Peer Community In
Evolutionary Biology

A new powerful tool to easily encode the geo-spatial dimension in population genetics simulations

Emiliano Trucchi *based on peer reviews by* ***Liisa Loog*** *and 2 anonymous reviewers*

Martin Petr, Benjamin C. Haller, Peter L. Ralph, Fernando Racimo (2023) slendr: a framework for spatio-temporal population genomic simulations on geographic landscapes. bioRxiv, ver. 5, peer-reviewed and recommended by Peer Community in Evolutionary Biology. https://doi.org/10.1101/2022.03.20.485041

Submitted: 14 September 2022, Recommended: 30 May 2023

Cite this recommendation as:

Trucchi, E. (2023) A new powerful tool to easily encode the geo-spatial dimension in population genetics simulations. *Peer Community in Evolutionary Biology*, 100630. 10.24072/pci.evolbiol.100630

Published: 30 May 2023

Models explaining the evolutionary processes operating in living beings are often impossible to test in the real world. This is mainly because of the long time (i.e., the number of generations) which is necessary for evolution to unfold. In addition, any such experiment would require a large number of individuals and, more importantly, many replicates to account for the inherent variance of the evolutionary processes under investigation. Only organisms with fast generation times and favourable rearing conditions can be used to explicitly test for specific evolutionary hypotheses.

Computer simulations have filled this gap, revolutionising experimental testing in evolutionary biology by integrating genetic models into complex population dynamics, which can be run for (potentially) any length of time. Without going into an extensive description of the many available approaches for population genetics simulations (an exhaustive review can be found in Hoban et al 2012), three main aspects are, in my opinion, important for categorising and choosing one simulation approach over another. The first concerns the basic distinction between coalescent-based and individual-based simulators: the former being an efficient approach, which simulates back in time the coalescence events of a sample of homologous DNA fragments, while the latter is a more computationally intensive approach where all of the individuals (and their underlying genetic/genomic features) in the population are simulated forward-in-time, generation after generation. The second aspect concerns the simulation of natural selection. Although natural selection can be integrated into backward-in-time simulations, it is more realistically implemented as individual-based fitness in forward-in-time simulators. The third point, which has been often overlooked in evolutionary simulations, is about the possibility to design a simulation scenario where individuals and populations can exploit a physical (geographical) space.

Amongst the coalescent-based simulators, SPLATCHE (Currat et al 2004), and its derivatives, is one of the few simulation tools deploying the coalescence process in sub-demes which are all connected by migration, thus getting as close as possible to a spatially-explicit population. On the other hand, individual-based simulators, whose development followed the increasing power of computational machines, offer a great opportunity to include spatio-temporal dynamics within a genomic simulation model. One of the most realistic and efficient individual-based forward-in-time simulators available is SLiM (Haller and Messer 2017), which allows users to implement simulations in arbitrarily complex spaces. Here, the more challenging part is encoding the spatially-explicit scenarios using the SLiM-specific EIDOS language.

The new R package *slendr* (Petr et al 2022) offers a practical solution to this issue. By wrapping different tools into a well-known scripting language, *slendr* allows the design of spatiotemporal simulation scenarios which can be directly executed in the individual-based SLiM simulator, and the output stored with modern tree-sequence analysis tools (tskit; Kellerer et al 2018). Alternatively, simulations of non-spatial models can be run using a coalescent-based algorithm (msprime; Baumdicker et al 2022). The main advantage of *slendr* is that the whole simulative experiment can be performed entirely in the R environment, taking advantage of the many libraries available for geospatial and genomic data analysis, statistics, and visualisation. The open-source nature of this package, whose main aim is to make complex population genomics modelling more accessible, and the vibrant community of SLiM and tskit users will very likely make *slendr* widely used amongst the molecular ecology and evolutionary biology communities.

Slendr handles real Earth cartographic data where users can design realistic demographic processes which characterise natural populations (i.e., expansions, displacement of large populations, interactions among populations, migrations, population splits, etc.) by changing spatial population boundaries across time and space. All in all, *slendr* is a very flexible and scalable framework to test the accuracy of spatial models, hypotheses about demography and selection, and interactions between organisms across space and time.

