Stein LD. Pan-cancer analysis of whole genomes. bioRxiv. 2017:162784.
13. Melhem A, Muhanna N, Bishara A, Alvarez CE, Ilan Y, Bishara T, et al. Anti-fibrotic activity of NK cells in experimental liver injury through killing of activated HSC. J Hepatol. 2006; 45(1):60–71.
14. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. 2003; 12 (5):1007–17.

15. Hayashi S, Yamaguchi R, Mizuno S, Komura M, Miyano S, Nakagawa H, et al. ALPHLARD: a Bayesian method for analyzing HLA genes from whole genome sequence data. BMC genomics. 2018; 19(1):790. https://doi.org/10.1186/s12864-018-5169-9 PMID: 30384854

16. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. Widespread monoallelic expression on human autosomes. Science. 2007; 318(5853):1136–40. https://doi.org/10.1126/science.1148910 PMID: 18006746

17. Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol. 2015; 33(11):1152–8. https://doi.org/10.1038/nbt.3344 PMID: 26372948

18. Marty Pyke R, Thompson WK, Salem RM, Font-Burgada J, Zanetti M, Carter H. Evolutionary Pressure against MHC Class II Binding Cancer Mutations. Cell. 2018; 175(7):1991. https://doi.org/10.1016/j.cell.2018.11.050 PMID: 30550793

19. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2018; 173(7):1823.

20. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017; 171(5):1029–41 e21. https://doi.org/10.1016/j.cell.2017.09.042 PMID: 29056346

21. Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, et al. A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics. 2004; 16(2):184–93. https://doi.org/10.1152/physiolgenomics.00163.2003 PMID: 14583597

22. Olivier M. From SNPs to function: the effect of sequence variation on gene expression. Focus on "a survey of genetic and epigenetic variation affecting human gene expression". Physiol Genomics. 2004; 16 (2):182–3. https://doi.org/10.1152/physiolgenomics.00194.2003 PMID: 14726602

23. Kreiter S, Vormehr M, van de Roemer N, Diken M, Lower M, Diekmann J, et al. Mutant MHC class II epitopes drive therapeutic immune responses to cancer. Nature. 2015; 520(7549):692–6. https://doi.org/10.1038/nature14426 PMID: 25901682

24. Kreiter S, Vormehr M, van de Roemer N, Diken M, Lower M, Diekmann J, et al. Erratum: Mutant MHC class II epitopes drive therapeutic immune responses to cancer. Nature. 2015; 523(7560):370.

25. Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford). 2011; 2011:bar030.

26. Center BITGDA. Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run. Broad Institute of MIT and Harvard2016.