# A Practical Guide to Estimating the Heritability of Pathogen Traits - SUPPLEMENTARY INFORMATION

Venelin Mitov,[*,1,2] Tanja Stadler,[1,2]
[1]Department of Biosystems, Science and Engineering (D-BSSE)
[2]Swiss Federal Institute of Technology (ETH), Zürich, Switzerland
***Corresponding author:** E-mail: vmitov@gmail.com

In the sections below, we provide additional details and evidence in support of the statements made in the main text. In section "Approximations in equations 4 and 5", we clarify the approximations used in the main text. In secton "Why does PMM underestimate the correlation between PPs in the UK-data?", we investigate in some depth the observed bad fit of the PMM model to the UK data. In section "Analysis of bias in $H^2$-estimates in the toy-model simulations", we explain in detail the causes of bias in $H^2$-estimators, which were encountered in the toy model simulations. In section "Covariance between donor and recipient values in the toy model", we show analytically that in a neutral drift scenario, when all pathogen strains are encountered at equal frequencies, the covariance between donor and recipient values decays exponentially with the evolutionary time, $d_{ij}$ between the moments of trait measurement. In section "Choosing the threshold phylogenetic distance $d_{ij}{}'$ in ANOVA-CPP", we discuss the choice of threhold on $d_{ij}$ (e.g. $d_{ij}{}'=D_1$ and $d_{ij}{}'=10^{-4}$) when defining closest phylogenetic pairs. Supplementary tables and figures are provided at the end of this document.

SI

## Approximations in equations 4 and 5

To express the correlation in phylogenetic pairs under the PMM and the POUMM ML fits as functions of $d_{ij}$ (eq. 4 and 5), we applied three approximations:

- In eq. 2 and 3, we replaced $t$ by the mean root-tip distance in the tree, $\bar{t}$. This approximation was reasonable, because the mean root-tip distance did not vary substantially between different strata (fig. 3). The mean root-tip distance was 0.15 in the left-most decile going gradually down to 0.14 in the right-most decile. We also performed linear regression of the root-tip distance, $t$, on the phylogenetic distance, $d_{ij}$ in the 1917 PPs. This was significant but with negligible slope and coefficient of determination ($\hat{(t)}=0.15-0.13*d_{ij}$, $p<0.01$, $R^2_{adj}=0.01$), showing that PPs of all phylogenetic distances were spread nearly uniformly across the tree. Substituting $t$ with its linear regression on $d_{ij}$ instead of $\bar{t}$ did not result in any noticeable difference and is not reported.

- In eq. 2 and 3, we used the relationship between $t_{ij}$ and $d_{ij}$. This was the only way to incorporate $d_{ij}$ in eq. 2. In an ultrametric tree, $t_{ij}$ is an exact linear function of $d_{ij}$, namely, $t_{ij}=t-0.5d_{ij}$, where $t$ is the root-tip distance. In the non-ultrametric UK tree, the OLS regression of $t_{ij}$ on $d_{ij}$ was $\hat{t_{ij}}=0.15-0.63d_{ij}$, $p<10^{-16}$, $R^2_{adj}=0.24$.
- In eq. 3, we approximated $\exp(-8.35+36.47d_{ij})$ with 0, which was a valid approximation on the scale of the other terms in the equation and for the range of phylogenetic distances ($d_{ij}\in[0,0.14]$) in the UK tree.

The above approximates were validated visually by comparing the analytical curves corresponding to equations 4 and 5 with the corresponding brown and green points and error-bars on fig. 3.

**Why does PMM underestimate the correlation between PPs in the UK data?**

We have shown in the main text that, the phenotypic correlation between members of phylogenetic pairs depends on their phylogenetic distance, $d_{ij}$: members of pairs with small $d_{ij}$ tend to have higher phenotypic correlation compared to members of pairs with big $d_{ij}$ (fig. 3). For PMM, the only way to incorporate this information is indirect, namely, through the relationship between $d_{ij}$ and the root-mrca distance, $t_{ij}$. In the non-ultrametric UK tree, this relationship is rather weak: the slope of the OLS regression of $d_{ij}$ on $t_{ij}$ equals -0.37 and is significant ($p<0.01$) but the coefficient of determination of this regression, $R^2_{adj}$, is (only) 0.24 (fig. S2A). Thus, the principal source of information for fitting the PMM parameters, $\sigma^2$ and $\sigma^2_e$, is the assumed linear relationship between the observable correlation between pairs of tips and the two distances involved in eq. 2: the root-mrca distance $t_{ij}$ and the root-tip distance, $t$. Noticing that the correlation between the lg(spVL)-values in phylogenetic pairs is a covariance to variance ratio (eq. 2 and 3), we analyze how PMM fits to these two components in the UK data (fig. S2 B and C). The panels B and C on fig. S2 show that the covariance and the variance progress at different rates with $t_{ij}$ and $t$ respectively. PMM is not able to model this difference in the rates, because it uses a single parameter, $\sigma^2$, to model both of them. We notice that the ML estimate for $\sigma^2$ fits well to the linear increase in the variance (parallel brown and black lines on fig. S2C) but underestimates the increase in the covariance (non-parallel brown and black lines on fig. S2B). This indicates that a linear model of the covariance as a function of $t_{ij}$ is rather inappropriate and no particular value for $\sigma^2$ could result in a better fit (higher likelihood). As a result, the penalty on the PMM likelihood is minimized when the parameter $\sigma^2$ is fit to the increase in the variance, neglecting the covariance. Finally, this leads to the observed underestimate of the correlation in the closest phylogenetic pairs.

## Analysis of bias in $H^2$-estimates in the toy-model simulations

In order to understand the origin of the bias in the different toy-model scenarios, we used variance decomposition into the heritable component, $\sigma_G^2$ and the non-heritable component $\sigma_e^2$. Most of the biases observed on fig. 4 could be explained by a bias in one or both of these two components. The main source of these biases was the within-host evolution causing a decrease in the measured covariance between donor-recipient partners or phylogenetic pairs. Also, we identified various sampling biases introduced by within-/between-host selection and filtering of the data. We clarify these sources of bias in the following subsections.

### Neutral evolution of the trait within hosts

This phenomenon consists in a random change of the trait value caused by pathogen mutation. As a result, the phenotypic correlation between donors and recipients tends to decrease. We show later that, in a neutral scenario, this correlation decay is expected to be exponential in the phylogenetic distance, $d_{ij}$. As a result, all $H^2$-estimators neglecting or improperly modeling this decay are negatively biased. The most affected estimators are $b_{d_{ij}}$, $r_{A,d_{ij}}$, $H^2_{BM}(\bar{t})$ and $H^2_{BMe}$ (fig. 4); see also the decreasing sample donor-recipient covariance $s(z_{don},z_{rcp})$ on fig. S4.

### Directional selection within a host

This phenomenon consists in mutant strains contributing to a higher trait value, e.g. strains with higher reproductive capacity in the case of viral load, getting selected within each host. As a result a population of newly infected hosts tends to have higher genotypic variance than a population of hosts which have undergone within-host evolution (notice $s^2(G_{rcp,0}) > s^2(G)$ on fig. S4B). This explains the positive bias of $b_0$ with respect to $H^2$ on fig. 4B in the main text. Another possible effect of within-host selection is a convergent evolution in donors and recipients towards strains, which have higher fitness on average in the population. Intuitively, this could lead to a slight increase in phenotypic covariance and, therefore, a positive bias in $b$. Such a bias was not obvious in the toy-model simulations ($s(z_{don,0},z_{rcp,0}) > s(z_{don},z_{rcp})$ in all simulations, fig. S4B), because the convergent evolution was leading to a decreasing overall genetic and phenotypic variance in the population (see decreasing $s^2(G)$ and $s^2(z)$ with $d_{ij}$ on fig. S4B and S5B).

### Stabilizing selection between hosts

In case the trait is positively correlated with pathogen load, virulence and per contact transmission rate, hosts with very high pathogen load tend to be more infectious but stay infectious for a short period of time due to earlier diagnosis or death; hosts with very low pathogen load are infectious for a longer time but transmit very rarely. Thus, hosts with intermediate values of pathogen load have the highest

transmission potential on average (Fraser et al. 2007). This leads to a sampling bias in donor-recipient estimators - the donors have a narrower distribution than the overall population ($s^2(z_{don}) < s^2(z)$ on Fig. S5C). Intuitively this should lead to a positive bias in $b_0$ with respect to $H^2$ (because the denominator ($s^2(z_{don})$ is smaller). However, this was not confirmed by the toy-model simulations because the genotypic variance in the donor-recipient values at the moment of transmission was also smaller than that at the population level ($s(z_{don,0}, z_{rcp,0}) \approx s^2(G_0) < s^2(G)$ on fig. S4C).

