# A high-throughput yeast display approach to profile pathogen proteomes for MHC-II binding

Brooke D. Huisman[1,2], Zheng Dai[3,4], David K. Gifford[2,3,4], Michael E. Birnbaum[1,2,5,*]

[1] Koch Institute for Integrative Cancer Research, Cambridge, MA, USA
[2] Department of Biological Engineering, MIT, Cambridge, MA, USA
[3] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
[4] Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA
[5] Ragon Institute of MIT, MGH, and Harvard, Cambridge, MA, USA

* Corresponding author: mbirnb@mit.edu

**Author contributions**

Conception of project: B.D.H. and M.E.B. Conducting experiments: B.D.H. Data analysis: B.D.H. and Z.D. Supervision of work: D.K.G. and M.E.B. Writing manuscript: B.D.H. and M.E.B. Editing manuscript: all authors.

**Competing Interest Statement**

D.K.G. is a founder of ThinkTx. M.E.B. is an equity holder in 3T Biosciences, and is a co-founder of Viralogic Therapeutics and Abata Therapeutics. The other authors declare no competing interests.

**Keywords:** yeast surface display, antigen prediction, peptide-MHC

**Abstract**

T cells play a critical role in the adaptive immune response, recognizing peptide antigens presented on the cell surface by Major Histocompatibility Complex (MHC) proteins. While assessing peptides for MHC binding is an important component of probing these interactions, traditional assays for testing peptides of interest for MHC binding are limited in throughput. Here we present a yeast display-based platform for assessing the binding of tens of thousands of user-defined peptides in a high throughput manner. We apply this approach to assess a tiled library covering the SARS-CoV-2 proteome and four dengue virus serotypes for binding to human class II MHCs, including HLA-DR401, -DR402, and -DR404. This approach identifies binders missed by computational prediction, highlighting the potential for systemic computational errors given even state-of-the-art training data, and underlines design considerations for epitope identification experiments. This platform serves as a framework for examining relationships between viral conservation and MHC binding, and can be used to identify potentially high-interest peptide binders from viral proteins. These results demonstrate the utility of this approach for determining high-confidence peptide-MHC binding.

## Introduction

Major histocompatibility complex (MHC) proteins play a critical role in adaptive immunity by presenting peptide fragments on the surface of cells. Peptide-MHCs (pMHCs) are then surveilled by T cells via their T cell receptors (TCRs), enabling immune cells to sense dysfunction, such as the presence of pathogen-derived peptides (Chaplin, 2010; Hennecke and Wiley, 2001). Class II MHC molecules (MHC-II) are expressed primarily on professional antigen presenting cells, and are recognized by antigen-specific $CD4^+$ T cells that drive the coordination of innate and adaptive immune responses (Chaplin, 2010; Swain et al., 2012). MHC-II molecules have an open peptide-binding groove, allowing for display of long peptides, consisting of a 9 amino acid 'core' flanked by a variable number of additional residues on each side (Jones et al., 2006).

Generating reliable and rapid data on peptide-MHC binding is beneficial for understanding the underlying biology of adaptive immunity and for clinical applications, including for optimized T cell epitopes in vaccine design (Dai et al., 2021; Keskin et al., 2019; Liu et al., 2020; G. Liu et al., 2021; Moise et al., 2015; Ott et al., 2017; Patronov and Doytchinova, 2013; Rosati et al., 2021). In fact, therapeutics to generate antigen-specific T cell responses have shown great promise in cancer (Keskin et al., 2019; Ott et al., 2017) and infectious disease (Gambino et al., 2021). Since understanding peptide-MHC binding is critical for identifying and engineering T cell epitopes, there have been sustained efforts to produce high-quality experimental data and predictive algorithms.

Initial experimental methods for determining peptide binding to MHC relied upon the analysis of synthesized candidate peptides via MHC stability or functional assays, and can produce high-confidence data, but can be difficult to scale beyond a small number of candidate peptides (Altmann and Boyton, 2020; Justesen et al., 2009; Mateus et al., 2020; Sidney et al., 2010; Yin and Stern, 2014). More recently, mass spectrometry-based approaches have been demonstrated for determining the MHC-presented peptide repertoire of cells. These approaches include monoallelic mass spectrometry, which allows for the unambiguous assignment of presented peptides to a given MHC allele. However, mass spectrometry-based approaches are not necessarily quantitative measures of presented peptide affinity or abundance, although there have been advances in quantitation using internal standards (Stopfer et al., 2021, 2020). Additionally, the peptides endogenously expressed by a cell can crowd out exogenously examined peptides of interest, and mass spectrometry approaches typically require large numbers of input cells (Abelin et al., 2019, 2017; Parker et al., 2021; Purcell et al., 2019).

A wave of higher throughput approaches have been recently developed for studying peptide-MHC interactions, including yeast display (Jiang and Boder, 2010; R. Liu et al., 2021; Rappazzo et al., 2020) and mammalian display-based methods (Obermair et al., 2021). Several of these approaches circumvent the bottlenecks of synthesizing or identifying peptides by utilizing DNA-based inputs and outputs (Jiang and Boder, 2010; Obermair et al., 2021; Rappazzo et al., 2020). These assays rely upon libraries that are often generated via DNA oligonucleotide synthesis, and use peptide stabilization and surface expression (Jiang and Boder, 2010; R. Liu et al., 2021; Obermair et al., 2021) or peptide dissociation (Rappazzo et al., 2020) to assess peptide-MHC binding.

In addition to experimental advances, computational approaches for peptide-MHC binding prediction have advanced markedly over the past decade. These developments are due to algorithmic advances (O'Donnell et al., 2020; Racle et al., 2019; Reynisson et al., 2020; Zeng and Gifford, 2019) and the availability of large, high-quality training data (Abelin et al., 2019, 2017; Rappazzo et al., 2020; Reynisson et al., 2020). However, despite the improvements in predicting peptide binding to MHC in a broad sense, the predictive power for individual peptides often remain imperfect relative to experimental measurements (Rappazzo et al., 2020; Zhao and Sher, 2018).

Here we present a yeast display approach to directly assess peptide-MHC binding for large collections of defined peptide antigens to screen whole viral proteomes for MHC-II binding in high-throughput. We utilize this approach to screen the full proteome of SARS-CoV-2, a present, global threat to public health, and identify SARS-CoV-2-derived MHC binders missed by computational prediction. We additionally apply this approach to screen proteomes from serotypes 1-4 of dengue viruses, in which antibody dependent enhancement results in more severe disease upon second infection with a different dengue virus serotype (Guzman et al., 2016), and thus represents a potential important application area for T cell-directed therapeutics. Our approach enables exploration of peptide binding to MHCs in the context of serotype-specific mutations, identifying homologous, pan-serotype regions of interest that are capable of MHC binding and thus may represent desirable targets for immune interventions.

## Results

*Generation of yeast display libraries for profiling the SARS-CoV-2 proteome*

Previous studies have reported the use of yeast-displayed MHC-II for characterizing peptide-MHC and pMHC-TCR interactions (Birnbaum et al., 2017, 2014; Rappazzo et al., 2020). We adapted MHC-II yeast display constructs (Rappazzo et al., 2020) to generate a defined library of peptides that cover the SARS-CoV-2 proteome to assess them for MHC binding. To compare SARS-CoV-2 with a related coronavirus, we also included peptides from the spike and nucleocapsid proteins from SARS-CoV.

Each protein was windowed into peptides of 15 amino acids in length, with a step size of 1 to cover every possible 15mer peptide in the protein **(Figure 1a)**. Each peptide was encoded in DNA and cloned in a pooled format into yeast vectors containing MHC-II proteins. The generated library was linked to three MHC-II alleles: HLA-DR401 (HLA-DRA1*01:01, HLA-DRB1*04:01), HLA-DR402 (HLA-DRA1*01:01, HLA-DRB1*04:02), and HLA-DR404 (HLA-DRA1*01:01, HLA-DRB1*04:04). Yeast were formatted with a flexible linker connecting the peptide and MHC, containing a 3C protease site and a Myc epitope tag, which can be used for selections **(Figure 1a)** (Rappazzo et al., 2020). The final library contained 11,040 unique peptides, with 99% of the designed peptides present in each cloned yeast library, as assessed by next-generation sequencing.

*Strategies for selecting defined libraries*

To enrich for peptide binders, iterative selections were performed (**Figure 1a**): the library is first incubated with competitor peptide and 3C protease, which cleaves the covalent linkage between peptide and MHC, followed by the addition of HLA-DM at lower pH. These conditions allow for the encoded peptide to be displaced from the peptide-binding groove. The

129 Myc epitope tag is proximal to the peptide, which can be identified via incubation with an anti-
130 epitope tag antibody followed by enrichment via magnetic bead selection if the yeast-
131 expressed peptide remains bound to the MHC after the peptide exchange reaction.
132 Three rounds of selection were iteratively performed. Representative enrichment of
133 yeast expressing Myc-tagged peptides can be seen in **Figure 1c** ("undoped library"), for the
134 library displayed by HLA-DR401. Here the pre-selection Myc-positive population starts at 29.3%
135 and quickly converges, with 65.0% positive in the pre-selection Round 2 population and 74.1%
136 in the pre-selection Round 3 population.
137 Given the rapid convergence of the library, we performed a second set of selections in
138 which we doped the defined library into a randomized, null library to enable a greater degree of
139 enrichment as compared to non-binding peptides. The null library was generated by fully
140 randomizing ten amino acids in the peptide region of the peptide-MHC-II construct while fixing
141 three amino acids to encode stop codons. This library provides a baseline population of yeast
142 which should not express pMHC, and therefore not enrich in our selections. We doped our
143 defined peptide library into a 500-fold excess of null library, such that each peptide member
144 was represented at approximately the same frequency (**Figure 1b**). The null library provides
145 baseline competition, which true binders must enrich beyond, and increases the stringency of
146 the enrichment task.
147 We performed four rounds of selection on the doped library. Because of the excess of
148 null yeast, the initial pre-selection stain is low (1.6%) compared to the initial undoped library
149 (**Figure 1c**). This staining enriched over the first three rounds of selection, reflective of the
150 stringency of the task and clarity of enrichment. This is in contrast to the initial undoped library,
151 which began with a much higher pre-selection stain, with a lower fold-change in staining over
152 rounds of selection. The low frequency of each member in the starting doped library, however,
153 increases the likelihood of stochastic dropout for any given member.
154
155 *Analysis of selection data*
156 After selections, peptide identities were determined through deep sequencing of
157 enriched yeast populations, providing us with a dataset comprised of positive enrichment over
158 four rounds of selection from the doped library and both positive and negative enrichment for
159 three rounds of selection from the undoped library (**Supplemental Data**). **Supplemental Figure**
160 **1** shows the correlation between defined library members on HLA-DR401. As expected, the
161 unselected library correlated poorly with post-selection rounds. Consistent with the observed
162 staining (**Figure 1c**), the doped library essentially converged after Round 3. Similarly, the
163 undoped library appears converged following Round 2.
164 Next, we established metrics for enrichment for each mode of selection. Given the high
165 starting frequency of members in the undoped library, we classify enrichment based on fold
166 change between Round 1 and Round 2, and we define criteria for enriched yeast in the
167 undoped library as making up a higher fraction of reads following Round 2 compared to Round
168 1. In contrast, in the doped library, members start at low frequencies, and we define
169 enrichment based on presence above a threshold in Round 3 of selection, specifically as having
170 greater than or equal to 10 reads following Round 3. **Figure 2b** illustrates the correspondence
171 between enrichment metrics in the doped and undoped library for the library on HLA-DR401.
172 Of the 11,040 peptides in the library, 2,467 enriched in both the doped and undoped libraries

