Response to comments on ‘Exploring the Effects of Ecological Parameters on the Spatial Structure of Genealogies’

Mariadaria Ianni-Ravn, Martin Petr and Fernando Racimo

Dear Dr Ortega-Del Vecchyo,

We have been working to revise our manuscript, “Exploring the Effects of Ecological Parameters on the Spatial Structure of Genealogies”, in response to your comments, and here upload an updated version. We thank you, and all the referees for your time and work - we strongly believe that the critique provided has led to significant improvements in the manuscript.

We have responded to each of the comments, and provide point-by-point responses below. The line numbers refer to the revised version. Additionally, we have amended the order of some of the figures, and shifted two to be supplementary material. For simplicity, we keep these in the same document.

Furthermore, to comply with PCI requirements, we have added a conflict of interest disclosure section, and have supplied a perennial URL (https://zenodo.org/doi/10.5281/zenodo.10402648) for the scripts used in the analysis.

All the best,
Mariadaria Ianni-Ravn

Diego Ortega-Del Vecchyo

Round #1

by Diego Ortega-Del Vecchyo, 13 Jun 2023 23:25

Manuscript: https://doi.org/10.1101/2023.03.27.534388 version 2

Great contribution to understand factors changing patterns of spatial genetic variation

Dear Mariadaria K. Ianni-Ravn and coauthors,

Four reviewers and myself have read your paper and we have found that your work is a great contribution to understand the impact of various phenomena on patterns of spatial genetic variation. The reviewers and I have some comments that would be good to address. If you are interested, I would like to consider a revised version of the manuscript for recommendation after taking into account all the comments into account.

All the best,

Diego

Comments:

- ☐ The number of the figures should be mentioned on the text in an ascending order starting from Figure 1. Figure 2 is referenced before Figure 1 on the text. Thank you for pointing this out - the reference has been corrected.
- ☐ Figure 2.- “the right one gives the shape of the corresponding dispersal distributions.” The legend at the top of the right panel figures state that these are “Theoretical”. Unsure about what are the right panels showing. The right hand panels show parent-offspring distances sampled from the SLiM simulations. This was unclear, and we have amended the title of the right hand panels to reflect this.
- ☐ Figure 2.- Could the authors state more precisely what is the difference between the two top panels and the two bottom panels? The bottom two panels are merely a small section of the top plot - specifically, a zoom-in on the tails. This has been clarified in the caption of the Figure 2 by adding the following: “Bottom: a zoom-in on the tails of curves; the height of the tails of the distributions correspond to those of the corresponding dispersal functions…”
- ☐ Line 120.- “The shape of the DDd was related to that of the theoretical DF”. It would be nice to give a line or two explaining how the theoretical DF was derived and pointing to

the reader to the place on the manuscript showing where they can find the derivation of these results. In this sentence, we were referring to the theoretical distributions from which p1-offspring distances were drawn, rather than a theoretical model including both dispersal and mate choice (which we do in a later section). To make this clearer, we amended the wording to: “We compared parent-offspring distances sampled from the simulations (the $\widehat{DD}$) to the theoretical probability distributions from which p1-offspring distances were drawn (the DF). The shape of the $\widehat{DD}$ tended to be related to that of the DF. For example, 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. \ref{fig:1}). This is consistent with the heavy tail of the theoretical Cauchy distribution, compared to other distributions (uniform, half-normal, exponential or Rayleigh).”

☐ Figure 7.- I would suggest to have results for alternative values of sigma to check the accuracy of the estimate for alternative values of this parameter. We have added another figure (Figure S2, also pasted below) where we check alternate values of sigma. We find that the same patterns hold - the results from the unsimplified tree are relatively accurate, while simplified genealogies and tip-only estimates are biased downwards.

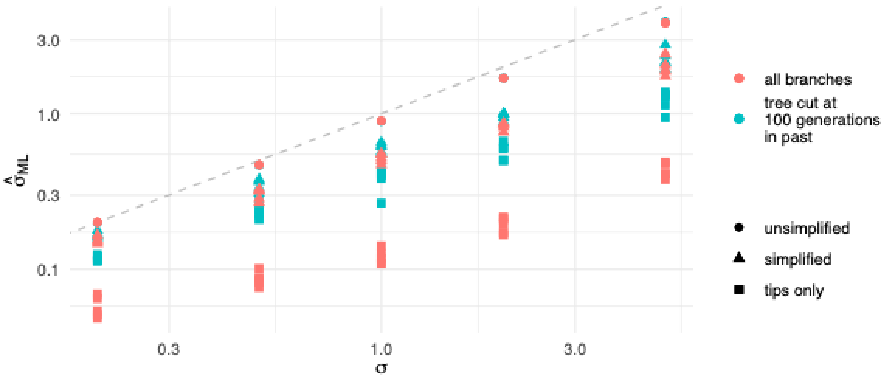


Figure S2: **Estimating the dispersal distance with Brownian dispersal, across a range of σ values.** The grey line shows the true σ. We found that the pattern of bias shown in (a) was replicated across the range of σ values tested. In these simulations, the mating distance was 0.2 and the competition distance was 0.2.

☐ Figure 8 is not referenced on the text. Again, thank you for pointing this out - there was a typo in the figure reference on line 203, which has been amended.

Anonymous Reviewer #1

In this manuscript, the authors seek to understand how different population genetic and ecological parameters affect the geographic properties of the distribution of gene genealogies (i.e., patterns of isolation by distance) across the genome for a sample of individuals. Specifically, they are interested in how the mechanics of mating and dispersal affects the distribution of parent-offspring distances. They investigate this question using a simulation-based approach, and implement a maximum-likelihood (ML) estimator of the mean parent-offspring distance. They demonstrate (1) a correspondence between the shape of the dispersal kernel specified as well as the magnitude of the simulated dispersal and the resultant distribution of parent-offspring distances and (2) that competition-based density regulation, as well as mate-choice radius, both affect parent-offspring distances. They further demonstrate that their ML estimator of dispersal works pretty well and is impacted by the mating radius simulated.
This is a timely manuscript that contains lots of information that will be useful to many researchers who are interested in spatial population genetics (also landscape genetics, phylogeography, etc.). Overall, the paper is well-written and the details of the analyses are, for the most part, clear. However, I have some minor comments on the scholarship (I think some relevant work is overlooked) and a more major concern with the authors' interpretation and presentation of their results.

Major Comments:
In the Results, the authors present the lack of perfect correspondence between \widehat{DD} (the empirical parent-offspring distance distribution) and DF (the specified dispersal function) as something unexpected, a framing that reappears in the discussion. For me, this framing was both confusing and a little bit misleading. Using genetic methods, it is only ever possible to learn about effective parameters (e.g., effective dispersal, effective density). Because the SLiM model includes a spatial mate choice component, the DF specified in SLiM will should not be the same as the \widehat{DD}. That is to say, in these simulations, effective dispersal differs from the

specified DF not only because the DF describes the distribution of dispersal distances from in the census population, rather than just in nodes of the tree sequence with the sampled individuals as its tips, but also because the distribution of distances from offspring to parents (backward in time) will necessarily incorporate the spatial mechanics of mate choice. Focusing on the discrepancy between DF and \widehat{DD} as a main result therefore seems like a confusing choice on the part of the authors, because it feels like a feature of their simulation (rather than an emergent property or a surprising result). Note also that this discrepancy has been discussed elsewhere in the spatial population genetic literature; see Smith et al 2023 ("Dispersal inference from population genetic variation using a convolutional neural network"), in which the authors derive the "mean squared directional displacement between a child and a randomly chosen parent" as a function of the mother-offspring and mother-father spatial mechanics. (Although note that their model, which simulated both dispersal and mate choice from truncated Gaussian distributions with the same variance, is somewhat different from the one implemented in this manuscript). To be clear, I absolutely think that it's interesting to explore the effects of mating and dispersal forward-time parameters of effective dispersal observed in, or inferred from, the tree-sequence. I just think the manuscript would be more effective if it's framing of these points was clearer.