### Combined within- and between-host selection

This results in a combination of the sampling biases due to each of the two selection phenomena (previous subsections).

### Non-stationary trait distribution during the epidemic

The density of the trait values evolves during the epidemic due to continuous change in the frequencies of the different pathogen strains, introduction of new strains through de-novo mutation, change in the frequencies of infected host types and a number of demographic factors such as migration, prevention, diagnosis and treatment. Thus, the broad-sense heritability, $H^2$, is a dynamic property of the population which changes through time. The direct estimator $R^2_{adj}$ obtained over a grouping by identical strain in patients sampled at different times has the meaning of a summary statistic averaging over the time of the epidemic. Plotting the phylogenetic estimators $H^2_{BM}(\sigma, \sigma_e, t)$ and $H^2_{OU}(\alpha, \sigma, \sigma_e, t)$ over time can help understanding the above dynamics. This, however, depends strongly on the goodness of fit of the phylogenetic model (e.g. BM or OU) to the data.

### Violation of phylogenetic model assumptions

The phylogenetic estimates of heritability are valid only if the model assumptions are at least partially met. For example, in this article, we have shown how an inaccurate assumption about the form of the correlation between two tips in the PMM model can lead to a significant negative bias in phylogenetic heritability.

### Clarifying the observed positive bias in $r_A$[id]

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 $K$.

```r
library(data.table)
library(patherit)
# grand mean and variance of group effects
mu <- 3.5
sigma2a <- .2
# 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, 96)) {

  test[[as.character(K)]] <- t(sapply(1:nIter, function(iter) {
    # sample group means at each iteration from a normal distribution
    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], sd=sqrt(sigma2e)), by=g]
    data[, G:=mean(z), by=g]
    data[, e:=z-G]
    rAValues <- rA(epidemic=NULL, data=data, GEValues=NULL, by='g', report=TRUE)
    with(rAValues, 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)])
  }))
}

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

