**Cell Systems, Volume *14***

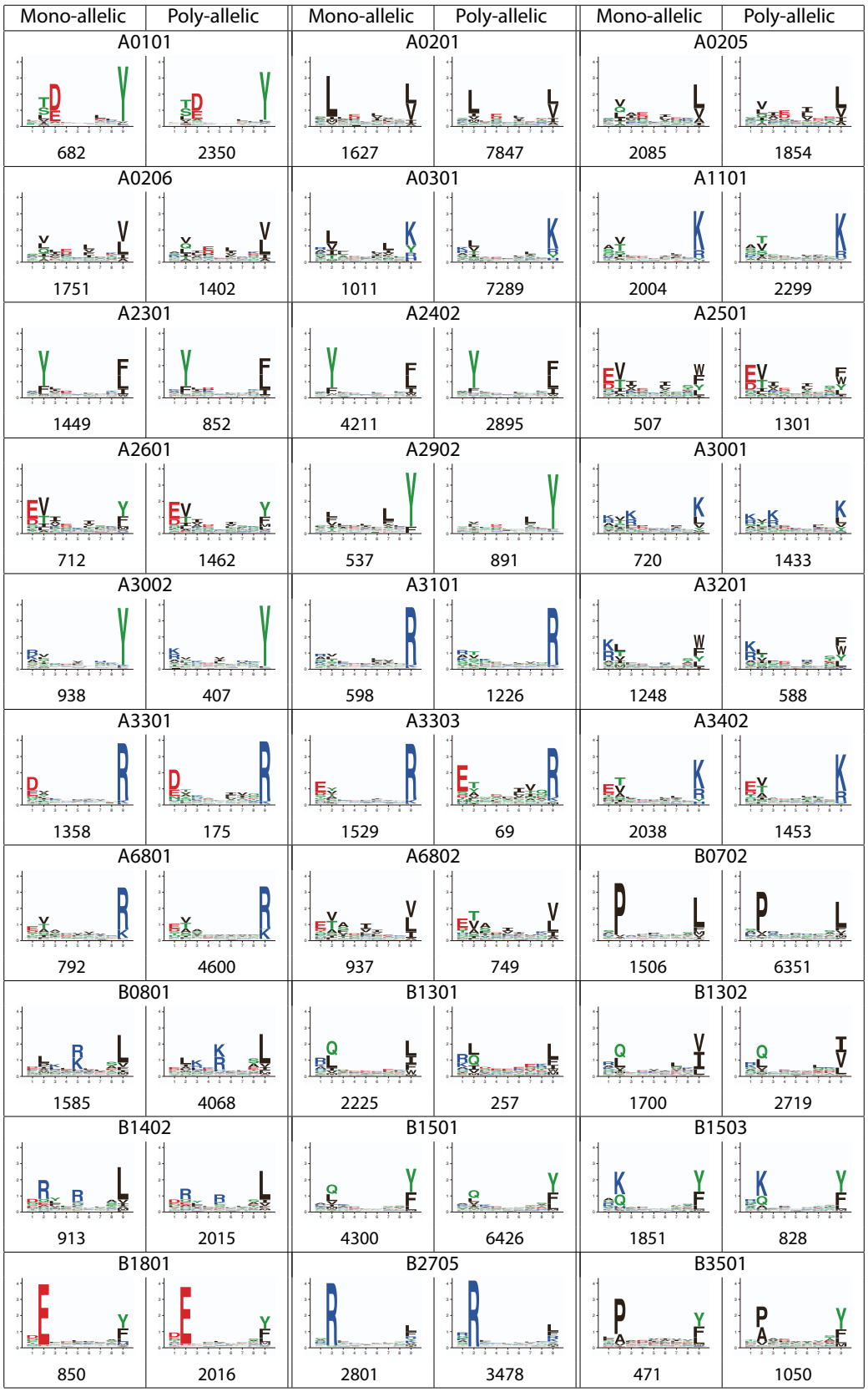
## Supplemental information

# Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 $CD8^+$ T-cell epitopes

