Simulating genetic diversity in heterogenous landscapes: why, how, limits and perspectives

Brazilian Webinars on Bioinformatics - Chair: Cecilia Fiorini

By Dr. Arnaud Becheler

October 15th, 2021

University of Michigan, USA

Why heterogeneous landscapes?

Heterogeneity with 2 dimensions:

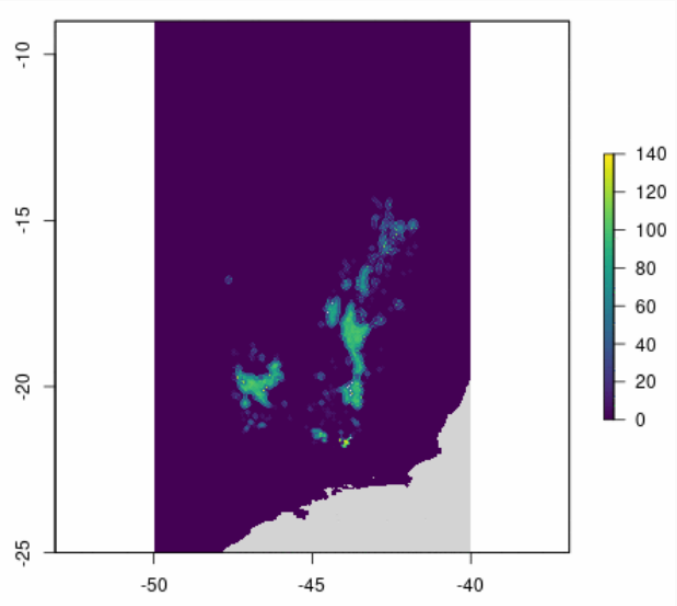
- **spatial autocorrelation:** areas that are close together tend to have similar values (*e.g., elevation*)
- **temporal autocorrelation:** events happening in quick successions tend to be more similar than events happening after a longer separation (*e.g., climate changes, vegetation dynamics*)

This can have profound effects on the different scales and levels of biodiversity.

Why heterogeneous landscapes?

Example: how to explain the biodiversity of *campos rupestres*?

- Higher elevation (above 900m)
- Poor rocky soils
- Disjuncted mosaic of sky-islands
- 15% of Brazil plants species
  but ... 1% of its area!
- Climatically buffered (stable)



Can we infer past demographic pulses based on genetic data?

Tournebize et al. (2017); He et al. (2017); Estoup et al. (2010)

Why genetic data?

- **The past is long gone:** historical data about past distribution are (at best) scarce and biased.
- **The present may be unreachable:** some ecological/evolutionary processes are not easily measured, even in a lab (e.g., the dispersal of a mosquito).
- **Promise of demogenetics**: signals of past demographic events may still be conserved in neutral areas of the genome, under the form of patterns and distribution of genetic diversity .

How to use such an indirect source of information to gain knowledge about distal processes?

Environmental demogenetics

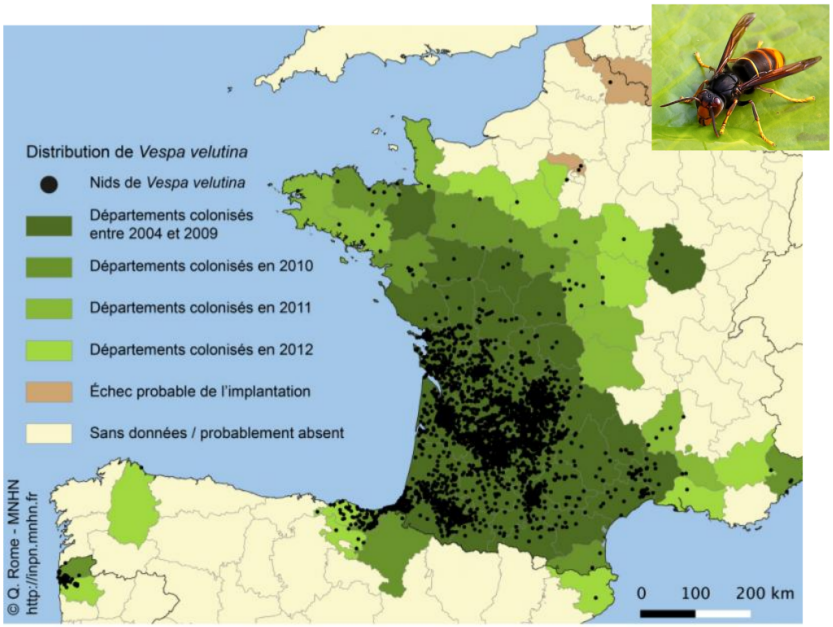
Environmental demogenetics aims to analyze geographic patterns of genetic diversity to inform the influence of **environment** on **demography**.

