A phylogeny-aware GWAS framework to correct for heritable pathogen effects on infectious disease traits

Sarah Nadeau[1,2], Christian W. Thorball[3], Roger Kouyos[4,5], Huldrych F. Günthard[4,5], Jürg Böni[4], Sabine Yerly[6], Matthieu Perreau[7], Thomas Klimkait[8], Andri Rauch[9], Hans H. Hirsch[8,10,11], Matthias Cavassini[12], Pietro Vernazza[13], Enos Bernasconi[14], Jacques Fellay[2,3,15], Venelin Mitov[†,1,2], Tanja Stadler[†,*,1,2], and the Swiss HIV Cohort Study (SHCS)

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]Precision Medicine Unit, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[4]Institute of Medical Virology, University of Zurich, Zurich, Switzerland

[5]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

[6]Division of Infectious Diseases, Laboratory of Virology, Geneva University Hospital, Geneva, Switzerland

[7]Division of Immunology and Allergy, University Hospital Lausanne, Lausanne, Switzerland

[8]Department of Biomedicine, University of Basel, Basel, Switzerland

[9]Department of Infectious Diseases, Bern University Hospital and University of Bern, Bern, Switzerland

[10]Division of Clinical Virology, University Hospital Basel, Basel, Switzerland

[11]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland

[12]Division of Infectious Diseases, University Hospital Lausanne, Lausanne, Switzerland

[13]Division of Infectious Diseases, Cantonal Hospital St. Gallen, St. Gallen, Switzerland

[14]Division of Infectious Diseases, Regional Hospital Lugano, Lugano, Switzerland

[15]Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

†Co-last authors

∗Corresponding author: tanja.stadler@bsse.ethz.ch

Abstract

Infectious diseases are a unique challenge for genome-wide association studies (GWAS) because pathogen, host, and environmental factors can all affect disease traits. Previous GWAS have successfully identified several human genetic variants associated with HIV-1 set point viral load (spVL), among other important infectious disease traits. However, these GWAS do not account for potentially confounding or extraneous pathogen effects that are heritable from donor to recipient in transmission chains. We propose a new method to consider the full genome of each patient's infecting pathogen strain, remove strain-specific effects on a trait based on the pathogen phylogeny, and thus better estimate the effect of human genetic variants on infectious disease traits. In simulations, we show our method can increase GWAS power to detect truly associated host variants when pathogen effects are highly heritable, with strong phylogenetic correlations. When we apply our method to HIV-1 subtype B data from the Swiss HIV Cohort Study, we recover slightly weaker but qualitatively similar signals of association between spVL and human genetic variants in the *CCR5* and major histocompatibility complex (MHC) gene regions compared to standard GWAS. Our simulation study confirms that based on the estimated heritability and selection parameters for HIV-1 subtype B spVL, standard GWAS are robust to pathogen effects. Our framework may improve GWAS for other diseases if pathogen effects are even more phylogenetically correlated amongst individuals in a cohort.

Introduction

A key goal of genome-wide association studies (GWAS) is to understand the genetic basis of phenotypic variation among individuals. In a typical GWAS, millions of genetic variants from across the human genome are screened for statistical association with a trait of interest. Ideally, this procedure identifies variants that are located in, or are in linkage disequilibrium with, alleles that directly affect the trait. If GWAS finds a variant strongly associated with a disease trait, the gene product may be a good drug target (Okada *et al.*, 2014). Even if no single variant has a strong association, many small associations can be aggregated into a polygenic risk score to identify high-risk individuals (Dudbridge, 2013).

For HIV, GWAS have used a trait called set point viral load (spVL) to identify human variants associated with the severity of disease course. spVL is generally defined to be the average concentration of viral RNA copies in host plasma during the asymptomatic phase of infection in the absence of treatment (see e.g. Alizon *et al.* (2010)). In untreated individuals, spVL is predictive of the duration of asymptomatic infection (Mellors *et al.*, 1996) and infectiousness (Quinn *et al.*, 2000). If viral load can be reduced to undetectable levels, an individual is effectively uninfectious and the risk of disease progression is massively reduced (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2019). Notably, spVL varies by orders of magnitude between individuals (Mellors *et al.*, 1996). Thus, spVL measurements point to a wide range in natural HIV control amongst individuals.

To understand HIV pathogenicity, it is important to understand to what extent spVL is determined by host genetic factors (Bartha *et al.*, 2013; Dalmasso *et al.*, 2008; Fellay *et al.*, 2007, 2009; McLaren *et al.*, 2012; Pelak *et al.*, 2010; Pereyra *et al.*, 2010; van Manen *et al.*, 2011). Heritability is a key measure of how genetically-determined a trait is. Here we distinguish between two different heritability measures that are used in different contexts in the study of infectious diseases. Broad-sense heritability H^2 measures the fraction of total trait variance that is heritable, i.e. due to inherited differences. In the infectious disease case, broad-sense heritability from pathogen factors, which are inherited by recipients from their infection partners, is typically measured. On the other hand, narrow-sense heritability h^2 measures the fraction of total trait variance due specifically to additive genetic effects, i.e. the sum of independent effects from all genetic variants. GWAS for infectious disease traits typically measure the narrow-sense heritability of a trait based on human genetic variants.

Several GWAS have been done to measure the narrow-sense heritability of spVL and identify associated host genetic variants (Bartha *et al.*, 2013; Dalmasso *et al.*, 2008; Fellay *et al.*, 2007, 2009; McLaren *et al.*, 2012; Pelak *et al.*, 2010; Pereyra *et al.*, 2010; van Manen *et al.*, 2011). The largest study to-date by McLaren *et al.* (2015) estimated the narrow-sense heritability of spVL from human genetic variants to be approximately 25%. All but 5% of this was attributed to two regions in the human genome, the major histocompatibility complex (MHC) and C-C motif chemokine receptor 5 (*CCR5*). Both associations are biologically relevant: the MHC encodes proteins that present viral epitopes at the cell surface and *CCR5* encodes a co-receptor for HIV-1 cell entry. In other words, MHC proteins match bits of the virus like puzzle pieces and display these to signal that a cell is infected. CCR5 proteins help the virus infect target cells.

In addition to these human genetic factors, it is well-recognized that viral genetic factors affect spVL. As mentioned, heritability from the viral side is typically measured using broad-sense heritability. Estimates differ depending on the methods employed and the cohort studied (see Mitov and Stadler (2018) for a discussion of this uncertainty). Estimates using phylogenetic methods on large UK and Swiss cohorts by Mitov and Stadler (2018) and Bertels *et al.* (2018) measured the broad-sense heritability of spVL from the virus to be 21% - 29%. However, variation in the MHC is known to exert strong selective pressure on the virus (Kløverpris *et al.*, 2016; Nguyen *et al.*, 2021). If the virus can change its "puzzle piece" shape to escape MHC-presentation, infected cells can go undetected. This means that MHC variants affect spVL largely via selection on the virus (Bartha *et al.*, 2017). In summary, human genetic factors play a role in determining spVL, but these effects may be due to interaction with specific viral genetic variants.

Most of the GWAS for human genetic determinants of spVL conducted so far (Dalmasso *et al.*, 2008; Fellay *et al.*, 2007, 2009; McLaren *et al.*, 2012, 2015; Pelak *et al.*, 2010; Pereyra *et al.*, 2010; van Manen *et al.*, 2011) do not explicitly consider any viral effect on spVL. In these GWAS, viral genetic effects are lumped in with residual variance due to other, non-genetic factors. This

has several potential negative consequences. (i) Viral effects may be confounding or extraneous variables that bias estimates of host genetic effects. (ii) Variability due to viral effects would make it more challenging to identify human variants of small effect. Finally, (iii) spVL values from a cohort are not truly independent samples, given that patients closer in the transmission chain have more similar strains and therefore more similar spVL values.

