Characters of neoantigens in cancer immunotherapy

Peng Bai[1], Yongzheng Li[1], Qiuping Zhou[1], Jiaqi Xia[1], Min Wu[2], Hexiang Deng[3], John W. Kappler[4,5,6,7], Philippa Marrack[4,6,8], Yu Zhou[1], Lei Yin[1*]

1 State Key Laboratory of Virology, Hubei Key Laboratory of Cell Homeostasis, College of Life Sciences, Wuhan University, Wuhan, 430072, Hubei, China

2 Hubei Key Laboratory of Cell Homeostasis, College of Life Sciences, Wuhan University, Wuhan, 430072, Hubei, China

3 Key Laboratory of Biomedical Polymers, Ministry of Education, the Institute for Advanced Studies, College of Chemistry and Molecular Sciences, Wuhan University, Wuhan, 430072, China

4 Department of Biomedical Research, National Jewish Health, Denver, CO 80206, USA

5 Barbara Davis Center for Childhood Diabetes, University of Colorado, Aurora, CO 80045, USA

6 Department of Immunology and Microbiology, University of Colorado School of Medicine, Aurora, CO 80045, USA

7 Structural Biology and Biochemistry program, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

8 Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

Abstract

Cancer mutations lead to the production of mutated proteins from which can be processed and presented to $CD8^+$ T cells as neoantigens. Despite the promise of immunotherapy targeting neoantigens, which mutant peptides are likely to induce immune responses remain elusive. Here, we determined two sets of MHC-I bound neoantigen structures and performed bioinformatics investigation based on neoantigen data. Structural and bioinformatic analyses reveal that some types of mutations in immunogenic neoantigens are preferentially selected. We also developed a strategy to expand the usage of a known neoantigen by scanning potential MHC alleles. Overall, our results will extend the ability to discover neoantigens that may be useful for cancer immunotherapy.

Introduction

The immune system can sometimes recognize and destroy cancer cells*(1)*. $CD8^+$ T cells have been demonstrated to be a major component in this process by both recognizing and destroying the target cells*(2–5)*. For T cells to react with tumor cells two components are required, major histocompatibility complex class I and possibly class II proteins (MHC-I, MHC-II) on the tumor cells and peptides which processed by the tumors bound to these MHC proteins. $CD8^+$ T cells can then recognize the combination of MHCI and tumor derived peptides on the cancer cells. Thus, it is crucial to identify effective tumor antigenic peptides with their relevant MHCs.

During the last 25 years, great efforts have been made to identify tumor antigens*(6)*. These tumor antigens can be classified into two broad categories, self-antigens which are seldom expressed in normal tissue or expressed at much higher levels on the cancer cell versus normal tissues (like NY-ESO-1, MART-1) and non-self-antigens called neoantigens which are derived from somatic mutations in the tumor*(7)*. The neoantigens are not presented by normal tissues, hence the immune system is not tolerant to them and views them, correctly, as foreign antigens and responds appropriately*(8, 9)*. Recently, next-generation sequencing (NGS) provide formidable identification of tumor-specific non-synonymous mutations, frameshift mutations, and gene rearrangements, which have driven clinical immunotherapeutic methods such as vaccination, adoptive cell transfer (ACT) and blockade of the proteins that inhibit T cell activation, have demonstrated that immunotherapy targeting tumor neoantigens can induce objective clinical responses in many patients with metastatic cancers*(10–13)*. It

appears that T cell targeting of tumor neoantigens might represent the “final common pathway” to treat cancer*(14)*. However, neoantigen discovery and validation remains a daunting problem. Classical cDNA expressing screening methods to identify tumor neoantigens without knowledge of the properties needed for a mutated peptide to be a successful immunotherapeutic is not efficient*(15)*. To solve this problem, recently, a study identified mouse neoantigens by combination of NGS and prediction algorithms for antigen presentation by MHCI*(16)*. In brief, they compared mouse tumor and normal DNA for identifying non-synonymous mutations and used peptide/MHCI binding prediction algorithms to find mutated peptides which could bind MHC to be potential effective neoantigens. These advances of discovering neoantigens bring us a step closer to personalize immunotherapy.

Indeed, neoantigen prediction lighten the burden of immunogenicity testing by reducing the number of candidate peptides*(17–27)*. Nevertheless these algorithms are not robust enough to identify immunogenic neoantigens accurately*(14, 28)*. There are several reasons for this. Firstly, many of these algorithms do not consider the likelihood that the peptide in question will be properly processed and transported to the endoplasmic reticulum, where they bind MHCI. This problem can be solved by adding other methods, such as NetChop*(29)* (for antigen processing prediction) and NetCTL*(30)* (for peptide transport prediction) to the prediction process. Secondly, the MHC binding prediction methods based on artificial neural network (ANN) models need a lot of data ,such as those available in NetMHC 4.0, for training*(31)*. However, large amounts of these input data are lacking for many human MHC alleles. Also, in some cases, most peptides in the training data were of non-tumor origin such as cultured cell lines and microbial foreign peptides. Thus, it is unclear whether these non-self peptidome data accurately represent the profile of mutated-self neoantigens and elicit T cell response*(32)*. Thirdly, among neoantigens binding MHC actually, only a small portion of them are immunogenic and have therapeutic effects*(33)*. Thus, it is crucial to investigate the details of neoantigen immunogenicity and to improve the performance of *in silico* neoantigen prediction algorithm.

Recently, Tran *et al*. reported a remarkable case of a patient whose metastatic colorectal cancer regressed after infusion of HLA-C*08:02-restricted tumor-infiltrating lymphocytes (TILs) targeting KRAS G12D neoantigens. *Kras* encodes a small GTPase that mediates downstream signaling from growth factor receptors, and it is one of the most infamous driver genes led to oncogenesis. The most common KRAS mutation is a gain-of-function substitution of the glycine to aspartic acid at codon 12 (KRAS G12D), a change that is often present in colorectal or pancreatic cancer*(34–36)*. The KRAS

G12D mutant 9mer peptide GADGVGKSA and 10mer peptide GADGVGKSAL can both stimulate autologous T cells in the context of HLA-C*08:02, while their wild type counterparts cannot*(23)*. Adoptive cell transfer (ACT) of T cells specific for the single mutant neoantigen led to tumor disappearance in patient. Thus, the KRAS G12D neoantigens have been acknowledged to be useful and typical neoantigen that might be used as a model to illustrate the secret of therapeutic tumor neoantigens.

Here we took structural analyses of several neoantigens with their relevant MHCs, combined with a broad examination of immunotherapeutic and non-immunogenic tumor antigens. We concluded that a universal type of tumor neoantigen peptide mutations from non-preferential residues to preferential residues with their relevant MHCs at anchor positions is unexpectedly a good indicator to predict neoantigen immunogenicity. Combined with other findings about effective neoantigens we thus propose an optimized strategy for neoantigen identification.

Results

The structures of HLA-C*08:02 in complex with KRAS G12D neoantigens

As described above, Tran and colleagues reported a clinical case which showed substantial tumor regression after adoptive T cell transfer of T cells specific for HLA-C*08:02 bound to the mutant KRAS G12D 9mer and 10mer peptides*(23)*. HLA-C*08:02 is expressed in approximately 8% of whites*(23)* and 7% of African Americans*(37)* in the U.S. The KRAS G12D mutation is most common in pancreatic and colorectal cancers, being present in approximately 45% and 13% of these two tumor types, respectively*(38)*. Thus, overall about 2.3% patients with colorectal cancer share the HLA-C*08:02 allele and KRAS G12D mutation (Supplementary Table 1) and these results suggest that thousands of patients per year in the United States alone could potentially be eligible for the therapy targeting HLA-C*08:02-restricted KRAS G12D neoantigens.

With this in mind we decided it would be useful to know the structure of HLA-C*08:02 in complex with four peptides: wild type KRAS G12D 9mer GAGGVGKSA (wt9m); wild type KRAS G12D 10mer peptide GAGGVGKSAL (wt10m); mutant KRAS G12D 9mer peptide GA<u>D</u>GVGKSA (mut9m, mutation site is indicated with the amino acid underlined) and mutant KRAS G12D 10mer peptide

108 GADGVGKSAL (mut10m). HLA-C*08:02 heavy chain and light chain β2-microglobulin (β2M) were
109 expressed in *E. coli* as inclusion bodies and refolded with peptides. C08-mut9m and C08-mut10m
110 complex were successfully refolded and detected on SDS-PAGE gels. Unfortunately, we failed to
111 refold C08-wt9m complex and C08-wt10mer complex even with a ten-fold increase in concentration of
112 the peptide (data not shown). This result suggested that the glycine to aspartic acid mutation of these
113 neoantigens is important to maintain the pMHC stability. However, the artificial neural network (ANN)
114 based peptide binding algorithm NetMHC4.0*(39)* predicted that binding affinity of KRAS G12D
115 mutated neoantigens (both 9mer and 10mer) and their wild type counterparts for HLA-C*08:02 would
116 all be very low (less than 15,000 nM, Supplementary Table 2). This could be due to relatively less
117 training data of HLA-C*08:02.

118 We solved the structures at 2.4 angstroms (Å) for the C08-mut9m complex and 1.9 Å for the C08-
119 mut10m complex (Supplementary Table 3). The high-quality resolution and unbiased electron density
120 at the region involving the peptides (Fig.1a, b) allowed reliable model building. The C08-mut9m
121 complex and the C08-mut10m complex shown highly conserved conformation except the peptide
122 region (Fig.1 c, d).

123 Peptide amino acid binding motif of HLA-C*08:02 shows that smaller residues such as alanine and
124 serine are preferably selected at P1 and P2 position (Fig.1e). This preference is due to the narrow cleft
125 formed by several C*08:02 aromatic residues (Tyr7, Phe33, Tyr67, Tyr99, Tyr59, Tyr171, Tyr159 and
126 Trp167), which limit the size of the P1 and P2 side chains that can bind at that site (Fig.1f). C08-
127 mut9m and C08-mut10m are relatively similar at the position 1 (P1) and position 2 (P2) of the
128 peptides. In C08-mut9m and C08-mut10m the N-terminal P1G (Glycine at position 1) residue of the
129 peptides form hydrogen bonds with Tyr7, Tyr159 and Tyr171 and the P2A residues interact with Glu63
130 and Lys66 (Fig. 2a).

131 Differences appeared at the P3-PΩ (PΩ: C-terminal position of peptide) regions of the peptides.
132 Although P3D in C08-mut9m and C08-mut10m both act as anchor residues, the side chain carboxyl
133 group of P3D forms an intra-chain hydrogen bond with P4G in mut9m peptide, while this hydrogen
134 bond is absent in the mut10m peptide structure. The hydrogen bond in the mut9m structure pulls down
135 the arch of the 9mer peptide, causing the region of the 9mer peptide that is accessible to TCRs to differ
136 from that of the 10mer peptide. The greatest difference between the two structures is apparent in the

137 center of the peptide (P4-P7), where each complex differs in the contacts it offers for TCR recognition.
138 The P8S has different orientations in the two structures and acts as an auxiliary anchor in C08-mut10m.
139 Differences were also observed at the C-termini of the peptides (Fig. 1c; Fig. 2b, c). The P9A of C08-
140 mut9m forms hydrogen bonds with Ser77, Asn80, Tyr84, Thr143 and Lys146, and acts as an anchor
141 residue (Fig. 2b). The P9A of C08-mut10m is totally different from the P9A of C08-mut9m with the
142 side chain up. However, the P10L of C08-mut10m has similar orientation towards HLA-C*08:02 and
143 plays a similar role as P9A in C08-mut9m (Fig. 1c; Fig. 2b, c).

144

145 **HLA-C*08:02 presents KRAS G12D neoantigens with an unconventional MHC anchor motif**

146 Typically, peptides always use anchor positions at P2 and PΩ to bind the B and F pockets respectively
147 of HLA class I alleles. The structures of the C08-mut9m and C08-mut10m reveal that G12D 9mer and
148 10mer peptides all use unconventional P3 and PΩ residues for anchors to form stable complexes with
149 HLA-C*08:02 (Fig.1c; Fig. 2b, c) thus differing from the canonical HLA Class I binding pattern (using
150 P2 and PΩ residues for anchors). In this case, the B pocket of HLA-C*08:02 binds the P3D anchor
151 residue via interactions with Arg97 and Arg156. Although the 9mer and 10mer peptides occupy the
152 same F pocket with their PΩ residues as anchors, the P10L side chain from the 10mer peptide is buried
153 more deeply than P9A in the 9mer structure (Fig.1c; Fig. 2b, c). Also, instead of the upward residue,
154 P8S, in 9mer peptide, the 10mer peptide uses P8S as an auxiliary anchor, binding the HLA protein via
155 hydrogen bonds with Glu152 and Arg156. These interactions squeeze the P4 to P7 residues together
156 with residue P3D (Fig. 2c). Also, together with P3D, the existence of auxiliary anchor residue P8S
157 arches the P4-P7 residues of the 10mer peptide to a higher position than that in the 9mer peptide (Fig.
158 2b, c). This alteration likely increases the binding affinity of the 10mer peptide with HLA and changes
159 the pHLA binding surface for the T cell receptor (TCR).

160 KRAS G12D 9mer and 10mer peptides were previously reported to stimulate different T-cell clones
161 with different TCR variable chains*(10)*. Our structural analyses are in line with such a report since
162 C08-mut9m and C08-mut10m provide different binding surfaces for T cell recognition (Fig.1b; Fig. 2b,
163 c). Furthermore, both wild type counterparts of 9mer and 10mer peptides failed to refold with HLA-
164 C*08:02, which means that the G12D mutation is important for forming the pMHC complex. Structural
165 analysis shows that G12D mutations of the KRAS neoantigens act as anchors and stabilize contacts

166 with the MHC. Thus, in this case, the KRAS mutation shows the likelihood that anchor mutated
167 neoantigens will exist to stimulate T cells.

