Cell Reports Medicine, Volume *2*

Supplemental Information

Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting

Julien Schmidt, Angela R. Smith, Morgane Magnin, Julien Racle, Jason R. Devlin, Sara Bobisse, Julien Cesbron, Victor Bonnet, Santiago J. Carmona, Florian Huber, Giovanni Ciriello, Daniel E. Speiser, Michal Bassani-Sternberg, George Coukos, Brian M. Baker, Alexandre Harari, and David Gfeller

Supplemental Information

Supplementary Figures

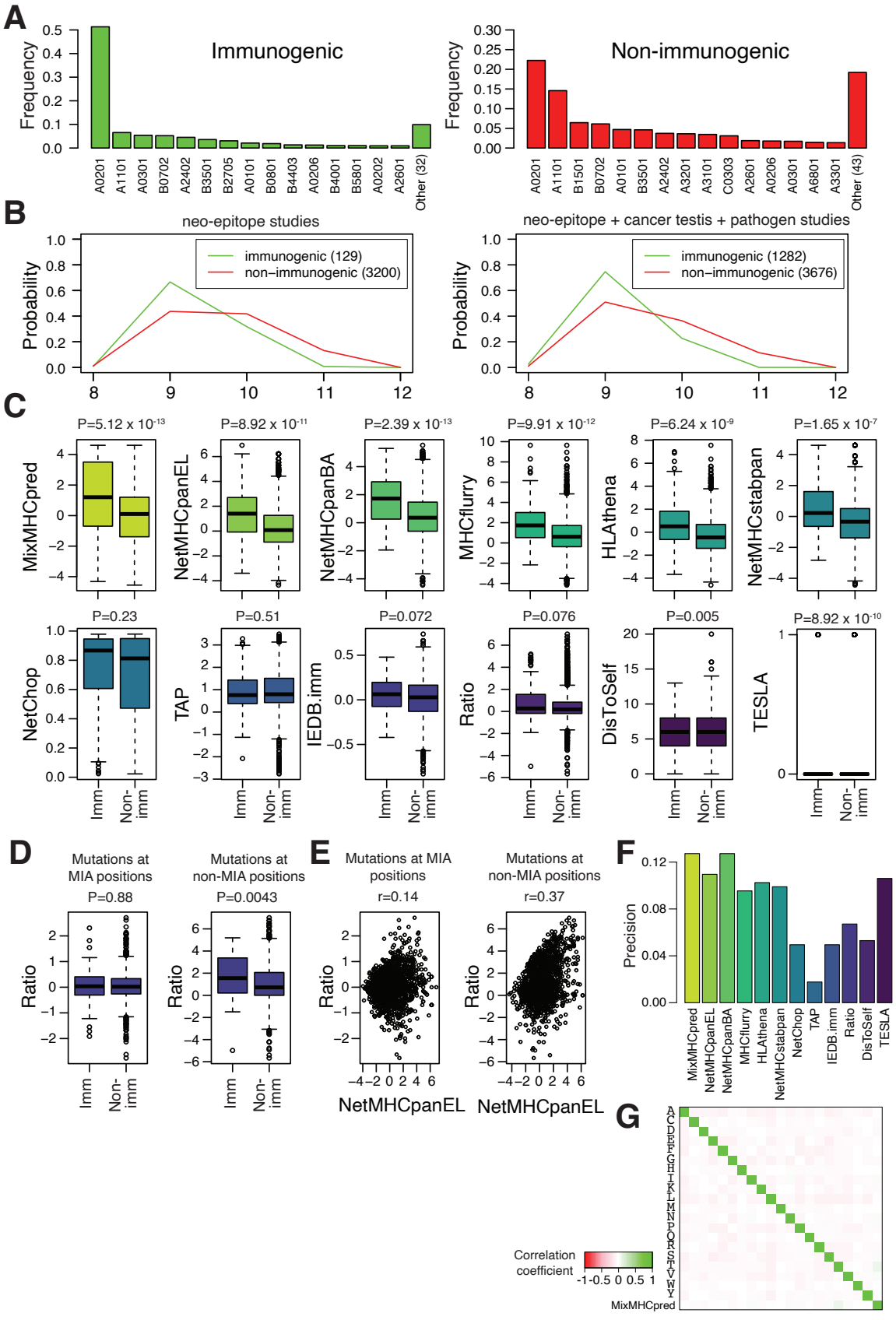


Figure S1: Immunogenicity predictions beyond presentation on HLA-I molecules, related to Figure 1.
(A) Allele coverage across immunogenic (green) and non-immunogenic (red) peptides from mutated proteins in cancer, cancer testis antigens and pathogen proteins.
(B) Peptide length distribution of immunogenic and non-immunogenic peptides analyzed in cancer neo-epitope studies (left) and cancer neo-epitope + cancer testis antigen + pathogen epitope studies (right).
(C) Distribution of the scores of existing predictors for immunogenic ("Imm", n=129) and non-immunogenic ("non-Imm", n=3'200) cancer mutated peptides used in our benchmark. P-values are computed based on Wilcoxon test and are also shown in Figure 1C.
(D) Ratio of binding affinity between the mutated and the wt for immunogenic and non-immunogenic cancer mutated peptides for mutations falling at positions with minimal impact on binding affinity (MIA positions, STAR Method) (left) or other positions (right).
(E) Correlation between the predicted affinity and the Ratio of affinity between mutant and wt for immunogenic and non-immunogenic mutated peptides with mutations falling at MIA positions (left) or other positions (right).
(F) Precision among the top *n* predicted mutated peptides for each predictor, where *n*=283 corresponds to the number of mutated peptides predicted to be immunogenic by the TESLA method.
(G) Spearman correlation between the different input features of PRIME across all peptides used in our training set.

