SORBONNE UNIVERSITÉ

PROGRAMME INTITUTS ET INITIATIVES

Appel à projet – campagne 2021

Proposition de projet de recherche doctoral (PRD)

ISCD-Institut des Sciences du calcul des Données

Intitulé du projet de recherche doctoral (PRD): Inférence des processus démographiques, culturels et adaptatifs à partir des données genomiques

Directeur.rice de thèse porteur.euse du projet (titulaire d'une HDR) :

NOM : **Austerlitz** Prénom : **Frédéric**
Titre : Directeur de Recherche ou
e-mail : frederic.austerlitz@mnhn.fr
Adresse professionnelle : Musée de l'Homme, 17 Place du Trocadéro, 75116 Paris
(site, adresse, bât., bureau)

Unité de Recherche :
Intitulé : Eco-anthropologie
Code *(ex. UMR xxxx)* : UMR 7206

École Doctorale de rattachement de l'équipe (future école doctorale du.de la doctorant.e) : **ED227-Sciences vie homme : évolution écolo**

Doctorant.e.s actuellement encadré.e.s par la.e directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) : 2. une étudiante inscrite depuis 2017 (thèse prolongée pour raison médicale), un étudiant inscrit depuis 2019

Co-encadrant.e :

NOM : **Pouyet** Prénom : **Fanny**
Titre : Maître de Conférences des Universités ou HDR ☐

e-mail : fanny.pouyet@universite-paris-saclay.fr

Unité de Recherche :
Intitulé : Laboratoire Interdisciplinaire des Sciences du Numérique
Code *(ex. UMR xxxx)* : UMR9015

École Doctorale de rattachement : **Choisissez un élément :**
Ou si ED non Alliance SU : **ED580 STIC (Sciences et technologies de l'information et de la communication) de l'Université Paris Saclay**

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) : 0

Co-encadrant.e :

NOM : Prénom :
Titre : Choisissez un élément : ou HDR ☐
e-mail :

Unité de Recherche :
Intitulé :
Code *(ex. UMR xxxx)* :

École Doctorale de rattachement : **Choisissez un élément :**
Ou si ED non Alliance SU :

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) :

Cotutelle internationale : ☒ Non ☐ Oui, précisez Pays et Université :

Selon vous, ce projet est-il susceptible d'intéresser une autre Initiative ou un autre Institut ?
☐ Non ☒ Oui, précisez SCAI - Sorbonne Center for Artificial Intelligence

Description du projet de recherche doctoral *(en français ou en anglais) :*

Ce texte sera diffusé en ligne : il ne doit pas excéder 3 pages et est écrit en interligne simple.

Détailler le contexte, l'objectif scientifique, la justification de l'approche scientifique ainsi que l'adéquation à l'initiative/l'Institut.

Le cas échéant, préciser le rôle de chaque encadrant ainsi que les compétences scientifiques apportées. Indiquer les publications/productions des encadrants en lien avec le projet.
Préciser le profil d'étudiant(e) recherché.

Context

Cultural transmission of reproductive success (CTRS) has been observed in several human populations: children from large families have more children on average than children from small families (Heyer, et al. 2005). This non-genetic transmission of reproductive success affects the genetic evolution of populations, leading to a decrease in genetic diversity and an increase in frequency of severe genetic diseases (Austerlitz and Heyer 1998). We showed (Sibert, et al. 2002; Blum, et al. 2006) that CTRS yields an imbalance of the coalescent tree, the genealogical tree of a sample of genes in a population (Kingman 1982), in particular on the uniparentally transmitted Y chromosomes and mitochondrial DNA (Brandenburg et al. 2012), allowing to infer patrilineal and matrilineal transmission, respectively (Heyer, et al. 2015).

Human populations are also affected by polygenic selection, i.e. selection on a trait coded by many loci distributed along the genome. Natural selection occurring at such a trait should lead to a correlation of reproductive success between parents and offspring, as the offspring will inherit at least partially the favourable phenotype of their parents. It may thus leave on the genome a signal similar to that of CTRS. Demographic processes, such as expansions, contractions or migrations, also affect genomic diversity. The question remains whether it is possible to distinguish all these phenomena using current population genomics data.

