Accurate neoantigen prediction depends on mutation position relative to patient allele-specific MHC anchor location

Authors: Huiming Xia[1,2], Joshua F. McMichael[2], Suangson Supabphol[3,4], Megan M. Richters[1,2], Anamika Basu[2], Cody A. Ramirez[1,2], Cristina Puig-Saus[5,6,7] , Kelsy C. Cotto[1,2], Jasreet Hundal[1,2], Susanna Kiwala[2], S. Peter Goedegebuure[3,9], Tanner M. Johanns[1], Gavin P. Dunn[8], Todd A. Fehniger[1], Antoni Ribas[5,6,7], Christopher A. Miller[1,2,9], William E. Gillanders[3,9] , Obi L. Griffith[1,2,9,10,†] , Malachi Griffith[1,2,9,10,†]

Affiliations:

1. Division of Oncology, Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA
2. McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA
3. Department of Surgery, Washington University School of Medicine, St. Louis, MO, USA
4. The Center of Excellence in Systems Biology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand
5. Division of Hematology/Oncology, Department of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
6. Jonsson Comprehensive Cancer Center, Los Angeles, CA, USA.
7. Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA.
8. Department of Neurological Surgery, Washington University School of Medicine, St. Louis, Missouri, USA
9. Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO, USA
10. Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

Abstract

Neoantigens are novel peptide sequences resulting from somatic mutations in tumors that upon loading onto major histocompatibility complex (MHC) molecules allow recognition by T cells. Accurate neoantigen identification is thus critical for designing cancer vaccines and predicting response to immunotherapies. Neoantigen identification and prioritization relies on correctly predicting whether the presenting peptide sequence can successfully induce an immune response. As the majority of somatic mutations are SNVs, changes between wildtype and mutant peptide are subtle and require cautious interpretation. An important yet potentially underappreciated variable in neoantigen-prediction pipelines is the mutation position within the peptide relative to its anchor positions for the patient's specific HLA alleles. While a subset of peptide positions is presented to the T-cell receptor for recognition, others are responsible for anchoring to the MHC, making these positional considerations critical for predicting T-cell responses. We computationally predicted high probability anchor positions for different peptide lengths for over 300 common HLA alleles and identified unique anchoring patterns among them. Analysis of 923 tumor samples shows that 7-41% of neoantigen candidates are potentially misclassified and can be rescued using allele-specific knowledge of anchor positions.

Introduction

Neoantigens are short peptide sequences resulting from somatic mutations specifically found in tumor cell populations. They can be loaded onto major histocompatibility complex (MHC) class I or II molecules to allow recognition by cytotoxic or helper T cells. Upon recognition, T cells are then able to signal cell death for an anti-tumor response. Multiple studies have shown the efficacy of neoantigen based immunotherapy treatments for cancer[1–3] and numerous clinical trials are underway. Accurate neoantigen prediction and prioritization is critical for the design of personalized vaccines[4] and several bioinformatic pipelines have been developed in an attempt to aid this process[5–8].

The effectiveness of a neoantigen-based vaccine relies in part on whether the neoantigen sequences presented to T cells have previously been exposed to the immune system and would be subject to central tolerance (where immune response to antigens is limited as a result of clonal deletion of autoreactive B cells and T cells). While a variety of mutation types are being explored as neoantigen sources[9–13], the vast majority of somatic mutations found in tumors are single nucleotide variants (SNVs). Amino acid sequence changes between the wildtype (WT) and mutant (MT) peptides are subtle and mutant peptides remain similar to native sequences of the host. Additionally, only a subset of positions on the loaded peptide sequence are potentially presented to the T-cell receptor for recognition, and another subset of positions are responsible for anchoring to the MHC, making these positional considerations critical for predicting T-cell responses (**Figure 1**). Thus, subtle amino acid changes must be interpreted cautiously. Multiple factors should be considered when prioritizing neoantigens, including mutation location, anchor position, predicted MT and WT binding affinities, and WT/MT fold change, also known as agretopicity[14]. Examples of four different scenarios involving these factors are illustrated in **Figure 1** where hypothetical wildtype peptides are presented on the left and mutant peptides on the right. The first scenario shows the cases where the WT is a poor binder and the MT peptide, a strong binder, contains a mutation at an anchor location. Here, the mutation results in a tighter binding of the MHC and allows for better presentation and potential for recognition by the TCR. As the WT does not bind (or is a poor binder), this neoantigen remains a good candidate since the sequence presented to the TCR is novel. The second and third scenarios both have strong binding WT and MT peptides. In the second scenario, the mutation of the peptide is located at a non-anchor location, creating a difference in the sequence participating in TCR recognition compared to the WT sequence. In this case, although the WT is a strong binder, the neoantigen remains a good candidate that should not be subject to central tolerance. However, as shown in scenario three, there are neoantigen candidates where the mutation is located at the anchor position and both peptides are strong binders. Although anchor positions can themselves influence TCR recognition[15], a mutation at a strong anchor location generally implies that both WT and MT peptides will present the same residues for TCR recognition. As the WT peptide is a strong binder, the MT neoantigen, while also a strong binder, will likely be subject to central tolerance and should not be considered for prioritization. The last scenario is similar to the first scenario where the WT is a poor binder. However, in this case, the mutation is located at a non-anchor position, likely resulting in a different set of residues presented to the TCR and thus making the neoantigen a good candidate. Recent studies on neoantigens for both mouse and human models confirmed the importance of anchor location when predicting the overall immunogenicity of a given peptide[16,17].

Failing to account for these positional considerations may result in susceptibility to central tolerance and potentially induce auto-immunity. Despite this, the mutation's position within the peptide relative to its anchor positions for the patient's human leukocyte antigen (HLA) alleles, is currently overlooked by neoantigen predicting pipelines. Many recently published neoantigen studies have used simple filtering strategies with either only binding affinity filters[3,18] (e.g. MT peptide IC50 < 500 nM) or with an additional agretopicity filter[19–21], all without specifying whether they account for anchor and mutation locations during their selection process. Researchers have previously discussed how anchor locations can affect our interpretation of other factors considered in neoantigen prioritization (e.g. MT, WT binding affinities)[22]. However, a systematic method for determining anchor locations for the wide range of HLA alleles present in the population and application of these to evaluate MT/WT peptide pairs arising in tumors has not been reported. As a result, many neoantigen studies have either failed to adequately consider this crucial factor or have used conventional assumptions to guide their neoantigen identification process.

Here, we provide a computational workflow for predicting anchor locations for a wide range of HLA alleles using a reference dataset generated from clinical and The Cancer Genome Atlas (TCGA) patient samples. Analysis of results showed clusters of different anchor trends among the HLA alleles analyzed and a subset of these HLA anchor results were orthogonally validated using protein crystallography structures. Using additional TCGA samples, we further evaluated how prioritization results may change when provided with additional anchor information, emulating steps in the neoantigen selection process by an immunotherapy tumor board, tasked with prioritizing vaccine candidates. By incorporating our results into neoantigen prediction pipelines, we hope to formalize and streamline the identification process for relevant clinical studies.

Results

Computational and quantitative prediction of HLA-specific anchor positions. In order to predict anchor locations for a wide range of HLA alleles, we assembled a reference HLA-peptide dataset of strong binding peptides with a median predicted IC50 of less than 500 nM across 8 MHC class I algorithms (NetMHC[23], NetMHCpan[24], MHCnuggets[25], MHCflurry[26], SMM[27], Pickpocket[28], SMMPMBEC[29], NetMHCcons[30]). These peptides were obtained from TCGA and supplemented with additional patient datasets from our own neoantigen study cohorts including lymphoma, glioblastoma, breast cancer, and melanoma (**Methods**). A total of 609,807 peptides were identified with the majority being 9-mers and 10-mers (**Supplemental Figure 1**). It should be noted that such peptides need not be generated from cancer patient data but must simply meet the criteria of being a strong binder for an HLA allele of interest. However, since our focus in this study is on cancer neoantigens and the comparison between wildtype and mutant versions of potentially immunogenic peptides we chose to select peptides from among those actually observed in tumors.