168 **Structural basis of non-therapeutic mouse neoantigen DPAGT1 V213L in complex with H-2 Kb**

169 Having verified that the P3D residue of KRAS G12D neoantigens plays an important role in formation
170 of the pMHC complex, we studied whether other cancer neopeptides with anchor mutations had similar
171 properties. One such neopeptide, which is not immunotherapeutic, is the mutation of the C terminal
172 amino acid anchor residue of the mouse mutant DPAGT1 V213L peptide*(7)*. To study this we solved
173 the structure of mouse MHCI H-2 Kb in complex with mutant DPAGT1 V213L 8mer peptide
174 (SIIVFNLV, mutation site is indicated with the amino acid underlined) and the wild type 8mer
175 counterpart (SIIVFNLL) (Supplementary Table 3). The last residue of the peptide, a valine in the wild
176 type 8mer peptide (wt8mV) was replaced by leucine in the mutant 8mer peptide (mut8mL). Soluble
177 complexes of Kb-wt8mV and Kb-mut8mL were separately expressed, refolded and purified for
178 crystallization trials. Crystal diffraction data of Kb-wt8mV and Kb-mut8mL are processed to 2.4 Å and
179 2.5 Å resolution respectively (Supplementary Table 3) and provides unambiguous electron density for
180 each peptide (Supplementary Fig.1a).

181 The overall structure of the Kb-wt8mV complex closely resembles that of Kb-mut8mL with the
182 exception of a slight difference at PΩ (P8) (Supplementary Fig.1b). The C-terminal PΩ residue acts as
183 an anchor in both the Kb-wt8mV and Kb-mut8mL complexes (Supplementary Fig.1b). Moreover, both
184 PΩ valine and leucine were preferably selected by H-2 Kb in a peptide library analysis (Supplementary
185 Fig.1c). Both of the PΩ residues formed hydrogen bonds with Asp77, Tyr84, Thr143 and Lys146 of Kb
186 but the side chain of leucine in the mut8mL peptide is insert more deeply into Kb than valine in wt8mV
187 peptide because of longer side chain of the Leucine residue. Thus, P8L is able to make Van der Waals
188 contact with Tyr116 on the floor of the F pocket (Supplementary Fig.1d, e). These alterations did not
189 significantly affect the formation of peptide/Kb complex and change the TCR binding surface,
190 suggesting that, as far as T cell recognition is concerned the mut8mL neoantigen would be treated as a
191 "self-peptide", thus minimizing mut8mL neoantigen immunogenicity *in vivo*. All these indicate the
192 wild type and mutant DPAGT1 V213L peptides interact with Kb in a similar way, thus the mutation
193 might provide neither a higher concentration of peptide/Kb on the cell surface nor a TCR interface that
194 differs. This is unlike the case of the KRAS G12D mutant peptides which appear to interact more

intensively with HLA, via the mutant acid residue P3D interaction with basic residues in the B pocket of C*08:02, versus than the wild type counterpart.

Taken together, all these suggest that whether wild type or mutant amino acids at anchor positions interacted differently with relevant HLA alleles might be key to neoantigen immunogenicity.

Curated datasets of human immunogenic neoantigens from publications

To extend these analyses, bioinformatic studies were performed on all available immunogenic and non-immunogenic neoantigen peptide/MHC complexes. We manually gathered immunogenic neoantigens associated with various tumor types from recently published papers. This immunogenic neoantigen dataset (IND) encompassed 128 neoepitopes (neoantigens and their relevant HLAs) from 15 types of human cancer. Each of them was well tested by at least tested *in vitro* T cell assays. Less well characterized neoantigens and the neoantigens without minimal length were not included. Each entry included information such as the cancer type, HLA allele type, gene name, peptide sequence and other characteristics which are important in cancer immunotherapy. Melanoma and non-small cell lung cancer contain higher mutational burdens and have been more extensively studied than others, thus nearly 80% entries in our dataset came from them.

We also gathered 11739 non-immunogenic neopeptides (peptides splice from mutant proteins) from 13 published papers termed raw non-immunogenic neopeptide dataset (RNND). All of these mutant peptides were tested in *in vitro* T cell assays or clinical trials and did not elicit effective immune responses. Notably, the vast majority of them were not the minimal peptides known to induce $CD8^+$ T cells response (most of them are 25mer peptides). Thus, we use the prediction algorithm NetMHCpan 4.0*(40)* to predict the conventional minimal length (9mers and 10mers) non-immunogenic neopeptides with relevant HLAs under a cut-off IC50=500 nM. This dataset was termed non-immunogenic neopeptide dataset (NND). This dataset contains 2883 putative entries. It turns out that many neoantigens predicted to bind patients' HLAs very well are experimentally tested to be non-immunogenic and do not have the ability to evoke effective immune response. This fact compromised the usage of predicting immunogenic neoantigen through prediction of neoantigen binding to patients' HLAs.

Evident differences of mutation distribution between immunogenic and non-immunogenic peptides

It is thought that immunogenic neoantigens would bind MHC well and the mutations in the immunogenic neoantigens could increase the affinity for binding MHC. With our immunogenic neoantigen data collected, we preformed binding prediction (NetMHCpan 4.0) of wild type and mutant peptides in IND and showed the distribution of affinity ratio between them (Fig.3a). About 25.5% of immunogenic neoantigens shows the predicted binding affinity above 500nM and about 24% do not exhibit increased prediction binding affinity (WT/MUT affinity ratio <1). Thus, more characters of immunogenic neoantigens which could help immunogenic neoantigen predictions are needed.

It is suspected that the position in the peptides of mutated amino acids could affects the immunogenicity of neoantigens. To further investigate the mutation distribution at each peptide position between immunogenic and non-immunogenic data, we calculated the mutation occurrence frequency at each position of 9mer and 10mer peptides. Interestingly, both immunogenic 9mer and 10mer neoantigens show frequency fluctuation at each position with their mutations. The fluctuation of non-immunogenic 9mer and 10mer neoantigens seems not as high as those from immunogenic dataset (Fig. 3b; Supplementary Fig. 2). However the difference is not reaching significance probably due to the limited size of so far available immunogenic data. Mutations in 9mer neoantigens were more frequent at P2, P4, P5 and P8 position and relatively lower frequency at P3 and P7 positions, relative to non-immunogenic 9mer data (Fig. 3b). Conceptually, when bound to HLA, peptides usually use P2 and PΩ as anchor residues and use P4, P5 and P8 to contact TCR. Other residues in the peptide may contact with MHC but usually contribute less to the affinity of the peptide-HLA interaction*(41)*. Also to be taken into this account is the fact that our structural studies above also show that mutations at anchor positions can evoke effective immune responses. Thus, these observations enlightened us to classify the mutation based on whether it located on the anchor (contribute primary binding force to MHC), MHC contact (contact MHC but contribute much less for binding MHC) or TCR contact positions for further study.

To find out the mutation location (anchor, MHC contact position and TCR contact position) of each peptide in IND an NND respectively, a strategy was developed to define the mutations location of

neoantigens by combining the SYFPEITHI database*(42)*, antigen binding motif from NetMHCpan*(40)* and solved pMHC structures from Protein Data Bank (see **Methods**). We chose 9mer length of neoantigens to do these studies since 9mer peptides are most frequent in both the IND and NND collections. Peptides bound to MHC alleles for which structure models were not available in PDB were excluded from these analyses.

The 9mer neoantigens (or neopeptides) with mutations from IND and NND were divided into three functional classes based on the following: anchor mutation, MHC contact position mutation or TCR contact position mutation (Fig. 3c). After calculating the frequency of each class, we found that the immunogenic neoantigens were significantly more likely than NND to express mutations at TCR contact region (Fig. 3d, $p<0.05$, Z-test) and less likely to include mutations at MHC contact regions (excluding those at anchor positions) (Fig. 3d, $p<0.01$, Z-test). These results can partially support proposals of others that neopeptides with TCR contact position mutation will preferentially be immunogenic*(7)*. However, the odds of mutations at the anchor positions were similar between INDs and NNDs, therefore the importance of anchor position mutations in immunogenicity is not clear (Fig. 3d). These results are consistent with our structural studies showed that the immunogenic neoantigen KRAS G12D and the non-immunogenic neopeptide DPAGT1 V213L can both present by MHC.

Certain type of mutation change at anchor position could be a clear indicator to distinguish immunogenic neoantigens and non-immunogenic neopeptides

It is thought that the nature of mutated amino acids affects the neoantigen's immunogenicity. To find out if this is so, we calculated the frequency with which different amino acids appeared as the mutations in either IND or NND neoantigens. So far Leu and Tyr are likely to show up in the mutation position from immunogenic neoantigens (Supplementary Fig. 3). This could be due to the idea that mutation to Leu and Tyr might supply good contacts with MHC to stabilize the binding of neoantigens to MHC or good contacts with TCR to stimulate T cells. This is consistent with our previously studies that Tyr is important in TCR-peptide/MHC recognition*(43–46)*.

Next, to check the similarity between wild type and the mutant amino acids of neoantigens and non-immunogenic neopeptides, we defined similarity of those amino acids from IND and NND using the normalized BLOSUM62 substitution matrix*(47)*. In this matrix, bigger normalized score represents

more similarity between the wild type and the mutant peptides. For the situation of immunogenic data compare to non-immunogenic data, the peptides with immunogenicity did not shown significant difference between wild type and the mutant amino acids (Fig. 4a, p=0.2077, student t-test). For the situation of anchor mutated peptides, the similarity of immunogenic neoantigens to their wild type counterparts turn out to be lower than the similarity of the non-immunogenic peptides (Fig. 4b, p=0.001, student t-test). For the situation of 9mer peptides with TCR contact position mutation, similar result occurs between immunogenic and non-immunogenic data (Fig. 4c, p=0.019, student t-test). We did not test for mutations at MHC contact regions, because the available data was too limited so far. These results demonstrated that through the mutation position classification quantifiable amino acid differences between mutation and their wild type counterparts do indeed exist at anchor and TCR contact positions between immunogenic and non-immunogenic data.

To test whether the mutant peptides in IND are more preferentially presented by HLAs than wild type peptides, we generated amino acids half-bits of each HLA from P1 to P9 (we test 9mer peptides only due to their relatively abundant quantity) which present in IND, based on the binders collected from IEDB (see **Methods**). Then, merged bits of mutated neoantigens and their wild type peptides with their relevant HLAs were calculated respectively (Fig. 4d-f). To our knowledge, peptide with higher merge bits means preferentially presented by HLA. We first tested peptide merge bits in all 9mer neoantigens and their wild type counterparts. The higher merged were shown in mutant neoantigens compare to their wild type counterparts, which means immunogenic neoantigens more likely to be presented by their relevant HLAs (Fig. 4d, p=0.0088, student t-test). We next tested this character from the data of 9mer anchor mutated neoantigen and 9mer TCR contact position mutated neoantigen. Surprisingly, anchor mutated neoantigen showed significantly higher preference to be present by their relevant HLAs than their wild type counterparts (Fig. 4e, ****p<0.0001, student t-test), while TCR contact position mutated data did not shown any significance (Fig. 4f, p=0.9358, student t-test). From above results, it is evident that mutations of immunogenic neoantigens in the anchor positions and in the TCR contact positions shows different characters and should be studied separately. Moreover, the mutations of immunogenic neoantigens in the anchor positions were needed for further studies.

Inspired by our structure studies of two sets of anchor position mutated neoantigens, we therefore decided to find out how the mutations and wild type amino acids in immunogenic neoantigens were related to the known residue preferences of different HLA alleles, and whether the associations were

311 similar for immunogenic and non-immunogenic peptides. The cut-off for preferential amino acids was
312 applied to be above 10% (include 10%) as described previouslly*(48)* based on the data from peptide
313 library (Supplementary Fig. 4). Neoantigen mutations changes from wild type to mutant at anchor
314 positions were divided into 4 types: non-preferential to non-preferential residues (N to N), non-
315 preferential to preferential residues (N to P), preferential to non-preferential residues (P to N) and
316 preferential to preferential residues (P to P) (Supplementary Table 5). Mutations at MHC contact
317 position mutation were not examined because, as above, the numbers of these types of mutations were
318 too small to achieve statistical significance. Mutations at TCR contact regions was also not examined
319 because no residue preferences were found at these sites (see above).

320 We analyzed the proportion of these 4 types of mutation in the immunogenic and non-immunogenic
321 datasets respectively. Interestingly, almost all (26 in 27) of the immunogenic neoantigens appear to be
322 in the class of N to P preference (Supplementary Table 5; Fig. 4e). Statistical analysis shown that the N
323 to P preference frequency of immunogenic neoantigens with mutations at anchor positions was
324 significant higher (Fig. 4e, $p<0.001$) than that for non-immunogenic neoantigens'. The other classes of
325 preference did not show significant differences between the immunogenic and non-immunogenic
326 neoantigens. Analyzing our results in more depth we found that one immunogenic neoantigen,
327 MYADM, in the immunogenic peptide collection did not follow the N to P preference at its C-terminal
328 binding site for MHC. However, the original paper describing this immunogenic neoantigen showed
329 that both the wild type and the mutant peptides of MYADM could elicit self T cell responses*(49)*,
330 suggesting that the wild type peptide, with its C terminal anchor, was binding the patient's MHC
331 protein and immunogenic anyway. Thus, this neoantigen may not be a good target for further clinical
332 trial and can be regarded as an exception in our study.