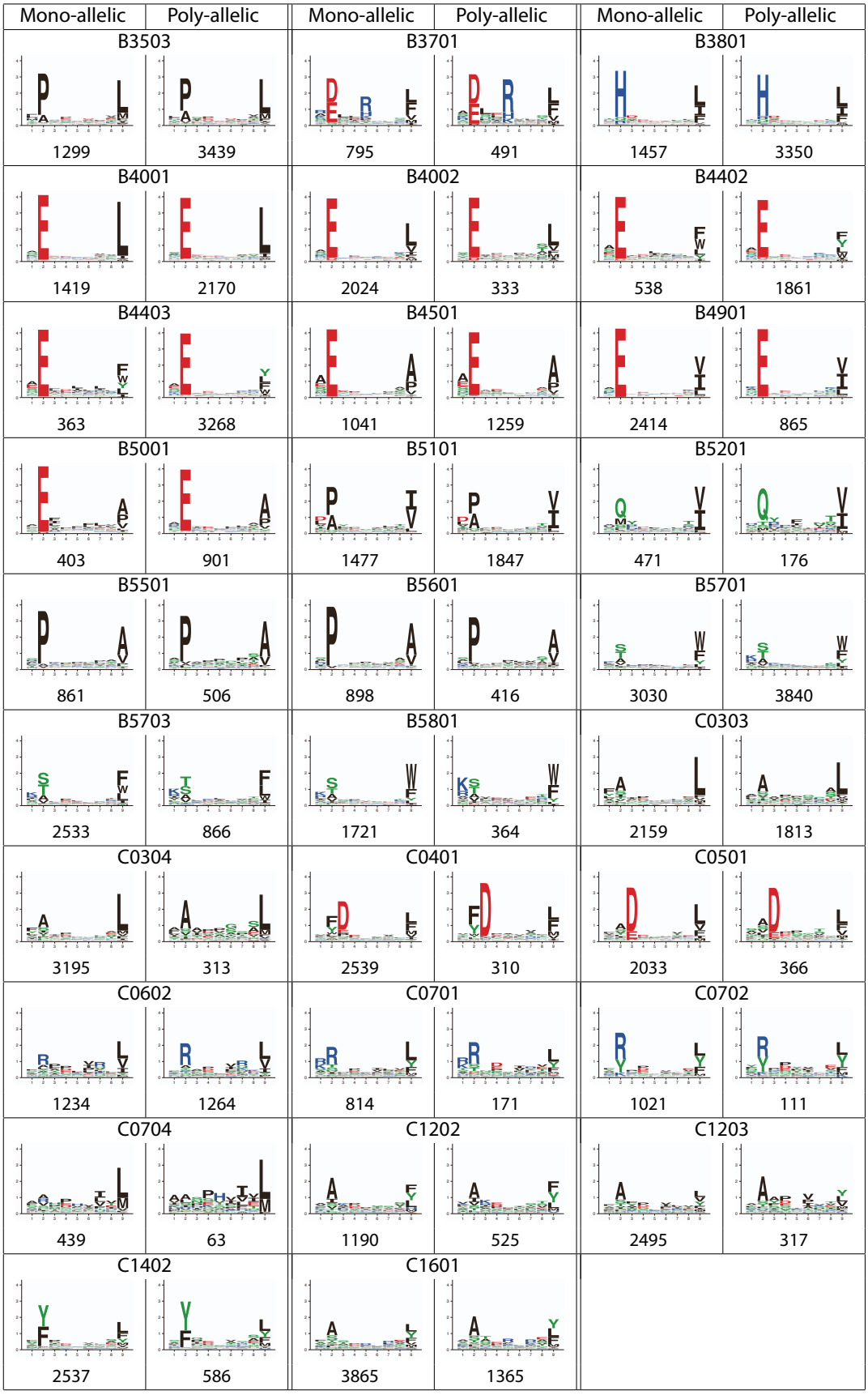
**David Gfeller, Julien Schmidt, Giancarlo Croce, Philippe Guillaume, Sara Bobisse, Raphael Genolet, Lise Queiroz, Julien Cesbron, Julien Racle, and Alexandre Harari**