173 displayed by HLA-DR401 (**Figure 2a**). An additional 1,252 enriched in the doped library only and
174 797 enriched in the undoped library only.
175 Because the library is designed with a step size of one, we next utilized overlap between
176 adjacent peptides to determine high-confidence binders. This analysis allows us to address the
177 potential that peptide sequences could register shift in such a way that invariant portions of the
178 linker sequences could inadvertently be incorporated into the peptide-binding groove. To do
179 this, we develop and implement a smoothing method, examining overlapping peptides for
180 shared enrichment behavior. Classically, the strongest determinant of peptide affinity for an
181 MHC is the nine amino acid stretch sitting within the peptide-binding groove (Jones et al., 2006;
182 Stern, 1994), although proximal peptide flanking residues can also affect binding (Lovitch et al.,
183 2006; O'Brien et al., 2008; Zavala-Ruiz et al., 2004). In our libraries, a given 9mer is present in
184 seven overlapping 15mer peptides, and we calculate how many of these seven 15mers have
185 enriched. This calculation is shown schematically in **Supplemental Figure 2a** with toy sequences
186 and applied to enrichment data for SARS-CoV-2 nucleocapsid on HLA-DR401 in **Supplemental**
187 **Figure 2b**. Sequences with good 9mer cores should enrich along with neighboring sequences
188 with the same 9mer sequence. In contrast, sequences which enrich spuriously or due to linker
189 sequence in the peptide groove or other stochastic factors should have few neighbor sequences
190 also enriching. Thus, we define a cutoff for high confidence 9mer enrichment of five out of
191 seven 9mer-containing sequences enriching. This cutoff tolerates some stochastic dropout,
192 while still disallowing any cores that may solely enrich by register shifting the Gly-Ser linker
193 residues into the Position 9 pocket, which are favorable for each MHC allele in our study.
194 (Abelin et al., 2019; Rappazzo et al., 2020; Reynisson et al., 2020). Of the 2,467 peptides which
195 enriched in both the doped and undoped libraries for HLA-DR401, 1,791 also contain a 9mer
196 sequence which enriched in five or more peptides of the seven neighboring sequences
197 containing it (**Figure 2a**), with 676 peptides enriching in both doped and undoped libraries but
198 not containing a 9mer core enriched in five or more peptides, and 788 15mers containing a
199 9mer which enriched in five or more peptides but enriched in zero or one of the doped and
200 undoped libraries. These full relationships are captured in Venn diagrams in **Supplemental**
201 **Figure 3** for all three MHC alleles studied here.
202
203 *Sequence motifs of enriched peptides are consistent with known binders and highlight*
204 *considerations for designing epitope identification experiments*
205 To examine the 9mer core motifs of enriched peptides, we utilized a position weight
206 matrix method to infer the peptide register and generated visualizations of the 9mer cores
207 using Seq2Logo (Thomsen and Nielsen, 2012). **Figure 2c** shows a sequence logo of the aligned
208 9mer cores from the 2,467 15mer peptides which enriched on HLA-DR401 in both doped and
209 undoped libraries. The peptide motif is consistent with previously reported motifs for HLA-
210 DR401 (Abelin et al., 2019; Rappazzo et al., 2020): hydrophobic amino acids are preferred at P1,
211 acidic residues at P4, polar residues at P6, and small residues at P9. We also observe some
212 preference for glycine at P8 in the sequence logo, which is potentially an artifact of non-native
213 registers with linker at P8 and P9.
214 The other alleles used in the study, HLA-DR402 and HLA-DR404, have polymorphisms in
215 their peptide binding groove sequences as compared to HLA-DR401, which affect binding
216 preferences. HLA-DR401 differs from HLA-DR402 at four amino acids and from HLA-DR404 at

217 two amino acids, with all polymorphisms located in the beta chain. HLA-DR402 and HLA-DR404
218 share an amino acid distinct from HLA-DR401 affecting the P1 pocket (Gly86Val), resulting in a
219 preference for smaller hydrophobic residues (**Figure 3a**). Three polymorphisms in HLA-DR402
220 affect P4, P5, and P7 compared to HLA-DR401 (Leu67Ile, Gln70Asp, and Lys71Glu), while HLA-
221 DR404 has only one (Lys71Arg). Sequence logos for HLA-DR402 and HLA-DR404 are consistent
222 with previously reported motifs and MHC polymorphisms (**Supplemental Figure 4**). For HLA-
223 DR402, we observe less P4 preference compared to the motif of HLA-DR402 binders enriched
224 from a randomized yeast display peptide library (Rappazzo et al., 2020), albeit consistent with
225 mass spectrometry-generated motifs which also showed minimal P4 preference for HLA-DR402
226 (Abelin et al., 2019).
227 To explore differences between mass spectrometry, defined libraries, and random
228 libraries, and to probe the differing strengths of P4 peptide preference observed for HLA-DR402
229 between these modalities, we examined the compositions of randomized and defined libraries.
230 We hypothesized that skewed amino acid abundances in nature, which are reflected in the
231 defined library, could result in an apparent diminished amino acid preference. Indeed, three of
232 the most preferred P4 residues for binding HLA-DR402, Trp, His, and Met (Rappazzo et al.,
233 2020), are all low abundance in the SARS-CoV-2 proteome (Trp 1.1%, His 1.9%, Met 2.2%). In
234 comparison, a randomized peptide library for HLA-DR402 (Rappazzo et al., 2020) had a higher
235 representation of these amino acids (Trp 3.8%, His 2.9%, Met 3.8%). Additionally, the
236 randomized library had approximately nine thousand-fold more members than the defined
237 library, providing more instances of all amino acids. The low abundance and
238 underrepresentation of these amino acids likely underlies the apparent lack of amino acid
239 consensus at P4 in enriched peptides. Interestingly, Arg and Lys, which have also been reported
240 as preferred HLA-DR402 P4 residues, are more abundant than Trp, His, and Met in the SARS-
241 CoV-2 proteome (Arg 3.4% and Lys 5.9%; compare to Arg 9.7%, Lys 4.0% in the random library),
242 but still show less representation at P4 in the defined library enriched peptides compared to
243 the random library-enriched peptides. These differences in motifs between randomized and
244 defined libraries highlight the utility of randomized libraries for downstream applications such
245 as training prediction algorithms. Approaches influenced by amino acid abundance in nature,
246 such as defined libraries and mass spectrometry approaches, could inadvertently bias against
247 possible binders because of absence of amino acids in their null distribution, rather than true
248 binding preference.
249 Next, we wanted to examine the distribution of peptides among the possible 9mer
250 registers along each 15 amino acid sequence. Based on our register inference, of the 2,467
251 enriched peptides from the HLA-DR401 library, 1,610 peptides bound native 9mer cores
252 without using any linker sequence residues in the 9mer core, which is consistent with
253 theoretical ratios of possible native and non-native cores for a given 9mer (**Supplemental**
254 **Data**). The peptides with predicted native 9mer cores were approximately equally distributed
255 between possible registers, with the exception of the N-terminal register, which had one-third
256 fewer peptides. This register had only a single N-terminal flanking residue (a fixed Ala), which is
257 likely disfavored.
258 Because the library was designed with step size of one, many of the 9mer cores will be
259 repeated among neighboring peptides. Of the 1,610 HLA-DR401 peptides which enriched using
260 a native 9mer core, there are 563 unique 9mer cores identified through register-inference.

261 **Table 1** summarizes enrichment for each protein included in the library, highlighting the
262 number of 15mers which enriched in both the doped and undoped libraries, the number of
263 unique native 9mer cores, and the number of 15mers containing a 9mer enriched in at least
264 five of seven overlapping peptides.
265
266 *Examining relationships between MHC-specific binding and spike proteins from SARS-CoV-2 and*
267 *SARS-CoV*
268 To further explore relationships between the MHCs studied here and their virally-
269 derived peptide repertoires, we compared the binding of SARS-CoV-2 and SARS-CoV spike
270 proteins to all three MHC alleles. Sequence alignment of these three MHC alleles is shown in
271 **Figure 3a**, with polymorphic regions highlighted on an HLA-DR401 structure (adapted from PDB
272 1J8H). Interplay between viral conservation and binding are illustrated in **Figure 3b**, highlighting
273 conserved regions of the proteome in black and binders to each allele in grey, red, and blue.
274 Regions are highlighted where sequences enrich in overlapping peptides; that is, for each 9
275 amino acid stretch along the proteome, we calculated how many of the seven 15mer peptides
276 enrich in the yeast display assay, and if a 9mer enriched five or more times, it is marked as a hit.
277 Specific examples of these relationships are probed in **Figure 3c, d, and e**, where individually
278 enriched 15mer sequences are represented as horizontal lines above 15mer stretches in the
279 proteome. Bolded 9mers are identified through register inference as consensus binding cores
280 for these peptides. Only 15mers which contain the bolded 9mer are included in this
281 representation. Non-conserved amino acids within this 9mer are highlighted in yellow.
282 **Figure 3c** illustrates a region that is not conserved between SARS-CoV-2 and SARS-CoV,
283 where the SARS-CoV-2 peptides containing the core IYQAGSTPC are enriched for binding to all
284 three MHCs, but mutations, including at both P1 and P4 to Proline, discourage binding of the
285 aligned SARS-CoV peptide. **Figure 3e** illustrates a core that is conserved between SARS-CoV and
286 SARS-CoV-2, which can bind only to HLA-DR401, but not to HLA-DR402 or HLA-DR404, likely due
287 to the size of the P1 hydrophobic residue and, for HLA-DR402, the acidic P4 residue. **Figure 3d**
288 illustrates relationships between both viral conservation and MHC preference. In **Figure 3d**, the
289 SARS-CoV peptides containing the core IKNQCVNFN can bind to all three alleles. However, the
290 aligned SARS-CoV-2 peptides containing the core VKNKCVNFN do not bind to HLA-DR401, likely
291 because of the less preferable P1 Valine and basic P4 Lysine, but can bind to HLA-DR402, which
292 prefers these residues. These peptides can bind to HLA-DR404, although only four of the
293 adjacent peptides containing this core enrich, which is below the cutoff of five or more, and
294 since no other adjacent peptides enriched, this would not have been classified as a binder
295 (reflected in **Figure 3b**). This marginal, but below-threshold binding is logical, given that the P4
296 pocket for HLA-DR404 is similar to HLA-DR401, which does not prefer P4 Lysine, but HLA-DR404
297 has the same P1 binding pocket as HLA-DR402, which both prefer the P1 Valine in the SARS-
298 CoV-2 peptide.
299
300 *Identifying peptide binders missed by computational prediction*
301 Next, we compared our direct experimental assessments with results from
302 computational MHC binding predictions. Prediction algorithms allow for rapid computational
303 screening of potential peptide binders (Abelin et al., 2019; Reynisson et al., 2020), although
304 they can contain systemic biases (Rappazzo et al., 2020). To test the outputs of our direct

