OXFORD

Subject Section

HAMdetector: A Bayesian regression model that integrates information to detect HLA-associated mutations

Corresponding Author [1,*], Co-Author [2] and Co-Author [2,*]

[1]Department, Institution, City, Post Code, Country and
[2]Department, Institution, City, Post Code, Country.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: A key process in anti-viral adaptive immunity is that the human leukocyte antigen system (HLA) presents viral peptide fragments in MHC I protein-peptide complexes on cell surfaces and in this way alerts CD8$^+$ cytotoxic T-Lymphocytes (CTLs). This pathway exerts a strong selection pressure on viruses, favoring viral mutants that escape recognition by the HLA/CTL system, e.g. by point mutations that decrease binding of viral peptides to MHC I. Naturally, such immune escape mutations often emerge in highly variable viruses, e.g. HIV or HBV, as HLA-associated mutations (HAMs), specific to the host HLA alleles and its MHC I proteins. The reliable identification of HAMs is not only important to understand viral genomes and their evolution, but it also impacts the development of broadly effective anti-viral treatments and vaccines against variable viruses.

By their very nature HAMs are amenable to detection by statistical methods in paired sequence / HLA data. However, HLA alleles are very polymorphic in the human host population which makes the available data relatively sparse and noisy. Under these circumstances, one way to optimize HAM detection is to integrate all relevant information in a coherent model. Bayesian inference offers a principled approach to achieve this.

Results: Here, we present a new regression model for the detection of HAMs. As we choose a Bayesian approach we can include the novel sparsity-inducing priors, and we obtain easily interpretable quantitative information on HAM candidates. The basic model can be extended to include prior information relevant to HAM detection, which we demonstrate by integrating predictions of epitope affinities to MHC I, predictions of epitope peptide processing, and computation of phylogenetic background. This integrative method improves performance in HAM detection considerably over state-of-the-art methods.

Availability: The source code of this software is available at ... under a permissive MIT license.

Contact: name@bio.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

1.1 The HLA system

One way of how the human immune system recognizes viral infections is through the human leukocyte antigen system (Germain, 1994): Proteins, including viral proteins, in cells are often degraded in proteasomes to peptide fragments (Goldberg *et al.*, 2002). A small subset of these peptides is presented on the cell surface by MHC I receptors, encoded by HLA genes (in humans HLA-A, -B, -C). HLA genes are very polymorphic with more than 20,000 known alleles in the human population (Robinson *et al.*, 2014). The resulting gene products differ in their binding properties, which means that cells from different individuals in general present different peptides on the cell surface.

Cytotoxic T-Lymphocytes (CTLs) are selected during maturation to only weakly bind to complexes of MHC I with self-peptides, but CTLs can bind strongly to complexes of MHC I with peptides from viral proteins (Murata *et al.*, 2007). Upon activation, CTLs become cytotoxic and recruit other immune cells so that infected cells can be eliminated (Harty *et al.*, 2000).

1.2 HLA escape

In this way, the HLA system exerts strong selection pressure towards virus variants that escape T cell recognition (Borrow *et al.*, 1997). Such variants could, for example, carry mutations that result in reduced binding of viral peptides to MHC I or T cell receptors, or that alter peptide processing so that peptides are no longer presented (Yewdell and Hill, 2002).

HLA diversity drives viral evolution in individuals where a virus adapts to specific host features, and in populations and regions that in general differ in HLA frequencies (Kawashima *et al.*, 2009). Upon transmission to a new host with different HLA alleles, HLA escape mutations may revert, as they are often associated with a reduction in viral replication capacity (Matthews *et al.*, 2008), but they also may be irreversible (Kawashima *et al.*, 2009).

Whether and how quickly a given escape mutation is selected in a host depends, e.g., on the viral genomic background, the magnitude of the reduction in viral replication, the availability of compensatory mutations that recover fitness, or the strength of selection pressure (Kløverpris *et al.*, 2016).

Immune escape is a driver of viral evolution in populations and patients, and therefore of eminent importance for the understanding of many aspects of highly variable viruses such as HIV or HBV (Alizon *et al.*, 2011; Allen *et al.*, 2005; Rousseau *et al.*, 2008; Lumley *et al.*, 2018). This includes the development of treatments and vaccines that rely on effective immune responses. This makes detection of immune escape mutations critical.

1.3 Identifying HLA escape mutations

There are several experimental methods available to study HLA escape (Czerkinsky *et al.*, 1983; Brunner *et al.*, 1968; Lamoreaux *et al.*, 2006; Altman *et al.*, 1996). However, these methods are relatively slow and costly, especially for screening purposes. A promising approach that makes efficient use of frequently available data is to combine viral genome sequencing, host HLA determination, computational identification by statistical association analysis, and targeted experimental validation (Carlson *et al.*, 2012).

As the selection pressure exerted by cytotoxic T cells depends on successful recognition of viral peptides on the cell surface and thus on binding to the HLA encoded MHC I protein, escape mutations are usually HLA allele specific and can therefore be detected as HLA allele dependent "footprints" in sequence alignments of viral proteins (Moore, 2002), i.e. as HLA associated mutations (HAM) or replacements enriched in viral sequences from hosts with a specific HLA allele.

One way of quantifying this enrichment is Fisher's exact test (Fisher, 1922): For a given replacement R_i at alignment position i and HLA allele H, a 2-by-2 contingency table is constructed containing the absolute counts of the number of sequences in the four possible categories (R_i, H), $(R_i, \neg H)$, $(\neg R_i, H)$ and $(\neg R_i, \neg H)$, where $\neg R_i$ denotes any replacement except R_i, and $\neg H$ denotes any HLA allele except H.

Fisher's exact test is a conventional null hypothesis significance test (NHST) that generates p-values. In this case, the null hypothesis is that HLA allele H and replacement R_i are independent, and the p-value is the probability of observing a deviation from independence that is at least as extreme as in the data at hand under the assumption that the null hypothesis is true.

Fisher's exact test has the advantage of being fast and easy to apply (Budeus *et al.*, 2016), but it also has several disadvantages (Carlson *et al.*, 2008). The most striking one is that viral sequences share a common phylogenetic history, and, therefore, treating sequences as independent and identically distributed samples may under- or overestimate effect sizes. In the context of hypothesis testing, this leads to increased false positive and false negative rates (Osborne and Waters, 2002; Scariano and Davenport, 1987).

Another issue with Fisher's exact test is the genomic proximity of human HLA class I loci (on chromosome 6 (Francke and Pellegrino, 1977)) leading to linkage disequilibrium – inheritance of HLA alleles can be correlated. Therefore, spurious associations can occur if associations of replacements

with individual HLA alleles are tested: if HLA allele H_1 is associated with an amino acid replacement R because of immune escape, but H_1 is in linkage disequilibrium with allele H_2, this leads to an association of R and H_2, even without being an escape mutation from H_2.

Carlson *et al.* (2008) developed the Phylogenetic Dependency Network, a method that accounts for several of the aforementioned problems, in particular phylogenetic bias and HLA linkage disequilibrium. However, it is based on null hypothesis significance testing.

1.4 Issues with p-values for screening

There are fundamental statistical issues with p-values as a screening tool (Amrhein and Greenland, 2017): with small effect sizes and high variance between measurements, as is often the case with biological data, statistically significant results can be misleading, have the wrong direction (type S error), or greatly overestimate an effect (type M error) (Gelman and Carlin, 2014). Such problems are more and more appreciated in the context of the current "replication crisis", which describes that scientific claims with seemingly strong statistical evidence fail to replicate (Ioannidis, 2005; Begley and Ellis, 2012; Baker, 2016).