# Supplemental Figures

## A



# A (second page)



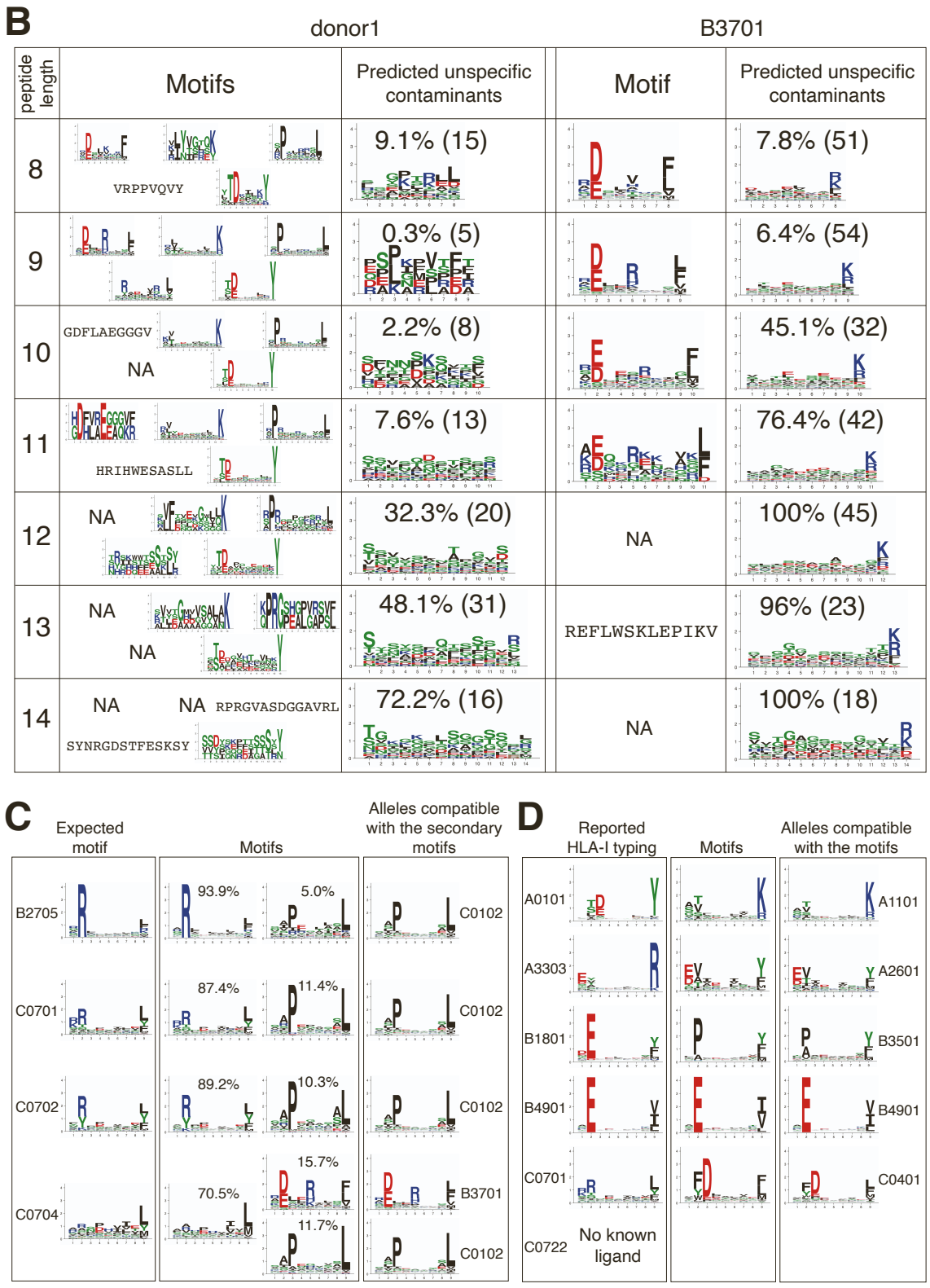



**Figure S1: Motif deconvolution reveals motifs consistent between mono- and ploy-allelic HLA-I peptidomics data and identifies different types of contaminants, related to Figure 1.** (A) Comparison between motifs identified in mono-allelic and poly-allelic HLA-I peptidomics data. All alleles found in both mono-allelic and poly-allelic samples are shown (59 in total). Numbers below each motif indicate the number of 9-mer ligands. (B) Examples of predicted unspecific contaminants (i.e., flat motif in MixMHCp) for different peptide

lengths in the poly-allelic sample ‘donor1’ using five motifs[1] and the mono-allelic sample ‘B3701’ using one motif[2]. The number of predicted unspecific contaminants is shown for each length, together with the corresponding percentage. When only one peptide is found, the sequence of the peptide is written instead of the motif. Cases without any peptide are indicated with ‘NA’. (C) Examples of predicted allele-specific contaminations in mono-allelic samples from Sarkizova et al.[2], with secondary motifs identified by MixMHCp showing high similarity with those of other alleles. The left box shows the expected motif based on data from other studies. The middle box shows the motifs obtained with MixMHCp in each sample (for clarity the flat motif is not shown, which is why the percentages do not sum up to one). The right box shows the motifs of the HLA-I alleles which the contaminating peptides are predicted to come from. (D) Example of erroneous HLA-I typing (sample ‘1180157F’ in [3]). The left box shows the motifs of the alleles reported in the original study. The middle box shows the motifs identified by motif deconvolution with MixMHCp. The right box shows the motifs of alleles compatible with those predicted by motif deconvolution.