For each HLA allele of which data was obtained, peptides were separated by their respective lengths, ranging from 8 to 11 (**Figure 2**). These peptides were then mutated in silico at all possible positions to all possible amino acids. Predicted binding affinities for each individual peptide were then obtained using the same set of algorithms as previously described. These binding affinities were compared to the median binding affinity of the original strong binding peptide sequence. This comparison enables us to evaluate how mutations occurring at each individual position change the predicted binding interaction between the strong binding peptide and the MHC molecule. A significant change observed at a particular location indicates a higher probability of the amino acid at the position acting as an anchor. On the other hand, little to no change in binding affinities when a position is mutated would indicate a lower probability of the position acting as an anchor. An overall score per position was obtained by summing across all peptides analyzed for an individual HLA allele (**Figure 2; Methods)**.

Prediction results show distinct patterns of HLA anchor locations. We generated anchor prediction scores for the 328 HLA alleles with a sufficient number of strong binding peptides in our reference dataset (**Supplemental Table 1**). These HLA alleles include 95 HLA-A alleles (representing 99.2% observed in the population), 175 HLA-B alleles (representing 97.9% of the population) and 58 HLA-C alleles (representing 98.5% of the population)[31](**Methods**). Results were separated based on peptide lengths (8-11) and the anchor prediction scores across all HLA alleles were visualized using hierarchical clustering with average linkage (**Figure 3; Supplemental Figure 3**). We observed different anchor patterns across HLA alleles, varying in both the number of anchor positions as well as the location. These anchor position patterns could be roughly clustered into at least 6 distinct groups.

Previously, anchor locations have generally been assumed to be at the second and terminal position of the peptide with equal weighting (with the exception of HLA-B*08:01)[32]. Our 9-mer clustering results confirm that the majority of HLA alleles predicted show positions 2 and 9 as likely anchor locations. However, three distinct cluster groups can be further identified within the larger group. The 2S-9W cluster represents HLA alleles with a strong anchor predicted at position 2 and a weak anchor predicted at position 9 (2S-9W; **Figure 3**). The 2W-9S cluster shows those with a strong anchor predicted at position 9 and a weak anchor predicted at position 2 (2W-9S; **Figure 3**). Additionally, we observe a smaller cluster of HLA alleles with moderate anchor predictions for both positions (2M-9M; **Figure 3**) and another cluster with strong anchor predictions for only position 9 (9S; **Figure 3**). We also discovered other patterns differing from the previous anchor assumptions of 2 and 9. In particular, we observed a clustered group of HLA-C alleles that have a moderate anchor at 3 and 9 accompanied by a weaker signal at 2 (2W-3M-9M; **Figure 3**). A smaller group of HLA-B alleles also show an additional anchor at position 5 (2W-3W-5W-9M; **Figure 3**). Our results indicate that a conventional anchor assumption putting equal weights on positions 2 and 9 does not capture the significant heterogeneity in anchor usage between different HLA alleles. These anchor considerations can affect neoantigen prioritization decisions and HLA allele-specific anchor predictions should allow ranking of neoantigens with greater accuracy.

Protein structural analysis confirms predicted anchor results. To validate our anchor predictions, we collected X-ray crystallography structures for MHC molecules with bound peptides (**Supplemental Table 2**). The 166 protein structures collected corresponded to 33 HLA alleles with the majority of them containing 9-mer peptides (8-mers: 6, 9-mers: 110, 10-

mers: 39, 11-mers: 11). These structures were analyzed using two methods: 1) measuring the physical distance between the peptide and the MHC binding groove and 2) calculating the solvent-accessible surface area (SASA) of the peptide residues (**Figure 4a,b**; **Methods**). These methods were selected to validate predicted anchor positions based on the assumptions that if a certain peptide position is designated as an anchor then it is 1) more likely to be closer to the HLA molecule and 2) more secluded from solvent surrounding the peptide-MHC complex compared to non-anchor positions. This is because non-anchor peptide residues should be accessible to the TCR for recognition, thus in the peptide-MHC structures collected where a TCR is not present, peptide surface area available to the surrounding solvent roughly mimics the area that would be accessible by the TCR. HLA-A*02:01, shown as an example, was found to have the greatest number of qualifying structures, with 47 of them containing a 9-mer peptide (**Figure 4c**). These x-ray crystallography structures each capture a snapshot of a dynamic protein structure in constant movement. By overlaying the distance and SASA scores across all 47 complexes respectively, we observe that positions 2 and 9 are the ones most consistently close to the HLA molecule while also being secluded from the solvent. This observation corresponds well with our prediction of strong anchors at both positions 2 and 9 for HLA-A*02:01 and a 9-mer peptide.

To evaluate how the distance and SASA metric correlates with our prediction results across different HLA alleles, we calculated spearman correlations between our prediction scores and distance/SASA results for each peptide position. The distribution of these correlations was compared to that of a randomized dataset where positions of the peptide were randomly shuffled (**Figure 4d; Supplemental Table 3; Methods**). Two sample t-tests showed the distributions were significantly different from the randomized dataset with statistical values of -9.9795 (p value = 1.3757e-18) and -14.7322 (p value = 8.7472e-30) for distance and SASA respectively. Results show that 91.95% of our prediction scores are inversely correlated with the distance metric and 100% of them are inversely correlated with the SASA scores. Furthermore, we analyzed 61 protein structures that contained both the peptide-MHC complex and an additional binding TCR molecule. The distance between the TCR and the peptide showed high correlation with our prediction scores (**Supplemental Figure 4**). Two-sample t-tests showed significant differences between the randomized dataset and both the HLA-peptide distance (p = 8.8023e-08) and the TCR-peptide distance (p=2.4915e-13). These results together strongly suggest that our anchor prediction workflow is returning valid results.

Neoantigen prioritization results are influenced by accounting for anchor locations.

Current pipelines fail to take into account HLA allele-dependent effects on anchor locations and immunotherapy tumor boards lack specific tools and databases to make use of such information. While the decision of whether a neoantigen should be prioritized over others involves many aspects not discussed here (including variant allele frequencies, gene expression, and manufacturability considerations to name a few), we used a straightforward approach to evaluate the effects of introducing improved anchor information on neoantigen prioritization. Depending on the MT and WT binding affinity, agretopicity, mutation position, and anchor location(s), various scenarios can arise when you have a strong binding MT peptide, leading to different choices when prioritizing neoantigens (**Figure 5a**). If the mutation is not at an anchor location, regardless of the WT peptide binding affinity, the MT peptide should be prioritized. In this case, the sequence for TCR recognition contains a mutation and will not be subject to central tolerance (**Figure 5a**; scenario 1,2). Additionally, if the WT peptide is a weak binder, the MT peptide should be accepted regardless of whether the mutation is at an anchor location since both the MT and WT sequences have not previously been exposed to the immune system and therefore not subject to tolerance (**Figure 5a**; scenario 4). However, if the WT binds strongly, regardless of agretopicity, and the mutation is at an anchor location, then this neoantigen will likely be subject to central tolerance and should be rejected from prioritization (**Figure 5a**; scenario 3). These scenarios are considered when performing the anchor position impact analysis.