These problems are exacerbated if the p-values are used for screening purposes (multiple testing problem). The probability of obtaining a statistically significant result increases with each additional test, even in absence of any real effect. When using p-values as a filter, it is therefore likely to obtain significant effects that are in fact not real. To circumvent this problem, a common strategy is to control the false discovery rate (Benjamini and Hochberg, 1995). These adjustment procedures have the problem that, when performing many of such comparisons, none but the very largest effects remain.

Instead of performing many hypothesis tests and trying to adjust for them, we prefer to fit a single, multilevel model that contains all comparisons of interest. When using multilevel models, the problem of multiple comparisons can disappear entirely and yield more valid estimates (Gelman *et al.*, 2012).

2 Materials and methods

Our general approach for HAMdetector is to fit Bayesian regression models that captures relationships between host HLA alleles and replacements in viral proteomes. This Bayesian approach is advantageous because it allows use of: (1) prior information (e.g. knowledge of effect magnitudes), (2) relevant additional information (phylogeny, epitope information), (3) a problem-specific structure, (4) partial pooling (Gelman, 2010).

2.1 Model backbone

We chose a logistic regression model as backbone because it is easily extensible, and because coefficients can be interpreted in the familiar way as summands on the log-odds scale.

$$y_{ik} \sim \text{Bernoulli}(\theta_{ik}) \tag{1}$$

$$\theta_{ik} = \text{logistic}\left(\beta_{0_k} + \sum_{j=1}^{D} X_{ij}\beta_{jk}\right), \tag{2}$$

where y_{ik} is the binary encoded observation of replacement k in viral sequence i (each observed amino acid state k contributes a separate column to y_{ik}); θ_{ik} is the estimated probability that we observe replacement k in sequence i; X_{ij} is 1 if sequence i comes from host individual HLA allele j and 0 otherwise; β_{jk} is the HLA regression coefficient of HLA allele j for replacement k; D is the number of HLA alleles in the dataset; the logistic inverse link function transforms the linear model in parentheses to the probability scale of θ_{ik}.

The main parameters of interest for HAMdetector are the regression coefficients β_{jk}, as they quantify the strength of association between the occurrence of replacement k and each of the observed HLA alleles. The β_{jk} are on the log-odds scale, i.e. if we go from viral sequences from hosts without HLA allele j to those from hosts with j, the log-odds $\log(p_k/(1 - p_k))$ of observing replacement k increase by addition of β_{jk}.

2.2 Inclusion of additional information

On top of the paired data of viral sequences and host HLA alleles modeled by the backbone (Eq 1), we extend the model to include further information of relevance to improve HAM detection, namely phylogenetic information and predictions of epitope peptide processing and MHC I affinity, as described in the following.

2.2.1 Phylogeny

One problem that occurs when analyzing sequence data is that species (here: viral strains) share a common phylogenetic history. Thus replacements are not independently and identically distributed, and therefore violate a common assumption of standard statistical methods. In fact, Bhattacharya *et al.* (2007) demonstrated the importance of correcting for the phylogenetic structure in identifying HLA associations.

The most popular approach in phylogeny-aware regression of binary variables is to estimate an additional multivariate normally distributed intercept, where the covariance matrix is based on the branch lengths of a given phylogenetic tree (Ives and Garland, 2009, 2014). This approach turned out to be too computationally expensive in our model, hence we chose a strategy similar to the one in Carlson *et al.* (2008):

Consider a phylogenetic tree Ψ obtained from standard maximum likelihood methods for a given multiple sequence alignment. We are interested in estimating $P(y_{ik} = 1|\Psi)$, that is, the probability of observing the replacement k in sequence i based on the underlying phylogenetic model. A quantity that can be readily computed using phylogenetic software like RAxML-NG (Kozlov *et al.*, 2019) is $P(\Psi|y_{ik} = 1)$. For this, we keep the tree topology fixed, annotate the tree with the binary observations y_{ik} at its leaves and optimize the branch lengths. $P(\Psi|y_{ik} = 1)$ is then the likelihood of the annotated phylogenetic tree. Similarly, we can also compute $P(\Psi|y_{ik} = 0)$ by flipping the annotation of sequence i from 1 to 0 (keeping all other observations). With $P(\Psi|y_{ik} = 1)$ and $P(\Psi|y_{ik} = 0)$ known and the relative frequencies of 0 an 1 as priors, we can estimate $P(y_{ik} = 1|\Psi)$ by applying Bayes' theorem. The estimated probabilities based on phylogeny are then included in the model as additional intercepts (second term of logistic argument):

$$
\begin{aligned}
y_{ik} &\sim \text{Bernoulli}(\theta_{ik}) \\
\theta_{ik} &= \text{logistic}\Big(\beta_{0_k} + \gamma\text{logit}\left(P(y_{ik} = 1|\Psi)\right) \\
&\quad + \sum_{k=1}^{D} X_{ik}\beta_{jk}\Big)
\end{aligned}
\tag{3}
$$

The logit transform is used because it cancels out with the logistic inverse link function. The phylogeny term acts as a baseline in absence of any HLA effects. As this baseline itself is not certain but subject to errors of the phylogenetic probabilities $P(y_{ik} = 1|\Psi)$, we introduce an additional parameter γ.

2.2.2 Inclusion of CTL epitope predictions

As outlined earlier, escape mutations often appear as HAMs. Given the underlying mechanism, it is not surprising that escape mutations are enriched in CTL epitopes, i.e. in those viral peptides presented by MHC I to T cell receptors (Bronke *et al.*, 2013). This suggests that knowledge of epitope regions can be used to boost HAM detection. Fortunately, availability of large experimental data sets (Vita *et al.*, 2019) has enabled the development of

computational tools that predict with good accuracy the binding of peptides to MHC I molecules encoded by various HLA alleles (Mei *et al.*, 2020).

Not only mutations in CTL epitopes can lead to failure of MHC I to present epitopes, but also mutations at epitope-flanking positions that interfere with pre-processing of peptides, notably proteasomal cleavage of viral proteins (Milicic *et al.*, 2005; Gall *et al.*, 2007).

In HAMdetector we use MHCFlurry 2.0 (O'Donnell *et al.*, 2020) to predict epitopes that are properly processed and presented by MHC I. For this, we create an input matrix of dimensions $R \times D$, where R is the number of evaluated replacements and D is the number of observed HLA alleles in the dataset. The elements of this matrix are binary encoded and contain a 1 if that position is predicted to be in an epitope, and 0 otherwise. Given an amino acid sequence, MHCFlurry provides a list of possible epitopes (9-13 mers) and HLA allele pairs and calculates a rank based on comparisons with random epitope and HLA pairs. For the binarization we use the rank threshold of 0.2% suggested by MHCFlurry.

We use epitope prediction as information about the expected degree of sparsity, i.e. if we know that a certain position is restricted by a certain HLA allele, we expect that this HLA allele is more likely to be associated with replacements at that position than the other HLA alleles. This idea is implemented by increasing the scale of the local shrinkage parameters λ_{jk} depending on epitope information:

$$
\begin{aligned}
\lambda_{jk} &\sim \text{Cauchy}^+(0, \sigma_j\exp(Z_{jk}\beta_{\text{epi}})) \\
\beta_{\text{epi}} &\sim \text{Normal}^+(1, 2),
\end{aligned}
\tag{4}
$$

where Z_{jk} is 1 if HLA allele j is predicted to restrict the alignment position corresponding to replacement k, and 0 otherwise. The parameter β_{epi} governs the increase in scale of the corresponding local shrinkage parameters. The larger the estimated values of β_{epi} are, the more likely it is to see non-zero regression coefficients for these HLA alleles.