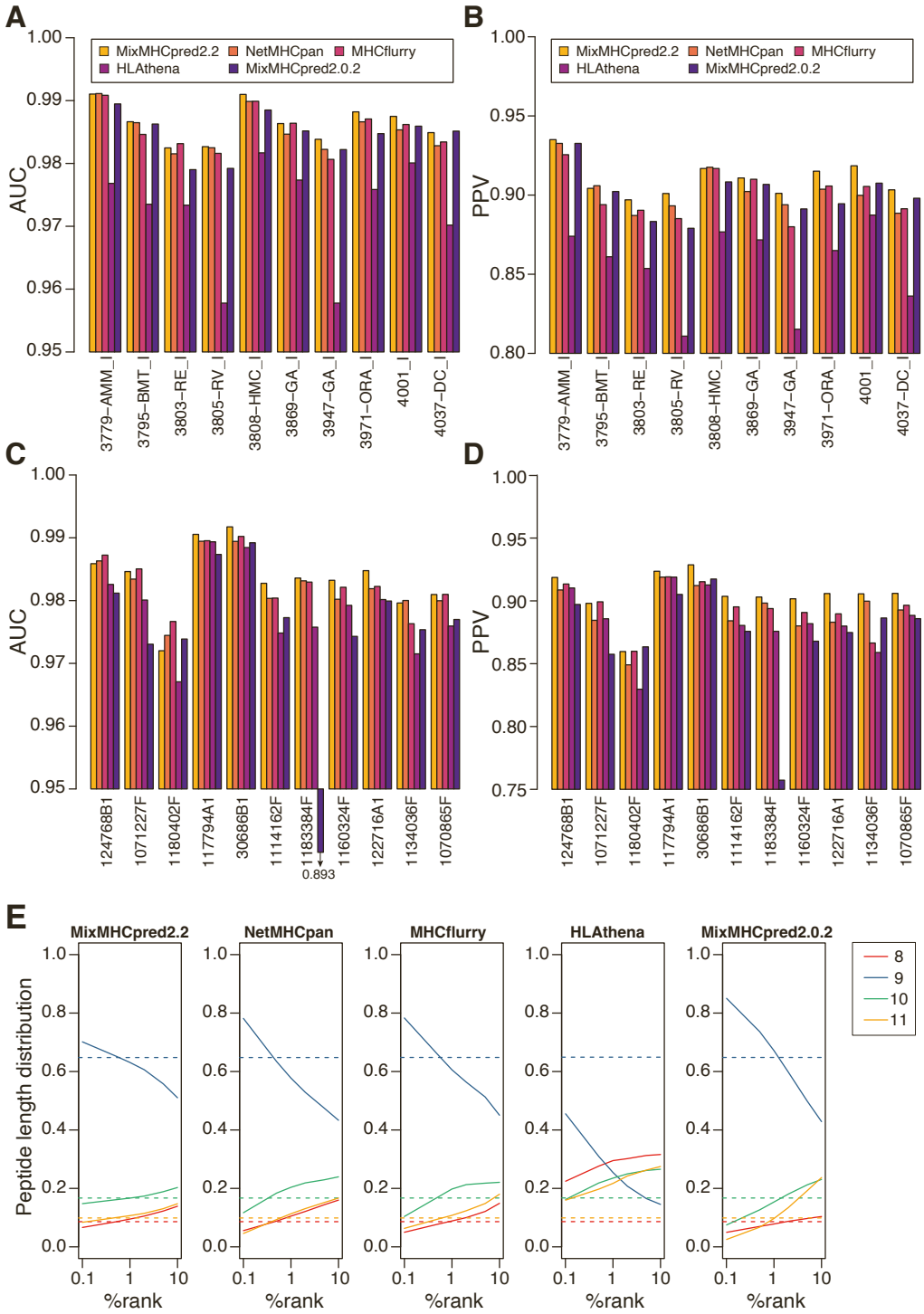



**Figure S2: Models of HLA-I binding specificities and peptide length distributions improve predictions of naturally presented HLA-I ligands, related to Figure 2.** (A-B) AUC (A) and PPV (B) values for different predictors applied on ten HLA-I peptidomics samples from Gfeller et al.[4]. (C-D) AUC (C) and PPV (D) values for different predictors applied on eleven HLA-I peptidomics samples from Pyke et al.[3]. (E) Predicted peptide length distributions at multiple %rank thresholds for each HLA-I ligand predictor. The values were computed by averaging over all alleles with mono-allelic HLA-I peptidomics data. Dashed lines show the peptide length distributions observed in naturally presented HLA-I ligands from mono-allelic samples.