References:

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., ... & Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. Genetics, 220(3), iyab229. https://doi.org/10.1093/genetics/iyab229

Currat, M., Ray, N., & Excoffier, L. (2004). SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. Molecular Ecology Notes, 4(1), 139-142. https://doi.org/10.1046/j.1471-8286.2003.00582.x

Haller, B. C., & Messer, P. W. (2017). SLiM 2: flexible, interactive forward genetic simulations. Molecular biology and evolution, 34(1), 230-240. https://doi.org/10.1093/molbev/msw211

Hoban, S., Bertorelle, G., & Gaggiotti, O. E. (2012). Computer simulations: tools for population and evolutionary genetics. Nature Reviews Genetics, 13(2), 110-122. https://doi.org/10.1038/nrg3130

Kelleher, J., Thornton, K. R., Ashander, J., & Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. PLoS computational biology, 14(11), e1006581. https://doi.org/10.1371/journal.pcbi.1006581

Petr, M., Haller, B. C., Ralph, P. L., & Racimo, F. (2023). slendr: a framework for spatio-temporal population genomic simulations on geographic landscapes. bioRxiv, 2022.03.20.485041, ver. 5 peer-reviewed and recommended by Peer Community in Evolutionary Biology. https://doi.org/10.1101/2022.03.20.485041

Reviews

Evaluation round #2

DOI or URL of the preprint: https://doi.org/10.1101/2022.03.20.485041
Version of the preprint: 4

Authors' reply, 15 May 2023

Dear Dr. Trucchi,

Thank you for your positive assessment of our revised manuscript. We greatly appreciate the time and effort you and the reviewers have dedicated to examining and improving our work since the initial submission.

We have submitted the final version of the manuscript which now includes two new references suggested by one reviewer (change highlighted on page 21 of the tracked changes document) and further clarifies the potential confusion of *slendr* simulating one species but multiple populations of that species pointed out by another reviewer (change highlighted on page 22).

We are looking forward to contributing to the wonderful PCI initiative through our work.

Best regards,

Martin Petr and co-authors

Download tracked changes file

Decision by Emiliano Trucchi, posted 28 March 2023, validated 29 March 2023

Final minor changes

Dear authors,

Two reviewers are fully satisfied with the new version of the manuscript taking into account their comments. Unfortunately, I could not get a second evaluation from the third reviewer but I positively assessed your replies to her/his comments.

I am therefore very happy to recommend this preprint in PCI Evolutionary Biology.

I am attaching two minor comments from the reviewers you may want to consider while submitting the final version of your manuscript. In the meantime, I will prepare my recommendation letter.

Thanks for submitting your preprint to PCI Evol Biol!

Best regards,

Emiliano Trucchi

Reviewed by anonymous reviewer 1, 23 March 2023

I am pleased with the changes made by the authors as they take into account my comments and those of the other reviewers and greatly clarify what is currently available with slendr and what is not yet available. In addition, the calculation times mentioned now allow potential users to get a better idea of whether their own project can be realized with slendr.

There is one last minor remark that the authors could take into account in the final version:

- In the last paragraph of line 2 (last line), it says that the SPLATCHE program only considers 2 coexisting ”populations”, but on page 22 it says ”species” to describe the same thing: *”At the moment, slendr can only produce genome sequences from a single ”species” (although with an arbitrary number and spatial arrangement of population groups)...”*. So for consistency and clarity, I would replace ”population” with ”species” on page 2, which would thus read "*...allows simulation of no more than two species co-existing at a time (divided in an arbitrary number of demes) ,...*"

I am therefore in favour of publishing this version of the article.

Reviewed by Liisa Loog, 21 March 2023

The authors have addressed my all my comments and concerns. I am happy to recommend this work for publication.

As for literature on best inferential practices, publications on this topic are in short supply but the following works also come into mind: Beaumont 2010, Annu Rev Ecol Evol Syst (doi.org/10.1146/annurev-ecolsys-102209-144621) and Gerbault et al. 2014, PNAS (doi.org/10.1073/pnas.14004251)

Evaluation round #1

DOI or URL of the preprint: `https://doi.org/10.1101/2022.03.20.485041`

Version of the preprint: 3

Authors' reply, 17 February 2023

Download author's reply

Download tracked changes file

Decision by Emiliano Trucchi, posted 02 November 2022, validated 02 November 2022

Revision recommended

Dear Dr Petr and co-authors,

Three reviewers have now read and commented on your manuscript.