Our cohort impact analysis involved an additional set of TCGA patient samples where neoantigens were predicted for 923 selected patient-HLA allele pairings. Two different methods were utilized when selecting patient-HLA paired samples: 1) balanced HLA allele distribution and 2) population-based selection (**Methods**). The former method intends to reflect the distribution of HLA alleles in a representative population of patients (TCGA) and the latter method attempts to give a more balanced view of the impact across all HLA alleles without overt bias for the most common alleles. All potential neoantigens were filtered according to three different criteria: A) mutant IC50 < 500 nM and agretopicity > 1 (no anchor filter), B) supplementing this with a conventional anchor assumption (conventional filter), or C) using our computationally predicted anchor locations (allele-specific filter). Peptide results from method 1 showed that under the no anchor filter 57.9% of neoantigens are accepted compared to 92.3% under the conventional filter and 93.2% under the allele-specific filter, showing an overall net gain in the number of peptides when taking anchor considerations into account. When comparing filtered data sets under different criteria, over 41.0% of neoantigens are potentially misclassified using the no anchor filter, and approximately 7.6% of candidates are potentially misclassified between the conventional filter and the allele-specific filter (**Figure 5b**). Method 2 produced similar results where under the no anchor filter 62.4% of neoantigens are accepted compared to 92.3% under the conventional filter and 93.7% under the allele-specific filter. Method 2 results show that over 36.3% of neoantigens are potentially misclassified using the no anchor filter, and approximately 7.2% of candidates are potentially misclassified between the conventional filter and the allele-specific filter (**Figure 5c**). These misclassifications include the inclusion of peptides that are likely to be subject to tolerance (and could lead to false positives) and exclusion of peptides that could be strong candidates (false negatives). By comparing peptides prioritized using the allele-specific anchor filter and those from the no anchor/ conventional anchor filters, we highlighted the potential sources for false positive and false negatives (**Figure 5b,c**). Examples of each scenario were pulled from our dataset to show how peptides passed or failed individual filters (**Figure 5d**).

We additionally performed a patient-level impact analysis using 100 randomly selected TCGA samples, and predicted neoantigens each with their full set of class I HLA alleles (up to 6) (**Methods; Supplementary Table 5; Supplementary Figure 5a,b**). The neoantigen candidates were prioritized using the same set of criterion applied in the previous cohort analysis. We observed a significant impact on neoantigen prioritization results depending on the chosen filtering criteria. Specifically between the no anchor filter and the allele-specific filter, 99% of the patients analyzed had at least one neoantigen decision changed with a median of 11 peptides per patient with altered decisions. Similarly, between the no anchor filter and the conventional filter, 98% of the patients analyzed had at least one decision changed (median: 11 peptides) and between the conventional filter and the allele-specific filter, 65% of the patients had at least one decision changed (median: 1 peptide) (**Supplementary Figure 5c,d,e**). These results show the widespread effect of anchor considerations on patient-level prioritization results.

Discussion

We developed a computational workflow for predicting probabilities of anchor positions for a wide range of the most common HLA alleles. Our results show that anchor positions vary substantially between different HLAs. A subset of our prediction results were confirmed by analyzing available crystallography structures of peptide-MHC complexes. The underlying quantitative scores from our anchor prediction workflow are available for incorporation into neoantigen prediction workflows and we believe this will improve their performance in predicting immunogenic tumor specific peptides. It is important to note that for simplicity reasons our illustrations have depicted peptide residues as either anchoring or potentially participating in TCR recognition. However, previous research has shown that heteroclitic peptides can alter TCR binding and T-cell recognition[15]. Hence, anchor residues and TCR recognition sites should not be considered mutually exclusive and should ideally be interpreted quantitatively where the anchor scores provided reflect the probability of a peptide position participating in binding.

Using an independent pool of TCGA samples, previously excluded from the computational prediction process, we show that consideration of anchor prediction results can have a significant impact on neoantigen prioritization results. However, it is important to note that the choices of whether to accept or reject a prioritization decision were based on hard cutoffs. In most neoantigen characterization workflows, numerous other aspects are taken into account to arrive at an overall prioritization decision, which may further increase differences between filtering strategies. These results not only impact the selection process of neoantigens for personalized cancer vaccines, but also change the way neoantigen load estimation is currently defined. Neoantigen load estimation is commonly defined as the number of peptides whose binding affinity passes a certain threshold. However, this threshold, meant to limit to the approximate number of strong binding neoantigens, should also take into account the mutation position, HLA specific anchor locations, and agretopicity for more precise estimation. Our anchor impact analysis demonstrates the effect of this alteration on estimation results. Moreover, our analysis results show that there is a net gain of neoantigen candidates when taking anchor considerations into account compared to the commonly used agretopicity filters. This becomes important in the context of neoantigen prioritization, particularly when the minimum number of peptide vaccine candidates cannot be met for patients due to low tumor mutational burden.

As previously stated, the neoantigen selection process requires careful consideration of numerous aspects, which have been discussed extensively[4]. In general, neoantigen-based vaccines act by stimulating the patient's immune system for the production of activated cytotoxic T cells. However, compared to viral antigens where the protein sequence is entirely foreign, neoantigens, particularly those developed from SNVs, have merely subtle differences from the individual's wildtype proteome. Thus, the need for a WT versus MT peptide comparison, while considering anchor locations, is an aspect specific to tumor neoantigens that previous other vaccine development pipelines disregarded. Though neoantigens derived from in-frame or frameshift indels diverge more from the WT sequence and are generally less influenced by our findings, cases where such mutations are located towards the beginning or end of a neoantigen may still cause anchor disruption in an allele-specific manner. Additionally, more work should be done to characterize the similarity of neoantigen candidates and the patient's wildtype proteome for an overall accurate prioritization process.

Recently, the Tumor Neoantigen Selection Alliance (TESLA), a group of 25 teams that independently predicted and ranked neoantigens from a common data set, published their findings on features important for neoantigen prediction[17]. They made the unexpected observation that among the 37 positively validated neoantigen candidates, none of the peptides had a mutation at position 2, a common anchor position for a range of HLA alleles, despite a high number of prioritized neoantigens with a position 2 mutation. While this is an interesting finding, it would be premature to exclude such neoantigen candidates with position 2 mutations in future prioritization schemes without additional research. However it does suggest several hypotheses that are relevant to this work and require further analysis. One explanation for the lack of position 2 candidates could be that neoantigens with the mutant residue at a strong anchor position have a disadvantage over those present at TCR sites as they require their WT counterpart to be a poor binder and the threshold for determining sufficiently weak binding of the WT peptide is unclear. Thus highlighting the importance of considering allele-specific anchors in prioritization algorithms as we propose here. Other explanations for why a large number of false positives with a mutation at position 2 exist might include: 1) filters implemented by the TESLA participants that were based only on agretopicity rather than additional considerations of anchor locations and WT binding affinities, 2) bias among neoantigen predictions for specific HLA alleles (the TESLA finding corresponds to data from only 6 individuals) and 3) statistical randomness given limited neoantigen pool. Overall, these studies highlight the importance of anchor positions and further investigation is required to address questions raised by this observation.

In addition to the limitations of being applicable to a subset of neoantigens derived from SNVs and certain indels, our work also needs to be expanded to a wider range of HLA alleles. A larger HLA-peptide reference dataset could be achieved through a wide-scale prediction of strong binders for rare HLA alleles by mutating the wildtype proteome. Furthermore, while x-ray crystallography structures show support for our anchor location predictions, experimental validation with neoantigens designed to induce T-cell activation is needed to explicitly showcase the importance of our results in clinical settings. Although numerous clinical trials using neoantigen-based vaccines are underway, results published show a low accuracy for current neoantigen prediction pipelines[33]. By accounting for additional positional information, we hope to significantly reduce the number of false positive candidates and rescue false negative neoantigens to increase prediction accuracy. A prioritization strategy utilizing anchor results will be incorporated into the visual reporting of our neoantigen identification pipeline pVACseq[5]. Furthermore, machine learning algorithms have been widely applied in the context of neoantigen binding predictions. However, machine learning models trained on experimentally validated data with T-cell activation results are lacking and identifying features for these models is an active area of research. Anchor location probabilities may serve as an additional feature in machine learning model training on clinical data. These results and tools will help streamline the prioritization of candidates for neoantigen vaccines by immunotherapy tumor boards that currently play a central role in the vetting of candidates and may aid in the design of more effective cancer vaccines.

