Lost in the woods: shifting habitats can lead phylogeography astray

Richard A. Neher

Swiss Institute of Bioinformatics, and Biozentrum, University of Basel, Switzerland

(Dated: July 3, 2024)

Continuous phylogeographic inference is a popular method to reconstruct the spatial distribution of ancestral populations and estimate parameters of the dispersal process. While the underlying probabilistic models can be complex and their parameters are often computationally demanding to infer, these models typically ignore that replication and population growth are tightly coupled to spatial location: populations expand into fertile uninhabited areas and contract in regions with limited resources. Here, I first investigate the sampling consistency of popular summary statistics of dispersal and show that estimators of "lineage velocities" are ill-defined. I then use simulations to investigate how local density regulation or shifting habitats perturb phylogeographic inference and show that these can result in biased and overconfident estimates of ancestral locations and dispersal parameters. These, sometimes dramatic, distortions depend in complicated ways on the past dynamics of habitats and underlying population dynamics and dispersal processes. Consequently, the validity of phylogeographic inferences, in particular when involving poorly sampled locations or extrapolations far into the past, is hard to assess and confidence can be much lower than suggested by the inferred posterior distributions.

As organisms replicate, they also disperse in space, resulting in mixing of the population and exploration of new habitats. The reconstruction of ancestral locations and past migrations from sampling locations of extant individuals, typically in combination with genome sequence information to infer the phylogeny, is known as phylogeography. Phylogeographic methods are implemented in popular evolutionary analysis software such as BEAST (Lemey *et al.*, 2009, 2010).

In phylogeographic inference, one can either assume that individuals migrate between discrete locations or disperse in continuous space. Here, we consider migration in continuous space, where the underlying model is typically assumed to be diffusive, that is the probability of sampling a descendant at position $\mathbf{r}_c$ after time t when the parent was at position $\mathbf{r}_p$ is given by

$$P(\mathbf{r}_c|\mathbf{r}_p, t) = \frac{1}{4\pi D t} e^{\frac{|r_c - r_p|^2}{4Dt}} \tag{1}$$

where D is a diffusivity with dimensions length^2/time, we have assumed a two-dimensional space, and we have assumed dispersal is isotropic. Such a diffusive process is the simplest and most natural choice in absence of directed motion or long range migration. It is also mathematically convenient as unknown ancestral locations can be integrated in closed form and the marginal distribution of each position can be calculated exactly for a fixed tree along with the spatial likelihood. But this model assumes dispersal of one lineage is independent of other lineages, that dispersal properties are homogeneous across the habitat, and that the habitat does not change over time. Furthermore, it assumes that the tree topology and the branching rates of the tree are independent of the spatial location. This independence assumption is for example manifest in coalescent models assuming "exchangeability" of individuals, or birth-death models assuming identical rates on contemporary branches of the tree.

However, population dynamics are often tightly coupled to spatial location. Changes of the environment could mean that a habitat of a population is shifting, resulting in population growth and rapid branching at the expanding edge of the habitat and/or contraction in other areas. Similarly, invasive species or pathogens spread into areas that support high population densities but were previously unoccupied. Examples of such shifting habitats are common and range from geological time scales, over decades (climate change), to weeks (seasonal fluctuations). Such changes are not restricted to the physical environment, but ecological shifts can similarly affect the habitat. In all such cases, the replication process determining the phylogeny of the sample and the spatial location are strongly coupled.

Such coupling can be accommodated in inference methods based on structured models that divide the population into discrete demes that could correspond to spatial locations that exchange migrants. Such models are flexible enough to allow for different population dynamics in different demes (Vaughan *et al.*, 2014), though in practice can struggle to represent the population in sufficient detail or suffer from identifiability problems (Layan *et al.*, 2023). But the extent to which violations of the independence assumption affect continuous phylogeographic inferences is not well understood.

In a typical phylogeographic inference problem, the input data are genome sequences as well as their sampling times and sampling locations. Software like BEAST will then sample the posterior distribution of the phylogenetic tree topology, the sampling times and locations of ancestral nodes in the tree, and model parameters like the diffusion constant D and the evolutionary rate μ. The inferred quantities of interest could be estimates of the ancestral location of a particular groups of samples, or summary statistics of the dispersal process. Here, I

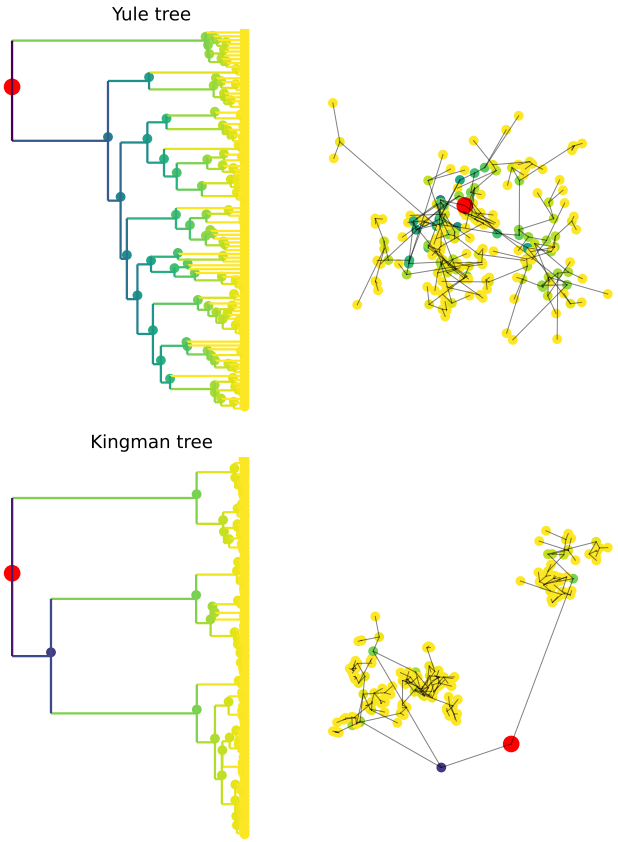

FIG. 1 **Illustration Yule and Kingman trees with $N = 100$ tips and the spatial location of the tips.** The total tree length of Yule trees is proportional to the number of tips n and the time to the MRCA is $\log(n)$. Kingman trees have a total tree length that is proportional to $\log(n)$ and are characterized by rapid merging close to the present. Temporal and spatial scales are determined by parameter choice, only tree shape and spatial patterns are meaningful.