2.2.3 Sparsity-inducing priors

Sparsity-promoting priors (Piironen and Vehtari, 2017b) can drastically improve predictive performance, because the model is better able to differentiate between signal and noise. These priors convey the a priori expectation that most coefficients in a regression model are close to 0, i.e. that non-zero coefficients are sparse. This assumption is likely correct for HLA-associated mutations, because the number of epitopes that are restricted by a given HLA allele is typically small compared to the number of all possible epitopes. As presentation on MHC I is a necessary condition for the HLA system to function, most associations between an HLA allele and a replacement are likely truly zero, with the confounding factors HLA linkage disequilibrium and codon covariation, e.g. in the form of compensatory mutations that lie outside an epitope (Ruhl *et al.*, 2011).

There is a range of sparsity-promoting priors with slightly different properties. They share the common structure of placing most probability mass very close to 0, with heavy tails to accommodate the non-zero coefficients. For our model, we use the so-called regularized horseshoe prior (Piironen and Vehtari, 2017b), which is an improvement of the original horseshoe prior presented by Carvalho *et al.* (2010), in that it additionally allows some shrinkage for the non-zero coefficients. The original horseshoe prior is given by:

$$
\begin{aligned}
\beta_{jk} &\sim \text{Normal}(0, \tau^2\lambda_{jk}^2) \\
\lambda_{jk} &\sim \text{Cauchy}^+(0, 1) \\
\tau &\sim \text{Cauchy}^+(0, \tau_0),
\end{aligned}
\tag{5}
$$

where β_{jk} are the regression coefficients; τ and λ_{jk} are the so-called global and local shrinkage parameters, respectively;, Cauchy$^+$ is the positively constrained Cauchy distribution; τ_0 is the overall degree of sparsity. Shrinkage of the non-zero coefficients in the regularized horseshoe prior is achieved by replacing λ_{jk}^2 with $\tilde{\lambda}_{jk}^2 = \frac{c_k^2\lambda_{jk}^2}{c_k^2 + \tau_k^2\lambda_{jk}^2}$, where the additional parameter c governs the magnitude of shrinkage for the non-zero coefficients.

With Eq 5 the global shrinkage parameter τ is typically very small and shrinks most of the regression coefficients close to 0, whereas the local shrinkage parameters λ_{jk} can occasionally be very large to allow some coefficients to escape that shrinkage.

The overall degree of sparsity τ_0 can be chosen based on the expected number of non-zero coefficients (Piironen and Vehtari, 2017a).

2.3 Full model specification

The full specification of the HAMdetector model is:

$$
\begin{aligned}
y_{ik} &\sim \text{Bernoulli}(\theta_{ik}) \\
\theta_{ik} &= \text{logistic}(\beta_{0_k} + \gamma_k \text{logit}(P(y_{ik} = 1|\Psi)) \\
&\quad + \sum_{k=1}^{D} X_{ij}\beta_{jk}) \\
\beta_{0_k} &\sim \text{Normal}(0, 100^2) \quad\quad\quad\quad (*) \\
\gamma_k &\sim \text{Normal}(\mu_{\text{phy}}, \sigma_{\text{phy}}^2) \\
\mu_{\text{phy}} &\sim \text{Normal}(1, 1) \quad\quad\quad\quad\quad (*) \\
\sigma_{\text{phy}} &\sim \text{Normal}^+(0, 0.5) \quad\quad\quad\quad (*) \\
\beta_{\text{epi}} &\sim \text{Normal}^+(1, 2) \quad\quad\quad\quad\ (*) \quad\quad (6) \\
\beta_{jk} &\sim \text{Normal}(0, \tau_k^2 \tilde{\lambda}_{jk}^2) \\
\tilde{\lambda}_{jk}^2 &= \frac{c_k^2 \lambda_{jk}^2}{c_k^2 + \tau_k^2 \lambda_{jk}^2} \\
c_k^2 &\sim \text{Inv-Gamma}(3.5, 3.5) \quad\quad\quad (*) \\
\lambda_{jk} &\sim \text{Cauchy}^+(0, \sigma_j \exp(Z_{jk}\beta_{\text{epi}})) \\
\tau_k &\sim \text{Cauchy}^+(0, \tau_{0k}) \quad\quad\quad\quad (*) \\
\tau_{0k} &= \frac{10}{D - 10} \frac{2}{\sqrt{N}}
\end{aligned}
$$

where N is the number of available annotated sequences.

The full model specification includes some aspects that were not covered in the previous sections. In particular, the overall phylogeny-weight γ in Eq 1 is replaced in the full model by hierarchically modelled γ_k, which allows partial pooling across replacements (even with a global parameter γ the model works reasonably well). All other additions ($\tilde{\lambda}_{jk}^2$ and τ_{0k}) are explained in detail in Piironen and Vehtari (2017b). Briefly, the addition of $\tilde{\lambda}_{jk}^2$ allows some regularization for the non-zero coefficients and the parameterization of τ_{0k} allows to place a prior on the expected number of non-zero coefficients. This is particularly useful for logistic regression models, as some shrinkage helps to deal with issues of separability and collinearity that commonly occur with logistic regression models.

2.3.1 Prior justification

Prior distributions are labeled with an asterisk in Eq 6. They are weakly informative, which means that they effectively limit posteriors to realistic magnitudes of parameters. One exception to this are the intercepts β_{0_k}, which are essentially flat because they are well identified by the data alone.

The hierarchical mean and standard deviation of the phylogeny coefficients γ_k place most probability mass on γ_k values around 1. In absence of any HLA effects, a $\gamma_k = 1$ would mean that the estimate for the probability of observing replacement k is identical to the probability based on the phylogenetic model. This treats phylogeny as a baseline, and any observations not attributed to phylogeny must be explained by HLA alleles or noise.

The prior on c_k^2 implies a Student-t prior with 7 degrees of freedom and a scale of 1 on the non-zero HLA regression coefficients β_{jk}. A Student-t prior with these parameters is a reasonable default choice for logistic regression models (Piironen and Vehtari, 2017b).

The value of τ_{0k} implies 10 effective non-zero HLA regression coefficients per replacement. The rationale behind this parameterization is again outlined in Piironen and Vehtari (2017b). The value of 10 corresponds to a generously estimated magnitude based on available HIV epitope maps (Yusim *et al.*, 2018). The model is also parameterized in a way that assumes an equal degree of sparsity across all alignment positions a priori. We also tried to model τ_k hierarchically, but observed sampling issues due to the resulting unfavorable geometry of the posterior.

2.4 Model implementation

All models were implemented in Stan 2.23 (Stan Development Team, 2021), a probabilistic programming language and Hamiltonian Monte Carlo sampler for efficient numerical computation of posterior distributions. The Stan code is available at ... in two versions: One optimized for readability and one optimized for speed by utilizing Stan's multithreading and GPU capabilities. A Julia (Bezanson *et al.*, 2017) package is available at ... to run the model on custom data. Due to restrictions of dependencies (MHCFlurry and RAxML-ng), HAMdetector is only available on Linux.