1. what data ? What scales?
2. how to formalize these processes (what model)?
3. how to extract information from data (what inference method)?
4. how to implement the method (what tools)?

What data? What scales?

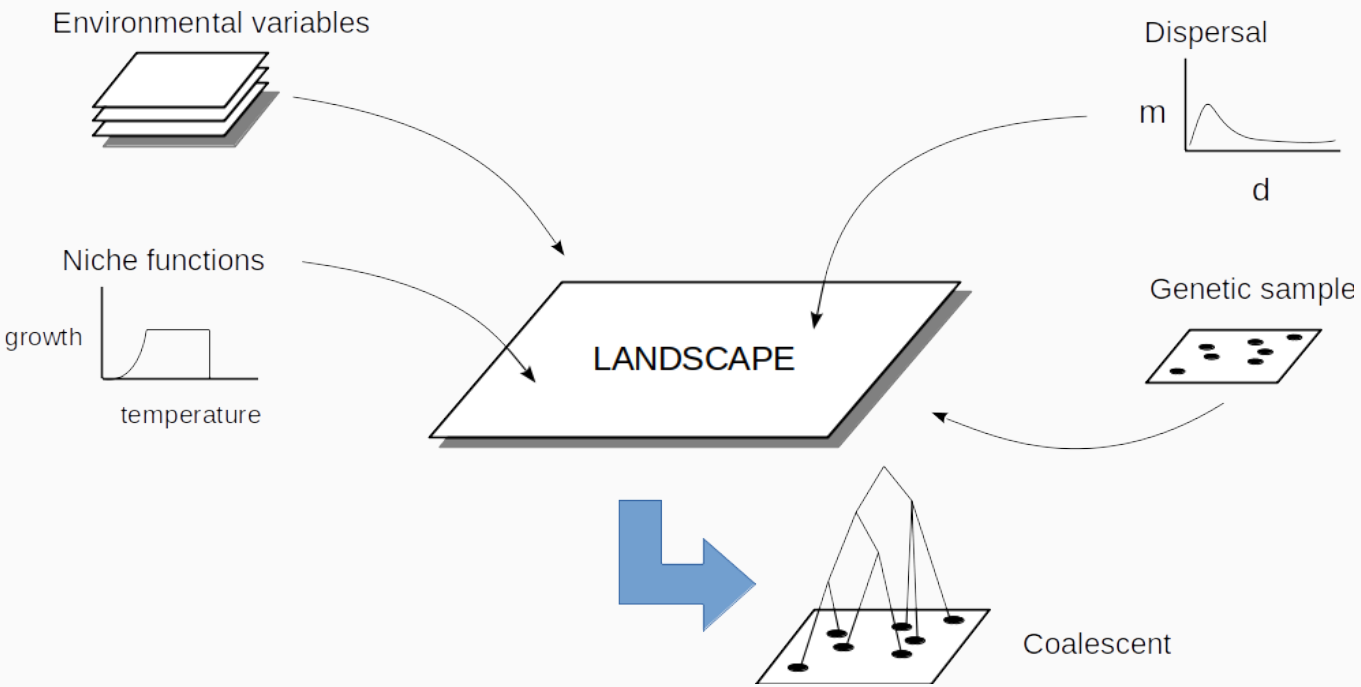
Vespa velutina nigritorax (yellow-legged hornet) invasion in Europe.

