Bioinformatics, 38(9), 2022, 2428–2436
https://doi.org/10.1093/bioinformatics/btac134
Advance Access Publication Date: 3 March 2022
Original Paper

Sequence analysis

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

Daniel Habermann [1,*], Hadi Kharimzadeh[2], Andreas Walker[3], Yang Li[4], Rongge Yang[4], Rolf Kaiser[5], Zabrina L. Brumme[6,7], Jörg Timm[3], Michael Roggendorf[8] and Daniel Hoffmann [1,9,10,*]

[1]Bioinformatics and Computational Biophysics, Faculty of Biology, University of Duisburg-Essen, Essen 45117, Germany, [2]Division of Clinical Pharmacology, University Hospital, LMU Munich, Munich, Germany, [3]Institute of Virology, Medical Faculty, University Hospital Düsseldorf, Heinrich-Heine-Universität, Düsseldorf 40225, Germany, [4]AIDS and HIV Research Group, State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Science, Wuhan, China, [5]Institute of Virology, University of Cologne, Faculty of Medicine and University Hospital of Cologne, Cologne 50935, Germany, [6]Faculty of Health Sciences, Simon Fraser University, Burnaby, Canada, [7]British Columbia Centre for Excellence in HIV/AIDS, Vancouver, Canada, [8]Institute of Virology, School of Medicine, Technical University of Munich/Helmholtz Zentrum München, Munich, Germany, [9]Center of Medical Biotechnology, University of Duisburg-Essen, Essen, Germany and [10]Center for Computational Sciences and Simulation, University of Duisburg-Essen, Essen, Germany

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on July 22, 2021; revised on November 21, 2021; editorial decision on February 11, 2022; accepted on February 28, 2022

Abstract

Motivation: A key process in anti-viral adaptive immunity is that the human leukocyte antigen (HLA) system presents epitopes as major histocompatibility complex I (MHC I) protein–peptide complexes on cell surfaces and in this way alerts CD8+ cytotoxic T-lymphocytes (CTLs). This pathway exerts strong selection pressure on viruses, favoring viral mutants that escape recognition by the HLA/CTL system. Naturally, such immune escape mutations often emerge in highly variable viruses, e.g. HIV or HBV, as HLA-associated mutations (HAMs), specific to the hosts MHC I proteins. The reliable identification of HAMs is not only important for understanding 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: We present a new Bayesian regression model for the detection of HAMs that integrates a sparsity-inducing prior, epitope predictions and phylogenetic bias assessment, and that yields easily interpretable quantitative information on HAM candidates. The model predicts experimentally confirmed HAMs as having high posterior probabilities, and it performs well in comparison to state-of-the-art models for several datasets from individuals infected with HBV, HDV and HIV.

Availability and implementation: The source code of this software is available at https://github.com/HAMdetector/Escape.jl under a permissive MIT license. The data underlying this article were provided by permission. Data will be shared on request to the corresponding author with permission of the respective co-authors.

Contact: daniel.habermann@uni-due.de or daniel.hoffmann@uni-due.de

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

1 Introduction

1.1 The human leukocyte antigen system

The human immune system recognizes viral infections through two pathways: The innate and adaptive immune response. T-cell, or 'cellular', immunity, which represents one major arm of the adaptive immune system, is modulated by human leukocyte antigen (HLA) molecules (Germain, 1994): Briefly, proteins that are synthesized within the cell—which will include viral proteins if the cell is infected—are degraded in proteasomes to peptides (Goldberg *et al.*, 2002). Some of these peptides are presented as epitopes on the cell surface by HLA class I molecules. These viral peptide-HLA complexes can then be recognized by circulating $CD8^+$ cytotoxic T-lymphocytes (CTLS) through their T-cell receptor (Murata *et al.*, 2007). Following this recognition, the CTL can eliminate the infected cell (Harty *et al.*, 2000).

HLA class I molecules are encoded at three loci, HLA-A, -B and -C and these genes are very polymorphic with more than 20 000 known alleles in humans (Robinson *et al.*, 2015). HLA molecules vary drastically in their affinities to given epitopes so that cells from different individuals, in general, present different peptides on the cell surface. In other words, the HLA class I alleles expressed by a given individual will determine their CTL response to a given viral pathogen.

