Exploring the effects of ecological parameters on the spatial structure of genetic tree sequences

Mariadaria K. Ianni-Ravn[1,2], Martin Petr[1,3], and Fernando Racimo[1,3]

[1]Section for Molecular Ecology and Evolution, Globe Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.
[2]University of Chicago, Chicago, 60637, IL, USA
[3]Lundbeck Foundation GeoGenetics Centre, Globe Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

March 27, 2023

1 Abstract

Space is a fundamental dimension of evolutionary change, determining how individuals disperse and interact with each other. Consequently, space has an important influence on the structure of genealogies and the distribution of genetic variants over time. Recently, the development of highly flexible simulation tools and computational methods for genealogical inference has greatly increased the potential for incorporating space into models of population genetic variation. It is now possible to explore how spatial ecological parameters can influence the distribution of genetic variation among individuals with unprecedented detail. In this study, we explore the effects of three specific parameters (the dispersal distance, competition distance and mate choice distance) on the spatial structure of genealogies relating individuals across a genome. We carry out a series of *in silico* experiments using forwards-in-time simulations, to determine how these parameters influence the distance between closely- and distantly-related individuals in these genealogies. We also assess the accuracy of the maximum likelihood estimation of the dispersal distance in a Gaussian model of dispersal from tree-sequence data. Our results showcase the potential for linking phylogeography, population genetics and ecology, in order to answer fundamental questions about the nature of spatial interactions across a landscape.

2 Introduction

From nutrient-fixing bacteria in the digestive system, to pollen carried on the legs of bees, all living organisms must deal with the particularities of the range that they inhabit. At each generation, individuals tend to disperse from their parents, often carrying their genes across great geographic distances. Space is also a major determinant of mate choice and competition patterns, which can, in turn, influence how genetic relatedness decays as

a function of the distance between individuals (Wright 1943). The connection between genetic differentiation and space has indeed been the focus of numerous theoretical models (B. Charlesworth, D. Charlesworth, and Barton 2003; Hardy and Vekemans 1999) and empirical studies (Sexton, Hangartner, and Hoffmann 2014; Jenkins et al. 2010; Aguillon et al. 2017). Overall, we now know that genetic data can hold important information about the geographic distribution of individuals in the past (Aguillon et al. 2017).

Biologists often seek to understand the rate at which individuals move across space. One way to approach this problem is by focusing on the "dispersal distribution": a probability distribution over the parent-offspring distance (Kot, Lewis, and Driessche 1996) i.e. how far away a particular offspring mates compared to its birthplace. The shape of the dispersal distribution for different species has been of great interest in ecology, since long-distance dispersal is predicted to strongly affect ecological processes, such as allele surfing, as well as the spread of invasive species and host-parasitoid interactions (Paulose and Hallatschek 2020; McCann et al. 2000; Clark 1998).

The dispersal distribution is often summarized via a "dispersal distance" parameter, σ, which governs how far away an offspring tends to be from its parents. Over multiple generations - for example, over branches in a phylogeny - this determines the speed at which two lineages move away from one another after descending from a common ancestor (Rousset 2001). Populations with large σ are expected to mix rapidly across their habitat compared to those with a smaller σ.

One way to estimate the parameters of the dispersal distribution in a real population, is to track the exact locations of all individuals in a pedigree. However, this is often difficult or expensive (Cayuela et al. 2018). While non-recombining genetic sequences can be easily represented as a phylogeny (Miles et al. 2009; Markov et al. 2009; Castillo et al. 2011), the full geographic history of recombining genomes cannot be represented in a single tree. The full history of coalescent events and recombination events can be represented as a network, known as the Ancestral Recombination Graph (ARG) (Hudson et al. 1990; Griffiths and Marjoram 1996; Griffiths and Marjoram 1997). This network fully encodes how a set of sampled genomes in a population are related to each other. Though notoriously difficult to infer from sequence data, recent computational developments have made it tractable to approximately infer the ARG for a given genome panel (Kelleher, Wong, et al. 2019; Speidel et al. 2019; Hubisz and Siepel 2020).

With some loss of information, an ARG can be further simplified into a sequence of trees along the genome, which are more liable for analysis using phylogeographic theory. Recently, Osmond and Coop developed a method to provide a probabilitistic estimate of the location of ancestors in a sequence of trees, using branching Brownian motion (Osmond and Coop 2021). In another recent example, a recursive method was used to simultaneously infer the locations of all ancestors on a tree sequence inferred from a large-scale present-day dataset of worldwide human genomes (Wohns et al. 2021).

The dispersal distribution implicitly includes several steps, such as mate choice or juvenile-to-adult movement. Such stages are often ignored, and little work has been done to assess how spatial parameters, such as the strength of competition between individuals and the scale of mate choice, can affect a tree sequence. This is particularly an issue when using forwards-in-time simulators, where users are required to specify these dynamics explicitly. Furthermore, little is known about how these spatial parameters might affect widely-used statistics, such as F_{ST}.

In recent years, new software for generating spatially explicit, forwards-in-time simula-

tions have enabled researchers to explore genetic variation under a wide range of population histories. The recently developed software *slendr* (Petr et al. 2022), which uses the powerful software *SLiM* as one of its simulation engines (Haller and Messer 2019; Haller, Galloway, et al. 2019) provides a particularly approachable way to model, visualize and simulate mate choice, dispersal and spatial interactions in continuous space. These simulators can bridge the gap between a lack of theoretical results and the desire to build realistic spatial models of species.

Here, we set out to understand properties of geographically annotated sequences of trees along a genome, using a simulation-based approach. We leverage *slendr* to study how ecological parameters affect the spatial distribution of individuals, and the structure of genealogies relating them over time.

First, we explore the effects of varying the mode and scale of mating and dispersal on the realised distances between parents and their offspring. We show that, in some cases, these distances closely match their theoretical distribution. Then, we illustrate a case in which the realised distribution closely matches a theoretical model which explicitly includes the radius of mate choice. Finally, we test the efficacy of a maximum likelihood estimator of the mean distance between parent and offspring, using distances recorded in the branches of a phylogeny under a commonly used Gaussian mode of dispersal.

Our work serves to show that a sound understanding of the geographic parameters of a species, with respect to the dispersal distribution and to ecological factors (such as competition for resources and mate choice), is key to carrying out reliable phylogeographic inference in real populations.

3 Results

3.1 Dispersal patterns in spatially-tagged trees

We were interested in determining the extent to which the observed parent-offspring distances in a (perfectly inferred) tree sequence provided information about the underlying dispersal function in a population. In our simulations, a dispersal function (DF) and its scale/variance parameter, σ, determine how the simulator decides where to place offspring compared to the gestating parent (p1). The range of DFs and their parametrization are summarised in Table 1 and plotted in Fig. 2. Although the meaning of σ with respect to the DF varied for each distribution, our parametrization was such that increasing σ increased the variance of parent-offspring displacement (Fig. 3). In essence, the larger σ, the further an offspring tends to be from its parents and the faster the population spreads across the habitat. More details on this process are described in the Methods section 5.1.1, and a schematic of these mechanics is shown in Fig. 1.

Two other important parameters in our simulations are the competition and mating distances. The competition distance serves to parametrize competition for resources within a neighbourhood. Essentially, the simulator counts the number of neighbours an individual has within a radius of the competition distance and down-scales their fitness proportionally to this number. The mate choice radius, or mating distance, determines the maximum radius within which a parent can choose a mate. In *slendr* V0.5.1, mates are chosen uniformly at random from within this distance.