- first encountered in South-West France in 2004
- fast expansion
- economical/ecological impact (honey bee predator)
- 84 females genotyped for 22 SSR loci in 2008



lat	lon	List2003		VMA8	
43,50	2,21	163	163	262	264
43,50	2,21	163	169	262	264
43,56	1,46	163	169	260	264
43,97	3,68	177	177	262	264
44,14	4,17	177	177	264	267
44,22	0,55	177	177	264	264
44,22	0,55	163	177	262	264
44,22	0,55	169	177	262	264
44,29	0,69	177	177	264	264
44,29	0,69	163	163	246	264
44,37	3,41	165	169	0	0
44,57	0,23	163	169	264	264
44,57	0,23	163	177	249	264
44,57	0,23	163	163	249	264
44,57	0,23	153	163	249	264
44,57	0,23	169	177	249	264
44,57	0,19	163	163	264	264
44,57	0,19	163	169	264	264
44,57	0,19	163	169	249	264
44,57	0,19	163	177	249	264
44,67	2,16	169	177	264	264

What model(s)?



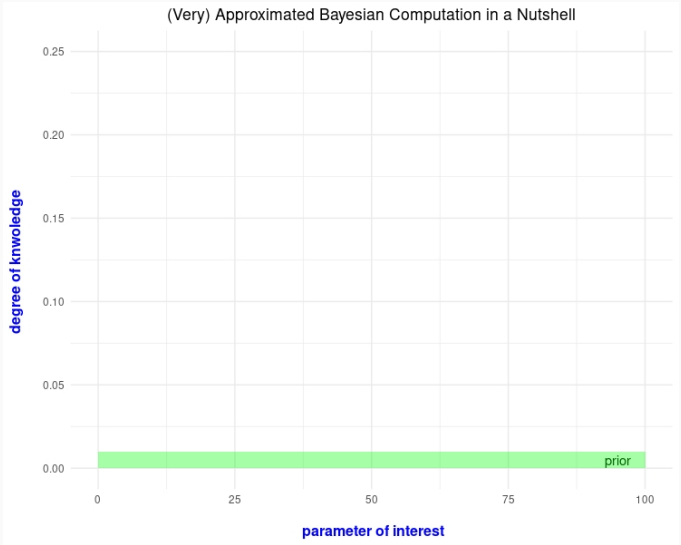
The main challenges are:

- chose submodels (**model selection**): technical challenges
- estimate parameters from genetic data (**statistical inference**)

What inference method?

- You have some knowledge or intuition *a priori* on parameters: *Bayesian framework*
- You can't do much mathematical work with the model: *intractable likelihood*

- You can't describe *mathematically* parameters of interest as a function of observed data.
- But you always can describe it *programatically* using **simulations**!

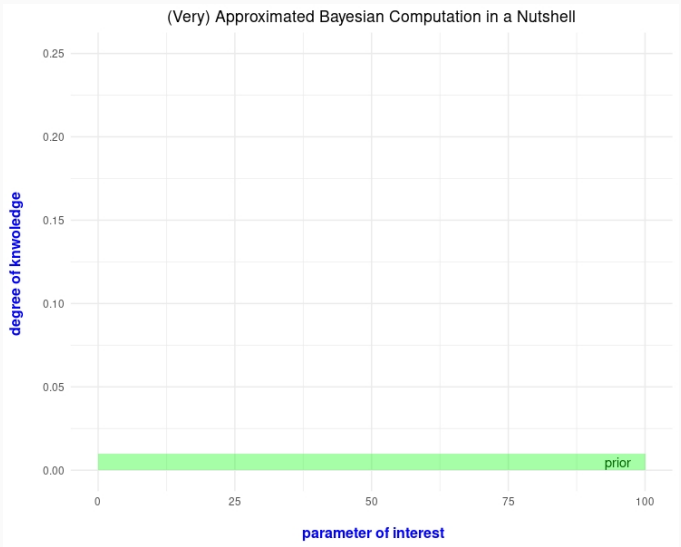


Use Approximate Bayesian Computation to explore the parameter space

How do you filter simulations?

You can use a (very basic) simulation/reject algorithm:

- sample parameters in a prior distribution: $\theta' \sim p(\theta)$
- simulate data from the model: $y' \sim p(y|\theta')$
- weigh (θ', y') as a function of the distance $||y' - y_{obs}||$



Wait ... you said *distance*?