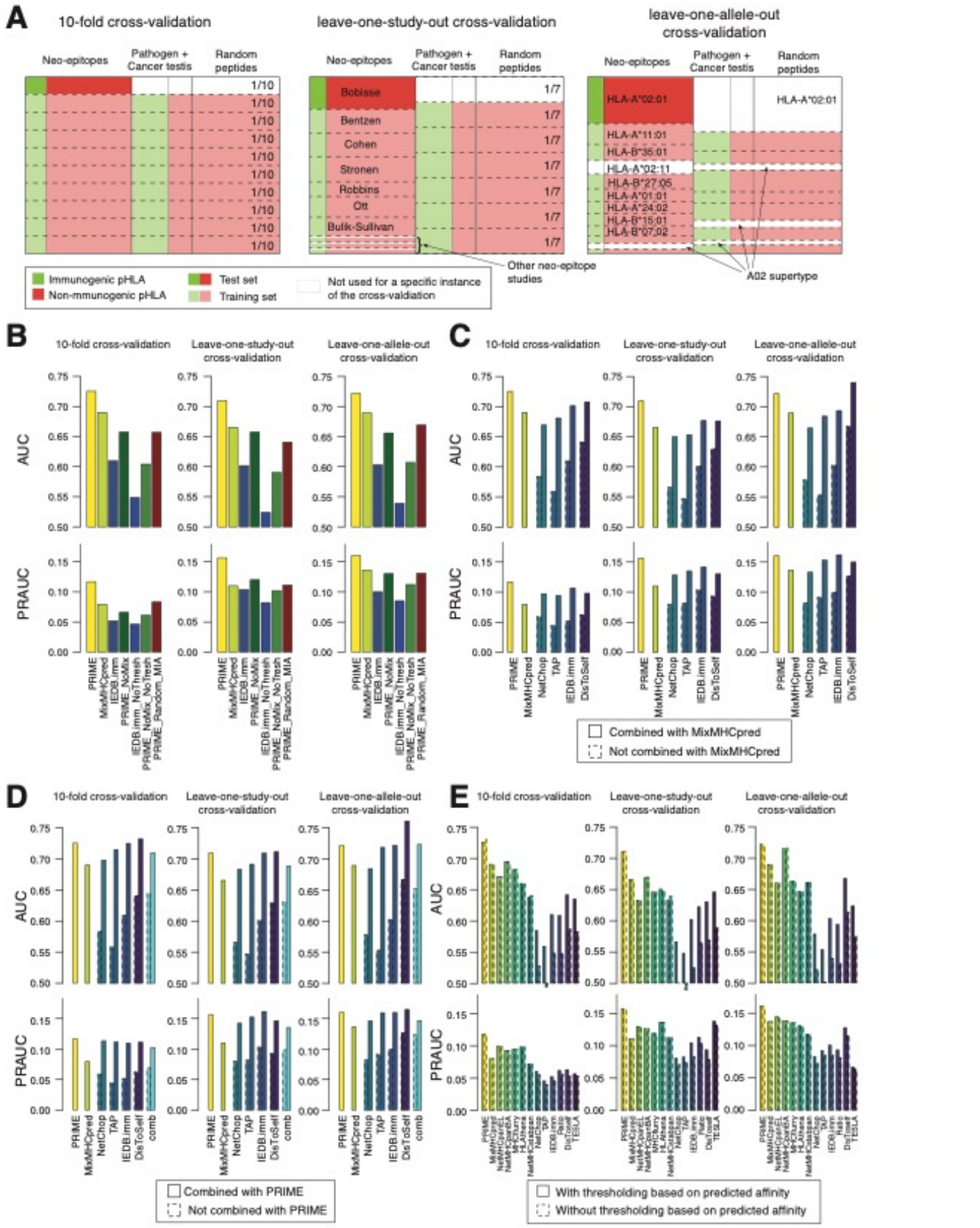


Figure S2: **Cross-validation and additional combinations of input features in PRIME, related to Figure 1.** (A) Description of the different cross-validation frameworks. Left: Standard 10-fold cross-validation where the algorithm is trained on 9/10th of the data and tested on the remaining 1/10th of neo-epitopes. Middle: Leave-one-study-out cross-validation across all seven neo-epitope studies with at least five immunogenic and five non-

immunogenic peptides. Each of the seven studies was iteratively used as test set. The other six studies were used for training, together with 6/7th of the other data (i.e., other neo-epitope studies + Pathogen/cancer testis + random). Right: Leave-one-allele-out cross-validation across the nine HLA-I alleles with at least five immunogenic and five non-immunogenic peptides in neo-epitope studies. Peptides restricted by HLA-I alleles of the same supertype as the one used in the test set (A02 in this example) were excluded from the training.
(B) Benchmarking variants of PRIME trained without predicted affinity ("PRIME_NoMix", dark green), trained without predicted affinity and without threshold on affinity values ("PRIME_NoMix_NoThresh", green) and trained after randomizing MIA positions ("PRIME_Random_MIA", dark red). For comparison, results for PRIME (yellow), MixMHCpred (light green), IEDB.imm with thresholding (dark blue) and IEDB.imm without thresholding (blue) are shown.
(C) Effect of combining NetChop, TAP, IEDB.imm and DisToSelf with MixMHCpred.
(D) Effect of combining NetChop, TAP, IEDB.imm and DisToSelf with PRIME. "comb" (cyan bars) stands for the full combination (NetChop + TAP + IEDB.imm + DisToSelf w/o PRIME).
(E) Effect of the threshold on predicted binding to HLA-I (T=5%rank, based on MixMHCpred).