1.2 Immune escape is reproducible based on host HLA allele expressed

Virus variants arise continuously through mutation. Because the HLA system modulates CTL responses through viral epitope presentation, it exerts strong selection pressure toward virus variants that escape CTL recognition (Borrow *et al.*, 1997). Such variants could, for example, carry mutations that reduce binding of viral epitopes to HLA, or that reduce recognition of the epitope/HLA complex by the CTL's T-cell receptor, or that alter peptide processing so that epitopes are no longer presented on the infected cell surface (Yewdell *et al.*, 2002). The latter type of mutation can occur within (Yokomaku *et al.*, 2004) or outside (Draenert *et al.*, 2004) CTL epitopes.

Immune escape is a major driver of viral evolution, particularly for highly variable viruses such as HIV or HBV (Alizon *et al.*, 2011; Allen *et al.*, 2005; Lumley *et al.*, 2018; Rousseau *et al.*, 2008). Whether and how quickly a given escape mutation is selected in a host depends on a number of factors including the viral genomic background, the magnitude of the reduction in viral replication caused by changes in the viral proteins, the selection of compensatory mutations that recover fitness, and the strength of immune response targeting the presented epitope (Kløverpris *et al.*, 2015). Despite the complexity of these factors, the mutational pathways of immune escape in certain viruses such as HIV are nevertheless broadly reproducible, and thus predictable, based on the HLA alleles expressed by the host. For example, about 75% of people living with HIV who carry the HLA-B*57 allele, will select a T242N substitution in the HIV structural protein Gag in the first weeks to months of infection (Brumme *et al.*, 2008b; Leslie *et al.*, 2004).

In addition to driving viral evolution at the individual level, HLA pressures also drive viral evolution in human populations, as circulating viruses adapt to HLA alleles commonly expressed in that population (Kawashima *et al.*, 2009). Upon transmission to a new host with different HLA alleles, HLA escape mutations may revert, particularly if they are associated with a reduction in viral replication capacity (Matthews *et al.*, 2008), but they can also persist, leading to their population-level accumulation (Kawashima *et al.*, 2009).

Methods to accurately and comprehensively identify HLA-associated immune escape mutations in HIV and other viruses are therefore critical for the study of viral evolution and immune escape. An improved understanding of immune escape can aid in the development of treatments and vaccines that rely on effective immune responses.

1.3 Identifying HLA escape mutations

There are several experimental methods available to study HLA escape (Altman *et al.*, 1996; Brunner *et al.*, 1968; Czerkinsky *et al.*, 1983; Lamoreaux *et al.*, 2006). However, these methods are relatively slow and costly, especially for screening purposes. Theoretically, an option to identify escape mutation could be the use of epitope prediction tools (Mei *et al.*, 2020). At their core, these tools identify epitopes as peptides with high predicted affinities to HLA molecules. One could envisage applying such tools to combinatorially mutated epitopes to identify substitutions that reduce predicted affinities significantly and therefore would be good candidates for escape mutations. However, these tools can be rather insensitive to escape mutations (Acevedo-Sáenz *et al.*, 2015), which is not unexpected because they have been developed to recognize epitopes as a whole. A more 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 bound to HLA molecules on the infected cell surface, escape mutations are HLA allele specific and can therefore be detected as HLA allele dependent amino acid substitutions, or 'footprints,' in sequence alignments of viral proteins (Moore, 2002). Amino acid substitutions enriched in viral sequences from hosts with a specific HLA allele are termed HLA-associated mutations (HAM).

One way of quantifying this enrichment is Fisher's exact test (Fisher, 1922): For a given substitution S_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 (S_i, H), $(S_i, \neg H)$, $(\neg S_i, H)$ and $(\neg S_i, \neg H)$, where $\neg S_i$ denotes any substitution except S_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 substitution S_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 *et al.*, 2002; Scariano *et al.*, 1987).

Another issue with Fisher's exact test is the genomic proximity of human HLA class I loci (Francke *et al.*, 1977) leading to linkage disequilibrium—inheritance of HLA alleles can be correlated. Therefore, spurious HAMs can occur if associations of substitutions with individual HLA alleles are tested: if HLA allele H_1 is associated with an amino acid substitution R because of immune escape, but H_1 is in linkage disequilibrium with allele H_2, then 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 et al., 2018): with small effect sizes and high variance between measurements, as is often the case with biological data, statistically significant results can be misleading, can have the wrong direction (type S error), or can greatly overestimate an effect (type M error) (Gelman *et al.*, 2014). Such problems are more and more appreciated in the context of the current 'replication crisis'—in the life sciences scientific claims with seemingly strong statistical

support often fail to replicate (Baker, 2016; Begley *et al.*, 2012; Ioannidis, 2005).