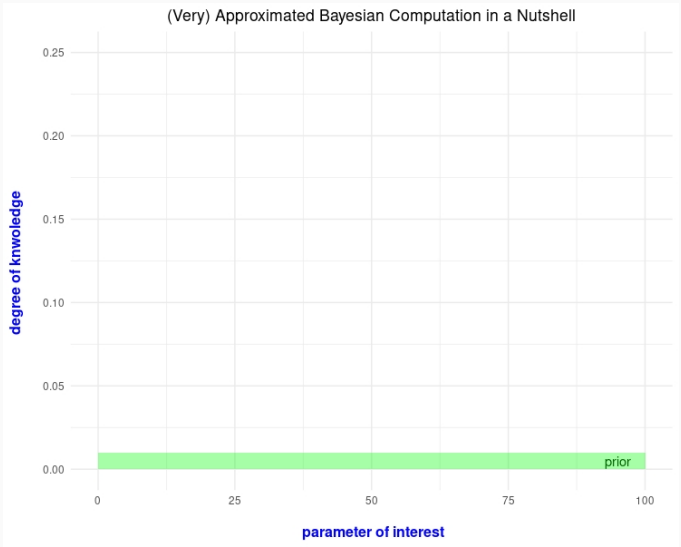
How do you define a distance?

Genetic data are:

- heterogeneous, with missing data
- highly multidimensional (location, individuals, loci...)
- distances in high dimension are weeeeird (Aggarwal et al., 2001)

In demogenetics you generally:

- compute summary statistics
- FST, heterozygosity etc
- using Arlsumstat Excoffier et al. (2005)

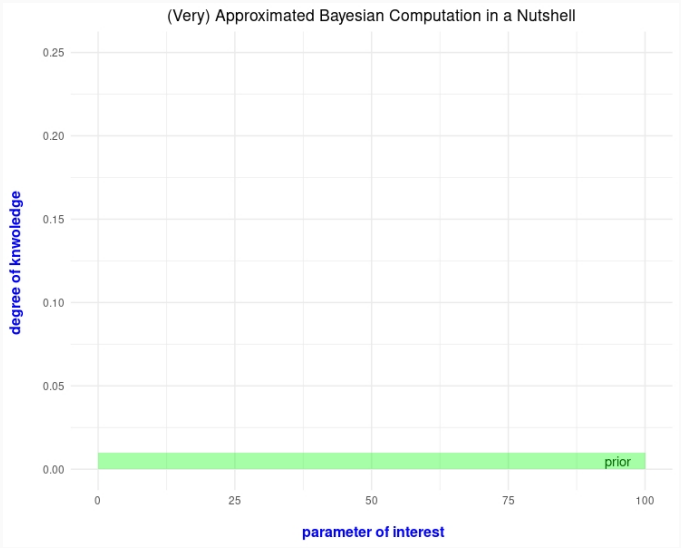


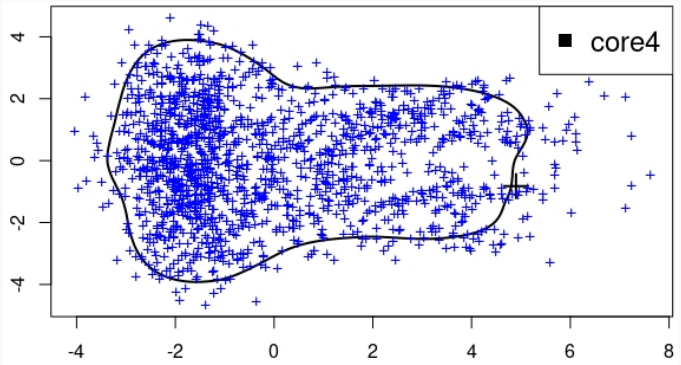
What do I gain form that, in terms of knowledge?

Posterior distribution estimation

As ABC algorithm progress

- density of accepted parameters increases in a region (hopefully)
- posterior density converges towards a target (hopefully)
- you updated your knowledge by switching probability mass from unlikely regions to more likely regions of the parameter space
- you can come up with point estimates (mean, median ...)





OK, but What simulators do I use?

