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

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

Abstract

Motivation
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), a key component of adaptive immunity. 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 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 approaches to achieve this.

Results
Here, we present a Bayesian regression model that is easily extensible to include information from different sources, and that makes use of recent advances in Bayesian inference, e.g. by using a sparsifying prior. We show that this integrative approach 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, 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 intracellular viral infections is through the human leukocyte antigen system [1]: In cells with active protein biosynthesis, proteins are continuously synthesized and also degraded by proteasomal degradation, which cleaves proteins into peptides [2].

A small subset of these peptides is presented on the cell surface by receptors called MHC class I encoded by HLA genes, e.g. in humanc HLA-A, -B, and -C. HLA genes are very polymorphic between human individuals, with more than 20,000 different described HLA alleles [3]. The resulting gene products differ in their binding properties, which means that cells from different individuals present a highly diverse set of peptides on the cell surface.

Cytotoxic T-Lymphocytes (CTLs) cells are selected during maturation to only weakly bind to MHC I protein-peptide complexes if the peptide originates from proteins of the self-proteome, but might be able to strongly bind to MHC I protein-peptide complexes with peptides generated from viral proteins [4]. Upon activation, T cells induce cytolytic activity and recruit other immune cells [5].

1.2 HLA escape

In this way, the HLA system exerts strong selection pressure towards virus variants that escape T cell recognition [6], for example through a point mutation that results in reduced binding of an immunogenic peptide to MHC I or through a set of mutations that alters the viral protein in such a way that it is cleaved into different peptides that are not recognized by the host's T cell repertoire [7].

The evolutionary events are complex and occur not only on the level of individuals, where a virus adapts to specific features of the host, but also on the population level, for example because HLA alleles differ in their frequency across geographic regions [8]. Upon transmission to a new host, HLA escape mutations can revert to their wild type, as they are typically associated with a reduction in viral replicative capacity [9]. Kawashima et al. [8] describe an escape mutation that is selected by HLA allele HLA-B*51, does not strongly affect viral replicative capacity, and therefore slowly enriches over time in Japan, where HLA-B*51 commonly occurs.

How quickly a given escape mutation is selected upon transmission in a host depends on the magnitude of the reduction in viral replicative capacity, on the strength of selection pressure, and also on the genetic background, e.g. some escape mutations require compensatory mutations which partly attenuate the negative impact on viral replicative capacity.

Studying HLA escape provides a unique opportunity to gain insight into viral evolution, on the host level, but also on the population level. Unfortunately, identifying HLA escape mutations is difficult in practice.

1.3 Identifying HLA-escape mutations

There are several experimental methods available to study HLA escape: Recombinant MHC I molecules can be used in binding assays. Upon complex formation with a peptide, a change in conformation can be detected with conformation-specific antibodies. This method is relatively fast, but only measures binding affinity of a peptide to MHC I and does not account for antigen processing or immunodominance, which describes the observation that a peptide may be presented via MHC I on the cell surface, but does not induce an immune response.

An experimental setup that resembles the conditions in-vitro more closely but is also more time-consuming is to measure CD8$^+$ T cell responses instead. This is usually done by stimulating peripheral blood mononuclear cells with prototype and variant peptides and measuring the secretion of IFN-γ by intracellular cytokine staining and fluorescence-assisted cell sorting.

To analyze CD8$^+$ T cell responses against endogenously processed antigens, it is necessary to generate cell-lines stably expressing the antigen in question and adding antigen-specific CD8$^+$ T cells. This method scales poorly as it requires transfection of cell lines and antigen-specific expansion of CD8$^+$ T cells.

1.4 Computational methods

Because the currently available experimental methods do not scale well enough to analyze whole viral genomes, a

useful strategy might be to use annotated sequence data to identify candidate HLA escape mutations that can then be verified experimentally.

As the selection pressure exerted by cytotoxic T cells depends on successful recognition of viral peptides on the cell surface, escape mutations are often HLA-allele specific and can therefore be detected as HLA-allele dependent footprints in sequence alignments of viral proteins [10]: At certain alignment positions, a replacement might be more frequently observed in sequences from hosts with a specific HLA allele than in sequences from hosts without that HLA allele. By quantifying this difference for all replacement and HLA allele pairs, it is possible to identify replacements that are enriched in sequences coming from hosts with certain HLA alleles, and thus are likely to be HLA escape mutations.