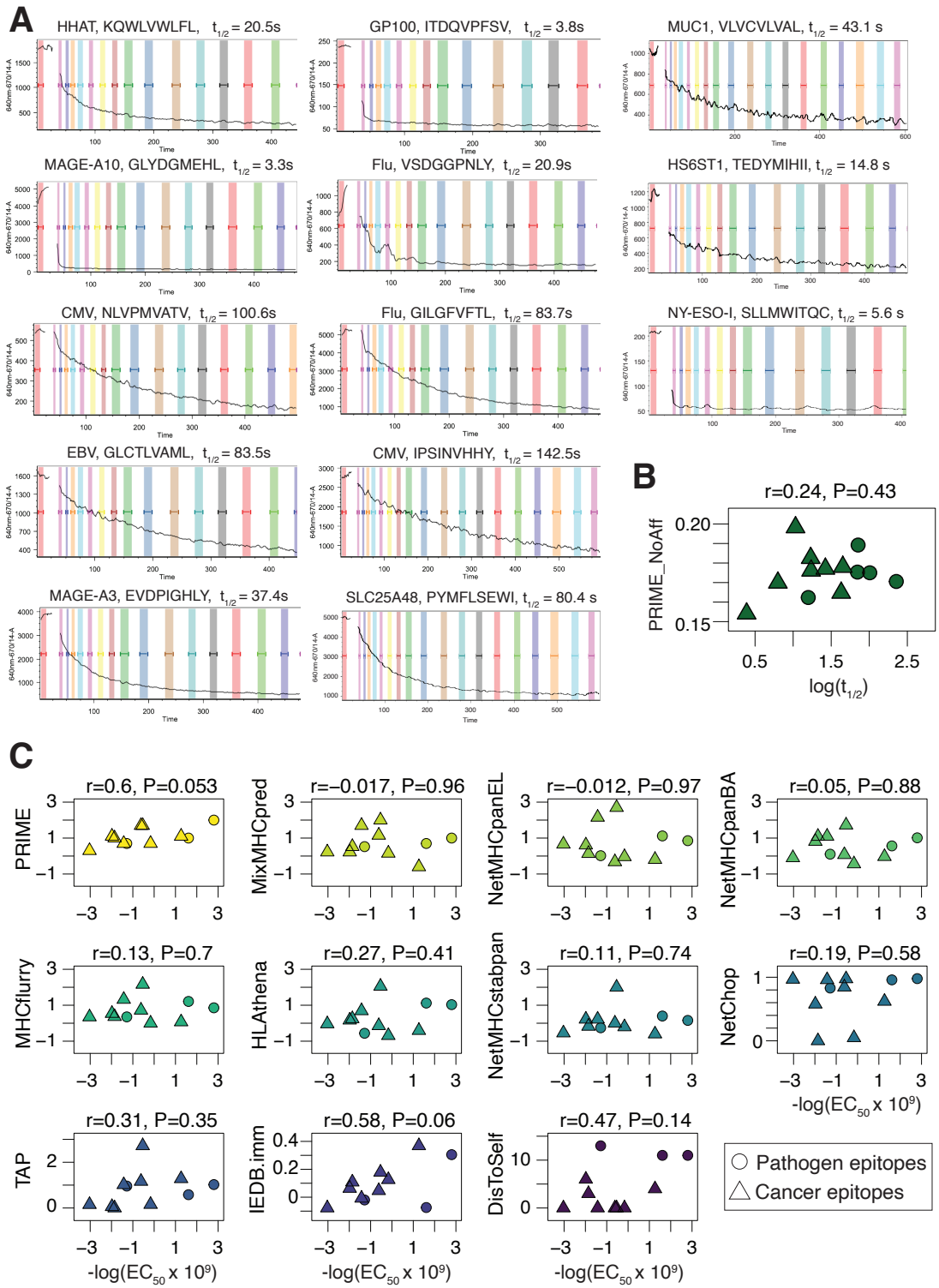


Figure S3: **PRIME correlates with structural avidity, related to Figure 3.**

(A) Representative results of dissociation assays (one for each epitope in Figure 3) for the off-rate measurements. The fluorescence intensity in the colored regions was averaged and used to fit a single-phase exponential decay.

(B) Correlation between structural avidity ($t_{1/2}$) and the predictions of PRIME trained without affinity predictions (PRIME_NoAff).

(C) Correlation between killing (i.e., -log(EC_{50}), averaged over multiple clones, see Table S4B) measured for eleven epitopes and the scores of the different predictors. Pearson correlation coefficients and P-values are shown above each plot.