Thank you for this insightful comment. We agree that our wording was unhelpful. Therefore, we have worked to shift our focus from emphasising the fact that the DD and DF did not agree, to instead "learning the relationship" between the two throughout the manuscript. See, for example, a change in the first sentence of the results
"We were interested in learning the relationship between the observed parent-offspring distances in a (perfectly inferred) genealogy and the underlying dispersal function in a population." We think that this change helps to communicate the main focus of our results - how the backwards-in-time distribution of distances between parents and offspring is generated by a combination of all the dispersal mechanics you mention, and how this manifests from a time-scale of single generations up to that of genealogical branches.

Minor Comments:

- □ L96: This is nitpicky but I think it would help with the clarity of the introduction of the simulations if you specified here (rather than just in the detailed methods) that generations are non-overlapping. Thank you for this comment, we have added this information to the line 99: “We leverage \textit{slendr} to carry out forwards-in-time simulations with non-overlapping generations, and study how ecological parameters affect the spatial distribution of individuals, and the structure of genealogies relating them over time”

☐ L27-29: also lots of simulation-based studies and statistical inference. See Battey, Kern, & Ralph “Space is the Place: effects of continuous spatial structure on analysis of population genetic data” as well as Bradburd & Ralph “Spatial population genetics: it’s about time,” both of which contain many references that might be appropriate to cite here (especially work by Malecot and Rousset). Thank you for this pointer, we have added some key references to this section.

☐ L38-39: see also: Smith & Weissman “Isolation by Distance in Populations with Long-Range Dispersal” Thank you for the helpful reference, this has been added to the relevant paragraph.

☐ Figure 4: I’m a little confused about the simulations and what’s being shown in this Figure. Are the dots all the individuals present in the simulation over 10 generations, or simply following a single lineage? If the former, I think it might be a more generalizable spatial model to have competition induce stronger density-dependence so that individuals occupy a larger portion of the available range. If increasing the strength of competition (to induce a more spatially homogeneous density) leads to higher rates of simulation failure because of local Allee effects, you can avoid this by increasing the total population size. What is being shown in the figure are the individuals in a single lineage (the “unsimplified” tree), over 10 generations. The aim of this figure is to illustrate that both density-dependence and mate choice range (particularly the latter, in our case) strongly affect the distribution of individuals throughout the habitat. We suspect that the “clumping behaviour”, at small mate choice scales, is due to phenomena such as those described in Felsenstein 1975 for constant size populations - and we have added a paragraph describing this in the discussion (line 281): “ The mate choice radius caused distinctive patterns in the distribution of a population within its landscape. In particular, close-range mating led to clustered groups of individuals, which may be a practical nuisance to simulation users, and lead to unwanted geographic structure. We suggest that this is the same phenomenon described in \cite{felsenstein1975pain}. As Felsenstein describes, the intuition behind this behaviour is that, when either mate choice or dispersal distances are small, individuals each seed a \say{clump} of descendants. Due to the constraint of constant population size, several of these clumps are destined to die out. The small mating distance forbids mating between these clumps, so the remaining ones become larger and further apart. This is particularly cumbersome because relatively small mating distances are required for the average parent-offspring dispersal to match p1-offspring dispersal. Although not possible in the most recent version of \textit{slendr} (\cite{petr2022slendr}), allowing for less generally constrained simulations with fluctuating population size might alleviate these factors. However, this would require the development of dedicated software for the analysis of tree sequences produced by such dynamics (known as \say{non-Wright-Fisher} in \textit{slendr}."

Anthony Wilder Wohns

Ianni-Ravn, Petr, and Racimo have contributed a timely and insightful manuscript that seeks to disentangle the relationship between theoretical dispersal distributions, mate choice and spatial competition. The forward simulators SLiM and slendr have allowed researchers to perform complex spatial simulations, but have also highlighted gaps in our understanding of the influence of ecological parameters on spatial dynamics. The authors thus undertook a carefully designed simulation study to understand the interaction between these parameters.

The authors provide useful theoretical results with sensible parameterizations of parent-offspring dispersal distance functions. They explore the effects of the dispersal variance parameter in simulations, with expected results. Varying the scale of competition is observed to affect clustering, while mate choice boundaries modify dispersal distances. They also make the important observation that long branch lengths can lead to inference biases in the setting of finite habitat size.

The complexities of these problems are not to be underestimated, and I appreciate that the authors took care to explore a well-defined and approachable slice of the parameter space. There is clearly significant scope for more work in this area.

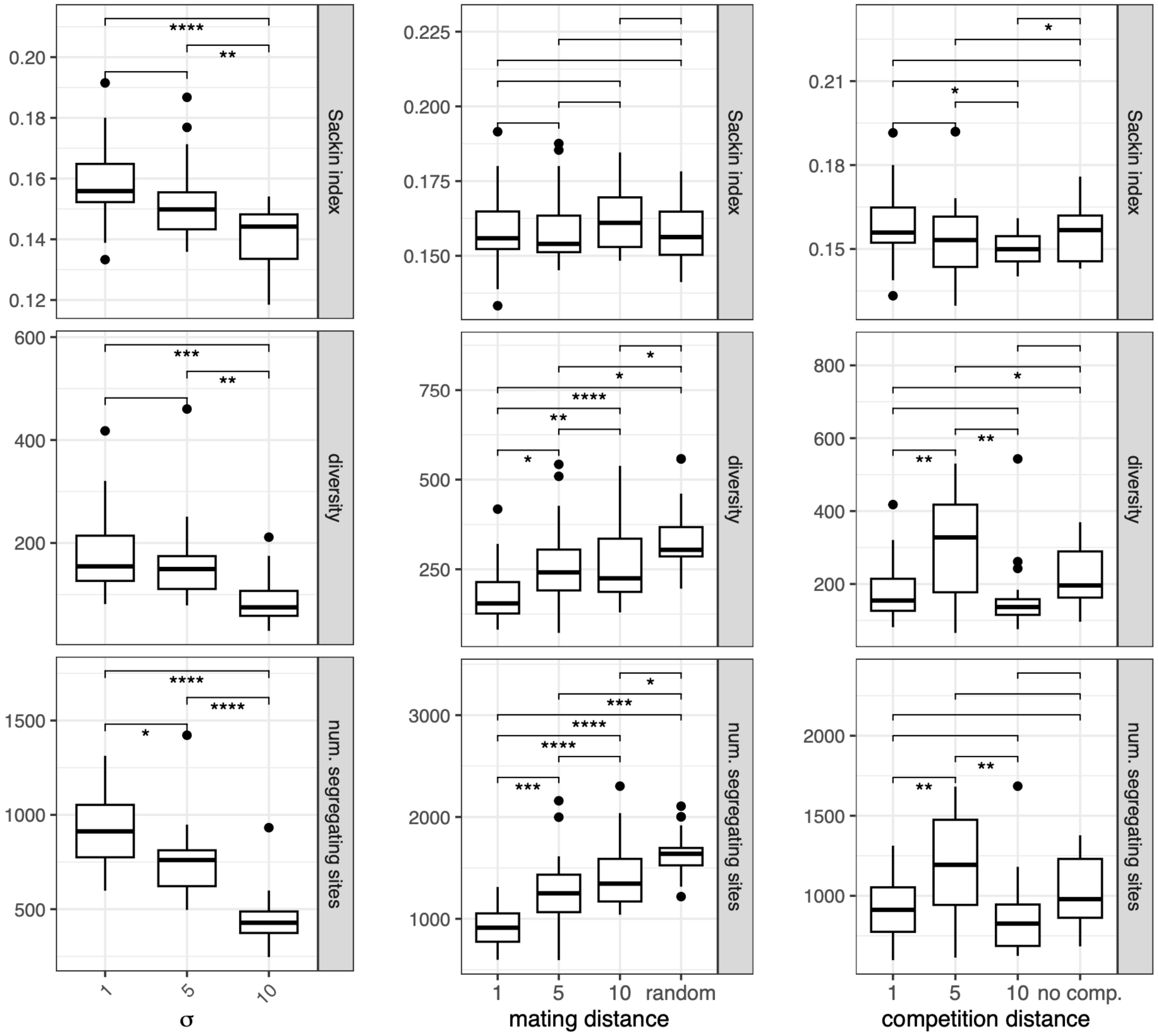
I have the following specific comments for the authors:

