Exercices Génétique des populations et démographie

flora.jay@lri.fr

Exercice : Ancêtre commun le plus récent *(à faire chez vous)*
TMRCA= Time to Most Recent Common Ancestor
A partir du cours sur le coalescent donner l'espérance du TMRCA: E[TMRCA de n lignées]
Indice: Longueurs de branches, Tk, ...

Exercice : hétérozygotie
Soit un échantillon de taille 2 (s1, s2)
Hétérozygotie attendue en fonction de theta ?
Indices : Considérer la mutation et la coalescence comme deux processus concurrents ayant chacuns leur propre taux de succès. Si s1 et s2 coalescent avant qu'une mutation que peut-on dire de s1 et s2 ?

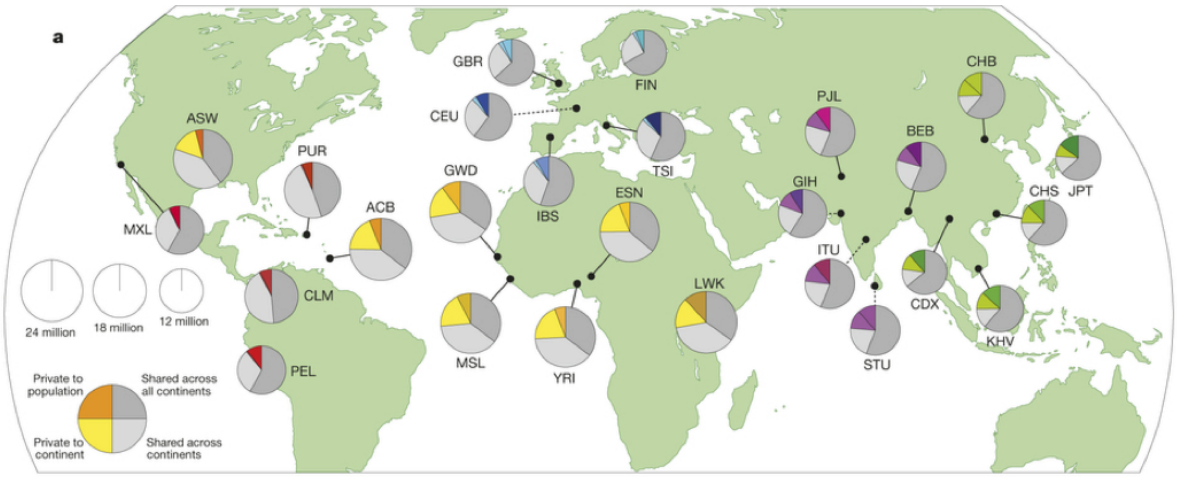
Exercice : Détecter le métissage entre populations (D-stats)
Construction par étapes d'un test pour détecter le flux de gène de Néandertal vers les populations non Africaines.

TP: Simuler la dérive génétique (si on a le temps)
(1) Implémenter une fonction genetic_drift qui simule la dérive génétique pour g générations à une position polymorphique ayant 2 allèles a/A
Arguments :
n taille de population
p fréquence initiale de A
g nombre de generations
Return :
vecteur de taille g+1 contenant la fréquence allélique de A à chaque pas de temps (le 1er élément étant la fréquence initiale)
Indice ?rbinom
(2) a. Tracer plusieurs courbes montrant la dérive génétique pendant 500 générations avec pour fréquence initiale p=0.5 et comme taille de population n=10000
b. Idem pour n=100. Discuter la différence des résultats

TP : Populations humaines de CG-Diversity Panel
Chromosome nº22 issu des données du Panel Diversity de CompleteGenomics
http://www.completegenomics.com/public-data/
Données de séquençage (génomes complets, couverture ~50X)
54 individus de populations réparties dans le monde
Détails sur les populations et pipelines voir
http://www.completegenomics.com/documents/PublicGenomes.pdf

