frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 21 November 2023
DOI 10.3389/fimmu.2023.1236080

OPEN ACCESS

EDITED BY
Jonathan Kaye,
Cedars Sinai Medical Center, United States

REVIEWED BY
Philip Bradley,
Fred Hutchinson Cancer Center, United States
Johannes Schetelig,
University Hospital Carl Gustav Carus, Germany

*CORRESPONDENCE
Yoram Louzoun
louzouy@math.biu.ac.il

RECEIVED 07 June 2023
ACCEPTED 26 October 2023
PUBLISHED 21 November 2023

CITATION
Levi R, Levi L and Louzoun Y (2023) Bw4 ligand and direct T-cell receptor binding induced selection on HLA A and B alleles. *Front. Immunol.* 14:1236080. doi: 10.3389/fimmu.2023.1236080

Bw4 ligand and direct T-cell receptor binding induced selection on HLA A and B alleles

Reut Levi, Lee Levi and Yoram Louzoun*

Department of Mathematics, Bar-Ilan University, Ramat Gan, Israel

Introduction: The HLA region is the hallmark of balancing selection, argued to be driven by the pressure to present a wide variety of viral epitopes. As such selection on the peptide-binding positions has been proposed to drive HLA population genetics. MHC molecules also directly binds to the T-Cell Receptor and killer cell immunoglobulin-like receptors (KIR).

Methods: We here combine the HLA allele frequencies in over six-million Hematopoietic Stem Cells (HSC) donors with a novel machine-learning-based method to predict allele frequency.

Results: We show for the first time that allele frequency can be predicted from their sequences. This prediction yields a natural measure for selection. The strongest selection is affecting KIR binding regions, followed by the peptide-binding cleft. The selection from the direct interaction with the KIR and TCR is centered on positively charged residues (mainly Arginine), and some positions in the peptide-binding cleft are not associated with the allele frequency, especially Tyrosine residues.

Discussion: These results suggest that the balancing selection for peptide presentation is combined with a positive selection for KIR and TCR binding.

KEYWORDS

selection, HLA, balancing, machine learning, allele, Bw4, T cell receptor

1 Introduction

A major challenge in understanding the evolutionary forces that act on species and affect their genetic variation is the identification of loci and positions under selection. In a simple model of directional selection, a novel mutation is favored if it confers a selective advantage to the organism (positive selection) (1). However, in some loci, balancing selection has been proposed to favor a large number of alleles in the same locus (2).

A hallmark of balancing selection is the MHC (See Table 1 for all abbreviations) region, encoding the MHC molecule that presents peptides to T lymphocytes (3), denoted HLA in humans. The HLA region is the most diverse loci in the human genome (4). The selection has been argued to emerge from the need to bind peptides from different pathogens. As

TABLE 1 List of acronyms used in the current analysis.

AA	Amino Acid
TCR	T-Cell Receptor
MHC	Major Histocompatibility Complex
HLA	Human Leukocyte Antigen
CDR	Complementarity-Determining Region
KIR	Killer cell Immunoglobulin-like Receptor
NK	Natural Killer
PB	Peptide-Binding
NPB	Non-Peptide-Binding
LILR	Leukocyte Immunoglobulin-Like Receptor
TSP	Trans-Species Polymorphism
ESP	Electrostatic Surface Potential
SVR	Support Vector Regression
RBF	Radial Basis Function

such, it is centered on peptide-binding positions in the MHC molecule (5). Classical HLA genes include two main groups - A, B and C denoted class I presenting intra-cellular peptides, and DR and DQ denoted class II, typically presenting extracellular peptides. Most of the variations among alleles are indeed concentrated in the peptide-binding regions in the second and third exons of the class I loci and the second exon of the class II loci (6). We currently have limited accuracy of DP allele frequencies. Thus, DP was not studied in the current analysis.

The main evidence for balancing selection in HLA are trans-species polymorphism (TSP) and high diversity. Many distinct mechanisms have been proposed to induce this balancing selection (7), including direct selection by pathogens, heterozygote advantage (8, 9), MHC-dependent mate choice (assortative mating) and sexual selection, including MHC dependence on mother-fetal interactions and the apparent olfactory recognition of MHC haplotypes (10).

However, in humans, MHC-I also has direct interactions with three other molecules that could affect the selection of HLA alleles. The MHC-I molecule has direct interaction with the TCR and plays a role in TCR-HLA peptide binding. Recently, the direct interaction of the TCR and the HLA was shown to be affected by the V gene and CDR3 sequence of the TCR β chain (11–14). NK cells also bind MHC-I molecules via two distinct groups of receptors, killer immuno-globulin-like receptors (KIRs) and CD94:NKG2. Natural killer cells are lymphocytes of the innate immune response that provide an important defense against infection, particularly viral infections (15–17). KIRs are inhibitory and activating receptors expressed mostly on the surface of NK cells and some T-cells. KIRs recognize broad groups of HLA class I molecules, mainly through the Bw4 binding domain in the A and B HLA alleles (18). Bw4 is a public epitope present on a subset of HLA-B and on some HLA-A alleles. NK cells can induce cell death in cells lacking Bw4.

MHC-I molecules are also the ligands for the leukocyte immunoglobulin-like receptors (LILR) of which LILRB1 and LILRB2 are the best characterized (18). A variety of HLA allotypes bind LILRB1 and LILRB2 with varying affinities, especially LILRB2, which shows considerable variation across HLA alleles (19). The LILRB1 and LILRB2 receptors are inhibitory receptors found mainly on myeloid cells such as dendritic cells and macrophages; signaling via LILR influences their activation (20). We here show that in humans, the direct interaction of MHC-I with TCRs and KIR molecules has a direct signature of selection in the HLA region.

Several methods were proposed for the identification of positions associated with selection (21), including among others, the examination of surplus in heterozygous genotypes (22), identification of local uplifted genetic variance (23), polymorphisms (24), changes in the range of sites frequencies toward common frequencies (25–27), deviation of genetic diversity from neutral models (28), presence of trans-species polymorphism (29, 30), explicit models of polymorphism patterns (31, 32), correlation of environmental features and allele frequencies (33), and others. Most of these methods are based on the distribution of nucleotides and amino acids at the appropriate position. As such, they are indirect evidence for selection. Recently some frequency-based methods were also developed (34–37). Such methods are based on the principle that non-neutral evolution leaves a signature of selection on the allele frequencies.

A more direct measure of selection would be to measure the effect of each amino acid in each position on the allele frequency. While in most genes the sampling depth and the polymorphism do not allow for such a direct measurement, the HLA locus is polymorphic enough (over 24,000 alleles in A, B, and C, and more than 7,000 as defined by the amino acid sequence of exons 2 and 3 only), and has a large enough coverage (over 39 million typed donors worldwide) (38, 39). We have recently demonstrated the validity of the frequency estimates of HLA haplotypes and their adequacy for population structure modeling (37, 40, 41). We here show for the first time that allele sequence can be used to predict allele frequency on a test set. We then show that the coefficients of the prediction algorithm highlight a strong additional selection induced on the HLA locus on regions not binding the peptide, but rather NK cells or directly the T-Cell Receptor (TCR). To the best of our knowledge, this is the first prediction of allele frequency in the human population from their sequence in any locus.

