**Classification:** Biological Sciences / Genetics

# Title: The Heritability of Pathogen Traits - Definitions and Estimators

**Authors:** Venelin Mitov [a,b]*, Tanja Stadler [a,b]

**Affiliations:**

[a]Department of Biosystems, Science and Engineering (D-BSSE), Swiss Federal Institute of Technology (ETH), Zürich, Switzerland.

[b]Swiss Institute of Bioinformatics, Lausanne, Switzerland.

**Corresponding Author:** Venelin Mitov: ETH Zürich, Department of Biosystems Science & Engineering (D-BSSE), Mattenstrasse 26, 4058 Basel, Switzerland, office: 7.48, tel.: +41 61 38 73411, e-mail: venelin.mitov@bsse.ethz.ch.

**Keywords:** HIV virulence, donor-recipient regression, Phylogenetic Mixed Model, Ornstein-Uhlenbeck, ANOVA

## Abstract

Pathogen traits, such as the virulence and the transmissibility of an infection, can vary significantly between patients. A major challenge is to measure the extent to which the genetic differences between infecting strains explain the observed variation of the trait. This is quantified by the so-called broad-sense heritability $H^2$ – a term borrowed from quantitative genetics of sexual species. A recent discrepancy between estimates of HIV-virulence-heritability has opened a debate on the accuracy of the estimators. Here, we show that the discrepancy originates from model limitations and important lifecycle differences between sexually reproducing organisms and transmittable pathogens. In particular, current quantitative genetics methods are prone to underestimate $H^2$, because they do not account for rapid within-host mutation combined with natural selection on the trait. We introduce two independent approaches correcting these errors: ANOVA-CPP and our POUMM method. Empirical analyses reveal that at least 20% of the variation in virulence is explained by the virus genome both for European and African data. These results should terminate the ongoing discussion whether the virus affects virulence at all, and should motivate further genome-wide association studies on the virus, as well as studies on the interaction between host- and viral factors for virulence. Beyond HIV, we discuss that ANOVA-CPP is ideal for slowly evolving protozoa, bacteria and DNA-viruses, while POUMM is ideal for rapidly evolving RNA-viruses, thus, enabling heritability estimation for a broad range of pathogens.

## Significance statement

Pathogen traits, such as the virulence of an infection, can vary tremendously between patients. To what extent the pathogen rather than the host determines these traits remains a mystery for many infectious diseases. This is quantified by the traits' "heritability" – a term borrowed from quantitative genetics of sexual species. Recently, a discrepancy between studies of HIV-virulence has opened a debate on the appropriate estimators of pathogen trait heritability. We find the origin of this discrepancy in the inability of current quantitative genetics methods to account for rapid pathogen mutation in combination with natural selection on the trait. We introduce two independent approaches correcting these errors and report agreeing heritability estimates of these approaches on synthetic and empirical HIV-data.

## Introduction

Pathogens transmitted between donor and recipient hosts are genetically related much like children are related to their parents through inherited genes. This analogy between transmission and biological reproduction has inspired the use of heritability ($H^2$) – a term borrowed from quantitative genetics (1-3) – to measure the contribution of pathogen genetic factors to pathogen traits, such as virulence, transmissibility and drug-resistance of infections.

Two families of methods enable estimating the heritability of a pathogen trait in the absence of knowledge about its genetic basis:

a. Resemblance estimators measuring the relative trait-similarity within groups of transmission-related patients. Common methods of that kind are linear regression of donor-recipient pairs (DR) (4, 5) and analysis of variance (ANOVA) of patients linked by (near-)identity of carried strains (6, 7).

b. Phylogenetic comparative methods measuring the association between observed trait values from patients and their (approximate) transmission tree inferred from carried pathogen sequences. Common examples of such methods are the phylogenetic mixed model (PMM) (8) and Pagel's $\lambda$ (9).

Most of these methods have been applied in studies of the viral contribution to virulence of an HIV-1 infection (4, 5, 7, 10-16), quantified by $\log_{10}$ set point viral load – $\lg(\mathrm{spVL})$ – the amount of virions per blood-volume stabilizing in HIV-1 patients at the beginning of the asymptomatic phase and best-predicting its duration (17). In the view of discrepant reports of $\lg(\mathrm{spVL})$-heritability, several authors have questioned the methods' accuracy (4, 5, 7). Shirreff et al. 2012 used simulation of trait-values on existing HIV-1 transmission trees to reveal that phylogenetic comparative methods report strongly under- or over- estimated values depending on the true heritability value used in the simulation (7). Later, Fraser et al. 2014 claimed that DR is unbiased w.r.t. $\lg(\mathrm{spVL})$-heritability and is robust to trait-based selection for transmission (4). Finally, Leventhal & Bonhoeffer (5) simulated Wright-Fisher generations of transmission confirming that DR outperforms PMM in terms of robustness and accuracy and suggesting that current phylogenetic methods are compromised by questionable assumptions - such as ultrametricity of trees (all measurements collected at the same time) and neutral evolution of the trait. These three studies assume that once the trait value is set in the recipient upon infection, it remains constant throughout its infectious time. This assumption is partially acceptable for lg(spVL), see (18) and references therein, but it is arguable in general for traits of chronic infections due to continuous within-host adaption. Moreover, the theory of heritability, which was developed by quantitative geneticists to study populations of animals and plants (1-3), does not account for individual gradual evolution and other lifecycle differences between pathogens and mating species. This reveals the need for a careful transfer of the quantitative genetics terminology and methods to the domain of pathogen traits.

Herein, we review the definitions of heritability in sexually reproducing species, transfer these definitions to pathogens and point out the principle sources of bias in commonly used heritability estimators, in particular, PMM, DR and ANOVA. Using simulations, we validate two independent approaches to counter these sources of bias and show that these approaches yield agreeing $\lg(\mathrm{spVL})$-heritability estimates across HIV cohorts.

## Theory

**Heritability in sexual species.** Jacquard, 1983 (19) noticed that the term "heritability" has been used by quantitative geneticists to serve three different concepts: (i) the genetic determination of a trait; (ii) the resemblance between relatives; (iii) the efficiency of selection. Due to this terminological discrepancy and the fact that most estimators are sample statistics matching or approximating the target definition of heritability only in the validity of a number of assumptions, it is confusing to use the term "heritability" without an accompanying definition or a qualifier like "narrow-sense", "broad-sense" and "realized". We briefly introduce this terminology, referring to equations in SI for the formal definitions.

Considering a real-valued (quantitative) trait, the degree to which the genes of individuals determine their trait-values is quantified in a statistical sense by the **broad-sense heritability**, $H^2$ (eq. S2). $H^2$ summarizes how much of the observed trait variance in a population can be explained by genetic differences between individuals in the population. In the world of animals and plants, it is usually hard to measure $H^2$ directly. Thus, quantitative genetics focuses on estimating its lower bound – the **narrow-sense heritability**, $h^2$, summarizing how much of the trait variance is attributable to single-locus additive genetic effects (eq. S4). Sexual reproduction ensures that additive effects constitute most of the heritable genetic effects. Thus, $h^2$ is estimable from measures of the trait-resemblance between genetically related individuals in the population.

Relatives resemble each other not only for carrying similar genes but also for living in similar environments. Hence, it is necessary to disentangle the concept of resemblance from that of genetic determination. Considering an ordered relationship such as parent-offspring, their resemblance is usually measured by the **regression slope**, $b$, of least squares regression of expected offspring values on mean parental values (eq. S5). Considering members of unordered relationships, such as identical twins, sibs and cousins, their relative resemblance is quantified by the one-way analysis of variance (ANOVA), which estimates the so-called **intraclass correlation** (ICC) denoted here as $r_A[\text{type of relationship}]$ (eq. S6).

The last of the three concepts is that of the efficiency of selection for breeding of the individuals with "best" trait-values. This is quantified by the **realized heritability**, $h_R^2$, defined in (3) as the response to selection relative to the selection differential (eq. S7).

**Connecting the dots.** The success of quantitative genetics in the pre-genomic era relies on the insight that "*inferences concerning the genetic basis of quantitative traits can be extracted from phenotypic measures of the resemblance between relatives* (1)". Mathematically, this quote is expressed as a set of approximations, which have become dogmatic in quantitative genetics:

$$H^2 \simeq r_A[\text{identical twins}] \qquad (1)$$

$$h^2 \simeq b \simeq 4r_A[\text{half sibs}] \simeq h_R^2, \qquad (2)$$

Equation 1 above is valid in general, provided there is no strong maternal effect on the trait, the observed twins have been separated at birth and raised in independent environments and the assumptions of ANOVA such as normality and homoscedasticity are at least approximately met. Equation 2, though, relies on genetic segregation and

recombination during sexual reproduction and is, therefore, provable only for diploid sexually reproducing species (1, 3). The reason is that during sexual reproduction a child inherits random halves of its parents' alleles but, due to meiotic segregation and chromosomal crossover, couples and higher order combinations of alleles are inherited at decreasing proportions (1, 2). Thus, it can be assumed that the phenotypic resemblance between relatives originates predominantly from inherited single-locus additive effects, rather than multi-locus (epistatic) interactions.

In summary, in sexually reproducing populations, heritability is used to quantify to what extent the genetics explain a trait (broad-sense heritability, $H^2$) as well as to measure or predict the response to trait-based selection for reproduction (realized heritability, $h_R^2$). Since it is practically hard to measure $H^2$, one often uses empirical measures of the resemblance between relatives (i.e. parent-offspring regression, $b$, or ICC from half sibs, $r_A$) to estimate the extent, to which single-locus additive effects determine the trait ($h^2$). It turns out that $h^2 \simeq h_R^2$, justifying the dual role of $h^2$ as a measure of genetic determination and a measure for the rate of trait-evolution resulting from selection.

**Transfer to pathogen traits.** The transfer of the above terminology from traits of diploid organisms to pathogen traits is almost verbatim and only requires substituting "pathogen genes" for "organism genes", "donor value" for "mean parental value" and "recipient value" for "offspring value". However, three important differences between the lifecycles of diploid organisms and pathogens alter the interpretation and the connections between the definitions:

- **Asexual haploid nature of pathogen transmission**

The first difference is that, unlike reproduction of diploid organisms, the transmission of a pathogen from a donor to a recipient is more similar to asexual reproduction in haploid organisms, because, typically, whole pathogens get transferred between hosts. Importantly, in the absence of genetic segregation and recombination at transmission, there is no preference in transmitting single-locus over multi-locus genetic effects.

- **Partial quasispecies transmission**

The second difference is that, due to transmission bottlenecks, typically, only a tiny sample of the large and genetically diverse pathogen population in the donor (aka quasispecies) penetrates and survives in the recipient (20). The transmitted proportion of genetic information characterizing the quasispecies is unknown and varying between transmission events.

- **Within-host pathogen evolution**

The third difference involves the change in phenotypic value due to within-host pathogen mutation and recombination. While genetic change is rare during the lifetime of animals and plants and its phenotypic effects are typically delayed to the offspring generations, it constitutes a hallmark in the lifecycle of pathogens and causes an immediate phenotypic change such as increased virulence, immune escape or drug resistance.

Due to the asexual nature of transmission, the donor-recipient regression (DR) estimates the broad-, rather than the narrow-sense heritability. Fraser et al. 2014 (4) have used Wright's method of path coefficients (21) to prove that $b$ estimates $H^2$ under

several assumptions including linear dependence of recipient- on donor trait-values and matching pathogen strains in donor- and recipient hosts. Leventhal & Bonhoeffer performed simulations where DR is unbiased in the case of minute evolution in the recipient host upon infection (5). In their simulation, partial quasispecies transmission and gradual within-donor/-recipient evolution throughout the infection is ignored. However, these two phenomena cause a negative bias in $b$ as estimator of $H^2$, because they co-act for decreased trait-resemblance. Thus, it is necessary to regard $b$ as a statistic summarizing the resemblance in transmission couples observable after partial quasispecies transmission and delay between transmission and measurements. Further in the text, we use the symbol $b[\tau]$ to emphasize that it has been calculated on a sample of donors and recipients with (variable) periods $\tau_d$ and $\tau_r$ between transmission and measurements, $\tau = \tau_d + \tau_r$ denoting the total amount of time (Fig. 1). By contrast, we use $b[0]$ to emphasize that the calculation has been done on the immediate trait-values right after transmission.

In summary, for pathogen traits, measures of resemblance, such as $b[0]$, estimate $H^2$ rather than $h^2$. However, due to delayed diagnosis, data from transmission couples for estimating $b[0]$ is rarely available in practice, while $b[\tau]$ is a negatively biased estimator of $H^2$. In the absence of genetic segregation and recombination at transmission, $h^2$ loses its double role as an accessible measure of genetic determination and as a predictor for the rate of evolution.

By the above logic, all resemblance-based estimators of $H^2$ are negatively biased by partial quasispecies transmission and measurement delays. In principle, both of these sources of bias can be addressed in a genome-wide association study (GWAS) on pathogen sequence- and phenotypic data. However, this approach is usually technically involved, because it needs the collection of deep sequence data from multiple host-tissues (see e.g. (22)) from numerous patients and its statistical power is often compromised by multiple testing (23). We now discuss two simpler methods that allow overcoming measurement delays, while using increasingly available pathogen consensus sequences.

**Filtering the data.** The first approach is to minimize the bias from measurement delays by limiting the analysis to the observations with minimal delays. For example, Hecht et al 2010 (11) used Pearson correlation (a measure similar to $b$) in donor-recipient couples with recent transmission and other authors (6, 7) used ANOVA ($r_A$) on patients grouped by phylogenetic proximity. In particular, Shirreff et al. 2012 (7) defined the method of phylogenetic pairs (PP) as ANOVA on pairs of tips in the transmission tree that are mutually nearest to each other by phylogenetic distance, $\tau$. We will consider this method in two forms: (i) the original one as proposed in (7) and denoted here as ANOVA-PP with estimate $r_A[\mathrm{PP}:\tau]$ and (ii) ANOVA on the closest phylogenetic pairs (CPP) defined as PPs that are not farther apart than a cut-off phylogenetic distance $\tau'$ and denoted here as **ANOVA-CPP** with estimate $r_A[\mathrm{PP}:\tau \le \tau']$. The main drawback of these filtering techniques is their reduced statistical power due to fewer observations.

**Phylogenetic heritability.** A different approach allowing overcoming measurement delays is to use the transmission tree connecting all available observations. Like a species tree tracing the common ancestry of a set of species, the transmission tree connecting a set of infected hosts is a tree structure representing transmission events as

branching points and evolutionary time between events as branch-lengths (Fig. 1). For rapidly evolving pathogens, such as RNA viruses, it is possible to infer the approximate transmission tree from pathogen sequences sampled at the moment of trait measurement (24). This has inspired the use of phylogenetic comparative methods, such as the phylogenetic mixed model (PMM) (8, 25), to estimate virulence heritability in HIV (7, 14, 15). The PMM method decomposes the trait value into a non-heritable component, $e \sim \mathcal{N}(0, \sigma_e^2)$, and a genetic component, $G$, which (i) evolves continuously according to a Brownian Motion (BM) process along branches; (ii) gets inherited by the two daughter branches descending from each internal node. Thus, PMM allows inferring the so-called **phylogenetic heritability** as the expected genetic variance relative to the total variance at the tree tips (eq. 4). In biological terms, $G$ is a genetic contribution to the trait value that evolves according to random drift, $e$ is the non-genetic contribution such as the host immune response. By staying free of assumptions about the behavior of $e$ during infections (i.e. along tree branches), PMM is suitable to different biological scenarios, such as a stable or a variable effect of the immune system throughout infections. However, as noticed also by other authors, an important issue with the PMM method is that the BM-assumption of linearly increasing variance through time rarely holds for pathogen traits, which are typically constrained by some natural limits or by balancing selection (5, 16). Later, we show that such BM violations lead to negative bias in the phylogenetic heritability w.r.t. $H^2$. To remedy this, inspired by the work of other authors (26-28), we develop the **POUMM** method – an extension of PMM, in which we replace the BM-assumption with an assumption of an Ornstein-Uhlenbeck (OU) process acting on $G$ (Methods). As a final note, we mention that the standard definition of phylogenetic heritability depends on the tree-length, which is not defined in the case of non-ultrametric transmission trees. Thus, we prefer a time-independent formulation, denoted as $H^2_{BMe}$ and $H^2_{OUe}$ for PMM and POUMM respectively (eq. 5).

## Results

**A toy-model of an epidemic.** To test the different estimators of heritability introduced above, we developed a toy-model of an epidemic, in which an imaginary pathogen trait, $z$, was determined by the interaction between the alleles at two loci in the pathogen genotype and one of two immune system types encountered at equal frequencies in the susceptible population (Fig. 2A). This toy-model was embedded into a stochastic Susceptible-Infected-Recovered (SIR) epidemic model with demography and frequency dependent transmission (29), implementing "neutral" and "selection" modes of within- and between-host dynamics (Fig. 2B, Methods).

Using different contact-rates between individuals, $\kappa \in \left\{\frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{12}\right\}$, we performed 240 simulations, of which 175 resulted in epidemic outbreaks of at least 1,000 diagnosed individuals. A detailed analysis of the different heritability estimates on samples between 1,000 and 10,000 diagnosed individuals (see Fig. 3C,E, Fig. S1, Fig. S2 and Supporting Text) confirmed the negative bias due to measurement delays in the resemblance-based estimators (mostly pronounced in $b[\tau]$ and $r_A[\mathrm{PP}:\tau]$, but also noticeable in $b[\tau \le D_1]$ and $r_A[\mathrm{PP}:\tau \le D_1]$, $D_1$ denoting the first decile ($10^{\text{th}}$ quantile) of $\tau$). In addition, the simulations showed that a worsening fit of the BM model on longer transmission trees without growth of phenotypic variance (Fig. 3D) caused an inflated estimate of the

environmental variance, $\sigma_e^2$, in the PMM method and, therefore, a negative bias in $H^2_{BMe}$ (Fig. 3E). Several other sources of bias, such as non-linear dependence of recipient on donor-values and deviation from normality were identified and are summarized in Table S3.

We conclude that, apart from the practically inaccessible immediate donor-recipient regression ($b[0]$) and ICC of patients grouped by identity of carried strain ($r_A[\mathrm{id}]$), the most accurate estimator of $H^2$ in the toy-model simulations is $H^2_{OUe}$ followed by estimators minimizing measurement delays such as $b[\tau \le D_1]$ and $r_A[\mathrm{PP}:\tau \le D_1]$.

**Analysis of HIV-data.** We performed ANOVA-CPP and POUMM on data from the UK HIV-1 cohort comprising $\lg(\mathrm{spVL})$ measurements and a tree of viral (pol) sequences from 8,483 patients inferred previously in (15). The goal was to test our conclusions on a real dataset and compare the $H^2$-estimates from ANOVA-CPP and POUMM to previous PMM/ReML-estimates on exactly the same data (15). A scatter plot of the phylogenetic distances of tip-pairs against the absolute phenotypic differences, $|\Delta\lg(\mathrm{spVL})|$, reveals a small set of 116 PPs having $\tau \le 10^{-4}$ while the phylogenetic distance in all remaining tip-pairs is more than an order of magnitude longer, i.e. $\tau > 10^{-3}$ (Fig. 4A). The random distribution of these PPs along the transmission tree suggests that these phylogenetic pairs correspond to randomly occurring early detections of infection (Fig. 4B). Based on the observed gap of $\tau$, we defined these PPs as closest ones (CPP). By applying the $1.5\times IQR$-rule on $|\Delta\lg(\mathrm{spVL})|$ in CPPs (Methods), we identified five outlier CPPs shown as blue bullets on Fig. 4.

We compared the following estimators of $H^2$, with and without inclusion of these outlier CPPs in the data:

- ANOVA on CPPs/PPs;
- POUMM/PMM on the whole tree (including tips belonging to CPPs);
- POUMM/PMM on the tree obtained after dropping tips belonging to CPPs;

Excluding outlier CPPs, ANOVA-CPP (222 patients) reported lg(spVL)-heritability estimates of 0.31, 95% CI [0.19, 0.43]. POUMM (8,473 patients) reported agreeing estimates of 0.25, 95% CI [0.16, 0.36] and 0.22, CI [0.13, 0.35] upon omitting all 222 patients belonging to CPPs. The slightly lower POUMM estimates could be explained by errors in the transmission tree, which are not present in CPPs. These results show first, that ANOVA-CPP and POUMM agree on disjoint subsets of the UK data and, second, that POUMM provides an alternative to resemblance-based methods in the absence of early-diagnosed cases.

Figure 5 compares these estimates to previous $\lg(\mathrm{spVL})$ studies (estimates including outliers written in Table S4). In agreement with the toy-model simulations, estimates of $H^2$ using PMM or other phylogenetic methods (i.e. $K$ and $\lambda$) are notably lower than all other estimates, suggesting that current phylogenetic comparative methods underestimate $H^2$ due to violation of the BM-assumption (see also Fig. 3D and Fig. 4B); resemblance-based estimates are down-biased by measurement delays (compare early vs late on Fig. 5).

In summary, POUMM and ANOVA-CPP yield agreeing estimates for $H^2$ in the UK data and these estimates agree with DR-based estimates in datasets with short measurement delay. Similar to the toy-model simulations, we notice a well-pronounced pattern of negative bias for the other estimators, PMM and ANOVA-PP, as well as for the previous DR-studies on data obtained under long measurement delay.

## Discussion

**Clarifying the terminology and notation.** The first task of this study was the transfer of quantitative genetics terminology to the domain of pathogen traits. Due to important lifecycle differences between pathogens and mating organisms, it is essential to disentangle the concepts of relative resemblance and genetic determination. In essence, the estimators of trait resemblance between transmission-related patients, such as DR and ICC, and the phylogenetic heritability, must be regarded as estimators of the broad-sense heritability, $H^2$, compromised by partial quasispecies transmission, within-host evolution and various violations of model assumptions (Table S3). A few examples from recent studies of HIV-1 demonstrate the need for a careful consideration of these concepts. For example, in (15) and (5) the authors introduce the PMM/ReML and the DR methods for estimating heritability after a definition of the heritability in the narrow sense, $h^2$. This can leave a confusing impression that the reported values are estimates of $h^2$ rather than $H^2$, because these methods are popular for estimating narrow-sense heritability for sexual species. As another example, in (4, 7), the authors use the lower-case notation "$h^2$" to denote estimates of $H^2$. In fact, there are historical reasons to associate the symbol "$h^2$" with the regression slope, $b$ (4, 21). However, "$h^2$" is the standard symbol for narrow-sense heritability and $b$ is, most of all, a measure of phenotypic resemblance. To avoid confusion, we recommend using the standard symbol "$H^2$" for broad-sense heritability(1, 3) and different symbols for its indirect estimators.

**A disagreement between simulation studies.** Using simulations of a classical epidemiological model, we have shown that two methods based on phenotypic and sequence data from patients - ANOVA-CPP and POUMM - provide more accurate heritability estimates compared to previous approaches like DR and PMM. However, we should not neglect the arising discrepancy between our and previous simulation reports advocating either PMM (15) or DR (5) as unbiased heritability estimators. Compared to these simulations, the toy-model presented here has several important advantages: (i) it is biologically motivated by phenomena such as pathogen mutation during infection, transmission of entire pathogens instead of proportions of trait values and within-/between-host selection; (ii) it is a fair test for all estimators of heritability, because it doesn't obey any of the estimators' assumptions, such as linearity of recipient- on donor values, normality of trait values, OU or BM evolution, independence between pathogen and host effects; (iii) it generates transmission trees that reflect the between-host dynamics, e.g. clades with higher trait values exhibit denser branching in cases of between-host selection. As a criticism, we note that the toy-model does not allow strain coexistence within a host. Although it may be exciting from biological point of view, the inclusion of strain coexistence comes with a series of conceptual challenges, such as the definition of genotype and clonal identity, the formulation of the trait value as a function of a quasispecies- instead of a single strain genotype, etc. At the same time, such an extension is unlikely to change our understanding of partial quasispecies transmission as a cause of negative bias in estimators of $H^2$. To conclude, the discrepancy between

simulation studies teaches that no method suits all simulation setups *ergo* biological contexts. Thus, rather than proving universality of a particular method, simulations should be used primarily to study how particular biologically relevant sources of bias affect the methods on table.

**The heritability of HIV set-point viral load is at least 20%.** Applied to data from the UK, ANOVA-CPP and POUMM reported four to five times higher point estimates and non-overlapping CIs compared to a previous PMM/ReML-based estimate on the same data (0.06, 95% CI [0.02, 0.09]) (15). Our PMM implementation confirmed this estimate. However, these results are based on assuming BM, which is clearly violated by the data. Based on our simulation study, these estimates are thus underestimates of the true heritability. To give an intuition about this negative bias, we note that BM assumes a linear increase in trait variance through time. Thus, fitting PMM to a trait with constant variance through time results in a vanishing time unit variance increment (speed) of the BM process leading to $\sigma_e^2$ explaining all the observed variance in the trait. Overall, our analyses yield an unprecedented agreement between estimates of donor-recipient resemblance and phylogenetic heritability in a large European dataset and African cohorts (11, 13) (Fig. 5A). All datasets support the hypothesis of HIV influencing spVL ($H^2 > 0.2$). The particular estimates provided here should be interpreted as lower bounds for $H^2$, because the partial quasispecies transmission, the noises in spVL measurements and transmission trees are included implicitly as environmental (non-transmittable) effects. These results motivate further HIV whole-genome sequencing (30) and genome-wide studies of the viral genetic association with viral load and virulence.

**Outlook.** Beyond HIV, ANOVA-CPP and POUMM have great potential to become widely used tools in the study of pathogens. ANOVA-CPP works on pairs of trait values from carriers of nearly identical strains and can be easily extended to groups of variable size (1, 6). Thus, ANOVA-CPP is ideal for slowly evolving pathogens such as DNA-viruses, bacteria and protozoa, where clusters of patients carrying identical-by-descent (IBD) strains are frequently found. For example, Anderson et al. 2010 identified 27 clusters of two to eight carriers of IBD strains in a small set of 185 malaria patients, i.e. 41% of the patients participated in clusters (6). On the other hand, IBD-pairs are rare for rapidly evolving RNA-viruses, such as HIV and HCV. For instance, we identified only 116 CPPs in a large dataset of 8483 HIV-sequences, i.e. less than 3% of the patients involved in IBD-pairs. However, the rapidly accumulating sequence diversity of RNA-viruses allows building large-scale phylogenies, which approximate transmission trees between patients. Thus, RNA-viruses should make the ideal scope for the POUMM. We believe that, together, the two methods should enable accurate and robust heritability estimation in a broad range of pathogens.

# Methods

## The phylogenetic Ornstein-Uhlenbeck mixed model (POUMM)

In this section we describe technical details concerning the PMM method and its extension advocated in this paper, the Phylogenetic Ornstein-Uhlenbeck Mixed Model (POUMM).

The PMM method (8, 25) is based on the simplest decomposition of the trait value in which the observed value of an individual is represented as a sum of a genetically determined (heritable) component, $G$, and an environmental (non-heritable) component, $e$:

$$z = G + e \qquad (3)$$

While quantitative genetics theory defines $G$ in a statistical sense as a mean observed phenotype of the carriers of a given genotype, i.e. configuration of alleles at quantitative trait loci in the genome (aka QTLs, defined in Supporting Information), the PMM method is ignorant about genotypes and defines $G$ as an unobserved value assigned at any point on the tree.

It is assumed that $G$ evolves according to a BM process. This means that at any point on any branch of the tree $G$ is a normally distributed random variable with mean $g_0$ and variance $\sigma^2_{BM} t$, where $g_0$ denotes the value of $G$ at the root, $\sigma^2_{BM}$ denotes the unit-time variance of the BM process and $t$ denotes the branch-distance from the point to the root. The covariance of the values $G$ at any two distinct points on the tree is equal to $\sigma^2_{BM} t_a$, where $t_a$ is the branch-distance from the most recent common ancestor of the two points to the root of the tree. Thus, the conditional likelihood of an observed vector of tip-values $\mathbf{z}$, given a phylogeny $\mathrm{T}$ and model parameters, is given by a multivariate normal probability density function $\mathcal{N}[\mathbf{z}; g_0, \Sigma_{BM}(\mathrm{T}, \sigma^2_{BM}, \sigma^2_e)]$. The variance-covariance matrix $\Sigma_{BM}(\mathrm{T}, \sigma^2_{BM}, \sigma^2_e)$ has diagonal elements equal to $\left(\sigma^2_{BM} t_i + \sigma^2_e\right)$ and off-diagonal elements equal to $\left(\sigma^2_{BM} t_{ij}\right)$, $t_i$ denoting the distance from the root to tip $i$ and $t_{ij}$ denoting the distance from the root to the most recent common ancestor of tips $i$ and $j$.

The **phylogenetic heritability**, $H^2_{BM}$, is defined as the proportion of phenotypic variance attributable to $G$ at the tips of the tree (8):

$$H^2_{BM}(t) = \sigma^2_{BM} t / (\sigma^2_{BM} t + \sigma^2_e) \qquad (4)$$

Note that $h^2_{BM}(t) \to 1$ as $t \to \infty$. As mentioned previously, in the case of a non-ultrametric phylogeny, the choice of $t$ in eq. 4 is not obvious. Thus, we prefer the following tree-independent formulation of phylogenetic heritability, assuming that the sample phenotypic variance at the tips, $s^2(z)$, is a good approximation of the phenotypic variance in the population:

$$H^2_{BMe} = 1 - \sigma^2_e / s^2(z) \qquad (5)$$

POUMM is defined similarly to PMM after replacing the BM assumption with an assumption of OU process acting on $G$ along the phylogeny and parameterized by a single global optimum value $\theta$, selection strength $\alpha > 0$ and unit-time variance $\sigma^2_{OU}$. Moreover, POUMM can be seen as a generalization of PMM as in the limit $\alpha \to 0$ the OU process is equivalent to a BM process with the same unit-time variance. Similarly to PMM, the conditional likelihood is a multivariate normal probability density function $\mathcal{N}[\mathbf{z}; \mu_{OU}(T,\alpha,\theta,g_0), \Sigma_{OU}(T,\sigma^2_{OU},\sigma^2_e)]$. Based on equations 1 and 6 in (28), we derive the expressions for the mean vector $\mu_{OU}$ and for the variance covariance matrix $\Sigma_{OU}$. The elements of the mean vector are defined by substituting the corresponding root-tip distance for $t$ in eq. 6 below:

$$\mu_{OU}(t;\alpha,\theta,g_0) = \exp(-\alpha t) g_0 + \left[1 - \exp(-\alpha t)\right]\theta \qquad (6)$$

The matrix $\Sigma_{OU}$ is defined by

$$\Sigma_{OU,(ii)} = \left[1 - \exp(-2\alpha t_i)\right]\sigma^2_{OU} \Big/ (2\alpha) + \sigma^2_e \qquad (7)$$

for the diagonal elements $(ii)$ and by

$$\Sigma_{OU,(ij)} = \exp(-\alpha \tau_{ij})\left[1 - \exp(-2\alpha t_{ij})\right]\sigma^2_{OU} \Big/ (2\alpha) \qquad (8)$$

for the off-diagonal elements $(ij)$, $t_i$ and $t_{ij}$ defined as above and $\tau_{ij}$ denoting the phylogenetic distance between tips $i$ and $j$.

The key difference between OU and BM processes is the fact that in the limit $t \to \infty$ OU converges to a stationary (equilibrium) normal distribution with mean $\theta$ and finite variance $\sigma^2_{OU}/(2\alpha)$, aka the "dispersion parameter" of the OU process (28).

For a given combination of parameters, the phylogenetic heritability estimated by the POUMM method is defined by eq. 9 as a function of time or distance from the root of the phylogeny.

$$H^2_{OU}(t) = \frac{\sigma^2_{OU}\left[1 - \exp(-2\alpha t)\right]}{\sigma^2_{OU}\left[1 - \exp(-2\alpha t)\right] + 2\alpha\sigma^2_e} \qquad (9)$$

At equilibrium, that is when $t \to \infty$, $H^2_{OU}(t)$ converges to a value, which we call **equilibrium heritability** and denote as $H^2_{eq}$ (eq. 10).

$$H^2_{eq} = \sigma^2_{OU} / (\sigma^2_{OU} + 2\alpha\sigma^2_e) \qquad (10)$$

Like in PMM, we defined a tree-independent version of the phylogenetic heritability and denoted it as $H^2_{OUe}$ (see eq. 5).

**Likelihood calculation**

To calculate the conditional likelihood of an observed vector of trait-values $\mathbf{z}_0$ at the tips of a given transmission tree $T$ and POUMM parameters $\Theta = (\alpha, \theta, \sigma_{OU}, \sigma_e)$, we implemented a variant of the Felsenstein's "pruning" algorithm (31), which relies on the

recursive factorization of the likelihood along $T$. In comparison with the standard approach to calculate multivariate normal densities based on equations 6, 7 and 8, our approach is considerably faster and more memory efficient, calculation time and memory consumption scaling linearly with the number of tips in the tree.

Now, we describe the technical details using Fig. 1 as illustration. We denote by $z_i$, $g_i$, $t_i$ the trait-value, heritable component and the length of the branch ending at node $i$. We denote by $\mathbf{z}_i$ the ensemble of trait-values of the tips descendant from node $i$ in the tree, e.g. $\mathbf{z}_4 = \{z_2, z_3\}$ on Fig. 1. Assuming for a moment that $g_0$ is known, the conditional likelihood is expressed as a product of two definite integrals:

$$\ell(\Theta, g_0) = f(\mathbf{z}_0 \mid \Theta, g_0) = \int_{-\infty}^{\infty} f(g_1 \mid \Theta, t_1, g_0)\, f(z_1 \mid \Theta, g_1)\, \mathrm{d}g_1 \times \int_{-\infty}^{\infty} f(g_4 \mid \Theta, t_4, g_0)\, f(\mathbf{z}_4 \mid \Theta, g_4)\, \mathrm{d}g_4 \tag{11}$$

The term $f(g_1 \mid \Theta, t_1, g_0)$ is a univariate normal density with mean and variance expressed as in equations 6, 7. This density can be written as an exponential of a polynomial of degree two of $g_1$ (eq. 12):

$$\begin{cases} f(g_1 \mid \Theta, t_1, g_0) = \exp(p_1 g_1^{\,2} + q_1 g_1 + r_1)\ , \ \text{where} \\ \\ p_1 = -\dfrac{\alpha \exp(2\alpha t_1)}{\sigma_{OU}^2 \left[\exp(2\alpha t_1) - 1\right]} \\ \quad = \dfrac{\exp(2\alpha t_1)\, a(\alpha, 2t_1)}{\sigma_{OU}^2} \quad , \text{substituting } \dfrac{\alpha}{1 - \exp(2\alpha t_1)} \text{ by } a(\alpha, 2t_1) \\ \\ q_1 = \dfrac{2\alpha \exp(\alpha t_1)\left\{g_0 + \theta\left[\exp(\alpha t_1) - 1\right]\right\}}{\sigma_{OU}^2 \left[\exp(2\alpha t_1) - 1\right]} \\ \quad = -\dfrac{2a(\alpha, 2t_1)\left\{g_0 + \theta\left[\exp(\alpha t_1) - 1\right]\right\}}{\sigma_{OU}^2} \\ \\ r_1 = -\dfrac{\alpha\left\{g_0 + \theta\left[\exp(\alpha t_1) - 1\right]\right\}^2}{\sigma_{OU}^2 \left[\exp(2\alpha t_1) - 1\right]} - \dfrac{1}{2}\ln\left(\dfrac{\pi \sigma_{OU}^2 \left[1 - \exp(-2\alpha t_1)\right]}{\alpha}\right) \\ \quad = \dfrac{a(\alpha, 2t_1)\left\{g_0 + \theta\left[\exp(\alpha t_1) - 1\right]\right\}^2}{\sigma_{OU}^2} - \dfrac{1}{2}\ln\left[-\dfrac{\pi \sigma_{OU}^2}{a(\alpha, 2t_1)\exp(2\alpha t_1)}\right] \end{cases} \tag{12}$$

Noticing that eq. 12 is not defined for $\alpha = 0$ (the case of Brownian Motion), we take the limit at $\alpha \to 0$, by defining the function $a(\alpha, t)$ as follows:

$$a(\alpha,t) = \begin{cases} \dfrac{\alpha}{1-\exp(\alpha t)}, & \text{for } \alpha > 0 \\ -1/t, & \text{for } \alpha = 0 \end{cases} \qquad (13)$$

The term $f(z_1 | \Theta, t_1, g_1)$ in eq. 11, which is a univariate normal density with mean $g_1$ and variance $\sigma_e^2$, can also be expressed as an exponential of a polynomial of degree two of $g_1$:

$$\begin{cases} f(z_1 | \Theta, g_1) = u_1 g_1^2 + v_1 g_1 + w_1, & \text{where} \\ u_1 = -\dfrac{1}{2\sigma_e^2} \\ v_1 = \dfrac{z_1}{\sigma_e^2} \\ w_1 = -\dfrac{z_1^{\,2}}{2\sigma_e^2} - \dfrac{1}{2}\ln\left(2\pi\sigma_e^2\right) \end{cases} \qquad (14)$$

If $p_1$ (eq. 12) and $u_1$ (eq. 14) satisfy $(p_1 + u_1) < 0$, the first integral in eq. 11 can be solved in closed form, which, again, is an exponential of a polynomial of degree two of $g_0$ :

$$
\begin{cases}
\int_{-\infty}^{\infty} \exp\left[(p_1+u_1)g_1^2+(q_1+v_1)g_1+(r_1+w_1)\right]\mathrm{d}g_1 \\
= \exp\left[\dfrac{-(q_1+v_1)^2}{4(p_1+u_1)}+(r_1+w_1)+\ln\left(\sqrt{\dfrac{\pi}{-(p_1+u_1)}}\right)\right] \\
= \exp(u_{01}g_0^2+v_{01}g_0+w_{01}) \quad , \quad \text{where} \\
\\
u_{01} = \dfrac{u_1 a(\alpha,2t_1)}{a(\alpha,2t_1)-\alpha+\sigma_{OU}^2 u_1} \\
\\
v_{01} = \dfrac{a(\alpha,2t_1)(\exp(\alpha t_1)(2\theta u_1+v_1)-2\theta u_1)}{a(\alpha,2t_1)-\alpha+\sigma_{OU}^2 u_1} \\
\\
w_{01} = w_1+t_1\alpha-\dfrac{0.25 v_1^{\,2}\sigma_{OU}^2}{-\alpha+u_1\sigma_{OU}^2+a(\alpha,2t_1)}-\dfrac{\ln\left[-\alpha+u_1\sigma^2+a(\alpha,2t_1)\right]}{2a(\alpha,2t_1)} \\
\qquad +\dfrac{\alpha\theta\left[u_1\theta-(v_1+u_1\theta)\exp(\alpha t_1)\right]}{a(\alpha,t_1)+(-\alpha+u_1\sigma^2)\left[1+\exp(\alpha t_1)\right]}
\end{cases}
\tag{15}
$$

By applying the same factorization and integration technique to the sub-tree from node 4, we find:

$$
\begin{aligned}
f(\mathbf{z}_4|\Theta,g_4) &= \int_{-\infty}^{\infty} f(g_2|\Theta,\tau_2,g_4)\, f(z_2|\Theta,g_2)\mathrm{d}g_2 \times \\
&\quad \int_{-\infty}^{\infty} f(g_3|\Theta,\tau_3,g_0)\, f(z_3|\Theta,g_3)\mathrm{d}g_3 \\
&= \exp\left(u_{42}g_4^2+v_{42}g_4+w_{42}\right)\times \\
&\quad \exp\left(u_{43}g_4^2+v_{43}g_4+w_{43}\right) \\
&= \exp\left(u_4 g_4^2+v_4 g_4+w_4\right), \ \text{where} \\
&\quad u_4 = u_{42}+u_{43},\ v_4 = v_{42}+v_{43},\ \mathrm{w}_4 = w_{42}+w_{43}
\end{aligned}
\tag{16}
$$

Then, we express the term $f(g_4|\Theta,t_4,g_0)$ in the same way as in eq. 12:

$$
f(g_4|\Theta,t_4,g_0) = \exp(p_4 g_4^2+q_4 g_4+r_4) \tag{17}
$$

Multiplying the exponentials from eq. 16 and eq. 17 and repeating the integration (eq. 15), we obtain the second integral in eq. 11 in the form of an exponential of a polynomial of degree two of $g_0$. Thus, the likelihood is:

$$\ell(\Theta, g_0) = f(\mathbf{z}_0 \mid \Theta, g_0) = \exp(u_0 g_0^2 + v_0 g_0 + w_0), \text{ where}$$
$$u_0 = u_{01} + u_{04}, \mathrm{v}_0 = v_{01} + v_{04}, \mathrm{w}_0 = w_{01} + w_{04} \quad (18)$$

Recall that $g_0$ is an unknown parameter. Thus, we maximize $\ell(\Theta, g_0)$ over $g_0$ by setting $g_0 = -0.5 v_0 / u_0$.

This recursive algorithm was implemented using dynamic programming maximizing the use of vector over scalar operations. The implementation can be found in the function lik.poumm of the accompanying R-package "patherit". The accuracy was validated against the multivariate normal density implementation using equations 6-8, as well as an alternative implementation based on the function "make.ou" in the R-package diversitree(32) (see also file CompareOUPackages.Rmd, Supporting programs). This validation procedure showed that the patherit package is numerically stable in cases when the other implementations returned infinite or NA values and is also about 30 times faster than the diversitree-based implementation using the C-backend (more than 100 times faster if using the R-backend).

**Maximum likelihood (ML) inference**

In the POUMM ML-fits, the conditional likelihood of the data was maximized over the parameters $\alpha$, $\theta$, $\sigma_{OU}$, $\sigma_e$ and $g_0$ (function ml.poumm of the patherit package).

In the PMM ML fits the conditional likelihood of the data was redefined as its corresponding limit for $\alpha \to 0$ and was maximized over the parameters $\sigma_{OU}$, $\sigma_e$ and $g_0$ (ignoring $\theta$, which cancels out in the case $\alpha \to 0$).

The values for $H^2_{BM}$ and $H^2_{OU}$ were calculated from eq. 4 and eq. 9 after setting $t$ to the maximum root-tip distance in the tree. The values for $H^2_{BMe}$ and $H^2_{OUe}$ were calculated using eq. 5.

**Bayesian inference**

For HIV-1 data, we performed a Markov Chain Monte Carlo (MCMC) fit (function mcmc.poumm of the patherit package) using an adaptive Metropolis algorithm with coerced acceptance rate (33) written in R (34).

The MCMC sampling was performed on the POUMM parameters $\alpha$, $\theta$, $\sigma^2_{OU}$ and $\sigma^2_e$. The prior was specified as a joint distribution of four independent variables: $(\alpha, \theta, \sigma^2_{OU}, \sigma^2_e) \sim \mathrm{Exp}(0.01) \times \mathcal{U}(0,100) \times \mathrm{Exp}(0, 10^{-4}) \times \mathrm{Exp}(0.01)$. The exponential rates above and the interval of the uniform distribution have been chosen such to ensure that the prior is uninformed, both, for the sampled parameters $\alpha$, $\theta$, $\sigma^2_{OU}$, $\sigma^2_e$ and for the inferred heritability estimates $H^2_{OU}$, $H^2_{OUe}$ (prior densities denoted as blue curves on Fig. S3B). The initial values for the parameters were set to $(\alpha, \theta, \sigma^2_{OU}, \sigma^2_e)_0 = (0,0,1,1)$.

The adaptive Metropolis MCMC was run for 4.2E+06 iterations, of which the first 2E+05 were used for warm-up and adaptation of the jump distribution variance-covariance matrix. The target acceptance rate was set to 0.01 and the thinning interval was set to 1,000. The convergence and mixing of the MCMC was validated by visual analysis

(Fig. S3A). The presence of signal in the data was confirmed by the observed significant difference between prior (blue) and posterior (black) densities (see Fig. S3B). Calculation of 95% CI has been done using the function "HPDinterval" from the coda package.

## Direct measurement of $H^2$ in simulated data

To measure $H^2$, we used the direct estimate $R^2_{adj}$ (eq. S3) after grouping the patients in the data by their (currently carried) pathogen genotype and estimating the genotypic values as the group means (implemented as function R2adj in the patherit package).

## Calculating $b[0]$, $b[\tau \leq D_1]$ and $b[\tau]$

The value of the donor-recipient regression slope was calculated using eq. S5, implemented as a function called "b" in the patherit package.

## Calculating $r_A$

To estimate $r_A$ we implemented one-way ANOVA as a function "rA" in the package patherit. As a reference we used the description in chapter 18 of (1). To calculate confidence intervals, we used the R-package "boot" to perform 1,000-replicate bootstraps, upon which we called the package function boot.ci() with type="basic". These confidence intervals were fully contained in the standard ANOVA confidence intervals based on the F-distribution (see (1)), which were slightly wider (not reported).

## Identifying outlier CPPs

Outlier CPPs were identified as CPPs having absolute phenotypic difference below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$, $Q_1$ and $Q_3$ denoting the 25th and 75th quantile of $|\Delta \lg(\text{spVL})|$ in CPPs and IQR denoting the interquartile range Q3-Q1.

## Computer simulations of the toy epidemiological model

We simulated six possible genotypes (pathogen strains) defined as combinations of $M_1 = 3$ and $M_2 = 2$ possible alleles at each locus and denoted 1:11, 2:12, 3:21, 4:22, 5:31, 6:32 (Fig. 2A). We assumed absence of strain coexistence within a host, so that at any moment one strain represented the pathogen quasispecies in a host. At a time $t$, the value $z_i(t)$ of an infected individual $i$ was modeled as a function of its immune system type $\mathbf{y}_i \in \{1,2\}$, the currently carried strain $\mathbf{x}_i(t) \in \{1,\ldots,6\}$, and the individual's specific effect for this strain $e_i[\mathbf{x}_i(t)] \sim \mathcal{N}(0,0.36)$ drawn at random for each strain (in each infected individual). We call a (type **y**-**x**) **general effect** the expected trait value of type-$\mathbf{y}$ carriers of strain $\mathbf{x}$ in an infected population: $GE[\mathbf{y},\mathbf{x}] = \mathrm{E}[z|\mathbf{y},\mathbf{x}]$. For a set of fixed general effects, $z_i(t)$ was constructed according to the equation:

$$z_i(t) = GE[\mathbf{y}_i, \mathbf{x}_i(t)] + e_i[\mathbf{x}_i(t)] \qquad (19)$$

We used a fixed set of general effects drawn from the uniform distribution $\mathcal{U}(2,4)$ for the twelve **y**-**x** combinations (Fig. 2A).

We embedded this toy-model into a stochastic Susceptible-Infected-Recovered (SIR) model of an epidemic with demography and frequency dependent transmission as described in (29), ch. 1. Each infected individual, *i*, had a variable trait value $z_i(t)$ constructed as in eq. 19. Within-host phenomena (strain mutation and substitution) and between-host phenomena (natural birth, contact, transmission, diagnosis, recovery and death) occurred at random according to Poisson processes (Fig. 1). The rate parameters defining these processes are written in Table S1 and described shortly.

For each group of parameters (within- and between-host), we considered the following two modes of dynamics:

- neutral: rates were defined as global constants mimicking neutrality (i.e. lack of selection) w.r.t. $z$ (black lines on Fig. 2B-D). For within-host phenomena, it was assumed that a mutation of the pathogen was followed by instantaneous substitution of the mutant for the current dominant strain, regardless of the induced change in $z$ (black line on Fig. 2E);
- select: borrowing the approach from (35), the rates of transmission and within-host pathogen mutation were defined as increasing Hill functions of $10^z$, while the infected death rate was defined as an inverse decreasing Hill function of $10^z$, thus mimicking increasing per capita transmission- and pathogen-induced mortality for higher $z$ (red lines on Fig. 2B-D). Within hosts, it was assumed that a mutation of the pathogen was followed by instantaneous substitution only if it resulted in a higher $z$. Otherwise, the mutation was considered deleterious (red line on Fig. 2E).

At the between-host level, the phenomena of birth, contact, transmission, recovery and death define the dynamics between the compartments of susceptibles, infected and recovered individuals - *X*, *Y* and *Z*. The natural birth rate, $\alpha_{\text{nat}}$, and the natural per capita death rate, $\delta_{\text{nat}}$, have been defined as constants satisfying $\alpha_{\text{nat}} = \delta_{\text{nat}} N_0$, so that the average lifespan of an uninfected individual equals $1/\delta_{\text{nat}} = 850$ (arbitrary) time units and in a disease-free population the total number of alive individuals equilibrates at $N_0 = 10^5$. An epidemic starts with the migration of an individual with random immune system type carrying pathogen strain 1:11 to a fully susceptible population of $N_0$ individuals. Each individual has contacts with other individuals occurring randomly at a constant rate, $\kappa$. A transmission can occur upon a contact involving an infected and a susceptible individual, here, called a "risky" contact. It is assumed that the probability of transmission per risky contact, $\gamma$, is either a constant (black on Fig. 2B) or a function of the value $z$ (red on Fig. 2B) of the infected host and does not depend on the uninfected individual. Once infected, a host starts transmitting its currently dominant pathogen strain at a rate defined as the product of $\gamma$, $\kappa$, and the current proportion of susceptible individuals in the population, $S = X/N$. Thus, for fixed $\kappa$, the transmission rate of an infected host is a function of the global variable $S$ and the constant or variable $\gamma$. This transmission process continues until recovery or death of the host. Recovery has the meaning of a medical check occurring at a constant per capita rate, $\rho$, followed by immediate therapy and immunity. Due to the virulence of the pathogen, an infected host has an increased (per capita) death rate, $\delta$, which is defined either as a constant or as a function of $z$. Based on their scope of action, we call "between-host" the parameters $\alpha_{\text{nat}}$, $\delta_{\text{nat}}$, $\kappa$, $\gamma$, $\rho$ and $\delta$.

Within a host, mutants of the dominant strain can appear at any time as a result of random single-locus mutations, which occur at a constant or $z$-dependent rate, $\nu$. It is important to make a distinction between a mutation and a substitution of a mutant strain for a dominant strain within a host, because a mutation doesn't necessarily lead to a substitution. For example, when $z$ is (or correlates with) the within-host reproductive fitness of the pathogen, substitutions would result only from mutations causing an increase in $z$. The rate of substitution of a mutant strain $\mathbf{x}_j$ for a dominant strain $\mathbf{x}_i$, differing by a single nucleotide at a locus $l$, is denoted $\xi_{l,i\leftarrow j}$ and defined as a function of $\nu$, the number of alleles at the locus, $M_l$, and the presence or absence of within-host selection with respect to $z$. No substitution can occur between strains differing at more than one locus, although, the same effect can result from two or more consecutive substitutions. Based on their scope of action, we call "within-host" the parameters $\nu$ and $\xi$.

The parameters $\alpha_{\text{nat}}$, $\delta_{\text{nat}}$, $\kappa$ and $\rho$ were kept as global constants as written in Table S1. The simulations were run for four times the time until reaching 10,000 recovered patients, hereafter denoted as $t_{10k}$, but not longer than 2400 time units. The transmission history as well as the history of within-host strain substitutions was preserved during the simulations in order to reproduce exact transmission trees at different time-points and to extract donor and recipient values at moments of transmission for the calculation of $b[0]$.

By combining "neutral" and "select" dynamics for the strain mutation and substitution rates at the within-host level, and the virus-induced per capita death rate and per contact transmission probability at the between-host level, we created the following four scenarios (Fig. 3E):

- Within: neutral / Between: neutral;
- Within: select / Between: neutral;
- Within: neutral / Between: select;
- Within: select / Between: select;

For each of these scenarios and mean contact interval $1/\kappa \in \{2,4,6,8,10,12\}$ (arbitrary time units), we performed ten simulations resulting in a total of $4\times6\times10=240$ simulations, of which 175 simulations resulted in outbreaks with more than 1,000 diagnosed cases. In each of these 175 simulations we analyzed the population of the first up to 10,000 diagnosed individuals. We denote this set of individuals by $Z_{10\text{k}}$ and the corresponding transmission tree – by $T_{10\text{k}}$. The direct estimate of broad-sense heritability, $R^2_{adj}$ (defined in Supporting Information) has been compared to the following: $b[0]$ in all transmission couples found in $Z_{10\text{k}}$; $b[\tau]$ in the same transmission couples; $b[\tau \le D_1]$ in transmission couples in $Z_{10\text{k}}$ having $\tau$ not exceeding the first decile ($10^{\text{th}}$ quantile) of $\tau$'s - $D_1$; $r_A[\text{id}]$ based on grouping by identity of carried strain in $Z_{10\text{k}}$; $r_A[\text{PP}:\tau]$ based on phylogenetic pairs in $T_{10\text{k}}$; $r_A[\text{PP}:\tau \le D_1]$ based on closest phylogenetic pairs (CPP) defined as PPs in $T_{10\text{k}}$ having $\tau$ not exceeding the first decile of $\tau$'s among all PPs in

$T_{10k}$; $H^2_{BMe}$ and $H^2_{OUe}$ based on the maximum likelihood (ML) fit of the PMM and POUMM methods on $T_{10k}$. To calculate $b[0]$, we used the immediate trait-values at moments of transmission (usually not available in practice). All other estimators were calculated using trait-values at diagnosis.

The toy-model SIR simulation is implemented in the function "simulateEpidemic" of the patherit package; the extraction of diagnosed donor-recipient couples – in the function "extractDRCouples"; the extraction of a transmission tree from diagnosed individuals – in the function "extractTree".

## Acknowledgements

We thank Dr. Emma Hodcroft for sending us the UK phylogeny in Newick format together with the spVL measurements. We thank Dr. Gabriel Leventhal and prof. Sebastian Bonhoeffer for valuable insights on donor-recipient regression. We thank Dr. David Rasmussen for a careful review of the manuscript.

## Funding

V.M. and T.S. thank ETH Zürich for funding. T.S. is supported in part by the European Research Council under the 7th Framework Programme of the European Commission (PhyPD: Grant Agreement Number 335529).

## References

1. Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits* (Sinauer Associates Incorporated).

2. Falconer DS (1996) *Introduction to Quantitive Genetics* (San Val, Incorporated).

3. Hartyl DL, Clark AG (2007) *Principles of population genetics* (Sinauer Associates).

4. Fraser C, et al. (2014) Virulence and Pathogenesis of HIV-1 Infection: An Evolutionary Perspective. *Science* 343(6177):1243727–1243727.

5. Leventhal GE, Bonhoeffer S (2016) Estimating the heritability of viral traits using phylogenetic trees. *bioRxiv*:046797.

6. Anderson TJC, et al. (2010) Inferred relatedness and heritability in malaria parasites. *Proc Biol Sci* 277(1693):2531–2540.

7. Shirreff G, et al. (2013) How effectively can HIV phylogenies be used to measure heritability? *Evolution, Medicine, and Public Health* 2013(1):209–224.

8. Housworth EA, Martins EP, Lynch M (2004) The phylogenetic mixed model. *Am Nat* 163(1):84–96.

9. Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence. *Am Nat* 160(6):712–726.

10. Tang J, et al. (2004) HLA allele sharing and HIV type 1 viremia in

seroconverting Zambians with known transmitting partners. *AIDS Res Hum Retroviruses* 20(1):19–25.

11. Hecht FM, et al. (2010) HIV RNA level in early infection is predicted by viral load in the transmission source. *AIDS* 24(7):941–945.

12. van der Kuyl AC, Jurriaans S, Pollakis G, Bakker M, Cornelissen M (2010) HIV RNA levels in transmission sources only weakly predict plasma viral load in recipients. *AIDS* 24(10):1607–1608.

13. Hollingsworth TD, et al. (2010) HIV-1 transmitting couples have similar viral load set-points in Rakai, Uganda. *PLoS Pathog* 6(5):e1000876.

14. Alizon S, et al. (2010) Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathog* 6(9):e1001123.

15. Hodcroft E, et al. (2014) The Contribution of Viral Genotype to Plasma Viral Set-Point in HIV Infection. *PLoS Pathog* 10(5):e1004112.

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

17. Mellors JW, et al. (1996) Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* 272(5265):1167–1170.

18. Geskus RB, et al. (2007) The HIV RNA setpoint theory revisited. *Retrovirology* 4(1):65.

19. Jacquard A (1983) Heritability: One Word, Three Concepts. *Biometrics* 39(2):465.

20. Keele BF, et al. (2008) Identification and characterisation of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci USA* 105(21):7552–7557.

21. Wright S (1934) The Method of Path Coefficients. *The Annals of Mathematical Statistics* 5(3):161–215.

22. Lorenzo-Redondo R, et al. (2016) Persistent HIV-1 replication maintains the tissue reservoir during therapy. *Nature* 530(7588):51–56.

23. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature reviews Genetics* 12(9):628–640.

24. Hu S, Clewley JP, Cane PA, Pillay D (2004) HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 18(5):719–728.

25. Lynch M (1991) Methods for the Analysis of Comparative Data in Evolutionary

Biology. *Evolution* 45(5):1065–1080.

26. LANDE R (1976) Natural-Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution* 30(2):314–334.

27. Felsenstein J (1988) Phylogenies And Quantitative Characters. *Annual Review of Ecology and Systematics* 19(1):445–471.

28. Hansen TF (1997) Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution* 51(5):1341–1351.

29. Keeling MJ, Rohani P (2011) *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press).

30. Metzner KJ (2016) HIV Whole-Genome Sequencing Now: Answering Still-Open Questions. *J Clin Microbiol* 54(4):834–835.

31. Felsenstein J (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet* 25(5):471–492.

32. FitzJohn RG (2012) Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution* 3(6):1084–1092.

33. Vihola M (2012) Robust adaptive Metropolis algorithm with coerced acceptance rate. *Stat Comput* 22(5):997–1008.

34. Scheidegger A adaptMCMC: Implementation of a generic adaptive Monte Carlo Markov Chain sampler. Available at: http://CRAN.R-project.org/package=adaptMCMC.

35. Fraser C, Hollingsworth TD, Chapman R, de Wolf F, Hanage WP (2007) Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proc Natl Acad Sci USA* 104(44):17441–17446.

36. Fisher RA (1925) *Statistical Methods For Research Workers* (Genesis Publishing Pvt Ltd).

37. Donner A (1986) A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model. *International Statistical Review* 54(1):67–82.

38. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20(2):289–290.

39. Dowle M, Short T, Liangolou S, Srinivasan A (2014) data.table: Extension of data.frame. 9.

40. Maechler M Rmpfr: R MPFR - Multiple Precision Floating-Point Reliable. Available at: http://CRAN.R-project.org/package=Rmpfr.

41. Plummer M, Best N, Cowles K, Vines K CODA: Convergence Diagnosis and

Output Analysis for MCMC. Available at: http://CRAN.R-project.org/doc/Rnews/.

**Supporting Information:**

Supporting text

Supporting programs

Figures S1-S3

Tables S1-S4

## Figures in the main text

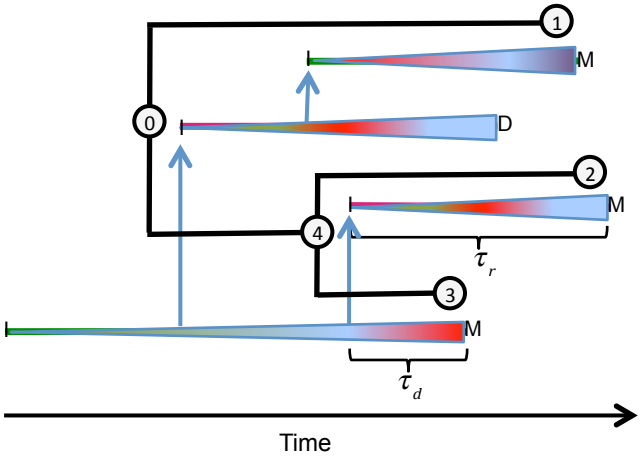


**Fig. 1**

A schematic representation of an epidemic. Colored rectangles represent infectious periods of hosts, different colors corresponding to different host-types. Triangles inside hosts represent pathogen quasispecies, change of color indicating substitution of dominant strains. Capital letters denote host-events as follows: I: beginning of infection, M: phenotype measurement and recovery D: host death. Arrows show the time and direction of transmission events. The within-donor and within-recipient measurement delays, $\tau_d$ and $\tau_r$, are shown for one donor-recipient couple. The transmission tree connecting the measured hosts is drawn in black, numbered circles indicating the root 0, the internal node 4 and the tips 1, 2 and 3.

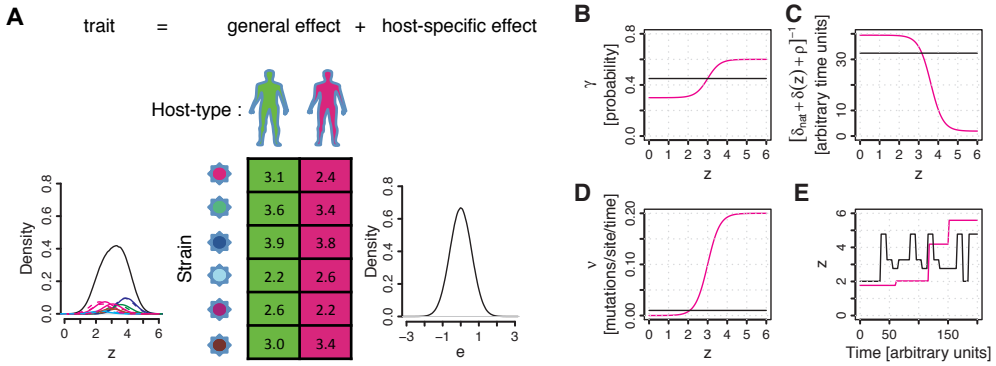


**Fig. 2**

A toy model of an epidemic. **A**, schematic representation of a pathogen trait formed from a general <host-type × carried-strain> effect and a host-specific effect (eq. 19). The density of the trait-values in a population represents a mixture of normal densities corresponding to each one of twelve host-type × strain combinations, scaled by their frequencies (dashed-lines depict host-type 2); **B-E**, SIR dynamics, color indicating selection modes w.r.t. $z$ , black – neutral, magenta – select (as specified in Table S1): **B**, per risky contact transmission probability; **C**, expected infectious period if no mutation happens; **D**, per site mutation rate; **E**, example within-host evolution;

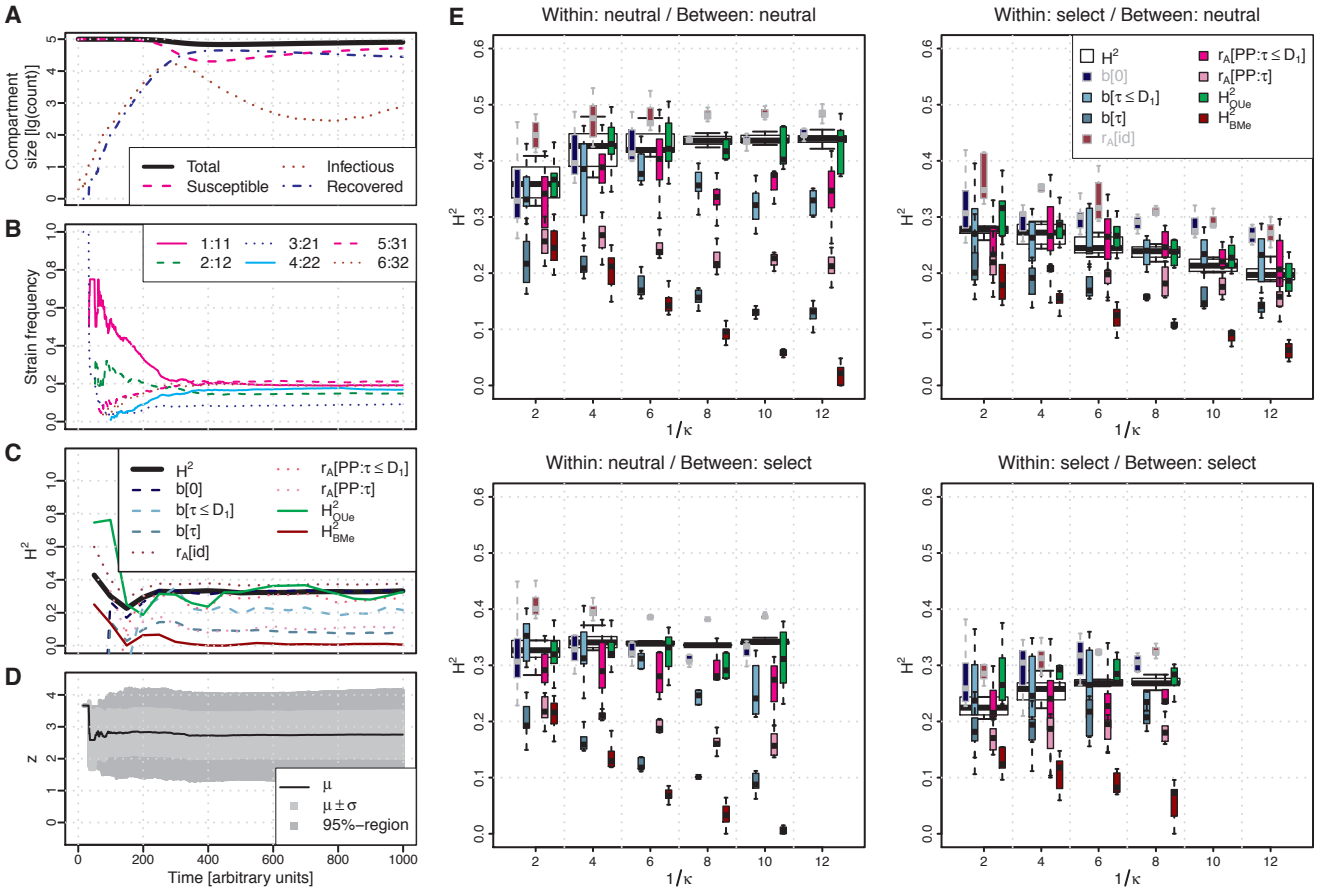


**Fig. 3**

Toy model simulations. **A-D**, time-profile of one simulation: **A**, sizes of the Susceptible, Infectious and Recovered (diagnosed) compartments; **B**, strain frequencies in samples of up to 10,000 most recently diagnosed hosts w.r.t. time; **C**, estimators of $H^2$ in samples of up to 10,000 most recently diagnosed patients w.r.t. 20 equidistant time points (for notation, see text); **D**, trait-distribution in samples as above; **E**, $H^2$-estimates in simulations of "neutral" and "select" within-/between-host dynamics. Each box-group summarizes simulations (first up to 10,000 diagnoses) at a fixed contact rate, $\kappa$ (Methods); white boxes (background) denote true heritability, colored ones denote estimates (foreground).

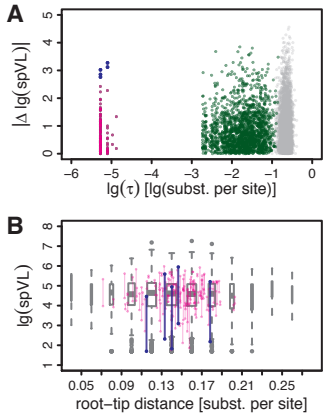


**Fig. 4**

A graphical analysis of HIV-data from UK. **A**, a scatter plot of the phylogenetic distances between pairs of tips against their absolute phenotypic differences: grey – random pairs; green – PPs ($\tau > 10^{-4}$); magenta: CPPs ($\tau \leq 10^{-4}$); blue – outlier CPPs (CPPs, for which $|\Delta\lg(\text{spVL})| > Q_3 + 1.5\times(Q_3 - Q_1)$, $Q_1$ and $Q_3$ denoting the 25$^{\text{th}}$ and 75$^{\text{th}}$ quantile of $|\Delta\lg(\text{spVL})|$ among CPPs); **B**, a box-plot representing the trait-distribution along the transmission tree. Each box-whisker represents the $\lg(\text{spVL})$-distribution of patients grouped by their distance from the root of the tree measured in substitutions per site. Wider boxes indicate groups bigger in size. Bullet-ending segments denote $\lg(\text{spVL})$-values in CPPs.

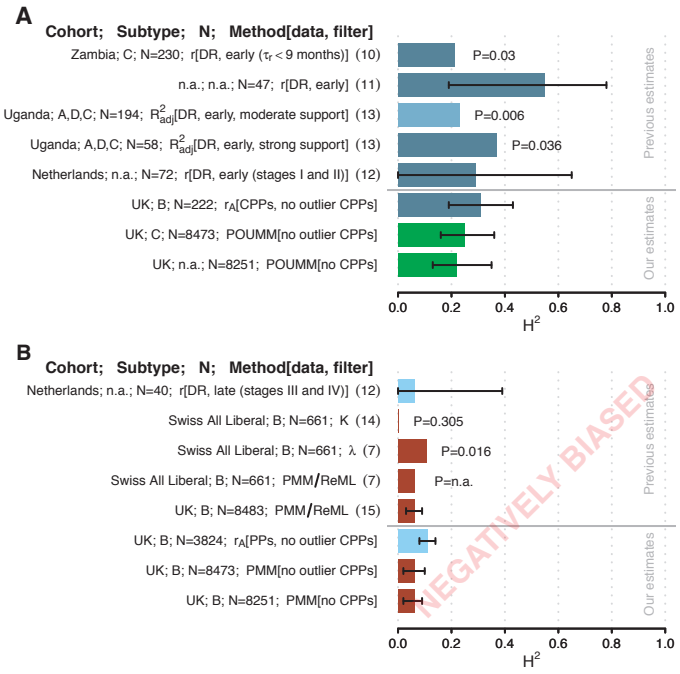


**Fig. 5**

Comparing our $H^2$ estimates from the UK HIV-cohort with previous estimates on African and Dutch cohorts. **A**, estimates with minimized measurement delay (dark cadet-blue) and POUMM estimates (green); **B**, down-biased estimates due to higher measurement delays (light-blue) or violated BM-assumption (brown). Confidence is depicted either as segments indicating estimated 95% CI or P-values in cases of missing 95% CIs. For clarity, the figure does not include estimates from the UK data including the five outlier CPPs (see Table S4) and estimates from previous studies, which are not directly comparable (e.g. previous results from Swiss MSM/strict datasets (14)).