Issue (iii) is closely related to a well-known problem in standard GWAS. Shared (human) ancestry, especially between close relatives, can also give rise to spurious genetic correlations with a trait. Corrections for these correlations are well-developed and widely accepted (Astle and Balding, 2009; Price *et al.*, 2006). More recently, Power *et al.* (2017) emphasized the need to do similar corrections for shared pathogen ancestry in microbial GWAS. Two state-of-the-art methods exist for this (Collins and Didelot, 2018; Earle *et al.*, 2016). However, these approaches are only suitable to quantify effects from *pathogen* genetic variants on a trait. In contrast, we want to estimate effects of *human* genetic variants on a trait, while accounting for pathogen effects. Naret *et al.* (2018) developed a relevant method for this in the context of a genome-to-genome GWAS framework. The authors suggest adding principle components derived from the pathogen phylogeny as covariates to the linear regression models for association testing. This should correct for trait correlations due to shared pathogen ancestry. However, the top principle components capture only some of the information from the full pathogen phylogeny. Furthermore, we would like to simultaneously address issues (i) and (ii).

In this work, we draw from the field of phylogenetic comparative methods to develop a new GWAS framework that estimates and removes trait variability due to the pathogen using information from the full pathogen phylogeny. Our approach should help identify human genetic variants that affect disease traits and more accurately estimate their effects.

In the following we describe a statistical model for the spVL trait, derive a maximum likelihood estimate for the viral part of spVL under this model, and describe a new infectious disease GWAS framework using this information. In simulations, we show that this framework can improve GWAS power to detect host genetic variants that affect disease traits. Finally, we apply our framework to human and viral genome data from the Swiss HIV Cohort Study (SHCS) and show that associations with spVL are robust to a correction for viral effects. Although we developed our framework in the context of HIV-1 spVL, this approach can readily be applied to other heritable infectious disease traits.

New Approaches

A statistical model for spVL

Variation in spVL comes from several sources: direct host genetic effects, pathogen effects, interaction effects between the host and the pathogen, and other environmental effects. Of these, only pathogen effects are heritable from one transmission partner to another (Leventhal and Bonhoeffer, 2016). To characterize these effects, we use a phylogenetic mixed model (PMM) (Housworth *et al.*, 2004). PMMs assume continuous traits like spVL are the sum of independent heritable and non-heritable parts. In our case, pathogen effects comprise the heritable part and all other effects comprise the non-heritable part. The heritable part is modeled by a random process occurring in continuous time along the branches of the pathogen phylogeny, as in Figure 1A. The non-heritable part is modeled as Gaussian noise added to sampled individuals at the tips of the phylogeny.

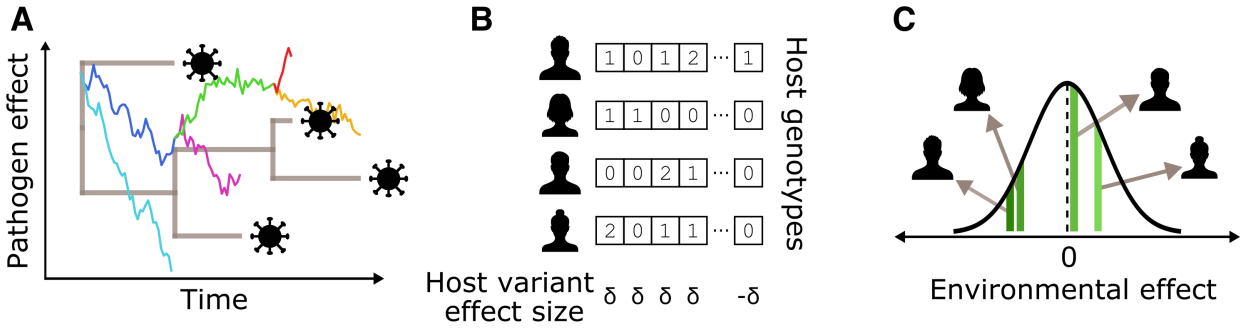

Figure 1: A high-level schematic of our POUMM-based simulation framework. (A) shows how pathogen genetic effects on spVL evolve along the pathogen phylogeny according to an Ornstein-Uhlenbeck process. (B) shows how host genetic effects are the sum of independent effects from several causal variants. Each variant can be present in 0, 1, or 2 copies. Half the variants have a positive effect of size δ and half have a negative effect of size δ. (C) shows how environmental effects are independently drawn from a Gaussian distribution centered at 0. These three effects sum to the trait value for each simulated individual.

So far, PMMs with two types of random processes have been used to model spVL evolution. The Brownian Motion (BM) process assumes unbounded trait values, i.e. spVL can attain any value. The Ornstein-Uhlenbeck (OU) process assumes trait values fluctuate around an optimal value, i.e. extreme spVL values are unlikely. Mitov and Stadler (2018) and Bertels *et al.* (2018) previously showed the OU process has higher statistical support for spVL. This makes sense given that spVL is likely under stabilizing selection to maximize viral transmission potential (Fraser *et al.*, 2014). Therefore, we assume the OU process. The full model is called the phylogenetic Ornstein-Uhlenbeck mixed model (POUMM) and is described in detail by Mitov and Stadler (2018). Here, we review

the main points in the spVL context.

Under the POUMM, the spVL trait z is the sum of viral effects g_v, host genetic effects g_h, and other environmental or interaction effects ϵ. We can group the non-heritable effects g_h and ϵ into a broader category of "environmental" effects e:

$$z = g_v + e \tag{1}$$

g_v is a viral trait that evolves along the phylogeny according to an OU process. The OU process is defined by a stochastic differential equation with two terms. The first term represents a deterministic pull towards an optimal trait value and the second term represents stochastic fluctuations modelled by Brownian motion (Butler and King, 2004):

$$\begin{aligned} dg_v(t) &= \alpha[\theta - g_v(t)]dt + \sigma dW_t \\ g_v(0) &= g_0 \end{aligned} \tag{2}$$

Here the parameter α represents selection strength towards an evolutionarily optimal value represented by parameter θ. The parameter σ measures the intensity of stochastic fluctuations in the evolutionary process. Finally, dW_t is the Wiener process underlying Brownian motion. The OU process is a Gaussian process, meaning that $g_v(t)$ is a Gaussian random variable. Assuming $g_v(t)$ starts at initial value g_0 at time $t = 0$ at the root of the phylogeny, we can write the expectation for $g_v(t)$ at time t:

$$E[g_v(t)] = g_0 e^{-\alpha t} + (1 - e^{-\alpha t})\theta \tag{3}$$

and the variance in $g_v(t)$ if we were to repeat the random evolutionary process many times (Butler and King, 2004):

$$Var[g_v(t)] = \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t}) \tag{4}$$

g_v evolves independently in descendent lineages after a divergence event in the phylogeny. The covariance between $g_v(t)$ in a lineage i at time t_i and another lineage j at time t_j, $Cov\big(g_{v_i}(t_i), g_{v_j}(t_j)\big)$, increases with the amount of time between t_0 and the divergence of the two lineages, $t_{0(ij)}$, and decreases with the total amount of time the lineages evolve independently, d_{ij} (Butler and King, 2004):

$$Cov\big(g_{v_i}(t_i), g_{v_j}(t_j)\big) = \frac{\sigma^2}{2\alpha}[e^{-\alpha d_{ij}}(1 - e^{-2\alpha t_{0(ij)}})] \tag{5}$$

Next, we remember that e is the non-heritable, environmental part of spVL. e is modeled as a Gaussian random variable that is time- and phylogeny-independent. The expectation of e is 0,

meaning environmental effects are equally likely to raise or lower spVL from the virus-determined level. The parameter σ_e^2 measures the between-host variance of the environmental effect.

$$\begin{aligned} E(e) &= 0 \\ Var(e) &= \sigma_e^2 \end{aligned} \tag{6}$$

Finally, broad-sense trait heritability can be calculated as the fraction of total trait variance that is heritable:

$$H_t^2 = \frac{Var[g_v(t)]}{Var[g_v(t)] + Var(e)} = \frac{\frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t})}{\frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t}) + \sigma_e^2} \tag{7}$$

Teasing apart pathogen and non-pathogen effects on spVL

Given the assumptions of the POUMM, we can estimate a heritable pathogen effect on spVL and a non-heritable, host and environmental effect on spVL. Here, we derive a maximum-likelihood estimate for these values for individuals in a cohort, given measured spVL values and a pathogen phylogeny linking the infecting strains.