2.5 Model diagnostics

2.5.1 Convergence diagnostics

We use the split-$\hat{R}$ convergence diagnostic to identify Markov chain convergence issues (Gelman and Rubin, 1992; Gelman *et al.*, 2013). We require a value of $\hat{R}$ below 1.01 for all model parameters. Additionally, we require that the effective sample size N_{eff} (Stan Development Team, 2021) is above 500 for all model parameters and that sampling occurs without any divergent transitions (Betancourt, 2017).

2.5.2 Posterior predictive checks

In posterior predictive checks, we simulate new data from the inferred posterior distribution and the likelihood, and we compare these simulated data with representative real data (Gabry *et al.*, 2019). A good model should predict data that are consistent with real data. This general idea was employed in two ways to test our models.

For a first posterior predictive check we used *calibration plots*: two binned quantities were plotted against each other, the observed relative frequencies of replacements $f(y_{ik} = 1)$, and the predicted probabilities $P(y_{ik} = 1|\text{model})$. In such a plot, a well-calibrated model should yield points following the diagonal. Technically, all observations were first sorted by increasing estimated probability $P(y_{ik} = 1|\text{model})$ and grouped into n bins. For each bin, the fraction of observations with $y_{ik} = 1$ (observed event percentage) was then plotted against the midpoint of each bin. The cutpoints of the bins are indicated by error bars.

Second, we assessed the abilities of different models and methods to discover HAMs with *HAM enrichment plots*. These plots are based on the observation that CTL escape mutations are enriched in epitopes (Bronke *et al.*, 2013). Hence, the degree by which methods for HAM prediction recover this trend is a measure of model performance. To implement this measure, we first ranked all evaluated replacements according to their respective credibility of being a HAM, computed as integral of the marginal posterior $P(\beta_{jk} > 0)$. For comparison with established methods, namely Fisher's exact test and Phylogenetic Dependency Network (Carlson *et al.*, 2008), ranked lists based on p- and q-values, respectively, were computed. Then we calculated for each rank r the accumulated number $N_e(r)$ of predictions of this rank or better ranks were located inside known epitopes. The higher the curve $N_e(r)$, the higher the enrichment of predicted HAMs in epitopes, see e.g. Fig. 2.

2.5.3 Leave-one-out cross-validation

Another performance measure is the ability to generalize to unseen data. To examine this ability for the different model variants we performed leave-one-out cross-validation (LOOCV), using the efficient Pareto-smoothed LOOCV Vehtari *et al.* (2016).

From the LOOCV we obtain the Expected Log-Predictive Density (ELPD) $\sum_{i=1}^{n} \log(\int p(y_i|\theta)p(\theta|y_{i-1})d\theta))$ for samples $i = 1, \ldots, n$, ith observation y_i, data y_{i-1} with i left out, and model parameters θ. Thus, the ELPD is the average log predictive density of the observed data points based on the leave-one-out posterior distributions. This measure has the advantage over other performance measures like classification accuracy, that it not only takes the location of the predictive distribution (the number of correct predictions) into account, but also the width, i.e. how confident the model is in its predictions.

2.6 Data

The model was fit with several data sets:

- A large HIV dataset consisting of a subset of sequences from the HOMER (Brumme *et al.*, 2007, 2008) cohort, the Western Australian HIV Cohort Study (WAHC, Moore (2002); Bhattacharya *et al.* (2007)) and the US AIDS Clinical Trials Group (ACTG) protocols 5142 participants (John *et al.*, 2008) who also provided Human DNA under ACTG protocol 5128 (Haas *et al.*, 2003) (total $N = 1383$). These data were kindly provided by Zabrina Brumme and collaborators and were in part also used in the Phylogenetic Dependency Network study (Carlson *et al.*, 2008). The dataset contains sequences spanning the *gag*, *pol*, *env*, *nef*, *vif*, *vpr*, *vpu*, *tat* and *rev* genes.
- A set of 351 HIV sequences mostly spanning the *pol* gene from the Arevir database (Roomp *et al.*, 2006).
- A set of 544 Hepatitis-B-Virus sequences that were kindly provided by Jörg Timm. The dataset contains sequences of the HBe, LHBs, Pol and HBx proteins.
- A set of 104 Hepatitis-D-Virus sequences containing the HDV-antigen that were kindly provided by Michael Roggendorf (Karimzadeh *et al.*, 2018).
- A set of 41 HIV sequences spanning the *gag* and *pol* genes that were kindly provided by Rongge Yang.

Lists of known epitopes were gathered from the Immune Epitope Database (IEDB, Vita *et al.* (2019)) for HIV sequences, and the Hepitopes database **?** for Hepatitis-B-Virus sequences.

2.7 Data preparation

For all sequences, we applied the following preparation steps:

1. For each dataset, the sequences were split into subsequences, either by protein or gene.
2. If not already present in this format, sequences were translated into their amino acid representation.
3. RAxML-NG (Kozlov *et al.*, 2019) version 1.0.0 was used to generate a maximum likelihood phylogenetic tree for each gene/protein using the `-model GTR+G+I` option with all other parameters set to default values. If available, we used RNA or DNA sequences for this step, rather than protein sequences.

3 Results

In order to understand what the different building blocks of HAMdetector contribute, we applied four different Bayesian models of increasing complexity to each dataset, starting with the standard logistic regression model (Eq 1), and adding then the further components, i.e. the horseshoe prior (Eq 5), phylogeny (Eq 3), and epitope prediction, resulting in the full model (Eq 6). For comparisons to existing methods, we also applied Fisher's exact test and the Phylogenetic Dependency Network Carlson *et al.* (2008) to the same data.
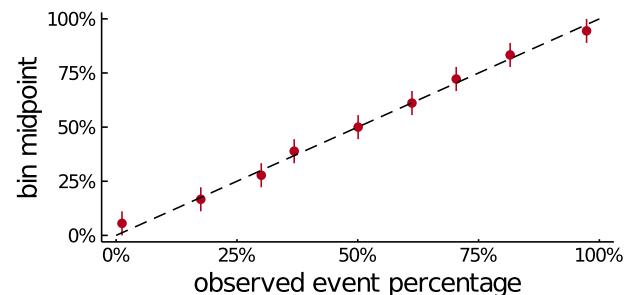

Fig. 1. Calibration plot for the HBV core protein.

3.1 Run times and convergence

For a standard officie computer, run times of HAMdetector on the smaller HDV dataset were of the order of minutes and on the order of hours for the Hepatitis B dataset. For the large HIV dataset, the models were run overnight. Run times scale approximately linearly with the product NK, where N is the number of sequences and K is the number of replacements. All model fits showed no signs of inference issues. In total, samples were drawn from four Hamiltonian Markov chains with 1000 iterations each after 300 warm-up iterations. The effective sample size exceeded 500 for all model parameters, $\hat{R}$ convergence diagnostic values were below 1.01 in all cases.

3.2 Posterior predictive checks

The model yields well-calibrated posterior predictive probabilities of replacements. This is exemplified in Figure 1 for HBV core protein, but also holds true for the other datasets (Supplementary Figures ...-...).