We simulated a population of 50 individuals under a range of dispersal functions and σ values, and also varied the mating and competition distances. After 50 generations, we

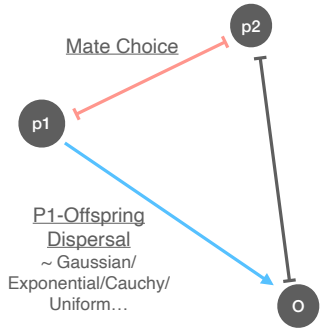

Figure 1: **The mechanics of dispersal in our simulations.** In our forwards-in-time simulations, two parents p1 and p2 are chosen. The distance between them (red line) must be less than the user-specified mating distance. The offspring ('o') is then dispersed from p1 (blue line) according to a specified mode of dispersal parametrized via a dispersal function (DF) and distance (which we call σ). These mechanics imply that a given one-generation dispersal may either be a direct observation of a draw from the DF (the p1-o distance, blue line) or it may be a composite of mate choice and dispersal (the p2-o distance, grey line).

sampled all individuals and reconstructed the tree connecting them. For each condition, we ran 3 replicates. We will call the distribution of realised parent-offspring distances in these trees the $\widehat{DD}$ (empirical distance distribution).

The shape of the $\widehat{DD}$ was related to that of the theoretical DF. For example, the DF distribution can be set to be a Cauchy distribution, which is particularly heavy-tailed compared to other distributions (uniform, half-normal, exponential or Rayleigh). When parameterizing the DF as Cauchy, we observe a higher frequency of long $\widehat{DD}$ dispersal values, compared to other DF distributions, when the parameter σ is kept constant (see Fig. 2).

However, there was not a perfect correspondence between $\widehat{DD}$ and DF, as the other ecological parameters in the simulation also influenced the realized distance between parent and offspring. To investigate this, we examined the positions of individuals throughout the simulations. When the competition distance was zero, individuals tended to group together and move cohesively throughout the landscape (as shown in Fig. 4). As the competition distance increased, the population broke into discrete clusters which appeared to "repel" each other. Varying the competition distance had little effect on the $\widehat{DD}$ (Fig. 6), and we found that a competition distance of approximately 0.2 avoided excessive "clumping" or "scattering" - we note that this is not an general threshold, but rather a function of our fleet of parameters, in this section 50 individuals in a world size of 50x50 units. A competition distance of 0.2 also led to the least excess variance in the $\widehat{DD}$ compared to the theoretical DF (Fig. 6).

Increasing the mating distance tended to allow a greater level of scattering, and allowed mating to occur across clustered populations. This is shown by the long branches connecting clusters in Fig. 4. Increasing the mating distance had a strong effect on the excess variance of the $\widehat{DD}$. It also altered the shape of the $\widehat{DD}$: as we increased the mating distance, the $\widehat{DD}$ generally acquired a flat "shoulder" which decayed around the value of the mating distance. Overall, the marked discrepancy between the theoretical and realized parent-offspring distributions - under varying competition and mating distances - suggests

that these ecological parameters may be important determinants of the scale of dispersal of individuals in the wild.

Dispersal Function	Parametrization	Theoretical Mean	Theoretical Variance
Brownian	Distance in x and y dimensions drawn independently from N(0,σ^2). Distance follows Rayleigh(σ)	$\sigma\sqrt{\pi/2}$	$\sigma^2(4-\pi)/2$
Cauchy	Angle drawn uniformly, distance drawn from Cauchy(scale=σ,location=0)	undefined	undefined
Exponential	Angle drawn uniformly, distance drawn from Exp($1/\sigma$)	σ	σ^2
Half-Normal	Angle drawn uniformly, distance drawn from N(0,σ^2). Distance follows folded normal distribution	$\sigma\sqrt{2/\pi}$	$\sigma^2 - \frac{2}{\pi}\sigma^2$
Uniform	Angle drawn uniformly, distance drawn from U(0,σ)	$\sigma/2$	$\sigma^2/12$

Table 1: **The parametrization of parent-offspring distances via the dispersal distance.** We parametrized the dispersal distribution through a parameter σ, such that the theoretical variance increased with σ^2, and the mean with σ (this does not apply to the Cauchy distribution, which has undefined mean and variance; here, σ was the scale parameter). Further details are given in Methods section 5.1.1.

3.2 Modelling dispersal patterns

Inspired by these observations, we developed a theoretical model of parent-offspring distances combining σ and the scale of mating, given a mode of dispersal where distances were drawn from a Gaussian distribution (which here we term "Brownian", see Methods section 5.1.1) using the uniform model of mate choice implemented in *slendr*. This also represents a more general example of a species for which mate choice and dispersal distances are not drawn from the same distribution, or at the same scale.

The distribution of parent-offspring distances is an equally weighted mixture of dispersals from a "gestating" parent (often conceptualised as the mother) and a non-gestating parent. If the parent-offspring distance is y, its density given a dispersal distance parameter σ and a mate choice radius r_b is therefore

$$g_{y|\sigma,r_b}(y|\sigma,r_b) = \frac{1}{2}\frac{y}{\sigma^2}e^{-\frac{y^2}{2\sigma^2}} + \frac{1}{2}f_{y|\sigma,r_b}(y|\sigma,r_b) \tag{1}$$

The first term reflects the density given a standard Rayleigh distribution (between the gestating parent and its offspring) with scale σ, while the second term models the distance between the non-gestating parent and the offspring.

If we assume a uniformly distributed mate choice radius, then the density function of

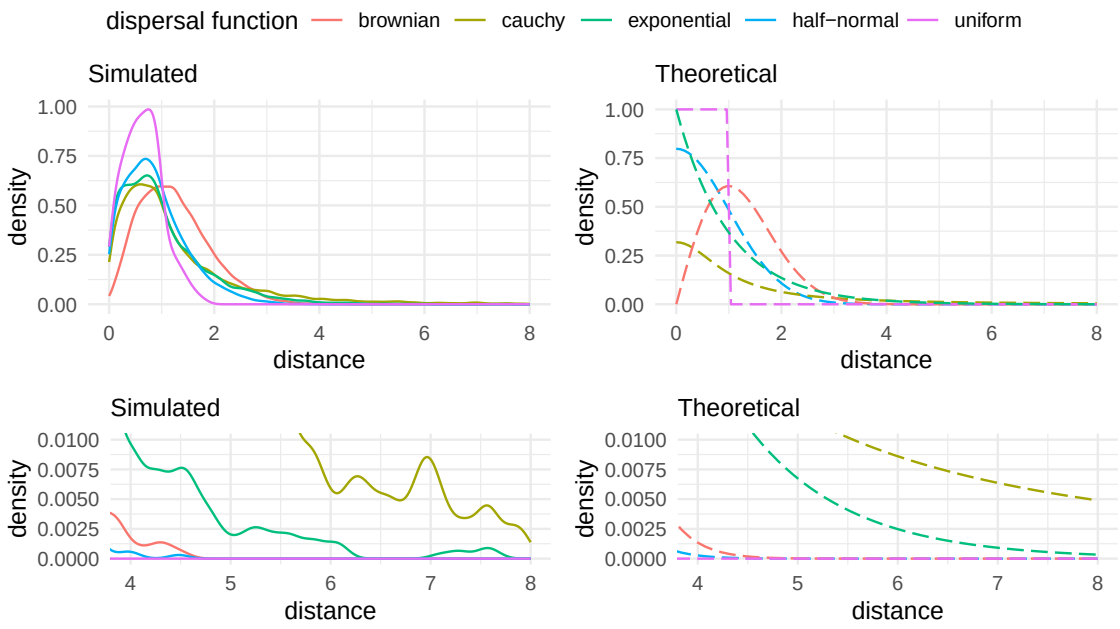

Figure 2: **Distributions of parent-offspring distances reflect the underlying dispersal function.** The left panel shows the simulated distances, while the right one gives the shape of the corresponding dispersal distributions. Bottom: the height of the tails of the distributions corresponded to those of the corresponding dispersal functions, with the Cauchy having the most heavy tail, followed by the exponential, Brownian, half-normal and then uniform. These densities were obtained from three replicates of a simulation of 50 individuals over 50 generations. The mating distance was 1 and the competition distance 0.