2 Results

To show that the HLA allele amino acid composition can be used to predict the frequency of an unseen allele, we regressed the log allele frequency on the amino acid composition represented as a one-hot per position (see Online Methods for formalism and training-test division and Figure 1A for a schematic scheme). We used the HLA allele frequencies imputed from the HLA typings from 6.59 million donors of the National Marrow Donor Program registry. The frequencies are divided into 21 detailed and 5 broad sub-population across the US (42) (see Supp. Mat. Table S1 for

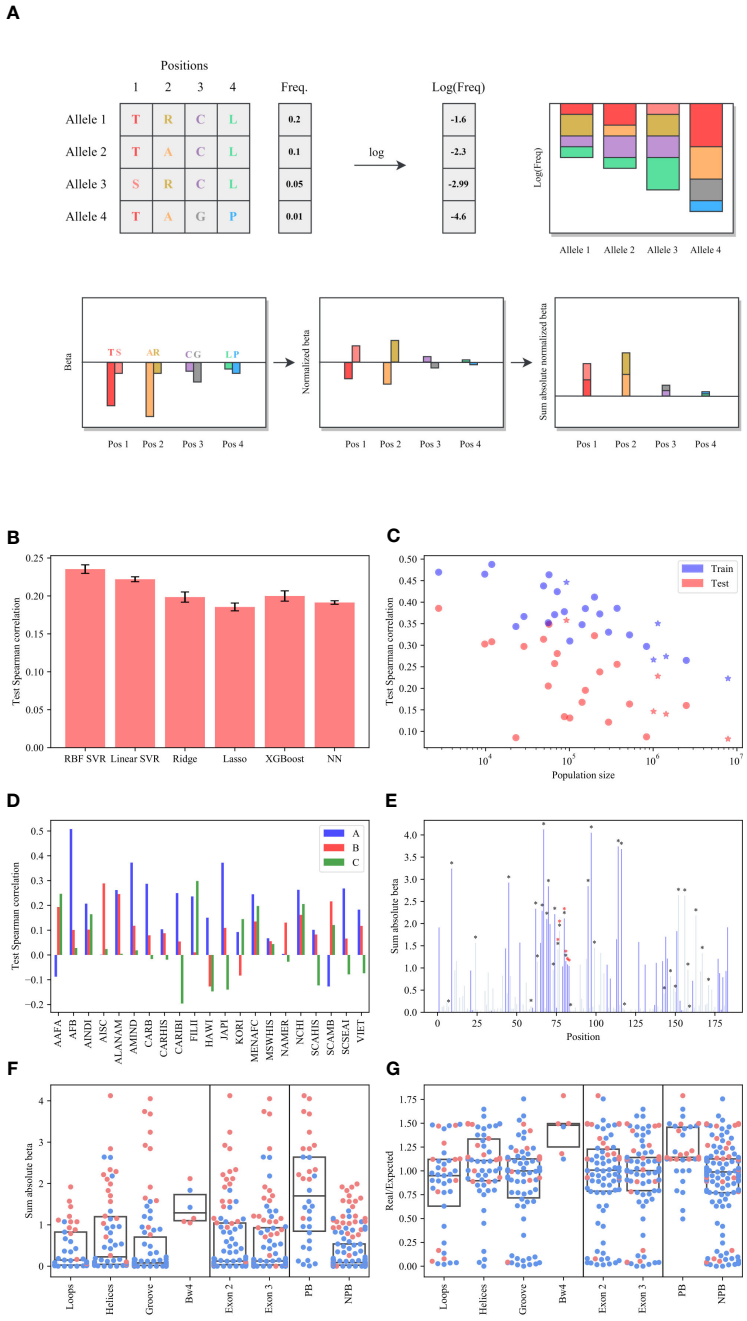

FIGURE 1
Frequency prediction. **(A)** Schematic description of S_β estimate. We regress the log frequency on the AA in all positions and obtain a coefficient β per AA per position. We normalize the sum of β to be 0 in each position. S_β is the sum of the absolute of β values per position. **(B)** The average Spearman correlation test between the predicted and real log frequencies over all populations for different models. Complex models actually have a worse prediction than linear models. **(C)** Spearman correlation as a function of the population. The stars represent the broad populations (AFA, API, CAU, HIS and NAM). **(D)** The Spearman correlation between the log real y values (allele's frequencies) and the predicted ones by SVR on the test set (the amino acid sequence of each allele) for each locus separately and each population, where the blue bars represent the A locus, the pink bars represent the B locus and the green bars represent the C locus. The A and B loci have consistent positive correlations, while the C locus has no correlation. **(E)** The sum of the absolute of β values per position (the β values are defined as the regression coefficient of the SVR), where the black/redstars represent the peptide-binding or the Bw4 positions respectively, and the dark blue bars represent the significant positions. **(F, G)** The distribution of the S_β values **(F)** and the d_n/d_s values **(G)** for each region.Pink dots are significantly different from the null model.

details). We tested multiple linear and non-linear regression methods for each population. Formally, for the linear regressors, each population j, and each allele i in locus L (A, B or C), with frequency $a_{i,j}$, we mark:

$$loss = \sum \left(log_{10}(a_{i,j}) - \sum_k \beta_{j,k} \star x_{i,k}\right)^2 + g(\beta), \quad (1)$$

where $x_{i,k}$ is a one-hot representation of the allele sequence, $\beta_{j,k}$ are the coefficients for the appropriate population and $g(\beta)$ is a

regularization term that varies among methods (e.g. Ridge LASSO). The formalism is similar in non-linear methods (see Methods). We have also tested the possibility of regression of all loci simultaneously. In such cases, an additional term was added to the regression representing the locus. Finally, we also performed a similar regression on all populations simultaneously. In this case, an additional one-hot term was added for the population (see Methods).

The RBF Support-Vector Regression (SVR) produced the highest average test correlation (Figure 1B), but a linear SVR had almost similar scores (ANOVA test between all the models $p < 9.74e-27$, T-test between the RBF SVR and the linear SVR $p < 0.001$). Thus, to get a simple explanation of the coefficients, we trained the SVR model with the linear kernel on all loci together. We thus used the linear SVR model for all loci together.

Note that more precise results can be obtained for specific loci and populations using other models (Supp. Mat. Table S9), but as further mentioned their coefficients fail to detect previously reported selection, and were thus not used. Moreover, the samemodel is mainly predicting the difference between populations, and not the direct effect of the sequence on the log-frequency.

