Inference of the Cultural Transmission of Reproductive Success from human genomic data: ABC and machine learning methods

Arnaud Quelin, Jérémy Guez, Ferdinand Petit, Flora Jay, Frederic Austerlitz

Inference of the Cultural Transmission of Reproductive Success from human genomic data: ABC and machine learning methods

Arnaud Quelin, Jérémy Guez, Ferdinand Petit, Flora Jay, Frédéric Austerlitz

Inference of the Cultural Transmission of Reproductive Success from human genomic data: ABC and machine learning methods

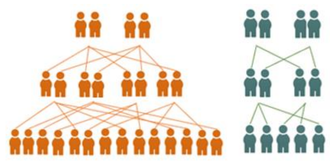
Arnaud Quelin[1,2], Jérémy Guez[1,2], Ferdinand Petit[1,2], Flora Jay[2], Frédéric Austerlitz[1]

[1] UMR 7206 Eco-Anthropologie, Université Paris Cité, CNRS, MNHN, [2]LISN, Université Paris-Saclay, CNRS, INRIA

Introduction

The **Cultural Transmission of Reproductive Success (CTRS)** is one of the various cultural processes which can impact human genetic evolution.

Under CTRS, there is a **positive correlation** between **the progeny size** of parents and children.



Genealogical trees along the genome are imbalanced under this process, which we can measure through **imbalance indices**.

Objective: develop and evaluate methods to infer the intensity of the CTRS (α) from genomic data.

Data Simulation

In order to create a labelled database to train our models, we sample 1000 scenarios from the prior of α. The forward-in-time simulator SLiM [1] is used to **simulate populations** while integrating the strength of the CTRS (α) given these scenarios.

The probability p_i for a given couple i of being chosen as parent for an individual of the new generation is given by:

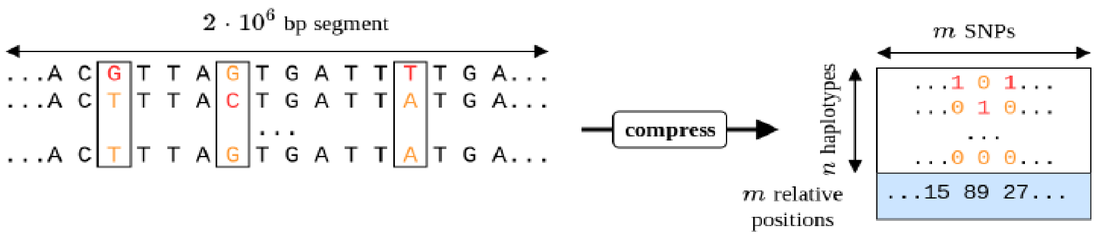
$$p_i = \frac{\gamma_i(b) \times {s_i}^{\alpha}}{\sum_{j=1}^{N_c} \gamma_j(b) \times {s_j}^{\alpha}}$$

where s_i is the average sibship size of the two members of couple i, α the parameter controlling the intensity of CTRS and b the parameter controlling the variance in reproductive success.

The other scenario parameters are:

- Genome size : 2×10^{-6} bp
- Mutation rate : 1.45×10^{-8} per bp
- Recombination rate : 1×10^{-8} per bp
- Sample size : 20 individuals
- N_e : 1000 individuals
- CTRS length : 10 generations