305 assessment approach and computational prediction algorithms, we assessed binding of several
306 peptides using a fluorescence polarization competition assay to determine $IC_{50}$ values, as
307 described previously (Rappazzo et al., 2020; Yin and Stern, 2014). Yeast-formatted peptides
308 (Ala+15mer+Gly+Gly+Ser) from SARS-CoV-2 spike protein were run through NetMHCIIpan4.0
309 for binding to HLA-DR401, with binders defined as having $\leq$ 10% Rank (Eluted Ligand mode).
310 Yeast display binders to HLA-DR401 were defined via the stringent criteria of 1) enriching in
311 both in doped and undoped selections, and 2) containing a 9mer that enriched in five or more
312 of the overlapping seven 15mers. 15mers were selected such that they could contain a
313 maximum overlap of 8 amino acids with other selected peptides, to avoid selecting peptides
314 with redundant 9mer cores. An length-matched version of the commonly studied Influenza A
315 $HA_{306-318}$ peptide (APKYVKQNTLKLATG) known to bind HLA-DR401 (Hennecke and Wiley, 2002;
316 Rappazzo et al., 2020) was included as a positive control, along with sequences that yeast
317 display and NetMHCIIpan4.0 both classified as either binders or non-binders. **Supplemental**
318 **Figure 5** shows a comparison of yeast-enriched and NetMHCpan4.0 predicted binders, with
319 boxed sequences selected for testing by fluorescence polarization.
320 The resulting fluorescence polarization $IC_{50}$ data from the native 15mer peptides are
321 shown in **Table 2** and **Supplemental Figure 6**. Peptides which both enriched in yeast display and
322 were predicted by NetMHCIIpan4.0 to bind ('Agreed Binders') all showed $IC_{50}$ values consistent
323 with binding, each with $IC_{50}$ < 2.2 µM. Similarly, peptides which were agreed non-binders
324 showed no affinity for HLA-DR401, with $IC_{50}$ > 50 µM.
325 All 8 'Yeast-Enriched Binders', which enriched in the yeast display assay but were not
326 predicted to bind via NetMHCIIpan4.0, showed some degree of binding, with $IC_{50}$ values
327 distributed from 14 nM (higher affinity than the HA control peptide) to 18 µM (weak, but
328 measurable, binding). Retrospectively, the weakest two binders appear to be enriching in the
329 yeast display assay using the peptide linker or have a binding core offset from center.
330 Interestingly, NetMHCIIpan4.0 predictions on the peptides identified via yeast display proved
331 highly sensitive to the length or content of the flanking sequences: if we repeat predictions on
332 only the antigen-derived 15mer sequences without the flanking sequences, NetMHCIIpan4.0
333 recovers four of its former false negative peptides (**Table 3**; peptides listed at the top in each
334 section of the table). We will refer to these four peptides as 'flank-sensitive centered peptides',
335 as they each have the consensus 9mer core centered in the peptide.
336 To further investigate the relationship with flanking residues, we selected five additional
337 peptides ('offset peptides') matching three criteria; these offset peptides were 1) enriched in
338 the yeast display assay, 2) share an overlapping core with the four flank-sensitive centered
339 peptides, but are 3) not predicted by NetMHCIIpan4.0 to be binders (either with or without
340 invariant flanking sequence added). All five offset peptides have their predicted cores offset by
341 1-2 amino acids from center, leaving at minimum 1 amino acid on both ends of the 9mer core
342 for each peptide. All five offset peptides exhibit some binding, with $IC_{50}$ values below 13 µM.
343 Each peptide is lower affinity than its overlapping centered counterpart, illustrating effects of
344 flanking residues on peptide binding, although some over-estimation of these effects in
345 NetMHCIIpan4.0 predictions are present.
346 We tested three 'NetMHC-Predicted Binders', which were predicted to bind by
347 NetMHCIIpan4.0, but were not enriched (nor did any neighboring sequences within an offset of
348 4 amino acids) in the yeast display assay (**Table 2**). Of these, one bound to HLA-DR401 ($IC_{50}$ 475

349 nM), while two showed minimal binding with IC50 > 35 µM, which is above the maximum 20
350 µM concentration tested. All three were predicted by NetMHCIIpan4.0 to bind with or without
351 the invariant flanking sequences (Eluted ligand mode % Rank: 5.7, 4.1, 8.7 (with flanking
352 residues) and 2.3, 0.6, 7.0 (without flanking residues), for ELDKYFKNHTSPDVD,
353 LQSYGFQPTNGVGYQ, and KTQSLLIVNNATNVV, respectively).
354 Of the eight 'Yeast-Enriched Binders' in **Table 2**, six contain cysteine residues, which
355 have been shown to be systematically absent from other datasets, including those from mono-
356 allelic mass spectrometry (Abelin et al., 2019; Barra et al., 2018), yet present in yeast display-
357 derived datasets (Rappazzo et al., 2020). To test for non-specific binding due to cysteine, two
358 cysteine-containing 'Agreed Non-Binders' were also tested and showed no affinity for HLA-
359 DR401, suggesting that cysteine itself is not causing non-specific binding. In the fluorescence
360 polarization dataset, the highest affinity binder (14 nM) contained cysteine and was missed by
361 NetMHCIIpan4.0 predictions (Eluted ligand mode % Rank: 71 (with flanking residues) and 28
362 (without flanking residues)).
363 The relationship between measured $IC_{50}$ values and NetMHCIIpan4.0 predicted values
364 for all 15mer SARS-CoV-2 spike peptides tested is shown in **Figure 4** and **Supplemental Figure 7**.
365
366 *Comparing whole dengue serotype proteomes for common MHC-binding peptides*
367 Defined yeast display libraries can generate data for diverse objectives. Dengue viruses
368 typically cause most severe disease after a second infection with a serotype different from the
369 first infection, due to antibody dependent enhancement (Guzman et al., 2016), which makes T
370 cell-directed therapeutics a potentially attractive means of combatting disease. To profile and
371 compare MHC binding across serotypes, we generated libraries containing 12,672 dengue-
372 derived peptides, covering the entire proteomes of dengue serotypes 1-4. These libraries were
373 on HLA-DR401 and HLA-DR402 and had coverage of 98% and 96% of the dengue library
374 members after construction, respectively.
375 Peptides from homologous regions of the four dengue serotypes have different MHC
376 binding ability, as illustrated in **Figure 5a** for binding to HLA-DR401. The proteins encoded in the
377 dengue genome are indicated along the horizontal axis (C: capsid; M: membrane; E: envelope;
378 NS: nonstructural proteins). Peptides that enriched in the yeast display assay are marked by a
379 line (serotype 1 in blue, serotype 2 in purple, serotype 3 in red, and serotype 4 in grey). The
380 proteome is smoothed to 9 amino acid stretches (as in **Figure 3b**), with a given 9 amino acid
381 region marked as a hit if five or more of the seven adjacent peptides enrich. For each 9mer, the
382 maximum number of serotypes with a conserved identical 9mer at that position is indicated at
383 the top in black.
384 These data can reveal relationships between conservation and binding ability. **Figure 5b-**
385 **d** shows enrichment data for individual 15mer peptides, with consensus inferred 9mer cores in
386 bold and non-conserved amino acids in these cores highlighted in yellow, as in **Figure 3c-e**.
387 Conserved cores which show binding ability (**Figure 5c**) may be ideal T cell targets. However,
388 the permissiveness of the binding groove allows for peptides to bind that have mutations at the
389 anchors, such as in NS5 (**Figure 5d**), where P4 Asn and P4 Met both allow binding. Interestingly,
390 the serotype 3 core (LASNAICSA) only enriched in four peptides, which is below our described
391 cutoff for high-confidence peptide cores. However, three adjacent peptides enriched and
392 register-inference for these peptides identifies the non-native, linker-containing version of the

393 LASNAICSA core as binding in the MHC-binding groove. This results in an adjacent 9mer being
394 highlighted as a binder in this region (**Figure 5a**) because overlapping 15mers enrich in five or
395 more of the seven adjacent peptides. With this in mind, care must be taken for core
396 identification in enriched regions and can be aided by coupling enrichment with register-
397 inference of enriched peptides. Further, we can also see relationships between conservation
398 and binding in non-conserved regions, such as in the envelope protein (**Figure 5b**) with the
399 mutations in serotype 3 enabling binding.
400
401 **Discussion**
402 $CD4^+$ T cell responses play important roles in infection, autoimmunity, and cancer. By
403 extension, understanding peptide-MHC binding is critical for identifying and engineering T cell
404 epitopes. Here we present an approach to directly assess defined libraries of peptides covering
405 whole pathogen proteomes for binding to MHC-II proteins. We examine alternative modes of
406 selection and utilize overlapping peptides to determine high-confidence binders. We
407 demonstrate the utility of this approach by identifying binders that are missed by prediction
408 algorithms, highlighting a prediction algorithm bias against cysteine-containing peptides and
409 sensitivity to peptide flanking residues (**Table 2** and **Table 3**). Finally, this approach can be
410 utilized for different objectives, including comparing binding to multiple MHC alleles (**Figure 3**)
411 or comparing peptides from related pathogen sequences for MHC-II binding (**Figure 5**). Whole
412 protein- or proteome-scale analysis across related viruses provides insight into relationships
413 between conserved epitopes and MHC binding (**Figure 3b, 5a**) and specific examples validate
414 the consistency with the underlying biophysics of peptide-MHC binding (**Figures 3c-e** and **5b-d**).
415 This approach for direct assessment shows benefit compared to prediction algorithms
416 for identifying binders, particularly for finding weak peptide binders. The overlapping peptides
417 in our library were useful for identifying enriched cores, especially when combined with our
418 register inference to identify consensus cores shared between these overlapping peptides.
419 NetMHCIIpan4.0 exhibits a sensitivity to length and register, which may cause users to miss
420 binders, albeit potentially of lower affinity. Of the overlapping peptides we tested to study this
421 phenomenon, NetMHCIIpan4.0 correctly ranked the affinities of the overlapping peptides
422 (**Table 3**), but missed binders. **Supplemental Figure 5** also highlights the sensitivity of
423 NetMHCIIpan4.0 to flanking sequences, where neighboring peptides with shared cores often
424 are not predicted to bind, resulting in fewer clusters of peptides in **Supplemental Figure 5**.
425 Our work reveals insights on the design of epitope identification experiments, including
426 the utility of overlapping peptides and considerations for comparing libraries of unbiased and
427 proteome-derived peptides. Design of defined libraries with sources of redundancy, such as
428 overlapping peptides, was critical for determining binders with higher degrees of confidence
429 and allowed us to apply stringent cutoffs for individual peptides. Overlapping peptides allowed
430 us to account for construct-specific confounding effects, such as the peptides binding using
431 non-native residues in the linker. Future iterations can change the sequence of the linker, such
432 as defining favorable P(-1) and P10 anchors to fix the register (Rappazzo et al., 2020), although
433 these adaptations would likely require MHC-specific knowledge in advance and may need to be
434 altered for different MHCs. Additionally, the engineered redundancy and multiple modes of
435 selection result in hyperparameters that can be tuned to meet users' stringency requirements,
436 such as defining different thresholds for calling individual 15mer binders or alternative

437 integration of overlapping binders. Additionally, our comparison of unbiased and proteome-
438 derived libraries highlights how aggregate motifs may be affected by underlying amino acid
439 preferences found in protein sequences themselves, which may inadvertently disfavor
440 sequences that can bind strongly to MHC molecules yet consist of amino acid covariates that
441 are not as commonly found in proteins.
442 Further, this approach can be used to study MHC binding between similar viruses, as
443 done with the dengue proteomes and the spike proteins from SARS-CoV-2 and SARS-CoV,
444 highlighting regions where mutations disrupt binding as well as regions where binding is
445 unperturbed. This method can also be rapidly adapted to study future sequences if pathogens
446 evolve over time.
447 As experimental approaches and computational approaches continue to co-develop,
448 they present complementary benefits. Though this platform allows for rapid assessment of
449 peptide-MHC binding, the speed of computational prediction surpasses experimental
450 approaches. NetMHCIIpan4.0 prediction and yeast display selections identified sets of non-
451 overlapping misses, highlighting a utility for both. Additionally, all agreed binders and non-
452 binders matched fluorescence polarization results, suggesting a consensus of yeast display
453 enrichment and algorithmic prediction provide high-confidence results. Approaches such as
454 yeast display assessment can be used to complement computational approaches, such as for
455 identifying cysteine-containing peptides which are still under-predicted by algorithms. Similarly,
456 prediction algorithms can be trained using large, quality datasets to account for biases. In
457 another application, our platform to assess peptide-MHC binding can be used to design high-
458 throughput assays to test peptide immunogenicity in clinical samples (Klinger et al., 2015;
459 Snyder et al., 2020).
460 Defined yeast display peptide libraries can also be readily applied to identification of T
461 cell ligands and present an opportunity for identifying unknown ligands from orphan TCRs
462 known to respond to a proteome of interest (Birnbaum et al., 2014; Gee et al., 2018). Indeed, as
463 DNA synthesis and sequencing continue to advance, defined peptide libraries expanding
464 beyond viral proteomes to covering whole bacterial or human proteomes will be possible, and
465 could present opportunities for investigating autoimmune diseases, which frequently have
466 strong MHC-II associations (Karnes et al., 2017). Such tools would be rich resources for
467 identifying both peptide-MHC binders and TCR ligands.