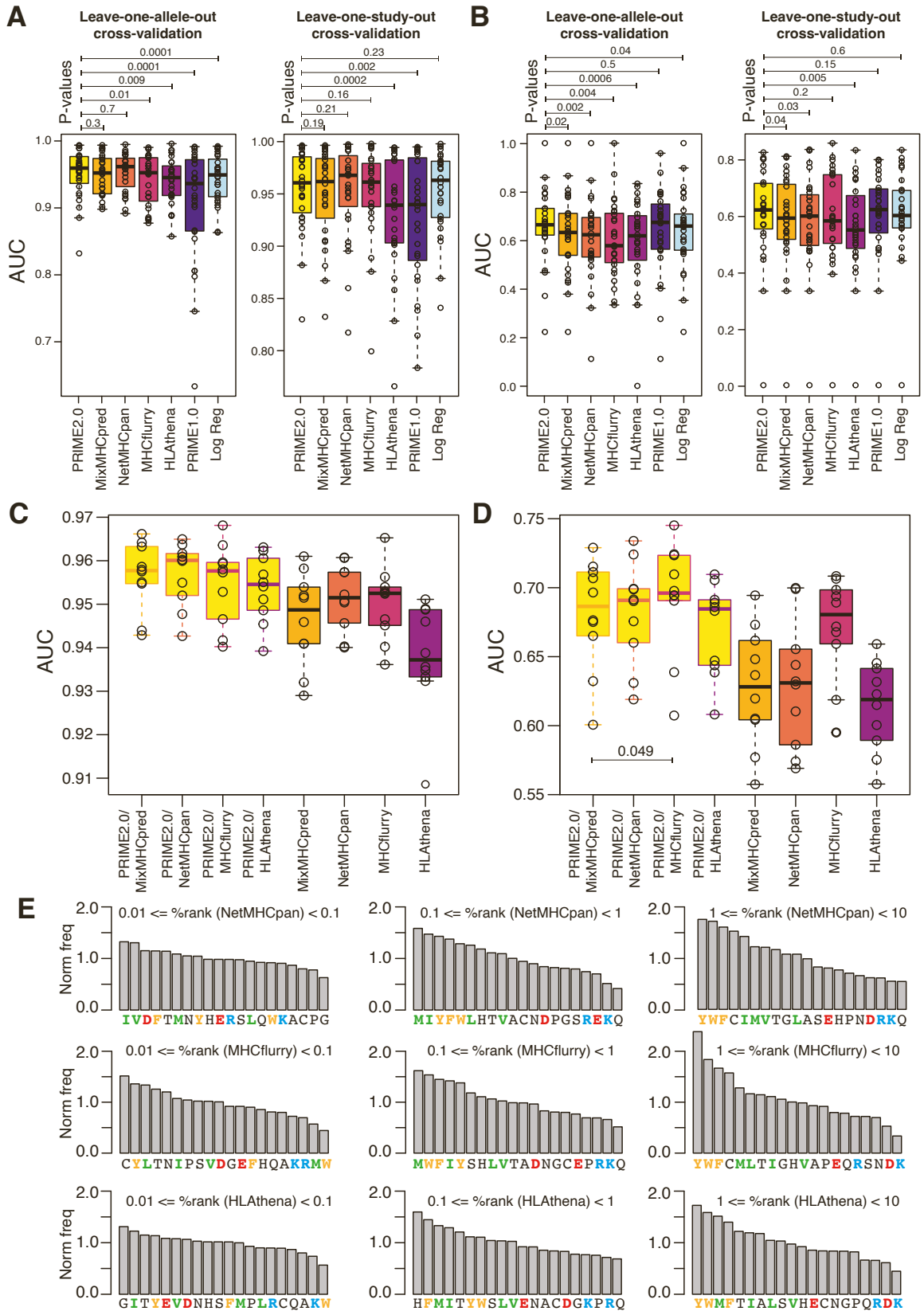



**Figure S3: Models of TCR recognition propensity improve predictions of neo-epitopes, related to Figure 3.** (A) Benchmarking of PRIME2.0 based on leave-one-allele-out cross-validation and leave-one-study-out cross-validation. (B) Same validation as in (A) after excluding randomly generated negatives in the test sets. (C) 10-fold cross-validation of PRIME2.0 trained with different HLA-I ligand predictors. For clarity, only P-values smaller

than 0.05 within the first four yellow bars are indicated. (D) Same validation as in (C) after excluding randomly generated negatives in the test sets. (E) Normalized amino acid frequencies at positions with minimal impact on predicted affinity to HLA-I for immunogenic versus non-immunogenic peptides used to train PRIME2.0 within different ranges of predicted HLA-I binding (%rank) computed with different tools (NetMHCpan, MHCflurry and HLAthena). Boxplots in panels A-D show the median and lower/upper quartiles. P-values are computed with paired Wilcoxon test.

