Inferring the history of human populations from polymorphism data of ancient and modern genomes by ABC and machine learning methods

Context

ABC and machine learning methods in population genomics are based on very large simulated datasets. These simulations (or a fraction of them sufficiently close to real data) are used to train predictive models. Regarding ABC, training is performed on summary statistics, computed on the simulated and real data [1]. These training methods generally rely on a correction step that learns the local relationship between the summary statistics and the target parameters, in the vicinity of the real data. This can be done through simple penalized linear regressions, through shallow neural networks [2] or random forests [3]. In end-to-end deep learning methods, the raw data (the DNA sequences) are processed directly by a deep neural network that automatically computes informative features for a given task (model classification or parameter estimation) [4]. The learned predictive function is global (i.e., a single predictive model is learned and applicable to all inputs). All these approaches enable to determine which demographic scenario among a set of possible historical scenarios is the most consistent with the data and to infer the parameters of this scenario (population sizes, migration rates, splitting times, ...). We have developed methods mainly in the context of isolated populations [4-6], without taking into account possible migrations between populations. In addition, our methods were based solely on modern DNA, while more and more ancient genomes become available [7]. Methods integrating migrations between populations have also been developed in our teams [8, 9] but they were based on traditional markers (microsatellites, short DNA sequences).

Objectives

The project aims to develop Approximate Bayesian Computation (ABC) [1] and deep learning methods to infer the demographic history of populations, from current and ancient DNA sequences. We will study, in particular, the case of populations that separated at a given moment in the past, continuing or not to exchange migrants afterwards and going through demographic events such as expansions or contractions. The objective will be to determine which are the most efficient methods and, for the ABC methods, the most efficient summary statistics. The methods will then be applied to a set of human populations with different lifestyles: farmers, herders and hunter-gatherers. These populations coexist in various part of the World. For instance, farmers and hunter-gatherers live near each other in Equatorial Africa, and farmers and herders coexist in Central Asia (these populations are well studied in UMR7206). The methods that we will develop will allow inferring the joint demographic history of these neighbouring populations. We aim in particular to determine the timing of their splitting, and to assess whether they have been under expansion and/or contraction events since then, and if they have exchanged some migrants. The history of splitting of populations from different continents will also be studied. Finally, we will also investigate the contribution of ancient DNA samples to the quality of estimates, a question that has been little studied so far.

Methods

The first step will consist in finalizing the development of the ABC approach based on summary statistics, by integrating multi-population statistics into our previous approaches [5, 6], defining the prior distributions of the demographic histories of interest, and simulating on HPC clusters the training database. We will then adapt the neural network developed by UMR9015 for single-population demographic inference to the processing of multi-population data [4] and compare it to an approach that handles multiple populations [10] but does not scale to several real (or test) datasets (because its training is tedious and done for each target set, as in ABC). In particular, we will adapt (i) our exchangeable neural network (invariant to the permutation of sampled individuals within a population) and (ii) our new solution for integrating multiple genomic regions for demographic prediction. The latter is a case of multiple instance learning where multiple examples are leveraged to make a single global prediction (as it is the case in demographic inference contrary to local selection inference).This task will be carried out with dnadna [11], the software (developed by UMR9015) that facilitates the use of deep statistical learning for population genetics. The performances of the different algorithms (after optimization of their hyper-parameters) will be evaluated on test datasets, i.e. simulated data whose true demographic history is known. This will allow gauging the ability of each method to choose the right scenario among several proposed and to infer its parameters.

The most effective method(s) will then be applied to complete genome data, such as those of public databases [12, 13] or those available in UMR 7206. This will make it possible to reconstruct the history of separation and exchange between human populations. Ancient DNA data will then be integrated into the methods, including the specificities of ancient genomes such as low coverage and high rate of sequencing errors. This has been proposed recently in an ABC framework, however for a single population only [14]. It will be necessary to study through simulations the behaviour of the summary statistics in mixed samples of ancient and modern DNA, given in particular that the latter will have accumulated fewer mutations than the former. It will then be possible to assess the increase in precision in demographic estimates provided by the inclusion of ancient DNA. The developed methods will then be applied to datasets including both ancient and modern DNA data, for example for populations in Central Asia, where such data become available.

Expected results

The thesis will make it possible to develop powerful methods to reconstruct the history of populations. We aim to apply these methods to human populations, but they can then be adapted to populations of other species (animals, plants, …). Application to human populations will allow us to infer in particular finely the joint history of populations with contrasting lifestyles that leave nearby. We will be able to infer when these populations separated, if they have since experienced demographic events and gene flow between populations. The methods will also be applied on a larger scale to determine in particular the histories of separation and possible reconnections between the populations of the different continents.

Material conditions

The project does not require the acquisition of new data. The student will benefit from a powerful computing station and will have free access to the MNHN computing cluster.

Planned collaborations

Fanny Pouyet UMR9015 LISN Orsay on the development of summary statistics.

Role and percentage of supervision of each supervisor on this project

Frederic Austerlitz (50%). In particular, will supervise the development, testing and application of ABC methods.
Flora Jay (50%). In particular, will oversee the development, testing and application of machine learning methods.
The other tasks will be supervised equally by the two co -supervisors.
Weekly meetings will be organized between the student and his co -supervisors. The thesis committee will meet annually and informal interactions may take place in the meantime with committee members.