The correlation between the predicted and real log frequencies decreases with the population size (Figure 1C, the broad populations (AFA, API, CAU, HIS and NAM) are marked with a star), as a result, the correlation for the broad groups is lower than for the detailed groups in general. We thus focus on the detailed groups in the remainder of the analysis. The correlation is highest in the A locus (Figure 1D), followed by B (0.178 vs 0.102). The average correlation in the C locus is almost null (0.03) and non-significant.

This may be due to the sequence differences and failure to learn from A and B to C which has fewer alleles. To check that this is not the case, we performed a regression on each of the loci separately. Again, in the C locus, the prediction models fail to predict the frequencies of the alleles (Supp. Mat. Figure S3). Therefore, the lack of prediction at the C locus is not due to its difference from the A and B loci, nor is it because of the number of alleles, which is similar among loci (2,196 in C vs 2,477 in A and 3,219 in B).

In the linear models, each amino acid at each position is associated with a coefficient (β), we computed the sum of the absolute of β values (S_β) per position. A low S_β implies that mutations in this position have a minimal effect on the allele frequencies, and a high S_β implies that some AAs in this position are strongly correlated with a high or low allele frequency. As is the case for most selection measures, this is no proof of causality, since different positions may be in Linkage Disequilibrium (LD).

The regression coefficients were consistent among the different populations, with an average correlation of 0.3 ± 0.009 over the large enough coefficients ($\sum_i |\beta_{i,j,k}| > 10^{-2}$, when computing the correlation on all positions, it is closeto 1, but this is because many positions have values near 0).

S_β is per definition biased toward positions with a more diverse amino acid composition (Supp. Mat. Figure S4). This is expected since such positions are also the ones most associated with selection. Still, we have examined several possible methods of S_β estimation, including the sum of absolute values (as above), the average of absolute values, and the average of absolute values weighted by the frequency of each AA at the appropriate position. The sum of absolute values best reproduces known results on the selection affecting the peptide binding domain and was thus kept (Supp. Mat. Figure S5A vs S5B, C). Similarly, a single model trained on all the populations together had less distinctive S_β values in the peptide binding region than outside (Supp. Mat. Figure S5D), and was thus ignored.

As expected, the positions with the highest S_β are the peptide-binding region positions. However, surprisingly, those are followed by the Bw4 KIR ligand (see Online Methods for the definition of HLA positions, Figure 1E and Supp. Mat. Figure S8), where the black/red stars represent the peptide-binding or the Bw4 positions respectively). The dark blue bars represent the significant positions. A significant position is defined as S_β larger than the 95th percentile of the S_β values in the null model (where all the frequencies are mixed - see Methods). Note that there is some overlap between PB and Bw4. However, there is a clear selection for Bw4. Positions 80-83 are significantly selected, and only 80,81 are PB, while 76 and 77 are not selected and are PB.

We further divided the 183 positions of exons 2 and 3 into 4 regions (the loop, helices, groove, and the Bw4 region), two exons (exon 2 and exon 3 regions) and peptide-binding/non-peptide-binding regions (PB vs NPB - see Table 1 for all abbreviations and Supp. Mat. Table S5 for all groups' positions). We computed S_β for each region (Figure 1F). The S_β values in the Bw4 region and the PB region are the highest, and significantly different than others for A and Bloci. No difference was detected between the other divisions (Kruskal Wallis test and U-test p-values results are shown in Supp. Mat. Table S4).

To validate the selection in positions outside the peptide binding domain, we compared the S_β based prediction to a more classical (albeit indirect) method - the ratio of non-synonymous to synonymous substitutions ($\omega = d_n/d_s$) 43 population. ω measures selection pressures by comparing the rate of synonymous (d_S) and non-synonymous substitutions (d_N) at each codon. The expected ratio d_N/d_S is computed assuming an equal mutation rate at all positions, but different rates between or within purines and pyrimidines. If selection favors new mutations affecting the phenotypes, a higher ratio is expected, and vice versa. This intuitive interpretation of d_N/d_S is supported by theoretical work on the relationship between the d_N/d_S statistic and the underlying selection pressure in a Wright-Fisher model (44, 45).We compared the S_β based results with the ω based results and obtained a similar trend, but a much clearer signal of S_β (Figure 1F) than for d_n/d_s (Figure 1G for all loci and Supp. Mat. Figure S1 for A, B, and C separated) in Bw4 and PB. Note that many NPB positions also have high and significant S_β values and some PB have low S_β values, as further discussed. We have repeated the results here with a Kimura model (46), with similar results (Supp. Mat. Figure S6).

High S_β values may simply represent the appearance time during HLA evolution. AA appearing at some position early in the HLA evolution can be expected to be associated with frequent alleles. To test if high β values only represent evolution time, the time of appearance of each amino acid in each position in the HLA phylogeny was compared with S_β.

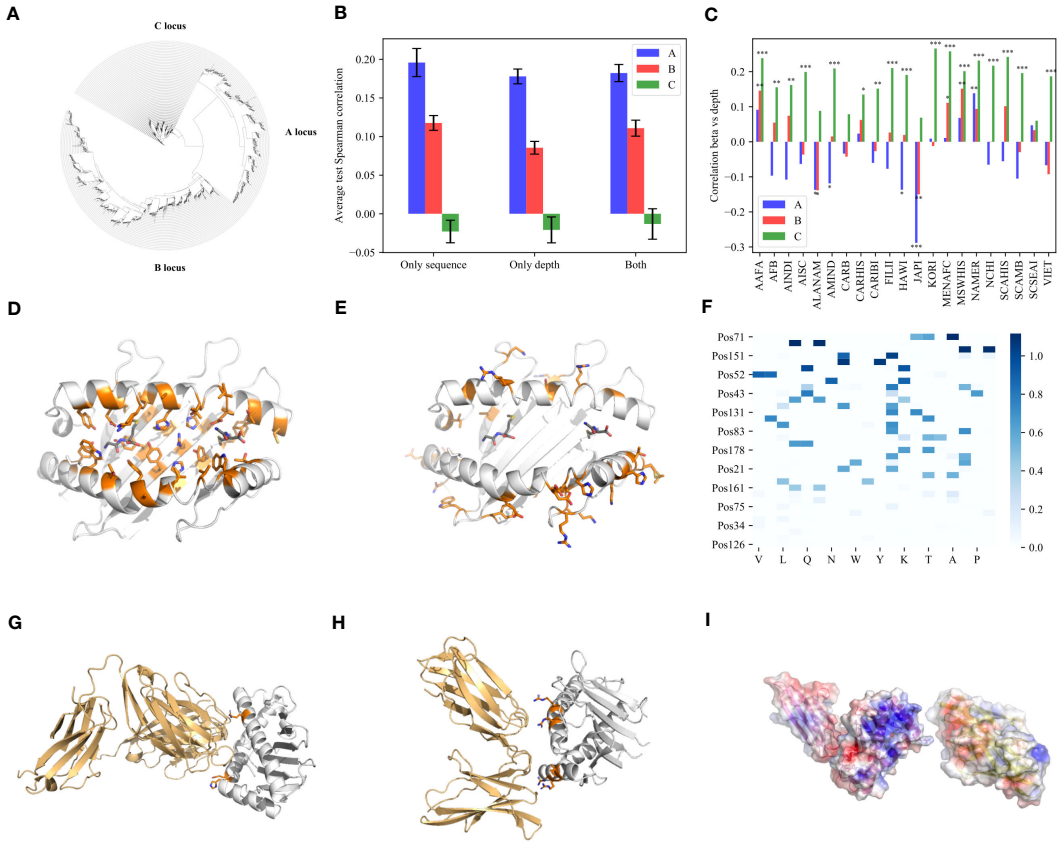

FIGURE 2
Depth. **(A)** A phylogenetic tree built with all the HLA class I nucleotide sequences, using RDP4 and the maximum likelihood algorithm on A, B and C loci together. Except for very rare cases, A, B and C are clustered separately. **(B)** The average test correlation over all populations for each of the three models (only sequence, only depth and both). The blue bars represent the A locus, the pink bars represent the B locus and the green bars represent the C locus. **(C)** The correlation between the weights (β) and depth (H) vectors over all positions and AA for each population separately. **(D, E)** PyMOL visualization of the positions side chain in the PB region **(D)** and in the significant NPB region **(E)**. **(F)** Heatmap of $\beta(AA, Pos)$ values of all the significant positions in the NPB region. One can clearly see a dominant effect of Arginine (R). **(G, H)** PyMOL visualization of the MHC class I and the positions in the significant NPB region binding to TCR **(G)** and KIR molecules **(H)**. The orange color represents the positions and their side chain. **(I)** PyMOL visualization for the electrostatic surface potential (ESP) of MHC (left) and KIR (right). The blue color represents positive EPS and the red color represents negative EPS.

The HLA locus is known to have passed recombination and gene conversion events (47). Thus, standard phylogeny may fail to capture the HLA locus evolution. To detect such events, we first built a tree of all class I alleles together (Figure 2A), using RDP4 (48). RDP4 builds phylogenies and in parallel, detects recombination events. When the phylogeny of A, B, and C HLA alleles was computed on the same tree, a clear separation into the A, B, and C loci appears, except for B*07:13, B*67:02, B*73:01, and B*73:02, which appeared on different branches of the tree than the other alleles in their locus (Figure 2A, for a higher resolution view of the tree, please refer to the following link: https://itol.embl.de/tree/1096723843363115470235 41). We repeated the analysis per locus (A, B, and C) without these alleles to detect within locus recombination events, which were further removed from the analysis (37 out of 7,892 alleles to have passed recombinations or gene conversion within exon 2 or 3 - Supp. Mat. Table S2). We thenanalyzed exons 2 and 3 separately to avoid between exon recombinations (49) using the PHYLIP package (50), without the removed alleles mentioned in Supp. Mat. Table S3. Some of these events were previously reported and others are new. The phylogeny was performed at the amino acid level to be consistent with the regression analysis. Thus, amino acid conserving convergent evolution events (the same amino acid with different nucleotides (51)) were ignored.

We defined $H(AA, Pos)$ to be the average depth of each AA in the phylogenic tree of all HLA alleles (see Online Methods) at each position. We then tested whether the allele frequencies can be predicted using $H(AA, Pos)$. Three models were tested: A) Only sequence The sequence model used above. B) **Only depth** prediction using only $H(AA, Pos)$, and 0 when an AA was observed less than 3 times in a position, as in the first model. C) **Both** Both values as input (Figure 2B). Note that the depth model contains the sequence information since it also has 0 for AA not in the sequence, but it also contains information of the depth of each AA. One can see adding depth does not improve the prediction accuracy. Thus, the allele frequency is not strongly affected by the appearance time of its amino acids if at all. To further test for association between $\beta(AA, Pos)$ and $H(AA, Pos)$, we computed the

correlation between the weights in the depth independent model (A) and the depth vectors: β vs H over all positions and AA for each population separately (Figure 2C). In the A and B loci, correlations are weak and around 0, while in the C locus, correlations tend to be positive and very significant ($p < 0.001$) (Figure 2C). Thus, the allele frequency in the C locus is strongly associated with their appearance time, in contrast with the A and B loci (Figure 2B). The time of appearance is not generalizable to new alleles. As such it cannot be used to predict frequency. This is consistent with the lack of prediction of the C allele frequencies using β. Note that if allele frequency would be fully driven by peptide-binding, one would expect no difference between A, B and C.

Beyond the selection induced by Bw4 and PB domain, there are positions in the PB region with low S_β value, and positions not in PB and not in Bw4 with high beta value. Out of the 9 insignificant positions in the PB region with $S_\beta < 1$, 6 of them (66%) are Tyrosine (Supp. Mat. Table S7 for all AAs of these positions), Tyrosine is known for its low evolution rate, among others, because of the neighboring stop codon (52).

In contrast, there are also high S_β values in the NPB region (Supp. Mat. Table S6). We used PyMOL (53) to compute the positions of their side chain. Interestingly, all these positions are predicted to face outside of the binding cleft toward the T-cell itself or other binding cells (Figure 2E), in contrast to the positions in the PB region that face the binding cleft (Figure 2D), suggesting a selection mediated by the direct interaction with other cells rather than the peptide.

There are two natural candidates for inducing this selection - T-cells and NK cells. To compare those, we used 3 TCR-MHC-I structures and 3 MHC-I KIR interactions with 3DL1, 2DL1 and 2DL2 receptors. We then computed the positions on the MHC molecule closest to the KIR or the TCR. 4 out of 34 significant positions were found to directly bind the TCR (positions 65, 151, 154 and 161) (Figure 2G and Supp. Mat. Figures S2A, B). 6 out of the 34 were computed by PyMOL to directly bind KIR molecules (positions 151, 145 and 79 were found to be common among all the structures, but in addition positions 142, 75 and 83 were also found in specific structures) (Figure 2H and Supp. Mat. Figures S2C, D). Most of these positions are Arginines. Some of those were found to bind two different KIR receptors. These results suggest a strong charge-mediated effect of KIR binding positions beyond the Bw4 domains, not only in B, but also in A HLA alleles. Note that the TCR variability is large. Thus, the three tested TCRs here may not represent the full variability, and the significant SPB that point outside may bind different TCRs.