```
##      K    H2true        R2     R2adj        rA
## 3    3 0.3571429 0.2304160 0.2147992 0.2768490
## 6    6 0.3571429 0.2987824 0.2816688 0.3181378
## 12  12 0.3571429 0.3475622 0.3298251 0.3494137
## 24  24 0.3571429 0.3622040 0.3441696 0.3539454
## 48  48 0.3571429 0.3655049 0.3472988 0.3522200
## 96  96 0.3571429 0.3720558 0.3537375 0.3562131
```

The results show that $r_A$ dominates $R^2_{adj}$ on average, in particular for small values of $K$, i.e. $K \leqslant 12$. For bigger $K$, the two estimators are asymptotically equal.

Clarifying the observed difference between $H^2_{BM}(\bar{t})$ and $H^2_{BMe}$ in toy-model simulations

As another detail, we notice that the expected correlation under the PMM ML fit, $r_{BM}$, was significantly positively biased with respect to the correlation, $r_A$, measured in PPs (brown line versus brown dots on fig. S7). Investigating these cases, we found that these positive biases were due to the use of the mean root-tip distance $\bar{t}$ in the formulation of $r_{BM}$ (eq. 2), the bias being less pronounced if using the median or a higher quantile of the root-tip distance. Since, at $d_{ij}=0$, $r_{BM}$ is equal to the phylogenetic heritability, $H^2_{BM}(\bar{t})$, in these cases, we observe a value of $H^2_{BM}(\bar{t})$ closer to the true heritability value, $R^2_{adj}$ (black horizontal line on fig. S7). This could lead to a wrong conclusion that $H^2_{BM}(\bar{t})$ is less biased than $H^2_{BMe}$. In fact though, this is merely the effect of cancelling out two biases with opposite directions. Compared to the PMM, the POUMM produced a better fit to the decaying correlation in all simulations (green dots and error-bars on fig. S7).

**Covariance between donor and recipient values in the toy model**

One of the main results of this article is the observation that the accuracy of a heritability estimator depends strongly on how it accounts for the within-host evolution of the pathogen taking place between transmission events and measurement. This becomes obvious from the fact that both, real data and the toy model simulations, showed a pattern of decaying correlation between phylogenetic pairs as a function of their phylogenetic distance, $d_{ij}$ (fig. 3 and fig. S7). Is this pattern of decaying correlation a general characteristic of epidemics? Here, we use a simplified version of the toy model allowing an analytical approach to this question.

We consider a version of the toy model, in which there is one SNP in the pathogen genotype with two possible alleles and there are two possible host-types. We denote the four genotype×host-type combinations as subscripts 00, 01, 10, 11, where the first index denotes the pathogen genotype and the second index denotes the host-type. The trait values are denoted as $z_{00}$, $z_{01}$, $z_{10}$ and $z_{11}$; the frequencies of the four genotype×host-type combinations in the populations are denoted as $f_{00}$, $f_{01}$, $f_{10}$ and $f_{11}$. We use the symbol $f_{\cdot 0}=f_{00}+f_{10}$ to denote the total frequency of host-type 0 and $f_{\cdot 1}=f_{01}+f_{11}$ to denote the total frequency of host-type 1. We assume that the evolution of the pathogen strain within a host follows random drift - at time $d_{ij}/2$ after infection, the strain infecting a host has been substituted by a mutant strain with probability $\nu$, regardless of the trait value before and after substition; the strain has remained unchanged with probability $1-\nu$.

This mechanism of within-host mutation is summarized on fig. S9A. For simplicity, we assume a generation-based dynamics, in which transmission to new susceptible hosts occurs at fixed moments in

time separated by a period $d_{ij}/2$. At every generation, each member of the infected population transmits his/her currently carried pathogen to a random susceptible individual and becomes uninfectious, (although, he/she remains infected with the pathogen). The recipient host transmits his infection at the next generation and becomes uninfectious on his turn. We assume an infinite susceptible pool with fixed frequencies of the two host-types. Given that there is no selection with respect to host-type, we can assume that the frequencies of the host-types in the infected population equals the host-type frequencies in the susceptible population. The frequencies of the two pathogen strains in the infected population can evolve as a result of within-host mutation. However, in the absence of within-host selection, the strain frequencies conditioned on host-type would equalize several generations after the onset of epidemic.

With the above simplified version of the toy model, it is possible to express the covariance between a donor and recipient trait value at time $d_{ij}/2$ after the transmission has taken place. We do this in two steps: first, we express the covariance in terms of the substitution probability $\nu$; then, we use a 2-nucleotide form of the Jukes-Cantor 69 substitution model to express $\nu$ in terms of evolutionary time. Denoting the donor value by $z_{don}$ and the recipient value by $z_{rcp}$, we start from a known property for the covariance:

$$Cov(z_{don}, z_{rcp}) = E[z_{don} z_{rcp}] - E[z_{don}]E[z_{rcp}]$$

In the case of neutral evolution and sufficiently large population size, we can assume that donors and recipients share the same frequencies of genotype $\times$ host-type combinations. Thus, we write:

$$E[z_{don}] = E[z_{rcp}] = f_{00} z_{00} + f_{01} z_{01} + f_{10} z_{10} + f_{11} z_{11}$$

To obtain the expectation of the product $z_{don} z_{rcp}$, it suffices to sum up all possible products of donor and recipient values weighted by their expected frequencies (fig. S9A):

$$
\begin{aligned}
E[z_{don}z_{rcp}] = f_{00}\Big( & (1-\nu)f_{\cdot 0}(1-\nu)z_{00}z_{00}+(1-\nu)f_{\cdot 0}\nu z_{00}z_{10}+(1-\nu)f_{\cdot 1}(1-\nu)z_{00}z_{01}+(1-\nu)f_{\cdot 1}\nu z_{00}z_{11}+ \\
& \nu f_{\cdot 0}(1-\nu)z_{10}z_{00}+\nu f_{\cdot 0}\nu z_{10}z_{10}+\nu f_{\cdot 1}(1-\nu)z_{10}z_{01}+\nu f_{\cdot 1}\nu z_{10}z_{11}\Big)+ \\
f_{01}\Big( & (1-\nu)f_{\cdot 0}(1-\nu)z_{01}z_{00}+(1-\nu)f_{\cdot 0}\nu z_{01}z_{10}+(1-\nu)f_{\cdot 1}(1-\nu)z_{01}z_{01}+(1-\nu)f_{\cdot 1}\nu z_{01}z_{11}+ \\
& \nu f_{\cdot 0}(1-\nu)z_{11}z_{00}+\nu f_{\cdot 0}\nu z_{11}z_{10}+\nu f_{\cdot 1}(1-\nu)z_{11}z_{01}+\nu f_{\cdot 1}\nu z_{11}z_{11}\Big)+ \\
f_{10}\Big( & (1-\nu)f_{\cdot 0}(1-\nu)z_{10}z_{10}+(1-\nu)f_{\cdot 0}\nu z_{10}z_{00}+(1-\nu)f_{\cdot 1}(1-\nu)z_{10}z_{11}+(1-\nu)f_{\cdot 1}\nu z_{10}z_{01}+ \\
& \nu f_{\cdot 0}(1-\nu)z_{00}z_{10}+\nu f_{\cdot 0}\nu z_{00}z_{00}+\nu f_{\cdot 1}(1-\nu)z_{00}z_{11}+\nu f_{\cdot 1}\nu z_{00}z_{01}\Big)+ \\
f_{11}\Big( & (1-\nu)f_{\cdot 0}(1-\nu)z_{11}z_{10}+(1-\nu)f_{\cdot 0}\nu z_{11}z_{00}+(1-\nu)f_{\cdot 1}(1-\nu)z_{11}z_{11}+(1-\nu)f_{\cdot 1}\nu z_{11}z_{01}+ \\
& \nu f_{\cdot 0}(1-\nu)z_{01}z_{10}+\nu f_{\cdot 0}\nu z_{01}z_{00}+\nu f_{\cdot 1}(1-\nu)z_{01}z_{11}+\nu f_{\cdot 1}\nu z_{01}z_{01}\Big)
\end{aligned}
$$

Taking the difference of $E[z_{don}z_{rcp}]-E[z_{don}]E[z_{rcp}]$ and grouping on the degrees of $\nu$, we obtain a polynomial of degree two of $\nu$:

$$Cov(z_{don},z_{rcp}) = A\nu^2+B\nu+C$$

The coefficients $A$, $B$ and $C$ are algebraic expressions of the frequencies and trait values:

$$
\begin{aligned}
A &= (f_{00}(z_{00}-z_{10})+f_{10}(z_{00}-z_{10})+(f_{01}+f_{11})(z_{01}-z_{11}))(f_{\cdot 0}(z_{00}-z_{10})+f_{\cdot 1}(z_{01}-z_{11})) \\
B &= 2f_{10}f_{\cdot 0}z_{00}z_{10}+f_{11}f_{\cdot 0}z_{01}z_{10}+f_{10}f_{\cdot 1}z_{01}z_{10}-2f_{10}f_{\cdot 0}z_{10}^2+f_{11}f_{\cdot 0}z_{00}z_{11}+ \\
&\quad f_{10}f_{\cdot 1}z_{00}z_{11}+2f_{11}f_{\cdot 1}z_{01}z_{11}-2f_{11}f_{\cdot 0}z_{10}z_{11}-2f_{10}f_{\cdot 1}z_{10}z_{11}-2f_{11}f_{\cdot 1}z_{11}^2+ \\
&\quad f_{01}(2f_{\cdot 1}z_{01}(-z_{01}+z_{11})+f_{\cdot 0}(-2z_{00}z_{01}+z_{01}z_{10}+z_{00}z_{11}))+ \\
&\quad f_{00}(-2f_{\cdot 0}z_{00}(z_{00}-z_{10})+f_{\cdot 1}(-2z_{00}z_{01}+z_{01}z_{10}+z_{00}z_{11})) \\
C &= f_{00}f_{\cdot 0}z_{00}^2+f_{01}f_{\cdot 0}z_{00}z_{01}+f_{00}f_{\cdot 1}z_{00}z_{01}+f_{01}f_{\cdot 1}z_{01}^2+f_{10}f_{\cdot 0}z_{10}^2+f_{11}f_{\cdot 0}z_{10}z_{11}+ \\