- ☐ - Section 5.1.2 states "We simulated a single locus in order to focus on fundamental geographic dynamics which act on single trees". It thus appears that all simulation work was performed on single locus trees, not ARGs or tree sequences. I would suggest the title be amended to reflect this salient point, as recombination can introduce complexities which are not explored in this work. The results are still quite relevant for understanding the spatiotemporal dynamics of recombining organisms, but it could be misleading for readers to assume that tree sequences, rather than single trees, are explored here. Thank you for this comment. We agree that this was a misleading statement. We have changed the title and added a paragraph to the introduction to address this point, in line 71: “Genome-wide tree sequences are an ideal object on which to perform phylogeographic inference, and are already beginning to be used for such analyses (for example, \cite{wohns2022unified}). Recent computational developments have made it tractable to approximately infer tree sequences for a given genome panel (\cite{kelleher2019inferring,speidel2019method,hubisz2020inference}). However, both estimating and working with full tree sequences comes with substantial computational burden. One approach to this problem, which has been used in recent work (\cite{osmond2021estimating}), is to \say{thin} the full sequence of trees covering the entire chromosome into a set of approximately independent genealogies. Although these genealogies do not wholly capture the complexity of the full tree sequence, we believe that the insights obtained from them are an important basis for understanding how spatial dispersal affects recombining genomes. ”

☐ - The abstract states the paper will examine how the parameters studied "influence the distance between closely- and distantly-related individuals in these genealogies". However, the results focus on the distribution of parent-offspring relations only. Similarly, the structure of the genealogies themselves are not a focus of the work, aside from the observation that long-range competition led to clustering. I would suggest the abstract be modified to better reflect the scope of the work performed. To approach this point, we have modified the introduction. We believe that the “distantly-related individuals” statement remains relevant because of the section on estimating dispersal distance, since the ML estimate of sigma can be seen as a parameter conveying the rate of ‘long-term diffusion’. To investigate different aspects of tree structure, we added a figure (Fig. S1, pasted below) showing the effect of our spatial parameters on statistics describing the genealogies. We computed the Sackin index (a measure of tree balance), as well as the average number of pairwise differences (Tajima’s estimator) and the number of segregating sites. Here is the relevant paragraph, starting at line 179: “We next examined how the dispersal, competition and mating distance affected a set of summary statistics for the genealogies (Fig. \ref{fig:7}). We computed Sackin's index, as well as two measures of diversity: the average number of pairwise differences (Tajima’s estimator of diversity) and the number of segregating sites for each of the trees, as described in Methods section \ref{treestats}.

The average number of pairwise differences decreased with the dispersal distance, and the number of segregating sites showed the same pattern. This suggests that limited dispersal range preserved diversity in the population, although it appears to be inconsistent with the well-known Wahlund effect, the decrease in diversity brought about by population structure (the Wahlund effect, \cite{wahlund1928zusammensetzung}). Interestingly, increasing the mating distance instead led to an increase in diversity and the number of segregating sites, which is instead in agreement with the Wahlund effect. The average number of pairwis and number of segregating sites showed no clear pattern with increasing competition distance.

Furthermore, the Sackin index exhibited a reduction with increasing dispersal distance, while it remained constant when altering mating and competition distances. Sackin's index, a measure of tree balance, is defined as the sum of the number of ancestors for each tip of a tree (\cite{lemant2022robust}). A higher Sackin index signifies a less balanced tree, indicating that certain clades tended to give rise to more descendants than others. Consequently, this pattern suggests that short-range dispersal introduced some imbalance into the branching structure of the genealogies. “



□ Furthermore, it would be helpful to highlight some of the concrete theoretical and empirical results of the work in the abstract, for instance the effect of competition and mate choice on the dispersal distances, and/or the importance of habitat size in inference. We have added a sentence to the abstract (lines 9-22), which addresses these results “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. We carry out a series of \textit{in silico} experiments using forwards-in-time simulations to determine how these parameters influence the distance between closely- and distantly-related individuals. We also assess the accuracy of the maximum likelihood estimation of the dispersal distance in a Gaussian model of dispersal from tree-sequence data, and highlight how it is affected by realistic factors such as finite habitat size and limited data. We find overall that the scale of mate choice in particular has marked patterns on short and long terms patterns of dispersal, as well as on the positions of individuals within a habitat. 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”

- ☐ - Regarding the introduction, it should be clarified that the work was not performed with ARGs/tree sequences, though it has important relevance to work in that area. Furthermore, the definitions and usage of the terms phylogenies, pedigrees, genealogical trees, ancestral recombination graphs, and tree sequences could be more precise. For instance, paragraph 5 of the introduction states "With some loss of information, an ARG can be further simplified into a sequence of trees along the genome". What specific information are the authors referring to, and how is this relevant? It's unclear from the text why the authors consider tree sequences to be theoretically more amenable for inference than ARGs. Additionally the brief description of tree sequences leaves the impression that they are simply a sequence of unrelated local genealogical trees. These subtleties are especially important since the work in this manuscript does not operate on tree sequences, so its relevance here should be more clearly argued.Our choice to deal with independent trees approximates the practise of thinning the ARG, and mitigates the complexity of dealing with correlated local genealogies. As the reviewer states, this was not clarified properly in the text, so we have lengthened the section leading from the description of the ARG to our choice to use single genealogies (see the section added as a response to the previous comment). We agree that this leaves much interesting work to be done, and will defer this aspect of spatial structure to future studies.
- ☐ - There is a minor error in the citation of Kelleher, Wong, et al. 2019., which should be Kelleher et al. 2019. Also, Wohns et al. 2021 refers to the preprinted version of a manuscript published in 2022. Thank you for spotting this - we have amended the citations.
- ☐ - Two very minor points on simulation size: I don't think this would change any results, but the tails of the simulated distance distribution in Figure 2 would appear cleaner with a few more replicates. This shouldn't be too onerous as it appears the simulations were not very computationally intensive. Also, using different icons for the replicates in figure 7 is not necessary as the data points are overplotted anyway. We have both re-made Figure 2 with double the number of replicates (indeed it does look cleaner), and

amended Figure 7 (now Figure 6) so that each simulation replicate has the same icon. Thank you for the suggestions.

