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], Zabrina L. Brumme[5,6], Jörg Timm[3], Michael Roggendorf[7], Daniel Hoffmann[1,8,9]

Abstract

Motivation

A key process in anti-viral adaptive immunity is that the Human Leukocyte Antigen system (HLA) 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, 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 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 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 https://github.com/HAMdetector/Escape.jl under a permissive MIT license.

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

[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, P. R. China*
[5] *Faculty of Health Sciences, Simon Fraser University, Burnaby, Canada*
[6] *British Columbia Centre for Excellence in HIV/AIDS, Vancouver, Canada*
[7] *Institute of Virology, School of Medicine, Technical University of Munich/Helmholtz Zentrum München, Munich, Germany*
[8] *Center of Medical Biotechnology, University of Duisburg-Essen, Essen, Germany*
[9] *Center for Computational Sciences and Simulation, University of Duisburg-Essen, Essen, Germany*

Contents

1. Introduction

1.1 The HLA 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., 2014). 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 HLA escape

Virus variants arise continuously through mutation. Because the HLA system modulates CTL responses through viral epitope presentation, it exerts strong selection pressure towards 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 and Hill, 2002).

HLA diversity drives viral evolution in individuals where a virus adapts to the specific HLA alleles expressed in the host, and in human populations, where 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).

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 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., 2016).

Immune escape is a driver of viral evolution in individuals and populations, particularly for highly variable viruses such as HIV or HBV (Alizon et al., 2011; Allen et al., 2005; Rousseau et al., 2008; Lumley et al., 2018). Methods to accurately detect immune escape mutations are therefore critical. More broadly, 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 (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 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 en-

riched 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 and Waters, 2002; Scariano and Davenport, 1987).

Another issue with Fisher's exact test is the genomic proximity of human HLA class I loci (Francke and Pellegrino, 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 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, can have the wrong direction (type S error), or can 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" – in the life sciences scientific claims with seemingly strong statistical support often fail to replicate (Ioannidis, 2005; Begley and Ellis, 2012; Baker, 2016).

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 and Hochberg, 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 Bayesian regression models that captures relationships between host HLA alleles and substitutions 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 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 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 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 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 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 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 (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 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 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 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., 2011) 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 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:

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

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 substitutions (even with a global parameter γ the model works reasonably well). The final additional parameters $\tilde{\lambda}_{jk}^2$ and τ_{0k} are explained in detail in Piironen and Vehtari (2017b). Briefly, $\tilde{\lambda}_{jk}^2$ 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 substitution 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 substitution. 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

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.

2.5 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* (Fig. 1): 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. 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 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.

2.6 Data

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

- 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 (WAHCS, Moore (2002); Bhattacharya et al. (2007)) 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 and Walker, 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 gathered from the Immune Epitope Database (IEDB, Vita et al. (2019)).

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 representations.
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.

2.8 Data availability

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.

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),

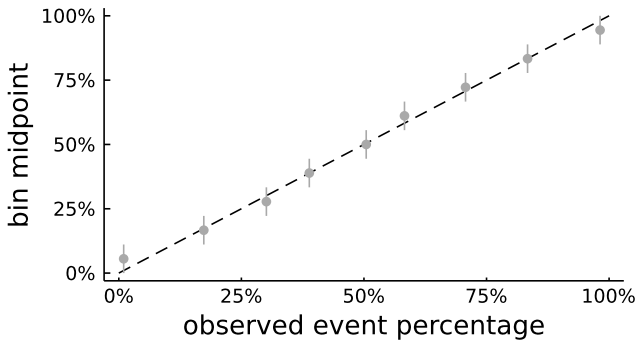

Figure 1. Calibration plot for the HBV PreC/core protein.

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.

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 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 substitutions. This is exemplified in Figure 1 for HBV core protein, but also holds true for the other datasets (Supplementary Figures "Calibration plots").

The predictions of the tested models are enriched in epitopes over baselines for almost all tested datasets (Fig. 2 for HBV preC/core protein and Supplementary Figures "HAM enrichment plots" for other datasets). Although the relative and absolute performance varies by

protein (see supplementary figure "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. 2). 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 3 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 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.

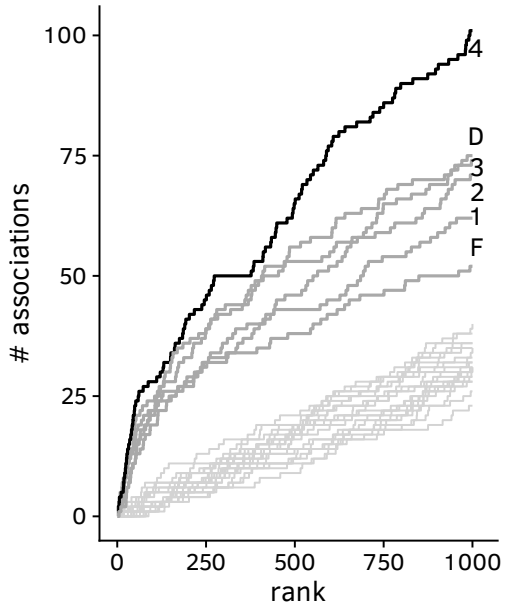

Figure 2. HAM enrichment plot for HBV preC/core protein: number N_e of associations inside the boundary of known epitopes vs. rank r. D: Phylogenetic Dependency Network; F: Fisher's exact test; 1: simple logistic regression model with broad Student-t priors; 2: logistic regression model with horseshoe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

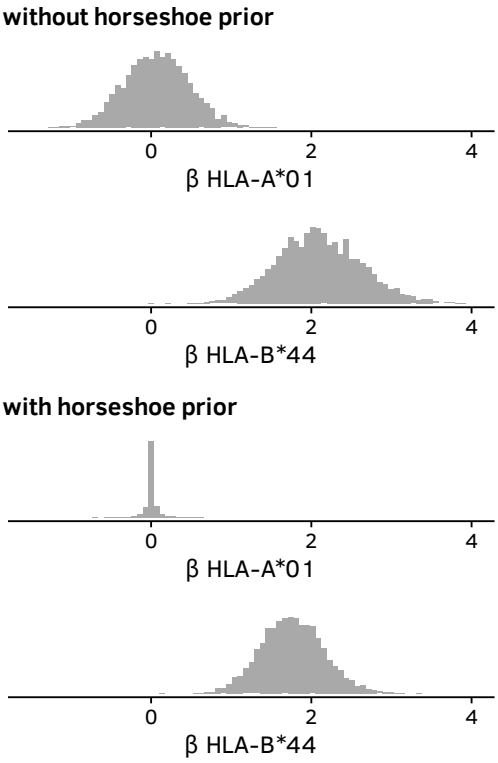

Figure 3. Marginal posterior distributions of regression coefficients for the association of substitution 11D of the HIV integrase with HLA alleles A*01 and B*44. Top half: inferred with logistic regression model, bottom half: inferred with logistic regression with sparsifying horseshoe prior.

Table 1. ELPD changes as HAMdetector components are added. Data computed for HBV preC/core protein. 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 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.

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) 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} (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 strongly supports 15 substitution - allele pairs that have previously not been identified (question marks in last column of Table 2). 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 (Table 2), 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 coeffi-

Table 2. List of HAMs predicted by Fisher's exact test (FET). In the last column "+" and "-" mark experimentally confirmed or rejected HAMs, respectively; "?" below the horizontal line indicate untested bona fide predictions. "post.prob." are posterior probabilities for positive associations computed with HAMdetector.

substitution	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	?

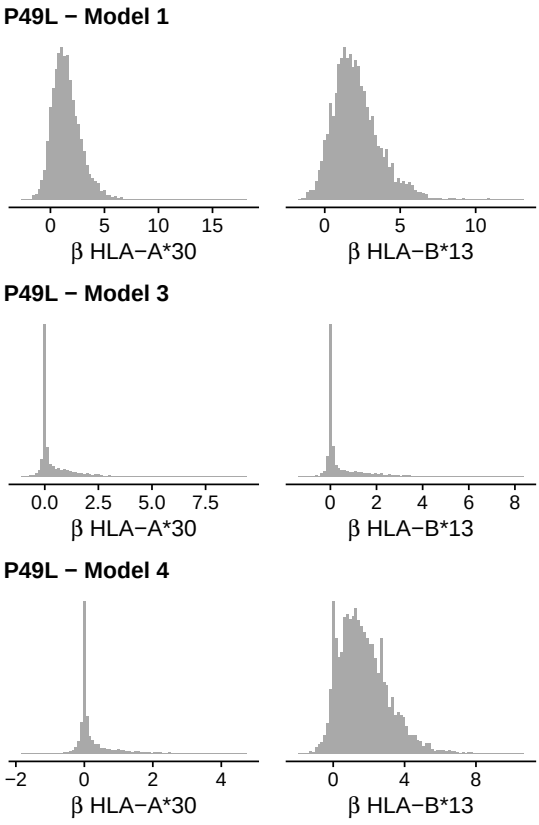

Figure 4. Marginal posteriors for the regression coefficients of A*30 (left column) and B*13 (right column) for substitution P49L with different model stages (rows).

cient $\rho = 0.5$). 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 strong linkage disequilibrium (Brumme et al., 2007, supplementary table 2).

Figure 4 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 towards 0.

3.6 HAMs outside epitopes

The epitope and processing predictions that HAMdetector uses are imperfect, as the underlying tools extrapolate binding affinities for new epitopes based on necessarily incomplete experimental data. Bayesian statistics provides a coherent framework to make use of imperfect data. In HAMdetector, this is achieved by an additional parameter β_{epi} that governs how strongly the model takes an apparent association between a substitution and the corresponding HLA allele into account. By default, the regression coefficients that quantify the strength of association between allele and substitution are shrunk towards 0, and only in the presence of considerable evidence in favor of an association (e.g. because the substitution often co-occurs with a certain HLA allele), this shrinkage is overcome by the observed data.

If the epitope prediction happens to be reliable, i.e. when the presence of a predicted epitope correlates strongly with the probability observing the substitution in a host with the respective HLA allele, the parameter β_{epi} is estimated to be large and less evidence by the sequence data is enough to escape the shrinkage and estimate a non-zero association between allele and substitution, compared to associations that do not lie inside a predicted epitope. Likewise, if the epitope prediction turns out to be non-reliable, β_{epi} is estimated to be close to 0 and the presence of a predicted epitope does not strongly affect the conclusions drawn from the sequence data.

However, it is important to consider that biologically relevant HAMs do not necessarily have to lie within or close to the boundary of an epitope. For instance, compensatory mutations can occur far away from the epitope they are associated with, as they might be the result of improved physical interactions with another amino acid in the folded, three-dimensional protein (Ruhl et al., 2011). Such compensatory mutations (Kelleher et al., 2001; Ruhl et al., 2011; Neumann-Haefelin et al., 2011; Schneidewind et al., 2008) can confer a strong selection advantage, e.g. by partially restoring replicative capacity that would otherwise be impaired by the exclusive presence of a certain HLA escape mutation.

We therefore also expect HAMs outside epitopes and one possible concern is that the model focuses too strongly on associations with substitutions that lie within the boundary of predicted epitopes.

Figure 5 shows posterior probabilities $P(\beta_{jk} > 0)$ for

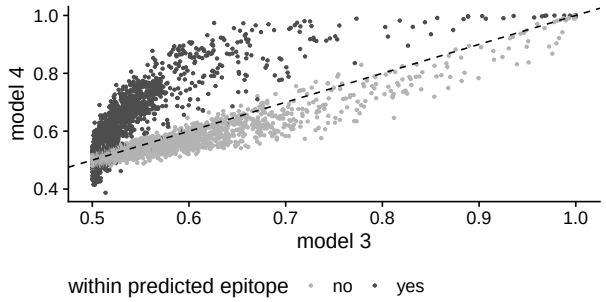

Figure 5. Integral of the marginal posterior $P(\beta_{jk} > 0)$ for the HAMdetector model with epitope prediction (model 4) and without epitope prediction (model 3) for all substitutions in the preC/core protein (HBV dataset).

substitution–HLA allele pairs as calculated by HAMdetector with (model 4) and without (model 3) epitope prediction. Each substitution–HLA allele pair is represented by a dot and colored according to whether or not that position lies within a predicted epitope. For substitutions that do not lie within a predicted epitope, both models provide similar estimates (points along the diagonal). However, some substitutions–HLA pairs that have only weak evidence of association in model 3 have strong support in model 4, which is explained by the additional evidence provided by epitope prediction. The figure shows that the model is still able to identify associations outside predicted epitopes and that epitope information augments evidence obtained from sequence data.

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 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 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

This work has been 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).

Acknowledgements

We thank Drs. Mina John and Simon Mallal for providing data.

References

Samuel Alizon, Fabio Luciani, and Roland R. Regoes. Epidemiological and clinical consequences of within-host evolution. *Trends in Microbiology*, 19(1):24–32, jan 2011. doi: 10.1016/j.tim.2010.09.005.

Todd M. Allen, Marcus Altfeld, Shaun C. Geer, Elizabeth T. Kalife, Corey Moore, Kristin M. O'Sullivan, Ivna DeSouza, Margaret E. Feeney, Robert L. Eldridge, Erica L. Maier, Daniel E. Kaufmann, Matthew P. Lahaie, Laura Reyor, Giancarlo Tanzi, Mary N. Johnston, Christian Brander, Rika Draenert, Jurgen K. Rockstroh, Heiko Jessen, Eric S. Rosenberg, Simon A. Mallal, and Bruce D. Walker. 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, nov 2005. doi: 10.1128/jvi.79.21.13239-13249.2005.

J. D. Altman, P. A. H. Moss, P. J. R. Goulder, D. H. Barouch, M. G. McHeyzer-Williams, J. I. Bell, A. J. McMichael, and M. M. Davis. Phenotypic analysis of antigen-specific T Lymphocytes. *Science*, 274(5284):94–96, oct 1996. doi: 10.1126/science.274.5284.94.

Valentin Amrhein and Sander Greenland. Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, 2(1):4–4, sep 2017. doi: 10.1038/s41562-017-0224-0.

M. Baker. 1500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016. doi: 10.1038/533452a.

C G Begley and L M Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–3, Mar 2012. doi: 10.1038/483531a.

Yoav Benjamini and Yosef Hochberg. 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, jan 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. 2017.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, jan 2017. doi: 10.1137/141000671.

T. Bhattacharya, M. Daniels, D. Heckerman, B. Foley, N. Frahm, C. Kadie, J. Carlson, K. Yusim, B. McMahon, B. Gaschen, S. Mallal, J. I. Mullins, D. C. Nickle, J. Herbeck, C. Rousseau, G. H. Learn, T. Miura, C. Brander, B. Walker, and B. Korber. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science*, 315(5818):1583–1586, mar 2007. doi: 10.1126/science.1131528.

Persephone Borrow, Hanna Lewicki, Xiping Wei, Marc S. Horwitz, Nancy Peffer, Heather Meyers, Jay A. Nelson, Jean Edouard Gairin, Beatrice H. Hahn, Michael B.A. Oldstone, and George M. Shaw. 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, feb 1997. doi: 10.1038/nm0297-205.

C. Bronke, C. M. Almeida, E. McKinnon, S. G. Roberts, N. M. Keane, A. Chopra, J. M. Carlson, D. Heckerman, S. Mallal, and M. John. HIV escape mutations occur preferentially at HLA-binding sites of CD8+ T-cell epitopes. *AIDS*, 27:899–905, 2013. doi: 10.1097/QAD.0b013e32835e1616.

Zabrina L Brumme, Chanson J Brumme, David Heckerman, Bette T Korber, Marcus Daniels, Jonathan Carlson, Carl Kadie, Tanmoy Bhattacharya, Celia Chui, James Szinger, Theresa Mo, Robert S Hogg, Julio S. G Montaner, Nicole Frahm, Christian Brander, Bruce D Walker, and P. Richard Harrigan. Evidence of differential HLA Class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. *PLoS Pathogens*, 3(7):e94, jul 2007. doi: 10.1371/journal.ppat.0030094.

Zabrina L Brumme, Iris Tao, Sharon Szeto, Chanson J Brumme, Jonathan M Carlson, Dennison Chan, Carl Kadie, Nicole Frahm, Christian Brander, Bruce Walker, David Heckerman, and P Richard Harrigan. Human leukocyte antigen-specific polymorphisms in HIV-1 gag and their association with viral load in chronic untreated infection. *AIDS*, 22(11):1277–1286, jul 2008. doi: 10.1097/qad.0b013e3283021a8c.

K. T. Brunner, J. Mauel, J. C. Cerottini, and B. Chapuis. 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, February 1968. ISSN 0019-2805.

Bettina Budeus, Jörg Timm, and Daniel Hoffmann. SeqFeatR for the discovery of feature-sequence associations. *PLOS ONE*, 11(1):e0146409, jan 2016. doi: 10.1371/journal.pone.0146409.

J. M. Carlson, J. Listgarten, N. Pfeifer, V. Tan, C. Kadie, B. D. Walker, T. Ndung'u, R. Shapiro, J. Frater, Z. L. Brumme, P. J. R. Goulder, and D. Heckerman. Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *Journal of Virology*, 86(9):5230–5243, feb 2012. doi: 10.1128/jvi.06728-11.

Jonathan M. Carlson, Zabrina L. Brumme, Christine M. Rousseau, Chanson J. Brumme, Philippa Matthews, Carl Kadie, James I. Mullins, Bruce D. Walker, P. Richard Harrigan, Philip J. R. Goulder, and David Heckerman. Phylogenetic dependency networks: Inferring patterns of CTL escape and codon covariation in HIV-1 gag. *PLoS Computational Biology*, 4(11):e1000225, nov 2008. doi: 10.1371/journal.pcbi.1000225.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, apr 2010. doi: 10.1093/biomet/asq017.

Cecil C. Czerkinsky, Lars-Åke Nilsson, Håkan Nygren, Örjan Ouchterlony, and Andrej Tarkowski. A solid-phase enzyme-linked immunospot (ELISPOT) assay for enumeration of specific antibody-secreting cells. *Journal of Immunological Methods*, 65(1-2):109–121, dec 1983. doi: 10.1016/0022-1759(83)90308-3.

R. A. Fisher. On the interpretation of X^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87, jan 1922. doi: 10.2307/2340521.

U. Francke and M. A. Pellegrino. 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, mar 1977. doi: 10.1073/pnas.74.3.1147.

Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, jan 2019. doi: 10.1111/rssa.12378.

Sylvie Le Gall, Pamela Stamegna, and Bruce D. Walker. Portable flanking sequences modulate CTL epitope processing. *Journal of Clinical Investigation*, 117(11):3563–3575, nov 2007. doi: 10.1172/jci32047.

Andrew Gelman. Bayesian Statistics then and now. *Statistical Science*, 25(2):162–165, may 2010. doi: 10.1214/10-sts308b.

Andrew Gelman and John Carlin. Beyond power calculations. *Perspectives on Psychological Science*, 9(6):641–651, nov 2014. doi: 10.1177/1745691614551642.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, nov 1992. doi: 10.1214/ss/1177011136.

Andrew Gelman, Jennifer Hill, and Masanao Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012. doi: 10.1080/19345747.2011.618213.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Ronald N. Germain. MHC-dependent antigen processing and peptide presentation: Providing ligands for T lymphocyte activation. *Cell*, 76(2):287–299, jan 1994. doi: 10.1016/0092-8674(94)90336-0.

Alfred L Goldberg, Paolo Cascio, Tomo Saric, and Kenneth L Rock. The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides. *Molecular Immunology*, 39(3-4):147–164, oct 2002. doi: 10.1016/s0161-5890(02)00098-6.

David W Haas, Grant R Wilkinson, Daniel R Kuritzkes, Douglas D Richman, Janet Nicotera, Laura F Mahon, Cara Sutcliffe, Sue Siminski, Janet Andersen, Kristine Coughlin, et al. A multi-investigator/institutional DNA bank for AIDS-related human genetic studies: AACTG Protocol A5128. *HIV clinical trials*, 4(5):287–300, 2003.

John T. Harty, Amy R. Tvinnereim, and Douglas W. White. CD8+ T Cell effector mechanisms in resistance to infection. *Annual Review of Immunology*, 18(1):275–308, apr 2000. doi: 10.1146/annurev.immunol.18.1.275.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, aug 2005. doi: 10.1371/journal.pmed.0020124.

Anthony R. Ives and Theodore Garland. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology*, 59(1):9–26, nov 2009. doi: 10.1093/sysbio/syp074.

Anthony R. Ives and Theodore Garland. Phylogenetic regression for binary dependent variables. In *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, pages 231–261. Springer Berlin Heidelberg, 2014. doi: 10.1007/978-3-662-43550-2_9.

M John, D Heckerman, L Park, S Gaudieri, and A Chopra. 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), 2008.

Hadi Karimzadeh, Muthamia M. Kiraithe, Anna D. Kosinska, Manuel Glaser, Melanie Fiedler, Valerie Oberhardt, Elahe Salimi Alizei, Maike Hofmann, Juk Yee Mok, Melanie Nguyen, Wim J. E. van Esch, Bettina Budeus, Jan Grabowski, Maria Homs, Antonella Olivero, Hossein Keyvani, Francisco Rodríguez-Frías, David Tabernero, Maria Buti, Andreas Heinold, Seyed Moayed Alavian, Tanja Bauer, Julian Schulze zur Wiesch, Bijan Raziorrouh, Daniel Hoffmann, Antonina Smedile, Mario Rizzetto, Heiner Wedemeyer, Jörg Timm, Iris Antes, Christoph Neumann-Haefelin, Ulrike Protzer, and Michael Roggendorf. 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, apr 2018. doi: 10.1128/jvi.01891-17.

Hadi Karimzadeh, Muthamia M. Kiraithe, Valerie Oberhardt, Elahe Salimi Alizei, Jan Bockmann, Julian Schulze zur Wiesch, Bettina Budeus, Daniel Hoffmann, Heiner Wedemeyer, Markus Cornberg, Adalbert Krawczyk, Jassin Rashidi-Alavijeh, Francisco Rodríguez-Frías, Rosario Casillas, Maria Buti, Antonina Smedile, Seyed Moayed Alavian, Andreas Heinold, Florian Emmerich, Marcus Panning, Emma Gostick, David A. Price, Jörg Timm, Maike Hofmann, Bijan Raziorrouh, Robert Thimme, Ulrike Protzer, Michael Roggendorf, and Christoph Neumann-Haefelin. 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, may 2019. doi: 10.1053/j.gastro.2019.02.003.

Yuka Kawashima, Katja Pfafferott, John Frater, Philippa Matthews, Rebecca Payne, Marylyn Addo, Hiroyuki Gatanaga, Mamoru Fujiwara, Atsuko Hachiya, Hirokazu Koizumi, Nozomi Kuse, Shinichi Oka, Anna Duda, Andrew Prendergast, Hayley Crawford, Alasdair Leslie, Zabrina Brumme, Chanson Brumme, Todd Allen, Christian Brander, Richard Kaslow, James Tang, Eric Hunter, Susan Allen, Joseph Mulenga, Songee Branch, Tim Roach, Mina John, Simon Mallal, Anthony Ogwu, Roger Shapiro, Julia G. Prado, Sarah Fidler, Jonathan Weber, Oliver G. Pybus, Paul Klenerman, Thumbi Ndung'u, Rodney Phillips, David Heckerman, P. Richard Harrigan, Bruce D. Walker, Masafumi Takiguchi, and Philip Goulder. Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*, 458(7238):641–645, feb 2009. doi: 10.1038/nature07746.

Anthony D. Kelleher, Chad Long, Edward C. Holmes, Rachel L. Allen, Jamie Wilson, Christopher Conlon, Cassy Workman, Sunil Shaunak, Kara Olson, Philip Goulder, Christian Brander, Graham Ogg, John S. Sullivan, Wayne Dyer, Ian Jones, Andrew J. McMichael, Sarah Rowland-Jones, and Rodney E. Phillips. Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27–restricted cytotoxic T lymphocyte responses. *Journal of Experimental Medicine*, 193(3):375–386, feb 2001. doi: 10.1084/jem.193.3.375.

Henrik N. Kløverpris, Alasdair Leslie, and Philip Goulder. Role of HLA adaptation in HIV evolution. *Frontiers in Immunology*, 6, jan 2016. doi: 10.3389/fimmu.2015.00665.

Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455, may 2019. doi: 10.1093/bioinformatics/btz305.

Laurie Lamoreaux, Mario Roederer, and Richard Koup. Intracellular cytokine optimization and standard operating procedure. *Nature Protocols*, 1(3):1507–1516, aug 2006. doi: 10.1038/nprot.2006.268.

Sheila F. Lumley, Anna L. McNaughton, Paul Klenerman, Katrina A. Lythgoe, and Philippa C. Matthews. Hepatitis B virus adaptation to the CD8+ T Cell response: Consequences for host and pathogen. *Frontiers in Immunology*, 9, jul 2018. doi: 10.3389/fimmu.2018.01561.

Philippa C. Matthews, Andrew Prendergast, Alasdair Leslie, Hayley Crawford, Rebecca Payne, Christine Rousseau, Morgane Rolland, Isobella Honeyborne, Jonathan Carlson, Carl Kadie, Christian Brander, Karen Bishop, Nonkululeko Mlotshwa, James I. Mullins, Hoosen Coovadia, Thumbi Ndung'u, Bruce D. Walker, David Heckerman, and Philip J. R. Goulder. Central role of reverting mutations in HLA associations with human immunodeficiency virus set point. *Journal of Virology*, 82(17):8548–8559, jul 2008. doi: 10.1128/jvi.00580-08.

S. Mei, F. Li, A. Leier, T. T. Marquez-Lago, K. Giam, N. P. Croft, T. Akutsu, A. I. Smith, J. Li, J. Rossjohn, A. W. Purcell, and J. Song. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief. Bioinformatics*, 21:1119–1135, 2020. doi: 10.1093/bib/bbz051.

Anita Milicic, David A. Price, Peter Zimbwa, Bruce L. Booth, Helen L. Brown, Philippa J. Easterbrook, Kara Olsen, Nicola Robinson, Uzi Gileadi, Andrew K. Sewell, Vincenzo Cerundolo, and Rodney E. Phillips. CD8+ T Cell epitope-flanking mutations disrupt proteasomal processing of HIV-1 nef. *The Journal of Immunology*, 175(7):4618–4626, sep 2005. doi: 10.4049/jimmunol.175.7.4618.

C. B. Moore. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science*, 296(5572):1439–1443, may 2002. doi: 10.1126/science.1069660.

S. Murata, K. Sasaki, T. Kishimoto, S. i. Niwa, H. Hayashi, Y. Takahama, and K. Tanaka. Regulation of CD8+ T cell development by thymus-specific proteasomes. *Science*, 316(5829):1349–1353, jun 2007. doi: 10.1126/science.1141915.

Christoph Neumann-Haefelin, Cesar Oniangue-Ndza, Thomas Kuntzen, Julia Schmidt, Katja Nitschke, John Sidney, Célia Caillet-Saguy, Marco Binder, Nadine Kersting, Michael W. Kemper, Karen A. Power, Susan Ingber, Laura L. Reyor, Kelsey Hills-Evans, Arthur Y. Kim, Georg M. Lauer, Volker Lohmann, Alessandro Sette, Matthew R. Henn, Stéphane Bressanelli, Robert Thimme, and Todd M. Allen. Human leukocyte antigen B27 selects for rare escape mutations that significantly impair hepatitis C virus replication and require compensatory mutations. *Hepatology*, 54(4):1157–1166, sep 2011. doi: 10.1002/hep.24541.

Timothy J. O'Donnell, Alex Rubinsteyn, and Uri Laserson. MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Systems*, 11(1):42–48.e7, jul 2020. doi: 10.1016/j.cels.2020.06.010.

Jason W. Osborne and Elaine Waters. Four assumptions of multiple regression that researchers should always test. 2002. doi: 10.7275/R222-HV23.

Juho Piironen and Aki Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. 2017a.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017b. doi: 10.1214/17-ejs1337si.

James Robinson, Jason A. Halliwell, James D. Hayhurst, Paul Flicek, Peter Parham, and Steven G. E. Marsh. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research*, 43(D1):D423–D431, nov 2014. doi: 10.1093/nar/gku1161.

Kirsten Roomp, Niko Beerenwinkel, Tobias Sing, Eugen Schülter, Joachim Büch, Saleta Sierra-Aragon, Martin Däumer, Daniel Hoffmann, Rolf Kaiser, Thomas Lengauer, and Joachim Selbig. Arevir: A secure platform for designing personalized antiretroviral therapies against HIV. In *Lecture Notes in Computer Science*, pages 185–194. Springer Berlin Heidelberg, 2006. doi: 10.1007/11799511_16.

Christine M. Rousseau, Marcus G. Daniels, Jonathan M. Carlson, Carl Kadie, Hayley Crawford, Andrew Prendergast, Philippa Matthews, Rebecca Payne, Morgane Rolland, Dana N. Raugi, Brandon S. Maust, Gerald H. Learn, David C. Nickle, Hoosen Coovadia, Thumbi Ndung'u, Nicole Frahm, Christian Brander, Bruce D. Walker, Philip J. R. Goulder, Tanmoy Bhattacharya, David E. Heckerman, Bette T. Korber, and James I. Mullins. 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, apr 2008. doi: 10.1128/jvi.02455-07.

M. Ruhl, P. Chhatwal, H. Strathmann, T. Kuntzen, D. Bankwitz, K. Skibbe, A. Walker, F. M. Heinemann, P. A. Horn, T. M. Allen, D. Hoffmann, T. Pietschmann, and J. Timm. 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, nov 2011. doi: 10.1128/jvi.05603-11.

Stephen M. Scariano and James M. Davenport. The effects of violations of independence assumptions in the one-way ANOVA. *The American Statistician*, 41(2):123–129, may 1987. doi: 10.1080/00031305.1987.10475459.

Arne Schneidewind, Mark A. Brockman, John Sidney, Yaoyu E. Wang, Huabiao Chen, Todd J. Suscovich, Bin Li, Rahma I. Adam, Rachel L. Allgaier, Bianca R. Mothé, Thomas Kuntzen, Cesar Oniangue-Ndza, Alicja Trocha, Xu G. Yu, Christian Brander, Alessandro Sette, Bruce D. Walker, and Todd M. Allen. Structural and functional constraints limit options for cytotoxic t-lymphocyte escape in the immunodominant HLA-b27-restricted epitope in human immunodeficiency virus type 1 capsid. *Journal of Virology*, 82(11):5594–5605, jun 2008. doi: 10.1128/jvi.02356-07.

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

Jörg Timm and Andreas Walker, 2021. GenBank Accession Numbers: Genotype A: MZ043025 - MZ043097, Genotype B: MW845286 - MW845312, Genotype C: MW887641 - MW887652, Genotype D: MZ097624 - MZ097884, Genotype E: MW926548 - MW926566.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5): 1413–1432, aug 2016. doi: 10.1007/s11222-016-9696-4.

R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, and B. Peters. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.*, 47:D339–D343, 2019. doi: 10.1093/nar/gky1006.

Jonathan W. Yewdell and Ann B. Hill. Viral interference with antigen presentation. *Nature Immunology*, 3(11):1019–1025, nov 2002. doi: 10.1038/ni1102-1019.

Karina Yusim, Elizabeth-Sharon David-Fung, Bette T. M. Korber, Christian Brander, Dan Barouch, Rob de Boer, Barton F. Haynes, Richard Koup, John P. Moore, Bruce D. Walker, and David I. Watkins, editors. *HIV Molecular Immunology*. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico, 2018.