&\quad f_{10}f_{\cdot 1}z_{10}z_{11}+f_{11}f_{\cdot 1}z_{11}^2-(f_{00}z_{00}+f_{01}z_{01}+f_{10}z_{10}+f_{11}z_{11})^2
\end{aligned}
$$

In the case of neutral drift, $f_{00}=f_{10}$ and $f_{01}=f_{11}$. Substituting $1-f_{\cdot 0}$ for $f_{\cdot 0}$, the expression for the covariance simplifies to:

$$Cov(z_{don},z_{rcp}) = \frac{1}{4}(1-2\nu)^2(z_{01}-z_{11}+f_{\cdot 0}(z_{00}-z_{01}-z_{10}+z_{11}))^2$$

Assuming a two-nucleotide Jukes Cantor 69 model with mutation rate $\lambda$, the probability of mutation at a site in the genetic sequence is expressed as a function of evolutionary time $\nu(t)=0.5-0.5\exp(-\lambda t)$ (Yang, 2006). Thus, in the case of neutral drift, the covariance between the donor and the recipient

value at time $d_{ij}/2$ after the transmission can be expressed in terms of the total evolutionary time, $d_{ij}$, separating the two hosts:

$$Cov(z_{don}, z_{rcp}) = \exp(-\lambda d_{ij}) \frac{1}{4} (z_{01} - z_{11} + f_{\cdot 0}(z_{00} - z_{01} - z_{10} + z_{11}))^2$$

The above expression for the covariance between donor and recipient values represents an exponential decay function of $d_{ij}$ - it has a non-negative value at $d_{ij} = 0$ and converges exponentially towards 0 as $d_{ij} \to \infty$. It is interesting to ask wether the above pattern of exponentially decaying covariance is preserved in the case of multiple loci (many possible pathogen genotypes) as well as in cases of within- and between-host selection. An analytical treatment of this question is beyound the scope of this article. However, using simulations of the toy model, we have shown that the pattern of exponential decay seems to be preserved in the case of the neutral/neutral scenario, that is, when each pathogen genotype is encountered at equal frequency for each host-type (fig. S7). Biologically, this reflects a situation, where the donor and recipient host exhibit similar trait values shortly after transmission, but later on tend to have uncorrelated values as a result of random mutation in the two hosts. In infinite time after the transmission, the correlation between the two hosts' trait values should converge to 0. In the cases of within- or between-host selection, the covariance between the trait values of a donor and a recipient would be influenced by additional factors such as similar age, race or habitat. This can result in convergent evolution of the pathogens within the two hosts towards strains which are best adapted to the shared environmental conditions. In this case, the covariance would deviate substantially from an exponential decay function of $d_{ij}$ and is even not guaranteed to converge to 0. This reaffirms that any parametric model of the covariance (and therefore, correlation) between transmission couples needs to be validated against empirical estimates (see fig. 3 and Fig. S2).

**Choosing the threshold phylogenetic distance $d_{ij}{}'$ in ANOVA-CPP**

Choosing an appropriate value for the threshold $d_{ij}{}'$ is one of the tricky aspects of ANOVA-CPP. This choice is a trade-off between minimizing the negative bias due to within-host evolution ($d_{ij}{}'$ close to 0) and maximizing the precision in terms of narrow confidence interval. While it is impossible to measure the bias in the absence of knowledge about the true value, there are ways to measure the precision, e.g. by taking the length of the 95% confidence interval. Thus, one way to define an optimality criterion is "the minimum value of $d_{ij}{}'$, for which the 95% confidence interval is narrower than some predefined length". A practical way to do this is to consider different stratifications of the phylogenetic pairs as shown on fig. S1. In the toy-model simulations, we have chosen the first decile, i.e. $d_{ij}{}' = D_1$, because this threshold

was suitable for demonstrating the negative bias due to within-host evolution (i.e. a difference between $b_{D_1}$ and $b_{d_{ij}}$ and loss of precision). In the real HIV data, the choice $d_{ij}{}'=10^{-4}$ was based on empirical observations (see text and fig. 5).

## SUPPLEMENTARY TABLES

**Table S1. PMM and POUMM fit to lg(spVL) data from the UK HIV cohort.**

| $N$ | Model | AICc | Type | $g_0$ | $\alpha$ | $\theta$ | $\sigma$ | $\sigma_e$ | $H^2(\bar{t})$ | $H_e^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 8,483 | PMM | 21,487 | MLE | 4.49 | - | - | 0.65 | 0.84 | 0.08 | 0.06 |
| | | | Mean | 4.49 | - | - | 0.67 | 0.83 | 0.08 | 0.06 |
| | | | HPD | [4.31, 4.66] | - | - | [0.5, 0.84] | [0.82, 0.85] | [0.05, 0.12] | [0.02, 0.1] |
| | POUMM | 21,455 | MLE | 5.54 | 28.78 | 4.45 | 2.97 | 0.77 | 0.21 | 0.2 |
| | | | Mean | 5.44 | - | - | 3.11 | 0.77 | 0.21 | 0.21 |
| | | | HPD | [4.06, 7.25] | [16.64, 46.93] | [4.41, 4.49] | [1.95, 4.37] | [0.73, 0.8] | [0.14, 0.29] | [0.13, 0.29] |

**Table S2.** Within- and between-host dynamics of the toy epidemiological model.

| Scope | Parameter | neutral | select |
|---|---|---|---|
| Between-host | Natural birth rate | $\lambda = 117.6$ | |
| | Natural per capita death rate | $\mu = 1/850$ | |
| | 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 | $\delta_{\text{neutral}} = .01$ | $\delta(z) = \mu + \frac{10^{zD_k} + (D_{50})^{D_k}}{D_{\min}10^{zD_k} + D_{\max}(D_{50})^{D_k}}$, 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 $z_j$ | $\xi_l = \frac{\nu_{\text{neutral}}}{M_l - 1}$ | $\xi_{l,i \leftarrow j}(z_i, z_j) = \begin{cases} \frac{\nu(z_i)}{M_l - 1} \text{ if } \nu(z_i) < \nu(z_j) \\ 0 \quad \text{, otherwise} \end{cases}$ |

## SUPPLEMENTARY FIGURES

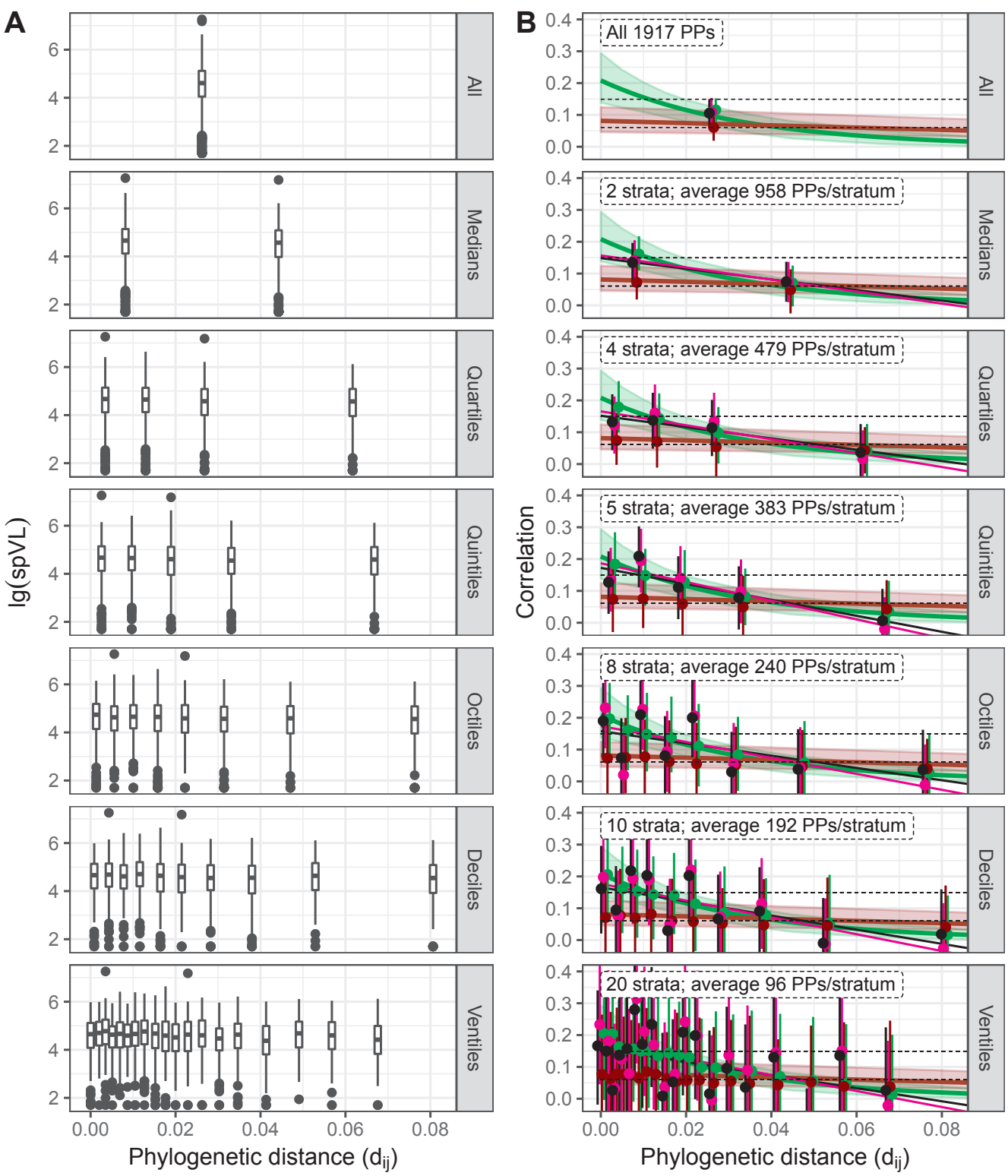


**FIG. S1. Different stratifications of the phylogenetic pairs in the UK tree.** A - box-plots of the trait values show nearly identical distributions (equal mean and interquartile range) in the different strata. B - correlation profiles in different stratifications. Black and magenta points with error-bars denote the estimated $r_A$ and $r_{Sp}$ in the real data. Dashed horizontal bars denote the 95% CI for $r_A$ evaluated on all phylogenetic pairs. A black and a magenta inclined line denote the least squares linear regression of $r_A$ and $r_{Sp}$ on the mean phylogenetic distance, $\bar{d_{ij}}$, in each decile. Brown and green points with error bars denote the estimated values of $r_A$ obtained after replacing the real trait values on the tree by values simulated under the maximum likelihood fit of the PMM and the POUMM methods respectively (mean and 95% CI estimated from 100 replications). A brown and a green line show the expected correlation between pairs of tips at distance $d_{ij}$, as modeled under the ML-fit of the PMM and the POUMM (eq. 2 and 3). A light-brown and a light-green region depict the 95% high posterior density (HPD) intervals inferred from Bayesian fit of the two models ("Materials and methods").

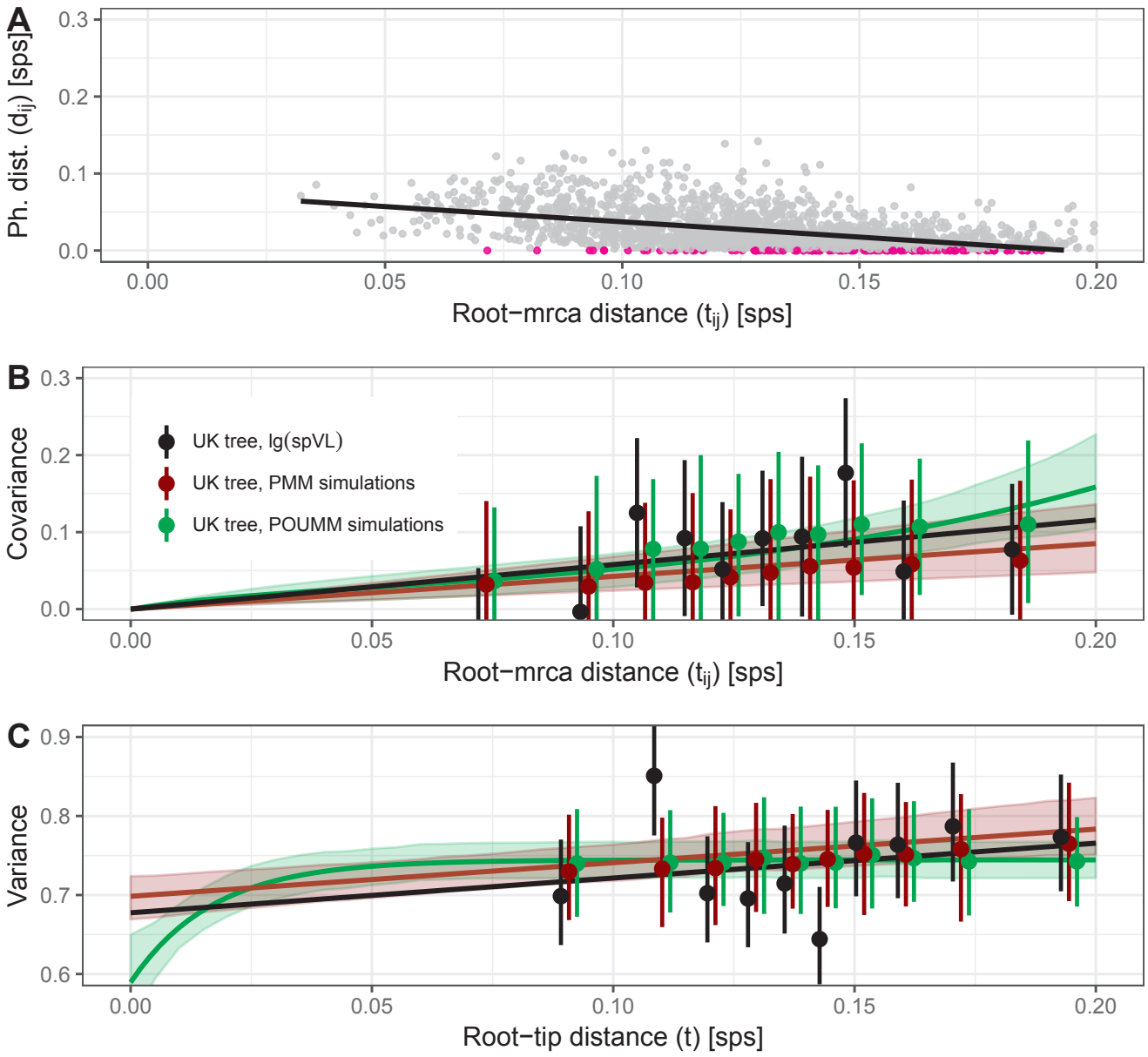


**FIG. S2. Bias in the PMM estimate for the correlation in phylogenetic pairs** A: A scatter plot and OLS regression of $d_{ij}$ on $t_{ij}$ (slope -0.34 ($p<0.01$), $R^2_{adj}=0.24$). Points in magenta denote CPPs ($d_{ij}<10^{-4}$); B: covariance modeled as a function of $t_{ij}$. Black points and error-bars denote the sample covariance and 95% CI upon a stratification in deciles of $t_{ij}$. Brown and green points and error-bars denote the mean and 95% CI upon replacing the lg(spVL)-values with values simulated under the ML fit of the PMM and the POUMM (100 replications). A black line going through the origin denotes the OLS regression with 0 intercept to the real data. A brown and a green line with brighter surrounding regions denote the covariance and its 95% HPDs under the PMM and the POUMM respectively. The latter have been obtained from the expressions for the nominator in eqs. 2 and 3 using the ML estimates and posterior samples for the model parameters. In the case of the POUMM, the phylogenetic distance $d_{ij}$ has been replaced by the linear regression of $d_{ij}$ on $t_{ij}$ from the phylogenetic pairs (panel A). The slope of the brown line equals the parameter $\sigma^2$ of the PMM. Notice the negative bias with respect to the OLS fit (black line). C: variance of the trait values at the tips of the UK tree modeled as a function of the root-tip distance, $t$. Black points and error-bars denote the sample variance and its 95% CI in the real data, upon a stratification in deciles. Brown and green points and error-bars denote the mean and 95% CI upon replacing the lg(spVL)-values with values simulated under the ML fit of the PMM and the POUMM (100 replications). A black line denotes the OLS regression of the variance in the real data on $t$. A brown and a green line with brighter surrounding regions denote the variance and its 95% HPDs under the PMM and the POUMM respectively. The latter have been obtained from the expressions for the denominator in eqs. 2 and 3 using the ML estimates and posterior samples. As in panel B, the slope of the brown line equals the parameter $\sigma^2$ of the PMM. The distances $t_{ij}$ and $d_{ij}$ are measured in substitutions per site (sps).

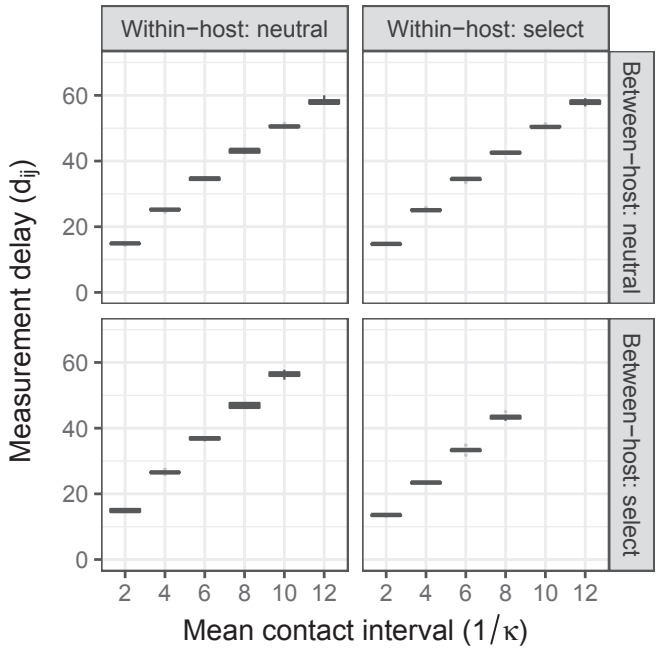


**FIG. S3. Mean phylogenetic distance $d_{ij}$ between PPs in the toy-model simulations**

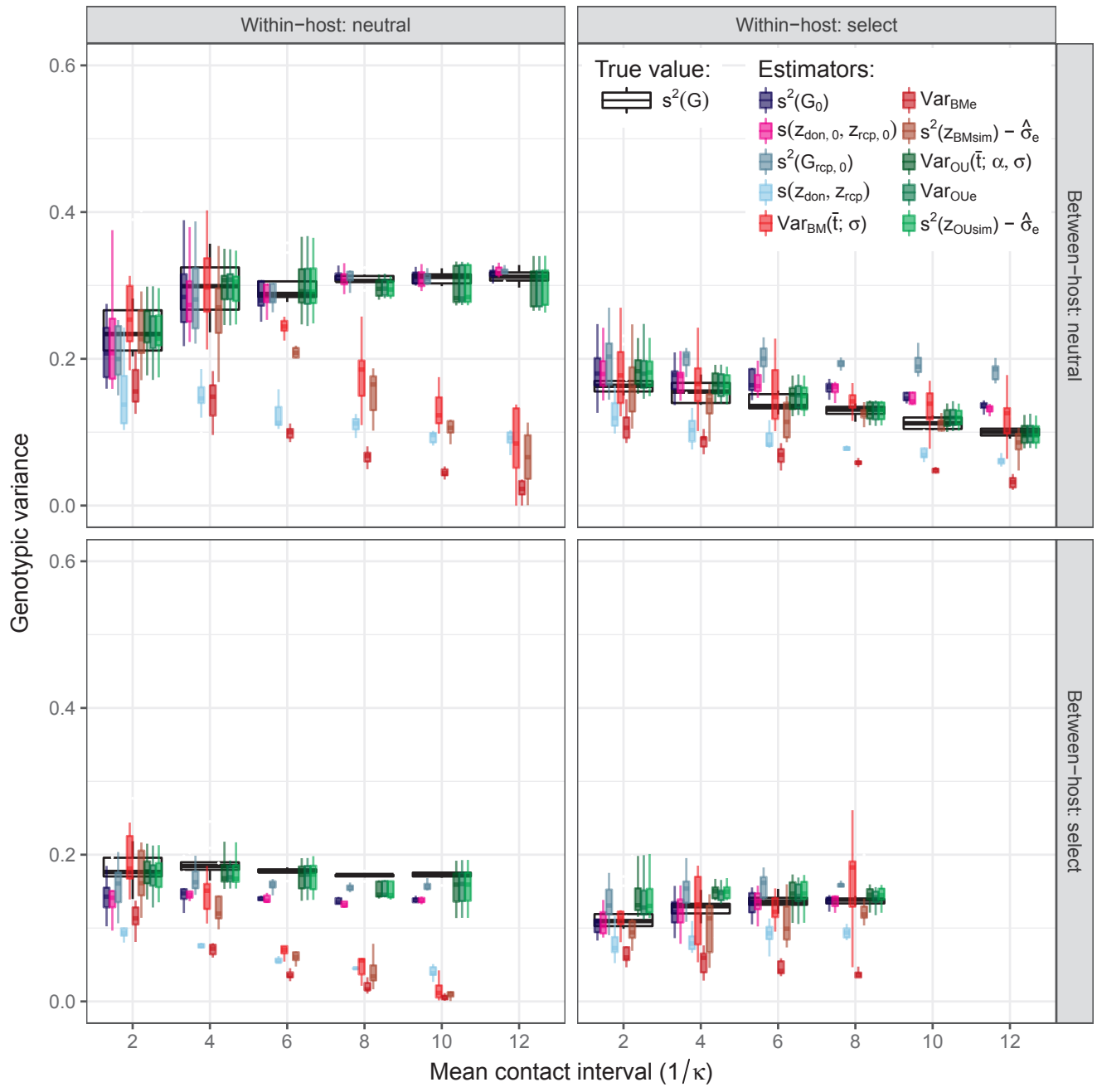


**FIG. S4. Estimating the genotypic variance in toy-model simulations.** $s^2(G)$: true genotypic variance calculated from grouping by identical genotype; $s^2(G_0)$: true genotypic variance calculated from grouping by identical genotype in the sample of known donor-recipients, taking their genotype and trait values at the moment of transmission; $s(z_{don,0},z_{rcp,0})$: empirical covariance between donors and recipients at the moment of transmission; $s^2(G_{rcp,0})$: true genotypic variance in recipients at the moment of getting infected; $s(z_{don},z_{rcp})$: donor-recipient covariance at moment of diagnosis (including measurement delay); $\mathrm{Var}_{BM}(\bar{t};\sigma)$: estimated PMM genotypic variance at $\bar{t}$ according to eq. 16; $\mathrm{Var}_{BMe}$: estimated PMM genotypic variance based on the difference $s^2(z)-\hat{\sigma}_e^2$ in the ML fit of the PMM; $s^2(z_{BMsim})-\hat{\sigma}_e^2$: estimated PMM genotypic variance based on the difference of the mean trait variance in 100 simulations of the ML PMM fit on the tree and the ML value of the parameter $\sigma_e^2$; $\mathrm{Var}_{OU}(\bar{t};\alpha,\sigma)$: estimated POUMM genotypic variance at $\bar{t}$ according to eq. 17; $\mathrm{Var}_{OUe}$: estimated POUMM genotypic variance based on the difference $s^2(z)-\hat{\sigma}_e^2$ in the ML fit of the POUMM; $s^2(z_{OUsim})-\hat{\sigma}_e^2$: estimated POUMM genotypic variance based on the difference of the mean trait variance in 100 simulations of the ML POUMM fit on the tree and the ML value of the parameter $\sigma_e^2$.

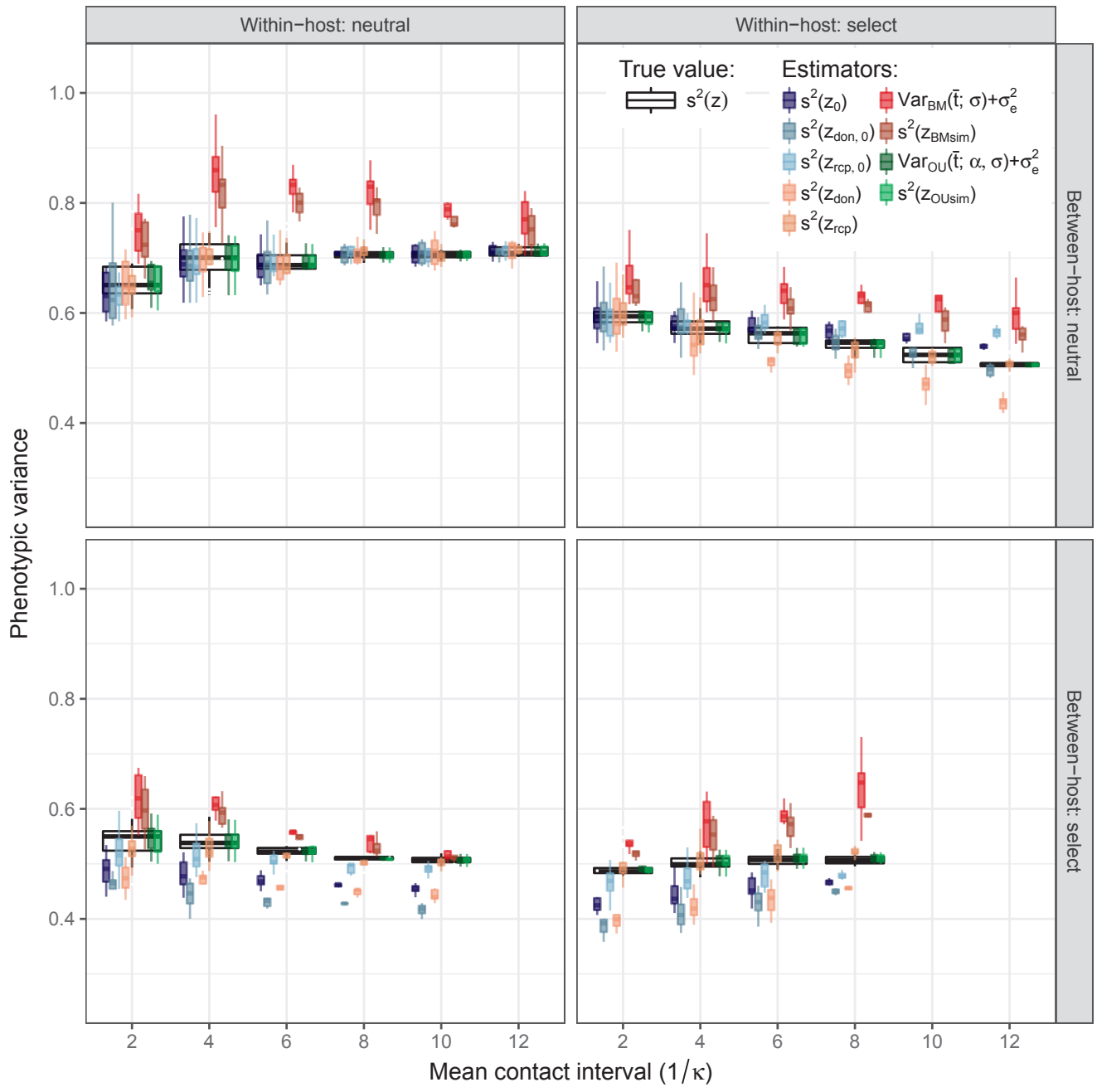


**FIG. S5. Phenotypic variance in the toy-model simulations.** $s^2(z)$: sample variance of the trait value in the entire sampled population; $s^2(z_0)$: sample variance in the sampled donor-recipient couples taking the trait values at moment of infection; $s^2(z_{don,0})$: sample variance in the donors from donor-recipient couples, taking the trait values at moment of infection; $s^2(z_{rcp,0})$: sample variance in the recipients from donor-recipient couples, taking the trait values at moment of infection; $s^2(z_{don})$: sample variance in the donors from donor-recipient couples, taking the trait values at moment of diagnosis; $s^2(z_{rcp})$: sample variance in the recipients from donor-recipient couples, taking the trait values at moment of diagnosis; $\mathrm{Var}_{BM}(\bar{t};\sigma,\sigma_e)=\sigma^2\bar{t}+\sigma_e^2$: expected phenotypic variance under the ML fit of the PMM at the mean root-tip distance $\bar{t}$; $s^2(z_{BMsim})$: mean sample trait variance from 100 simulations of the ML PMM fit on the tree; $\mathrm{Var}_{OU}(\bar{t};\alpha,\sigma,\sigma_e)=\frac{\sigma^2}{2\alpha}(1-\exp(-2\alpha t)+\sigma_e^2$: expected phenotypic variance under the ML fit of the POUMM at the mean root-tip distance $\bar{t}$; $s^2(z_{OUsim})$: mean sample trait variance from 100 simulations of the ML POUMM fit on the tree.

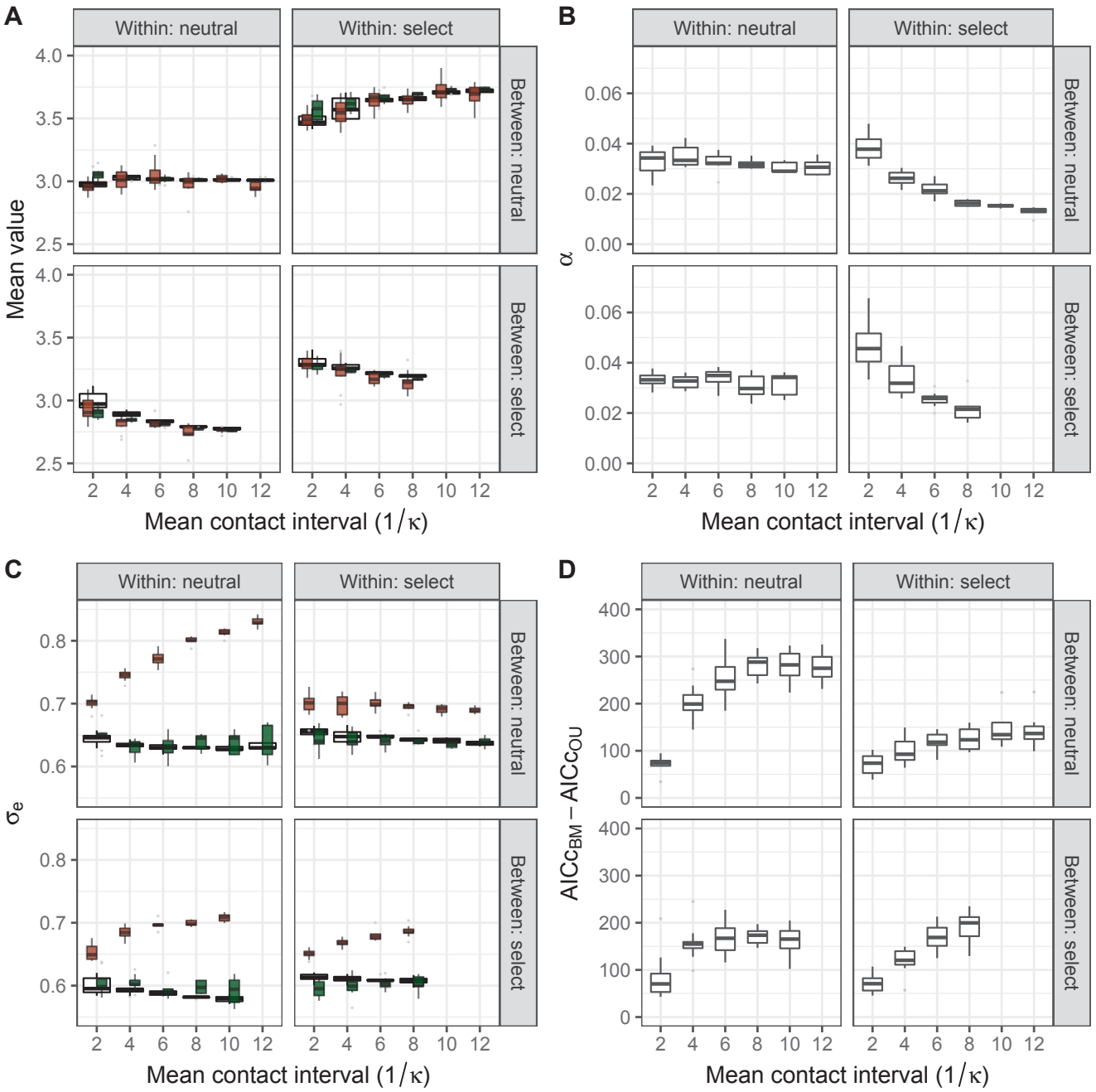


**FIG. S6. Details of the PMM and POUMM ML fits to the toy-model simulations.** A: comparison beween the true population mean (wide boxes in the background) to the mean-value expected under the PMM method (brown) and the long-term mean value, $\theta$ expected under the POUMM method (green); B: Estimates for the parameter $\alpha$ in the toy-model simulations; C: estimates for the parameter $\sigma_e$ of the PMM (brown) and the POUMM (green) compared to the non-heritable standard deviation estimated from grouping by identical genotype; D: comparison of the corrected Akaike information criterion for the PMM and the POUMM fits - positive values indicate lower (better) AICc for the POUMM method.

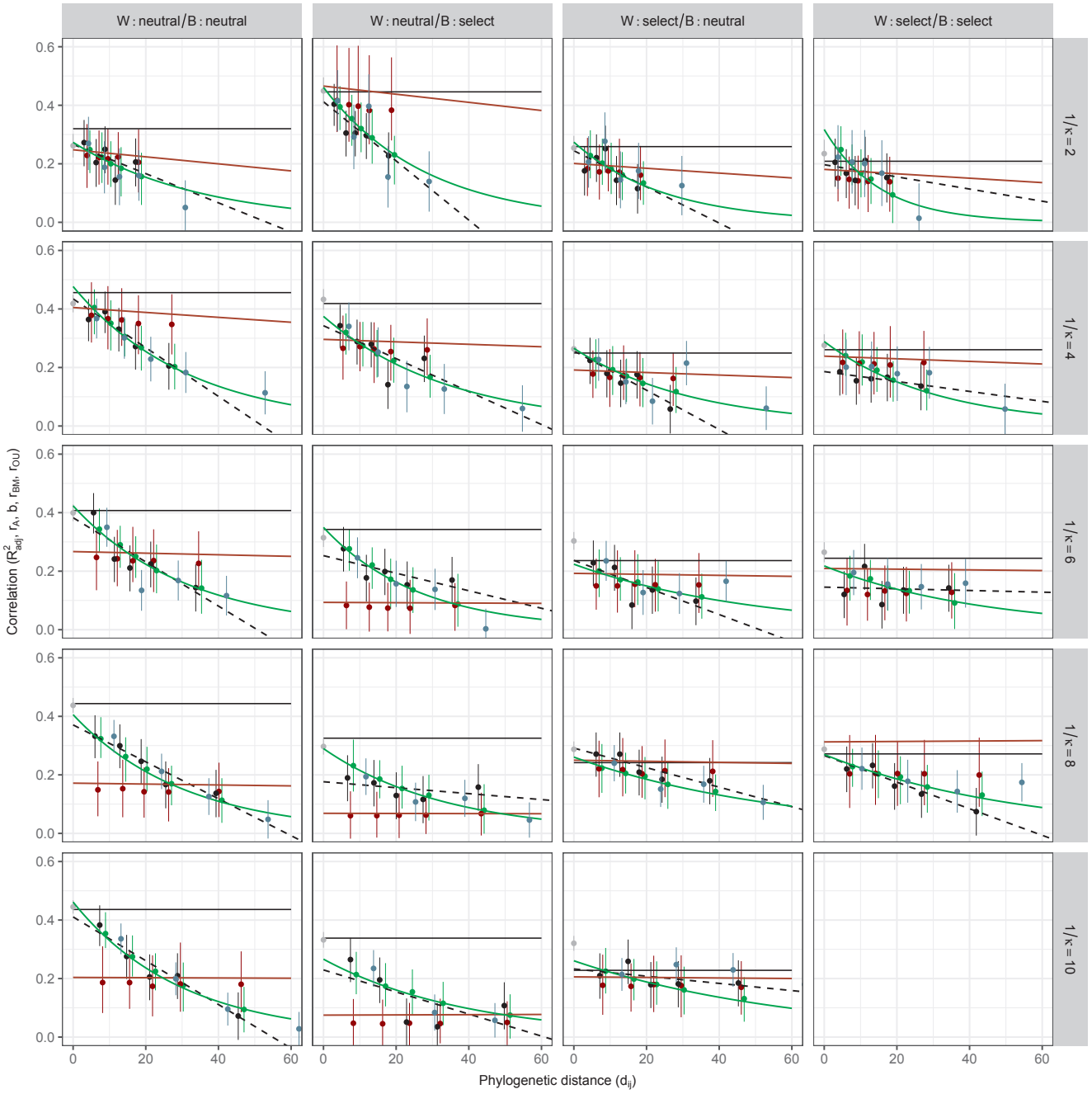


**FIG. S7. Correlation in phylogenetic pairs and donor-recipient couples in toy-model simulations.** Each panel displays the correlation as a function of $d_{ij}$ in a randomly chosen epidemic for a given scenario and mean contact interval, $1/\kappa$. In each simulated epidemic, we consider the population of the first 10,000 diagnozed individuals. In this population, the exact transmission tree and transmission couples are known. A black horizontal line represents the true value of $H^2$ measured by the direct estimator $R^2_{adj}$. Dots and vertical bars display point-estimates and 95% CIs of $r_A$ in the PPs and of $b$ in the donor-recipient couples upon a stratification into quintiles of $d_{ij}$. Black: $r_A$ in PPs; brown: $r_A$ in PPs after replacing the trait-values simulated under the toy-model with values simulated under the ML fit of the PMM; green: $r_A$ in PPs after replacing the trait-values simulated under the toy-model with values simulated under the ML fit of the POUMM; cadet-blue: $b$ in donor-recipient's; grey (only for $d_{ij} = 0$): $b_0$ in donor-recipient's based on trait-values at moment of infection.

A brown and a green line indicate the correlation between tip-pairs in the tree expected under the ML fit of the PMM and POUMM respectively.

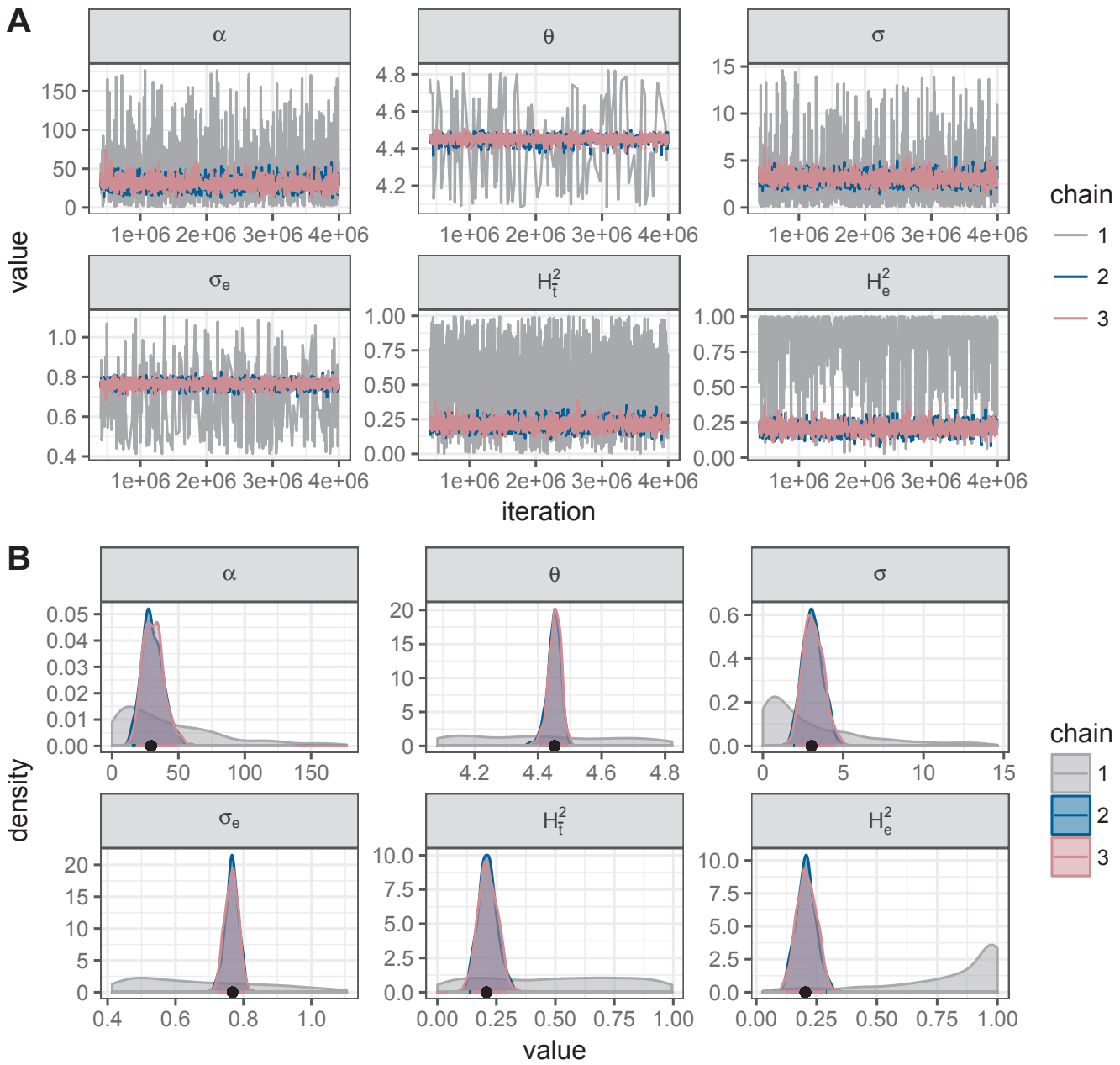


**FIG. S8. Trace-plots and posterior densities from the POUMM MCMC-fits to HIV from the UK cohort (8483 patients).** Three MCMC chains have been executed: 1 - sampling from the prior distribution; 2 and 3 - sampling from the posterior distribution. (A) Trace-plots - the randomness and the lack of time-correlation in the traces show the correct mixing of the MCMC chain; (B) Inferred posterior densities. The clear distinction between prior and posterior densities proves the presence of informative signal in the data. The match between the densities from chain 2 and 3 proves the convergence of the MCMCs towards the posterior distribution. This convergence was also validated through the Gelman-Rubin statistic being nearly equal to 1 (results not shown).

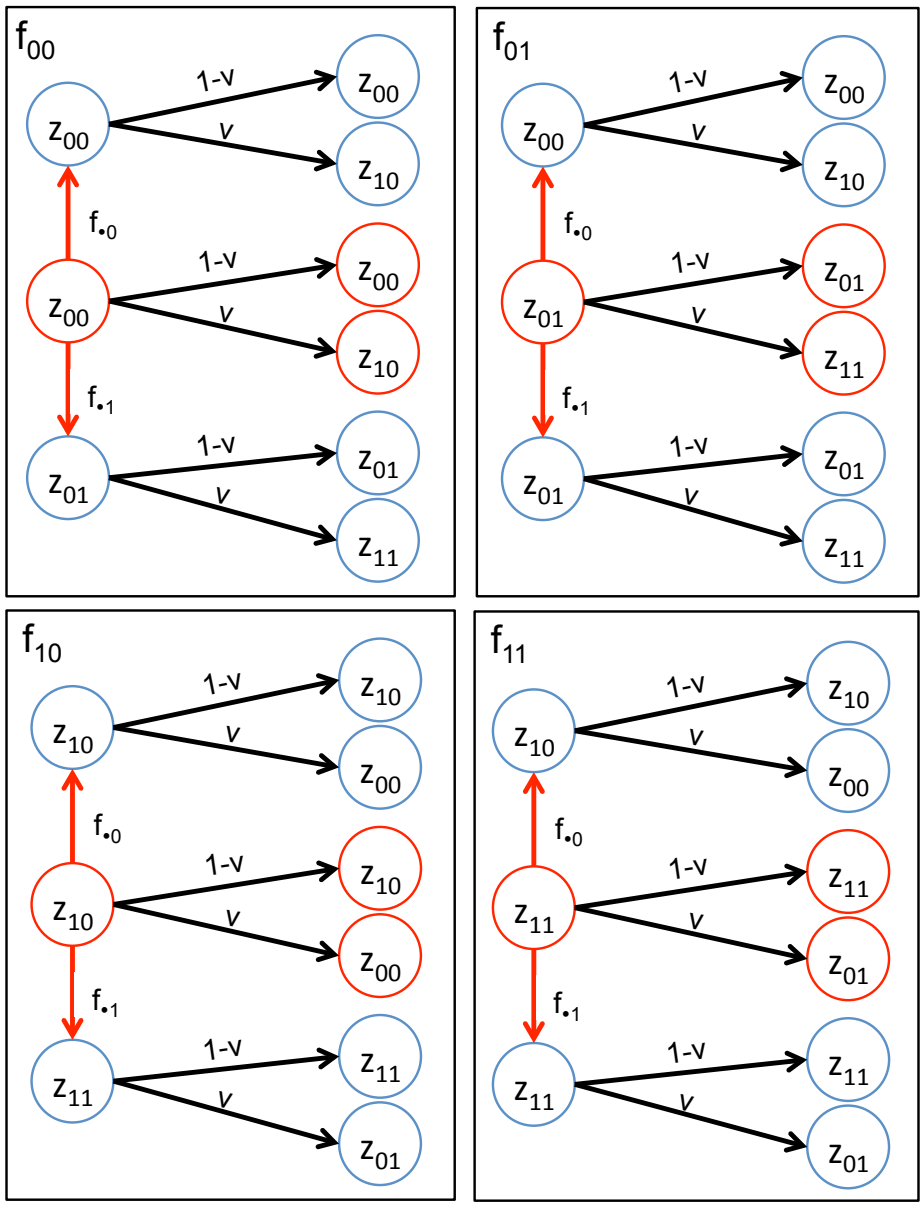


**FIG. S9. Expected couples of donor-recipient trait values at $d_{ij}/2$ past transmission** Red circles denote donors, blue circles denote recipients. Red vertical arrows denote transmission. Black left-to-right arrows denote mutation during the time from transmission to measurement in the donors and the recipients. The weights above the arrows denote the probability of the transmission or mutation happening. The diagram can be read in the following way (example): at the moment of a generation, a type 00 infected host transmits its pathogen to a susceptible individual of host-type 0 with probability $f_{.0}$. After the transmission event the strain in each of the two hosts has a chance $\nu$ to be substituted by a mutant strain. Thus, the probability of having a donor recipient couple, in which both hosts have a state 00 at the moment of measurement, given that the donor was type 00 at the moment of transmission, is equal to the product $\nu f_{.0} \nu$. It remains to multiply this by the frequency of encountering a type 00 donor, to obtain the overall probability of the event.

## References

Yang, Z. 2006. *Computational Molecular Evolution*. OUP Oxford.