333 With the classification of 4 types of neoantigen mutations changes at anchor positions, 96% (100% if
334 the exception MYADM is not counted) of the immunogenic neoantigen mutations belong to the class of
335 N to P preference and our detailed study of the KRAS G12D therapeutic neoantigen with the mutation
336 from the non-preferential residue G to the preferential residue D (N to P) is one of the clear examples.
337 That is, neoantigens with mutations at anchor positions not belonging to the class of N to P preference
338 would be very unlikely to be immunogenic and our detailed study of the DPAGT1 V213L non-
339 therapeutic neoantigen with the mutation from preferential residue V to the preferential residue L is one
340 of the clear examples. Thus it could be a clear indicator for the exclusion of non-immunogenic

341 neoantigen in the process of neoantigen prediction. From our NND dataset containing non-
342 immunogenic neopeptides predicted to bind relevant HLAs, about 42% of non-immunogenic
343 neoantigens with mutations at anchor positions were not N to P preference and could be eliminated
344 during immunogenic neoantigen prediction with this indicator (Fig. 4f). Thus, we suggest that the
345 processes we describe here could greatly help prediction of functional neoantigens for cancer
346 immunotherapy.

347

348 **HLA-C*05:01 might be able to present KRAS G12D 10mer peptide**

349 Because the frequency of KRAS mutations in tumors is so high, mutations in this protein are attractive
350 targets for immunotherapy. So far, it is thought that targeting the KRAS G12D neoantigen would be
351 useful only for patients expressing HLA-C*08:02. To expand use of the KRAS mutant we developed a
352 strategy to screen other HLA-C alleles for their ability to bind KRAS G12D and thus have the potential
353 to elicit antitumor responses.

354 Our strategy to identify new HLA candidates is divided into four steps. First, phylogenetic analysis was
355 performed for the 15 different HLA-C alleles with relatively high frequency in human*(50)*
356 (Supplementary Table 6; Supplementary Fig. 5). In the phylogenetic tree, three alleles (C*07:01,
357 C*07:02 and C* 17:01) were distantly related to the reference allele, C*08:02, and were excluded from
358 further analyses (Supplementary Fig. 5). The relative distance of the remaining alleles to C*08:02 was
359 calculated based on the phylogenetic tree. Secondly, their relative distance from C*08:02 was
360 normalized as shown on the heatmap (Fig. 5a). This showed that C*05:01 is closely related to C*08:02
361 and might therefore be able to present the KRAS neoantigen. Third, sequence alignment combined with
362 structure analyses was used to compare HLA-C*08:02 and other HLA-C alleles. Comparison of P3 and
363 PΩ binding motif of HLA-C alleles shown that C*05:01 is the only allele have the same sequence as
364 C*08:02 at its P3 anchor binding motif (Supplementary Table 7; Supplementary Fig. 6). Moreover
365 C*05:01 differed from C*08:02 by only two amino acids (at positions 77 and 80 on the α1 helix) at its
366 PΩ binding motif. Because of these observations we decided to test whether the changes would affect
367 the ability of C*05:01 to present the KRAS G12D peptide. We performed structural analyses based on
368 our solved HLA-C*08:02 structure (C08-mut9m).

369 Three residues from HLA -C*08:02 (position 97, position 99 and position 156) bind P3 of the peptide.

370 Two of these are arginine amino acids which promote engagement of the peptide acidic residue (Fig.
371 5b; Supplementary Table 7) and thus may play important roles in presenting acidic P3 mutated
372 neoantigen. These three amino acids are the same in C*05:01. The PΩ of peptide contacts amino acids
373 77, 80, 84, 143 and 146 in C*08:02 (Fig. 2c). Three of these (positions 84, 143, and 146) are invariable
374 in most HLA-C alleles (Supplementary Table 7). The two amino acids at positions 77 and 80 are also
375 the same in many of the HLA-C alleles but C*05:01 and several other HLA-C alleles differ from
376 C*08:02 by S77N and N80K. Amino acids at these positions pointed toward the TCR region and
377 peptide respectively (Fig. 5c). However, these five residues may not affect the presentation of peptides
378 which use leucine as a c-terminal residue because, regardless of the amino acids they express at these
379 five positions, most of HLA-C alleles select leucine as a preferential residue at the PΩ position of
380 peptide (Supplementary Fig. 4). Thus, after these structural comparisons of HLA-C*05:01 and HLA-
381 C*08:02, we inferred that HLA-C*05:01 might also present mutant KRAS G12D peptide and therefore
382 might be of potential use for T cell-based immunotherapy. We came to this conclusion in spite of the
383 fact that peptide binding predictions suggest that HLA-C*05:01 would bind with only low affinity to
384 both KRAS G12D 9mer and 10mer mutated peptides (Supplementary Table 2).

385 We next evaluated whether KRAS G12D can really be presented by C*05:01. We expressed C*05:01
386 heavy chain and β2M in E. coli and refolded them with 4 peptides: GA<u>D</u>GVGKSA (mut9m, mutation
387 site is indicated with the amino acid underlined); KRAS G12D 10mer peptide GA<u>D</u>GVGKSAL
388 (mut10m); wild type KRAS G12D 9mer peptide GAGGVGKSA (wt9m) and KRAS G12D 10mer
389 peptide GAGGVGKSAL (wt10m). Only the mutant KRAS G12D 10mer peptide successfully refolded
390 with HLA-C*05:01, suggesting that HLA-C*05:01 and HLA-C*08:02 are partially differ in their
391 antigen presentation abilities. We solved the crystal structure of C*05:01 in complex with the KRAS
392 G12D 10mer peptide to a resolution of 1.9 angstroms (Supplementary Table 3). High quality resolution
393 and electron density maps shown reliable placement of the 10mer peptide in C*05:01 complex (Fig.
394 5d). The TCR contact positions P5, P6, P7 and P9 of the peptides are flexible and can't form stable
395 conformations (Fig.5e). Nevertheless, the structure of the C05-mut10m closely resembles that of C08-
396 mut10m at the P1-P4 peptide region (Fig. 5e-g). Overlay of the C05-mut10m and C08-mut10m peptide
397 regions shows that an exception occurs at the P5-PΩ region of the peptides and position 77 and 80 of
398 the HLA proteins (Fig. 5e). The peptide P8S positions in both the C05-mut10m and C08-mut10m act
399 as axillary anchors. Differences also occur at the P10L position. Although C*05 and C*08 differ at

400 position 77, the Asn77 of C05 and the Ser77 of C08 form hydrogen bonds with P10L in a similar way.
401 However, Asn80 from C*05 binds the oxygen of P10L while Lys80 from C*08 does not with the
402 oxygen of P10L shifting towards the a2 helix and thus losing one of the hydrogen bonds from Tyr84
403 (Fig. 5h). These results explain to some extent why the GADGVGKSA (mut9m) peptide did not
404 successfully refold with C*05:01. The most likely cause is the change at position 80, which will
405 weaken the interaction between the 9mer peptide and C*05:01, thus reducing the likelihood of forming
406 a stable complex. Excluding these differences, the peptide binding grooves of the HLA-C proteins had
407 relatively conserved interactions with the KRAS G12D 10mer peptide. It is, however, possible that
408 position 80 is located in the interface of the pMHC and T cell receptor, and there may change the
409 morphology of the epitope and impact T cell selection.

410 In line with our hypothesis, the structure of HLA-C*05:01 in complex with KRAS G12D indicates that
411 C*05:01 can also present the KRAS G12D neoantigen but in a slightly different way than C*08:02
412 does. Therefore HLA-C*05:01 might also be a potential target for T cell-based cancer immunotherapy.

413

414 **An optimized procedure of neoepitope prediction**

415 Nowadays, most in silico neoepitope (neoantigen and corresponding HLAs) prediction methods have
416 used neopeptide binding affinity to discover neoantigens for cancer immunotherapy. Although binding
417 to MHC-I is currently the most popular computational filter for removing nonantigenic neopeptides, it
418 still not robust enough to eliminate most of these candidates.

419 Our results emphasize the importance of anchor mutated neopeptides in cancer immunotherapy. We
420 demonstrated that almost all of anchor mutated neoantigens which have induced autologous T cell
421 response and are clinically effective have a mutation the demonstrates an N to P preference at anchor
422 positions. Thus, we suggest an optimized procedure that will improve the accuracy of neoepitope
423 prediction methods. The workflow of this procedure is shown in Fig. 6a. Tumor and normal samples
424 are first obtained from cancer patient for genomic DNA and RNA sequencing and tumor infiltrating
425 lymphocytes (TILs) isolation. Whole-exome sequencing analysis allows discovery of somatic
426 mutations present in the tumor cells, but not in normal cells. *In silico* binding prediction is performed
427 and can output a list of candidate peptides. The novelty of our procedure is in the following steps.
428 Neoantigen candidates will be classified into 3 types: MHC anchor mutation (contact MHC and is vital

429 for binding MHC), MHC contact position mutation (contact MHC but contribute much less) or TCR
430 contact position mutation. Each type may use a different filtration strategy during neoantigen candidate
431 prediction (Fig. 6b). This procedure may eliminate at least 42% of the anchor position mutated
432 candidates. Also, the similarity score could be used to help judge the possibility that the TCR contact
433 position mutated neopeptides will be immunogenic. These candidates can be further tested *in vitro* and
434 later used in clinical trials. Meanwhile, a validated neoantigen can be used to identify other HLA alleles
435 that might present it, as shown in our analyses of KRAS G12D and HLA-C*05:01.

436

437 **Discussion**

438 Recently T cells targeting tumor antigens, especially neoantigens, have proved to be one of the most
439 promising new methods of treating cancers. However, few investigations have reported the structures
440 of T cells reacting with cancer neoantigens. To fill this gap, here we solved the structure of one of the
441 most successful clinical cancer immunotherapy neoantigens, KRAS with the G12D mutation, presented
442 by HLA-C*0802. We found that the G12D mutation in these neoantigens (9mer and 10mer) drive
443 intensive interactions to MHC via acting as an unconventional anchor residue. This observation of the
444 mutation at the anchor position which cause the epitope to be immunogenic illustrates the complexity
445 of neoantigens beyond the previous evidences that mutations which were solvent-exposed and
446 therefore accessible to TCRs were more likely to be immunogenic. We next studied the structure of a
447 non-immunogenic neoantigen, DPAGT1 V213L discovered from previously publication*(19)*. Although
448 this peptide also had a mutation at the anchor position, it did not change the conformation of the TCR
449 contact region. Thus different types of mutations in the anchor position of neoantigens would lead to
450 different fates for immunogenicity.

451 To expand our findings more broadly and deeply we combined our structural work with the
452 bioinformatics research in which the latter examined all available clinical data comparing cancer
453 neopeptides that had proved to be immunogenic or non-immunogenic. Both types of mutated peptides
454 from IND and NND were assigned to one of three different categories, those that contained mutations
455 in anchor positions for MHC, or mutations in TCR contact regions or those that were mutated in
456 contact regions with MHC that were not in anchor positions. Based on our results, mutations prefer to
457 occur at TCR contact positions rather than occur at MHC contact position (contact MHC but not anchor

position) in IND. However, mutation frequency at anchor position did not show significant difference between IND and NND. And in fact, whether mutations in peptide anchor positions are important for the immunogenicity of the neoantigen was previously unclear. Different pieces of evidence led to different conclusions*(19, 20, 51, 52)*. These assignments took into account the fact that the peptide anchor amino acids differ according to MHC allele with which they are engaged and the MHC anchor mutations were further categorized by whether or not the changed amino acids would improve binding of the peptide to MHC (N to P changes), or reduce binding to MHC (P to N changes) or would not affect the ability of the peptide to bind MHC (N to N or P to P changes). Our analyses revealed a remarkable finding, for the immunogenic anchor position mutated neoantigens nearly all the changes fall into the non-preferential to preferential, N to P, category (100% if the exception one MYADM, which is always reactive no matter it has the mutation or not, was not counted). Thus this indicator are clear and practical for identifying anchor mutated neoantigens, called as "Non-preferential to Preferential Conversion (NPC)". The amino acid turning from non-preferential to preferential may create a new epitope which never seen by self-immune system. Given the fact that during the prediction of viral antigen, most amino acids in the peptide are different from self-peptide, such indicator above is not necessary for consideration. Hence, we believe NPC is an effective supplement indicator to eliminate non-immunogenic peptides from cancer mutaome *in silico*.

The way of neoantigen study and the rules for immunogenic neoantigens could be very different with normal antigens. Neoantigens are the altered-self between self-peptides and totally different foreign peptides. In conventional antigen immunogenicity research, how amino acids of multiple positions get together to interact with MHC and be recognized by TCR are main points. However, in neoantigen research, how one residue change could reshape the peptide interactions with MHC and/or reshape TCR contact region as to break the immune tolerance are main points. Thus, mutation positions, mutation types, surrounding elements affecting mutations or affected by mutations and etc. should be paid more attentions. More specific bioinformatic strategies and more specific rules for immunogenic neoantigens are needed for research. In our study, certain rules were revealed only after large scale studies on human tumor mutant peptides and careful classification according to the mutation position in the peptides, which shows that the simple presence of a peptide mutated amino acid at an MHC binding site is not predictive. Only if the mutation follows the NPC indicator does the change at an anchor position become valuable.

488 On a related subject, here we tackled the issue of whether MHC alleles other than C*08:02 might be
489 useful in presenting the common KRAS G12D mutation as a neoantigen. We found that an MHC allele
490 that is closely related to C*08:02, C*05:01 might be such a possibility. Our structural work showed that
491 C*05:01 could indeed bind the KRAS G12D 10mer peptide and the configuration of the structure we
492 observed was quite similar to that of the same peptide bound to C*08:02. Thus, KRAS G12D mutation
493 could also be a good neoantigen in HLA-C*05:01 positive patients and the strategy we used to discover
494 this could be used to help enlarge knowledge of other suitable HLA alleles thus increasing the size of
495 the patient population in which such neoantigens could be used.