Scientific objectives

We aim first at performing a simulation study to assess to which extent these processes shape the genomic diversity of populations. Based on these simulations, we will develop methods assessing whether a given population was affected specifically by one of these processes or a combination of them. These methods will then be applied to human populations, in particular the population of Saguenay Lac Saint Jean, for which CTRS has been identified using demographic data (Austerlitz & Heyer, 1998), and which is also known to have been under demographic expansion, though a collaboration with Simon Gravel (Université McGill, Canada). This will allow assessing to which extent these phenomena can indeed be detected. Then, the methods will be applied to public databases such as the 1000 Genomes project (The 1000 Genomes Project Consortium. 2015) or the HGDP-CEPH panel (Bergström et al, 2020), to determine the intensity of these phenomena in populations with contrasted lifestyles such as farmers, herders and hunter-gatherers.

Scientific approach

We will first develop a simulation approach based on an existing program (Slim 3.0, Haller & Messer, 2019). It will allow simulating populations connected by migration submitted to CTRS, polygenic selection and demographic processes. Individuals will be characterised by their genotypes at many loci. Most loci will be neutral, but some will be involved in one or several quantitative traits. Polygenic selection will occur through the setting of a phenotypic optimum for these traits. CTRS will be modeled by an increased probability of having children for individuals from large families, independently of their genotype. Populations will also be submitted to demographic expansions or contractions, and variations of the levels of migration through time.

The simulations will first predict the impact of these phenomena on genomic diversity, measured with summary statistics such as Tajima's D, among-population differentiation (FST), within- and among-population linkage disequilibrium and site frequency spectrum (SFS), as well as the coalescent tree shape, inferred with tsinfer (Kelleher, et al. 2019) or relate (Speidel, et al. 2019). We will also use the derived allele frequency per individual (DAFi, Pouyet et al. 2018). It should help disentangling the effect of CTRS versus demography or selection, as it is insensitive to within-population demographic events and sensitive to natural selection in a predictable way.

Then these simulations will be used to develop methods aiming at detecting whether one or several of these phenomena occurred in a given population, and at which intensity. The methods will be based either on Approximate Bayesian Computation (ABC, Beaumont, et al. 2002) or Machine Learning tools, such as neural networks (Ripley 1996) or random forests (Breiman 2001). All methods rely on performing numerous simulations, used as a training set. This training set allows inferring which model best fits on real genomic data and estimating its parameter values. Regarding ABC, training will be performed on the above-described summary statistics, computed on the simulated and real data. For machine learning, the raw data (the DNA sequences) will be processed directly by a neural network that automatically computes informative features for a given task. Deep neural networks tailored to genetic data will be trained on simulated datasets to learn how to predict CTRS presence and its parameters. We already developed such networks for demographic inferences or generating genomic data (Sanchez, et al. 2020; Yelmen, et al. 2021). These networks will be the starting point for our project. A cross-validation procedure will be performed on the simulated data, in order to assess which method performs best depending on the conditions. Then, the most accurate methods will be applied to the real data sets.

This project will allow us to develop new machine learning methods for the inference of CTRS together with polygenic selection and demographic processes, in the case of ABC by including new statistics, and in the case of deep learning methods by testing the performance of existing neural architectures and, if needed, by developing new architectures. Developing new summary statistics and studying the generalization of neural architectures is of interest to the whole population genetics community (as they could be applied to other questions). The application to real human populations is of interest for social sciences, as it will allow in particular for the first time assessing precisely the generality of cultural transmission of reproductive success in human populations.

The student will thus have to combine in-depth knowledge of population genetics with computer and statistical skills. S/He will develop new modelling and statistical tools by adapting simulations programs and developing the necessary scripts to analyse the outputs. S/He will also perform the application of these tools to human genomics data, developing bioinformatic pipelines allowing to perform the necessary analyses.

Frédéric Austerlitz is a theoretical population geneticist specialised in the modelling of the impact of demographic, adaptive and cultural processes on genomic diversity, and conversely on the development of statistical methods aiming at inferring these processes from genomic diversity. He

will supervise in particular the demographic and cultural aspects in the simulation study and the development of statistics for ABC methods, and the application of these methods.