explore what dispersal quantities can be robustly estimated and how sensitive inferences are to the fact that coupling between population dynamics and spatial location is ignored. My focus here is only on estimates of dispersal parameters and ancestral locations and I will use simulations with perfect knowledge of the tree and the sampling locations. So no tree reconstruction will be necessary and sampling is uniform across the population. This simplification is justified since the tree tends to depend much more strongly on sequence information and spatial information and will allow clearer conclusions. In practice, tree reconstruction and sampling biases can create additional uncertainty and biases. Furthermore, I will assume strictly diffusive dispersal without any long-range migrations.

Consistency of summary statistics of dispersal parameters.

Before investigating the more complex issues of coupling between spatial locations and reproduction, I will first explore the properties of summary statistics in simple models where spatial location and population dynamics are independent. Popular summaries of the dispersal process are empirical estimates of the diffusion constant D_w (Pybus *et al.*, 2012; Trovão *et al.*, 2015) and a so-called *dispersal velocity* v_w (Dellicour *et al.*, 2017; Lemey *et al.*, 2010). These are calculated from the displacements along branches of the trees sampled from the posterior. For each branch i with parent p_i, the displacement $d_i = |\mathbf{r}_{p_i} - \mathbf{r}_i|$ and the elapsed time Δt_i are used to calculate

$$D_w = \frac{\sum_{i\in B} d_i^2}{4\sum_{i\in B}\Delta t_i} \quad \text{and} \quad v_w = \frac{\sum_{i\in B} d_i}{\sum_{i\in B}\Delta t_i} \tag{2}$$

where B is the set of all branches of the tree. Effectively D_w divides the sum of all observed squared displacements by the total time elapsed on the tree to obtain an estimate of D. It is known as the 'weighted' estimate of the diffusion constant. The weighted dispersal velocity estimate compares the sum of absolute values of observed displacements to the total time, which has dimensions of a velocity. Alternative formulations use the mean of fractions $D_b = |B|^{-1}\sum_{i\in B} d_i^2/4\Delta t_i$ or $v_b = |B|^{-1}\sum_{i\in B} d_i/\Delta t_i$ instead of the ratio of sums. We refer to these here as 'by branch' estimates. The latter tend to be dominated by short branches and thus noisier (Trovão *et al.*, 2015).

The estimators in Eq. 2 are commonly used to summarize dispersal of lineages. While the estimator D_w directly estimates a parameter of the model and its behavior has been studied in simulations (Pybus *et al.*, 2012), it is unclear what v_w is measuring and how it behaves. I simulated diffusion along trees from Kingman coalescent (Kingman, 1982) and Yule model (Yule, 1925) tree ensembles and evaluated the different estimators using the known true ancestral locations. Kingman trees correspond to the classical neutral model of a constant size population, while Yule trees capture features of exponentially growing populations (see Fig. 1). Examples for each ensemble are shown in Fig. 1.

Fig. 2 shows empirical estimates of diffusion constants and velocity as defined in Eq. 2 from simulated data using freely diffusing locations along Yule and Kingman trees. The diffusion constant D can be estimated robustly from these data and the estimates are compatible with the diffusion constant used to simulate the data $D = 1$. Values of v_w and v_b estimated from Kingman trees, however, depend strongly on the sample size and are incompatible with each other. That the velocity estimates are ill-defined is not surprising: By changing the sample size, the average length of branches in Kingman coalescent trees changes. Diffusion along each branch results in a displacement of $d_i \sim \sqrt{4D\Delta t_i}$ such that a quantity that depends on $d_i/\Delta t_i \sim 1/\Delta t_i^{1/2}$ will change with the time Δt_i that elapsed along the branch. With more samples

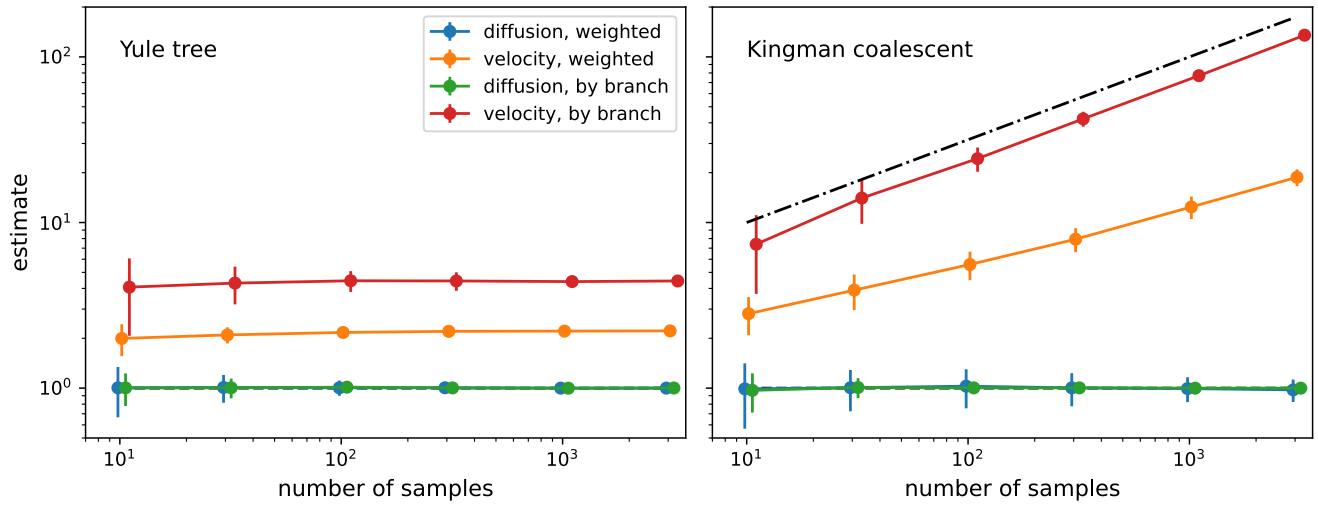

FIG. 2 **Empirical summary statistics of D and "dispersal velocity" for different sample sizes.** While the diffusion constant D (equal to 1 in this example) can be robustly estimated from simulation data of the underlying model, the dispersal rate v_w (weighted) and v_b (by branch) are not well-defined quantities. For Kingman coalescent tree their numerical value depends strongly on sample size and there is no ground truth to compare to. The expected scaling $v_w \sim \sqrt{n}$ of v estimates is indicated by the dashed line.

in the tree, the average branch length decreases, and the velocity estimate increases. This expected scaling $\sim \sqrt{n}$ is indicated by the dashed line in right panel of Fig. 2. For Yule trees, the average branch length is independent of the number of samples and the velocity estimates are more stable. But as for the Kingman case, the estimates for v_w and v_b are not consistent with each other.

In either case, estimates of v_w or v_b are not meaningful. The underlying model is diffusive and has no parameters that individually or combined would have dimensions of a velocity. There is no ground truth to compare the estimates to and v_w and v_b are just descriptive summaries of spatial spread that can not be compared between datasets.

While the estimator D_w as defined in Eq. 2 is well-behaved when the model is exactly diffusive, long range dispersal can affect D_w dramatically. If the probability of long jumps with distance $x < d < x + dx$ decays more slowly than $\sim dx/x^3$, the expectation value of the D_w diverges. In this case, the values of the above estimators for a dataset with long range dispersal are dominated by the longest displacements and a single number like D_w is no longer a useful summary of the typical dispersal process. Estimators like D_w might also become problematic under the assumptions of relaxed random walk models with long-tailed prior distributions for D.

Effects of density regulation on population structure and dispersal

Phylogeographic inference typically assumes that the spatial diffusion process is independent of the branching pattern of the tree and ignores coupling between the two. Such a coupling, however, is expected since organisms compete for resources and growth rates naturally depend on local population density. It has been known for a long time that ignoring such coupling leads to unrealistically heterogeneous population densities (Felsenstein, 1975), as is evident in the example of the Kingman tree in Fig. 1. If one simulates a diffusive spatial process in a square of length L with periodic boundary conditions, the population density shows strong fluctuations whenever the diffusion constant $D \ll L^2/T_c$, where T_c is the coalescent timescale of the population ($T_c = N$ in the case of a neutral model with a population of size N). Fig. 3A (black line) quantifies this heterogeneity as the standard deviation of population density in two-dimensional bins of size $L/5$ relative to the expectation in the case of a spatially well-mixed population when each individual picks its location at random. For $D < L^2/2T_c$, diffusion is too slow to explore the entire habitat during the time since the MRCA, leading to clustering of individuals. This heterogeneity increases substantially for even lower D, when the population is essentially a cloud of width $\sqrt{4DT_c} \ll L$. The comparison of D to the ratio L^2/T_c will continue to be important below, and I therefore use diffusion constants in units of L^2/T_c.

Biological populations spread into areas that support them and will quickly populate fertile areas that are below their carrying capacity. A simple model for the population density $\phi(\mathbf{r}, t)$ would be

$$\frac{\partial \phi(\mathbf{r},t)}{\partial t} = D\Delta_{\mathbf{r}}\phi(\mathbf{r},t) + \alpha\phi(\mathbf{r},t)(1 - \phi(\mathbf{r},t)/\phi_0) \quad (3)$$

where $\Delta_{\mathbf{r}}$ is the two-dimensional Laplace operator, ϕ_0 is the carrying capacity, and α is the population growth

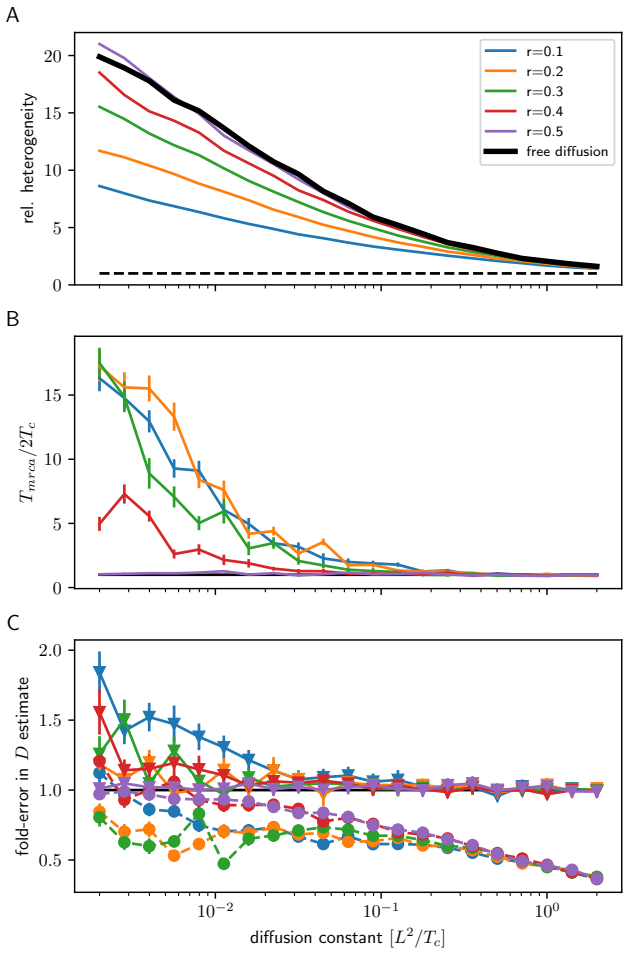

FIG. 3 **Effects of population density regulation.** When spatial motion is independent of population density and coalescence is sufficiently rapid, the realized population density fluctuates strongly across the habitat. **Panel A** shows the standard deviation of the population density in bins of linear dimension $L/5$ as a function of the diffusion coefficient relative to the expected value if individuals were independently distributed across the habitat. The thick black line shows the case of free diffusion, the colored lines density regulation with different interaction radii. For $r \ll L$, density fluctuations are strongly suppressed. At the same time, the time to the most recent common ancestor increases dramatically to values several fold higher than the neutral coalescence timescale $T_c = N = 1000$ (**panel B**), indicating that the population fragments in subpopulations that don't interact. **Panel C** shows the fold-error of the dispersal estimator D_w for periodic boundary conditions (triangles, solid lines, like panels A&B) and reflecting boundaries (circles, dashed lines).

rate at zero density. Such logistic density regulation was proposed by Bolker and Pacala (1997) and further analyzed by Etheridge (2004), but goes back to much earlier work by Fisher (1937) and Kolmogorov *et al.* (1937). The density has a stable fixed point at $\phi = \phi_0$ we thus expect the population density to settle at a constant density ϕ_0 and density fluctuations to decay with rate α.

While the dependence of growth (and thus tree structure) on density is a key component of real-world population dynamics, it is difficult to implement in phylogeographic inference frameworks. It is, however, fairly straightforward to include density dependence in forward simulations and these simulations can be used to investigate the effect of such coupling between growth and spatial location on phylogeographic inferences.

Instead of the coalescent trees used above, I now used an individual based simulation forward in time. Individuals are assigned a fitness $f = \alpha(1 - \hat{\phi}/\phi_0) + (1 - N/N_0)$ and produce a Poisson distributed number of offspring with mean $1 + f$. The first term determines the fitness due to the local population density and has the tendency to stabilize this density at ϕ_0, the second term $(1-N/N_0)$ keeps the overall population close to the target population size N_0. The population density is calculated using a Gaussian kernel with interaction radius r

$$\hat{\phi}(\mathbf{r}, t) = \sum_{i \in N} \frac{e^{-|\mathbf{r}_i - \mathbf{r}|^2/2r^2}}{\sqrt{4\pi r^2}} \tag{4}$$

where the sum is over all individuals i in the population of size N and $\mathbf{r}_i$ is location of individual i at time t. The spatial locations of the offspring are chosen from a 2-D Gaussian centered on the parent with a variance to match the diffusion constant. The simulation records the tree of population, including their ancestral location and times. In addition to the true location of ancestral nodes, I also infer their locations from the locations of the extant samples. For this inference, I make the common assumption of continuous phylogeography that growth and location are uncoupled and that dispersal is diffusive. In this case, the marginal distributions of ancestral locations and their mean and variance can be calculated exactly using recursion relations.

The radius r of the density estimate $\hat{\phi}(\mathbf{r}, t)$ in Eq. 4 determines the distance over which different individuals compete for resources. As soon as r is much smaller than the habitat size, this density regulation leads to a more homogeneous population density, see Fig. 3A. But with decreasing diffusion, the population fragments into smaller and smaller subpopulations that interact only weakly and coalesce slowly. This manifests in an increasing time to the MRCA, see Fig. 3B, which at low D is limited by the time the population was simulated for. Spatial location and the coalescence process are thus strongly coupled.

Despite the substantial reduction in density fluctuations and strong population structure, the estimates D_w of the diffusion constant using Eq. 2 are not dramatically affected by density regulation, see Fig. 3C. If boundary conditions are periodic, that is the habitat has the topology of a torus, estimates D_w of the diffusion constant are accurate for most values of D but slightly inflated and noisy when the population is heavily fragmented (the increased noise is due to a smaller number of effective samples when tMRCA is large). If instead of periodic

boundary conditions reflective boundary conditions are used, diffusion constants are underestimated once $\sqrt{2DT_c}$ becomes comparable to the habitat size. This underestimation of diffusion constants will then result in overconfident inference of ancestral locations, in particular for recent ancestral nodes.

These simulations reveal that both at high and at low diffusion constants, assumptions of phylogeographic inference can be problematic: if diffusion constants are much smaller than L^2/T_c, the population is fragmented into effective subpopulations and this fragmentation violates the assumptions of phylodynamic models. A structured model would be more appropriate in this case. If $D \gg L^2/T_c$, the boundaries of the habitat restrict free diffusion, resulting in underestimated dispersal parameters, while estimates of deeper ancestral locations have uncertainties comparable to the habitat size and are thus uninformative.

How do changing habitats affect phylogeographic inferences?

As discussed above, even a static carrying capacity in an equilibrium population can strongly affect population structure, though this structure has limited effects on estimates of ancestral locations and diffusion coefficients beyond boundary effects. In reality, carrying capacities vary through time. Such variation happens on all time scales, ranging from geological timescales over millions of years, environmental changes over millennia to seasonal fluctuations. The circulation of seasonal influenza viruses in temperate climates, for example, varies by many orders of magnitude between summer and winter.

Fluctuations in carrying capacity mean that populations not only spread because individuals move, but also because populations grow in regions where the population density is below the carrying capacity. In such situations, the accuracy and interpretation of phylogeographic inferences are unclear. In Fig. 4 we consider a case where the region with highest carrying capacity shifts from the left to the right in a periodic manner with period T. The carrying capacity is illustrated in Fig. 4A for 20 time points covering half a period. In between the two extreme left/right concentration of the carrying capacity, the carrying capacity is uniform. To avoid extinction at low dispersal rates, selection is soft, meaning that the overall population size is kept approximately constant even if the entire population is stuck in the unfavorable region. If the dispersal rate is sufficiently high, however, the population follows the shifting carrying capacity and lineages in the tree undergo directed back-and-forth motion. Inference of ancestral locations in such situations can be misleading as illustrated in Fig. 4C. The population was sampled at a time when the carrying capacity is flat after having contracted to the left and the positions of older nodes are not estimated correctly. Their true locations are to the right of the current population, but the inferred locations are central in the shifted population. This is expected as there is no information in the current sample to suggest otherwise, but it highlights the potential of misleading inference, in particular when extrapolating far into the past.

The severity of such biases depends on the period T at which the carrying capacity oscillates and on the dispersal rate and tends to be strongest when the period is of the same magnitude as the coalescent time scale. Fig. 4D shows the estimated diffusion constant (using Eq. 2) in x (solid lines) and y (dashed) direction relative to the true value used in the simulations. At very high diffusion constants, both D_x and D_y are underestimated due to the spatial constraints of carrying capacity as also discussed above (see Fig. 3). With decreasing D, D_x starts to be overestimated and peaks at a value that depends on the period T, before decreasing to (noisy) estimates that are broadly compatible with the true value. In this range of dispersal rates, the population can no longer follow the shifting carrying capacity and the T_{mrca} increases sharply (not shown). This decoupling happens at lower values of D for longer periods of environmental change. With the parameters used in the simulation, the estimates of D_x deviate by a factor of 2 from its true value, which is not dramatic compared to the variation of dispersal rates in nature, but clearly shows that spatial inference can be substantially affected if the habitat changes on time scales comparable to the T_{mrca}.

Along with problematic inference of diffusion coefficients, estimates of ancestral locations are over-confident. Panel Fig. 4E shows $z_x^2 = \langle(\hat{x} - x)^2/\sigma_x^2\rangle$ and $z_y^2 = \langle(\hat{y} - y)^2/\sigma_y^2\rangle$ for the oldest 20% of internal nodes, where $\hat{x}$ and σ_x are the mean and standard deviation of the posterior of x, respectively (and analogously for y). If the model estimated ancestral locations consistently, z_x^2 should be 1 on average. Instead, it is as high as 20 for the simulations shown and these over-confident and biased estimates depend in non-monotonic ways on the true diffusion constant and the period of environmental change. With decreasing D, z_x^2 increases largely because the expected uncertainty decreases and the strongest over-confidence is observed when the dispersal capacity of the population is just about sufficient to follow the shifting carrying capacity. A peak value of $z_x^2 \approx 10 - 20$ indicates that the typical ancestral location of the deep nodes is 3-5 standard deviations away from the true location, an example of which is shown in Fig. 4C. Conversely, at high dispersal rates, the average z_x^2 and z_y^2 are both smaller than one, i.e. the inference is not confident enough. In this range, the estimated uncertainty is larger than the true uncertainty since the diffusion is constrained by the boundaries of the habitat.

The above results show that behavior of a population in shifting habitats and the interpretation of phylogeographic inferences can depend strongly on the rate of environmental change and how it compares to the dispersal rate. If the population can not follow shifting habitat, the situation is similar to the static case often leads to

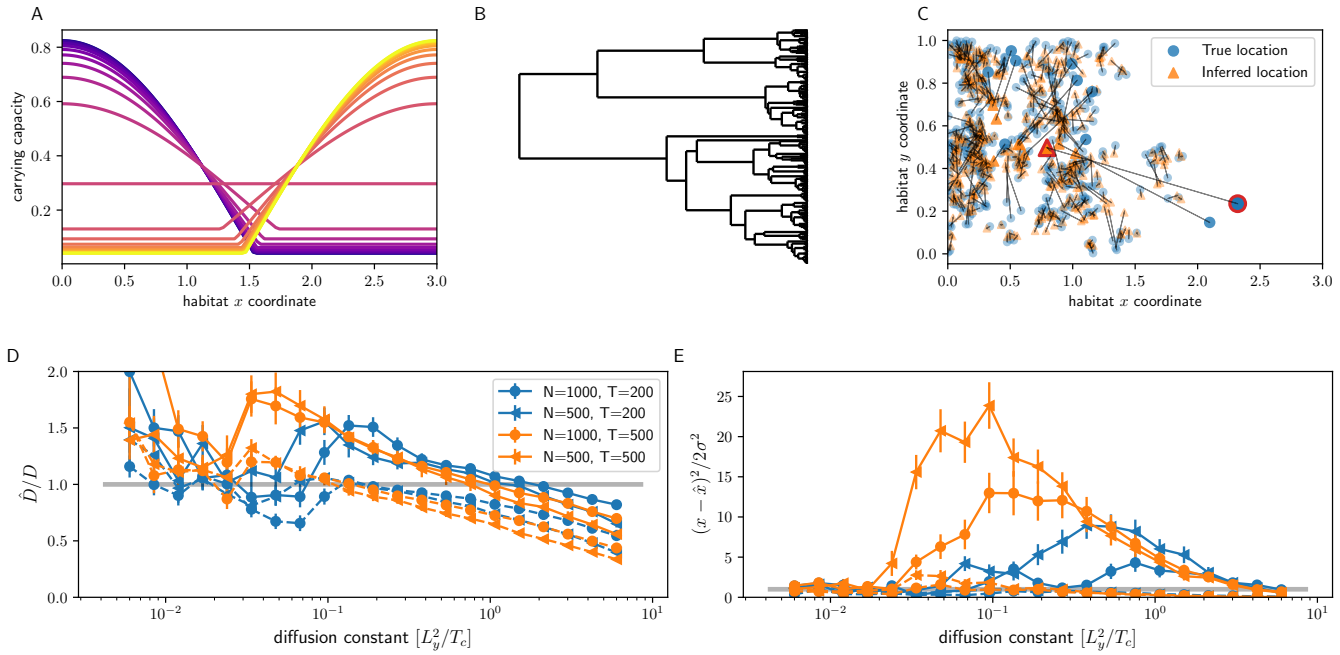

FIG. 4 **Phylogeographic estimates in rapidly changing environments. Panel A** shows the time course of the carrying capacity through half a period. The carrying capacity first peaks on the left, becomes flat, and then peaks on the right before shifting back in the second half of the period. The habitat size is $L_x = 3$, $L_y = 1$. **Panel B** shows a tree from a simulation under these conditions with $T = 500$, $D = 0.0008$ and $N = 500$, while **panel C** on the right shows the inferred (triangles, orange) and true positions (circles, blue) of internal nodes, highlighting the root with large markers and a red border. The population is sampled at a time when the carrying capacity is flat after having moved the left, that is the population is about to spread to the right. **Panels D** show estimates of diffusion coefficients in directions x (solid lines) and y (dashed lines) relative to their true values as a function of the true value of D. **Panel E** shows the average z_x^2 and z_y^2 for the oldest 20% of internal nodes. Inference of the x-coordinate of ancestral nodes is drastically overconfident for some values of D.

fragmented populations. Once the population starts following, inferred ancestral locations of deep nodes become unreliable, while at high dispersal rates expected uncertainties are larger than the habitat size and there is little information on deep ancestral locations.

To study the interplay of moving habitats and phylogeographic inference more systematically, I will now consider a situation where the population is constrained to a stripe with a Gaussian profile that moves at a constant velocity v along the x-axis (Fig. 5). The dispersal rates necessary to keep up with the shifting carrying capacity can be estimated from the Fisher-KPP equation (Eq. 3) above. This equation admits solutions with a traveling front, i.e. the population "invades" the empty territory with velocity $v_f = 2\sqrt{D\alpha}$ (Fisher, 1937; Hallatschek and Nelson, 2010; Kolmogorov *et al.*, 1937). Note that the model itself does not have an explicit parameter that has dimensions of a velocity. The speed at which the population invades space is given as compound of diffusion and accelerated growth in regions of low density and only grows like the square root of diffusion constant.

Fig 5B&C show simulation results for this traveling habitat. In Panel B, estimates of D_x using Eq. 2 are shown as a function of the ratio of the FKPP velocity v_f to the external velocity v. As expected, estimates of $\hat{D}_x$ are often dramatically too high (by up to factors of 1000). They peak right around when the external velocity matches the FKPP velocity v_f, corresponding to the situation where the population dynamics is maximally affected by the moving carrying capacity. If the external velocity was lower, undirected dispersal contributes more and more to the displacement of lineages, while the population would be unable to follow the moving carrying capacity if the external velocity was higher.

I showed above that estimates of v using Eq. 2 are ill-defined in models with diffusive dispersal, but it is worth exploring how they behave in cases when one can define a velocity within the model. Fig. 5C shows the estimator v_w in x direction relative to velocity of the habitat. This ratio is monotonically increasing with the diffusion coefficient and shows a range where it is close to 1, i.e. where the estimated velocity of lineages agrees with the external velocity. This range coincides with the location of the peak of D_w.

These behaviors are explained comparing the speed v_f at which the population can invade new territory with the external speed v, the ratio of which is used as the x-axis in Fig. 5 B&C. For $v_f \ll v$, the population cannot keep up with the moving carrying capacity and is saved from extinction only because selection is soft. In

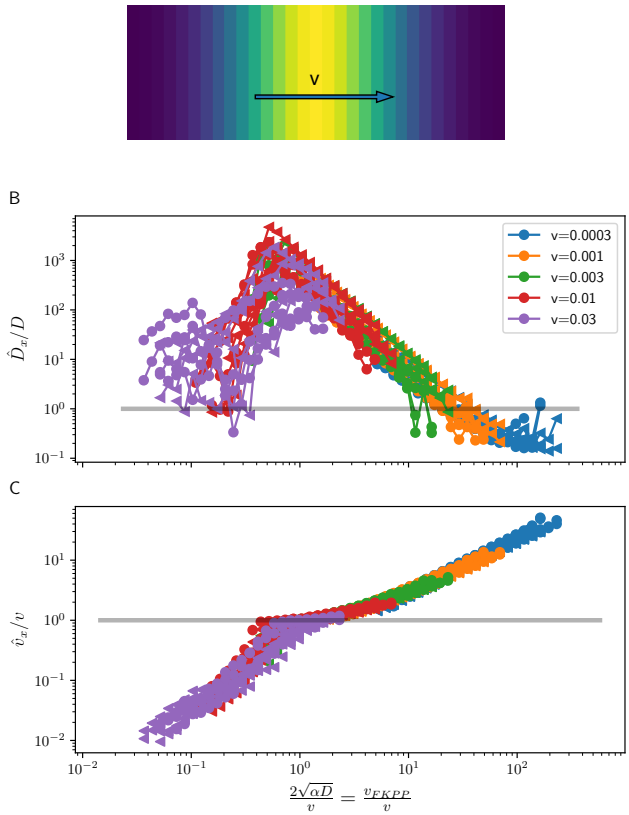

FIG. 5 **Phylogeographic inference in a moving habitat. Panel A** shows an illustration of the simulated set-up: a carrying capacity with a Gaussian profile with $\sigma = 0.5$ that moves along the x-axis with velocity v on a patch with $L_x = 3$, $L_y = 1$, and periodic boundary conditions. **Panels B&C** show estimates of diffusion coefficients and velocities, relative to their true values, as a function of the ratio of the FKPP velocity to the external velocity v.

this regime, the estimated velocity of lineages is below v and both the estimated v and D increase with v_f. Once $v \sim v_f$, the estimated lineage velocity coincides with v_f, while the estimated D peaks at values orders of magnitude above the true value. Once v_f increases beyond the narrow range of $v \approx v_f$, the velocity estimate v_w increases further. In this latter regime, diffusive motions dominates over directed motion and the estimates of v suffer from the problems of sample size dependence and lack interpretation as shown in Fig. 2.

Discussion

Phylogeography aims to reconstruct locations and migrations of ancestors from the spatial distribution of sampled individuals and their genome sequences. The latter allow the reconstruction of the phylogenetic tree which encodes how closely different individuals are related to each other. Together with the spatial location of the samples – the leaves of the tree – one can reconstruct the likely locations of ancestors and estimate parameters of dispersal. Popular tools for such inference, however, make strong assumptions about the dispersal process, including assuming that dispersal of individuals is diffusive (i.e. characterized by small displacements in random directions) and that the replication process is independent of spatial location. Here, I have explored the robustness of such inference to violations of the latter assumption while keeping microscopic dispersal strictly diffusive, that is without any long-range migration.

The common practice to estimate dispersal "velocities" using displacements of lineages along the tree (Dellicour *et al.*, 2021) is problematic. Firstly, the numerical value of such estimates depends strongly on the sample size and tree ensemble. Secondly, the underlying model is diffusive and has no parameters that have dimensions of a velocity and even if the quantity could be estimated robustly, there is no interpretation of the quantity within the model. In cases where a population invades a novel habitat, e.g. the invasion of North America by the West-Nile virus (Pybus *et al.*, 2012), the moving front of the population defines a bona fide velocity. However, as discussed above, this velocity is determined by the interplay of dispersal and population growth at the front. Estimators of a "lineage velocity" as defined in Eq. 2 only agree with invasion speed in a narrow, a priori unknown, parameter regime. The speed of invasion into a new habitat is better determined by modeling the position of the front (Pybus *et al.*, 2012), rather than by tracking individual lineages along the tree.

A more complex set of questions concerns how coupling between growth rate of the population and spatial density affects phylogeographic inferences and how robust these inferences are to shifting habitats. As has long been known, ignoring the coupling between population growth and spatial location can lead to unrealistic population densities (Felsenstein, 1975) and that density regulation is necessary. However, once density regulation is in place, the population dynamics depends strongly on whether dispersal is fast enough to explore the entire habitat during the coalescence time T_c of the population in absence of spatial structure. If $D \ll L^2/T_c$, the population fragments into subpopulations and the time to the most recent common ancestor increases dramatically compared to a neutral panmictic population. The resulting strong coupling between spatial location and coalescence process violates typical tree priors used in such inference. In the other case of $D \gg L^2/T_c$, dispersal is rapid enough that uncertainties of ancestral locations of deep nodes are large compared to the size of the habitat and such inferences thus uninformative. At the same time, this will lead to under-estimation of dispersal parameters since ancestral locations are constrained by the boundaries of the habitat. I did not consider additional effects that emerge from static carrying capacities that vary strongly in space, which cab lead to unexpected effective lineage dynamics corresponding to sources where

population densities are high and sinks where they are low (Wilkins and Wakeley, 2002).

Interpretation of phylogeographic inference becomes even more challenging when habitats shift in time. Depending on parameters, dispersal rates D can be overestimated or underestimated, sometimes by large factors, and estimates of ancestral locations can be overconfident or too uncertain. Simulations reveal that this behavior depends on the relative magnitude of the rate at which the habitat shifts and the FKPP velocity which depends on the diffusion constant D and the growth rate in empty regions α. The parameter range where both ancestral locations and dispersal parameters are inferred correctly is narrow and a priori unknown.

Phylogeographic methods are often studied using the spread of West Nile virus in North America and Rabies virus. West-Nile virus was first detected on the East Coast of the US in 1999 and reached the West Coast five years later in 2004 (Pybus *et al.*, 2012), translating to a wave front velocity of about 1000 km/year. Pybus *et al.* (2012) estimated a diffusion coefficient between 200 and 5000 km^2/day, where the latter refers to branches with particularly high diffusion constant. These extremely high estimates of diffusive dispersal are implausible. Alternatively, one can interpret the spread from east to west as a range expansion. To this end, we need an estimate of the growth rate α of the population at low density, which we can roughly obtain from the fact that the virus is seasonal: over the course of a few months, its prevalence increases by orders of magnitude, which typically requires a doubling time of about a week or a growth rate of $\alpha = 0.1$/day. Using $v_f = 2\sqrt{D\alpha}$ and equating it to 1000 km/year ($\approx$ 3km/day), we obtain estimates of D of $D \approx \frac{9\mathrm{km}^2}{4\alpha\mathrm{day}^2} \approx 22\mathrm{km}^2/\mathrm{day}$ – around 10 to 200 times smaller than estimates from phylogeography. This lower estimate of D implies a daily exploration radius of about 5 km. This picture is also consistent with the apparent "slowing down" of lineages after the initial expansion across North America (Dellicour *et al.*, 2020): Once the habitat was fully explored, directed range expansion with v_f ceases. Furthermore, rare long range dispersal could have strongly influenced the spread of the virus in ways not captured by Brownian or relaxed random walks (Hallatschek and Fisher, 2014).

Diffusion constants of rabies virus are typically estimated to be between 500 and 1500 km^2/year (Dellicour *et al.*, 2017). The most recent common ancestor of various populations is between 30 and 100 years in the past. This would translate into rather limited spread of 150 to 500km of individual clades from their common ancestor, suggesting that these populations are highly structured (isolated by distance) and that their long range dispersal is dominated by rare introductions of the virus into new populations (Brunker *et al.*, 2015).

The results presented here and the discussion of the examples above suggested that phylogeographic methods should be used and interpreted with caution. In many cases the habitat of a population has undergone dramatic and recurrent changes since the $MRCA$ and such changes will affect inferences in ways that are not captured by the models. In addition, uneven sampling, or complete lack of samples from some regions, can undermine phylogeographic estimates (Kalkauskas *et al.*, 2021; Layan *et al.*, 2023). The ability to infer ancestral locations is further limited by long-range dispersal (Hallatschek and Fisher, 2014). To capture such deviations from simple diffusive motions, many inference models assume a "relaxed random walk", where diffusion constant can vary across the tree according to a broad prior distribution (Dellicour *et al.*, 2021). However, while such models will often generate a better fit, they don't capture complex interactions between spatial location and population dynamics.

Why does phylogeography then often generate sensible results? On short time scales, a diffusive model will mostly reduce to a parsimonious reconstruction of ancestral locations which is robust to details of the model. Sampling at multiple time points further constrains ancestral locations sufficiently strongly that there is limited uncertainty, at least within the period for which samples are available. But inferences for deeper nodes in the phylogeny before the earliest samples can be problematic as the chance that the environment has changed and shifted, or that the ancestral locations are in poorly sampled regions, increases. And while phylogeographic inferences will still often yield plausible results by virtue of their tendency to produce parsimonious solutions that aggregate available spatial information, they can also be highly unreliable and misleading. In particular confidence statements should be treated with extreme caution, because the tails of the inferred distributions are determined by properties of the model and the underlying assumptions typically do not reflect the wide spectrum of possible environmental histories. In absence of a good understanding of the dispersal properties and how population dynamics varied over time and space, simple models that allow transparent interpretation of how results are informed by features of the data are preferable to complex and computationally demanding models that nevertheless lack relevant history.

Materials and methods

All simulation were performed in Python and the associated code is available in the github repository github.com/neherlab/phylogeography.

Acknowledgements

I am grateful to Louis de Plessis, Oskar Hallatscheck, Emma Hodcroft, Simon Dellicour, and Philippe Lemey for stimulating discussions and useful feedback on earlier versions of this manuscript.

References

Bolker, B., and S. W. Pacala, 1997, Theoretical Population Biology **52**(3), 179, ISSN 0040-5809, URL https://www.sciencedirect.com/science/article/pii/S0040580997913319.

Brunker, K., D. A. Marston, D. L. Horton, S. Cleaveland, A. R. Fooks, R. Kazwala, C. Ngeleja, T. Lembo, M. Sambo, Z. J. Mtema, L. Sikana, G. Wilkie, *et al.*, 2015, Virus Evolution **1**(1), vev011, ISSN 2057-1577.

Dellicour, S., M. S. Gill, N. R. Faria, A. Rambaut, O. G. Pybus, M. A. Suchard, and P. Lemey, 2021, Molecular Biology and Evolution **38**(8), 3486, ISSN 1537-1719, URL https://doi.org/10.1093/molbev/msab031.

Dellicour, S., S. Lequime, B. Vrancken, M. S. Gill, P. Bastide, K. Gangavarapu, N. L. Matteson, Y. Tan, L. du Plessis, A. A. Fisher, M. I. Nelson, M. Gilbert, *et al.*, 2020, Nature Communications **11**(1), 5620, ISSN 2041-1723, publisher: Nature Publishing Group, URL https://www.nature.com/articles/s41467-020-19122-z.

Dellicour, S., R. Rose, N. R. Faria, L. F. P. Vieira, H. Bourhy, M. Gilbert, P. Lemey, and O. G. Pybus, 2017, Molecular Biology and Evolution **34**(10), 2563, ISSN 0737-4038, URL https://doi.org/10.1093/molbev/msx176.

Etheridge, A. M., 2004, The Annals of Applied Probability **14**(1), 188, ISSN 1050-5164, publisher: Institute of Mathematical Statistics, URL https://www.jstor.org/stable/4140494.

Felsenstein, J., 1975, The American Naturalist **109**(967), 359, ISSN 0003-0147, publisher: The University of Chicago Press, URL https://www.journals.uchicago.edu/doi/10.1086/283003.

Fisher, R. A., 1937, Annals of Eugenics **7**(4), 355, ISSN 2050-1439, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1937.tb02153.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1937.tb02153.x.

Hallatschek, O., and D. S. Fisher, 2014, Proceedings of the National Academy of Sciences **111**(46), E4911, publisher: Proceedings of the National Academy of Sciences, URL https://www.pnas.org/doi/abs/10.1073/pnas.1404663111.

Hallatschek, O., and D. R. Nelson, 2010, Evolution **64**(1), 193, ISSN 0014-3820, URL https://doi.org/10.1111/j.1558-5646.2009.00809.x.

Kalkauskas, A., U. Perron, Y. Sun, N. Goldman, G. Baele, S. Guindon, and N. D. Maio, 2021, PLOS Computational Biology **17**(1), e1008561, ISSN 1553-7358, publisher: Public Library of Science, URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008561.

Kingman, J. F. C., 1982, Stochastic Processes and their Applications **13**(3), 235, ISSN 0304-4149, URL https://www.sciencedirect.com/science/article/pii/0304414982900114.

Kolmogorov, A., I. Petrovskii, and N. Piscunov, 1937, Byul. Moskovskogo Gos. Univ. **1**(6), 1, URL http://books.google.com/books?id=ikN59GkYJKIC&lpg=PP1&dq=A.N.%20Kolmogorov%3A%20Selected%20Works&client=firefox-a&pg=PA242#v=onepage&q=&f=false.

Layan, M., N. F. Müller, S. Dellicour, N. De Maio, H. Bourhy, S. Cauchemez, and G. Baele, 2023, Virus Evolution **9**(1), vead010, ISSN 2057-1577, URL https://doi.org/10.1093/ve/vead010.

Lemey, P., A. Rambaut, A. J. Drummond, and M. A. Suchard, 2009, PLOS Computational Biology **5**(9), e1000520, ISSN 1553-7358, publisher: Public Library of Science, URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000520.

Lemey, P., A. Rambaut, J. J. Welch, and M. A. Suchard, 2010, Molecular Biology and Evolution **27**(8), 1877, ISSN 0737-4038, URL https://doi.org/10.1093/molbev/msq067.

Pybus, O. G., M. A. Suchard, P. Lemey, F. J. Bernardin, A. Rambaut, F. W. Crawford, R. R. Gray, N. Arinaminpathy, S. L. Stramer, M. P. Busch, and E. L. Delwart, 2012, Proceedings of the National Academy of Sciences **109**(37), 15066, publisher: Proceedings of the National Academy of Sciences, URL https://www.pnas.org/doi/abs/10.1073/pnas.1206598109.

Trovão, N. S., M. A. Suchard, G. Baele, M. Gilbert, and P. Lemey, 2015, Molecular Biology and Evolution **32**(12), 3264, ISSN 0737-4038, URL https://doi.org/10.1093/molbev/msv185.

Vaughan, T. G., D. Kühnert, A. Popinga, D. Welch, and A. J. Drummond, 2014, Bioinformatics **30**(16), 2272, ISSN 1367-4803, URL https://doi.org/10.1093/bioinformatics/btu201.

Wilkins, J. F., and J. Wakeley, 2002, Genetics **161**(2), 873, ISSN 0016-6731.

Yule, G. U., 1925, Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character **213**(402-410), 21, publisher: Royal Society, URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1925.0002.