These problems are exacerbated if *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. A common strategy to mitigate this problem is to control the false discovery rate (Benjamini *et al.*, 1995). The downside of such adjustment procedures is that only the very largest effects remain if large datasets are screened.

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. Multilevel models can make the problem of multiple comparisons disappear entirely and yield more valid estimates (Gelman *et al.*, 2012).

2 Materials and methods

Our general approach for HAMdetector is to fit a Bayesian regression model that captures relationships between host HLA alleles and substitutions in viral proteomes.

This Bayesian approach is advantageous because it allows use of: (i) prior information (e.g. knowledge of effect magnitudes), (ii) relevant additional information (phylogeny, epitope information), (iii) a problem-specific structure and (iv) 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. This is the core of HAMdetector, which models the strength of association between substitutions in viral sequences and host HLA alleles.

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

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

where y_{ik} is the binary encoded observation of substitution 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 substitution k in sequence i; β_{0k} is an intercept for substitution k, corresponding to the overall log-odds for substitution k; X_{ij} is 1 if sequence i comes from host individual with HLA allele j and 0 otherwise; β_{jk} is the HLA regression coefficient of HLA allele j for substitution 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 substitution 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 substitution k increase by addition of β_{jk}.

Reasoning about coefficients on the log-odds scale can sometimes be unintuitive. A useful approximation to interpret logistic regression coefficients on the probability scale is the so-called divide-by-4 rule, which means that a regression coefficient of 2 corresponds to an expected increase on the probability scale of up to $2/4 = 50\%$.

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 major histocompatibility complex I (MHC I) affinity, as described in the following.

2.2.1 Phylogeny

Viral strains have a common phylogenetic history. Thus substitutions 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.

A 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 *et al.*, 2010, 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 substitution 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 and 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):

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

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 TCRs (Bronke *et al.*, 2013). This suggests that knowledge of epitope regions can be used to boost HAM detection. Fortunately, availability of large experimental datasets (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 to present epitopes to T cell receptors, but also mutations at epitope-flanking positions that interfere with pre-processing of peptides, notably proteasomal cleavage of viral proteins (Le Gall *et al.*, 2007; Milicic *et al.*, 2005).

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 substitutions 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 pairs of epitopes and HLA alleles. For the binarization, we use the rank threshold of 2% suggested by MHCflurry.

We use epitope prediction as information about the expected degree of sparsity, i.e. if we know that there is an epitope restricted by

a given HLA allele at that location, we expect that this HLA allele is more likely to be associated with substitutions 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 substitution 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 *et al.*, 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 HAMs: the dominating mechanism that leads to HLA association of mutations is probably selection of mutations that mediate escape from MHC I presentation of epitopes; however, we know that these epitopes are sparse, i.e. the number of actual epitopes that are restricted by a given HLA allele is typically small compared to the number of all conceivable epitopes. Thus, for most pairs of HLA allele and substitution, the association is likely truly zero. Note that this reasoning does not preclude associations outside of epitopes as sometimes observed for compensatory mutations (Ruhl *et al.*, 2012) but just implies that these are more rare.

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 *et al.*, 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 $\tilde{\lambda}_{jk}^2 = \frac{c_k^2\lambda_{jk}^2}{c_k^2+\tau^2\lambda_{jk}^2}$, 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 *et al.*, 2017a).

The full model specification together with a prior justification is given in Supplementary Information.

2.3 Model implementation

A Julia (Bezanson *et al.*, 2017) package is available at https://github.com/HAMdetector/Escape.jl to run the model on custom data. Due to restrictions of dependencies (MHCflurry and RAxML-ng), HAMdetector is currently only available on Linux, but can be run on Windows using the Windows Subsystem for Linux (WSL2). 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 in two versions: One optimized for readability and one optimized for speed by utilizing Stan's multithreading and GPU capabilities.

2.4 Model diagnostics

2.4.1 Convergence diagnostics

We use the split-$\hat{R}$ convergence diagnostic to identify Markov chain convergence issues (Gelman *et al.*, 1992, 2013). We require a value of $\hat{R}$ below 1.1 for all model parameters. Additionally, we require that the effective sample size N_{eff} (Stan Development Team, 2021) is above 200 for all model parameters and that sampling occurs without any divergent transitions (Betancourt, 2018).

2.4.2 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* (Supplementary Fig. S1): two binned quantities were plotted against each other, the observed relative frequencies of substitutions $f(y_{ik} = 1)$, 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 substitutions 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-values 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. Figure 1.