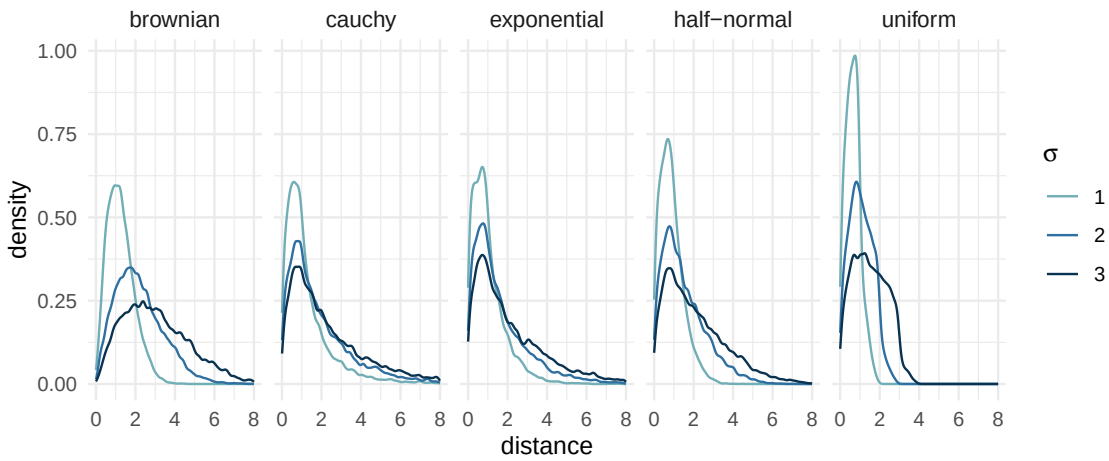

Figure 3: **Increasing σ leads to longer range parent-offspring dispersal.** The figure shows the empirical density of the simulated parent-offspring distances under each of our modes of dispersal, with increasing σ. We chose to parametrize σ for each distribution such that it led to longer parent-offspring distances on average (see Table 1). This is evidenced by the horizontal stretch of the densities with larger σ. These curves were obtained from three replicates of a simulation of 50 individuals over 50 generations. The mating distance was 1 and the competition distance 0.2.

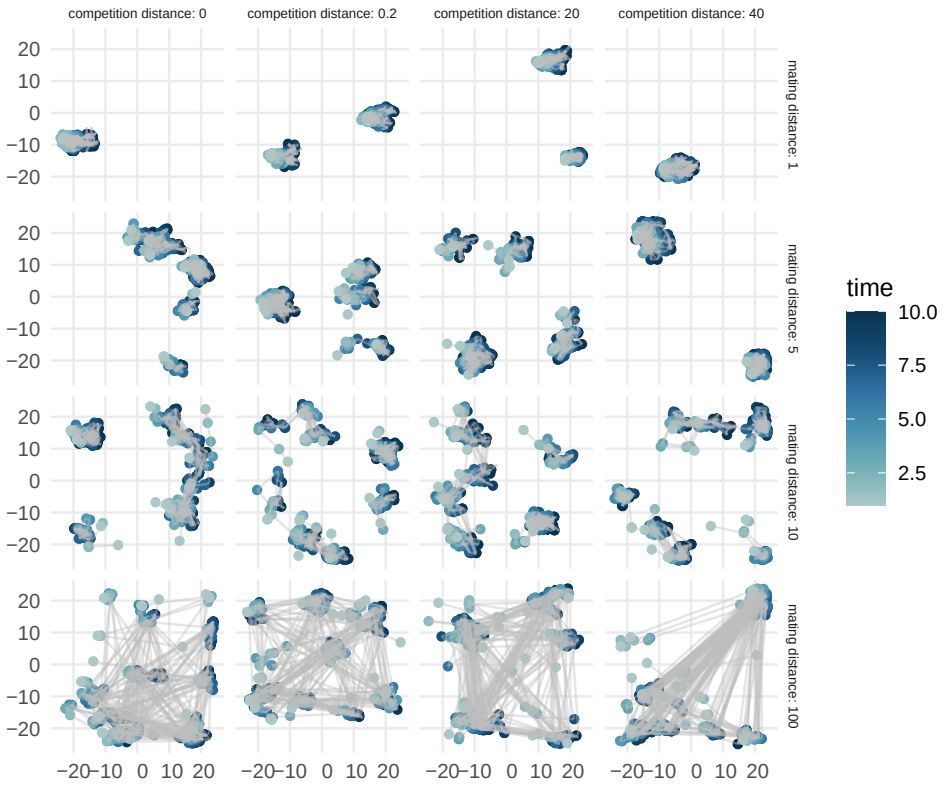

Figure 4: **The effects of the mating and competition distance on the placement of individulas in a population.** The positions of individuals over 10 generations are shown and coloured by time. The connections between parents and offspring are shown as grey lines. When the competition distance was zero (left column), we observed a strong clumping behaviour - which was somewhat rescued by allowing a large scale of mate choice (bottom row). As we increased the competition and mating distances, the population tended to scatter into discrete clusters. In these simulations, the mode of dispersal was 'Brownian' and σ was 1.

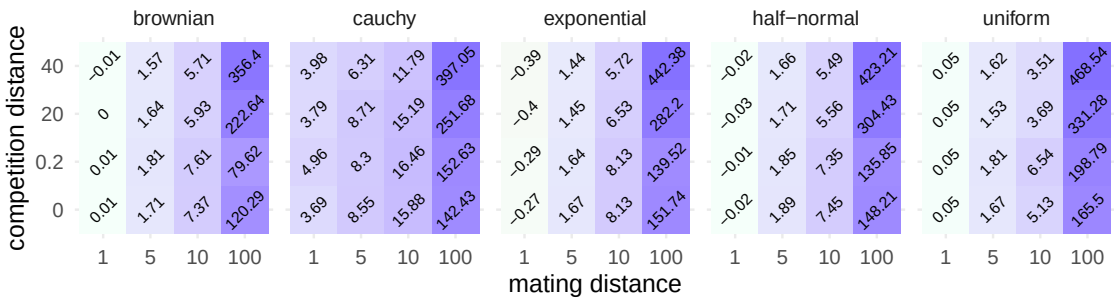

Figure 5: **The effect of mate choice and competition radius on realized parent-offspring distances.** Each tile shows the excess variance of the empirical dispersal distribution compared to the theoretical one - as given by Table 1. Since the Cauchy distribution has undefined variance, the 'excess' is relative to zero. Increasing the competition distance tended to have little effect on the variance of parent-offspring distances, but altering the scale of mate choice had a very strong effect. However, when the mating distance was 100, the excess variance was the lowest when the competition distance was 0.2 - and not zero.

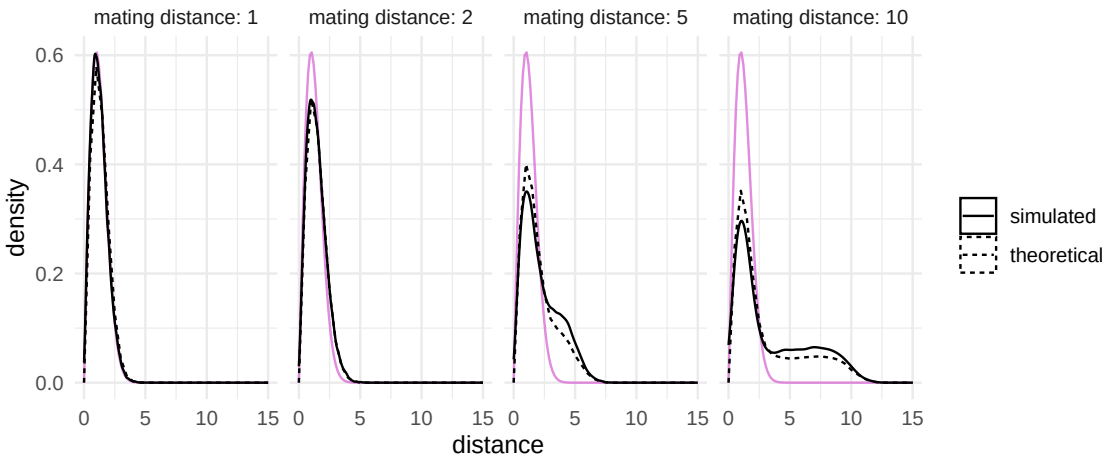

Figure 6: **Modelling dispersal and mate choice.** The distribution of parent-offspring distances against the theoretical distribution given in Equation (1) (dashed lines) and the Rayleigh distribution (purple). There was a close match between the theoretical and simulated distributions across a range of mating distances. As the mating distance increased, the distributions acquired a flat shoulder compared to the Rayleigh distribution from long father-offspring dispersals. In all simulations, the dispersal distance parameter σ was 1 and the competition distance was 0.2.