Simulation tools, see Yannic et al. (2020)

Program	Simulator	Level	Lg.	Reference
Splatche3	Backward	Population	C++	Currat et al. (2019)
PhyloGeoSim	Backward	Population	Java	Dellicour (2013)
IBDsim	Backward	Population	C	Leblois et al. (2009)
SLIM3	Forward	Individual	C++	Haller and Messer (2019)

There is an abundance of simulation **programs**.

But sometimes it is impossible to get the desired behavior just with configuration files.

What if you want a different dispersal kernel ?

Our community needs to refine the **granularity of software resources**.

Where do I think we are going?

We need more simulation models, and we need them quickly.

There are too much variation in ecological processes to hope for a one-fit-all tool. We need to be able to increase the complexity of simulation models without losing years coding them.

This is being eased by the emergence of new domain-specific programming tools (libraries) that we can reuse and combine:

- the Quetzal framework (Becheler et al., 2019): good for demographic aspects, but currently assumes independent markers/trees
- Increasing density of markers has statistical impacts.
- TSKIT (Kelleher et al., 2018): efficient simulation of correlated trees along the genome

Can we interface these resources?

Quetzal

Programming choices for QuetzalCoalTL:

- C++ for a good compromise between design and performances
- Give as much information as possible to the compiler to make the code faster/safer
- Large use of generic programming (templates, metaprogramming)
- Focus on modularity and user experience

We need easier workflows, that can run anywhere at scale

It is still very difficult to understand, gather, install, configure, run the required resources. We need to decrease/hide the complexity of simulation-based inference frameworks.

This is being eased by new recent progresses:

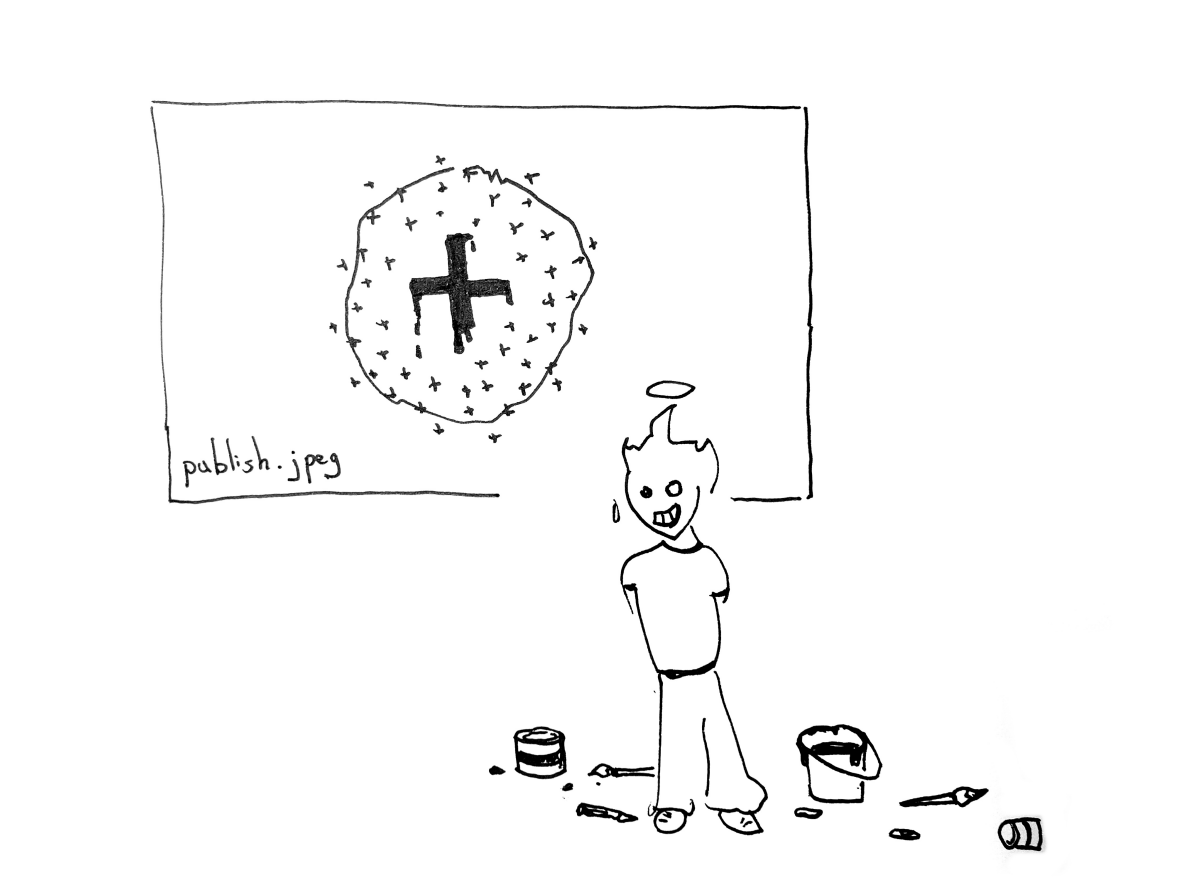
- **Random Forest ABC** (Raynal et al., 2019): bypass calibration, alleviate computational burden, standardize statistical workflow
- **Docker**: allows to package and run anywhere **entire environments**. Simplifies distribution and dependency management of scientific softwares.
- **Singularity** (Kurtzer et al., 2017): makes containers available for High Throughput Computing
- **OpenScienceGrid**: (Pordes et al., 2007): large scale computations on very heterogeneous grids of nodes.

