Analyse de la distance génétique entre populations à l'échelle du génome

Contexte

En génétique des populations on s'intéresse à la caractérisation des populations (humaines, animales ou végétales) à partir de la distribution des allèles au sein de celles-ci. Cette distribution peut en particulier refléter le degré d'isolement des populations les unes par rapport aux autres, et/ou le niveau de pression évolutive subie par une population. Dès les années 40, Wright [1] et Malécot [2] ont proposé des indices statistiques pour mesurer la différenciation entre populations à partir de leur polymorphisme génétique. Le F_{ST} est actuellement l'indice le plus utilisé pour le calcul de distances ou de similarités entre populations. Il existe toutefois plusieurs définitions de cet indice, et il a récemment été montré que ces différentes définitions ne sont pas consistantes et reflétent des caractéristiques différentes de l'histoire de différentiation des populations [3]. Il existe un cadre statistique rigoureux dans lequel il est possible de formaliser les différentes notions de F_{ST} et de proposer des estimateurs de ces indices basés sur un ensemble de marqueurs génétiques. On peut ainsi réaliser l'inférence du F_{ST} à partir d'un échantillon d'individus issus des différentes populations et pour lesquels l'information au marqueur est disponible. Toutefois, la méthode utilise l'ensemble des marqueurs (quelle que soit leur localisation le long du génome) pour réaliser l'inférence, ce qui suppose que les différentes régions du génome reflètent la même histoire de différenciation, hypothèse peu vraisemblable en pratique.

Objectif du stage

L'objectif du stage est donc de développer une méthode d'inférence du F_{ST} locale. On cherche ainsi à réaliser simultanément deux tâches : d'une part identifier des régions du génome partageant un même historique de différenciation, et d'autre part caractériser en terme de différenciation les régions identifiées. Cette détection/caractérisation conjointe peut être réalisée à l'aide de méthodes de segmentation de signal [4]. L'application de ce type de méthodes à l'analyse de la différenciation pose toutefois des problèmes en terme de temps de calcul : le nombre de populations à analyser conjointement et le nombre de marqueurs caractéristant les différentes populations (de l'ordre de 10^6 marqueurs par chromosome chez l'humain par exemple) peuvent être importants.

Le stagiaire aura à réaliser les tâches suivantes :
- participer à la formalisation statistique du problème,
- implémenter et/ou appliquer les méthodes de segmentation existantes,
- proposer des améliorations algorithmiques afin d'accélérer les temps de calcul,
- réaliser l'analyse d'un jeu de données de populations humaines.

Compétences recherchées

Ce stage s'adresse aux étudiants en M2 de statistique/biostatistique. Une connaissance des méthodes d'inférence classiques en grande dimension (critères pénalisés et/ou méthodes régularisées) est requise. Les analyses seront réalisées à l'aide du logiciel R (ou Rcpp si besoin).

Unité d'accueil, ressources mises à disposition

Le stage se déroulera au sein de l'équipe Statistique & Génome du laboratoire de statistique d'AgroParisTech (site Claude Bernard, Paris 05). Il sera encadré par Tristan Mary-Huard et Guillem Rigaill. Le stagiaire disposera d'un ordinateur personnel et pourra utiliser les ressources informatiques (serveurs + cluster) de l'unité. Le stagiaire travaillera sur les données publiques issues du projet international "1000 Genomes" [5]. Le stagiaire percevra la gratification INRA. La durée du stage (entre 5 et 6 mois) et la date de commencement (mars/avril) peuvent être adaptées en fonction des contraintes du candidat.

Contact

Tristan Mary-Huard, UMR 518 AgroParisTech / INRA MIA, maryhuar@agroparistech.fr
Guillem Rigaill, UMR 9213/UMR1403, Institute of Plant Sciences Paris-Saclay guillem.rigaill@inra.fr

Références

[1] Wright(1949), *The genetical structure of populations*. Annals of Human Genetics.
[2] Malecot (1948), *Les Mathématiques de l'Hérédité*. Masson, Paris.
[3] Balding & Mary-Huard, *Estimation of Fst and tree inference under hierarchical population structure*, SMPGD 2018
[4] Auger and Lawrence (1989), *Algorithms for the identification of Segment Neighborhood*, Bulletin of Math. Biol.
[5] 1000 Genomes Project Consortium (2012), *An integrated map of genetic variation from 1,092 human genomes*. Nature.