the distance between the non-gestating parent and the offspring is given by

$$f_{y|\sigma,r_b}(y|\sigma,r_b) = \int_0^{r_b}\int_0^{\infty} \frac{1}{\pi} \times \frac{1}{r_b} \times \frac{a}{\sigma^2} e^{\frac{-a^2}{2\sigma^2}} \times \frac{y}{ab\sqrt{1-\left(\frac{a^2+b^2-y^2}{2ab}\right)^2}}\, da\, db \tag{2}$$

Where a is the distance between the gestating parent and its offspring, and b is the distance between parents. This derivation is elaborated in the Methods section 5.3. We verified that these equations matched the simulated distances (Fig. 6) across the parameter range we examined.

If the mate choice distance is instead modeled more simply as a Rayleigh distribution (see Methods section 5.3), the density function between the offspring and the (unknown) parent can be analytically solved:

$$g_{y|\sigma,\tau}(y|\sigma,\tau) = 0.5\frac{y}{\sigma^2}e^{-\frac{y^2}{2\sigma^2}} + 0.5\frac{y}{\sigma^2+\tau^2}e^{-\frac{y^2}{2(\sigma^2+\tau^2)}} \tag{3}$$

where τ is scale of the Rayleigh distribution governing the mate choice distance.

This formulation also leads to a simple result for the mean parent-offspring distance. Since the expected mother-offspring distance is $\sigma\sqrt{\pi/2}$ and the expected father-offspring distance is $\sqrt{\sigma^2+\tau^2}\sqrt{\pi/2}$, the expected parent-offspring distance is $\frac{\pi\left(\sigma+\sqrt{\sigma^2+\tau^2}\right)}{2\sqrt{2}}$. If we were to measure the distances along branches of a genealogy, we would eventually expect to see generation-scaled distances follow a Gaussian with mean $\frac{\pi\left(\sigma+\sqrt{\sigma^2+\tau^2}\right)}{2\sqrt{2}}$. This may be interpreted as a many-generation "effective" dispersal distance parameter.

3.3 Estimating dispersal distances from spatially tagged trees

Finally, we sought to test how accurately σ could be estimated, given a perfectly inferred spatial tree sequence. Under a Gaussian mode of dispersal (what we term "Brownian"), the maximum likelihood estimator of σ is

$$\hat{\sigma}_{ML} = \sqrt{\frac{1}{2N}\sum_{i=1}^{N}\left(\frac{d_i}{l_i}\right)^2} \tag{4}$$

where the index i denotes each of N branches in a genealogy, with geographic distance d_i and branch-length in generations l_i (see Methods section 5.4).

We sampled 100 genomes across 5 simulation replicates from a population dispersing with a σ of 1, and obtained maximum likelihood estimates of σ from the set of all parent-offpsring distances (Fig. 7). We next emulated a situation where we know the geographic positions of tree tips and internal nodes, but the positions of all individuals in the tree are not known (labelled "*simplified*" in our plots). Lastly, for a more realistic scenario, we extracted the distances between all pairs of tips, to emulate a situation where only present-day individuals have a known location assigned to them ("*tips only*").

While the estimates of σ from the full set of parent-offspring distances were accurate, the estimates from longer tree branches generally underestimated the true parameter. To see whether limited world size was responsible for this observation, we adopted the approach detailed in Osmond and Coop 2021 and eliminated branches which were more than 100 generations long. This caused the distribution of branch-wise distances to more closely resemble that of the simplified tree, and correspondingly caused an increase in the estimated σ. This indicates that accumulated distances over long branches in a given tree tended to be shorter than expected: a phenomenon probably caused by the fact that long-range dispersal is limited in a finite habitat.

Increasing the maximum distance between parents (the mating distance) increases the average parent-offspring distance, as shown in Fig. 7. To investigate the effects of this on the accuracy of the estimates of σ, we simulated under the conditions above but with a range of mating distances. In line with previous results, increasing the mating distance caused the $\hat{\sigma}$ inferred from parent-offspring distances to increase.

We also tested whether assuming an incorrect dispersal kernel could affect estimates of σ. This might be applicable in a situation where, for example, a population follows power-law dispersal, but we assume parent-offspring distances to be Gaussian and attempt to estimate the variance parameter of this dispersal function using an incorrect estimator. To mimic this situation, we simulated under a mode of dispersal where a random angle was drawn from a uniform distribution and a distance from a Cauchy distribution with scale and location 1. The Cauchy distribution is more heavy-tailed compared to a Rayleigh distribution with the same scale. In agreement with this, the estimated σ was larger than the true parameter (Fig. 7), and was also more sensitive to increasing mating distances.

4 Discussion

In this study, we have explored the effects of three important ecological parameters (dispersal distance, competition distance and mate choice distance) on the geographic distances captured in a geographically tagged genealogy.