ASW: African ancestry in Southwest USA

CEU: Utah residents with Northern and Western European ancestry from the CEPH collection
CHB: Han Chinese in Beijing, China
GIH: Gujarati Indian in Houston, Texas, USA
JPT: Japanese in Tokyo, Japan
LWK: Luhya in Webuye, Kenya
MKK: Maasai in Kinyawa, Kenya
MXL: Mexican ancestry in Los Angeles, California
TSI: Toscans in Italy
YRI: Yoruba in Ibadan, Nigeria
PUR: Puerto Rican in Puerto Rico



1000 Genomes Project Nature 2015

(1) Prise en main des données

Décompressez le jeu de données filename.tgz
dans le terminal : tar -xzf filename.tgz
Chargez les données dans un objet snp (utiliser read.table avec nrows=1000 pour commencer).
Si vous ouvrez CG_54genomes_indiv.txt vous verrez qu'il a une en-tête. Chargez le dans sample.info en utilisant read.table, l'argument header=T et stringAsFactors =F
Posez-vous des questions : head(snp)
Que représente ce jeu de données, qu'il y a-t-il sur une ligne ? A quoi correspondent les 5 premières colonnes ?
Combien de positions le long du génome sont stockées dans ce jeux ? tous des SNPs ?
Combien d'individus ? Pourquoi 5+108 colonnes ?
Que contient sample.info ?
Etc..

Enlever de *snp* les 5 premières colonnes contenant des informations supplémentaires (vous pouvez les stocker dans un autre objet). Désormais *snp* contient uniquement les marqueurs génétiques. Le codage de ces marqueurs :
0 = allèle ancestral
1 = allèle dérivé
L'ordre des colonnes (individus) correspond à l'ordre des lignes dans sample.info.

(2) Palette de couleur + Astuces R

(a) Copier dans la console R:

```
#### Setting up color palettes
upop = unique(sample.info[,"POP"])   # populations présentes dans le jeu
require(RColorBrewer)  # if possible otherwise replace brewer.pal by rainbow
# Construire un vecteur avec une couleur par population
colpop =  brewer.pal(length(upop),"Paired")  #rainbow(length(upop))
names(colpop) = upop
# Vecteur de couleurs: pour chaque individu, la couleur correspondant à sa population
sample.colpop = colpop[sample.info[,"POP"]]

# idem pour les régions:
ureg = unique(sample.info[,"REGION"])
colreg = brewer.pal(length(ureg),"Set1")       #rainbow(length(ureg))
names(colreg) = ureg
sample.colreg = colreg[sample.info[,"REGION"]]
```

(b) Astuces utiles pour le reste du TP

- En R vous pouvez construire des masques et sélectionner une partie seulement d'une matrice ou data.frame
  eg mat[, info==”EUR”] ou mat[, info %in% c(“JPT,”GIH”)]

- Afficher du texte plutôt que des points
  eg le nom de la population de chaque individu i pour sa valeur x_i
  #Soit x, un vecteur de taille n le nombre d'individu (ou une matrice de dim nx2)
  x=runif(nrow(sample.info)) # exemple
  plot(x, type="n")
  text(x, sample.info[,"POP"], col=sample.colpop)
  legend("topleft",names(colpop),pch=19,col=colpop,ncol=5)

(3) Hétérozygotie

- Calculer l'hétérozygotie pour chaque individu diploïde (individu 1 = colonne 1 et 2, individu = colonne 3 et 4, etc). Une fois les tests faits, pensez à recharger les données avec nrows=100000 par exemple.

- Comparer l'hétérozygotie dans différentes populations. Quelle information cela donne sur la démographie ? (cf Le 1er exercice sur l'hétérozygotie).
- Vous pouvez aussi les ordonner (*?order*) : il semble qu'il y ait une tendance avec les populations asiatiques ayant l'hétérozygotie la plus basse, puis les européens et enfin les populations africaines. Quel(s) modèle(s) pourrai(en)t expliquer cela ?