The predictions of all tested models are strongly enriched in epitopes over baselines, irrespective of dataset (Fig. 2 for HBe protein and Supplementary Figures ...-... for other datasets). For the best ranked HAMs, Fisher's exact test performs about as well as the HAMdetector backbone logistic regression model (model 1 in Fig. 2). Each of the following three model stages of HAMdetector increases HAM enrichment further, independently of the dataset. The horseshoe prior alone (model 2) is a drastic improvement over model 1, even though it does not include any external information. The logistic regression model with horseshoe prior works roughly as well as the Phylogenetic Dependency Network Carlson *et al.* (2008), which includes much more information. Note that the line for the Phylogenetic Dependency Network is shorter because it only outputs associations with a q-value lower than 0.2 by default. Model 3 with its additional inclusion of phylogeny has higher enrichment than model 2, and finally, the full model 4 with the inclusion of epitope prediction leads to a further improvement. Note that model 4 only uses epitope *prediction* software and does not use any information of experimentally confirmed epitopes. These are here only used for model evaluation.

The Bayesian approach lends itself to addition of prior knowledge, and it is not too surprising that exploiting this feature to incorporate information from phylogeny or epitope prediction improves HAM detection. It may be more surprising that the sparsifying horseshoe prior has such an impact. However, this is in principle the same mechanism as for the other information components: it is known that HAMs are sparse per HLA allele, and therefore supplying this information to the inference improves predictions. Figure 3 illustrates the effect of the sparsifying prior with a concrete example, the replacement 11D in HIV integrase (Arevir dataset). There is no evidence for an association of HLA-A*01 with this replacement, whereas for HLA-B*44 the data is consistent with a strong association. The horseshoe prior has the effect of shrinking towards 0 specifically those regression coefficients with weak evidence of an association (A*01 in Fig. 3). This reduces the standard error for the remaining coefficients, leading in our example to narrowed histogram for the association with B*44 in the model with horseshoe prior.
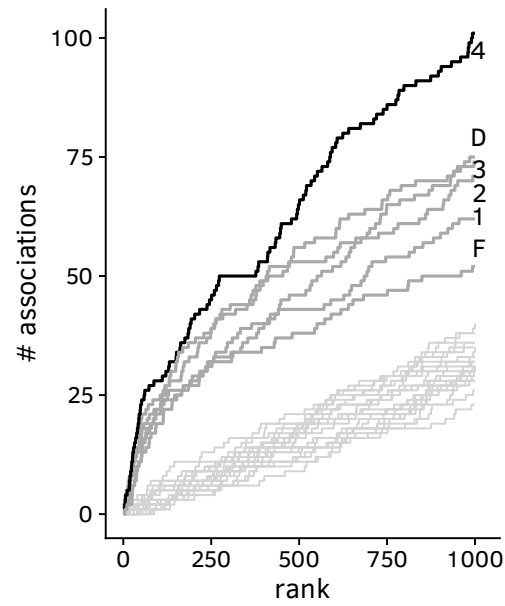
Fig. 2. HAM enrichment plot for HBe protein: number N_e of associations inside the boundary of known epitopes vs. rank r. D: Phylogenetic Dependency Network; F: Fisher's exact test; 1: simple logistic regression model with broad Student-t priors; 2: logistic regression model with horseshoe prior; 3: logistic regression model with horseshoe prior and phylogeny; 4: full model with epitope prediction. Unannotated gray lines at the bottom of the graph are HAM enrichment curves for random permutations of the list of HLA allele - replacement pairs and act as baselines.
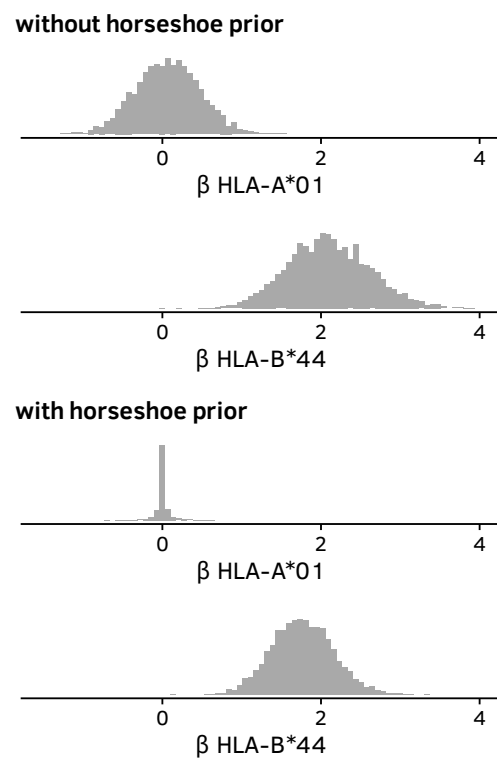

Fig. 3. Marginal posterior distributions of regression coefficients for the association of replacement 11D of the HIV integrase with HLA alleles A*01 and B*44. Top half: inferred with logistic regression model, bottom half: inferred with logistic regression with sparsifying horseshoe prior.

Table 1. ELPD changes as HAMdetector components are added. Data computed for HBe protein of the HBV dataset. All differences in ELPD are larger than several times the estimated standard error (column se$_{diff}$), indicating that models that include more information have better predictive performance.

	ELPD$_{diff}$	se$_{diff}$
logistic regression (baseline)	0.0	0.0
+ horseshoe prior	949.8	65.2
+ phylogeny	4440.9	94.4
+ epitope prediction	63.1	18.9

3.3 Leave-one-out cross-validation

To quantify the ability of the four different model stages of HAMdetector to generalize to unseen cases, we computed the ELPD with Pareto-smoothed leave-one-out cross-validation. Table 1 shows results for the HBe protein of the HBV dataset in terms of ELPD changes with each new model stage. Each new model stage adds ELPD, i.e. is better at generalizing than the simpler model stages.

The model with horseshoe prior alone already has a much higher ELPD than the standard logistic regression model, even though it does not use any specific external data. This is because including the sparsity assumption allows the model to better separate signal from noise and the uncertainty of the close-to-zero coefficients does not propagate into uncertainty of predictions.

Including phylogeny further improves model performance a lot, as the assumption of independent and identically distributed data is replaced with specific information from the shared phylogenetic history.

While addition of sparsity and phylogeny has an effect on all replacements and samples, epitope prediction only influences those replacements that are restricted by a given HLA allele and only those samples that are annotated with the allele. Therefore, inclusion of epitope prediction does not improve ELPD as much as inclusion of phylogeny and the sparsity assumption. However, inclusion of epitope prediction is highly useful for determining which HLA alleles are associated with a replacement, as shown in the previous section.

3.4 HAMs in HDV as test case

The Hepatitis D Virus (HDV) dataset (Karimzadeh *et al.*, 2019) is an excellent test case: we have (1) a set of paired HDV sequences and patient HLA alleles, (2) HAM predictions by Fisher's exact test as implemented in SeqFeatR (Budeus *et al.*, 2016), and (3) experimental tests of the predicted HAMs for immune escape activity (IFN-γ production assays, Karimzadeh *et al.* (2019)). This allows us to see whether HAMdetector decreases the false positive rate in comparison to the simpler Fisher's exact test, and we can make *bona fide* predictions on previously undetected HAMs. We have 15 HAMs predicted in HDV by Fisher's exact test at significance level 5×10^{-3} (Table 2) as published (Karimzadeh *et al.*, 2019). The corresponding p-values have no clear relation to experimental confirmation, i.e. p-values for confirmed HAMs are not generally lower than those of non-confirmed ones.

For HAMdetector, we use in Table 2 the posterior probability of a positive regression coefficient ($P(\beta_{jk} > 0)$) as measure for the confidence in having detected a HAM. HAMs with strong support have a posterior probability close to 1, associations with no support a probability close to 0.5 (corresponding to a regression coefficient centered around 0). The five predicted HAMs with top posterior probabilities (all ≥ 0.90) have all been experimentally confirmed. There is only one outlier with posterior probability 0.75 (P89T and B*37).