One way of quantifying this enrichment is Fisher's exact test [11]: 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 $(\neg R_i, H)$, 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 easy to apply [12], but also has several disadvantages that are outlined in Carlson et al. [13]. 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 of applying Fisher's exact test is that HLA class I loci are located in proximity on chromosome 6 and are therefore in linkage disequilibrium, which means that HLA alleles are not inherited completely independent of each other, i.e. inheritance of one HLA allele correlates with inheritance of another HLA allele. When using a statistical method that tests each HLA allele individually without considering the whole set of alleles present, spurious associations might occur: If HLA allele H_1 is associated with an amino acid replacement R, but H_1 is in linkage disequilibrium with another allele H_2, this also means that we observe an association between replacement R and H_2, even in absence of any underlying escape mechanism.

Correlations can not only occur between HLA alleles, but also between replacements. This kind of codon covariation occurs for example in compensatory mutations that attenuate the negative impact of immune escape mutations. For instance, a compensatory mutation might lead to a conformational change in such a way that the mutated protein resembles the original wild type more closely.

Carlson et al. [13] developed a method called Phylogenetic Dependency Network that accounts for phylogenetic bias, HLA linkage disequilibrium, and codon covariation and is based on null hypothesis significance testing.

1.5 Issues with using p-values as a screening tool
In addition to the aforementioned biological reasons why Fisher's exact test is not suited for the analysis of sequence data, there are also more fundamental statistical issues that universally occur when using p-values as a screening tool [14]:

In the presence of small effect sizes and high variance between measurements, as it is typically the case when working with biological data, statistically significant results can often be misleading and are likely to be in the wrong direction (a so-called type S error) or greatly overestimate an effect (a so-called type M error) [15]. This problem has recently been widely appreciated in the literature in the context of the current "replication crisis", which describes the circumstance that scientific claims with seemingly strong statistical evidence fail to replicate [16].

Another issue that occurs when using p-values as a screening tool is the problem of multiple comparisons. When applying a statistical test, 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, which is the expected proportion of false positives [17]. 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 instead 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 also yield more valid estimates [18].

2. Material and methods

When possible, we choose to fit Bayesian models. By using prior information, adding problem-specific structure and partial pooling, the accuracy of estimates can often be noticeably improved [19]. Prior information does not necessarily mean to use external data, even a rough idea about the expected magnitude of estimates is often surprisingly effective.

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 [20].

In the context of identifying HLA-associated mutations we propose to improve existing methods by the following additions, which can be broadly divided into additional information and model structure:

Additional information
Binding affinity prediction

HLA-associated mutations are expected to lie more frequently in regions of known epitopes. 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 [21] 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 binding affinity and antigen processing data in order to predict HLA-associated mutations.

Model structure
Sparsity-inducing priors

Recent advantages in Bayesian inference include so-called sparsity-promoting priors. 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 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.

Partial pooling of 4-digit HLA alleles

It has recently been appreciated that binding specificities can vary drastically across HLA alleles from the same allele group, e.g. between HLA-B*51:01 and HLA-B*51:03. 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: Instead of either treating all HLA alleles from the same allele group as identical or as completely separate, partial pooling of the estimates allows something in between: With partial pooling, estimates

do share some information across each other, but they are also allowed to vary if necessary.

In the context of our model this means that observing a relevant association between a replacement and an HLA allele also makes the model more inclined to estimate relevant associations between that replacement and all other alleles from that allele group. The degree of this partial pooling can be estimated from the data, so if we do observe strong similarity of HLA alleles in a given allele group the estimates are more influenced by each other than if they do not.

3. Implementation

Model backbone
We choose 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. We also choose to fit the model in a Bayesian fashion using Stan, mostly because it allows us to make use of the hierarchical structure in the data and the possibility to use posterior predictive checks as a way of model checking. Consider the following notation:

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

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

,

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 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%.