496 Our data suggest three possible immunogenic neoantigen models (Supplementary Fig. 7). Model A
497 represents the situation which peptide mutation occurs at TCR contact region and therefore appears
498 directly foreign to T cells. In this case, the neoantigens might be strongly immunogenic if the
499 corresponding wild type peptide is greatly dissimilar at the mutated site. Important differences might be
500 amino acid properties such as hydrophobicity, charge and polarity. Model B represents the situation in
501 which peptide mutation occurs at an MHC contact region (not including anchor sites). We found that
502 mutations at these sites are less likely to render the neoantigen immunogenic. However, our curated
503 data did show some bias for mutation at such regions, due to the relatively limited sample size
504 available now for this category, the relationship for such between mutation and immunogenicity is not
505 clear yet. Such mutants may still be immunogenic under certain conditions, if the mutant fits into an
506 MHC pocket of dramatically changes the TCR contact surface of the pMHC complex. Model C
507 represents the situation which the peptide mutation occurs at an MHC anchor position and increases
508 MHC binding.

509 Overall, more structural and biochemical experiments, with inclusion of more MHC alleles are needed
510 to improve bioanalytical predictive power in the important search for useful cancer neoantigens to treat
511 these devastating diseases. We hope the methods proposed here will help guide such future studies.

512

513 **Materials and methods**

514 **Study design**

515 This study was designed to use x-ray crystallography to examine the difference between immunogenic

516 and non-immunogenic mutant peptides bound to MHCI. On the basis of these results, we also took
517 bioinformatic analysis of our manually curated data form publications to see whether immunogenic
518 peptides have a certain character for neoantigen prediction. Moreover, we also performed phylogenetic
519 and structural analyses to extended neoantigen application.

520

521 **Protein Expression, refolding and purification**

522 The DNA encoding MHC heavy chain (HLA-C*08:02, HLA-C*05:01 and H-2 Kb) and light chain
523 (human β2M and mouse β2M) were synthesized (Idobio) and cloned into pET-22(b) vector (Novagen).
524 The vectors were transformed into the E. coli strain BL21 DE3 (Novagen). Transformants were
525 selected from Lurian broth (LB) agar plates containing ampicillin. A single colony was selected and
526 cultured in LB fluid medium with the antibiotics listed above at 37°C. Upon reaching an optical density
527 OD600 of 0.6, expression was induced with the addition of 1mM IPTG. Incubation continued at 37°C
528 for 5h. The cells were harvested by centrifugation and then resuspended in PBS buffer with 1 mM
529 PMSF at 4°C. The cells were lysed, and the lysate was clarified by centrifugation at 10,000 g to collect
530 inclusion bodies. Inclusion bodies were harvested and solubilized in 20 mM Tris (Vetec) pH 8.0, 8 M
531 urea (Vetec), 1 mM EDTA (BBI life sciences), 1 mM DTT (Sinopharm chemical reagent) and 0.2 mM
532 PMSF (Sinopharm chemical reagent).

533 Refolding was performed in the presence of MHC heavy chain, β2M and peptides as described
534 previously*(53)*. Briefly, the resolubilized heavy chain and light chain in the presence of the
535 corresponding peptide were added into 1 liter of refolding buffer. Refolding buffer contains 100 mM
536 Tris (pH 8.4), 0.5 mM oxidized glutathione (BBI life sciences), 5 mM reduced glutathione (BBI life
537 sciences), 400mM L-arginine (Vetec), 2mM EDTA (BBI life sciences). After 48h refolding, the 1 L
538 mixture dialyzed against 15 liters of 10 mM Tris buffer (pH 8.0) at 4 °C. for 24 h.

539 Soluble proteins were first purified by anion exchange chromatography with Q Sepharose HP (GE
540 Healthcare) column and Mono Q column (GE Healthcare). The purified protein sample was
541 concentrated by tangential flow filtration using Amicon Ultra centrifugal filters (Merck). For
542 desalination and purification, sample was loaded onto a Superdex 200 increase 10/300 GL column (GE
543 Healthcare) for size exclusion chromatography. Chromatography performed with BioLogic DuoFlow
544 system (Bio-rad) at a flow rate of 1 ml/min. Peak analysis was performed using the ASTRA software

package (BioLogic Chromatography Systems).

Crystallization, data collection and processing

Purified pMHC complex were concentrated to 10 mg/ml before the crystallization trials. For the HLA complex, crystallizations were done by the sitting drop vapor diffusion technique at 4 °C. Crystallization screens were performed using a series of kits from Hampton Research. Single crystal of C08-mut9m, C08-mut10m and C05-mut10m complex were obtained in the condition of 0.2 M ammonium acetate, 0.1 M HEPES (pH 7.5), 25% w/v polyethylene glycol 3,350. For the H-2 Kb complex, single crystal of Kb-8mV and Kb-8mL complex were obtained when 4% v/v Tacsimate (pH 7.0), 12% w/v Polyethylene glycol 3,350 was used as the reservoir buffer.

Before data collection, crystals were transferred to crystallization buffer containing 20% (w/v) glycerol. Next, the crystals were flash-cooled in liquid nitrogen immediately. The diffraction data were collected at the Shanghai Synchrotron Radiation Facility (Shanghai, China) on beam line BL17U1/BL18U1/BL19U1. The diffraction data were processed using the iMosflm program*(54)*, and data reduction was performed with Aimless Pointless in the CCP4 software suite*(55)*. All structures were determined by molecular replacement using the program Phaser*(56)*. The models from the molecular replacement were built using the COOT (Crystallographic Object-Oriented Toolkit) program*(57)* and subsequently subjected to refinement using Phenix software*(58)*. Data collection, processing, and refinement statistics are summarized in (Supplementary Table 3). All the structural figures were prepared using PyMOL program. The atomic coordinates and structure factors for the reported crystal structures have been deposited on the Protein Data Bank (PDB; http://www.rcsb.org/pdb/).

Neoantigens and non-immunogenic neopeptides data collection and correction.

Neoantigens and non-immunogenic neopeptides were manually curated with their corresponding HLA alleles from 34 published literatures and recorded in immunogenic neoantigen dataset, IND and non-immunogenic neopeptide dataset, NND.

To construct the immunogenic dataset IND, we collected 128 minimal peptides come from cancer

somatic mutations and their corresponding HLA class I alleles that reported to elicit T cell response in humans. Those tumor neopeptides with uncertain response or not provided with minimal sequence were not collected in our dataset. Each entry includes following information: tumor type, gene symbol, wild type/ mutant peptide sequence, wild type/mutant peptide predicted binding affinity*(31)*, single amino acid change, HLA allele, and PubMed ID from NCBI.

For non-immunogenic dataset NND, we collected 11739 peptides (not the minimal peptide) from cancer somatic mutations and their corresponding HLA class I alleles that reported not elicit T cell response in humans. Each entry includes following information: tumor type, gene symbol, wild type/ mutant peptide sequence, wild type/mutant peptide predicted binding affinity*(31)*, single amino acid change, HLA allele, and PubMed ID from NCBI.

Construction of non-immunogenic neopeptide dataset (NND)

In order to generate minimal non-immunogenic neopeptide dataset, binding prediction was taken by NetMHCpan 4.0 Server*(40)* based on the non-immunogenic neopeptide dataset (NND). Briefly, different length of mutant peptides from NND were input in NetMHCpan 4.0 Server with custom python scripts. The mutant peptides in 9- or 10-mer length with binding affinity under 500 nM (IC50) were considered to be binders and then recorded in the IND. Related peptide information was added to each entry like IND dataset.

Similarity comparison

Similarity comparison of wild type and mutant amino acids were taken based on normalized BLOSUM62 matrix as described previously*(47)*. This matrix provides a value from 0 to 1 for representing the similarity of amino acid pairs (higher value indicates more similar). Tails of the violins to the range of the data were not trimmed.

HLA present tendentiousness determination

Amino acid half-bit log-odds score in every position for different HLA alleles was calculated as

600 described by Morten *et al(59)*. Briefly, frequency of amino acid on distinct position was firstly
601 computed using Hobonm 1 algorithm to decrease the weight of similar sequence before calculation.

602 It is worth to note that, in some cases, if the sample size is too large, it may include similar sequences,
603 even duplicate sequences. Thus, clustering should be performed at first. For example, if one cluster
604 include 10 similar sequences, the weight of every sequence in this cluster will be 1/10. The similarity
605 score was the maximum percent of the same amino acids between one current sequence and every
606 sequence in one cluster, and the threshold of default similarity to decide whether current sequence
607 belongs to a cluster is 0.63.

608 In other cases, small sample related to amino acid depletion. Thus, we make use of pseudo counts to
609 solve this problem. The formula to calculate the pseudo counts is

610 $$g_i = \sum_j f_j \times q_{i|j}$$

611 where f_j is the real frequency of the amino acid j and $g_{i|j}$ is frequency of amino acid i replace amino
612 acid j based on BLOSUM62 matrix*(47)*.

613 After that, the estimated frequency is calculated according to equation

614 $$P_i = \frac{\alpha \cdot f_i + \beta \cdot g_i}{\alpha + \beta}$$

615 where α is the number of clusters subtracted by one. The default value β is 200.

616 Next, the half-bit log odds score is defined as

617 $$W_{i,a} = 2 \cdot \log_2 \frac{p_{i,a}}{q_a}$$

618 where q_a is the background frequency of amino acids. In our case, an equiprobable background amino
619 acid distribution is used.

620 The merged half-bits of one peptide is the summing half-bit score of amino acids in this peptide.
621 Statistical test of wild type peptides and mutant peptides merged half-bits from IND were tested using
622 t-test.

623

Nonameric peptide library obtained, peptide library sequence logo generation and HLA anchor position preferential determination

The nonameric peptide library of 30 HLA alleles was downloaded from IEDB*(60)*. Sequence logos were generated base on peptides library using the sequence logo generator*(61)*. Threshold for preferential amino acids at anchor position of each HLA alleles was applied to be above 10%*(48)* (include 10%) based on the data from nonameric peptide library.

HLA-C allele's frequency verification, phylogenetic analysis and heat map comparison

HLA-C allele frequency was verified from Allele Frequency Net Database (AFND; http://www.allelefrequencies.net)*(50)*. HLA-C alleles sequences were obtained from The IPD and IPD-IMGT/HLA Database*(62)*. Multiple sequence alignment was performed using MUSCLE (v3.8.31 for windows)*(63)*. The input file and output file format were defaulted. We constructed phylogenetic trees using MrBayes (v3.2.5 for windows) by following the instruction of 'Quick Start Version' section in manual*(64)*.

For display the allele similarities on the heat map, we read the phylogenetic tree using R package ape, and deleted some distantly related alleles (C*07:01, C*07:02, C*17:01). Similarity calculation was performed according to the distance of each HLA allele to C*08:02. The similarity was normalized between [0,1] and the heat map was plotted using R package heat map.

Statistical analysis

All data handling and nonlinear regression analyses were done using R, Microsoft Office Excel 2016 and GraphPad prism 6. P-values for difference were calculated using student's t test and Z-test.

References and Notes

1. M. D. Vesely, M. H. Kershaw, R. D. Schreiber, M. J. Smyth, Natural Innate and Adaptive Immunity

650 to Cancer, *Annu. Rev. Immunol.* **29**, 235–271 (2011).

651 2. Y. Sykulev, M. Joo, I. Vturina, T. J. Tsomides, H. N. Eisen, Evidence that a Single Peptide–MHC
652 Complex on a Target Cell Can Elicit a Cytolytic T Cell Response, *Immunity* **4**, 565–571 (1996).

653 3. P. C. Tumeh, C. L. Harview, J. H. Yearley, I. P. Shintaku, E. J. M. Taylor, L. Robert, B. Chmielowski,
654 M. Spasic, G. Henry, V. Ciobanu, A. N. West, M. Carmona, C. Kivork, E. Seja, G. Cherry, A. J. Gutierrez,
655 T. R. Grogan, C. Mateus, G. Tomasic, J. A. Glaspy, R. O. Emerson, H. Robins, R. H. Pierce, D. A. Elashoff,
656 C. Robert, A. Ribas, PD-1 blockade induces responses by inhibiting adaptive immune resistance,
657 *Nature* **515**, 568–571 (2014).

658 4. T. N. Schumacher, R. D. Schreiber, Neoantigens in cancer immunotherapy, *Science* **348**, 69–74
659 (2015).

660 5. N. A. Rizvi, M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P.
661 Wong, T. S. Ho, M. L. Miller, N. Rekhtman, A. L. Moreira, F. Ibrahim, C. Bruggeman, B. Gasmi, R.
662 Zappasodi, Y. Maeda, C. Sander, E. B. Garon, T. Merghoub, J. D. Wolchok, T. N. Schumacher, T. A.
663 Chan, Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,
664 *Science* **348**, 124–128 (2015).

665 6. R.-F. Wang, H. Y. Wang, Immune targets and neoantigens for cancer immunotherapy and
666 precision medicine, *Cell Res.* **27**, 11–37 (2017).

667 7. M. Yadav, S. Jhunjhunwala, Q. T. Phung, P. Lupardus, J. Tanguay, S. Bumbaca, C. Franci, T. K.
668 Cheung, J. Fritsche, T. Weinschenk, Z. Modrusan, I. Mellman, J. R. Lill, L. Delamarre, Predicting
669 immunogenic tumour mutations by combining mass spectrometry and exome sequencing, *Nature*
670 **515**, 572–576 (2014).

671 8. P. G. Coulie, B. J. Van den Eynde, P. van der Bruggen, T. Boon, Tumour antigens recognized by T
672 lymphocytes: at the core of cancer immunotherapy, *Nat. Rev. Cancer* **14**, 135–146 (2014).

673 9. N. R. J. Gascoigne, V. Rybakin, O. Acuto, J. Brzostek, TCR Signal Strength and T Cell Development,
674 *Annu. Rev. Cell Dev. Biol.* **32**, 327–348 (2016).

675 10. E. Tran, P. F. Robbins, Y.-C. Lu, T. D. Prickett, J. J. Gartner, L. Jia, A. Pasetto, Z. Zheng, S. Ray, E. M.
676 Groh, I. R. Kriley, S. A. Rosenberg, T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer, *N. Engl.*
677 *J. Med.* **375**, 2255–2262 (2016).

678 11. P. A. Ott, Z. Hu, D. B. Keskin, S. A. Shukla, J. Sun, D. J. Bozym, W. Zhang, A. Luoma, A. Giobbie-
679 Hurder, L. Peter, C. Chen, O. Olive, T. A. Carter, S. Li, D. J. Lieb, T. Eisenhaure, E. Gjini, J. Stevens, W.
680 J. Lane, I. Javeri, K. Nellaiappan, A. M. Salazar, H. Daley, M. Seaman, E. I. Buchbinder, C. H. Yoon, M.
681 Harden, N. Lennon, S. Gabriel, S. J. Rodig, D. H. Barouch, J. C. Aster, G. Getz, K. Wucherpfennig, D.
682 Neuberg, J. Ritz, E. S. Lander, E. F. Fritsch, N. Hacohen, C. J. Wu, An immunogenic personal
683 neoantigen vaccine for patients with melanoma, *Nature* **547**, 217–221 (2017).

684 12. U. Sahin, E. Derhovanessian, M. Miller, B.-P. Kloke, P. Simon, M. Löwer, V. Bukur, A. D. Tadmor,
685 U. Luxemburger, B. Schrörs, T. Omokoko, M. Vormehr, C. Albrecht, A. Paruzynski, A. N. Kuhn, J. Buck,

686 S. Heesch, K. H. Schreeb, F. Müller, I. Ortseifer, I. Vogler, E. Godehardt, S. Attig, R. Rae, A. Breitkreuz,
687 C. Tolliver, M. Suchan, G. Martic, A. Hohberger, P. Sorn, J. Diekmann, J. Ciesla, O. Waksmann, A.-K.
688 Brück, M. Witt, M. Zillgen, A. Rothermel, B. Kasemann, D. Langer, S. Bolte, M. Diken, S. Kreiter, R.
689 Nemecek, C. Gebhardt, S. Grabbe, C. Höller, J. Utikal, C. Huber, C. Loquai, Ö. Türeci, Personalized
690 RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer, *Nature* **547**,
691 222–226 (2017).

692 13. P. F. Robbins, Y.-C. Lu, M. El-Gamil, Y. F. Li, C. Gross, J. Gartner, J. C. Lin, J. K. Teer, P. Cliften, E.
693 Tycksen, Y. Samuels, S. A. Rosenberg, Mining exomic sequencing data to identify mutated antigens
694 recognized by adoptively transferred tumor-reactive T cells, *Nat. Med.* **19**, 747–752 (2013).

695 14. E. Tran, P. F. Robbins, S. A. Rosenberg, "Final common pathway" of human cancer
696 immunotherapy: targeting random somatic mutations, *Nat. Immunol.* **18**, 255–262 (2017).

697 15. T. Wolfel, M. Hauer, J. Schneider, M. Serrano, C. Wolfel, E. Klehmann-Hieb, E. De Plaen, T.
698 Hankeln, K. Meyer zum Buschenfelde, D. Beach, A p16INK4a-insensitive CDK4 mutant targeted by
699 cytolytic T lymphocytes in a human melanoma, *Science* **269**, 1281 (1995).

700 16. H. Matsushita, M. D. Vesely, D. C. Koboldt, C. G. Rickert, R. Uppaluri, V. J. Magrini, C. D. Arthur,
701 J. M. White, Y.-S. Chen, L. K. Shea, J. Hundal, M. C. Wendl, R. Demeter, T. Wylie, J. P. Allison, M. J.
702 Smyth, L. J. Old, E. R. Mardis, R. D. Schreiber, Cancer exome analysis reveals a T-cell-dependent
703 mechanism of cancer immunoediting, *Nature* **482**, 400–404 (2012).

704 17. N. van Rooij, M. M. van Buuren, D. Philips, A. Velds, M. Toebes, B. Heemskerk, L. J. A. van Dijk,
705 S. Behjati, H. Hilkmann, D. el Atmioui, M. Nieuwland, M. R. Stratton, R. M. Kerkhoven, C. Keşmir, J.
706 B. Haanen, P. Kvistborg, T. N. Schumacher, Tumor Exome Analysis Reveals Neoantigen-Specific T-
707 Cell Reactivity in an Ipilimumab-Responsive Melanoma, *J. Clin. Oncol.* **31**, e439–e442 (2013).

708 18. P. F. Robbins, Y.-C. Lu, M. El-Gamil, Y. F. Li, C. Gross, J. Gartner, J. C. Lin, J. K. Teer, P. Cliften, E.
709 Tycksen, Y. Samuels, S. A. Rosenberg, Mining exomic sequencing data to identify mutated antigens
710 recognized by adoptively transferred tumor-reactive T cells, *Nat. Med.* **19**, 747–752 (2013).

711 19. M. Yadav, S. Jhunjhunwala, Q. T. Phung, P. Lupardus, J. Tanguay, S. Bumbaca, C. Franci, T. K.
712 Cheung, J. Fritsche, T. Weinschenk, Z. Modrusan, I. Mellman, J. R. Lill, L. Delamarre, Predicting
713 immunogenic tumour mutations by combining mass spectrometry and exome sequencing, *Nature*
714 **515**, 572–576 (2014).

715 20. F. Duan, J. Duitama, S. Al Seesi, C. M. Ayres, S. A. Corcelli, A. P. Pawashe, T. Blanchard, D.
716 McMahon, J. Sidney, A. Sette, B. M. Baker, I. I. Mandoiu, P. K. Srivastava, Genomic and bioinformatic
717 profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity, *J. Exp.*
718 *Med.* **211**, 2231–2248 (2014).

719 21. C. J. Cohen, J. J. Gartner, M. Horovitz-Fried, K. Shamalov, K. Trebska-McGowan, V. V. Bliskovsky,
720 M. R. Parkhurst, C. Ankri, T. D. Prickett, J. S. Crystal, Y. F. Li, M. El-Gamil, S. A. Rosenberg, P. F. Robbins,
721 Isolation of neoantigen-specific T cells from tumor and peripheral lymphocytes, *J. Clin. Invest.* **125**,
722 3981–3991 (2015).

723 22. B. M. Carreno, V. Magrini, M. Becker-Hapak, S. Kaabinejadian, J. Hundal, A. A. Petti, A. Ly, W.-R.
724 Lie, W. H. Hildebrand, E. R. Mardis, G. P. Linette, A dendritic cell vaccine increases the breadth and
725 diversity of melanoma neoantigen-specific T cells, *Science* **348**, 803 (2015).

726 23. E. Tran, P. F. Robbins, Y.-C. Lu, T. D. Prickett, J. J. Gartner, L. Jia, A. Pasetto, Z. Zheng, S. Ray, E. M.
727 Groh, I. R. Kriley, S. A. Rosenberg, T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer, *N. Engl.*
728 *J. Med.* **375**, 2255–2262 (2016).

729 24. E. Stronen, M. Toebes, S. Kelderman, M. M. van Buuren, W. Yang, N. van Rooij, M. Donia, M.-L.
730 Boschen, F. Lund-Johansen, J. Olweus, T. N. Schumacher, Targeting of cancer neoantigens with
731 donor-derived T cell receptor repertoires, *Science* **352**, 1337–1341 (2016).

732 25. V. Anagnostou, K. N. Smith, P. M. Forde, N. Niknafs, R. Bhattacharya, J. White, T. Zhang, V. Adleff,
733 J. Phallen, N. Wali, C. Hruban, V. B. Guthrie, K. Rodgers, J. Naidoo, H. Kang, W. Sharfman, C.
734 Georgiades, F. Verde, P. Illei, Q. K. Li, E. Gabrielson, M. V. Brock, C. A. Zahnow, S. B. Baylin, R. B.
735 Scharpf, J. R. Brahmer, R. Karchin, D. M. Pardoll, V. E. Velculescu, Evolution of Neoantigen Landscape
736 during Immune Checkpoint Blockade in Non–Small Cell Lung Cancer, *Cancer Discov.* **7**, 264–276
737 (2017).

738 26. U. Sahin, E. Derhovanessian, M. Miller, B.-P. Kloke, P. Simon, M. Löwer, V. Bukur, A. D. Tadmor,
739 U. Luxemburger, B. Schrörs, T. Omokoko, M. Vormehr, C. Albrecht, A. Paruzynski, A. N. Kuhn, J. Buck,
740 S. Heesch, K. H. Schreeb, F. Müller, I. Ortseifer, I. Vogler, E. Godehardt, S. Attig, R. Rae, A. Breitkreuz,
741 C. Tolliver, M. Suchan, G. Martic, A. Hohberger, P. Sorn, J. Diekmann, J. Ciesla, O. Waksmann, A.-K.
742 Brück, M. Witt, M. Zillgen, A. Rothermel, B. Kasemann, D. Langer, S. Bolte, M. Diken, S. Kreiter, R.
743 Nemecek, C. Gebhardt, S. Grabbe, C. Höller, J. Utikal, C. Huber, C. Loquai, Ö. Türeci, Personalized
744 RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer, *Nature* **547**,
745 222–226 (2017).

746 27. N. Zacharakis, H. Chinnasamy, M. Black, H. Xu, Y.-C. Lu, Z. Zheng, A. Pasetto, M. Langhan, T.
747 Shelton, T. Prickett, J. Gartner, L. Jia, K. Trebska-McGowan, R. P. Somerville, P. F. Robbins, S. A.
748 Rosenberg, S. L. Goff, S. A. Feldman, Immune recognition of somatic mutations leading to complete
749 durable regression in metastatic breast cancer, *Nat. Med.* **24**, 724–730 (2018).

750 28. E. Stronen, M. Toebes, S. Kelderman, M. M. van Buuren, W. Yang, N. van Rooij, M. Donia, M.-L.
751 Boschen, F. Lund-Johansen, J. Olweus, T. N. Schumacher, Targeting of cancer neoantigens with
752 donor-derived T cell receptor repertoires, *Science* **352**, 1337–1341 (2016).

753 29. M. Nielsen, C. Lundegaard, O. Lund, C. Keşmir, The role of the proteasome in generating
754 cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage,
755 *Immunogenetics* **57**, 33–41 (2005).

756 30. M. V. Larsen, C. Lundegaard, K. Lamberth, S. Buus, O. Lund, M. Nielsen, Large-scale validation
757 of methods for cytotoxic T-lymphocyte epitope prediction, *BMC Bioinformatics* **8**, 424 (2007).

758 31. M. Andreatta, M. Nielsen, Gapped sequence alignment using artificial neural networks:
759 application to the MHC class I system, *Bioinformatics* **32**, 511–517 (2016).

760 32. M. M. Davis, P. Brodin, Rebooting Human Immunology, *Annu. Rev. Immunol.* **36**, 843–864 (2018).

761 33. A. Vitiello, M. Zanetti, Neoantigen prediction and the need for validation, *Nat. Biotechnol.* **35**,
762 815–817 (2017).

763 34. C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M.
764 A. Wyczalkowski, M. D. M. Leiserson, C. A. Miller, J. S. Welch, M. J. Walter, M. C. Wendl, T. J. Ley, R.
765 K. Wilson, B. J. Raphael, L. Ding, Mutational landscape and significance across 12 major cancer
766 types, *Nature* **502**, 333–339 (2013).

767 35. C. H. June, Drugging the Undruggable Ras — Immunotherapy to the Rescue?, *N. Engl. J. Med.*
768 **375**, 2286–2289 (2016).

769 36. T. Kamisawa, L. D. Wood, T. Itoi, K. Takaori, Pancreatic cancer, *The Lancet* **388**, 73–85 (2016).

770 37. F. F. González-Galarza, L. Y. C. Takeshita, E. J. M. Santos, F. Kempson, M. H. T. Maia, A. L. S. da
771 Silva, A. L. T. e Silva, G. S. Ghattaoraya, A. Alfirevic, A. R. Jones, D. Middleton, Allele frequency net
772 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction
773 associations, *Nucleic Acids Res.* **43**, D784–D788 (2015).

774 38. T-Cell Transfer Therapy Targeting Mutant KRAS, *N. Engl. J. Med.* **376**, e11 (2017).

775 39. M. Andreatta, M. Nielsen, Gapped sequence alignment using artificial neural networks:
776 application to the MHC class I system, *Bioinformatics* **32**, 511–517 (2016).

777 40. V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters, M. Nielsen, NetMHCpan-4.0: Improved
778 Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity
779 Data, *J. Immunol.* **199**, 3360–3368 (2017).

780 41. M. Matsumura, D. Fremont, P. Peterson, I. Wilson, Emerging principles for the recognition of
781 peptide antigens by MHC class I molecules, *Science* **257**, 927–934 (1992).

782 42. H.-G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, S. Stevanović, SYFPEITHI:
783 database for MHC ligands and peptide motifs, *Immunogenetics* **50**, 213–219 (1999).

784 43. K. Christopher Garcia, J. J. Adams, D. Feng, L. K. Ely, The molecular basis of TCR germline bias
785 for MHC is surprisingly simple, *Nat. Immunol.* **10**, 143 (2009).

786 44. L. Yin, J. Scott-Browne, J. W. Kappler, L. Gapin, P. Marrack, T cells and their eons-old obsession
787 with MHC, *Immunol. Rev.* **250**, 49–60 (2012).

788 45. L. Yin, E. Huseby, J. Scott-Browne, K. Rubtsova, C. Pinilla, F. Crawford, P. Marrack, S. Dai, J. W.
789 Kappler, A Single T Cell Receptor Bound to Major Histocompatibility Complex Class I and Class II
790 Glycoproteins Reveals Switchable TCR Conformers, *Immunity* **35**, 23–33 (2011).

791 46. J. J. Adams, S. Narayanan, M. E. Birnbaum, S. S. Sidhu, S. J. Blevins, M. H. Gee, L. V. Sibener, B.
792 M. Baker, D. M. Kranz, K. C. Garcia, Structural interplay between germline interactions and adaptive
793 recognition determines the bandwidth of TCR-peptide-MHC cross-reactivity, *Nat. Immunol.* **17**, 87

794 (2015).

795 47. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks., *Proc. Natl.*
796 *Acad. Sci.* **89**, 10915–10919 (1992).

797 48. K. Falk, O. Rötzschke, S. Stevanovié, G. Jung, H.-G. Rammensee, Allele-specific motifs revealed
798 by sequencing of self-peptides eluted from MHC molecules, *Nature* **351**, 290–296 (1991).

799 49. N. McGranahan, A. J. S. Furness, R. Rosenthal, S. Ramskov, R. Lyngaa, S. K. Saini, M. Jamal-
800 Hanjani, G. A. Wilson, N. J. Birkbak, C. T. Hiley, T. B. K. Watkins, S. Shafi, N. Murugaesu, R. Mitter, A.
801 U. Akarca, J. Linares, T. Marafioti, J. Y. Henry, E. M. Van Allen, D. Miao, B. Schilling, D. Schadendorf,
802 L. A. Garraway, V. Makarov, N. A. Rizvi, A. Snyder, M. D. Hellmann, T. Merghoub, J. D. Wolchok, S. A.
803 Shukla, C. J. Wu, K. S. Peggs, T. A. Chan, S. R. Hadrup, S. A. Quezada, C. Swanton, Clonal neoantigens
804 elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade, *Science* **351**, 1463–
805 1469 (2016).

806 50. F. F. González-Galarza, L. Y. C. Takeshita, E. J. M. Santos, F. Kempson, M. H. T. Maia, A. L. S. Da
807 Silva, A. L. Teles E Silva, G. S. Ghattaoraya, A. Alfirevic, A. R. Jones, D. Middleton, Allele frequency
808 net 2015 update: New features for HLA epitopes, KIR and disease and HLA adverse drug reaction
809 associations, *Nucleic Acids Res.* **43**, D784–D788 (2015).

810 51. E. F. Fritsch, M. Rajasagi, P. A. Ott, V. Brusic, N. Hacohen, C. J. Wu, HLA-Binding Properties of
811 Tumor Neoepitopes in Humans, *Cancer Immunol. Res.* **2**, 522–529 (2014).

812 52. M. Yarchoan, B. A. Johnson, E. R. Lutz, D. A. Laheru, E. M. Jaffee, Targeting neoantigens to
813 augment antitumour immunity, *Nat. Rev. Cancer* **17**, 209–222 (2017).

814 53. W. A. Macdonald, Z. Chen, S. Gras, J. K. Archbold, F. E. Tynan, C. S. Clements, M. Bharadwaj, L.
815 Kjer-Nielsen, P. M. Saunders, M. C. J. Wilce, F. Crawford, B. Stadinsky, D. Jackson, A. G. Brooks, A.
816 W. Purcell, J. W. Kappler, S. R. Burrows, J. Rossjohn, J. McCluskey, T Cell Allorecognition via
817 Molecular Mimicry, *Immunity* **31**, 897–908 (2009).

818 54. T. G. G. Battye, L. Kontogiannis, O. Johnson, H. R. Powell, A. G. W. Leslie, *iMOSFLM* : a new
819 graphical interface for diffraction-image processing with *MOSFLM*, *Acta Crystallogr. D Biol.*
820 *Crystallogr.* **67**, 271–281 (2011).

821 55. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B.
822 Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton,
823 H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, Overview of the *CCP* 4 suite and current
824 developments, *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).

825 56. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, *Phaser*
826 crystallographic software, *J. Appl. Crystallogr.* **40**, 658–674 (2007).

827 57. P. Emsley, K. Cowtan, *Coot* : model-building tools for molecular graphics, *Acta Crystallogr. D Biol.*
828 *Crystallogr.* **60**, 2126–2132 (2004).

58. P. D. Adams, R. W. Grosse-Kunstleve, L.-W. Hung, T. R. Ioerger, A. J. McCoy, N. W. Moriarty, R. J. Read, J. C. Sacchettini, N. K. Sauter, T. C. Terwilliger, *PHENIX* : building new software for automated crystallographic structure determination, *Acta Crystallogr. D Biol. Crystallogr.* **58**, 1948–1954 (2002).

59. M. C. F. Thomsen, M. Nielsen, Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion, *Nucleic Acids Res.* **40**, W281–W287 (2012).

60. R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, B. Peters, The immune epitope database (IEDB) 3.0, *Nucleic Acids Res.* **43**, D405–D412 (2015).

61. M. C. F. Thomsen, M. Nielsen, Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion, *Nucleic Acids Res.* **40**, W281–W287 (2012).

62. J. Robinson, J. A. Halliwell, J. D. Hayhurst, P. Flicek, P. Parham, S. G. E. Marsh, The IPD and IMGT/HLA database: allele variant databases, *Nucleic Acids Res.* **43**, D423–D431 (2015).

63. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.* **32**, 1792–1797 (2004).

64. F. Ronquist, M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, J. P. Huelsenbeck, MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space, *Syst. Biol.* **61**, 539–542 (2012).

65. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language, *Bioinformatics* **20**, 289–290 (2004).

Figure Legends

Figure 1. Structural comparison of HLA-C*08:02 in complex with therapeutic KRAS G12D 9mer and 10mer neoantigens.

a-b, Unambiguous 2Fo-Fc electron density maps of (**a**) KRAS G12D 9mer (GADGVGKSA, green) and (**b**)10mer (GADGVGKSAL, yellow) peptides from solved complex structures. The underlined amino acids represent the mutation in the peptides.

c, Overlay of the KRAS G12D 9mer and 10mer peptides.

d, Overlay of Cα traces of HLA-C*08:02 in complex with 9mer and 10mer peptides. These two structures show conformational differences at their peptide regions.

e, Sequence logo shows amino acid preference for 9mer peptide bounds to HLA-C*08:02. The peptide library obtained from IEDB, n=2321. Sequence logo was generated from Seq2Logo server.

f, Aromatic residues (green) from HLA-C*08:02 accommodating P1 and P2 residues of KRAS G12D 9mer peptide.

Figure 2. Different presentation patterns between C08-mut9m and C08-mut10m

a, Polar interactions between the HLA-C*08:02 molecule (grey) and 9mer (green), or HLA-C*08:02 interacts with 10mer (yellow) peptide at P1G and P2A positions.

b, KRAS G12D 9mer peptide interact with HLA-C*08:02 with the side chains P3D and P9A pointing down.

c, KRAS G12D 10mer peptide interact residues of HLA-C*08:02 with the side chains of P3D, P8S and P9A pointing down.

875

876 **Figure 3. Systematic analysis of IND and NND dataset.**

877 **a,** Neoantigen predicted affinity ratio. Dots represents the ratio of predicted affinity of wild type
878 peptide and predicted affinity of mutant peptide. All tested neoantigens were from the IND (n=128).
879 Affinity prediction was performed using NetMHC 4.0pan. Orange dots (n=26) represent the
880 neoantigens with less than (or equal to) 500 nM predicted affinity. Green dots (n=102) represent the
881 neoantigens with more than 500 nM predicted affinity.

882 **b,** Comparation of the mutation occurrence frequency at each position of 9mer peptides from IND (red
883 line) and NND (dark blue line). The grey dotted line represents the mutation theoretical frequency (1/9)
884 of 9mer peptide. Positions of 9mer peptide show on horizontal axis.

885 **c,** Donut plots showing the percentage of 9mer peptide mutation distribution from IND and NND based
886 on the three different classes: anchor position mutation, MHC contact position mutation and TCR
887 contact position mutation. (immunogenic data, n =51; non-immunogenic data, n =763).

888 **d,** 9mer peptides mutation distribution of three different classes (anchor position mutation, MHC
889 contact position mutation and TCR contact position mutation) from IND and NND. The percentage of
890 mutation distribution at anchor position did not shown significant difference between immunogenic and
891 non-immunogenic data (ns, no significance). The frequency of mutation distribution at TCR contact
892 region and MHC contact region shown significant difference (*$p<0.05$, **$P<0.01$; immunogenic data,
893 n =51; non-immunogenic data, n =763, Z-test).

894

895 **Figure 4. Similarity and preference of data from IND and NND.**

896 **a-c, a,** Similarity between wild type and mutant amino acids from immunogenic and non-immunogenic
897 peptides. Normalized similarity score was calculated based on the BLOSUM62 matrix (ns, p=0.3077, n
898 (immunogenic data) =128; n (non-immunogenic data) =2883, student t-test). **b,** Similarity between
899 wild type and mutant amino acids from anchor position mutated immunogenic and non-immunogenic
900 peptides (**p=0.010, n (immunogenic data) =29; n (non-immunogenic data) =426, student t-test). **c,**
901 Similarity of amino acids from 9mer TCR contact position mutated immunogenic and non-
902 immunogenic peptides (**p<0.05, n (immunogenic data) =40; n (non-immunogenic data) =426, student
903 t-test).

904 **d-f, d,** Comparation of HLA present tendentiousness between wild type and mutant 9mer neoantigens
905 from IND (n=69, **p=0.0088, student t-test); **e,** Comparation of HLA present tendentiousness between
906 wild type and mutant 9mer anchor mutated neoantigens from IND (n=17, ****p<0.0001, student t-
907 test); **f,** Comparation of HLA present tendentiousness between wild type and mutant 9mer TCR contact
908 position mutated neoantigens from IND (n=39, p=0.9358, student t-test).

909 **g,** Frequency of 4 different anchor mutation classes (non-preferential to non-preferential residues (N to
910 N), non-preferential to preferential residues (N to P), preferential to non-preferential residues (P to N)
911 and preferential to preferential residues (P to P)) of immunogenic data (from IND) and non-
912 immunogenic data (from NND). Each class is representing the preference alteration from wild type
913 counterparts to mutant peptides. The frequency between immunogenic and non-immunogenic data on
914 N to P class shows significant difference (***p<0.001, n (immunogenic data) = 27; n (non-
915 immunogenic data) =425, Z-test).

916 **h,** Pie charts represent the percentage of different anchor mutation classes. N to P class (n=270 and not

917 N to P class (n=425).

918

919 **Figure 5. HLA-C*05:01 is a potential allele to present KRAS G12D 10mer peptide.**

920 **a,** Heatmap illustrating the relative sequence similarity of HLA-C alleles. HLA-C*08:02 acts as the

921 reference allele.

922 **b,** Top (left panel) and side (right panel) view of HLA-C*08:02 molecule. Three residues (Arg97,

923 Tyr99 and Arg156) side chains from HLA-C*08:02 molecule contact to peptide P3 were shown as

924 spheres.

925 **c,** Top (left panel) and side (right panel) view of HLA-C*08:02 molecule. Two residues (Ser77 and

926 Asn80) side chains contact to peptide PΩ, which different from HLA-C*05:01 molecule were shown as

927 spheres.

928 **d,** Unambiguous 2Fo-Fc electron density maps of KRAS G12D 10mer peptide (GA<u>D</u>GVGKSAL) from

929 solved C05-mut10m complex structures.

930 **e,** Overlay of the peptide region of C05-mut10m (cyan) and C08-mut10m (yellow) complexes.

931 **f,** Overlay of C05-mut10m (cyan) and C08-mut10m (yellow) complexes.

932 **g,** Similar peptide binding pattern at P1-P3 region of C05-mut10m and C08-mut10m complexes.

933 Similar bonds between these two complexes were shown in red dot lines.

934 **h,** Comparation of C05-mut10m and C08-mut10m complexes at peptides P10L. The upper panel shows

935 P10L from C08-mut10m. The peptide and bonds show in yellow. The lower panel shows P10L C05-

mut10m. The peptide and bonds show in cyan.

Figure 6. Optimized procedure of neoepitope (neoantigen and corresponding HLA) and new HLA candidate discovery.

a, Workflow of neoepitope discovery. Conventional procedures were indicated in solid box. Propositional procedures in our study were indicated in dotted box.

b, Interpretation of neoantigen candidate filtration procedures based on ours studies. Different classes of neoantigen candidates filtrate with different strategies during neoantigen candidate filtration.

Supplementary Figure 1. Structure of H-2 Kb in complex with DPAGT V213L wild type peptide wt8mV and mutant mut8mL.

a, Unambiguous 2Fo-Fc electron density maps of DPAGT1 wild type 8mer peptide (wt8mV peptide SIIVFNLV, magenta) and mutant 8mer peptide (mut8mL peptide SIIVFNL<u>L</u>, orange) from solved structures.

b, Overlay of the DPAGT1 wt8mV (magenta) and mutant mut8mL (orange) peptides.

c, Sequence logo based on the data from IEDB showing amino acid preferences for 8mer peptides bound to H-2 Kb. The peptide library was obtained from IEDB (n=4141).

d, Anchor residue P8V of wt8mV peptide interacts with H-2 Kb (grey).

e, Anchor residue P8L of mut8mLpeptide interacts with H-2 Kb (grey).