468 **Methods**

469 *Library design and creation*

470 Yeast display libraries were designed to cover all 15mer sequences within a given
471 proteome, with step size one. Reference proteomes used in creating defined libraries were
472 accessed from Uniprot, with the following Proteome IDs. SARS-CoV-2: UP000464024, SARS-CoV:
473 UP000000354, dengue serotype 1: UP000002500, dengue serotype 2: UP000180751, dengue
474 serotype 3: UP000007200, dengue serotype 4: UP000000275. The dengue proteome is
475 expressed as a single polypeptide, and peptides were generated from that contiguous stretch.

476 Each library peptide is encoded in DNA space, with specific codons selected randomly
477 from possible codons, with probabilities matching yeast codon usage (GenScript Codon Usage
478 Frequency Table). The DNA-encoded peptide sequences were flanked by invariant sequences
479 from the yeast construct for handles in amplification and cloning, and the DNA oligonucleotide
480 sequences were ordered from Twist Bioscience (South San Francisco, CA), with maximum
481 length of 120 nucleotides. The DNA oligo pool was amplified in low cycle PCR, followed by
482 amplification with construct DNA using overlap extension PCR. This extended product was
483 assembled in yeast with linearized pYal vector at a 5:1 insert:vector via electroporation with
484 electrocompetent RJY100 yeast.

485 HLA-DR401 and HLA-DR402 libraries were generated using previously described vectors
486 (Rappazzo et al., 2020) which contain mutations from wild type Metα36Leu, Valα132Met,
487 Hisβ33Asn, and Aspβ43Glu to enable proper folding without disrupting TCR or peptide contact
488 residues (Birnbaum et al., 2017). HLA-DR404 was generated using the same stabilizing
489 mutations. As previously described (Rappazzo et al., 2020), the peptide C-terminus is connected
490 to the MHC construct via a Gly-Ser linker (**Figure 1a**), and the N-terminus of the peptide
491 includes an extra alanine to ensure consistent cleavage between the construct and its signal
492 peptide.

493 The previously described null library (Dai et al., 2021) was generated with a peptide
494 encoded as "NNNTAANNNNNNNNNTAGNNNNNNNNNNNNTGANNNNNN", where "N" indicates
495 any nucleotide and encodes ten random amino acids and three stop codons. This library was
496 similarly generated in yeast using electrocompetent RJY100 yeast.

497

498 *Peptide visualizations and predictions*

499 Data visualizations of viral conservation and enrichment were generated using custom
500 scripts. For each 9mer stretch in a protein of interest, there are seven 15mer sequences that
501 overlap and contain that 9mer. We calculate how many of these seven 15mers enriched in both
502 the doped and undoped libraries. If five or more of the seven 15mers enriched, that stretch is
503 marked as a 'hit'. To examine conservation between viruses, viral proteins are aligned using
504 ClustalOmega (Madeira et al., 2019). Aligned 9mer stretches are compared between viruses
505 and identical stretches are considered conserved. Hits are determined individually for each
506 virus before merging, such that gaps in sequence alignments do not affect calculations of
507 enrichment for a given virus.

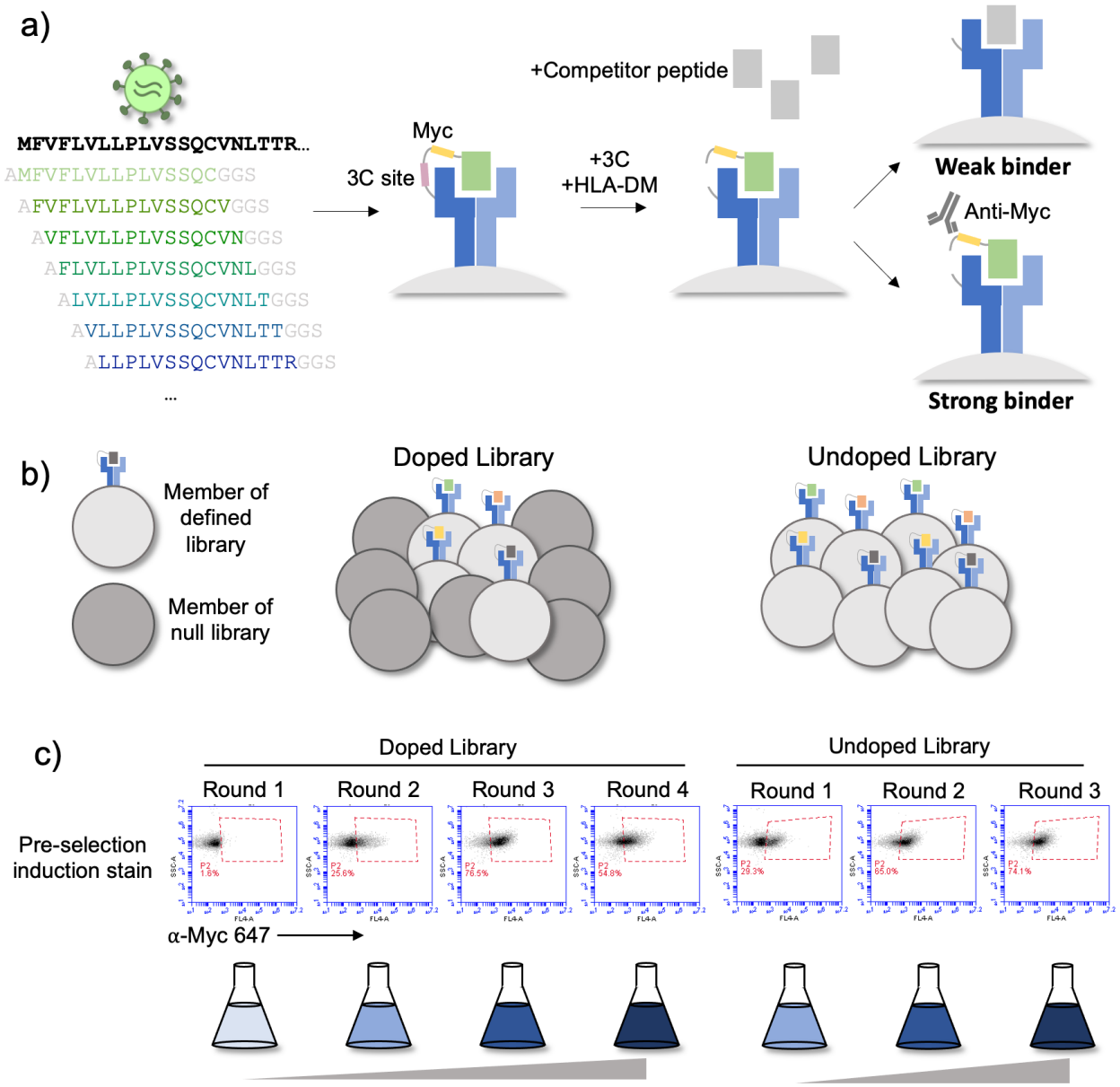
508 Representations of 15mer hits (as in **Figure 3**, **Figure 5** and **Supplemental Figure 5**) were
509 generated using in-house scripts, such that a 15mer that enriched in both the doped and
510 undoped library was marked as a horizontal line above the relevant 15mer sequence. Only
511 15mers containing the bolded 9mer in **Figure 3** and **Figure 5** were included.

512 NetMHCIIpan4.0 webserver was used for computational predictions (Reynisson et al.,
513 2020), where a binder is defined as having a predicted percent rank ≤ 10%, as defined in the
514 webserver instructions.
515
516 *Yeast library selections*
517 Library selections were consistent with previous peptide-MHC-II yeast display
518 dissociation studies (Dai et al., 2021; Rappazzo et al., 2020). Yeast were washed into pH 7.2 PBS
519 with 1 µM 3C protease and incubated at room temperature for 45 minutes. Yeast were then
520 washed into 4 °C acid saline (150mM NaCl, 20mM citric acid, pH5) with 1 µM HLA-DM and
521 incubated at 4 °C overnight. Each step takes place in the presence of competitor peptide (HLA-
522 DR401: $HA_{306-318}$ PKYVKQNTLKLAT, 1 µM; HLA-DR402: $CD48_{36-51}$ FDQKIVEWDSRKSKYF, 5 µM;
523 HLA-DR404: NKVKSLRILNTRRKL, 5 µM (Vita et al., 2019)). Non-specific binders are removed by
524 incubating yeast with anti-AlexaFluor647 magnetic beads and flowed over a magnetic Milltenyi
525 column at 4 °C. A positive selection follows, comprised of incubation with anti-Myc-
526 AlexaFluor647 antibody (1:100 volume:volume) and anti-AlexaFluor647 magnetic beads (1:10
527 volume:volume) and flowed over a Milltenyi column on a magnet at 4 °C, such that yeast with
528 bound peptide are retained on the column. These yeast are eluted, grown to confluence in at
529 30 °C in SDCAA media (pH 5), and sub-cultured in at 20 °C SGCAA media (pH 5) at OD600=1 for
530 two days. The first round of selections of doped libraries were conducted on 180 million yeast
531 (SARS-CoV-2 library) or 400 million yeast (dengue library) to ensure at least 20-fold coverage or
532 peptides. Subsequent rounds of doped library selection, and all rounds of undoped library
533 selections, were performed on 20-25 million yeast.
534
535 *Library sequencing and analysis*
536 Libraries were deep sequenced to determine their composition after each round of
537 selection. Plasmid DNA was extracted from ten million yeast from each round of selection using
538 the Zymoprep Yeast Miniprep Kit (Zymo Research), following manufacturer instructions.
539 Amplicons were generated through PCR, covering the peptide sequence through the 3C cut site.
540 A second PCR round was performed to add i5 and i7 sequencing handles and in-line index
541 barcodes unique to each round of selection. Amplicons were sequenced on an Illumina MiSeq
542 using paired-end MiSeq v2 300bp kits at the MIT BioMicroCenter.
543 Paired-end reads were assembled using PandaSeq (Masella et al., 2012). Peptide
544 sequences were extracted by identifying correctly encoded flanking regions, and were filtered
545 to ensure they matched designed members of the library or the randomized null construct
546 encoding, providing a stringent threshold for contamination and PCR and read errors.
547 The resulting data are analyzed for convergence, as described in the main text. Once a
548 library has converged, it is likely that changes in subsequent rounds of selection are due to
549 stochastic variation rather than improved binding.
550
551 *Register inference and sequence logos*
552 The 9mer core of enriched sequences was inferred using an in-house alignment
553 algorithm. In this approach, we utilize a 9mer position weight matrix (PWM), which we assess
554 at different offsets along the peptide. We one-hot encode sequences and pad with zeros on the
555 C-terminus of the peptide; to assess seven native registers and four non-native registers, we