2.4.3 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.*, 2017).

From the LOOCV, we obtain the Expected Log-Predictive Density (ELPD) $\sum_{i=1}^{n} \log\left(\int p(y_i|\theta)p(\theta|y_{i-1})d\theta\right)$ for samples $i = 1, \ldots, n$, ith observation y_i, data y_{i-1} with the ith data point 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 of not only taking into account the location of the predictive distribution (the number of correct predictions) but also the width, i.e. how confident the model is in its predictions. For a description of ELPD in the context of LOOCV, see Vehtari *et al.* (2017) and Gneiting *et al.* (2007).

2.5 Data

The model was fit with several datasets consisting of viral sequences paired to host HLA class I data:

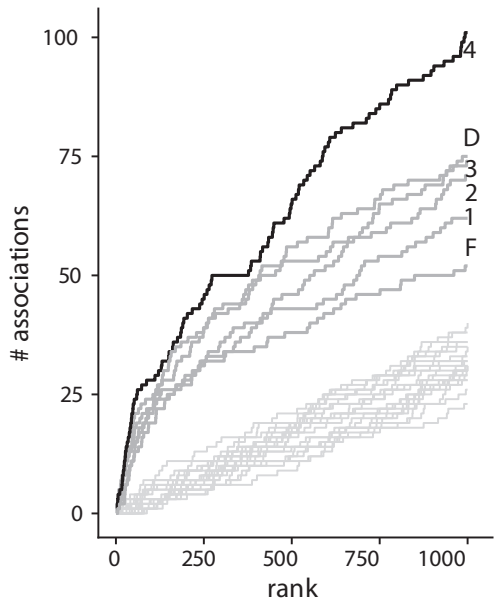

Fig. 1. HAM enrichment plot for HBV preC/core protein: number N_e of associations inside the boundary of known epitopes versus rank r. D: Phylogenetic Dependency Network; F: Fisher's exact test; 1: simple logistic regression model with broad Student-t priors; 2: logistic regression model with horseshoe 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—substitution pairs and act as baselines

- A large HIV dataset consisting of a subset of sequences from the HOMER (Brumme *et al.*, 2007, 2008a) cohort, the Western Australian HIV Cohort Study (WAHCS, Bhattacharya *et al.*, 2007; Moore, 2002) and participants of the US AIDS Clinical Trials Group (ACTG) protocol 5142 (John *et al.*, 2008) who also provided Human DNA under ACTG protocol 5128 (Haas *et al.*, 2003) (total $N = 1383$). These data 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 (Timm *et al.*, 2021). The dataset contains sequences of the preC/core, LHBs, Pol and HBx proteins.
- A set of 104 Hepatitis-D-Virus sequences containing the HDV-antigen (Karimzadeh *et al.*, 2018).
- A set of 41 HIV sequences spanning the *gag* and *pol* genes.

Lists of known epitopes were collected from the Immune Epitope Database (IEDB, Vita *et al.* (2019)). For HBV and HIV, we added data from the Hepitopes database (Lumley *et al.*, 2016) and the Los Alamos HIV Molecular Immunology Database (Yusim *et al.*, 2018), respectively. In total, we obtained 20 epitopes for HDV, 339 for HBV and 2684 epitopes for HIV. The counts refer to unique pairs of epitope and HLA alleles.

2.6 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 representations.
3. Multiple sequence alignments were produced with MAFFT (Katoh *et al.*, 2013) (default parameters). In the few cases when the alignment introduced frameshifts, these were corrected manually.
4. 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 (Equation 1) and adding then the further components, i.e. the horseshoe prior (Equation 5), phylogeny (Equation 3) and epitope prediction, resulting in the full model (Supplementary Equation S1). 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.

3.1 Run times and convergence

For a standard office 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 substitutions. 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 200 for all model parameters, $\hat{R}$ convergence diagnostic values were below 1.1 in all cases.

3.2 Posterior predictive checks

The model yields well-calibrated posterior predictive probabilities of substitutions. This is exemplified in Supplementary Figure S1 for HBV core protein, but also holds true for the other datasets (Supplementary Fig. 'Calibration plots').