HAMdetector shows strong support for 15 replacement - allele pairs that have previously not been identified. All of them exceed the significance level of 5×10^{-3} used in Karimzadeh *et al.* (2019). This shows that HAMdetector

Table 2. List of HAMs predicted Fisher's exact test (FET). In the last column a "+" marks experimental confirmation and a "-" a failed confirmation. Column "post.prob." shows the corresponding posterior probabilities for positive associations computed with HAMdetector. Bona fide predictions are separated by a horizontal line. These are replacement - allele pairs we expect to to be verified in future experimental tests. The replacement annotated with a (+) lies within a known A*02 epitope.

replacement	allele	p-value (FET)	post. prob.	confirmed
S170N	B*15	$3 \cdot 10^{-8}$	0.99	+
D101E	B*37	0.0002	0.96	+
R105K	B*27	0.0011	0.93	+
R139K	B*41	0.0034	0.92	+
E47D	B*18	0.0027	0.90	+
D33E	B*13	0.0001	0.86	-
T134A	A*68	0.0045	0.82	-
K43R	B*13	0.0021	0.77	-
P89T	B*37	0.0011	0.75	+
D47E	A*30	0.0010	0.76	-
K113R	B*13	0.0043	0.76	-
A107T	B*14	0.0028	0.70	-
P49L	A*30	0.0031	0.63	-
Q100L	B*13	0.0018	0.60	-
D96E	B*13	0.0035	0.51	-
E46D	A*02	0.0054	0.97	?
V81I	A*68	0.0073	0.97	?
K113N	B*08	0.0063	0.96	?
A71T	B*41	0.0065	0.96	?
L188I	A*68	0.0632	0.94	?
T95S	A*01	0.0285	0.93	?
D33E	A*03	0.0226	0.93	?
P49L	B*13	0.0035	0.92	?
A74S	A*68	0.0123	0.91	?
E29D	B*44	0.0559	0.91	?
D46E	B*57	0.0190	0.91	?
R88K	A*68	0.0123	0.91	?
T149P	B*52	0.0281	0.91	?
K43R	A*02	0.2158	0.90	(+)
N22S	B*08	0.0405	0.90	?

is able to identify possible HAMs that would not be detected otherwise. The most striking example for this is K43R - A*02, which using Fisher's exact test has a p-value of 0.22. While this replacement has not been experimentally verified in a IFN-γ production assay yet, it lies within a known A*02 epitope and thus it is highly likely that K43R conveys some form of immune escape.

3.5 Linkage disequilibrium

Table 2 shows that for 3 of the false positives identified with Fisher's exact test, HAMdetector identifies associations with the same replacement but a different allele (P49L - B*13 instead of P49L - A*30; K43R - A*02 instead of K43R - B*13; and D33E - B*13 instead of D33E - A*03). One possible explanation for this observation is HLA-Linkage disequilibrium and accordingly correlation between HLA alleles in the dataset due to random sampling variation. Indeed, out of the 12 times P49L is observed in sequences annotated with A*30, B*13 is also present in 5 of those cases (Spearman's rank correlation coefficient $\rho = 0.5$). A similar observation can be done for K43R and D33E, albeit the correlation between the respective alleles is much weaker.

Figure 4 shows regression coefficients of the HLA alleles A*30 and B*13 for replacement P49L. When using a simple logistic regression model without any additions (model 1), both A*30 and B*13 show medium evidence of being associated with replacement P49L. However, when also taking phylogeny into account and using a sparsity-promoting
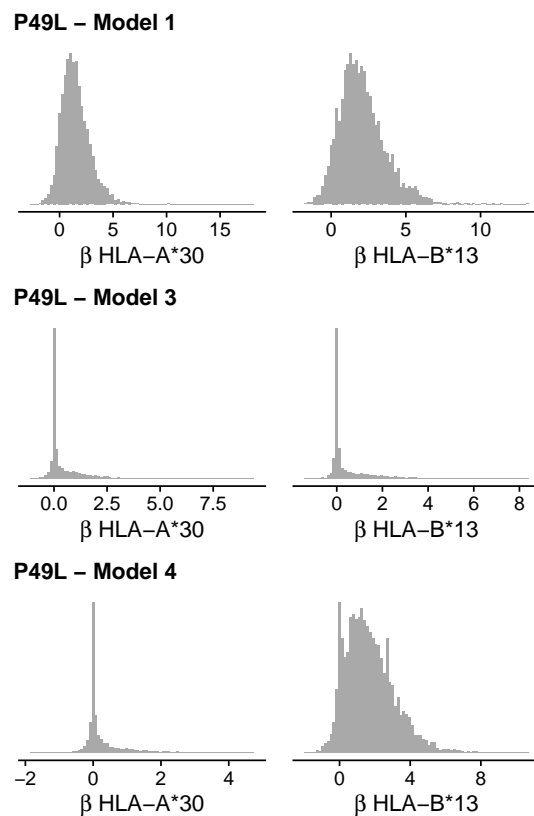
P49L – Model 1

β HLA–A*30   β HLA–B*13

P49L – Model 3

β HLA–A*30   β HLA–B*13

P49L – Model 4

β HLA–A*30   β HLA–B*13

Fig. 4. Marginal posteriors for the regression coefficients of A*30 and B*13 for replacement P49L. The association between P49L and A*30 identified by Fisher's exact test turned out to be a false positive.

prior (model 3), both regression coefficients get shrunk to 0 because the association is not strongly enough identified by the data alone. Using epitope prediction as an additional source of information (model 4) allows to disentangle the association of the correlated alleles with P49L and identify B*13 as associated with P49L. The false-positive association between P49L and A*30 remains being shrunk towards 0.

4 Discussion

HAMdetector follows a general paradigm of Bayesian modelling, namely to map all information that is available about a system of interest onto a probabilistic model, and then to apply Bayesian inference to learn about probable parameter values of that model, e.g. about β_{jk}, the association of HLA j with replacement k. The more relevant information we infuse into the model, the sharper the inference. As HAMdetector includes an unprecedented amount of relevant information, it is only natural that it outperforms other methods.

We have demonstrated that the logistic regression backbone is a platform that can be extended by model components that contribute new information. We have selected such modules guided by widely accepted knowledge, such as phylogeny or epitope location. However, even knowledge that is rarely stated explicitly may be helpful in inference, as in the case of sparsity of HLA associations. Since the included knowledge is generic for interactions of variable viruses with CTL immunity, HAMdetector performance does not depend on the virus.

Yet, HAMdetector is far from perfect. For instance the outlier in Table 2 could point to missing information in HAMdetector. Another deficiency is that it currently works only with two-digit HLA alleles. We are currently

exploring models for 4-digit HLA alleles that exploit partial pooling so that we can attenuate effects of the increased data fragmentation.

Another extension of our model would be to better account for phylogenetic uncertainty by using a Bayesian method to estimate a posterior distribution over possible tree topologies. The uncertainty over the tree topologies and the underlying parameters of the phylogenetic model would then propagate into uncertainty of the estimated probabilities $P(y_{ik} = 1|\Psi)$.

However, the good performance of the current version of HAMdetector makes it already a valuable tool for the study of interactions between viruses and T cell immunity.

