The impact of sampling bias on viral phylogeographic reconstruction

Pengyu Liu[1], Yexuan Song[1], Caroline Colijn[1], Ailene MacPherson[1*],

1 Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada

* ailenem@sfu.ca

Abstract

Genomic epidemiology plays an ever-increasing role in our understanding of and response to the spread of infectious pathogens. Phylogeography, the reconstruction of the historical location and movement of pathogens from the evolutionary relationships among sampled pathogen sequences, can inform policy decisions related to viral movement among jurisdictions. However, phylogeographic reconstruction is impacted by the fact that the sampling and virus sequencing policies differ among jurisdictions, and these differences can cause bias in phylogeographic reconstructions. Here we assess the potential impacts of geographic-based sampling bias on estimated viral locations in the past, and on whether key viral movements can be detected. We quantify the effect of bias using simulated phylogenies with known geographic histories, and determine the impact of the biased sampling and of the underlying migration rate on the accuracy of estimated past viral locations. We then apply these insights to the geographic spread of Ebolavirus in the 2014-2016 West Africa epidemic. This work highlights how sampling policy can both impact geographic inference and be optimized to best ensure the accuracy of specific features of geographic spread.

Introduction

Genomic epidemiology refers to the use of pathogen genomes to understand how infectious diseases are transmitted across a range of scales, from local outbreaks to global geographic spread. Due to improvements in sequencing technology and decreasing cost, genomic epidemiology has increased greatly in scale in many jurisdictions and for many viruses to the extant that sequencing now plays an important role in pathogen surveillance [1]. Sequencing allows us to identify new viruses, understand their origins and natural reservoirs, to characterize their transmission dynamics in human populations, and to understand their evolution [2,3]. Interest in genomic surveillance for viruses has grown, in the context of the high utility of genomic data for these tasks and the increasing awareness of the severe threat that emerging

viruses can represent for global public health. At the time of writing, the Global initiative on sharing all influenza data (GISAID) repository [4,5] has over 10 million SARS-CoV-2 sequences and over 350,000 influenza virus sequences. Nearly 3000 Zaire Ebolavirus sequences are accessible through ViruSurf along with over 12,000 Dengue virus sequences and sequences of MERS, other Ebolaviruses and Severe Acute Respiratory Syndrome coronavirus 1 (SARS-CoV-1).

Virus sequences are a rich potential source of information about the origins, past population dynamics and geographic movements of viruses [6]. This is in part because sequences reveal information about viral evolution and population dynamics at times and in places where the viruses were not directly observed. Sequences are interpreted with the aid of phylogenetic trees, which represent the evolutionary relationships between a set of taxa. Here we focus on the fact that the geographic locations of the virus' ancestors in the past can be reconstructed by extrapolating the location of the observed sequence back in time from the present day [7–11]. There are a number of methods that perform this extrapolation. These treat the geographic location as either a discrete or continuous trait evolving on the phylogenetic tree; some methods use Bayesian approaches to simultaneously reconstruct the location and the phylogeny and others reconstruct the geographic location on a fixed phylogeny [12,13].

Indeed, phylogeography is one of the main applications of large-scale virus sequence datasets. In 2007, phylogeographic analysis revealed Guangdong as a likely source of Avian H5N1 influenza viruses, and Indonesia a likely sink [9]. Phylogeographic analysis was used to understand transmission dynamics of Ebolavirus in space and time in the Democratic Republic of Congo in 2018-2020, and was part of the genomic surveillance system informing the response to Ebolavirus in real time [14]. In human influenza, phylogeographic analysis is used to characterize the emergence and global spread of human seasonal influenza viruses and compare their circulation patterns [15]. In the SARS-CoV-2 pandemic, phylogeography has been used at a range of scales, to uncover the virus' early dissemination from Europe in 2020 [16], to demonstrate repeated introductions and localized spread in Spain [17], Peru [18], Canada [19] and other settings, and to visualize geographic movements (e.g. in nextstrain) [20].

A number of authors have noted that phylogeographic reconstructions depend on the fraction of infections that are sampled, sequenced and shared, and that phylogeographic estimates can be biased as a result [15,21–24]. These authors have taken different approaches to compensating for this bias. de Maio *et al.* [23] propose a method called the BAyesian STructured coalescent Approximation (BASTA) to approximate the structured coalescent, having found that discrete trait analysis (DTA), in which migration among locations is treated in the same manner as mutation, gives biased estimates of the migration rates if the locations are strongly non-representatively sampled. Perhaps more fundamentally for viral phylogeography, they also point out that the relative sampling intensities of the locations in DTA-based phylogeography are treated as data and inform the migration estimates. This makes phylogeographic estimates sensitive to sampling choices and can lead to erroneous inferences and

erroneously small apparent uncertainties; the structured coalescent approach does not have this issue, and parameter estimates in simulations were better than those from DTA models [23]. Magee and Scotch characterize the impact of two sampling schemes and a range of generalized linear models on the estimation of the root state (point of origin) and on the variables (in the linear model) that are deemed to be important [25]. Bias can occur both in estimates of the parameters of the process (e.g. migration or dispersal rates) and in the reconstructed ancestral geographic locations themselves. Approaches to adjust for bias are in development. In the context of the structured coalescent model, De Maio et al. account for the effect of sampling bias on the estimation of migration rates by integrating over all possible migration histories [23]. Guindon and De Maio address non-uniform sampling in a diffusion model of movement in continuous space by using a Bayesian approach to include the sampling process in the posterior distribution, but this requires an exchange algorithm that introduces substantial computational complexity [26]. Their correction leads to differences in the inferred growth rate, effective population size, dispersal parameter and the time of the most recent common ancestor (compared to not correcting for the sampling process). Kalkauskas *et al.* explore adding "empty" viral sequences in a continuous-space model [24], finding that this approach can bias in estimates of the root location, and diffusion rate, but not eliminate it. Lemey et al. incorporate both "empty" viral sequences and individual travel history in part to overcome sampling bias [8] in phylogeography. Indeed, the role of travellers is of course important in shaping viral geographic movements themselves, but also potentially important in reconstructing them. In some jurisdictions, travellers' samples are prioritized for sequencing as part of monitoring viral diversity [27]. The information about which sequences are from traveller samples is not typically shared to public databases such as GISAID, which could additionally confound phylogeographic inferences.

Here we focus on maximum-likelihood phylogeographic inference with location modeled as a discrete quantity. Discrete space is appropriate for many epidemiological applications where data is both collected and policies implemented at a regional level. While methods have primarily been developed with Bayesian tools, and so have inherited the disadvantage that limited numbers of sequences can be included if the analysis is to be computationally tractable. An additional challenge is that Bayesian phylogenetic reconstruction produces multiple (posterior) phylogenies with distinct internal nodes. Without comparable internal nodes from tree to tree, it would not be a well-posed task to summarize the nodes' locations, and thus to infer the phylogeographic information. In practice, users of phylogeography who have thousands of sequences often cannot use state-of-the-art Bayesian methods due to computational limitations of these methods for large datasets, and instead use a simpler maximum-likelihood approach [19]: construct a phylogeny, and then reconstruct the geographic locations of the past nodes on that fixed phylogeny, both with maximum-likelihood methods [11].

The impact of location-based sampling bias on the phylogeographic reconstructions of the past locations of pathogens has not been thoroughly characterized, in either

Bayesian or maximum likelihood methods. Yet this has important practical implications for the users of phylogeography, particularly where large datasets are used to reconstruct a virus' global geographic movement. The implications of sampling bias for the design of genomic surveillance and data sharing strategies are not well understood. Here we characterize the impact of sampling bias on error in the maximum-likelihood phylogeographic reconstruction of a virus' location in the past. We begin by using simulations of viral branching, cure (or host death), and migration to quantify the impact of sampling bias on the overall reconstruction accuracy, on the types of errors that result, and on the sources of variability in the quality of reconstruction. This exploration includes the impact of sampling bias on location of individual ancestral nodes, the inference of the root location (origin), and the impact of over/under representing recent travelers. To illustrate the potential impact of sampling in a natural context, we then perform phylogeographic analysis using Ebolavirus sequences from West Africa. Throughout, we offer practical comments on when geographic sampling bias is likely to impact maximum-likelihood phylogeographic reconstructions.

Materials and methods

Assumptions and tree simulation

We simulate pathogen diversification under the binary-state speciation and extinction (BiSSE) model [28] implemented in the R package `diversitree` [29]. Here we will consider the case where the two states represent distinct geographical locations. Each node of the resulting rooted binary-state phylogenetic tree is in either location A or location B. Without loss of generality, the diversification is simulated with the initial lineage (the root of the tree) in location A. We assume that speciation and extinction rates are independent of a lineage's location (i.e. location is a neutral character), though we relax this assumption in the Supplement. While this assumption is unlikely to be true in natural systems, it allows us to focus on geographical differences in sampling intensity in isolation. As with the speciation and extinction rate, we assume that migration is symmetrical between the two locations. The resulting diversification model has three parameters: the speciation rate λ, the extinction rate μ, and the migration rate α between location A and location B. In the context of a pathogen phylogenetic tree, "extinction" represents the end of the infectious period (which could be death, or recovery, successful treatment, effective quarantine, etc). Similarly, "speciation" is usually assumed to be coincident with pathogen transmission between hosts.

Observed at the present day, a phylogenetic tree generated by this model has two types of tips: "extinct" tips associated with extinction events some of which, importantly for us here, are assumed to be coincident with sampling and sequencing of the pathogen lineage and "extant" tips representing lineages that exist and remain unsequenced at the present day. As we assume that the sampling and sequencing of a pathogen is coincident with treatment or quarantine the result is a heterochronous time

tree consisting of a subset of the extinct tips. To obtain this tree, we drop all extant (continuing past the present day, and not sampled, as sampling occurs through time in the past) lineages of the simulated tree. We call the resulting tree consisting of only "extinct" tips the "true tree", because it has all of the tips that could possibly have been sampled. The geographical location of the internal nodes of this (simulated) true tree are known and will be used as the reference for quantifying the accuracy of ancestral state reconstruction. We select n tips from the true tree, where n reflects sequencing capacity. We call the resulting subtree with only the n selected tips a "downsampled tree". In order to ensure that all n sampled lineages may be in either location, the initial simulation for a true tree is terminated only once it contains at least n tips in location A and n tips in location B.

