Computational prediction of MHC anchor locations guides neoantigen identification and prioritization

Huiming Xia[1,2], **Joshua McMichael**[2], **Michelle Becker-Hapak**[1], **Onyinyechi C. Onyeador**[1], **Rico Buchli**[3], **Ethan McClain**[1], **Patrick Pence**[1], **Suangson Supabphol**[4,5], **Megan M. Richters**[1,2], **Anamika Basu**[2], **Cody A. Ramirez**[1,2], **Cristina Puig-Saus**[6,7,8], **Kelsy C. Cotto**[1,2], **Sharon L. Freshour**[1,2], **Jasreet Hundal**[1,2], **Susanna Kiwala**[2], **S. Peter Goedegebuure**[4,9], **Tanner M. Johanns**[1], **Gavin P. Dunn**[10], **Antoni Ribas**[6,7,8], **Christopher A. Miller**[1,9], **William E. Gillanders**[4,9], **Todd A. Fehniger**[1], **Obi L. Griffith**[1,2,9,11,*], **Malachi Griffith**[1,2,9,11,*]

[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]Pure Protein LLC, Oklahoma City, OK 73104, USA.

[4]Department of Surgery, Washington University School of Medicine, St. Louis, MO, USA.

[5]The Center of Excellence in Systems Biology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.

[6]Division of Hematology/Oncology, Department of Medicine, University of California, Los Angeles, Los Angeles, CA, USA.

[7]Jonsson Comprehensive Cancer Center, Los Angeles, CA, USA.

[8]Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA.

[9]Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO, USA.

[10]Department of Neurosurgery, Washington University School of Medicine, St. Louis, MO, USA.

[11]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA.

Abstract

Neoantigens are tumor-specific peptide sequences resulting from sources such as somatic DNA mutations. Upon loading onto major histocompatibility complex (MHC) molecules, they can trigger recognition by T cells. Accurate neoantigen identification is thus critical for both designing cancer vaccines and predicting response to immunotherapies. Neoantigen identification and prioritization relies on correctly predicting whether the presenting peptide sequence can

*Corresponding author. mgriffit@wustl.edu (M.G.); obigriffith@wustl.edu (O.L.G.).

successfully induce an immune response. Because most somatic mutations are single-nucleotide variants, changes between wild-type and mutated peptides are typically subtle and require cautious interpretation. A potentially underappreciated variable in neoantigen prediction pipelines is the mutation position within the peptide relative to its anchor positions for the patient's specific MHC molecules. Whereas a subset of peptide positions are 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 anchor positions for different peptide lengths for 328 common HLA alleles and identified unique anchoring patterns among them. Analysis of 923 tumor samples shows that 6 to 38% of neoantigen candidates are potentially misclassified and can be rescued using allele-specific knowledge of anchor positions. A subset of anchor results were orthogonally validated using protein crystallography structures. Representative anchor trends were experimentally validated using peptide-MHC stability assays and competition binding assays. By incorporating our anchor prediction results into neoantigen prediction pipelines, we hope to formalize, streamline, and improve the identification process for relevant clinical studies.

INTRODUCTION

Neoantigens arise from short peptide sequences specifically found in tumor cell populations resulting from sources such as somatic mutations, RNA editing (1), alternative splicing (2), etc. 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 antitumor response. Multiple studies have shown the potential of neoantigen-based immunotherapy treatments for cancer (3–5), and numerous clinical trials are underway. Accurate neoantigen prediction and prioritization is critical to understanding tumor immunology and response to checkpoint blockade therapy and for the design of personalized vaccines (6) and T cell therapies. Several bioinformatic tools and pipelines have been developed to facilitate neoantigen identification (7–10).

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). Whereas a variety of mutation types are being explored as neoantigen sources (11–15), the vast majority of somatic mutations currently identified as sources of neoantigens are single-nucleotide variants (SNVs) (4, 16), although emerging research suggests that we are underestimating the potential for neoantigen generation from more complex forms of variation (17–20). Amino acid sequence changes between the wild-type (WT) and mutant (MT) peptides are subtle (often a single–amino acid substitution), and MT peptides remain similar to native sequences of the host. In addition, only a subset of positions on the loaded peptide sequence are primarily presented to the T cell receptor (TCR) for recognition, and another subset of positions are primarily responsible for anchoring to the MHC, making these positional considerations critical for predicting T cell responses (Fig. 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 (21). Examples of the four distinct possible scenarios for a predicted strong MHC binding peptide involving these factors are illustrated in Fig. 1. There are other possible scenarios where the MT is a poor binder; however, those are not listed because they would not pertain to our goal of neoantigen identification. 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. Because the WT does not bind (or is a poor binder), this neoantigen remains a good candidate because 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 with 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 the third scenario, 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 (22), a mutation at a strong anchor location generally implies that both WT and MT peptides will present the same residues for TCR recognition. Because the WT peptide is a strong binder, the MT neoantigen, although 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 have confirmed the importance of anchor location when predicting the overall immunogenicity of a given peptide (23, 24). However, it is important to note that these scenarios are not absolute. Whereas a position that acts as a strong anchor is less likely to be TCR facing (and vice versa), certain positions of the peptide may interact to varying degrees with both the MHC and TCR.

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 prediction pipelines. However, failing to account for these positional considerations may result in susceptibility to central tolerance and potentially induce autoimmunity. Many recently published neoantigen studies have used simple filtering strategies with either only binding affinity filters [e.g., MT peptide half-maximal inhibitory concentrations (IC_{50}) < 500 nM] (5, 25) or with an additional agretopicity filter (26–28), 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 and WT binding affinities) (29). 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 have not been reported. As a result, many neoantigen studies have either failed to adequately consider this crucial factor or have used simplified and arbitrary 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 seed dataset generated from a collection of patient samples from local tumor sequence studies (30) combined with samples from the Cancer Genome Atlas (TCGA). 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. Representative anchor trends were experimentally validated using cell-based stability assays and IC_{50} binding assays. Using additional TCGA samples, we further evaluated how prioritization results may change when provided with additional anchor information, emulating a subset of steps in the neoantigen selection process by an immunotherapy tumor board, tasked with prioritizing vaccine candidates. By sharing our results for incorporation into neoantigen prediction pipelines, we hope to improve neoantigen prioritization for relevant experimental and clinical studies.

RESULTS

Computational and quantitative prediction of HLA-specific anchor positions

To predict anchor locations for a wide range of HLA alleles, we assembled a seed HLA-peptide dataset of strong binding peptides with a median predicted IC_{50} of less than 500 nM across (up to) eight MHC class I algorithms [NetMHC (31), NetMHCpan (32), MHCnuggets (33), MHCflurry (34), SMM (35), Pickpocket (36), SMMPMBEC (37), NetMHCcons (38)]. These peptides were obtained from TCGA and supplemented with additional patient datasets from our own neoantigen study cohorts including lymphoma, glioblastoma (30), breast cancer, and melanoma (Materials and Methods and data file S1). A total of 609,807 peptides were identified, with most being 9-mers and 10-mers (fig. S1). In total, these peptides corresponded to 328 HLA alleles across 1443 tumor samples.

For each individual HLA allele of which data were obtained, peptides were sorted and separated by their respective lengths, ranging from 8 to 11 (Fig. 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 with 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 substantial 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 (Fig. 2, fig. S2, and Materials and Methods).

Prediction results show distinct patterns of HLA anchor locations

We generated anchor prediction scores for the 328 HLA alleles with strong binding peptides in our seed dataset (data file S2). These HLA alleles include 95 HLA-A alleles (representing 99.2% of the HLA-A alleles observed in the population according to the Allele Frequency Net Database), 175 HLA-B alleles (representing 97.9% of HLA-B alleles), and 58 HLA-C

alleles (representing 98.5% of HLA-C alleles; Materials and Methods) (39). Results were separated on the basis of peptide lengths (8–11), and the anchor prediction scores across all HLA alleles were visualized using hierarchical clustering with average linkage (Fig. 3 and fig. S3). We observed different anchor patterns across HLA alleles, varying in both the number of anchor positions and the location. These anchor position patterns could be roughly clustered into six distinct groups. In addition, whereas we used peptides generated from cancer patient data, we wanted to see whether changing the source of the peptides would affect our predicted anchor patterns (fig. S4 and Materials and Methods). After averaging calculations over 1000 peptides from each source, we observed high convergence of the anchor pattern for HLA-A*02:01 across all four different sources (random, reference proteome, viral proteome, and cancer mutations). This demonstrated the possibility of expanding results to a wider span of HLA alleles in the future by using different sources of peptides.

Previously, anchor locations have generally been assumed to be at the second and terminal positions of the peptide with equal weighting (with the notable exception of HLA-B*08:01) (40). Our 9-mer clustering results confirm that most of HLA alleles predicted show positions 2 and 9 as likely anchor locations (95% fall into the 2S-9W, 2W-9S, and 2M-9M clusters combined). 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; Fig. 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; Fig. 3). In addition, we observe a smaller cluster of HLA alleles with moderate anchor predictions for both positions (2M-9M; Fig. 3) and another cluster with strong anchor predictions for only position 9 (9S; Fig. 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; Fig. 3). A smaller group of HLA-B alleles also show an additional anchor at position 5 (2W-3W-5W-9M; Fig. 3). Our results indicate that a conventional anchor assumption putting equal weights on positions 2 and 9 does not capture the notable heterogeneity in anchor usage between different HLA alleles. These anchor considerations can affect neoantigen prioritization decisions, and we hypothesized that HLA allele–specific anchor predictions would allow ranking of neoantigens with greater accuracy.

To analyze the possibility of our clusters arising because of bias from prediction algorithms and their training datasets, we examined the training data used by NetMHCpan4.0 for the 328 HLA alleles that we have included (data file S3). In addition, for alleles with limited data (hence predicted through a pan-allele strategy), the ability to effectively use information from other alleles (for those lacking training data) also depends on how similar the allele of interest is to other alleles that do have robust amounts of data. Thus, we also looked at the neighboring HLA alleles that are used in cases where training data were not available and the distance (calculated by NetMHCpan4.0) between each HLA allele and its nearest neighbor. Overall, most of the HLA alleles (99.7%) had binding predictions from at least four different algorithms, with 54 HLA alleles being supported by all eight (fig. S5A). The amount of training data for each HLA allele also correlated with how consistent binding predictions were across algorithms (fig. S5B). This analysis confirms our expectation that

more training data lead to more consistent predictions across algorithms. For the smaller clusters such as 2W-3W-5W-9M (blue), 2W-3W-9M (green), and 9S (gray), we can see that, in general, they fall along the same level as other clusters when comparing their nearest neighbor distances and how much variation is seen across prediction algorithms (fig. S5, C and D). In addition, two of the three alleles in the 2W-3W-5W-9M cluster, four of seven alleles in the 2W-3W-9M, cluster and six of six alleles in the 9S cluster all have some amount of training data available (fig. S5E and data file S3). When looking at network graphs showing HLA alleles and their neighbors, those with 2W-3W-5W-9M (blue) and 2W-3W-9M (green) patterns are scattered among different clusters with some alleles even acting as the center node (indicating nonzero amount of training data and, thus, no need to estimate binding based on similar alleles; fig. S5, F and G). Thus, we concluded that the less common anchor clusters do not correlate with lack of training data or similarity to a limited set of nearest neighbors. Hence, we excluded the possibility of clusters arising because of bias in algorithms and their training data and believe that allele-specific differences are the true underlying reason for our observations.

Protein structural analysis confirms predicted anchor results

To validate our anchor predictions, we collected x-ray crystallography structures for MHC molecules with bound peptides (data file S4 and movie S1). The 166 protein structures collected corresponded to 33 HLA alleles, with most of them containing 9-mer peptides (8-mers, 6; 9-mers, 110; 10-mers, 39; and 11-mers, 11). These structures were analyzed using two methods: (i) measuring the physical distance between the peptide and the MHC binding groove and (ii) calculating the solvent-accessible surface area (SASA) of the peptide residues (Fig. 4, A and B, and Materials and 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 (i) more likely to be closer to the HLA molecule and (ii) more secluded from solvent surrounding the peptide-MHC complex compared with 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. Thus, we expect an inverse correlation between our anchor prediction scores and the distance/SASA metrics. 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 (Fig. 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, we observed 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. We also performed a targeted analysis on HLA-B*08:01 (2W-3W-5W-9M; blue cluster) with the limited data available and observed that positions 6 and 7 consistently bulged out, whereas other positions tended to be closer to the HLA molecule while also being secluded from solvent (fig. S6).

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 across 87 Protein Data Bank (PDB) structures. These structures were determined by randomly selecting at most five structures per HLA-length combination. The distribution of these correlations was compared with that of a randomized dataset where positions of the peptide were randomly shuffled (Fig. 4D, data file S5, and Materials and Methods). Two sample t tests showed that the distributions were significantly different from the randomized dataset, with statistical values of $-9.9795 \left(P = 1.3757 \times 10^{-18} \right)$ and $-14.7322 \left(P = 8.7472 \times 10^{-30} \right)$ for distance and SASA metrics, respectively. Results showed that 91.95% of our prediction scores were inversely correlated with the distance metric and that 100% of them were 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 (fig. S7). Two-sample t tests showed significant differences between the randomized dataset and both the HLA-peptide distance $\left(P = 8.8023 \times 10^{-8} \right)$ and the TCR-peptide distance $\left(P = 2.4915 \times 10^{-13} \right)$. These results together strongly suggest that our anchor prediction workflow is accurately predicting allelespecific anchor sites.

Experimental validation shows similar anchor trends as our prediction across distinct HLA alleles

To further validate the predicted anchor patterns that we observed, we selected a range of HLA alleles representing these patterns for experimental validation. We used two different experimental validation methods across our peptide-HLA combinations: in vitro IC_{50} binding assays and cell-based stabilization assays (Materials and Methods). Validations were performed on a total of 136 peptide-HLA combinations across eight different HLA alleles selected to represent varying anchor patterns. For each HLA allele selected, we first validated peptides that were predicted to be strong binders. These peptides were originally MT peptides from either clinical datasets or TCGA samples paired with matching HLA alleles from the patient. For a validated strong binding peptide-HLA combination, we then synthesized peptides mutated in multiple ways at different positions, which each had a varying predicted score of acting as an anchor. These mutated peptides were then evaluated experimentally for MHC binding and/or stability and compared with the original strong-binding peptide (data file S6). For example, we performed an in-depth analysis for HLA-B*07:02 with a strong binding 9-mer peptide RPDVKHSKM. Overall, for 9-mer peptides, HLA-B*07:02 was predicted to have a high anchor score at position 2 and a lower anchor score at position 9 (2S-9W; Fig. 5A). After our computational prediction workflow, we also calculated the specific anchor trend for peptide RPDVKHSKM, which agrees with the overall 9-mer trend with positions 2, 9, and 1 as anchors in descending scores (Fig. 5B). We mutated positions 1, 2, 5, and 9 to four different amino acids, each with varying predicted influence on the binding affinity, and measured their stability with regard to HLA-B*07:02 (Materials and Methods and Fig. 5C). For non-anchor position P5, all mutated peptides remained strong binders. For the two weak anchor positions P1 and P9, certain amino acid changes disturbed the binding, whereas others had little influence. For the strong anchor position P2, all amino acid changes led to a nonbinder. We further performed IC_{50} binding assays and observed a similar trend to that of the stability experiments (Fig.

5D). Similar mutation experiments on HLA-B*08:01, HLA-A*68:01, and HLA-A*23:01 were conducted (figs. S8 and S9). These additional experiments confirmed the varying strengths of individual positions acting as anchors for MHC alleles comprising anchor patterns 2W-3W5W-9M, 2W-9S, and 9S.

Overall, across 136 peptide-HLA combinations, experimental validation results were consistent with the average predicted binding affinities calculated from eight prediction algorithms. For peptides validated using IC_{50} binding assays, reasonable correlations were observed between the predicted and measured IC_{50} values, with respect to individual HLA alleles (Fig. 6A). When comparing the measured binding category (determined by the measured IC_{50}) with predicted binding affinities, we saw a trend with high-affinity binders having the lowest predicted IC_{50} values (Fig. 6B). A few outliers were observed where peptides were predicted as strong binders, but when validating, these peptides were categorized as nonbinding. This may be explained by difficulties in manufacturing and/or solubilizing these peptides, preventing detection in our validation experiments. As expected, an inverse correlation was observed between the mean fluorescence intensity (MFI) average values and the predicted binding affinities (Fig. 6C). Similarly, peptides with measured high affinities showed the highest MFI values. Peptides with measured binding affinities among the "medium," "low," "very low," and "no binder" categories were similar in their range of MFI values, indicating a difference in the sensitivity between the IC_{50} binding assays and cell stabilization experiments (Fig. 6D and Materials and Methods).

We further examined how each prediction algorithm performed when compared with the measured IC_{50} binding values (data file S7 and fig. S10). Of the eight algorithms used, HLA-A*02:01 had the highest correlation scores for six of the eight algorithms, consistent with our observation that it has the largest training dataset (data file S3) and thus was expected to yield the best accuracy. Overall, NetMHCpan came out on top, showing the highest correlation values in all four HLA alleles where sufficient data were available (fig. S11). These experimental validation results combined help show the distinct anchor trends that exist among HLA alleles.

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 researchers lack specific tools and databases to make use of such information. Whereas the decision of whether a neoantigen should be prioritized over others involves many aspects not discussed in detail 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. Interpretation of a strong binding MT peptide candidate may depend on other factors such as WT binding affinity, agretopicity, mutation position, and anchor location(s), leading to different choices when prioritizing neoantigens (Fig. 7A). If the mutation is not at an anchor location, regardless of the WT peptide binding affinity, then 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 (Fig. 7A, scenarios 1 and 2). In addition, if the WT peptide is a weak binder, then the MT peptide should be accepted regardless of whether the mutation

is at an anchor location because both the MT and WT sequences have not previously been exposed to the immune system and therefore are not subject to tolerance (Fig. 7A, 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 (Fig. 7A, scenario 3). These scenarios were considered when performing the following 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. When selecting patient-HLA paired samples, we chose from a balanced HLA allele distribution (Materials and Methods). Our intent was to give a more balanced view of the impact across all HLA alleles without overt bias for the most common alleles. We first examined the proportion of SNV-induced neoantigens that went into each of the four scenarios described in Fig. 7A and found that 17, 57, 8, and 18% of all peptide candidates fell into scenarios 1, 2, 3, and 4, respectively (fig. S12). All potential neoantigens were filtered according to three different criteria: (i) MT IC_{50} < 500 nM and agretopicity > 1 ("no anchor filter"), (ii) supplementing this with a conventional anchor assumption ("conventional filter"), or (iii) using our computationally predicted anchor locations ("allele-specific filter") (Materials and Methods). Peptide results showed that under the no anchor filter, 57.9% of neoantigens were accepted compared with 93.2% under the conventional filter and 94.8% under the allele-specific filter, showing an overall net gain in the number of peptides when taking anchor considerations into account. When comparing filtered datasets under different criteria, 38.3% of neoantigens were potentially misclassified using the no anchor filter, and about 5.7% of candidates were potentially misclassified between the conventional filter and the allele-specific filter (Fig. 7B). These misclassifications involve 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). We repeated this analysis to see whether our observations were consistent when (i) using different binding affinity cutoffs for strong-binding peptides (<100 nM and <50 nM) and (ii) looking exclusively at peptide-MHC pairings that have a nonconventional anchor pattern (Fig. 7C and Materials and Methods). Our results showed that when comparing results between the no anchor (A) and any anchor considerations (B or C), differences of accepted candidates were similar across the four criteria (<500 nM, <100 nM, <50 nM, and HLA filter). When comparing results between the allele-specific considerations (C) and the more naive approaches (A or B), we saw an increase in percentage of difference with our HLA allele–filtered dataset, indicating a greater impact on neoantigen prioritization when using our allele-specific anchor method. This was expected because peptides analyzed in this dataset corresponded to HLA alleles with nonconventional anchor patterns. We also saw a decrease in percentage when applying a binding affinity cutoff of 50 nM, which we believe is in part due to fewer HLA alleles passing the stricter binding cutoffs (81 versus 98 unique HLA alleles found in datasets with cutoffs of 50 and 500 nM, respectively). 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 positives and false negatives (Fig. 7B). Examples of each scenario were pulled from our dataset to show how peptides passed or failed individual filters (Fig. 7D).

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 six; Materials and Methods, data file S8,and fig. S13, A and B). The neoantigen candidates were prioritized using the same set of criteria applied in the previous cohort analysis. We observed a substantial impact on neoantigen prioritization results depending on the chosen filtering criteria. Specifically, between the no anchor filter and the allelespecific filter, 99% of the patients analyzed had at least one neoantigen decision changed, with a median of 11 peptides with altered decisions per patient. 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; fig. S13, C to E). These results show the potential widespread effect of anchor considerations on patient-level prioritization results.

DISCUSSION

We developed a computational workflow for predicting anchor position scores 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. We further experimentally validated HLA allele anchor patterns using binding assays and cell-based stabilization assays. The underlying quantitative scores from our anchor prediction workflow are available for incorporation into neoantigen prediction workflows, and we believe that this will improve their performance in predicting immunogenic tumor-specific peptides. For simplicity, our illustrations have depicted peptide residues as either anchoring or potentially participating in TCR recognition, although previous research has shown that heteroclitic peptides can alter both simultaneously (22). Hence, anchor residues and TCR recognition sites should not be considered mutually exclusive and should ideally be interpreted quantitatively, where the anchor scores (provided in data file S2) 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 notable impact on neoantigen prioritization. In this study, the choices of whether to accept or reject a prioritization decision were based on hard cutoffs determined using an objective strategy across all HLA alleles. However, when making clinical decisions, we recommend using the actual anchor scores as guidance when prioritizing candidates. In most neoantigen characterization workflows, numerous other factors are taken into account to arrive at an overall prioritization decision, which may further increase differences between filtering strategies. In addition, although our anchor results represent overall averaged scores across multiple neoantigens, slight variations of patterns do exist among different peptides for the same HLA allele. It would generally be computationally prohibitive to perform our workflow on a large scale for all possible neoantigen candidates. However, as a compromise, one could repeat our detailed process to generate peptide-specific anchor predictions after performing other filtering strategies (e.g., genomic information including variant allele frequency (VAF) and expression) and arriving at a shortlist of candidates.

Anchor results not only affect 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 with 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 because of low tumor mutational burden. However, we also acknowledge that although our approach improves sensitivity, it could come at the cost of slightly decreased specificity for patients with high neoantigen burden.

The neoantigen selection process requires careful consideration of numerous aspects, which have been discussed extensively (6). In general, neoantigen-based vaccines act by stimulating the patient's immune system for the production of activated cytotoxic T cells. However, compared with viral antigens where the protein sequence is entirely foreign, neoantigens, particularly those developed from SNVs, have merely subtle differences from the individual's WT proteome. Thus, the need for a WT versus MT peptide comparison, while considering anchor locations, is an aspect specific to tumor neoantigens that other vaccine development pipelines could generally ignore. Although 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 toward the beginning or end of a neoantigen may still cause anchor disruption in an allele-specific manner. In addition, more work should be done to characterize the similarity of neoantigen candidates and the patient's WT proteome for an overall accurate prioritization process.

Recent work, such as that from the Tumor Neoantigen Selection Alliance (24), has hinted at the potential importance of anchor locations. In that study, researchers 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. Possible explanations could include the fact that neoantigens with the MT residue at a strong anchor position have a disadvantage over those present at TCR sites because they require their WT counterpart to be a poor binder, and the threshold for determining sufficiently weak binding of a peptide is unclear. Although a 500 nM cutoff is widely used, it is not universally accepted, and allele-specific binding thresholds continue to expand for a larger group of HLA alleles. However, further investigation is required to address questions raised by such observations.

In addition to the limitations of being applicable to a subset of neoantigens derived from SNVs and certain indels, our work could be expanded to a wider range of HLA alleles. A larger HLA-peptide seed dataset could be achieved through a wide-scale prediction of strong binders for rare HLA alleles by mutating the WT proteome. There have also

been publications noting the sequence motifs in HLA class II binding peptides (41), as well as others using molecular docking approaches (42) in an attempt to estimate anchor positions for HLA class II alleles. However, class II alleles differ substantially from class I because of the difference in length of binding peptides and the binding pocket being more open and composed of a protein dimer. Additional data generation and further research will be required for the expansion of our workflow to address class II, although we hypothesize that the same principles should be applicable. Furthermore, although 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, the results published show a low accuracy for current neoantigen prediction pipelines (43). By accounting for additional positional information, we hope to substantially reduce the numberof false-positive candidates and rescue false-negative neoantigens to increase prediction accuracy. A prioritization strategy using anchor results has been incorporated into the visual reporting of our neoantigen identification pipeline pVACseq (7).

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 scores may serve as an additional feature in machine learning model training on clinical data. This will allow for a more nuanced approach where anchor scores may be weighted for each allele-peptide pairing accordingly. These results and tools will help streamline the prioritization of candidates for neoantigen vaccines and may aid in the design of more effective cancer vaccines.

MATERIALS AND METHODS

Study design

The objective of this study was to predict allele-specific MHC anchor locations and investigate whether the use of this information can substantially influence neoantigen identification and prioritization. We established a computational approach based on an ensemble of MHC binding prediction algorithms and exhaustive in silico mutation of peptides. By application of this approach to 609,807 distinct peptide sequences, we were able to predict anchor locations for 328 common HLA alleles with peptides of varying lengths (8- to 11-mers). Anchor location scores were clustered using hierarchical clustering, and representative trends were identified and experimentally validated using various methods, including x-ray crystal-lographystructures, IC_{50} competition binding assays, and cell-based stabilization assays. All stabilization assays were completed two or three times in duplicate ($N = 4$ to 6). Use of allele-specific anchor predictions was compared against conventional prioritization methods to demonstrate the potential influence on neoantigen prioritization results.

Input data for identifying strong binding seed 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 brain tumor samples, and 6 melanoma samples. In addition, we mined data from 9216 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, 1356 TCGA samples were used to generate seed 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 two of four variant callers as previously described (44). Peptide lengths considered ranged from 8- to 11-mers. All selected data samples were run through pVACseq with the following options with all available class I prediction algorithms: -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. In total, these datasets corresponded to 1443 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 (fig. S1 and data file S1).

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 seed 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 more than 1500 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 (fig. S4). 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 (by mutation to all 19 other amino acid identities) will affect the overall binding affinity

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

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) (7), in which the following algorithms were selected: NetMHC (31), NetMHCpan (32), MHCnuggets (33), MHCflurry (34), SMM (35), Pickpocket (36), SMMPMBEC (37), and NetMHCcons (38). The median binding affinity across all eight of these algorithms was

used both to nominate strong binder peptides for the seed dataset and to assess the impact of in silico mutation at each position of these peptides. For each strong binding peptide in the seed dataset, this systematic in silico mutation led to [length × amino acid identity × algorithm] binding predictions (e.g., 1368 binding predictions for a single 9-mer peptide).

Each position was assigned a score based on how much binding affinity values were influenced by mutations at that position. These scores were then 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 relative overall binding affinity change were assigned as anchor locations for further impact analysis.

Evaluating different seed peptide sources and their impact on the anchor patterns predicted

Three additional sources were explored and compared with peptides from our seed dataset to investigate whether the source of these strong binding peptides would influence our results. For comparison, we selected 1000 strong binding peptides (median predicted binding affinity <500 nM) from the following four sources: randomly generated peptides, peptides randomly selected from the WT proteome, peptides randomly selected from a viral genome, and peptides from our seed dataset. To obtain randomly generated peptides, we first generated a random list of amino acids (100,000,000 amino acids in length) using a random choice function and tested 50,000 9-mer peptides to reach 1000 predicted strong binders. For peptides generated from the WT human proteome (Ensembl) and the viral proteome (variola virus), we randomly selected 30,000 9-mer peptides from proteins with valid sequences longer than nine amino acids.

For each strong binding peptide, we followed our computational workflow for anchor predictions and mutated each position of the peptide to all possible amino acids and accessed the change to binding, respectively. Normalized scores across all peptide positions for 1000 peptides are plotted in fig. S4.

Bias analysis for HLA allele–specific anchor patterns

To evaluate whether our anchor patterns are influenced by the training data available for each HLA allele, we downloaded the training dataset of NetMHCpan4.0. In addition, for HLA alleles with no training data available, we determined which HLA allele was its closest neighbor defined by NetMHCpan4.0 and what the distance measurements were between the neighboring alleles. With these data obtained, we used Cytoscape to visualize the relationships between all alleles using network graphs. Each center node represents an HLA allele with training data (size of dataset correlates with the size of each node), and each connected node represents a neighboring HLA allele that uses the center node allele for its binding estimations. Distances are marked on the edges of each graph. Colors of each node reflect the anchor cluster assigned as shown in Fig 3.

HLA coverage and population frequency

Global HLA allele frequencies were generated using data from the Allele Frequency Net Database (39). 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 (i) aggregating all 195 sample populations, (ii) summing HLA allele counts over all sample populations, and (iii) dividing HLA allele counts over total counts of the HLA gene across all populations. The HLA frequencies calculated do not reflect true global HLA frequencies because 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.

Selection of validation peptides

To select a group of peptide-HLA combinations for validation experiments, we first examined our seed database of predicted strong binding combinations of peptides and HLA alleles. The database was then filtered for preselected HLA alleles with varying anchor patterns. For each HLA allele within an anchor subgroup, we then selected three to five peptides and performed validation experiments to measure their binding affinity. Validated peptides were then mutated at various positions (including those predicted to be anchors and non-anchors) to evaluate amino acid changes and their effect on binding affinity. Because of the limited resources and high cost of synthesizing peptides and performing validation experiments, we strategically chose four different amino acid mutations for each position evaluated. Both positions and amino acid mutations were chosen to optimize informativeness. The amino acid mutations were selected on the basis of their predicted influence on binding affinities by two methods. For our first batch of peptides and their in-depth analysis, we selected two of the most/least disturbing mutations each according to our predictions. For our second batch of peptides, this strategy was further optimized to select one mutation from each quartile on the basis of predictions of how much they disturb the binding interaction. These mutated peptides were then synthesized by GenScript (Piscataway, NJ, USA) for further validation. The peptides were ordered with the following specifications: quantity, 4 mg; weight, gross; purity, 75%; delivery format, lyophilized; and aliquoting to vials, 4.

Peptide:MHC IC$_{50}$ binding assays

As previously described (45), peptide binding validation assays to determine IC$_{50}$ values were carried out by Pure Protein LLC using the method of fluorescence polarization (FP). The technique is unique among methods used to analyze molecular binding events because it allows the instantaneous measurement of the ratio between free and bound labeled ligands in solution without any separation steps. FP is based on the principle that if a fluorescent-labeled peptide binds to the soluble HLA molecule of higher molecular weight, then polarization values will increase because of the slower molecular rotation of the bound probe. In this competition assay, a reference fluorescent-labeled peptide is incubated with activated sHLA in the presence of the neoantigen-derived peptide competitor, and peptide/HLA interaction is monitored over time. A positive response occurs when the peptide of interest outcompetes the labeled peptide tracer. A negative response takes place when the peptide of interest has no binding characteristics and only the tracer is

assembling with the sHLA. Competition experiments were analyzed by plotting FP values as a function of the log-arithms of competitor concentrations. The binding affinity of each competitor peptide was expressed as the concentration that inhibits 50% binding of the fluorescein isothiocyanate–labeled reference peptide. Observed IC_{50} were determined by nonlinear curve–fitting to a dose-response model with a variable slope using the specific software Prism (GraphPad Software Inc., San Diego, CA). To prioritize epitopes with the greatest potential, ranked peptide IC_{50} values were classified as log-dependent values (x) on the basis of preset affinity categories as described by Buchli *et al.* (45) into high ($x < 3.7$), medium ($3.7 < x < 4.7$), low ($4.7 < x < 5.5$), very low ($5.5 < x < 6.0$), and no binder ($x > 6.0$). Binding assays were performed for all (or part of) the peptides with the following matching HLA alleles: HLA-B*07:02, HLA-B*08:01, HLA-A*02:01, HLA-A*24:02, and HLA-A*68:01.

Cell-based peptide:MHC stabilization assays

The recipient of all HLAs of interest in this study was a transporter associated with antigen processing (TAP)–deficient, class I–negative cell line that was created by the introduction of HSV-ICP-47 (gift from T. Hansen) using the pMIP (puro^r) retrovirus and selection of puromycin-resistant cells into the class I–negative, lymphoblastoid cell line 721.221 (gift from M. Colonna, Washington University Saint Louis; also available from MilliporeSigma, catalog number SCC275) cultured in Iscove's MEM, 10% (v/v) fetal bovine serum (FBS) with puromycin (0.6 μg/ml). HLA cDNA (IDP-IMGT/HLA) was prepared synthetically (Blue Heron, Bothell, WA) and shuttled into pMIG [green fluorescent protein (GFP)] retroviral vector. All HLA-expressing cells were enriched by sorting GFP-positive cells to greater than 95% using a Sony Synergy Flow Cell Sorter. For peptide stabilization assays, the cell line expressing the HLA was washed twice in phosphate-buffered saline and then serum-starved for 1 hour in RPMI 1640, 10 mM Hepes, 1× non-essential amino acids, 1 mM sodium pyruvate, and 2 mM glutamine before plating, washed twice, and incubated with the peptide (100 μg/ml) of interest previously solubilized in 10% dimethyl sulfoxide and sterile-filtered through a 0.2-μm Centrex filter. The cell suspensions were maintained at ambient temperature for 1.5 to 3 hours and then shifted to 37°C overnight at 5% CO_2. To quantitate peptide-stabilized complexes, the cell suspensions were then washed twice to remove unbound peptide and then incubated with allophycocyanin (APC)–conjugated W6/32 Mab (BioLegend, 311410) for 30 min at ambient temperature. Cells were washed twice, and complexes were detected in the presence of 7-AAD to remove dead cells from the analysis. The median fluorescence intensity of 7-AAD negative, GFP/APC-positive cells was quantitated by FlowJo v.10.6.1 after collection on a Beckman Coulter Navios three-laser flow cytometer. Each HLA tested used a flu viral positive control peptide known to stabilize the HLA of interest (data file S6) and a negative control peptide that stabilized a different class I molecule. Each assay was completed two or three times in duplicate ($N = 4$ to 6) and graphed in GraphPad Prism v.9.2.0 after the subtraction of the MedFI of the no-peptide control. The corrected MFI values are included in data file S6. Stabilization assays were performed for all of the peptides with the following matching HLA alleles: HLA-A*02:01, HLA-A*23:01, HLA-A*24:02, HLA-A*31:01, HLA-A*68:01, HLA-B*07:02, and HLA-B*18:01.

Additional methods regarding orthogonal validation and impact analysis

Please refer to Supplementary Methods for details of these methods.

Statistical analysis

Statistical analysis was performed using the python SciPy package. Nonpaired two-sample *t* tests assuming unequal variance were performed to evaluate the differences among distributions for Fig. 4D. *T* tests were performed using the scipy.stats.ttest_ind function. The normal distribution of variables was calculated using the ShapiroWilk test.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

Funding:

Data and materials availability:

Sequence data for each cohort analyzed in this study are available through the Database of Genotypes and Phenotypes (dbGaP) at the following accession IDs: phs000178 for TCGA cohorts, phs002612 for GBM, phs001229 for lymphoma, and phs003153.v1.p1 for melanoma. Code used in our prediction pipeline and analyses are organized and available at https://github.com/griffithlab/anchor_huiming_etal_2023 (10.5281/zenodo.7672092).

REFERENCES AND NOTES

1. Zhang M, Fritsche J, Roszik J, Williams LJ, Peng X, Chiu Y, Tsou C-C, Hoffgaard F, Goldfinger V, Schoor O, Talukder A, Forget MA, Haymaker C, Bernatchez C, Han L, Tsang Y-H, Kong K, Xu X, Scott KL, Singh-Jasuja H, Lizee G, Liang H, Weinschenk T, Mills GB, Hwu P, RNA editing derived epitopes function as cancer antigens to elicit immune responses. Nat. Commun 9, 3919 (2018). [PubMed: 30254248]

2. Slansky JE, Spellman PT, Alternative Splicing in Tumors - A Path to Immunogenicity? N. Engl. J. Med 380, 877–880 (2019). [PubMed: 30811916]

3. Robbins PF, Lu Y-C, El-Gamil M, Li YF, Gross C, Gartner J, Lin JC, Teer JK, Cliften P, Tycksen E, Samuels Y, Rosenberg SA, Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. Nat. Med 19, 747–752 (2013). [PubMed: 23644516]

4. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, Zhang W, Luoma A, Giobbie-Hurder A, Peter L, Chen C, Olive O, Carter TA, Li S, Lieb DJ, Eisenhaure T, Gjini E,Stevens J, Lane WJ, Javeri I, Nellaiappan K, Salazar AM, Daley H, Seaman M,Buchbinder EI, Yoon CH, Harden M, Lennon N, Gabriel S, Rodig SJ, Barouch DH, Aster JC, Getz G, Wucherpfennig K, Neuberg D, Ritz J, Lander ES, Fritsch EF, Hacohen N, Wu CJ, An immunogenic personal neoantigen vaccine for patients with melanoma. Nature 547, 217–221 (2017). [PubMed: 28678778]

5. Keskin DB, Anandappa AJ, Sun J, Tirosh I, Mathewson ND, Li S, Oliveira G, Giobbie-Hurder A, Felt K, Gjini E, Shukla SA, Hu Z, Li L, Le PM, Allesøe RL, Richman AR, Kowalczyk MS, Abdelrahman S, Geduldig JE, Charbonneau S, Pelton K, Iorgulescu JB, Elagina L, Zhang W, Olive O, McCluskey C, Olsen LR, Stevens J, Lane WJ, Salazar AM, Daley H, Wen PY, Chiocca EA, Harden M, Lennon NJ, Gabriel S, Getz G, Lander ES, Regev A, Ritz J, Neuberg D, Rodig SJ, Ligon KL, Suvà ML, Wucherpfennig KW, Hacohen N, Fritsch EF, Livak KJ, Ott PA, Wu CJ, Reardon DA, Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. Nature 565, 234–239 (2019). [PubMed: 30568305]

6. Richters MM, Xia H, Campbell KM, Gillanders WE, Griffith OL, Griffith M, Best practices for bioinformatic characterization of neoantigens for clinical utility. Genome Med. 11, 56 (2019). [PubMed: 31462330]

7. Hundal J, Kiwala S, McMichael J, Miller CA, Xia H, Wollam AT, Liu CJ, Zhao S, Feng Y-Y, Graubert AP, Wollam AZ, Neichin J, Neveau M, Walker J, Gillanders WE, Mardis ER, Griffith OL, Griffith M, pVACtools: A computational toolkit to identify and visualize cancer neoantigens. Cancer Immunol. Res 8, 409–420 (2020). [PubMed: 31907209]

8. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, Griffith M, pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med. 8, 11 (2016). [PubMed: 26825632]

9. Rubinsteyn A, Hodes I, Kodysh J, Hammerbacher J, Vaxrank: A computational tool for designing personalized cancer vaccines, doi:10.1101/142919.

10. Bjerregaard A-M, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC, MuPeXI: Prediction of neo-epitopes from tumor sequencing data. Cancer Immunol. Immunother 66, 1123–1130 (2017). [PubMed: 28429069]

11. Xu C, A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput. Struct. Biotechnol. J 16, 15–24 (2018). [PubMed: 29552334]

12. Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, Wong YNS, Rowan A, Kanu N, Al Bakir M, Chambers T, Salgado R, Savas P, Loi S, Birkbak NJ, Sansregret L, Gore M, Larkin J, Quezada SA, Swanton C, Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: A pan-cancer analysis. Lancet Oncol. 18, 1009–1021 (2017). [PubMed: 28694034]

13. Yang W, Lee K-W, Srivastava RM, Kuo F, Krishna C, Chowell D, Makarov V, Hoen D, Dalin MG, Wexler L, Ghossein R, Katabi N, Nadeem Z, Cohen MA, Tian SK, Robine N, Arora K, Geiger H, Agius P, Bouvier N, Huberman K, Vanness K, Havel JJ, Sims JS, Samstein RM, Mandal R, Tepe J, Ganly I, Ho AL, Riaz N, Wong RJ, Shukla N, Chan TA, Morris LGT, Immunogenic neoantigens derived from gene fusions stimulate T cell responses. Nat. Med 25, 767–775 (2019). [PubMed: 31011208]

14. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, Fugmann T, Wong K-K, Van Allen EM, Intron retention as a novel source of cancer neoantigens, doi:10.1101/309450.

15. Laumont CM, Vincent K, Hesnard L, Audemard É, Bonneil É, Laverdure J-P, Gendron P, Courcelles M, Hardy M-P, Côté C, Durette C, St-Pierre C, Benhammadi M, Lanoix J, Vobecky S, Haddad E, Lemieux S, Thibault P, Perreault C, Noncoding regions are the main source of targetable tumor-specific antigens. Sci. Transl. Med 10, eaau5516 (2018).

16. Hilf N, Kuttruff-Coqui S, Frenzel K, Bukur V, Stevanovi S, Gouttefangeas C, Platten M, Tabatabai G, Dutoit V, van der Burg SH, Straten PT, Martínez-Ricarte F, Ponsati B, Okada G, Lassen U, Admon A, Ottensmeier CH, Ulges A, Kreiter S, von Deimling A, Skardelly M, Migliorini D, Kroep JR, Idorn M, Rodon J, Piró J, Poulsen HS, Shraibman B, McCann K, Mendrzyk R, Löwer M, Stieglbauer M, Britten CM, Capper D, Welters MJP, Sahuquillo J, Kiesel K, Derhovanessian E, Rusch E, Bunse L, Song C, Heesch S, Wagner C, Kemmer-Brück A, Ludwig J, Castle JC, Schoor O, Tadmor AD, Green E, Fritsche J, Meyer M, Pawlowski N, Dorner S, Hoffgaard F, Rössler B, Maurer D, Weinschenk T, Reinhardt C, Huber C, Rammensee H-G, Singh-Jasuja H, Sahin U, Dietrich P-Y, Wick W, Actively personalized vaccination trial for newly diagnosed glioblastoma. Nature 565, 240–245 (2019). [PubMed: 30568303]

17. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. Nature 578, 82–93 (2020). [PubMed: 32025007]

18. Sakamoto Y, Sereewattanawoot S, Suzuki A, A new era of long-read sequencing for cancer genomics. J. Hum. Genet 65, 3–10 (2020). [PubMed: 31474751]

19. Aganezov S, Goodwin S, Sherman RM, Sedlazeck FJ, Arun G, Bhatia S, Lee I, Kirsche M, Wappel R, Kramer M, Kostroff K, Spector DL, Timp W, McCombie WR, Schatz MC, Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. Genome Res. 30, 1258–1273 (2020). [PubMed: 32887686]

20. Capietto A-H, Hoshyar R, Delamarre L, Sources of cancer neoantigens beyond single-nucleotide variants. Int. J. Mol. Sci 23, 10131 (2022). [PubMed: 36077528]

21. Duan F, Duitama J, Al Seesi S, Ayres CM, Corcelli SA, Pawashe AP, Blanchard T, McMahon D, Sidney J, Sette A, Baker BM, Mandoiu II, Srivastava PK, Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. J. Exp. Med 211, 2231–2248 (2014). [PubMed: 25245761]

22. Cole DK, Edwards ESJ, Wynn KK, Clement M, Miles JJ, Ladell K, Ekeruche J, Gostick E, Adams KJ, Skowera A, Peakman M, Wooldridge L, Price DA, Sewell AK, Modification of MHC anchor residues generates heteroclitic peptides that alter TCR binding and T cell recognition. J. Immunol 185, 2600–2610 (2010). [PubMed: 20639478]

23. Capietto A-H, Jhunjhunwala S, Pollock SB, Lupardus P, Wong J, Hänsch L, Cevallos J, Chestnut Y, Fernandez A, Lounsbury N, Nozawa T, Singh M, Fan Z, de la Cruz CC, Phung QT, Taraborrelli L, Haley B, Lill JR, Mellman I, Bourgon R, Delamarre L, Mutation position is an important determinant for predicting cancer neoantigens. J. Exp. Med 217, e20190179 (2020).

24. Wells DK, van Buuren MM, Dang KK, Hubbard-Lucey VM, Sheehan KCF, Campbell KM, Lamb A, Ward JP, Sidney J, Blazquez AB, Rech AJ, Zaretsky JM, Comin-Anduix B, Ng AHC, Chour W, Yu TV, Rizvi H, Chen JM, Manning P, Steiner GM, Doan XC; Tumor Neoantigen Selection Alliance, Merghoub T, Guinney J, Kolom A, Selinsky C, Ribas A, Hellmann MD, Hacohen N, Sette A, Heath JR, Bhardwaj N, Ramsdell F, Schreiber RD, Schumacher TN, Kvistborg P, Defranoux NA, Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. Cell 183, 818–834.e13 (2020). [PubMed: 33038342]

25. Zamora AE, Crawford JC, Allen EK, Guo X-ZJ, Bakke J, Carter RA, Abdelsamed HA, Moustaki A, Li Y, Chang T-C, Awad W, Dallas MH, Mullighan CG, Downing JR, Geiger TL, Chen T, Green DR, Youngblood BA, Zhang J, Thomas PG, Pediatric patients with acute lymphoblastic leukemia generate abundant and functional neoantigen-specific CD8 T cell responses. Sci. Transl. Med 11, eaat8549 (2019).

26. Chen C, Zhou Q, Wu R, Li B, Chen Q, Zhang X, Shi C, A comprehensive survey of genomic alterations in gastric cancer reveals recurrent neoantigens as potential therapeutic targets. Biomed. Res. Int 2019, 2183510 (2019).

27. Perumal D, Imai N, Laganà A, Finnigan J, Melnekoff D, Leshchenko VV, Solovyov A, Madduri D, Chari A, Cho HJ, Dudley JT, Brody JD, Jagannath S, Greenbaum B, Gnjatic S, Bhardwaj N, Parekh S, Mutation-derived neoantigen-specific T-cell responses in multiple Myeloma. Clin. Cancer Res 26, 450–464 (2020). [PubMed: 31857430]

28. Zhang J, Caruso FP, Sa JK, Justesen S, Nam D-H, Sims P, Ceccarelli M, Lasorella A, Iavarone A, The combination of neoantigen quality and T lymphocyte infiltrates identifies glioblastomas with the longest survival. Commun. Biol 2, 135 (2019). [PubMed: 31044160]

29. Fritsch EF, Rajasagi M, Ott PA, Brusic V, Hacohen N, Wu CJ, HLA-binding properties of tumor neoepitopes in humans. Cancer Immunol. Res 2, 522–529 (2014). [PubMed: 24894089]

30. Schaettler MO, Richters MM, Wang AZ, Skidmore ZL, Fisk B, Miller KE, Vickery TL, Kim AH, Chicoine MR, Osbun JW, Leuthardt EC, Dowling JL, Zipfel GJ, Dacey RG, Lu H-C, Johanns TM, Griffith OL, Mardis ER, Griffith M, Dunn GP, Characterization of the genomic and immunologic diversity of malignant brain tumors through multisector analysis. Cancer Discov. 12, 154–171 (2022). [PubMed: 34610950]

31. Andreatta M, Nielsen M, Gapped sequence alignment using artificial neural networks: Application to the MHC class I system. Bioinformatics 32, 511–517 (2016). [PubMed: 26515819]

32. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M, NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J. Immunol 199, 3360–3368 (2017). [PubMed: 28978689]

33. Shao XM, Bhattacharya R, Huang J, Sivakumar IKA, Tokheim C, Zheng L, Hirsch D, Kaminow B, Omdahl A, Bonsack M, Riemer AB, Velculescu VE, Anagnostou V, Pagel KA, Karchin R, High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. Cancer Immunol. Res 8, 396–408 (2020). [PubMed: 31871119]

34. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J, MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Syst. 7, 129–132.e4 (2018). [PubMed: 29960884]

35. 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). [PubMed: 17608956]

36. 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). [PubMed: 19297351]

37. 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). [PubMed: 19948066]

38. 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). [PubMed: 22009319]

39. Gonzalez-Galarza FF, McCabe A, Santos EJMD, Jones J, Takeshita L, Ortega-Rivera ND, Cid-Pavon GMD, Ramsbottom K, Ghattaoraya G, Alfirevic A, Middleton D, Jones AR, 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). [PubMed: 31722398]

40. Yarzabek B, Zaitouna AJ, Olson E, Silva GN, Geng J, Geretz A, Thomas R, Krishnakumar S, Ramon DS, Raghavan M, Variations in HLA-B cell surface expression, half-life and extracellular antigen receptivity. eLife 7, e34961 (2018). [PubMed: 29989547]

41. Chen B, Khodadoust MS, Olsson N, Wagar LE, Fast E, Liu CL, Muftuoglu Y, Sworder BJ, Diehn M, Levy R, Davis MM, Elias JE, Altman RB, Alizadeh AA, Predicting HLA class II antigen presentation through integrated deep learning. Nat. Biotechnol 37, 1332–1343 (2019). [PubMed: 31611695]

42. Patronov A, Dimitrov I, Flower DR, Doytchinova I, Peptide binding prediction for the human class II MHC allele HLA-DP2: A molecular docking approach. BMC Struct. Biol 11, 32 (2011). [PubMed: 21752305]

43. Linette GP, Carreno BM, Neoantigen Vaccines Pass the Immunogenicity Test. Trends Mol. Med 23, 869–871 (2017). [PubMed: 28867556]

44. Cotto KC, Feng Y-Y, Ramu A, Skidmore ZL, Kunisaki J, Richters M, Freshour S, Lin Y, Chapman WC, Uppaluri R, Govindan R, Griffith OL, Griffith M, RegTools: Integrated analysis of genomic and transcriptomic data for the discovery of splicing variants in cancer, doi:10.1101/436634.

45. Buchli R, VanGundy RS, Hickman-Miller HD, Giberson CF, Bardet W, Hildebrand WH, Development and validation of a fluorescence polarization-based competitive peptide-binding

assay for HLA-A*0201--a new tool for epitope discovery. Biochemistry 44, 12491–12507 (2005). [PubMed: 16156661]

46. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, Christie C, Dalenberg K, Duarte JM, Dutta S, Feng Z, Ghosh S, Goodsell DS, Green RK,Guranovic V, Guzenko D, Hudson BP, Kalro T, Liang Y, Lowe R, Namkoong H, Peisach E, Periskova I, Prlic A, Randle C, Rose A, Rose P, Sala R, Sekharan M, Shao C, Tan L, Tao Y-P, Valasatava Y, Voigt M, Westbrook J, Woo J, Yang H, Young J, Zhuravleva M, Zardecki C, 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). [PubMed: 30357411]

47. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, Schwantes CR, Wang L-P, Lane TJ, Pande VS, MDTraj: A modern open library for the analysis of molecular dynamics trajectories. Biophys. J 109, 1528–1532 (2015). [PubMed: 26488642]
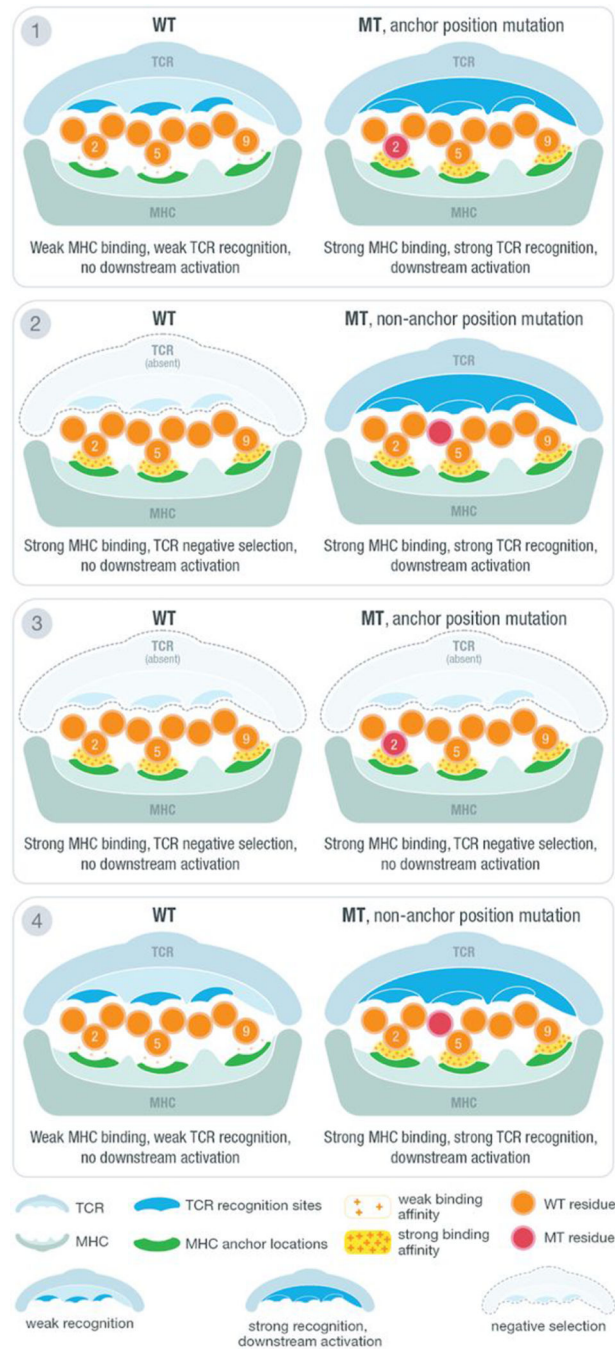
Fig. 1. Anchor and mutation position scenarios at the MHC-peptide-TCR interface.
Illustration of MHC-peptide-TCR interface using an example structure with anchors at
positions 2, 5, and 9. At the contact interface between the peptide-loaded MHC and the
recognizing T cell receptor, certain positions are 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 MT and WT peptides produce four distinct scenarios for interpreting candidate
neoantigens. Example TCR recognition sites are shown in blue, whereas MHC anchor

locations are shown in green. The peptide residues are shown in orange, whereas the MT 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 because of 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.
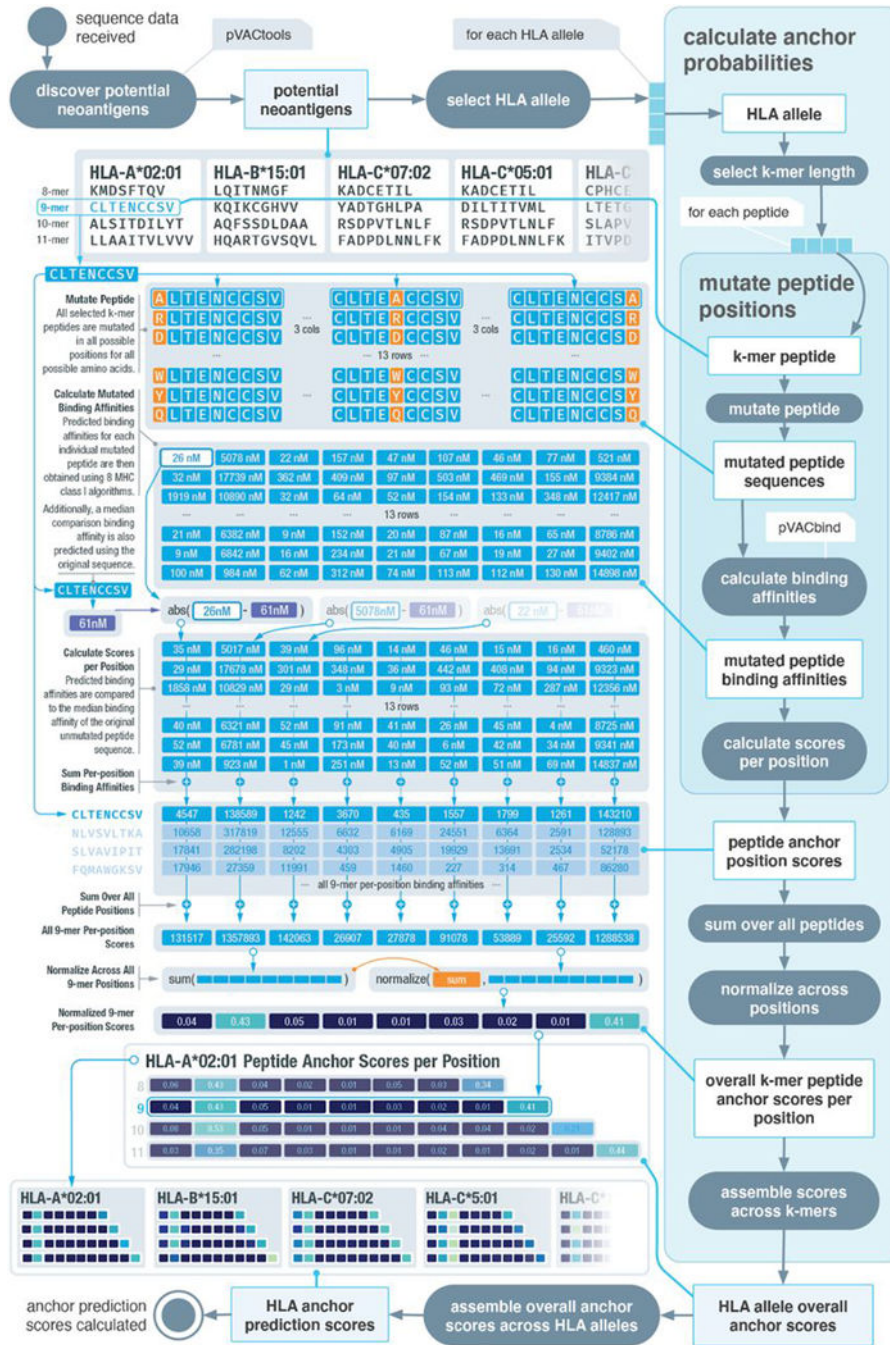
Fig. 2. Overview of computational workflow for anchor position prediction.
Schematic of our computational workflow to simulate the effect of mutation position on MHC binding affinities of peptides for individual HLA alleles. HLA and peptide pairings are selected from a reference dataset of putative strong binders. All possible amino acid changes are applied to all possible positions, and the 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.
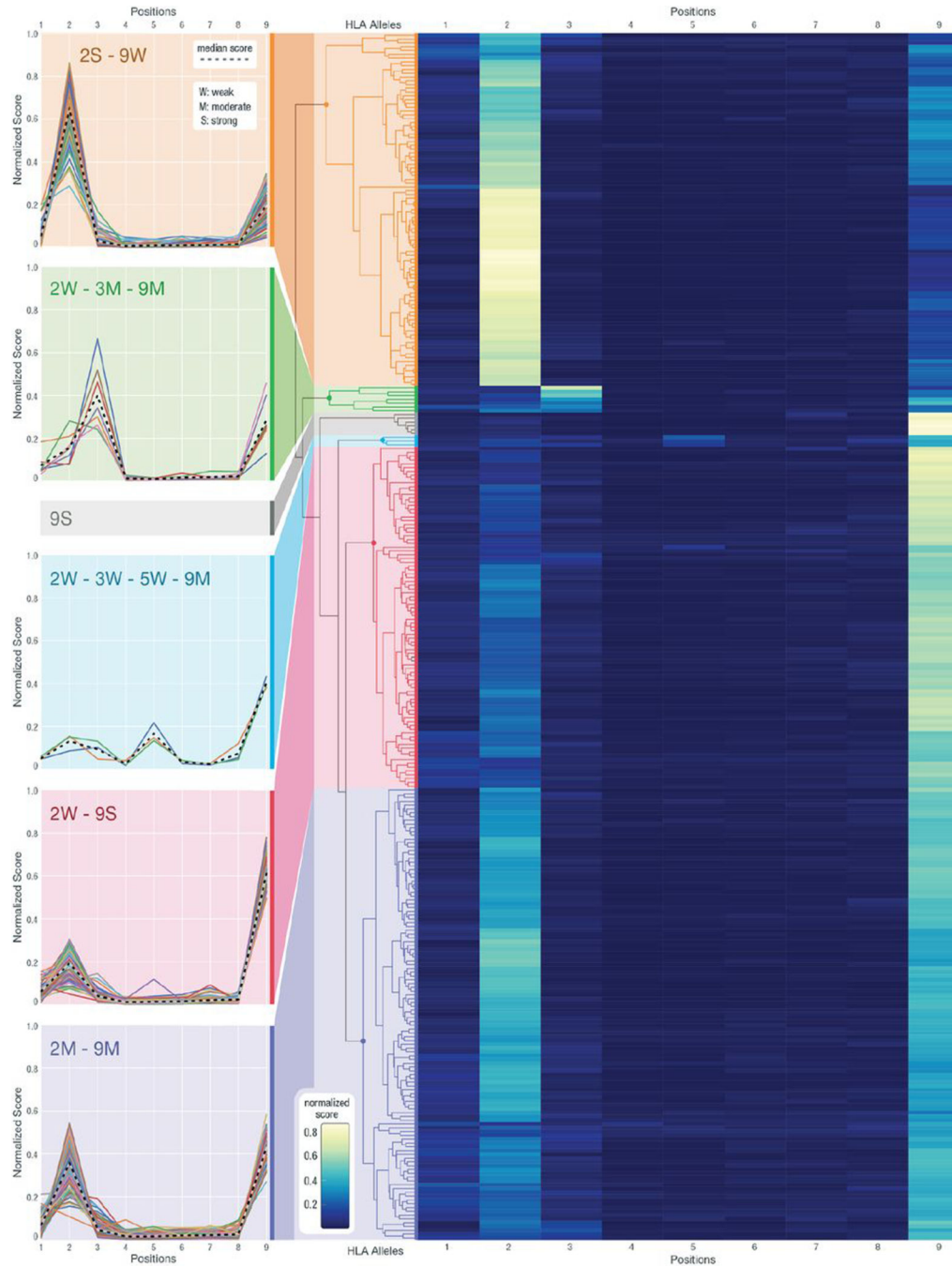
Fig. 3. Hierarchical clustering of anchor prediction scores across all 9-mer peptides.
Anchor prediction scores clustered using hierarchical clustering with average linkage for all
318 HLA alleles for which 9-mer peptide data were collected (of the 328 HLA alleles, 10
did not have corresponding 9-mer data). For the heatmap, the x axis represents the nine
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. In the cluster line plots on the left, the x axis shows the peptide positions,
whereas 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 with other anchors (S, strong; M, moderate; and W, weak). The median scores for each cluster are presented with a dashed line.
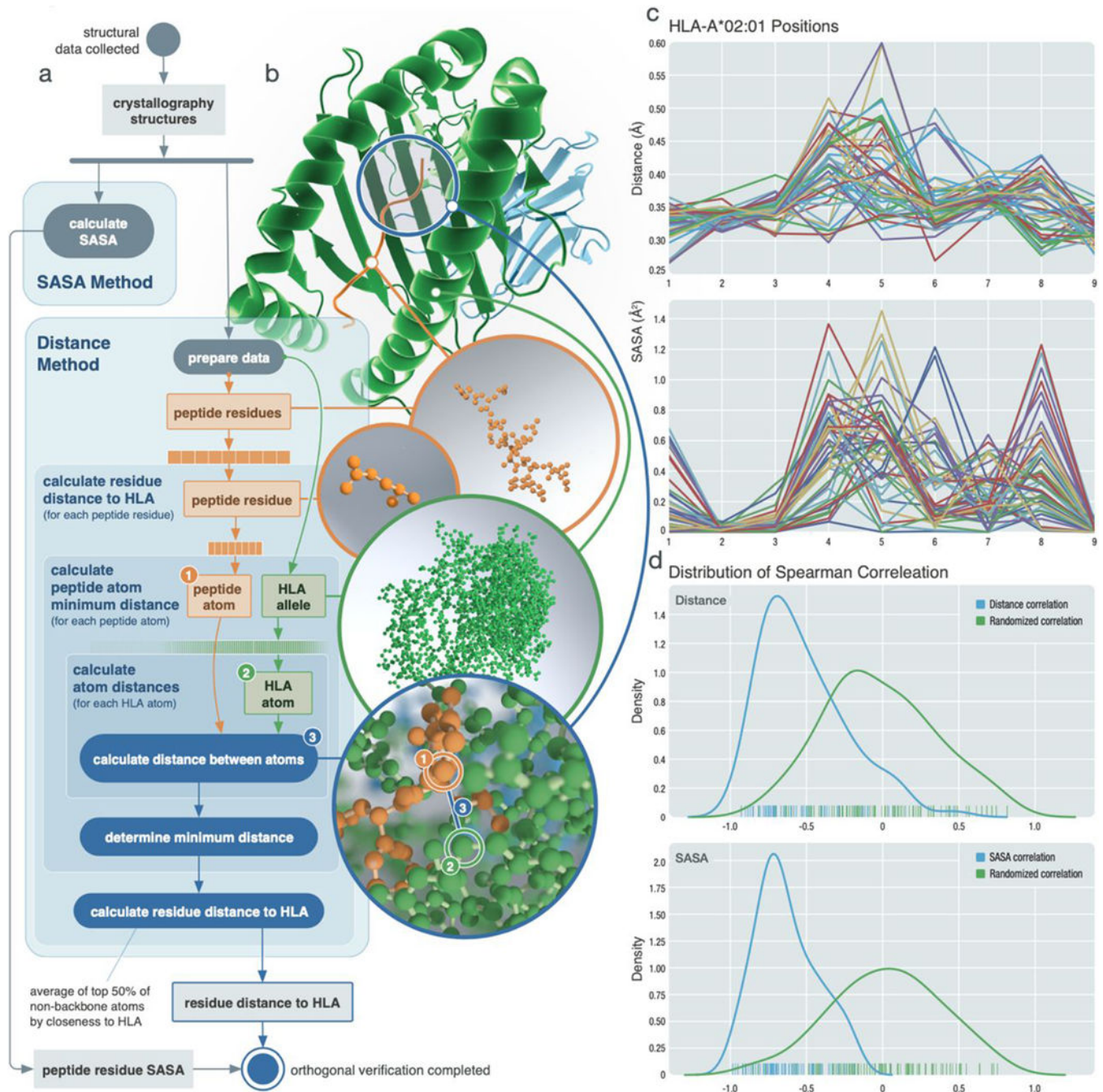
Fig. 4. Orthogonal validation using protein crystallography structures.
Orthogonal validation of predicted anchor scores using x-ray crystallography structures. (**A**) Schematic of analysis workflow for each HLA-peptide structure collected. For the distance metric, backbone atoms were excluded, with 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) Distance measurements for each position. (Bottom) SASA measurements. (**D**) Distribution of Spearman correlations calculated between distance and prediction scores (top) and

SASA and prediction scores (bottom). The blue line represents each respective correlation distribution, whereas the green line shows the distribution of Spearman correlation values obtained from randomly shuffled peptide positions.
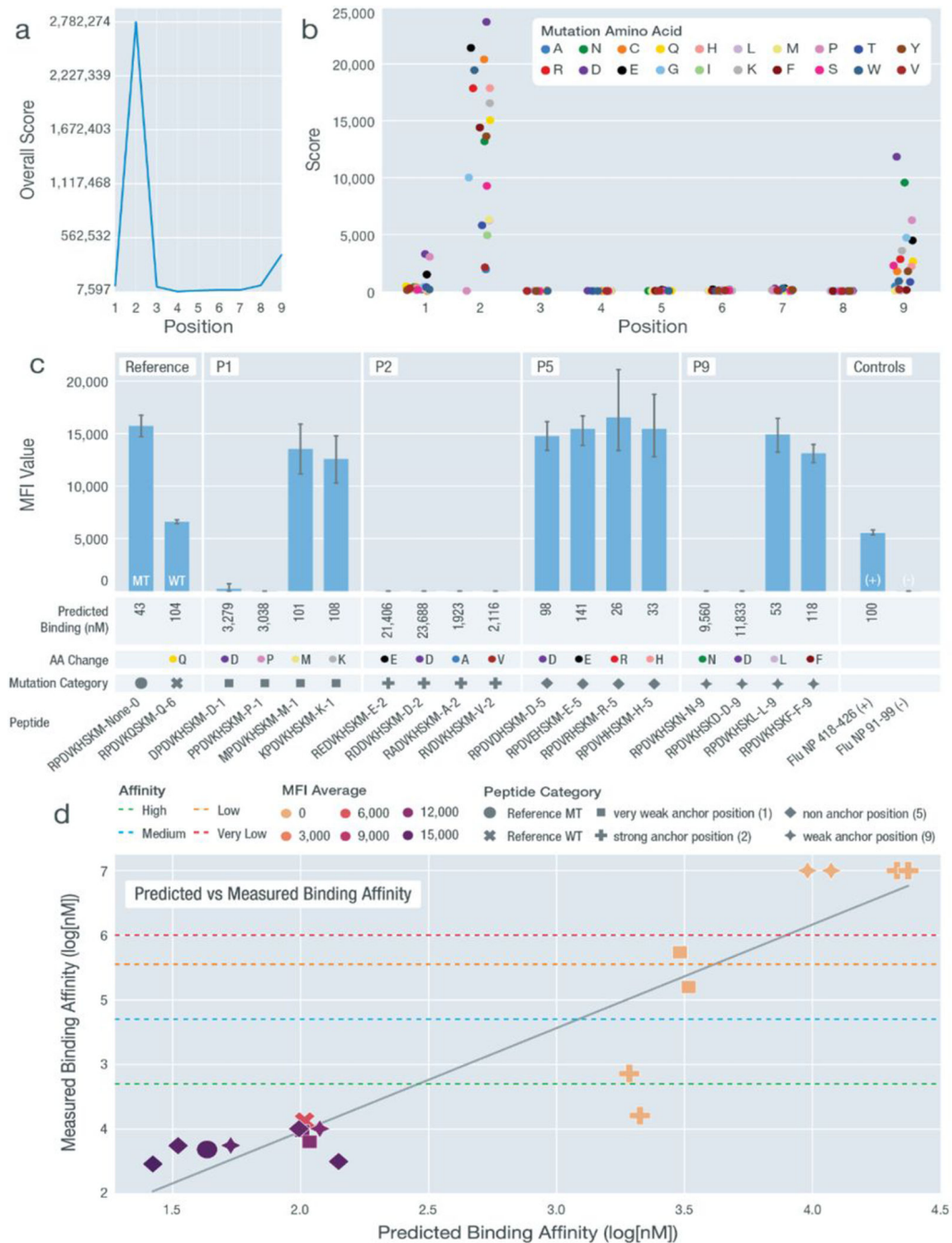
Fig. 5. Experimental validation of anchor pattern for HLA-B*07:02.
Validation results for HLA-B*07:02 with a predicted 2S-9W anchor pattern using peptide sequence RPDVKHSKM. (**A**) Overall anchor prediction scores for each position of 9-mer peptides when binding to HLA-B*07:02. Higher scores indicate a higher probability of acting as an anchor. (**B**) Binding affinity changes (*y* axis) plotted for each position of the specific 9-mer peptide RPDVKHSKM. Each peptide position was mutated to 19 other amino acids to evaluate influence on binding affinity. (**C**) MFI values measured from cell stabilization assays are plotted for both unmutated and mutated peptides. Peptides

are marked by their mutation positions (P1, P2, P5, and P9), predicted binding affinity values, amino acid changes [color coordinated with (B)], and mutation category [shape coordinated with (D)]. (**D**) Predicted binding affinity scores ($\log_{10}[nM]$) plotted against measured binding affinity values ($\log_{10}[nM]$) from IC_{50} binding assays. Binding categories based on measured binding affinity values are marked using horizontal lines. MFI values are overlaid using heatmap coloring for all data points. Peptide categories are marked using different shapes [coordinated (C)]. All nonbinders were plotted with a measured binding affinity value of 7.
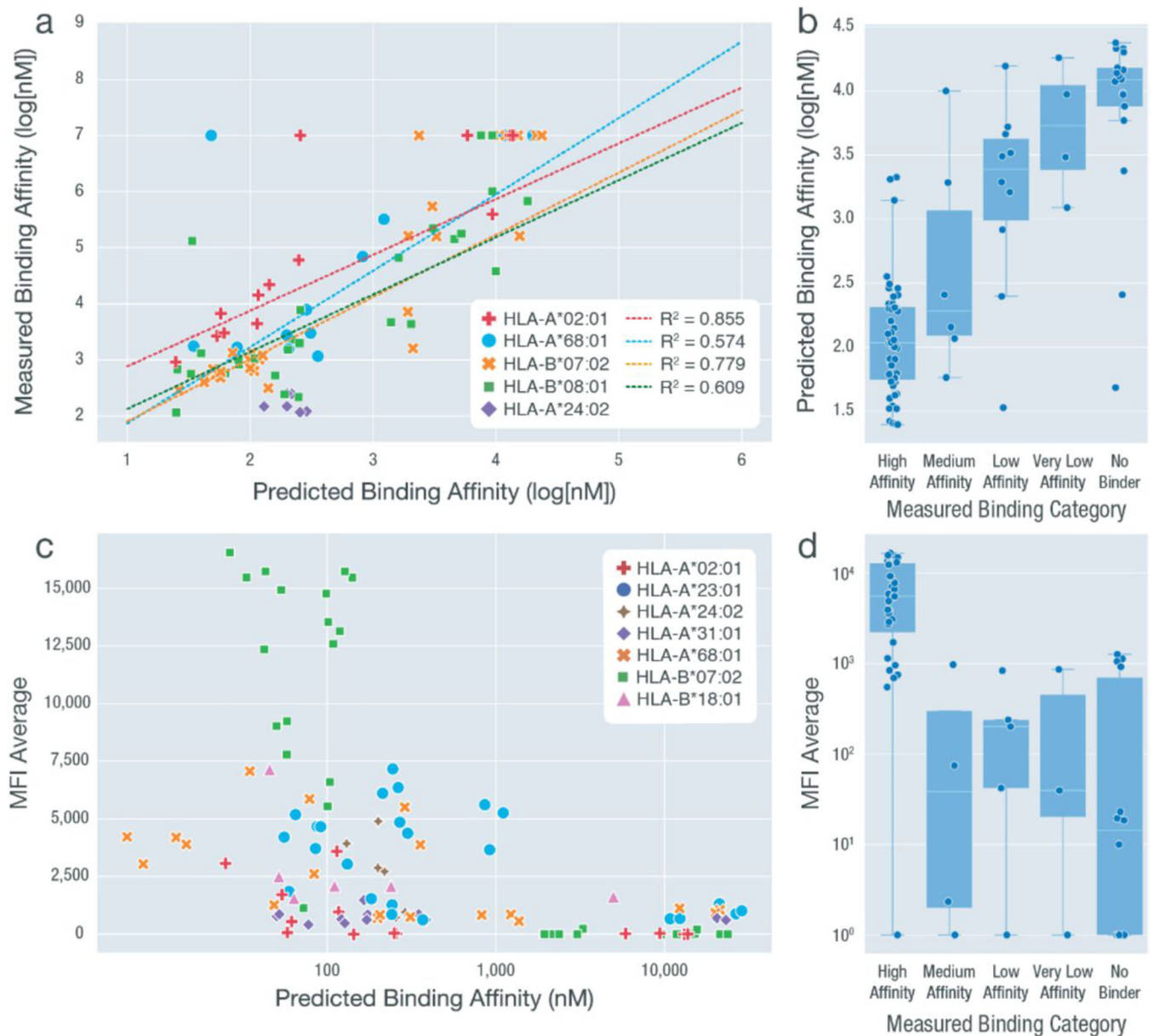
Fig. 6. Summary of experimental validation results across all HLA allele-peptide combinations.
We performed two types of experimental validations, including cell-based stabilization assays (MFI values) and IC_{50} binding assays (log[nM]), for a total of 136 peptide-HLA combinations. The results from these assays were compared between each other and against the predicted binding affinity (averaged across eight different algorithms). Measured binding categories were determined on the basis of the measured IC_{50} values. (**A**) Predicted binding affinity values plotted against the measured binding affinity (log[nM]). HLA-A*68:01, HLA-B*07:02, HLA-B*08:01, and HLA-A*02:01 were subject to linear fitting with R^2 values shown. HLA-A*24:02 was excluded because of the limited range of data available. Nonbinding peptide-HLA combinations that have no exact measured binding affinity value available are plotted at $y = 7$. These data points were not included when performing linear regression. (**B**) Measured binding categories plotted against predicted binding affinity

values. (**C**) Predicted binding affinity values plotted against average MFI values (measured at 100 nM concentration). (**D**) Measured binding categories plotted against MFI average value (measured at 100 nM concentration).
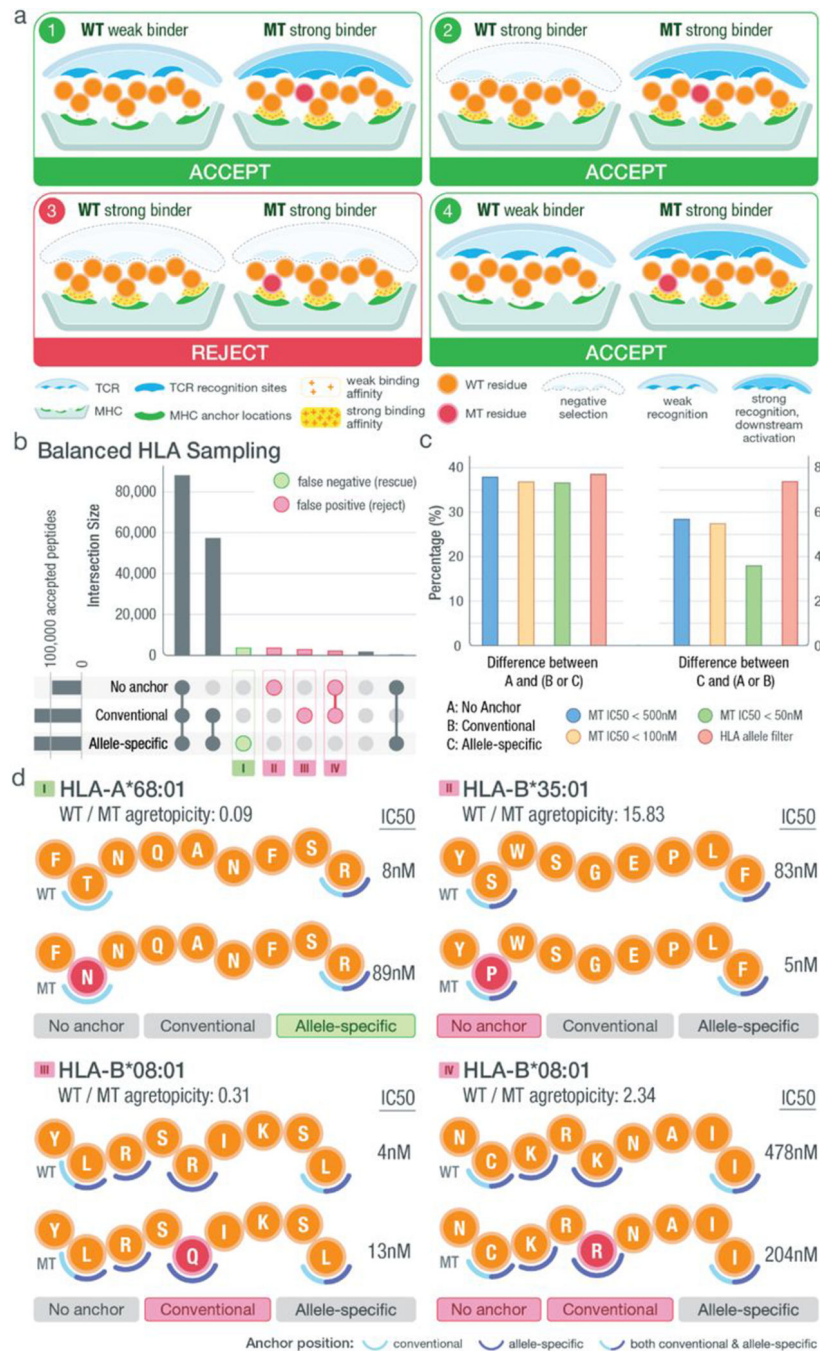
Fig. 7. Impact of anchor position information on neoantigen prioritization decisions.
(**A**) Illustration of different scenarios that could be encountered when prioritizing neoantigens. Each circle represents a peptide residue with WT residues indicated by orange and MT residues by red. Anchor locations of the MHC are marked in green, whereas 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. For putative strong binding MT peptides (IC$_{50}$ < 500 nM), the four different scenarios are depicted with the proposed varying classification

(accept/reject). The classification in each scenario depends on the WT binding and location of the MT residue, with respect to predicted anchor position(s). Given weak WT binding ($IC_{50} > 500$ nM), regardless of the mutation's anchor status, a strong MT binder is accepted as a neoantigen candidate (scenarios 1 and 4). However, given a strong WT binder, a peptide is accepted in the case where the MT residue is not at an anchor (scenario 2) and rejected when the MT residue is at a predicted anchor (scenario 3) because these peptides may be subject to central tolerance. (**B**) Upset plot showing the number of intersecting peptides based on those prioritized with no anchor filter (binding affinity < 500 nM and agretopicity > 1), conventional filter (filtering based on conventional anchor assumptions), or allelespecific filter (filtering based on our 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**) Bar plot showing the percentage difference among accepted neoantigen candidates based on filters discussed in (B). The filters in the legend represent how each starting dataset was filtered: (i) filtered by a strong binding cutoff of either 500, 100, or 50 nM and (ii) filtered on the basis of the anchor patterns of the corresponding HLA allele. Our HLA allele filter excluded all peptides binding to HLA alleles with a canonical anchor pattern (e.g., [2,9] for 9-mer peptides). (**D**) Examples of false-positive and false-negative peptides from each of the four subsets as marked in (B). Matching HLA allele, peptide sequence, mutation position (red), median WT/MT IC_{50} values, and fold changes are shown accordingly. Two sets of anchor locations are depicted for each scenario using semicircles: 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 semicircle.