Viral Simulation Reveals Overestimation Bias in Within-Host Phylodynamic Migration Rate Estimates Under Selection

Nicolas Ochsner[*†] Judith Bouman[*‡] Timothy Vaughan[§] Tanja Stadler[§]
Sebastian Bonhoeffer[†] Roland Regoes[†]

Dec 2024

[*]These two authors contributed equally to this work
[†]ETH Zurich, Institute of Integrative Biology, Zurich, Switzerland
[‡]University of Bern, Institute of Social and Preventive Medicine, Bern, Switzerland
[§]ETH Zurich, Department of Biosystems Science and Engineering, Basel, Switzerland

Abstract

Phylodynamic methods are widely used to infer the population dynamics of viruses between and within hosts. For HIV-1, these methods have been used to estimate migration rates between different anatomical compartments within a host. The methods typically assume that there is no selective pressure acting on the virus, even though it is known that viruses often experience strong selection pressures. In this study, we investigate how selection affects phylodynamic migration rate estimates. To this end, we developed a novel agent-based simulation tool, `virolution`, to simulate the evolution of virus within two anatomical compartments of a host. Using this tool, we generated viral sequences and genealogies assuming both, neutral evolution and purifying selection that is concordant in both compartments. We found that, under the selection regime, migration rates are significantly overestimated with a stochastic mixture model and a structured coalescent model in the Bayesian inference framework BEAST2. Our results reveal that commonly used phylogeographic methods, which assume neutral evolution, can significantly bias migration rate estimates in selective regimes. This study underscores the need for assessing the robustness of phylodynamic analysis with respect to more realistic selection regimes.

1 Introduction

Phylodynamic methods have transformed pathogen research by enabling the estimation of population dynamic parameters from genetic data alone. These methods leverage the principle that phylogenetic tree topology is shaped by the underlying population dynamics. Phylodynamic approaches simultaneously estimate the population dynamical parameters and the phylogeny from viral sequences. The parameters commonly inferred include population sizes (Ewing et al., 2004), growth rates (Rasmussen et al., 2011), and migration rates between subpopulations (Lemey et al., 2009). This versatility allows phylodynamic methods to address various questions, relating to the between-host transmission dynamics of viruses (e.g. Nadeau et al. (2021)), the evaluation of public health policies (Ratmann et al., 2017), and also the within-host pathogen dynamics, which could lead to better treatment options (Lorenzo-Redondo et al., 2016). However, despite their widespread application, the robustness of these methods for specific tasks under realistic scenarios has rarely been tested. To our knowledge the only paper addressing this thoroughly is Ratmann et al. (2017).

One application of phylodynamic methods focuses on the migration of pathogens between different subpopulations (Suchard and Rambaut, 2009). For example, Lemey et al. (2009) used phylodynamic inference to study the spatial dispersion patterns of Avian influenza A-H5N1 in Eurasia (Lemey et al., 2009).

These phylogeographic approaches can also be applied to investigate the spread and differentiation of viruses between the anatomical compartments within a single infected host (Poss et al., 1998; Salemi and Rife, 2016; Bons and Regoes, 2018). For instance, the brain constitutes a sanctuary site, where the blood-brain barrier limits the exchange of viruses between anatomical compartments (Salemi and Rife, 2016; Nickle et al., 2003; Zárate et al., 2007).

Beyond the brain, other tissues such as the genital tract (Poss et al., 1998), the spleen, the intestinal mucosa, $CD4^+$ T cells and follicular dendritic cells (Salemi and Rife, 2016) have been identified as compartments where the exchange of viruses can be limited. The limited exchange between compartments can lead to genetic differentiation between the viral populations because of local adaptation or genetic drift. To study the exchange / migration of viruses between different anatomical compartments a new field of within-host phylogeographic analysis has emerged, applying phylodynamic methods (Lorenzo-Redondo et al., 2016; Cybis et al., 2013; Ewing et al., 2004;

Chaillon et al., 2014). This new field is sometimes referred to as phyloanatomy (Salemi and Rife, 2016; Lorenzo-Redondo et al., 2016; Shirreff et al., 2013; Bons and Regoes, 2018).

To analyze within-host migration dynamics, phylodynamic methods rely on population genetic models that allow for subpopulations at various locations with migration between them. A variety of such models have been developed, including discrete traits with character mappings (Lemey et al., 2009), isolation-with-migration models (Hey, 2010), structured coalescent models (Kingman, 1982; Vaughan et al., 2014; Müller et al., 2017), and multi-type birth-death models (Kühnert et al., 2016). While each of these models has its own strengths and weaknesses, all of them impose assumptions on the underlying processes that shape the phylogeny. With few exceptions (e.g. Rasmussen and Stadler (2019)), these models assume that both the population dynamic process and the evolutionary process are independent and neutral (unless specifically accounted for).

In contrast to the assumptions, the dynamics of many viruses are subject to strong selection. For instance, in SIV infections Cytotoxic T cells (CTL) have been shown to exert very strong selection pressures (Mandl et al., 2007; Fernandez et al., 2005). HIV in humans is under weaker, but still strong, selection pressure from CTL (Asquith et al., 2006). Additionally, there is evidence for purifying and positive selection (Zanini et al., 2017; Edwards et al., 2006). This is further supported by evidence for parallel evolution in HIV infected individuals (Keele et al., 2008; Bertels et al., 2017) and even in cell culture (Bertels et al., 2019a) in the absence of the immune system. The *in vivo* HIV sequence data of Keele et al. (2008) and its patterns of parallel evolution (Bertels et al., 2017) have been used to infer the mutational fitness effect distribution (MFED) of the *env* gene. Specifically, they found that 4.5% of mutations are beneficial (Bons et al., 2018). Thus, HIV dynamics is subject to strong purifying and diversifying selection.