Downsampling schemes

To assess the impact of location-biased sampling on ancestral state reconstruction we consider two sampling schemes. In the first scheme, we construct a downsampled tree S_k with n tips from a true tree T. The fraction of tips in the downsampled tree that are from location A is k (up to rounding, e.g. if $k = 1/3$ and $n = 100$ we have 33 location-A tips in S_k. Mathematically: the tips in S_k are uniformly selected at random from the extinct tips of T such that $\lfloor kn \rfloor$ (rounded to the smaller integer of kn) tips are in location A and $n - \lfloor kn \rfloor$ tips are in location B. In our experiments with this downsampling scheme, k ranges from 0.05 to 0.95 in increments of 0.05 ($5 - 95\%$).

To assess the impact of sampling travellers or those with recent travel-related exposure on phylogeographic reconstruction, we implement a downsampling scheme in which lineages with changes in location near the tips are over- or under- represented. Specifically, we classify a tip in the true tree T as a "recent migrant" if the tip has a different state from its parent node. Suppose T has m_A recent migrants in location A and m_B recent migrants in location B. We construct a downsampled tree S_p with n tips, for which we select tips from T, attempting to have a fraction p of the location-A tips be "recent migrants". For example, if $n = 100$, $k = 0.4$ and $p = 0.8$, we want 80% (or 32) of the 40 location-A tips in the downsampled tree to be recent migrant tips. There may be fewer than 32 recent migrant tips in the true tree to begin with. In this case we include all of them. If there are more than 32, we include 32 sampled at random. Mathematically: we uniformly select $\lfloor kn \rfloor$ location-A tips at random such that $\min(m_A, \lfloor kpn \rfloor)$ of the tips are recent migrants, and uniformly select $n - \lfloor kn \rfloor$ location-B tips at random such that $\min(m_B, \lfloor kpn \rfloor)$ of the tips are recent migrants. In our experiments with this downsampling scheme, we set k to be 25%, 50% or 75%, and p ranges from 5% to 95% in increments of 5%.

Ancestral state reconstruction

To reconstruct the ancestral states at the internal nodes of the downsampled trees, we use the maximum likelihood method first described by Pagel [30] and implemented in

the *ace* function in the R package `ape` [31]. This method reconstructs the states of internal nodes for a fixed unlabelled tree with known tip states, and gives the likelihood that each internal node (including the root) was in each of the two states (locations). This method also estimates the migration rates of the downsampled trees under the assumption of equal migration rates between the two locations.

Reconstruction accuracy

We use two quantities to assess the quality of our phylogeographic reconstructions: absolute accuracy and relative accuracy. The absolute accuracy is the fraction of internal nodes for which the reconstructed location is correct. However, if most of the tips and most of the internal nodes are from the same location, the absolute accuracy can be high in a trivial way, as the reconstruction can simply estimate that all the nodes were in that location. The relative accuracy accounts for this. It is the improvement in the absolute accuracy, compared to a null model based only on the locations of the tips.

Mathematically, we define absolute accuracy and relative accuracy as follows. Consider a true tree T and its downsampled tree S with an internal node i. Let c_i represents the location of internal node i in the true tree T, where $c_i = 1$ if i is in location A and $c_i = 0$ if i is in location B, and $\tilde{c_i}$ be the corresponding likelihood of node i being in location A as inferred by the ancestral state reconstruction method. We define the absolute accuracy of node i as $a_i = 1 - |\tilde{c_i} - c_i|$ and the absolute accuracy of the downsampled tree S, a_S, as the absolute accuracy averaged over all internal nodes in S.

To define the relative accuracy of the downsampled tree S, we compare the absolute accuracy a_S to a null expectation e_S assuming that the probability that an internal node is in a given state is proportional to the frequency of the two locations at the tips and hence agnostic of phylogenetic structure. Suppose S has n_A location-A tips and n_B location-B tips. We define the probability that an internal node of S is in location A to be $p = n_A/(n_A + n_B)$ and the expected accuracy of node i as $e_i = 1 - |\tilde{c_i} - p|$. The expected accuracy of the downsampled tree S, e_S is the expected accuracy averaged over all internal nodes of S, and the relative mean reconstruction accuracy of the downsampled tree S is then given by $r_S = a_S - e_S$.

Key migration events

To understand the impact of sampling on the reconstruction of "key migration events", events that ultimately lead to large outbreaks following introduction, here we define a procedure for selecting such events from a true tree and measures for quantifying the accuracy of their reconstruction. Noting that the root of a true tree is always in location A, we define a key migration event (KME) in a true tree T to be any edge connecting a location-A internal (parent) node to a location-B internal (child) node which subsequently has at least 15 location-B tips as its descendants. These location-B tips are called the "tips of the KME".

To quantify the different mechanisms resulting in inaccurate reconstruction of KMEs,

we characterize the state of a KME in a downsampled tree with two measures: the presence of the internal nodes (the parent node and the child node of the KME) and the presence of sufficient (at least 5) tips of the KME. The presence of the internal nodes of a KME indicates if we can correctly infer spatial-temporal information about the introduction (migration) event, and the presence of sufficient tips of the KME determines if we consider the there exists a large outbreak led by the introduction event. The reconstructed KME characterized by these two measures is summarized in Table 1 and explained below. The presence of internal nodes has three states: if both parent and child internal nodes of a KME appear in a downsampled tree, then we say the KME is "observed via internal nodes" in the downsampled tree; if only the parent internal node of a KME appears in a downsampled tree, then we say the KME is "erred via internal nodes" in the downsampled tree, and we measure the error by the branch length between the original child internal node of the KME and the apparent child internal node in the downsampled tree; if the parent internal node of a KME is not present in a downsampled tree, then we consider the spatial-temporal information of the introduction event can not be correctly reconstructed, and we say the KME is "obscured via internal nodes". The presence of sufficient tips of the KME simply has two states: if at least 5 tips of the KME appear in a downsampled tree, then we say the KME is "observed via tips"; if fewer than 5 tips of the KME appear in a downsampled tree, then we say the KME is "obscured via tips".

State of a KME	**Parent node**	**Child node**	**≥ 5 tips of the KME**
Observed via internal nodes Observed via tips	Present	Present	Present
Erred via internal nodes Observed via tips	Present	Absent	Present
Obscured via internal nodes Observed via tips	Absent	Present or absent	Present
Observed via internal nodes Obscured via tips	Present	Present	Absent
Erred via internal nodes Obscured via tips	Present	Absent	Absent
Obscured via internal nodes Obscured via tips	Absent	Present or absent	Absent

Table 1. Definitions of key migration event (KME) states in a downsampled tree.

To assess the KME reconstruction accuracy for a true tree, we construct 100 downsampled trees for each fraction k of location-A tips in downsampled trees; we count the number of downsampled trees in which each KME from the true tree is observed, erred and obscured via internal nodes and the number of downsampled trees in which each KME from the true tree is observed and obscured via tips.

2014-2015 Ebola Epidemic

We illustrate the impacts of sampling differences using data from an Ebolavirus outbreak in Africa in 2014–2015 [32]. The data consists of 262 taxa collected from 4 countries: 152 from Guinea, 84 from Sierra Leone, 22 from Liberia, 2 from Mali, and 2 from unknown locations. The time-scaled tree is generated using BEAST with a log-normal distributed relaxed molecular clock and a non-parametric coalescent prior. More details can be found in the original paper [32].

Naturally, the published data only include some of the infections that occurred, but we can explore the impact of sampling by further downsampling, selecting among the tips that are present. We consider the reconstructed phylogeny with 262 tips as the true tree here; we reconstruct the location for the internal nodes of the tree with all 262 tips using `ace` and consider the reconstructed location of the internal nodes as true ancestral locations. Similarly, if a tip in the true tree has a different location as its parent node, then we say the tip is a "recent migrant". We apply two sampling schemes: randomly downsampling tips in one location at a time and preferentially downsampling recent migrants in the true tree. These tips are a proxy for travelers (and/or contacts of travelers). For random downsampling, since Liberia and Mali have significantly fewer tips than Guinea and Sierra Leone in the phylogeny, we only consider downsampling Guinea and Sierra Leone tips. We downsample by dropping 80% of the tips in each of the two locations randomly. In addition, for the second downsampling scheme we also randomly downsample 80% recent migrants. Once we have a downsampled tree, we use `ace` to reconstruct the ancestral location for internal nodes included in the downsampled tree and compare reconstructed ancestral location to the true ancestral location (reconstructed with all 262 tips).

Results

We consider the impact of the sampling scheme on the inference of two broad categories of phylogeographic inference: geography (the inference of ancestral states/locations) and migration (the inference of migration rates and key migration events). For each of these categories, we explore the effect of geographically biased sampling and sampling of tips for which there is recent migration in the true tree.

Geography

Absolute and relative reconstruction accuracy

We find that the overall absolute accuracy of phylogeographic reconstruction is often high, particularly when the migration rate is low (Fig. 1 and Fig. S1). Under a low migration rate, changing the fraction of the taxa that are from location A to half of what would be representative (from approximately 0.68 to 0.35 in Fig. 1B) reduces the accuracy from being consistently above 95% to being in the range of $90-100\%$; while most nodes are correct, biased sampling greatly increases the number of errors. Under a

low-to-moderate migration rate ($\alpha = 0.3$; Fig. 1), halving the number of location-A samples in the tree reduces the average accuracy from above 90% to approximately 80%, more than doubling the number of internal nodes with an incorrect location estimate. Furthermore, the estimates are quite variable. Under higher migration rates, the accuracy is lower, but it is less sensitive to sampling.

In contrast to our initial expectation that absolute accuracy would peak under unbiased sampling (i.e. when the proportion of location-A tips in the downsampled tree was the same as the proportion in the true tree), we found that the absolute accuracy increased as the proportion of location-A tips in the downsampled tree increased (Fig. 1B,D and Fig. S1). This is a result of the starting point that the root node is in location A, so that over half of the internal nodes are expected to be in location A (recall that migration from A to B and back are equal in our simulations). Without directly accounting for sampling bias in the ancestral state reconstruction, the absolute accuracy will therefore be higher when location A is oversampled. The higher absolute accuracy is simply because almost all of the tips and internal nodes of the downsampled tree are in location A, making it easy for the reconstruction to obtain correct locations. This is relevant particularly when the migration rate is relatively low. In this case, there are few migration events, and the true tree consists of larger monomorphic (same location) clades. This greatly improves the ancestral state reconstruction by increasing the probability that internal nodes have the same location as all their immediate descendants, and this raises the absolute accuracy.

