bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Potential pitfalls in estimating viral load heritability

Gabriel E. Leventhal[1,2] **and Sebastian Bonhoeffer**[2]
1. Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA, 02139, U.S.A.
2. Institute of Integrative Biology, ETH Zürich, 8092 Zürich, Switerland

GEL: gaberoo@mit.edu
SB: sebastian.bonhoeffer@env.ethz.ch

Keywords: heritability, phylogenetics, virus evolution, HIV, set-point viral load

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

In HIV patients, the set-point viral load (SPVL) is the most widely used predictor of disease severity. Yet SPVL varies over several orders of magnitude between patients. The heritability of SPVL quantifies how much of the variation in SPVL is due to transmissible viral genetics. There is currently no clear consensus on the value of SPVL heritability, as multiple studies have reported apparently discrepant estimates. Here we illustrate that the discrepancies in estimates are most likely due to differences in the estimation methods, rather than the study populations. Importantly, phylogenetic estimates run the risk of being strongly confounded by unrealistic model assumptions. Care must be taken when interpreting and comparing the different estimates to each other.

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Searching for set-point viral load heritability in HIV infection

During the asymptomatic phase of HIV infection, the viral load within a patient fluctuates around a relatively stable value known as the set-point viral load. Set-point viral load (SPVL) has proven to be highly relevant, as untreated patients with higher SPVL tend to progress to AIDS faster than those with low SPVL, and consequentially SPVL is one of the most widely used predictors of disease severity [1, 2]. What is particularly striking, however, is the amount of variation in SPVL between patients: SPVL varies over several orders of magnitude between patients [1, 3–5]. Understanding the source of this wide variation in SPVL in the patient population is key to understanding HIV pathogenicity and why certain patients progress to AIDS rather quickly, while others progress much more slowly or not at all.

The potential contributions to SPVL variation can generally be split into four categories, determined by:

(i) the specific genetic viral variant that infects a host;
(ii) the host genetics;
(iii) any interactions between host and viral genetics;
(iv) other extrinsic factors, both independent of and interacting with host and viral genetics.

From a virus-centric view, we can group contributions (ii)-(iv) together, and hence the question becomes to what degree the variation in HIV genetics explains the variation in SPVL in the patient population, and by proxy how much viral genetics control the variation HIV pathogenicity.

New HIV infections can occur after a virus is transmitted from a donor host to a recipient host. Thus, from an evolutionary point of view, viral genetic information is passed on and conserved from one infection to the next, whereas donor and recipient host genetics are typically unrelated. In evolutionary theory, the concept of heritability quantifies precisely the question at hand [6]: How much of the observed variation in a trait is explained by variation in the genetics that are passed on to the next generation? A heritability of 100% means that all of the trait variation is explained by transmissible genetic information, while a heritability of 0% means that transmissible genetics explain none of the trait variation. A number of recent publications have reported heritability estimates based on different methods that range from as low as 5% to as high as 50% [7–14]. These studies used partly different methods to estimate heritability, depending on the available data. The crucial question thus arises as to whether these discrepant estimates reflect true differences in the study populations, or whether the differences may rather be artifacts of the estimation method used.

Every statistical model comes with a set of assumptions that must be fulfilled in order for the results to be meaningful. Here, we aim to illustrate how some of these estimates of heritability might be confounded by model assumptions that do not completely conform to the dynamics of HIV transmission. Nevertheless, even when the assumptions are fulfilled, the choice of estimation method alone can potentially lead to differences in the heritability estimates.

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Estimating the heritability of a viral trait

The concept of heritability was initially developed to measure the degree of correspondence between the trait value in an offspring and the trait value of its parents, and can be traced back to Galton [15], though the first use of the term "heritability" remains elusive [16]. Most of the credit for developing the methods to estimate heritability go back to Fisher [17] and Wright [18], although the common contemporary use corresponds to "narrow sense heritability" as defined by Lush [19]: Heritability is the ratio of additive genetic variance, σ_g^2, and the sum of additive genetic and environmental variance, σ_e^2, in a population,

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2},$$

where the total phenotypic variance, $\sigma_\phi^2 = \sigma_g^2 + \sigma_e^2$, is the sum of all contributions to variance. Heritability can then be estimated from data of related individuals (e.g. parents and their offspring) using regression methods [6, 17, 18]. For the case of SPVL in HIV, σ_g^2 corresponds to the transmissible contribution (i) and σ_e^2 comprises the non-transmissible contributions (ii)–(iv).

Analogous to these methods that were developed for sexually reproducing individuals, the trait variance of clonally reproducing organisms (such as viruses) can be partitioned into variance components. In viral infections, only the virus genotype is passed on from one infection to the next, and thus the heritability of a viral trait as defined by h^2 can be used to quantify the relative influence of the variation in transmissible viral factors to the trait of interest in a population [8, 20]. While the general framework of partitioning trait variance is the same in sexual and clonal populations, the underlying processes that generated the observed variance are different. Thus, although the methods to estimate heritability in sexual populations can generally be directly applied to viral heritability, care must be taken when evaluating the applicability of the methods.

Furthermore, whenever heritability is estimated, it is also important to remember that even though h^2 can be used to characterize the propensity of virus genetics to influence a trait, the actual measured heritability depends on both the expressed variation of this trait in a given population, as well as the variation in other factors that influence the trait. Therefore, h^2 strongly depends on the population it is measured in, and can consequently vary between study populations. This does not invalidate the usefulness of measuring h^2, but rather means that predictions on the basis of h^2 are primarily valid in the study population at hand and care must be taken when using h^2 for predictions in other populations.

Parent-offspring regression

Parent-offspring regression is tightly linked to the original definition of heritability and is thus the most straightforward estimation method [17, 18]. The idea is to compare trait values in parents to trait values in their respective offspring, or in the case of viral infections compare traits from donors and recipients within transmission pairs. The applicability of "donor-recipient" regression (DR) in the context of viral heritability has been recently reviewed in Fraser *et al.* [8]. In short, DR assumes that viral traits are imperfectly

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

passed on from one infection to the next and determines the predictive ability of SPVL in a donor of SPVL in the recipient. Differences in SPVL between donors and recipients must then either be due to non-transmissible contributions of type (ii)-(iv), or due to genetic changes that accumulate as a result of intra-host evolution or genetic bottlenecks at transmission [20]. Viral heritability of SPVL is then the regression slope of recipient SPVL on donor SPVL, akin to parent-offspring regression [6].

The biggest difficulty in estimating heritability using donor-recipient regression lies in observing direct transmission between pairs. Currently, the best source of such transmission pairs are discordant couple study cohorts [8]. But even then, there is the possibility that infection of the seronegative partner occurred from an unknown third-party. Furthermore, such studies are restricted to specific study populations that may not be representative of the general global HIV epidemic. Non-independence between different donor-recipient pairs such as a mutual infector of the donors may further distort estimates.

Phylogenetic methods

To overcome the shortage of appropriate study populations, considerable hope has been placed in estimating heritability from much more easily available sequence information using phylogenetic methods. Various phylogenetic methods have previously been employed to estimate the heritability of SPVL [7, 9, 21]. All of these methods have been developed in the context of ecological species and have yet to be validated for use on viral populations. The methods used include non-parametric estimates of phylogenetic signal [22], as well parametric regression methods such as independent contrasts [23], the phylogenetic mixed model [24], and Pagel's λ [25]. Non-parametric methods can be indicative of the level of signal in a dataset, but are generally less useful when the goal is a quantitative estimate of heritability. The parametric regression methods are in essence an extension of donor-recipient regression that can be applied to datasets where the relation between individuals goes beyond parents and offspring or direct transmission pairs. A consequence of this extension is that a process model is required that links the observed genetic differences to the number of transmission events and the time subjected to intra-host evolution. All of the above methods assume that the genetic differences accumulate neutrally through time, irrespective of the number of transmission events or selection on the trait of interest, which generally is not the case for HIV. Hence, the reported estimates must therefore be interpreted together with such model assumptions. Despite being developed from different first principles, in certain cases Pagel's λ is equivalent to the Phylogenetic Mixed Model (Box 1). In the following sections we thus only consider the Phylogenetic Mixed Model (PMM), which is, as shown in Box 1, the most generic and appropriate method for viral heritability estimation.

Heritability estimation in simulated populations

In order to illustrate the effects of incorrect model assumptions about the evolutionary process underlying the phylogenetic methods, we compare the different estimation methods using mock data from simulations (Fig. 1). While simulated data have previously been

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

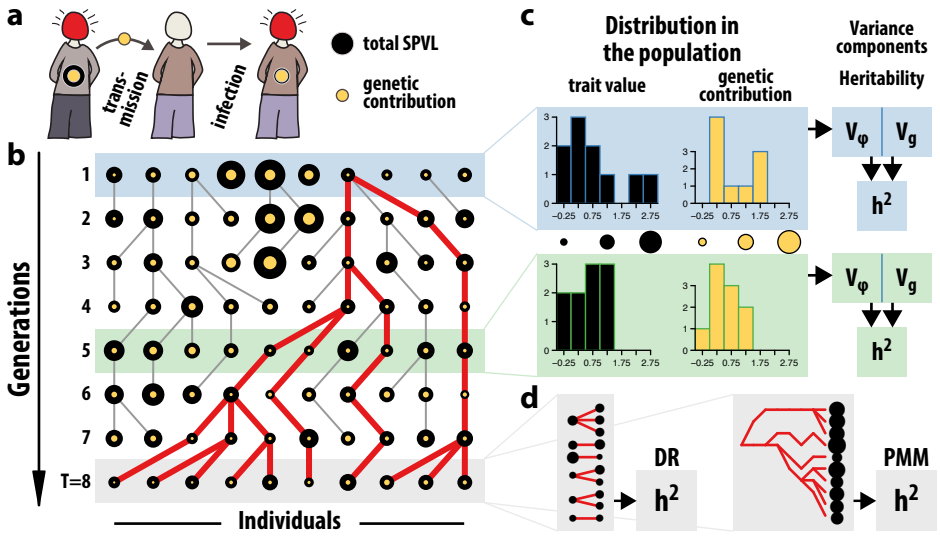

Figure 1: **Estimating heritability in a simulated Wright-Fisher population. (a)** At transmission, only the genetic contribution to SPVL is passed on from donor to recipient. Differences in host contribution result in differences in SPVL between donors and recipients with the same genetic contribution. **(b)** At every generation, parents are chosen from the previous generation. The circle size of each individual is representative of its trait value. The genetic component (yellow circles) is passed on to the offspring including potential genetic changes due to intra-host evolution. The red lines show the full ancestry of the individuals in the last generation. **(c)** The distribution of the trait in the population may change from generation to generation and is shown for two particular generations. The heritability in a population at a specific generation time is the genetic variance divided by the trait variance. **(d)** Alternatively, the heritability in the final generation can be estimated using either a donor-recipient regression or a phylogenetic method.

used to validate the respective heritability estimates for SPVL in HIV [7, 9, 26], the simulation models differ between publications and generally do not adequately represent key aspects of the evolution of SPVL across infections. Here, we use a simple previously published model for HIV transmission across many generations in a Wright-Fisher population that accounts for transmission bottlenecks and intra-host evolution [20].

An individual's log SPVL is assumed to be a linear combination of both a transmissible component, g, and a non-transmissible component, e, such that SPVL of individual i at generation t is $\phi_i(t) = g_i(t) + e_i(t)$. For each individual i in the subsequent generation, a donor, $d(i)$, is chosen from the current generation, either uniformly (drift) or based on their value of SPVL (selection). The genetic contribution of the donor is then transferred from to the recipient, $g'_i(t+1) = g_{d(i)}(t)$, including an potential additional change due to intra-host evolution that follows a normal distribution with variance σ_i^2. Because the population is finite, there is a generation T at which the whole population will have descended from the same individual (Fig. 1), and the phylogeny or pedigree of the population can be traced back to the common ancestor. Thus we can compare heritability estimates in this final generation obtained using DR regression or the PMM.

We first consider the case where SPVL does not influence the likelihood of transmission, and hence there is no selection on SPVL between hosts, since this is a key assumption underlying the PMM. Note that the distribution in the genetic contribution to SPVL

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

may still change from generation to generation due to drift. The heritability estimated by donor-recipient regression is consistent with the fraction of variances in the final generation, $h^2 = \sigma_g^2(T)/\sigma_\phi^2(T)$ (Fig. 2a). Heritability estimates found using the PMM, however, do not estimate the ratio of variances, but rather the amount of genetic variance that would be expected from a drift process that has run for T generations, with an increment in variance per generation, σ_i^2, equal to that produced by intra-host evolution (Fig. 2b). The expected amount of genetic variance at generation T thus is $\sigma_g^2 = T\,\sigma_i^2$, which is generally not equal to the variance estimate in the final generation. In addition to estimating differing quantities, PMM also strongly underperforms in in precision compared to DR regression when only a fraction of the population is sampled (Fig. 2c).

Next we consider the case where SPVL influences the likelihood of transmission of an individual, as is known for HIV [5]. We compare both the PMM and DR to the population heritability, i.e. the fraction of variance components, as there is no equivalent process based interpretation of heritability for a non-neutral trait. Because DR regression only depends on a single transmission event, the estimate is robust to the selection of donors based on SPVL (Fig. 2d; see also [8]). Heritability estimates obtained using the PMM generally break down if the population evolved under selection on the trait of interest. This can lead to an unpredictable underestimation or overestimation bias, and also misleading narrow confidence intervals (Fig. 2d). Such overconfidence in biased estimates is typical of fitting parametric models. A certain parameter value might very clearly give the best fit of the underlying process to the data and hence that parameter value has narrow confidence intervals. However, the actual model process might still be a poor fit to the data. For a population evolving under selection on the trait of interest, this generally leads to an underestimation of the genetic contribution to variance, σ_g^2, when using the PMM (Fig 2e). However, depending on the esimate of the overall variance, σ_ϕ^2, under the PMM (Box 1), this unpredictably leads to either overestimation or underestimation of heritability under the PMM (Fig 2d).

Comparing donor-recipient and phylogenetic estimates

As we have illustrated, estimates from DR regression and the PMM cannot be immediately compared to each other. Donor-recipient regression estimates the heritability in the current population under the assumption that all individuals are independent. This corresponds to calculating the sample variance of the genetic contributions of all individuals in the populations, and dividing by the sample variance of the phenotype of all individuals in the population. It is therefore a measure of the sampled population at the current point in time, irrespective of the evolutionary history. In this sense, donor-recipient regression is robust to variations in the evolutionary model that generated the genetic variance in the population. The phylogenetic mixed model estimates the process heritability under the assumption that the genetic contribution to the variance in the population increased linearly through time, i.e. through drift. This corresponds to calculating the expected variance in a population that has evolved from a common ancestor at some time T in the past. This latter heritability is therefore a measure of the process that gave rise to the current population and

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

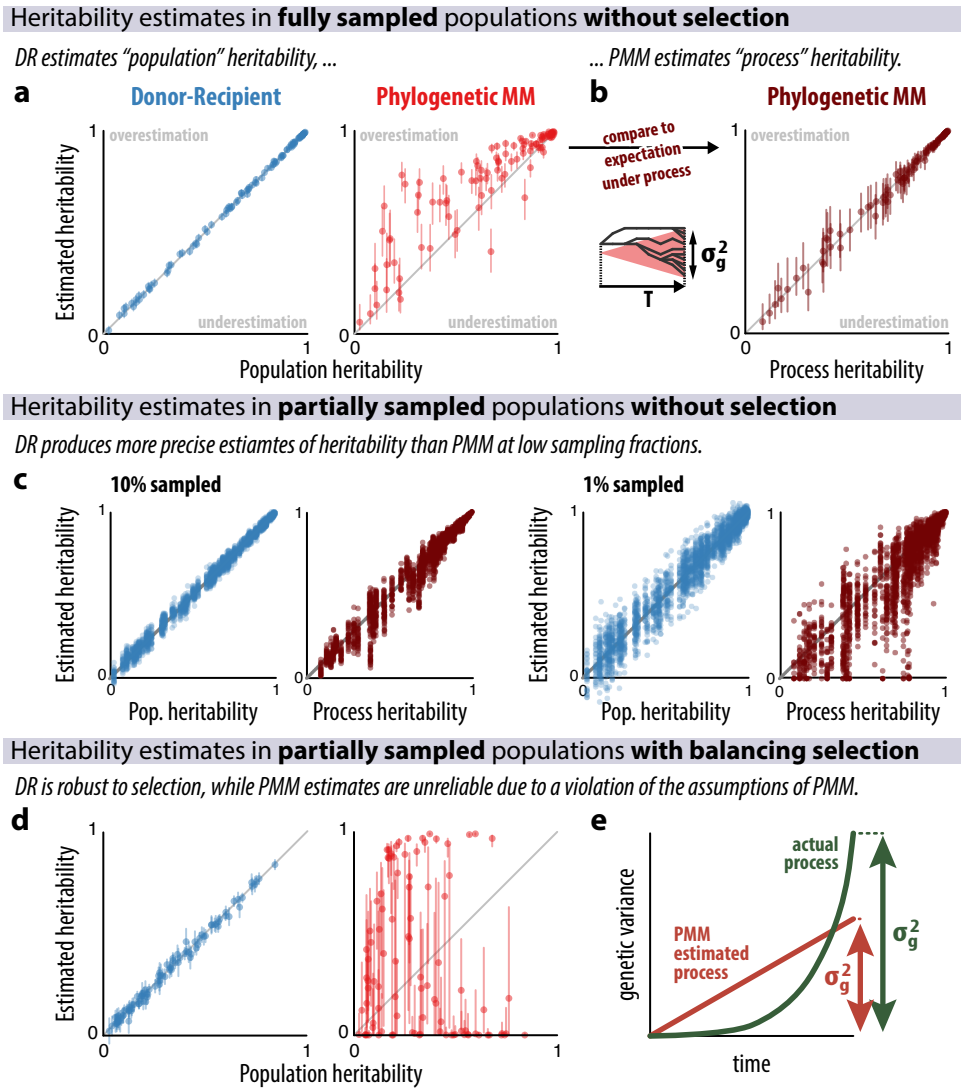

Figure 2: **Heritability estimation using donor-recipient regression and the Phylogenetic Mixed Model.** Wright-Fisher populations of 10000 individuals were simulated as described in Fig. 1. At each generation, the genetic contribution in recipients is chosen from a normal distribution centered on the value of the donor with variance σ_i^2 (see also [20]). **a.** Maximum likelihood heritability estimates of DR regression (blue) and the PMM (red) compared to population heritability as defined by the ratio of variances. Vertical lines indicate the profile likelihood confidence intervals. **b.** The PMM consistently estimates the process heritability based on the expected genetic variance at generation T, given a per-generation incremental increase of σ_i^2 (dark red). **c.** Estimates from DR regression remain robust even at low sampling levels of 10% and 1%, compared to the PMM. DR and PMM estimates are compared to the population and process heritability, respectively. **d.** Heritability estimates using DR and PMM when SPVL influences the likelihood of transmission. **e.** The PMM estimates a linear variance generating process, which can be different from the actual process that generated the variance. For **a-c**, 100 random parameter sets were uniformly chosen from $\log \sigma_g^2 \in [-0.3, 1]$, $\log \sigma_e^2 \in [-0.3, 0.3]$, $\log \sigma_i^2 \in [-2.3, -1]$. For **d**, 100 random parameter sets were chosen acording to the true transmission potential for HIV [5]: $\sigma_o^2 \in [0.5, 1.5]$, $\log \sigma_i^2 \in [-1.3, 0]$, $\log \sigma_e^2 \in [-0.3, 0.3]$, $\log \sigma_g^2 \in [0.7, 2]$.

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

thus depends on the evolutionary history of the population. Hence, a mismatch between the evolutionary process that is assumed in the PMM and the true process that occurred can strongly confound the estimates in an unpredictable way. Viral phylogenies reconstructed from real data are also mostly non-ultrametric and consequently the process time T has to be chosen arbitrarily, which can further affect heritability estimates (see Box 2).

Which interpretation of heritability we are interested in depends on how h^2 is used for prediction. One common use of h^2 is to determine the amount of genetic variation that can be potentially found. A large h^2 increases the statistical power of GWAS studies to detect common genetic variants that influence a trait [27, 28]. Another common use of h^2 is to predict the response to selection [6]. A classic result in population genetics relates the magnitude of change in phenotype in the subsequent generation to the product of the heritability and the strength of selection. The larger h^2, the larger the effect of selection on a population. Both of these uses are only concerned with genetic variance in the current population, irrespective of its evolutionary history. Therefore, h^2 estimates from DR studies are most appropriate. We can also make use of heritability estimates to understand the evolution of HIV. From one infection to the next, two main processes can contribute to genetic changes: First, intra-host evolution can result in the accumulation of genetic changes in the viral population that may then be transmitted. Second, transmission bottlenecks may select a minority variants from the viral population which is then transmitted. When the assumptions about the process are correct, then phylogenetic mixed model can estimate the rate at which such transmissible genetic changes take place.

Concluding remarks and outlook

The discrepancy in reported estimates has lead to some confusion about the reliability and usefulness of estimating heritability of SPVL. On the one hand, estimates using DR regression in various study populations from Africa and Europe produce consistent estimates (Tab. 1). On the other hand, estimates based on phylogenetic trees suggest potentially large differences between studies. However, these latter phylogenetic estimates are difficult to compare, as only Hodcroft *et al.* [9] employ the PMM, while the remaining studies report Pagel's λ and other measures [7, 21, 26], from which heritability cannot readily be derived for non-ultrametric trees (see Box 1). Furthermore, the simulation studies used to validate these other measures do not produce the actual heritability that results from explicitly accounting for the transmissible genetic contribution to SPVL as we do here.

This leads us to speculate that most of the differences in heritability estimates come from differences in the estimation methods, rather than actual differences in the data. One must thus decide, which of the methods produces estimates that are more relevant to the research questions being asked. The increasing amount of available viral sequence data presents an intriguing opportunity to radically increase the number of data points. However, larger n does not guarantee better estimates if phylogenetic methods are used instead of donor-recipient regression. In our mock data, we see that phylogenetic methods potentially need at least one to two orders of magnitude more data to achieve similar confidence in the estimates. More importantly, the model assumptions that accompany the phylogenetic methods cannot be

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Study	N	h^2	95% CI
DR: discordant couples cohorts			
Hollingsworth *et al.* [13]	97	0.36^*	$[0.06, 0.66]$
Lingappa *et al.* [11]	141	0.44	$[0.19, 0.69]$
Yue *et al.* [14]	195	0.26	$[0.08, 0.44]$
DR: pairs identified by interview			
Hecht *et al.* [12]	23	$0.41^\dagger$	$[0.07, 0.75]$
van der Kuyl *et al.* [10]	56	$0.23^\dagger$	$[-0.02, 0.49]$
PMM			
Hodcroft *et al.* [9]	8384	0.06	$[0.03, 0.09]$

Table 1: **Reported heritability estimates.** *Here we report the re-estimate of heritability based on the DR regression slope from [8]. †The studies by Hecht *et al.* [12] and van der Kuyl *et al.* [10] reported Pearson correlation coefficients, rather than regression slopes. For these studies, we extracted the donor and recipient $\log_{10}$ SPVL values for the published figures (Fig. 2 in [12] and Fig. 1 in [10]), and recalculated the estimates of the regression slope. The precision of the estimates is thus dependent on the precision of the plotted points in the figures.

blindly ignored.

In HIV infection, a higher viral load results in a higher likelihood of transmission. The viral load in a patient varies over time, with the largest overall difference occurring between the relatively high viral load during the acute phase compared to the set-point viral load at the beginning of the chronic asymptomatic phase. Thus, whether this link between viral load and transmission rate translates into between-host selection on *set-point* viral load is not yet fully resolved and may depend on the relative contributions of onward transmission during the acute and the chronic phases. Yet, there is nevertheless evidence of a certain degree of balancing selection for optimal levels of SPVL in HIV on a between host level [5, 29, 30].

A proper estimation of heritability using phylogenetic trees would require extensions to the PMM that incorporate the relevant aspects of the true evolutionary process. For example, constant stabilizing selection on SPVL could be accounted for using an Ornstein-Uhlenbeck (OU) process [31]. This has previously been applied to study macro-evolution on ultrametric trees [32], and current efforts are underway to develop an OU-based method for non-ultrametric pathogen trees (personal communication, Venelin Mitov & Tanja Stadler, Apr 2016). As stabilising selection on SPVL has shown to occur [5, 29, 30], this may represent a major step forward in estimating heritability using phylogenetic methods. Yet in cases where the selective process is unknown or may have fluctuated over the evolutionary history, constructing an appropriate evolutionary model may nevertheless be challenging. Likelihood-free methods such as Approximate Bayesian Computation (ABC) have the advantage of being rather flexible in terms of the underlying evolutionary process. Such an approach has previously been applied to constrain the magnitude of selection and genetic change that would be compatible with the observed distribution of SPVL in patient populations without taking the phylogenetic relationships into account [20]. Incorporating phy-

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

logenetic information into such an ABC approach is currently not possible, as we still lack sufficiently large systematic HIV sequence data. Current large-scale collaborative projects such as BEEHIVE[(i)] promise to fill this data gap.

Finally, throughout this paper we have assumed that the phylogenetic tree exactly represents the genetic relatedness of viral isolates from patients. In reality, phylogenetic trees are reconstructed from sequence data and are thus only estimates of the true phylogenetic tree. It is difficult to predict in which way errors in the tree might bias heritability estimates, but incorporating the uncertainty in the tree into the uncertainty in the heritability estimates is advisable and can readily be done in a Bayesian framework [21]. Overall, estimates from phylogenetic methods cannot be readily preferred over estimates from donor-recipient studies without properly scrutinizing the adequacy of the model assumptions that underlie the estimation method and taking into account additional noise and biases that may arise from the phylogenetic tree reconstruction.

Acknowledgements

The authors thank Venelin Mitov, Tanja Stadler, Christophe Fraser, Samuel Alizon, Roland Regoes and Frederic Bertels for constructive discussions.

Resources

(i) BEEHIVE: Bridging the Evolution and Epidemiology of HIV in Europe, `https://erc.europa.eu/projects-and-results/erc-funded-projects/beehive`. Last accessed: 12 Apr 2016.

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Glossary

Asymptomatic phase HIV infections are generally split into three characteristic stages: (I) Primary infection/Acute phase; (II) Chronic asymptomatic phase; (III) AIDS phase.

Viral load The density of virus in the blood of a patient. It is a proxy for the amount of virus in the rest of the body.

Set-point viral load The viral load during the asymptomatic phase fluctuates around a remarkably stable level, the set-point viral load.

Genetic variance Amount of variance in the trait (e.g. SPVL) that is due to differences in transmissible viral genetics.

Environmental variance Amount of variance in the trait (e.g. SPVL) that is due to anything other then viral genetics.

Viral heritability The fraction of phenotypic variance that is explained by transmissible genetic factors.

Donor-recipient regression Method to estimate heritability by regressing the trait values in the recipients on the traits values of the donors.

Seronegative Negative for HIV infection.

Serodiscordant couples Sexual partnerships where only a single individual is infected.

Pedigree Ancestral tree linking parents to their offspring.

Genetic bottlneck Sudden decrease in population size where only a few genetic variants are selected. This leads to a drastic decrease in genetic variation in the population.

Genetic drift Change in the genotype distribution over time due to the finite size of a population.

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Box 1: Equivalence between phylogenetic methods

For a single trait value evolving along a tree, it is straightforward to derive the maximum likelihood estimators for Pagel's λ (PL) and the Phylogenetic Mixed Model (PMM) at a given level of heritability. To do this, both PL and the PMM assume that the sampled trait values are distributed according to a multivariate normal distribution with covariance matrix $\mathbf{\Sigma}(h^2)$ or $\mathbf{\Sigma}(\lambda)$ for PMM and PL, respectively. Multivariate normal maximum likelihood estimators for both the mean, μ, and the overall variance, σ_ϕ^2, of a fixed covariance structure, $\mathbf{\Sigma}(h^2)$, for a specific value of h^2 are easily derived (Box 3). The maximum likelihood estimate of h^2 can then be found by subsequently numerically maximizing the likelihood over values of h^2 (or λ for PL). Despite following the same statistical framework, there are small differences in the derivation of $\mathbf{\Sigma}$ for PL and the PMM.

The Phylogenetic Mixed Model The idea of PMM is that the phenotype, ϕ_i, of individual i, consists of a genetic, g_i, environmental, e_i, and a mean, μ, contribution,

$$\phi_i = \mu + g_i + e_i.$$

Under a linear statistical model, we can interpret the environmental contribution as a random error. We assume that the environmental contributions, $\boldsymbol{e}$, are drawn from the same distribution for all individuals,

$$\boldsymbol{e} \sim \mathcal{N}(0, \sigma_e^2 \mathbb{I}_n),$$

where $\mathbb{I}_n$ is the identity matrix with dimension n. We further assume that the genetic contributions, $\boldsymbol{g}$, are random samples from a multidimensional normal distribution with a covariance matrix equal to the genetic relatedness matrix $\mathbf{G}$ (Fig. I),

$$\boldsymbol{g} \sim \mathcal{N}(0, \sigma_g^2 \mathbf{G}).$$

The phenotype distribution then is,

$$\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma_g^2 \mathbf{G} + \sigma_e^2 \mathbb{I}_n).$$

The variance of this normal distribution can also be written in terms of the total phenotypic variance, $\sigma_\phi^2 = \sigma_g^2 + \sigma_e^2$, and the heritability, $h^2 = \sigma_g^2/\sigma_\phi^2$, resulting in a final covariance matrix

$$\mathbf{\Sigma}_{\text{PMM}} = \sigma_g^2 \mathbf{G} + \sigma_e^2 \mathbb{I}_n = \sigma_\phi^2 \left(h^2 \mathbf{G} + \left(1 - h^2\right) \mathbb{I}_n \right). \tag{1}$$

The entries g_{ij} of $\mathbf{G}$ are proportional to the joint evolutionary time of the individuals i and j relative to some evolutionary origin T (Fig. I). If a_{ij} is the time of the most recent common ancestor of individuals i and j, then $g_{ij} = 1 - a_{ij}/T$. The origin T is the time of the most recent common ancestor of all individuals, such that the smallest entry in $\mathbf{G}$ is 0 and the diagonal elements are 1.

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pagel's Lambda The idea behind the PL method is to find an optimal rescaling parameter, λ, of the internal branches of the tree, such that the new ancestor times are $b_{ij} = \lambda a_{ij}$, but the diagonal elements are unchanged, $b_{ii} = a_{ii}$. Under this rescaling, the covariance matrix is

$$\mathbf{\Sigma}_{\mathrm{PL}} = \sigma_\phi^2 \left(\lambda \mathbf{G}' + \mathbf{G_D}\right), \tag{2}$$

where $\mathbf{G}'$ is equal to $\mathbf{G}$ with the diagonal removed, $\mathbf{G}' = \mathbf{G} - \mathbf{G_D}$.

Comparing PL and PMM A comparison between the PMM and PL becomes more clear by rewriting Eq. 1 in terms of $\mathbf{G}'$ and $\mathbf{G}_D$,

$$\mathbf{\Sigma}_{\mathrm{PMM}} = \sigma_\phi^2 \left(h^2 \boldsymbol{G'} + h^2 \boldsymbol{G_D} + (1 - h^2)\mathbb{I}_n\right),$$

For ultrametric trees, all individuals are sampled at the same time $t_i = 0$, and thus all the $g_{ii} = 1$, and $\boldsymbol{G_D} = \mathbb{I}_n$. In this case,

$$\mathbf{\Sigma}_{\mathrm{PMM}} = \sigma_\phi^2 \left(h^2 \boldsymbol{G'} + \mathbb{I}_n\right),$$
$$\mathbf{\Sigma}_{\mathrm{PL}} = \sigma_\phi^2 \left(\lambda \boldsymbol{G'} + \mathbb{I}_n\right),$$

for which the equivalence between PL and PMM becomes evident by setting $h^2 = \lambda$.

For non-ultrametric trees, not all g_{ii} are the same and thus the two methods can no longer be made identical by simply setting $\lambda = h^2$. The rescaling in PL only applies to the internal branches, which can be interpreted such that individuals that are sampled later and thus have larger entries in $\mathbf{G}_D$ also have larger environmental contributions. This, however, would only make sense if the environmental contribution increases linearly with time of sampling. A more sensible assumption is that the environmental contribution is independent of the time of sampling, and hence the PMM is the proper formulation of a neutrally evolving trait along a tree.

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

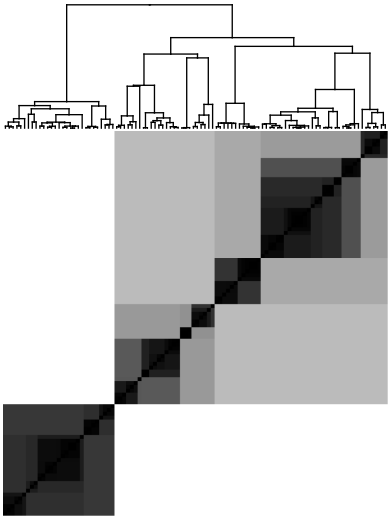

Figure I: **Graphical representation of the genetic relationship matrix for a Wright-Fisher population.** The strength of the covariance between two individuals depends on the amount of shared ancestry in the tree. White entries corresponds to a covariance of 0 and black to a covariance of 1.

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

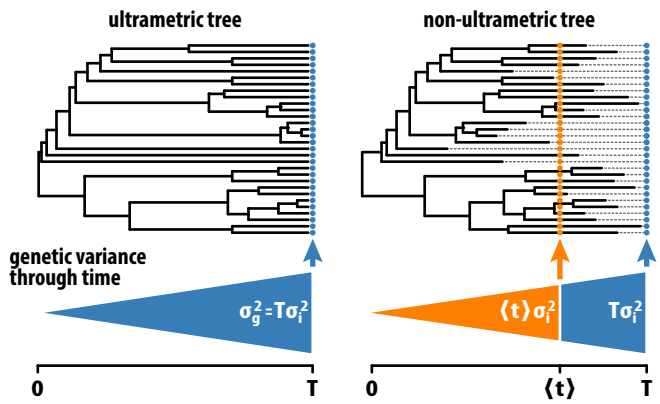

Figure II: **The PMM estimates a per-unit increase in genetic variance.** For ultrametric trees, the estimated genetic varaince in the population is equal to the total height of the tree, T, times the estimated per-unit increase in genetic variance, σ_i^2. For non-ultrametric trees, the choice of the process time determines the amount of estimated genetic variance and hence the heritability. Common choices are the mean sampling time of all the tips, $\langle t \rangle$, or the total height of the tree, T.

Box 2: Care must be taken when applying the PMM to non-ultrametric trees

An issue with applying the PMM to non-ultrametric trees is that the ancestry matrix, $\boldsymbol{A}$, needs to be rescaled by an arbitrary parameter. The entries, a_{ij}, of the ancestry matrix represent the total time of shared ancestry of two individuals between the present and the root of the tree at some time T in the past. For ultrametric trees, all the tips of the tree align to the present, and hence the diagonal elements of $\boldsymbol{A}$ are all T. Thus, an obvious choice of rescaling is $\boldsymbol{G} = \boldsymbol{A}/T$. The genetic variance component, σ_g^2, estimated using the PMM is then interpreted as the expected amount of genetic variance that accumulated during the time T. The per unit time increase of genetic variance, $\sigma_i^2 = \sigma_g^2/T$, is the parameter of the underlying model of linear increase in variance in the PMM (Fig. II).

For non-ultrametric trees, the tips in the tree no longer all align to the present. Thus, the choice of normalization for $\boldsymbol{A}$ is less evident. We can write Eq. 1 from Box 1 using the ancestry matrix and σ_i^2 directly,

$$\boldsymbol{\Sigma}_{\text{PMM}} = \sigma_i^2 \boldsymbol{A} + \sigma_e^2 \mathbb{I}_n.$$

The mathematical framework to estimate variance components using the PMM thus remains valid for covariance matrices from non-ultrametric trees and can be used to estimate the variance component σ_e^2 and the per unit time variance component σ_i^2. However, in order to estimate the heritability, σ_i^2 must be multiplied by an arbitrary time τ,

$$h^2 = \frac{\sigma_i^2 \tau}{\sigma_i^2 \tau + \sigma_e^2}. \tag{3}$$

Equation 3 gives the heritability estimate for a population at time τ since the beginning of the process at the root of the tree. For ultrametric trees, choosing $\tau = T$ is the same as

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

rescaling the ancestry matrix, $\boldsymbol{A}$, to the genetic relatednes matrix, $\boldsymbol{G}$. For non-ultrametric trees, common choices of τ are the present time, T, or the mean sampling time of all the tips (Fig. II). Importantly, the specific choice of τ will determine the quantitative esimate of heritability and results in slghlty different interpretations. By choosing the present time, $\tau = T$, the estimated heritability applies to the overall (potentially unknown) patient population that exists at the present time. Choosing τ as the mean sampling time across all individuals returns an estimate of heritability that applies to the pooled sampled population, represented by a hypothetical individual that was sampled at time τ.

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Box 3: Obtaining maximum likelihood estimates for heritability

Here, we recapitulate [33] on how to obtain heritability estimates by finding the values of $(\mu, \sigma_\phi^2, h^2)$ that maximize the likelihood for the SPVL data under a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma} = \sigma_\phi^2 \boldsymbol{V}$. Using the expressions for the covariance matrix $\boldsymbol{\Sigma}(h^2)$ from Box 1, the log-likelihood of the SPVL data, $\boldsymbol{\phi}$, is simply

$$\ln \mathcal{L} = -\frac{1}{\sigma_\phi^2}(\boldsymbol{\phi} - \boldsymbol{\mu})^T \boldsymbol{V}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu}) - \frac{n}{2}\ln\left(2\pi\sigma_\phi^2\right) - \frac{1}{2}\left|\boldsymbol{V}\right|,$$

where $\boldsymbol{\mu} = (\mu, \ldots, \mu)^T$. The maximum likelihood (ML) estimates as a function of h^2 for the mean and variance are easily found by setting the first-order derivatives of the log-likelihood equal to zero,

$$\hat{\mu}(h^2) = \frac{\mathbf{1}^T \boldsymbol{V}^{-1}\boldsymbol{\phi}}{\mathbf{1}^T \boldsymbol{V}\mathbf{1}},$$
$$\hat{\sigma}_\phi^2(h^2) = \frac{(\boldsymbol{\phi} - \boldsymbol{\mu})^T \boldsymbol{V}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu})}{n}.$$

The ML estimate of h^2 can then be found numerically by optimizing the log-likelihood over values of h^2. Similarly, confidence intervals for h^2 are derived using the profile likelihood method.

These ML estimates require computing the matrix inverse, $\boldsymbol{V}^{-1} = \left(h^2\boldsymbol{G} + (1-h^2)\mathbb{I}_n\right)^{-1}$, which can be computationally intensive. Historically, diagonalizing or inverting even moderately sized matrices was computationally prohibitive. Henderson [34] presented a trick that avoids matrix inversion by including unobserved ancestors as latent variables. This approach to estimating heritability is used for restricted maximum likelihood estimation in the proprietary software ASReml [35], and for Bayesian MCMC heritability estimation in the open source R package MCMCglmm [36].

However, diagonalizing even large matrices is much less prohibitive these days, allowing for a much more straightforward calculation of h^2 estimates using the above expressions. Importantly, Housworth *et al.* [33] showed that estimates of h^2 could be rapidly obtained by only diagonalizing the genetic relatedness matrix once, $\boldsymbol{G} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^T$, where $\boldsymbol{Q}$ is the matrix of eigenvectors of $\boldsymbol{G}$, and $\boldsymbol{D}$ is the diagonal matrix of eigenvalues. The normalized covariance matrix is then,

$$\boldsymbol{V} = \boldsymbol{Q}\left(h^2\boldsymbol{D} + \left(1-h^2\right)\mathbb{I}_n\right)\boldsymbol{Q}^T = \boldsymbol{Q}\boldsymbol{D}'\boldsymbol{Q}^T,$$

where $\boldsymbol{D}'$ is a diagonal matrix with elements $d'_{ii} = h^2\xi_i + (1-h^2)$, and ξ_i is the i-th eigenvalue of $\boldsymbol{G}$. This allows for the rapid computation of the matrix inverse, $\boldsymbol{V}^{-1} = \boldsymbol{Q}\boldsymbol{D}'^{-1}\boldsymbol{Q}^T$, where $\boldsymbol{D}'^{-1} = \mathrm{diag}(d_{ii}^{-1})$. Finally, the last missing term to compute the likelihood is the determinant, $|\boldsymbol{V}| = \prod_{i=1}^n d_{ii}$.