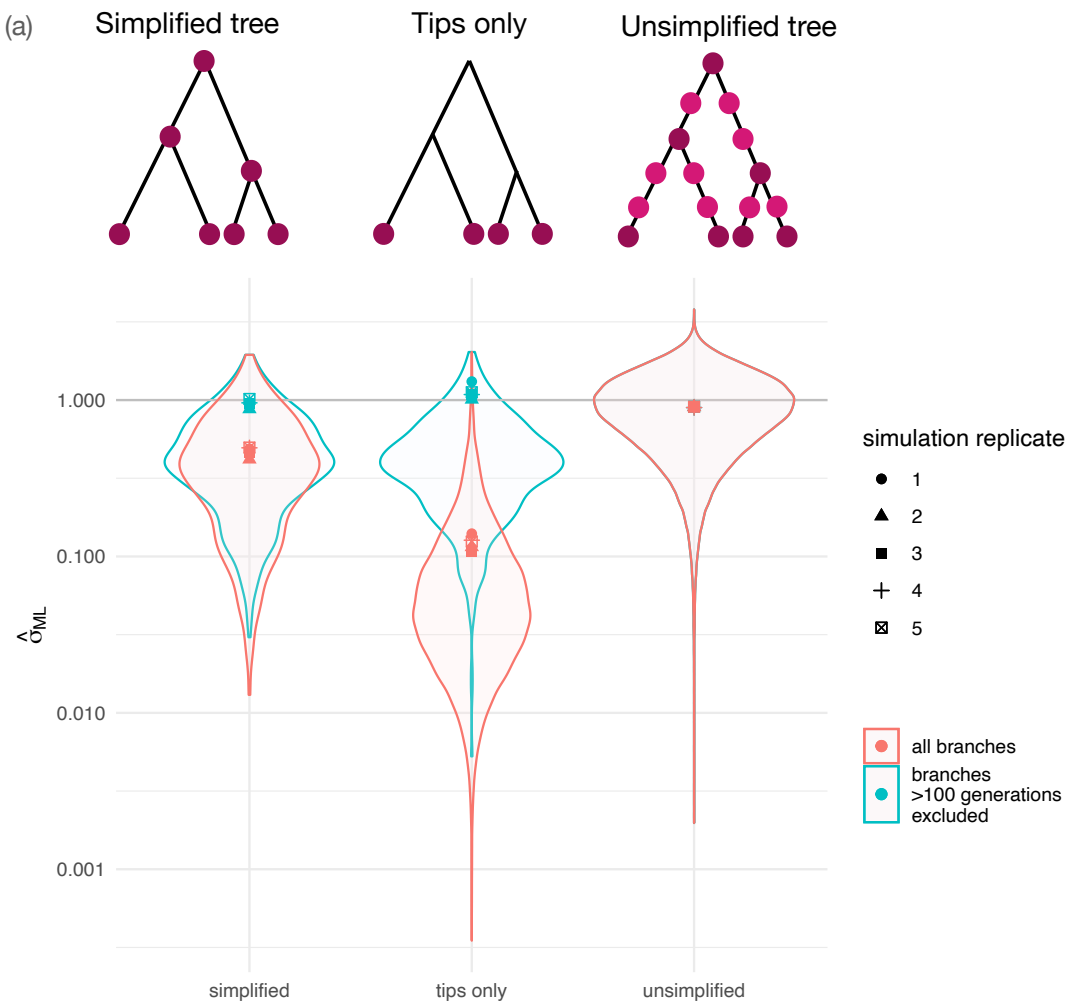

Figure 7: **Estimating the dispersal distance under the Brownian mode of dispersal.** The plots show results from across 5 simulations, and each dot the ML $\hat{\sigma}$ from the whole tree in one replicate. The grey lines show the true σ, 1. (a) The violin plots were produced from all branch-wise $\hat{\sigma}$ values. All branches are length 1 in the unsimplified tree, so 'cutting' (excluding all branches of more than 100 generations from the estimation procedure) has no effect on the distribution of distances. However, cutting long branches increased the estimated dispersal distance for the simplified tree and the tip-only distances.

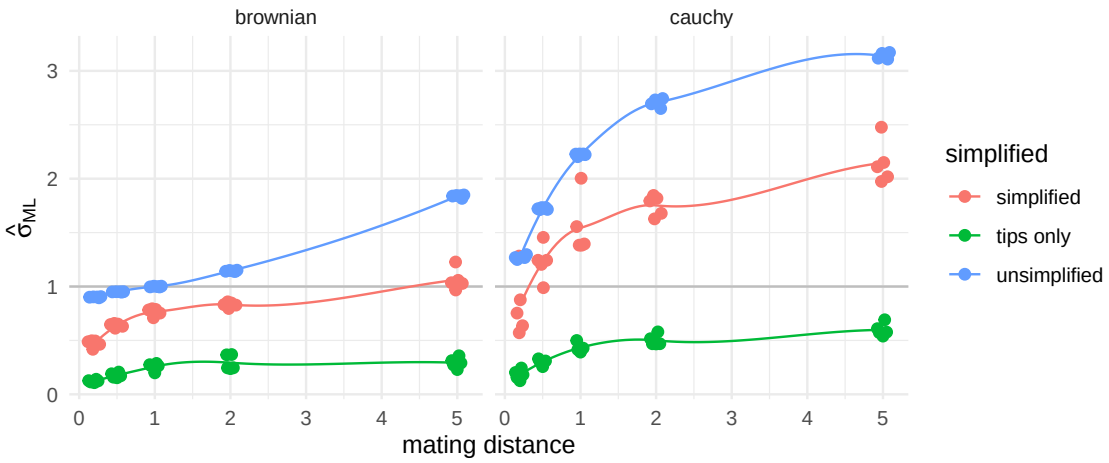

Figure 8: **Estimating the mean dispersal distance under the Brownian mode of dispersal.** Left: increasing the mating distance increases the effective dispersal distance. Right: in these simulations, the dispersal function was *Cauchy* with scale and location 1, but we naively used maximum likelihood estimator of σ for the Brownian mode of dispersal. In this setting, increasing the mating distance led to an even greater inflation of $\hat{\sigma}$. The dots show the result of each of five replicates, and the lines are smoothed rolling means.

We show that altering the kernel of parent-offspring dispersal can have profound effects on the diffusion captured within a genealogy, and in particular on the tails of this dispersal distribution. We were able to manipulate these distributions in order to produce different diffusion rates across the habitat - for example, the Cauchy, a text-book example of a "heavy-tailed" distribution did indeed produce a greater proportion of long-distance dispersals.

There was some difficulty in choosing a common parametrization for these dispersal distributions, especially since *slendr* implements two different mechanics of parent-offspring dispersal (one where a random distance and angle are chosen, and another where latitudinal and longitudinal distances are chosen, see Methods section 5.1.1). We suggest that a pragmatic solution for the sake of simulation might be to encode a dispersal distribution where the "heavy-tailed"-edness of the distribution may be controlled independently from the variance.

The scale of competition had little effect on these distances overall, but caused distinctive patterns in the distribution of a population within its landscape. In particular, long-range competition led to a clustered distribution of individuals, which may be a practical nuisance to simulation users, and lead to unwanted geographic structure.

We observed that the distances within a genealogy increased dramatically if the scale of mate choice was large. Mating is often not explicitly modelled - yet the step of mate allocation is essential in forwards-in-time, agent-based genetic simulators such as *slendr* and *SLiM*. Furthermore, the dynamics of mate choice and parent-offspring dispersal may differ starkly in natural populations: for example, the same model of dispersal may not apply to the dynamics with which pollen and seeds spread. Our results support that this is an important parameter, and absorbing mating and parent-offspring dispersal dynamics into one step may not always be appropriate.

Aside from changing the distances in the genealogy, the scale of mate choice also

changed the shape of the distribution of parent-offspring distances. To illustrate a case where this may be modelled, we described the theoretical distribution of parent-offspring distances under uniform mate choice, and found a close match between the this and simulated distances. The natural next step would be to use these results in an inference framework, by deriving analytical solutions for the maximum-likelihood or method-of-moments estimators for the dispersal and mating distances.

Rather than the theoretical dispersal distance itself, a parameter that may be more liable to inference is an effective dispersal distance parameter, which incorporates both the mate choice and dispersal processes. The distance between parents and offspring over many generations should follow a normal distribution in the limit of infinite generations, due to the central limit theorem. Therefore, if we were to take the distances along branches of a phylogeny and scale them by the respective number of generations (as inferred from genetic data), the distribution of distances would approach a Gaussian distribution, centred around this effective dispersal distance. Specifically, this is an equally weighted mean of the expected distances of the offspring from either parent (see Methods section 5.3). For example, in the Methods section 5.3.1, I show that under a model with Gaussian dispersal (with scale σ) and mate choice (with scale τ), this effective dispersal distance can be easily calculated as $\frac{\pi\left(\sigma+\sqrt{\sigma^2+\tau^2}\right)}{2\sqrt{2}}$.

This compound parameter is in effect what is estimated when mate choice dynamics are not explicitly modelled in phylogeographic studies. We therefore motivate distinguishing between spatial models intended for few generations, where the stages of mating and dispersal should be treated as distinct, from those for phylogenetic time-scales, where they may be absorbed into one parameter. We also note that, over long time-scales, dispersal was limited by finite population ranges. In our results, this led to estimates of the mean dispersal distance which were smaller than expected, illustrating that deep coalescent branches should only be used with caution for inference, as illustrated by Osmond and Coop 2021.

In cases where we are interested in untangling the mating and dispersal distances, uniparentally inherited genetic material could be of use. Mitochondrial DNA only moves via mother-offspring dispersals, the direct manifestation of the dispersal function (when the mother is $p1$). Conversely, the Y-chromosome always moves according to a convolution of mating and dispersal distances. Comparing their respective rates of diffusion could help us identify cases in which the between-parent distance might be masking the underlying mother-offspring dispersal dynamic.

At the moment, *slendr* is not able to model sex differences. Yet, mother-offspring dispersal and mate choice may span different scales if dispersal is strongly sex-biased. Theoretical results across a range of animals suggest that this is the case when the limiting resource differs between males and females (Li and Kokko 2019). In line with this, field observations and genetic data have pointed to a breadth of matrilocal and patrilocal behaviours across animal species (for example Liebgold, Brodie III, and Cabe 2011; Oota et al. 2001; Schubert et al. 2011). These sex-biased processes might be an intriguing direction for further investigation.