The predictions of the tested models are enriched in epitopes over baselines for almost all tested datasets (Fig. 1 for HBV preC/core protein and Supplementary Fig. 'HAM enrichment plots' for other datasets). Although the relative and absolute performance varies by protein (see Supplementary Fig. 'HAM enrichment summary'), HAMdetector consistently outperforms all other methods in all but two datasets, and performs on-par with the other methods in these two cases. For the best ranked HAMs, Fisher's exact test performs about as well as the HAMdetector backbone logistic regression model (model 1 in Fig. 1). Each of the following three model stages of HAMdetector increases HAM enrichment further. The horseshoe prior alone (model 2) is a drastic improvement over model 1, even though it does not include any specific 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. 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. The latter are here only used for model evaluation.

The Bayesian approach lends itself to incorporation of prior knowledge which usually helps in accurate modeling and prediction. In fact, a considerable effect is confirmed by the HAM enrichment plots with their ladder of improvements with increasing inclusion of information. It may be particularly surprising that the sparsifying horseshoe prior has such an impact although it does not use specific prior information. 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 2 illustrates the effect of the sparsifying prior with an example, the substitution 11D in HIV integrase (Arevir dataset). There is no evidence for an association of HLA-A*01 with this substitution, whereas for HLA-B*44 the data is consistent with a strong association. The horseshoe prior has the effect of shrinking toward 0 specifically those regression coefficients with weak evidence of an association (A*01 in Fig. 2). 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.

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 HBV preC/core protein 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 substitutions and samples, epitope prediction only influences those substitutions 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 substitution, as shown in the previous section.

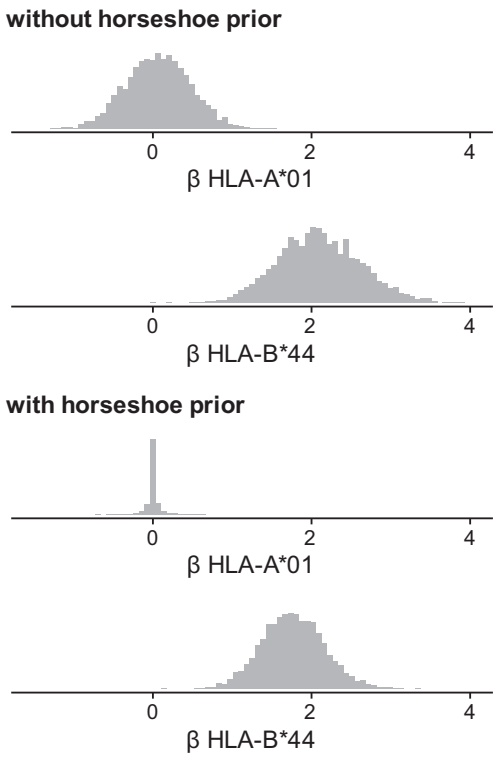

Fig. 2. Marginal posterior distributions of regression coefficients for the association of substitution 11D of the HIV integrase with HLA alleles A*01 and B*44. Top half: inferred with logistic regression model, bottom half: inferred with logistic regression with sparsifying horseshoe prior

Table 1. Prediction performance changes in terms of ELPD as HAMdetector components are added

	$ELPD_{diff}$	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

Note: Values computed for HBV preC/core protein. Each value in the column $ELPD_{diff}$ is the ELPD difference to the model in the previous row, e.g. the ELPD difference between the model with epitope prediction and the previous one is 63.1. Models with larger ELPD have better predictive performance, e.g. the model with all components, including epitope prediction, has better predictive performance than the model lacking epitope prediction. All differences are several times the estimated standard error (column se_{diff}) away from zero, indicating that models that include more information have better predictive performance.

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 (i) a set of paired HDV sequences and patient HLA alleles, (ii) HAM predictions by Fisher's exact test as implemented in SeqFeatR (Budeus *et al.*, 2016) and (iii) an *in vitro* assay to quantify the effect of the predicted HAMs on IFN-γ release of $CD8^+$ T cells (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} (Supplementary Table S1) 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 Supplementary Table S1 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 strongly supports 15 substitution—allele pairs that have previously not been identified (question marks in last column of Supplementary Table S1). All of them have association probabilities of 0.90 or higher, while their *P*-values from Fisher's exact test exceed the significance level of 5×10^{-3} used in Karimzadeh *et al.* (2019). Given the superior performance of HAMdetector on the experimentally tested HAMs, these 15 bona fide predictions suggest that most true HAMs may still to be discovered. A striking example is K43R—A*02 with a *P*-value of 0.22 in Fisher's exact test but a HAM-probability of 0.90 and location inside an A*02 restricted epitope.