To further understand the possible effect of charge on the difference in S_β among HLA positions with side chains toward other binding cells, we analyzed the $\beta(AA, Pos)$ values of all the significant positions in the NPB region (Figure 2F). We performed a Chi-Square test between the sum of the S_β values for each amino acid and the sum of the S_β values when mixing all these values (see Online Methods). The top 4 AA are R, G, H and K, with R the most significant, suggesting again that selection is strongly associated with a positive charge. A selection for charge may be simply the result of an opposite charge on the 2DL1 binding site. Indeed, when computing the electrostatic surface potential of 2DL1 molecules in front of the positions computed to bind 3DL1 in the MHC a clear negative chargecan be observed (Figure 2I, and detailed view in Supp. Mat. Figure S7). Note that a positive charge was previously reported to be crucial in Immunoglobulin binding, especially in the context of autoimmunity (54, 55). We here suggest that selection for positive charge in binding TCR and KIR may also be crucial.

3 Discussion

Most population genetics methods use indirect measures to explain the gene diversity in present populations and the allele and genotype frequencies and identify selection pressures. We have here analyzed the Human Leukocyte Antigen (HLA) genes and shown that the sequence of HLA A and B alleles can be used to predict the appropriate alleles log frequency with a linear model, where each amino acid at each position contributes a constant value to the allele log frequency. The linear model has been found to be much better than the tested non-linear model suggesting that epistatic effects are limited. Interestingly, the relation between AA sequence and frequency was only present in A and B alleles suggesting a mechanism beyond peptide binding, which is similarin A, B, and C loci.

The relation between AA at a given position and the allele frequency can be explained by either selection or the time since the AA's first appearance in the phylogeny. An AA can be associated with a large allele frequency, either because it contributes to the fitness of the phenotype, or because it is ancient. We have previously addressed this problem through the branch imbalance following mutations (36). Given the very large number of alleles with measured frequencies, we could here compare directly the depth of each AA with its contribution to the allele frequency. We have shown that in the C locus, depth and contribution to size are highly correlated, but not in A and B.

To measure selection, we defined a novel score S_β for the relation between sequence and frequency, based on the sum of the regression coefficients' absolute values. Applying this score to the MHC class I shows a clear selection in PB positions. However, there were many significant positions in the NPB region, with the strongest selection occurring at the Bw4 ligand. We computed the orientation of the AA side chains and showed that many of them bind directly to KIR even beyond the Bw4 regions. Some of the remaining positions bind directly to the TCR. We found no evidence for selection in LILR binding positions. While there are some sources in the literature of HLA positions that are reported to be bound to the LILR receptor (18), the current analysis was limited to exon 2 and 3, and the LILR binding region being farther away from the peptide binding cleft may affect other loci.

HLA allele frequencies have been argued to be mainly selected by a balancing selection for peptide-binding (56). However, our recent results suggest that the selection affecting the HLA region may be much more complex and dominated by a purifying selection at the haplotype level (37, 41). We have here shown at the AA level that a very strong selection is induced by charge-mediated interactions between KIR and TCR and the MHC molecules. Such a selection may favor specific haplotypes in parallel with the binding peptide-induced balancing selection on alleles.

Multiple caveats have to be considered when analyzing these results. The most significant is the known Linkage Disequilibrium (LD) between HLA genes (57). Selection in the HLA locus may not limited to single genes, but may work on full haplotypes. Thus, the frequency of a gene in a population may actually be affected by other genes. This may explain the limited accuracy of the prediction based only on each gene sequence. A combined haplotype-based score may improve the accuracy of the current predictor and will be further studied. Another important caveat is the effect of AA diversity. The current selection score is affected by the number of AA candidates in each position. We have tested different score combinations. A score that would avoid this dependence may further improve the accuracy of the selection estimation.

An interesting conclusion from the current study would be that some new alleles may have a higher probability of emerging in the population. To predict such alleles, one would need beyond the current results, a model for the generation probability of alleles from the existing ones.

4 Methods

4.1 Data

For the lineage analysis, we used the HLA class I allele's exon 2 and exon 3 sequences from the IMGT/HLA Database (58). To compute the allele's frequencies, we used the data of 6.59 million donor HLA typing from the National Marrow Donor Program Registry (37, 42, 59). It consists of the abundances of all different HLA haplotypes in the registry. Allele frequencies were derived as marginal sums of the haplotype frequencies. For example, to compute the one-locus A frequencies for a given allele, we merged all extended A C B DRB1 DQB1 haplotypes with the appropriate A allele into an A allele frequency (42).

4.2 Training test split

We divided the data into training and test sets using the *train_test_split* method from the python scikit-learn library (60). The first group constitutes 80% of the data and was used for training and finding the best hyperparameters. The second group constitutes 20% of the data and was used as an external test. All the results are reported on the test group.

4.3 Neural Network Intelligence

NNI (61) was used for parameter hypertuning. For each algorithm, NNI was used in two steps for a broad hyperparameter tuning. First, a grid search of a wide range of parameters was performed to get the amplitude of the regularization. The second step was to refine the outcome by setting the tuner to Tree-structure Paezen Estimator (TPE) and running another search, while considering historical measurements. We then found the hyperparameters that produce the highest Spearman correlation on the internal validation set (our metric). The search space of the hyperparameters for the best model, SVR, is presented in Supp. Mat. Table S8.

4.4 Prediction model

One-hot (OH) vectors were used to represent amino acid sequences in $\mathbb{R}^d$. Each vector describes the AA positions of the HLA of one population. These vectors, X_{ij}, are used as an input to the regression learning models, and the predicted values y_i are each allele log frequencies in the appropriate population.

$$y_i = Log(f_i), \tag{2}$$

where f_i is the frequencies vector of the i-th population. Positions with less than 3 AA differing from the majority AA were ignored.

The OH vectors were the input to an SVR learning algorithm, for each population by itself. When the model was trained on A, B and C loci together, we added a OH vector at the end of the input sequence in order to separate the different alleles. When the model was trained on all the populations together, we added a OH vector at the end of the sequence (after the one-hot vector that separates the alleles) in order to differentiate between alleles that came from different populations.

For each training test division, the Spearman correlations were averaged across all ten trials. The SVR (Support Vector Regression - python scikit-learn library (62)) gave the highest correlation on the test set. We optimized each algorithm separately using NNI (61) on an internal validation set. The parameters for the best algorithm, SVR (the linear and non-linear), are presented in Table 2.

The S_β score assigned to each position was calculated as a sum of the absolute value of the SVR coefficients attribute, which assigned a weight to the features.

4.5 Definition of HLA positions

The peptide-binding and the Bw4 positions are shown in Supp. Mat. Table S5, as defined in (63, 64).

4.6 Estimate of amino acid depth

To create the phylogenetic tree, we split our data into two exons (exon 2 - positions 1-90 and exon 3 - positions 91-183), removed

TABLE 2 SVR models hyperparameters.

	SVR	Linear SVR
Normalization	z_score	–
Kernel	RBF	Linear
C	1	0.01
Epsilon	1	2

the alleles mentioned in Supp. Mat. Table S3 and built phylogenic trees for each of the exons separately using the PHYLIP package (50). The trees were built using a Maximum Parsimony algorithm for each locus by itself. We added a single gene from another locus to each such tree (A*01:02 for B and C, and B*07:03 for A). We checked that the root is indeed between the outgroup and all the alleles within this locus for all three loci. A Fitch algorithm was then used to estimate the sequence of the internal nodes in the tree.

Then, for each node in the tree, we calculated its level in the tree (l) so that the level of the node is set to be the level of its son node plus 1, and the level of the leaves is set to be 0 as follows:

$$\begin{aligned} l_{leaf} &= 0 \\ l_i &= l_{i-1} + 1 \end{aligned} \tag{3}$$

Node $i-1$ is the descendant of the node i. Note that all the leaves in the tree are the alleles and the edges are the sequences composed of the amino acids. For each position in the sequence and each level, we calculated the probability of each amino acids at this level in the tree.

To compare the weights vector (β) to the phylogenetic results, we defined H as the age matrix as follows:

$$H_{k,j} = \frac{\sum_{i=1}^{l_k} i \cdot p_{ij}}{\sum_{i=1}^{l_k} p_{ij}}, \tag{4}$$

where j is the amino acid, l_k is the number of levels on the tree at the k-th position and p_{ij} is the probability of amino acid j in level i. We consider each value in the matrix as the depth of the j-th amino acid at the k-th position.

4.7 Correlation between beta and depth

To compare the weights vector, β to the depth above, we compared the two matrices: H^{20x183} representing the age rate of each amino acid in each position, and β^{20x183} representing the coefficient of each amino acid on each position (foreach population). We checked the correlation between the beta and the depth values for each locus and each population (as a single flattened vector). We ignored AA absent from the data at any positions.

4.8 *DN/DS* based estimates of selection

We used the nucleotide sequence of all the alleles for each of the loci from the (65) site (we ignored alleles containing non AA codes). We separated the sequence into codons, 3 nucleotides in each codon and converted each codon to its corresponding amino acid. For each column (each codon), we calculated the number of mutants in this column, and counted the number of mutants whose amino acid differs from the amino acid of the consensus in that column (*diff_aa_mutants*), where the consensus was based on the most frequent nucleotide in the same position among all loci. Then, for each column, we took the consensus codon, and changed each of its three nucleotides to the three remaining nucleotides. We converted these 3 nucleotides to amino acid and counted the number of amino acids different from the original amino acid (we divided by 9 to get a number between 0 and 1). We multiplied that number by the number of mutants in each codon to get the expected number of different amino acids. In order to get the real number of different amino acids for each codon, we divided the *diff_aa_mutants* value by the number of mutants in that codon.

Finally, for each codon, we calculated the ratio between the expected number of different amino acids and the real number of different amino acids and calculated the Chi-Square value by the following formula:

$$\begin{aligned} &Chi - Square \\ &= \frac{(NumRealChangeAA - NumExpectedChangeAA)^2}{NumExpectedChangeAA} \end{aligned} \tag{5}$$

and extracted the corresponding p-value for each codon.

4.9 MHC-I structures

We have analyzed several structures: 6TDQ (66) for visualization of PB and NPB regions, 1AO7 (67), 3WOW (68) and 1BD2 (69) for TCR-MHC-I structures, 1EFX (70), 1IM9 (71) and 5T6Z (72) for KIR-MHC-I structures, and 4NO0 (73), 1P7Q (74) for LILR-MHC-I structures. For each structure, we used (75) for calculating the distances between the MHC positions and the TCR or the KIR positions. We used PyMOL (53) for computing the positions of their side chain, for calculating the electrostatic surface potential (ESP) and for visualization.

4.10 Statistical test for selection

To test for the significant deviation of S_β at a given position in all loci combined from a null model, we compared the S_β value to the one obtained in the same position (with the same sequence), when the frequencies of each population were scrambled - i.e. the frequency of a given allele was assigned to a different allele over all loci. Significant positions were defined as positions where S_β is higher than 95% of the S_β values in the scrambled model. Note that S_β is defined as the sum over all the absolute values of the coefficients over all populations and all AA at the appropriate position. The null model was computed over 100 Cross Validations (CV).

To test for AA consistently selected at $\beta(AA, Pos)$ values of all the significant positions in the NPB region, we performed a Chi-Square test between the sum of the S_β values for each amino acid and the sum of the S_β values when mixing all these values over 100 random mixings. We computed how often each AA with a $p-value < 0.05$ appears in these cross-validations.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

YL supervised the work and wrote a part of the manuscript RL performed the analysis and wrote the manuscript LL performed part of the analysis. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The work of RL and YL was funded by ISF grant 870/20, a Vatat DSI grant and an Israel MOH grant.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1236080/full#supplementary-material

References

1. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PloS Biol* (2006) 4(3):e72. doi: 10.1371/journal.pbio.0040072

2. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. *PloS Genet* (2006) 2(4):e64. doi: 10.1371/journal.pgen.0020064

3. Hedrick PW, Thomson G. Evidence for balancing selection at HLA. *Genetics* (1983) 104(3):449–56. doi: 10.1093/genetics/104.3.449

4. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M. Complete sequence and gene map of a human major histocom patibility complex. *Nature* (1999) 401 (6756):921–3. doi: 10.1038/44853

5. Hughes AL, Ota T, Nei M. Positive darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol Biol Evol* (1990) 7(6):515–24. doi: 10.1093/oxfordjournals.molbev.a040626

6. Slatkin M. Joint estimation of selection intensity and mutation rate under balancing selection with applications to HLA. *bioRxiv*. (2021). doi: 10.1101/2021.11.18.469194

7. Pierini F, Lenz TL. Divergent allele advantage at human MHC genes: Signatures of past and ongoing selection. *Mol Biol Evol* (2018) 35(9):2145–58. doi: 10.1093/molbev/msy116

8. Klein J, Sato A, Nikolaidis N. MHC, TSP, and the origin of species: From immunogenetics to evolutionary genetics. *Annu Rev Genet* (2007) 41:281–304. doi: 10.1146/annurev.genet.41.110306.130137

9. Radwan J, Babik W, Kaufman J, Lenz TL, Winternitz J. Advances in the evolutionary understanding of MHC polymorphism. *Trends Genet* (2020) 36(4):298–311. doi: 10.1016/j.tig.2020.01.008

10. Lenz TL, Mueller B, Trillmich F, Wolf JB. Divergent allele advantage at MHC-DRB through direct and maternal genotypic effects and its consequences for allele pool composition and mating. *Proc R Soc B: Biol Sci* (2013) 280(1762):20130714. doi: 10.1098/rspb.2013.0714