Funding

References

Alizon, S., Luciani, F., and Regoes, R. R. (2011). Epidemiological and clinical consequences of within-host evolution. *Trends in Microbiology*, **19**(1), 24–32.

Allen, T. M., Altfeld, M., Geer, S. C., Kalife, E. T., Moore, C., O'Sullivan, K. M., DeSouza, I., Feeney, M. E., Eldridge, R. L., Maier, E. L., Kaufmann, D. E., Lahaie, M. P., Reyor, L., Tanzi, G., Johnston, M. N., Brander, C., Draenert, R., Rockstroh, J. K., Jessen, H., Rosenberg, E. S., Mallal, S. A., and Walker, B. D. (2005). Selective escape from cd8+ t-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. *Journal of Virology*, **79**(21), 13239–13249.

Altman, J. D., Moss, P. A. H., Goulder, P. J. R., Barouch, D. H., McHeyzer-Williams, M. G., Bell, J. I., McMichael, A. J., and Davis, M. M. (1996). Phenotypic analysis of antigen-specific t lymphocytes. *Science*, **274**(5284), 94–96.

Amrhein, V. and Greenland, S. (2017). Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, **2**(1), 4–4.

Baker, M. (2016). 1500 scientists lift the lid on reproducibility. *Nature*, **533**, 452–454.

Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, **483**(7391), 531–3.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**(1), 289–300.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, **59**(1), 65–98.

Bhattacharya, T., Daniels, M., Heckerman, D., Foley, B., Frahm, N., Kadie, C., Carlson, J., Yusim, K., McMahon, B., Gaschen, B., Mallal, S., Mullins, J. I., Nickle, D. C., Herbeck, J., Rousseau, C., Learn, G. H., Miura, T., Brander, C., Walker, B., and Korber, B. (2007). Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science*, **315**(5818), 1583–1586.

Borrow, P., Lewicki, H., Wei, X., Horwitz, M. S., Peffer, N., Meyers, H., Nelson, J. A., Gairin, J. E., Hahn, B. H., Oldstone, M. B., and Shaw, G. M. (1997). Antiviral pressure exerted by HIV-l-specific cytotoxic t lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nature Medicine*, **3**(2), 205–211.

Bronke, C., Almeida, C. M., McKinnon, E., Roberts, S. G., Keane, N. M., Chopra, A., Carlson, J. M., Heckerman, D., Mallal, S., and John, M. (2013). HIV escape mutations occur preferentially at HLA-binding sites of CD8 T-cell epitopes. *AIDS*, **27**, 899–905.

Brumme, Z. L., Brumme, C. J., Heckerman, D., Korber, B. T., Daniels, M., Carlson, J., Kadie, C., Bhattacharya, T., Chui, C., Szinger, J., Mo, T., Hogg, R. S., Montaner, J. S. G., Frahm, N., Brander, C., Walker, B. D., and Harrigan, P. R. (2007). Evidence of differential HLA class i-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. *PLoS Pathogens*, **3**(7), e94.

Brumme, Z. L., Tao, I., Szeto, S., Brumme, C. J., Carlson, J. M., Chan, D., Kadie, C., Frahm, N., Brander, C., Walker, B., Heckerman, D., and Harrigan, P. R. (2008). Human leukocyte antigen-specific polymorphisms in HIV-1 gag and their association with viral load in chronic untreated infection. *AIDS*, **22**(11), 1277–1286.

Brunner, K. T., Mauel, J., Cerottini, J. C., and Chapuis, B. (1968). Quantitative assay of the lytic action of immune lymphoid cells on 51-cr-labelled allogeneic target cells in vitro; inhibition by isoantibody and by drugs. *Immunology*, **14**, 181–196.

Budeus, B., Timm, J., and Hoffmann, D. (2016). SeqFeatR for the discovery of feature-sequence associations. *PLOS ONE*, **11**(1), e0146409.

Carlson, J. M., Brumme, Z. L., Rousseau, C. M., Brumme, C. J., Matthews, P., Kadie, C., Mullins, J. I., Walker, B. D., Harrigan, P. R., Goulder, P. J. R., and Heckerman, D. (2008). Phylogenetic dependency networks: Inferring patterns of CTL escape and codon covariation in HIV-1 gag. *PLoS Computational Biology*, **4**(11), e1000225.

Carlson, J. M., Listgarten, J., Pfeifer, N., Tan, V., Kadie, C., Walker, B. D., Ndung'u, T., Shapiro, R., Frater, J., Brumme, Z. L., Goulder, P. J. R., and Heckerman, D. (2012). Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *Journal of Virology*, **86**(9), 5230–5243.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**(2), 465–480.

Czerkinsky, C. C., Nilsson, L.-Å., Nygren, H., Örjan Ouchterlony, and Tarkowski, A. (1983). A solid-phase enzyme-linked immunospot (ELISPOT) assay for enumeration of specific antibody-secreting cells. *Journal of Immunological Methods*, **65**(1-2), 109–121.

Fisher, R. A. (1922). On the interpretation of chi2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, **85**(1), 87.

Francke, U. and Pellegrino, M. A. (1977). Assignment of the major histocompatibility complex to a region of the short arm of human chromosome 6. *Proceedings of the National Academy of Sciences*, **74**(3), 1147–1151.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **182**(2), 389–402.

Gall, S. L., Stamegna, P., and Walker, B. D. (2007). Portable flanking sequences modulate CTL epitope processing. *Journal of Clinical Investigation*, **117**(11), 3563–3575.

Gelman, A. (2010). Bayesian statistics then and now. *Statistical Science*, **25**(2), 162–165.

Gelman, A. and Carlin, J. (2014). Beyond power calculations. *Perspectives on Psychological Science*, **9**(6), 641–651.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**(4), 457–472.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, **5**(2), 189–211.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Germain, R. N. (1994). MHC-dependent antigen processing and peptide presentation: Providing ligands for t lymphocyte activation. *Cell*, **76**(2), 287–299.

Goldberg, A. L., Cascio, P., Saric, T., and Rock, K. L. (2002). The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides. *Molecular Immunology*, **39**(3-4), 147–164.

Haas, D. W., Wilkinson, G. R., Kuritzkes, D. R., Richman, D. D., Nicotera, J., Mahon, L. F., Sutcliffe, C., Siminski, S., Andersen, J., Coughlin, K., *et al.* (2003). A multi-investigator/institutional dna bank for aids-related human genetic studies: Aactg protocol a5128. *HIV clinical trials*, **4**(5), 287–300.

Harty, J. T., Tvinnereim, A. R., and White, D. W. (2000). CD8+ t cell effector mechanisms in resistance to infection. *Annual Review of Immunology*, **18**(1), 275–308.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, **2**(8), e124.

Ives, A. R. and Garland, T. (2009). Phylogenetic logistic regression for binary dependent variables. *Systematic Biology*, **59**(1), 9–26.

Ives, A. R. and Garland, T. (2014). Phylogenetic regression for binary dependent variables. In *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, pages 231–261. Springer Berlin Heidelberg.

John, M., Heckerman, D., Park, L., Gaudieri, S., and Chopra, A. (2008). Genome-wide hla-associated selection in hiv-1 and protein-specific correlations with viral load: An actg5142. 15th conference on retroviruses and opportunistic infections (croi)(abstract 312).

