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

Daniel Habermann[1], ..., Daniel Hoffmann[1]

Abstract

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

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

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

Availability and Implementation
The source code of this software is available at http://github.com under a permissive MIT license.

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

Keywords
human leukocyte antigen system, HLA, multiple sequence alignment, escape mutations, viral escape, Bayesian inference, sparsity, horseshoe prior, epitope prediction

[1]*Bioinformatics & Computational Biophysics, Faculty of Biology, University of Duisburg-Essen, 45117 Essen, Germany*

Contents

1. Introduction

1.1 The HLA system

One way of how the human immune system recognizes viral infections is through the human leukocyte antigen system [1]: Proteins, including viral proteins, in cells are often degraded in proteasomes to peptide fragments [2]. A small subset of these peptides is presented on the cell surface by MHC I receptors, encoded by HLA genes (in humans HLA-A, -B, -C). HLA genes are very polymorphic with more than 20,000 known alleles in the human population [3]. The resulting gene products differ in their binding properties, which means that cells from different individuals in general present different peptides on the cell surface.

Cytotoxic T-Lymphocytes (CTLs) are selected during maturation to only weakly bind to complexes of MHC I with self-peptides, but CTLs can bind strongly to complexes of MHC I with peptides from viral proteins [4]. Upon activation, CTLs become cytotoxic and recruit other immune cells so that infected cells can be eliminated [5].

1.2 HLA escape

In this way, the HLA system exerts strong selection pressure towards virus variants that escape T cell recognition [6]. Such variants could, for example, carry mutations that result in reduced binding of viral peptides to MHC I or T cell receptors, or that alter peptide processing so that peptides are no longer presented [7].

HLA diversity drives viral evolution in individuals where a virus adapts to specific host features, and in populations and regions that in general differ in HLA frequencies [8]. Upon transmission to a new host with different HLA alleles, HLA escape mutations may revert, as they are often associated with a reduction in viral replication capacity [9], but they also may be irreversible [8].

Whether and how quickly a given escape mutation is selected in a host depends, e.g., on the viral genomic background, the magnitude of the reduction in viral replication, the availability of compensatory mutations that recover fitness, or the strength of selection pressure [10].

Immune escape is a driver of viral evolution in populations and patients, and therefore of eminent importance for the understanding of many aspects of highly variable viruses such as HIV or HBV [11, 12, 13, 14]. This includes the development of treatments and vaccines that rely on effective immune responses. This makes detection of immune escape mutations critical.

1.3 Identifying HLA escape mutations

There are several experimental methods available to study HLA escape [15]. DaHo: use key references, e.g. those that have introduced a technique; alternatively: cite review like articles that give an overview of several experimental methods 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 [16].

As the selection pressure exerted by cytotoxic T cells depends on successful recognition of viral peptides on the cell surface and thus on binding to the HLA encoded MHC I protein, escape mutations are usually HLA allele specific and can therefore be detected as HLA allele dependent "footprints" in sequence alignments of viral proteins [17], i.e. as HLA associated mutations (HAM) or replacements enriched in sequences from hosts with a specific HLA allele.

One way of quantifying this enrichment is Fisher's exact test [18]: For a given replacement R_i at alignment position i and HLA allele H, a 2-by-2 contingency table is constructed containing the absolute counts of the number of sequences in the four possible categories (R_i, H), $(R_i, \neg H)$, $(\neg R_i, H)$ and $(\neg R_i, \neg H)$, where $\neg R_i$ denotes any replacement except R_i, and $\neg H$ denotes any HLA allele except H.

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

Fisher's exact test has the advantage of being fast and easy to apply [19], but it also has several disadvantages [20]. 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.

Another issue with Fisher's exact test is the genomic proximity of human HLA class I loci (on chromosome 6) leading to linkage disequilibrium – inheritance of HLA alleles can be correlated. Therefore spurious associations can occur if associations of replacements with individual HLA alleles is tested: if HLA allele H_1 is associated with an amino acid replacement R because of immune escape, but H_1 is in linkage disequilibrium with allele H_2, this leads to an association of R and H_2, even without being an escape mutation from H_2.

Carlson et al. [20] 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 using p-values as a screening tool
There are fundamental statistical issues with p-values as screening tool [21]: with small effect sizes and high variance between measurements, as is often the case with biological data, statistically significant results can be misleading, have the wrong direction (type S error), or greatly overestimate an effect (type M error) [22]. Such problems are more and more appreciated in the context of the current "replication crisis", which describes that scientific claims with seemingly strong statistical evidence fail to replicate [23, 24, 25].

These problems are exacerbated if the p-values are used for screening purposes (multiple testing problem). The probability of obtaining a statistically significant result increases with each additional test, even in absence of any real effect. When using p-values as a filter, it is therefore likely to obtain significant effects that are in fact not real. To circumvent this problem, a common strategy is to control the false discovery rate [26]. These adjustment procedures have the problem that, when performing many of such comparisons, none but the very largest effects remain.

Instead of performing many hypothesis tests and trying

to adjust for them, we prefer to fit a single, multilevel model that contains all comparisons of interest. When using multilevel models, the problem of multiple comparisons can disappear entirely and yield more valid estimates [27].

2. Material and methods

DaHo: Da es ein Methoden-Paper ist, sollte dies der Hauptteil sein

Our general approach is to fit a Bayesian regression model that captures relationships between host HLA alleles and replacements in viral proteomes. This Bayesian approach is advantageous because it allows use of: (1) prior information (even if only knowledge of effect magnitudes), (2) relevant additional information (phylogeny, epitope information), (3) a problem-specific structure, (4) partial pooling [28]. With these features, HAM detection is improved markedly over the state-of-the-art.

2.1 Model backbone
We chose a logistic regression model as a backbone for two reasons: First, because it is easily extensible and secondly, because the coefficients are interpretable as expected increases on the log-odds scale, identical to the familiar interpretation of standard logistic regression models.

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

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

where y_{ik} is the binary encoded observation at sequence i for replacement k, θ_{ik} is the estimated probability that we observe replacement k in sequence i, X_{ij} is 1 if sequence i is annotated with HLA allele j and 0 otherwise, $\beta_{\text{HLA}_{jk}}$ is the HLA regression coefficients for HLA allele j at replacement k, D is the number of HLA alleles in the dataset, and inv_logit is the logistic inverse link function. Note that, if we observe N different amino acid states at a certain alignment position, each of those are considered in turn to be the "replacement".

The regression coefficients $\beta_{\text{HLA}_{jk}}$ are the main focus of interest, as they quantify the association between the occurrence of replacement k and each of the observed HLA alleles. They are in units of log-odds, so a regression coefficient of somewhere around 2 means that when

comparing sequences with and without that particular HLA allele, the log-odds of observing replacement k in presence of that HLA allele is increased by 2.

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 about 2/4 = 50%.

2.2 Inclusion of additional information

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

2.2.1 Phylogeny

One problem that occurs when analyzing sequence data is that species share a common phylogenetic history. Standard statistical methods usually assume samples to be independent and identically distributed, which may lead to wrong conclusions when this assumption is strongly violated. In the context of identifying HLA associations, Bhattacharya et al. [36] demonstrated the importance of correcting for the phylogenetic structure. There are many approaches described in the literature for phylogenetic regression for binary dependent variables [37], where the most popular approach is to estimate an additional multivariate normally distributed intercept, where the covariance matrix is based on the branch lengths of a given phylogenetic tree [38].

This approach showed to be too computationally expensive to include in our model, so we chose a strategy that is similar to the one in Carlson et al. [20] for the Phylogenetic Dependency Network:

Consider a phylogenetic tree Ψ obtained from standard maximum likelihood methods for a given multiple sequence alignment. We are interested in estimating $P(y_{ik} = 1|\Psi)$, that is, the probability of observing the replacement k in sequence i based on the underlying phylogenetic model. A quantity that can be readily computed using phylogenetic software like RAxML-NG [39] is $P(\Psi|y_{ik} = 1)$. For this, we keep the tree topology fixed, annotate the tree with the binary observations y at its leaves and optimize the branch lengths. $P(\Psi|y_{ik} = 1)$ is then the likelihood of the annotated phylogenetic tree. We

can then use Bayes' rule to invert that conditional probability by additionally computing $P(\Psi|y_{ik} = 0)$, which is done by altering the annotation for sequence i and keeping all other observations constant.

In order to check if the inferred probabilities $P(y_{ik} = 1|\Psi)$ faithfully reflect the observed data we use calibration plots (see fig. 1). All observations are binned according to their inferred probability $P(y_{ik} = 1|\Psi)$. If these probabilities are calibrated correctly, we expect that in a group of observations with an inferred probability $P(y_{ik} = 1|\Psi)$ of around x% we really do observe $y_{ik} = 1$ around x% percent of the time.
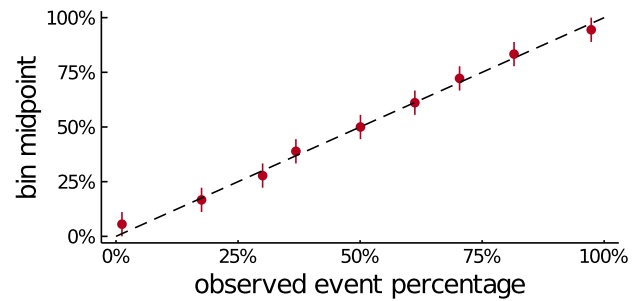

Figure 1. Calibration plot for the HBV dataset. For a description of all datasets used in this study see section 4. Calibration plots for the other datasets are shown in the supplementary. All observations are first sorted by increasing estimated probability $P(y_{ik} = 1|\Psi)$ and then grouped into n bins. For each bin, the fraction of observations with $y_{ik} = 1$ (observed event percentage) is compared to the midpoint of each bin (the value in the center between the lowest and highest probability). The error bars show the cutpoints for each bin. If the probabilities are calibrated correctly, each dot is supposed to scatter closely around the diagonal line.

The estimated probabilities are then included in the model as additional intercepts:

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

The logit transform is used because is cancels out with the logistic link function. This means that the phylogeny information acts as a baseline in absence of any HLA

effects. Note that we also include an additional parameter γ, that is constrained to be positive. This helps because the inferred probabilities $P(y_{ik} = 1|\Psi)$ might not necessarily reflect the true underlying phylogenetic signal, for example because the phylogenetic tree does not match the observed data well enough.

A straight-forward extension of our model would be to also account for these sources of uncertainty, for example by using a Bayesian method to estimate a posterior distribution over possible tree topologies. The uncertainty over the tree topology and the underlying parameters of the phylogenetic model would then propagate into uncertainty of the estimated probabilities $P(y_{ik} = 1|\Psi)$. However, in order to not increase the runtime of the model further we use standard maximum likelihood estimates and then include these in a way that additionally allows for measurement error.

Binding affinity prediction

DaHo: Mein Verständnis von Materials and Methods ist, dass hier die Methoden aufgeführt werden, z.B. die Methoden für die Berechnung von Affinitäten, siehe z.B. https://academic.oup.com/bioinformatics/article/36/21/5159/5874440

HAMs are expected to lie more frequently in T-cell epitopes, i.e. the peptide fragments presented by MHC to T-cell receptors. There are vast experimental binding affinity data available of different peptides and MHC I molecule pairs, and there are well-established computational methods that use these data to extrapolate HLA binding for untested peptides. We show that, by including the outcome of these computational tools as input for a probabilistic model, the prediction of HLA-associated mutations can be improved.

Antigen processing prediction

Similarly, there are also HLA-allele independent effects like antigen processing that influence presentation on MHC I. Second generational tools like MHCFlurry 2.0 [29] use both binding affinity and antigen processing data to improve epitope prediction. For our tool HAMdetector, we use the output of MHCFlurry to benefit from this data in order to predict HLA-associated mutations.

2.2.2 Epitope predictions
There is vast experimental data on epitope binding motifs available, e.g. through the use of mass spectrometry [41]. Epitope prediction software like MHCFlurry [29] use these data to extrapolate MHC I binding for new, untested

peptides. As HLA-associated mutations are much more likely to lie in epitopes (e.g. because of selection pressure to escape MHC I binding), using epitope prediction software can provide valuable external data for identifying HLA-associated mutations.

For our model, we use MHCFlurry 2.0 for binding prediction, as it does not only predict HLA binding but also antigen processing. For this, we create an input matrix of dimensions $R \times D$, where R is the number of evaluated replacements and D is the number of observed HLA alleles in the dataset. The elements of this matrix are binary encoded and contain a 1 if that position is either expected to be inside a predicted epitope or related to antigen processing, and 0 otherwise. Given an amino acid sequence, MHCFlurry 2.0 provides a list of possible epitopes (9-13 mers) and HLA allele pairs and calculates a rank based on comparisons with random epitope and HLA pairs. For the binarization we use a rank threshold of 0.2% (as suggested by MHCFlurry).

We typically try to avoid thresholds whenever possible, but in this case we deemed it to be preferable to including the predicted binding affinities, as we expected the binary values to be more reliable than the predicted numeric values. Note that the output of these epitope prediction tools may be imprecise or even wrong, but by including it as an input for a probabilistic model, the model is able to learn how much to trust this data if parameterized in a way to allow for measurement error.

We use epitope prediction as information about the expected degree of sparsity, i.e. when knowing a certain position is restricted by a certain HLA allele, we expect that this HLA allele is more likely to be associated with the replacement than the other HLA alleles. This can be reflected by increasing the scale of the local shrinkage parameters λ_{jk}:

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

where Z_{jk} is 1 if HLA allele j is predicted to restrict the alignment position corresponding to replacement k or is associated antigen processing 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.

2.3 Sparsity-inducing priors

Recent advantages in Bayesian inference include so-called sparsity-promoting priors [30]. Sparsity-promoting priors convey the a priori expectation that most coefficients in a regression model are close to 0. This assumption also applies to HLA-associated mutations, because the number of epitopes that are restricted by a given HLA allele is typically small compared to the number of all possible epitopes. If no epitope spanning a given alignment position is presented on the MHC I molecule, any association between a replacement and the respective HLA allele is likely to be due to random variation alone (or HLA linkage disequilibrium).

It has been shown DaHo: where? that using a sparsity-promoting prior when non-zero coefficients are sparse can drastically improve predictive performance, because the model is better able to differentiate between signal and noise.

We show that a simple logistic regression model with a sparsity-promoting prior on the regression coefficients for the HLA alleles alone already performs roughly on-par with the much more elaborate Phylogenetic Dependency Network approach from Carlson et al. when using the fraction of identified escapes that lie in known HLA epitopes as a benchmark.

In the context of identifying HLA associations we expect most of the HLA coefficients to be very close to 0, as a given alignment position is typically only restricted by none or few HLA alleles. This is important prior knowledge that should be incorporated in the model.

For Bayesian models, there are a variety of different sparsity-promoting priors with slightly different properties. They share the common structure of placing most probability mass very close to 0, with large to tails to accommodate the non-zero coefficients. For our model, we use a so-called horse prior [40], which is defined as a scale-mixture of gaussians:

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

, where β_{jk} are the regression coefficients, τ is the so-called global shrinkage parameter, λ_{jk} are the so-called local shrinkage parameters and Cauchy$^+$ is the positively constrained Cauchy distribution.

Intuitively speaking, the global shrinkage parameter τ is typically very small and shrinks most of the regression coefficients to 0, whereas the local shrinkage parameters can occasionally be very large to allow some coefficients to escape that shrinkage.

A useful addition to the horseshoe prior is given in Piironen et al. [30], which describes a way to choose the parameter governing the overall degree of sparsity τ_0 based on the expected number of non-zero coefficients.

For our HLA model, we utilize an extension to the horseshoe prior called the regularized horseshoe prior [30], which also provides some regularization for the 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 when fitting logistic regression models.

2.4 Full model specification

The full specification of the model is given by:

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

The full model specification includes some aspects that were not covered in the previous sections. In particular, the parameters γ_k are modelled hierarchically, which allows information to be partially pooled across replacements. However, the model would also work reasonably

well with just a single global parameter γ. All other additions ($\tilde{\lambda}_{jk}^2$ and τ_{0k}) are explained in detail in [30]. Briefly, the addition of $\tilde{\lambda}_{jk}^2$ allows some regularization for the non-zero coefficients and the parameterization of τ_{0k} allows to place a prior on the expected number of non-zero coefficients.

2.4.1 Prior justification
Prior distributions are labeled with an asterisk in equation 6. They are meant to be weakly informative, which means that the goal is to provide some modest amount of shrinkage while still allowing for large absolute values given enough support by the observed data. 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 are chosen to place most probability mass at values of γ_k of around 1. In the absence of any HLA effects, a parameter value of 1 would mean that the estimate for the probability of observing replacement k is identical to the probability based on the phylogenetic model. Intuitively speaking, We therefore treat the phylogeny as a baseline and any observations that cannot be attributed to phylogeny are attributed to the 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 $\beta_{\mathrm{HLA}_{jk}}$. A Student-t prior with these parameters is a reasonable default choice for logistic regression models. The theoretical justification behind this regularized variant of the horseshoe prior is explained in detail in [30].

The value of τ_{0k} is chosen to imply 10 effective non-zero HLA regression coefficients per replacement. The rationale behind this parameterization is again outlined in [30]. The value of 10 is approximately based on available HIV epitope maps. Please note that this is only an approximate statement about the general magnitude, i.e. we would expect maybe a couple of HLA alleles to be associated with a single replacement, not much more. 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.

2.4.2 Model implementation
All models were implemented in Stan [32]. The Stan code is available online in two versions: One optimized for readability and one optimized for speed by utilizing Stan's multithreading and GPU capabilities. A Julia [42] package is available <mark>DaHo: at ...</mark> to run the model on custom data. Due to restrictions of dependencies (MHCFlurry and RAxML-ng), HAMdetector is only available on Linux.

2.5 Model diagnostics
2.5.1 Posterior predictive checks
Additionally, Bayesian statistics provides an accessible way to test models: By comparing data generated under the model's assumptions to the actually observed data, it is possible to identify important aspects of the dataset that the models fails to capture and subsequently improve the model until it is consistent with the observed data [31].

Here, we use these so-called posterior predictive checks to compare the calibration of the predicted probabilities of observing the replacement to the actual observed data.

2.5.2 Convergence diagnostics
We use Stan (version 2.23) [32] to specify and fit our Bayesian models. Stan provides an implementation of a dynamic Hamilitonian Monte Carlo sampler, which is a powerful technique to approximate expectations with respect to the posterior distribution [33]. One advantage of using such methods is that they provide sensitive diagnostics to identify a wide range of possible sampling issues. We use the split-$\hat{R}$ convergence diagnostic [34] to identify convergence issues. We require a value of $\hat{R}$ below 1.01 for all model parameters. Additionally, we require that the effective sample size N_{eff} is above 500 for all model parameters and that sampling occurs without any divergent transitions. A detailed explanation of these quantities is provided in the Stan reference manual.

2.5.3 Leave-one-out cross-validation
In addition to testing our model on real-world data, we also performed leave-one-out cross-validation to examine the relative importance of the additions of epitope prediction, sparsity and phylogeny with respect to predictive performance. For this, we create several variations of the model that are all special cases of the full HAMdetector model. By comparing the model performance of these variations, we can gain insight on the relative importance of each aspect of the model.

Leave-one-out cross-validation has historically been chal-

lenging to apply for Bayesian models, because fitting a model once for each left out data-point is often infeasible because of computational constrains. Recently, Pareto-smoothed importance sampling has gained popularity to perform approximate leave-one-out cross-validation for Bayesian models. This method approximates the leave-one-out posterior by re-weighting existing draws from the posterior distribution and has the advantage of requiring to fit the model only once. Additionally, the method provides good diagnostics to identify cases where the approximation is unreliable [35].

Several utility (or cost) functions are used in practice to compare the predictive performance of models. For frequentist approaches, classification accuracy or receiver operator characteristic curves (ROC curves) are most common. For Bayesian models, expected log-predictive density (elpd) is used most frequently.

elpd is defined as $\sum_{i=1}^{n} \log(\int p(y_i|\theta)p(\theta|y_{i-1})d\theta)$, i.e. the average log predictive density of the observed data points based on the leave-one-out posterior distributions. This has the advantage over other performance measures like classification accuracy, that it not only takes the location of the predictive distribution into account (i.e. the number of correct predictions), but also the width (i.e. how confident the model is in its predictions). elpd cannot be computed for frequentist models because the parameters of the model are viewed as fixed values.

2.6 Data and data preparation

DaHo: das Material in Materials and Methods sind die Daten; die sollten hier kurz beschrieben werden, möglichst mit Referenzen

The model was fit to several data sets:

- A large HIV dataset consisting of 1383 sequences, that were kindly provided by Zabrina Brumme and collaborators and were also used in the Phylogenetic Dependency Network study [20].

- A set of 351 HIV sequences mostly spanning the Pol gene from the Arevir database [43].

- A set of 544 Hepatitis-B-Virus sequences that were kindly provided by Jörg Timm.

- A set of 104 Hepatitis-D-Virus sequences that were kindly provided by Michael Roggendorf. DaHo: bitte von Hadi geeignetes Zitat erfragen

- A set of 41 HIV sequences that were kindly provided by Rongge Yang.

DaHo: zu jedem Datensatz die Proteine nennen

For all sequences, we applied the following data preparation steps:

1. For each dataset, the sequences were split into subsequences, either by protein or gene.

2. If not already present in this format, sequences were translated into their amino acid representation.

3. RAxML-NG [39] 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. Implementation

4. Results

In order to evaluate each building block of HAMdetector, we applied 4 different Bayesian models to each dataset, each of which includes the previous model as a special case: A standard logistic regression model, a logistic regression model with horseshoe prior, a model additionally including phylogeny and the full model with horseshoe prior, phylogeny and epitope prediction. As a comparison to existing methods, we also applied Fisher's exact test and the Phylogenetic Dependency Network.

Convergence diagnostics

All model fits showed no signs of inference issues. In total, samples were drawn from 4 different chains with 1000 iterations each. The effective sample size exceeded 500 for all model parameters, $\hat{R}$ convergence diagnostic values were below 1.01 in all cases. DaHo: citation Gelman + Rubin original paper or later versions if improved $\hat{R}$ have been used

Posterior predictive checks

Figure 2 compares the calibration of the posterior predictive probability of observing the replacement to the actual observed data for the Core protein of the HBV dataset. Plots for the other models and datasets are shown in the supplementary.

The plots are similar to the one showed in figure 1, however, they do not evaluate the probabilities based on the phylogenetic model alone, but instead show the calibration of the posterior predictive probabilities based on the full model.
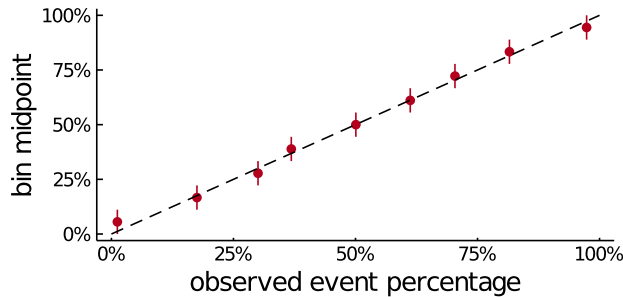

Figure 2. Calibration plot for the HBV dataset. Calibration plots for the other datasets are shown in the supplementary. All observations are first sorted by increasing estimated probability $P(y_{ik} = 1)$ and then grouped into n bins. For each bin, the fraction of observations with $y_{ik} = 1$ (observed event percentage) is compared to the midpoint of each bin (the value in the center between the lowest and highest probability). The error bars show the cutpoints for each bin. If the probabilities are calibrated correctly, each dot is supposed to scatter closely around the diagonal line.

Comparison to a list of known epitopes

When testing models, we prefer to evaluate them on real-world data whenever possible. However, comparing against real-world data is not straight forward when identifying HLA-associated mutations, because if a tool identifies an HLA-associated mutation which is not yet documented, this could be either because it is a false positive, or a yet unknown HLA-associated mutation that has not been documented before. Also, viewing every non-identified HLA-associated mutation that is listed in the literature as a false negative is not correct either, because documented escape mutations do not necessarily have to show in each sequence alignment. A third issue when comparing models is that HLA-associated mutations are not necessarily binary, so the true positive / false negative framework does not work as well.

To circumvent these issues and still compare against real-world data, we chose the following strategy: For each model, all evaluated replacements are ranked by decreasing confidence of being an HLA associated mutation. That is, for Bayesian models we calculate the probability of the corresponding regression coefficient exceeding

zero $P(\beta_{HLA} > 0)$, for Fisher's exact test we sort HLA allele - replacement pairs by increasing p-values and for the Phylogenetic Dependency Network, we rank the associations by increasing q-values. Then, a list of known epitopes is used to create a plot of the cumulative number of replacements inside the boundary of a known epitope vs. the corresponding rank. The underlying assumption for this kind of model check is that we expect to see an enrichment of replacements that lie inside the region of known epitopes and that this enrichment is stronger for better performing models.
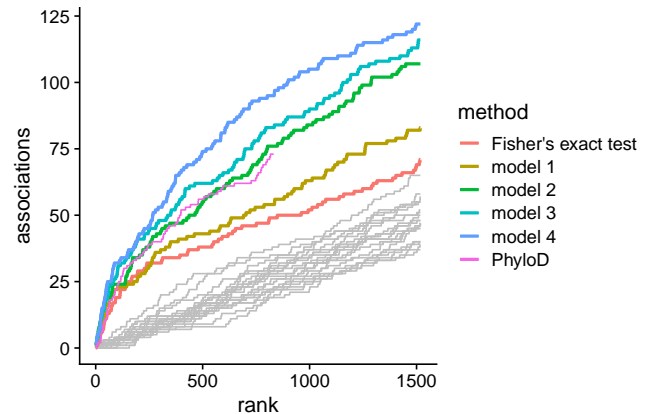

Figure 3. Number of associations inside the boundary of known epitopes vs. rank. For each model, the pairs of HLA alleles and replacements are sorted by decreasing confidence of being an HLA associated mutation. Then, each replacement is compared to a list of known epitopes and a line of the cumulative number of associations inside known epitopes vs. the current rank is plotted for each model. Better performing models are expected to have an enrichment of associations that are inside of known epitopes. PhyloD: Phylogenetic Dependency Network, Model 1: simple logistic regression model with broad Student-t priors, Model 2: logistic regression model with horseshoe prior, Model 3: logistic regression model with horseshoe prior and phylogeny, Model 4: Full model with epitope prediction. The thin gray lines denote random permutations of the list of HLA allele - replacement pairs and act as a baseline.

Figure 3 shows a comparison of different models on the HBe protein of the HBV dataset. Plots for all other datasets are shown in the supplementary, but show identical patterns.

All models show a strong enrichment of HLA-associated mutations compared to a randomly permuted list of HLA allele - replacement pairs (gray lines). Fisher's exact test

performs about as well as the simple logistic regression model, as from a statistical perspective, they are both similar. The horseshoe prior alone is a drastic improvement over Fisher's exact test and the logistic regression model with non-sparsifying priors, even though it does not include any external information. The benefit comes from including the assumption that most regression coefficients are expected to be very close to 0. The logistic regression model with horseshoe prior works roughly as well as the Phylogenetic Dependency Network, which includes much more information. Note that the line for the Phylogenetic Dependency Network is shorter because it only outputs associations with a q-value lower than 0.2 by default.

The remaining models are ranked in order of the number of additional features they include. Additionally including phylogeny is a small improvement over the model with the horseshoe prior, while adding epitope prediction provides another strong improvement.

Note that this model only uses epitope *prediction* software and does not use any information of experimentally confirmed epitopes. These are here only used for model evaluation.

Figure 4 shows a comparison of marginal posteriors for regression coefficients with and without horseshoe prior.

Supplying information about the expected degree of sparsity with the horseshoe prior is a considerable improvement over a standard logistic regression model (section 4).
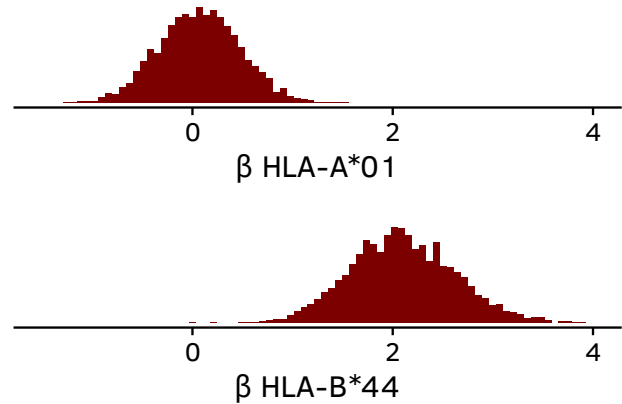
4.1 Leave-one-out cross-validation
Table 1 shows a comparison of models with increasing complexity run on the HBe protein in the HDV dataset.

Predictive accuracy increases as the models include more and more information. Note that 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 additional 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 the estimated y.

Including phylogeny further improves model performance, as the assumption of independent and identically distributed data does not hold for sequence data that share a common phylogenetic history.
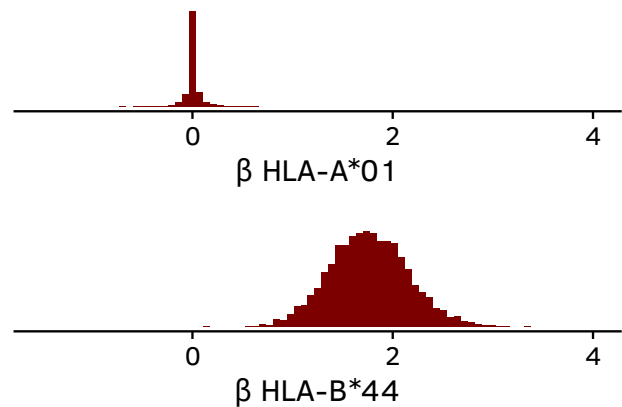

Figure 4. Marginal posterior distributions of a subset of regression coefficients for replacement 11D of the HIV integrase, based on the Arevir dataset. The graphic shows regression coefficients for two HLA alleles: For HLA-A*01, the data does not show evidence of an association with replacement 11D, whereas for HLA-B*44, the data is consistent with a strong association. The horseshoe prior has the effect of shrinking regression coefficients with weak evidence of an association with the replacement strongly towards 0, while leaving the coefficients that are far away from 0 mostly unshrunk. Shrinking the irrelevant regression coefficients to 0 has the advantage of reducing the standard error for the remaining coefficients, which can be observed by the much smaller width of the histogram of HLA-B*44 in the model with horseshoe prior.

Table 1. Differences in elpd for models with increasing model complexity (HBe protein of the HBV dataset). The numbers in column elpd$_{\text{diff}}$ show increases in elpd compared to the previous model (higher is better), e.g. the difference in elpd of the model with horseshoe prior is 949.8 compared to the baseline model, and the difference in elpd for the model with phylogeny and horseshoe prior is 4440.9 compared to the model just with the horseshoe prior. All differences in elpd are larger than several times the estimated standard error (column se$_{\text{diff}}$), indicating that models which account for more information are expected to have a better predictive performance.

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

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

4.2 Generalization
<mark>DaHo: Es gibt fast keinen Text in dieser subsection</mark>

The models were applied on sequence data from a variety of different viruses and show a considerable improvement over currently applied methods that are consistent across datasets. These results are not due to overfitting, because leave-one-out cross-validation approximates the model performance on unseen data.

4.3 Example Hepatitis Delta Virus
We had the opportunity to test our model against a set of associations identified by Fisher's exact test where some of them turned out to be false positives [44]. This data is rare because negative results are usually not published in the literature, but provide a useful way to access model performance. A comparison with a preliminary

version of HAMdetector was published recently. Here we provide a comparison with results from Fisher's exact test [19] and the mature version of HAMdetector used in this work.

Fisher's exact test and HAMdetector were applied to the same data (table 2). The posterior probabilities given by HAMdetector denote the probability that the corresponding regression coefficient is positive. Associations with strong support in our model 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 results show low posterior probabilities for most of the false-positives and high posterior probabilities for most replacements that could be experimentally validated in IFN-γ production assays. One important exception is P89T - HLA-B*37, which was experimentally validated but does not have strong support in our model.

Table 2. List of HLA associated mutations that were identified using Fisher's exact test (FET). Experimental validation showed that 9 out of 15 replacements could not be experimentally validated (column "confirmed"). The data were re-analyzed using HAMdetector. Except for P89T, all validated replacements have high posterior probabilities, while the replacements that could not be experimentally validated have lower evidence of being HLA associated. Note that all replacement - HLA allele pairs had low p-values using Fisher's exact test.

replacement	allele	p-value (FET)	post. prob. (HAMdetector)	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	0.91	+
E47D	B*18	0.0027	0.89	+
D33E	B*13	0.0001	0.86	-
T134A	A*68	0.0045	0.82	-
K43R	B*13	0.0021	0.77	-
D47E	A*30	0.0010	0.76	-
K113R	B*13	0.0043	0.76	-
P89T	B*37	0.0011	0.75	+
A107T	B*14	0.0028	0.70	-
P49L	A*30	0.0031	0.63	-
Q100L	B*13	0.0018	0.61	-
D96E	B*13	0.0035	0.50	-

4.4 Individual examples

In the following we illustrate with ... examples the effects of the inclusion of additional relevant information on the detection of HAMs.

DaHo: bitte Beispiel mit Linkage Disequilibrium zeigen

DaHo: bitte konkretes Bsp mit Auswirkung Phylogenie zeigen

5. Discussion

We have presented HAMdetector, a Bayesian model to identify HLA-associated mutations in a multiple sequence alignment. We propose to use models that include as much information as possible and have shown that including sparsity and epitope prediction software can achieve better performance than existing methods. The logistic regression backbone can be easily extended, which allows the model to also be used in other contexts, for example in identifying associations between sequence data and other features.

If the quality of the prediction is the same for all viruses, as it appears to be in our results, this method can be used to probe the frequency of HLA immune escape across different viruses.

In further research we plan to use HAMdetector to study circulating recombinant forms in HIV. One possible explanation for the occurrence of these recombinant forms is improved fitness by evading several immune epitopes at once.

While our model is an improvement over existing methods, a further addition that we plan to implement is partial pooling across 4-digit HLA alleles. It has recently been appreciated that binding specificities can vary drastically across HLA alleles from the same allele group. For predicting HLA associated mutations, there has consequently been a shift to use 4-digit resolution data whenever available.

This is not without downsides however, because overall, binding specificities are more similar for alleles in the same group, and, therefore, treating them as completely separate might unnecessarily fragment the available data.

In Bayesian statistics, it is not required to chose one extreme or the other, With partial pooling, estimates from the same allele group do share some information across each other, but they are also allowed to vary if necessary. Therefore, implementing partial pooling might be

another possibility to further increase the reliability of computational methods to study HLA escape.

References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] Jonathan W. Yewdell and Ann B. Hill. Viral interference with antigen presentation. *Nature Immunology*, 3(11):1019–1025, nov 2002.

[8] 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.

[9] 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.

[10] Henrik N. Kløverpris, Alasdair Leslie, and Philip Goulder. Role of HLA adaptation in HIV evolution. *Frontiers in Immunology*, 6, jan 2016.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Joerg Timm, Georg M. Lauer, Daniel G. Kavanagh, Isabelle Sheridan, Arthur Y. Kim, Michaela Lucas, Thillagavathie Pillay, Kei Ouchi, Laura L. Reyor, Julian Schulze zur Wiesch, Rajesh T. Gandhi, Raymond T. Chung, Nina Bhardwaj, Paul Klenerman, Bruce D. Walker, and Todd M. Allen. CD8 epitope escape and reversion in acute HCV infection. *Journal of Experimental Medicine*, 200(12):1593–1604, dec 2004.

[16] 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.

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

[18] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87, jan 1922.

[19] Bettina Budeus, Jörg Timm, and Daniel Hoffmann. SeqFeatR for the discovery of feature-sequence associations. *PLOS ONE*, 11(1):e0146409, jan 2016.

[20] 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.

[21] Valentin Amrhein and Sander Greenland. Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, 2(1):4–4, sep 2017.

[22] Andrew Gelman and John Carlin. Beyond power calculations. *Perspectives on Psychological Science*, 9(6):641–651, nov 2014.

[23] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, aug 2005.

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

[25] M. Baker. 1500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016.

[26] 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.

[27] 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.

[28] Andrew Gelman. Bayesian statistics then and now. *Statistical Science*, 25(2):162–165, may 2010.

[29] 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.

[30] Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors.

[31] 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.

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

[33] Michael Betancourt, Simon Byrne, Sam Livingstone, and Mark Girolami. The geometric foundations of hamiltonian monte carlo. *Bernoulli*, 23(4A):2257–2298, nov 2017.

[34] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, nov 1992.

[35] 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.

[36] 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.

[37] 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.

[38] Anthony R. Ives and Theodore Garland. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology*, 59(1):9–26, nov 2009.

[39] 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.

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

[41] D. Hunt, R. Henderson, J Shabanowitz, K Sakaguchi, H Michel, N Sevilir, A. Cox, E Appella, and V. Engelhard. Characterization of peptides bound to the class i MHC molecule HLA-a2.1 by mass spectrometry. *Science*, 255(5049):1261–1263, mar 1992.

[42] 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.

[43] 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.

[44] 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.