556 pad the peptides with four zeros. Three of the non-native registers utilize the linker at the P9
557 anchor but not the P6 anchor, and the addition of a fourth register captures a minority set of
558 peptides which utilize Gly-Gly-Ser-Gly of the linker at P6 through P9 in the groove. Register-
559 setting is performed with zero-padded 15mers, rather than 15mers flanked by invariant
560 flanking residues, because the PWM would otherwise align all sequences to the invariant
561 region.
562 At the start, we randomly assign peptides to registers and generate a 9mer PWM. Over
563 subsequent iterations, peptides are assigned to new registers and the PWM was updated.
564 Assignments are random but biased, such that clusters corresponding to registers that match
565 the PWM are favored. Specifically, at each assignment we first take out the sequence under
566 consideration from the PWM. The PWM then defines an energy value for each register shift of a
567 given peptide, which is then used to generate a Boltzmann distribution from which we sample
568 the updated register shift. The stochasticity is decreased over time by raising the inverse
569 temperature linearly from 0.05 to 1 over 60 iterations, simulating 'cooling' (Andreatta et al.,
570 2017). A final deterministic iteration was carried out, where the distribution concentrates
571 entirely on the optimal register shift.
572 After register inference, sequence logo visualizations of the 9mer cores were generated
573 using Seq2Logo-2.0 with default settings, except using background frequencies from the SARS-
574 CoV-2 proteome and SARS-CoV spike and nucleocapsid proteins (Thomsen and Nielsen, 2012).
575 For registers with the C-terminus utilizing the C-terminal linker, the relevant linker sequence
576 was added to achieve a full 9mer sequence for visualizing the full 9mer core. For HLA-DR401,
577 distribution among registers, starting from N-terminally to C-terminally aligned in the peptide,
578 is: 161, 237, 227, 238, 231, 279, 237, 266, 271, 202, 118.
579
580 *Recombinant protein expression*
581 HLA-DM and HLA-DR401 were expressed recombinantly in High Five insect cells (Thermo
582 Fisher) using a baculovirus expression system, as previously described (Birnbaum et al., 2014;
583 Rappazzo et al., 2020). Ectodomain sequences of each chain were formatted with a C-terminal
584 poly-histidine purification tag and cloned into pAcGP67a vectors. Each vector was individually
585 transfected into SF9 insect cells (Thermo Fisher) with BestBac 2.0 linearized baculovirus DNA
586 (Expression Systems; Davis, CA) and Cellfectin II Reagent (Thermo Fisher), and propagated to
587 high titer. Viruses were co-titrated for optimal expression to maximize balanced MHC
588 heterodimer formation, co-transduced into Hi5 cells, and grown for 48-72 hours at 27 °C. The
589 secreted protein was purified from pre-conditioned media supernatant with Ni-NTA resin and
590 purified via size exclusion chromatography with a S200 increase column on an AKTA PURE FPLC
591 (GE Healthcare). To improve protein yields, the HLA-DRB1*04:01 chain was expressed with a
592 $CLIP_{87-101}$ peptide (PVSKMRMATPLLMQA) connected to the N-terminus of the MHC chain via a
593 flexible, 3C protease-cleavable linker.
594
595 *Fluorescence polarization experiments for peptide $IC_{50}$ determination*
596 Peptide $IC_{50}$ values were determined following a protocol modified from Yin & Stern (Yin
597 and Stern, 2014), as in Rappazzo et al (Rappazzo et al., 2020). In the assay, recombinantly
598 expressed HLA-DR401 is incubated with fluorescently labelled modified $HA_{306-318}$ (APRFV{Lys(5,6
599 FAM)}QNTLRLATG) peptide and a titration series for each unlabeled competitor peptide is

600 added (1.28 nM – 20 uM). A change in polarization value resulting from displacement of
601 fluorescent peptide from the binding groove is used to determine $IC_{50}$ values.
602 Relative binding at each concentration is calculated as $(FP_{sample} – FP_{free})/(FP_{no_comp} –$
603 $FP_{free})$. Here, $FP_{free}$ is the polarization value for the fluorescent peptide alone with no added
604 MHC, $FP_{no_comp}$ is polarization value for MHC with no competitor peptide added, and $FP_{sample}$ is
605 the polarization value with both MHC and competitor peptide added. Relative binding curves
606 were then generated and fit in Prism 9.3 to the equation $y = 1/(1+[pep]/IC_{50})$, where [pep] is the
607 concentration of un-labelled competitor peptide, in order to determine the concentration of
608 half-maximal inhibition, the $IC_{50}$ value.
609 Each assay was performed at 200 uL, with 100 nM recombinant MHC, 25 nM fluorescent
610 peptide, and competitor peptide (GenScript). This mixture co-incubates in pH 5 binding buffer
611 at 37 °C for 72 hours in black flat bottom 96-well plates. Competitor peptide concentrations
612 ranged from 1.28 nM to 20 µM, as a five-fold dilution series. Three replicates are performed for
613 each peptide concentration. Fluorescent peptide-only, no competitor peptide, and binding
614 buffer controls were also included. Our MHC was expressed with a linked CLIP peptide, so prior
615 to co-incubation, the peptide linker is cleaved by addition of 3C protease at 1:10 molar ratio at
616 room temperature for one hour; the residual cleaved 100 nM CLIP peptide is not expected to
617 alter peptide binding measurements.
618 Measurements were taken on a Molecular Devices SpectraMax M5 instrument. G-value
619 was 1.1 for each plate, as calculated per manufacturer instructions for each plate based on
620 fluorescent peptide-only wells minus buffer blank wells, with 35 mP reference for 5,6FAM
621 (Fluorescein setting). Measurements were made with 470 nm excitation and 520 nm emission,
622 10 flashes per read, and default PMT gain high.
623
624 **Data Availability**
625 All deep sequencing data are deposited on the Sequence Read Archive (SRA), with accession
626 codes PRJNA806475 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA806475] and
627 PRJNA708266 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA708266]
628
629 **Code Availability**
630 Scripts used for data processing and visualization are publicly available at
631 https://github.com/birnbaumlab/Huisman-et-al-2022.
632
633 **Acknowledgements**
634 We would like to thank Patrick Holec for feedback on this manuscript, and the MIT BioMicro
635 Center for library sequencing. This work was supported in part by the Koch Institute Support
636 (core) Grant P30-CA14051 from the National Cancer Institute. This work was supported by
637 National Institute of Health (U19-AI110495), the Packard Foundation to M.E.B., a Schmidt
638 Futures grant to D.K.G. and M.E.B., and a National Science Foundation Graduate Research
639 Fellowship to B.D.H.

640 **Figures**




641

642 **Figure 1. Overview of library and selections. a)** The defined library contains pathogen
643 proteome peptides (length 15, sliding window 1). Poor binding peptides are displaced with
644 addition of protease, competitor peptide, and HLA-DM. **b)** Schematic of doped and undoped
645 libraries: in the doped selection strategy, the library is added to a library of null, non-expressing
646 constructs. **c)** Representative flow plots showing enrichment of MHC-expressing yeast over
647 rounds of selection for the library containing SARS-CoV-2 and SARS-CoV peptides on HLA-
648 DR401.

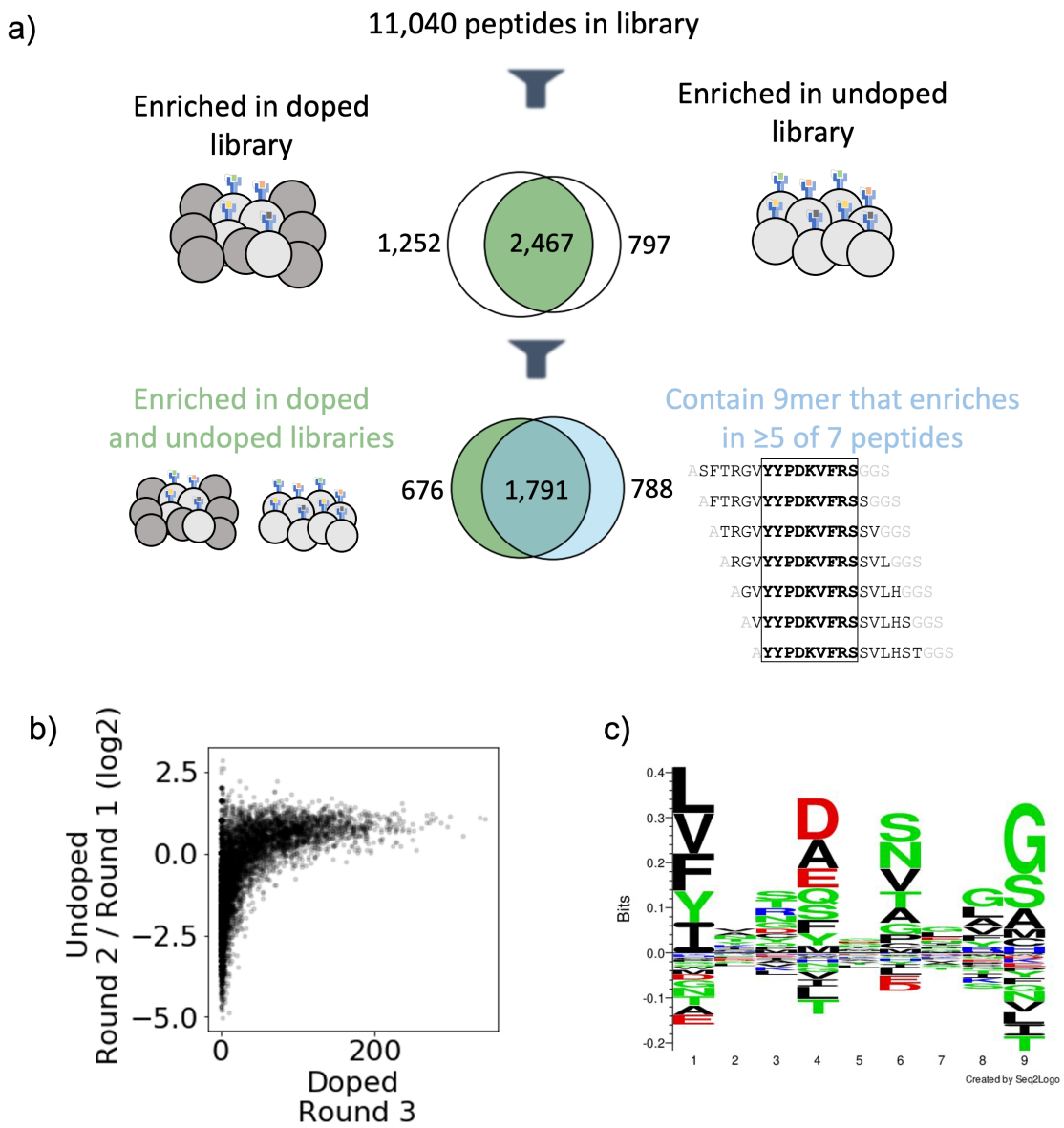



**Figure 2. Output of selections and analysis of selection data. a)** Overview of filtering peptides and correspondence between selection strategies for SARS-CoV and SARS-CoV-2 library on HLA-DR401. Peptides are filtered for enrichment in both doped and undoped libraries. Further, the relationship between these peptides and peptides which contain a 9mer that is enriched in five or more of the seven peptides containing it is shown. **b)** Relationships between enrichment in doped and undoped libraries. Absolute counts following Round 3 of selection of the doped library are plotted against the log2 fold change between read fraction for peptides in Round 2 and Round 1. Data are shown for the library on HLA-DR401. **c)** Sequence logo of 2,467 peptides that enriched in both doped and undoped selected libraries for HLA-DR401. Registers are inferred with a position weight matrix-based alignment method. Logos were generated with Seq2Logo-2.0.

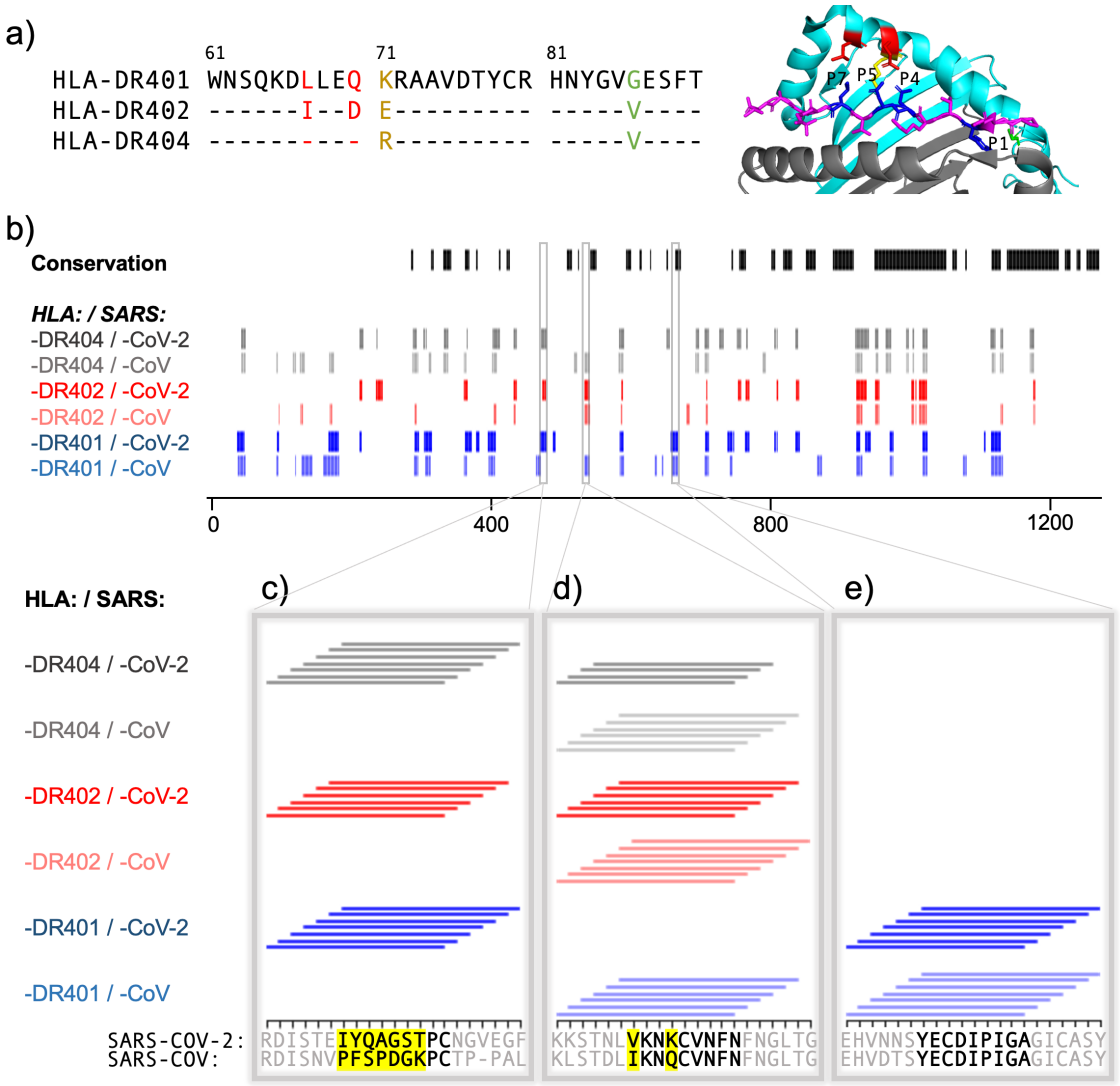
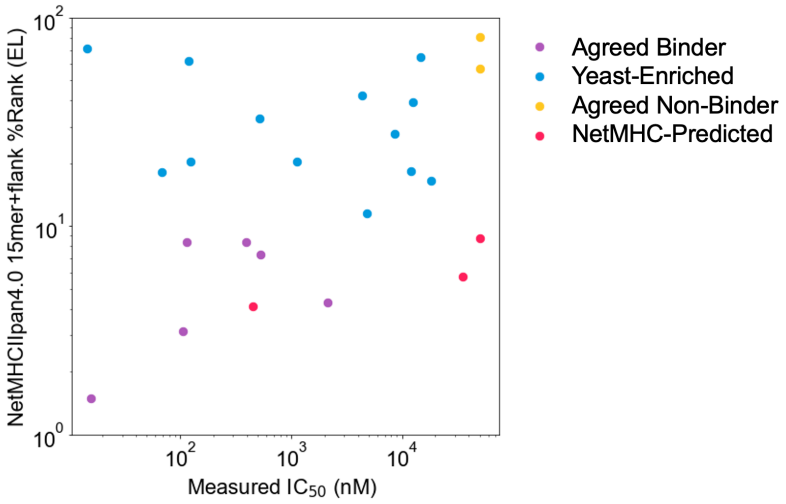



**Figure 3. Comparing HLA-DR401, HLA-DR402, and HLA-DR404 for binding to related Spike proteins from SARS-CoV-2 and SARS-CoV. a)** Sequence alignment showing sequence differences in HLA-DR402 and HLA-DR404 compared to HLA-DR401 and highlighted on HLA-DR401 structure (PDB 1J8H). Colors are: red for amino acids shared between HLA-DR401 and HLA-DR404, green for amino acids shared between HLA-DR402 and HLA-DR404, and yellow for amino acids different in all 3 alleles. Affected peptide positions (P1, P4, P5, P7) are colored in blue and labeled on the structure. **b)** Conservation and enrichment of 9mer peptides from SARS-CoV-2 and SARS-CoV Spike proteins. Conserved 9mers are indicated in black. If a 9mer along the proteome enriched in 5 or more of the adjacent peptides containing it, its enrichment is indicated with a vertical line with color for allele (HLA-DR401: blue; HLA-DR402: red; HLA-DR404: grey) and opacity for virus (SARS-CoV-2: dark; SARS-CoV: light). **b-e)** Zoomed regions show enrichment of individual 15mer peptides. Only peptides containing the bolded 9mer sequence are shown. Amino acids in the bolded 9mer that are not conserved between SARS-CoV-2 and SARS-CoV are highlighted in yellow.




676
677 **Figure 4. Comparing measured $IC_{50}$ values and computational prediction.** Relationship
678 between measured $IC_{50}$ values and NetMHCIIpan4.0 predicted ranks in Eluted Ligand mode (EL)
679 on invariant-flanked sequences. Data points are colored by label, and $IC_{50}$ values ≥50 µM are set
680 to 50 µM.

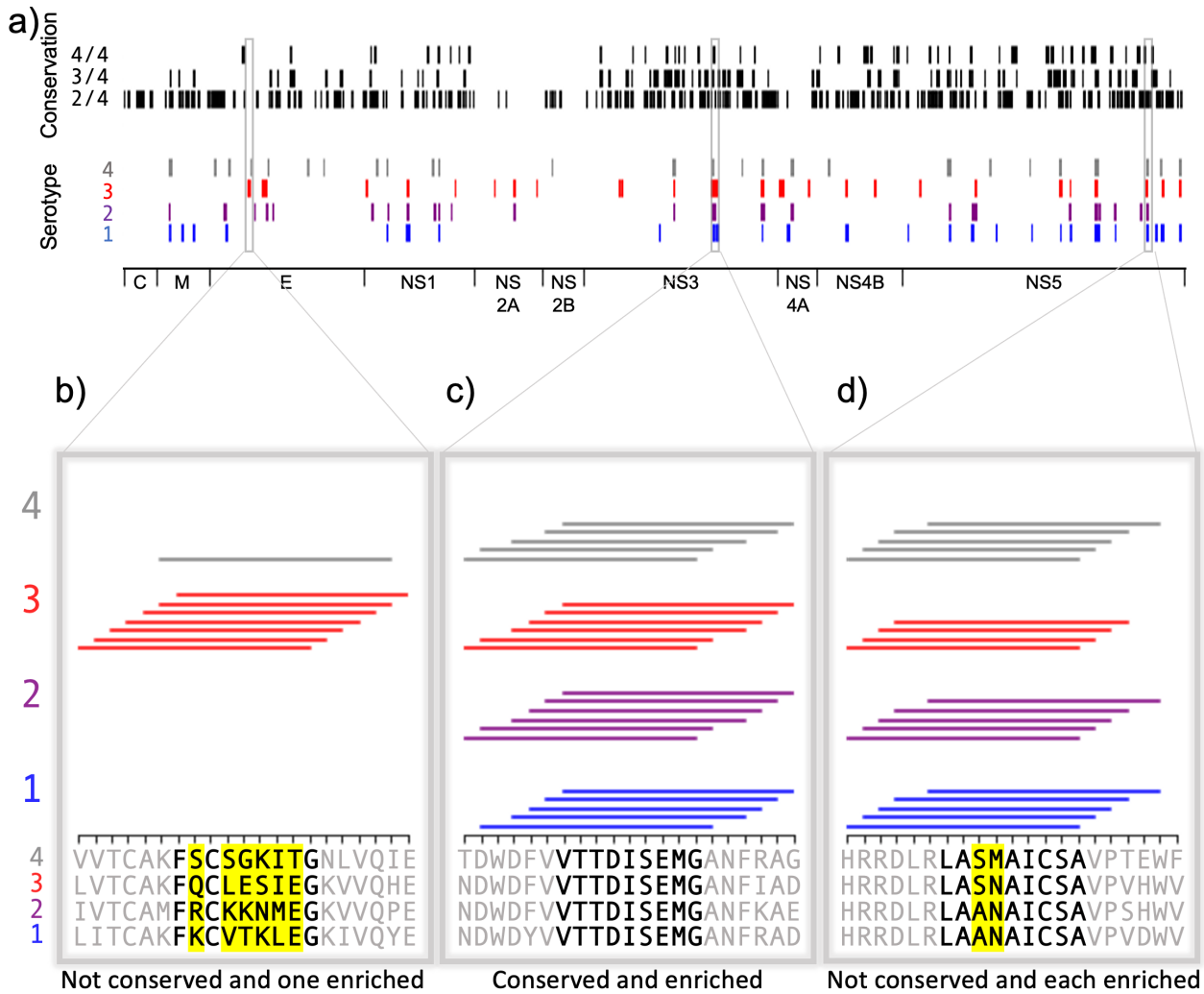



**Figure 5. Conservation and enrichment of dengue virus serotypes 1-4. a)** Conservation and enrichment of 9mer peptides along four aligned dengue serotypes. All stretches of 9 amino acids are compared across the four serotypes and conservation is indicated with a black vertical line (i.e. 2, 3, or 4 of 4 serotypes conserved). 9mers which enriched on HLA-DR401 are also indicated, colored by virus serotype. **b-d)** Zoomed regions, showing enrichment for individual 15mer peptides to HLA-DR401. Only peptides which contain the bolded 9mer sequence are shown. Amino acids in the bolded 9mer that are not conserved between serotypes are highlighted in yellow. Insets show regions which are differently conserved and enriched: **b)** non-conserved sequences with peptides from one serotype enriched; **c)** conserved sequences enriched across all serotypes; **d)** non-conserved sequences which are enriched.