All of them found your work very interesting, scientifically sound and well presented.

They also provide some comments which I think could further improve your work.

Therefore, I recommend to carefully consider their suggestions and submit a revised version of your manuscript alongside a point-by-point rebuttal letter.

Best regards,

Emiliano Trucchi

Reviewed by anonymous reviewer 1, 20 October 2022

This article by Petr et al presents a new R package called slendr, which is wrapper aiming at facilitating the simulation of genomic data distributed in space and time. This simulated data can then be used to make inferences by comparing it to observed data. The package is divided in three parts, which are intended to be used one after the other but can also be used independently. The first part allows the user to design a spatiotemporal simulation framework of population dynamics and genomic diversity. The second part can be used to call two different already existing simulators (SLiM or msprime) based on the virtual world created during the first part. These two simulators have different characteristics, including individual-based vs coalescent-based. The third part allows to analyze data generated during the second part by directly computing statistics or outputting files to be analyzed with other analytical programs.

I fully agree with the authors that the spatial dimension of population genomics is important and often neglected or considered in a simplistic way, due to lack of available tools. From this point of view, I see the interest of the R wrapper developed by the authors, whose aim is to facilitate the use of approaches that consider the spatiotemporal dynamics of populations when studying genomics data. The main strength of slendr is that it encapsulates in the commonly used R language a whole series of programs written in different programing languages. Using R is therefore the only requirement for slendr users, without having to know/learn other languages.Overall, I find that the manuscript and accompanying documentation are well written and clearly present the use of slendr.

In my opinion, there are several points that need to be improved before publication:

- I can see the value of slendr and the possible future developments, but these developments are not trivial and I think the authors should distinguish even more clearly between what is currently feasible with slendr and what is for future developments, perhaps with separate sections. For instance, the authors cite different approaches or programs currently available and their limitations, which they aim to overcome. However, some of these improvements have not yet been achieved. For example, they cite in introduction the program SPLATCHE as being limited to two interacting populations, but the current version of slendr only allows the simulation of one population (!) as I understood it in reading the discussion. In this respect, the current version of slender does not overcome the limitations of alternative approaches, although future developments may do so.

- Although SLiM and msprime are very flexible, large migration matrices between many populations are difficult to implement for models with complex geographical features. I am not sure how much the use of slendr simplifies the creation of these population interaction matrices compared to the original programs. As far as I understand, one instruction per population is required, which may be a limitation in the complexity of the models that can be implemented. To my opinion, this should be better explained. How many populations can be realistically considered with slendr?

- More importantly, if there are a few tools to simulate spatiotemporal genetic data, there is especially a lack of equivalent tools to simulate genomic data, which is becoming the standard of data produced in the literature. The problem lies mainly in the computing power needed to generate data at the genomic scale jointly to complex population dynamics models. slender aims to fill this gap, but I think that this manuscript is missing one major information useful for the potential users, which is the computation time needed to simulate the example scenarios. Giving examples of computation times will allow the readers to get a better feeling about the potential applications of slendr. Being able to simulate complex models is great, but if the computation time does not allow for a satisfactory exploration of the parameter space, especially for genomic data, this can be a strong limitation.

Reviewed by Liisa Loog, 24 October 2022

Martin Petr and colleagues present an R package -*slendr*- for generating and analysing simulated genomic data under spatiotemporally explicit demographic scenarios. More specifically, the framework provides a single easy-to-use front end that integrates with widely used, powerful and flexible genetic data simulation and analyses frameworks – SLIM, *msprime* and *tskit*. As such, the *slendr* package has great potential for simplifying and facilitating population genetics tool development and testing, with a much-needed functionality for explicitly incorporating spatial aspects into demographic models. Due to its single interface implemented in a popular R scripting and statistical analyses environment, the *slendr* package has the potential to make the field of computational population genetics more accessible to researchers and students with little computational or analytical background, as well as, improving overall reproducibility of research in the field.