□ - I believe there is an error in the reference to Fig. 7 in the last sentence of the results, it appears the analysis described in this paragraph should refer to Fig. 8? Yes, this is correct – thank you for catching this typo, which has now been amended.

□ - I greatly enjoyed reading the discussion in particular, which I thought had a number of very important and practical points. It might be helpful to outline other important, but currently underexplored questions. One is the possible consequences of using tree sequences rather than marginal trees in these analyses. Another is the biases which could be introduced by using inferred rather than simulated trees. These are very interesting points. We have added some paragraphs to the discussion addressing them, including reference to recent work on biases arising from estimating trees (Line 332): "This study has focused on single-locus genealogies, which is comparable to studying approximately independent genealogies from a tree sequence. Such an approach, followed for example in \cite{osmond2021estimating}, greatly reduces the computational burden of analysing the full tree sequence, yet retains the ability to uncover variation in dispersal and geographic ancestry across the genome. However, we expect that ignoring the correlation structure which exists between trees in a tree sequence leads to some loss of information --- specifically, in a fully annotated tree sequence, it is possible to identify nodes which are shared between trees. This information could be used, for example, to constrain the positions of shared internal nodes based on information coming from several trees. We note that, since \textit{SLiM} and \textit{slendr} are able to run spatial simulations of recombining genomes, they might be valuable tools to begin to investigate how much information is lost when we \say{thin} tree sequences.
Another aspect of complexity which we have not investigated is the bias which might arise from using estimated genealogies, rather than known ones. There is recent evidence that currently available methods (Argweaver, Relate and tsinfer + tsdate) tend to underestimate the time of deep coalescences, and vice versa (\cite{yc2022evaluation}). This is a form of a well-known phenomenon in phylogenetics called \say{long branch attraction}. We expect that would lead to biases in inferences of dispersal (longer-range than reality towards internal nodes, and shorter than expected at the tips). Again, this could be aptly studied in \textit{slendr} by \textit{post-hoc} adding mutations onto the simulated genealogies, and adding a genealogy estimation step to the analyses. "

- □ - Typo in line 295 "ecah". Thank you, this has been corrected.

It is my opinion that the technical aspects of this work are sound and that it will prove valuable to the growing number of researchers interested in spatial population genetics. I congratulate the authors on their work!

Anonymous Reviewer #2

Reviewed by anonymous reviewer, 20 May 2023 1ti:30

In this paper the authors examine the effect of three functions -- the dispersal kernel, the mate choice kernel, and the competition kernel (fig 1) -- on the distance a sampled genetic lineage moves from one generation to the next (fig 2, 3, 5, 6, hereafter "realized parent-offspring distances"), spatial clumping (fig 4), and maximum likelihood estimates of the dispersal rate (fig 7, 8). They do this via individual-based simulations (section 5.1) and math (section 5.3). One key finding is that while local competition tended to decrease the probability of small realized parent-offspring distance (fig 2) and affect spatial clumping (fig 4), it had a much smaller effect on the variance in realized offspring-parent distances than the mate choice kernel (fig 5), which dramatically affected the distribution of realized parent-offspring distances (fig 6) and the resulting dispersal estimates (fig 8). A second key finding is that habitat boundaries cause underestimates of the dispersal rate, but this can be corrected by ignoring long branches (fig 7). And a third key finding is that the shape of the dispersal kernel affects estimates of the dispersal rate (fig 8). Together, the authors show here that all three kernels affect spatial patterns in relatedness, which is important to keep in mind when interpreting estimates of "effective" dispersal rates.

- ☐ L58-64: You introduce the dispersal distribution, the idea that we can estimate a dispersal distance, and that the ARG might be useful. You then say that two recent methods use the ~ARG to locate genetic ancestors. The idea of locating ancestors seems to come out of nowhere to me, and is not examined in this paper. I think it would be more appropriate to here discuss that dispersal distance can be estimated from a tree (and tree sequences). If you want to discuss locating ancestors then a bit more context is needed, I think, to connect to the rest of the paper, especially since under a model of Brownian motion the most likely location of an ancestor doesn't actually depend on the dispersal distance (though the confidence interval does).

  We agree that this is likely to be confusing for the reader and out-of-scope for this work, and have removed the reference to locating ancestors.

□ L66-68: The comment "... little work has been done to assess how spatial parameters ... affect a tree sequence" is a little vague to me, making it hard to determine how much previous work has been done. (You could say something more specific instead, like the distance between parents and offspring in sample lineages and/or estimates of the dispersal distance from a tree.) While they don't use trees, Smith et al (https://www.biorxiv.org/content/10.1101/2022.08.25.50532tiv5) do look at the effect of mate and competition kernels on dispersal estimates (their supp fig 1), coming to the same conclusion that competition has a much weaker effect than mate choice (they use SLiM too). Thank you for this helpful reference - we have added it in the text, and have re-written the paragraph with clearer wording (line 90): “Two types of interactions which people often use to model populations in geographic space are mate choice and competition for resources. Both of these can be understood via a distance parameter. The mate choice distance controls the scale at which individuals tend to find each other to produce an offspring. The competition distance determines how far individuals can be separated for them to compete for resources. The effects of these parameters on dispersal and genetic diversity have not been the main focus on previous studies. However, there is some evidence from simulations that the scale of mating has more impact on effective dispersal than that of competition (\cite{smith2023dispersal}). The lack of work in this area is particularly troublesome for any users of forwards-in-time simulators such as SLiM, where they are required to specify these dynamics explicitly. ”

□ L81: Why did you choose to use slendr and not just SLiM? My understanding is that slendr is a great help when working with complex spatial habitats (eg, real world maps) and population structure (eg, population splits and admixture), but here the simulations are a homogenous square with a single population. Wouldn't using SLiM directly make it easier to vary kernels, like mate choice (whereas it sounds like slendr forces this to be uniform, L113-114)?. I'm guessing the answer has something to do with sf (section 5.1.3)? Indeed, the main reason for using *slendr* it in this manuscript was its built-in support for analyzing spatially-annotated tree sequence data in the R environment, as it not only exposes all *tskit* tree sequence tables and individual phylogenetic tree objects as native R objects but automatically converts all relevant metadata in the spatial *sf* data format. We believe that this flexibility of data analysis and statistics outweighed the limitations of *slendr* in terms of specifying arbitrary spatial interaction kernels. As *slendr*’s support for arbitrary spatial tree sequences matures (and will eventually fully support even non-slendr “pure” SLiM tree sequences), future studies will certainly be able to use it to discover more general properties of parent-offspring distributions in more flexible dispersal and mating scenarios.