Supplementary Figure 2. Comparation of the mutation distribution of 10mer peptides from IND and NND.

Comparation of the mutation occurrence frequency at each position of 10mer peptides from IND (red line) and NND (dark blue line). The grey dotted line represents the mutation theoretical frequency (1/10) of 10mer peptide. Positions of 10mer peptide show on horizontal axis.

Supplementary Figure 3. Amino acids frequency of mutant peptides from IND and NND.

The horizontal axis depicts the amino acid codes of mutations. The vertical axis indicates the frequency of mutations (***$p<0.001$, n (immunogenic data) =129; n (non-immunogenic data) =11739, Z-test).

Supplementary Figure 4. 9mer peptides binding profile of 29 HLA alleles.

Sequence logos showing amino acid binding profile for 9mer peptides bound to 29 HLA alleles. The peptide libraries were obtained from IEDB. The sample size of each allele is shown on the figure. Sequence logos were generated from peptide-binding matrices using the Seq2Logo server.

Supplementary Figure 5. Phylogenetic analysis of 15 HLA-C alleles.

15 HLA-C sequences were aligned and then used to make phylogenetic relationship. Alleles in the red box represent the relatively far genetic relationship from HLA-C*08:02.

Supplementary Figure 6. Multiple sequence alignment of HLA-C*08:02 and HLA-C*05:01

Sequence Alignment was done using Clustal Omega. Position 77 (Ser to Asn) and position 80 (Asn to Lys) are shown without underpainting.

Supplementary Figure 7. Different peptide-MHC models with their potential immunogenicity.

Model A demonstrates the peptides which have mutations at TCR contact positions and presented by MHC. The blue circle represents wild type residues. The red triangle represents the mutant residues with far similarities with the wild type. Model B demonstrates the peptides which have mutations at MHC contact position and presented by MHC. The red triangle represents the mutant residues with far similarities with the wild type. Model C demonstrates the peptides which have mutations at anchor position and are presented by MHC. The yellow rectangle represents wild type residues which are not preferential in anchor position for related MHC alleles and thus hardly to bind and present by HLA. The purple circle represents mutated residues which are preferential selected by related MHC alleles at anchor position.

Figure 1

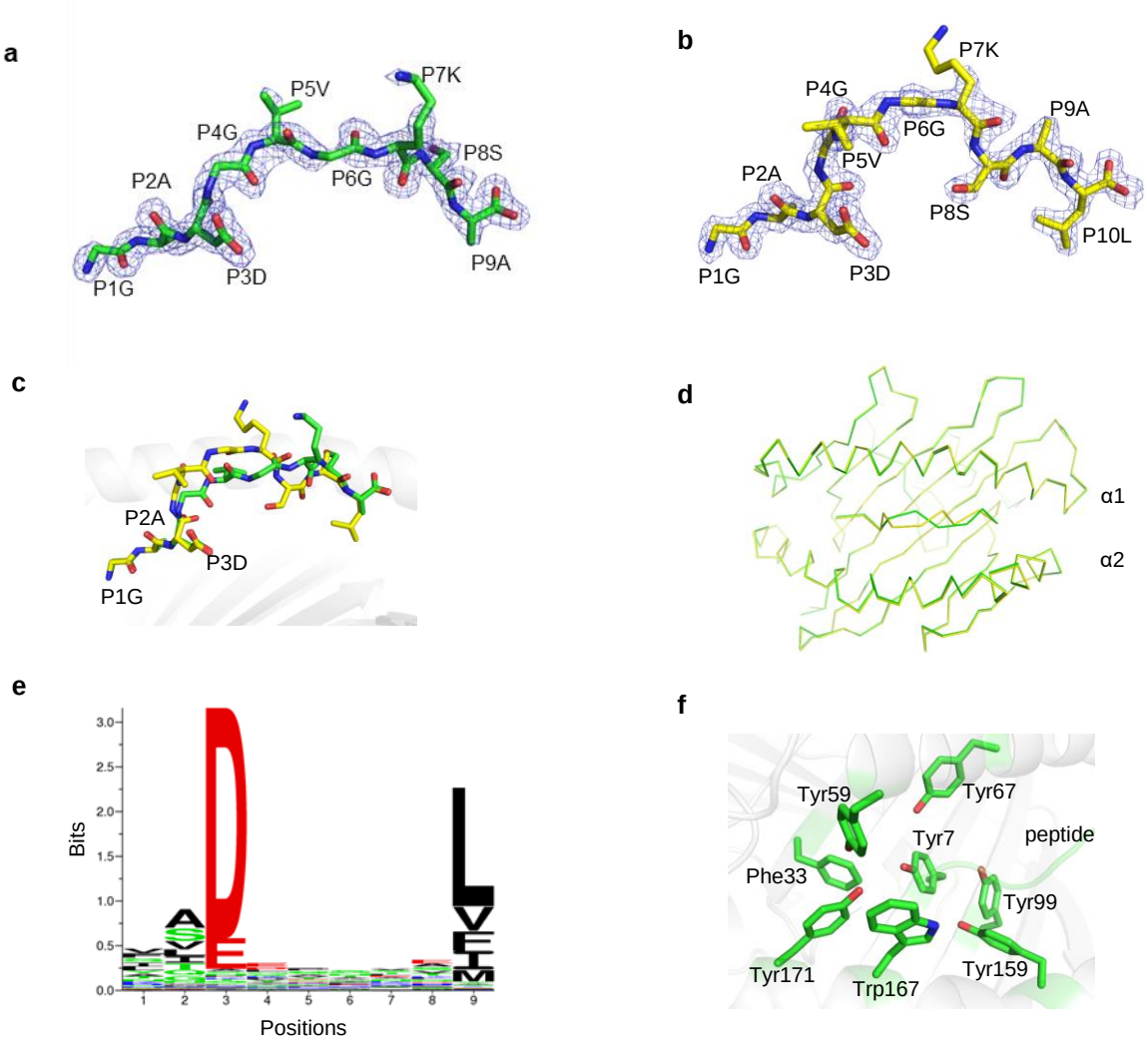
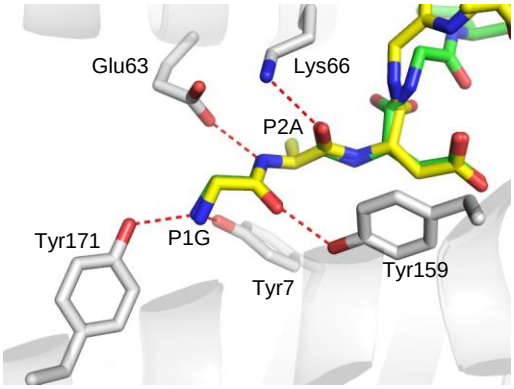
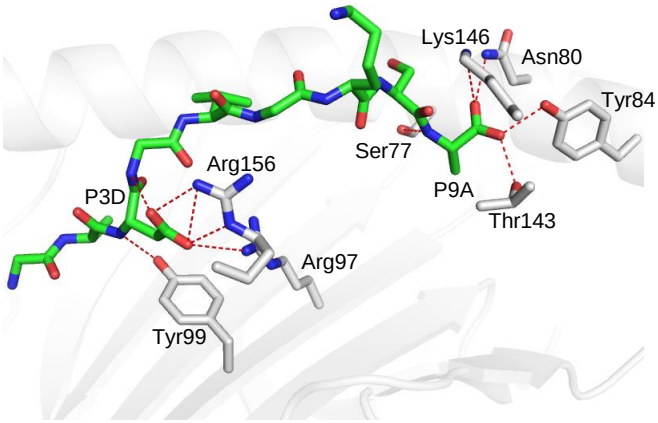

a
P7K
P5V
P4G
P8S
P2A
P6G
P9A
P1G
P3D
b
P7K
P4G
P9A
P6G
P5V
P2A
P8S
P10L
P1G
P3D
c
P2A
P3D
P1G
d
α1
α2
e
Bits
Positions
f
Tyr59
Tyr67
Tyr7
peptide
Phe33
Tyr99
Tyr171
Trp167
Tyr159

Figure 2

a


Glu63
Lys66
P2A
Tyr171
P1G
Tyr7
Tyr159

b


Lys146
Asn80
Tyr84
Ser77
Arg156
P3D
P9A
Thr143
Arg97
Tyr99

c

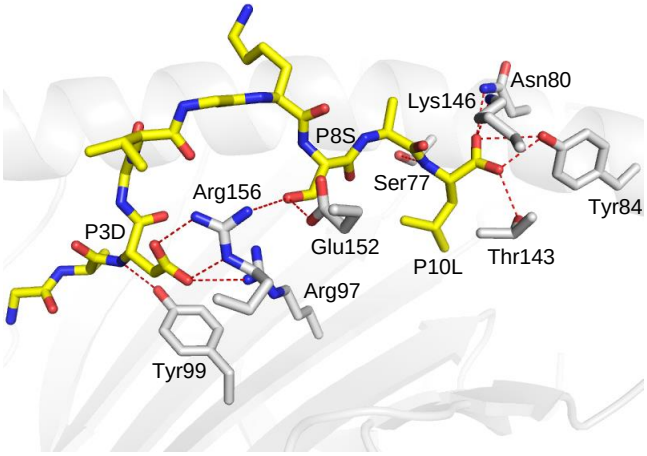

Asn80
Lys146
P8S
Ser77
Arg156
Tyr84
P3D
Glu152
P10L
Thr143
Arg97
Tyr99

Figure 3

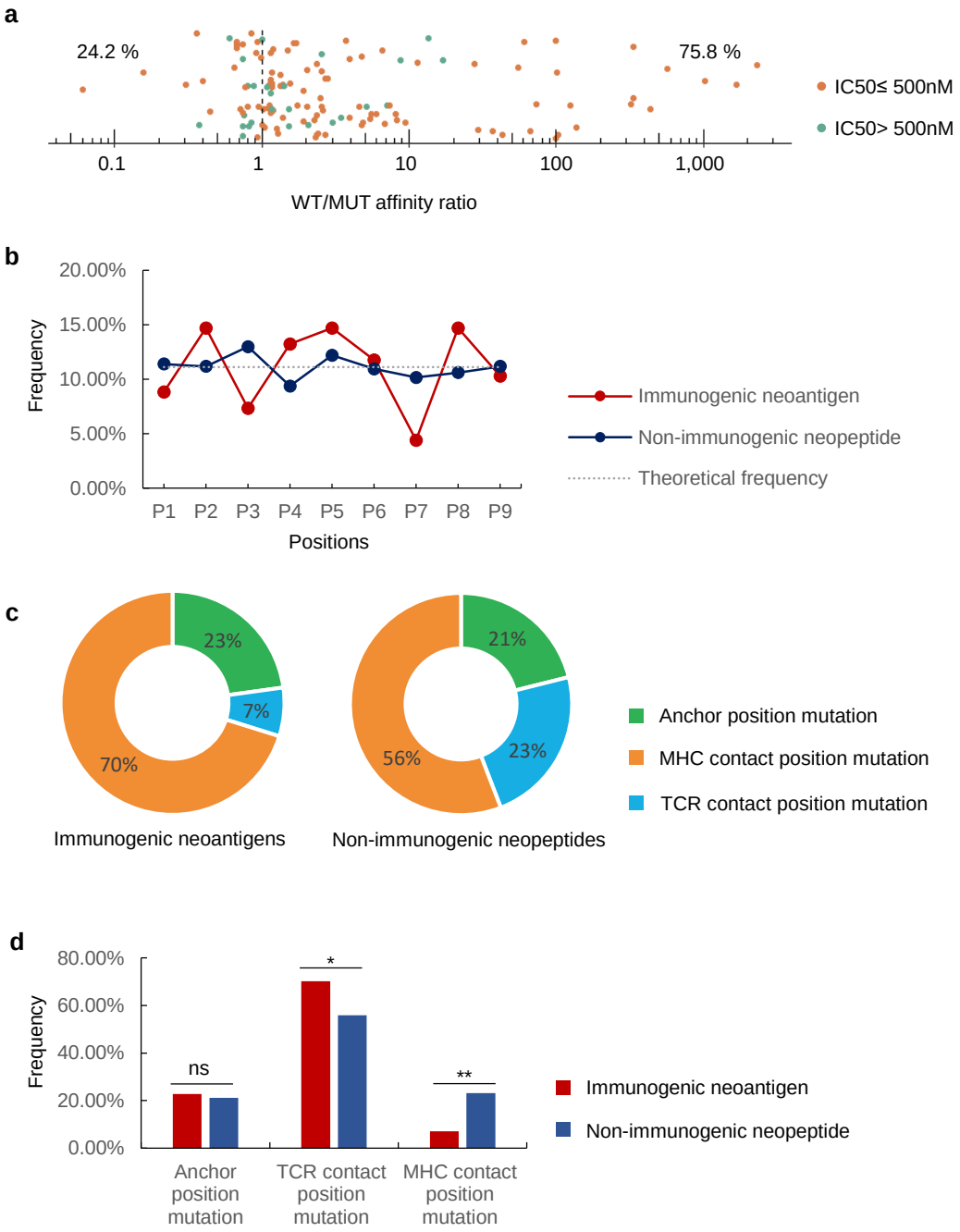

a
24.2 %
75.8 %
IC50≤ 500nM
IC50> 500nM
0.1
1
10
100
1,000
WT/MUT affinity ratio
b
20.00%
15.00%
10.00%
5.00%
0.00%
Frequency
P1 P2 P3 P4 P5 P6 P7 P8 P9
Positions
Immunogenic neoantigen
Non-immunogenic neopeptide
Theoretical frequency
c
23%
7%
70%
21%
23%
56%
Anchor position mutation
MHC contact position mutation
TCR contact position mutation
Immunogenic neoantigens
Non-immunogenic neopeptides
d
80.00%
60.00%
40.00%
20.00%
0.00%
Frequency
ns
*
**
Anchor position mutation
TCR contact position mutation
MHC contact position mutation
Immunogenic neoantigen
Non-immunogenic neopeptide