3.5 Linkage disequilibrium

For three of the false positives proposed by Fisher's exact test (Supplementary Table S1), HAMdetector identifies associations with the same substitution 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: If a certain HLA allele selects for a specific HAM and there is another HLA allele that co-occurs with that HLA allele, any method that relies on the

statistical analysis of pairs of HLA allele and substitution alone will also detect these associations. Due to random sampling variation, the HLA allele that selects for a mutation might not necessarily have the strongest correlation. Inclusion of additional information like epitope prediction can help to identify associations that are otherwise confounded by noise.

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 made for K43R and D33E, although the correlation between the respective alleles is much weaker. A*30 and B*13 have been shown to be in strong linkage disequilibrium (Brumme *et al.*, 2007, Supplementary Table S2).

Figure 3 shows regression coefficients of the HLA alleles A*30 and B*13 for substitution P49L. With the simplest logistic regression model (model 1), both A*30 and B*13 have medium evidence of being associated with substitution P49L. However, with phylogeny and sparsity-promoting prior (model 3) both regression coefficients shrink close to 0—the associations are not convincingly supported by the data. Using epitope prediction as additional source of information (model 4) allows to disentangle the association of the correlated alleles with P49L and identify B*13 as likely associated with P49L. The association between P49L and A*30 (predicted by Fisher's exact test) remains shrunk toward 0.

3.6 HAMs outside epitopes

It is important to consider that biologically relevant HAMs do not necessarily have to lie within or close to the boundary of an epitope. In Supplementary Section S5, we outline that the model is still able to identify associations outside predicted epitopes and that epitope information augments evidence obtained from sequence data.

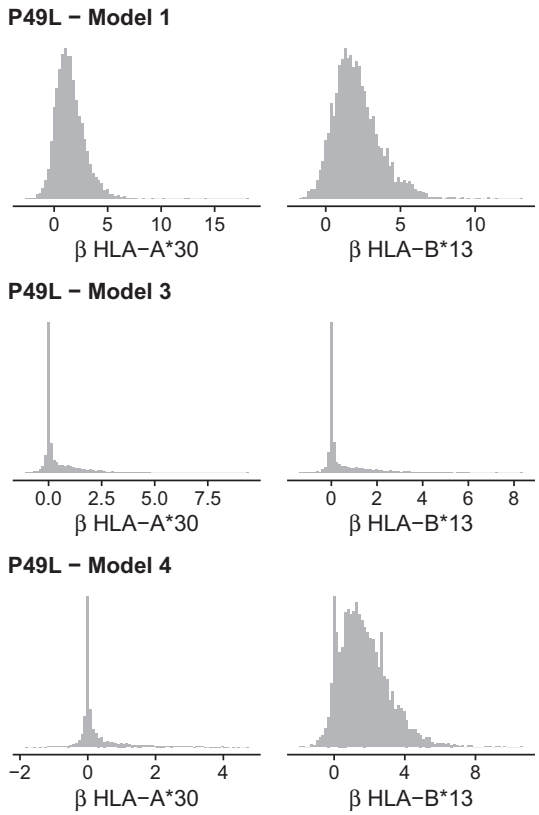

Fig. 3. Marginal posteriors for the regression coefficients of A*30 (left column) and B*13 (right column) for substitution P49L with different model stages (rows)

4 Discussion

HAMdetector follows a general paradigm of Bayesian modeling, 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 substitution k. The more relevant information we infuse into the model, the sharper the inference. HAMdetector outperforms other methods as it includes an unprecedented amount of relevant information.

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 Supplementary Table S2 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.

Generally, the platform character of HAMdetector model allows optimization of prediction performance by replacing components by more powerful ones, for example replacing a single epitope predictor by an ensemble predictor (Hu *et al.*, 2010). 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.

Acknowledgements

The authors thank Drs. Mina John and Simon Mallal for providing data.

Funding

This work was supported by Deutsche Forschungsgemeinschaft (grant HO 1582/10-1). ZLB is supported by the Canadian Institutes for Health Research (through project grant PJT-148621) and by the Michael Smith Foundation for Health Research (through a Scholar Award).

Conflict of Interest: none declared.

References

Acevedo-Sáenz,L. *et al.* (2015) Selection pressure in CD8+ T-cell epitopes in the pol gene of HIV-1 infected individuals in Colombia. A bioinformatic approach. *Viruses*, **7**, 1313–1331.

Alizon,S. *et al.* (2011) Epidemiological and clinical consequences of within-host evolution. *Trends Microbiol.*, **19**, 24–32.

Allen,T.M. *et al.* (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. *J. Virol.*, **79**, 13239–13249.

Altman,J.D. *et al.* (1996) Phenotypic analysis of antigen-specific T lymphocytes. *Science*, **274**, 94–96.

Amrhein,V. *et al.* (2018) Remove, rather than redefine, statistical significance. *Nat. Hum. Behav.*, **2**, 4–4.

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

Begley,C.G. *et al.* (2012) Drug development: raise standards for preclinical cancer research. *Nature*, **483**, 531–533.

Benjamini,Y. *et al.* (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.

Betancourt, M. (2017). *A conceptual introduction to Hamiltonian Monte Carlo*. arXiv preprint arXiv:1701.02434.

Bezanson,J. *et al.* (2017) Julia: a fresh approach to numerical computing. *SIAM Rev.*, **59**, 65–98.

Bhattacharya,T. *et al.* (2007) Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science*, **315**, 1583–1586.

Borrow,P. *et al.* (1997) Antiviral pressure exerted by HIV-l-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat. Med.*, **3**, 205–211.

Bronke,C. *et al.* (2013) HIV escape mutations occur preferentially at HLA-binding sites of CD8+ T-cell epitopes. *AIDS*, **27**, 899–905.

Brumme,Z.L. *et al.* (2007) Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. *PLoS Pathogens*, **3**, e94.

Brumme,Z.L. *et al.* (2008a) Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection. *AIDS*, **22**, 1277–1286.

Brumme,Z.L. *et al.* (2008b) Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 (HIV-1) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection. *J. Virol.*, **82**, 9216–9227.

Brunner,K.T. *et al.* (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. *et al.* (2016) SeqFeatR for the discovery of feature-sequence associations. *PLoS One*, **11**, e0146409.

Carlson,J.M. *et al.* (2008) Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Comput. Biol.*, **4**, e1000225.

Carlson,J.M. *et al.* (2012) Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *J. Virol.*, **86**, 5230–5243.

Carvalho,C.M. *et al.* (2010) The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

Czerkinsky,C.C. *et al.* (1983) A solid-phase enzyme-linked immunospot (ELISPOT) assay for enumeration of specific antibody-secreting cells. *J. Immunol. Methods*, **65**, 109–121.

Draenert,R. *et al.* (2004) Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *J. Exp. Med.*, **199**, 905–915.

Fisher,R.A. (1922) On the interpretation of X^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.*, **85**, 87.

Francke,U. *et al.* (1977) Assignment of the major histocompatibility complex to a region of the short arm of human chromosome 6. *Proc. Natl. Acad. Sci. USA*, **74**, 1147–1151.

Gabry,J. *et al.* (2019) Visualization in Bayesian workflow. *J. R. Stat. Soc. Ser. A (Stat. Soc.)*, **182**, 389–402.

Gelman,A. (2010) Bayesian statistics then and now. *Stat. Sci.*, **25**, 162–165.

Gelman,A. *et al.* (1992) Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–472.

Gelman,A. *et al.* (2012) Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Effect.*, **5**, 189–211.

Gelman,A. *et al.* (2013) *Bayesian Data Analysis*. CRC Press, London.

Gelman,A. *et al.* (2014) Beyond power calculations. *Perspect. Psychol. Sci.*, **9**, 641–651.

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

Gneiting,T. *et al.* (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, **102**, 359–378.

Goldberg,A.L. *et al.* (2002) The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides. *Mol. Immunol.*, **39**, 147–164.

Haas,D.W. *et al.*; Adult AIDS Clinical Trials Group. (2003) A multi-investigator/institutional DNA bank for AIDS-related human genetic studies: AACTG Protocol A5128. *HIV Clin. Trials*, **4**, 287–300.

Harty,J.T. *et al.* (2000) CD8+ T cell effector mechanisms in resistance to infection. *Annu. Rev. Immunol.*, **18**, 275–308.

Hu,X. *et al.* (2010) MetaMHC: a meta approach to predict peptides binding to MHC molecules. *Nucleic Acids Res.*, **38**, W474–W479.

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

Ives,A.R. *et al.* (2010) Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.*, **59**, 9–26.

Ives,A.R. *et al.* (2014) Phylogenetic regression for binary dependent variables. In: Garamszegi, L.Z. (ed.) *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Springer, Berlin, Heidelberg, pp. 231–261.

John,M. *et al.* (2008) Genome-wide HLA-associated selection in HIV-1 and protein-specific correlations with viral load: an ACTG5142. In: *15th Conference on Retroviruses and Opportunistic Infections (CROI) (Abstract 312)*, *Boston, MA*.