# Supporting Information:

## Supporting Text

### ST1. Definitions of heritability and its resemblance-based estimators

Here, we review the formal definitions of heritability in sexually reproducing populations based on the general linear model of quantitative traits (1) and the three concepts introduced in the main text: the genetic determination of a trait, the resemblance between relatives, and the efficiency of selection (19).

**The general linear model of a quantitative trait**

A principal goal of quantitative genetics is to partition the observed phenotypic variance in a population into components attributable to genetic and environmental factors. Fundamental for the study of the genetic and environmental sources of variance is the general linear model for the phenotype (see (1), ch. 6), in which, for a given trait of interest, the observed phenotypic value, $z$, of an organism is represented as a sum of effects of the organism's genes, $G$, general (macro-) environmental effects, $E$, gene by (macro-) environment interaction, $I$, and special (micro-) environmental effects $e$:

$$z = G + I + E + e \qquad \text{(S1)}$$

It is assumed that the trait is influenced by a number of genes whose locations in the species' reference genetic sequence are called quantitative trait loci (QTL). In an individual, the configuration of alleles found at the trait's QTLs is called genotype and, for a population, the genotypic value, $G_{\mathbf{x}}$, of a genotype $\mathbf{x}$ is defined as the expected trait value of its carriers: $G_{\mathbf{x}} = \mathrm{E}(z \mid \text{genotype} = \mathbf{x})$. The remaining terms in eq. S1 are "defined in a least-squares sense as deviations from lower order expectations" (1). It is worthy to note that $G_{\mathbf{x}}$ depends on the distribution of $\mathbf{x}$ across environments in the population and that, by construction, the residuals $z - G = I + E + e$ have zero mean and are uncorrelated with $G$ (ch. 6 in (1)). Thus, the total phenotypic variance observed in the population can be partitioned into a component that is purely genetic and a component that is attributable to both, non-genetic (purely environmental) factors as well as gene-by-environment interactions: $\sigma^2(z) = \sigma^2(G) + \sigma^2(z - G)$.

**Measuring the genetic determination of a trait**

**Heritability in the broad sense**, aka degree of genetic determination (2), is defined as the ratio of the variance of genotypic values to total phenotypic variance in the population:

$$H^2 = \sigma^2(G) / \sigma^2(z) \qquad \text{(S2)}$$

A direct estimation of $H^2$ would require that all QTLs were known and that for each genotype there was a sample of measurements from individuals who were: (i) genetically identical at the QTLs; (ii) raised in randomly and independently assigned environments; (iii) present in the final dataset according to the population-specific environment-genotype frequencies. Given such a dataset of $N$ independent measurements from carriers of all $K$ distinct genotypes in the population ($K \ll N$), $H^2$ can be estimated by the ratio of sample

variances $s^2(\hat{G})/s^2(z)$, where $\hat{G}$ denotes the individuals' genotypic values estimated by the mean value of their corresponding group and $s^2(.)$ denotes sample variance. Though, intuitive, this formula is slightly positively biased in the case of finite sample size. Thus, we prefer its correction for finite degrees of freedom, aka as adjusted coefficient of determination:

$$R^2_{adj} = 1 - \frac{N-1}{N-K}\frac{s^2(z-\hat{G})}{s^2(z)} \qquad \text{(S3)}$$

In the absence of full QTL information and data from independently grown clones, direct estimation of $H^2$ is rarely possible. Instead, quantitative geneticists focus on estimating its lower bound defined below.

**Heritability in the narrow sense** is defined as the ratio of variance of additive genetic values to total phenotypic variance:

$$h^2 = \sigma^2(A)/\sigma^2(z) \qquad \text{(S4)}$$

The additive genetic value, $A$, of an organism is defined as the sum of additive effects of its alleles at the trait's QTLs. We provide the technical definition of additive effect later on and note here that $h^2$ represents the largest proportion of phenotypic variance that can be explained by linear regression on the allele contents at single QTLs, ignoring epistatic (inter-locus) and dominance interactions (1). As discussed shortly, for sexually reproducing species, $h^2$ has two main advantages to $H^2$: (i) it can be estimated from empirical data of genetically related (but not identical) organisms; (ii) it can be used to predict the response to selection for traits associated with reproductive fitness.

**Measuring the resemblance between relatives**

Relatives resemble each other not only for carrying similar sets of alleles but also for living in similar environments. Thus, it is necessary to disentangle the concept of resemblance from that of genetic determination.

Considering an ordered relationship such as parent-offspring, the least squares regression slope of offspring values on mean parental values is defined as

$$b = s(z_o, z_{mp})/s^2(z_{mp}), \qquad \text{(S5)}$$

where $z_o$ and $z_{mp}$ denote observed offspring and mean parent values, and $s(.,.)$ denotes sample covariance among observed couples of values (1). Assuming no systematic dissimilarity between parents and offspring, $b$ is a value between 0 and 1, higher values indicating closer resemblance between the expected phenotype of offspring and the mid-phenotype of their parents.

Considering members of unordered relationships, such as identical twins, sibs and cousins, the resemblance between members within groups is measured by the intraclass correlation (ICC) defined as the ratio of the "between group" variance over the total variance, $r = \sigma^2(c)/\sigma^2(z)$, $c$ denoting the observed within-group means (1, 36). Given a dataset of measurements grouped by a factor such as twinship, the standard estimation procedure for $r$ is the one-way analysis of variance - ANOVA (see, e.g. (37) or ch. 18 in (1)). ANOVA uses

mean squares to find estimators for the between- and within-group variances, $\hat{\sigma}^2(c)$ and $\hat{\sigma}^2(z-c)$ and reports ICC as the ratio:

$$r_A = \frac{\hat{\sigma}^2(c)}{\hat{\sigma}^2(c)+\hat{\sigma}^2(z-c)}. \qquad \text{(S6)}$$

We notice that both, $R^2_{adj}$ (eq. S3) and $r_A$ (eq. S6), are estimators of ICC, but there is a key difference in their assumptions: $R^2_{adj}$ assumes that all possible groups, i.e. genotypes, are present in the data but makes no explicit assumption about the distribution of group means (i.e. genotypic values); $r_A$ is aware that only a subset of all possible groups is present in the data but assumes that the observed group means, are an iid sample from a normal distribution.

**Measuring the efficiency of selection**

In breeding experiments the goal is to optimize a trait by repetitive artificial selection for reproduction of the "best" individuals in a generation. A textbook example is truncation selection in which only individuals with measurements above a given threshold are allowed to reproduce. For a generation, the difference $\Delta_s = \mu_s - \mu$ between the mean value of individuals selected for reproduction, $\mu_s$, and the mean of the generation, $\mu$, is called the selection differential. Denoting by $\mu_o$ the mean of the offspring generation, the difference $R = \mu_o - \mu$, is called the response to selection. Then, the efficiency of the truncation selection is measured by the **realized heritability** (3), defined as the ratio:

$$h^2_R = R / \Delta_s \qquad \text{(S7)}$$

**Definition of additive genetic effect and additive genetic value**

So far, we have skipped the more technical definition of additive genetic effect, which is the basis of the definitions of additive genetic value and narrow-sense heritability. Here we provide these definitions in the context of haploid organisms, noting that the definitions for diploid organisms found in textbooks (1, 2) are conceptually the same but somewhat more complicated for they treat dominance interactions separately from epistatic interactions.

We will assume that a trait has a finite number of QTLs, *L*, with a finite number of alleles $M_l \geq 2$ for each locus $l = 1,\ldots,L$. Denoting by $x_{lm}$ the content (0 or 1) of allele $m$ at locus $l$, $l = 1,\ldots,L$, $m = 1,\ldots,M_l$, we can describe an individual's genotype by a binary vector $\mathbf{x}$ of length $\sum_{1\leq l\leq L} M_l$. The products of allele contents for different loci signify the presence or absence of allele combinations in a genotype. This representation results in the system of equations S8, in which the genotypic value of each genotype $\mathbf{x}$ is written as a sum of the population mean, $\mu$, and the effects $\eta_{lm}$, $(\eta\eta)_{l_1 m_1 l_2 m_2}$ and so on, associated with each allele, couple of alleles at two loci and higher order- (up to order *L*) multi-locus configurations of alleles, present in the genotype.

$$G_{\mathbf{x}} = \mu + \sum_{l \le L} \sum_{m \le M_l} \eta_{lm} x_{lm} + \sum_{l_1 \neq l_2} \sum_{\substack{m_1 \le M_{l_1} \\ m_2 \le M_{l_2}}} (\eta\eta)_{l_1 m_1 l_2 m_2} x_{l_1 m_1} x_{l_2 m_2} + \ldots \qquad \text{(S8)}$$

If for a moment we imagine that in system S8 $G_{\mathbf{x}}$, $\mu$, and $\mathbf{x}$ are known while the $(\eta\ldots)$'s are unknown, from an algebraic point of view, there exist infinitely many combinations of $(\eta\ldots)$'s solving the system, because there are more unknowns than equations. From the point of view of genetics, however, useful solutions are only those that maximize the proportion of variance in the genotypic values explained by the effects of single alleles or low-order allele combinations. This reasoning finds a mathematical reflection in the ordinary least squares (OLS) solution for the linear regression of $G_{\mathbf{x}}$ on single-locus allele contents $\mathbf{x}$ (system S8 taken without the grey-shaded higher order terms on the right). Denoting by $f_{\mathbf{x}}$ the frequency of genotype $\mathbf{x}$ among individuals in the population, the vector of OLS coefficients, $\eta^*$, is found as a solution to the optimization task S9:

$$\eta^* = \arg\min_{\eta} \sum_{\mathbf{x}} f_{\mathbf{x}} (G_{\mathbf{x}} - \mu - \sum_{l \le L} \sum_{m \le M_l} \eta_{lm} x_{lm})^2 \qquad \text{(S9)}$$

The elements $\eta^*_{lm}$ of any vector $\eta^*$ solving this optimization task are called **additive allele effects** and the sum $A_{\mathbf{x}} = \sum_{l \le L} \sum_{m \le M_l} \eta^*_{lm} x_{lm}$ is called **additive genetic value** of the genotype $\mathbf{x}$. As a detail, we clarify that for multiple QTLs ($L > 1$) the vector $\eta^*$ solving S9 is not uniquely defined because for each locus one of the allele contents can be expressed as a function of the others, i.e. the design matrix of the linear model is not of full rank. However the additive genetic values are invariant to the exact choice of $\eta^*$.

## ST2. Analysis of bias in broad-sense heritability estimates in the toy-model SIR simulations

Here, we report an in-depth analysis of several groups of SIR simulations corresponding to fixed Within/Between scenarios and average between-contact interval, i.e. group of box-plots above a single a value of $1/\kappa$ on Fig. 3E in the main text. The purpose of this report is to provide details in support of the statements made in the main text pointing out various sources of bias in estimating the broad-sense heritability, $H^2$. In order to gain statistical power, some of the reported statistics have been taken on the union of recovered populations ($\cup Z_{10k}$) of all simulated epidemics within a $1/\kappa$-group. The total number of individuals in such a union is indicated in parentheses and denoted by "#all" and the total number of donor-recipient pairs is denoted by "#DR". Cases in which, the total number (#all) is lower than $10^5$ indicate that some of the epidemic simulations did not result in an outbreak.

The analysis comprises the following items:

- Broad-sense heritability $H^2 = R^2_{adj}$ calculated on $\cup Z_{10k}$;

- Sample variance of observed $z$-values at moment of recovery (sampling) denoted $s^2(z)$ and calculated on $\cup Z_{10k}$;
- Sample variance of $z$-values at moment of transmission in donors from donor-recipient couples in $\cup Z_{10k}$, denoted $s^2(z_d)$;
- Sample variance of $z$-values at moment of getting infected in recipients from sampled donor-recipient couples in $\cup Z_{10k}$, denoted $s^2(z_r)$;
- Decomposition of the variance in $z$-values at the moment before recovery/sampling calculated on $\cup Z_{10k}$;
- Donor-recipient regression slope at moment of transmission $b[0]=s(z_d,z_r)/s^2(z_d)$ and donor-recipient covariance at moment of transmission, $s(z_d,z_r)$, from transmission couples in $\cup Z_{10k}$;
- ANOVA heritability estimate, $r_A[\mathrm{id}]$, based on individuals grouped by carried strain at moment of recovery in $\cup Z_{10k}$;
- Average phylogenetic Brownian Motion (PMM) heritability, $\overline{H^2_{BMe}}$, and estimate of the environmental variance $\overline{\sigma^2_e}$, from ML PMM fits on all $Z_{10k}$ populations in a $1/\kappa$-group;
- Average phylogenetic Ornstein-Uhlenbeck (POUMM) heritability, $\overline{H^2_{OUe}}$, and estimate of the environmental variance $\overline{\sigma^2_e}$, from ML PMM fits on all $Z_{10k}$ populations in a $1/\kappa$-group;

In addition to the above statistics, we have generated a graphical report for each $1/\kappa$-group under each Within/Between scenario including the following items:

- A scatter plot of donor and recipient $z$-values at moment of transmission from donor-recipient couples in $\cup Z_{10k}$. The points on that plot represent a random sample of 10% of the transmission couples in $\cup Z_{10k}$ and are colored corresponding to the transmitted strain as follows: red – 1:11, green - 2:12, blue - 3:21, cyan - 4:22, violet – 5:31, brown – 6:32. Black dash-dotted line represents the least-squares linear regression of recipient- on donor value. Red dash-dotted line represents the linear regression of recipient- on donor value if the true broad-sense heritability, $H^2$, replaces the least-squares regression slope, $b[0]$. A mismatch between these two lines indicates a bias in $b[0]$ as an estimator of $H^2$. With black circles are shown the means of recipient $z$-values at moment of getting infected corresponding to donor-values, which group together upon the transformation "$\mathrm{round}(z_d/4,1)*4$" in R. When these circles are not aligned with the least-squares regression line, we say that there is a non-linear relationship between donor-values and expected recipient values.

- Density plots of the whole population $\cup Z_{10k}$ - black line, the donor values - black dashed line, the recipient values – black dotted line, and the corresponding normal approximations (grey lines with the same line-patterns).

**Within: neutral / Between: neutral**

In the neutral/neutral case (Fig. 3E) the true $H^2$ (wide white boxes) has values between 0.3 and 0.5 consolidating around 0.45 as $1/\kappa$ approaches 12. This convergence corresponds to an equilibrium state of the infected population, in which all pathogen genotypes are encountered at equal frequencies and are uncorrelated with the immune system types and the special environmental effects.

With the exception of $1/\kappa \in \{2,4\}$, $H^2$ is very well approximated by $b[0]$. For $1/\kappa \in \{2,4\}$, we observe some bias in these two estimators due to differing strain distributions in $Z_{10k}$ and the sampled transmission couples (strain frequencies written on Fig. S1A).

For each $1/\kappa$, $b[\tau]$, are negatively biased with values below 0.3 and going further down as $1/\kappa$ increases. This increasingly negative bias is due to the increasing expected evolutionary distances $\tau$ (see also Fig. S2A).

The ANOVA-id estimator is positively biased due to overestimating the genotypic variance $\hat{\sigma}^2(G)$. This overestimation is due to the violation of normality assumption of the small set of six genotypic values. Thus, the positive bias is omnipresent in all cases and for all values of $1/\kappa$. In real situations with numerous genotypes and approximately normally distributed genotypic values, this bias should not exist (validated in additional simulations at the end of this section).

The PMM estimator, $H^2_{BMe}$, also has an increasingly negative bias. By comparing the values of $\sigma^2(z)$ for $1/\kappa = 2$ and $1/\kappa = 10$ on Fig. S1B and the PMM estimates of the environmental variance, $\sigma^2_e$, we noticed an increasingly positive bias of $\sigma^2_e$ caused by a worsening fit of the BM model on longer transmission trees with very slow growth of phenotypic variance. This bias inflicts a strong negative bias of $H^2_{BMe}$, as an estimator of the broad-sense heritability $H^2$.

The green boxplots of $H^2_{OUe}$ center around the true values of $H^2$ with apparently widening 95% whiskers as the average time between contacts, $1/\kappa$, approaches 12. This can be explained by loss of phylogenetic signal, as the branches in the transmission tree get longer and accumulate more mutations due to less frequent transmission events. We note, however, that such an effect should be less pronounced in cases of within-host selection. At an average around -0.01, the deviation of $H^2_{OUe}$ from $H^2$ for all neutral/neutral simulations was statistically insignificant ($p=0.13$), contrasting with significant ($p<10^{-22}$) negative biases for $\beta[\tau]$, and $H^2_{BMe}$ (Table S2).

**Within: select / Between: neutral**

In the select/neutral case (Fig. 3E) $H^2$ varies between 0.2 and 0.3 with a tendency to decrease as $1/\kappa$ is increasing. This is due to fixation of the fittest strains as a result of longer within-host selection (compare frequencies of strain 3:21 on Fig. S1B).

$b[0]$ shows a tendency to have a positive bias, more pronounced for bigger values of $1/\kappa$, which can be explained by the non-linear relationship between donor-values and expected recipient value (Fig. S1B).

Similarly to the neutral/neutral case, the estimators $b[\tau]$, have significant negative bias increasing with $1/\kappa$, although this bias is compensated to some extent by the positive bias in the corresponding estimators at moment of transmission.

Similarly to neutral/neutral case, the PMM estimator, $H^2_{BMe}$, has an increasing negative bias due to the worsening fit of the BM process.

A small positive bias (mean 0.01 at significant p-value of 3.3E-03, Table S2) is noticeable for the phylogeny-based estimator $H^2_{OUe}$. This bias is due to a departure from normality of the distribution of trait values in the population (see density plots on Fig. S1B).

**Within: neutral / Between: select**

The neutral/select case (Fig. 3E) looks similar to the neutral/neutral case except for the noticeable negative bias of $H^2_{OUe}$ at $1/\kappa \geq 8$. The negative bias of the POUMM estimator for longer waiting times between transmission events is due to the loss of signal in phylogenies with very long branches. No epidemic outbreak could be simulated at $1/\kappa = 12$, due to very low rate of risky contacts.

**Within: select / Between: select**

In the select/select case (Fig. 3E), $H^2$ stabilizes around 0.27 for $1/\kappa = 8$.

The regression slope at transmission, $b[0]$, has a significant positive bias (means 0.04 and 0.1, p<1.0E-12, Table S2). This bias results from two sources: smaller variance of donor-values due to selection for transmission (values of $\sigma^2(z_d)$ on Fig. S1D); non-linear dependence of recipient expected phenotype on donor phenotype (black circles on donor-recipient scatter plot on Fig. S1D).

Although less pronounced than the neutral/neutral and neutral/select cases, there is a significant negative bias of $b[\tau]$ due to the accumulating within-host evolution between moments of transmission and recovery. It appears, though, that this bias is weaker compared to the within-host neutral cases. The reason for that is two-fold: First, this bias is slightly compensated by the positive bias in the corresponding estimators at moment of transmission. Second, the within-host selection for higher $z$-values tends to slow-down the strain-substitution rate after a few beneficial mutations.

Similarly to the neutral/neutral case, the PMM estimator, $H^2_{BMe}$, has an increasing negative bias due to the worsening fit of the BM process (see Fig. 3D).

At an average of 0.02 (p=6.2E-05) (Table S2), the deviation of $H^2_{OUe}$ from $H^2$ for all select / select simulations was significant, yet very small compared to the bias of $b[\tau]$ and $H^2_{BMe}$.

### Clarifying the observed positive bias in ANOVA-CPP ($r_A$)

Here we demonstrate a positive bias in $r_A$ with respect to $R^2_{adj}$ for small number of groups (genotypes) $K$. We show that this bias vanishes for bigger values of $K$, i.e. $K>24$, given that the genotypic values are sampled from a normal distribution. For each $K \in \{3,6,12,24,48\}$ we simulate 100 datasets with $K$ genotypes and varying number of carriers for each genotype. We draw genotypic values from a normal distribution and add random (white) noise to them to construct the phenotype. After estimating $R^2$, $R^2_{adj}$ and $r_A$ for each dataset, we report the average values for each $K$.

```
# grand mean and variance of group effects
mu <- 3.5
sigma2a <- 1.44

# within-class variance
sigma2e <- 0.36

#number of simulated data-sets with K groups and ni individuals per group
nIter <- 100

# make results reproducible
set.seed(20)

test <- list()
# number of classes/groups
for(K in c(3, 6, 12, 24, 48)) {
  test[[as.character(K)]] <- t(sapply(1:nIter, function(iter) {
    # sample group means at each iteration
    ai <- rnorm(K, mean=mu, sd=sqrt(sigma2a))

    # numbers of sampled individuals per group
    ni <- sample(20:50, K, replace=TRUE)

    # generate data
    data <- data.table(g=do.call(c, lapply(1:K, function(k) rep(k, ni[k]))), key='g')
    data[, z:=rnorm(ni[g], mean=ai[g], sqrt(sigma2e)), by=g]

    data[, G:=mean(z), by=g]
    data[, e:=z-G]

    anova <- estimH2aov(NULL, data, NULL, by='g', report=TRUE)

    with(anova, data[, c(K=K, H2true=sigma2a/(sigma2a+sigma2e),
                         R2=var(G)/var(z), R2adj=1-(N-1)/(N-K)*var(z-G)/var(z),
                         rA=H2aov)])
  }))
}
```

The results show that $r_A$ dominates $R^2_{adj}$ on average, in particular for small values of $K$, i.e. $K \leq 12$. The reason is that ANOVA treats the observed $K$ group means in the data as a small iid sample from a normal distribution and makes an estimate of its variance, $\hat{\sigma}^2(G)$, while $R^2_{adj}$ is free of normality assumption and uses the sample variance $s^2(G)$ with a correction for finite degrees of freedom (eq. S3).

```
t(sapply(names(test), function(K) {
  colMeans(test[[K]])
}))
```

```
##  K H2true       R2    R2adj       rA
##  3    0.8 0.553072 0.543951 0.614931
##  6    0.8 0.712762 0.705719 0.740026
## 12    0.8 0.775991 0.769897 0.785147
## 24    0.8 0.789886 0.783941 0.791199
## 48    0.8 0.796711 0.790880 0.794446
```

## Supporting programs

This study relies on the accompanying R-package "patherit". The used version of these package, together with all program-code used for the toy-model simulations and the analysis of HIV-data, are provided in the attached file SP.zip. Inside it, a file named ReadMe.txt contains further instructions on how to run the code.

In addition the following third-party R-packages were used: ape v3.4 (38), data.table v1.9.6 (39), adaptMCMC v1.1 (34), Rmpfr v0.6-0 (40), and coda v0.18-1 (41). All programs have been run on R v3.2.4.

## Figures S1-S3

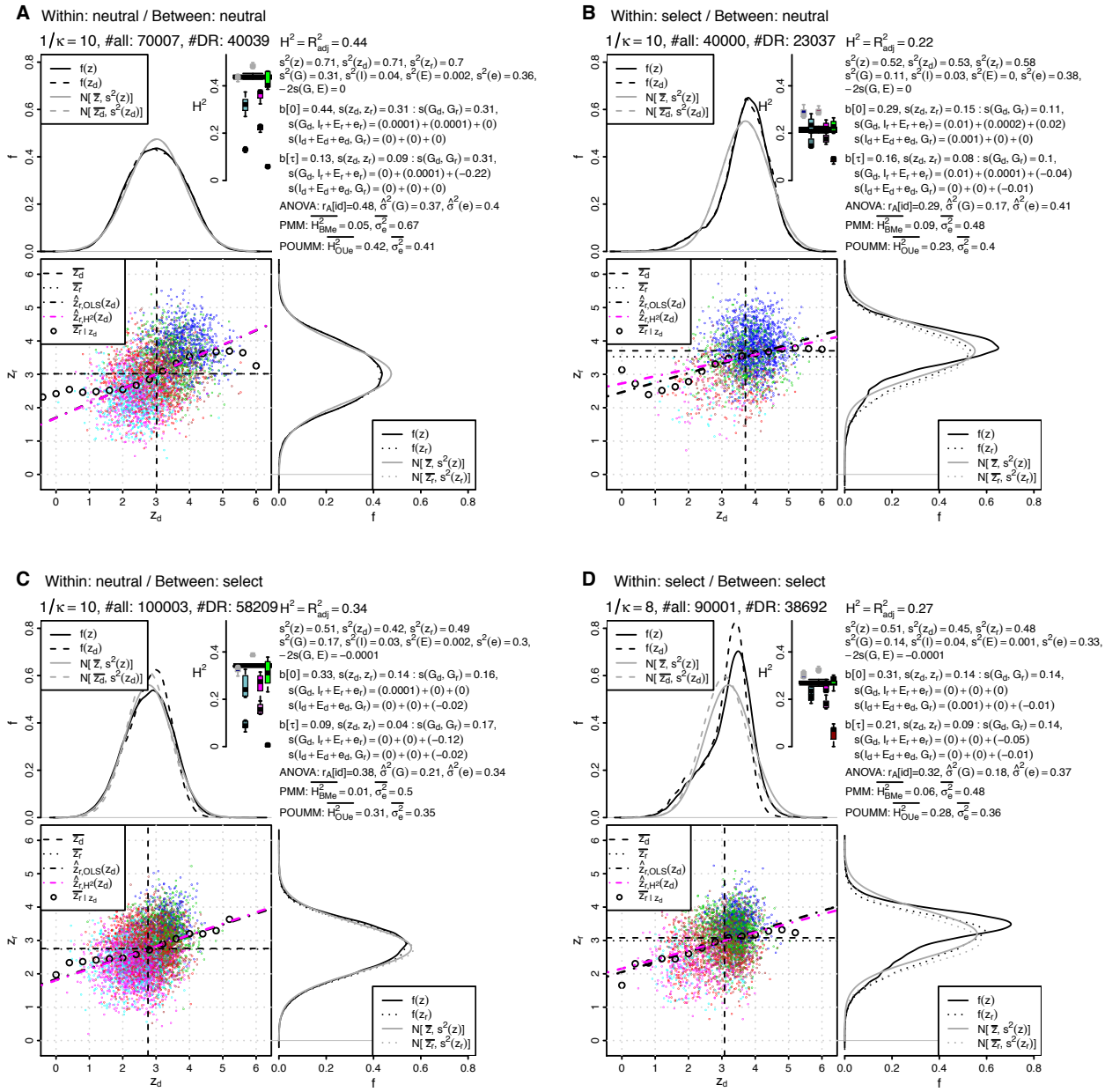


**Fig. S1 Analysis of bias of broad-sense heritability estimates in the toy-model simulations.** Notation: f: probability density function; $z_d$/$z_r$: donor/recipient values at the moments of transmission; N[mean, variance]: normal distribution; $\hat{z}_{r,OLS}(z_d)$: ordinary least

squares regression of recipient- on donor-values at the moments of transmission (using $b[0]$ as the slope of regression); $\hat{z}_{r,H^2}(z_d)$ : regression of recipient- on donor-values upon replacing the OLS-slope with the true heritability. A mismatch between black and red dash-dotted lines indicates bias in $b[0]$ w.r.t $H^2$ due to nonlinear relationship between $z_d$ and $z_r$. For a further description, see section "Analysis of bias in broad-sense heritability estimates" in Supporting Information.

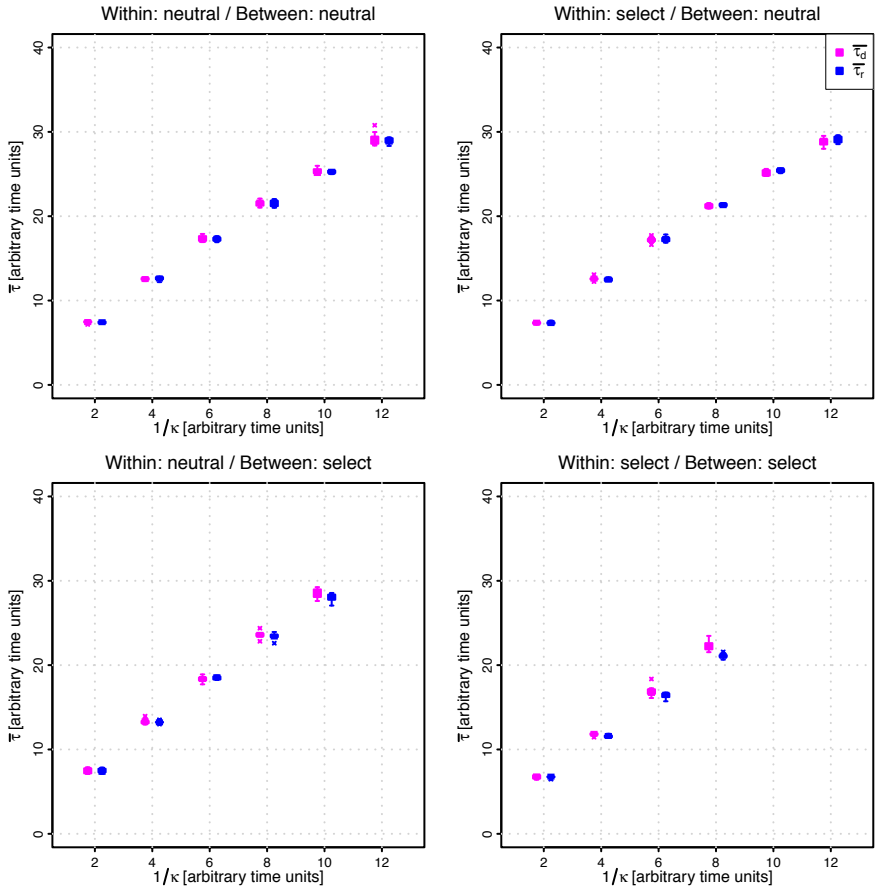


**Fig. S2 Mean donor- ($\tau_d$) and recipient ($\tau_r$) measurement delays in the toy model simulations.** Each box-whisker represents the average time-period between transmission and measurement events in diagnosed donor-recipient couples from up to ten simulated epidemic outbreaks under the same contact-rate $\kappa$ and between-/within-host scenario. Red (donor) and blue (recipient) measurement delays are very similar because both, donors and recipients get diagnosed at the same rate. Comparing this figure with Fig 3e in the main text confirms that longer measurement delays cause stronger negative bias in resemblance-based estimators of heritability.

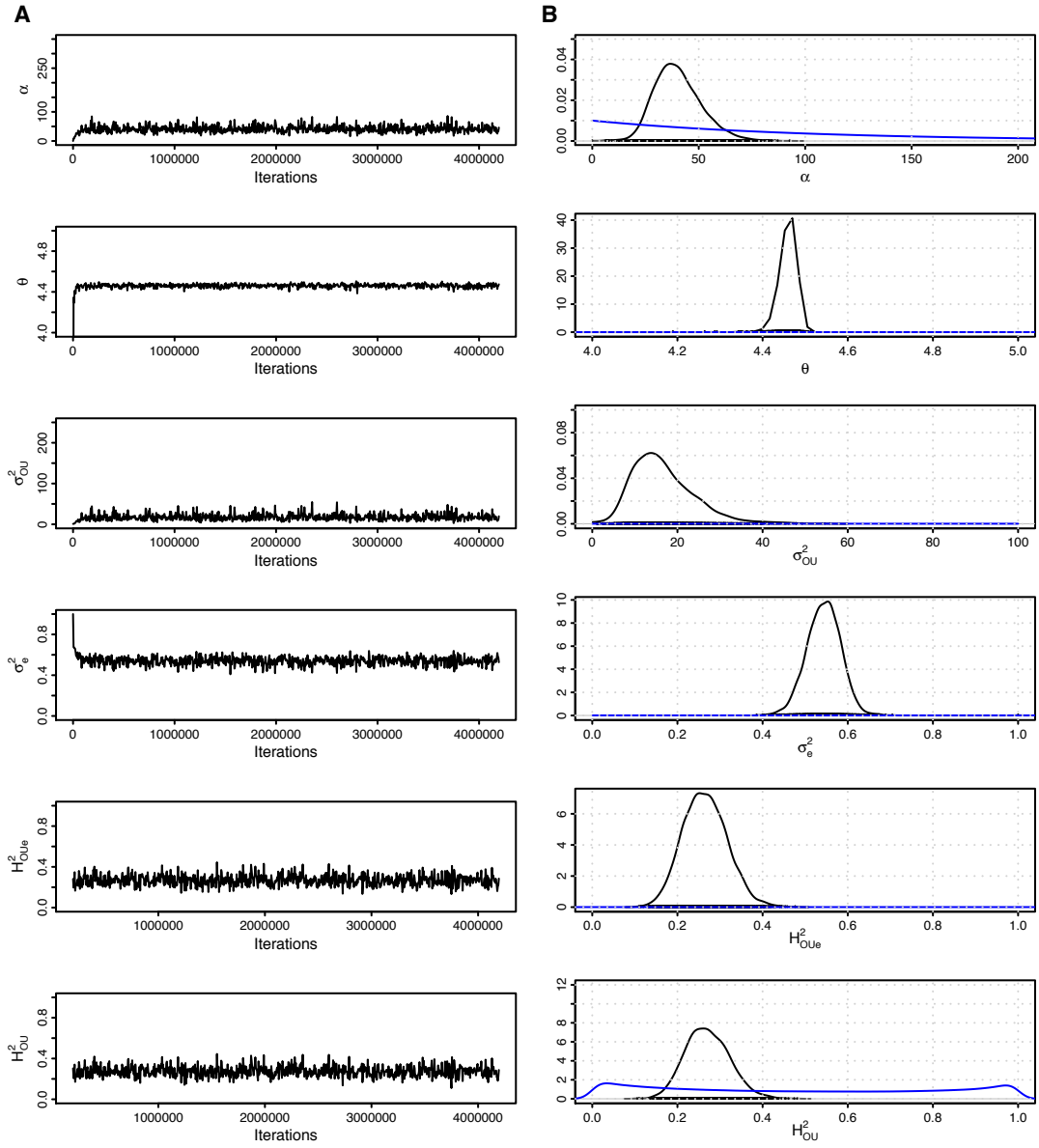


**Fig. S3 | Trace-plots and posterior densities from the POUMM MCMC-fits to HIV-1 from the UK cohort (8473 patients)**

The parameters $\alpha$, $\theta$, $\sigma^2_{OU}$ and $\sigma^2_e$ were sampled via MCMC (as described in Methods), while the parameters $H^2_{OU}$ and $H^2_{OUe}$ were calculated from the latter by eq. 9 (setting $t$ to the maximum root-tip distance) and eq. 5. **A**, trace-plots - the randomness and the lack of time-correlation in the traces show the correct mixing of the MCMC chain; **B**, Inferred posterior densities, black: MCMC sample density functions; blue: prior density functions. The clear distinction between prior and posterior densities proves the presence of informative signal in the data.

# Tables S1-S4

## Table S1 Within- and between-host SIR dynamics

| Scope | Parameter | neutral | select |
|---|---|---|---|
| Between-host | Natural birth rate | | $\alpha_{\text{nat}} = 117.6$ |
| | Natural per capita death rate | | $\delta_{\text{nat}} = 1/850$ |
| | Per capita recovery rate | | $\rho = 1/48$ |
| | Per capita contact rate | | $\kappa \in \{\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \frac{1}{10}, \frac{1}{12}\}$ |
| | Per capita risky contact rate (S: current proportion of susceptible in the pop.) | | $S \times \kappa$ |
| | Per risky contact transmission probability | $\gamma_{\text{neutral}} = .45$ | $\gamma(z) = \gamma_{\min} + \frac{(\gamma_{\max} - \gamma_{\min})(\gamma_{50})^{\gamma_k}}{10^{z\gamma_k} + (\gamma_{50})^{\gamma_k}}$, where $\gamma_{\min} = .3, \gamma_{\max} = .6, \gamma_{50} = 10^3, \gamma_k = 1.4$ |
| | Per capita death rate for infected individuals | $\delta_{\text{neutral}} = .01$ | $\delta(z) = \delta_{\text{nat}} + \frac{10^{zD_k} + (D_{50})^{D_k}}{D_{\min} 10^{zD_k} + D_{\max} (D_{50})^{D_k}}$, where $D_{\min} = 2, D_{\max} = 300, D_{50} = 10^3, D_k = 1.4$ |
| Within-host | Per locus pathogen mutation rate | $\nu_{\text{neutral}} = .01$ | $\nu(z) = \frac{\nu_{\max}(\nu_{50}) 10^{z\nu_k}}{10^{z\nu_k} + (\nu_{50})^{\nu_k}}$, where $\nu_{\max} = .2, \nu_{50} = 10^3, \nu_k = 1.4$ |
| | Rate of substitution of strain $\mathbf{x}_j$ for $\mathbf{x}_i$, where $\mathbf{x}_i \neq \mathbf{x}_j$ at a single locus, $l$, $M_l$ is the number of alleles at locus $l$, and the corresponding values are $z_i$ and $z_j$ | $\xi_l = \frac{\nu_{\text{neutral}}}{M_l - 1}$ | $\xi_{l,i \leftarrow j}(z_i, z_j) = \begin{cases} \frac{\nu(z_i)}{M_l - 1} & \text{if } \nu(z_i) < \nu(z_j) \\ 0 & \text{, otherwise} \end{cases}$ |

For description of each parameter, see Methods.

**Table S2 Mean difference $\overline{\widehat{H^2} - H^2}$ from the SIR simulations grouped by scenario**

| Within / Between | $b[0]$ | $b[\tau \le D_1]$ | $b[\tau]$ | $r_A[\mathrm{id}]$ | $r_A[\mathrm{PP}:\tau \le D_1]$ | $r_A[PP:\tau]$ | $H^2_{BMe}$ | $H^2_{OUe}$ |
|---|---|---|---|---|---|---|---|---|
| neutral/neutral | -0.01** | -0.07*** | -0.25*** | 0.05*** | -0.05*** | -0.18*** | -0.28*** | -0.01 |
| select/neutral | 0.05*** | 0 | -0.07*** | 0.08*** | 0 | -0.06*** | -0.12*** | 0.01** |
| neutral/select | -0.02*** | -0.05*** | -0.2*** | 0.05*** | -0.06*** | -0.15*** | -0.24*** | -0.02*** |
| select/select | 0.04*** | -0.01 | -0.06*** | 0.06*** | -0.03** | -0.08*** | -0.16*** | 0.03*** |

Statistical significance was estimated by Student's t-tests, p-values denoted by an asterisk as follows: * $p<0.05$; ** $p<0.01$; *** $p<10^{-3}$. Values in grey indicate that these estimates are typically unavailable in practice.

**Table S3 Sources of bias in estimators of $H^2$**

| Source of Bias \ Estimator of $H^2$ | $b[0]$ | $b[\tau \le D_1]$ | $b[\tau]$ | $r_A[\mathrm{id}]$ | $r_A[\mathrm{PP}{:}\tau \le D_1]$ | $r_A[\mathrm{PP}{:}\tau]$ | $H^2_{BMe}$ | $H^2_{OUe}$ |
|---|---|---|---|---|---|---|---|---|
| Gradual within-host pathogen evolution | 0 | - | --- | 0 | - | -- | 0 | 0 |
| Finite range of trait value (violated BM assumption) | 0 | 0 | 0 | 0 | 0 | 0 | --- | 0 |
| Nonlinear dependence of expected recipient value on donor value | ++/-- | ++/-- | ++/-- | 0 | 0 | 0 | 0 | 0 |
| Non-normality of $z$ and $G$ | +/- | +/- | +/- | +/- | +/- | +/- | +/- | +/- |
| Loss of phylogenetic signal due to scarcity of the transmission tree | 0 | 0 | 0 | 0 | 0 | 0 | -/+ | -/+ |
| Presence of outliers in the data, such as close PPs with high phenotypic difference | +/- | +/- | +/- | -- | -- | -- | - | - |
| Non-homogeneity of the evolutionary process over the tree (i.e. demographic changes, migrations, etc)(5) | 0 | - | --- | 0 | - | --- | +/- | +/- |
| Random error in the inferred transmission tree | 0 | 0 | 0 | 0 | - | - | - | - |
| Bias in transmission tree due to QTLs under selection in the region used for phylogenetic inference | 0 | 0 | 0 | 0 | ? | ? | ? | ? |
| Partial quasi-species transmission | -- | -- | -- | 0 | -- | -- | -- | -- |

The direction of the bias is indicated by a "+" or a "-", separated by a "/" when both directions are possible. The number of signs indicates the relative intensity of the bias that was observed in the simulations or in the analysis of the HIV-1 data. A zero indicates no bias observed. A "?" indicates unknown (and probably context-specific) direction. Lines separate sources that were identified in the SIR simulations (top) from sources identified in the analysis of the HIV-1 data (middle) and sources suggested by this or previous works that were not tested (bottom). Written in grey are estimators, which are not available in practice.

**Table S4 ANOVA-CPP and POUMM estimates of $\lg(\text{spVL})$-heritability in HIV-1 data from UK**

| Method | All tips in the phylogeny | | | Without outlier CPPs | | |
|---|---|---|---|---|---|---|
| | $N$ | $\hat{H}^2$ | 95% CI | $N$ | $\hat{H}^2$ | 95% CI |
| ANOVA-CPP ( $r_A[\text{PP}{:}\tau \le 10^{-4}]$ ) | 232[c] | 0.16[c] | [0.01, 0.30][c] | 222 | 0.31 | [0.19, 0.43] |
| ANOVA-PP ( $r_A[\text{PP}{:}\tau]$ )[a] | 3834[c] | 0.11[c] | [0.07, 0.14][c] | 3824 | 0.11 | [0.08, 0.14] |
| POUMM ( $H^2_{OUe}$ ) | 8483[c] | 0.20[c] | [0.13, 0.29][c] | 8473 | 0.25 | [0.16, 0.36] |
| POUMM, no CPP ( $H^2_{OUe}{:}\tau > 10^{-4}$ ) | 8251 | 0.22 | [0.13, 0.35] | | = | |
| PMM ( $H^2_{BMe}$ )[b] | 8483[c] | 0.06[c] | [0.02, 0.09][c] | 8473 | 0.06 | [0.02, 0.10] |
| PMM, ReML (15)[b] | 8483[c] | 0.06[c] | [0.03, 0.09][c] | | - | |

Also written are the results from a previous analysis on the same dataset (15). '=': the input data (and MCMC prior) is not altered by filtering out outlier CPPs; '-': the analysis not done by the cited study. Grey: $H^2$-estimates that are considered unreliable due to: a: negative bias due to measurement delays; b: negative bias due to BM violation; c: presence of outlier CPPs;