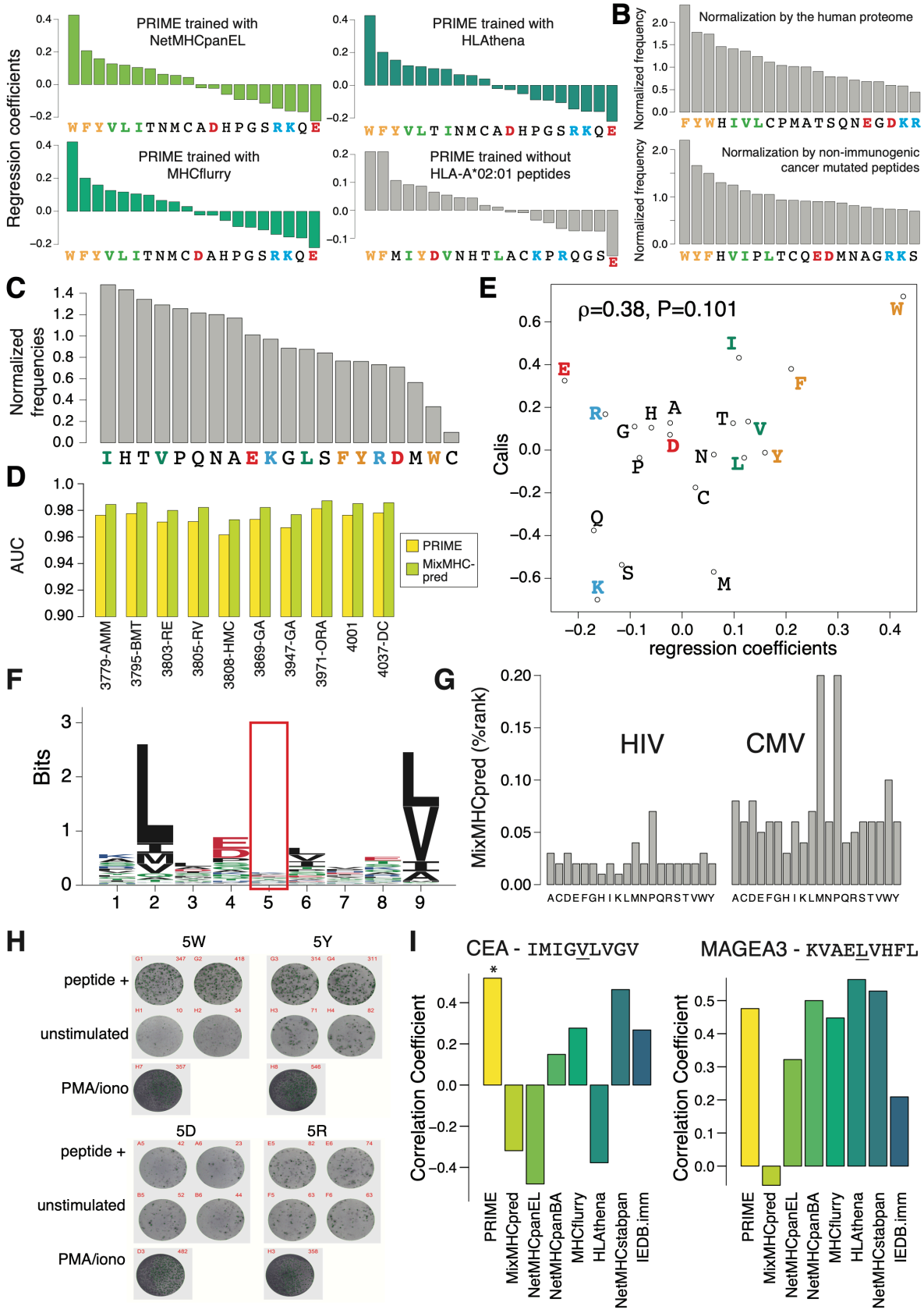


Figure S4: **PRIME reveals determinants of TCR recognition, related to Figure 4.**
(A) Coefficients of the logistic regression of PRIME using affinity to HLA-I (%rank) predicted by NetMHCpanEL, MHCflurry or HLAthena, or excluding peptides restricted to HLA-A*02:01.

(B) Normalized amino acid frequencies at MIA positions in neo-epitopes. Normalization is done either by amino acid frequencies in the human proteome (taking for each protein residues between the fifth and the second-to-last positions to mimic the definition of MIA positions) (upper panel), or by amino acid frequencies at MIA positions in non-immunogenic cancer mutated peptides with the same distribution of predicted affinity to HLA-I, same HLA-I allele distribution and same peptide length distribution as the neo-epitopes (lower panel).
(C) Normalized amino acid frequencies at MIA positions for a set of HLA-I ligands identified by MS with the same allele/length distribution as neo-epitopes. Normalization is done by the amino acid frequencies in proteins seen in HLA-I peptidomics studies (taking for each protein residues between the fifth and the second-to-last positions).
(D) AUC obtained for the predictions with PRIME and MixMHCpred of naturally presented HLA-I ligands identified in the ten samples measured in Gfeller et al. (2018) (i.e., not included of the training of MixMHCpred).
(E) Comparison between the regression coefficients of PRIME (Figure 4A) and the coefficients of amino acids reported in Calis et al. (2013) The Spearman correlation coefficient and the corresponding P-value are indicated.
(F) Motif of HLA-A*02:01. The red box shows the fifth position, which displays very low specificity.
(G) Predicted binding of all the P5 variants of the HIV (ALIRILQQL) and CMV (NLVPMVATV) epitopes with MixMHCpred.
(H) Representative results of the IFNγ ELISpot assays of Figure 4B for four different P5 variants of the CMV epitope (NLVPMVATV) with donor d2. The second line corresponds to unstimulated wells and the third line to the positive control (PMA/iono).
(I) Spearman correlation coefficient between IFNγ ELISpot signals from Tangri et al. (2001) for CEA (`IMIGVLVGV`) P5 analogs (n=17) or MAGEA3 (`KVAELVHFL`) P5 analogs (n=11) and the scores of different HLA-I ligand and immunogenicity predictors. Stars indicate P-values smaller than 0.05.