Even though non-neutral evolution is common for HIV-1 and other pathogens, it is not accounted for in phylodynamic methods that are applied to within-host dynamics (Roje, 2014). Some studies do mention that selection could potentially affect their results, but argue that it is either compensated for by using several different genes to build the genealogy or captured by the effective populations size (Bedford et al., 2011). Roje (2014) and Castoe et al. (2009) illustrate that including multiple genes into the analysis can help, but that phylogenetic techniques are, despite this multiplicity, vulnerable to selection (Roje, 2014; Castoe et al., 2009). Another method that is frequently used to counterbalance the effect of selection allows for varying mutation rates along

tree branches, i.e. an uncorrelated log-normal relaxed molecular clock model (Drummond et al., 2006; Chaillon et al., 2014). However, we are not aware of any studies that tested to what extent this model actually accounts for the effects of non-neutral selection.

In this study, we test the effect of selection on phylodynamic inference, in particular on migration rate estimates between two within-host compartments. For this purpose, we simulate an infection, mutation and replication cycle within a host using a detailed agent-based simulation. The simulation relies on the mutational fitness effects distribution (MFED) estimated from the HIV-1 sequence data of Keele et al. (2008) (Bons et al., 2018; Keele et al., 2008; Lee et al., 2009) to account for the effects of selection. We simulate the evolution of a viral genome in two compartments with migration. We assume that the MFEDs in the two compartments are identical. The sequences and genealogies obtained from the simulation are then analyzed with BEAST2 using DTA as in Lemey et al. (2009) and MASCOT as in Müller et al. (2018). Our results indicate that selection leads to an overestimation of migration rates and that the relaxed molecular clock model — which has been suggested as a potential solution to account for the effects of selection (Drummond et al., 2006) — is particularly susceptible to this effect.

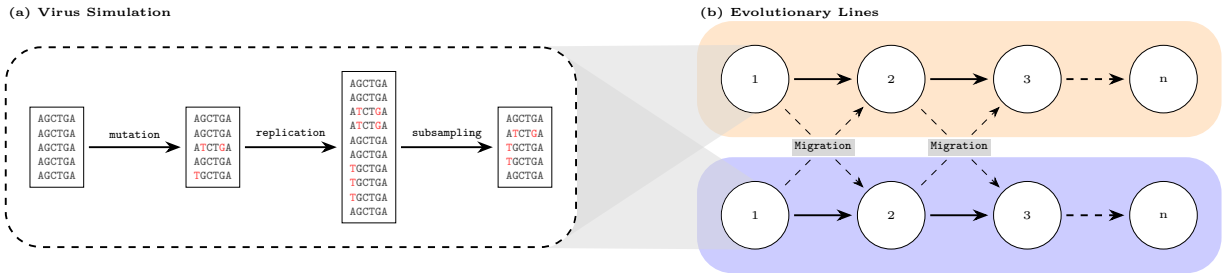

Figure 1: **Schematic representation of the simulation.** (a) shows the simulation of a single generation of viral evolution. The virus mutates, replicates (based on the reproductive fitness), and is sub-sampled. (b) shows the evolution of two compartments over multiple generations. There is migration between the compartments.

2 Results

Simulation of Within-Host Viral Evolution in Two Compartments with Selection

We simulated the evolution of viral populations using a newly-developed agent-based simulation tool (‘virolution‘). The simulation features viral genomes (agents) that replicate and mutate in cells in two distinct anatomical compartments. The genomes are assumed to migrate between the two compartments at the same rates.

We considered two selection scenarios: neutral evolution as the baseline, and a selection scenario in which the replicative fitness differs between genotypes. The fitness differences are calculated on the basis of an empirically-supported mutational fitness effects distribution (MFED) for HIV-1 (Bons et al., 2018). We assumed that the MFED is identical in both compartments.

The output of these simulations are the sequences and genealogy of the viral genomes. In our simulations, we assumed a population size of 100 or 1000 virions per compartment, to accommodate the computational requirements of the BEAST2 analysis, and varied the migration rate from 0 to 0.01 (which corresponds to a migration of up to 1% of the virions per generation). We simulated the evolution of the virus for 1000 generations. A detailed description of the simulation can be found in Section 4.

We then used DTA and MASCOT in BEAST2 to estimate the migration rate for all simulated scenarios from a fixed number of samples. Each method was applied in two ways: once with the true genealogy from the simulation and once with the sequence data, where the genealogy was reconstructed during the analysis. Finally, we compared these estimates with the true migration

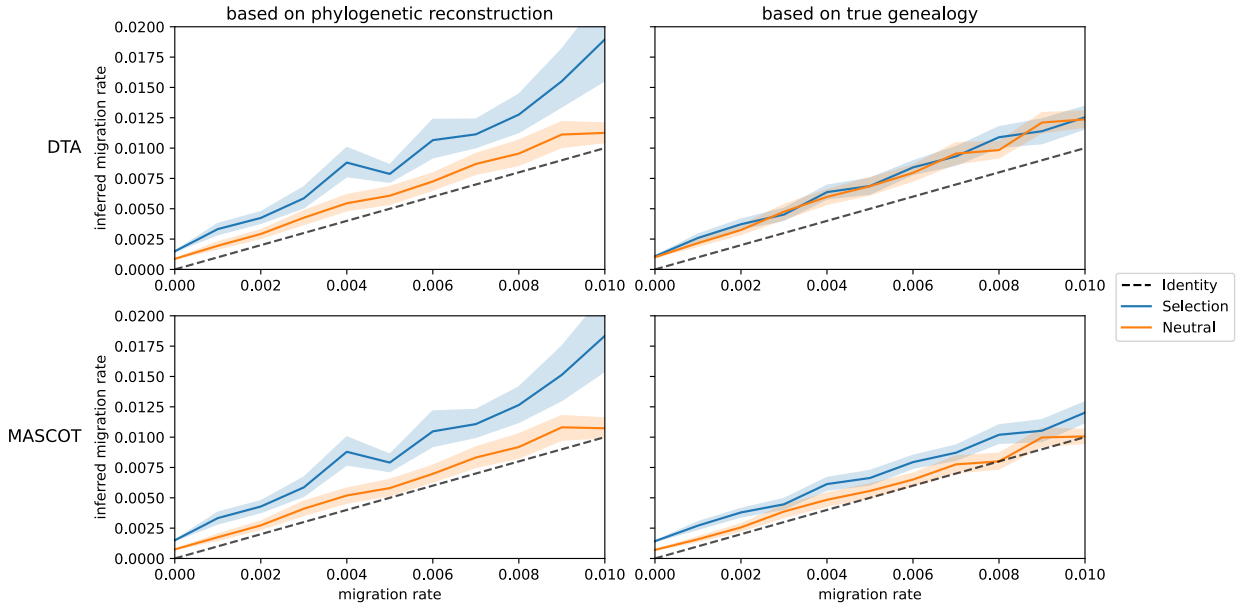

Figure 2: **Selection Bias in Phylogenetic Estimates of HIV-1 Migration Rates Using Relaxed Clock Models.** Average mean migration rate estimate with DTA and MASCOT in BEAST2. For each migration rate, 50 simulations were conducted, maintaining a constant population of 100 virions in each compartment, with a mutation rate of 2.16×10^{-5}, and spanning a total of 1,000 generations. All sequences were sampled and analyzed in BEAST2 with a relaxed clock and fixed effective population size. The left figure illustrates the inference while sampling trees, while the right figure displays the migration rate estimates based on the true–simulated–genealogy. The lighter colored bands illustrate the 95% confidence interval.

rate, i.e. the migration rate that has been used in the simulation. We expected a bias in the migration rate estimates with selection that is confounded within phylogenetic inference.

Selection-Driven Overestimation of Migration Rates in Phylogenetic Reconstructions

In both phylogeographic methods, DTA and MASCOT, we find evidence of structural overestimation biases in the inferred migration rates. In Figure 2, we show the inferred migration rates based on the phylogenetic reconstruction and the true genealogy. The estimates are inferred from simulations with a population size of 100 virions per compartment and a mutation rate of 2.16×10^{-5} (estimated for HIV-1). In many of the scenarios with and without selection, the migration rates are significantly overestimated (see Table S2). We therefore distinguish between two types of biases: biases that affect both regimes (with / without selection) and biases that only affect the selection regime.

			migration rate	0e+00	1e-03	2e-03	3e-03	4e-03	5e-03	6e-03	7e-03	8e-03	9e-03	1e-02
population size	**mutation rate**	**method**	**tree**											
100	2.16e-5	DTA	recon	0.69	0.67	0.64	0.62	0.67	0.62	0.64	0.62	0.63	0.59	0.66
			true	0.53	0.57	0.56	0.50	0.54	0.49	0.54	0.49	0.56	0.46	0.51
		MASCOT	recon	0.73	0.71	0.67	0.63	0.70	0.65	0.66	0.64	0.65	0.60	0.68
			true	0.74	0.70	0.67	0.58	0.63	0.59	0.62	0.58	0.65	0.52	0.61
	2.16e-4	DTA	recon	0.71	0.76	0.72	0.74	0.75	0.70	0.73	0.71	0.65	0.72	0.66
			true	0.59	0.65	0.58	0.60	0.58	0.56	0.56	0.55	0.47	0.54	0.50
		MASCOT	recon	0.74	0.77	0.73	0.75	0.76	0.71	0.74	0.72	0.66	0.73	0.68
			true	0.74	0.77	0.74	0.74	0.75	0.71	0.74	0.71	0.67	0.75	0.69
1000	2.16e-5	DTA	recon	0.77	0.75	0.75	0.80	0.75	0.77	0.81	0.81	0.74	0.76	0.72
			true	0.51	0.55	0.51	0.53	0.57	0.48	0.57	0.45	0.46	0.41	0.44
		MASCOT	recon	0.76	0.75	0.76	0.81	0.76	0.78	0.81	0.82	0.75	0.77	0.73
			true	0.71	0.61	0.67	0.65	0.67	0.67	0.66	0.59	0.54	0.51	0.37
	2.16e-4	DTA	recon	0.87	0.82	0.85	0.86	0.81	0.89	0.86	0.82	0.83	0.84	0.83
			true	0.62	0.66	0.66	0.68	0.63	0.75	0.66	0.58	0.65	0.57	0.59
		MASCOT	recon	0.86	0.81	0.86	0.86	0.82	0.89	0.86	0.83	0.83	0.84	0.85
			true	0.83	0.79	0.78	0.83	0.72	0.86	0.78	0.71	0.69	0.70	0.68

Figure 3: **Probability of Overestimation of Migration Rates from Selection with Relaxed Clock Model.** This table shows the average posterior probability of overestimation for each migration rate corrected for the neutral baseline. The average is calculated over 50 paired simulations where the posterior probability of overestimation is derived from 1000 migration rate samples. The values are colored according to the posterior probability, with red indicating a high probability of overestimation and blue indicating a high probability of underestimation.

In both regimes the compartments are initiated with the same wildtype sequence. This choice is motivated by the initial dynamics of an HIV-1 infection, where the majority of infections are initiated by a single genotype (Keele et al., 2008). During phylogenetic reconstruction, the genealogies of the compartments still need to be connected by some migration event, hence causing a structural bias in absence of migration. This effect — caused by a discrepancy between the inference and simulation model — does not limit itself to the absence of migration, but maintains a consistent bias through all migration rates (see Figure 2).

Additionally, we find that the estimates based on the phylogenetic reconstruction are increasingly overestimated in the presence of selection (see Figure 2). In Table 3, we show the average posterior probability of overestimation caused by selection for each migration rate for 50 simulations. The posterior probability of overestimation caused by selection is calculated as the proportion of samples where the inferred migration rate of a simulation with selection is higher than a sample

from a neutral simulation (see Figure S10 for the distributions). Values higher than 0.5 indicate that the migration rate is more likely to be overestimated in the presence of selection than estimates without selection.

We find that for the simulation parameters used in Figure 2, selection introduces a significant overestimation bias to the migration rate estimates based on the phylogenetic reconstruction, which is reflected in the average posterior probability of overestimation averaging at 0.642 for DTA and 0.666 for MASCOT.

While this effect does not limit itself to the estimates based on the phylogenetic reconstruction, it is more pronounced in these compared to the estimates based on the true genealogy, where the average posterior probability of overestimation is 0.524 for DTA and 0.628 for MASCOT.

Indeed, during phylogenetic reconstruction, the evolutionary model assumes neutral evolution. Therefore, mutations that have occurred independently in both compartments can result in incorrect phylogenetic reconstructions, which will consequently be misinterpreted as migration events.

Increased Bias from Selection in Relaxed Clock Models

Relaxed clock models have been suggested to be more robust and potentially counteract the effect of selection (Drummond et al., 2006; Chaillon et al., 2014). Here, we cannot confirm this. The migration rate estimates shown in Figure 2 have been obtained using a relaxed clock model. In fact, when using a strict clock model, we find less evidence for structural overestimation biases for the same simulated data as the posterior probability of overestimation is lower for both DTA (0.566) and MASCOT (0.631) (see Figures S4, S10, S12, and Table S1). Evidently, the flexibility of the relaxed clock model does not counteract, but instead reinforce the effect of selection on the migration rates.

Elevated Mutation Rates Amplify Overestimation Bias in Migration Inference with Selection

We also investigated the robustness of the inference to changes in mutation rate. In Figure 4, we show the results of simulations with a 10-fold higher than empirical estimates. Without selection, there is no significant difference in the migration rate estimates to the simulation with lower mutation rate (2.16×10^{-5}). However, when affected by selection, we observe that the overestimation

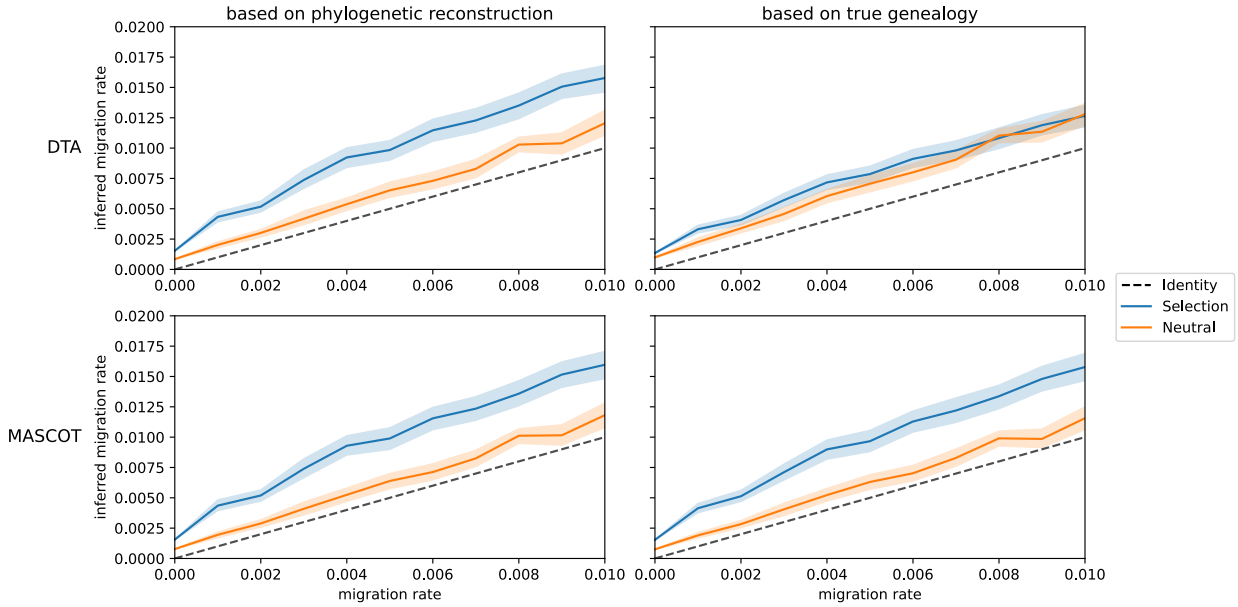

Figure 4: **Increased Overestimation of HIV-1 Migration Rates with Increased Mutation Rates with Relaxed Clock.** Average mean migration rate estimate with DTA and MASCOT in BEAST2. For each migration rate, 50 simulations were conducted, maintaining a constant population of 100 virions in each compartment, with a mutation rate of 2.16×10^{-4}, and spanning a total of 1,000 generations. All sequences were sampled and analyzed in BEAST2 with a relaxed clock and fixed effective population size. The left figure illustrates the inference while sampling trees, while the right figure displays the migration rate estimates based on the true–simulated–genealogy. The lighter colored bands illustrate the 95% confidence interval.

bias increases in both methods as mutation rate increases with the relaxed clock model.

Here, both the inference based on the phylogenetic reconstructions and the true genealogy are affected. Specifically, we start to see a strong bias in MASCOT even when the tree is fixed to the true genealogy. This observation suggests that the complex interplay of selection effects in the evolutionary and population dynamic process can negatively affect the inference of population dynamic parameters beyond the phylogenetic reconstruction.

It should be noted that the estimates obtained with the strict clock do not show the same increase in overestimation bias due to selection with the increased mutation rate (Figure S5). However, using the strict clock model, estimates will show a general trend towards underestimation with higher migration rates, which is not observed with the relaxed clock model (see Table S3).

Increased Population Size Breaks Accuracy of Phylodynamic Inference

Given that the overestimation bias was strongly affected by the mutation rate, it is expected that an increase in population size would have a similar effect on the estimates of migration rates. Figures S6 and S7 show the results of simulations with a population size of 1000 virions per compartment. When compared to the migration rate estimates for simulations with a smaller population size of 100 virions per compartment, both methods performed significantly worse.

Figure S6 shows inference on simulations with a mutation rate of 2.16×10^{-5} and a population size of 1000 virions per compartment. In this scenario, there is a similar supply of mutations as in the simulations with a smaller population size, but higher mutation rate. Therefore, one might expect the estimates to show similar biases. However, this is not the case. In fact, we observe that the estimates based on the phylogenetic reconstruction perform better than the estimates based on the true genealogy for both, DTA and MASCOT (Figures S6 and S8).

This finding contrasts sharply with the results from simulations with smaller population size of 100 virions per compartment, where estimates based on the true genealogy outperformed those based on the phylogenetic reconstruction (see Figures 2 and 4, and Table S2). This suggests that the increased population size and the consequent need for sub-sampling are major factors contributing to the decreased accuracy of the estimates.

Indeed, sub-sampling can reduce the representativeness of the sample and affect the inferred phylogenetic relationships, leading to increased uncertainty in the estimates. Specifically, the decreased performance of both methods on the true genealogy compared to the phylogenetic reconstruction may be explained by the sampling bias introduced when sub-sampling. In this case, when phylogenetic uncertainty is not accounted for, the sampling bias may directly translate into decreased accuracy of the estimates. This could be supporting evidence, that the ability to account for phylogenetic uncertainty is crucial for accurate phylodynamic inference.

Overall, the conflicting results and variety in biases between different simulation parameters, inference methods, and clock models, do not permit conclusive explanations for all observed biases and more detailed investigations are necessary to understand when phylodynamic inference results are reliable.

3 Discussion

In this study, we have investigated the performance of phylogeographic methods, DTA and MASCOT, to estimate migration rates between anatomical compartments with realistic selection effects. Although phylodynamic methods have been used to study the compartmentalization of HIV-1 (Chaillon et al., 2014; Lorenzo-Redondo et al., 2016), a systematic evaluation of the methods specific to the task of estimating within-host migration rates has not been performed. There is a discrepancy between the assumptions of the inference models and characteristics of the within-host dynamic of HIV: while the inference models, especially the phylogenetic reconstruction on which they rely, assume neutral evolution, HIV-1 is subject to strong selection pressures within its host. To study how this discrepancy affects phylodynamic analysis, we have created a simulation of HIV-1 diversification in two within-host compartments, in which we assume concordant selection. Our results show that both methods are sensitive to selection and parallel evolution, leading to an overestimation of the migration rates.

In scenarios in which migration occurs, we found that migration rates are overestimated under selection. When selection causes parallel evolution mutations may independently arise and fixate in both compartments. Such instances of parallel evolution are not accounted for in the model, and will be wrongly attributed to migration events by the inference algorithm. This will cause migration rates to be overestimated. Additionally, we have found that the relaxed clock model (Drummond et al., 2006), which is frequently used to account for selection effects (Chaillon et al., 2014), does not decrease the bias, but instead increases it for both methods, DTA and MASCOT. We consider this overestimation bias that is caused by selection to be a major concern for phylodynamic analysis. In our analysis, we have found an overestimation bias that is caused by the specific implementation of selection: we assumed identical purifying selection in both compartments. However, other forms of selection — e.g. divergent evolution — may also lead to biases in the estimation of migration rates, not only overestimation biases (Bons and Regoes, 2018).

Moreover, we have observed a consistent overestimation bias for migration rates at zero, i.e. in a scenario in which migration does not occur. This bias is caused by a discrepancy between the model used for the inference and the one we used for the simulation. In our simulation, the initial population in both compartments is identical, a modeling choice motivated by the initial dynamics

of HIV-1 infection where the majority of infections are initiated by a single genotype (Keele et al., 2008). Here, even in absence of migration, the two compartments initiated with the same ancestor will be connected through the genealogy. This cannot be accurately modeled by any of the two methods that we have tested in this study, however may be addressed by other population dynamic models, such as Isolation-with-Migration models (Hey, 2010). We do not consider this to be a major concern, as the methods are designed to estimate migration rates, and not to detect the absence of migration. However, it is important to note, that even when data does not violate the assumption of neutral evolution, such effects can still create a significant structural bias on inference of migration rates.

In our phylodynamic analyses of the simulated data, rather than estimating all parameters, we have fixed all population biological and evolutionary parameters, except for the migration rate. While this strategy is difficult to apply to real world data because most of these parameters have not been reliably estimated independently, it was chosen to obtain tight migration rate estimates and to isolate the effect of selection and phylogenetic reconstruction on the estimate. Interestingly, we found that restraining the inference to the true genealogy does not necessarily improve the performance of the methods. In Section 2, both methods fail to recover the true migration rate if we fix the genealogy, while maintaining better performance when the genealogy is inferred. This suggests that accounting for phylogenetic uncertainty may be necessary to avoid biases in sampling to translate into the derived estimates. And while not having to know the genealogy *a priori* for reliable migration estimation may sound like good news, it adds a note of caution to how prior knowledge affects phylodynamic inference.

In this study, we assessed phylodynamic inference using a detailed simulation model that explicitly incorporates details on the replication, diversification, and selection of viruses and outputs genetic sequences. This approach is rarely adopted. More commonly, phylodynamic methods are tested by simulating trees using the exact same models that are inverted for the inference (see for example Stadler (2011); Louca and Doebeli (2018); Vaughan and Drummond (2013)). Our study clearly shows that the common approach falls short in capturing potential structural biases that arise from more realistic settings (that, for example, feature selection) and simulations that depart from the inference assumptions in meaningful ways are needed. To our knowledge, only one study has taken a similar approach based on a detailed simulation of an HIV-1 epidemic to evaluate

phylodynamic methods (Ratmann et al., 2017), also identifying significant estimating biases.

Phylodynamic methods that are based on a posterior distribution largely rely on Markov chain Monte Carlo (MCMC) algorithms to explore the parameter space. While this model generally allows us to characterize complex, even non-convex spaces (such as tree space), it requires relatively large computation times to achieve convergence. To allow an analysis with DTA and MASCOT in BEAST2, we had to restrict the number viral sequences we used for inference. For example, when approaching larger population sizes ($N >= 1000$), we had to resort to subsampling the data. The performance of the phylodynamic methods we have tested degraded as we increased the population sizes beyond 1000 until eventually, we were unable to recover the migration rates. Additionally, to achieve convergence, we were forced to limit the biological realism of the simulation: we chose to limit the number of compartments, the maximum population size, and excluded processes such as coinfection and recombination. While the focus of the present study was to investigate the influence of selection on phylodynamics inference, our simulation can integrate a wide range of complex, biologically relevant mechanisms. A thorough evaluation of how these mechanisms impact phylodynamic inference is a promising avenue for future research.

In conclusion, we have established biases of phylodynamic inference in selective regimes. While we specifically focused on the estimation of migration rates between compartments within HIV-infected hosts we argue that these results also apply in epidemiological setting in which selection is also known to operate. Our study illustrates the value of detailed, empirically-supported simulations that are not necessarily consistent with the assumptions of the inference method for its assessment.

4 Methods

Data Simulation

Evolution and migration of virions is simulated using the tool ‘virolution‘. ‘virolution‘ is a simulation tool written in the programming language Rust. It is designed to simulate large virus populations in a multitude of complex scenarios. Within its simulation loop, ‘virolution‘ will simulate the infection of host cells by virions, the replication and mutation of virions, and the sampling process that determines which virions will be present in the next generation. ‘virolution‘ is open-source and available on GitHub: https://github.com/sirno/virolution.

Here, we have used ‘virolution‘ to simulate the evolution within two compartments. The compartments start with identical populations of virions, which evolve largely independent of each other. The rate of replication of the virions is determined by the fitness associated with their genomic sequence. To keep the population size constant, the virions are sub-sampled at the end of each generation. We simulated migration between the compartments by exchanging a fraction of the virions at each generation.

Figure S19 and S20 show the genetic distance between the two compartments for selected parameters and illustrate the compartmentalized evolution of the virions in the simulation under different conditions.

For the analysis in this study, we have varied the parameters that the simulation is based on:

Mutation rates and substitution model All simulations use the same GTR substitution model with a fixed substitution matrix. However, we ran simulations with two different mutation rates, one which resembles a typical mutation rate for HIV-1 (2.16×10^{-5}) and one that maintains a 10-fold higher mutation rate (2.16×10^{-4}). The slightly higher mutation rate, was intended to increase the evolutionary signal in the trees and data, and therefore potentially improve the predictive power of phylodynamic methods.

Selection based on Mutational Fitness Effects Distribution Selection is simulated using variation in reproductive fitness of the virions. We simulate reproduction by sampling from a Poisson-distribution with $\lambda = 6$, which is the basic reproduction rate of the wild-type virus. We

distinguish between simulations with and without selection effects. When selection effects are enabled, we use a mutation fitness effects distribution (MFED) that has been estimated for HIV-1 in Bons et al. (2018). The MFED is a piece-wise log-normal distribution, with 4.5% lethal mutations, and otherwise log-normal distributed fitness values ($\mu = -0.248$, $\sigma = 0.148$). From this distribution, we assign a specific fitness value to any possible mutation and then multiply the fitness values of all mutations to obtain the reproductive fitness of a genotype. The reproductive fitness is then multiplied by the basic reproduction rate to obtain the number of viable virions for a specific genotype. The variation in reproductive fitness with subsequent sub-sampling will cause selective pressure on the virions.

In this study, we have used the same assignment of fitness values to mutations for both compartments.

Fixed constant population sizes We enforce fixed population sizes in each compartment by sub-sampling each compartment after reproduction. We simulate 100 and 1000 virions per compartment.

Migration between compartments After sub-sampling, virions can migrate between the two compartments at rates $0, 0.001, 0.002, \ldots, 0.01$. We choose the number of migrants by multiplying the number of virions in the compartment by the migration rate. We exchange the integer value of this product and add the result of a Bernoulli-distributed variable based on the remainder.

Repeated runs to account for stochastic variation For each combination of settings that has been described in this section, we run the simulation a 50 times, leading to a total of 8400 independent simulations. Each simulation will produce a full sample of all sequences that are present after 1000 generation, including the exact genealogy of the final sequences as they evolved during the simulation.

Phylodynamic analysis

To infer migration rates from the simulated data, we use DTA as described in Lemey et al. (2009) and MASCOT as described in Müller et al. (2018). For both analyses we fix the population size and migration rate to the values that were used for simulation. Further, values of the GTR substitution

model were fixed to the known rates that were used during the simulations. For DTA we set the non-zero rate prior to a Poisson-distribution with $\lambda = 0.693$ and the relative geo rates prior to a Gamma-distribution with $\alpha = 1$ and $\beta = 1$. For MASCOT no informative priors were set during the analysis.

The inference was done based either on the sequences (where phylogenetic reconstruction is necessary) or the true genealogy. In both cases, we used 100 individuals from each compartment, which in the case of the larger population size of 1000 individuals per compartment were chosen randomly.

All inference runs used a maximum chain length of 10^8. The resulting ESS values all achieved values well above 1000.

Evaluation of Structured Biases

We distinguish between two types of biases that we have observed in the inferred migration rates: (1) biases that are confounded by the simulation design and (2) biases that are related to selection effects.

To evaluate the first type of bias we compared the mean inferred migration rates of 50 simulations to the true migration rates. To this end, we calculated the posterior probability of overestimation for each posterior sample after an initial burn-in of 10% of the samples, counting the number of samples that overestimate the true migration rate.

For the second type of bias, we compared the mean inferred migration rates between 50 simulations with and without selection effects. Here, we calculated the posterior probability of overestimation for pairs of posterior samples after an initial burn-in of 10% of the samples. For each pair of samples, we counted the number of samples, where the sample with selection is larger than the sample without selection.

Additionally, error bars on the figures represent the 95% confidence interval of the mean inferred migration rates.

Distance Between Compartments

To quantify the genetic distance between the two compartments in the model, the following relation is used:

$$D = \sum_{k=1}^{n} \sum_{p \in \{A,T,C,U\}} |f_{kp}^{1} - f_{kp}^{2}|. \tag{1}$$

In this equation, f_{kp}^{i} stands for the frequency of nucleotide p at position k in compartment i and n is the sequence length. Practically, the distance is thus defined as the sum over all absolute differences in nucleotide frequencies over all locations in the genotype. The definition is from Bertels et al. (2019b). The distance is calculated within the 'virolution' simulation tool.

Code Availability

All scripts and software elements necessary to reproduce our experiments are available online. An overview of the code repositories and data can be found at `github.com/sirno/overestimation_data`. The simulation tool 'virolution' is available at `github.com/sirno/virolution`, for the simulations in this study we used the version tagged as 'overestimationspaper'. Any common code used for the analysis and dispatch in this study is available at `github.com/sirno/phynalysis`.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgement

This work was supported by the Swiss National Science Foundation (SNSF) under grant number 179170. Additionally, we would like to thank Eva Bons for her initial conception of a Python-based simulation that our simulation is based on and Hannelore Macdonald for providing valuable feedback on the manuscript.

References

Greg Ewing, Geoff Nicholls, and Allen Rodrigo. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics*, 168(4):2407–2420, 2004.

David A. Rasmussen, Oliver Ratmann, and Katia Koelle. Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series. *PLOS Computational Biology*, 7(8):e1002136, August 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002136. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002136. Publisher: Public Library of Science.

Philippe Lemey, Andrew Rambaut, Alexei J Drummond, and Marc A Suchard. Bayesian phylogeography finds its roots. *PLoS computational biology*, 5(9):e1000520, 2009.

Sarah A. Nadeau, Timothy G. Vaughan, Jérémie Scire, Jana S. Huisman, and Tanja Stadler. The origin and early spread of SARS-CoV-2 in Europe. *Proceedings of the National Academy of Sciences*, 118(9):e2012008118, March 2021. doi: 10.1073/pnas.2012008118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2012008118. Publisher: Proceedings of the National Academy of Sciences.

Oliver Ratmann, Emma B. Hodcroft, Michael Pickles, Anne Cori, Matthew Hall, Samantha Lycett, Caroline Colijn, Bethany Dearlove, Xavier Didelot, Simon Frost, A.S. Md Mukarram Hossain, Jeffrey B. Joy, Michelle Kendall, Denise Kühnert, Gabriel E. Leventhal, Richard Liang, Giacomo Plazzotta, Art F.Y. Poon, David A. Rasmussen, Tanja Stadler, Erik Volz, Caroline Weis, Andrew J. Leigh Brown, Christophe Fraser, and on behalf of the PANGEA-HIV Consortium. Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison. *Molecular Biology and Evolution*, 34(1):185–203, January 2017. ISSN 0737-4038. doi: 10.1093/molbev/msw217. URL https://doi.org/10.1093/molbev/msw217.

Ramon Lorenzo-Redondo, Helen R Fryer, Trevor Bedford, Eun-Young Kim, John Archer, Sergei L Kosakovsky Pond, Yoon-Seok Chung, Sudhir Penugonda, Jeffrey G Chipman, Courtney V

Fletcher, et al. Persistent hiv-1 replication maintains the tissue reservoir during therapy. *Nature*, 530(7588):51, 2016.

Marc A. Suchard and Andrew Rambaut. Many-core algorithms for statistical phylogenetics. *Bioinformatics*, 25(11):1370–1376, June 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp244. URL https://doi.org/10.1093/bioinformatics/btp244.

Mary Poss, Allen G Rodrigo, John J Gosink, Gerald H Learn, Dana de Vange Panteleeff, Harold L Martin Jr, Job Bwayo, Joan K Kreiss, and Julie Overbaugh. Evolution of envelope sequences from the genital tract and peripheral blood of women infected with clade a human immunodeficiency virus type 1. *Journal of virology*, 72(10):8240–8251, 1998.

Marco Salemi and Brittany Rife. Phylogenetics and phyloanatomy of hiv/siv intra-host compartments and reservoirs: the key role of the central nervous system. *Current HIV research*, 14(2): 110–120, 2016.

Eva Bons and Roland R Regoes. Virus dynamics and phyloanatomy: Merging population dynamic and phylogenetic approaches. *Immunological reviews*, 285(1):134–146, 2018.

David C Nickle, Mark A Jensen, Daniel Shriner, Scott J Brodie, Lisa M Frenkel, John E Mittler, and James I Mullins. Evolutionary indicators of human immunodeficiency virus type 1 reservoirs and compartments. *Journal of virology*, 77(9):5540–5546, 2003.

Selene Zárate, Sergei L Kosakovsky Pond, Paul Shapshak, and Simon DW Frost. Comparative study of methods for detecting sequence compartmentalization in human immunodeficiency virus type 1. *Journal of virology*, 81(12):6643–6651, 2007.

Gabriela B Cybis, Janet S Sinsheimer, Philippe Lemey, and Marc A Suchard. Graph hierarchies for phylogeography. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368 (1614):20120206, 2013.

Antoine Chaillon, Sara Gianella, Joel O Wertheim, Douglas D Richman, Sanjay R Mehta, and Davey M Smith. Hiv migration between blood and cerebrospinal fluid or semen over time. *The Journal of infectious diseases*, 209(10):1642–1652, 2014.

G Shirreff, V Garcia, TH Vanderford, G Silvestri, and RR Regoes. The phyloanatomy of early siv infection. *Mathematical and Computational Evolutionary Biology*, pages 27–31, 2013.

Jody Hey. Isolation with Migration Models for More Than Two Populations. *Molecular Biology and Evolution*, 27(4):905–920, April 2010. ISSN 0737-4038. doi: 10.1093/molbev/msp296. URL https://doi.org/10.1093/molbev/msp296.

John FC Kingman. On the genealogy of large populations. *Journal of applied probability*, 19(A): 27–43, 1982.

Timothy G Vaughan, Denise Kühnert, Alex Popinga, David Welch, and Alexei J Drummond. Efficient bayesian inference under the structured coalescent. *Bioinformatics*, 30(16):2272–2279, 2014.

Nicola F Müller, David A Rasmussen, and Tanja Stadler. The structured coalescent and its approximations. *Molecular biology and evolution*, 34(11):2970–2981, 2017.

Denise Kühnert, Tanja Stadler, Timothy G Vaughan, and Alexei J Drummond. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Molecular biology and evolution*, 33(8):2102–2116, 2016.

David A Rasmussen and Tanja Stadler. Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. *Elife*, 8:e45562, 2019.

Judith N Mandl, Roland R Regoes, David A Garber, and Mark B Feinberg. Estimating the effectiveness of simian immunodeficiency virus-specific cd8+ t cells from the dynamics of viral immune escape. *Journal of virology*, 81(21):11982–11991, 2007.

Caroline S Fernandez, Ivan Stratov, Robert De Rose, Katrina Walsh, C Jane Dale, Miranda Z Smith, Michael B Agy, Shiu-lok Hu, Kendall Krebs, David I Watkins, et al. Rapid viral escape at an immunodominant simian-human immunodeficiency virus cytotoxic t-lymphocyte epitope exacts a dramatic fitness cost. *Journal of Virology*, 79(9):5721–5731, 2005.

Becca Asquith, Charles TT Edwards, Marc Lipsitch, and Angela R McLean. Inefficient cytotoxic t lymphocyte–mediated killing of hiv-1–infected cells in vivo. *PLoS biology*, 4(4):e90, 2006.

Fabio Zanini, Vadim Puller, Johanna Brodin, Jan Albert, and Richard A Neher. In vivo mutation rates and the landscape of fitness costs of hiv-1. *Virus evolution*, 3(1):vex003, 2017.

CTT Edwards, EC Holmes, OG Pybus, DJ Wilson, RP Viscidi, EJ Abrams, RE Phillips, and AJ Drummond. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics*, 174(3):1441–1453, 2006.

Brandon F Keele, Elena E Giorgi, Jesus F Salazar-Gonzalez, Julie M Decker, Kimmy T Pham, Maria G Salazar, Chuanxi Sun, Truman Grayson, Shuyi Wang, Hui Li, et al. Identification and characterization of transmitted and early founder virus envelopes in primary hiv-1 infection. *Proceedings of the National Academy of Sciences*, 105(21):7552–7557, 2008.

Frederic Bertels, Karin J. Metzner, and Roland R. Regoes. Convergent evolution as an indicator for selection during acute hiv-1 infection 2. 2017.

Frederic Bertels, Christine Leemann, Karin J Metzner, and Roland R Regoes. Parallel Evolution of HIV-1 in a Long-Term Experiment. *Molecular Biology and Evolution*, 36(11):2400–2414, 07 2019a. ISSN 0737-4038. doi: 10.1093/molbev/msz155. URL https://doi.org/10.1093/molbev/msz155.

Eva Bons, Frederic Bertels, and Roland R Regoes. Estimating the mutational fitness effects distribution during early hiv infection. *Virus evolution*, 4(2):vey029, 2018.

Dawn M Roje. Evaluating the effects of non-neutral molecular markers on phylogeny inference. *PLoS One*, 9(2):e87428, 2014.

Trevor Bedford, Sarah Cobey, and Mercedes Pascual. Strength and tempo of selection revealed in viral gene genealogies. *BMC Evolutionary Biology*, 11(1):220, 2011.

Todd A Castoe, AP Jason de Koning, Hyun-Min Kim, Wanjun Gu, Brice P Noonan, Gavin Naylor, Zhi J Jiang, Christopher L Parkinson, and David D Pollock. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences*, 106 (22):8986–8991, 2009.

Alexei J Drummond, Simon YW Ho, Matthew J Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS biology*, 4(5):e88, 2006.

Ha Youn Lee, Elena E Giorgi, Brandon F Keele, Brian Gaschen, Gayathri S Athreya, Jesus F Salazar-Gonzalez, Kimmy T Pham, Paul A Goepfert, J Michael Kilby, Michael S Saag, et al. Modeling sequence evolution in acute hiv-1 infection. *Journal of theoretical biology*, 261(2): 341–360, 2009.

Nicola F Müller, David Rasmussen, and Tanja Stadler. MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, 34(22):3843–3848, November 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty406. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6223361/.

Tanja Stadler. Simulating Trees with a Fixed Number of Extant Species. *Systematic Biology*, 60 (5):676–684, October 2011. ISSN 1063-5157. doi: 10.1093/sysbio/syr029. URL https://doi.org/10.1093/sysbio/syr029.

Stilianos Louca and Michael Doebeli. Efficient comparative phylogenetics on large trees. *Bioinformatics*, 34(6):1053–1055, March 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx701. URL https://doi.org/10.1093/bioinformatics/btx701.

Timothy G. Vaughan and Alexei J. Drummond. A Stochastic Simulator of Birth–Death Master Equations with Application to Phylodynamics. *Molecular Biology and Evolution*, 30(6):1480–1493, June 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst057. URL https://doi.org/10.1093/molbev/mst057.

Frederic Bertels, Karin J Metzner, and Roland R Regoes. Convergent evolution as an indicator for selection during acute hiv-1 infection. 2019b.