Can we use these resources to build entire reproducible workflows?

Thank you!

Thank you to Cecilia Fiorini for inviting me, to the NSF for funding me, to the Open Science Grid school 2021 organizers for forming me, and to the Lacey Knowles lab!

Keep targetting!



References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer.

Becheler, A., Coron, C., and Dupas, S. (2019). The quetzal coalescence template library: a c++ programmers resource for integrating distributional, demographic and coalescent models. *Molecular ecology resources*.

Currat, M., Arenas, M., Quilodràn, C. S., Excoffier, L., and Ray, N. (2019). Splatche3: simulation of serial genetic data under spatially explicit evolutionary scenarios including long-distance dispersal. *Bioinformatics*, 35(21):4480–4483.

References ii

Dellicour, S. (2013). Phylogeosim 1.0: a simulator of dna sequences evolution under a demographic and geographic model of coalescence. In *5th CÉCI Scientific Meeting, Location: Namur, Belgium*.

Estoup, A., Baird, S. J., Ray, N., Currat, M., Cornuet, J.-M., Santos, F., Beaumont, M. A., and Excoffier, L. (2010). Combining genetic, historical and geographical data to reconstruct the dynamics of bioinvasions: application to the cane toad Bufo marinus: RECONSTRUCTING BIOINVASION DYNAMICS. *Molecular Ecology Resources*, 10(5):886–901.

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary bioinformatics*, 1:11769343050010003.

References iii

Haller, B. C. and Messer, P. W. (2019). Slim 3: forward genetic simulations beyond the wright–fisher model. *Molecular biology and evolution*, 36(3):632–637.

He, Q., Prado, J. R., and Knowles, L. L. (2017). Inferring the geographic origin of a range expansion: Latitudinal and longitudinal coordinates inferred from genomic data in an ABC framework with the program x - origin. *Molecular Ecology*, 26(24):6908–6920.

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

References iv

Kurtzer, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5):e0177459.

Leblois, R., Estoup, A., and Rousset, F. (2009). IBDSim: a computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources*, 9(1):107–109.

Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., Roy, A., Avery, P., Blackburn, K., Wenaus, T., Würthwein, F., et al. (2007). The open science grid. In *Journal of Physics: Conference Series*, volume 78, page 012057. IOP Publishing.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2019). Abc random forests for bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728.

References v

Tournebize, R., Manel, S., Vigouroux, Y., Munoz, F., de Kochko, A., and Poncet, V. (2017). Two disjunct pleistocene populations and anisotropic postglacial expansion shaped the current genetic structure of the relict plant amborella trichopoda. *PloS one*, 12(8):e0183412.

Yannic, G., Hagen, O., Leugger, F., Karger, D. N., and Pellissier, L. (2020). Harnessing paleo-environmental modeling and genetic data to predict intraspecific genetic structure. *Evolutionary Applications*, 13(6):1526–1542.