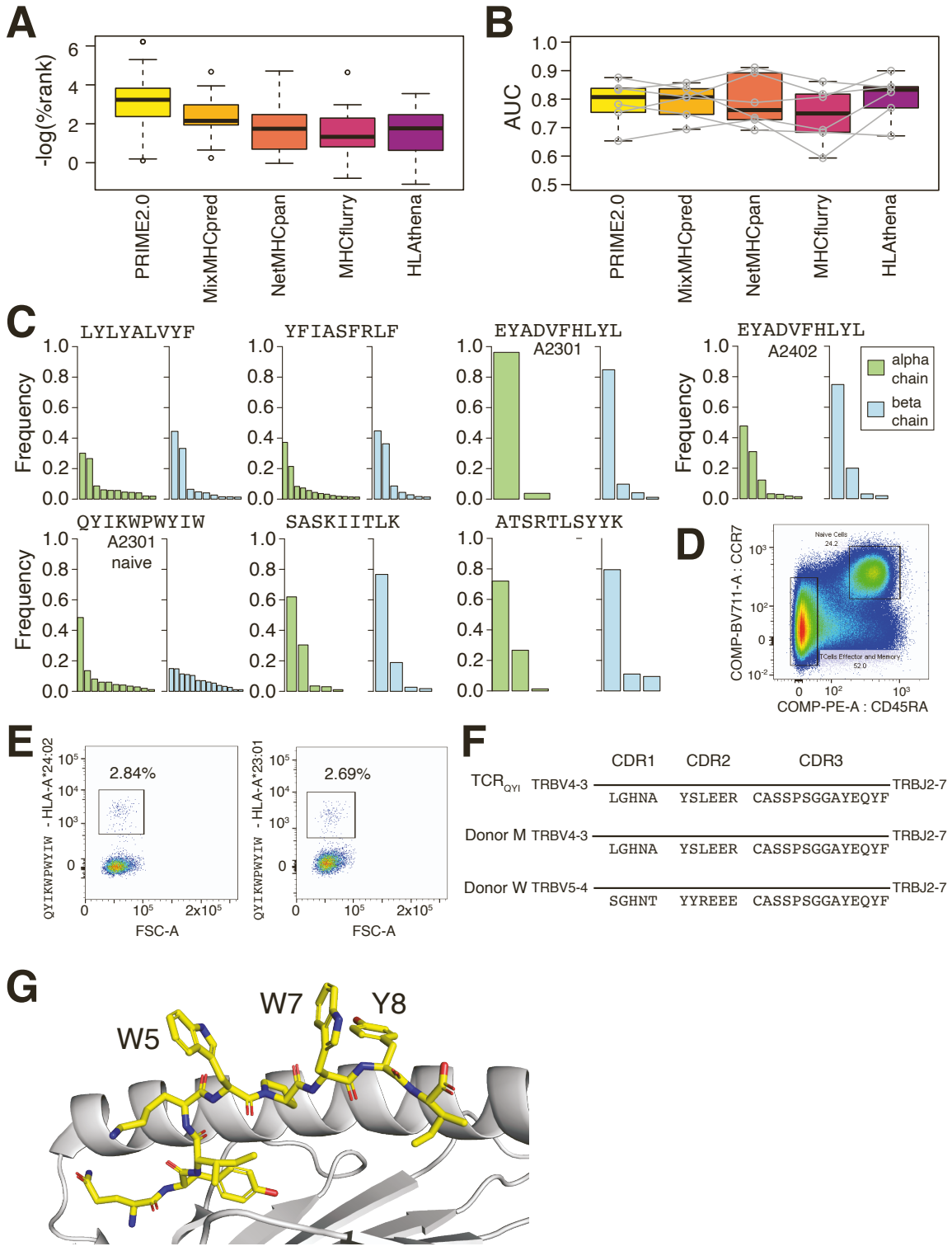



**Figure S4: PRIME2.0 identifies SARS-CoV-2 CD8$^+$ T-cell epitopes, related to Figure 4.** (A) Distribution of the scores of different predictors for the immunogenic peptides identified in each sample (best %rank across the alleles of each sample). (B) AUC values obtained in each sample with different predictors when re-predicting the immunogenic peptides identified in our ELISpot assays (see STAR Methods). No comparison between the different methods displayed statistical significance. (C) Frequency of alpha and beta chains found for each epitope with more than one alpha and beta chain in Figure 4D. (D) Sorting of naïve and effector/memory CD8$^+$ T cells in Leu184. (E) Staining of effector/memory CD8$^+$ T

cells from Leu184 donor with the QYIKWPWYIW – HLA-A*24:02 and QYIKWPWYIW – HLA-A*23:01 multimers. (F) Comparison between the sequence of the TCRβ chain found in the monoclonal $CD8^+$ T cell population specific for the QYIKWPWYIW epitope ($TCR_{QYI}$) and the closest TCRβ chains found in the two SARS-CoV-$2^+$ HLA-A*24:$02^+$ patients analyzed in Minervina et al.[5] (both donors had TCRα chains identical to the one in $TCR_{QYI}$). (G) Crystal structure of the 9-mer QYIKWPWYI epitope (yellow) in complex with HLA-A*24:02 (grey) (PDB: 7EJL)[6]. The aromatic residues at non-anchor positions (W5, W7 and Y8) point outside of the HLA-I binding pocket and towards the TCR binding interface. For clarity, the α2 helix of the HLA-I is not shown.

# Supplemental Tables

**Table S1**

**Table S2**

**Table S3**

**Table S4**

**Table S5**

**Table S6**

# References


1. Ritz, D., Gloger, A., Neri, D., and Fugmann, T. (2017). Purification of soluble HLA class I complexes from human serum or plasma deliver high quality immuno peptidomes required for biomarker discovery. Proteomics *17*. 10.1002/pmic.201600364.

2. Sarkizova, S., Klaeger, S., Le, P.M., Li, L.W., Oliveira, G., Keshishian, H., Hartigan, C.R., Zhang, W., Braun, D.A., Ligon, K.L., et al. (2019). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat Biotechnol. 10.1038/s41587-019-0322-9.

3. Pyke, R.M., Mellacheruvu, D., Dea, S., Abbott, C.W., Zhang, S.V., Phillips, N.A., Harris, J., Bartha, G., Desai, S., McClory, R., et al. (2021). Precision Neoantigen Discovery Using Large-scale Immunopeptidomes and Composite Modeling of MHC Peptide Presentation. Mol Cell Proteomics *20*, 100111. 10.1016/j.mcpro.2021.100111.

4. Gfeller, D., Guillaume, P., Michaux, J., Pak, H.-S., Daniel, R.T., Racle, J., Coukos, G., and Bassani-Sternberg, M. (2018). The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. J. Immunol. *201*, 3705–3716. 10.4049/jimmunol.1800914.

5. Minervina, A.A., Komech, E.A., Titov, A., Bensouda Koraichi, M., Rosati, E., Mamedov, I.Z., Franke, A., Efimov, G.A., Chudakov, D.M., Mora, T., et al. (2021). Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. Elife *10*, e63502. 10.7554/eLife.63502.

6. Shimizu, K., Iyoda, T., Sanpei, A., Nakazato, H., Okada, M., Ueda, S., Kato-Murayama, M., Murayama, K., Shirouzu, M., Harada, N., et al. (2021). Identification of TCR repertoires in functionally competent cytotoxic T cells cross-reactive to SARS-CoV-2. Commun Biol *4*, 1365. 10.1038/s42003-021-02885-6.