In contrast to absolute accuracy, the relative accuracy (which takes this into account, and captures the improvement over expected accuracy from a null model; see Methods) peaks at a sampling proportion of 50% location-A tips, and the overall relative accuracy depends on the geographic sampling bias (Fig. 1 and Fig. S1). When the migration rate is higher, the accuracy of the phylogeographic reconstruction is lower, and the dependence on sampling is less pronounced. Both the impact of sampling proportion and migration rate hold across a large number of true trees (Fig. S2).

Figure S3 illustrates *how* the phylogenetic reconstructions are inaccurate. Reconstruction of migration events, or the lack there of, between a parent and child node is particularly sensitive to state of the parental node. If the parental state is over(under)-sampled, we are likely to greatly over(under)-represent those types of edges in the tree. In contrast, over- or under- sampling of the child state has a limited impact on geographic inference. This result is most pronounced when the migration rate is high. To interpret these results: suppose we are located in a region with comparatively high sequencing (location B; left side of the panels in Figure S3). Sampling bias will impact our understanding of both migration between jurisdictions and the extent of sustained transmission within a jurisdiction. Specifically we will substantially underestimate the migration rate from the under-sampled jurisdiction (A-B edges) but fairly accurately capture migration to that jurisdiction from the highly sampled region (B-A edges). In addition, we are likely to conclude there are long sustained transmission chains in the over-sampled jurisdiction (B-B edges) and short transmission chains in the

under-sampled jurisdiction (A-A edges).

We also explored how sensitive our results are to the assumption of equal 'speciation' rates. For example, transmission may be occurring at a higher rate in location A or B due to different public health measures or population contact rates. When transmission (or speciation, in the language of our model) is location-dependent, the overall patterns of bias are similar to what we have found in the neutral case. However, the maximum relative accuracy is obtained by sampling the location with the lower speciation rate more than would be representative of the locations' frequencies in the true tree; see Fig. S5.

Oversampling recent migrants

We find that both the absolute and the relative accuracy decrease as more recent migrants (recall that recent migrants are tips in a true tree that have a different location from its parent node) are sampled (Fig. 2A,B and Fig. S6). We use recent migrants as models for either travelers or members of their contact networks. This means if we sample more travel-associated infections, mark the corresponding taxa with the location at which they were sampled, and use the taxa to infer ancestral states or migration patterns, then the inference accuracy may be reduced. We note (Fig. 2C) that there exist tips with migration in their very recent ancestry in a downsampled tree that are not recent migrants in the true tree. We call such tips of the downsampled tree the "apparent" recent migrants. Fig. 2D illustrates how the presence of recent migrants affects the both the absolute accuracy of their ancestors, and that the apparent recent migrants tips observed in downsampled trees also contribute to the inaccuracy in ancestral state reconstruction. The error caused by recent migrants or apparent recent migrants can cascade up the tree and affect the overall absolute and relative accuracy. For example, in Fig. 2F on the right hand side, note the progression of error (red node) due to incorrect inference of location B instead of A, following on from including two tips with recent migration (rightmost tips in Fig. 2E). If recent migrants are heavily sampled (Fig. 2E,F), then the absolute accuracy of internal nodes can be much reduced.

Migration

Migration rate inference

The maximum likelihood method that we use for ancestral state reconstruction gives estimates of migration rates assuming that migration is symmetric between the two locations. We find that at the least biased geographic sampling proportions, the estimates of migration rates are close to the migration rates of the true trees, but the best migration rate estimates do not always occur when the geographic sampling is least biased (Fig. 3). Migration rates are on average overestimated when the true migration rate is low (Fig. 3 top panels) whereas rate estimates are highly variable but on average closer to the truth when the true migration rate is high (Fig 3 bottom panels). Migration rate estimation is presumably less biased because higher migration provides

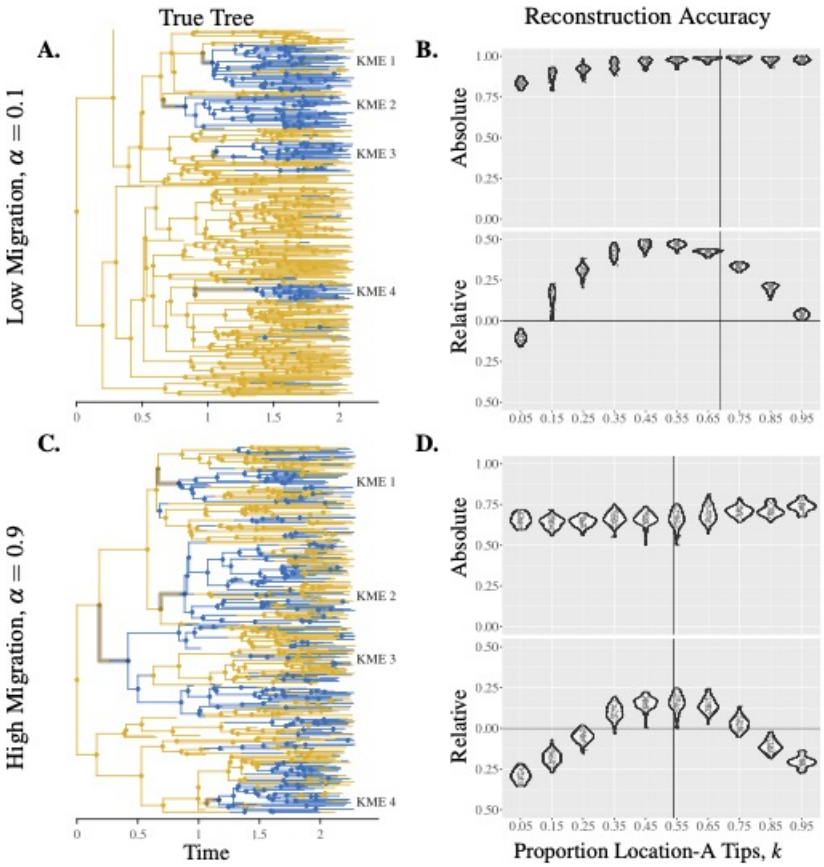


Fig 1. Reconstruction accuracy depends on sampling bias and migration rate. Panels A & C: True trees simulated with low or high migration rates, KMEs are highlighted. Panels B & D: Absolute and relative accuracy of downsampled trees. Each data point represents a downsampled tree from the true tree on the left. The vertical lines indicate the proportion of location-A tips in the true trees. Parameters: $\lambda = 4$ and $\mu = 1$.

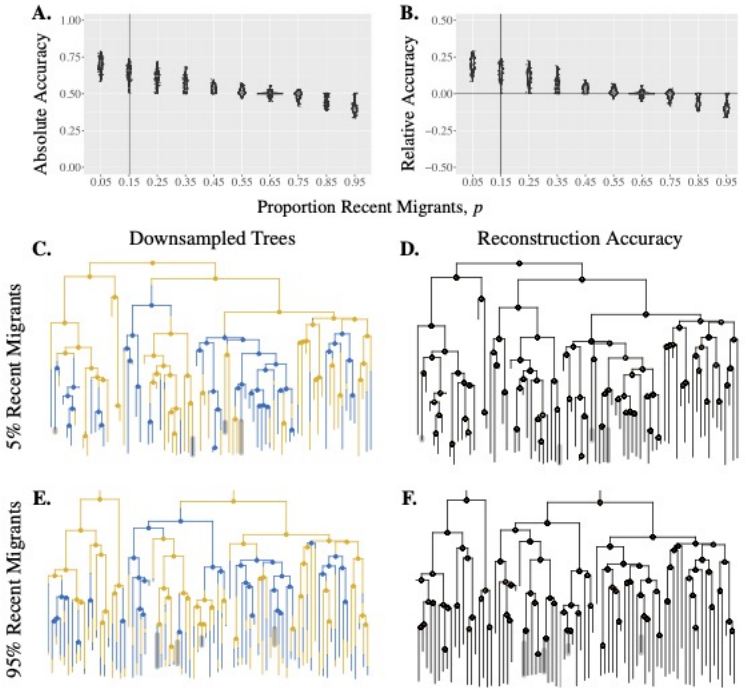


Fig 2. Reconstruction accuracy depends on the sampling of recent migrants. Absolute (Panel A) and relative (Panel B) accuracy of downsampled trees with varying p. The corresponding true tree is shown in Fig.1C with the vertical line showing the proportion of recent migrants in this true tree. Panels C & E: Examples of downsampled trees from the true tree in Fig. 1C with a low (Panel C: $p = 0.05$; recent migrants highlighted) or high (Panel E: $p = 0.95$; non-migrants highlighted) proportion of recent migrants tips. Panels D & F: The absolute accuracy of each internal node (green fraction) of the downsampled trees in panels C and E respectively. Parameters: $\lambda = 4$, $\mu = 1$, $\alpha = 0.9$, and $k = 0.5$.

more migration events on which to base the estimates. However, as the number of migration events increases, so too does sensitivity to biases sampling (more migration events are incorrectly inferred). Note that the y-axis in Fig. 3 is on a logarithmic scale, the migration rate is often underestimated under extreme bias and overestimated when sampling is unbiased, especially when the true migration rate is high. When sampling is highly biased, the downsampled tree contains far fewer migration events than the true tree (clades become more monomorphic) which results in an underestimate of the migration rate. This is particularly true when the migration rate is high and there are more migration events that can be excluded by biased sampling. In contrast, the overestimation of migration rate when sampling is unbiased results from the fact that downsampling makes clades less monomorphic than in the true tree.

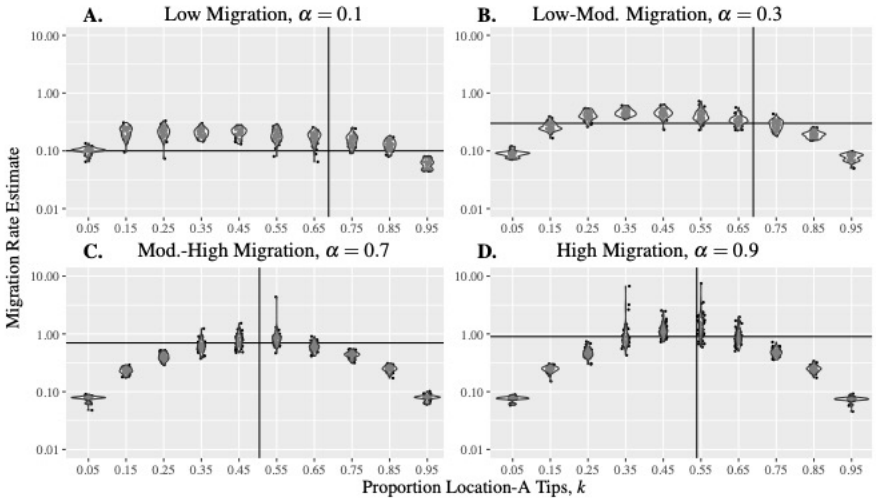


Fig 3. Sampling bias affects the migration rate estimate. Estimated migration rates from trees with varying migration rate. The true trees for each panel are shown in Fig.1A (Panel A), Fig.1C (Panel B),Fig. 4A (Panel C), and Fig. 1C (Panel D). Vertical lines indicate the true proportion of location-A tips and the horizontal lines indicate the true migrations rates. Parameters: $\lambda = 4$ and $\mu = 1$.

Key migration events

Here we assess how the reconstruction key migration events (KMEs) in downsampled trees is impacted by sampling bias. Recall that a KME is an edge in the true tree with a location-A parent internal node a location-B child internal node, and the subsequent location-B tips of the child node of the KME are the tips of the KME. Whether the edge can be reconstructed in a downsampled tree indicates whether we can learn about the introduction event from the sample and whether we sample sufficient location-B tips shows whether we can accurately infer that the introduction lead to onward transmission in the destination. The number of KMEs that are "observed via tips" decreases as the number of location-A tips in the downsampled tree increases (Fig. 4

and Fig. S7) because, if we do not sample enough location-B tips, we will not know there is ongoing transmission after an introduction. Note that there are always two child lineages arising from the parental node of a KME, the focal child node which is in location B and a sister clade which is, at least initially, in location A. For a KME to be "observed via internal nodes" requires that the parental node is in the downsampled tree. This in turn, depends on whether tips in the sister clade of the KME are sampled (Fig. 4 and Fig. S7).

The sampling policy most likely to reconstruct KMEs depends on the migration rate. When the migration rate is high sampling more location-B tips helps reconstruct KMEs, since the true tree has few monomorphic clades and the KMEs would rarely be "erred" or "obscured via internal nodes". When the migration rate is low, unbiased sampling gives better results because the true tree has monomorphic clades and heavily biased sampling can drop entire clades and cause the KMEs to be erred or obscured. Biased sampling comes at an additional cost to reconstruction of KMEs. If sampling is biased enough such that both parent and child internal nodes of a KME are reconstructed with the same location, then we may not identify the KME even though both parent and child internal nodes and the sufficient number of tips of the KME appear in the downsampled tree(Fig. S5).

Application to the 2014-2015 Ebola Epidemic

We apply our analysis to the phylogenetic tree reconstructed from sequences of Ebola virus. The sequences are sampled at five locations, but most of them are from two locations, namely Guinea and Sierra Leone. The reconstructed maximum clade credibility (MCC) phylogenetic tree from BEAST is used as the "true tree", which has an estimated migration rate of 0.40 per unit time, from `ace` under the assumption of neutral migration rates between locations.

Relative to the simulations, the Ebolavirus tree has a relatively low migration rate compared to the branching rates, resulting in large monomorphic clades. We would therefore expect the absolute accuracy to be high overall, in keeping with Fig. 1, but that the relative accuracy would depend on the geographic sampling bias. We examine the absolute accuracy of the internal nodes in downsampled trees, especially the internal nodes representing an introduction event in the true tree. Downsampling 80% of Sierra Leone tips does not greatly reduce the absolute accuracy of the downsampled tree (Fig. S9). This is because the Sierra Leone tips form a monomorphic clade. When downsampling 80% of Guinea tips (less arranged in monomorphic clades), the overall absolute accuracy of the internal nodes is also high (Fig. S10), but we observe a group of internal nodes (red box in Fig. S10) that have inaccurate reconstructed locations. This does not have a large impact on overall absolute accuracy, but it significantly affects our inferences about the introduction event. Specifically, in this case we would infer that Ebolavirus was introduced to Guinea via Liberia, which is not the case in the true tree.

When we downsample 80% of the "recent migrant" tips (reflecting more likely recent-travel-related sequences) in addition to dropping 80% of tips independent of

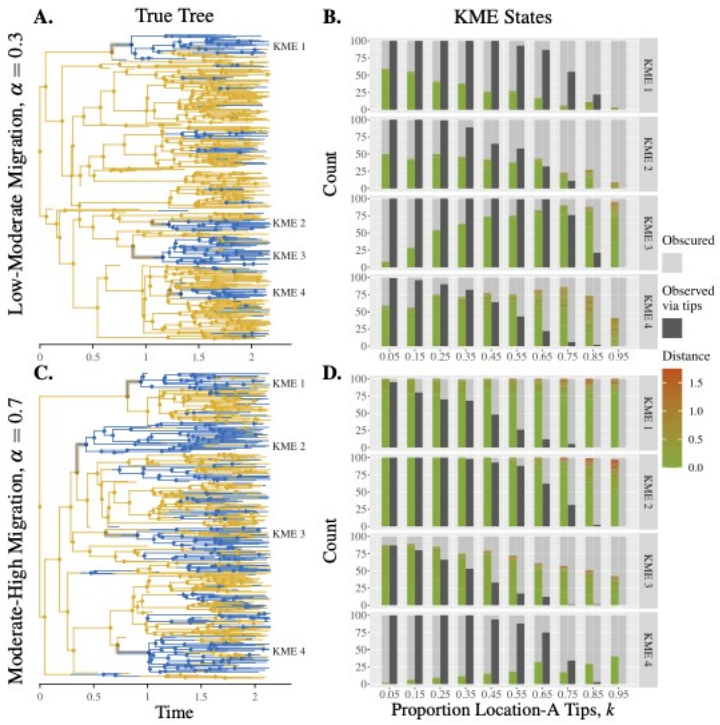


Fig 4. Reconstruction of KMEs depends on sampling bias and migration rate. Panels A & C: True trees with intermediate migration rates with key migration events highlighted. Panels B & D: The KME states (see Table 2) for 100 downsampled trees with low-moderate. Number of downsampled trees in which a KME is erred (coloured bars) or observed via internal nodes (left-hand transparent-grey bars) and weather KMEs are observed via tips (solid-grey bars) or obscured via tips (right-hand transparent-grey bars). Parameters: $\lambda = 4$ and $\mu = 1$.

location, the ancestral state reconstruction is accurate except at the migration event in the red box (Fig. 5). This is because the reconstructed MCC tree has relatively few recent migrants (32/262), and removing most of these tips does not affect the absolute accuracy. However, if we were particularly interested in the information in the red box in Fig. 5, that is, whether the geographic spread is directly from Sierra Leone to Guinea or from Sierra Leone to Liberia to Guinea, then the impact of sampling travelers (or those with recent travel contact) would be of concern. Table 2 summaries how inference of the likely source (parent location) and destination (child location) of introductions is impacted by down-sampling of recent migrants. After downsampling, the introduction event appears to be most likely from Sierra Leone to Guinea, whereas the introduction event in the true tree is from Liberia to Guinea.

Location	Guinea	Liberia	Mali	Sierra Leone	Unknown
Parent node (before)	0.00	0.99	0.00	0.01	0.00
Child node (before)	0.99	0.01	0.00	0.00	0.00
Parent node (after)	0.08	0.29	0.01	0.61	0.01
Child node (after)	0.62	0.12	0.02	0.22	0.02

Table 2. Effect of downsampling on inferred introduction events. Likelihood of the parent node and the child node of a migration event in the red box being at each location before and after downsampling recent migrants.

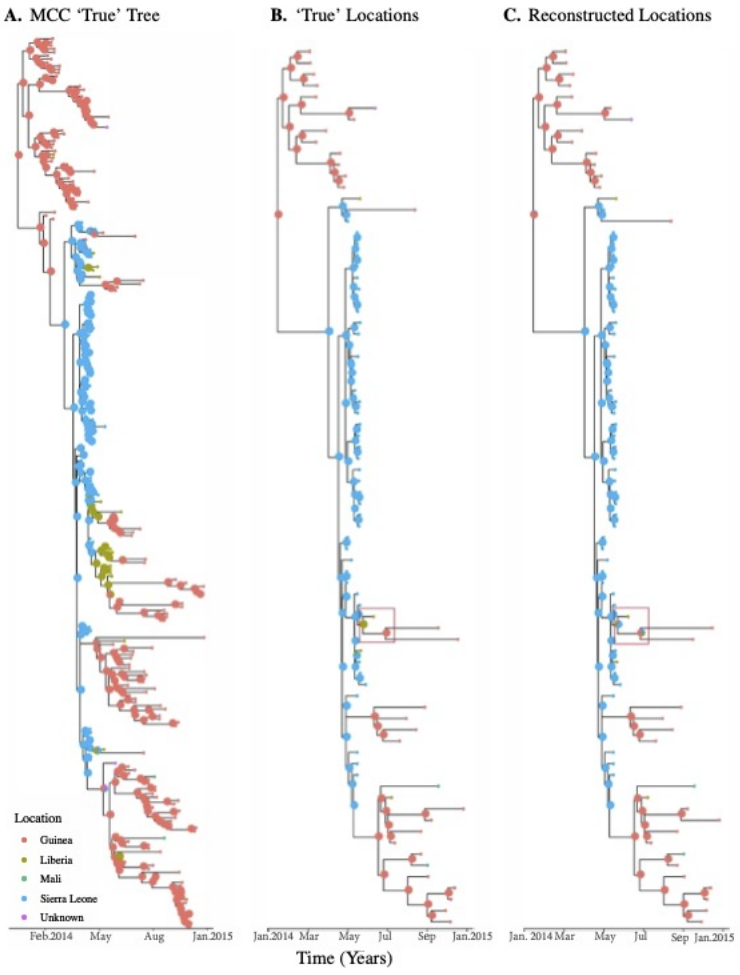


Fig 5. Ancestral state reconstruction of Ebolavirus trees with recent migrants downsampled. Panel A: The Maximum Clade Credibility (MCC) tree of Ebolavirus sequences generated using BEAST, which is used as the true tree. Downsampled trees obtained by dropping 80% recent migrants with either the true ancestral locations (Panel A) or ancestral locations reconstructed after downsmapling migrants (Panel C).

Discussion

We performed a simulation study to characterize the impact of geographic sampling bias on reconstructions of past viral locations, and to investigate how and why sampling bias affects the ancestral state reconstruction. Downsampling – and sampling bias in particular – can create two types of error in phylogeographic inference. First, the proportion of tips in each location can affect the overall accuracy of reconstructed ancestral locations. The error created near the tips can cascade up the tree, as node locations are estimated using their descendants' locations. Second, with biased sampling, the reconstructed tree shape can omit nodes and edges related to key migration events, making inference about important introduction events difficult. The eventual accuracy depends on the underlying migration pattern, the migration rate, the rate of sampling bias and the rate of sampling travelers. We found that the relative accuracy is worsened with increased sampling bias and when travel-associated tips are over-represented. If the underlying migration rate is higher , the relative accuracy of phylogeographic reconstructions after sampling can be lower, but inferences about migration rates and about key migration events can be more accurate than if the underlying migration rate is lower.

We only simulated two locations in our model, and in the main text we used a birth-death model without location-depending speciation rate or migration rate. We showed in the supplementary material that location-dependant speciation rate can also impact phylogeographic reconstruction, and in this case oversampling the location with a lower speciation rate (compared to that location's representation in the true tree) can help to compensate. We did not vary the size of downsampled trees, and we assumed that a true tree is large enough to contain both at least n location-A tips and at least n location-B tips, which may be an unrealistic assumption. We chose to generate true trees first and simulate geographic sampling by downsampling the true trees after they are generated. However, we note that there are methods to simulate geographic sampling along a birth-death process, for example in [33]. We compared these two approaches to simulate geographic sampling and construct downsampled trees in the supplementary material, and we found the two approaches provide similar results in terms of absolute reconstruction accuracy, hence relative reconstruction accuracy. See Fig. S11.

In the phylogenetic tree reconstructed from sequences of Ebolavirus, the true locations of the virus in the past are not known, and we used the reconstructed ancestral states from `ace` as the 'true' ancestral states. The reconstructed phylogenetic tree has a relatively small migration rate and monomorphic clades. Dropping tips from one location does not much affect the overall absolute accuracy, but can significantly reduce the quality of inference about introduction events. Similarly, the phylogenetic tree does not contain many recent migrants, and dropping most of the recent migrants does not reduce the absolute accuracy by much but can heavily affect inference about introduction events.

Furthermore, we simulated trees and either downsampled those trees or simulated birth, death and sampling through time; we did not simulate sequences evolving on the trees, nor reconstruct trees from these simulated sequences. Accordingly, our results are for the idealized circumstance where the reconstruction of the tree itself is essentially perfect. In practice, tree uncertainty would add additional noise to phylogeographic reconstruction, and the accuracy of reconstructions would be expected to be lower than we have found here.

We have found that even under idealized circumstances, the portion of infections from the different locations that are sequenced and represented in the data, and the extent to which they are representative samples of the circulating infections in their jurisdictions, can strongly shape the reliability of phylogeographic inferences. While this was known in its impact on estimates of migration rates and the originating sequence (root), the impact on the estimated virus locations themselves, or on the ability to detect important viral movements, has not previously been characterized. Methods are being developed to compensate for this bias, but currently, these are Bayesian methods with high computational costs, and they are are not suitable for the analysis of the high volumes of viral sequences that are being generated. Furthermore, while incorporating past locations into Bayesian phylogenetic reconstruction is appealing, the fact that the internal nodes in the posterior sample of phylogenetic trees are not the same from tree to tree hinders interpretation of the reconstructed locations. This analysis is therefore more suited to estimating rates than viral movements. One approach is to proceed as in Lemey *et al.* [8], summarizing the posterior trees to capture viral movements without focusing on individuals nodes (though this does not resolve the issue of computational capacity for Bayesian tree reconstruction beyond hundreds of sequences). Another would be to use the BASTA framework of de Maio *et al.* [23], but if there are too many sequences for Bayesian phylogenetic estimation to be practical, fix the phylogenetic tree (estimating it first by maximum likelihood), and take the structured coalescent approach to the phylogeographic reconstruction. This may be limited by the assumptions of the structured coalescent (like fixed effective population sizes for the locations, or demes).

Adjustment for sampling differences is likely to be feasible only when those differences are known. Sharing data on the fraction of cases that are sampled and sequenced, the fraction of infections that are reported as cases, and the strategy by which samples are prioritized for sequencing, will help in making these adjustments. Ultimately this will aid in correctly inferring pathogens' geographic movements. Our results suggest that pathogens are likely to be overly estimated to be in jurisdictions that contribute more data (compared to the numbers of infections), and that over-representation of travel-associated samples can drive inaccurate estimates of past locations, incorrectly placing nodes in the location of sampling. If the reason for sequencing (e.g. travel-associated case; dense sampling due to a large outbreak) were known, this could be reduced. If the relative sampling fractions were known, a representative sample could be obtained by downsampling taxa from the relevant jurisdictions (though developing methods that can use all of the data is to be preferred).

Meanwhile, we would caution that in real datasets from public databases, covering multiple jurisdictions sequencing viruses at potentially very different rates and in non-representative ways (due to targeting outbreaks, travellers, health-care workers, specific variants and so on for prioritized sequencing but then not making the reason for sequencing available), the results of phylogeographic reconstructions should be taken with caution.

Implementation

Code and data for simulations and analyses conducted in this paper are available at `https://github.com/pliumath/sampling-bias`

References

1. du Plessis L, Stadler T. Getting to the root of epidemic spread with phylodynamic analysis of genomic data. Trends in Microbiology. 2015;23(7):383–386.

2. Robishaw JD, Alter SM, Solano JJ, Shih RD, DeMets DL, Maki DG, et al. Genomic surveillance to combat COVID-19: challenges and opportunities. Lancet Microbe. 2021;2(9):e481–e484.

3. Cyranoski D. Alarming COVID variants show vital role of genomic surveillance. Nature. 2021;589(7842):337–338.

4. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22(13).

5. GISAID - Initiative;. `https://www.gisaid.org/`.

6. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews Genetics. 2009;10(8):540–550.

7. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5(9):e1000520.

8. Lemey P, Hong SL, Hill V, Baele G, Poletto C, Colizza V, et al. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. Nat Commun. 2020;11(1):5110.

9. Wallace RG, Hodac H, Lathrop RH, Fitch WM. A statistical phylogeography of influenza A H5N1. Proc Natl Acad Sci U S A. 2007;104(11):4473–4478.

10. Faria NR, Suchard MA, Rambaut A, Lemey P. Toward a quantitative understanding of viral phylogeography. Curr Opin Virol. 2011;1(5):423–429.

11. Nielsen R. Mapping mutations on phylogenies. Syst Biol. 2002;51(5):729–739.

12. Joy JB, Liang RH, McCloskey RM, Nguyen T, Poon AFY. Ancestral Reconstruction. PLOS Computational Biology. 2016;12(7):e1004763–.

13. Barido-Sottani J, Vaughan TG, Stadler T. Detection of HIV transmission clusters from phylogenetic trees using a multi-state birth–death model. Journal of The Royal Society Interface. 2018;15(146):20180512.

14. Kinganda-Lusamaki E, Black A, Mukadi DB, Hadfield J, Mbala-Kingebeni P, Pratt CB, et al. Integration of genomic sequencing into the response to the Ebola virus outbreak in Nord Kivu, Democratic Republic of the Congo. Nat Med. 2021;27(4):710–716.

15. Bedford T, Riley S, Barr IG, Broor S, Chadha M, Cox NJ, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature. 2015;523(7559):217–220.

16. Rito T, Richards MB, Pala M, Correia-Neves M, Soares PA. Phylogeography of 27,000 SARS-CoV-2 Genomes: Europe as the Major Source of the COVID-19 Pandemic. Microorganisms. 2020;8(11).

17. Gómez-Carballa A, Bello X, Pardo-Seco J, Pérez Del Molino ML, Martinón-Torres F, Salas A. Phylogeography of SARS-CoV-2 pandemic in Spain: a story of multiple introductions, micro-geographic stratification, founder effects, and super-spreaders. Zool Res. 2020;41(6):605–620.

18. Juscamayta-López E, Carhuaricra D, Tarazona D, Valdivia F, Rojas N, Maturrano L, et al. Phylogenomics reveals multiple introductions and early spread of SARS-CoV-2 into Peru. J Med Virol. 2021;93(10):5961–5968.

19. McLaughlin A, Montoya V, Miller RL, Mordecai GJ, Worobey M, Poon AFY, et al. Early and ongoing importations of SARS-CoV-2 in Canada. medRxiv. 2021; p. 2021.04.09.21255131.

20. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121–4123.

21. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 2014;10(2):e1003932.

22. Magee D, Suchard MA, Scotch M. Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction. PLoS Comput Biol. 2017;13(2):e1005389.

23. De Maio N, Wu CH, O'Reilly KM, Wilson D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. PLOS Genetics. 2015;11(8):e1005421.

24. Kalkauskas A, Perron U, Sun Y, Goldman N, Baele G, Guindon S, et al. Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. PLoS Comput Biol. 2021;17(1):e1008561.

25. Magee D, Scotch M. The effects of random taxa sampling schemes in Bayesian virus phylogeography. Infect Genet Evol. 2018;64:225–230.

26. Guindon S, De Maio N. Accounting for spatial sampling patterns in Bayesian phylogeography. Proc Natl Acad Sci U S A. 2021;118(52).

27. Genome Canada Canadian COVID-19 Genomics Network (CanCOGeN) and the Canadian Public Health Laboratory Network CanCOGeN Working Group. Canadian national COVID-19 genomics surveillance priorities for existing and emerging variants of concern. Can Commun Dis Rep. 2021;47(3):139–141.

28. Maddison WP, Midford PE, Otto SP. Estimating a Binary Character's Effect on Speciation and Extinction. Systematic Biology. 2007;56(5):701–710.

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

30. Pagel M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society of London Series B: Biological Sciences. 1994;255(1342):37–45.

31. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018;35(3):526–528.

32. Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. Nature. 2015;524(3):97–101.

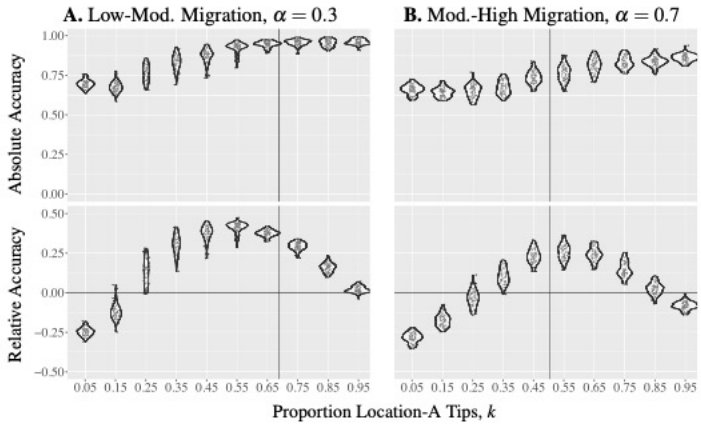
33. Stadler T, Bonhoeffer S. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Philosophical Transactions of the Royal Society B: Biological Sciences. 2013;368(1614):20120198.

Supplementary Material

Absolute and Relative Reconstruction Accuracy

Here we present complementary results and sensitivity analyses in which we choose different values for fixed parameters and trees.

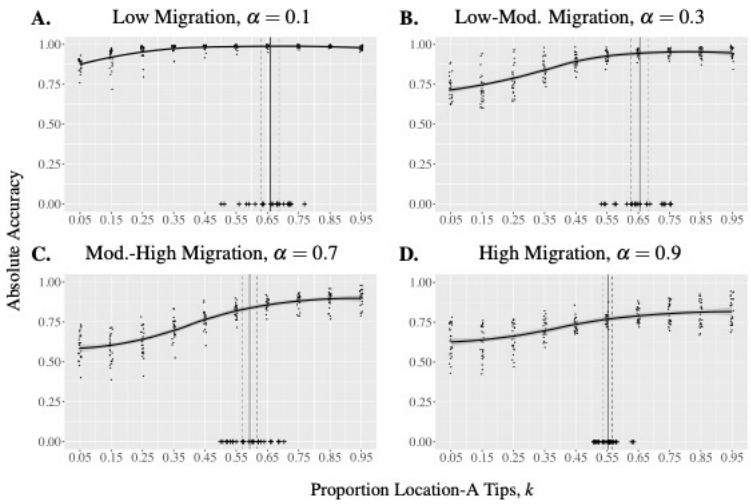
Fig. S1 shows the absolute and relative accuracy for the true trees displayed in Fig. 4A and Fig. 4C, in parallel with Fig. 1B and Fig.S1D.



Supp Fig S1. Reconstruction accuracy for intermediate migration rates. Absolute and relative accuracy for intermediate migration rates. True trees are shown in Fig. 4A (Panel A) and Fig. 4C (Panel B). The vertical lines indicate the true proportions of location-A tips. Parameters: $\lambda = 4$ and $\mu = 1$.

Whereas in the main text we downsampled many times from the same fixed true tree, here in Fig. S2 we show the absolute reconstruction accuracy and how it varies with sampling bias using 25 true trees for each choice of the migration rate; we downsample 50 times for each tree.

Each edge in the true tree has two endpoints, that is, the parent node of the edge near the root and the child node near the tips. For an edge in a downsampled tree of a true tree, each node has a true location from simulation and a reconstructed location (a likelihood of being location A) from the function `ace`. The state of the edge with the true locations at the endpoints can be A-A, A-B, B-A or B-B. We count an edge with a location-A parent node and a location-A child node as one edge in the state A-A and analogously for the other states. Whereas the state of the edge with reconstructed locations at the endpoints is computed by multiplying the likelihood of the parent node being at location A (or B) and the the likelihood of the child node being at location A (or B). For example, if the parent node has a likelihood 0.3 of being at location A and the child node has a likelihood 0.1 of being at location A, then we count the edge as 0.03 edge in the state A-A, 0.27 edge in the state A-B, 0.07 edge in the state B-A and 0.63 edge in the state B-B. We compute the number of edges with true locations at endpoints in each state in downsampled trees with different proportions of location-A



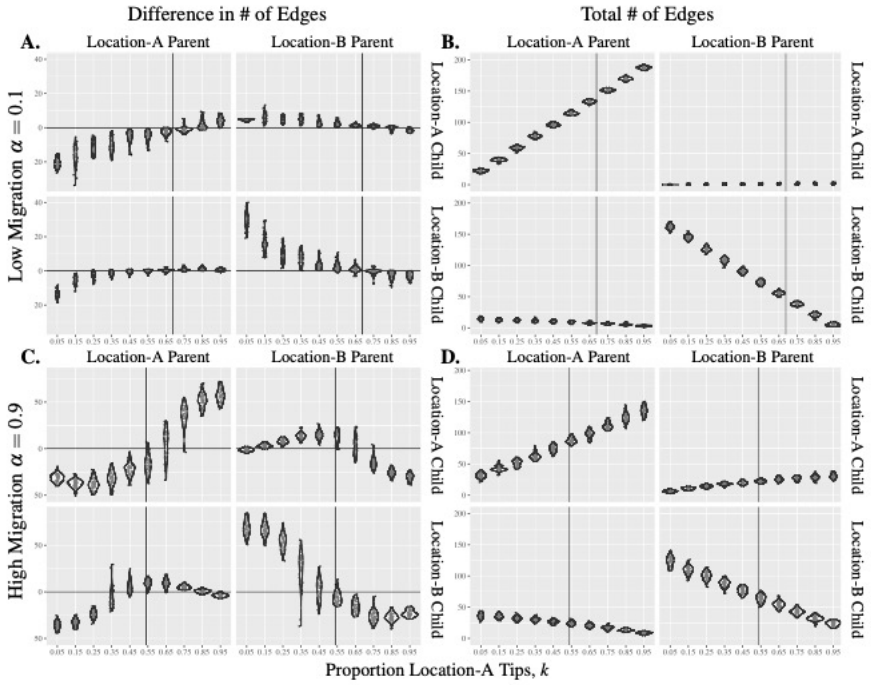
Supp Fig S2. Reconstruction accuracy across true trees. Reconstruction accuracy for each of 25 true trees calculated by averaging the absolute accuracy of 50 downsampled trees per true tree. Curve displays the mean accuracy across true trees, shaded band represents the 95% confidence interval of the mean. The plus signs indicates the true proportion of location-A tips of each of the 25 true trees and the vertical lines represent the mean (solid) proportion of location-A tips in the 25 true trees and the corresponding 95% confidence interval of the mean (dashed). Parameters: $\lambda = 4$ and $\mu = 1$.

tips for the true trees simulated with different migration rates. We also compute the difference of subtracting the number of edges with true locations at endpoints from the number of edges with reconstructed locations at endpoints in each state for the true trees. The results are displayed in Fig. S3 and Fig. S4

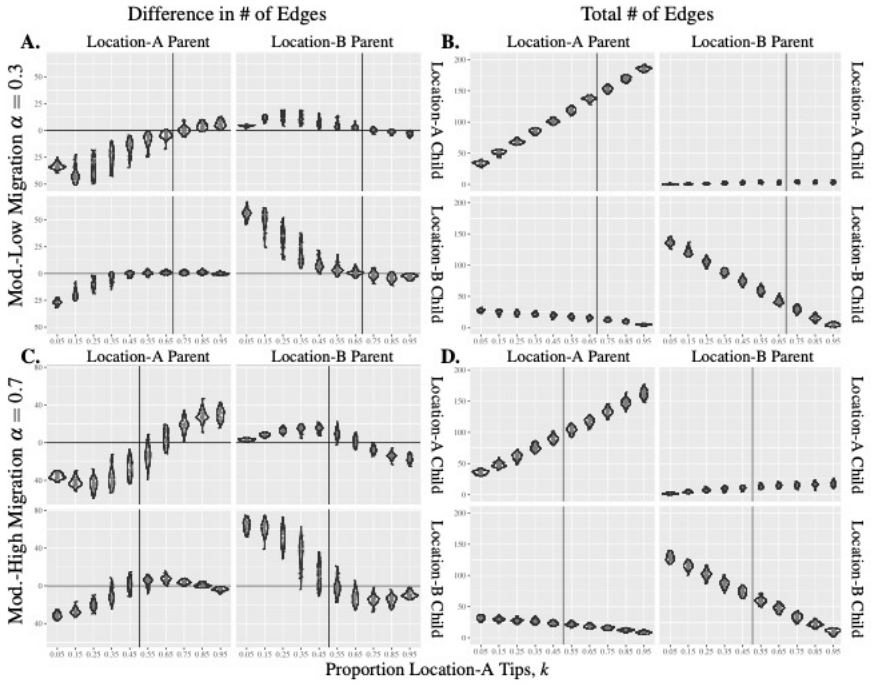
Furthermore, in the main text, we used a neutral branching process in which the branching and death rates did not depend on the location. Here, we simulate under with non-neutral speciation rates. Fig. S5 displays the absolute and relative accuracy for two true trees simulated with $\lambda_A = 8, \lambda_B = 4$ and $\lambda_A = 4, \lambda_B = 8$ respectively. We find that the relative accuracy is highest when the location with a lower speciation rate is oversampled compared to its representation in the true tree (few of the higher-speciation-rate location-A tips in Fig. S5A, and more of the now lower-speciation-rate location-A tips in Fig. S5B). In both cases the relative accuracy is highest when the fraction of tips from the lower speciation rate location is 45-65%, whereas due to having a lower speciation rate, that location has fewer tips in the true tree.

Oversampling Recent Migrants

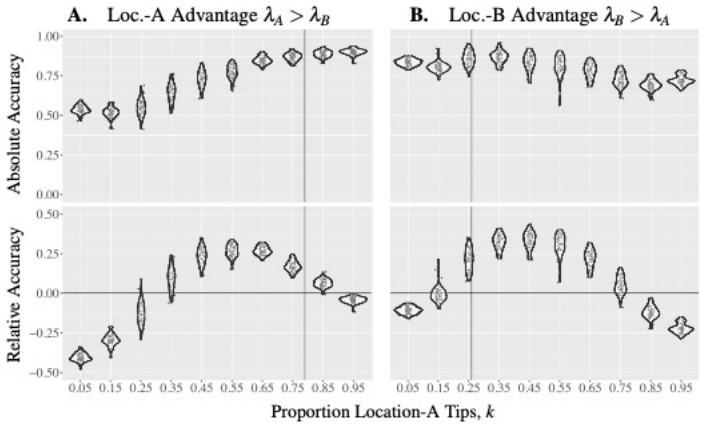
We demonstrate additional results regarding absolute and relative accuracy of downsampled trees with different proportions or recent migrants, where the downsampled trees have 25% and 75% location-A tips instead of 50%. Fig. S6 shows



Supp Fig S3. Characterization of tree edges with low versus high migration rates. Panel A & C: Difference between the number of edges with reconstructed versus true parent and child locations. Panel B & D: The number of edges with a given parent and child locations in the true tree.

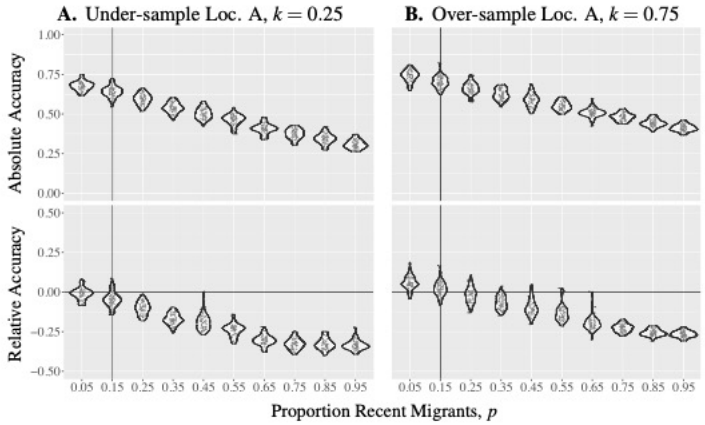


Supp Fig S4. Characterization of tree edges with intermediate migration rates. Panel A & C: Difference between the number of edges with reconstructed versus true parent and child locations. Panel B & D: The number of edges with a given parent and child locations in the true tree.



Supp Fig S5. Reconstruction accuracy for with location-dependent speciation. Absolute and relative accuracy of downsampled trees from true trees simulated with non-neutral speciation rates, where each data point represents the absolute or the relative accuracy of a downsampled tree from the true tree simulated with $\lambda_A = 8, \lambda_B = 4, \mu = 1, \alpha = 0.3$ (Panel A) or the true tree simulated with $\lambda_A = 4, \lambda_B = 8, \mu = 1, \alpha = 0.3$ (Panel B). The vertical lines indicate the proportions of location-A tips in the true trees.

the results compared to Fig. 2A and Fig. 2B, and we observe the same pattern that the accuracy decreases as the proportion of recent migrants increases.



Supp Fig S6. Effect of over/under-sampling recent migrants with different proportions of location-A tips. Absolute and relative accuracy of downsampled trees with different proportions of recent migrants p and either 25% location-A tips (Panel A: $k = 0.25$) or 75% of location-A tips (Panel B: $k = 0.75$). Each data point represents the absolute or the relative accuracy of a downsampled tree from the true tree displayed in Fig. 1C. The vertical lines indicate the proportion of recent migrants in the true tree.

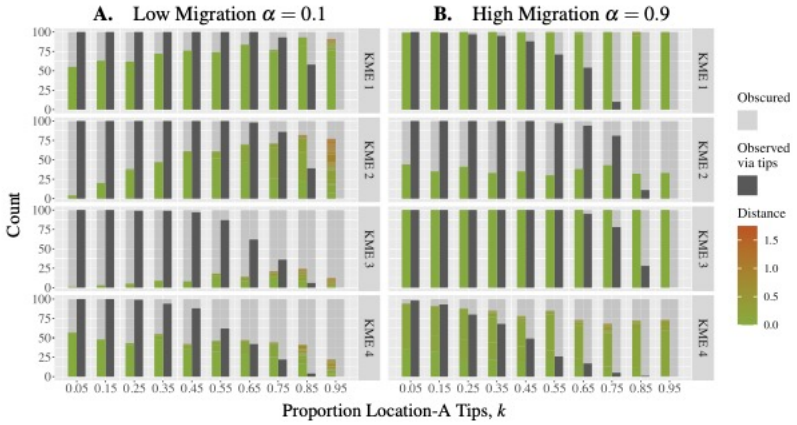
Key Migration Events

We show complementary results about key migration events (KMEs) to Fig. 4B and Fig.4AD in Fig. S7, where we count the number of downsampled trees in which a KME of the true tree is obscured, observed or erred for true trees displayed in Fig. 1A and Fig. 1C respectively.

We perform additional experiments to investigate the absolute accuracy of the parent and the child internal nodes of KMEs. Fig. S8 shows the absolute and relative accuracy of the internal nodes of KMEs. We observe that if we sample more location-A tips, then the child nodes of KMEs are more likely to be reconstructed with the wrong location, and the KMEs are likely to be obscured via internal nodes. Similarly, if we sample more location-B tips, then the parent nodes of KMEs are more likely to be reconstructed with the wrong location, and the KMEs are also likely to be obscured via internal nodes. Therefore, the unbiased geographic sampling performs better in identifying key migration events.

Application to the 2014-2015 Ebola Epidemic

Fig. S9 and Fig. S10 shows the results of downsampled 80% of the Sierra Leone and Guiena tips respectively. In both cases, thee overall absolute accuracy is high. We obtained 10 downsampled trees for each location and calculate average absolute



Supp Fig S7. The KME count for the true tree with migration rate 0.1 displayed in Fig. 1A and the true tree with migration rate 0.9 displayed in Fig. 1C respectively, where the left bars show the number of downsampled trees in which a KME of the true tree is observed (zero-distance green bars), erred (non-zero-distance color bars) and obscured (transparent grey bars) via internal nodes, and the right bars show the number of downsampled trees in which a KME of the true tree is observed (solid grey bars) and obscured (transparent grey bars) via tips.

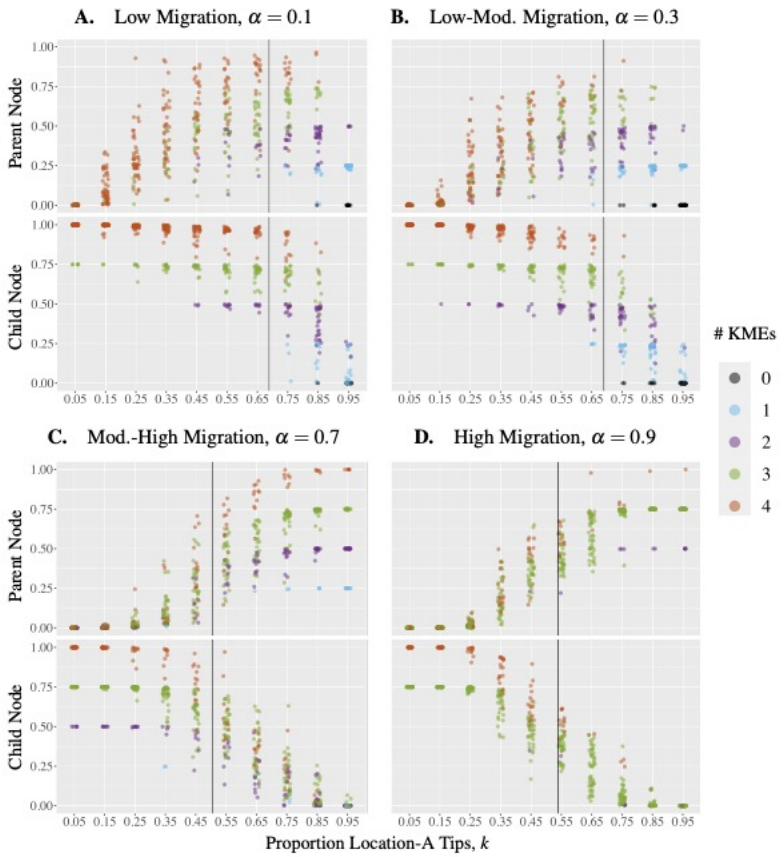
accuracy. The average absolute accuracy is 99% when downsampling Sierra Leone tips and 94% when downsampling Guinea tips.

Alternative Simulation

We show that the two approaches to simulating geographic sampling, generating true trees then downsampling and downsampling along birth-death processes, give similar results for the absolute accuracy. Fig. S11A shows the results of simulating geographic sampling with the first approach. Specifically, we simulate a single true tree with parameters $\lambda = 4$, $\mu = 1$, $\alpha = 0.7$, and we downsample 50 times from the true tree for each proportion of location-A tips. Fig. S11B shows the results of the second approach, where we run 50 birth-death processes with sampling using parameters $\lambda = 4$, $\mu = 1$, $\alpha = 0.7$ for each proportion of location-A tips. The function *sim.bdtypes.stt.taxa* in the R package *TreeSim* is used to realize the simulations, and the function produces a downsampled tree for each birth-death process, so we have in total 50 downsampled trees. We observe that the two approaches to simulating geographic sampling produce similar results regarding absolute accuracy.

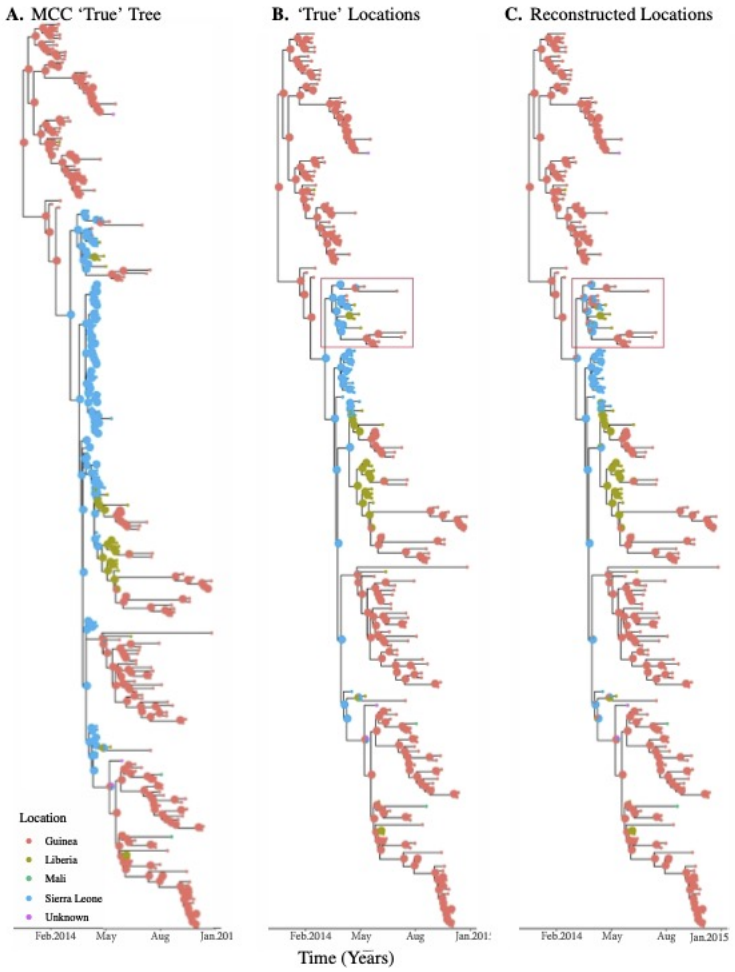
Epidemic Origin

We also explored how geographic sampling impacts the reconstruction of the location and the time of the root (Fig. S12). We find that geographic sampling bias can have a dramatic impact on the shows that the reconstructed state of the root depends on the proportion of location-A tips in downsampled trees. Since we begin simulations with a

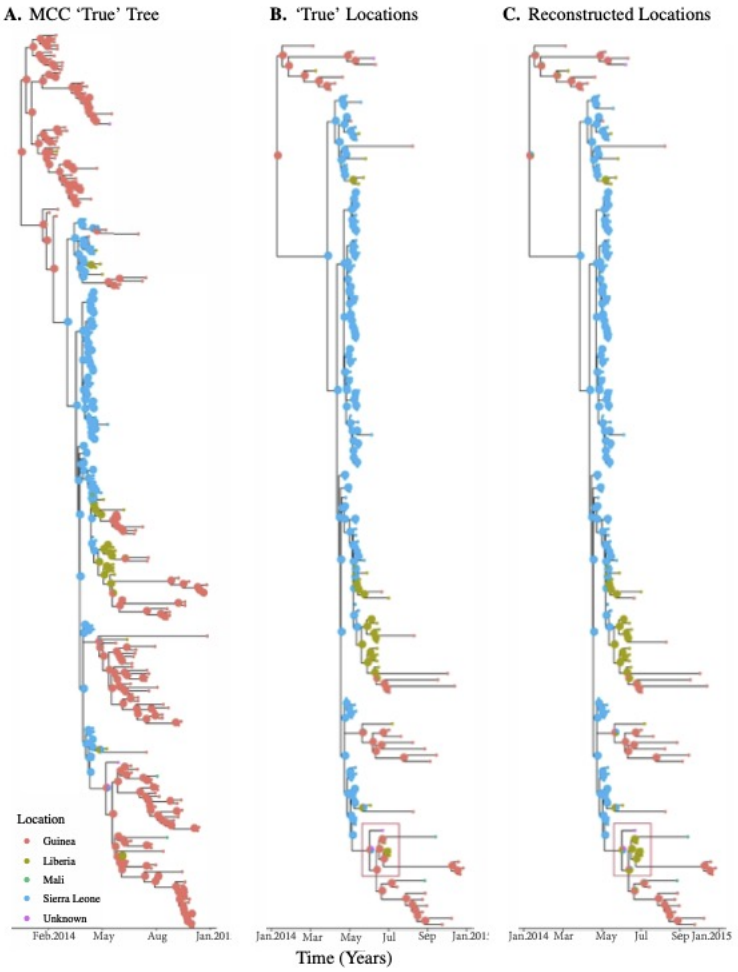


Supp Fig S8. Reconstruction accuracy of KME. Absolute accuracy of the parent and the child nodes over all observed (via internal nodes) KMEs in the true tree in Fig. 1A with migration rate 0.1 (Panel A), the true tree in Fig. 4A with migration rate 0.3 (Panel B), the true tree in Fig. 3C with migration rate 0.7 (Panel C), and the true tree in Fig. 1C with migration rate 0.9 (Panel D).

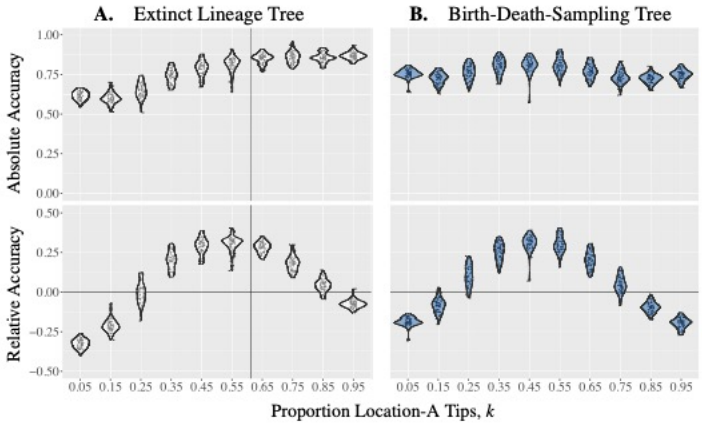
root at location A, the more location-A tips are sampled, the more accurate (i.e. location A) the root location is. Fig. S13 also shows that the absolute accuracy of the root location is more accurate when the migration rate of the true tree is low, presumably because more location A tips occur in more monomorphic clades. The reconstructed time of the root also depends on whether there are extinction events or small clades near the root. If there are extinction events (hence tips) or small clades near the root and these are unsampled, then the reconstructed time of the root can be inaccurate.



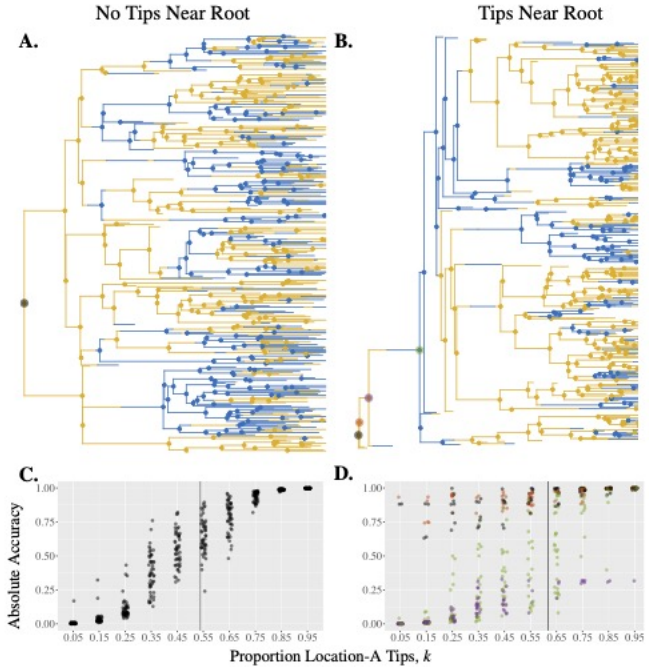
Supp Fig S9. Ancestral state reconstruction of Ebolavirus trees with Sierra Leone downsampled. The Maximum Clade Credibility (MCC) phylogenetic tree the Ebolavirus outbreak generated using BEAST, which is used as a true tree (Panel A). Downsampled tree obtained by dropping 80% Sierra Leone samples with true ancestral location (Panel B). The downsampled tree with reconstructed ancestral location after dropping Sierra Leone samples (Panel C).



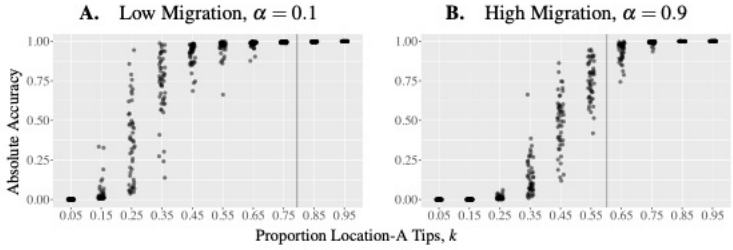
Supp Fig S10. Ancestral state reconstruction of Ebolavirus trees with Guinea downsampled. The Maximum Clade Credibility (MCC) phylogenetic tree for the Ebolavirus outbreak generated using BEAST, which is used as a true tree (Panel A). Downsampled tree obtained by dropping 80% Guinea samples with true ancestral location (Panel B). The downsampled tree with reconstructed ancestral location after dropping Guinea samples (Panel C).



Supp Fig S11. Reconstruction accuracy of trees of extinct lineages trees versus trees simulated from a birth-death-sampling process. Absolute and relative accuracy of trees with different proportions of location-A tips. Panel A: Each data point in represents the absolute or the relative accuracy of a downsampled tree from a single true tree simulated with $\lambda = 4$, $\mu = 1$, $\alpha = 0.7$. The downsampled trees are obtained after the true tree is simulated, and the vertical lines indicate the proportion of location-A tips in the true tree. Panel B: Each data point in represents the absolute or the relative accuracy of tree obtained along a birth-death process with $\lambda = 4$, $\mu = 1$, $\alpha = 0.7$.



Supp Fig S12. Accuracy of root reconstruction depends on tree shape. True trees with no tips near the root (Panel A) and with tips near the root (Panel B), where possible reconstructed root nodes are highlighted with different colors. Both true trees are simulated with $\lambda = 4$, $\mu = 1$, $\alpha = 0.9$. Reconstructed root node (color) and absolute accuracy of the root node in downsampled trees from the true trees are displayed in (Panel C) and (Panel D) respectively, where each data point represents the absolute accuracy of the root in a downsampled tree. The vertical lines indicate the proportions of location-A tips in the true trees.



Supp Fig S13. Accuracy of root reconstruction depends on migration rate. Absolute accuracy of the root in downsampled trees with different proportions of location-A tips. Each data point represents the absolute or the relative accuracy of a downsampled tree from true trees simulated with $\lambda = 4$, $\mu = 1$, $\alpha = 0.3$ (Panel A) or $\alpha = 0.7$ (Panel B). The true trees have no tips near the roots, and the vertical lines indicate the proportions of location-A tips in the true trees.