Publications linked to the project:

Verdu, P., Becker, N. S. A., Froment, A., Georges, M., Grugni, V., Quintana-Murci, L., . . . Austerlitz, F. (2013). Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. Molecular Biology and Evolution, 30(4), 918-937. doi:10.1093/molbev/mss328

Boitard, S., Rodriguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data - an approximate Bayesian computation approach. PLoS Genetics, 12(3), e1005877. doi:10.1371/journal.pgen.10058707

Jay, F., Boitard, S., & Austerlitz, F. (2019). An ABC method for whole-genome sequence data: Inferring Paleolithic and Neolithic human expansions. Molecular Biology and Evolution, 36(7), 1565-1579. doi:10.1093/molbev/msz038

Sanchez, T., Cury, J., Charpiat, G., & Jay, F. (2021). Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Molecular Ecology Resources, 21(8), 2645-2660. doi: 10.1111/1755-0998.13224

Sanchez T., Madison Bray E., Jobic P., Guez J., Letournel A.-C., Charpiat G., . . . Jay F. (2021) dnadna: Deep neural architectures for DNA - a deep learning framework for population genetic inference. https://hal.archives-ouvertes.fr/hal-03352910 (preprint)

Timetable

Year 1: Adaptation of the simulation program and development of scripts to calculate summary statistics. Writing a methodological article on the impact of the different parameters of the model on these statistics and the relevance of the information retained for inference (i.e. accuracy of estimations with ABC-ML based on these summary statistics).

Year 2: Adaptation of our single population DNN to multiple populations. Comparison of summary-statistic-based methods to the end-to-end DNN approach. Application to real data (modern DNA). Writing an article on these methods.

Year 3: Study of the impact of adding ancient DNA data. Testing the robustness of methods to the quality of ancient DNA. Writing an article on the subject and the rest of the thesis manuscript (thesis on article).

Adequation to SCAI

Our project is a strongly interdisciplinary project that aims at developing, optimising and applying machine learning methods to genomic data (Approximate Bayesian Computation combined with machine-learning-based adjustments; random forest; exchangeable convolutional neural networks), in order to reconstruct the demographic history of human populations with different lifestyles. It has thus a strong implication in several fields such as population genomics, demography and social sciences.

Profile and skills required

Ideally, the student would have received a training in bioinformatics that includes statistical learning. We will favour candidates with strong expertise in genetics/population genetics, machine learning and programming (python, pytorch, HPC scripts, etc.). S/He will develop new modelling and statistical tools by adapting simulation programs, implementing inference models, and developing the necessary scripts to analyse the outputs and benchmark methods. S/He will also perform the application of these tools to human genomics data, which requires developing bioinformatics pipelines processing large genomic datasets.

References

1. Beaumont, M.A., W. Zhang, and D.J. Balding, *Approximate Bayesian computation in population genetics.* Genetics, 2002. **162**(4): 2025-2035.
2. Blum, M.G.B. and O. François, *Non-linear regression models for Approximate Bayesian Computation.* Statistics and Computing, 2010. **20**(1): 63-73.
3. Raynal, L., et al., *ABC random forests for Bayesian parameter inference.* Bioinformatics, 2019. **35**(10): 1720-1728.
4. Sanchez, T., et al., *Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation.* Molecular Ecology Resources, 2021. **21**(8): 2645-2660.
5. Boitard, S., et al., *Inferring population size history from large samples of genome-wide molecular data - an approximate Bayesian computation approach.* PLoS Genet, 2016. **12**(3): e1005877.
6. Jay, F., S. Boitard, and F. Austerlitz, *An ABC method for whole-genome sequence data: Inferring Paleolithic and Neolithic human expansions.* Molecular Biology and Evolution, 2019. **36**(7): 1565-1579.
7. Orlando, L., et al., *Ancient DNA analysis.* Nature Reviews Methods Primers, 2021. **1**(1): 14.
8. Verdu, P., et al., *Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies.* Mol Biol Evol, 2013. **30**(4): 918-37.
9. Palstra, F.P., E. Heyer, and F. Austerlitz, *Statistical inference on genetic data reveals the complex demographic history of human populations in central Asia.* Mol Biol Evol, 2015. **32**(6): 1411-24.
10. Wang, Z., et al., *Automatic inference of demographic parameters using generative adversarial networks.* Molecular Ecology Resources, 2021. **21**(8): 2689-2705.
11. Sanchez, T., et al., *dnadna: Deep neural architectures for DNA - a deep learning framework for population genetic inference*. 2021. https://hal.archives-ouvertes.fr/hal-03352910.
12. The 1000 Genomes Project Consortium, *A global reference for human genetic variation.* Nature, 2015. **526**(7571): 68-74.
13. Bergström, A., et al., *Insights into human genetic variation and population history from 929 diverse genomes.* Science, 2020. **367**(6484): eaay5012.
14. Pavinato, V.A.C., et al., *Joint inference of adaptive and demographic history from temporal population genomic data.* bioRxiv, 2021: 2021.03.12.435133.