| Virus | Protein | Protein length (# of amino acids) | MHC Allele | # of 15mers | # of 9mer cores | # of smoothed 15mers |
|---|---|---|---|---|---|---|
| SARS-CoV | Spike | 1255 | HLA-DR401 | 324 | 74 | 221 |
| | | | HLA-DR402 | 217 | 65 | 110 |
| | | | HLA-DR404 | 289 | 61 | 193 |
| SARS-CoV | Nucleocapsid | 422 | HLA-DR401 | 40 | 8 | 34 |
| | | | HLA-DR402 | 34 | 13 | 12 |
| | | | HLA-DR404 | 31 | 6 | 20 |
| SARS-CoV-2 | Spike | 1273 | HLA-DR401 | 305 | 67 | 221 |
| | | | HLA-DR402 | 230 | 62 | 130 |
| | | | HLA-DR404 | 290 | 64 | 217 |
| SARS-CoV-2 | Nucleocapsid | 419 | HLA-DR401 | 34 | 8 | 24 |
| | | | HLA-DR402 | 33 | 10 | 15 |
| | | | HLA-DR404 | 30 | 8 | 18 |
| SARS-CoV-2 | Replicase polyprotein 1ab | 7096 | HLA-DR401 | 1652 | 388 | 1204 |
| | | | HLA-DR402 | 1104 | 325 | 678 |
| | | | HLA-DR404 | 1368 | 350 | 890 |
| SARS-CoV-2 | Non-structural protein 8 | 121 | HLA-DR401 | 41 | 10 | 32 |
| | | | HLA-DR402 | 21 | 7 | 17 |
| | | | HLA-DR404 | 32 | 8 | 19 |
| SARS-CoV-2 | Protein 7a | 121 | HLA-DR401 | 27 | 8 | 18 |
| | | | HLA-DR402 | 7 | 3 | 0 |
| | | | HLA-DR404 | 13 | 2 | 6 |
| SARS-CoV-2 | Non-structural protein 6 | 61 | HLA-DR401 | 0 | 0 | 0 |
| | | | HLA-DR402 | 1 | 1 | 0 |
| | | | HLA-DR404 | 0 | 0 | 0 |
| SARS-CoV-2 | Membrane protein | 222 | HLA-DR401 | 40 | 7 | 29 |
| | | | HLA-DR402 | 26 | 6 | 19 |
| | | | HLA-DR404 | 23 | 7 | 21 |
| SARS-CoV-2 | Envelope small membrane protein | 75 | HLA-DR401 | 6 | 1 | 0 |
| | | | HLA-DR402 | 7 | 3 | 0 |
| | | | HLA-DR404 | 6 | 1 | 0 |
| SARS-CoV-2 | Protein 3a | 275 | HLA-DR401 | 22 | 4 | 11 |
| | | | HLA-DR402 | 13 | 4 | 10 |
| | | | HLA-DR404 | 10 | 2 | 0 |
| SARS-CoV-2 | Replicase polyprotein 1a | 4405 | HLA-DR401 | 948 | 228 | 658 |
| | | | HLA-DR402 | 657 | 196 | 409 |
| | | | HLA-DR404 | 865 | 222 | 582 |
| SARS-CoV-2 | ORF10 protein | 38 | HLA-DR401 | 6 | 1 | 6 |
| | | | HLA-DR402 | 2 | 0 | 0 |
| | | | HLA-DR404 | 5 | 1 | 5 |
| SARS-CoV-2 | Protein non-structural 7b | 43 | HLA-DR401 | 0 | 0 | 0 |
| | | | HLA-DR402 | 0 | 0 | 0 |
| | | | HLA-DR404 | 0 | 0 | 0 |
| SARS-CoV-2 | Uncharacterized protein 14 | 73 | HLA-DR401 | 8 | 4 | 6 |
| | | | HLA-DR402 | 20 | 5 | 16 |
| | | | HLA-DR404 | 22 | 4 | 21 |
| SARS-CoV-2 | Protein 9b | 97 | HLA-DR401 | 29 | 7 | 27 |
| | | | HLA-DR402 | 35 | 6 | 31 |
| | | | HLA-DR404 | 37 | 9 | 34 |

692
693 **Table 1.** Summary of enriched peptides for each source protein, including: the number of
694 unique 15mers which each enriched in both of the doped and undoped libraries; the number of
695 unique 9mer cores identified by register-inference in these enriched 15mers (native cores only,
696 so linker-containing inferred cores excluded); and the number of unique enriched 15mers that
697 contain 9mer sequences enriched in five or more of overlapping neighbors.

| | Spike Position | Peptide+flank (A+15mer+GGS) | NetMHCIIpan4.0 Predicted Core (A+15mer+GGS) | NetMHCIIpan4.0 %Rank (A+15mer+GGS) | 15mer Affinity from FP ($IC_{50}$, nM) |
|---|---|---|---|---|---|
| Agreed Binders | 34-48 | ARGVYYPDKVFRSSVLGGS | YYPDKVFRS | 1.49 | 15.8 |
| | 87-101 | ANDGVYFASTEKSNIIGGS | VYFASTEKS | 4.28 | 2117 |
| | 303-317 | ALKSFTVEKGIYQTSNGGS | FTVEKGIYQ | 8.41 | 396.9 |
| | 362-376 | AVADYSVLYNSASFSTGGS | YSVLYNSAS | 8.36 | 113.7 |
| | 1015-1029 | AAAEIRASANLAATKMGGS | IRASANLAA | 3.13 | 105.4 |
| | 1112-1126 | APQIITTDNTFVSGNCGGS | ITTDNTFVS | 7.32 | 527.0 |
| Yeast-Enriched Binders | 165-179 | ANCTFEYVSQPFLMDLGGS | YVSQPFLMD | 64.83 | 14,652 |
| | 172-186 | ASQPFLMDLEGKQGNFGGS | FLMDLEGKQ | 20.34 | 123.2 |
| | 286-300 | ATDAVDCALDPLSETKGGS | VDCALDPLS | 32.68 | 521.6 |
| | 373-387 | ASFSTFKCYGVSPTKLGGS | YGVSPTKLG | 16.59 | 18,452 |
| | 469-483 | ASTEIYQAGSTPCNGVGGS | IYQAGSTPC | 18.22 | 67.7 |
| | 580-594 | AQTLEILDITPCSFGGGGS | LEILDITPC | 62 | 119.9 |
| | 739-753 | ATMYICGDSTECSNLLGGS | YICGDSTEC | 70.91 | 14.4 |
| | 920-934 | AQKLIANQFNSAIGKIGGS | FNSAIGKIG | 20.47 | 1121 |
| NetMHC-Predicted Binders | 113-127 | AKTQSLLIVNNATNVVGGS | IVNNATNVV | 8.74 | >50,000 |
| | 492-506 | ALQSYGFQPTNGVGYQGGS | YGFQPTNGV | 4.11 | 454.7 |
| | 1151-1165 | AELDKYFKNHTSPDVDGGS | YFKNHTSPD | 5.74 | 35,510 |
| Agreed Non-Binders | 534-548 | AVKNKCVNFNFNGLTGGGS | FNFNGLTGG | 57.13 | >50,000 |
| | 1079-1093 | APAICHDGKAHFPREGGGS | ICHDGKAHF | 80.47 | >50,000 |

**Table 2.** Peptides selected for fluorescence polarization (FP) experiments for binding to HLA-DR401. NetMHCIIpan4.0 predictions for HLA-DR401 binding are performed on 15mers plus invariant flanking residues (N-terminal Ala, C-terminal Gly-Gly-Ser) and percent rank values generated using Eluted Ligand mode. Fluorescence polarization is performed on native 15mer peptides without invariant flanking residues.

704

| Spike Position | Sequence | NetMHCIIpan4.0 Predicted Core (A+15mer+GGS) | NetMHCIIpan4.0 %Rank (A+15mer+GGS) | NetMHCIIpan4.0 Predicted Core (15mer) | NetMHCIIpan4.0 %Rank (15mer) | 15mer Affinity from FP (IC50, nM) |
|---|---|---|---|---|---|---|
| 172-186 | **SQPFLMDLEGKQGNF** | FLMDLEGKQ | 20.34 | FLMDLEGKQ | 4.1 | 123.2 |
| 173-187 | **QPFLMDLEGKQGNFK** | FLMDLEGKQ | 27.73 | FLMDLEGKQ | 12.21 | 8613 |
| 286-300 | **TDAVDCALDPLSETK** | VDCALDPLS | 32.68 | VDCALDPLS | 9.8 | 1154 |
| 287-301 | **DAVDCALDPLSETKC** | VDCALDPLS | 42.42 | VDCALDPLS | 22.57 | 4393 |
| 469-483 | **STEIYQAGSTPCNGV** | IYQAGSTPC | 18.22 | IYQAGSTPC | 5.41 | 67.7 |
| 467-481 | **DISTEIYQAGSTPCN** | IYQAGSTPC | 11.47 | IYQAGSTPC | 12.61 | 4875 |
| 471-485 | **EIYQAGSTPCNGVEG** | YQAGSTPCN | 39.17 | YQAGSTPCN | 21.81 | 12519 |
| 920-934 | **QKLIANQFNSAIGKI** | FNSAIGKIG | 20.47 | IANQFNSAI | 7.89 | 1495 |
| 921-935 | **KLIANQFNSAIGKIQ** | FNSAIGKIQ | 18.3 | IANQFNSAI | 19.79 | 11937 |

705

706 **Table 3.** Effects of peptide flanking sequences on NetMHCIIpan4.0 predictions for HLA-DR401
707 binding and measured fluorescence polarization (FP) values for overlapping peptides. Yeast
708 display-enriched peptides that are predicted to bind by NetMHCIIpan4.0 when without flanking
709 residues, plus offset variants of these peptides, which are not predicted to bind, with or
710 without flanking sequence. Yeast display register-inferred consensus cores are highlighted in
711 green. NetMHCIIpan4.0 percent rank values are generated using Eluted Ligand mode.

712 **References**


713 Abelin JG, Harjanto D, Malloy M, Suri P, Colson T, Goulding SP, Creech AL, Serrano LR, Nasir G,
714 Nasrullah Y, McGann CD, Velez D, Ting YS, Poran A, Rothenberg DA, Chhangawala S,
715 Rubinsteyn A, Hammerbacher J, Gaynor RB, Fritsch EF, Greshock J, Oslund RC,
716 Barthelme D, Addona TA, Arieta CM, Rooney MS. 2019. Defining HLA-II ligand processing
717 and binding rules with mass spectrometry enhances cancer Epitope prediction.
718 *Immunity* **51**:766-779.e17.
719 Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, Stevens J, Lane W, Zhang GL,
720 Eisenhaure TM, Clauser KR, Hacohen N, Rooney MS, Carr SA, Wu CJ. 2017. Mass
721 spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more
722 accurate Epitope prediction. *Immunity* **46**:315–326.
723 Altmann DM, Boyton RJ. 2020. SARS-CoV-2 T cell immunity: Specificity, function, durability, and
724 role in protection. *Sci Immunol* **5**:eabd6160.
725 Andreatta M, Alvarez B, Nielsen M. 2017. GibbsCluster: unsupervised clustering and alignment
726 of peptide sequences. *Nucleic Acids Res* **45**:W458–W463.
727 Barra C, Alvarez B, Paul S, Sette A, Peters B, Andreatta M, Buus S, Nielsen M. 2018. Footprints of
728 antigen processing boost MHC class II natural ligand predictions. *Genome Med* **10**:84.
729 Birnbaum ME, Mendoza J, Bethune M, Baltimore D, Garcia KC. 2017. Ligand discovery for t cell
730 receptors. US20170192011A1.
731 Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, Ozkan E, Davis MM,
732 Wucherpfennig KW, Garcia KC. 2014. Deconstructing the peptide-MHC specificity of T
733 cell recognition. *Cell* **157**:1073–1087.
734 Chaplin DD. 2010. Overview of the immune response. *J Allergy Clin Immunol* **125**:S3-23.
735 Dai Z, Huisman BD, Zeng H, Carter B, Jain S, Birnbaum ME, Gifford DK. 2021. Machine learning
736 optimization of peptides for presentation by class II MHCs. *Bioinformatics*.
737 doi:10.1093/bioinformatics/btab131
738 Gambino F Jr, Tai W, Voronin D, Zhang Y, Zhang X, Shi J, Wang X, Wang N, Du L, Qiao L. 2021. A
739 vaccine inducing solely cytotoxic T lymphocytes fully prevents Zika virus infection and
740 fetal damage. *Cell Rep* **35**:109107.
741 Gee MH, Han A, Lofgren SM, Beausang JF, Mendoza JL, Birnbaum ME, Bethune MT, Fischer S,
742 Yang X, Gomez-Eerland R, Bingham DB, Sibener LV, Fernandes RA, Velasco A, Baltimore
743 D, Schumacher TN, Khatri P, Quake SR, Davis MM, Garcia KC. 2018. Antigen
744 identification for orphan T cell receptors expressed on tumor-infiltrating lymphocytes.
745 *Cell* **172**:549-563.e16.
746 Guzman MG, Gubler DJ, Izquierdo A, Martinez E, Halstead SB. 2016. Dengue infection. *Nat Rev*
747 *Dis Primers* **2**:16055.
748 Hennecke J, Wiley DC. 2002. Structure of a complex of the human alpha/beta T cell receptor
749 (TCR) HA1.7, influenza hemagglutinin peptide, and major histocompatibility complex
750 class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-
751 restriction and alloreactivity. *J Exp Med* **195**:571–581.
752 Hennecke J, Wiley DC. 2001. T cell receptor-MHC interactions up close. *Cell* **104**:1–4.
753 Jiang W, Boder ET. 2010. High-throughput engineering and analysis of peptide binding to class II
754 MHC. *Proc Natl Acad Sci U S A* **107**:13258–13263.