The presented description of the key workings and key features of the package is clear and concise. The authors also provide examples of varying complexity, as well as links to external sources of further description, guidance and help, including a dedicated webpage for the R package.

I am not able to provide any comments on potential errors in the code as I have not extensively tested the described package or familiarised myself with the underlying code. However, the open-source nature of this software facilitates efficient flagging (and fixing) of any problems by the user-community. The package is also already part of the CRAN R package repository.

My main concern is that, while it is high time for a framework that would allow researchers with various degrees of analytical background to explicitly simulate and consider spatial factors affecting patterns of genetic variation, these tools (when used for model comparison for demographic inference, as also proposed by the authors) present a great opportunity to introduce (implicit) biases that are not easy to detect without proper statistical controls. (Similarly to the frequent bad practices in use of agent-based modelling in social sciences.) This issue of special concern here because human population history research is of elevated popular interest and frequently (mis)used by groups with strong political agendas.

To that end it would be great if the authors could elaborate on the discussion of the basic requirements for the downstream use of simulated data for model comparison and model parameter estimation that would incorporate (1) an emphasis on including a realistic null-model, (2) exploring a wide range of demographic scenarios and (3) performing hypothesis testing by formal model comparison (e.g. using AIC within the Approximate Bayesian Computation (ABC) framework), as well as, as providing examples of best practices and/or citing research with some theoretical discussion on the topic.

Liisa Loog

Reviewed by anonymous reviewer 2, 19 October 2022

I was pleased to be offered the opportunity to comment on this preprint, having been excited by its initial release. The manuscript is very well-written, and I am convinced that slendr succeeds at what it sets out to achieve; principally improving accessibility to complex population genetic modelling. The target audience for slendr is molecular ecologists, who as the manuscript highlights are often familiar with R as opposed to python, bash and in particular Eidos (the bespoke SLiM language). The package also provides a range of general functions for interacting with tree-sequencing outputs within the R environment that will likely have widespread interest for anyone using SLiM and/or msprime. I am confident that slendr represents an important addition to the ever-expanding SLiM/msprime ecosystem, and it is reassuring to see that this manuscript is co-authored by the developers of SLiM and msprime. In general, I found the manuscript did an excellent job in justifying the case for slendr and documenting and demonstrating its functionality. I have a few minor comments that I hope the authors find insightful, but otherwise commend their excellent work and look forward to making use of slendr in the future.

Coming from the perspective of someone who has worked with SLiM in the past, something that I felt could have been expanded on in the manuscript was a more explicit discussion of functionality that is not available. The manuscript does a good job of explaining functionality that is available, which for those with limited experience of simulation is useful, however at times I was unsure of whether simulations I've run in the past would be possible with slendr. For example, defining mutation types, genome element types, recombination landscapes in the initiate phase, or defining fitness and mating callbacks. It might be interesting to know which of SLiM's recipes are reproducible in the slendr framework to further provide a sense of what is and isn't possible for those who have worked with SLiM previously. Expanding on these would be beneficial for those readers who wish to take advantage of slendr's more general functionality, i.e. improved reproducibility, whole analyses contained within R, interacting with tskit etc.

Related to the above, my take-away from the manuscript was that the absence of mutation types and fitness callbacks limits slendr to simulating neutral models (with the exception of fitness through competition). If this is the case it may be worth mentioning briefly at some point in the discussion.

I was curious whether there were any be benchmarking issues incurred as a result of running slim simulations through slendr as opposed to as standard on the command-line. Does it scale similarly as users increase pop size, mutation/recombination rate? In addition to this, does the slendr framework lend itself to running simulations in parallel across multiple CPU, or would this be handled with general R parallelisation e.g. running models within doParallel or similar?

Why would a pre-requisite to spatial raster models have to necessarily involve non-WF functionality (line 659)? Presumably a WF-model that could describe the expected spatial population structure observed when a WF population's dispersal or mating is limited by local environment would be useful for various applications of interest to molecular ecologists, including connectivity for example. I should add that this section (line 653-671) does a good job overall of highlighting future applications.