11. Rudolph MG, Stanfield RL, Wilson IA. How tcrs bind mhcs, peptides, and coreceptors. *Annu Rev Immunol* (2006) 24:419–66. doi: 10.1146/annurev.immunol.23.021704.115658

12. Marrack P, Scott-Browne JP, Dai S, Gapin L, Kappler JW. Evolutionarily conserved amino acids that control tcr-mhc interaction. *Annu Rev Immunol* (2008) 26:171–203. doi: 10.1146/annurev.immunol.26.021607.090421

13. Glazer N, Akerman O, Louzoun Y. Naive and memory T cells TCR-HLA binding prediction. *Oxford Open Immunol* (2022) 3. doi: 10.1093/oxfimm/iqac001

14. Wucherpfennig WK, Call MJ, Deng L, Mariuzza R. Structural alterations in peptide–MHC recognition by self-reactive T cell receptors. *Curr Opinion Immunol*. (2009) 21(6):590–5. doi: 10.1016/j.coi.2009.07.008

15. Andoniou CE, Andrews DM, Degli-Esposti MA. Natural killer cells in viral infection: More than just killers. *Immunol Rev* (2006) 214(1):239–50. doi: 10.1111/j.1600-065X.2006.00465.x

16. Khakoo SI, Carrington M. KIR and disease: A model system or system of models? *Immunol Rev* (2006) 214(1):186–201. doi: 10.1111/j.1600-065X.2006.00459.x

17. Moesta AK, Norman PJ, Yawata M, Yawata N, Gleimer M, Parham P. Synergistic polymorphism at two positions distal to the ligand-binding site makes KIR2DL2 a stronger receptor for HLA-C than kir2dl3. *J Immunol* (2008) 180(6):3969–79. doi: 10.4049/jimmunol.180.6.3969

18. Debebe BJ, Boelen L, Lee JC, Thio CL, Astemborski J, Kirk G, et al. Identifying the immune interactions underlying HLA class I disease associations. *Elife* (2020) 9: e54558. doi: 10.7554/eLife.54558

19. Jones DC, Kosmoliaptsis V, Apps R, Lapaque N, Smith I, Kono A, et al. HLA class I allelic sequence and conformation regulate leukocyte Ig-like receptor binding. *J Immunol* (2011) 186(5):2990–7. doi: 10.4049/jimmunol.1003078

20. Bashirova AA, Martin-Gayo E, Jones DC, Qi Y, Apps R, Gao X, et al. LILRB2 interaction with HLA class I cor435 relates with control of HIV-1 infection. *PloS Genet* (2014) 10(3):e1004196. doi: 10.1371/journal.pgen.1004196

21. Isildak U, Stella A, Fumagalli M. Distinguishing between recent balancing selection and incomplete sweep using deep neural networks. *Mol Ecol Resour* (2021) 21(8):2706–18. doi: 10.1111/1755-0998.13379

22. Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi G, Menozzi G, et al. A population genetics study of the familial Mediterranean fever gene: Evidence of balancing selection under an overdominance regime. *Genes Immun* (2009) 10 (8):678–86. doi: 10.1038/gene.2009.59

23. Cagliani R, Fumagalli M, Riva S, Pozzoli U, Fracassetti M, Bresolin N, et al. Polymorphisms in the CPB2 gene are maintained by balancing selection and result in haplotype-preferential splicing of exon 7. *Mol Biol Evol* (2010) 27(8):1945–54. doi: 10.1093/molbev/msq082

24. Soni V, Vos M, Eyre-Walker A. A new test suggests that balancing selection maintains hundreds of non-synonymous polymorphisms in the human genome. *bioRxiv*. (2021). doi: 10.1101/2021.02.08.430226

25. Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, et al. Targets of balancing selection in the human genome. *Mol Biol Evol* (2009) 26 (12):2755–64. doi: 10.1093/molbev/msp190

26. Siewert KM, Voight BF. Detecting long-term balancing selection using allele frequency correlation. *Mol Biol Evol* (2017) 34(11):2996–3005. doi: 10.1093/molbev/msx209

27. Bitarello BD, De Filippo C, Teixeira JC, Schmidt JM, Kleinert P, Meyer D, et al. Signatures of long-term balancing selection in human genomes. *Genome Biol Evol* (2018) 10(3):939–55. doi: 10.1093/gbe/evy054

28. Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi GP, Menozzi G, et al. The signature of long-standing balancing selection at the human defensin β-1 promoter. *Genome Biol* (2008) 9(9):1–11. doi: 10.1186/gb-2008-9-9-r143

29. Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* (2013) 339(6127):1578–82. doi: 10.1126/science.1234070

30. Teixeira JC, De Filippo C, Weihmann A, Meneu JR, Racimo F, Dannemann M, et al. Long-term balancing selection in lad1 maintains a missense trans-species polymorphism in humans, chimpanzees, and bonobos. *Mol Biol Evol* (2015) 32 (5):1186–96. doi: 10.1093/molbev/msv007

31. DeGiorgio M, Lohmueller KE, Nielsen R. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PloS Genet* (2014) 10(8):e1004561. doi: 10.1371/journal.pgen.1004561

32. Cheng X, DeGiorgio M. Flexible mixture model approaches that accommodate foot print size variability for robust detection of balancing selection. *Mol Biol Evol* (2020) 37(11):3267–91. doi: 10.1093/molbev/msaa134

33. Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, et al. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res* (2009) 19(2):199–212. doi: 10.1101/gr.082768.108

34. Hughes AL, Yeager M. Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet* (1998) 32(1):415–35. doi: 10.1146/annurev.genet.32.1.415

35. Ronen R, Udpa N, Halperin E, Bafna V. Learning natural selection from the site frequency spectrum. *Genetics* (2013) 195(1):181–93. doi: 10.1534/genetics.113.152587

36. Liberman G, Benichou JI, Maman Y, Glanville J, Alter I, Louzoun Y. Estimate of within population incremental selection through branch imbalance in lineage trees. *Nucleic Acids Res* (2016) 44(5):e46–6. doi: 10.1093/nar/gkv1198

37. Alter I, Gragert L, Fingerson S, Maiers M, Louzoun Y. HLA class I haplotype di421 versity is consistent with selection for frequent existing haplotypes. *PloS Comput Biol* (2017) 13(8):e1005693. doi: 10.1371/journal.pcbi.1005693

38. Sawai H, Nishida N, Khor S-S, Honda M, Sugiyama M, Baba N, et al. Genome-wide association study identified new susceptible genetic variants in HLA class I region for hepatitis B virus-related hepato581 cellular carcinoma. *Sci Rep* (2018) 8(1):1–8. doi: 10.1038/s41598-018-26217-7

39. Total number of donors and cord blood units. (2022). Available at: https://statistics.wmda.info/.

40. Slater N, Louzoun Y, Gragert L, Maiers M, Chatterjee A, Albrecht M. Power laws for heavy-tailed distributions: Modeling allele and haplotype diversity for the national marrow donor program. *PloS Comput Biol* (2015) 11(4):e1004204. doi: 10.1371/journal.pcbi.1004204