Methods

Input data for identifying strong binding reference peptides for anchor site prediction

We assembled peptide data from various sources where binding prediction data were available through clinical collaborations and supplemented these with TCGA datasets where necessary to achieve better representation of less common HLA alleles. Datasets from clinical collaborations that were incorporated include 7 triple-negative breast cancer samples, 54 lymphoma samples, 20 glioblastoma samples, and 6 melanoma samples. Additionally, we mined data from 9,216 TCGA samples to optimize the number of strong binding peptides matched to each HLA allele by adding 10 samples for insufficient (<10 strong binding peptides) and 15 samples for previously unseen HLA alleles. Of these, 1,356 TCGA samples were used to generate reference HLA-peptide combinations to be used for downstream simulations. High quality variants included from TCGA samples were obtained from the Genomic Data Commons and selected according to their filter status (pass only) and required to be called by at least 2 out of 4 variant callers as previously described[34]. Peptide lengths considered ranged from 8- to 11-mers. In total, these datasets corresponded to 1,443 tumor samples, representing 328 matching HLA alleles, with 737 of these having more than 10 strong binding peptides, and a grand total of 609,807 strong binding peptides for use in the following analyses (**Supplemental Figure 1; Supplemental Table 4**).

Computational prediction of anchor site positions for 328 HLA alleles

Peptides collected from input datasets were first filtered for strong binders using a binding affinity cutoff of 500 nM. These were used to build a reference dataset consisting of peptides predicted to be strong binders to individual HLA alleles. We first performed a saturation analysis to determine the appropriate number of random peptides needed to obtain a robust estimate of the likely anchor site locations of each HLA allele. This was done using peptides collected for HLA-A*02:01, where over 1,500 peptides were obtained for each peptide length. Random sampling with a subset size of 10 peptides showed consistently high correlation (> 0.95) with the largest subset size used where all 1000 peptides were incorporated (**Supplemental Figure 2**). Thus in downstream analysis, for each unique HLA and peptide length combination, 10 peptides were randomly selected from the database. For each of the 10 starting peptides at each position n, we obtained a score that reflects how much a mutation at this position will affect the overall binding affinity:

$$(\mathrm{abs}(Y_{(n,1)}\text{-}X)+..+\ \mathrm{abs}(Y_{(n,20)}\text{-}X))$$

where X is the binding affinity of the unmutated peptide, and $Y_{(i,j)}$ is the binding affinity of the peptide mutated at position i to amino acid number j (total of 20 possible amino acids to mutate to). All binding affinities were calculated using pVACbind from pVACtools (version 1.5.0)[5] in which the following algorithms were selected: NetMHC[23], NetMHCpan[24], MHCnuggets[25], MHCflurry[26], SMM[27], Pickpocket[28], SMMPMBEC[29], NetMHCcons[30]. The median binding affinity across all 8 of these algorithms was used both to nominate strong binder peptides for the reference dataset, and to assess the impact of in silico mutation at each position of these peptides.

Each position was assigned a score based on how binding was influenced by mutations. These scores were used to calculate the relative contribution of each position to the overall binding affinity of the peptide. Positions that together account for 80% of the overall binding affinity change were assigned as anchor locations for further impact analysis.

Input data for orthogonal evaluation of predicted anchor sites

To evaluate our anchor predictions, we collected 166 protein structures (pdb format) of peptide-MHC complexes and 61 peptide-MHC-TCR complexes from the Protein Data Bank[35] by querying for structures containing macromolecules matching class I HLAs. Structures were additionally reviewed to ensure valid peptide length (8-11) and those with TCRs attached were separated into a different list for downstream analysis to allow accurate solvent-accessible surface area (SASA) calculations. The HLA-peptide structures corresponded to 33 HLA alleles with peptides of varying lengths (8 to 11mer), while the HLA-peptide-TCR structures corresponded to 12 HLA alleles. A complete list of PDB ids selected for this analysis can be found in **Supplemental Table 2**.

Orthogonal validation of predicted anchor sites by analysis of pMHC structures

The structures of peptide-MHC molecules collected were analyzed to infer potential anchor locations/residues. All PDB structures were analyzed in python using the MDTraj package[36]. For each position of a peptide bound to an HLA, we utilized two different metrics: 1) minimum distance of non-backbone atoms to all HLA associated atoms and 2) estimated SASA of the residue. In method 1, we calculated the distances between each atom of each residue and all HLA associated atoms. Non-backbone atoms were ordered by their distance to the closest HLA-associated atom and the top 50% were used to calculate an average distance representing an entire residue (with the exception of glycine where all atoms were considered). In method 2, we directly calculated the SASA of each residue (shrake_rupley function in MDTraj), which was used to infer the likelihood of being able to be recognized by the T-cell receptor.

For an overall evaluation of how well our anchor predictions correlated with these metrics (distance and SASA), spearman correlations were determined. For example, for a 9-mer peptide, a spearman correlation was calculated for the 9 anchor prediction scores from the in silico mutation exercise compared to the 9 distance or SASA estimates obtained from the structure analysis. Out of 166 peptide-MHC structures collected, correlation values for 87 were plotted by randomly selecting at most 5 structures per HLA-length combination (**Supplemental Table 3**). For comparison, we also randomly shuffled distance and SASA scores across all positions of individual peptides and calculated correlation scores against this randomized dataset. The different sets of correlation values were then fit to Gaussian distributions (**Figure 4d**). Two sample t-tests were performed to evaluate the differences among distributions.

Additional analysis was performed on the 61 peptide-MHC-TCR structures collected. After randomly selecting at most 5 structures per HLA-length combination, spearman correlations derived from 31 structures were plotted. Correlations were calculated for 1) distance from peptide to HLA versus anchor prediction scores and 2) distance from peptide to TCR versus anchor prediction scores. Once again, the HLA-peptide distances were randomly shuffled and used as comparison and two sample t-tests were performed to evaluate the differences among distributions.

Input data for evaluating the impact of anchor site considerations

To evaluate how anchor site considerations might influence neoantigen prioritization decisions, we considered two different methods when selecting input data: 1) balanced HLA allele distribution and 2) population-based selection. For method 1, we randomly sampled up to 10 corresponding TCGA samples for each HLA allele with sufficient data (at least 3 out of 4 lengths have 10 or more matching peptides). In method 2, we randomly selected TCGA-HLA combinations from our database matching the total number of samples in method 1. For each method, 923 TCGA-HLA combinations were chosen from a total of 9,216 TCGA samples excluding the 1,356 used for the anchor site prediction data set described above. The 923 TCGA-HLA combinations corresponded to 863 and 853 TCGA patients for methods 1 and 2 respectively (**Supplemental Table 5**). To further evaluate impact of anchor considerations on a patient-specific level, an additional 100 TCGA patients were selected from the original 1,356 TCGA patient samples where we had neoantigen predictions for the patient's full set of HLA alleles (**Supplemental Table 5**).