Another exciting direction for further study is selection. A positively selected allele will often have more descendants than a neutral one, resulting in excess branching. This means that positively selected loci, and genomic regions in linkage disequilibrium with them, are expected to have more descendant lineages which can explore space and travel faster than neutral ones. This result is similar to Fisher's travelling wave model, where

the velocity of spread is proportional to the square root of the selection coefficient (Fisher 1937; Muktupavela et al. 2021; Steiner and Novembre 2022). For the purpose of inference, we often assume that the coalescent branching process and geographic location are independent (although this is not the case, see Wilkins and Wakeley 2002). How far do we deviate from this assumption, for example, when selection pressures are local?

Overall, it is clear that accurately modelling the dispersal of a given species may require sound understanding of a variety of ecological parameters. From our simulations, we observed that geographic distances captured within a geographically tagged genealogy captured these compound effects. These are not yet theoretically well-understood, and may become confounding factors in joint analyses of geographic space and genetic diversity. Simulations will be key to approaching these issues.

5 Methods

5.1 Spatial simulations

We used the software *SLiM* (Haller and Messer 2019) via its R interface *slendr* (Petr et al. 2022) to simulate populations in space and time. At ecah generation, individuals were chosen randomly at each generation to be the parents of the next generation. Individuals were hermaphroditic and generations were discrete and non-overlapping. There was no modelled age structure or sex-based differentiation. We chose to keep these at a constant size in order to focus on fundamental aspects of dispersal without confounding effects from demography.

In *slendr*, a user specifies a model and its parameters. These are passed to a *SLiM* backend, which executes the simulation. After this, among the data which can be recovered from a simulation are the locations of all individuals, the times at which they lived and the phylogeny and pedigree connecting them. A schematic overview of this is given in Figure 9.

5.1.1 Encoding dispersal

We simulated under several modes of mother-offspring dispersal, coming under two categories:

1. Angle-distance dispersal: in these, the absolute distance are controlled by a given distribution. An angle is drawn randomly from a uniform distribution between 0 and 2π, and a distance d was drawn from one of the following distributions:
    - *Uniform*: the p1-offspring distance is uniformly distributed between 0 and σ, $d \sim U(0, \sigma)$. The mean absolute distance is $\sigma/2$ and the variance is $(1/12)\sigma^2$.
    - *Half-Normal*: the p1-offspring distance is Gaussian distributed, with mean 0 and variance σ^2. When a distance is below zero, the offspring is effectively ejected backwards. The mean of the resulting folded normal distribution (specifically, a half-normal) is $\sigma(\sqrt{2/\pi})$ and the variance is σ^2.
    - *Exponential*: the p1-offspring distance is exponentially distributed, with rate parameter $1/\sigma$, $d \sim Exp(\sigma)$. The mean is σ and the variance is σ^2.

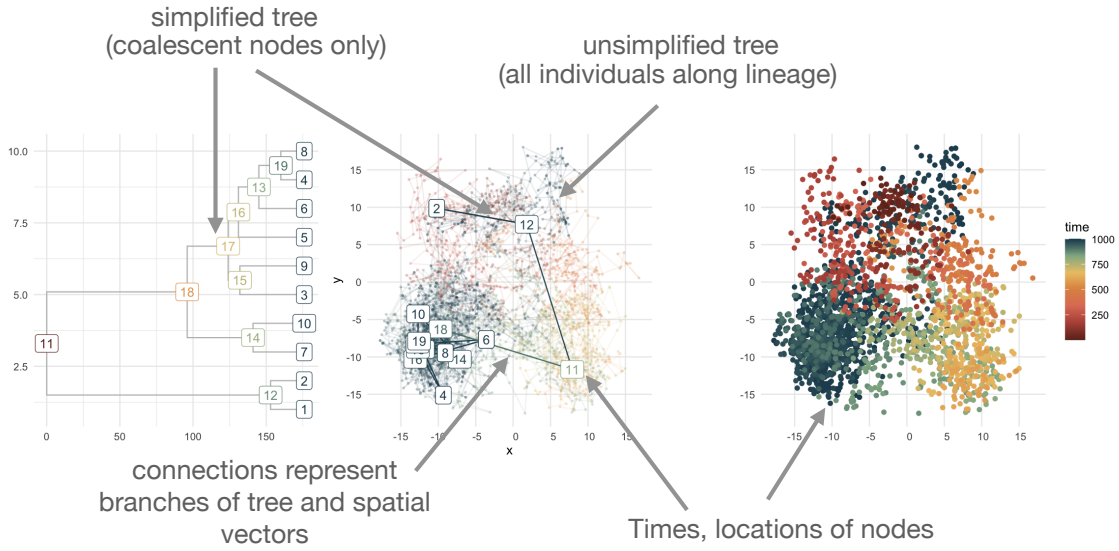

Figure 9: **An overview of the data which can be retrieved using the spatio-temporal population genomic simulator, *slendr*.** (a) A *slendr* user can retrieve both unsimplified and simplified trees. In an unsimplified tree, all the ancestors of sampled tips are recorded. In a simplified tree, only nodes which represent coalescence events are retained. (b) Each dot is a recorded node, and its colour represents the time at which it lived (the number of generations after which the individual was born, from the start of the simulation). The x and y axes represent Eastings and Northings. All nodes on simplified and unsimplified trees are associated with a time and location, and all edges (connections) with a geo-spatial vector, shown as a line. The transparent dots and lines in the middle panel show the individuals and connections in the unsimplified tree. The nodes in the simplified tree are overlaid on this, labelled according to id in the phylogenetic tree (left). (c) Each panel shows a slice of 250 generations, and each dot represents an individual in the unsimplified tree, coloured according to its age. The recording of times and locations allows the user to inspect nascent geographic structure in a simulation.

- *Cauchy*: the p1-offspring distance is Cauchy distributed, with location 0 and rate parameter σ, $d \sim Cauchy(0, \sigma)$. The mean and variance of this distribution are undefined.

2. *Brownian*: here, the axial distances are controlled. random distances in the x and y dimensions (d_x and d_y), each drawn from a Gaussian with mean 0 and variance σ^2, $d_x \sim \mathcal{N}(0, \sigma^2)$, $d_y \sim \mathcal{N}(0, \sigma^2)$. The absolute distance then follows a Rayleigh distribution with scale σ, which has mean $\sigma(\sqrt{\pi/2})$ and variance $\frac{4-\pi}{2}\sigma^2$. This mode is named "*Brownian*" due to its conceptual relation to a 2-dimensional Brownian motion.

5.1.2 Tree recording and manipulation

We simulated a single locus in order to focus on fundamental geographic dynamics which act on single trees. After a simulation run, we retrieved the simplified and unsimplified trees. Simplified trees consist of nodes representing coalescence events, and edges connecting them, which implicitly record many individuals. In contrast, an unsimplified tree records all individuals along edges. Such a tree is useful to directly observe the dispersals which occurred at every generation along a long branch. We processed and analysed these via the *slendr* interface to the *tskit* library (Kelleher, Thornton, et al. 2018). *tskit* is a powerful framework for storing and manipulating trees and tree-sequences with close-to-optimal space usage. We also converted these trees to the "phylo" R object class, which allowed us to analyse them via the phylogenetics package *ape* (Paradis and Schliep 2019).

5.1.3 Geo-spatial analyses

slendr integrates with the spatial package *sf* (Pebesma et al. 2018), and this allowed us to extract a variety of spatial features from the trees, including the positions of individuals, the vectors connecting nodes and the distances between them.

5.2 Statistics and Plotting

We calculated statistics in base *R*, as well as with the packages *VGAM* (T. W. Yee, M. T. Yee, and VGAMdata 2022) and *moments* (Komsta and Novomestky 2015). We evaluated numerical integrals in Mathematica (Wolfram 1991). We produced plots with *ggplot2* (Gómez-Rubio 2017) and auxiliary packages.

5.3 Derivation of the probability density of the distribution of parent-offspring distances

A diploid individual carries two genome copies, each inherited from a parent. These have a distinct genealogy and in any given tree, we follow the movement of one of these copies through individuals over time and space. We can therefore break down the dispersals which occur in one generation into two categories:

1. Genetic parent is the "mother", $p1$. We observe $p1$-offspring dispersal, (which in *slendr* is directly encoded).
2. Genetic parent is the "father", $p2$. We observe a convolution of $p1$-offspring dispersal and the $p1 - p2$ distance.

We can draw a triangle which connects both parents and offspring, as shown in Fig. 10. In case (1), we observe side $\tilde{a}$. In case (2), we observe side $\tilde{y}$. $\tilde{b}$ is the distance which separates the two parents, and the angle between sides $\tilde{a}$ and $\tilde{b}$ is $\tilde{\theta}$. $\tilde{a} \sim Rayleigh(\sigma)$, if we have *Brownian* dispersal. Since in *slendr*, parents are chosen with uniform probability from a specified radius r_b (the mating distance), $\tilde{b} \sim Unif(0, r_b)$ where r_b is the mating distance. The angle between these sides is free to range between zero and π, so $\tilde{\theta} \sim Unif(0, \pi)$.

We can calculate the length of the side y from a, b and θ:

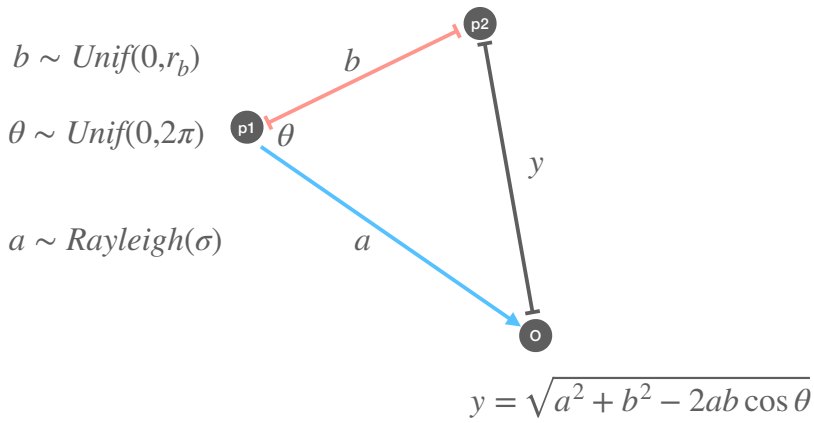

Figure 10: **A schematic of parent-offspring dispersal.** When we observe dispersal from $p2$, the observed parent-offspring distance (y) is a convolution of the distance between $p1$ and $p2$ (b, in red), and the dispersal between $p1$ and the offspring (a, in blue). The cosine rule gives us an expression for y in terms of a, b and the angle between them θ. If we know the probability distributions of a, b and θ, we can obtain that of y via a change of variables.

$$y = \sqrt{a^2 + b^2 - 2ab\cos\theta}$$

We aim to derive the probability density function (pdf) of y, using the pdfs of a, b and θ. This can be achieved with a change of variables:

$$f_{y,a,b}(y,a,b) = f_a(a)f_b(b)f_\theta(\theta) \times \frac{1}{\det(\mathbf{J})} \tag{5}$$

$\mathbf{J}$ is the jacobian matrix of partial derivatives:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial a} & \frac{\partial y}{\partial b} \\ \frac{\partial a}{\partial \theta} & \frac{\partial a}{\partial a} & \frac{\partial a}{\partial b} \\ \frac{\partial b}{\partial \theta} & \frac{\partial b}{\partial a} & \frac{\partial b}{\partial b} \end{bmatrix} = \begin{bmatrix} \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial a} & \frac{\partial y}{\partial b} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The determinant of this matrix is

$$\det(\mathbf{J}) = \frac{\partial y}{\partial \theta} = \frac{ab\sqrt{1 - \left(\frac{a^2+b^2-y^2}{2ab}\right)^2}}{y} \tag{6}$$

Which goes back into equation (5):

$$f_{y,a,b}(y,a,b) = f_a(a)f_b(b)f_\theta(\theta) \times \frac{1}{\det(\mathbf{J})} \tag{7}$$

$$= f_a(a)f_b(b)f_\theta(\theta) \times \frac{y}{ab\sqrt{1-\left(\frac{a^2+b^2-y^2}{2ab}\right)^2}} \tag{8}$$

This is the joint pdf of the three sides of the triangle. Now, we integrate out the parameters a and b in order to get a fully marginalised f_y.

$$f_y(y) = \int_b \int_a f_a(a)f_b(b)f_\theta(\theta) \times \frac{y}{ab\sqrt{1-\left(\frac{a^2+b^2-y^2}{2ab}\right)^2}}\, da\, db \tag{9}$$

This holds for any distribution of a and b. Let's consider the case where a is Rayleigh distributed (as it is under the *Brownian* mode of dispersal), and mate choice is random within a radius r_b (as encoded in *slendr*). θ and b are uniform random variables, so have a constant probability of $1/\pi$ and $1/r_b$ respectively. We also know that a has a Rayleigh pdf of $(a/\sigma^2)e^{(-a^2/2\sigma^2)}$. Replacing these in the function above:

$$f_y(y) = \int_0^{r_b} \int_0^{\infty} \frac{1}{\pi} \times \frac{1}{r_b} \times \frac{a}{\sigma^2} e^{\frac{-a^2}{2\sigma^2}} \times \frac{y}{ab\sqrt{1-\left(\frac{a^2+b^2-y^2}{2ab}\right)^2}}\, da\, db \tag{10}$$

This is the fully marginalised pdf of y. This integral is challenging to solve analytically, but we can obtain the approximate shape of the pdf by numerical integration.

Finally, we can write out the pdf of the distance between a randomly chosen parent and its offspring. Let's call this pdf $g_y(y)$. With probability $P = 0.5$, the parent is the mother ($p1$) and y simply follows a Rayleigh distribution with scale σ. When the genome is inherited from the father ($p2$), which again occurs with $P = 0.5$, the pdf of y is the distribution shown above. This leads to the final pdf $g_y(y)$ of the parent-offspring distance,

$$g_{y|\sigma,r_b}(y|\sigma,r_b) = \frac{1}{2}\frac{y}{\sigma^2}e^{-\frac{y^2}{2\sigma^2}} + \frac{1}{2}f_{y|\sigma,r_b}(y|\sigma,r_b) \tag{11}$$

Note that we could get the expectation of the distance y from this expression:

$$\mathbb{E}[y|\sigma,r_b] = \int_{-\infty}^{\infty} \left[\frac{y}{2}r_{y|\sigma}(y|\sigma) + \frac{1}{2}f_{y|\sigma,r_b}(y|\sigma,r_b)\right] dy \tag{12}$$

$$= \int_{-\infty}^{\infty} \left[\frac{y}{2}r_{y|\sigma}(y|\sigma)\right] dy + \int_{-\infty}^{\infty} \left[\frac{y}{2}f_{y|\sigma,r_b}(y|\sigma,r_b)\right] dy \tag{13}$$

which is a half-weighted average of the distance expected from the parent-offspring distance kernel, and from mate choice.

5.3.1 A simpler model with Gaussian mate choice

There are simple scenarios that lead to a more analytically tractable pdf. For example, let us suppose that the distance between parents is also generated in a similar way to *Brownian* dispersal, from independent normal distributions in x and y dimensions with variance τ^2. In this case, the father-offspring distance in each dimension is a sum of two Gaussian random variables and is itself normally distributed with variance $\sigma^2 + \tau^2$. This gives rise to a Rayleigh distribution with scale $\sqrt{\sigma^2 + \tau^2}$ for the norm of the distance, y. In that case, the final pdf is then:

$$g_{y|\sigma,\tau}(y|\sigma,\tau) = 0.5\frac{y}{\sigma^2}e^{-\frac{y^2}{2\sigma^2}} + 0.5\frac{y}{\sigma^2+\tau^2}e^{-\frac{y^2}{2(\sigma^2+\tau^2)}} \tag{14}$$

5.4 Maximum likelihood estimation of σ

When the mating distance is small, and dispersal is "Brownian", distances in latitude and longitude at each generation are drawn from independent $\mathcal{N}(0, \sigma^2)$, and the dispersal over many generations may be modelled as a Brownian motion. Given a genealogy with N branches i, of length l_i and geographic distance d_i, the maximum likelihood estimator of σ is given by

$$\hat{\sigma}_{ML} = \sqrt{\frac{1}{2N}\sum_{i=1}^{N}\left(\frac{d_i}{l_i}\right)^2} \tag{15}$$

We may also wish to survey how each branch is contributing to the estimate. Since $\mathbb{E}[d_i] = \sqrt{l_i}\sigma\sqrt{\frac{\pi}{2}}$, we define $\hat{\sigma}_{branch} = \frac{d_i}{\frac{l_i\pi}{2}}$.

5.5 Code availability

The functions used (which are not included in *slendr* or other packages) are available as an R package *treesinspace* (`https://github.com/mkiravn/treesinspace/`). We include all relevant scripts, with which the simulations and plots included may be reproduced.

5.6 Acknowledgements

We thank members of the Racimo group and the Novembre group for valuable comments on the manuscript, as well as Thomas Bataillon (Aarhus University) and John Novembre (University of Chicago).

FR was supported by a Villum Young Investigator Grant (project no. 00025300), a COREX ERC Synergy grant (ID 951385) and a Novo Nordisk Fonden Data Science Ascending Investigator Award (NNF22OC0076816). MP was supported by a Lundbeck Foundation grant (R302-2018-2155) and a Novo Nordisk Foundation grant (NNF18SA0035006) given to the GeoGenetics Centre.

References

Aguillon, Stepfanie M et al. (2017). "Deconstructing isolation-by-distance: The genomic consequences of limited dispersal". In: *PLoS genetics* 13.8, e1006911.

Castillo, XA Olarte et al. (2011). "Phylogeography and molecular epidemiology of Papaya ringspot virus". In: *Virus research* 159.2, pp. 132–140.

Cayuela, Hugo et al. (2018). "Demographic and genetic approaches to study dispersal in wild animal populations: A methodological review". In: *Molecular ecology* 27.20, pp. 3976–4010.

Charlesworth, Brian, Deborah Charlesworth, and Nicholas H Barton (2003). "The effects of genetic and geographic structure on neutral variation". In: *Annual Review of Ecology, Evolution, and Systematics* 34.1, pp. 99–125.

Clark, James S (1998). "Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord". In: *The American Naturalist* 152.2, pp. 204–224.

Fisher, Ronald Aylmer (1937). "The wave of advance of advantageous genes". In: *Annals of eugenics* 7.4, pp. 355–369.

Gómez-Rubio, Virgilio (2017). "ggplot2-elegant graphics for data analysis". In: *Journal of Statistical Software* 77, pp. 1–3.

Griffiths, Robert C and Paul Marjoram (1996). "Ancestral inference from samples of DNA sequences with recombination". In: *Journal of computational biology* 3.4, pp. 479–502.

— (1997). "An ancestral recombination graph". In: *Institute for Mathematics and its Applications* 87, p. 257.

Haller, Benjamin C, Jared Galloway, et al. (2019). "Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes". In: *Molecular ecology resources* 19.2, pp. 552–566.

Haller, Benjamin C and Philipp W Messer (2019). "SLiM 3: forward genetic simulations beyond the Wright–Fisher model". In: *Molecular biology and evolution* 36.3, pp. 632–637.

Hardy, Olivier J and Xavier Vekemans (1999). "Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models". In: *Heredity* 83.2, pp. 145–154.

Hubisz, Melissa and Adam Siepel (2020). "Inference of ancestral recombination graphs using ARGweaver". In: *Statistical Population Genomics*. Humana, New York, NY, pp. 231–266.

Hudson, Richard R et al. (1990). "Gene genealogies and the coalescent process". In: *Oxford surveys in evolutionary biology* 7.1, p. 44.

Jenkins, David G et al. (2010). "A meta-analysis of isolation by distance: relic or reference standard for landscape genetics?" In: *Ecography* 33.2, pp. 315–320.

Kelleher, Jerome, Kevin R Thornton, et al. (2018). "Efficient pedigree recording for fast population genetics simulation". In: *PLoS computational biology* 14.11, e1006581.

Kelleher, Jerome, Yan Wong, et al. (2019). "Inferring whole-genome histories in large population datasets". In: *Nature genetics* 51.9, pp. 1330–1338.

Komsta, Lukasz and Frederick Novomestky (2015). "Moments, cumulants, skewness, kurtosis and related tests". In: *R package version* 14.

Kot, Mark, Mark A Lewis, and Pauline van den Driessche (1996). "Dispersal data and the spread of invading organisms". In: *Ecology* 77.7, pp. 2027–2042.

Li, Xiang-Yi and Hanna Kokko (2019). "Sex-biased dispersal: A review of the theory". In: *Biological Reviews* 94.2, pp. 721–736.

Liebgold, Eric B, Edmund D Brodie III, and Paul R Cabe (2011). "Female philopatry and male-biased dispersal in a direct-developing salamander, Plethodon cinereus". In: *Molecular Ecology* 20.2, pp. 249–257.

Markov, Peter V et al. (2009). "Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa". In: *Journal of General Virology* 90.9, pp. 2086–2096.

McCann, Kevin et al. (2000). "Population outbreaks in a discrete world". In: *Theoretical Population Biology* 57.2, pp. 97–108.

Miles, MA et al. (2009). "The molecular epidemiology and phylogeography of Trypanosoma cruzi and parallel research on Leishmania: looking back and to the future". In: *Parasitology* 136.12, pp. 1509–1528.

Muktupavela, Rasa A et al. (2021). "Modelling the spatiotemporal spread of beneficial alleles using ancient genomes". In: *bioRxiv*.

Oota, Hiroki et al. (2001). "Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence". In: *Nature genetics* 29.1, pp. 20–21.

Osmond, Matthew M and Graham Coop (2021). "Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies". In: *bioRxiv*.

Paradis, Emmanuel and Klaus Schliep (2019). "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R". In: *Bioinformatics* 35.3, pp. 526–528.

Paulose, Jayson and Oskar Hallatschek (2020). "The impact of long-range dispersal on gene surfing". In: *Proceedings of the National Academy of Sciences* 117.14, pp. 7584–7593.

Pebesma, Edzer J et al. (2018). "Simple features for R: standardized support for spatial vector data." In: *R J.* 10.1, p. 439.

Petr, Martin et al. (2022). "slendr: a framework for spatio-temporal population genomic simulations on geographic landscapes". In: *bioRxiv*.

Rousset, Francois (2001). "Inferences from spatial population genetics". In: *Handbook of statistical genetics* 4, p. 23.

Schubert, Grit et al. (2011). "Male-mediated gene flow in patrilocal primates". In: *PLoS One* 6.7, e21514.

Sexton, Jason P, Sandra B Hangartner, and Ary A Hoffmann (2014). "Genetic isolation by environment or distance: which pattern of gene flow is most common?" In: *Evolution* 68.1, pp. 1–15.

Speidel, Leo et al. (2019). "A method for genome-wide genealogy estimation for thousands of samples". In: *Nature genetics* 51.9, pp. 1321–1329.

Steiner, Margaret C and John Novembre (2022). "Population genetic models for the spatial spread of adaptive variants: A review in light of SARS-CoV-2 evolution". In: *PLoS Genetics* 18.9, e1010391.

Wilkins, Jon F and John Wakeley (2002). "The coalescent in a continuous, finite, linear population". In: *Genetics* 161.2, pp. 873–888.

Wohns, Anthony Wilder et al. (2021). "A unified genealogy of modern and ancient genomes". In: *bioRxiv*.

Wolfram, Stephen (1991). *Mathematica: a system for doing mathematics by computer*. Addison Wesley Longman Publishing Co., Inc.

Wright, Sewall (1943). "Isolation by distance". In: *Genetics* 28.2, p. 114.

Yee, Thomas W, Maintainer Thomas Yee, and Suggests VGAMdata (2022). "Package 'VGAM'". In.