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Mellors, J. W. *et al.* Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* **272,** 1167–70 (1996).
2. De Wolf, F *et al.* AIDS prognosis based on HIV-1 RNA, CD4+ T-cell count and function: markers with reciprocal predictive value over time after seroconversion. *AIDS* **11,** 1799–806 (1997).
3. Bonhoeffer, S, Funk, G., Gunthard, H., Fischer, M & Muller, V. Glancing behind virus load variation in HIV-1 infection. *Trends Microbiol* **11,** 499–504 (2003).
4. Geskus, R. B. *et al.* The HIV RNA setpoint theory revisited. *Retrovirology* **4,** 65 (2007).
5. Fraser, C., Hollingsworth, T. D., Chapman, R., de Wolf, F. & Hanage, W. P. Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis. *Proc Natl Acad Sci U S A* **104,** 17441–17446 (2007).
6. Falconer, D. S. *Introduction to quantitative genetics.* (Longman., 1981).
7. Alizon, S. *et al.* Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathog* **6,** e1001123 (2010).
8. Fraser, C. *et al.* Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. *Science* **343,** 1243727 (2014).
9. Hodcroft, E. *et al.* The contribution of viral genotype to plasma viral set-point in HIV infection. *PLoS Pathog* **10,** e1004112 (2014).
10. Van der Kuyl, A. C., Jurriaans, S., Pollakis, G., Bakker, M. & Cornelissen, M. HIV RNA levels in transmission sources only weakly predict plasma viral load in recipients. *AIDS* **24,** 1607–1608 (2010).
11. Lingappa, J. R. *et al.* Estimating the impact of plasma HIV-1 RNA reductions on heterosexual HIV-1 transmission risk. *PLoS One* **5,** e12598 (2010).
12. Hecht, F. M. *et al.* HIV RNA level in early infection is predicted by viral load in the transmission source. *AIDS,* 1 (2010).
13. Hollingsworth, T. D. *et al.* HIV-1 transmitting couples have similar viral load setpoints in Rakai, Uganda. *PLoS Pathog* **6,** e1000876 (2010).
14. Yue, L. *et al.* Cumulative impact of host and viral factors on HIV-1 viral-load control during early infection. *J Virol* **87,** 708–15 (2013).
15. Bulmer, M. Galton's law of ancestral heredity. *Heredity* **81 (Pt 5),** 579–85 (1998).
16. Bell, A. E. Heritability in retrospect. *J Hered* **68,** 297–300 (1977).
17. Fisher, R. A. The correlation between relatives on the supposition of mendelian inheritance. *Trans R Soc Edinburgh* **52,** 399–433 (02 1918).
18. Wright, S. The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proc Natl Acad Sci U S A* **6,** 320–32 (1920).

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

19. Lush, J. L. *Genetic aspects of the Danish system of progeny-testing swine* in *Iowa Research Bulletin 204* (Agricultural Experiment Station, Iowa State College of Agriculture and Mechanic Arts, 1936).

20. Bonhoeffer, S., Fraser, C. & Leventhal, G. E. High heritability is compatible with the broad distribution of set point viral load in HIV carriers. *PLoS Pathog* **11,** e1004634 (2015).

21. Vrancken, B. *et al.* Simultaneously estimating evolutionary history and repeated traits phylogenetic signal: applications to viral and host phenotypic evolution. *Methods Ecol Evol* **6,** 67–82 (2015).

22. Blomberg, S. P., Garland, T. Jr & Ives, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57,** 717–45 (2003).

23. Felsenstein, J. Phylogenies and the Comparative Method. *Am Nat* **125,** 1–15 (1985).

24. Lynch, M. Methods for the Analysis of Comparative Data in Evolutionary Biology. *Evolution* **45,** 1065–1080 (1991).

25. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401,** 877–84 (1999).

26. Shirreff, G. *et al.* How effectively can HIV phylogenies be used to measure heritability? *Evol Med Public Health* **2013,** 209–24 (2013).

27. Bartha, I. *et al.* A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* **2,** e01123 (2013).

28. Bartha, I. *et al.* Estimating the respective contributions of human and viral genetic variation to HIV control. *bioRxiv.* doi:10.1101/029017 (2015).

29. Shirreff, G., Pellis, L., Laeyendecker, O. & Fraser, C. Transmission selects for HIV-1 strains of intermediate virulence: a modelling approach. *PLoS Comput Biol* **7,** e1002185 (2011).

30. Herbeck, J. T., Mittler, J. E., Gottlieb, G. S. & Mullins, J. I. An HIV epidemic model based on viral load dynamics: value in assessing empirical trends in HIV virulence and community viral load. *PLoS Comput Biol* **10,** e1003673 (2014).

31. Lande, R. Natural selection and random genetic drift in phenotypic evolution. *Evolution,* 314–334 (1976).

32. Hansen, T. F. Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution* **51,** 1341–1351 (1997).

33. Housworth, E. A., Martins, E. P. & Lynch, M. The phylogenetic mixed model. *Am Nat* **163,** 84–96 (2004).

34. Henderson, C. R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32,** 69–83 (1976).

35. Gilmour, A. R., Gogel, B., Cullis, B. & Thompson, R. ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK* (2009).

bioRxiv preprint doi: https://doi.org/10.1101/046797; this version posted April 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

36. Hadfield, J. D. MCMC Methods for multi-response generalized linear mixed models: The MCMCglmm R package. *J Stat Softw* **33,** 1–22 (2010).