Des Hommes et des mouches

Petit Guide, à l'usage des étudiants

Version 2015

Les données de polymorphismes au sein des populations font d'excellents marqueurs de la démographie et de la parenté des organismes qui les portent. Ainsi, nous tacherons d'inférer quelques conclusions sur la démographie et la parenté des populations d'humains et de mouches et à l'aide des polymorphismes caractérisés dans les projets

A) pilote des 1000 génomes humains (les données complètes étant trop grandes).
B) Les données drosophiles.

Humains : Le projet 1,000 génomes est accessible depuis http://www.1000genomes.org/. Sur le site ftp correspondant, dans le répertoire correspondant au projet pilote, allez chercher les données snp du sous-projet *exons* (/vol1/ftp/pilot_data/release/2010_07/exon/snps). Vous avez besoin de récupérer les fichiers *.sites.vcf, qui contiennent les informations sur les fréquences des différents polymorphismes pour diverses populations humaines (eg. CEU, CHB, YRI, etc…).

Drosophiles: Une partie des génomes de drosophiles séquencés est accessible depuis http://datadryad.org/resource/doi:10.5061/dryad.446sv.2/7.2/. Il faut récupérer le fichier vcf sur la page.

Dans une première partie, il faudra extraire les données génomiques et les résumer soit sous forme d'un spectre de fréquence (projet A), soit sous forme d'une matrice de distance (projet B).

Dans le projet A, vous ajusterez un modèle démographique aux données observées. Pour cela, vous utiliserez un simulateur qui permet de générer des spectres de fréquences dans différents scénarios. Vous ajusterez les spectres de fréquence théoriques à ceux observés pour inférer les paramètres du modèle.

Dans le projet B, vous reconstruirez la parenté moyenne qui existe entre les populations et vous la représenterez sous forme d'un arbre. Puis vous chercherez s'il existe des gènes qui présentent un patron différent de structuration (typiquement plus marqué) ou tenterez d'inférer les paramètres de migrations et de temps de divergence entre populations.

Projet A – démographie des populations.

Vous normaliserez les fréquences des SNPs par population pour 100 séquences et représenterez leurs distributions de fréquence dans chaque population.

Un peu de théorie : Soit SFS(i), la fraction des mutations ayant pour fréquence (i/n). Le spectre de fréquence est donc un vecteur ***SFS***, qui contient les abondances relatives de l'ensemble des fréquences des mutations présente dans

les génomes étudiés. Attention, si l'état ancestral est inconnu (ou suspect), il faut utiliser un spectre replié ou les mutations à fréquence i/n sont indiscernables des mutations à fréquences (n-i)/n : SFS(i)+SFS(n-i).

Rq : si la taille de la population est restée constante, l'attendu théorique est E[SFS(i)]=1/i. Ainsi, il y a beaucoup de singletons, deux fois moins de doubletons, etc. Ainsi l'interprétation du spectre de fréquence est difficile telle quelle, mais une petite transformation (sfs_plat(i) = i x SFS(i)) le rend « plat » si la taille de la population est de taille constante. S'il ne l'est pas, il faut proposer un modèle alternatif. Attention, si les fréquences i et (n-i/n) ne sont pas différenciées, il faut ajuster le facteur multiplicatif.

Observer les données et discuter des différences entre les populations d'hommes et de mouches.

Vous utiliserez ensuite un simulateur pour générer des spectres de fréquence sous différents modèles démographiques permettant d'expliquer les données. Vous pouvez utiliser in simulateur « maison » (archive SimulTrees). Compiler et utiliser le binaire *SiteFrequencySpectrum*. La documentation (brève) de ce binaire est donnée avec l'option –h

Vous ajusterez en priorité des modèles démographique à 1 seul paramètre. Soit des croissances linéaires ou exponentielles, soit une démographique instantanée ayant eu lieu il y a temps Tb (en temps coalescent, c'est à dire en Nc générations, où Nc est le nombre de chromosome). Des modèles à démographie plus complexe peuvent être envisagés.

Pour ajuster le modèle, vous pourriez minimiser par moindre carré pondérée en utilisant la métrique suivante :

$$\sum_i \frac{\left(SFS^{obs}(i) - SFS^{theo}(i)\right)^2}{SFS^{theo}(i)}$$

Examiner l'influence de la première case du spectre, SFS(1), qui pourrait contenir des erreurs de séquençage.

Discuter des paramètres inférés et de leur pertinence pour ce que l'on connaît de la démographie humaine.

Projet B – Parenté entre populations.

Comparer les fréquences des polymorphismes entre population et déduisez-en la parenté entre populations. Vous pourrez construire une distance entre les différentes populations en vous servant des différences de fréquence à chaque SNP. Refaire un 'arbre' de parenté entre les populations.

Des idées de distance ici :
http://www.uwyo.edu/dbmcd/molmark/lect06/lect6.html

Ensuite, vous pourrez entreprendre de rechercher les gènes qui présentent un patron de structuration différent du patron moyen. Pour cela, vous ferez la

mesure de distance entre populations gène par gène et chercherez une mesure adéquate permettant de trouver les intrus. L'idée du test de Lewontin-Krakauer (1973) basé sur le Fst est probablement une bonne première piste. Vous pourriez adapter la méthode à toute autre distance.

Alternativement, vous pourriez tenter de regarder la variance qu'il existe dans la distance entre les populations et en inférer des paramètres de migration en utilisant un simulateur. Si vous désirez entreprendre ce projet, je peux vous aiguiller sur un simulateur adéquat.