Including 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. [22] demonstrated the importance of correcting for the phylogenetic structure. There are many approaches described in the literature for phylogenetic regression for binary dependent variables [23], 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 [24].

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. [13] 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 [25] 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 proba-

bilities 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.
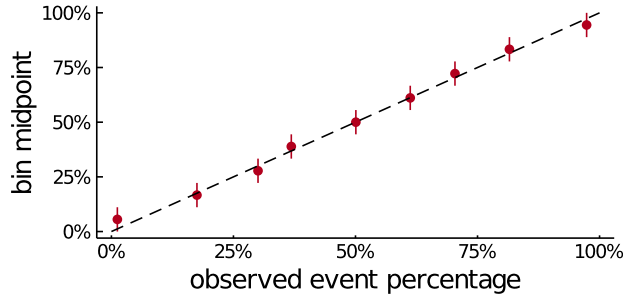

Figure 1. Calibration plot for the HBV dataset. For a description of all datasets used in this study see section 4. Calibration plots for the other datasets are shown in the supplementary. All observations are first sorted by increasing estimated probability $P(y_{ik} = 1|\Psi)$ and then grouped into 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:

$$
\begin{aligned}
y_{ik} &\sim \text{bernoulli}(\theta_{ik}) \\
\theta_{ik} &= \text{inv\_logit}(\beta_{0_k} + \gamma \log it(P(y_{ik} = 1|\Psi)) \\
&\quad + \sum_{k=1}^{D} X_{ik}\beta_{\text{HLA}_{jk}})
\end{aligned}
\tag{3}
$$

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.

Including HLA linkage disequilibrium and codon covariation

The other two important confounding effects that are recognized in Carlson et al. [13] are HLA linkage disequilibrium and codon covariation. HLA linkage disequilibrium describes the observation that HLA alleles that lie in proximity to each other on the chromosome are also more likely to be inherited together. When using a statistical method that evaluates each HLA allele individually, this may lead to spurious associations: If some HLA allele X is associated with replacement k, and HLA allele Y often co-occurs with HLA allele X, this means that we also identify an association between HLA allele Y and replacement k.

This association is correct in a statistical sense, but if the purpose of the model is to identify sites under HLA-mediated selection pressure, these associations due to linkage disequilibrium are most likely unwanted.

Note that these spurious associations only occur when testing each HLA allele individually. In a regression model where all HLA alleles of a sample are evaluated at once, the marginal poster distributions for correlated predictors become broader to accurately reflect that the model is unable to discern which of these predictors explain the observed association with a replacement.

We do not currently account for codon covariation in our model, as we are interested in identifying all HLA associated mutations (which include for example compensatory mutations). The general structure of the model, in particular the sparsity assumption introduced in section 3, would allow adding codon covariation by adding all other codons as additional predictors. This would generally not be possible in a frequentist setting because the predictor matrix would become singular and the inverse does not exist.

Including sparsity assumptions

One of the recent advances in Bayesian inference is the development of special prior distributions which can be used to give models additional useful properties. An example of these special prior distributions are the so-called sparsity-inducing priors, which can be used to convey the assumption that most of the predictors in a regression model are not associated with the outcome and are therefore 0.

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. Adding this sparsity assumption can lead to better inferences, because the uncertainty of the non-relevant HLA regression coefficients does not propagate further and therefore the model is able to better discern signal from noise.

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 [26], which is defined as a scale-mixture of gaussians:

$$
\begin{aligned}
\beta_{jk} &\sim \text{Normal}(0, \tau^2 \lambda_{jk}^2) \\
\lambda_{jk} &\sim \text{Cauchy}^+(0, 1) \\
\tau &\sim \text{Cauchy}^+(0, \tau_0)
\end{aligned}
\tag{4}
$$

,

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. [27], 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 [27], 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.

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

The improvement of the logistic regression model with the horseshoe prior is a considerable improvement over a standard logistic regression model (section 4). In our benchmark, the addition of the horseshoe prior alone shows to be more effective than a model withoutt horseshoe prior but with phylogeny included.

Including epitope prediction software

There is vast experimental data on epitope binding motifs available, e.g. through the use of mass spectrometry [28]. Epitope prediction software like MHCFlurry [21] 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}:

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

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.