Karimzadeh, H., Kiraithe, M. M., Kosinska, A. D., Glaser, M., Fiedler, M., Oberhardt, V., Alizei, E. S., Hofmann, M., Mok, J. Y., Nguyen, M., van Esch, W. J. E., Budeus, B., Grabowski, J., Homs, M., Olivero, A., Keyvani, H., Rodríguez-Frías, F., Tabernero, D., Buti, M., Heinold, A., Alavian, S. M., Bauer, T., zur Wiesch, J. S., Raziorrouh, B., Hoffmann, D., Smedile, A., Rizzetto, M., Wedemeyer, H., Timm, J., Antes, I., Neumann-Haefelin, C., Protzer, U., and Roggendorf, M. (2018). Amino acid substitutions within HLA-b*27-restricted t cell epitopes prevent recognition by hepatitis delta virus-specific CD8+ t cells. *Journal of Virology*, **92**(13), e01891–17.

Karimzadeh, H., Kiraithe, M. M., Oberhardt, V., Alizei, E. S., Bockmann, J., zur Wiesch, J. S., Budeus, B., Hoffmann, D., Wedemeyer, H., Cornberg, M., Krawczyk, A., Rashidi-Alavijeh, J., Rodríguez-Frías, F., Casillas, R., Buti, M., Smedile, A., Alavian, S. M., Heinold, A., Emmerich, F., Panning, M., Gostick, E., Price, D. A., Timm, J., Hofmann, M., Raziorrouh, B., Thimme, R., Protzer, U., Roggendorf, M., and Neumann-Haefelin, C. (2019). Mutations in hepatitis d virus allow it to escape detection by cd8+ t cells and evolve at the population level. *Gastroenterology*, **156**(6), 1820–1833.

Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., Gatanaga, H., Fujiwara, M., Hachiya, A., Koizumi, H., Kuse, N., Oka, S., Duda, A., Prendergast, A., Crawford, H., Leslie, A., Brumme, Z., Brumme, C., Allen, T., Brander, C., Kaslow, R., Tang, J., Hunter, E., Allen, S., Mulenga, J., Branch, S., Roach, T., John, M., Mallal, S., Ogwu, A., Shapiro, R., Prado, J. G., Fidler, S., Weber, J., Pybus, O. G., Klenerman, P., Ndung'u, T., Phillips, R., Heckerman, D., Harrigan, P. R., Walker, B. D., Takiguchi, M., and Goulder, P. (2009). Adaptation of HIV-1 to human leukocyte antigen class i. *Nature*, **458**(7238), 641–645.

Kløverpris, H. N., Leslie, A., and Goulder, P. (2016). Role of HLA adaptation in HIV evolution. *Frontiers in Immunology*, **6**.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, **35**(21), 4453–4455.

Lamoreaux, L., Roederer, M., and Koup, R. (2006). Intracellular cytokine optimization and standard operating procedure. *Nature Protocols*, **1**(3), 1507–1516.

Lumley, S. F., McNaughton, A. L., Klenerman, P., Lythgoe, K. A., and Matthews, P. C. (2018). Hepatitis b virus adaptation to the cd8+ t cell response: Consequences for host and pathogen. *Frontiers in Immunology*, **9**.

Matthews, P. C., Prendergast, A., Leslie, A., Crawford, H., Payne, R., Rousseau, C., Rolland, M., Honeyborne, I., Carlson, J., Kadie, C., Brander, C., Bishop, K., Mlotshwa, N., Mullins, J. I., Coovadia, H., Ndung'u, T., Walker, B. D., Heckerman, D., and Goulder, P. J. R. (2008). Central role of reverting mutations in HLA associations with human immunodeficiency virus set point. *Journal of Virology*, **82**(17), 8548–8559.

Mei, S., Li, F., Leier, A., Marquez-Lago, T. T., Giam, K., Croft, N. P., Akutsu, T., Smith, A. I., Li, J., Rossjohn, J., Purcell, A. W., and Song, J. (2020). A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief. Bioinformatics*, **21**, 1119–1135.

Milicic, A., Price, D. A., Zimbwa, P., Booth, B. L., Brown, H. L., Easterbrook, P. J., Olsen, K., Robinson, N., Gileadi, U., Sewell, A. K., Cerundolo, V., and Phillips, R. E. (2005). CD8+ t cell epitope-flanking mutations disrupt proteasomal processing of HIV-1 nef. *The Journal of Immunology*, **175**(7), 4618–4626.

Moore, C. B. (2002). Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science*, **296**(5572), 1439–1443.

Murata, S., Sasaki, K., Kishimoto, T., i. Niwa, S., Hayashi, H., Takahama, Y., and Tanaka, K. (2007). Regulation of c8+ t cell development by thymus-specific proteasomes. *Science*, **316**(5829), 1349–1353.

O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. (2020). MHCflurry 2.0: Improved pan-allele prediction of MHC class i-presented peptides by incorporating antigen processing. *Cell Systems*, **11**(1), 42–48.e7.

Osborne, J. W. and Waters, E. (2002). Four assumptions of multiple regression that researchers should always test.

Piironen, J. and Vehtari, A. (2017a). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior.

Piironen, J. and Vehtari, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, **11**(2), 5018–5051.

Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. E. (2014). The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research*, **43**(D1), D423–D431.

Roomp, K., Beerenwinkel, N., Sing, T., Schülter, E., Büch, J., Sierra-Aragon, S., Däumer, M., Hoffmann, D., Kaiser, R., Lengauer, T., and Selbig, J. (2006). Arevir: A secure platform for designing personalized antiretroviral therapies against HIV. In *Lecture Notes in Computer Science*, pages 185–194. Springer Berlin Heidelberg.

Rousseau, C. M., Daniels, M. G., Carlson, J. M., Kadie, C., Crawford, H., Prendergast, A., Matthews, P., Payne, R., Rolland, M., Raugi, D. N., Maust, B. S., Learn, G. H., Nickle, D. C., Coovadia, H., Ndung'u, T., Frahm, N., Brander, C., Walker, B. D., Goulder, P. J. R., Bhattacharya, T., Heckerman, D. E., Korber, B. T., and Mullins, J. I. (2008). HLA class i-driven evolution of human immunodeficiency virus type 1 subtype c proteome: Immune escape and viral load. *Journal of Virology*, **82**(13), 6434–6446.

Ruhl, M., Chhatwal, P., Strathmann, H., Kuntzen, T., Bankwitz, D., Skibbe, K., Walker, A., Heinemann, F. M., Horn, P. A., Allen, T. M., Hoffmann, D., Pietschmann, T., and Timm, J. (2011). Escape from a dominant HLA-b∗15-restricted CD8 t cell response against hepatitis c virus requires compensatory mutations outside the epitope. *Journal of Virology*, **86**(2), 991–1000.

Scariano, S. M. and Davenport, J. M. (1987). The effects of violations of independence assumptions in the one-way ANOVA. *The American Statistician*, **41**(2), 123–129.

Stan Development Team (2021). *Stan Modeling Language Users Guide and Reference Manual, 2.23. https://mc-stan.org*.

Vehtari, A., Gelman, A., and Gabry, J. (2016). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**(5), 1413–1432.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.*, **47**, D339–D343.

Yewdell, J. W. and Hill, A. B. (2002). Viral interference with antigen presentation. *Nature Immunology*, **3**(11), 1019–1025.

Yusim, K., David-Fung, E.-S., Korber, B. T. M., Brander, C., Barouch, D., de Boer, R., Haynes, B. F., Koup, R., Moore, J. P., Walker, B. D., and Watkins, D. I., editors (2018). *HIV Molecular Immunology*. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico.