Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies

Matthew M Osmond[1] and Graham Coop[2]

[1]Department of Ecology & Evolutionary Biology, University of Toronto

[2]Department of Evolution & Ecology and Center for Population Biology, University of California - Davis

Abstract

Spatial patterns in genetic diversity are shaped by individuals dispersing from their parents and larger-scale population movements. It has long been appreciated that these patterns of movement shape the underlying genealogies along the genome leading to geographic patterns of isolation by distance in contemporary population genetic data. However, extracting the enormous amount of information contained in genealogies along recombining sequences has, up till recently, not been computational feasible. Here we capitalize on important recent advances in gene-genealogy reconstruction and develop methods to use thousands of trees to estimate time-varying per-generation dispersal rates and to locate the genetic ancestors of a sample back through time. We take a likelihood approach in continuous space using a simple approximate model (branching Brownian motion) as our prior distribution of spatial genealogies. After testing our method with simulations we apply it to the 1001 Genomes dataset of over one thousand *Arabidopsis thaliana* genomes sampled across a wide geographic extent. We detect a very high dispersal rate in the recent past, especially longitudinally, and use inferred ancestor locations to visualize many examples of recent long-distance dispersal and admixture. We also use inferred ancestor locations to identify the origin and ancestry of the North American expansion and to depict alternative

geographic ancestries stemming from multiple glacial refugia. Our method highlights the huge amount of information about past dispersal events and population movements contained in genome-wide genealogies.

Introduction

Patterns of genetic diversity are shaped by the movements of individuals, as individuals move their alleles around the landscape as they disperse. Patterns of individual movement reflect individual-level dispersal; children move away from their parents village and dandelion seeds blow in the wind. These patterns also reflect large-scale movements of populations. For example, in the past decade we have learnt about the large-scale movement of different peoples across the world from ancient DNA (Slatkin and Racimo, 2016; Reich, 2018). Such large scale movements of individuals also occur in other species during biological invasions or with the retreat and expansion of populations in and out of glacial refugia, tracking the waxing and waning of the ice ages (Hewitt, 2000).

An individual's set of genealogical ancestors expands rapidly geographically back through time in sexually reproducing organisms (Kelleher et al., 2016a; Coop, 2017). Due to limited recombination each generation, more than a few tens of generations back an individual's genetic ancestors represent only a tiny sample of their vast number of genealogical ancestors (Donnelly, 1983; Coop, 2013). Yet the geographic locations of genetic ancestors still represent an incredibly rich source of information on population history (Bradburd and Ralph, 2019). We can hope to learn about the geography of genetic ancestors because individuals who are geographically close are often genetically more similar across their genomes; their ancestral lineages have only dispersed for a relatively short time and distance since they last shared a geographically-close common ancestor. This pattern is termed isolation by distance. These ideas about the effects of geography and genealogy have underlain our understanding of spatial population genetics since its inception (Wright, 1943; Malécot, 1948). Under coalescent models, lineages move spatially, as a Brownian motion if dispersal is random and local, spitting to give rise to descendent lineages till we reach the present day. Such models underlie inferences based on increasing allele frequency differentiation (such as F_{ST}) with geographic distance (Rousset, 1997) and the drop-off in the sharing of long blocks of genome shared identical by descent among pairs of

individuals (Ralph and Coop, 2013; Ringbauer et al., 2017). These models also are the basis of methods that seek violations of isolation-by-distance (Wang and Bradburd, 2014).

While spatial genealogies have proven incredibly useful for theoretical tools and intuition, with few exceptions they have not proven useful for inferences because we have not been able to construct these genealogies along recombining sequences. In non-recombining chromosomes (e.g., mtDNA and Y), constructed genealogies have successfully been used to understand patterns of dispersal and spatial spread (Avise, 2009). However, these spatial genealogy inferences are necessarily limited as a single genealogy holds only limited information about the history of populations in a recombining species (Barton and Wilson, 1995). Phylogenetic approaches to geography ('phylogeography'; Knowles, 2009) have been more widely and successfully applied to pathogens to track the spatial spread of epidemics, such as SARS-Cov-2 (Martin et al., 2021), but such approaches have yet to be applied to the thousands of genealogical trees that exist in sexual populations.

Here we capitalize on the recent ability to infer a sequence of genealogies, with branch lengths, across recombining genomes (Rasmussen et al., 2014; Speidel et al., 2019; Wohns et al., 2021). Equipped with this information, we develop a method that uses a sequence of trees to estimate dispersal rates and locate genetic ancestors in continuous, 2-dimensional space, under the assumption of Brownian motion. Using thousands of approximately unlinked trees, we multiply likelihoods of dispersal rates across trees to get genome-wide estimates and use the sequence of trees to predict a cloud of ancestral locations as a way to visualize geographic ancestries. We first test our approach with simulations and then apply it to *Arabidopsis thaliana*, using over 1000 genomes with a wide geographic distribution (Alonso-Blanco et al., 2016) and a complex history (Fulgione and Hancock, 2018; Hsu et al., 2019).

Results

Overview of approach

We first give an overview of our approach, the major components of which are illustrated in Figure 1. See Materials and Methods for more details.

The rate of dispersal, which determines the average distance between par-

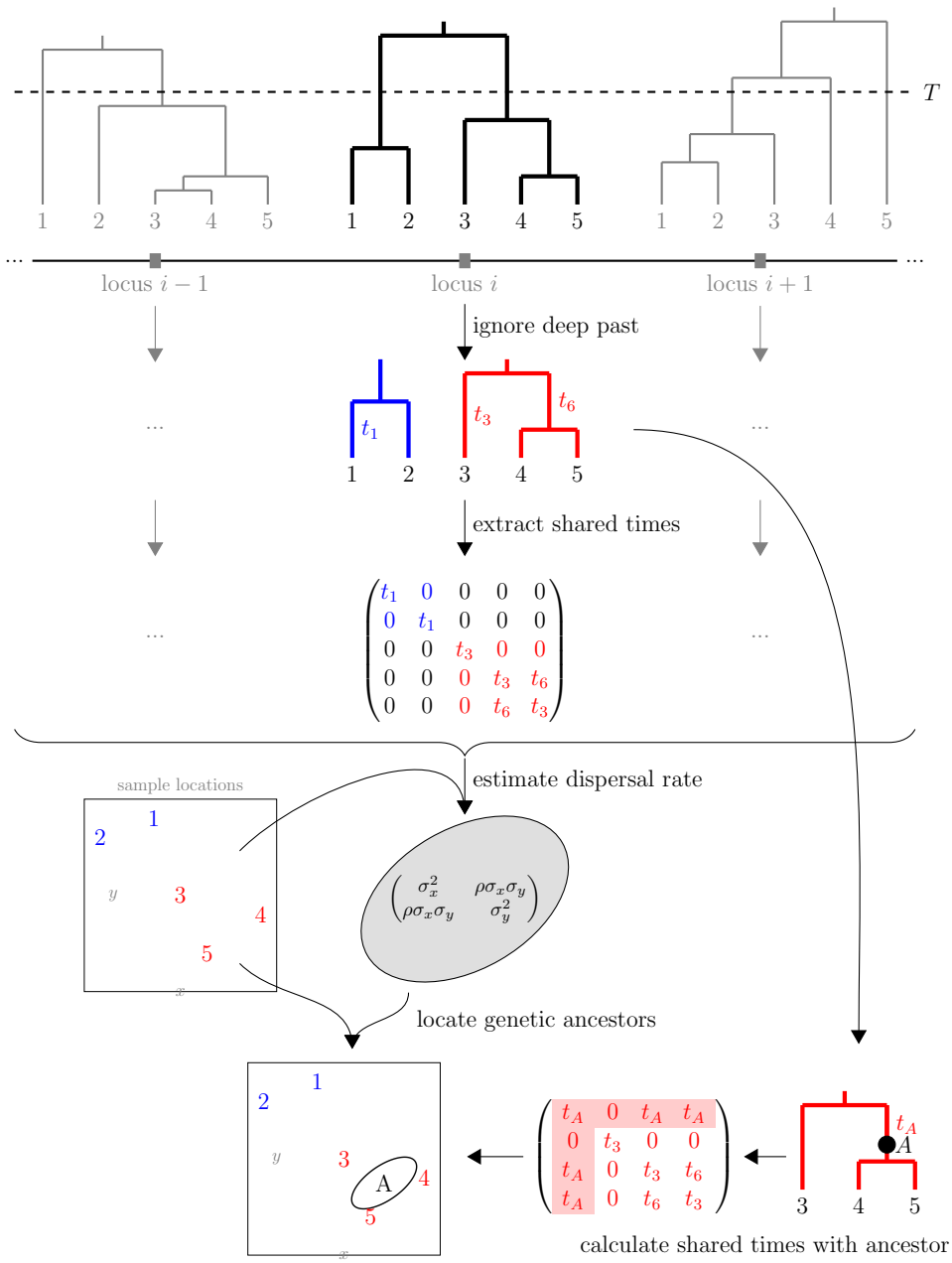


Figure 1: (Continued on the following page.)

Figure 1: **Conceptual overview of the approach.** From a sequence of trees covering the full genome, we downsample to trees at approximately unlinked loci. To avoid the influence of strongly non-Brownian dynamics at deeper times (e.g., glacial refugia, boundaries), we ignore times deeper than T, which divides each tree into multiple subtrees (here, blue and red subtrees at locus i). From these subtrees we extract the shared evolutionary times of each lineage with all others. In practice (but not shown here), we use multiple samples of the tree at a given locus, for importance sampling, and also extract the coalescence times for importance sample weights. Under Brownian motion, the shared times describe the covariance we expect to see in the locations of our samples, and so using the times and locations we can find maximum likelihood dispersal rate (a 2x2 covariance matrix). While we can estimate a dispersal rate at each locus, a strength of our approach is that we combine information across many loci, by multiplying likelihoods, to estimate genome-wide (or per-chromosome) dispersal rates. In practice (but not shown here), we estimate dispersal in multiple epochs, which allows for variation in dispersal rate over time and helps absorb non-Brownian movements further in the past. Finally, we locate a genetic ancestor at a particular locus (a point on a tree, here A) by first calculating the time this ancestor shares with each of the samples in its subtree, and then using this matrix and the dispersal rate to calculate the probability distribution of the ancestor's location conditioned on the sample locations. In practice (but not shown here), we calculate the locations of ancestors of a given sample at a given time at many loci, combining information across loci into a distribution of genome-wide ancestry across space.

ents and offspring, is a key parameter in ecology and evolution. To estimate this parameter we assume that in each generation the displacement of an offspring from its mother is normally distributed with a mean of 0 and covariance matrix $\mathbf{\Sigma}$. The covariance matrix is determined by the standard deviations σ_x and σ_y along the x (longitudinal) and y (latitudinal) axes, respectively, as well as the correlation, ρ, in displacements between these two axes. The average distance between mothers and offspring is then $\sqrt{2/\pi}\sigma_i$ in each dimension. We refer to the covariance matrix $\mathbf{\Sigma}$ as the (per-generation) dispersal rate.

Given this model, the path of a lineage from its ancestral location to the present day location is described by a Brownian motion. Lineages covary

in their locations because of shared evolutionary histories – lineages with a more recent common ancestor covary more. Given a tree at a locus we can calculate the covariance matrix of shared evolutionary times and compute the likelihood of the dispersal rate, $\mathbf{\Sigma}$, which is normally distributed given this covariance matrix (Equation (1)). At each locus we can estimate the likelihood of the dispersal rate given the tree at that locus and, given we sample loci far enough apart that they are essentially independent, multiply likelihoods across loci to derive a genome-wide likelihood, and thus a genome-wide maximum likelihood dispersal rate.

Under this same model we can also estimate the locations of genetic ancestors at a locus. Any point along a tree at any locus is a genetic ancestor of one or more current day samples. This ancestor's lineage has dispersed away from the location of the most recent common ancestor of the sample, and covaries with current day samples in their geographic location to the extent that it shares times in the tree with them. Under this model, the location of an ancestor is influenced by the locations of all samples in the same tree, including those that are not direct descendants (cf. Wohns et al., 2021). For example, in Figure 1 the ancestor's location is not be the midpoint of its two descendants (samples 4 and 5); the ancestor's location is also pulled towards sample 3 since the ancestor and sample 3 both arose from a common ancestor. Conditioning on the sample locations, and given the shared times and previously inferred dispersal rate, we can compute the probability the ancestor was at any location, which again is a normal distribution (Equation (23)). In contrast to dispersal, for ancestral locations we do not want to multiply likelihoods across loci since the ancestors at distant loci are likely distinct. Instead we calculate the maximum likelihood location of genetic ancestors at each locus to get a cloud of likely ancestral locations, and use these clouds to visualize the spatial spread of ancestry backwards through time.

We estimate marginal trees along the genome using `Relate` (Speidel et al., 2019). Relate infers a sequence of tree topologies and associated branch lengths, and can return a set of posterior draws of the branch lengths on a given tree. This posterior distribution of branch lengths is useful to us as the shared times in the tree are key to the amount of time that individual lineages have had to disperse away from one another and we wish include uncertainty in the times into our method. Relate gives us a posterior distribution of branch lengths that is estimated using a coalescent prior, which assumes a panmictic population of varying population size (the size changes are esti-

mated as part of the method), where any two lineages are equally likely to coalesce. This panmictic prior results in a bias in the coalescent times under a spatial model, where geographically proximate samples are more likely to coalesce. To correct for this bias we make use of importance sampling to weight the samples of branch lengths at each locus. We then calculate the weighted average likelihood over our draws of our sample of trees at a locus (or loci), so that it is as if they were drawn from a prior of branching Brownian motion (Meligkotsidou and Fearnhead, 2007). Branching Brownian motion, also known as the Brownian-Yule process, is a simple model of spatial genealogies in a continuous population; it does not describe the full complexities of spatial models such as the spatial coalescent, but it provides an analytically tractable model and a reasonable approximation over short-time scales (Edwards, 1970; Rannala and Yang, 1996; Meligkotsidou and Fearnhead, 2007; Novembre and Slatkin, 2009).

In practice we concentrate on the recent past history of our sample. For our estimates of dispersal rates in particular we do not want to assume that our model of Gaussian dispersal (and branching Brownian motion) holds deep into the past history of the sample. This is because the long-term movement of lineages is constrained by geographic barriers (e.g., oceans) and larger scale population movements may erase geographic signals over deep time scales. On theoretical level ignoring the deep past may also be justified because in a finite habitat the locations of coalescence events further back in time become independent of sampling locations as lineages have moved around sufficiently (Wilkins and Wakeley, 2002). Thus we only use this geographic model to some time point in the past (T), and at each locus we use the covariance of shared branch lengths based on the set of subtrees formed by cutting off the full tree T generations back.

To relax the assumption of a constant dispersal rate we extend our method to estimate dispersal rates in multiple epochs. Under Brownian motion this extension is fairly straight-forward as the covariance in sample locations is simply the sum, across epochs, of epoch-specific covariances (the Kronecker product of the shared times in a epoch and the dispersal rate in that epoch). An added benefit of estimating dispersal in epochs is that the estimates in more distant epochs can absorb some of the non-Brownian dynamics further back in time, increasing the accuracy of estimates in more recent epochs.

Simulations

We first wanted to test the performance of our method in a situation where the true answers were known. To do this we used a combination of spatially-explicit forward-time simulations (Haller and Messer, 2019), coalescent simulations (Kelleher et al., 2016b), and tree-sequence tools (Haller et al., 2019; Kelleher et al., 2019; Speidel et al., 2019) to compare our estimates of dispersal rates and ancestor locations with the truth (see Materials and Methods). This was also an opportunity to compare our estimates using the true trees vs. the Relate-inferred trees, to examine the influence of errors in tree inference.

Dispersal rates

Our method does a good job at capturing the simulated dispersal rate when using the true trees, especially at lower dispersal rates (Figure 2A). At larger dispersal rates the true trees tend to cause underestimates of the simulated dispersal rate, likely a consequence of the finite habitat we simulate (with reflecting boundaries). Here we use a time cutoff of only 100 generations in a 50x50 habitat, meaning that with any dispersal rate larger than $\sigma^2 \sim 50^2/100 = 5^2$ a lineage is reasonably likely to cross the entire habitat in that time. At higher dispersal rates or longer cutoff times (Figure S1), the simple Brownian motion model expects the samples to be more broadly distributed than the finite habitat allows, leading to larger underestimates. Regardless, we can interpret the dispersal rate inferred from the true trees as a true 'effective' dispersal rate, given the boundaries, local competition, biparental reproduction, etc. Encouragingly, the estimates from the inferred trees are highly correlated with these estimates, although with an upward bias. This upward bias is expected given the combination of isolation-by-distance and errors in inferring tree topologies, causing geographically distant samples to be mistakenly inferred to be close relatives. The bias (and variance across replicates) increases when we just use the panmictic coalescent prior from `Relate` (Figure S2), showing that our spatial prior implemented via importance sampling improves our inference.

In natural populations dispersal rates likely vary through time and so we also simulated a two-epoch scenario, where the dispersal rate switched 100 generations ago. Figure 2B,C shows the cases where the dispersal rate switched from $\sigma^2 = 0.25^2$ to $\sigma^2 = 0.5^2$, and vice-versa. As expected, our

estimates of the single dispersal rate fell in between the two different rates of dispersal (grey dots), averaging over the two epochs. We then applied our extension of the likelihood-based inference to allow for different dispersal rates in different epochs (with *a priori* switch times between epochs, see Materials and Methods) to these simulations. Our multi-epoch dispersal estimates using both the true and inferred trees captured the switch in dispersal rates (when supplied with the correct switch time), again with an upward bias when using the inferred trees. Ideally we would use likelihood ratio tests to test for significantly different dispersal rates between epochs, however, our simulations show that our dispersal estimates are not well calibrated under the null model of no change (Figure S3A,D, Supplementary Text). Instead

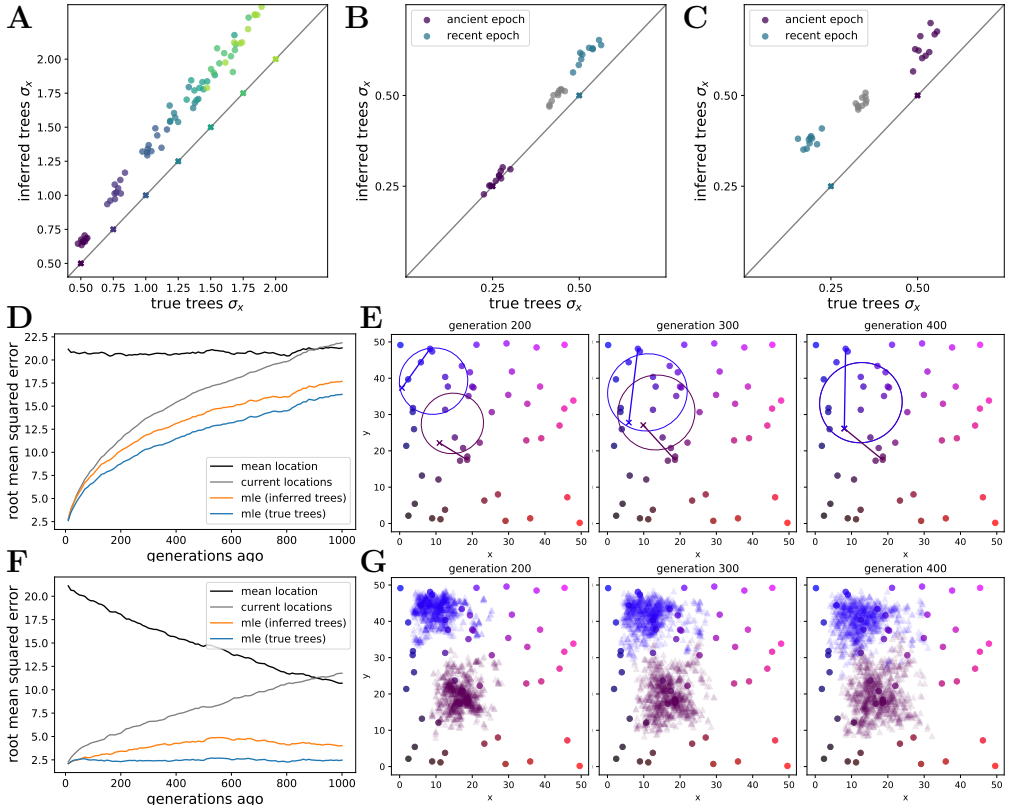


Figure 2: (Continued on the following page.)

Figure 2: **Simulations. (A) Accuracy of genome-wide dispersal rates.** Maximum composite likelihood estimates (over 10^2 evenly-spaced loci (trees)) of dispersal rate (only showing the standard deviation in the x dimension) using true trees vs. `Relate`-inferred trees (10 importance samples/locus, time cutoff of 10^2 generations). Each dot is a single simulation and the colours represent the simulated dispersal rates (value indicated by 'x' along diagonal; 10 replicates of each). **(B-C) Ability to detect time-varying dispersal rates.** Maximum composite likelihood estimates of dispersal rate using true trees vs. `Relate`-inferred trees in a two-epoch model (time cutoff of 10^3 generations). In B, the simulated dispersal rate switched from $\sigma^2 = 0.25^2$ in the deeper past to 0.5^2 in the most recent 10^2 generations. In C, the dispersal rates in the two epochs are flipped. Ten replicates are shown, with one dot of each colour for each epoch. The grey dots are the dispersal estimates under the assumption of a one-epoch model. **(D) Accuracy of locating genetic ancestors at individual loci.** Root mean squared errors between the true locations of ancestors (for 10^2 samples at 10^2 loci at 10^2 times) and the mean location of the samples, the current locations of the samples, and the maximum likelihood estimates from the inferred and true trees. **(E) Locating genetic ancestors at a particular locus.** 95% confidence ellipses for the locations of genetic ancestors for two samples at a single locus (using the true trees and the maximum likelihood dispersal rate). The 'x's are the true ancestral locations and the lines connect the true ancestral locations with the location of the contemporary descendant sample. **(F) Accuracy of mean genetic ancestor locations.** Root mean squared errors between the true mean (over 10^2 loci) location of ancestors (for 10^2 samples at 10^2 times) and the mean location of the samples, the current locations of the samples, and the mean (over 10^2 loci) maximum likelihood estimates from the inferred and true trees. **(G) Locating genetic ancestors at many loci.** Maximum likelihood estimates for the locations of genetic ancestors for the same two samples at 10^3 loci, here using true trees and the maximum likelihood dispersal rate. Panels D-G are simulated with $\sigma_x = \sigma_y = 0.5$ and $\rho = 0$.

we use the point estimates as a useful heuristic to detect broad patterns of dispersal rate change.

Locating ancestors

We next wanted to test our ability to locate the genetic ancestor of a sampled genome at a given locus and a given time. Our likelihood-based method gives both point estimates (maximum likelihood estimate, MLE) and 95% confidence ellipses (under the Brownian motion model), constructed based on the MLE of the genome-wide dispersal rate. We also have developed a best linear unbiased predictor (BLUP) of ancestral locations – importance sampling over analytically calculated MLE locations, rather than numerically finding the maximum of importance sampled likelihoods – that is faster to calculate, makes fewer assumptions, and is less reliant on the estimated dispersal rates (and completely independent of them when only one epoch), but this method does not give measures of uncertainty. Here we focus on the MLE-based method for estimating ancestral locations, but both methods are implemented in our software and both give essentially identical point estimates for ancestral locations in our simulations.

Figure 2D shows the error in the MLE ancestor locations, using the true or inferred trees, and compares this to sensible straw-man estimates (the current location of each sample and the mean location across samples). We see that the true trees give the best estimates and that the inferred trees do only slightly worse. That said, the mean squared error in our inferred location of the ancestor at a locus grows relatively rapidly back in time. To illustrate the cause of this increase in error, Figure 2D shows the 95% confidence ellipses for the locations of the ancestors of two samples at one particular locus at three different times in the past. In this example the lineages coalesce between generation 300 and 400, so the ellipses merge. While the ellipses do a good job of capturing true ancestral locations (the 'x's), the size of an ellipse at any one locus grows as we move back in time, meaning at deeper times (and higher dispersal rates) any one locus contains little information about an ancestor's location. Given the large uncertainty of an ancestor's location at any one locus, we combine information across loci and consider a cloud of MLE ancestor locations from loci across the genome for a particular sample at a particular time in the past (Figure 2G). The genome-wide mean of the MLE locations is able to predict the true mean location of genetic ancestors with much lower error (Figure 2F), even with the inferred trees, suggesting our method can successfully trace major geographic ancestries of a sample back into the past.

Empirical application: *Arabidopsis thaliana*

We next applied our method to 1135 *Arabidopsis thaliana* accessions from a wide geographic range (Alonso-Blanco et al., 2016). *A. thaliana* has a complex, and not yet fully resolved, population history (Fulgione and Hancock, 2018; Hsu et al., 2019). The samples in this 1001 Genomes dataset are thought to have descended from individuals from at least two glacial refugia, including one refuge in north Africa that contributed substantial ancestry to the 'Iberian Relict' samples (Alonso-Blanco et al., 2016; Durvasula et al., 2017; Fulgione et al., 2018) and one refuge near the Balkans that contributed substantial ancestry to the more weedy and cosmopolitan 'Non-Relict' samples (Lee et al., 2017). It is thought that populations from both refuges first expanded northwards, followed by a fast east and west expansion of the Balkan population across most of Eurasia (Alonso-Blanco et al., 2016; Lee et al., 2017; Fulgione and Hancock, 2018; Hsu et al., 2019). Extensive admixture between the expanding populations appears to have obscured the timing of the most recent east-west expansion, with some estimates before (Durvasula et al., 2017; Fulgione et al., 2018) and some after (Alonso-Blanco et al., 2016; Lee et al., 2017; Hsu et al., 2019) the last glaciation, which ended $\sim 11,000$ years ago. The 1001 Genomes dataset contains a relatively good spatial sampling of individuals with extensive Non-Relict ancestry while about 2% of the samples are considered Iberian Relicts, mostly in Spain but also two samples in Morocco (Alonso-Blanco et al., 2016). The dataset also contains one 'Relict' sample from each of Cabo Verde, the Canary Islands, Sicily, and Lebanon (Alonso-Blanco et al., 2016), all of which have been found to contain substantial Non-Relict ancestry (Alonso-Blanco et al., 2016; Lee et al., 2017; Zeng et al., 2017). The 1001 Genomes dataset does not include more recent samples from Madeira (Fulgione et al., 2018), Africa (Durvasula et al., 2017), and East Asia (Zeng et al., 2017; Zou et al., 2017), which likely contain ancestry from additional refugia (Fulgione and Hancock, 2018; Hsu et al., 2019). We did not attempt to include these more recent samples in this application because, while an important part of the puzzle, their spatial sampling is relatively sparse. Finally, the dataset includes 125 samples from across North America, a range expansion resulting of multiple recent human introductions (Exposito-Alonso et al., 2018; Shirsekar et al., 2021).

A. thaliana tree sequences

We used `Relate` (Speidel et al., 2019) to infer the tree sequence, estimate the panmictic effective population size through time, and resample branch lengths for importance sampling (see Materials and Methods). After dropping the 50% of trees with the fewest number of mutations, per chromosome, the tree sequence contains 213,481 trees across the 5 autosomes. The tree sequences (in both anc/mut and `tskit` formats) are publicly available at `https://doi.org/10.5281/zenodo.5099657`, which we hope will facilitate additional analyses (e.g., inferring selection; Stern et al., 2019, 2021).

A. thaliana is a selfer with a relatively low rate of outcrossing (Bomblies et al., 2010; Platt et al., 2010), thus it is worth taking a moment to consider the impact of selfing on our inferences. We chose the 1001 Genomes panel because of its large size and broad geographic sampling. Further, the availability of inbred accessions means that the samples have been well-studied and from our perspective remove the complications of obtaining phased haplotypes to run `Relate`. On the other hand, the high rate of selfing lowers the effective recombination rate and so is expected to increase the correlation in genealogies along the genome (Nordborg, 2000). However, in practice, linkage disequilibrium breaks down relatively rapidly in *A. thaliana*, on the scales of tens of kilobases (Kim et al., 2007), such that many trees along the genome should be relatively independent from each other. A related issue is that the individuals with recent inbreeding (selfing) in their family tree will have fewer genealogical and genetic ancestors than outbred individuals. Thus in any recent time-slice there are a reduced number of independent genetic ancestors of a individual from a selfing population, but even with relatively low rates of outbreeding the number of ancestors still grows rapidly (Lachance, 2009). Finally, while the effective recombination rate may vary through time along with rates of selfing, `Relate` uses a mutational clock to estimate branch lengths, and thus they should be well calibrated to a generational time scale.

Rapid recent east-west dispersal

We first used the tree sequence to estimate dispersal rates. In doing this we ignored the locations of 199 genetically near-identical samples ($< 10^3$ basepairs different, as in Alonso-Blanco et al., 2016), the 125 samples from North America, and the 2 samples from Japan (see Materials and Methods). The latter two groups are outliers for dispersal as they are thought to have

been carried large distances by humans in the recent past (Exposito-Alonso et al., 2018; Zou et al., 2017). The near-identical samples may also represent recent rare long-distance dispersal, but could alternatively be due to mis-assignments or mix-ups after collection (Alonso-Blanco et al., 2016), and so we remove them in case of the latter. Removing near-identical samples that are true long-distance migrants will cause us to underestimate dispersal rates. However, most near-identical samples are very near one another and are likely from the same inbred lineage (Alonso-Blanco et al., 2016); excluding these will have little effect on our inference. In estimating dispersal we use 10 importance samples of branch lengths per tree and a time cutoff of 10^4 generations (equivalently, years). Estimates with a cutoff of 10^3 were essentially identical, suggesting most of the information on dispersal comes from movements in the last 10^3 years, but having trees that go back 10^4 generations allows us to locate ancestors more reliably at deeper times.

We first estimate a single, constant dispersal rate, i.e., assuming one epoch. Figure 3A shows both the per-locus (dots) and the composite per-chromosome (horizontal lines, with gray denoting the standard error) estimates. We find that the per-generation rate of dispersal is ~ 10 times larger across longitude than across latitude (i.e., $\sigma^2_{\text{long}} \sim 10\sigma^2_{\text{lat}}$, with units km^2/generation). This high rate of longitudinal dispersal is consistent with the hypothesis of rapid expansion along the east-west axis of Eurasia from glacial refugia, facilitated by relatively weak environmental gradients and, potentially, human movements and disturbance (Alonso-Blanco et al., 2016; Lee et al., 2017). As discussed in Materials and Methods, our approach does not condition on the sample locations, which means that the distribution of sample locations may influence dispersal estimates. This is a particular worry here where we have samples from a wider range of longitudes than latitudes and we find a larger longitudinal dispersal rate. Fortunately, we can check our result by estimating dispersal separately in longitude and latitude (assuming no covariance), where conditioning on sample locations is much weaker (Meligkotsidou and Fearnhead, 2007). Doing this we find essentially identical dispersal rates as before (mean dispersal rates across chromosomes: $\sigma_{\text{long}} \approx 259$ km/$\sqrt{\text{gen}}$, $\sigma_{\text{lat}} \approx 81$ km/$\sqrt{\text{gen}}$), supporting our finding of faster dispersal across longitude ($\sigma^2_{\text{long}} \sim 10\sigma^2_{\text{lat}}$).

To explore the idea that environmental changes, such as human movements and disturbance, have affected the rate of spread of *A. thaliana*, we next estimated per-chromosome dispersal rates under a two-epoch model.

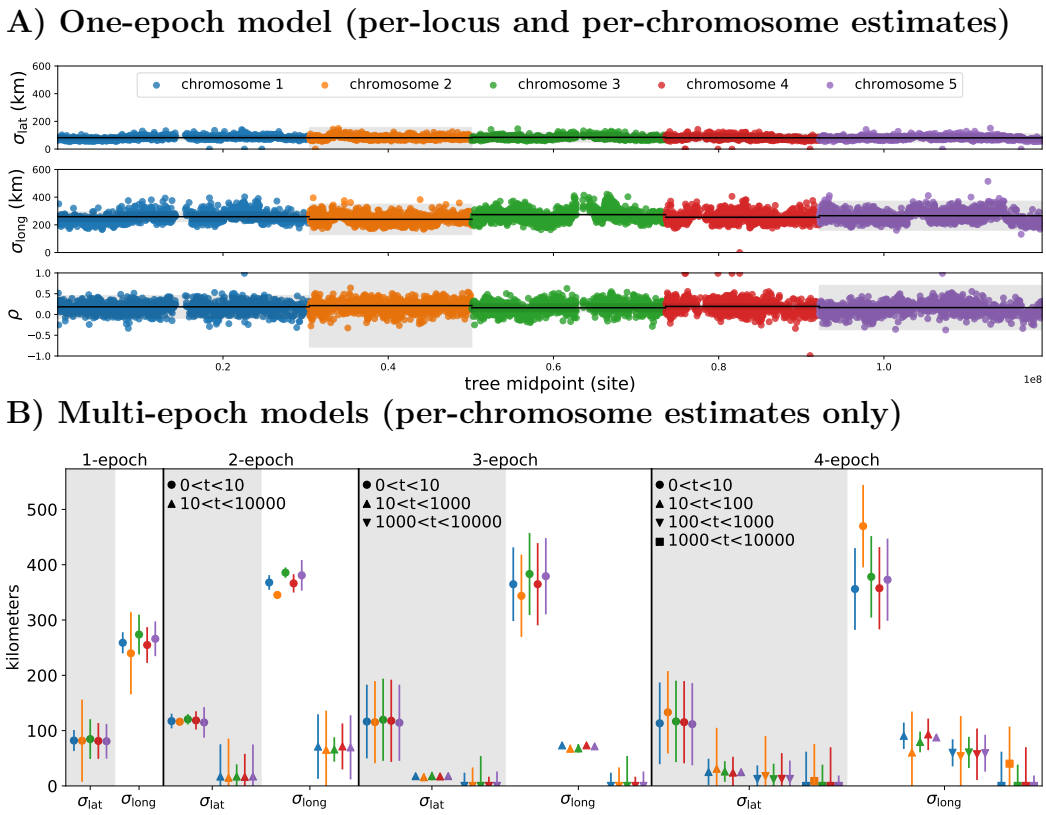


Figure 3: **Dispersal rate estimates in *Arabidopsis thaliana*. (A) One-epoch model.** The dots are maximum likelihood estimates of dispersal rate (σ, in units of km/$\sqrt{\text{generation}}$) at 10^3 evenly-spaced loci (trees) per chromosome. The black lines are maximum composite likelihood estimates for each chromosome (using 10^2 evenly-spaced loci) and the grey shading is the standard error (estimated from the Hessian at the maximum). **(B) Multi-epoch models.** The maximum composite likelihood dispersal rate estimates ($\pm$ standard error) across each chromosome (10^2 evenly-spaced loci) for models with multiple epochs (and the one-epoch model again for comparison). For two- and three-epoch models with alternative (and less likely) split times, see Figures S4 and S5.

Regardless of whether the more recent epoch ends 10, 10^2, or 10^3 generations ago, dispersal rates in the recent epoch are greater than those under the one-

epoch model (Figure S4). Figure 3B shows the dispersal estimates under the split time with the highest likelihood (10 generations ago), suggesting very rapid dispersal along the east-west axis in very recent times (on the order of decades). We also ran three-epoch models that allow dispersal to change at 10 and 10^2, 10 and 10^3, or 10^2 and 10^3 generations ago (Figure S5). The model with the highest likelihood (split times of 10 and 10^3 generations ago, Figure 3B) further supports the idea that the signal of very rapid dispersal comes from movements on the timescale of decades. Finally, we ran a four-epoch model with split times of 10, 10^2, and 10^3 generations ago (Figure 3B). This had the highest likelihood of any model we ran and reiterates the main conclusion – there appears to have been very rapid recent east-west dispersal. While it is tempting to compare dispersal rates across epochs, one caveat here is that finite habitat boundaries may disproportionally depress more ancient estimates; in essence, while we need a large recent dispersal rate to account for the large distances between relatively closely related samples, if such a dispersal rate continued far into the past the samples would be expected to cover a much wider geographic range than is possible given the constraints of habitat.

Identifying interesting dispersal outliers

We next used the tree sequence and dispersal estimates to locate genetic ancestors (see Materials and Methods). We can locate ancestors at every locus for every individual at any time, which represents an incredibly rich resource for understanding population movement. As a first step, we visualize the mean ancestral location, averaged over loci, for every individual to detect samples with unusual geographic ancestries. To do this we estimated the locations of recent ancestors of all samples at 100 evenly-spaced loci per chromosome and averaged over loci to give a mean ancestral 'displacement' (from the sample, backwards in time) for each sample.

Figure 4 shows these mean displacements as arrows, with colours emphasizing the length of the arrow, at 10 and 100 generations ago. As expected, most arrows point inwards, corresponding to the ancestors of the sample being geographically closer to one another than the samples are. There are a few exceptions however. For example, there is a sample in Romania (accession 9737) with a mean displacement far to the east. This outlier appears to coalesce at many loci ~ 100 generations ago with two samples in Russia near Kazakhstan's northeastern border (accessions 9627 and 9630), reflecting a

recent long-distance dispersal event. Taking a look at the coalescence times between accessions 9737 and 9627 along the tree sequence (Figure S6A) there are many large blocks of recent coalescence (< 100 generations), covering about 50% of the genome, while the remainder of the genome coalesces more deeply. This is consistent with the Romanian sample's previous placement in the 'Asia' admixture group (Alonso-Blanco et al., 2016). A second example is a sample from southern France (accession 9933) that quickly moves east to Romania/Ukraine. Taking a look at coalescence times between this sample and those further to the east, we find this sample coalesces < 100 generations ago with a sample from Afghanistan (accession 10015) over a few very large blocks of its genome, but is much further diverged elsewhere (Figure S6B). This is consistent with this French sample's previous 'Admixed' admixture assignment (including substantial ancestry from the 'Central Asia' group; Alonso-Blanco et al., 2016).

There are also some samples that are outliers in terms of distance traveled by their ancestors. For example, there are two samples, one from the UK (accession 7314) and the other from Belarus (accession 6981), that travel long distances towards one another in the past ~ 10 generations (see also Figure 3B in Alonso-Blanco et al., 2016). A look at the coalescence times between these two samples shows that these two individuals are extremely closely related along almost all of their genome except the first $\approx$ 700kb of chromosome 5 (Figure S6C). Other outliers include the sample from Cabo Verde (accession 6911), which is classified as a Relict but includes a large amount of European ancestry (Alonso-Blanco et al., 2016) and the most eastward Russian sample (accession 9622), which has substantial European ancestry (Alonso-Blanco et al., 2016). More generally, we also see greater rates of movement, on average, from samples from further east, consistent with the previous finding of reduced genetic distance between 'Asian' samples up to $\sim$250km apart (Alonso-Blanco et al., 2016), the signal of a rapid eastward expansion.

Detecting the existence and source of rare long-distant migrants and misplaced samples

We next looked in more detail at some of the largest outliers in recent ancestral movements. As mentioned above, the two accessions from Japan are thought to be recent long-distance migrants. When we include these samples' locations and locate their ancestors, these samples are the largest out-

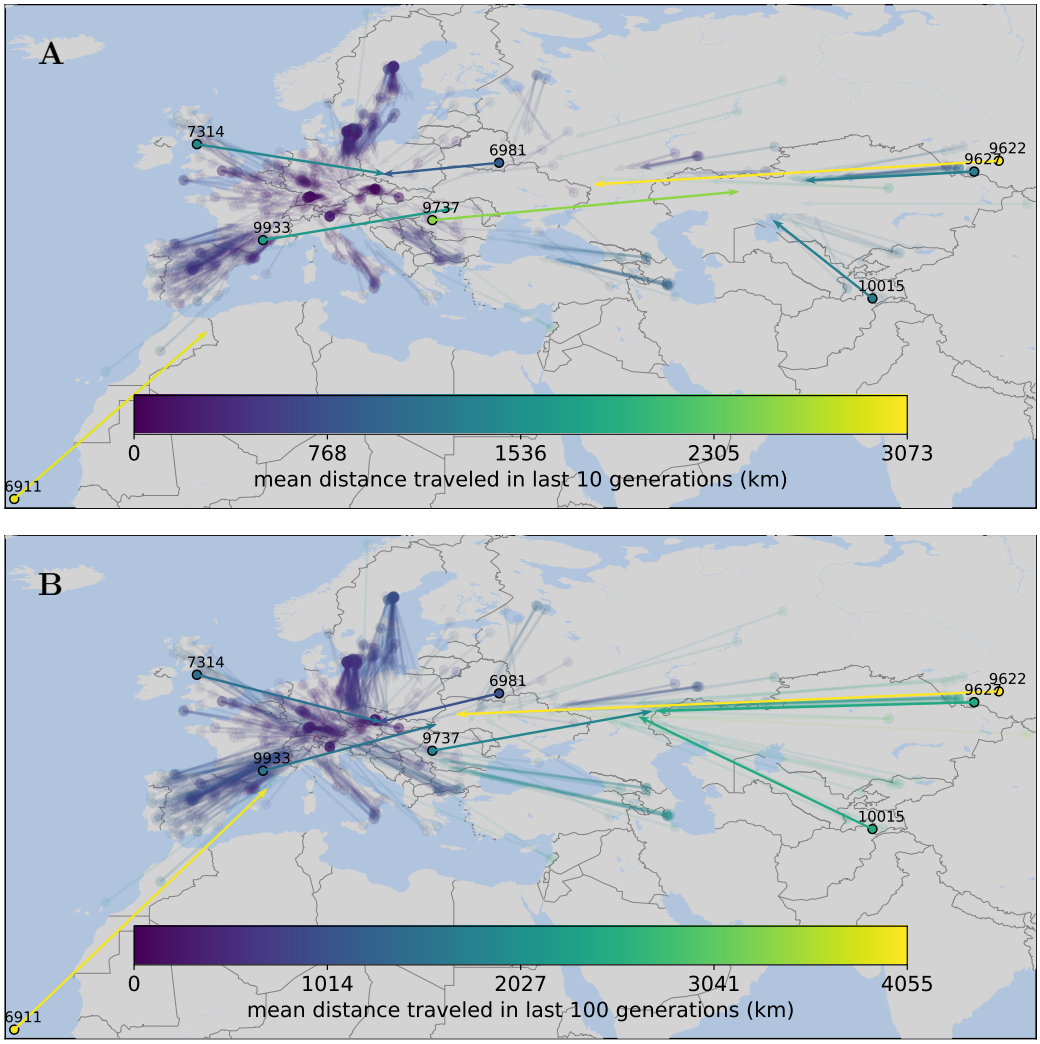


Figure 4: **Locating genetic ancestors to identify interesting outliers.** Mean maximum likelihood genetic ancestor locations (averaged over 10^2 evenly-spaced loci per chromosome) (**A**) 10 and (**B**) 100 generations ago. The samples discussed in the text are highlighted and labeled (accession numbers). Maximum likelihood locations calculated using per-chromosome four-epoch dispersal rates (Figure 3B).

liers in their inferred rate of recent movement (result not shown). Creating a smooth kernel from 10^3 inferred ancestral locations per chromosome, Figure 5A shows an extremely rapid transition between Japan and Europe in the last $\sim 10 - 100$ generations (only accession 7207 shown). This is consistent with the hypothesis that humans moved the ancestors of these samples from Europe to Japan in the last few hundred years (Zou et al., 2017).

The Japanese samples, while spatial outliers, are relatively closely related to the majority of the other samples (Alonso-Blanco et al., 2016). Thus, our method has the power to locate their ancestors and detect them as recent migrants. To see if the our method, and the data, have the power to detect a recent migrant from a less well represented ancestry, we deliberately misplaced the location of an Iberian Relict from Spain (accession 9533) to the mean location of the Eurasian samples after filtering (southern Czech Republic). Since this is an Iberian Relict, it is most closely related to the 21 other Iberian Relict samples (only $\approx 2\%$ of the samples), with much of its ancestry likely coming from a refuge in north Africa (Alonso-Blanco et al., 2016; Durvasula et al., 2017; Fulgione et al., 2018). Figure 5B shows that, despite the rarity of this ancestry in this dataset, our method correctly puts the misplaced Relict back into Spain at most loci in ~ 10 generations.

The examples above show that our method has the power to detect many recent long-distant migrants and misplaced samples, and identify their source. Note that, because we assume a model of continuous migration, the ancestors of recent migrants and misplaced samples have to migrate through intermediate locations to reach their likely source location. However, if, after inspection, an investigator suspects that a sample was likely misplaced, or dispersed a long distance suddenly, they could choose to ignore that sample's location and use the trees alone to estimate its source.

Inferring the origin and ancestry of a recent invasion

To illustrate how we can use our method to identify the source location of a subset of samples (e.g., recent migrants or misplaced samples), we used our method to infer the origin and ancestry of the 125 accessions from the recent expansion of *A. thaliana* into North America. We did this by ignoring the locations of these samples and predicting their current and ancestral locations based only on the trees and the locations of the remaining samples. Figure 6 shows the predicted current (purple) and ancestral locations, averaged over 10^2 loci per chromosome, for each North American accession.

This does two things. First, if we knew the time of colonization, we could read off where we expect the colonization to have originated from. Instead of simply asking where are the most closely related samples, we allow all the lineages to move backwards in time, naturally correcting for the movement

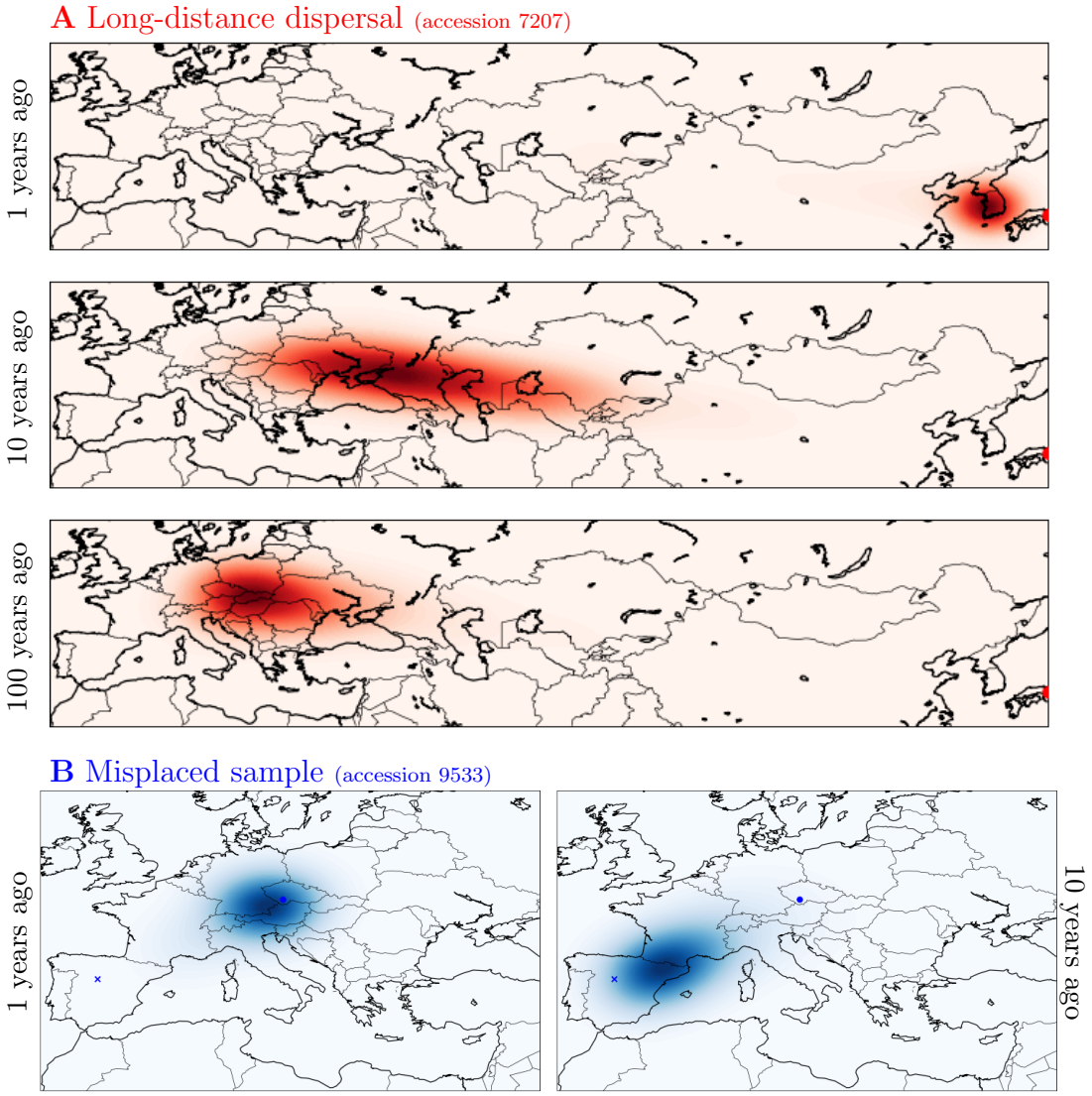


Figure 5: (Continued on the following page.)

Figure 5: **Locating genetic ancestors to detect the existence and source of (A) recent long-distance dispersal and (B) misplaced samples**. Shown are density kernels of the maximum likelihood ancestral locations at 10^3 evenly-spaced loci per chromosome for (**A**) a Japanese accession (100% 'Central Europe' ancestry) and (**B**) an Iberian Relict (100% Relict ancestry) from Spain that we misplaced to the mean location of all samples (Czech Republic). The dots show the assumed sample locations (at the edge of the panes in **A**) and the 'x' in **B** shows the true sample location. Maximum likelihood locations calculated using per-chromosome four-epoch dispersal rates (Figure 3B).

of lineages between the time of sampling and the time of colonization. In this case, the colonization of North America by *A. thaliana* is thought to be about 400 years ago (Exposito-Alonso et al., 2018), suggesting that the majority of the North American accessions came from southern Germany, despite being more closely related, in many cases, to samples in north eastern France today. Second, Figure 6 helps visualize the high degree of relatedness among the North American samples. For instance, a large group of closely related North American samples are also closely related to current-day samples in north eastern France, and appear to coalesce with one another at many loci in the last few hundred years, near the purported time of colonization. On the other hand, there is a handful of more distantly related North American samples, most closely related to present-day samples ranging from northern Germany to Russia, that do not coalesce at most loci in the past 10^3 years, suggesting multiple colonizations of North America, as previously hypothesized (Exposito-Alonso et al., 2018). Encouragingly, connecting the sample and inferred locations (Figure S7), we find that the samples placed in Russia and Slovakia are from near Manistee, Michigan (accessions 1890 and 2202) and the sample placed near northern Germany is the reference, Col-0, from Missouri (accession 6909), consistent with recent independent findings (Shirsekar et al., 2021).

Visualizing alternative geographic ancestries

As mentioned above, the 1001 Genomes dataset is thought to contain samples drawing ancestors from at least two different glacial refugia, including the Non-Relict samples with ancestry predominately from a Balkan refuge (Lee

et al., 2017) and the Iberian Relict samples with substantial ancestry from a refuge in north Africa (Alonso-Blanco et al., 2016; Durvasula et al., 2017;

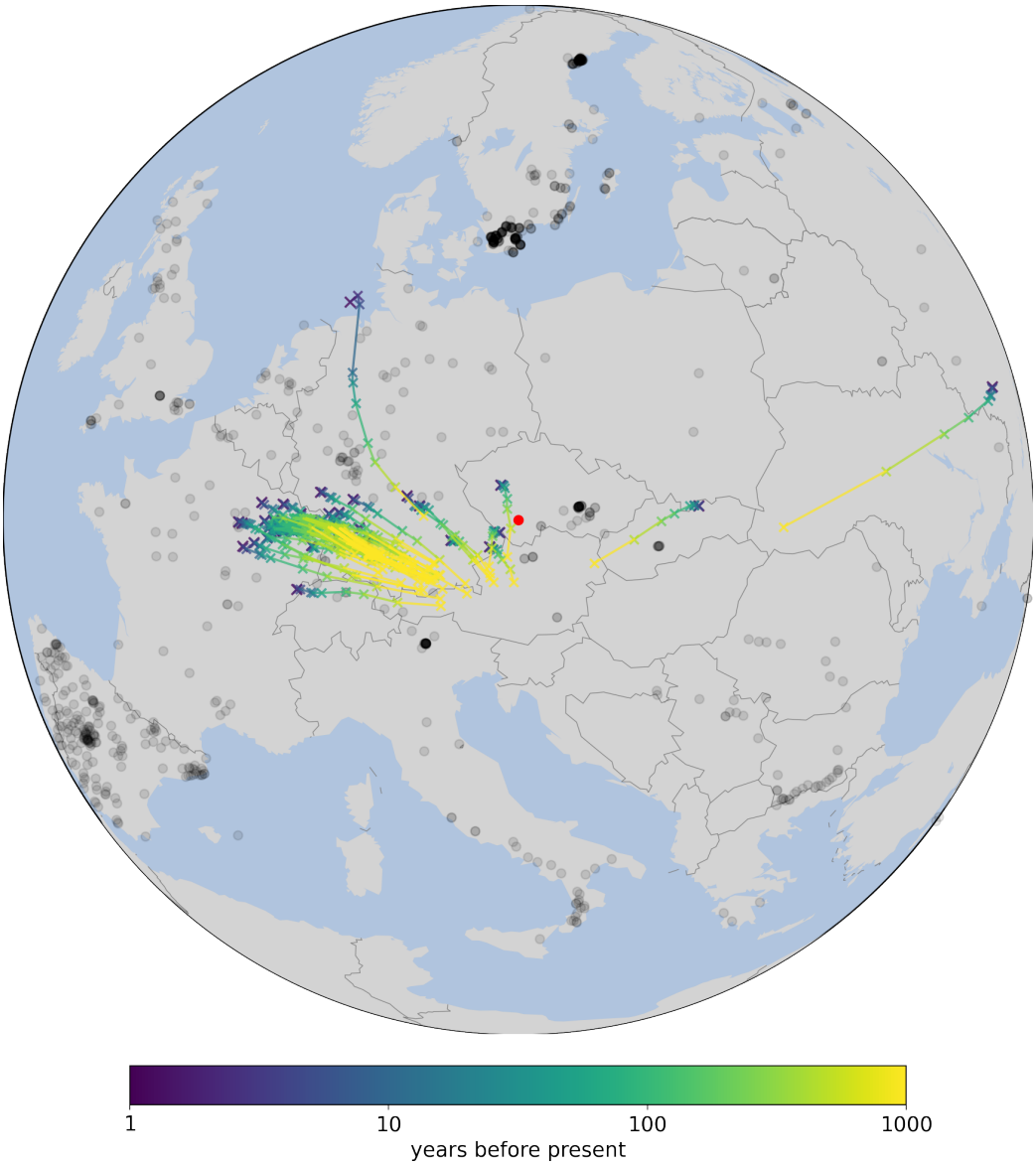


Figure 6: (Continued on the following page.)

Figure 6: **Locating genetic ancestors to identify the source and ancestry of recent expansions/colonizations.** Mean maximum likelihood genetic ancestor locations (averaged over 10^2 evenly-spaced loci per chromosome) of the 125 North American accessions over the last 10^3 years. The black dots are sample locations and the red dot is their mean. Maximum likelihood locations calculated using per-chromosome four-epoch dispersal rates (Figure 3B).

Fulgione et al., 2018). We therefore sought to use our method to visualize these alternative geographic ancestries.

We first compared the geographic ancestries of three representative accessions from Spain (Figure 7): one from a Non-Relict admixture group (accession 6933; 100% 'Spain' ancestry); one from the Relict admixture group (accession 9533; 100% Relict ancestry); and one drawing roughly equal ancestry from both these groups (accession 9530; 57% Relict and 43% 'Spain' ancestry). As expected given the glacial refuge of the Non-Relict lineages is thought to be near the Balkans (Lee et al., 2017), the Non-Relict sample's genetic ancestors move gradually and coherently north east, out of the Iberian Peninsula by 200 years ago and into northern Italy/Austria by 500 years ago. In contrast, a large proportion of the Relict sample's ancestors remain on the Iberian Peninsula for the last 500 years, where the rest of the closely related Iberian Relict samples are clustered. The Admixed sample's genetic ancestors display much more variability, with the rate of return to the east depending on the ancestry at each locus, with many ancestors still remaining in Spain but many already in Italy $\approx$ 200 years ago.

To confirm this is a general pattern beyond these three representative samples, we also looked at the mean ancestral displacements (over 10^2 loci per chromosome) of all mainland Spanish samples from each of the three groups (Relict, Non-Relict, and Admixed; Figure 8). We see that, on average, Non-Relict samples tend to move both further north and further east than the Relict samples, in the direction of the mean sample location and towards the Balkans (the same conclusions are reached if we compare just the 'Spain' and Relict groups, results not shown). The Admixed samples show much more variation in the amount of northward displacement, as expected given these samples could have ancestries from any two or more groups, including two Non-Relict groups (e.g., 'Spain' and 'North Sweden').

Eventually nearly all of an Iberian Relict sample's ancestors are pulled

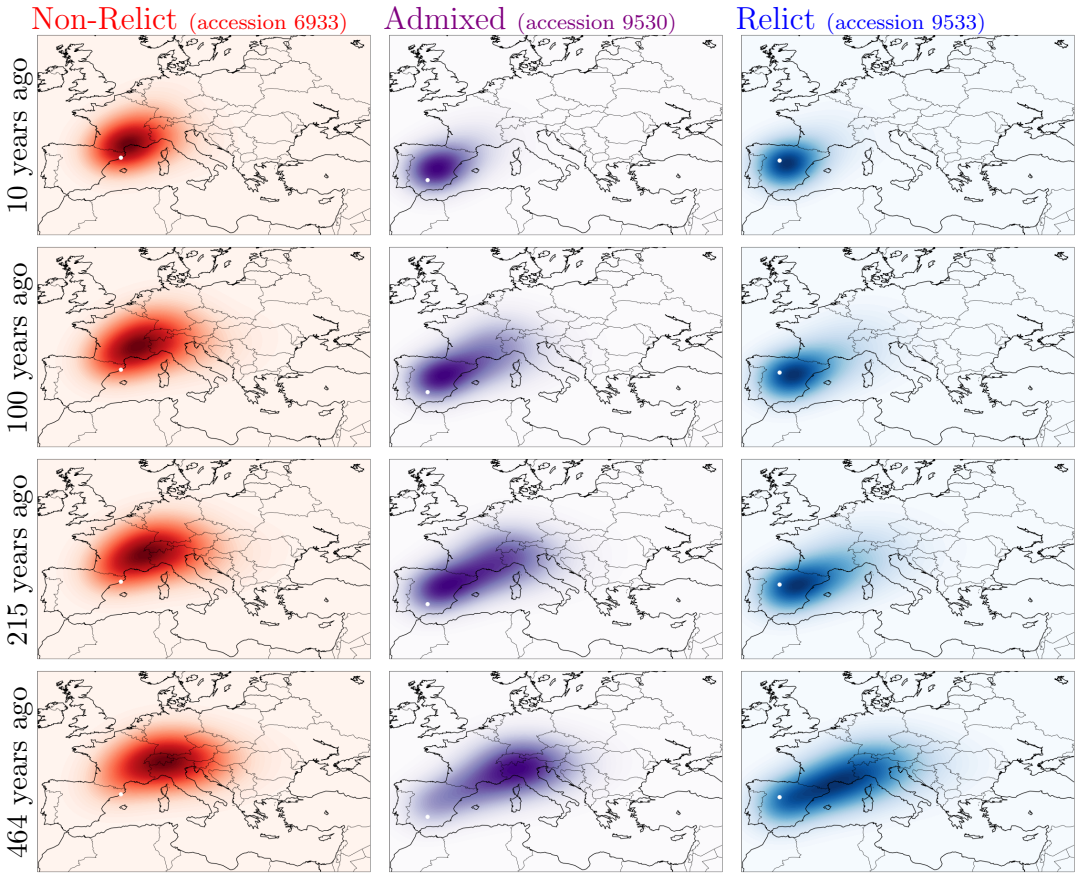


Figure 7: **Locating genetic ancestors to depict alternative geographic histories**. Genetic ancestors of three samples from Spain from three different admixture groups: a Non-Relict (100% 'Spain'), an Admixed (43% 'Spain', 57% Relict), and an Iberian Relict (100% Relict). Shown are density kernels of the maximum likelihood ancestral locations using 10^3 evenly-spaced loci per chromosome. White dots show the sample locations. Maximum likelihood locations calculated using per-chromosome four-epoch dispersal rates (Figure 3B).

east, perhaps because the Iberian Relict ancestry is only represented by $\sim 2\%$ of the samples, but perhaps also because the Iberian Relicts are through to have extensive admixture with the Non-Relicts (Fulgione et al., 2018). There

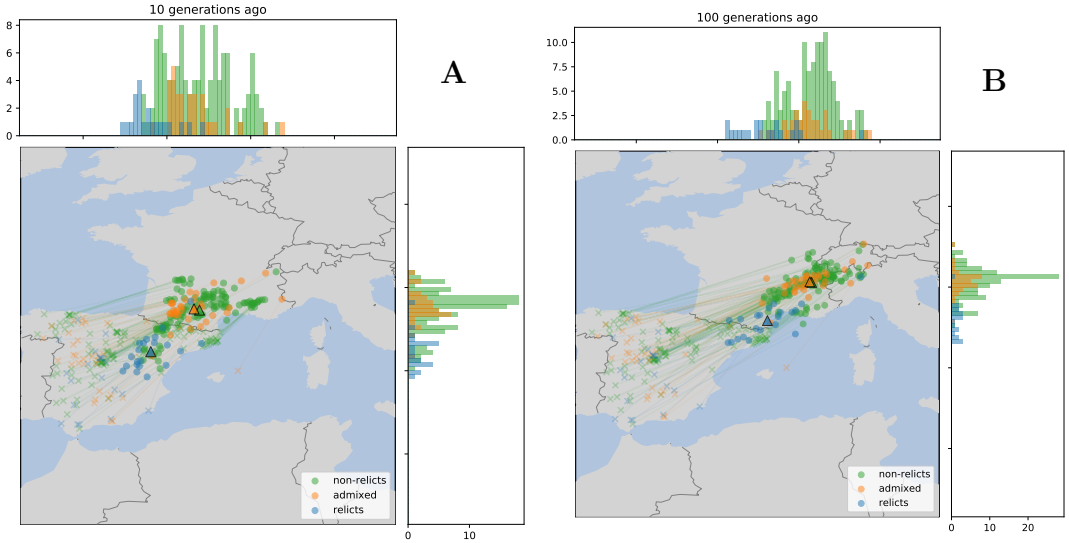


Figure 8: **Relict samples display an alternative geographic history.** Mean genetic ancestor locations of all samples from Spain (excluding the Canary Islands) by admixture group. Data as in Figure 4, with 'x's the sample locations and 'o's the mean ancestor location and lines connecting them. Colours indicate admixture group (Non-Relict is any group besides Relict or Admixed; Admixed is any sample with $< 60\%$ ancestry in all groups; Alonso-Blanco et al., 2016). Triangles indicate mean locations for each admixture group. Maximum likelihood locations calculated using per-chromosome four-epoch dispersal rates (Figure 3B).

are only three samples from Africa in the 1001 Genomes dataset (accessions 6911, 9606, and 9939), all of which show admixture with European groups and are inferred to have geographic ancestries eventually tracing north east (Figure 4). It would be interesting to see if adding the ~ 80 more African genomes currently available (Durvasula et al., 2017) would pull some of the ancestors of these Iberian Relicts towards a possible north African refuge (Durvasula et al., 2017; Fulgione et al., 2018), and help better visualize the Relict-Non-Relict admixture.

Visualizing the sources of admixture

As we have seen, the 1001 Genomes dataset contains a number previously identified Admixed samples (Alonso-Blanco et al., 2016). This provides a nice opportunity to explore how locating recent ancestors can help visualize admixed ancestry and its geographic sources. To do this we plot recent ancestral displacements, as in Figure 4, but this time without averaging over loci, instead plotting the displacement for each locus. Figure 9A shows the inferred displacements (in the past 10 generations) for 6 of the more striking Admixed samples, both in terms of the direction and magnitude of ancestral displacements. To help summarize the range of displacements from each sample, the 'windrose' insets show a histogram of displacement directions, weighted by displacement length and coloured by chromosome. In many cases there is a large spread in the direction of the displacements, illustrating multiple contributing geographic ancestries. For example, the sample from France discussed earlier (accession 9933) appears to coalesce recently with lineages from both relatively near north east and relatively far east (consistent with its previous admixture assignment containing substantial 'Germany' and 'Central Asia' ancestry) and the sample from Romania (accession 9743) appears to get ancestry from both the west and east (consistent with its previous admixture assignment containing substantial 'Italy, Balkans, & Caucasus' and 'Central Asia' ancestry).

There is also information contained in the correlation of ancestral movements across the genome. To demonstrate how geographic ancestries are correlated across the genome, Figure 9B shows each displacement emanating from it's location along a linear representation of the genome (rather than all emanating from the same place, as in the map) for the French and Romanian sample (see Figure S8 for the other 4 samples). With this visual we see that the French sample gets much of its most eastern ancestry from chromosomes 3 and 5 (green and purple) while the Romanian sample gets most of its eastern ancestry from chromosome 4 (red). This figure also serves to illustrate just how much information is being inferred – at any time in the past, we can estimate the locations of thousands of genetic ancestors for thousands of samples, providing a rich source of information to explore and use to test hypotheses.

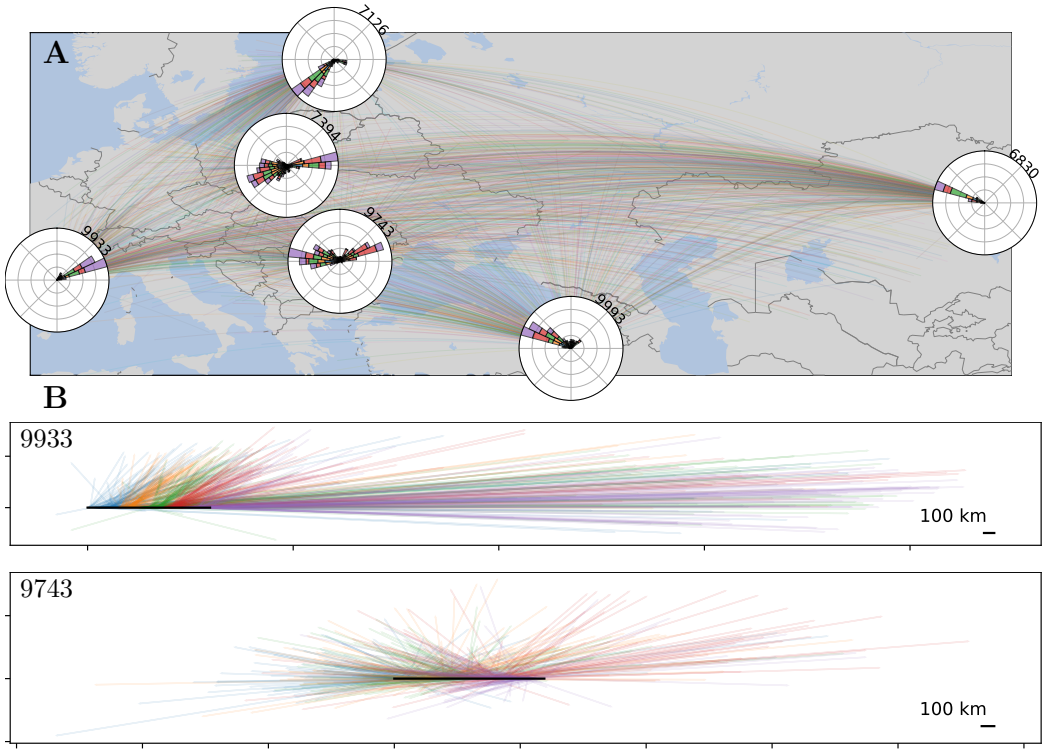


Figure 9: **Locating genetic ancestors to visualize admixture and its sources**. **(A)** The curves on the map are great circle paths connecting the sample locations (center of 'windrose' insets) to the maximum likelihood ancestor locations 10 generations ago at 10^2 evenly-spaced loci per chromosome for 6 striking Admixed samples (accession numbers as labels). The windrose insets show a histogram of directions from the sample to the ancestors, weighted by distance. Colours indicate chromosomes, as in Figure 3. **(B)** The black horizontal line represents the genome and the coloured lines show the ancestral displacements in longitude and latitude (converted to kms) from that position of the genome. See Figure S8 for the genome-view of ancestral displacements of all 6 samples. Maximum likelihood locations calculated using per-chromosome four-epoch dispersal rates (Figure 3B).

Discussion

Summary of main results

We have developed a method that uses a sequence of trees along a recombining genome – a genome-wide genealogy – to estimate individual-level dispersal rates and locate genetic ancestors (Figure 1). At the core of our method is a simple model of Brownian motion, allowing likelihoods to be quickly calculated from shared evolutionary times and sample locations. This also allows us to work in continuous space, negating the need to group individuals into discrete populations. On top of this we layer on importance sampling to correct for bias in inferred branch lengths, a time cutoff to ignore strong violations of the model in the deep past, and multiple epochs to allow the dispersal rate to vary through time. Simulation tests show that our method can accurately infer dispersal rate (with a slight overestimate caused by errors in tree inference), detect shifts in the dispersal rate over time, and trace major geographic ancestries hundreds of generations into the past (Figure 2). Applying our method to more than one thousand *Arabidopsis thaliana* genomes across a huge geographic expanse (Alonso-Blanco et al., 2016), we find strong evidence for very rapid recent dispersal (Figure 3), especially across longitude, and show how locating genetic ancestors can detect and visualize *i*) long-distance dispersal events (Figures 4-5), *ii*) the source of population expansions (Figure 6), *iii*) alternative geographic ancestries (Figures 7-8), and *iv*) admixture (Figure 9).

Comparison with previous methods

The idea of using trees to estimate continuous ancestral characters and their rate of change is an old one. This was originally applied to population-level characters, such as the frequency of genes in a population and their rate of genetic drift, in a phylogenetic context (Cavalli-Sforza et al., 1964; Cavalli-Sforza and Edwards, 1967; Edwards, 1970; Felsenstein, 1973, 1985; Grafen, 1989). DNA sequencing later allowed the inference of a single tree relating individual samples for a sufficiently long non-recombining sequence (as found on the human Y chromosome, in a mitochondrial genome, or in the nuclear genome of a predominately non-recombining species), which led to estimates of dispersal rates and ancestral locations under the banner of 'phylogeography' (Avise, 2009; Knowles, 2009). Phylogeography has proven in-

credibly useful, especially to infer the geographic origin and spread of viruses (Biek et al., 2007; Lemey et al., 2009, 2010; Bedford et al., 2010; Volz et al., 2013), such as SARS-CoV-2 (Worobey et al., 2020; Lemey et al., 2020; Dellicour et al., 2021a,b; Martin et al., 2021). Extending phylogeography to frequently-recombining sequences is not straightforward as there are then many true trees that relate the samples. Only recently has it become feasible to infer the sequence of trees, and their branch lengths, along a recombining genome (Rasmussen et al., 2014; Speidel et al., 2019; Wohns et al., 2021). Our method capitalizes on this advance to use some of the enormous amount of information contained in a tree sequence of a large sample in a recombining species.

A related method was recently demonstrated by Wohns et al. (2021), who inferred the geographic location of coalescent nodes in a `tskit` tree sequence (Kelleher et al., 2018), where information about nodes and branches are shared between trees. Our approach differs from theirs in a number of ways. First, they utilize the sharing of information about nodes and branches across trees to very efficiently geographically locate every node exactly once, by locating a 'parent' node at the midpoint between its two 'child' nodes geographic locations and iterating up the entire tree sequence simultaneously (rather than up each local tree individually). In addition to being fast this has the advantage of using information from all the trees in a tree sequence. In contrast, we locate ancestors independently at each local tree we consider. As some ancestors (represented as nodes and branches) are shared between nearby trees along the genome (though we don't know precisely for how long since we lose that information when converting the `Relate`-inferred genealogy into a tree sequence), we avoid locating the same ancestors multiple times by using only a sample of the trees (e.g., for *A. thaliana* we uniformly sampled 5000 of the ∼200,000 trees, $\sim 2.5\%$). Our approach therefore uses less of the information in the tree sequence, in this sense. (Note that while we choose trees that have low linkage disequilibrium with one another and are therefore essentially unlinked, in the very recent past they will share ancestors but will quickly become independent (Wakeley et al., 2012).) On the other hand, when locating an ancestor we not only use information 'below' this ancestor (i.e., its descendants' locations and relations) but also the information 'above' the ancestor, due to the ancestor's lineage sharing evolutionary time and a recent common ancestor with non-descendant lineages. Further advantages of our method include the ability to estimate dispersal rates and uncertainties in ancestor locations (as we have taken a parametric approach), the ability

to locate ancestors at any time (not just at coalescent nodes but also at any point along a branch), and accounting for uncertainty in the branch lengths (using importance sampling; this could be extended to capture uncertainty in the topologies as well once they can be efficiently sampled).

An exciting possibility is a merger of these two methods, to efficiently use information from all the trees in the tree sequence and all the information above and below ancestors. One approach might be to use something like the inside-outside algorithm in (Wohns et al., 2021), which they use for node times but not locations. A second approach would be to model Brownian motion on the full ancestral recombination graph (e.g. using a program like `ARGweaver`; Rasmussen et al., 2014). The latter approach would allow one to visualize the geographic splitting and coalescence of the ancestors of a genome backwards in time.

Future directions

We have chosen to stick with a very simple model of Brownian motion with a piecewise-constant dispersal rate over time. There are a number of extensions that could readily be applied. For instance, we could allow dispersal rates to vary between branches (O'Meara et al., 2006), e.g., to compare dispersal rates in different parts of a species' range. Or we could model dispersal under a more complex model, like the Early Burst (Harmon et al., 2010) or a Lévy process (Landis et al., 2013), which may help identify periods of range expansion or sudden long-range dispersal. Alternatively we could take a Bayesian approach, allowing much greater flexibility and the incorporation of many recent advances in phylogeography. For example, one could then model dispersal as a relaxed random walk (Lemey et al., 2010), which may be more appropriate for sample locations that are very non-normal and could incorporate habitat boundaries. Second, given that there is large variance in the inferred locations of distant ancestors at any one locus (Schluter et al., 1997), but very many loci, we could take an 'empirical Bayes' approach and use the posteriors on ancestral locations over many loci to set a prior for a given locus. This might be especially helpful at deeper times, e.g., tracing human ancestors back 100s of thousands of years, where the noise in the ancestral location at any one locus is large, yet we can be relatively certain that the majority of lineages are in Africa. We might alternatively set priors to test hypotheses, e.g., if we surmise there were multiple glacial refugia during the last glaciation we can set a prior on ancestral location with peaks

at these hypothesized locations and infer what percent of a sample's lineages descended from each. Models of ancestral locations based on past climatic- and ecological-niche models could provide a rich source of data for building such priors, and given the large amounts data available in recombining sequences these models could be subject to rigorous model choice. Finally, it might also be interesting to take a more model-agnostic approach and use machine learning. For example, `Locator` (Battey et al., 2020) uses deep neural networks to infer the locations of extant individuals from unphased genotype data. In essence this means `Locator` is both determining the relationships between samples and locating them. Separating these two steps by first inferring a tree sequence and then supplying information from this tree sequence to the deep neural network may both improve location estimates for extant individuals and allow such a method to locate genetic ancestors.

Our approach relies on the locations of the current-day samples. While we have shown that we can learn much about the geographical history of a species with this approach, its accuracy is necessarily limited. For example, if historical parts of the range are undersampled in a particular dataset the method will struggle to locate ancestors in these regions, particularly further into the past, as we saw with the Iberian Relicit samples with ancestry from the putative north African refugia. Similarly, if a species' range has shifted such that few present day individuals exist in portions of the historic range, we will often not infer ancestors to be in the currently sparsely-occupied portions. Other large scale population movements, such as one population replacing another, may also partially obscure the geographic locations of ancestors. Over the past decade we have learned about numerous large-scale movements of human populations alongside the expansions of archaeological cultures, a fact fairly hidden from view by contemporary samples that only ancient DNA could bring to light (Slatkin and Racimo, 2016; Reich, 2018). One obvious way to improve our method then, is to include ancient samples. Given that it is now possible to include high-coverage ancient genomes in tree sequences (Speidel et al., 2021; Wohns et al., 2021), it is straightforward to include this information in our likelihoods (ancient samples are treated as any other, we calculate the shared times of these lineages with themselves and with all other sample lineages), influencing both our dispersal estimates and inferred ancestral locations. This should help, in particular, in locating ancestors that are closely related to the ancient samples and for detecting population-scale movements, such as range shifts, contractions, and replacements.

Materials and Methods

Here we describe our methods to estimate dispersal rates and locate genetic ancestors and how we applied these to simulations and *Arabidopsis thaliana*.

The probability of the sampled locations given a tree and the dispersal rate

We first derive the probability of the sampled locations, $\mathbf{L}$, given a bifurcating tree topology (genealogy), $\mathcal{G}$, and associated branch lengths (times), $\mathcal{T}$, which describe the coalescent history of the sample, and the dispersal rate (covariance matrix), $\mathbf{\Sigma}$. We derive this probability, $\mathbb{P}(\mathbf{L}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma})$, compounding the normally-distributed dispersal events each generation to give Brownian motion down the tree in a similar manner to phylogenetic least squares regression (Grafen, 1989; Harmon, 2019).

Let $\boldsymbol{\ell}_i = [\ell_{i,1}\ \ell_{i,2}\ \ldots\ \ell_{i,m}]^\intercal$ be the m-dimensional location of sample i, $\mathbf{L} = [\boldsymbol{\ell}_1\ \boldsymbol{\ell}_2\ \ldots\ \boldsymbol{\ell}_n]^\intercal$ the $n \times m$ matrix of spatial locations for all n samples, and $\boldsymbol{\ell} = [\ell_{1,1}\ \ell_{2,1}\ \ldots\ \ell_{n,1}\ \ell_{1,2}\ \ell_{2,2}\ \ldots\ \ell_{n,2}\ \ldots\ \ell_{1,m}\ \ell_{2,m}\ \ldots\ \ell_{n,m}]^\intercal$ the matrix of locations represented as a vector of length nm. Let $\mathbf{S}_{\mathcal{G},\mathcal{T}}$ be the $n \times n$ matrix of shared evolutionary time (in generations) between each sample lineage in the tree defined by $\mathcal{G}$ and $\mathcal{T}$.

Then, assuming per generation dispersal is multivariate normal with mean displacement $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$, the probability of the locations given the tree is

$$\mathbb{P}(\mathbf{L}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma}) = \mathbb{P}(\boldsymbol{\ell}|\mathbf{S}_{\mathcal{G},\mathcal{T}}, \mathbf{\Sigma}) \sim \mathcal{N}\left(\mathbf{D}\widehat{\boldsymbol{\ell}_A}, \mathbf{S}_{\mathcal{G},\mathcal{T}} \otimes \mathbf{\Sigma}\right), \tag{1}$$

where

$$\widehat{\boldsymbol{\ell}_A} = [(\mathbf{1}^\intercal \mathbf{S}_{\mathcal{G},\mathcal{T}}^{-1} \mathbf{1})^{-1} (\mathbf{1}^\intercal \mathbf{S}_{\mathcal{G},\mathcal{T}}^{-1} \mathbf{L})]^\intercal \tag{2}$$

is the maximum likelihood estimate for the location of the most recent common ancestor given the tree, $\mathbf{1}$ is a column vector of n ones, and $\mathbf{D}$ is a $nm \times m$ design matrix whose i, j^{th} entry is 1 if $(j-1)n < i \leq jn$ and 0 otherwise.

Mean centering the locations

We can remove any dependence on the (unknown) location of the most recent common ancestor by mean centering the data. The mean centered locations

and associated matrix of shared times are $\mathbf{X} = \mathbf{ML}$ and $\mathbf{V}_{\mathcal{G},\mathcal{T}} = \mathbf{MS}_{\mathcal{G},\mathcal{T}}\mathbf{M}^{\intercal}$, where $\mathbf{M}$ is an $(n-1) \times n$ matrix with $n - 1/n$ on the diagonal and $-1/n$ elsewhere. We only use those (sub)trees (see below) with $n > 1$ samples (i.e., trees with only one sample contain no information about the dispersal rate if we do not know where the ancestor was). We then have

$$\mathbb{P}(\mathbf{X}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma}) = \mathbb{P}(\mathbf{x}|\mathbf{V}_{\mathcal{G},\mathcal{T}}, \mathbf{\Sigma}) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{V}_{\mathcal{G},\mathcal{T}} \otimes \mathbf{\Sigma}\right), \tag{3}$$

where $\mathbf{x}$ is the vector representation of $\mathbf{X}$ (as $\boldsymbol{\ell}$ is to $\mathbf{L}$).

Chopping the tree into subtrees

We will usually want to only consider dispersal more recently than some cut-off time, T, since deeper genealogical history may contain little geographic information (Wilkins, 2004). When this is the case we cut the full tree of at T, leaving us with $n_T \in [1, n]$ subtrees, where all coalescent times are $< T$, and we then calculate the shared times in each subtree independently. Letting $\mathbb{P}(\mathbf{x}_i|\mathbf{V}_{\mathcal{G}_i,\mathcal{T}_i}, \mathbf{\Sigma})$ be the probability of the (mean centered and vectorized) locations of the samples in subtree i given the (mean centered) matrix of shared evolutionary times in that subtree and the dispersal rate, the probability of the locations of all samples is

$$\mathbb{P}(\mathbf{X}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma}, T) = \prod_{i=1}^{n_T} \mathbb{P}(\mathbf{x}_i|\mathbf{V}_{\mathcal{G}_i,\mathcal{T}_i}, \mathbf{\Sigma}) \sim \prod_{i=1}^{n_T} \mathcal{N}\left(\mathbf{0}, \mathbf{V}_{\mathcal{G}_i,\mathcal{T}_i} \otimes \mathbf{\Sigma}\right). \tag{4}$$

This has the added benefit of expressing the probability as a function of smaller submatrices of shared evolutionary times, which are faster to invert.

Dividing time into epochs

We can extend this model to allow dispersal to vary through time by dividing time into epochs and assuming dispersal is only constant within each. If we divide time into K epochs (as described by a vector of split times, $\mathbf{t}$), so that the (mean centered) shared time matrix in the ith epoch is $\mathbf{V}^{(i)}_{\mathcal{G},\mathcal{T}}$, then

$$\begin{aligned}\mathbb{P}(\mathbf{X}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_K, \mathbf{t}) &= \mathbb{P}(\mathbf{x}|\mathbf{V}^{(1)}_{\mathcal{G},\mathcal{T}}, ..., \mathbf{V}^{(E)}_{\mathcal{G},\mathcal{T}}, \mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_E) \\ &\sim \mathcal{N}\left(\mathbf{0}, \sum_{i=1}^{K} (\mathbf{V}^{(i)}_{\mathcal{G},\mathcal{T}} \otimes \mathbf{\Sigma}_i)\right).\end{aligned} \tag{5}$$

Thus different dispersal epochs are easily accounted for.

Importance sampling

The calculations above, which give the likelihood of the dispersal rate given the sample locations, were all predicated on knowing the tree with certainty, which will not be the case when inferring trees from genetic data. Further, inferring a tree may involve assumptions (such as panmixia) that are inconsistent with the model we are using. To deal with the uncertainty and bias we calculate the likelihood of our parameters given the data using importance sampling, a likelihood ratio, and Monte Carlo approximation. Importance sampling corrects the expectation of the likelihood by reweighting draws from an 'incorrect' (proposal) distribution to match the 'correct' (target) distribution. These importance weights downweight draws from the proposal distribution that are likely under the proposal distribution but unlikely under the target distribution and upweight draws that are likely in the target distribution and unlikely under the proposal distribution. Importance sampling techniques to sample genealogies in population genetics have been applied a number of settings as coalescent models provide convenient priors on trees and it is often challenging to sample genealogies consistent with data (Griffiths and Tavare, 1997; Stephens and Donnelly, 2000; Meligkotsidou and Fearnhead, 2007). See Stern et al. (2019, 2021) for recent applications of these ideas to marginal trees inferred along a recombining sequence, whose general approach we follow below.

The data we have are the locations of the samples, $\mathbf{L}$, and their haplotypes, $\mathbf{H}$. From this data we want to infer two unknowns, a tree topology, $\mathcal{G}$, and associated branch lengths, $\mathcal{T}$, and use these to estimate the likelihood the dispersal rate, $\mathbf{\Sigma}$. Put another way, we want to estimate $\mathbb{E}_{\mathcal{G},\mathcal{T}|\mathbf{\Sigma}}\left[\mathbb{P}(\mathbf{L},\mathbf{H}|\mathcal{G},\mathcal{T},\mathbf{\Sigma})\right]$, which is the likelihood of our parameters given the data, integrated over the unknowns.

Current methods to infer tree topologies and times along a recombining sequence (Rasmussen et al., 2014; Speidel et al., 2019; Kelleher et al., 2019; Wohns et al., 2021) assume panmictic, well-mixed populations. This implies we cannot sample topologies and branch lengths, $\mathcal{G}$ and $\mathcal{T}$, under our spatial model, creating potential bias. To correct for this bias we use importance sampling,

$$\begin{aligned} \mathrm{L}(\mathbf{\Sigma}) &:= \mathbb{E}_{\mathcal{G},\mathcal{T}|\mathbf{\Sigma}}\left[\mathbb{P}(\mathbf{L},\mathbf{H}|\mathcal{G},\mathcal{T},\mathbf{\Sigma})\right] \\ &= \mathbb{E}_{\mathcal{G},\mathcal{T}|\mathbf{H},\text{panmixia}}\left[\frac{\mathbb{P}(\mathbf{L},\mathbf{H}|\mathcal{G},\mathcal{T},\mathbf{\Sigma})\mathbb{P}(\mathcal{G},\mathcal{T}|\mathbf{\Sigma})}{\mathbb{P}(\mathcal{G},\mathcal{T}|\mathbf{H},\text{panmixia})}\right], \end{aligned} \tag{6}$$

where $\mathbb{P}(\mathcal{G}, \mathcal{T}|\mathbf{H}, \text{panmixia})$ is the distribution of topologies and branch lengths we sample from.

The probabilities containing the genetic data, $\mathbf{H}$, are complicated to calculate. To simplify we divide by the likelihood of panmixia given the data, $\text{L}(\text{panmixia}) = \mathbb{P}(\mathbf{H}|\text{panmixia})$, to work with the likelihood ratio

$$\begin{aligned}
\text{LR}(\mathbf{\Sigma}) &:= \frac{\text{L}(\mathbf{\Sigma})}{\text{L}(\text{panmixia})} \\
&= \mathbb{E}_{\mathcal{G},\mathcal{T}|\mathbf{H},\text{panmixia}}\left[\frac{\mathbb{P}(\mathbf{L}, \mathbf{H}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma})\mathbb{P}(\mathcal{G}, \mathcal{T}|\mathbf{\Sigma})}{\mathbb{P}(\mathcal{G}, \mathcal{T}|\mathbf{H}, \text{panmixia})\mathbb{P}(\mathbf{H}|\text{panmixia})}\right] \\
&= \mathbb{E}_{\mathcal{G},\mathcal{T}|\mathbf{H},\text{panmixia}}\left[\frac{\mathbb{P}(\mathbf{L}, \mathbf{H}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma})\mathbb{P}(\mathcal{G}, \mathcal{T}|\mathbf{\Sigma})}{\mathbb{P}(\mathbf{H}, \mathcal{G}, \mathcal{T}|\text{panmixia})}\right] \\
&= \mathbb{E}_{\mathcal{G},\mathcal{T}|\mathbf{H},\text{panmixia}}\left[\frac{\mathbb{P}(\mathbf{H}|\mathcal{G}, \mathcal{T})\mathbb{P}(\mathbf{L}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma})\mathbb{P}(\mathcal{G}, \mathcal{T}|\mathbf{\Sigma})}{\mathbb{P}(\mathbf{H}|\mathcal{G}, \mathcal{T})\mathbb{P}(\mathcal{G}, \mathcal{T}|\text{panmixia})}\right] \\
&= \mathbb{E}_{\mathcal{G},\mathcal{T}|\mathbf{H},\text{panmixia}}\left[\frac{\mathbb{P}(\mathbf{L}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma})\mathbb{P}(\mathcal{G}, \mathcal{T}|\mathbf{\Sigma})}{\mathbb{P}(\mathcal{G}, \mathcal{T}|\text{panmixia})}\right],
\end{aligned} \tag{7}$$

where the third step assumes the genetic data ($\mathbf{H}$) is conditionally independent of the spatial parameters ($\mathbf{\Sigma}$ or panmixia) and locations ($\mathbf{L}$) given the tree ($\mathcal{G}$ and $\mathcal{T}$). We can approximate this expectation using Monte Carlo sampling

$$\widehat{\text{LR}(\mathbf{\Sigma})} = \frac{1}{M}\sum_{m=1}^{M}\frac{\mathbb{P}(\mathbf{L}|\mathcal{G}_m, \mathcal{T}_m, \mathbf{\Sigma})\mathbb{P}(\mathcal{G}_m, \mathcal{T}_m|\mathbf{\Sigma})}{\mathbb{P}(\mathcal{G}_m, \mathcal{T}_m|\text{panmixia})}, \tag{8}$$

where $\mathcal{G}_m, \mathcal{T}_m \sim \mathbb{P}(\mathcal{G}, \mathcal{T}|\mathbf{H}, \text{panmixia})$ is sampled using Markov chain Monte Carlo.

We make two final simplifications. First, we will use a model of branching Brownian motion for $\mathbb{P}(\mathcal{G}, \mathcal{T}|\mathbf{\Sigma})$ and the standard neutral coalescent for $\mathbb{P}(\mathcal{G}, \mathcal{T}|\text{panmixia})$. Under both of these models the probability of the topology, $\mathbb{P}(\mathcal{G}|\text{panmixia})$, is equivalent (and uniform). Second, we will use `Relate` (Speidel et al., 2019) to infer topologies and branch lengths. `Relate` returns a single topology and allows resampling over branch lengths conditional on this topology. We therefore take the topology as given and integrate only

over branch lengths. Putting these two simplifications together,

$$
\begin{aligned}
\widehat{\mathrm{LR}(\mathbf{\Sigma})} &= \frac{1}{M}\sum_{m=1}^{M}\frac{\mathbb{P}(\mathbf{L}|\mathcal{G}_m,\mathcal{T}_m,\mathbf{\Sigma})\mathbb{P}(\mathcal{G}_m,\mathcal{T}_m|\mathbf{\Sigma})}{\mathbb{P}(\mathcal{G}_m,\mathcal{T}_m|\text{panmixia})} \\
&= \frac{1}{M}\sum_{m=1}^{M}\frac{\mathbb{P}(\mathbf{L}|\mathcal{G}_m,\mathcal{T}_m,\mathbf{\Sigma})\mathbb{P}(\mathcal{T}_m|\mathcal{G}_m,\mathbf{\Sigma})\mathbb{P}(\mathcal{G}_m|\mathbf{\Sigma})}{\mathbb{P}(\mathcal{T}_m|\mathcal{G}_m,\text{panmixia})\mathbb{P}(\mathcal{G}_m|\text{panmixia})} \\
&= \frac{1}{M}\sum_{m=1}^{M}\frac{\mathbb{P}(\mathbf{L}|\mathcal{G}_m,\mathcal{T}_m,\mathbf{\Sigma})\mathbb{P}(\mathcal{T}_m|\mathcal{G}_m,\mathbf{\Sigma})}{\mathbb{P}(\mathcal{T}_m|\mathcal{G}_m,\text{panmixia})} \\
&\approx \frac{1}{M}\sum_{m=1}^{M}\frac{\mathbb{P}(\mathbf{L}|\mathcal{G},\mathcal{T}_m,\mathbf{\Sigma})\mathbb{P}(\mathcal{T}_m|\mathcal{G},\mathbf{\Sigma})}{\mathbb{P}(\mathcal{T}_m|\mathcal{G},\text{panmixia})},
\end{aligned}
\tag{9}
$$

where $\mathcal{T}_m \sim \mathbb{P}(\mathcal{T}|\mathbf{H},\text{panmixia},\mathcal{G})$ is sampled using Markov chain Monte Carlo. Note that because the probability of a topology is the same in the two models, and therefore cancels out, our method can immediately be extended to integrate over both topologies and branch lengths.

To find the maximum likelihood estimate of $\mathbf{\Sigma}$ we numerically search for the $\mathbf{\Sigma}$ that maximizes this likelihood ratio. We measure the variance around this estimate with the Hessian matrix of the likelihood surface returned by the numerical search. To use information from multiple loci, e.g., to estimate a per-chromosome or genome-wide dispersal parameter, we can multiply the likelihood ratios together to give a composite likelihood ratio. This is a composite likelihood because it ignores correlations in the trees between loci (Hudson, 2001; Larribe and Fearnhead, 2011; Varin et al., 2011). The maximum likelihood of genome-wide parameters from composite likelihood are known to be statistically consistent in the presence of such correlations (Wiuf, 2006), but the likelihood surface is overly peaked due to ignoring the correlations in the data. However, in practice we take loci that are far enough apart that their trees are only weakly correlated more than a few tens of generations back.

We next show how we derive the probability distributions of the branch lengths in the tree that we use as the weights in the importance sampler, $\mathbb{P}(\mathcal{T}|\mathcal{G},\text{panmixia})$ and $\mathbb{P}(\mathcal{T}|\mathcal{G},\mathbf{\Sigma})$.

The probability of the coalescence times given the tree topology and panmixia

`Relate` and other tree construction methods infer the times in the trees under a model of panmixia with variable population size through time (with this demographic history inferred genomes as part of the method). The probability of the coalescence times given the tree topology and panmixia, $\mathbb{P}(\mathcal{T}|\mathcal{G}, \text{panmixia})$, can be derived from the standard neutral coalescent allowing for (stepwise) changing effective population size (Griffiths and Tavare, 1994; Meligkotsidou and Fearnhead, 2007).

Let the number of samples be n and let $\mathbf{U}$ be the $n-1$ coalescence times, ordered from most recent to most historic. Then

$$\mathbb{P}(\mathcal{T}|\mathcal{G}, \text{panmixia}) = \prod_{i=1}^{n-1} \mathbb{P}(U_i = u_i | U_{i-1} = u_{i-1}), \tag{10}$$

where

$$\begin{aligned} &\mathbb{P}(U_i = u_i | U_{i-1} = u_{i-1}) \\ &= \binom{n-(i-1)}{2} \frac{1}{2N(u_i)} \exp\left[-\binom{n-(i-1)}{2} [\Lambda(u_i) - \Lambda(u_{i-1})] \right] \end{aligned} \tag{11}$$

with

$$\Lambda(u) = \int_0^u \frac{1}{2N(t)} \mathrm{d}t, \tag{12}$$

where $N(t)$ is the effective population size at time t in the past.

This says that the coalescence times are produced by a Markov process, so that the probability of the ith coalescence being at time u_i is independent of $\mathbf{U}$ once conditioned on the time of the previous coalescence, u_{i-1}. Between time u_{i-1} and u_i we have $n-(i-1)$ lineages. The probability of coalescence at time u_i is then approximately distributed as a time-inhomogeneous exponential random variable with instantaneous rate $\lambda(u) = \binom{n-(i-1)}{2} / [2N(u)]$.

When effective population size is piecewise constant, split into K epochs with effective population size N_k between time τ_k and τ_{k+1}, then (see Appendix A of Stern et al., 2021)

$$\Lambda(u) = \left[\sum_{k=0}^{b(u)} \frac{1}{2N_k} (\tau_{k+1} - \tau_k) \right] + \frac{1}{2N_{b(u)+1}} (u - \tau_{b(u)+1}), \tag{13}$$

where $b(u) = \max(k \in \{0, 1, 2, ..., K-1\} : u > \tau_{k+1})$ is the last epoch to end before time u.

If we only want the probability of coalescence times back as far as time T, we consider only these $m \leq n-1$ coalescence times and multiply the probability of these times by the probability of no coalescence between time u_m and T,

$$\begin{aligned} &\mathbb{P}(\mathcal{T}|\mathcal{G}, \text{panmixia}, T) \\ &= \exp\left[-\binom{n-m}{2}[\Lambda(T) - \Lambda(u_m)]\right] \prod_{i=1}^{m} \mathbb{P}(U_i = u_i | U_{i-1} = u_{i-1}). \end{aligned} \tag{14}$$

The probability of the coalescence times given the tree topology and spatial model

A number of different fully spatial models of the coalescent have been developed. However, for our purposes they are computational challenging to work with and so we pursue a simpler analytically-tractable approximation called branching Brownian motion (BBM), or the Brownian-Yule process. Such approximations have been used previously to approximate the short timescale dynamics of genealogies (Edwards, 1970; Rannala and Yang, 1996; Meligkotsidou and Fearnhead, 2007). We view the rate of branching (λ) in the BBM process as a nuisance parameter, but informally it is proportional to the inverse of the local population density and so it might be a useful parameter to explore for future work (for a recent application see Ringbauer et al., 2017).

Let T be the time in the past we wish to start the process (which can be no greater than the time to the most recent common ancestor in the tree). Let n_0 be the number of ancestral lineages at this time (which can be no smaller than 2). Taking a forward in time perspective, let $\mathbf{U}'$ be the birth times, $U'_i = T - U_{n-i}$, i.e., U'_i is the ith split time, measured forward in time from T, transitioning from i to $i+1$ lineages. Then under a pure birth (Yule) process with per capita birth rate λ the joint probability density of the $n - n_0$ split times and ending up with n samples (i.e., no birth once n

lineages existed) is

$$
\begin{aligned}
f(\mathbf{u}', n|\lambda, n_0, T) &= \left(\prod_{i=n_0}^{n-1} \lambda i \exp[-\lambda i(u'_i - u'_{i-1})]\right) \exp[-\lambda n(T - u'_{n-1})] \\
&= \lambda^{n-n_0} \frac{(n-1)!}{(n_0-1)!} \exp\left[-\lambda\left(nT + \sum_{i=n_0}^{n-1} u'_i\right)\right],
\end{aligned} \tag{15}
$$

where we have set $u'_{n_0-1} = 0$. Now, the probability that n_0 lineages produce exactly n lineages in time T under a pure birth process is given by a negative binomial with $n - n_0$ successes, n_0 failures, and success probability $1 - \exp(-\lambda T)$,

$$
\mathbb{P}(n|\lambda, n_0, T) = \binom{n-1}{n-n_0} \exp(-\lambda T n_0)[1 - \exp(-\lambda T)]^{n-n_0}. \tag{16}
$$

The probability of the coalescence times under the Yule process, which we take to be the probability of the times given the topology and spatial model, is therefore proportional to the ratio of Equations (15) and (16),

$$
\begin{aligned}
\mathbb{P}(T|G, \mathbf{\Sigma}) &= \mathbb{P}(\mathbf{U}' = \mathbf{u}'|\lambda, n_0, T, n) \\
&\propto \frac{f(\mathbf{u}', n|\lambda, n_0, T)}{\mathbb{P}(n|\lambda, n_0, T)} \\
&= (n-n_0)! \left(\frac{\lambda \exp(-\lambda T)}{1 - \exp(-\lambda T)}\right)^{n-n_0} \exp\left(-\lambda \sum_{i=n_0}^{n-1} u'_i\right).
\end{aligned} \tag{17}
$$

See Edwards (1970); Rannala and Yang (1996); Meligkotsidou and Fearnhead (2007) for more on the Yule process.

Conditioning on sampling locations

As many studies choose sampling locations before choosing samples, we would like to condition on sampling locations in our inferences of dispersal rate, i.e., use the conditional likelihood $\mathbb{P}(\mathbf{L}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma})/\mathbb{P}(\mathbf{L}|\mathbf{\Sigma})$. The denominator, $\mathbb{P}(\mathbf{L}|\mathbf{\Sigma})$, is the probability of the locations of the descendent tips of a branching Brownian motion given the dispersal rate. This is the numerator $\mathbb{P}(\mathbf{L}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma})$ after the genealogy and branch times have been integrated out. Calculating the integrals over genealogies and branch times could be achieved

by Monte Carlo simulation, but would be computationally intensive. However, as shown by (Meligkotsidou and Fearnhead, 2007), in one dimension (say x) the probability of the locations, $\mathbb{P}(\mathbf{L}|\sigma_x)$, is independent of dispersal, σ_x, if a scale invariant prior is placed on the branching rate. Intuitively, this follows from the fact that if we do not *a priori* know the scale of the coalescent rate then we do not know *a priori* how long the branch lengths are, and so we do not know how fast lineages need to disperse to get to where they are today. Therefore, in that case, conditioning on the locations does not change the likelihood surface of σ_x and so can be ignored in the inference of dispersal. This result also holds in two dimensions if we constrain $\sigma_x = \sigma_y$. Sadly it does not hold in general for arbitrary dispersal matrix $\mathbf{\Sigma}$. For example, if our samples (the end points of BBM) were distributed widely along the x axis but varied little in displacement along the y axis, this would be more likely if $\sigma_y/\sigma_x \ll 1$. Thus, while the overall magnitude of dispersal is not affected by conditioning on the sampling locations, the anisotropy of dispersal may be.

As one of our results is that rates of longitudinal dispersal are higher than latitudinal rates, we wanted to make sure that the lack of conditioning on sampling location did not drive this result. To do this we ran our method looking at only longitudinal dispersal, ignoring the samples latitudes, and then again looking at only latitudinal dispersal, ignoring the samples longitudes. This reduces the problem back to one dimension where the conditioning does not matter (as discussed above and in Meligkotsidou and Fearnhead, 2007). Doing this we find essentially identical dispersal rates as before, justifying our result of a larger longitudinal dispersal rate despite not conditioning on the sample locations.

Maximum likelihood estimates of the dispersal rate given a tree

If we take the tree as fixed, Equation (1) gives the likelihood of the dispersal rate given the locations. This implies the maximum likelihood estimate (MLE) of the dispersal rate given the tree is

$$\widehat{\mathbf{\Sigma}} = \frac{(\boldsymbol{\ell} - \widehat{\boldsymbol{\ell}_A}\mathbf{1}^\intercal)^\intercal \mathbf{S}_{\mathcal{G},\mathcal{T}}^{-1}(\boldsymbol{\ell} - \widehat{\boldsymbol{\ell}_A}\mathbf{1}^\intercal)}{n}. \tag{18}$$

This estimate is, however, biased. Using Equation (3), the restricted maximum likelihood estimate of the dispersal rate given the tree is

$$\widehat{\mathbf{\Sigma}}^* = \frac{\mathbf{x}^\intercal \mathbf{V}_{\mathcal{G},\mathcal{T}}^{-1} \mathbf{x}}{n-1}, \tag{19}$$

which is unbiased and equivalent to $n\widehat{\mathbf{\Sigma}}/(n-1)$.

In practice, when the tree has been chopped into subtrees, dispersal is allowed to vary through time, or we importance sample then we do not have an analytical expression for the MLE dispersal rate and instead numerically search for the dispersal rate which maximizes the likelihood (Equations (4), (5), (9)).

Likelihood of the location of a genetic ancestor given the tree

A point on a branch of a tree represents a genetic ancestor. Choosing one such point, if the ancestor's lineage shares $\mathbf{s}_a$ evolutionary time the sample lineages and s_a time with itself, then, by the properties of the conditional normal distribution, the probability that the ancestor was at location $\boldsymbol{\ell}_a$ is

$$\mathbb{P}(\boldsymbol{\ell}_a | \boldsymbol{\ell}, \mathbf{S}_{\mathcal{G},\mathcal{T}}, \mathbf{\Sigma}, \mathbf{s}_a) \sim \mathcal{N}\left(\widehat{\boldsymbol{\ell}_a}, \widehat{\mathbf{S}}\right), \tag{20}$$

where

$$\widehat{\boldsymbol{\ell}_a} = \widehat{\boldsymbol{\ell}_A} + \mathbf{s}_a^\intercal \mathbf{S}_{\mathcal{G},\mathcal{T}}^{-1} (\boldsymbol{\ell} - \widehat{\boldsymbol{\ell}_A} \mathbf{1}^\intercal) \tag{21}$$

is the maximum likelihood estimate of the ancestor's location and

$$\widehat{\mathbf{S}} = (s_a - \mathbf{s}_a^\intercal \mathbf{S}_{\mathcal{G},\mathcal{T}}^{-1} \mathbf{s}_a) \mathbf{\Sigma} \tag{22}$$

is the uncertainty (covariance) in the maximum likelihood estimate. Here $\mathbf{S}_{\mathcal{G},\mathcal{T}}^{-1}$ is the generalized inverse of $\mathbf{S}_{\mathcal{G},\mathcal{T}}$.

This approach gives the correct maximum likelihood estimate (Equation (21)) but the uncertainty (Equation (22)) is incorrect because the approach implicitly assumes we know with certainty where the most recent common ancestor was.

The correct uncertainty is derived by mean centering. The mean centered matrices of sample locations, $\mathbf{X}$, and shared times among the samples, $\mathbf{V}_{\mathcal{G},\mathcal{T}}$ are derived as above. The mean centered vector of shared times between

the ancestor and the samples is $\mathbf{v}_a = \mathbf{M}(\mathbf{s}_a - \mathbf{S}_{\mathcal{G},\mathcal{T}}\mathbf{1}/n)$, with M and n the mean centering matrix and number of samples, as above. The mean centered shared time of the ancestor with itself is $v_a = \mathbf{m}^\intercal \mathbf{S}_{\mathcal{G},\mathcal{T},a}\mathbf{m}$, where $\mathbf{m}$ is a $(n+1)$-column vector with 1 as the first entry and $-1/n$ elsewhere and $\mathbf{S}_{\mathcal{G},\mathcal{T},a} = \begin{pmatrix} s_a & \mathbf{s}_a^\intercal \\ \mathbf{s}_a & \mathbf{S}_{\mathcal{G},\mathcal{T}} \end{pmatrix}$.

The probability the ancestor was at location $\boldsymbol{\ell}_a$ is then

$$\mathbb{P}(\boldsymbol{\ell}_a|\mathbf{x}, \mathbf{V}_{\mathcal{G},\mathcal{T}}, \boldsymbol{\Sigma}, \mathbf{v}_a, v_a) \sim \mathcal{N}\left(\widehat{\boldsymbol{\ell}_a}, \widehat{\mathbf{S}}^*\right), \tag{23}$$

where the maximum likelihood estimate has not changed but can be rewritten

$$\widehat{\boldsymbol{\ell}_a} = \bar{\mathbf{L}} + \mathbf{v}_a^\intercal \mathbf{V}_{\mathcal{G},\mathcal{T}}^{-1}\mathbf{x}, \tag{24}$$

with $\bar{\mathbf{L}}$ the mean location of the samples. The uncertainty in this estimate is

$$\widehat{\mathbf{S}}^* = \left(v_a - \mathbf{v}_a^\intercal \mathbf{V}_{\mathcal{G},\mathcal{T}}^{-1}\mathbf{v}_a\right)\boldsymbol{\Sigma}. \tag{25}$$

This uncertainty is the correct uncertainty. For example, it produces a linear increase in the variance of the estimate as we move up from a sample when there is only a single sample, i.e., this is simple Brownian motion. When there are two samples with a recent common ancestor at time T, the variance in ancestor location along this tree back to the most recent common ancestor is maximized at the most recent common ancestor, $T/2$, i.e., this is a Brownian bridge. And so on.

As with the dispersal likelihood, to reduce dependencies on distant times we can chop the tree at some time T and use only the subtree containing the ancestor of interest. Often the most recent common ancestor of the resulting subtree occurs more recently than T but we want to infer the location all the way back to T, and so we need to infer the location of the common ancestral lineage past the most recent common ancestor. However, this method extends naturally to times beyond the most recent common ancestor, implicitly modelling simple Brownian motion (a fixed mean and linearly increasing variance) up the stem of the tree. Similarly, the ancestor need not have descendants in the sample – our method implicitly adds simple Brownian motion down the branch leading to the 'hanging' ancestor (which is useful when we want to ignore the location of a sample and locate it or its ancestors using the remaining sample locations and the trees). When there is only a single sample we cannot mean center and instead model simple Brownian motion up from the sample.

We want to allow dispersal rates to vary across epochs. To account for this in locating ancestors, instead of simply calculating the shared evolutionary times between lineages (and then multiplying by a constant dispersal rate) we must explicitly calculate the covariance between each lineage. The covariance between two lineages in any one epoch is the (Kronecker) product of their shared time in that epoch and the dispersal rate in that epoch. The total covariance between the lineages is then the sum of these covariances over epochs. We can then follow the approach above to get the probability distribution of ancestor locations (Equation (23)) after replacing $\mathbf{M}$ with $\mathbf{M} \otimes \mathbf{I}_d$, $\mathbf{m}$ with $\mathbf{m} \otimes \mathbf{I}_d$, $\mathbf{1}$ with $\mathbf{1} \otimes \mathbf{I}_d$, $\mathbf{x}$ (in Equation (24)) with $\mathbf{x} \otimes \mathbf{1}_d$, and $\mathbf{\Sigma}$ (in Equation (25)) with 1, where $\mathbf{I}_d$ is the identity matrix and $\mathbf{1}_d$ is a column vector of 1s, each with the same dimension (d) as $\mathbf{\Sigma}$.

Finally, to integrate over uncertainty and reduce bias we also want to use importance sampling. We do this by replacing $\mathbb{P}(\mathbf{L}|\mathcal{G}, \mathcal{T}, \mathbf{\Sigma})$ in Equation (9) with Equation (23),

$$\widehat{\mathrm{LR}(\boldsymbol{\ell}_a)} = \frac{1}{M} \sum_{m=1}^{M} \frac{\mathbb{P}(\boldsymbol{\ell}_a|\mathbf{x}, \mathbf{V}_{\mathcal{G},\mathcal{T}_m}, \mathbf{\Sigma}, \mathbf{v}_a, v_a)\mathbb{P}(\mathcal{T}_m|\mathcal{G}, \mathbf{\Sigma})}{\mathbb{P}(\mathcal{T}_m|\mathcal{G}, \text{panmixia})}, \tag{26}$$

and generally use the maximum composite likelihood estimates of $\mathbf{\Sigma}$ and λ (e.g., per-chromosome or genome-wide estimates). We measure the variance around this estimate with the Hessian matrix returned by the numerical search.

Best Linear Unbiased Predictions of Locations.

The above approach requires a numerical search for the maximum likelihood ancestor location because a sum of normal distributions (Equation (26)) does not, in general, follow a tractable distribution. An alternative approach, however, is to calculate the maximum likelihood ancestor location for each sampled tree (Equation (24)) and importance sample over these estimates. Writing the maximum likelihood ancestor location given a sampled tree as $\widehat{\boldsymbol{\ell}}_a(\mathcal{T}_m)$, we then have another estimate of the ancestor's location,

$$\widehat{\boldsymbol{\ell}}_a^{\,*} = \frac{1}{M} \sum_{m=1}^{M} \frac{\widehat{\boldsymbol{\ell}}_a(\mathcal{T}_m)\mathbb{P}(\mathcal{T}_m|\mathcal{G}, \mathbf{\Sigma})}{\mathbb{P}(\mathcal{T}_m|\mathcal{G}, \text{panmixia})}. \tag{27}$$

Equation (27) is a 'Best Linear Unbiased Predictor' (BLUP) averaged over the importance weights. We can calculate this directly and it is therefore, in

principle, faster to compute than the search over Equation (26). In practice we find that the MLE and weighted-BLUP estimators are very correlated. Throughout the paper we use the MLEs but both are implemented in our software.

Simulations

Spatially-explicit forward-time simulations

We performed simulations in `SLiM` v3.6 (Haller and Messer, 2019) with tree sequence recording (Haller et al., 2019). The `SLiM` code was adapted from the `pyslim` spatial vignette (https://pyslim.readthedocs.io/en/latest/vignette_space.html) to model non-overlapping generations.

Individuals are diploid for a single chromosome with $L = 10^8$ basepairs and per basepair recombination rate $r = 5 \times 10^{-9}$. Individuals exist in a two dimensional (2D) habitat, a square of width $W = 50$ with reflecting boundaries. Each generation begins with reproduction. Each individual acts once as a 'mother' and chooses a 'father' at random (individuals are hermaphrodites), weighted by their mating weights. The mating weight of an individual $d < 3\sigma$ distance from a mother follows a 2D normal distribution centered on the mother with variance σ^2 in both directions and no covariance. Individuals further than 3σ distance apart are ignored for efficiency (creating an Allee effect – i.e., a mother is not guaranteed to find a mate). Local density-dependence is modelled through competitive effects on fecundity. The strength of competitive interaction between two individuals $d < 3\sigma_c$ distance apart follows a 2D normal distribution centered on one of the individuals with variance $\sigma_c^2 = 0.5^2$ in both directions and no covariance (we again ignore more distant individuals). The number of offspring produced by a mating is Poisson, with mean $R/(1 + C/K)$, where C is the sum of interaction strengths the mother experiences, $R = 2$ is the mean number of offspring in the absence of competition, and $K = 2$ is the local carrying capacity. Each offspring disperses from its mother by a random 2D normal deviate with variance σ^2 in each dimension and no covariance. After reproduction all parents die and all offspring become adults in the next generation. We begin the population with $N_0 = W^2K = 5 \times 10^3$ individuals distributed uniformly at random across space. We end the simulation, and output the tree sequence, after $4N_0 = 2 \times 10^4$ generations.

True tree sequence of the sample

Using `pyslim` v0.6, we load the tree sequence into `Python`, sample $k/2 = 50$ present-day individuals ($k = 100$ chromosomes) at random, and simplify the tree sequence to lineages ancestral to the sample. We next use `pyslim` to 'recapitate' under the standard coalescent with recombination (Hudson, 2002), with effective population size N_0, to ensure all sampled lineages have coalesced. We then use `msprime` v1.0.1 (Kelleher et al., 2016b) to layer on neutral mutations with per basepair per generation mutation rate $U = 1.25 \times 10^{-8}$. This tree sequence represents the true genealogical history of the sample.

Inferred tree sequence of the sample

Because we will not know the true tree sequence of any natural sample, we also write this true tree sequence out as a VCF and use `Relate` v1.1.4 (Speidel et al., 2019) to infer the tree sequence. We give `Relate` the true uniform recombination map and mutation rate and an effective population size of $\theta/(4U)$, where θ is the observed mean genetic diversity (calculated from the tree sequence with `tskit` v0.3.5; Kelleher et al., 2018; Ralph et al., 2020). We then feed the resulting output to `Relate`'s 'EstimatePopulationSize' function to iteratively estimate a piece-wise constant effective population size and branch lengths, using 5 iterations and keeping all of the trees. We then use `Relate` to convert the anc/mut format to a `tskit` tree sequence for downstream processing.

True locations of ancestors

To analyze our ability to locate ancestors we also ran some `SLiM` simulations with `initializeTreeSeq(retainCoalescentOnly=F)` and `treeSeqRememberIndividuals(permanent=F)` options, meaning we remember all individuals (including their location data) that are ancestors of the final population. When simplifying the tree sequence we use the `keep_unary=True` option, to keep all nodes (and their locations) that are ancestral to the sample (rather than just the coalescent nodes).

Processing the tree sequence

Each simplified tree sequence contains on the order of 10^4 trees. Here we use only 10^2 of these trees, uniformly sampled from the tree sequence, so that each is relatively independent of the others. For each of these trees we use `Relate`'s 'SampleBranchLengths' function to sample branch lengths from the posterior $M = 10$ times. We load resulting newick trees with `DendroPy` v4.5.2 (Sukumaran and Holder, 2010) and process each sampled tree. In processing, we first chop the trees off at T generations to create subtrees and for each subtree record the coalescence times, the probability of these coalescence times under the neutral coalescent (given the estimated piecewise constant population size from `Relate`), and the shared evolutionary times between each sample. This pre-processing speeds up the numerical search for maximum likelihood parameters.

Dispersal estimates

We approximate the maximum likelihood estimate (MLE) of dispersal rate at each tree by numerically searching for the maximum of Equation (9). We use the L-BFGS-B method of `SciPy` v1.6.2 (Virtanen et al., 2020) to numerically find the MLEs, which allows us to put bounds on the parameters. We search for the MLEs in terms of the standard deviation of dispersal in x (latitude) and y (longitude) (with a lower bound of 10^{-6} to prevent non-positive estimates) and the correlation between these two axes (with a lower bound of -0.99 and upper bound of 0.99 to prevent estimates with absolute value greater than 1). In the process we also find the MLE of the birth rate in the branching process (with a lower bound of 10^{-6} to prevent non-positive estimates). As the likelihood was much more sensitive to a given change in birth rate than the other parameters, we instead search for 10^2 times the MLE birth rate and then rescale back to the original parameters, which makes the search for the MLE more efficient. We find the maximum composite likelihood estimate of dispersal and branching rate by searching for the parameters that maximize the product of the likelihoods over mulitple loci.

Locating genetic ancestors

To locate a genetic ancestor at a particular locus and time we numerically search for the location that maximizes Equation (26), following the same

approach as above with bounds of $(0, W)$ in both dimensions.

Arabidopsis thaliana data

Inferring the tree sequence

We first downloaded the VCF of SNPs for 1135 *Arabidopsis thaliana* individuals from https://1001genomes.org/data/GMI-MPI/releases/v3.1/ (Alonso-Blanco et al., 2016)) and used `PLINK 2.0` (www.cog-genomics.org/plink/2.0/; Chang et al., 2015) to find SNPs with minor allele frequency < 0.05. We then converted the matrix of imputed SNPs (https://1001genomes.org/data/GMI-MPI/releases/v3.1/SNP_matrix_imputed_hdf5) into a haploid 'haps' file (as required by `Relate`) for each chromosome, while filtering out those SNPs with minor allele frequency < 0.05. A multi-species alignment for *Arabidopsis thaliana*, *Boechera stricta*, *Arabidopsis lyrata*, and *Malcomia maritima* was kindly provided by Tyler Kent (University of Toronto). We used `bx-python` v0.8.9 to create a FASTA file containing the three outgroups from this alignment and, combined with the haps file described above, created the input file required by `est-sfs` (Keightley and Jackson, 2018), again for each chromosome separately. We then ran `est-sfs`, which incorporates both the within-population polymorphism data and outgroup sequences, on each chromosome separately using the Kimura 2-parameter model (Kimura, 1980) to get the probability that each reference allele is ancestral. We then used `bx-python` to extract the reference sequence for *Arabidopsis thaliana* from the alignment and created the ancestral chromosomes by making the alternate allele ancestral whenever the probability the reference allele was ancestral was < 0.5. We then used this ancestral genome to create polarized haps files for each chromosome with `Relate`'s 'FlipHapsUsingAncestor' function. The recombination map for each chromosome was downloaded from https://www.eeb.ucla.edu/Faculty/Lohmueller/data/uploads/ (Salomé et al., 2012). The polarized haps files and recombination maps were then used to infer the tree sequence with `Relate`, using per base pair per generation mutation rate $U = 7 \times 10^{-9}$ (Adrion et al., 2020) and haploid effective population size $2N_e = 1.7 \times 10^5$ (estimated from nucleotide diversity, $\pi = 4N_eU$, calculated with `tskit diversity()` statistic; Ralph et al., 2020). We then fed this output to `Relate`'s 'EstimatePopulationSize' function (with some customizing of the script to allow it to work with haploids and spit out the anc/mut files) to iteratively estimate a piece-wise

constant effective population size and branch lengths, assuming $U = 7 \times 10^{-9}$ and a single panmictic population, using 5 iterations and dropping half of the trees (the 50% with fewest mutations). We then converted the output to a `tskit` tree sequence using `Relate`'s 'ConvertToTreeSequence' function.

Nearly-identical samples

The 1001 Genomes dataset contains a number of nearly-identical samples as the result of selfing. A number of nearly identical pairs are separated by $>$ 1km and thus may represent long-distance migration or mis-assignment or mix-ups (Alonso-Blanco et al., 2016). We reduce the effect of these near identical samples by first calculating all pairwise genetic distances using `PLINK 1.9` (with the `--distance allele-ct` option; see Figure 3A in Alonso-Blanco et al., 2016). We then ignore all samples that differ at less than 10^3 base pairs from any other sample.

Removing dispersal outliers and samples without locations

We downloaded the metadata associated with the 1001 Genomes samples from https://raw.githubusercontent.com/hagax8/arabidopsis_viz/master/data/dataframe_1001G.csv (see also https://1001genomes.org/accessions.html), which includes location data and previous admixture assignments (Alonso-Blanco et al., 2016) (see https://1001genomes.github.io/admixture-map/ for ancestry proportions). We remove the outlier samples in North America ($n = 125$) and Japan ($n = 2$), as well as those without locations ($n = 4$), from the dispersal rate and genetic ancestor location likelihoods (they remain in the importance sample weights).

Dispersal estimates

We estimate the MLE of dispersal rate at each chosen tree as described above for simulations. To reduce computation time we search for maximum composite likelihoods across trees for each chromosome separately, rather than genome-wide.

We convert the estimates of dispersal rate (standard deviations) from degrees to kilometres by multiplying the latitudinal estimates by $\cos(x\pi/180)/111$, where x is the mean latitude of all (filtered) samples, and multiplying the longitudinal estimates by 110.

Locating genetic ancestors

We locate ancestors as described above for simulations, using bounds of $(-90, 90)$ for latitude and $(-180, 180)$ for longitude.

Data availability statement

All code used to perform the analyses in this study can be found at https://github.com/mmosmond/sparg-ms. A `Python` package of our method is available at https://github.com/mmosmond/sparg.

Acknowledgements

We thank Tyler Kent for the multispecies genome alignments, Yan Wong, Ben Haller, and Peter Ralph for the `retainCoalescentOnly` and `permanent` updates to `SLiM`/`pyslim`, Aaron Stern for making his importance sampling code freely available, Leo Speidel for helpful discussion regarding `Relate`, Vince Buffalo for the introduction to `Snakemake`, and Nick Barton for the healthy skepticism about per-locus estimates of ancestor locations at deep times. Funding provided by Banting (Canada) and Center for Population Biology (UC Davis) fellowships (awarded to MMO) and the National Institute of General Medical Sciences of the National Institutes of Health (NIH R01 GM108779 and R35 GM136290, awarded to GC). Computations were performed on the Niagara supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

References

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Ortega-Del Vecchyo, D., Racimo, F., Struck, T. J., Gravel, S., Gutenkunst, R. N., Lohmueller, K. E., Ralph, P. L., Schrider, D. R., Siepel, A., Kelleher, J., and Kern,

A. D. (2020). A community-maintained standard library of population genetic models. *eLife*, 9:e54967.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., et al. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, 166(2):481–491.

Avise, J. C. (2009). Phylogeography: retrospect and prospect. *Journal of Biogeography*, 36(1):3–15.

Barton, N. H. and Wilson, I. (1995). Genealogies and geography. *Philosophical Transactions of the Royal Society of London B*, 349(1327):49–59.

Battey, C. J., Ralph, P. L., and Kern, A. D. (2020). Predicting geographic location from genetic variation with deep neural networks. *eLife*, 9:e54507.

Bedford, T., Cobey, S., Beerli, P., and Pascual, M. (2010). Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS Pathogens*, 6(5):e1000918.

Biek, R., Henderson, J. C., Waller, L. A., Rupprecht, C. E., and Real, L. A. (2007). A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proceedings of the National Academy of Sciences*, 104(19):7993–7998.

Bomblies, K., Yant, L., Laitinen, R. A., Kim, S.-T., Hollister, J. D., Warthmann, N., Fitz, J., and Weigel, D. (2010). Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genetics*, 6(3):e1000890.

Bradburd, G. S. and Ralph, P. L. (2019). Spatial population genetics: it's about time. *Annual Review of Ecology, Evolution, and Systematics*, 50:427–449.

Cavalli-Sforza, L. L., Barrai, I., and Edwards, A. W. (1964). Analysis of human evolution under random genetic drift. In *Cold Spring Harbor symposia on quantitative biology*, volume 29, pages 9–20. Cold Spring Harbor Laboratory Press.

Cavalli-Sforza, L. L. and Edwards, A. W. (1967). Phylogenetic analysis. models and estimation procedures. *American Journal of Human Genetics*, 19:233–257.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(7):s13742–015–0047–8.

Coop, G. (2013). How many genetic ancestors do I have? https://gcbias.org/2013/11/11/how-does-your-number-of-genetic-ancestors-grow-back-over-time/. [Online; accessed 1-July-2021].

Coop, G. (2017). Where did your genetic ancestors come from? https://gcbias.org/2017/12/19/1628/. [Online; accessed 1-July-2021].

Dellicour, S., Durkin, K., Hong, S. L., Vanmechelen, B., Martí-Carreras, J., Gill, M. S., Meex, C., Bontems, S., André, E., Gilbert, M., et al. (2021a). A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. *Molecular Biology and Evolution*, 38(4):1608–1613.

Dellicour, S., Hong, S. L., Vrancken, B., Chaillon, A., Gill, M. S., Maurano, M. T., Ramaswami, S., Zappile, P., Marier, C., Harkins, G. W., et al. (2021b). Dispersal dynamics of SARS-CoV-2 lineages during the first epidemic wave in New York City. *PLoS Pathogens*, 17(5):e1009571.

Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, 23(1):34–63.

Durvasula, A., Fulgione, A., Gutaker, R. M., Alacakaptan, S. I., Flood, P. J., Neto, C., Tsuchimatsu, T., Burbano, H. A., Picó, F. X., Alonso-Blanco, C., et al. (2017). African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, 114(20):5213–5218.

Edwards, A. W. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society B*, 32(2):155–164.

Exposito-Alonso, M., Becker, C., Schuenemann, V. J., Reiter, E., Setzer, C., Slovak, R., Brachi, B., Hagmann, J., Grimm, D. G., Chen, J., et al. (2018). The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genetics*, 14(2):e1007155.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25(5):471.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15.

Fulgione, A. and Hancock, A. M. (2018). Archaic lineages broaden our view on the history of *Arabidopsis thaliana*. *New Phytologist*, 219(4):1194–1198.

Fulgione, A., Koornneef, M., Roux, F., Hermisson, J., and Hancock, A. M. (2018). Madeiran *Arabidopsis thaliana* reveals ancient long-range colonization and clarifies demography in Eurasia. *Molecular Biology and Evolution*, 35(3):564–574.

Grafen, A. (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London B*, 326(1233):119–157.

Griffiths, R. C. and Tavare, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London B*, 344(1310):403–410.

Griffiths, R. C. and Tavare, S. (1997). Computational methods for the coalescent. *IMA Volumes in Mathematics and its Applications*, 87:165–182.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., and Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, 19(2):552–566.

Haller, B. C. and Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution*, 36(3):632–637.

Harmon, L. J. (2019). *Phylogenetic comparative methods*. https://lukejharmon.github.io/pcm/.

Harmon, L. J., Losos, J. B., Jonathan Davies, T., Gillespie, R. G., Gittleman, J. L., Bryan Jennings, W., Kozak, K. H., McPeek, M. A., Moreno-Roark, F., Near, T. J., et al. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, 64(8):2385–2396.

Hewitt, G. (2000). The genetic legacy of the quaternary ice ages. *Nature*, 405(6789):907–913.

Hsu, C.-W., Lo, C.-Y., and Lee, C.-R. (2019). On the postglacial spread of human commensal *Arabidopsis thaliana*: journey to the east. *New Phytologist*, 222(3):1447–1457.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–1817.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.

Keightley, P. D. and Jackson, B. C. (2018). Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics*, 209(3):897–906.

Kelleher, J., Etheridge, A., Vber, A., and Barton, N. (2016a). Spread of pedigree versus genetic ancestry in spatially distributed populations. *Theoretical Population Biology*, 108:1–12.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016b). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5):e1004842.

Kelleher, J., Thornton, K. R., Ashander, J., and Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLoS Computational Biology*, 14(11):e1006581.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338.

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., and Nordborg, M. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, 39(9):1151–1155.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.

Knowles, L. L. (2009). Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, 40:593–612.

Lachance, J. (2009). Inbreeding, pedigree size, and the most recent common ancestor of humanity. *Journal of Theoretical Biology*, 261(2):238–247.

Landis, M. J., Schraiber, J. G., and Liang, M. (2013). Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. *Systematic Biology*, 62(2):193–204.

Larribe, F. and Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica*, 21(1):43–69.

Lee, C.-R., Svardal, H., Farlow, A., Exposito-Alonso, M., Ding, W., Novikova, P., Alonso-Blanco, C., Weigel, D., and Nordborg, M. (2017). On the post-glacial spread of human commensal *Arabidopsis thaliana*. *Nature Communications*, 8(1):1–12.

Lemey, P., Hong, S. L., Hill, V., Baele, G., Poletto, C., Colizza, V., OToole, Á., McCrone, J. T., Andersen, K. G., Worobey, M., et al. (2020). Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nature Communications*, 11(1):1–14.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5(9):e1000520.

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, 27(8):1877–1885.

Malécot, G. (1948). *Les mathématiques de l'hérédité*. Masson.

Martin, M. A., VanInsberghe, D., and Koelle, K. (2021). Insights from SARS-CoV-2 sequences. *Science*, 371(6528):466–467.

Meligkotsidou, L. and Fearnhead, P. (2007). Postprocessing of genealogical trees. *Genetics*, 177(1):347–358.

Nordborg, M. (2000). Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics*, 154(2):923–929.

Novembre, J. and Slatkin, M. (2009). Likelihood-based inference in isolation-by-distance models using the spatial distribution of low-frequency alleles. *Evolution*, 63(11):2914–2925.

O'Meara, B. C., Ané, C., Sanderson, M. J., and Wainwright, P. C. (2006). Testing for different rates of continuous trait evolution using likelihood. *Evolution*, 60(5):922–933.

Platt, A., Horton, M., Huang, Y. S., Li, Y., Anastasio, A. E., Mulyati, N. W., Ågren, J., Bossdorf, O., Byers, D., Donohue, K., et al. (2010). The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics*, 6(2):e1000843.

Ralph, P. and Coop, G. (2013). The geography of recent genetic ancestry across europe. *PLoS Biology*, 11(5):e1001555.

Ralph, P., Thornton, K., and Kelleher, J. (2020). Efficiently summarizing relationships in large samples: a general duality between statistics of genealogies and genomes. *Genetics*, 215(3):779–797.

Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, 43(3):304–311.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342.

Reich, D. (2018). *Who we are and how we got here: ancient DNA and the new science of the human past*. Oxford University Press.

Ringbauer, H., Coop, G., and Barton, N. H. (2017). Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–1351.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, 145(4):1219–1228.

Salomé, P., Bomblies, K., Fitz, J., Laitinen, R., Warthmann, N., Yant, L., and Weigel, D. (2012). The recombination landscape in *Arabidopsis thaliana* F2 populations. *Heredity*, 108(4):447–455.

Schluter, D., Price, T., Mooers, A. Ø., and Ludwig, D. (1997). Likelihood of ancestor states in adaptive radiation. *Evolution*, 51(6):1699–1711.

Shirsekar, G., Devos, J., Latorre, S. M., Blaha, A., Dias, M. Q., Hernando, A. G., Lundberg, D. S., Burbano, H. A., Fenster, C. B., and Weigel, D. (2021). Fine-scale population structure of north american *Arabidopsis thaliana* reveals multiple sources of introduction from across Eurasia. *bioRxiv*, doi:10.1101/2021.01.22.427575.

Slatkin, M. and Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences*, 113(23):6380–6387.

Speidel, L., Cassidy, L., Davies, R. W., Hellenthal, G., Skoglund, P., and Myers, S. (2021). Inferring population histories for ancient genomes using genome-wide genealogies. *bioRxiv*, doi:10.1101/2021.02.17.431573.

Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society B*, 62(4):605–635.

Stern, A. J., Speidel, L., Zaitlen, N. A., and Nielsen, R. (2021). Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *The American Journal of Human Genetics*, 108(2):219–239.

Stern, A. J., Wilton, P. R., and Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genetics*, 15(9):e1008384.

Sukumaran, J. and Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272.

Volz, E. M., Koelle, K., and Bedford, T. (2013). Viral phylodynamics. *PLoS Computational Biology*, 9(3):e1002947.

Wakeley, J., King, L., Low, B. S., and Ramachandran, S. (2012). Gene genealogies within a fixed pedigree, and the robustness of Kingmans coalescent. *Genetics*, 190(4):1433–1445.

Wang, I. J. and Bradburd, G. S. (2014). Isolation by environment. *Molecular Ecology*, 23(23):5649–5662.

Wilkins, J. F. (2004). A separation-of-timescales approach to the coalescent in a continuous population. *Genetics*, 168(4):2227–2244.

Wilkins, J. F. and Wakeley, J. (2002). The coalescent in a continuous, finite, linear population. *Genetics*, 161(2):873–888.

Wiuf, C. (2006). Consistency of estimators of population scaled parameters using composite likelihood. *Journal of Mathematical Biology*, 53(5):821–841.

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., and McVean, G. (2021). A unified genealogy of modern and ancient genomes. *bioRxiv*, doi:10.1101/2021.02.16.431497.

Worobey, M., Pekar, J., Larsen, B. B., Nelson, M. I., Hill, V., Joy, J. B., Rambaut, A., Suchard, M. A., Wertheim, J. O., and Lemey, P. (2020).

The emergence of SARS-CoV-2 in Europe and North America. *Science*, 370(6516):564–570.

Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114–138.

Zeng, L., Gu, Z., Xu, M., Zhao, N., Zhu, W., Yonezawa, T., Liu, T., Qiong, L., Tersing, T., Xu, L., et al. (2017). Discovery of a high-altitude ecotype and ancient lineage of *Arabidopsis thaliana* from tibet. *Science Bulletin*, 62(24):1628–1630.

Zou, Y.-P., Hou, X.-H., Wu, Q., Chen, J.-F., Li, Z.-W., Han, T.-S., Niu, X.-M., Yang, L., Xu, Y.-C., Zhang, J., et al. (2017). Adaptation of *Arabidopsis thaliana* to the Yangtze River basin. *Genome Biology*, 18(1):1–11.

Supplementary Text

Multi-epoch dispersal rate estimates

In both cases the two-epoch maximum negative log-likelihood was much larger than the maximum negative log-likelihood from the one-epoch model (mean differences of > 500 in B and > 2000 in C with true trees and > 700 and > 400 with inferred trees), providing strong statistical support for the two-epoch model. However, when we fit two-epoch models to simulations with a constant dispersal rate (Figure S3A,D), the two-epoch models still tended to have larger negative log-likelihoods than the one-epoch models (mean differences of ≈ 500 for $\sigma_x = \sigma_y = 0.25$ and ≈ 50 for $\sigma_x = \sigma_y = 0.5$ with true trees and ≈ 200 and ≈ 250 with inferred trees). Using the true trees, at higher dispersal rates ($\sigma_x = \sigma_y = 0.5$, Figure S3D) the estimates of dispersal in each of the two epochs were similar to one another, and to the one-epoch estimate. This likely accounts for the relatively small increase in likelihood under the two-epoch model. At lower dispersal rates ($\sigma_x = \sigma_y = 0.25$, Figure S3A) we see a different pattern, where the true (and inferred) trees now tend to overestimate dispersal in the more distant epoch. We suspect this to be a sampling bias; in cutting trees off at 10^3 generations we effectively give more weight to lineages that coalesce quickly, creating an upwards bias in dispersal rates. Supporting this hypothesis, using a deeper cutoff of 10^4 generations produced more accurate estimates at these lower dispersal rates (results not shown). Curiously, the inferred trees tell a different two-epoch story than the true trees under high dispersal (Figure S3D), but not under low dispersal (Figure S3A). This suggests tree inference is worse, or that errors have more impact on dispersal estimates, when the samples are more closely related.

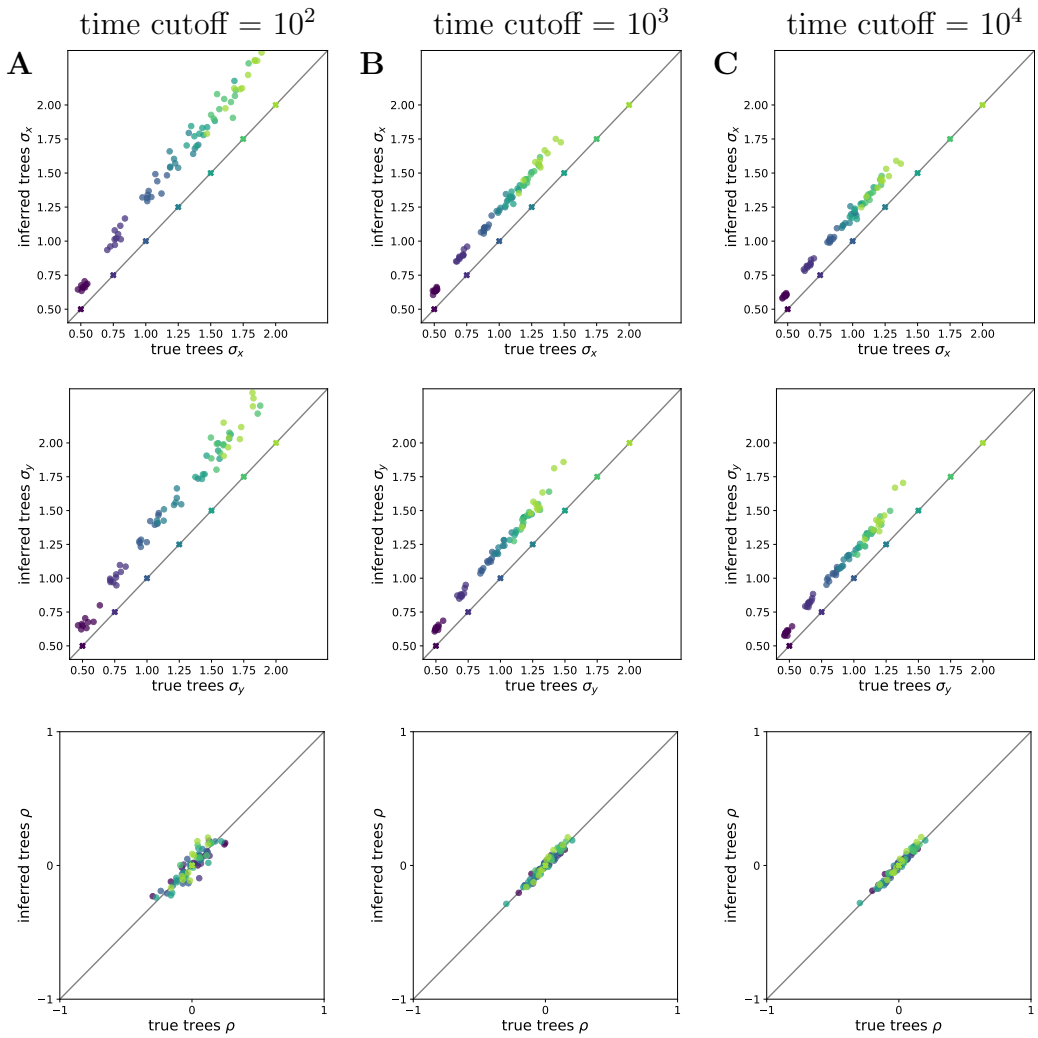


Figure S1: **The effect of the time cutoff on dispersal estimates in one-epoch simulations.** Panel **A** is as in the main text (Figure 2A), where we ignore everything beyond 10^2 generations ago (and here we show dispersal along the y-axis in the correlation in x and y). Panels **B** and **C** increase the time cutoff by one and two orders of magnitude, respectively. Dispersal estimates shrink as the time cutoff grows because smaller dispersal rates make it more likely all lineages have remained within the finite habitat (with reflecting boundaries). The variance in estimates seems to decline with the time cutoff, as larger cutoffs mean we use more information from the trees.

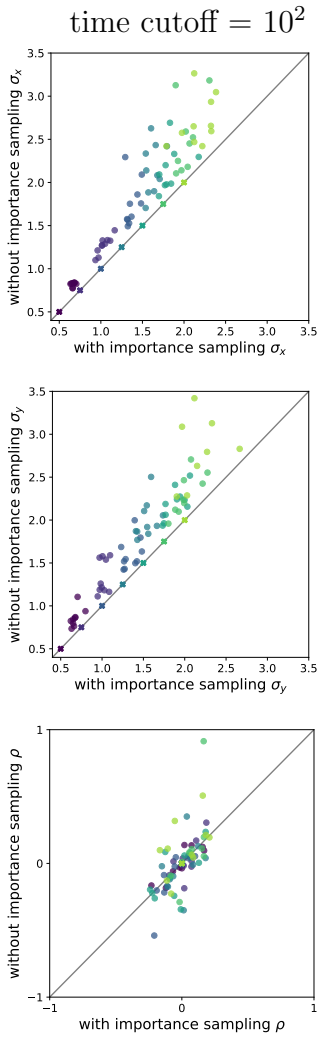


Figure S2: **The effect of importance sampling on dispersal estimates in one-epoch simulations.** In all cases importance sampling leads to smaller overestimates of the simulated dispersal rate (σ_x and σ_y) and generally exhibits lower variance in estimates across replicates (especially at high simulated dispersal rates). The 'without importance sampling' estimate comes from using only the first sample of branch lengths, while the 'with importance sampling' estimate uses all 10 samples.

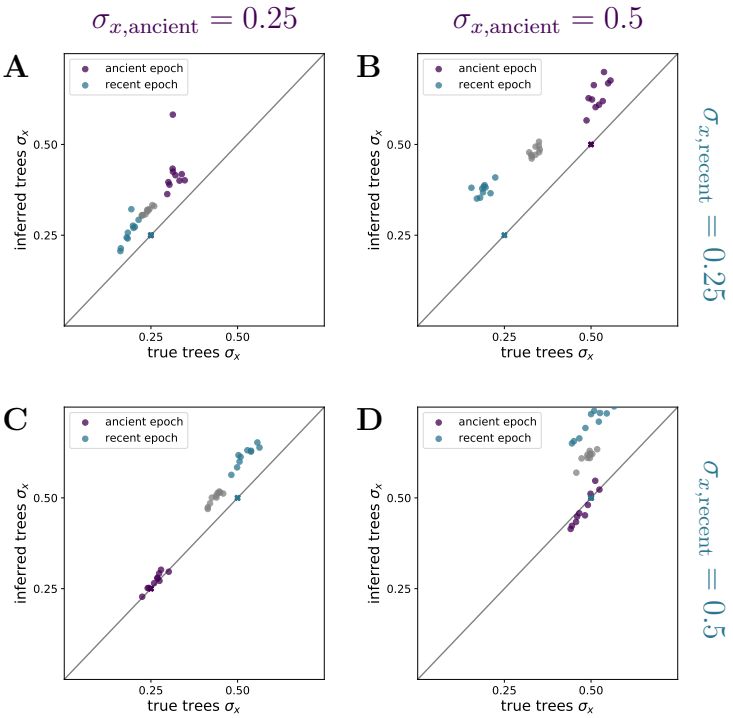


Figure S3: **Estimating two-epoch dispersal rates in simulations.** Panels **B** and **D** are panels **B** and **C**, respectively, in Figure 2. Panels **A** and **D** are simulated with a constant dispersal rate, $\sigma = 0.25$ and 0.5, respectively. The coloured dots show the estimates for the recent and ancient epochs. The grey dots show the estimates under a one-epoch model.

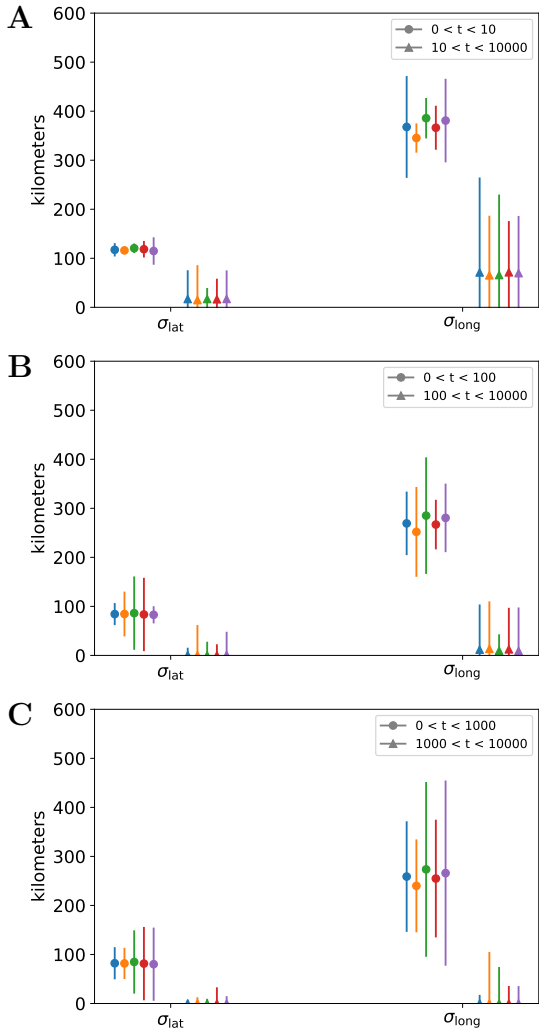


Figure S4: **Dispersal rate estimates in *Arabidopsis thaliana* under a two-epoch model.** Like Figure 3B, but here we also show the (less likely) split times of (**B**) 10^2 and (**C**) 10^3.

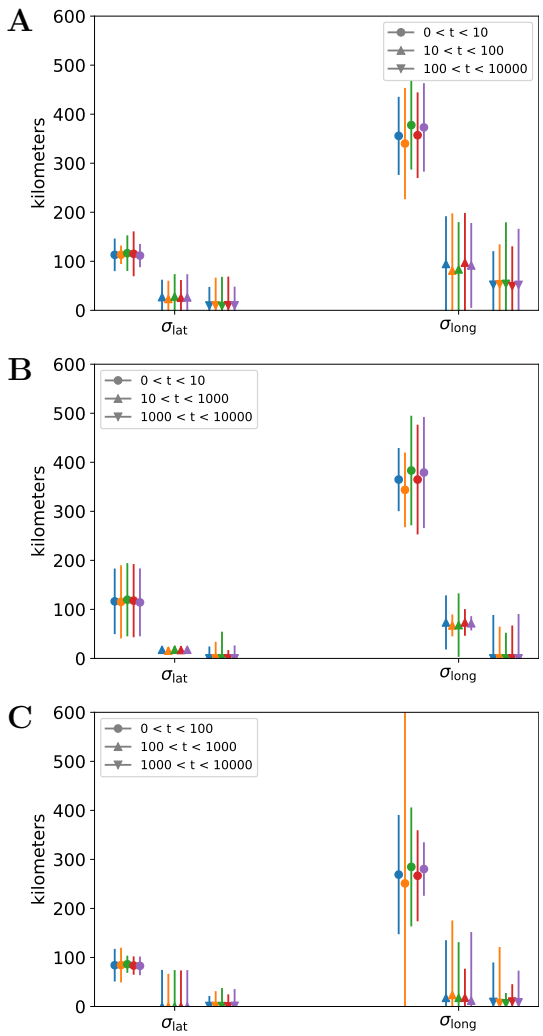
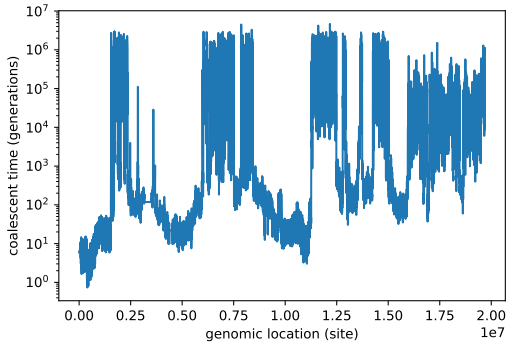
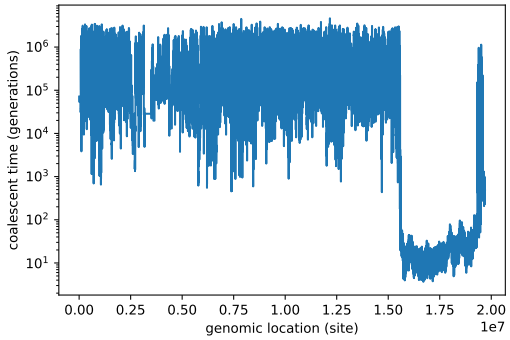


Figure S5: **Dispersal rate estimates in *Arabidopsis thaliana* under a three-epoch model.** Like Figure 3B, but here we also show the (less likely) split times of (**A**) [10,10^2] and (**C**) [10^2,10^3].

A) accessions 9737 & 9627 (chromosome 2)



B) accessions 9933 & 10015 (chromosome 2)



C) accessions 7314 & 6981 (chromosome 5)

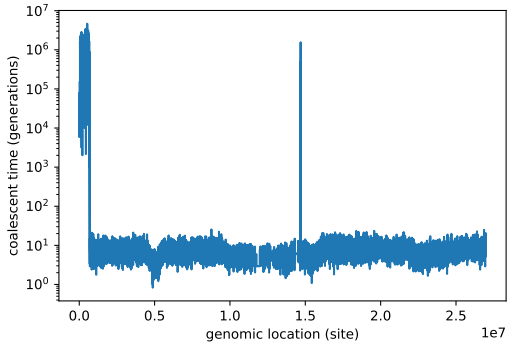


Figure S6: **Coalescence times between particular pairs of samples along a particular chromosome**. Calculated using `tskit`'s `divergence()` function.

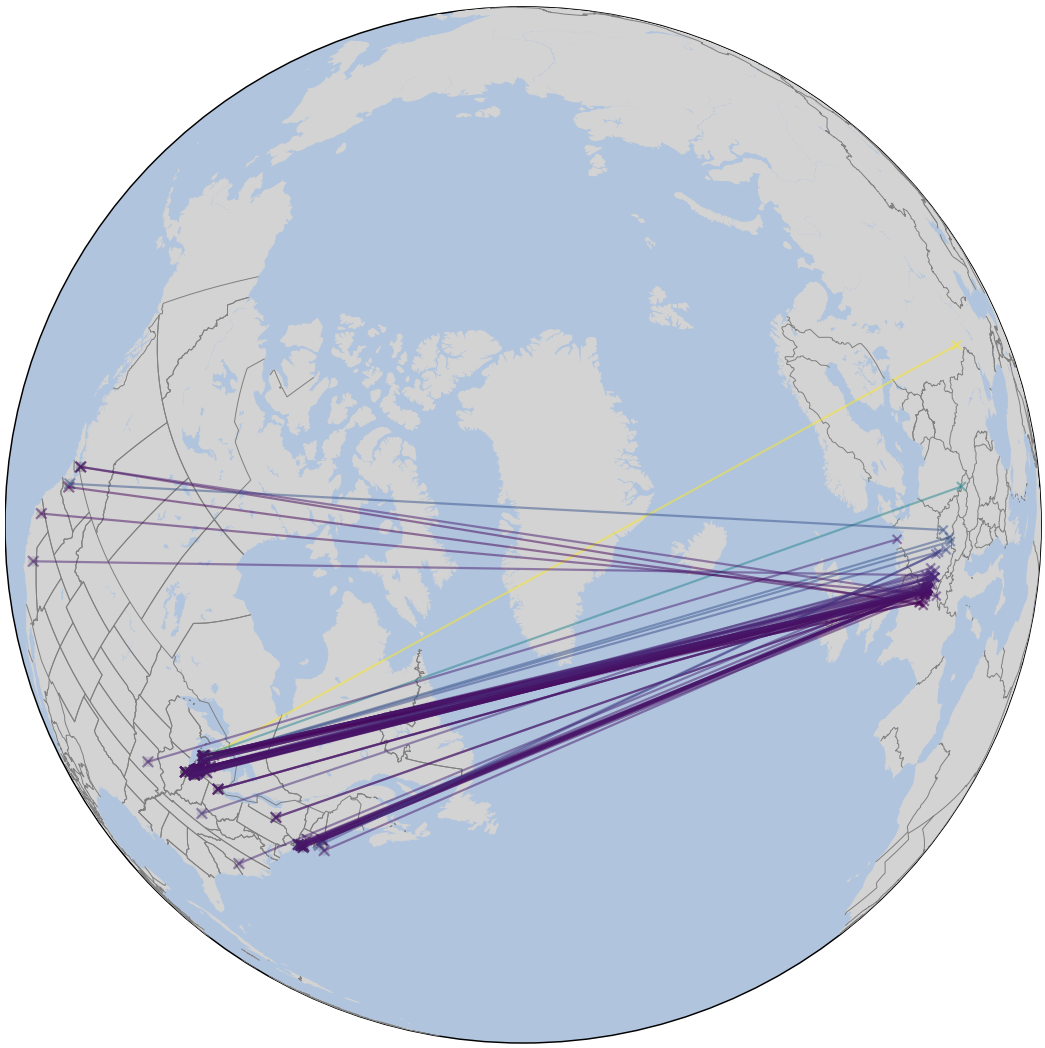


Figure S7: **Connecting the sample locations with the inferred present-day locations of the North American samples.** Color indicates longitude of the inferred location. See Figure 6 for more details.

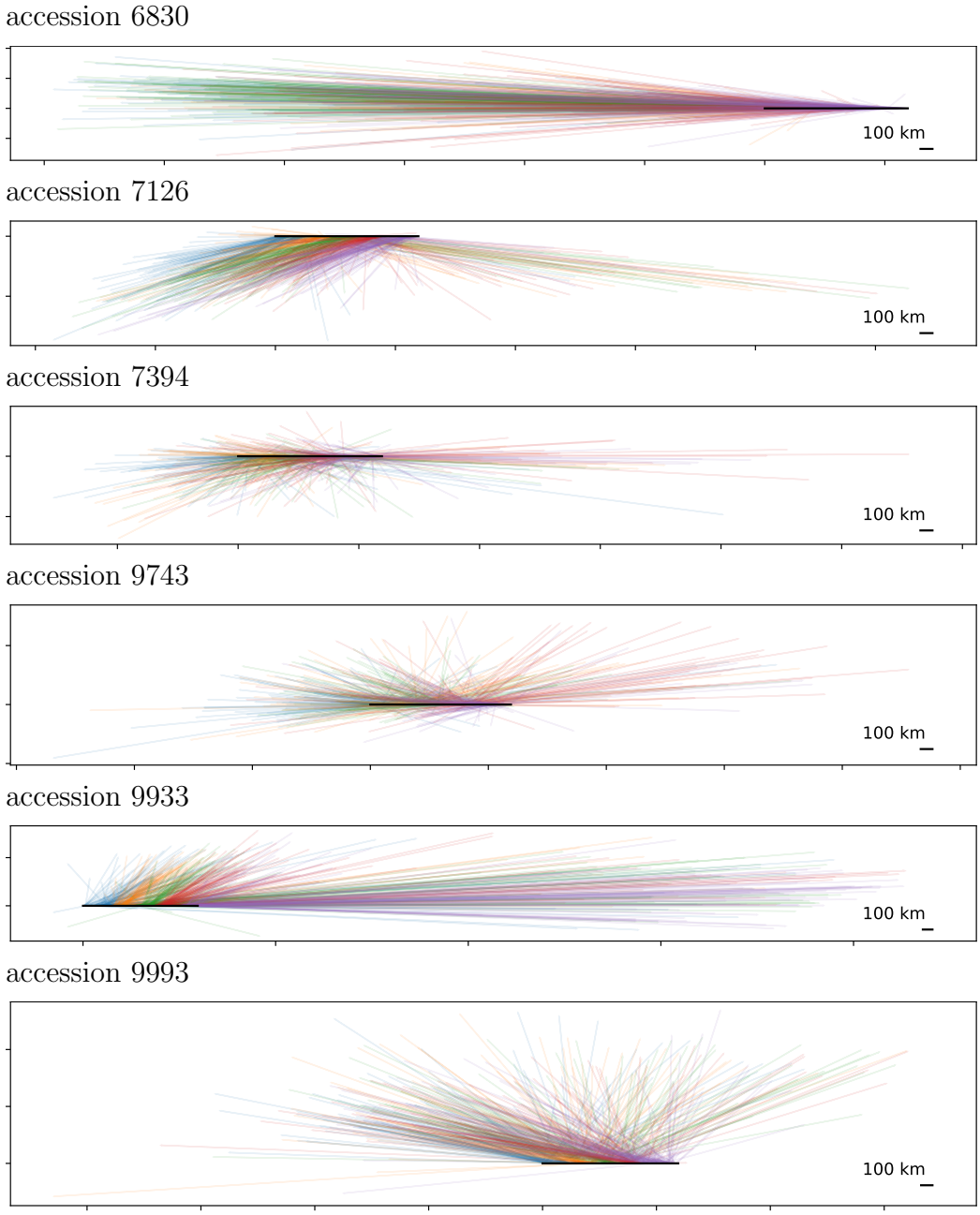


Figure S8: **Genome-view of ancestral displacements for all 6 samples in Figure 9.** See Figure 9 for more details.