Figure 4

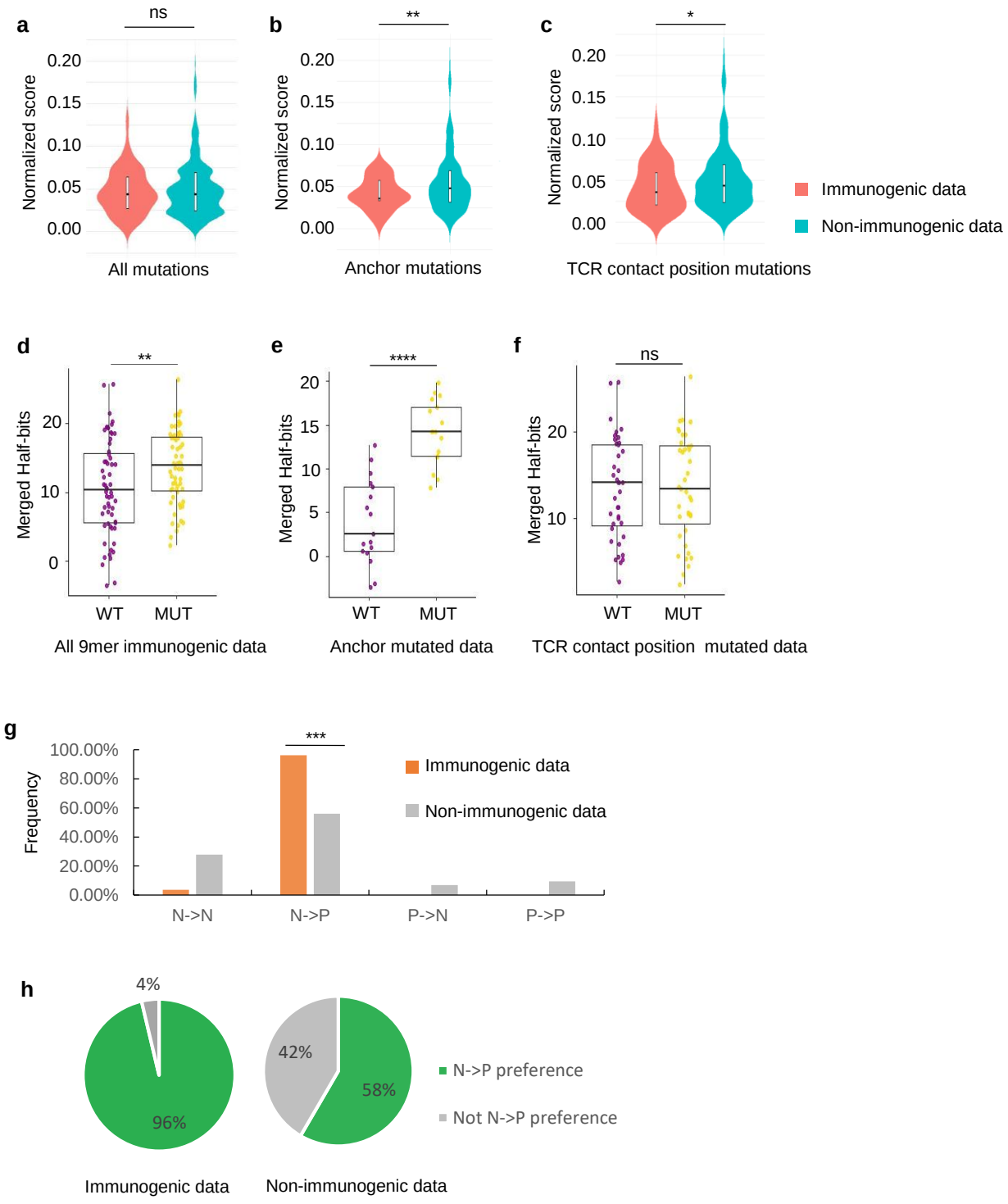

a
ns
Normalized score
0.20
0.15
0.10
0.05
0.00
All mutations
b
**
Anchor mutations
c
*
TCR contact position mutations
Immunogenic data
Non-immunogenic data
d
**
Merged Half-bits
20
10
0
WT
MUT
All 9mer immunogenic data
e
****
20
15
10
5
0
Anchor mutated data
f
ns
TCR contact position mutated data
g
***
Frequency
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
N->N
N->P
P->N
P->P
h
4%
96%
42%
58%
N->P preference
Not N->P preference
Immunogenic data
Non-immunogenic data

Figure 5

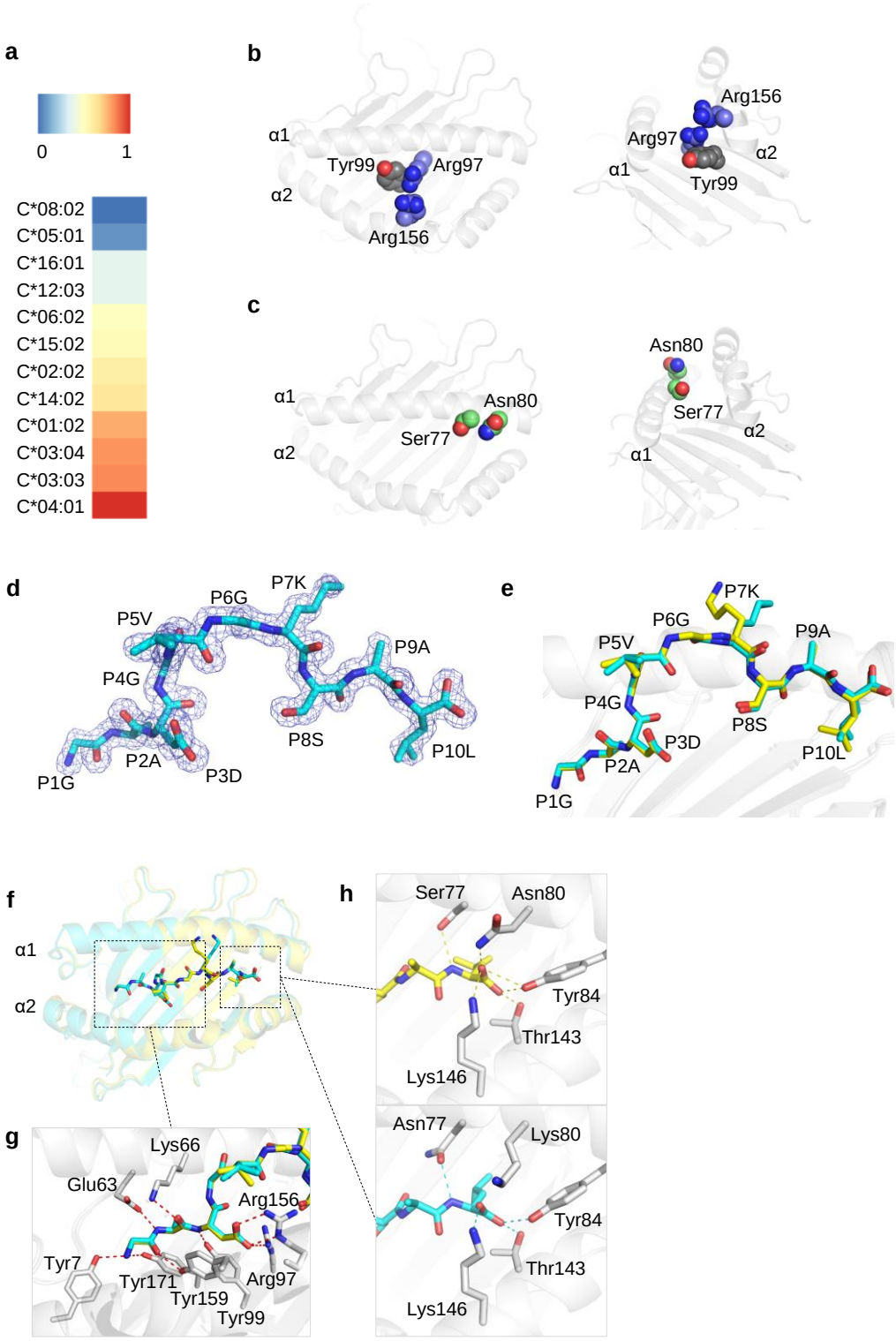
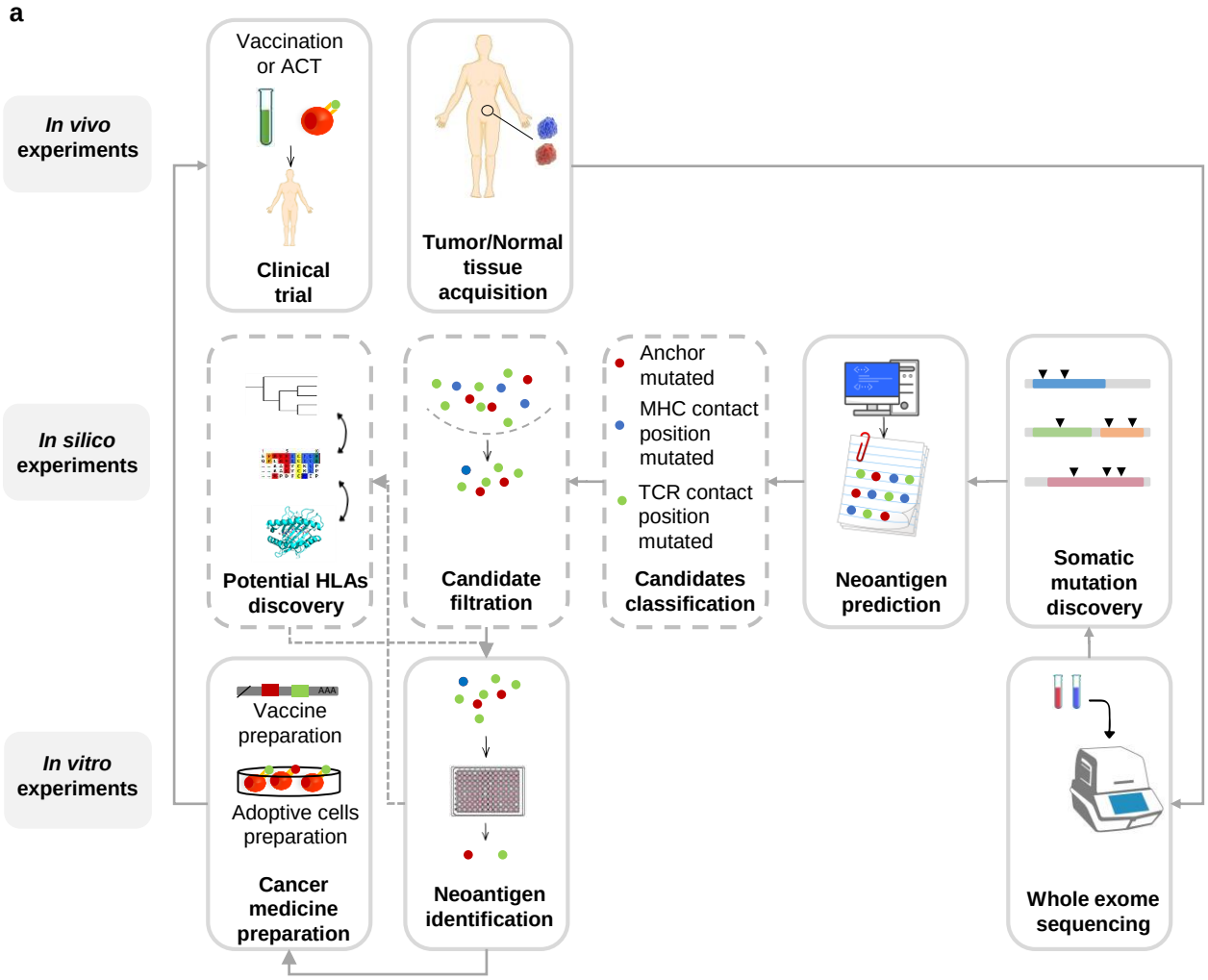

a
0
1
C*08:02
C*05:01
C*16:01
C*12:03
C*06:02
C*15:02
C*02:02
C*14:02
C*01:02
C*03:04
C*03:03
C*04:01
b
α1
α2
Tyr99
Arg97
Arg156
Arg156
Arg97
α2
α1
Tyr99
c
α1
α2
Asn80
Ser77
Asn80
Ser77
α2
α1
d
P1G
P2A
P3D
P4G
P5V
P6G
P7K
P8S
P9A
P10L
e
P1G
P2A
P3D
P4G
P5V
P6G
P7K
P8S
P9A
P10L
f
α1
α2
g
Lys66
Glu63
Arg156
Tyr7
Tyr171
Tyr159
Tyr99
Arg97
h
Ser77
Asn80
Tyr84
Thr143
Lys146
Asn77
Lys80
Tyr84
Thr143
Lys146

Figure 6

a



b

	Candidates classification	Candidates filtration methods	Candidates eliminated
Anchor mutated neopeptide	●●●	Preference selection	●● *Decrease ~42%
MHC contact position mutated neopeptide	●●	Importance selection Preference selection Morphologic selection	●
TCR contact position mutated neopeptide	●●●●●	Dissimilarity selection Morphologic selection	●●●●