Let $\boldsymbol{g_v}(\boldsymbol{t})$ be a vector of g_v values, one for each individual in the cohort. $\boldsymbol{t}$ are the sampling times of each individual relative to the root of the phylogeny. To simplify notation, we omit the $\boldsymbol{t}$ from here on. $\boldsymbol{g_v}$ is a realization of a Gaussian random vector $\boldsymbol{G_v} \sim \mathcal{N}(\boldsymbol{\mu}_{OU}, \boldsymbol{\Sigma}_{OU})$. The expectation $\boldsymbol{\mu}_{OU}$ is defined by equation 3, the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}_{OU}$ are defined by equation 4, and the off-diagonal elements of $\boldsymbol{\Sigma}_{OU}$ by equation 5.

Similarly, let $\boldsymbol{e}$ be a vector of the environmental part of spVL for each individual. $\boldsymbol{e}$ is a realization of a Gaussian random vector $\boldsymbol{E} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_E)$, where $\boldsymbol{\Sigma}_E$ is a diagonal matrix with diagonal elements equal to σ_e^2.

Considering that $\boldsymbol{G_v}$ and $\boldsymbol{E}$ are independent random vectors and that their realizations $\boldsymbol{g}_v$ and $\boldsymbol{e}$ must sum together to equal the observed spVL values $\boldsymbol{z}$, we can write the following proportionality for the joint probability density of $\boldsymbol{g}_v$ and $\boldsymbol{e}$:

$$f(\boldsymbol{g}_v, \boldsymbol{e}) \propto \mathcal{N}(\boldsymbol{g}_v; \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G) \tag{8}$$

where the expected value of $\boldsymbol{g}_v$ and the covariance matrix $\boldsymbol{\Sigma}_G$ are defined as:

$$Exp(\boldsymbol{g}_v) = \boldsymbol{\mu}_G = \boldsymbol{\Sigma}_G(\boldsymbol{\Sigma}_{OU}^{-1}\boldsymbol{\mu}_{OU} + \boldsymbol{\Sigma}_E^{-1}\boldsymbol{z}) \tag{9}$$

$$\boldsymbol{\Sigma}_G = (\boldsymbol{\Sigma}_{OU}^{-1} + \boldsymbol{\Sigma}_E^{-1})^{-1} \tag{10}$$

Proof.

$$\begin{aligned} f(\boldsymbol{g}_v,\ \boldsymbol{e}) &= f(\boldsymbol{g}_v|\ \boldsymbol{e}) \times f(\boldsymbol{e}) \\ &= f(\boldsymbol{g}_v) \times f(\boldsymbol{e}) \\ &= \mathcal{N}(\boldsymbol{g}_v;\ \boldsymbol{\mu}_{OU}, \boldsymbol{\Sigma}_{OU}) \times \mathcal{N}(\boldsymbol{e};\ \mathbf{0}, \boldsymbol{\Sigma}_E) \\ &= \mathcal{N}(\boldsymbol{g}_v;\ \boldsymbol{\mu}_{OU}, \boldsymbol{\Sigma}_{OU}) \times \mathcal{N}(\boldsymbol{z} - \boldsymbol{g}_v;\ \mathbf{0}, \boldsymbol{\Sigma}_E) \\ &= \mathcal{N}(\boldsymbol{g}_v;\ \boldsymbol{\mu}_{OU}, \boldsymbol{\Sigma}_{OU}) \times \mathcal{N}(\boldsymbol{g}_v;\ \boldsymbol{z}, \boldsymbol{\Sigma}_E) \end{aligned} \tag{11}$$

Equations 9 and 10 follow from eq. 11 and eq. 371, p. 42, section 8.1.8 "Product of Gaussian densities" in Petersen and Pedersen (2012). □

Importantly, equation 9 is the maximum likelihood estimate for $\boldsymbol{g}_v$, the viral effect on spVL, taking into account all available information - measured spVL, the pathogen phylogeny, and inferred POUMM parameters. This estimator is an inverse-variance weighted average of measured spVL ($\boldsymbol{z}$) and information from the POUMM evolutionary model ($\boldsymbol{\mu}_{OU}$). In other words, g_v will be closer to measured spVL if spVL is not very heritable. If spVL is highly heritable, g_v will be closer to the expected value under the POUMM, i.e. take more information from the phylogenetic relationships between infecting strains.

Given the estimator we just derived for $\boldsymbol{g}_v$, we can now estimate $\boldsymbol{e}$, the spVL value *without* pathogen effects:

$$\hat{\boldsymbol{e}} = \boldsymbol{z} - Exp(\boldsymbol{g}_v) \tag{12}$$

We will use this value to try to improve upon standard GWAS methods in infectious disease.

A POUMM-based GWAS framework for infectious disease

We propose to improve standard GWAS for infectious diseases by estimating and removing trait variability due to pathogen effects. Our new framework is as follows:

1. Sample host genotypes, trait values, and pathogen genome sequence data from a cohort.
2. Construct a pathogen phylogeny using the pathogen genome sequences.
3. Estimate the parameters of the POUMM based on the trait values and the pathogen phylogeny. This can be done with e.g. the R package POUMM (Mitov and Stadler, 2017).
4. Generate maximum-likelihood estimates for the pathogen and corresponding non-pathogen effects on the trait using equations 9 and 12.
5. Perform GWAS with only the non-pathogen effects on the trait as the response variable.

Results

Simulation study

To test the theoretical best-case performance of our method, we simulated data under the POUMM and applied our framework to the simulated data. Figure 1 shows a high-level schematic of our simulation framework and Table 2 gives the value or expression for each parameter. For a description of the full simulation scheme, see Figure S1. In a nutshell, we simulated independent, additive host genetic effects, independent environmental effects, and heritable pathogen genetic effects under different scenarios of trait heritability and selection strength. To maintain the same heritability while varying selection strength, we counter-balanced by varying the intensity of stochastic evolutionary fluctuations accordingly. We fixed other variables to plausible values based on the spVL literature.

Estimator accuracy

First, we evaluated how well our method estimated the additive host genetic effects from the simulated data. Additive host genetic effects represent an ideal (albeit unattainable) baseline for infectious disease GWAS. Figure 2A shows that our method incorporating phylogenetic information can more accurately estimate these value compared to the trait value. To ensure a fair comparison, we scaled trait values to have the same mean, zero, as host genetic effects so as not to bias the root mean squared error (RMSE) by a constant factor. In the supplemental material, we show why the scaled trait value is expected to have an RMSE of approximately 0.74 under our simulation scheme. By incorporating phylogenetic information, we can improve upon this error in scenarios where the trait is highly heritable, under low selection pressure, and with relatively moderate stochastic fluctuations compared to outbreak duration. This is because high heritability means the trait value is highly pathogen-dependent. Then, when the trait is under weak selection, the pathogen effects can drift far from the long-term optimum and stochastic fluctuations are low to maintain the same heritability. Thus, our method performs well when an infectious disease trait is highly heritable and trait values are strongly correlated amongst close transmission partners.

Theoretical GWAS improvement

Next, we characterized the evolutionary scenarios under which our framework can actually improve GWAS power. We used the true positive rate (TPR) to evaluate the fraction of simulated causal host genetic variants we could recover as being significantly associated with the trait. We performed three different GWAS for each simulated dataset: the first represents an ideal in which we can exactly know and remove pathogen effects from trait values, the second is using our method to estimate this value and remove it, and the third represents a standard GWAS using the scaled trait value. Figure 2B shows that our framework can improve the TPR in simulated scenarios

where selection strength < 10 time^{-1} and heritability $> 45\%$. If we were able to perfectly estimate and remove pathogen effects from a trait, the TPR would increase across all values of selection strength so long as the trait is more than marginally heritable. We estimate approximately 25% to be the heritability threshold above which GWAS power is negatively impacted by pathogen effects. In summary, we show it is theoretically possible to improve GWAS power for heritable infectious disease traits by estimating and removing pathogen effects using information from the pathogen phylogeny.

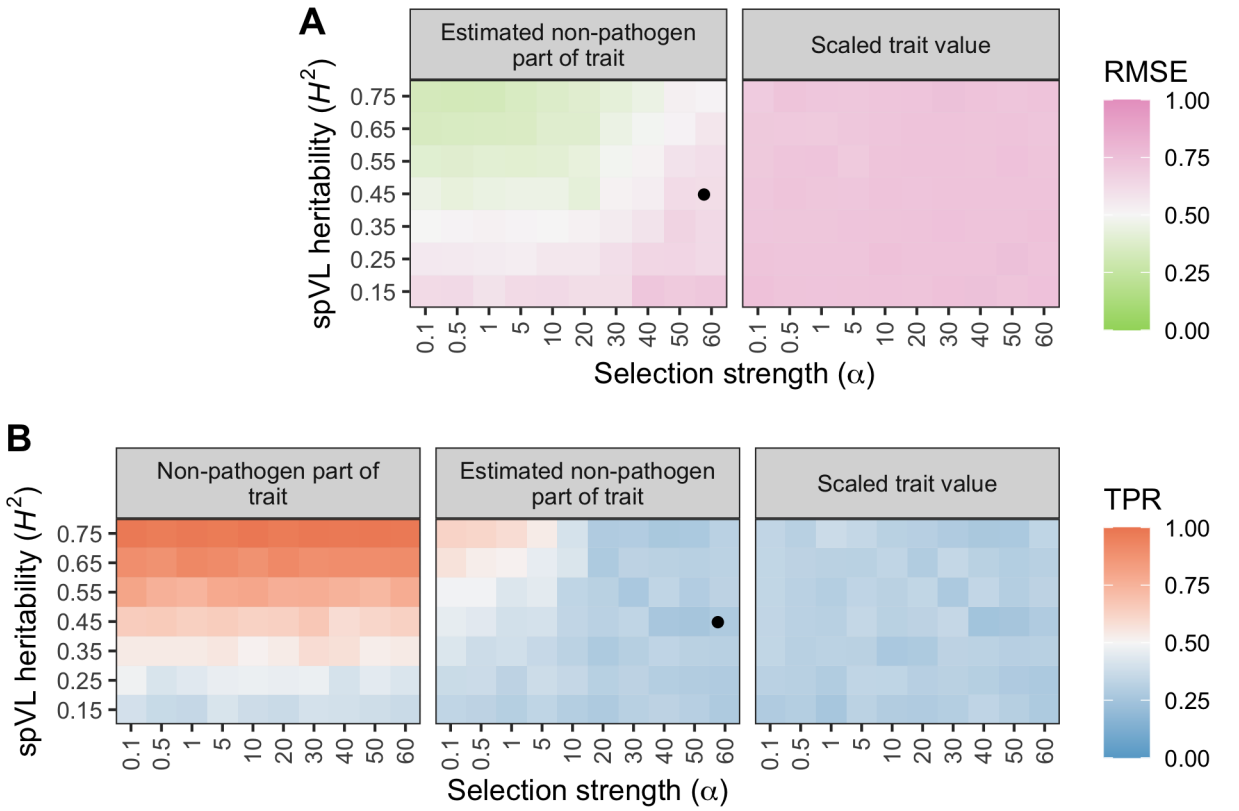

Figure 2: Results from the simulation study. We simulated host, pathogen, and environmental effects on a trait under the POUMM with different heritability (y-axis) and selection strength (x-axis) parameters. For each simulated dataset, we applied our method to estimate the non-pathogen effects and performed GWAS with these values. (A) shows that our method (left) can generate more accurate estimates of additive host genetic effects than the trait value, scaled by its mean (right). (B) shows how GWAS power can improve given the true, simulated non-pathogen effect on spVL (left) and using our estimate for this value (middle) compared to using the scaled trait value (right). Each tile's color corresponds to the average value across 20 simulated datasets of 500 samples. The black point represents our estimates for the heritability and selection strength of spVL based on Swiss HIV Cohort Study data. RMSE = Root mean square error, TPR = True positive rate.

GWAS on the Swiss HIV Cohort

Finally, we applied our framework to empirical data from the Swiss HIV Cohort Study (SHCS). We used data collected from 1,392 individuals in Switzerland infected with HIV-1 subtype B between 1994 and 2018. The SHCS provided viral load measurements, *pol* gene sequences, and human genotype data for these individuals. We followed the framework outlined above to estimate the pathogen and non-pathogen effects on spVL for the cohort from these data. Figure S2 shows the calculated (total) spVL values, which vary between approximately 1 and 6 log copies/mL in the cohort. Figure S3 shows that this trait is not strongly phylogenetically structured in the cohort, despite high heritability. Finally, figure S4 shows that the estimated non-pathogen effects on spVL correlate quite strongly with total spVL. We estimated spVL heritability in this cohort to be 45% (95% highest posterior density, HPD, 24 - 67%) and selection strength to be 58 time^{-1} (95% HPD 19 - 95) (Figure S5, Table S1). To put these values into the context of our simulation study, they are shown as black points on Figure 2.

We compared our proposed GWAS framework with a more standard approach by performing two different GWAS on the same SHCS human genotypes. In the "GWAS with standard trait value" we used the total trait value, our calculated spVL values, as the GWAS response variable. In the "GWAS with estimated non-pathogen part of trait" we used our estimates for the non-pathogen effects on spVL. Figure 3A shows that results are qualitatively similar between the two GWAS. Q-Q plots show the distribution of p-values are very similar as well (Figure S6). Figure 3B shows how the strength of association changed for some variants in the MHC and *CCR5* regions. Taking into account phylogenetic information slightly decreased association strength for most variants in the *CCR5* region. Association strength increased for some variants in the MHC, for example, SNP rs9265880 had the greatest increase in significance in the MHC region, from a p-value of 3.5×10^{-07} to 7.7×10^{-09}. However, the top-associated variants in the MHC and *CCR5* regions were consistent regardless of the GWAS response variable used (Table S2). Finally Table 1 shows how our GWAS results compare for the two top-associated SNPs identified by McLaren *et al.* (2015). In summary, there are no clear patterns that point to new regions of association in the human genome with spVL when we take into account the pathogen phylogeny.

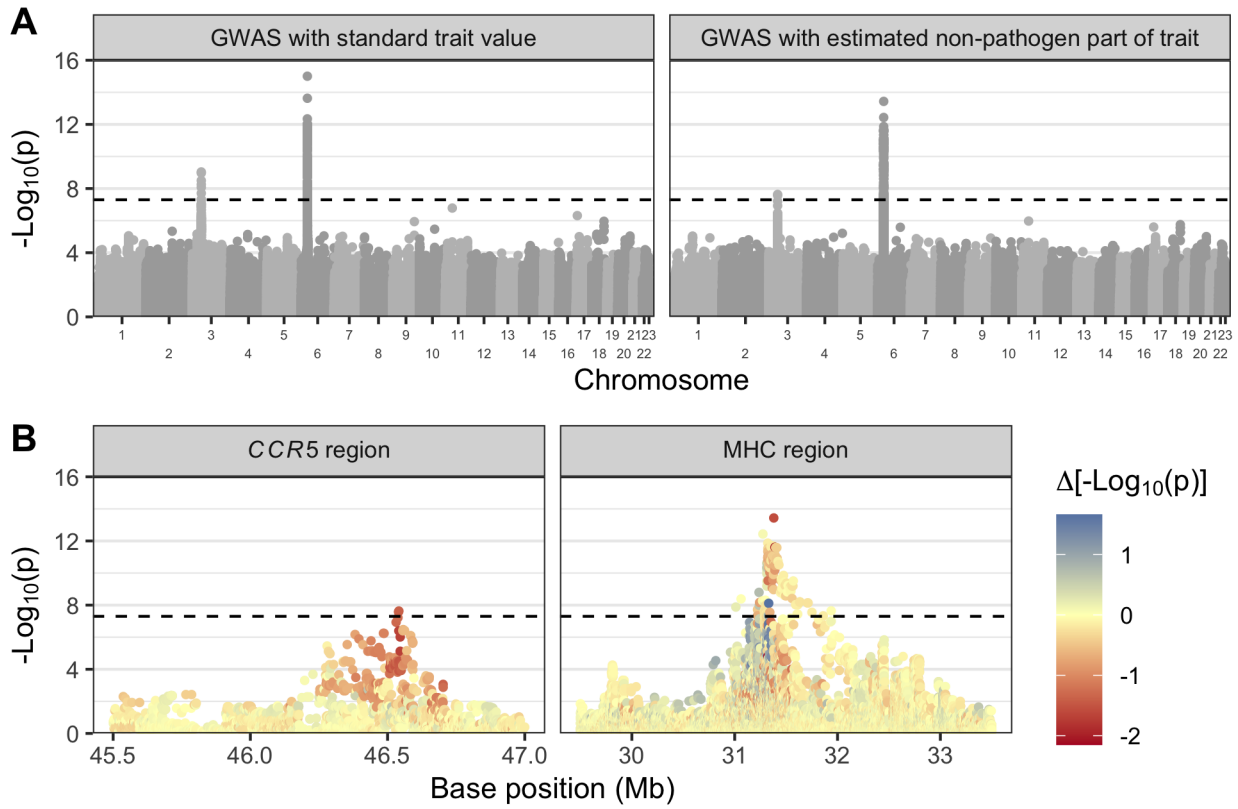

Figure 3: Results from comparative GWAS. (A) shows association p-values for the same host variants from the SHCS cohort in GWAS with two different response variables. On the left, we used unmodified (total) spVL values. On the right, we used our estimates for the non-pathogen effects on spVL. The alternating shades correspond to different chromosomes. (B) compares the strength of association for variants in the *CCR5* and MHC regions between the two GWAS (positions 45.4 - 47Mb on chromosome 3 and 29.5 - 33.5Mb on chromosome 6 for the *CCR5* and MHC, respectively). Base positions are with reference to genome build GRCh37. The color of each point represents the difference in -$\log_{10}$ p-value between the two GWAS. Red means taking into account phylogenetic information decreased the strength of association and blue means it increased it. The dashed lines show genome-wide significance at $p = 5 \times 10^{-8}$.

Table 1: Top association results from McLaren *et al.* (2015) compared to results from this study. Results from this study are for host variants from the SHCS in GWAS with two different response variables. "Standard trait value" means we used the unmodified (total) spVL value and "Estimated non-pathogen part of trait" means we used our estimates for the non-pathogen effects on spVL.

		McLaren et al.	Standard trait value		Estimated non-pathogen part of trait	
Region	Variant	p-value	Effect size	p-value	Effect size	p-value
MHC	rs59440261	2.0×10^{-83}	-0.4	3.3×10^{-11}	-0.22	2.6×10^{-10}
CCR5	rs1015164	1.5×10^{-19}	0.15	7.5×10^{-7}	0.078	8.5×10^{-6}

Discussion

In this paper, we presented a new phylogeny-aware GWAS framework to correct for heritable pathogen effects on infectious disease traits. By using information from the pathogen phylogeny, we show it is possible to improve GWAS power to detect host genetic variants associated with a disease trait. This should help us better understand which host factors are protective against a disease versus which increase susceptibility or disease severity.

Our method relies on the POUMM, a model of continuous trait evolution that accounts for heritable and non-heritable effects on a trait, as well as selection. Using this model, we estimated HIV-1 spVL heritability to be 45% (95% HPD 24 - 67%) in the Swiss HIV Cohort Study. Compared to previous studies, this estimate is at the higher end (see Mitov and Stadler (2018) and references therein). Also using the POUMM, Bertels *et al.* (2018) estimated a spVL heritability of 29% (N = 2014, CI 12 - 46%) from the same cohort and Blanquart *et al.* (2017) estimated 31% (N = 2028, CI 15 - 43%) from a pan-European cohort. We note that our sample size (N = 1493 individuals) is smaller than in these other studies. This might be because we restricted samples based on having *pol* gene sequences with at least 750 non-ambiguous bases. Our aim was to reconstruct a high-quality phylogeny, since the POUMM does not account for phylogenetic uncertainty and the POUMM parameter estimates are key to our downstream trait-correction method. Although our heritability estimate is rather high, the confidence interval largely overlaps that of other studies and we note that estimating heritability per se was not our primary focus.

Instead, the main novelty of our approach was to correct the spVL trait prior to performing the GWAS, thereby estimating and removing pathogen effects. In simulations, we show that when trait heritability amongst infection partners is greater than approximately 25%, GWAS power to detect host genetic variants associated with the same trait is reduced. Our method can correct for this effect in certain evolutionary scenarios by using information from the full pathogen phylogeny.

Based on our simulation results, our method is anticipated to be very useful for disease traits that are highly heritable from donor to recipient and maintain a high correlation between sampled individuals. In simulations, we showed this is the case when heritability is high, selection strength is low, and trait values are not subject to strong stochastic fluctuations. So, cohort-level, phylogenetically structured differences in the measured trait value are necessary for our approach to outperform state of the art methods.

Given our estimates for the heritability of spVL and the selection strength on this trait using Swiss HIV cohort data, our simulation results reveal that we cannot expect a significant improvement in GWAS power for human genetic determinants of spVL (Figure 2). Our method slightly decreases p-values for variants in $CCR5$ and slightly decreases some and increases other p-values for variants in the MHC (Figure 3B). Simulations show we shouldn't expect a net p-value decrease, but our simulations represent an ideal scenario since we simulate under the POUMM. In real life,

un-modeled evolutionary pressures like drug treatment and host-specific HLA alleles might cause the reduced p-values. However, the overall picture is consistent between the two GWAS (Figure 3A). Therefore, we conclude that GWAS for host determinants of HIV-1 subtype B spVL is robust to our correction for pathogen effects.

Our method is convenient for GWAS because it is simply a pre-processing step that produces an alternate response variable for GWAS association tests. It is still possible to use previously developed, well-documented, and fast tools for the actual association testing (we used PLINK (Chang *et al.*, 2015)). The method relies on the freely available R package POUMM (Mitov and Stadler, 2017) and all the code we wrote is available on the project GitHub at https://github.com/cevo-public/POUMM-GWAS. Future applications of our method might investigate other clinically significant disease traits and outcomes that are affected by both host and pathogen genetic factors, for instance Hepatitis B Virus-related hepatocellular carcinoma (An *et al.*, 2018), Hepatitis C treatment success (Ansari *et al.*, 2017), and susceptibility to or severity of certain bacterial infections, e.g. Donnenberg *et al.* (2015); Messina *et al.* (2016).

In summary, we argue that infectious disease GWAS should take the pathogen phylogeny into account when searching for host determinants of a disease trait. We give a practical threshold for identifying when GWAS suffer from pathogen effects (heritability of the trait amongst infection partners $> 25\%$) and provide a method that can help in scenarios where trait values are highly heritable and phylogenetically-structured amongst members of a cohort.

Materials and Methods

Simulation model

Whenever possible, we tried to parameterize our simulation model for spVL using empirical data. We set the total variance in spVL to 0.73 log copies^2 mL^{-2} based on UK cohort data (Mitov and Stadler, 2018). Other studies have estimated slightly lower values though (Table S3). After allotting 25% of this variance to the host part of spVL g_h based on results by McLaren *et al.* (2015), we partitioned the remaining variance between the viral part g_v and the environmental part ϵ in different ratios to assess estimator performance across a range of spVL heritabilities. g_h was simulated as the sum of contributions from 20 causal host genetic variants, 10 of which had an effect size of 0.2 log copies mL^{-1} and 10 of which had an effect size of -0.2 log copies mL^{-1}. Host genetic variants were generated from a binomial distribution with probability p calculated such that g_h had the appropriate variance (see Table 2). We generated a random viral phylogeny with branch lengths on the same time scale as a previously inferred UK cohort HIV tree (Hodcroft *et al.*, 2014) using the R package ape (Paradis and Schliep, 2018). g_v was simulated by running an OU process along the phylogeny using the R package POUMM (Mitov and Stadler, 2017) and sampling values

at the tips. For the OU parameters θ and g_0 we used 4.5 log copies mL^{-1} based on fitting the same model to SCHS data (Table S1). This is similar to values previously inferred for HIV (Table S4). To assess our estimator's performance under a range of evolutionary scenarios, we co-varied the OU parameters for selection strength, α, and intensity of random fluctuations, σ, so that different proportions of the variability in g_v were attributable to selection and drift, respectively. Finally, the environmental component of spVL ϵ was generated from a normal distribution with mean 0. For a full graphical model representation of the simulation scheme, see Figure S1.

Table 2: Simulation model parameters. For a full graphical model representation of the simulation scheme, including how these parameters are related, see Figure S1.

Variable	Expression	Definition
σ_z^2	0.73 log copies^2/mL^2	Total spVL variance
H_h^2	0.25	Host heritability of spVL
$H_{\bar{t}}^2$	varied	Viral heritability of spVL at $\bar{t}$
$\sigma_{g_h}^2$	$\sigma_{g_h}^2 = 0.25 * \sigma_z^2$	Variance in host part of spVL
$\sigma_{g_v}^2(\bar{t})$	$\sigma_{g_v}^2(\bar{t}) = H_{\bar{t}}^2 * \sigma_z^2$	Variance in viral part of spVL at $\bar{t}$
σ_ϵ^2	$\sigma_\epsilon^2 = \sigma_z^2 - \sigma_{g_v}^2 - \sigma_{g_h}^2$	Variance in environmental part of spVL
$\bar{t}$	0.14 substitutions site^{-1} yr^{-1}	Mean root-tip time in viral phylogeny
$\boldsymbol{g}_v$	$\boldsymbol{g}_v \sim Norm(\boldsymbol{\mu}_{OU}, \boldsymbol{\Sigma}_{OU})$	Viral part of spVL for all individuals
θ	4.5 log copies/mL	Optimal spVL value
g_0	4.5 log copies/mL	g_v at the root of the phylogeny
α	varied	Selection strength of OU process
σ	$\sigma = \sqrt{\frac{2\alpha\sigma_{g_v}^2(\bar{t})}{1-exp(-2\alpha\bar{t})}}$	Time-unit standard deviation of OU process
Ψ	branch lengths $\sim Exp(\bar{t})$	Viral phylogeny
g_{h_i}	$g_{h_i} = \delta \sum_{j=1}^{j=M/2} G_{ij} - \delta \sum_{j=M/2}^{j=M} G_{ij}$	Host part of spVL for individual i
$G_{N\times M}$	$G_{ij} \sim Binom(2, p)$ $\forall i \in 1...N, \forall j \in 1...M$	Host genotype matrix
p	$p = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{H_h^2\sigma_z^2}{2\delta^2 M}}$	Host variant allele frequency
δ	0.2	Host variant effect size
M	20	Number of causal host variants
ϵ_i	$\epsilon_i \sim Norm(0, \sigma_\epsilon^2)$	Environmental part of spVL for individual i
N	500	Number of simulated samples

Swiss HIV-1 data

Human genotypes, viral load measurements, and HIV-1 *pol* gene sequences from HIV-1 positive individuals were all collected in the context of other studies by the Swiss HIV Cohort Study (SHCS) (www.shcs.ch, Scherrer *et al.* (2021); Schoeni-Affolter *et al.* (2010)). All participants were HIV-1–infected individuals 16 years or older and written informed consent was obtained from all cohort participants. The anonymized data were made available for this study after the study proposal was approved by the SHCS.

For phylogenetic inference, we retained sequences from 1,493 individuals with non-recombinant subtype B *pol* gene sequences of at least 750 characters and paired RNA measurements allowing for calculation of spVL, as well as 5 randomly chosen subtype A sequences as an outgroup. We used MUSCLE version 3.8.31 (Edgar, 2004) to align the *pol* sequences with –maxiters 3 and otherwise default settings. We trimmed the alignment to 1505 characters to standardize sequence lengths. We used IQ-TREE version 1.6.9 (Nguyen *et al.*, 2014) to construct an approximate maximum likelihood tree with -m GTR+F+R4 for a general time reversible substitution model with empirical base frequencies and four free substitution rate categories. Otherwise, we used the default IQ-TREE settings. After rooting the tree based on the subtype A samples, we removed the outgroup. Viral subtype was determined by the SHCS using the REGA HIV subtyping tool version 2.0 (de Oliveira *et al.*, 2005). We calculated spVL as the arithmetic mean of viral RNA measurements made prior to the start of antiretroviral treatment. For a comparison of several different filtering methods, see Figure S2.

For GWAS, we retained data from 1,392 of the 1,493 SHCS individuals with European ancestry who were not closely related to other individuals in the cohort (Table S5). These were 227 females and 1165 males. Ancestry was determined by plotting individuals along the three primary axes of genotypic variation from a combined dataset of SHCS samples and HapMap populations (Figure S7). Kinship was evaluated using PLINK version 2.3 (Chang *et al.*, 2015); we used the –king-cutoff option to exclude one from each pair of individuals with a kinship coefficient > 0.09375. Initial host genotyping quality control and imputation were done as in Thorball *et al.* (2021). Subsequent genotyping quality control was performed using PLINK version 2.3 (Chang *et al.*, 2015). We used the options –maf 0.01, –geno 0.01, and –hwe 0.00005 to remove variants with minor allele frequency less than 0.01, missing call rate greater than 0.05, or Hardy-Weinberg equilibrium exact test p-value less than 5x10$^-$5. After quality filtering, approximately 6.2 million genetic variants from the 1,392 individuals were retained for GWAS (Table S6).

POUMM parameter inference

We used the R package POUMM version 2.1.6 (Mitov and Stadler, 2017) to infer the POUMM parameters $g_0, \alpha, \theta, \sigma,$ and σ_e from the approximate maximum-likelihood phylogeny and calculated

spVL values. The Bayesian inference method implemented in this package requires specification of a prior distribution for each parameter. We used the same, broad prior distributions as in Mitov and Stadler (2018), namely: $g_0 \sim \mathcal{N}(4.5,\ 3)$, $\alpha \sim Exp(0.02)$, $\theta \sim \mathcal{N}(4.5,\ 3)$, $H_t^2 \sim \mathcal{U}(0,\ 1)$, and $\sigma_e^2 \sim Exp(0.02)$. We ran two MCMC chains for $4\text{x}10^6$ samples each with a target sample acceptance rate of 0.01 and a thinning interval of 1000. The first $2\text{x}10^5$ samples of each chain were used for automatic adjustment of the MCMC proposal distribution. Figure S5 shows the posterior distributions for inferred parameters. Table S1 gives the posterior mean values used for subsequent calculations.

Phylogenetic spVL correction

We corrected calcualted spVL values using the method described in this paper. For each of the 1,392 individuals in the GWAS cohort, we estimated the viral part of spVL using equation 9 and the corresponding non-viral part using equation 12. For the POUMM parameters α, σ, θ, and σ_e, we used the posterior mean estimates generated as described above.

Association testing

We performed two GWAS using the same human genotype data from the SHCS. For the first "GWAS with standard trait value" we used total calculated spVL ($\boldsymbol{z}$) as the response variable for association testing, replicating prior GWAS for host genetic determinants of spVL. For the second "GWAS with estimated non-pathogen part of trait" we replaced total spVL with the estimated non-viral component of spVL ($\hat{\boldsymbol{e}}$) as the response variable. Association testing was performed using a linear association model in PLINK version 2.3 (Chang *et al.*, 2015) with sex and the top 5 principle components of host genetic variation included as covariates. The sex and principle components covariates were included to reduce residual variance in spVL and control for confounding from host population structure, respectively.

Data availability

The simulated data underlying this article can be re-generated using the code available on the project GitHub at https://github.com/cevo-public/POUMM-GWAS. The HIV pathogen genome sequences, clinical data, and human genotypes cannot be shared publicly due to the privacy of individuals who participated in the cohort study. The data may be shared on reasonable request to the Swiss HIV Cohort Study at http://www.shcs.ch.

Acknowledgments

This work was supported by ETH Zurich. We thank the patients who participate in the SHCS; the physicians and study nurses for excellent patient care; A. Scherrer, E. Mauro, and K. Kusejko from the SHCS Data Centre for data management; and D. Perraudin and M. Amstad for administrative assistance. We also thank Michael Landis, who shared a LaTeX template for graphical model drawing.

The members of the SHCS are: Abela I, Aebi-Popp K, Anagnostopoulos A, Battegay M, Bernasconi E, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H, Fux CA, Günthard HF (President of the SHCS), Hachfeld A, Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert CR (Chairman of the Mother Child Substudy), Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Kusejko K (Head of Data Centre), Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nemeth J, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (Chairman of the Scientific Board), Schmid P, Speck R, Stöckle M (Chairman of the Clinical and Laboratory Committee), Tarr P, Trkola A, Wandeler G, Yerly S.

The Swiss HIV Cohort Study is supported by the Swiss National Science Foundation (grant 201369), by SHCS project 858 and by the SHCS research foundation. Furthermore, the SHCS drug resistance database is supported by the Yvonne Jacob Foundation (to HFG). The data are gathered by the Five Swiss University Hospitals, two Cantonal Hospitals, 15 affiliated hospitals and 36 private physicians (listed in http://www.shcs.ch/180-health-care-providers).

References

Alizon, S., von Wyl, V., Stadler, T., Kouyos, R. D., Yerly, S., Hirschel, B., Böni, J., Shah, C., Klimkait, T., Furrer, H., *et al.* 2010. Phylogenetic Approach Reveals That Virus Genotype Largely Determines HIV Set-Point Viral Load. *PLoS Pathogens*, 6(9): e1001123.

An, P., Xu, J., Yu, Y., and Winkler, C. A. 2018. Host and viral genetic variation in HBV-related hepatocellular carcinoma. *Frontiers in Genetics*, 9: 261.

Ansari, M. A., Pedergnana, V., Ip, C. L., Magri, A., Von Delft, A., Bonsall, D., Chaturvedi, N., Bartha, I., Smith, D., Nicholson, G., *et al.* 2017. Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nature genetics*, 49(5): 666–673.

Astle, W. and Balding, D. J. 2009. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, 24(4): 451–471.

Bartha, I., Carlson, J. M., Brumme, C. J., McLaren, P. J., Brumme, Z. L., John, M., Haas, D. W., Martinez-Picado, J., Dalmau, J., López-Galíndez, C., *et al.* 2013. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *eLife*, 2: e01123.

Bartha, I., McLaren, P. J., Brumme, C., Harrigan, R., Telenti, A., and Fellay, J. 2017. Estimating the Respective Contributions of Human and Viral Genetic Variation to HIV Control. *PLoS Computational Biology*, 13(2): e1005339.

Bertels, F., Marzel, A., Leventhal, G., Mitov, V., Fellay, J., Günthard, H. F., Böni, J., Yerly, S., Klimkait, T., Aubert, V., *et al.* 2018. Dissecting HIV Virulence: Heritability of Setpoint Viral Load, CD41 T-Cell Decline, and Per-Parasite Pathogenicity. *Molecular biology and evolution*, 35(1): 27–37.

Blanquart, F., Wymant, C., Cornelissen, M., Gall, A., Bakker, M., Bezemer, D., Hall, M., Hillebregt, M., Ong, S. H., Albert, J., *et al.* 2017. Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. *PLoS Biology*, 15(6): e2001855.

Bonhoeffer, S., Fraser, C., and Leventhal, G. E. 2015. High Heritability Is Compatible with the Broad Distribution of Set Point Viral Load in HIV Carriers. *PLoS Pathogens*, 11(2): e1004634.

Butler, M. A. and King, A. A. 2004. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American naturalist*, 164(6): 683–695.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1): 7.

Collins, C. and Didelot, X. 2018. A Phylogenetic Method To Perform Genome-Wide Association Studies In Microbes That Accounts For Population Structure And Recombination. *PLoS Computational Biology*, 14(2): e1005958.

Dalmasso, C., Carpentier, W., Meyer, L., Rouzioux, C., Goujard, C., Chaix, M.-L., Lambotte, O., Avettand-Fenoel, V., Le Clerc, S., de Senneville, L. D., *et al.* 2008. Distinct Genetic Loci Control Plasma HIV-RNA and Cellular HIV-DNA Levels in HIV-1 Infection: The ANRS Genome Wide Association 01 Study. *PLoS ONE*, 3(12): e3907.

de Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E. J., Wensing, A. M. J., van de Vijver, D. A., *et al.* 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, 21(19): 3797–3800.