Evaluating the impact of anchor site consideration on neoantigen prioritization

To analyze the importance of positional information on prioritization of neoantigens, TCGA patient samples were used as input and run through pVACtools (version 1.5.2) using the following options: -e 8,9,10,11, --iedb-retries 50, --downstream-sequence-length 500, --minimum-fold-change 0, --trna-cov 0, --tdna-vaf 0, --trna-vaf 0, --pass-only. The neoantigen candidates were then filtered and prioritized according to different criteria: A) Basic Filter: mutant peptide IC50 < 500 nM and agretopicity > 1, B) Decision based on a conventional anchor assumption that anchors are located at position 2 and the C-terminal position, C) Decision based on computationally predicted allele-specific anchor locations. Filtered lists were then compared for overlap and differences. For our cohort analysis, all neoantigen candidates were considered with no additional filtering. For our patient-level analysis, neoantigen candidates were processed additionally using the top_score_filter ("pVACseq top_score_filter" command of pVACtools) to generate top neoantigen candidates for individual variants. These top candidates were compiled and the same filters A, B, and C were used to determine prioritization decisions. The percentage differences between filters were calculated based on decisions for all top candidates for each individual patient.

HLA coverage and population frequency

Global HLA allele frequencies were generated using data from the Allele Frequency Net Database[31]. The database contains HLA genotype data for Class I alleles across 197 distinct populations. Two populations in the database (“Chile Santiago” and “Russia Karelia”) did not have ambiguity-resolved HLA genotype data and were excluded from this analysis. Global HLA allele frequencies were calculated by (1) aggregating all 195 sample populations, (2) summing HLA allele counts over all sample populations, and (3) dividing HLA allele counts over total counts of the HLA gene across all populations. It should be noted that the HLA frequencies calculated do not reflect true global HLA frequencies since true population/region sizes were not considered. To calculate the percentage of population that our 328 HLA alleles affect, Class I alleles were split into respective subclasses of HLA-A, HLA-B and HLA-C. Global frequencies were summed in each subclass to obtain the percentage of population potentially affected by our HLA allele anchor results.

Acknowledgements

We thank the patients and their families for donation of their samples and participation in clinical trials. Malachi Griffith was supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under Award Number R00HG007940. Malachi Griffith and Obi Griffith were supported by the NIH National Cancer Institute (NCI) under Award Numbers U01CA209936, U01CA231844 and U24CA237719. Malachi Griffith was supported by the V Foundation for Cancer Research under Award Number V2018-007. Malachi Griffith, Obi Griffith, Huiming Xia, William Gillanders and Todd Fehniger were supported by the NCI under Award Number U01CA248235. Suangson Supabphol was supported by Prince Mahidol Award Youth Program of Prince Mahidol Award Foundation under the Royal Patronage of HM the king of Thailand.The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Conflicts of Interest

The authors declare that there is no conflict of interest.

References

1. Robbins, P. F. *et al.* Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* **19**, 747–752 (2013).
2. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).
3. Keskin, D. B. *et al.* Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* **565**, 234–239 (2019).
4. Richters, M. M. *et al.* Best practices for bioinformatic characterization of neoantigens for clinical utility. *Genome Med.* **11**, 56 (2019).
5. Hundal, J. *et al.* pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. *Cancer Immunol Res* **8**, 409–420 (2020).
6. Hundal, J. *et al.* pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* **8**, 11 (2016).
7. Rubinsteyn, A., Hodes, I., Kodysh, J. & Hammerbacher, J. Vaxrank: A computational tool for designing personalized cancer vaccines. doi:10.1101/142919.
8. Bjerregaard, A.-M., Nielsen, M., Hadrup, S. R., Szallasi, Z. & Eklund, A. C. MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol. Immunother.* **66**, 1123–1130 (2017).
9. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* **16**, 15–24 (2018).
10. Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* **18**, 1009–1021 (2017).
11. Yang, W. *et al.* Immunogenic neoantigens derived from gene fusions stimulate T cell responses. *Nat. Med.* **25**, 767–775 (2019).
12. Smart, A. C. *et al.* Intron retention as a novel source of cancer neoantigens. doi:10.1101/309450.

13. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, (2018).
14. Duan, F. *et al.* Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J. Exp. Med.* **211**, 2231–2248 (2014).
15. Cole, D. K. *et al.* Modification of MHC anchor residues generates heteroclitic peptides that alter TCR binding and T cell recognition. *J. Immunol.* **185**, 2600–2610 (2010).
16. Capietto, A.-H. *et al.* Mutation position is an important determinant for predicting cancer neoantigens. *J. Exp. Med.* **217**, (2020).
17. Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell* (2020) doi:10.1016/j.cell.2020.09.015.
18. Zamora, A. E. *et al.* Pediatric patients with acute lymphoblastic leukemia generate abundant and functional neoantigen-specific CD8 T cell responses. *Sci. Transl. Med.* **11**, (2019).
19. Chen, C. *et al.* A Comprehensive Survey of Genomic Alterations in Gastric Cancer Reveals Recurrent Neoantigens as Potential Therapeutic Targets. *Biomed Res. Int.* **2019**, 2183510 (2019).
20. Perumal, D. *et al.* Mutation-derived Neoantigen-specific T-cell Responses in Multiple Myeloma. *Clinical Cancer Research* vol. 26 450–464 (2020).
21. Zhang, J. *et al.* The combination of neoantigen quality and T lymphocyte infiltrates identifies glioblastomas with the longest survival. *Communications Biology* vol. 2 (2019).
22. Fritsch, E. F. *et al.* HLA-binding properties of tumor neoepitopes in humans. *Cancer Immunol Res* **2**, 522–529 (2014).
23. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511–517 (2016).
24. Jurtz, V. *et al.* NetMHCpan 4.0: Improved peptide-MHC class I interaction predictions

integrating eluted ligand and peptide binding affinity data. doi:10.1101/149518.

25. Shao, X. M. *et al.* High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. *Cancer Immunol Res* **8**, 396–408 (2020).

26. O'Donnell, T. J. *et al.* MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst* **7**, 129–132.e4 (2018).

27. Nielsen, M., Lundegaard, C. & Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* **8**, 238 (2007).

28. Zhang, H., Lund, O. & Nielsen, M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* **25**, 1293–1299 (2009).

29. Kim, Y., Sidney, J., Pinilla, C., Sette, A. & Peters, B. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinformatics* **10**, 394 (2009).

30. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64**, 177–186 (2012).

31. Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* **48**, D783–D788 (2020).

32. Yarzabek, B. *et al.* Variations in HLA-B cell surface expression, half-life and extracellular antigen receptivity. *Elife* **7**, (2018).

33. Linette, G. P. & Carreno, B. M. Neoantigen Vaccines Pass the Immunogenicity Test. *Trends in molecular medicine* vol. 23 869–871 (2017).

34. Cotto, K. C. *et al.* RegTools: Integrated analysis of genomic and transcriptomic data for the discovery of splicing variants in cancer. doi:10.1101/436634.

35. Burley, S. K. *et al.* RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464–D474 (2019).

36. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).