41. Lobkovsky AE, Levi L, Wolf YI, Maiers M, Gragert L, Alter I, et al. Multiplicative fitness, rapid haplotype discovery, and fitness decay explain evolution of human MHC. *Proc Natl Acad Sci* (2019) 116(28):14098–104. doi: 10.1073/pnas.1714436116

42. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire us donor registry. *Hum Immunol* (2013) 74(10):1313–20. doi: 10.1016/j.humimm.2013.06.025

43. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. *PloS Genet* (2008) 4(12):e1000304. doi: 10.1371/journal.pgen.1000304

44. Fisher RA. *The genetical theory of natural selection*. Oxford UK:Dover Pubns (1958).

45. Wright S. Evolution in Mendelian populations. *Genetics*. (1931) 16(2), p.97

46. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. *Genetics* (2000) 156(1):297–304. doi: 10.1093/genetics/156.1.297

47. Carrington M. Recombination within the human MHC. *Immunol Rev* (1999) 167(1):245–56. doi: 10.1111/j.1600-065X.1999.tb01397.x

48. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol* (2015) 1(1):1-5. doi: 10.1093/ve/vev003

49. Boegel S, Löwer M, Schäfer M, Bukur T, De Graaf J, Boisguérin V, et al. HLA typing from RNA-Seq sequence reads. *Genome Med* (2013) 4(12):1–12. doi: 10.1186/gm403

50. Felsenstein J. *Phylip* (2021). Available at: https://evolution.genetics.washington.edu/phylip.html.

51. Titus-Trachtenberg E, Rickards O, De Stefano G, Erlich H. Analysis of hla class ii haplotypes in the cayapa Indians of Ecuador: A novel drb1 allele reveals evidence for convergent evolution and balancing selection at position 86. *Am J Hum Genet* (1994) 55 (1):160.

52. Creixell P, Schoof EM, Tan CSH, Linding R. Mutational properties of amino acid residues: Implications for evolvability of phosphorylatable residues. *Philos Trans R Soc B: Biol Sci* (2012) 367(1602):2584–93. doi: 10.1098/rstb.2012.0076

53. DeLano WL. *PyMOL* (2022). Available at: https://pymol.org/2/.

54. Li Y, Louzoun Y, Weigert M. Editing anti-DNA B cells by V λx. *J Exp Med* (2004) 199(3):337–46. doi: 10.1084/jem.20031712

55. Chen C, Li H, Tian Q, Beardall M, Xu Y, Casanova N, et al. Selection of anti-double-stranded DNA B cells in autoimmune MRL-lpr/lpr mice. *J Immunol* (2006) 176 (9):5183–90. doi: 10.4049/jimmunol.176.9.5183

56. Hedrick PW. Balancing selection and MHC. *Genetica* (1998) 104(3):207–14. doi: 10.1023/A:1026494212540

57. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history and natural selection in the human major histocompatibility complex loci. *Genetics* (2006) 173(4):2121–42. doi: 10.1534/genetics.105.052837

58. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG. The imgt/hla database. *Nucleic Acids Res* (2012) 41(D1):D1222–7. doi: 10.1093/nar/gks949

59. Israeli S, Gragert L, Maiers M, Louzoun Y. Hla haplotype frequency estimation for heterogeneous populations using a graph-based imputation algorithm. *Hum Immunol* (2021) 82(10):746–57. doi: 10.1016/j.humimm.2021.07.001

60. Train test split. (2022). Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.

61. Neural Network Intelligence. (2021). Available at: https://nni.readthedocs.io/en/stable/.

62. Support vector regression. (2022). Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html.

63. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. Netmhcpan, a method for quantitative predictions of peptide binding to any HLA-A and-B locus protein of known sequence. *PloS One* (2007) 2(8):e796. doi: 10.1371/journal.pone.0000796

64. Gagne K, Busson M, Bignon J-D, Balère-Appert M-L, Loiseau P, Dormoy A, et al. Donor KIR3DL1/3DS1 gene and recipient Bw4 KIR ligand as prognostic markers for outcome in unrelated hematopoietic stem cell transplantation. *Biol Blood Marrow Transplant* (2009) 15(11):1366–75. doi: 10.1016/j.bbmt.2009.06.015

65. Ipd-imgt/hla database. (2022). Available at: http://hla.alleles.org/alleles/index.html.

66. Anjanappa R, Garcia-Alai M, Kopicki J-D, Lockhauserbäumer J, Aboelmagd M, Hinrichs J, et al. Structures of peptide-free and partially loaded MHC class I molecules reveal mechanisms of peptide selection. *Nat Commun* (2020) 11(1):1–11. doi: 10.1038/s41467-020-14862-4

67. Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* (1996) 384(6605):134–41. doi: 10.1038/384134a0

68. Shimizu A, Kawana-Tachikawa A, Yamagata A, Han C, Zhu D, Sato Y, et al. Structure of TCR and antigen complexes at an immunodominant CTL epitope in hiv-1 infection. *Sci Rep* (2013) 3(1):1–9. doi: 10.1038/srep03097

69. Ding Y-H, Smith KJ, Garboczi DN, Utz U, Biddison WE, Wiley DC. Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids. *Immunity* (1998) 8(4):403–11. doi: 10.1016/S1074-7613(00)80546-4

70. Boyington JC, Motyka SA, Schuck P, Brooks AG, Sun PD. Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand. *Nature* (2000) 405(6786):537–43. doi: 10.1038/35014520

71. Fan QR, Long EO, Wiley DC. Crystal structure of the human natural killer cell inhibitory receptor KIR2DL1–HLA-Cw4 complex. *Nat Immunol* (2001) 2(5):452–60. doi: 10.1038/87766

72. Pymm P, Illing PT, Ramarathinam SH, O'Connor GM, Hughes VA, Hitchen C, et al. MHC-I peptides get out of the groove and enable a novel mechanism of HIV-1 escape. *Nat Struct Mol Biol* (2017) 24(4):387–94. doi: 10.1038/nsmb.3381

73. Mohammed F, Stones DH, Zarling AL, Willcox CR, Shabanowitz J, Cummings KL, et al. The antigenic identity of human class I MHC phosphopeptides is critically dependent upon phosphorylation status. *Oncotarget* (2017) 8(33):54160. doi: 10.18632/oncotarget.16952

74. Willcox E, Thomas L, Bjorkman P. Crystal structure of HLA-A2 bound to LIR-1, a host and viral major histocompatibility complex receptor. *Nature immunology* (2003) 4(9):913–919.

75. Touw WG, Baakman C, Black J, Te Beek TA, Krieger E, Joosten RP, et al. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res* (2015) 43(D1):D364–8. doi: 10.1093/nar/gku1028