Donnenberg, M. S., Hazen, T. H., Farag, T. H., Panchalingam, S., Antonio, M., Hossain, A., Mandomando, I., Ochieng, J. B., Ramamurthy, T., Tamboura, B., *et al.* 2015. Bacterial Factors Associated with Lethal Outcome of Enteropathogenic Escherichia coli Infection: Genomic Case-Control Studies. *PLOS Neglected Tropical Diseases*, 9(5): e0003791.

Dudbridge, F. 2013. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9(3).

Earle, S. G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N. C., Walker, T. M., Spencer, C. C. A., Iqbal, Z., Clifton, D. A., Hopkins, K. L., *et al.* 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology*, 1: 16041.

Edgar, R. C. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5: 113.

Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., *et al.* 2007. A whole-genome association study of major determinants for host control of HIV-1. *Science*, 317(5840): 944–947.

Fellay, J., Ge, D., Shianna, K. V., Colombo, S., Ledergerber, B., Cirulli, E. T., Urban, T. J., Zhang, K., Gumbs, C. E., Smith, J. P., *et al.* 2009. Common Genetic Variation and the Control of HIV-1 in Humans. *PLoS Genetics*, 5(12): e1000791.

Fraser, C., Lythgoe, K., Leventhal, G. E., Shirreff, G., Hollingsworth, T. D., Alizon, S., and Bonhoeffer, S. 2014. Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. *Science*, 343(6177): 1243727.

Hodcroft, E., Hadfield, J. D., Fearnhill, E., Phillips, A., Dunn, D., O'Shea, S., Pillay, D., Leigh Brown, A. J., Study, o. b. o. t. U. H. D. R. D., and the UK CHIC 2014. The Contribution of Viral Genotype to Plasma Viral Set-Point in HIV Infection. *PLoS Pathogens*, 10(5): e1004112.

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., and Huelsenbeck, J. P. 2014. Probabilistic Graphical Model Representation in Phylogenetics. *Syst. Biol*, 63(5): 753–771.

Housworth, E. A., Martins, E. P., and Lynch, M. 2004. The Phylogenetic Mixed Model. *The American Naturalist*, 163(1): 84–96.

Kløverpris, H. N., Leslie, A., and Goulder, P. 2016. Role of HLA adaptation in HIV evolution. *Frontiers in Immunology*, 6: 665.

Leventhal, G. E. and Bonhoeffer, S. 2016. Potential Pitfalls in Estimating Viral Load Heritability. *Trends in Microbiology*, 24(9): 687–698.

McLaren, P. J., Ripke, S., Pelak, K., Weintrob, A. C., Patsopoulos, N. A., Jia, X., Erlich, R. L., Lennon, N. J., Kadie, C. M., Heckerman, D., *et al.* 2012. Fine-mapping classical HLA variation associated with durable host control of HIV-1 infection in African Americans. *Human Molecular Genetics*, 21(19): 4334–4347.

McLaren, P. J., Coulonges, C., Bartha, I., Lenz, T. L., Deutsch, A. J., Bashirova, A., Buchbinder, S., Carrington, M. N., Cossarizza, A., Dalmau, J., *et al.* 2015. Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47): 14658–63.

Mellors, J. W., Rinaldo, C. R., Gupta, P., White, R. M., Todd, J. A., and Kingsley, L. A. 1996. Prognosis in HIV-1 Infection Predicted by the Quantity of Virus in Plasma. *Science*, 272(5265): 1167–1170.

Messina, J. A., Thaden, J. T., Sharma-Kuinkel, B. K., and Fowler, V. G. 2016. Impact of Bacterial and Human Genetic Variation on Staphylococcus aureus Infections. *PLOS Pathogens*, 12(1): e1005330.

Mitov, V. and Stadler, T. 2017. POUMM: An R-package for Bayesian Inference of Phylogenetic Heritability. *ArXiv*.

Mitov, V. and Stadler, T. 2018. A Practical Guide to Estimating the Heritability of Pathogen Traits. *Molecular Biology and Evolution*, 35(3): 756–772.

Naret, O., Chaturvedi, N., Bartha, I., Hammer, C., and Fellay, J. 2018. Correcting for Population Stratification Reduces False Positive and False Negative Results in Joint Analyses of Host and Pathogen Genomes. *Frontiers in Genetics*, 9: 266.

Nguyen, H., Thorball, C. W., Fellay, J., Böni, J., Yerly, S., Perreau, M., Hirsch, H. H., Kusejko, K., Thurnheer, M. C., Battegay, M., *et al.* 2021. Systematic screening of viral and human genetic variation identifies antiretroviral resistance and immune escape link. *eLife*, 10.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1): 268–274.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., *et al.* 2014. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488): 376–381.

Panel on Antiretroviral Guidelines for Adults and Adolescents 2019. Guidelines for the Use of Antiretroviral Agents in HIV-1-Infected Adults and Adolescents Developed by the HHS Panel on Antiretroviral Guidelines for. Technical report, U.S. Department of Health and Human Services.

Paradis, E. and Schliep, K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35: 526–528.

Pelak, K., Goldstein, D., Walley, N., Fellay, J., Ge, D., Shianna, K., Gumbs, C., Gao, X., Maia, J., Cronin, K., *et al.* 2010. Host Determinants of HIV-1 Control in African Americans. *The Journal of Infectious Diseases*, 201(8): 1141–1149.

Pereyra, F., Jia, X., McLaren, P. J., Telenti, A., De Bakker, P. I., and Walker, B. D. 2010. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*, 330(6010): 1551–1557.

Petersen, K. B. and Pedersen, M. S. 2012. *The Matrix Cookbook*. Technical University of Denmark.

Power, R. A., Parkhill, J., and de Oliveira, T. 2017. Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics*, 18(1): 41–50.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8): 904–909.

Quinn, T. C., Wawer, M. J., Sewankambo, N., Serwadda, D., Li, C., Wabwire-Mangen, F., Meehan, M. O., Lutalo, T., and Gray, R. H. 2000. Viral Load and Heterosexual Transmission of Human Immunodeficiency Virus Type 1. *New England Journal of Medicine*, 342(13): 921–929.

Scherrer, A. U., Traytel, A., Braun, D. L., Calmy, A., Battegay, M., Cavassini, M., Furrer, H., Schmid, P., Bernasconi, E., Stoeckle, M., *et al.* 2021. Cohort Profile Update: The Swiss HIV Cohort Study (SHCS). *International Journal of Epidemiology*, 2021: 1–12.

Schoeni-Affolter, F., Ledergerber, B., Rickenbach, M., Rudin, C., Günthard, H. F., Telenti, A., Furrer, H., Yerly, S., and Francioli, P. 2010. Cohort profile: The Swiss HIV cohort study. *International Journal of Epidemiology*, 39(5): 1179–1189.

Thorball, C. W., Oudot-Mellakh, T., Ehsan, N., Hammer, C., Santoni, F. A., Niay, J., Costagliola, D., Goujard, C., Meyer, L., Wang, S. S., *et al.* 2021. Genetic variation near CXCL12 is associated with susceptibility to HIV-related non-Hodgkin lymphoma. *Haematologica*, 106(8): 2233–2241.

van Manen, D., Delaneau, O., Kootstra, N. A., Boeser-Nunnink, B. D., Limou, S., Bol, S. M., Burger, J. A., Zwinderman, A. H., Moerland, P. D., van 't Slot, R., *et al.* 2011. Genome-Wide

Association Scan in HIV-1-Infected Individuals Identifying Variants Influencing Disease Course. *PLoS ONE*, 6(7): e22208.