- □ L104-106: "... the faster the population spreads across the landscape" makes it sound like you are simulating a range expansion. I don't think you give any information about the initial conditions of the simulation, but my guess is that the first generation is placed uniformly at random, in which case this isn't an expansion. Perhaps, "... the faster genetic lineages spread across the landscape" would be more appropriate? Yes, this more precisely articulates the process, and we have corrected the line accordingly.
- □ L117: "... sampled all individuals and reconstructed the tree connecting them" makes it sound like you inferred the tree that relates individual to one another. I would have thought instead you simply saved the tskit tree sequence recorded by SLiM. If this is correct, there is also some imprecision in the word "them" (referring to individuals, which I presume are diploid), since a tree sequence connects haploid genomes (in this case alleles). The relations between the individuals can be described by a pedigree, but that is not generally a tree. Some clarification needed I think. We apologise that the description was unclear. We were actually describing the “unsimplified” genealogy - that is to say, a genealogy containing all nodes along each edge of the tree from generation to generation (in addition to the coalescent nodes). The tips of this genealogy correspond to all 2N genomes of the N sampled individuals. So, this is not a pedigree, since individuals which were not the most recent common ancestors of some of the sampled genomes are not recorded. We have clarified this in text: “After 50 generations, we sampled all individuals and reconstructed the genealogy connecting them. In these genealogies, we stored all individuals, rather than coalescent nodes only (this corresponds to a tskit \say{unsimplified} tree), so that we could observe dispersal at every generation.”
- □ L132-137: You give a specific value of the competition distance that you say reduces clumping and scattering and excess variance in dispersal distances (0.2) but then say this is not a general finding. Maybe remove the emphasis on the exact value (0.2) and instead refer to "an intermediate" value of competition distance? Thank you - we agree that emphasizing a particular value is confusing, so we have clarified that this is a function of our parameters, rather than a general finding.
- □ L12ti-137: I'm not sure I see the changes in clumping with competition distance that you describe in Fig 4. To me it looks like competition distance has very little impact on clumping and instead it is mating distance that determines how many clumps you can

have (and how connected they are). I think this might be connected to what Felsenstein (1975 Am Nat) showed, that you get clusters when you impose a constant population size (his eqn 13). This is long-range competition and is acting in all of your (Wright-Fisher) simulations, regardless of the competition distance. One way to avoid clumping, then, is to simulate a non-Wright-Fisher population with population size controlled dynamically by local competition. I'm not suggesting you re-do the simulations, I'm just looking for a clearer explanation of the patterns in fig 4. The fact that competition does not affect clumping as much as mate choice does may also help explain why competition has much less effect on realized parent-offspring distances and effective dispersal estimates? Indeed, it is mate choice radius which has a stronger effect on clumping, and we agree with you that this is related to the phenomenon described in Felsenstein 1975. We have amended the caption of figure 4 to reflect this. We have also worked to rephrase our descriptions of this clumping behaviour in text. Line 281: “ The mate choice radius caused distinctive patterns in the distribution of a population within its landscape. In particular, close-range mating led to clustered groups of individuals, which may be a practical nuisance to simulation users, and lead to unwanted geographic structure. We suggest that this is the same phenomenon described in \cite{felsenstein1975pain}. As Felsenstein describes, the intuition behind this behaviour is that, when either mate choice or dispersal distances are small, individuals each seed a \say{clump} of descendants. Due to the constraint of constant population size, several of these clumps are destined to die out. The small mating distance forbids mating between these clumps, so the remaining ones become larger and further apart.”

□ Fig 2: It would have helped me if you stated what "simulated distances" means right here in the caption (this is a general comment, since many readers will want to look at the figures and caption without having to repeatedly read the main text). Naively one might think these distances are just draws from the chosen theoretical distribution (given in the right panel), but I think they are instead the distances each sampled genetic lineage moved from one generation to the next, which is influenced by both competition and mate choice. This is correct, and we agree that the caption should be more informative. We have amended the caption of figure 2 to read: “The left panel shows the empirical distribution of effective parent-offspring distances drawn from the forwards-in-time spatial simulations, while the right one shows the PDF of the corresponding dispersal distributions. The effective distances are affected by the dispersal distribution, as well as competition and mate choice. Bottom: a zoom-in on the tails of curves; 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. The mating distance competition distances were both 1 unit.“

□ L161-163: As a side note, the fact that the math works out easier with Gaussians is one of the reasons (the other being biological realism) that I'm surprised slendr uses uniform mate choice and comeptition kernels. But I guess there is an argument to be made for computational speed? In developing the first version of slendr, its author had to strike a balance between how flexible its R interface to the underlying SLiM engine will be and how complex its R function API and its internal SLiM codebase need to be to support this flexibility. This particular limitation is present not so much for computational speed but for purely software engineering reason -- in other words, the line for flexibility (and complexity) had to be drawn somewhere, and this it was drawn in the first published slendr version. Again, as the package matures, future updates will very likely relax some of these limitations in a move from being a more of an aid in methods development (where slendr is now) to a higher biological realism.

□ L167 (and L16ti): I think the \pi should be \sqrt(\pi). Thank you for spotting this. You are correct, and we have amended this in text.

□ Eqn 4: I think there is an error in this equation: shouldn't the sum be of d_i^2/l_i, which gives the appropriate units of distance/sqrt(time)? This can be derived by setting \tau=0 in Eqn 3, replacing \sigma^2 with l_i\sigma^2 (ie, the variance increases linearly with time under Brownian motion), taking the product of g_{y_i}(y_i) over i from 1 to N, differentiating with respect to \sigma, setting the derivative equal to zero, and solving for \sigma. It looks like estimation.R uses the correct equation, so I think this is just a typo (and explains why the estimates do well in the figures). Again, you are correct - this was a typo, the correct summand is d_i^2/l_i.

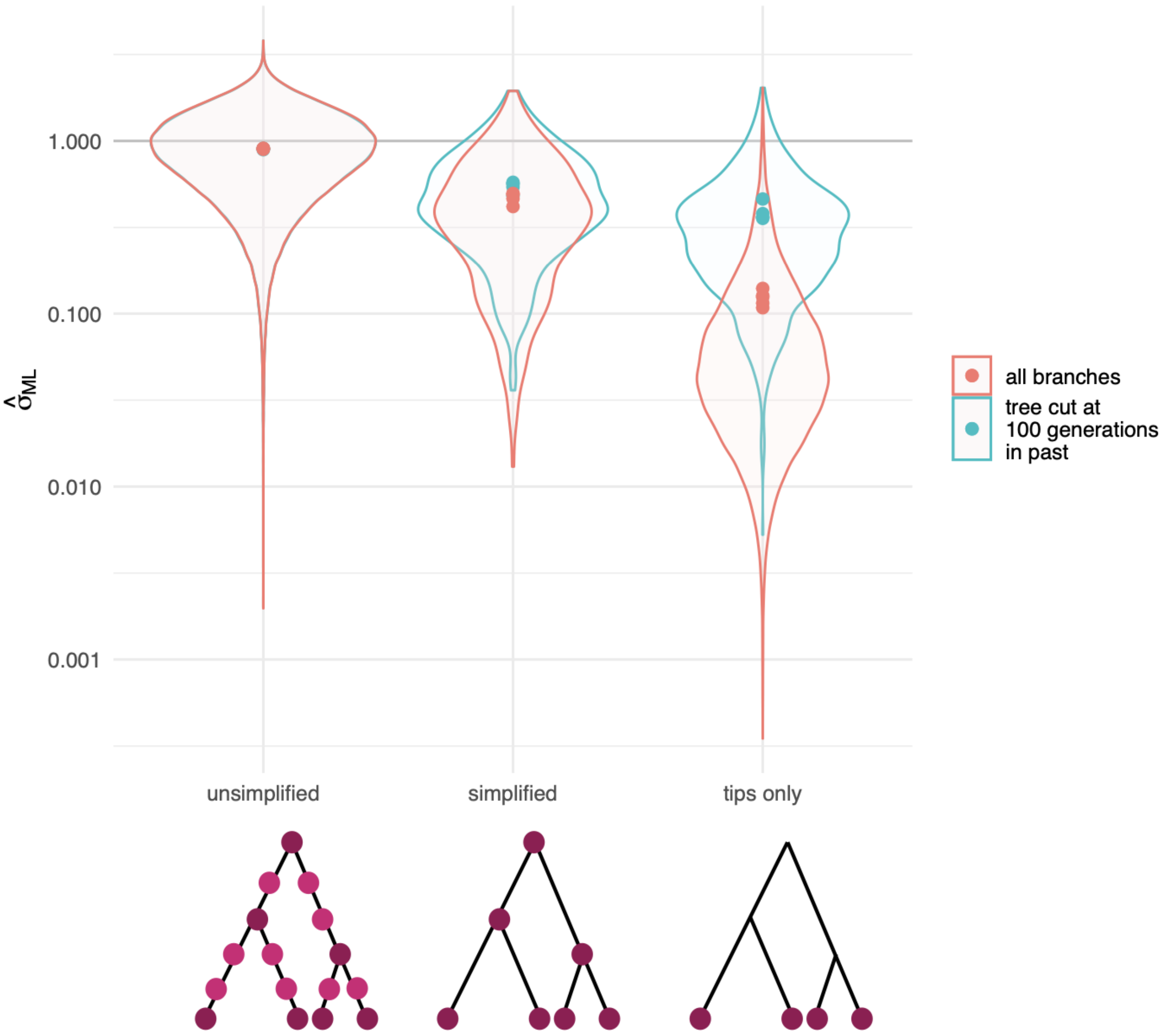
□ L174: Is this still a populatio of size 50 (ie, you are sampling every haploid genome)? estimation.R makes me think N=2000 here, which might be good to state in the main text. Indeed, we neglected to refer to this in the main text - the population size was 2000, and we have now added this to the corresponding paragraph.

□ L178-180: It might be helpful to state what l_i is for the tips-only case, since there are multiple branches involved. I think it is 2TMRCA? This would also help clarify what it means to 'cut' branches in the tips only case. One issue here is that, while the distance between two samples follows a Rayleigh distribution, that distance is not necessarily

independent of the distance between another pair of samples (if there are shared branches between the two pairs). This means that Eqn 4 maximizes the composite likelihood. However I think it also maximizes the true likelihood in this case, since the covariance between the distances from multiple pairs of samples is a nuisance parameter? I think this is one place in the paper where it would have helped me to see the connections between your work (in terms of absolute distances between pairs of nodes, which is Rayleigh when \tau=0) and previous approaches that instead model the locations of all nodes as a multivariate normal (eg, Osmond and Coop 2021, and the equivalent method of Felsenstein 1985). I guess they give equivalent dispersal estimates but the multivariate approach appropriately deals with covariance, and hence gives the correct likelihood function? Yes, we believe that this is correct, and we expect that the Osmond & Coop method deals better with covariance – whereas we treat pairs as independent. We have added more detail to line 220, stating this explicitly: “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 \say{Brownian}), and negligible mate choice distance, the maximum likelihood estimator of σ is
\begin{align}
 \hat{\sigma}_{ML}=\sqrt{\frac{1}{2N}\sum_{i=1}^N \frac{d_i^2}{l_i} }
\end{align}
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 \cref{math2}). It is worth noting that our method of estimation is na\"ive, since it ignores the fact that branches are shared between pairs containing the same individual --- indeed, we actually maximise the composite likelihood, rather than the full likelihood (as instead is used in \cite{osmond2021estimating}, where covariance between pairs is appropriately taken into account). However, with sufficient data, the maximum likelihood estimate of σ should be the same in both cases, and we use this as a simple bench-mark approach.”

We also clarified the connection between the pruning procedure and the tMRCA: “To investigate whether limited world size was responsible for this observation, we adopted the approach detailed in \cite{osmond2021estimating} and eliminated branches which were more than 100 generations old. In the \say{\textit{simplified}} and \say{\textit{tips only}} case, this amounted to retaining sub-trees for which the tMRCA lived less than 100 generations in the past.”

- □ L182-185: I would say "adopted an approach similar to" or "inspired by" Osmond and Coop 2021, as they didn't explicitly ignore long branches but instead cut the entire tree off at some time in the past. I guess the methods do something quite similar when using tips-only (ignoring pairs of samples with 2TMRCA above a cut-off) but reasonably different when using a simplified tree (Osmond and Coop removed old branches while here you remove long branches). Thank you for this helpful comment. We have amended our approach to match that of Osmond and Coop (excluding *old* rather than *long* branches). We agree that this is more intuitive. The results are similar - the unsimplified tree is barefly affected, while the estimates from both the simplified and tips only case increase, due to the fact that the older branches are indeed the longer ones, and those that are most biased by limited habitat size.



- ☐ L191: Fig 7 -> Fig 8 Indeed there was some figure mis-numbering - this is related to the next comment.
- ☐ L203: I would like to hear why you think the estimated dispersal distance is more sensitive to mating distance with a Cauchy dispersal distribution. This was in reference to Figure 8 - which was previously mis-labeled in line 203 as Figure 7. Here, we are referring to the fact that the increase in the estimate of sigma, with mating distance, was more stark when the DF was Cauchy – see the figure pasted below. We believe this is likely due to the estimator under the Brownian model being inappropriate in the Cauchy

case.

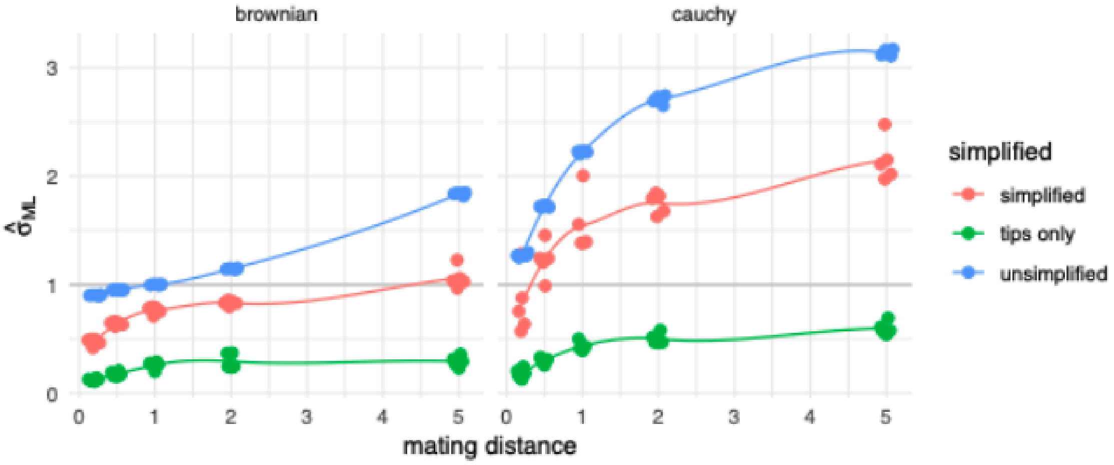
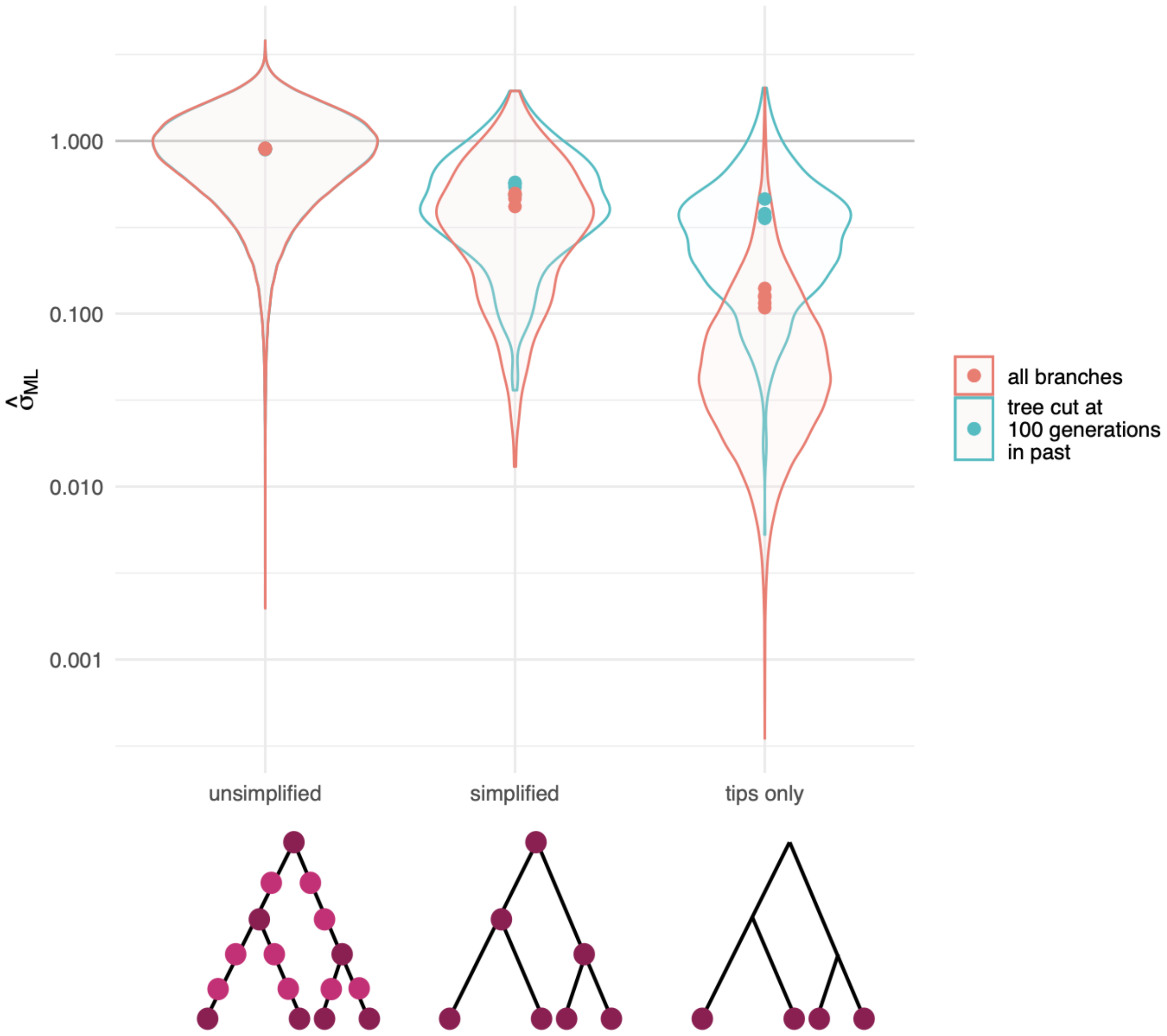


Figure 8: **Estimating the mean dispersal distance under the Brownian mode of dispersal.** Left: increasing the mating distance increases the effective dispersal distance. Right: in these simulations, the dispersal function was *Cauchy* with scale and location 1, but we naively used maximum likelihood 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.

□ Fig 7: I think it would be more intutive to order the figure, from left to right: unsimplified -> simplified -> tips only. This is the order discussed in the text, goes from most to least data, and shows a more or less linear decline in (naive) estimates of dispersal distance.

We agree that this is a helpful amendment - and we have reworked the figure to reflect this and other comments:



□ L208-210: "... altering the kernel of parent-offspring dispersal can have profound effects on the diffusion captured within a genealogy ...". I think the word "profound" is a little strong for a large but very expected result. We agree that this was an enthusiastic wording choice, and have altered the word to “stark”.

□ L217-220: Do you have a dispersal distribution in mind where the tail is independent of the variance? The distribution we had in mind was the Pareto distribution, where the tail can be controlled via the shape parameter - in contrast to, for example, to the normal distribution, where the tail behaviour is directly influenced by the variance. However, the wording “independent” is inappropriate, and we have added more detail to this sentence:

“We suggest that a pragmatic solution for the sake of simulation might be to encode a dispersal distribution where the height of the tail may be controlled more directly. An example may be the Pareto distribution, where the tail probability is particularly sensitive to the shape parameter, and does not directly depend on the variance (in contrast to, for example, the normal distribution).”.

- □ L248-251: The Smith et al paper I mentioned earlier calculates a similar effective dispersal distance (see the simulation section of their methods), and use it to train their machine learning algorithm to estimate the true dispersal distance, which might be interesting to compare to. Indeed, Smith et al. emphasize that they are setting out to estimate this effective dispersal, and clearly explain the meaning of this parameter. This is very related to one of the main points of this manuscript, the relationship between effective dispersal and parent-offspring dispersal. However, we have chosen not to compare their results to ours in this paper - both for the sake of brevity and since disperseNN is geared towards estimating dispersal from genotype data, which by-passes the tree-based focus we have here.
- □ L249: I -> we Thank you, this has been corrected.
- □ L253-256: I like this point, that a non-zero mating distance will make the distribution of distances non-Rayleigh (or, as I tend to think about it, the signed distances will be non-Gaussian), making parameter estimation more complicated, but over long time periods the central limit theorem will cause the distances to become Gaussian (as will the signed distances) with a readily calculable effective dispersal distance.
- □ L295: ecah -> each Thank you, this has been corrected.
- □ L296: randomly at each generation -> randomly Thank you, this has been corrected.
- □ L298: these -> populations Thank you, this has been corrected.
- □ L309: are -> is Thank you, this has been corrected.
- □ Fig 9: "(c) Each panel shows a slice of 250 generations" -- do you mean each color? Thank you for pointing this out - the caption referred to an earlier version of the figure. We have decided to remove Figure 9, since it does not directly illustrate the simulation techniques, or results.
- □ L323: grammar not right here. Thank you, this has been corrected.
- □ Eqn 10: Why is the first factor not 1/(2\pi)? I thought the angle was uniformly distributed over [0,2\pi] -- did a 2 get cancelled from elsewhere? The angle ranges from 0 to \pi

because angles of x and x-2\pi are equivalent in this model, which only ‘cares’ about absolute distance. Another way to say this is that the factor of 2 cancels out.

- □ L386: In fact, you can get any moment of y from that expression. Thank you, we have pointed this out in the text, below the relevant expression.
- □ Eqn 12: I think the f_y term should be multiplied by y as well. This is correct, and has been amended.
- □ Eqn 12: r_y is undefined. Thank you for pointing out these issues, we have amended equation 12 so that r_y (a carry-over from a previous version) has been removed and both terms are, as you say above, multiplied by y.
- □ L404: I'm confused by this 1-branch estimate of sigma. Why use a moment-based estimator to find out how much a branch contributes to the max likelihood estimator? Wouldn't it be more appropriate to set N=1 in Eqn 15, giving the max likelihood estimate d/\sqrt(2l)? Or better yet, if you look for the max likelihood estimate of \sigma^2, which is unbiased for a Rayleigh (Wikipedia), then d^2/(2Nl) is precisely how much each branch contributes to the overall max likelihood estimate. I believe the confusion here is related to a typo in equation 4, where we had mistakenly placed a ^2 in the summand (the term should be d_i^2/l_i instead of (di/l_i)^2) and carried forward the typo below. In the script estimation.R (https://github.com/mkiravn/treesinspace/blob/new_treesinspace/new_scripts/estimation.R), the correct formula is used for both. This means that the single-branch estimate is d/\sqrt(l * pi/2), which is the MOM estimator of sigma for a single branch.

Christian Huber

The manuscript by Ianni-Ravn et al. provides a simulation study investigating the effect of spatial demographic parameters (dispersal distance of offspring from the gestating parent, competition distance, mate choice distance) on the distribution of observed dispersal distance. They also assess the accuracy of an estimator of dispersal distance from genetic data, showing potentially relevant insights for empirical studies.
The study provides a new framework for how to think about dispersal that is including mating distance. I enjoyed reading the manuscript, the results are clearly described and easy to follow. I have a few comments that might lead to beCer correspondence with real biological systems.

Main comments:

- ☐ - I would appreciate a more detailed description of the simulation setup. In particular, it seems that the simulation is conditional on a constant population size of 50 individuals. I assume parents are selected at random, but weighted by their fitness? Yes, this is correct, and has been added to line 395 ("Individuals were chosen randomly, weighted by their fitness, to be the parents of the next generation."). How exactly is fitness down-scaled as a function of the number of competing individuals (i.e. provide the formula)? The fitness is scaled by 1/n where n is the number of individuals within a radius of the competition distance. How is it guaranteed that mating individuals are within the mating distance? E.g. if an individual is too far away from any potential mate, does it effectively have zero fitness? Would fitness increase with increasing number of potential mating partners? Insofar as an individual with zero neighbours within the mating distance would effectively have fitness zero, yes. We think that this is part of the reason for the clustering behaviour seen with small mate choice radius (see response to the next comment). Full details are available in sections 15.2 and 15.4 of the SLiM manual, (http://benhaller.com/slim/SLiM_Manual.pdf). These details should be added to section 5.1 (Spatial simulations). Thank you for highlighting the need for clarification in this section. We have lengthened section 5.1 to better describe the simulation set-up, and added these details.

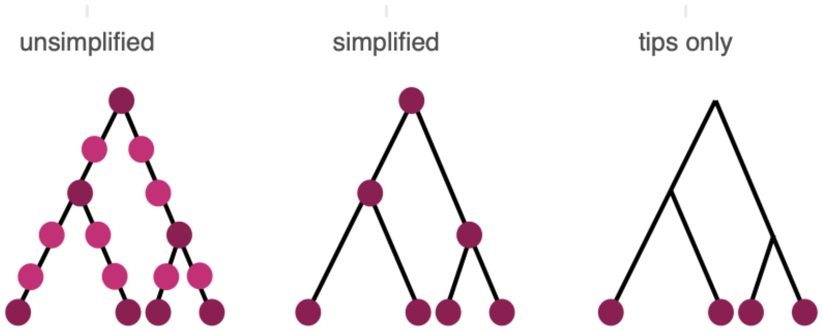
- ☐ - This type of relative fitness approach where the population size stays constant might not be very realistic (in reality population size fluctuates due to local competition and resources, there is no globally regulated population size). I wonder if it would be feasible to explore simulations that don't have this constraint (i.e. nonWF models in SLiM)? Yes, we agree that nonWF set-ups might be more realistic in terms of these factors. However, nonWF simulations come with their own fleet of extra parameters which need to be surveyed - and we believed a reductive approach was a good place to start. We have added a line commenting on this in the discussion (line 281): " The mate choice radius caused distinctive patterns in the distribution of a population within its landscape. In particular, close-range mating led to clustered groups of individuals, which may be a practical nuisance to simulation users, and lead to unwanted geographic structure. We suggest that this is the same phenomenon described in \cite{felsenstein1975pain}. As Felsenstein describes, the intuition behind this behaviour is that, when either mate choice or dispersal distances are small, individuals each seed a \say{clump} of descendants. Due to the constraint of constant population size, several of these clumps are destined to die out. The small mating distance forbids mating between these clumps, so the remaining ones become larger and further apart. This is particularly cumbersome because relatively small mating distances are required for the average parent-offspring dispersal to match p1-offspring dispersal. Although not possible in the most recent version of \textit{slendr} (\cite{petr2022slendr}), allowing for less generally constrained simulations with fluctuating population size might alleviate these factors. However, this would require the development of dedicated software for the analysis of tree sequences produced by such dynamics (known as \say{non-Wright-Fisher} in \textit{slendr}."
- ☐ - This also relates to the occurrence of clusters in the simulations with small mating distances - is there any intuition on why these clusters appear? Yes, we believe this is related to the phenomenon described by Felsenstein (1975 Am Nat), which he dubbed "the pain in the torus". Please see the text in the previous reply.
- ☐ - One assumption that is made is that the offspring's initial position is at the gestating parent's position. In reality, parents at some point move to the same location for mating and potentially for raising offspring together. I.e. it could be that the male moves to the female, the female moves to the male, or the offspring is raised somewhere between the original male and female parent location. I don't think this invalidates the simulations or any of the conclusions, but I think it would help if it is discussed more explicitly how these situations relate to the simulations or any of the results. Thank you for raising

these concerns. The first scenarios you mention (male moving to female, or female moving to male), are equivalent to the set-up in our simulations, since individuals are hermaphroditic. However, these differ a little from somewhere-in-between scenario. To clarify this, we have added a paragraph to the Methods section 5.1 (line 396): “Pairs of mates were chosen within a radius of the mating distance, with uniform probability. Within each of these pairs, one parent at random was set to be $p1$, which is sometimes called the \say{gestating parent}. However, note that this is purely a label --- it may also be that $p2$, whether it be the mother or the father, migrates to $p1$'s position to raise the offspring.

In this set-up, the location at which individuals mate is also that at which their fitness is evaluated. These are the coordinates recorded in our simulations. This means that $p1-o$ displacement can be seen as the net of parents moving to have the offspring, and the migration over the offspring’s lifetime from their birthplace to their mating location. “

Minor comments:

- ☐ - Commonly used terms should be used if possible. E.g. Coalescent tree instead of "simplified tree" or phylogeny. It was not clear to me if the "unsimplified tree" is the pedigree or if it just includes genetic ancestors at a locus. Thank you for highlighting this issue - we agree that the distinction between these is a confusing point. In terms of its topological structure, an unsimplified tree is equivalent to a coalescent tree, except that it includes genetic ancestors that are not themselves coalescent nodes. In the extreme case, an unsimplified tree can include every single genetic ancestor along every branch of that tree which are recorded as internal nodes along each branch - this is what we record. The distinction between ‘simplified’ and ‘unsimplified’ is important in section 3.3, so we have added a schematic of this concept to the Figure, which is now Figure 8.



Also, it should be stated more explicitly that a single locus is investigated, i.e. there is no recombination. We have added more references to the fact that we followed a single locus, throughout the text.

☐ - Please provide the derivation of the formula for the maximum likelihood estimate of sigma, or provide a citation. We have now provided a derivation of equation 14.

☐ - Figure 9: The labels (a, b, c) are missing. In the caption for (c) it says that "each panel shows a slice of 250 generations" -- I don't see what this refers to, there is just a single panel. We have decided that Figure 9 added relatively little to the text, so we have removed it.