Full model specification
The full specification of the model is given by:

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

The full model specification includes some aspects that were not covered in the previous sections. In particu-

lar, 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 [27]. 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.

All models were implemented in Stan [29]. 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 [30] package is available to run the model on custom data. Due to restrictions of dependencies (MHCFlurry and RAxML-ng), HAMdetector is only available on Linux. We additionally provide Docker images to easily run the model on any operating system in a containerized environment, assuming Docker is installed.

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 $\beta_{\text{HLA}_{jk}}$. 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 [27].

The value of τ_{0k} is chosen to imply 10 effective non-zero HLA regression coefficients per replacement. The ra-

tionale behind this parameterization is again outlined in[27]. 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.

4. Results

To show that our model works well in different real-world scenarios, we run versions of our model on several datasets:

- A large HIV dataset consisting of X sequences, that were kindly provided by X and were also used in the Phylogenetic Dependency Network study [13].

- A set of 351 HIV sequences mostly spanning the Pol gene from the Arevir database [31].

- 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

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

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 [25] 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.

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
For all the Bayesian models, Markov Chain Monte Carlo sampling was done using Stan [29]. All model fits showed no signs of inference issues. In total, 4 different chains with 1000 samples each were drawn from the posterior. The effective sample size for all model parameters exceeded 500, R-hat convergence diagnostic values were below 1.01 in all cases.

Posterior predictive checks
Bayesian modelling allows a simple yet effective strategy to test models: By simulating data under the model's assumptions and then comparing simulated and observed data, model misspecifications can be identified and highlight possible model improvements.

Here, we compare the calibration of the posterior predictive probability of observing the replacement to the actual observed data using calibration plots. The plots are similar to the one showed in figure 1 but this time do not evaluate the probabilities based on the phylogenetic model alone, but show the calibration of the posterior predictive probabilities based on the full model.

An example calibration plot of the HAMdetector model on the Core protein sequence data of the HBV dataset is shown in figure 2. Plots for the other models and datasets are shown in the supplementary.

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 a sequence alignment. A third issue when comparing models is that HLA-associated mutations are not necessarily binary, so the true positive / false negative
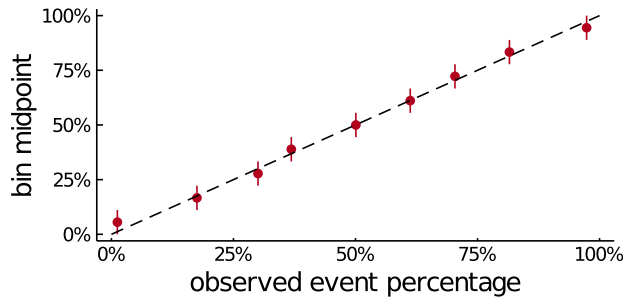
Figure 2. Calibration plot for the HBV dataset. Calibration plots for the other datasets are shown in the supplementary. All observations are first sorted by increasing estimated probability $P(y_{ik} = 1)$ and then grouped into n bins. For each bin, the fraction of observations with $y_{ik} = 1$ (observed 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.

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 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 logis-
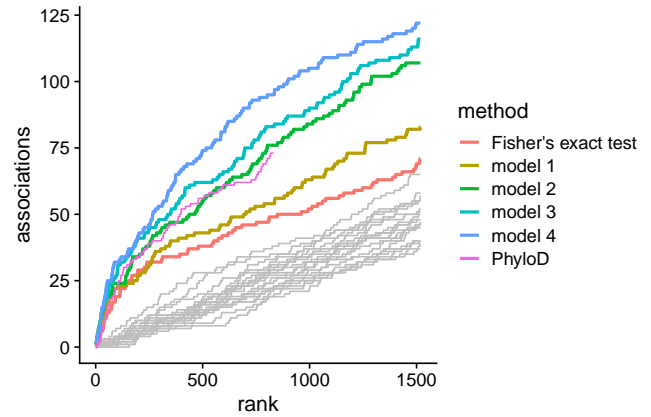

Figure 3. Number of associations inside the boundary of known epitopes vs. rank. For each model, the pairs of HLA alleles and replacements are sorted by decreasing confidence of being an HLA associated mutation. Then, each replacement is compared to a list of known epitopes and a line of the cumulative 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.

tic 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

shows to larger drastic improvement. Note that this model only uses epitope *prediction* software and does not utilize any information of experimentally confirmed epitopes. These are only used for model evaluation. Therefore, our method works even in absence of any experimental epitope information.