Karimzadeh,H. *et al.* (2018) Amino acid substitutions within HLA-B27-restricted T cell epitopes prevent recognition by hepatitis delta virus-specific CD8+ T cells. *J. Virol.*, **92**, e01891-17.

Karimzadeh,H. *et al.* (2019) Mutations in hepatitis D virus allow it to escape detection by CD8+ T cells and evolve at the population level. *Gastroenterology*, **156**, 1820–1833.

Katoh,K. *et al.* (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

Kawashima,Y. *et al.* (2009) Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*, **458**, 641–645.

Kløverpris,H.N. *et al.* (2015) Role of HLA adaptation in HIV Evolution. *Front. Immunol.*, **6**, 665.

Kozlov,A.M. *et al.* (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, **35**, 4453–4455.

Lamoreaux,L. *et al.* (2006) Intracellular cytokine optimization and standard operating procedure. *Nat. Protoc.*, **1**, 1507–1516.

Le Gall,S. *et al.* (2007) Portable flanking sequences modulate CTL epitope processing. *J. Clin. Investig.*, **117**, 3563–3575.

Leslie,A.J. *et al.* (2004) HIV evolution: CTL escape mutation and reversion after transmission. *Nat. Med.*, **10**, 282–289.

Lumley,S. *et al.* (2016) Hepitopes: a live interactive database of HLA class I epitopes in hepatitis B virus. *Wellcome Open Res.*, **1**, 9.

Lumley,S.F. *et al.* (2018) Hepatitis B virus adaptation to the CD8+ T cell response: consequences for host and pathogen. *Front. Immunol.*, **9**, 1561.

Matthews,P.C. *et al.* (2008) Central role of reverting mutations in HLA associations with human immunodeficiency virus set point. *J. Virol.*, **82**, 8548–8559.

Mei,S. *et al.* (2020) A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief. Bioinf.*, **21**, 1119–1135.

Milicic,A. *et al.* (2005) CD8+ T cell epitope-flanking mutations disrupt proteasomal processing of HIV-1 Nef. *J. Immunol.*, **175**, 4618–4626.

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

Murata,S. *et al.* (2007) Regulation of CD8+ T cell development by thymus-specific proteasomes. *Science*, **316**, 1349–1353.

O'Donnell,T.J. *et al.* (2020) MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst.*, **11**, 42–48.e7.

Osborne,J.W. *et al.* (2002) Four assumptions of multiple regression that researchers should always test. Pract. Assess. Res. Eval., 8, Article 2.

Piironen,J. *et al.* (2017a) On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 905-913). PMLR.

Piironen,J. *et al.* (2017b) Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.*, **11**, 5018–5051.

Robinson,J. *et al.* (2015) The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.*, **43**, D423–D431.

Roomp,K. *et al.* (2006) Arevir: a secure platform for designing personalized antiretroviral therapies against HIV. In: Leser, U., Naumann, F. and Eckman, B. (eds.), *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, pp. 185–194.

Rousseau,C.M. *et al.* (2008) HLA class I-driven evolution of human immunodeficiency virus type 1 subtype C proteome: immune escape and viral load. *J. Virol.*, **82**, 6434–6446.

Ruhl,M. *et al.* (2012) Escape from a dominant HLA-B15-restricted CD8 T cell response against hepatitis C virus requires compensatory mutations outside the epitope. *J. Virol.*, **86**, 991–1000.

Scariano,S.M. *et al.* (1987) The effects of violations of independence assumptions in the one-way ANOVA. *Am. Stat.*, **41**, 123–129.

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

Timm,J. *et al.* (2021) GenBank Accession Numbers: Genotype A: MZ043025 – MZ043097, Genotype B: MW845286 – MW845312, Genotype C: MW887641 – MW887652, Genotype D: MZ097624 – MZ097884, Genotype E: MW926548 – MW926566.

Vehtari,A. *et al.* (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.*, **27**, 1413–1432.

Vita,R. *et al.* (2019) The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.*, **47**, D339–D343.

Yewdell,J.W. *et al.* (2002) Viral interference with antigen presentation. *Nat. Immunol.*, **3**, 1019–1025.

Yokomaku,Y. *et al.* (2004) Impaired processing and presentation of cytotoxic-T-lymphocyte (CTL) epitopes are major escape mechanisms from CTL immune pressure in human immunodeficiency virus type 1 infection. *J. Virol.*, **78**, 1324–1332.

Yusim,K. *et al.* (eds.) (2018) *HIV Molecular Immunology*. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico.