Full title: Untangling the effects of within-host and between-host natural selection in viruses with applications to evolution of HIV-1 co-receptor usage

Short title: Quantifying within-host and between-host selection in virus evolution

Kieran O. Drake[1¶], Vinicius B. Franceschi[1¶], Fabrícia F. Nascimento[1¶], Erik M. Volz[1¶*], on behalf of the PANGEA Consortium[†]

[1]MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK

[†]A full list of the PANGEA consortium members appear at the end of the manuscript

[*]Corresponding author: e.volz@imperial.ac.uk

[¶]These authors contributed equally to this work.

28 April, 2025

Abstract

The fitness of a pathogen can be characterised at multiple levels corresponding to different stages of its life cycle, including replication within target cells (replicative capacity), immune evasion, and the capacity to infect new hosts (transmissibility). Pathogens evolve towards greater fitness within hosts, which can potentially compromise transmissibility. Typically, epidemiological studies for measuring pathogen fitness compute a single selection coefficient for pathogen variants that encompasses all stages of the pathogen life cycle. To address this problem, we develop a population-genetic model and software for the inference of multi-level selection coefficients which represent relative within-host fitness and transmissibility. Our approach builds on population genetic models previously applied to macro-evolution, whereby speciation and extinction rates are state-dependent. We relate the state-dependence in speciation and extinction rates to competing within- and between-host selection. Evolution of beneficial phenotypes within hosts may occur rapidly (higher 'speciation'), but reduce rates of dispersal between hosts (higher 'extinction'). This new coalescent-based method also accounts for highly nonlinear-population dynamics and variable sampling through time making it suitable for epidemiological studies. We derive the equilibrium proportion of a variant with multi-level selection under mutation-selection balance and show how this can be used in combination with the coalescent to greatly increase accuracy and precision of estimated selection coefficients. We characterise the accuracy of this method using a variety of birth-death and coalescent simulations before analysing a real dataset of HIV-1 genetic sequences from a large population-based sample of 4056 people living with HIV in Zambia. We quantified the multi-level relative fitness underlying HIV-1 'co-receptor switching'. In accordance with previous observations, we estimate that the relative transmissibility of X4-tropic virus is 9%, while within-host switching to X4-tropism is approximately seven-fold higher than switching to R5-tropism, indicating a strong within-host and between-host tradeoff of co-receptor usage. Our methods are available in an open-source R package `musseco`, providing a general-purpose tool for inferring multi-level selection coefficients for epidemiological sequence data.

Author Summary

The fitness of a pathogen can be characterised at each stage of its life cycle. However, epidemiological studies measuring pathogen fitness typically estimate a single value encompassing the entire life cycle. Pathogens evolve towards greater fitness within hosts, but this can compromise the capacity to infect new hosts. Our new method quantifies relative fitness at both levels. This method

accounts for rapidly changing pathogen populations and variable sampling through time making it suitable for epidemiological studies. We tested our method using simulations and demonstrate its application by analysing real HIV-1 genetic sequences from people living with HIV in Zambia. This method sheds new light on long-standing questions about the relative fitness of HIV-1 variants which utilise different host-cell co-receptors. Using our new method, we estimate that the capacity of X4 variants to infect new people is just 9% of that of R5 variants, while within-host switching from R5 to X4 usage is seven times greater than the reverse switch. This indicates a strong within-host and between-host trade-off in co-receptor usage.

1 Introduction

Microbial evolution often features episodic natural selection that stems from competing selective pressures in fluctuating environments. The life cycle of a pathogen exhibits such a fluctuating environment, and requires the successful completion of multiple steps: Crossing tissue barriers, evading innate immunity, binding to a target cell, replicating within host cells, evading adaptive immunity, and shedding infectious particles such that the cycle may repeat. The selective pressures at each stage can conflict. For example, pathogen phenotypes optimised for one step (such as immune evasion) may perform poorly at others (like cell binding). Natural selection within hosts necessarily involves compromising some steps of the infection cycle to advance others.

Within-host and between-host evolutionary compromises are well-documented across diverse pathogens. The evolution of antimicrobial resistance often comes at the cost of reduced replicative fitness within host cells. Similarly, mutations enabling escape from adaptive immunity can impair receptor binding capacity [1]. A particularly illustrative example, which we examine in detail below, concerns HIV-1 co-receptor switching [2, 3, 4], such that different viral phenotypes experience varying selective advantages during the course of a single host infection. The use of the R5 co-receptor by viral variants to enter host cells is predominant at transmission and during all stages of infection. However, HIV-1 evolves during the course of infection to switch to the X4 co-receptor (or both co-receptors, i.e. dual tropic/R5X4) in up to 50% of individuals [2, 4], which is associated with a poorer baseline clinical status [5] and faster progression to AIDS [6, 7]. Although early studies suggested that R5 viruses were preferentially selected at transmission, as a result of a single or

multiple leaky "gatekeepers" [8, 9], later transmission cluster analysis demonstrated that individuals infected by X4 viruses regularly transmit X4/R5X4, suggesting a stochastic process (random transmission hypothesis) [10, 11]. We address this question formally, and our findings support the existence of reduced transmissibility of X4 virus while indicating that limited transmission of X4 virus can still occur due to stochastic effects.

In quantitative terms, it is impossible to describe the fitness of a given phenotype using a single selection coefficient in the presence of multi-level selection. While multiple selection coefficients are needed to fully describe fitness variation across different environments, epidemiological studies have typically collapsed these into single estimates of variant fitness [12, 13]. And, despite the ubiquity of episodic selection in pathogens, there has been relatively little progress on methods to estimate multi-level selection coefficients of genomic variants. This is a challenging inference because of the large difference in timescales governing within-host pathogen replication and between-host transmission. Evolutionary rates within hosts are typically much higher than those measured between hosts, and the replication cycle of a pathogen between cells can be many orders of magnitude faster than the infection cycle between hosts. Another difficulty concerns the type of data available. Epidemiological surveillance data typically provides only one or a handful of sequences per host, but assessing the within-host fitness and transmissibility of particular variants would ideally make use of many sequences from hosts taken over the course of their infection which would itself be drawn from a population-based sample of cases. While deep sequencing data is enabling investigations into within-host microbial evolution [14, 15], such data are typically only collected at a single time point for each host.

In this paper, we present a general-purpose method for inferring multi-level selection coefficients for epidemiological sequence data. Our approach is similar to population genetic models previously applied in the macro-evolution setting, where multi-level pathogen selection parallels state-dependent speciation and extinction over evolutionary timescales [16]. In phylodynamic terms, competing within-host and between-host selection creates analogous state-dependence in speciation and extinction rates, but compressed into epidemiological timescales. Evolution of beneficial phenotypes within hosts may occur rapidly (higher 'speciation'), but face reduced rates of dispersal between hosts (higher 'extinction'). The key insight that enables inference of multi-level selection coefficients is that fitness depends on phenotypic state which changes over the phylogeny

of the pathogen. Estimation of selection coefficients is thus coupled to reconstruction of ancestral phenotypes given the observed traits in sampled lineages.

Some limitations of previously-developed state-dependent selection models preclude their use for analysis of infectious diseases. On the short timescales of pathogen evolution, epidemiological dynamics need to be taken into account. Epidemic dynamics often leads to highly nonlinear changes in effective population size that cover many orders of magnitude. Furthermore, in contrast to studies of macro-evolutionary processes, sampling of microbes is highly distributed through time, and pathogen phylogenies often consist of a mixture of sampled lineages near the present and close to the most recent common ancestor. Sampling through time is often non-random, and the rate of sampling can vary drastically and unpredictably over time, since sampling is often non-systematic and responsive to emerging public health incidents.

In this study, we develop a coalescent-based method for inferring multi-level selection that accounts for highly nonlinear population dynamics and variable sampling through time. Our approach provides a general-purpose tool for inference of multi-level selection coefficients. It can be applied to any dataset where a time-scaled phylogeny can be estimated and where the phenotype (e.g. drug resistance, immune escape, or co-receptor usage) is known for each sample. This method is most appropriate for situations where one pathogen genome is sampled per host. It is optimised for analyses of pathogen genomes by automatically estimating effective population size over time ($N_e(t)$), adjusting for nonlinear-population dynamics, allowing for heterochronous sampling, and conditioning on known sample dates.

2 Methods

2.1 The binary-state mutation and transmissibility model

In analogous fashion to the binary-state speciation and extinction (BiSSE) model [16], we specify the dynamics of two variants with different fitness at separate timescales. Our model is premised on the co-evolution of two variants $\mathcal{A}$ and $\mathcal{V}$ such that within hosts $\mathcal{V}$ has a selective advantage, while also having lower transmissibility between hosts. The number of infected hosts with each variant over time is $Y_a(t)$ and $Y_v(t)$ and we will assume unless otherwise stated that the variants are in a mutation-selection balance, such that $Y(t) = Y_a(t) + Y_v(t)$ and $Y_a(t) = p_a Y(t)$ and $Y_v(t) =$

$(1 - p_a)Y(t)$ where p_a is the proportion which is of the ancestral type (c.f. section 2.2.2). The variants have different rates of transmission between hosts, with time-dependent per-capita rates $\beta_a(t)$ and $\beta_v(t)$. We parametrise the relative fitness of variant $\mathcal{V}$ using

$$\beta_v(t) = \omega\beta_a(t), \tag{1}$$

and assuming that a fitness penalty for $\mathcal{V}$ such that $\omega < 1$. We sometimes also refer to the *selection coefficient* for $\mathcal{V}$ defined as $s = \omega - 1$ and require that $-1 < s < 0$. We also assume that the generation time (mean time between successive infections) is the same between variants.

This model will require that each infected host at each time point is colonised by either $\mathcal{A}$ or $\mathcal{V}$ and that mutation from one variant to another is instantaneous. This model is therefore premised on within-host genetic diversity being negligible. We parametrised the increased within-host fitness of $\mathcal{V}$ in terms of an imbalance between the rate of mutation from $\mathcal{V}$ to $\mathcal{A}$ and vice-versa. The mutation rate in a single host from $\mathcal{A}$ to $\mathcal{V}$ and from $\mathcal{V}$ to $\mathcal{A}$ are denoted μ_{av} and μ, respectively. Our model is premised on increased mutation from $\mathcal{A}$ to $\mathcal{V}$ by a factor α:

$$\mu_{av} = \alpha\mu, \tag{2}$$

such that $\alpha > 1$.

Describing multi-level selection therefore requires estimation of two parameters, ω and α. A compartmental diagram summarising the co-evolution of $\mathcal{A}$ and $\mathcal{V}$ is shown in Fig 1. In sections 2.2.2 and 2.2.1 we show how these parameters determined a mutation-selection balance between variants and how to account for epidemiological dynamics by appropriate choice of functional forms for $Y(t)$ and $\beta_x(t)$. Figure 1 also illustrates genealogies simulated under different scenarios of strong and weak within-host and between-host selection. In the weak-selection scenario, clades of $\mathcal{V}$ have longer persistence times and larger size, but under the constraint of a given mutation-selection balance, such a variant has correspondingly weak within-host selection leading to infrequent de-novo occurrences within hosts. In contrast, under strong within- and between-host selection, $\mathcal{V}$

occurs frequently throughout a genealogy, but clades of $\mathcal{V}$ rapidly go extinct.

2.2 The structured coalescent and modelling variant co-evolution

To model gene genealogies generated by the binary-state co-evolution model, we use the structured coalescent framework previously developed in [19] and [20]. This coalescent model specifies the rates of lineage migration and coalescence in terms of the corresponding population dynamics comprising three deterministic matrix-valued functions of time and the state of the system with m demes:

- Births: $F_{1:m,1:m}(t, Y)$. In this study, this represents between-host transmission rates for different pathogen variants.
- Mutation: $G_{1:m,1:m}(t, Y)$. In this study, this represents within-host mutation rates.
- Deaths: $\Gamma_{1:m}(t, Y)$. These terms may also correspond to recovery or removal from the infectious population in epidemiological models.

We order $m = 2$ demes corresponding to the co-evolving variants: $\mathcal{D} = (\mathcal{A}, \mathcal{V})$. Then, the *birth* matrix represents dispersal rates to new hosts and depends on the relative transmissibility parameter ω.

$$F(t) = \begin{bmatrix} \beta(t)p_a Y(t) & 0 \\ 0 & \omega\beta(t)(1-p_a)Y(t) \end{bmatrix}, \tag{3}$$

where $\beta(t)$ is the time-varying per-capita transmission rate and $Y(t)$ is the total number of infected hosts. We show below how $\beta(t)$ and $Y(t)$ are derived in terms of the effective population size over time. These birth rates appear on the diagonal since a host colonised with a particular variant may only generate new infections of the same type. In contrast, mutation terms appear on the off-diagonal and generate the simultaneous loss of a lineage of one type ($\mathcal{A}$ or $\mathcal{V}$) and the gain of the other type:

$$G(t) = \begin{bmatrix} 0 & \alpha\mu p_a Y(t) \\ \mu(1-p_a)Y(t) & 0 \end{bmatrix} \tag{4}$$

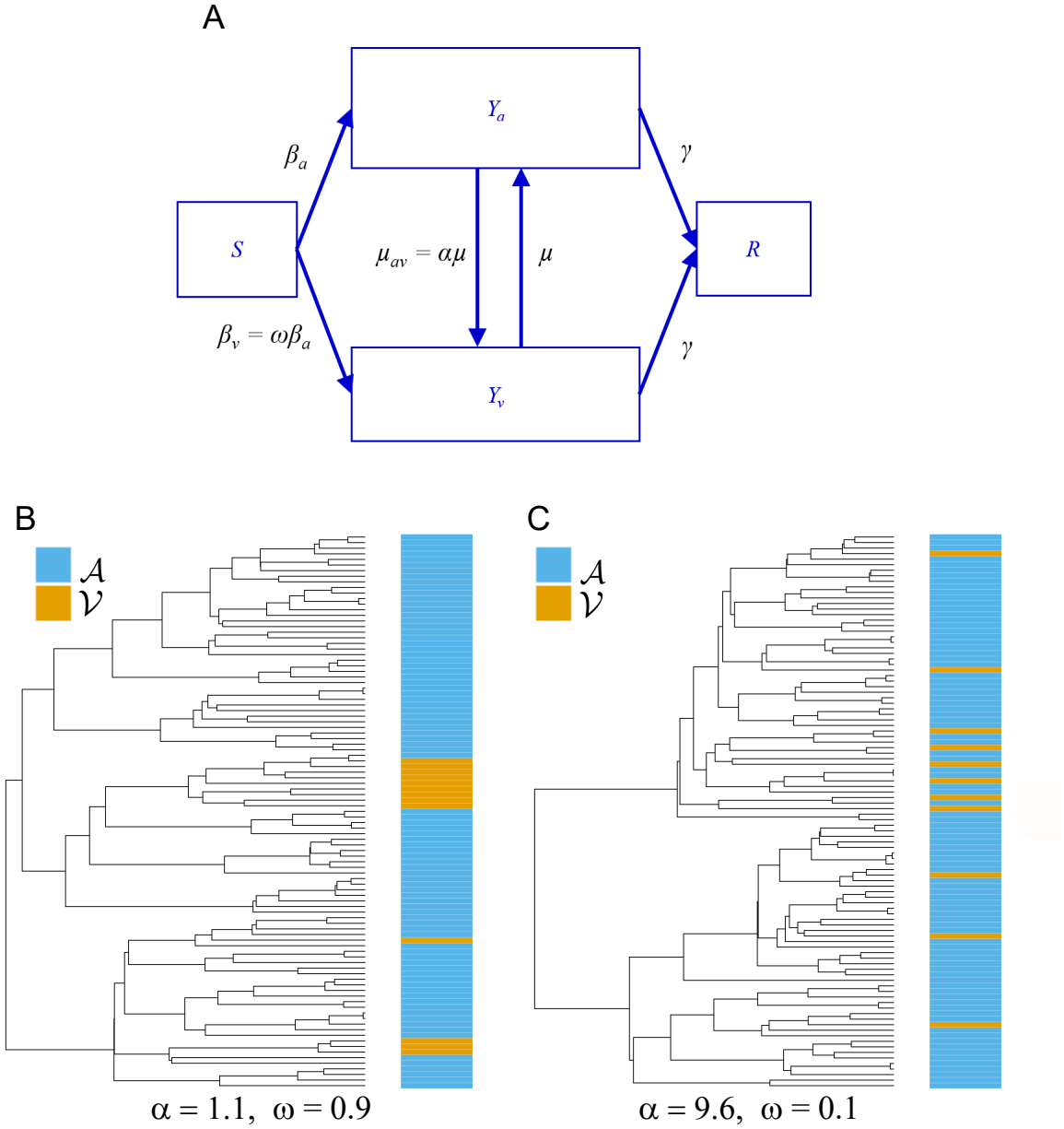

Figure 1: **Binary-state mutation and transmissibility model**: (A) Illustrative epidemiological compartmental model considering multi-level fitness. The model includes two variants, $\mathcal{A}$ and $\mathcal{V}$, which have different between-host transmission rates, with ratio ω, and asymmetric within-host mutation rates with ratio α. Panels (B) and (C) illustrate genealogies with identical mutation-selection balance of $\mathcal{V}$ ($\approx 15\%$), yet with different cladistic patterns stemming from different within- and between-host selection. Phylogenies with 100 tips were simulated with *diversitree* [17] and plotted using *ggtree* R package [18]. (B) Weak within- and between-host selection. The relative transmissibility of $\mathcal{V}$ is 90% ($\omega = 0.9$) and $\alpha = 1.1$. (C) Strong within- and between-host selection. The relative transmissibility of $\mathcal{V}$ is 10% ($\omega = 0.1$) and $\alpha = 9.6$.

Finally, death rates determine the generation time of the process and allow translation of genealogical time to calendar time:

$$\Gamma(t) = \begin{bmatrix} \gamma Y_a(t) \\ \gamma Y_v(t) \end{bmatrix}, \tag{5}$$

where the generation time $T_g = 1/\gamma$ is assumed constant between variants.

In the supporting information, we review how the likelihood of a tip-labeled genealogy is computed conditional on the epidemiological history specified by $\mathcal{M} = (F(t, Y), G(t, Y), \Gamma(t, Y))$. With $\mathcal{M}$ defined, the population size of each deme k can be computed by solving the system of ordinary differential equations:

$$\dot{Y}_k = -\Gamma_k(t) + \sum_{l=1}^{m} (F_{lk}(t) + G_{lk}(t) - G_{kl}(t)). \tag{6}$$

2.2.1 Effective population size

To adjust for variation in the coalescent rate that is unrelated to the coevolution of $\mathcal{A}$ and $\mathcal{V}$, we calibrate $\beta(t)$ and $Y(t)$ so that the total coalescent rate (averaged over all demes) matches an empirical estimate that is obtained separately. Specifically, suppose that we are provided with an estimate of $N_e(t)$ for the entire gene genealogy that neglects population structure in the form the co-circulating variants. Our software implementation will estimate $N_e(t)$ using a non-parametric maximum likelihood method by default [21] (c.f. section 2.5), but in principle, any method can be used and provided as an input to our algorithm. Note that the scale of $Y(t)$ is arbitrary, because the structured coalescent is over-determined, and many different forms of $\beta(t)$ and $Y(t)$ would yield the same $N_e(t)$. We need only to derive a definition such that changes in $Y(t)$ yield identical changes in the pre-estimated $N_e(t)$. Our approximation to $Y(t)$ is premised on the following:

1. Without loss of generality, $\beta_a(t) = 1$, and transmissions $\phi_a(t)$ per unit time from $\mathcal{A}$ will be equal to population size $Y_a(t)$. For $\mathcal{V}$, we will use $\beta_v(t) = \omega$ and transmissions $\phi_v(t)$ will be: $\phi_v(t) = \omega Y_v(t)$

2. The coalescent rate for a pair of lineages in a deme x is based on the coalescent model in [19]:

$\lambda_x(t) = 2\phi(t)/Y_x(t)^2$.

$$\begin{aligned}\lambda_a(t) &= 2\phi_a(t)/Y_a(t)^2 \\ &= \frac{2}{p_a Y(t)} \qquad (7)\\ \lambda_v(t) &= 2\phi_v(t)/Y_v(t)^2 \\ &= \frac{2\omega}{(1-p_a)Y(t)} \qquad (8)\end{aligned}$$

3. And, given the binomial probabilities p_x^2 of selecting two lineages from the same deme x, the approximate pairwise coalescent rate for the population as a whole will be

$$\begin{aligned}\lambda(t) &= p_a^2\lambda_a(t) + (1-p_a)^2\lambda_v(t) \qquad (9)\\ &= \frac{p_a}{Y(t)} + \frac{(1-p_a)\omega}{Y(t)}, \qquad (10)\end{aligned}$$

where the last line is obtained by substituting expressions from 7 and 8.

To solve for $Y(t)$, we equate $\lambda(t)$ (equation 10) to $\iota/N_e(t)$, where ι is an additional scaling parameter which is estimated to improve the overall estimate of the coalescent rate. This yields

$$Y(t) = 2\iota N_e(t)\left(p_a + (1-p_a)\omega\right) \qquad (11)$$

In practice, we observe that estimates of ι are very close to one (c.f. section 3.1), indicating that equation 11 is a good approximation in many situations.

2.2.2 Mutation-selection balance

If the variant and ancestral types are in equilibrium (mutation-selection balance) it is possible to construct a likelihood of sampling proportions that are of the ancestral type (p_a) and variant type ($p_v = 1 - p_a$) assuming random sampling. This likelihood can be used in combination with the structured coalescent likelihood to gain additional precision when estimating selection coefficients, although the conditions of random sampling and mutation-selection balance should be evaluated first.

Here we solve for p_a by solving a set of detailed balance equations involving transmission, death (or recovery) and mutation. The corresponding rates are:

	Ancestral	Variant
Transmissions:	$+\beta_a(t)Y_a(t)$	$+\beta_v(t)Y_v(t)$
		$= +(1+s)\beta_a(t)Y_v(t)$
		$= +\omega\beta_a(t)Y_v(t)$
Deaths:	$-\gamma Y_a(t)$	$-\gamma Y_v(t)$
Mutations:	$-\mu_{av}Y_a(t) + \mu Y_v(t)$	$-\mu Y_v(t) + \mu_{av}Y_a(t)$
	$= -\alpha\mu Y_a(t) + \mu Y_v(t)$	$= -\mu Y_v(t) + \alpha\mu Y_a(t)$

And, the following equations must be satisfied at equilibrium after dividing all expressions by $Y(t) = Y_a(t) + Y_v(t)$:

$$p_a\beta_a(t) = \gamma p_a + \mu\alpha p_a - \mu(1-p_a) \tag{12}$$

$$(1-p_a)\omega\beta_a(t) = \gamma(1-p_a) + \mu(1-p_a) - \mu\alpha p_a \tag{13}$$

Combining these equations and solving for p_a yields

$$0 = p_a(\gamma(1-p_a) + \mu(1-p_a) - \mu\alpha p_a) - (\omega(1-p_a))(\gamma p_a + \mu\alpha p_a - \mu(1-p_a)) \tag{14}$$

This quadratic equation in p_a has at most one root in the unit interval given the positive domain of all other parameters.

2.3 Simulations

We used three methods which support genealogical simulations in structured populations, *Coalescent.jl* [22], *TiPS* [23] and *diversitree* [17] to simulate trees in which we know the true values of the relative between-host fitness (ω) and relative within-host fitness (α) to test the new coalescent model. Two of these simulators, the *Coalescent.jl*[22] and *TiPS* [23], are based on the structured coalescent and require an epidemiological model describing the dynamics of infectious individuals. The *diversitree* method is based on a birth-death model and has previously been used to study the multi-state speciation and extinction process.

2.3.1 Epidemiological model

To carry out simulations with *Coalescent.jl* and *TiPS*, we developed a mathematical epidemiological model to describe the dynamics of ancestral and variant states. This model was defined as two ordinary differential equations (ODEs) describing the number of infectious individuals carrying sequences with the ancestral state $Y_a(t)$ and the number of infectious individuals carrying sequences with the variant state $Y_v(t)$:

$$\begin{aligned} Y_a(t) &= \beta Y_a(1 - \frac{N}{K}) + \mu Y_v - \mu_{av} Y_a - \gamma Y_a \\ Y_v(t) &= \beta(1+s) Y_v(1 - \frac{N}{K}) - \mu Y_v + \mu_{av} Y_a - \gamma Y_v \end{aligned} \tag{15}$$

The rate of new infections was defined as logistic growth controlled by the parameter β for $Y_a(t)$ and $\beta(1+s)$, where s is the selection coefficient, for $Y_v(t)$. The carrying capacity was defined by the parameter K and was fixed at 10,000 individuals. Infected individuals recover at the rate γ.

Individuals containing the ancestral state, $Y_a(t)$, mutate to the variant state, $Y_v(t)$, at a rate μ_{av}, and individuals carrying the variant state could revert to the ancestral state at a rate μ.

2.3.2 Genealogical simulations

The parameter values used to simulate genealogies with *Coalescent.jl* (v0.1.4) and *TiPS* (v1.2.3) are described in Table S1. Both methods employ Gillespie's direct algorithm with time-varying coalescent and migration rates. We obtained the value of the transition rate from Y_a to Y_v (μ_{av}) using the theoretical equation for the mutation-selection balance using fixed values for the proportion of individuals in the ancestral state, the molecular clock rate (here also defined as the transition rate from Y_v to Y_a), and the relative transmission fitness ($\omega = 1 + s$).

To simulate trees with *Coalescent.jl* and *TiPS*, we also provided a sampling date randomly sampled in the past 5 years using the Julia programming language. We then used the same sampling date from *Coalescent.jl* trees to simulate trees using the *TiPS* R package.

For *diversitree* (v0.10.1) simulations, we analysed different scenarios of population dynamics with a high and low R_0 (basic reproduction number). Note that R_0 depends on the proportion of

ancestral state and the birth rates. Each combination of parameter values analysed with *diversitree* had a different birth rate, but we grouped the results into two different values of R_0: A high value of approximately 1.85 and a low value of approximately 1.22 (See Table S2 for a complete list of parameter values and the R_0 for each combination of parameter values used with *diversitree*).

For all combinations of parameter values tested, we also sampled a total of 500 or 1000 individuals to understand how the sample size would influence the estimate of parameters.

2.3.3 Simulation and analysis of genetic sequences

We carried out an additional set of simulations to mimic the observed HIV data described in section 2.4. We used *Coalescent.jl* with the same epidemiological model as described in section 2.3.1 to simulate genealogies in which trajectories were simulated for 29 or 45 years. For a list of the parameter values used, see Table S3. We simulated genealogies with 4056 tips with sampling time from the past 4 years (which was observed for the PANGEA dataset, see section 2.4). These simulated genealogies will be referred to as "true trees".

We simulated sequence alignments of 105 bp and 2371 bp, based on the observed alignment length for the empirical dataset, using the simulated true trees and the program *AliSim* [24] to understand the impact of sequence alignment length on the estimation of relative fitness α and ω. We simulated sequence alignments without gaps or recombination that started from random starting sequences. We then estimated maximum likelihood (ML) trees using the simulated sequence alignments and the program *IQ-TREE* (v2.4.0) [25], followed by the estimation of time-scaled trees using the strict clock implemented in *treedater* (v.0.5.3) R package.

2.3.4 Statistical inference

Fitness parameters ω and α were estimated by maximum likelihood using prior knowledge of the generation time (equivalent to $1/\gamma$) and the molecular clock rate of evolution, which we assume is equivalent to the reversion rate μ. Inference was carried out using both the coalescent and the augmented-coalescent likelihoods (see Supplementary Information for details). Approximate confidence intervals were computed assuming asymptotic normality of the maximum likelihood estimate (MLE) and computing standard errors from the observed Fisher information (negative Hessian of the log likelihood).

Performance of the estimator is summarised by relative error ($\|(\hat{X} - X)/X\|$) and the coverage probability of the 95% confidence interval.

2.4 Multi-level evolution of HIV-1 co-receptor tropism

In order to disentangle the effects of within- and between-host selection related to HIV-1 co-receptor usage, we used data from a high-prevalence setting in Sub-Saharan Africa. The HPTN 071 study, commonly referred to as PopART (Population Effects of Antiretroviral Therapy to Reduce HIV Transmission) was a large-scale community-based trial conducted in South Africa and Zambia between 2013 and 2018. Its primary aim was to assess the effect of a universal test-and-treat strategy on population-level HIV incidence. During the study period, ≈5% (n>48,000) of the total population of >1 million people were surveyed, including blood samples, annually for three years [26]. The HPTN 071-2 PopART Phylogenetics Ancillary study used samples from HIV seroconverters within the PopART cohort from Zambia only and HIV-positive samples collected from health clinics across nine of the 12 Zambian PopART communities, which were then subjected to sequencing and genome assembly processing as previously described [27]. Access to HIV-1 genomes and appropriate metadata was granted by the PANGEA-HIV 2 (Phylogenetics And Networks for Generalised Epidemics in Africa) Steering Committee after the acceptance of a concept sheet proposal.

2.4.1 Data curation and phylogenetic tree reconstruction

Subtyping of all consensus sequences from this cohort was performed using *COMET* v2.4 [28]. Only the first sequence per individual and subtype C sequences (n=5460, 85.9%) were retained. The vast majority (85.3%) of those individuals were ART-naïve, and sequences from ART-experienced individuals were removed. Whole genome sequences (WGS) were aligned against the subtype B HXB2 reference using *MAFFT* v7.505 (auto and add-fragments options) [29]. Potential recombinants (<95% subtype C as determined by the Los Alamos National Laboratory [LANL] Recombinant Identification Program [*RIP*]) and hypermutated sequences were removed using the LANL Quality Control tool [30]. Finally, to ensure all data had a sufficient number of phylogenetically-informative sites, sequences shorter than 6500 ungapped nucleotides were removed, leading to n=4249 sequences.

Subsequently, HIV-1 genes were extracted. For the purposes of this work, which aimed to

estimate multi-level selection of HIV-1 co-receptor usage, we retained *env* and V3 loop alignment subsets (alignment sizes of 2371 and 126 columns, respectively). For the complete *env* analysis, the highly variable loops V1 to V5 of *env* were masked following previous recommendations [31]. As simulations demonstrated (see Section 2.3.3), an alignment of approximately 100 nucleotides did not contain sufficient information to allow a reliable inference of α and ω parameters. Therefore, we do not report such fitness estimates derived from V3 sequences.

After testing multiple evolutionary models using *ModelFinder* v2.2.2.6 [32], we built an initial maximum likelihood (ML) tree using *IQ-TREE* v2.2.2.6 [25]. This analysis employed the general time reversible (GTR) model [33] with a free-rate model for rate heterogeneity across four categories (R4), iterating 100 times. A few sequences (n=7 for *env*) with >50% gaps or erroneous long branches in the ML tree were removed. After these curation steps, the *env* ML tree was re-built using: (i) the GTR nucleotide substitution model with empirical state frequencies (F), a proportion of invariable sites (I), and ten free-rate categories (R10); (ii) 1000 ultrafast bootstrap replicates; (iii) a perturbation strength of 0.2; and (iv) 3000 minimum iterations. Time-scaled phylogenetic trees were estimated after removing the HXB2 reference using an additive uncorrelated model [34] as implemented in *treedater* v0.5.3 [35] with 10 maximum iterations. A sub-sampled version of the time-scaled phylogenetic tree is shown in Fig S13 as an example.

2.4.2 Co-receptor tropism and coalescent analyses

Nucleotide sequences spanning the V3 loop located in the gp120 genomic region inside *env* were extracted and submitted for co-receptor prediction using the *WebPSSM* server and a position- and subtype-C-specific SINSI scoring matrix [36, 37]. Subsequently, the distribution of *WebPSSM* scores, trends of co-receptor usage over time, distribution of CD4 counts and ranges for each predicted co-receptor, and the ratio of predicted R5 versus X4/R5X4 were computed.

For the coalescent analysis, the previously estimated *env* time-scaled tree was used alongside the co-receptor tropism predictions derived from the V3 loop region. A total of 4056 tips were considered, from which sample collection dates ranged from February 2014 to June 2018. The `fitbisseco` function from the *musseco* (v0.1.0) R package [38] was used. X4/R5X4 predicted sequences were treated as the variant state and R5 as the ancestral. The generation time (T_g) was defined as $\frac{1}{\gamma}$, using the same γ value as in the simulations (Table S1). The molecular clock rate

of evolution was based on the *treedater* estimate for *env* (3.2×10^{-3} nucleotide substitutions per site per year). The effective population size was internally estimated using the *mlesky* (v0.1.7) R package [21]. During this process, the smoothing parameter (τ) was optimised using cross-validation and uncertainty was quantified using parametric bootstrap (`parboot` function) with 100 replicates. The default initial values of the α, ω and $Y(t)$ scale parameters were used alongside the default augmented-coalescent likelihood.

2.5 Implementation

We have implemented the binary-state coalescent model as an open source R package `musseco` available at `https://emvolz-phylodynamics.github.io/musseco`. This includes a function, `fitbisseco`, that estimates multi-level selection coefficients by maximum likelihood and will estimate effective population size automatically if an estimate is not provided. The `fitbisseco` function requires a number of inputs in order to estimate the fitness parameters. A time-scaled phylogeny is required and can be estimated, for example, using the *treedater* R package [35]. This will also estimate the molecular clock rate of evolution, μ, which is another required input. Pathogen phylogenies are assumed to be reconstructed from population-based random samples of pathogen genomes and at most one sequence per host. The effective population size over time, $N_e(t)$, can also be input, although if it is not provided then it will be estimated with the *mlesky* R package [21] using the *skygrid* model [39] by default. This estimate of $N_e(t)$ is also provided in the returned BiSSECo fit using the `fitbisseco` function in the *musseco* R package [38]. Sample states, i.e. ancestral or variant, are also required to be defined for tips in the time-scaled phylogeny and included as an input to the `fitbisseco` function. An estimate of the generation time, T_g, is also required. Initial estimates of α, ω, and $Y(t)$ scaling parameter are also required as starting points for the estimation process, which is described in the Supporting Information.

Scripts used to simulate and analyse data are available at `https://github.com/thednainus/musseco_simulations`.

Scripts to analyse empirical data from HIV-1 are available at `https://github.com/vinibfranc/BiSSECo_co-receptor_PANGEA`.

3 Results

3.1 Simulations

Across most simulation studies, the augmented-coalescent outperformed the coalescent model that did not include a sampling likelihood based on mutation-selection balance. While the augmented-coalescent MLE generally had good precision (Fig 2A), the approximate confidence intervals did not always cover the true value, especially for birth-death simulations with high R_0.

3.1.1 Coverage

When estimating α, coverage was generally close to or equal to 100% for analyses carried out with *Coalescent.jl* or *TiPS*. However, when the true α value was 2.94 or 9.65, the augmented-coalescent likelihood performed better than the coalescent likelihood (Fig 2 and S5). For trees simulated with *diversitree*, it was clear that coverage was higher when R_0 was low and the true value of α was 0.36 or 2.94. When R_0 was high, coverage was close to zero when the true value of α was 9.65. In general, the augmented-coalescent likelihood also performed better than the coalescent likelihood (Fig 2 and S5).

When estimating ω, coverage was generally close to or equal to 100% when the true ω value was 0.9 for trees simulated using *Coalescent.jl* or *TiPS*. In most cases, the augmented-coalescent likelihood performed better than the coalescent likelihood (Fig 2 and S5). For analyses carried out with trees simulated with *diversitree*, we observed that coverage was close or equal to 100% when using the augmented-coalescent likelihood for the true ω value of 0.1 and independent of R_0 or sample size (Fig 2 and S5). Coverage was lower when the true ω value was 0.9 (Fig 2 and S5). The augmented-coalescent likelihood was, in general, better when ω was small.

3.1.2 Accuracy

For trees simulated with *Coalescent.jl* and *TiPS*, we observed that, in general, the estimates of α were very accurate showing lower values of relative error independent of sample size or true value of α. Note that when α was 2.94 or 9.65, the coalescent likelihood without augmentation produced very inaccurate results, showing relative error that was occasionally greater than 100-fold (Fig 2 and S5). For trees simulated with *diversitree*, accuracy was better for simulations in which the

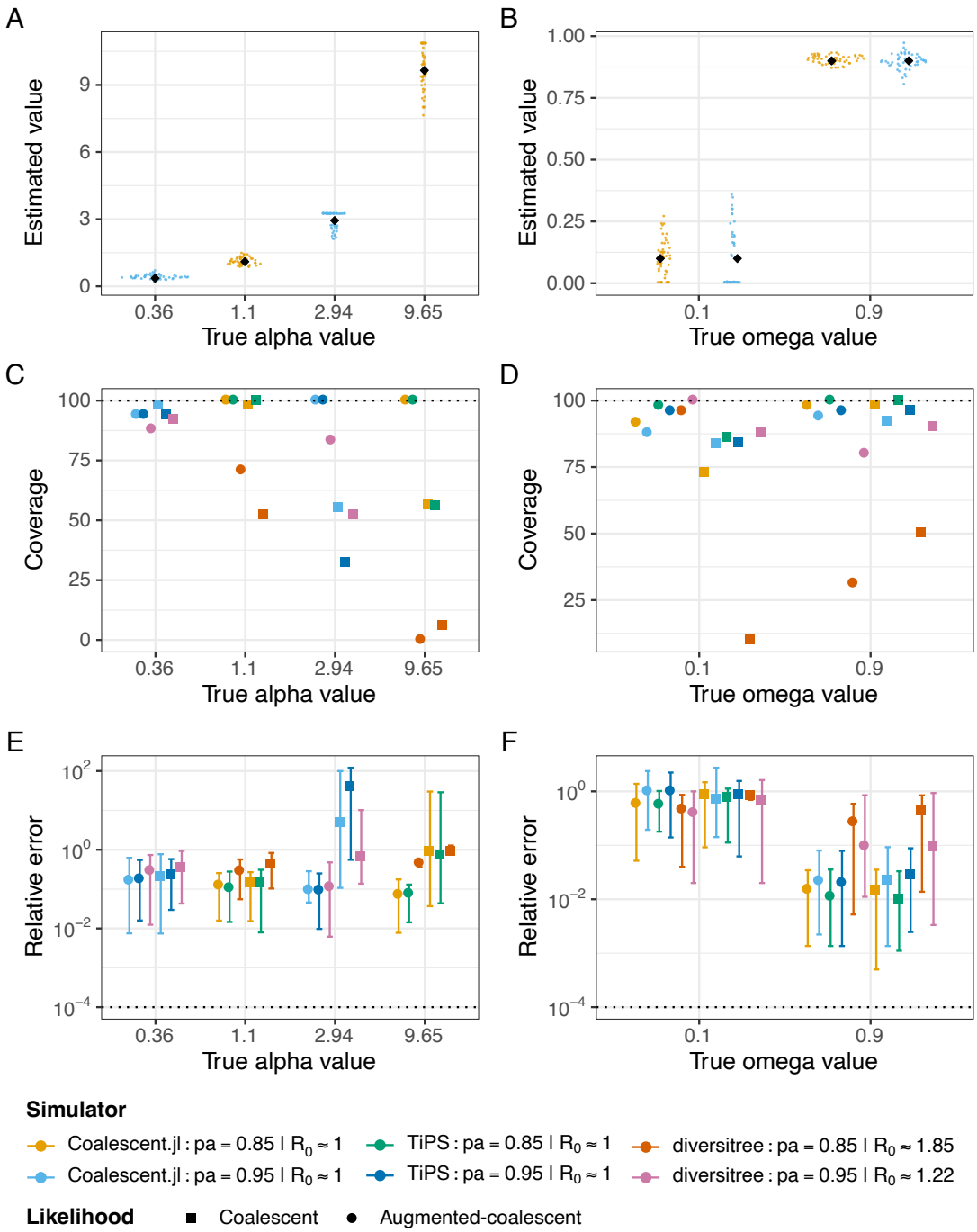

Figure 2: **Simulation results for sample size of 1000 tips**: (A) and (B) Sina plots showing the true and estimated values for α and ω for simulations carried out with *Coalescent.jl* and using the augmented-coalescent likelihood only. Black diamond shapes in the plot are the true values. (C) and (D) Coverage showing the percentage of replicates that contained the true α or ω values within the 95% credible interval for trees simulated with *Coalescent.jl*, *TiPS*, and *diversitree*. (E) and (F) Relative error for α and ω estimates. The relative error (y-axis) represents the computed median and the 0.025 and 0.975 quantiles for all replicates. Depending on the combination of parameter values, the proportion of individuals carrying the ancestral state (p_a) could be 0.85 or 0.95. For the *diversitree* analyses, we grouped the values of R_0 into two groups by averaging the observed R_0 to 1.22 and 1.85 (see supporting information). For the *TiPS*/*Coalescent.jl* analyses, we sampled at equilibrium and R_0 values are expected to be close to 1.0.

average R_0 was 1.22 and independent of sample size. However, accuracy was worse for estimating values of $\alpha = 2.94$ for lower R_0 and using the coalescent likelihood, and this was fixed by using the augmented-coalescent likelihood (Fig 2 and S5).

For trees simulated with *Coalescent.jl* and *TiPS*, we observed that when ω was set to the low value of 0.1, accuracy remained low and was unaffected by sample size. In addition, the use of the augmented-coalescent likelihood did not substantially improve the estimates. However, when ω was set to a high value of 0.9, the estimates remained highly accurate and unaffected by sample size (Fig 2 and S5). For trees simulated with *diversitree*, accuracy was better when the average R_0 was low. Furthermore, accuracy was slightly worse when ω was set to 0.1 (Fig 2 and S5).

Additional plots showing the estimated relative fitness parameters for individual replicates can be found in the supporting information (Fig S6 - S11). These figures span the range of tree simulators, numbers of tree tips, and both likelihood methods.

3.2 Simulation and analysis of sequence alignments

For analyses carried out with combination of parameters 1 (par1; Table S3), we were able to estimate both α and ω using sequence alignments of 2371 bp. The estimated time-scaled trees produced results similar to those obtained with the true trees (Fig 3). Furthermore, for par1 (Table S3), simulations were run for 45 years and the proportion of ancestral state reached 0.806 which was similar to the expected equilibrium value of 0.799 (Table S3).

For analyses carried out with combinations of parameters 2 and 3 (par2 and par3; Table S3), we were unable to estimate α and ω using the true or estimated trees when the TMRCA (time to the most recent common ancestor) was 29 years. When the TMRCA was 45 years, we were able to estimate α using the true trees, but were unable to estimate ω with true or estimated trees (Fig 3). However, when the simulations were run for long enough and the state proportions reached equilibrium, we could estimate both relative fitness α and ω (Fig S4). Furthermore, for par2 and par3 (Table S3), the proportion of ancestral state was 0.865 and 0.822, respectively, suggesting that it did not reach equilibrium (Table S3). We were also unable to recover the true TMRCA and substitution rates with *treedater* using trees reconstructed with alignments of 105 bp (Fig S3).

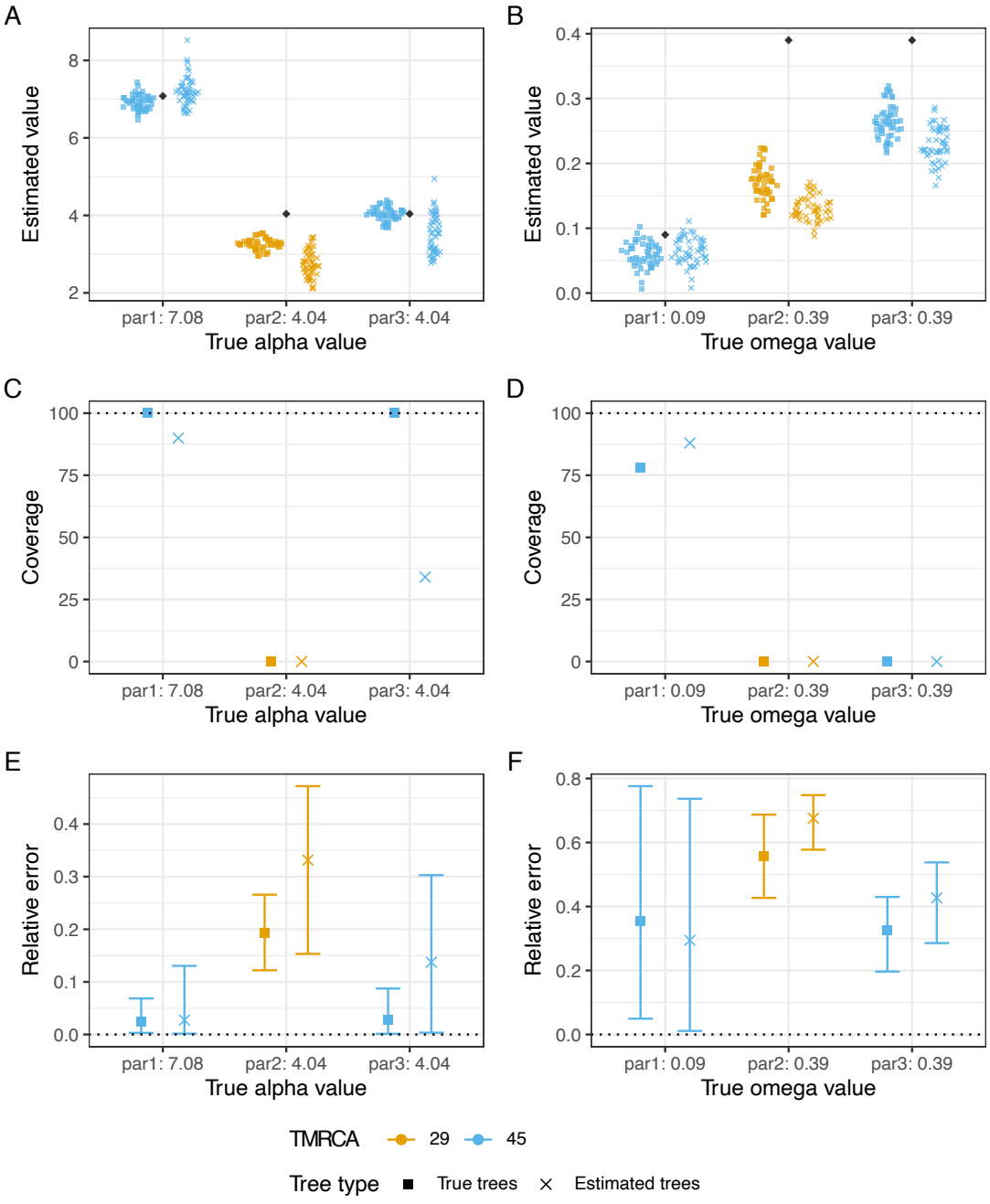

Figure 3: **Performance in simulations with n=4,056**: (A) and (B) Sina plots showing the true and estimated values for α and ω using the augmented-coalescent likelihood. Black diamond shapes in the plot are the true values. (C) and (D) Coverage showing the percentage of replicates that contained the true α or ω values within the 95% credible interval (E) and (F) Relative error for α and ω estimates. The relative error (y-axis) represents the computed median and the 0.025 and 0.975 quantiles for all replicates. All plots shows the estimated values for the true *Coalescent.jl* trees and estimated time-scaled trees. The par1, par2 and par3 refers to the combination of parameter values 1, 2 and 3 (Table S3). Sequence alignment length for par1 was 2371 bp and for par2 and par3 was 105 bp.

3.3 HIV-1 co-receptor tropism

The co-receptor tropism prediction of the subtype C PopART Phylogenetics sequences from Zambia revealed that 20.0% are capable of using the X4 co-receptor, associated with faster disease progression. This analysis showed that: (i) WebPSSM scores consistently split R5- and X4-capable viruses, (ii) there was a slight increase in the proportion of X4/R5X4 and decrease of R5 over the study period, (iii) the pre-treatment CD4 counts were similar for individuals infected by R5 and X4-capable viruses, and (iv) the proportion of X4/R5X4 tends to be higher for individuals in advanced stages of infection (low CD4) when compared with earlier stages (Fig S12).

As expected given high within-host selective pressures for co-receptor switching, the fold-change in within-host replicative fitness (α) was much higher than one (6.87, 95% CI: 6.08–7.76) for the *env* analysis, while the fold-change in between-host replicative fitness (ω) was much lower than one (0.09, 95% CI: 0.04–0.18). These values are consistent with the hypothesis that the within-host relative fitness of variant (X4/R5X4) is substantially higher than the ancestral type (R5), while the between-host fitness of X4/R5X4 is much lower than R5. In other words, the magnitude of the transmission penalty of X4-capable viruses is approximately -90%.

4 Discussion

We have developed a structured coalescent model which describes pathogen evolution of two co-circulating variants with multi-level (within- and between-host) selection. We further develop a maximum-likelihood method for inferring multi-level selection coefficients based on this model and a convenient software implementation. This approach builds on previous work on related structured birth-death models (e.g. the BiSSE model) and adds new capabilities: This model accounts for nonlinear dynamics in effective population size, variable sampling through time, and provides a novel integration of a sampling likelihood with the genealogical coalescent likelihood. In simulations, we find that the derived mutation-selection balance used in combination with the coalescent likelihood provides superior estimates of multi-level selection parameters. We further demonstrate that within-host selection coefficients can be estimated from population-based samples across a large range of values, from minor fitness costs to highly deleterious phenotypes. This provides an original avenue for estimating fitness cost of common deleterious variants from epidemiological

data which has not traditionally been used for this purpose. While our focus has been on evolution of infectious disease pathogens, the computational framework is quite general, and could also be applied to study macro-evolution.

In our simulations, we also demonstrated that when the genetic sequence alignment was sufficiently large and informative (similar to the empirical *env* gene), and the proportion of ancestral state and variant had reached equilibrium, we were able to estimate the relative fitness ω and α (Fig 3). Our empirical analysis of the evolution of HIV co-receptor tropism supports the commonly hypothesised reduction in transmission associated with the X4-tropic virus and contra-indicates the "random transmission" hypothesis [10, 11] according to which both X4 and R5 tropic virus are transmitted equally in proportion to their viral titre. Our study appears to be unique in estimating a specific value for the fitness cost. This analysis, performed using data from a real and high HIV prevalence setting, confirmed that, when compared to R5, the within-host replicative fitness (α) of X4/R5X4 variants was higher (seven-fold increase) and the between-host fitness (ω) was much lower ($\approx -90\%$). This estimate suggests that there are selective forces acting against X4 transmission and provides a magnitude for this transmission penalty [8, 9]. Therefore, the lower frequency of X4 viruses in genital secretions is not the only factor acting against their transmission, although it is still very possible that X4/R5X4 variants will be transmitted if they dominate the donor's virus population, consistent with X4-tropic virus being transmitted at 9% of the rate of R5-tropic virus. The α estimates demonstrate pronounced within-host selection acting on X4/R5X4.

There are several limitations in our study. First, we used a single value for the infection generation time. However, it is possible that for the co-receptor analysis, infections with different co-receptors may lead to different rates of disease progression. Second, current state-of-the-art bioinformatic methods for the classification of co-receptor usage do not distinguish between the dual-receptor R5X4 from the single receptor X4. Consequently, we were unable to test whether dual receptor usage would provide any advantage in replicative fitness compared to X4 usage only.

There is potential for this method to be expanded to incorporate more than two competing variants, such as adaptation driven by differing human leukocyte antigen (HLA) types in a host population. For example, the methods used by Palmer et al. [40] could be used to identify nucleotide sites that are linked to evolutionary pressure from HLA type and treatments that lead to drug resistance mutations (DRM), thereby allowing the definition of multiple phenotypes for analysis

using the model presented here. While in its current form, we expect there to be an issue with statistical-identifiability as the number of parameters in the model is increased, this is an area for future development. Other developments may be possible related to the incorporation of more than one sequence per host, which if available could greatly enhance within-host selection estimates, as well as the incorporation of additional metadata (e.g. structural data or replicative capacity assays) when estimating selection coefficients.

5 Funding

This work was supported by the Wellcome Trust [220885/Z/20/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. KOD, VBF, FFN and EMV acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), funded by the UK Medical Research Council (MRC). This UK funded award is carried out in the frame of the Global Health EDCTP3 Joint Undertaking.

6 Acknowledgements

The authors thank the Phylogenetics and Networks for Generalised HIV Epidemics in Africa (PANGEA) consortium for access to the PopART Zambian HIV-1 dataset. PANGEA is funded by the Bill & Melinda Gates Foundation (consecutive grants OPP1084362 and OPP1175094).

PANGEA consortium members

Christophe Fraser[1], Kate Grabowski[2], Sikhulile Moyo[3], Deenan Pillay[4], Andrew Rambaut[5], Oliver Ratmann[6], Deogratius Ssemwanga[7], Lucie Abeler-Dörner[1], Helen Ayles[8], David Bonsall[1], Rory Bowden[1], Max Essex[9], Sarah Fidler[10], Tanya Golubchik[1], Ravindra Gupta[11], Richard Hayes[12], Joshua Herbeck[13], Joseph Kagaayi[14], Pontiano Kaleebu[7], Jairam Lingappa[13], Vladimir Novitsky[15], Thomas Quinn[2], Janet Seeley[7,16], Frank Tanser[17], Maria Wawer[2], Myron Cohen[18], Vincent Calvez[19], Thumbi Ndung'u[17], Tulio D'Oliveira[20], Ann Dennis[18], Dan Frampton[4], Anne Hoppe[4], Paul Kellam[11],

Cissy Kityo[21], Andrew Leigh-Brown[5], Nick Paton[22].

[1]University of Oxford, Oxford, UK, [2]Johns Hopkins University, Baltimore, USA, [3]Botswana Harvard AIDS Institute Partnership, Botswana, [4]University College London, London, UK, [5]University of Edinburgh, Edinburgh, UK, [6]Imperial College London, London, UK, [7]Medical Research Council/Uganda Virus Research Institute, Entebbe, Uganda, [8]PopART/Zambart, Lusaka, Zambia, [9]Harvard Botswana, Gaborone, Botswana, [10]PopART/Imperial College London, UK, [11]University of Cambridge, Cambridge, UK, [12]PopART/London School of Hygiene and Tropical Medicine, London, UK, [13]University of Washington, Seattle, USA, [14]Rakai Health Sciences Program, Kalisizo, Uganda, [15]Harvard University, Boston, USA [16]London School of Hygiene and Tropical Medicine, London, UK, [17]Africa Health Research Institute, South Africa, [18]University of North Carolina, USA, [19]Institut Pasteur, Paris, FR, [20]University of KwaZulu-Natal, South Africa, [21]EARNEST/The Joint Clinical Research Centre, Kampala, Uganda, [22]EARNEST/University of Singapore, Singapore.

References

[1] Volz E. Fitness, growth and transmissibility of SARS-CoV-2 genetic variants. Nature Reviews Genetics. 2024;24(10):724-34.

[2] Schuitemaker H, Koot M, Kootstra N, Dercksen MW, De Goede R, Van Steenwijk R, et al. Biological phenotype of human immunodeficiency virus type 1 clones at different stages of infection: progression of disease is associated with a shift from monocytotropic to T-cell-tropic virus population. Journal of virology. 1992;66(3):1354-60.

[3] Berger EA, Murphy PM, Farber JM. Chemokine receptors as HIV-1 correceptors: Roles in viral entry, tropism, and disease. Annual Review of Immunology. 1999;17:657-700.

[4] Arif MS, Hunter J, Leda AR, Zukurov JPL, Samer S, Camargo M, et al. Pace of coreceptor tropism switch in HIV-1-infected individuals after recent infection. Journal of virology. 2017;91(19):10-1128.

[5] Brumme ZL, Goodrich J, Mayer HB, Brumme CJ, Henrick BM, Wynhoven B, et al. Molecular

and clinical epidemiology of CXCR4-using HIV-1 in a large population of antiretroviral-naive individuals. The Journal of infectious diseases. 2005;192(3):466-74.

[6] Koot M, Keet IP, Vos AH, de Goede RE, Roos MTL, Coutinho RA, et al. Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS. Annals of internal medicine. 1993;118(9):681-8.

[7] Connor RI, Sheridan KE, Ceradini D, Choe S, Landau NR. Change in coreceptor use correlates with disease progression in HIV-1–infected individuals. The Journal of experimental medicine. 1997;185(4):621-8.

[8] Margolis L, Shattock R. Selective transmission of CCR5-utilizing HIV-1: the'gatekeeper'problem resolved? Nature Reviews Microbiology. 2006;4(4):312-7.

[9] Grivel JC, Shattock RJ, Margolis LB. Selective transmission of R5 HIV-1 variants: where is the gatekeeper? Journal of translational medicine. 2011;9:1-17.

[10] Chalmet K, Dauwe K, Foquet L, Baatz F, Seguin-Devaux C, Van Der Gucht B, et al. Presence of CXCR4-using HIV-1 in patients with recently diagnosed infection: correlates and evidence for transmission. Journal of Infectious Diseases. 2012;205(2):174-84.

[11] Hedskog C, Mild M, Albert J. Transmission of the X4 phenotype of HIV-1: is there evidence against the "random transmission" hypothesis? Journal of Infectious Diseases. 2012;205(2):163-5.

[12] Bloom JD, Neher RA. Fitness effects of mutations to SARS-CoV-2 proteins. Virus Evolution. 2023;9(2):vead055.

[13] Neher RA, Russell CA, Shraiman BI. Predicting evolution from the shape of genealogical trees. Elife. 2014;3:e03568.

[14] Zanini F, Puller V, Brodin J, Albert J, Neher RA. In vivo mutation rates and the landscape of fitness costs of HIV-1. Virus Evolution. 2017;3:vex003.

[15] Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T, et al. Transmission

of Single HIV-1 Genomes and Dynamics of Early Immune Escape Revealed by Ultra-Deep Sequencing. PLoS ONE. 2010;5:e12303.

[16] Maddison WP, Midford PE, Otto SP. Estimating a Binary Character's Effect on Speciation and Extinction. Systematic Biology. 2007;56:701-10.

[17] FitzJohn RG. Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution. 2012;3(6):1084-92.

[18] Yu G, Smith D, Zhu H, Guan Y, Lam TTY. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017;8:28-36.

[19] Volz EM. Complex population dynamics and the coalescent under neutrality. Genetics. 2012;190(1):187-201.

[20] Volz EM, Siveroni I. Bayesian phylodynamic inference with complex models. PLOS Computational Biology. 2018;14(11):e1006546.

[21] Didelot X, Franceschi V, Frost SDW, Dennis A, Volz EM. Model design for nonparametric phylodynamic inference and applications to pathogen surveillance. Virus evolution. 2023;9(1):vead028.

[22] Volz E. Coalescent.jl; 2024. `https://emvolz.github.io/Coalescent.jl/dev/intro/`, visited 2025-02-03.

[23] Danesh G, Saulnier E, Gascuel O, Choisy M, Alizon S. TiPS: Rapidly simulating trajectories and phylogenies from compartmental models. Methods in Ecology and Evolution. 2023;14(2):487-95.

[24] Ly-Trong N, Naser-Khdour S, Lanfear R, Minh BQ. AliSim: A Fast and versatile phylogenetic sequence simulator for the genomic era. Molecula Biology and Evolution. 2022;39(5):msac092.

[25] Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution. 2020;37(5):1530-4.

[26] Hayes RJ, Donnell D, Floyd S, Mandla N, Bwalya J, Sabapathy K, et al. Effect of universal testing and treatment on HIV incidence—HPTN 071 (PopART). New England Journal of Medicine. 2019;381(3):207-18.

[27] Hall M, Golubchik T, Bonsall D, Abeler-Dörner L, Limbada M, Kosloff B, et al. Demographics of sources of HIV-1 transmission in Zambia: a molecular epidemiology analysis in the HPTN 071 PopART study. The Lancet Microbe. 2024;5(1):e62-71.

[28] Struck D, Lawyer G, Ternes AM, Schmit JC, Bercoff DP. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic acids research. 2014;42(18):e144-4.

[29] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013;30(4):772-80.

[30] Los Alamos National Laboratory. HIV-DB Sequence Quality Control Tool; 2024. `https://www.hiv.lanl.gov/content/sequence/QC/index.html`, visited 2025-02-03.

[31] Ratmann O, Wymant C, Colijn C, Danaviah S, Essex M, Frost S, et al. HIV-1 full-genome phylogenetics of generalized epidemics in sub-Saharan Africa: impact of missing nucleotide characters in next-generation sequences. AIDS Research and Human Retroviruses. 2017;33(11):1083-98.

[32] Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature methods. 2017;14(6):587-9.

[33] Tavaré S. Some probabilistic and statistical problems on the analysis of DNA sequence. Lecture of Mathematics for Life Science. 1986;17:57.

[34] Didelot X, Siveroni I, Volz EM. Additive uncorrelated relaxed clock models for the dating of genomic epidemiology phylogenies. Molecular Biology and Evolution. 2021;38(1):307-17.

[35] Volz E, Frost SD. Scalable relaxed clock phylogenetic dating. Virus evolution. 2017;3(2):vex025.

[36] Jensen MA, Li FS, van't Wout AB, Nickle DC, Shriner D, He HX, et al. Improved coreceptor usage prediction and genotypicmonitoring of R5-to-X4 transition by motif analysis of human-

immunodeficiency virus type 1 env V3 Loop sequences. Journal of virology. 2003;77(24):13376-88.

[37] Jensen MA, Coetzer M, Van't Wout AB, Morris L, Mullins JI. A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences. Journal of virology. 2006;80(10):4698-704.

[38] Volz E. musseco: Multi-State Speciation and Extinction Coalescent; 2025. R package version 0.1.0, https://github.com/emvolz-phylodynamics/musseco. Available from: `http://emvolz-phylodynamics.github.io/musseco/`.

[39] Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci. Molecular Biology and Evolution. 2013;30:713-24.

[40] Palmer DS, Turner I, Fidler S, Frater J, Goedhals D, Goulder P, et al. Mapping the drivers of within-host pathogen evolution using massive data sets. Nature Communications. 2019;10.

Supporting information captions

S1 Text. Supporting Information

Data availability statement

The method used here is available as an open source R package *musseco* at `https://github.com/emvolz-phylodynamics/musseco` and there is a tutorial at `https://emvolz-phylodynamics.github.io/musseco`. Scripts used to simulate and analyse data are available at `https://github.com/thednainus/musseco_simulations`. Scripts to analyse empirical data from HIV-1 are available at `https://github.com/vinibfranc/BiSSECo_co-receptor_PANGEA`. Access to the PANGEA PopART HIV-1 sequence dataset is through application. Details are available at `https://www.pangea-hiv.org/join-us`.