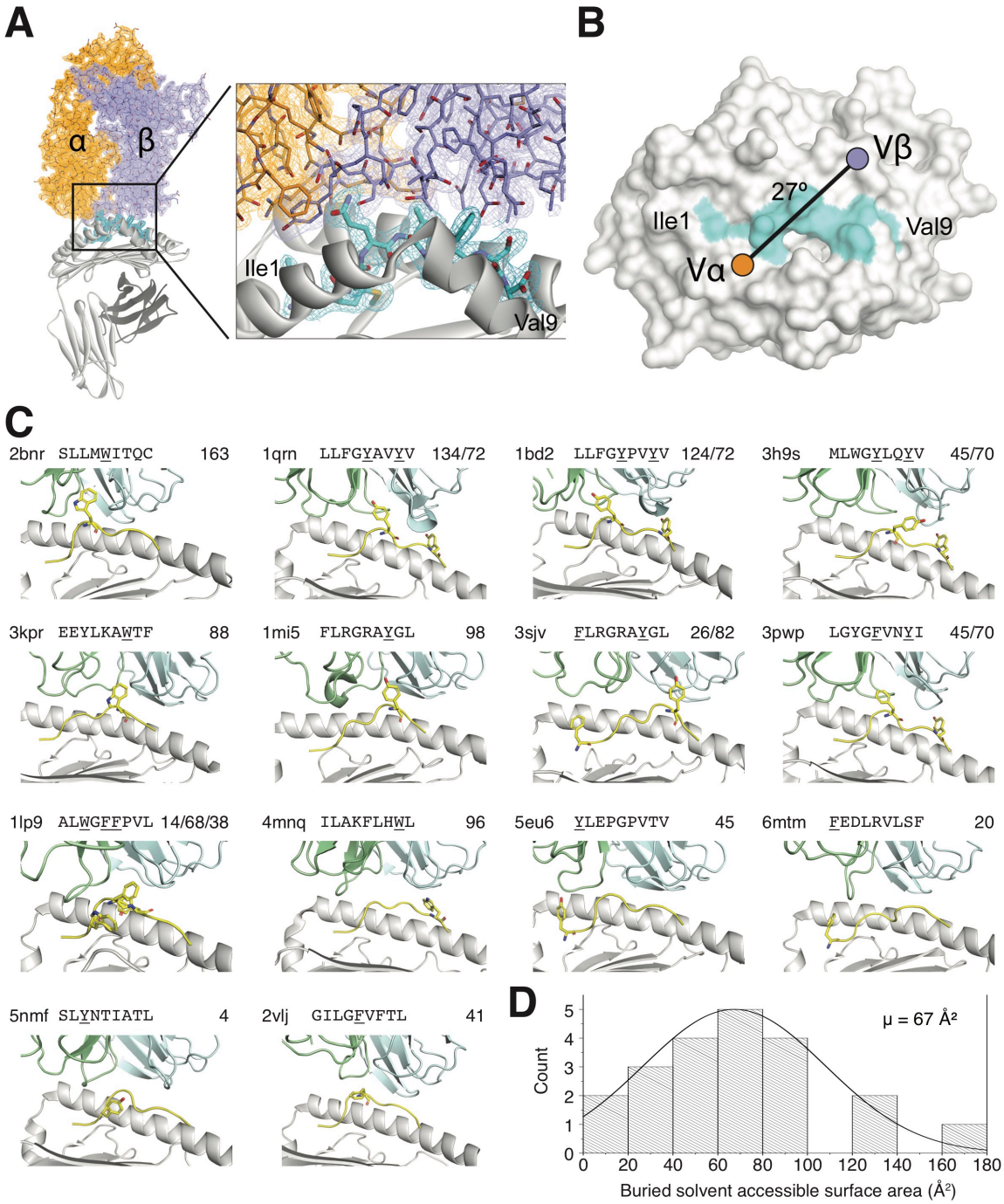


Figure S5: Structural interpretation of PRIME predictions, related to Figure 5.