(4) Theta

- Ecrivez deux fonctions implémentant les estimateurs de Watterson (et de Tajima) de theta. (Tajima si vous avez le temps).
- Donnez les estimations de theta pour les individus de populations Africaines, puis de populations Européennes. Commentez les différences de temps de calcul entre estimateurs (se remarque si nsnp grand).
- *On suppose que le taux de mutation pour l'ensemble du chromosome est de mu ~ 35e06 * 2.5e-08 = 0.875 par génération (Note: ceci est un chiffre très approximatif correspondant à la taille du chromosome étudié x le taux de mutations par génération et par base ADN). Si vous n'avez chargé que 100,000 lignes utilisez mu = 0.27 (pour 10,000 lignes mu=0.035 mais l'estimation est moins fiable).*
  Estimez les tailles de populations pour les groupes "Européens" et "Africains".
  Discuter les limitations de cette approche, en particulier par rapport aux hypothèses du coalescent. Comment pensez-vous que le taux de mutation est estimé ?
  Que dire de la taille de population récente, pensez-vous pouvoir détecter l'expansion liée à l'adoption de l'agriculture au néolithique ?

(5) Spectre de fréquences alléliques (SFS)

- Implémentez une fonction compute_allele_counts qui prend en entrée une matrice comme *snp* et qui renvoie le compte des allèles dérivés pour chaque site, <u>en ignorant les sites non polymorphiques</u> (donc renvoie un vecteur de taille potentiellement inférieur à nsnps). ?rowSums
- Affichez le SFS (histogrammes des comptes) pour l'ensemble du jeu. ?hist avec option prob=T
- Rappelez l'attendu neutre du SFS (sous le modèle coalescent avec taille constante). Calculer le nombre attendu de singletons et comparez au nombre observé.
- Utilisez votre fonction et le principe du masque (voir astuces plus haut) pour afficher les SFS des différentes régions. Calculez le sfs attendu.
  Rq : la taille d'échantillon *n* est égale à sum(mask)
- (Tracer sur une même figure le SFS observé pour "AFR" et l'attendu.)
- Commentez la forme des SFS par rapport à l'attendu neutre.
- En théorie qu'est-ce qui peut être à l'origine d'un excès de singletons ? Et dans la réalité ?

(6) Analyse en composante principale (ACP)

L'ACP est une méthode de statistique multidimensionnelle qui permet de résumer des données en éliminant la redondance qui peut exister au sein des variables (ici les marqueurs

génétiques). L'ACP projette les individus dans un espace dont les axes sont décorrélés (contrairement aux marqueurs génétiques) . En conséquence visualiser les individus dans le nouvel espace défini par (PC1, PC2) (=axes de l'ACP) est bien plus informatif que dans un espace défini par (marqueur 1, marqueur 2) ! Si les individus sont éloignés sur ce graphe, c'est qu'ils sont plus éloignés génétiquement que d'autres (analyse un peu simpliste qui peut être discutée..).
?prcomp

- Appliquez l'ACP aux données (attention prcomp prend en entrée une matrice avec en ligne les individus et en colonnes les variables, ici les marqueurs génétiques, utilisez la fonction transpose t()) .
- Affichez les deux premiers axes (les valeurs transformées se trouve dans l'attribut x de l'objet retourné par prcomp. Utilisez la palette de couleur des populations (puis des régions) pour identifier l'appartenance de chaque point de votre plot).
- Interprétez le résultat en terme de démographie. Quelles relations entre populations?
- Discuter les similarités et différences avec les objectifs et les hypothèses de l'algorithme STRUCTURE (cours précédent).

(7) Métissage

- Que pensez-vous des positions des individus Mexicains (MXL) et Porto-ricains (PUR) dans les analyses précédentes (eg en terme d'hétérozygotie, sur l'ACP, ..) ? Quelle explication proposeriez-vous pour ce signal ?

- Des études récentes ont montré que les Maasai (MKK) sont plus proches des européens que les autres populations africaines (ie d(MKK,EUR) dû à des flux de gènes plus importants entre européens et populations Maasai.
  Au vu des analyses précédentes pouvez-vous commenter ? Aviez-vous détecté cela ? sur l'ACP peut-être ?
  Quelles analyses proposeriez-vous pour détecter ce métissage ?
  → Remarque : je n'ai pas vérifié sur le jeu de données, le signal est potentiellement trop faible