4.1 Leave-one-out cross-validation
In addition to testing the model against experimentally validated epitopes, we also perform leave-one-out cross-validation. While the comparison against known epitopes compares inferences of the regression coefficients, leave-one-out cross-validation compares different models in regard to their predictive performance, i.e. how well they can predict the presence or absence of a replacement in a sequence based on the available data.

Unlike usual applications of cross-validation, we are not primarily interested in the absolute magnitude of prediction errors, because the main objective of the statistical inference is to quantify the contribution of the HLA alleles. However, as the compared models are all special cases of the complete model, we can use cross-validation to gain insight on the relative importance of each aspect of the model.

Performing leave-one-out cross-validation has historically been challenging 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. Pareto-smoothed importance sampling approximates the leave-one-out posterior by re-weighting existing draws from the posterior distribution and has the advantage of allowing to perform leave-one-out cross-validation by only fitting the model once. Additionally, the method provides good diagnostics to identify cases where the approximation is unreliable [32].

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.

Table x shows the LOO results for the HBV dataset. Using a simple logistic regression model similar to the one used in [10] as a baseline, we compare a logistic regression model with horseshoe prior, a logistic regression model with horseshoe prior and phylogeny, and the complete HAMdetector model with horseshoe prior, phylogeny and epitope prediction to each other. The table shows expected log-predictive density differences compared to the baseline model (higher values are better).

As the models include more and more information, the predictive accuracy increases. Note that this is not due to overfitting, because leave-one-out cross-validation approximates the model performance on unseen data. Additional tables are shown in the supplementary, the results are consistent across all datasets. 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. Perhaps surprisingly, the model with epitope prediction performs roughly as well as the model without. This is true from a perspective of predictive accuracy, however, in determining which HLA alleles are associated with a given replacement, adding epitope prediction is highly useful, as shown in the previous section. This observation is explained by the circumstance that leave-one-out cross-validation takes all positions and replacements into account, but a given HLA allele is only restricting a subset of all positions in a sequence alignment. Additionally, the benefit of accurately identifying an associated HLA allele is only relevant for the samples which do have that allele, which for most alleles is only a small subset of all available samples. Therefore, looking at predictive accuracy alone is misleading in that regard, because we are more interested in the predictors than in

the predictions themselves.

4.2 A real-world example

We had the opportunity to test our model against a set of associations identified by Fisher's exact test that turned out to be false positives. On the HDV dataset, Fisher's exact test was applied to a set of HLA annotated sequences and statistically significant results were reported. The identified positions were experimentally validated and 9 associations turned out to be false-positives. This kind of data is rate because negative results are usually not published in the literature. We ran HAMdetector on the same dataset, the posterior probabilities are shown in table **??**. They denote the probability that the corresponding regression coefficient is positive. Associations with strong support have a posterior probability close to 1, associations with no support a probability close to 0.5 (because the marginal posterior distribution is centered around 0 and allocated half of the probability mass to values greater than 0). The table shows that most of the false-positives in Fisher's exact test have low posterior probabilities in our model, which means that these alleles are correctly identified as not being associated with the replacement. The pairs of HLA alleles and replacements that were experimentally validated have the highest posterior probabilities, with the exception of P89T - HLA-B*37, which does not have strong support in our model. The results suggest that while our model is not perfect, it is a strong improvement over Fisher's exact test, which is frequently applied in this context. Please note that the results in this subsection have previously been published using a preliminary version of HAMdetector.

5. Discussion

We present 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.

replacement	allele	p-value (Fisher's exact test)	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	-

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

[11] 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.

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

[13] 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.

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

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

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

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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 poly-

morphisms and HLA allele associations. *Science*, 315(5818):1583–1586, mar 2007.

[23] 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.

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

[25] 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.

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

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

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.