Main Figures

Figure 1

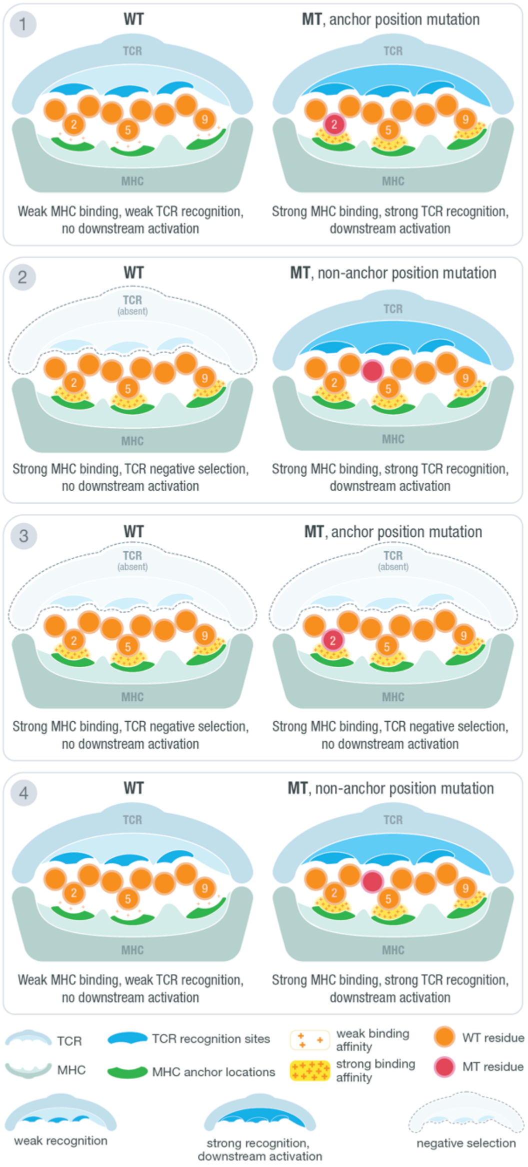
1
WT
MT, anchor position mutation
TCR
MHC
2
5
9
Weak MHC binding, weak TCR recognition, no downstream activation
Strong MHC binding, strong TCR recognition, downstream activation
2
WT
MT, non-anchor position mutation
TCR (absent)
Strong MHC binding, TCR negative selection, no downstream activation
Strong MHC binding, strong TCR recognition, downstream activation
3
WT
MT, anchor position mutation
Strong MHC binding, TCR negative selection, no downstream activation
Strong MHC binding, TCR negative selection, no downstream activation
4
WT
MT, non-anchor position mutation
Weak MHC binding, weak TCR recognition, no downstream activation
Strong MHC binding, strong TCR recognition, downstream activation
TCR
MHC
TCR recognition sites
MHC anchor locations
weak binding affinity
strong binding affinity
WT residue
MT residue
weak recognition
strong recognition, downstream activation
negative selection

Figure 1: Anchor and mutation position scenarios at the MHC-peptide-TCR interface

Illustration of MHC-peptide-TCR interface using an example structure with anchors at position 2, 5 and 9. At the contact interface between the peptide-loaded MHC and the recognizing T-cell receptor, certain positions are responsible for anchoring the peptide to the MHC molecule and/or potentially being recognized by the TCR. The position of tumor specific ("mutant") amino acids relative to anchor positions, and predicted binding affinity of mutant and wild type peptides produce four distinct scenarios for interpreting candidate neoantigens. Example TCR recognition sites are shown in blue while MHC anchor locations are shown in green. The peptide residues are shown in orange while the mutant residue is marked with red. A yellow force field with varying density is used to illustrate binding strength between peptide and the MHC molecule. Three different cases of TCR recognition level are depicted including: self-recognizing TCR absent due to negative selection, weak-recognizing TCR due to weak MHC binding of presented peptide, and strong-recognition of TCR triggering downstream activation of cytotoxic T-cells.

Figure 2

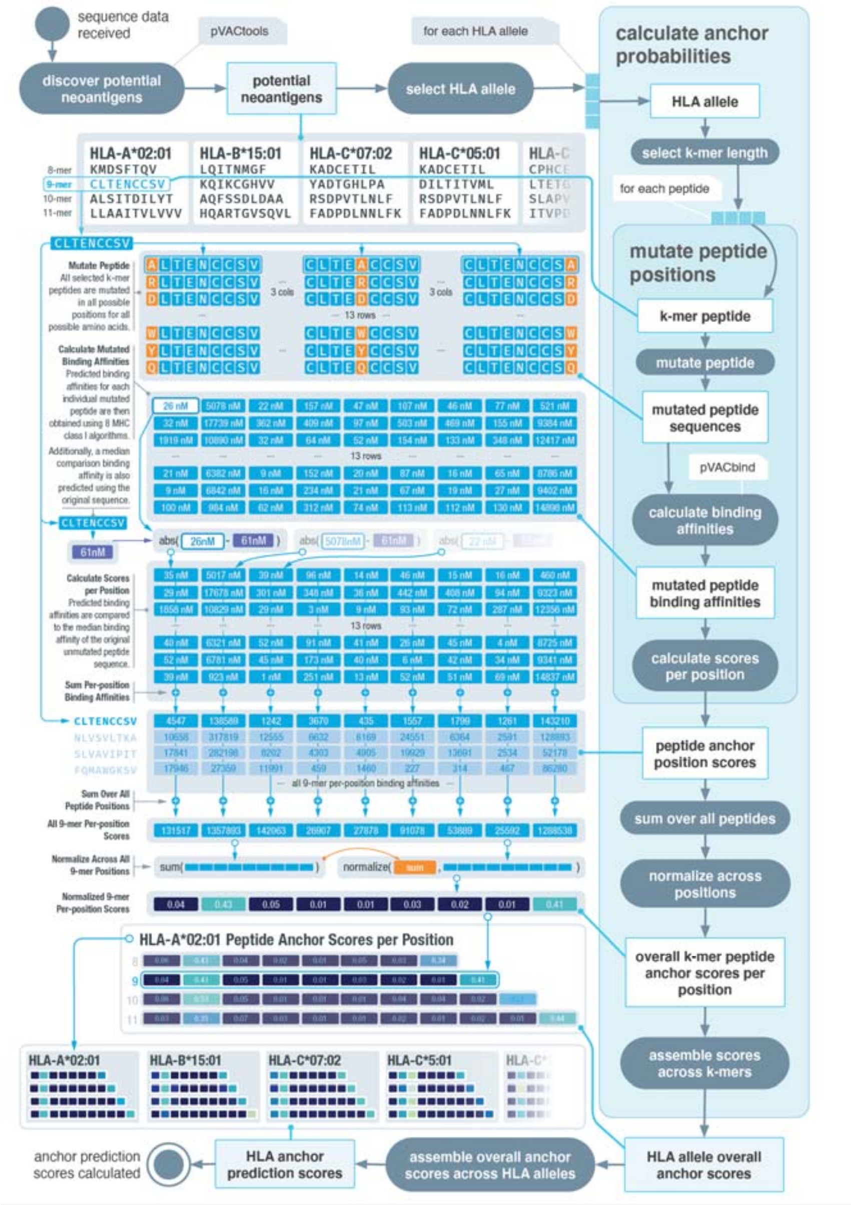
sequence data received
pVACtools
for each HLA allele
calculate anchor probabilities
discover potential neoantigens
potential neoantigens
select HLA allele
HLA allele
select k-mer length
for each peptide
HLA-A*02:01
HLA-B*15:01
HLA-C*07:02
HLA-C*05:01
8-mer
9-mer
10-mer
11-mer
KMDSFTQV
CLTENCCSV
ALSITDILYT
LLAAITVLVVV
LQITNMGF
KQIKCGHVV
AQFSSDLDAA
HQARTGVSQVL
KADCETIL
YADTGHLPA
RSDPVTLNLF
FADPDLNNLFK
KADCETIL
DILTITVML
RSDPVTLNLF
FADPDLNNLFK
mutate peptide positions
Mutate Peptide
All selected k-mer peptides are mutated in all possible positions for all possible amino acids.
3 cols
13 rows
k-mer peptide
mutate peptide
Calculate Mutated Binding Affinities
Predicted binding affinities for each individual mutated peptide are then obtained using 8 MHC class I algorithms.
Additionally, a median comparison binding affinity is also predicted using the original sequence.
mutated peptide sequences
pVACbind
calculate binding affinities
61nM
abs(26nM - 61nM)
Calculate Scores per Position
Predicted binding affinities are compared to the median binding affinity of the original unmutated peptide sequence.
mutated peptide binding affinities
calculate scores per position
Sum Per-position Binding Affinities
NLVSVLTKA
SLVAVIPIT
FQMANGKSV
peptide anchor position scores
all 9-mer per-position binding affinities
Sum Over All Peptide Positions
sum over all peptides
All 9-mer Per-position Scores
Normalize Across All 9-mer Positions
sum(
normalize(
sum
normalize across positions
Normalized 9-mer Per-position Scores
0.04
0.43
0.05
0.01
0.01
0.03
0.02
0.01
0.41
HLA-A*02:01 Peptide Anchor Scores per Position
overall k-mer peptide anchor scores per position
assemble scores across k-mers
HLA-A*02:01
HLA-B*15:01
HLA-C*07:02
HLA-C*5:01
anchor prediction scores calculated
HLA anchor prediction scores
assemble overall anchor scores across HLA alleles
HLA allele overall anchor scores

Figure 2: Overview of simulation workflow for anchor prediction

Schematic of our computational workflow to simulate the effect of mutation position on binding affinities of the peptides. HLA and peptide pairings are selected from a reference dataset of putative strong binders. All possible amino acid changes are applied to all possible positions and impact on binding affinity is assessed. An overall peptide anchor score is calculated for each position for all HLA-peptide length combinations. Higher scores indicate greater likelihood that a particular position in the peptide acts as an anchor residue for a given HLA.