(A) Structure of the complex with $2F_o$-F_c electron density contoured at 1σ shown for the TCR and the peptide.

(B) Positioning of the TCR over the pHLA. The TCR binds with a traditional crossing angle of 27°.

(C) Visualization of representative structures for 9-mer epitopes with aromatic residues directly contacting the TCR in X-ray structures from the PDB. The peptide is in yellow, the HLA-I in grey, the TCRα chain in light green and the TCRβ chain in light blue. The PDB code, peptide sequence and buried solvent accessible surface area ($Å^2$) of each underlined residue (same order as in the sequence) are given on top of each structure. Underlined peptide residues shown with sticks in the structures correspond to aromatic sidechains making direct contact (<4Å) with the TCR.

(D) Distribution of buried solvent accessible surface areas across residues highlighted in (C).

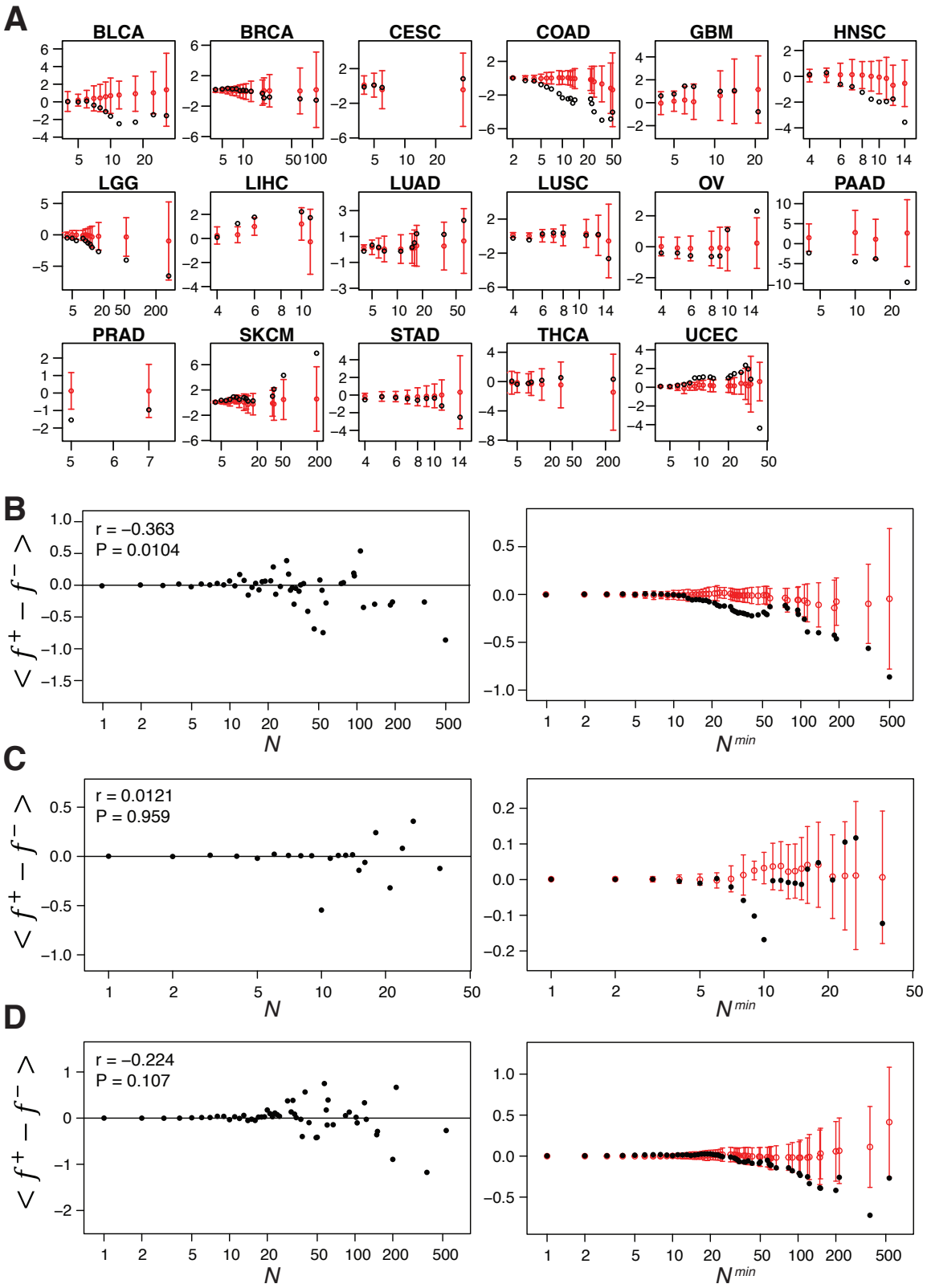


Figure S6: PRIME is consistent with immunoediting in human cancer, related to Figure 6.

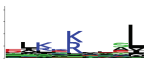
(A) Average value of $f^+ - f^-$ for mutations observed at least N^{min} times in each tumor type in the TCGA cohort. Red points and error bars indicate the results after randomly shuffling the HLA-I alleles among patients of each tumor type.

(B) Analysis of TCGA mutation frequencies in patients where they are predicted to be immunogenic (f^+) and patients where they are not (f^-), excluding from M (see Figure 6A) patients where a given mutation is found in

a poorly expressed gene, is predicted to be sub-clonal, or come from patients with deleterious alterations in antigen presentation genes.