755 Jones EY, Fugger L, Strominger JL, Siebold C. 2006. MHC class II proteins and disease: a
756 structural perspective. *Nat Rev Immunol* **6**:271–282.
757 Justesen S, Harndahl M, Lamberth K, Nielsen L-LB, Buus S. 2009. Functional recombinant MHC
758 class II molecules and high-throughput peptide-binding assays. *Immunome Res* **5**:2.
759 Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, Glazer AM, Mosley JD, Zhao S,
760 Raychaudhuri S, Mallal S, Ye Z, Mayer JG, Brilliant MH, Hebbring SJ, Roden DM, Phillips
761 EJ, Denny JC. 2017. Phenome-wide scanning identifies multiple diseases and disease
762 severity phenotypes associated with HLA variants. *Sci Transl Med* **9**.
763 doi:10.1126/scitranslmed.aai8708
764 Keskin DB, Anandappa AJ, Sun J, Tirosh I, Mathewson ND, Li S, Oliveira G, Giobbie-Hurder A, Felt
765 K, Gjini E, Shukla SA, Hu Z, Li L, Le PM, Allesøe RL, Richman AR, Kowalczyk MS,
766 Abdelrahman S, Geduldig JE, Charbonneau S, Pelton K, Iorgulescu JB, Elagina L, Zhang W,
767 Olive O, McCluskey C, Olsen LR, Stevens J, Lane WJ, Salazar AM, Daley H, Wen PY,
768 Chiocca EA, Harden M, Lennon NJ, Gabriel S, Getz G, Lander ES, Regev A, Ritz J, Neuberg
769 D, Rodig SJ, Ligon KL, Suvà ML, Wucherpfennig KW, Hacohen N, Fritsch EF, Livak KJ, Ott
770 PA, Wu CJ, Reardon DA. 2019. Neoantigen vaccine generates intratumoral T cell
771 responses in phase Ib glioblastoma trial. *Nature* **565**:234–239.
772 Klinger M, Pepin F, Wilkins J, Asbury T, Wittkop T, Zheng J, Moorhead M, Faham M. 2015.
773 Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of
774 Immune Assays and Immune Receptor Sequencing. *PLoS One* **10**:e0141561.
775 Liu G, Carter B, Bricken T, Jain S, Viard M, Carrington M, Gifford DK. 2020. Computationally
776 optimized SARS-CoV-2 MHC class I and II vaccine formulations predicted to target
777 human haplotype distributions. *Cell Syst* **11**:131-144.e6.
778 Liu G, Carter B, Gifford DK. 2021. Predicted cellular immunity population coverage gaps for
779 SARS-CoV-2 subunit vaccines and their augmentation by compact peptide sets. *Cell Syst*
780 **12**:102-107.e4.
781 Liu R, Jiang W, Mellins ED. 2021. Yeast display of MHC-II enables rapid identification of peptide
782 ligands from protein antigens (RIPPA). *Cell Mol Immunol* **18**:1847–1860.
783 Lovitch SB, Pu Z, Unanue ER. 2006. Amino-terminal flanking residues determine the
784 conformation of a peptide-class II MHC complex. *J Immunol* **176**:2958–2968.
785 Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC,
786 Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019.
787 *Nucleic Acids Res* **47**:W636–W641.
788 Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. 2012. PANDAseq: paired-end
789 assembler for illumina sequences. *BMC Bioinformatics* **13**:31.
790 Mateus J, Grifoni A, Tarke A, Sidney J, Ramirez SI, Dan JM, Burger ZC, Rawlings SA, Smith DM,
791 Phillips E, Mallal S, Lammers M, Rubiro P, Quiambao L, Sutherland A, Yu ED, da Silva
792 Antunes R, Greenbaum J, Frazier A, Markmann AJ, Premkumar L, de Silva A, Peters B,
793 Crotty S, Sette A, Weiskopf D. 2020. Selective and cross-reactive SARS-CoV-2 T cell
794 epitopes in unexposed humans. *Science* **370**:89–94.
795 Moise L, Gutierrez A, Kibria F, Martin R, Tassone R, Liu R, Terry F, Martin B, De Groot AS. 2015.
796 iVAX: An integrated toolkit for the selection and optimization of antigens and the design
797 of epitope-driven vaccines. *Hum Vaccin Immunother* **11**:2312–2321.

798 Obermair FJ, Renoux F, Heer S, Lee C, Cereghetti N, Maestri G, Haldner Y, Wuigk R, Iosefson O,
799 Patel P, Triebel K, Kopf M, Swain J, Kisielow J. 2021. High resolution profiling of MHC-II
800 peptide presentation capacity, by Mammalian Epitope Display, reveals SARS-CoV-2
801 targets for CD4 T cells and mechanisms of immune-escape. *bioRxiv*.
802 doi:10.1101/2021.03.02.433522
803 O'Brien C, Flower DR, Feighery C. 2008. Peptide length significantly influences in vitro affinity
804 for MHC class II molecules. *Immunome Res* **4**:6.
805 O'Donnell TJ, Rubinsteyn A, Laserson U. 2020. MHCflurry 2.0: Improved pan-allele prediction of
806 MHC class I-presented peptides by incorporating antigen processing. *Cell Syst* **11**:42-
807 48.e7.
808 Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, Zhang W, Luoma A, Giobbie-Hurder A,
809 Peter L, Chen C, Olive O, Carter TA, Li S, Lieb DJ, Eisenhaure T, Gjini E, Stevens J, Lane
810 WJ, Javeri I, Nellaiappan K, Salazar AM, Daley H, Seaman M, Buchbinder EI, Yoon CH,
811 Harden M, Lennon N, Gabriel S, Rodig SJ, Barouch DH, Aster JC, Getz G, Wucherpfennig
812 K, Neuberg D, Ritz J, Lander ES, Fritsch EF, Hacohen N, Wu CJ. 2017. An immunogenic
813 personal neoantigen vaccine for patients with melanoma. *Nature* **547**:217–221.
814 Parker R, Partridge T, Wormald C, Kawahara R, Stalls V, Aggelakopoulou M, Parker J, Powell
815 Doherty R, Ariosa Morejon Y, Lee E, Saunders K, Haynes BF, Acharya P, Thaysen-
816 Andersen M, Borrow P, Ternette N. 2021. Mapping the SARS-CoV-2 spike glycoprotein-
817 derived peptidome presented by HLA class II on dendritic cells. *Cell Rep* **35**:109179.
818 Patronov A, Doytchinova I. 2013. T-cell epitope vaccine design by immunoinformatics. *Open*
819 *Biol* **3**:120139.
820 Purcell AW, Ramarathinam SH, Ternette N. 2019. Mass spectrometry-based identification of
821 MHC-bound peptides for immunopeptidomics. *Nat Protoc* **14**:1687–1707.
822 Racle J, Michaux J, Rockinger GA, Arnaud M, Bobisse S, Chong C, Guillaume P, Coukos G, Harari
823 A, Jandus C, Bassani-Sternberg M, Gfeller D. 2019. Robust prediction of HLA class II
824 epitopes by deep motif deconvolution of immunopeptidomes. *Nat Biotechnol* **37**:1283–
825 1286.
826 Rappazzo CG, Huisman BD, Birnbaum ME. 2020. Repertoire-scale determination of class II MHC
827 peptide binding via yeast display improves antigen prediction. *Nat Commun* **11**:4414.
828 Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. 2020. NetMHCpan-4.1 and NetMHCIIpan-
829 4.0: improved predictions of MHC antigen presentation by concurrent motif
830 deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res*
831 **48**:W449–W454.
832 Rosati E, Pogorelyy MV, Minervina AA, Scheffold A, Franke A, Bacher P, Thomas PG. 2021.
833 Characterization of SARS-CoV-2 public CD4+ αβ T cell clonotypes through reverse
834 epitope discovery. *bioRxivorg*. doi:10.1101/2021.11.19.469229
835 Sidney J, Steen A, Moore C, Ngo S, Chung J, Peters B, Sette A. 2010. Divergent motifs but
836 overlapping binding repertoires of six HLA-DQ molecules frequently expressed in the
837 worldwide human population. *J Immunol* **185**:4189–4198.
838 Snyder TM, Gittelman RM, Klinger M, May DH, Osborne EJ, Taniguchi R, Zahid HJ, Kaplan IM,
839 Dines JN, Noakes MT, Pandya R, Chen X, Elasady S, Svejnoha E, Ebert P, Pesesky MW, De
840 Almeida P, O'Donnell H, DeGottardi Q, Keitany G, Lu J, Vong A, Elyanow R, Fields P,
841 Greissl J, Baldo L, Semprini S, Cerchione C, Nicolini F, Mazza M, Delmonte OM, Dobbs K,

842 Laguna-Goya R, Carreño-Tarragona G, Barrio S, Imberti L, Sottini A, Quiros-Roldan E,
843 Rossi C, Biondi A, Bettini LR, D'Angio M, Bonfanti P, Tompkins MF, Alba C, Dalgard C,
844 Sambri V, Martinelli G, Goldman JD, Heath JR, Su HC, Notarangelo LD, Paz-Artal E,
845 Martinez-Lopez J, Carlson JM, Robins HS. 2020. Magnitude and dynamics of the T-cell
846 response to SARS-CoV-2 infection at both individual and population levels. *medRxiv*.
847 doi:10.1101/2020.07.31.20165647
848 Stern LJ. 1994. Crystal structure of the human class II MHC protein HLA- DR1 complexed with an
849 influenza virus peptide. *Nature* **368**:215–221.
850 Stopfer LE, Gajadhar AS, Patel B, Gallien S, Frederick DT, Boland GM, Sullivan RJ, White FM.
851 2021. Absolute quantification of tumor antigens using embedded MHC-I isotopologue
852 calibrants. *Proc Natl Acad Sci U S A* **118**:e2111173118.
853 Stopfer LE, Mesfin JM, Joughin BA, Lauffenburger DA, White FM. 2020. Multiplexed relative and
854 absolute quantitative immunopeptidomics reveals MHC I repertoire alterations induced
855 by CDK4/6 inhibition. *Nat Commun* **11**:2760.
856 Swain SL, McKinstry KK, Strutt TM. 2012. Expanding roles for $CD4^+$ T cells in immunity to viruses.
857 *Nat Rev Immunol* **12**:136–148.
858 Thomsen MCF, Nielsen M. 2012. Seq2Logo: a method for construction and visualization of
859 amino acid binding motifs and sequence profiles including sequence weighting, pseudo
860 counts and two-sided representation of amino acid enrichment and depletion. *Nucleic*
861 *Acids Res* **40**:W281-7.
862 Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B.
863 2019. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* **47**:D339–
864 D343.
865 Yin L, Stern LJ. 2014. Measurement of peptide binding to MHC class II molecules by
866 fluorescence polarization. *Curr Protoc Immunol* **106**:5.10.1-5.10.12.
867 Zavala-Ruiz Z, Strug I, Anderson MW, Gorski J, Stern LJ. 2004. A polymorphic pocket at the P10
868 position contributes to peptide binding specificity in class II MHC proteins. *Chem Biol*
869 **11**:1395–1402.
870 Zeng H, Gifford DK. 2019. Quantification of uncertainty in peptide-MHC binding prediction
871 improves high-affinity peptide selection for therapeutic design. *Cell Syst* **9**:159-166.e3.
872 Zhao W, Sher X. 2018. Systematically benchmarking peptide-MHC binding predictors: From
873 synthetic to naturally processed epitopes. *PLoS Comput Biol* **14**:e1006457.