Figure 3

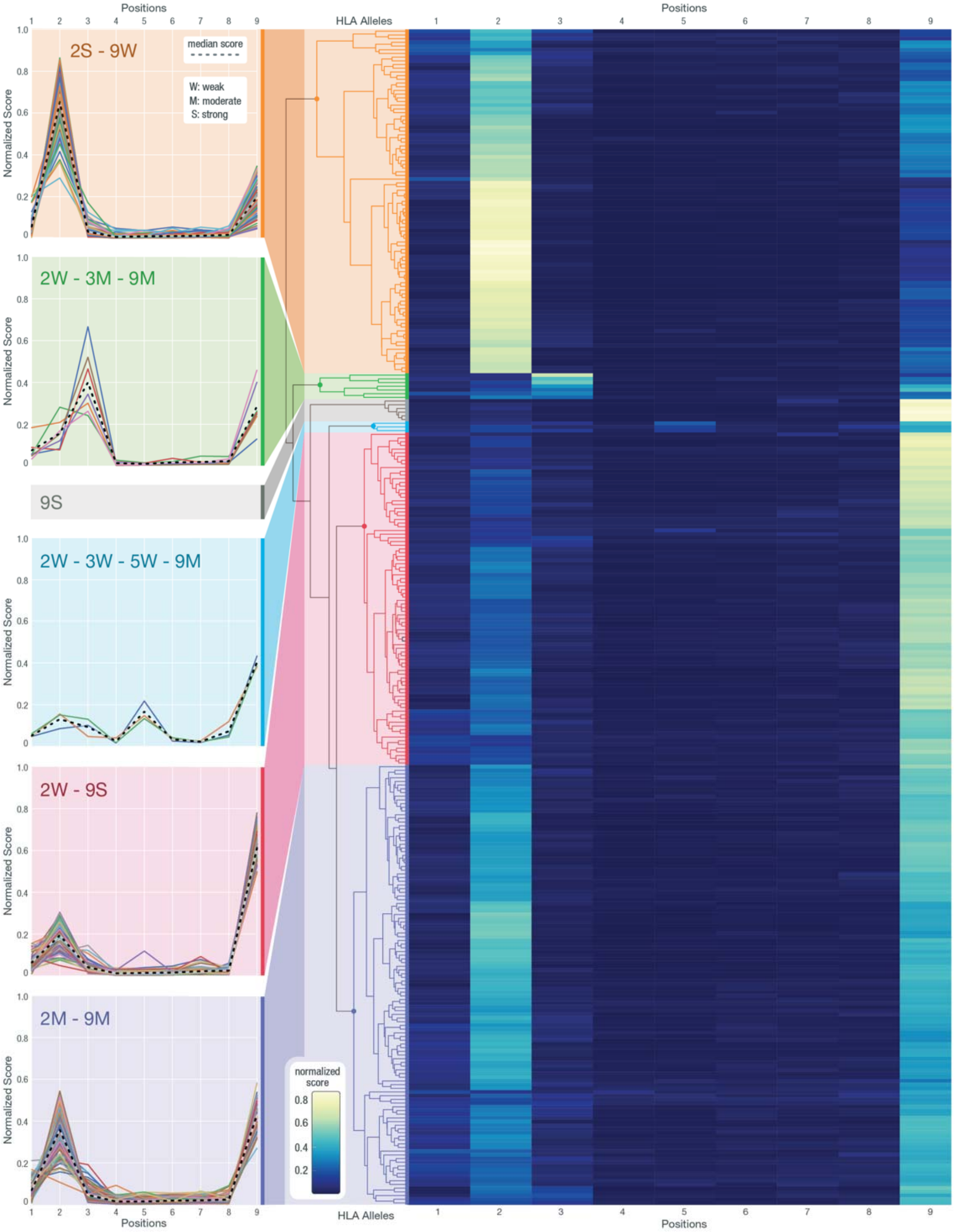

Figure 3: Hierarchical clustering of anchor prediction scores across all 9-mer peptides

Anchor prediction scores clustered using hierarchical clustering with average linkage for all 318 HLA alleles for which 9-mer peptide data were collected. For the heatmap, the x-axis represents the 9 peptide positions and the y-axis represents 318 HLA alleles. Example HLA clusters have been highlighted with various color bands and the score trends for individual HLA alleles are plotted for each. In the cluster line plots on the left, the x-axis shows the peptide positions while the y-axis corresponds to the anchor score, normalized across all peptide positions. Different annotations have been given to help summarize the trends observed in individual clusters, where numbers represent positions and letters represent its strength as a potential anchor in comparison to other anchors (S: strong, M: moderate, W: weak). The median scores for each cluster are presented with a dashed line.