(C) Analysis of TCGA mutation frequencies in patients where they are predicted to be immunogenic (f^+) and patients where they are not (f^-), restricting *M* (see Figure 6A) to patients where the mutation either come from a poorly expressed, is predicted to be sub-clonal, or patients with deleterious alterations in antigen presentation genes.
(D) Analysis of TCGA mutation frequencies in patients where they are predicted to be immunogenic (f^+) and patients where they are not (f^-), with predictions based on binding to HLA-I (MixMHCpred).

Supplementary Tables

Table S2: Positions with minimal impact on HLA-I affinity, related to Figure 1.
Positions with minimal impact on HLA-I affinity and potentially interacting with the TCR for each of the alleles used for training PRIME. Different values of '*g*' correspond to different groups of HLA-I alleles with distinct MIA positions (see STAR Method).

HLA-I	Motif	*g*	HLA-I	Motif	*g*	HLA-I	Motif	*g*
HLA-A01:01		1	HLA-A68:12		1	HLA-B44:03		1
HLA-A01:03		1	HLA-A69:01		1	HLA-B45:01		1
HLA-A02:01		2	HLA-B07:02		1	HLA-B49:01		1
HLA-A02:02		1	HLA-B08:01		6	HLA-B51:01		1
HLA-A02:03		3	HLA-B14:02		6	HLA-B53:01		1
HLA-A02:06		2	HLA-B15:01		1	HLA-B56:01		1
HLA-A02:11		2	HLA-B15:02		1	HLA-B57:01		1
HLA-A03:01		1	HLA-B15:17		1	HLA-B58:01		1
HLA-A11:01		1	HLA-B18:01		1	HLA-C01:02		1
HLA-A23:01		1	HLA-B27:02		1	HLA-C02:02		1
HLA-A24:02		1	HLA-B27:05		1	HLA-C03:03		1
HLA-A25:01		4	HLA-B35:01		1	HLA-C03:04		1
HLA-A26:01		4	HLA-B35:03		1	HLA-C04:01		1
HLA-A29:02		5	HLA-B37:01		6	HLA-C05:01		1
HLA-A30:01		1	HLA-B38:01		1	HLA-C06:02		1
HLA-A30:02		1	HLA-B39:01		1	HLA-C07:01		1
HLA-A31:01		1	HLA-B39:06		1	HLA-C07:02		1
HLA-A32:01		1	HLA-B40:01		1	HLA-C08:02		1
HLA-A33:01		1	HLA-B40:02		1	HLA-C12:03		1
HLA-A66:01		1	HLA-B41:01		1	HLA-C14:02		1
HLA-A68:01		1	HLA-B41:02		1	HLA-C15:02		5
HLA-A68:02		6	HLA-B44:02		1	HLA-C16:01		1