Fanny Pouyet is a population geneticist specialised on molecular evolutionary processes such as natural selection for this project. She will supervise the simulation study and the data analyses, in particular the implementation of polygenic selection and the comparison of summary statistics.

Flora Jay (collaborator) develops ABC and machine learning-based methods for population genetics questions, with major contributions to the inference of population demographic histories using modern and ancient genomic data. She will be strongly involved in particular in the design, choice, and application of machine-learning methods.

Adequation with the ISCD program

Our project fits quite well the objectives of the ISCD program since it relies on simulations and highly computational statistical methods to analyse large sets of data, namely genome-wide sequence polymorphism data. This project will provide insights of strong interest for life sciences, in which the question of polygenic adaptation and its extent is quite timely, and for social sciences, as we will investigate how a sociocultural factor differs

References (references by the carriers of the project are indicated by the mention [our work])

Austerlitz F, Heyer E. 1998. Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population PNAS 95:15140-15144. https://www.pnas.org/content/95/25/15140 [our work]

Beaumont et al 2002. https://www.genetics.org/content/162/4/2025

Bergström et al. 2020. Science 367:eaay5012. https://doi.org/10.1126/science.aay5012

Blum MGB, Heyer E, François O, Austerlitz F. 2006. Matrilineal fertility inheritance detected in hunter-gatherer populations using the imbalance of gene genealogies. PLoS Genetics 2:e122. https://doi.org/10.1371/journal.pgen.0020122

Breiman L. 2001. Machine Learning 45:5-32. https://doi.org/10.1023/A:1010933404324

Haller BC, Messer PW. 2019. Mol Biol Evol 36:632-637. https://doi.org/10.1093/molbev/msy228

Heyer E, Brandenburg JT, …, Austerlitz F. 2015. Patrilineal populations show more male transmission of reproductive success than cognatic populations in Central Asia, which reduces their genetic diversity. AJPA 157:537-543. https://doi.org/10.1002/ajpa.22739 [our work]

Heyer E, Sibert A, Austerlitz F. 2005. Cultural transmission of fitness: genes take the fast lane. Trends in Genetics 21:234-239. https://doi.org/10.1016/j.tig.2005.02.007 [our work]

Kelleher et al 2019 Nature Genetics 51:1330-1338. https://doi.org/10.1038/s41588-019-0483-y

Kingman JFC. 1982. Stoch Proc App 13:235-248. https://doi.org/10.1016/0304-4149(82)90011-4

Pouyet F, Aeschbacher S, Thiéry A and Excoffier L. 2018. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences . eLife 7:e36317 https://doi.org/10.7554/eLife.36317 [our work]

Ripley BD. 1996. Pattern recognition and neural networks. Cambridge; New York: Cambridge University Press.

Sanchez T, Cury J, Charpiat G and Jay F. 2020. Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Molecular Ecology Resources. Online early. https://doi.org/10.1111/1755-0998.13224 [our work]

Sibert A, Austerlitz F, Heyer E. 2002. Wright–Fisher revisited: The case of fertility correlation. Theoretical Population Biology 62:181-197. https://doi.org/10.1006/tpbi.2002.1609 [our work]

Speidel et al 2019. Nat. Genetics 51:1321-1329. https://doi.org/10.1038/s41588-019-0484-x

The 1000 Genomes Project Consortium. 2015. Nat. 526:68-74. https://doi.org/10.1038/nature15393

Yelmen B, ..., Jay F. 2021. Creating artificial human genomes using generative neural networks. PLoS Genetics 17(2):e1009303. https://doi.org/10.1371/journal.pgen.1009303 [our work]

Merci d'enregistrer votre fichier au format PDF et de le nommer :
«ACRONYME de l'initiative/institut – AAP 2021 – NOM Porteur.euse Projet »

Fichier envoyer simultanément par e-mail à l'ED de rattachement et au programme :
cd_instituts_et_initiatives@listes.upmc.fr avant le 20 février.