Figure 4

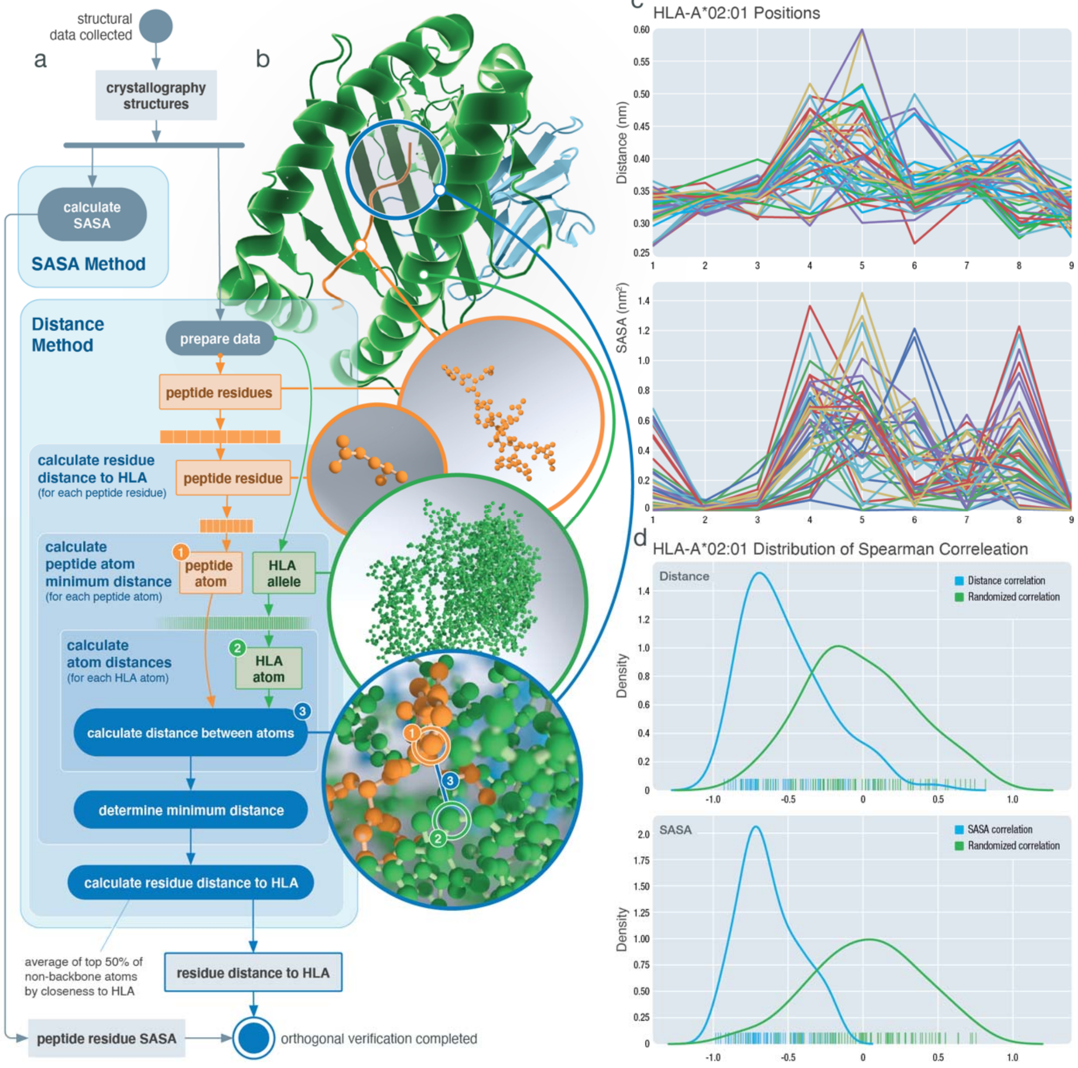

Figure 4: Orthogonal validation using protein crystallography structures
Orthogonal validation of predicted anchor scores utilizing X-ray crystallography structures. **a**, Schematic of analysis workflow for each HLA-peptide structure collected. For the distance metric, backbone atoms were excluded with the exception of glycine. **b**, Structural example of HLA-B*08:01 bound to peptide FLRGRAYGL (PDB ID: 3X13). **c**, Example results of 47 structures collected for HLA-A*02:01 with 9-mer peptides. Top panel corresponds to distance measurements for each position while the bottom panel corresponds to SASA measurements. **d**, Distribution of spearman correlations calculated between distance and prediction scores (top) and SASA and prediction scores (bottom). Blue line represents each respective correlation distribution while the green line shows the distribution of spearman correlation values obtained from randomly shuffled peptide positions.

Figure 5

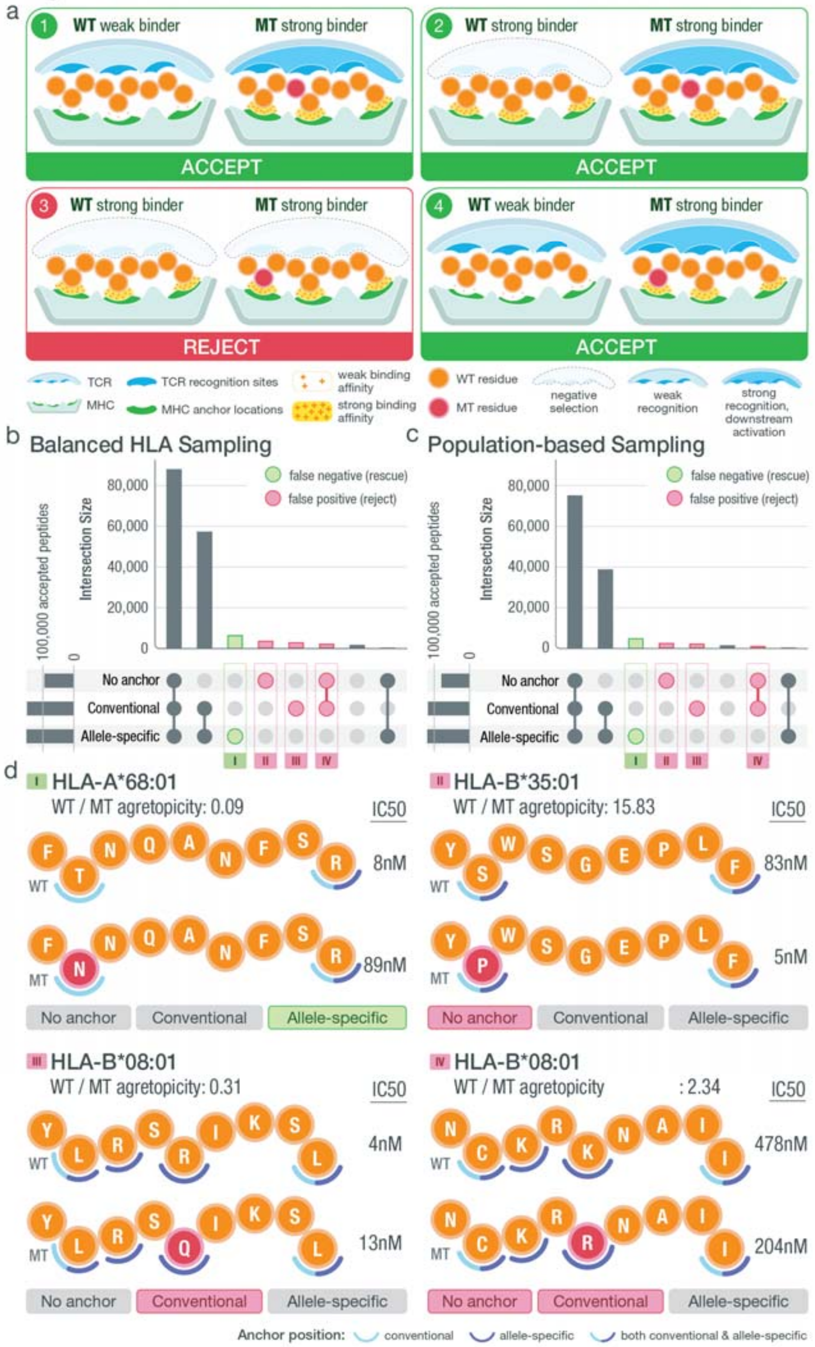

Figure 5: Impact of anchor position information on neoantigen prioritization decisions

a, Illustration of different scenarios that could be encountered when prioritizing neoantigens. Each circle represents a peptide residue with mutated residues marked in red. Anchor locations of the MHC are marked in green while TCR recognition sites are in blue. Predicted binding affinities of the MT/WT peptides are indicated using a yellow density field where higher density represents strong binding, and lower density represents weak binding. Three different scenarios of T-cell recognition are depicted. **b,** Upset plot showing number of intersecting peptides based on those prioritized with no anchor filter (binding affinity < 500nM and agretopicity > 1), conventional filter (filtering based on conventional anchor assumptions) or allele-specific filter (filtering based on computationally predicted anchor locations). Samples included for analysis were chosen such that HLA alleles were balanced appropriately. Peptides characterized differently between no anchor/conventional filter and allele-specific filter were categorized into false negatives (green circle) and false positives (red circle) with the assumption that the allele-specific filter produced more accurate results. **c,** Upset plot showing number of intersecting peptides based on those prioritized with no anchor filter, conventional filter or allele-specific filter. In contrast to the previous panel, samples included for this analysis were chosen by randomly sampling a large pool of TCGA samples. **d,** Examples of false positive and false negative peptides from each of the four subsets as marked in panel c. Matching HLA allele, peptide sequence, mutation position (red), median WT/MT IC50 values and fold changes are shown accordingly. Two sets of anchor locations are depicted for each scenario using semi-circles: conventional anchors are marked with light blue and allele-specific anchors are marked with dark blue. Positions where the two sets of anchors overlap are marked with split coloring of the semi-circle.