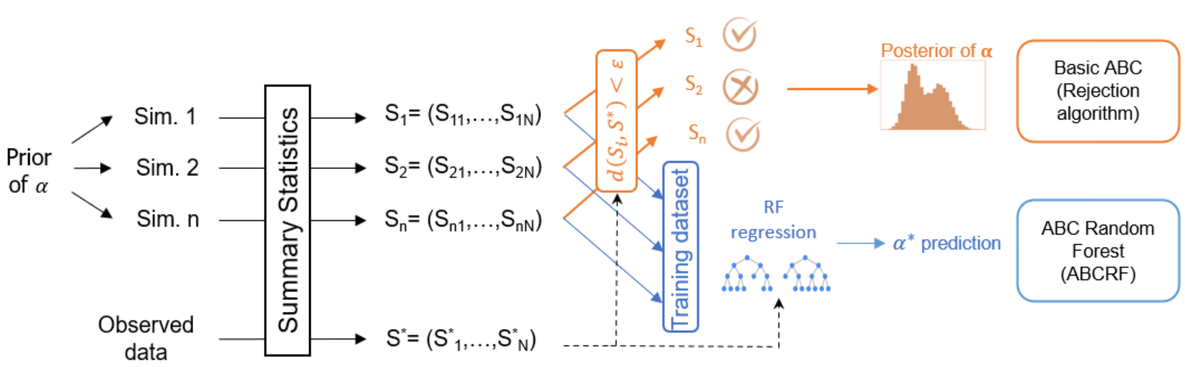
Simulations output the genomes in the form of **Single Nucleotide Polymorphism (SNP) matrices** (individuals in row and loci in column) where 0 and 1 correspond to the two possible alleles.



Approximate Bayesian Computation approach

Approximate Bayesian computation (ABC) is a frequently used **likelihood-free** inference method in population genetics. It relies on **summary statistics** that simplify our high-dimensional genomic data.

The ability to extract meaningful summary statistics in our data to infer the parameter of interest α is key. In our case we can use, among other statistics, the imbalance indices. We compute these indices on **inferred trees**, using tsinfer [2], a step that could yield biases in our indices.

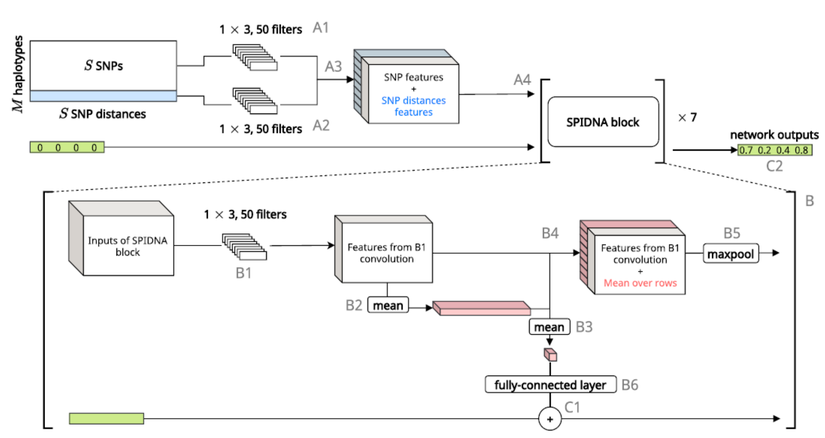


Deep Learning approach

An alternative to summary statistics is based on deep learning to automatically extract relevant information from raw genomic data.

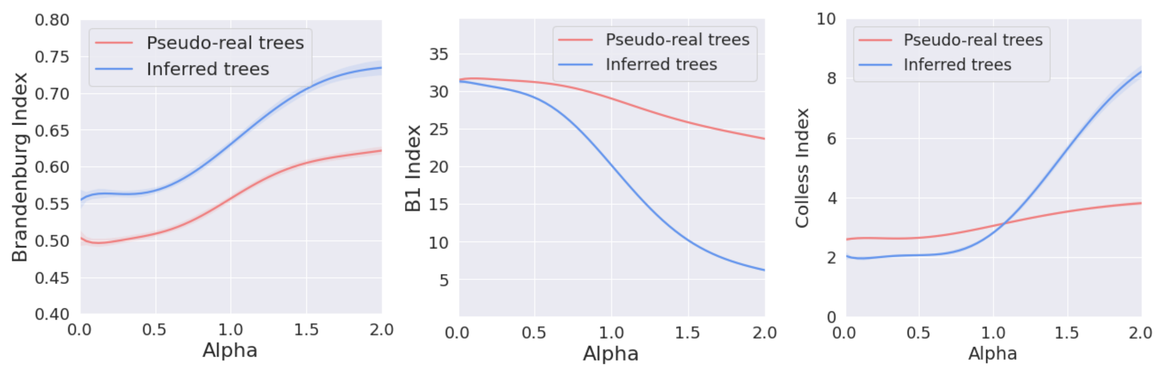
We use a neural network architecture called **SPIDNA** developed by Sanchez et al. [3] that is:

- **adaptive** to the number of SNPs and haplotypes
- **invariant** to haplotype permutation
- able to combine **relatives positions** and SNPs

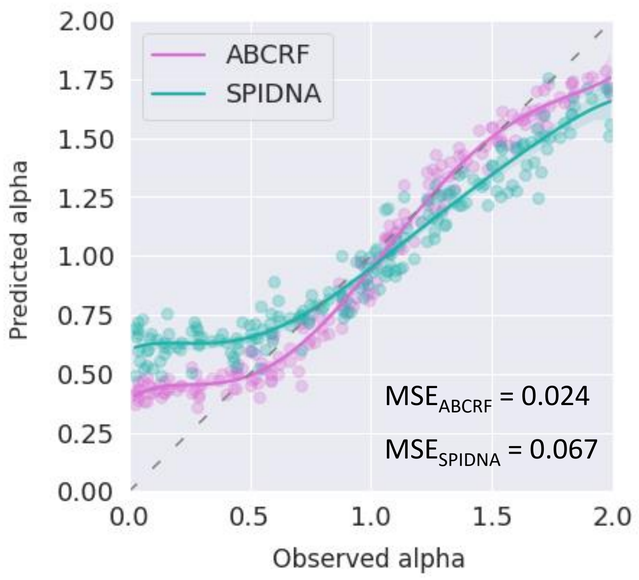


Results

We checked to what extent the tree inference process could introduce bias into our imbalance indices. Although an overestimation of the imbalance is produced, the statistics vary in the same direction between the inferred and pseudo-real trees.



We compared the predictions of the two models on a test dataset of 200 scenarios :



Both models show a good ability to infer the intensity of the CTRS on our simulations.

They exhibit increased difficulties in prediction for low values of α and to a lesser extent for high values.

The ABCRF model outperformed the SPIDNA model on our simulations, and is better on the low and high value ranges of α.

Conclusion

The summary statistics calculated on the inferred trees followed identical trends to those calculated on the pseudo-real trees, that is key to develop the ABC approach.

The two competing approaches show a good ability to infer CTRS on genomic data. Although ABCRF outperformed on simulations, they are both worth investigating, especially under more complex evolutionary histories.

Future works:

- Be able to co-infer demographic histories introduced in the simulations.
- Improve models by exploring further deep learning architectures or summary statistics.
- Explore other tree inference algorithms such as argweaver [4] and relate [5].
- For the model to be used on real human data, integrate other parameters such as natural selection or migration.

References

[1] Benjamin C Haller and Philipp W Messer. Slim 3 : forward genetic simulations beyond the wright–fisher model. Molecular biology and evolution, 36(3) :632–637, 2019.
[2] Jerome Kelleher, YanWong, AnthonyW Wohns, Chaimaa Fadil, Patrick K Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. Nature genetics, 51(9) :1330–1338, 2019.
[3] Théophile Sanchez, Jean Cury, Guillaume Charpiat, and Flora Jay. Deep learning for population size history inference : Design, comparison and combination with approximate bayesian computation. Molecular Ecology Resources, 21(8) :2645–2660, 2021.
[4] Leo Speidel, Marie Forest, Sinan Shi, and Simon R Myers. A method for genome-wide genealogy estimation for thousands of samples. Nature genetics, 51(9) :1321–1329, 2019.
[5] Matthew D Rasmussen, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. PLoS genetics, 10(5) :e1004342, 2014.