CURRENT DYNAMICS OF THE LAUSANNE POPULATION GENETICS

Etude de cas mathématiques appliqués à la biologie

Charlotte Daglish, Ema Janeckova, Jonathan Mignot, Loïc Brun, Nicolas Bonzon, Youri Markides

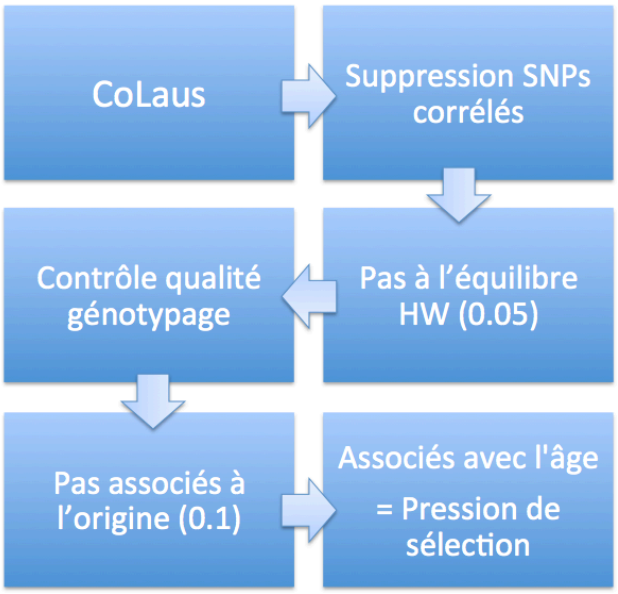
Supervisé par : Micha Hersch

INTRODUCTION

Contexte : L'étude CoLaus a génotypé plusieurs milliers d'individus habitants à Lausanne et alentours. Des données supplémentaires tels que l'origine, date de naissance et autres données médicales ont été collectées. Ces données ont essentiellement été utilisées pour étudier des aspects génétiques de certains traits physiologiques et médicaux, mais aussi pour certains aspects de la génétique de population locale.

Objectif : étudier différents SNPs pour déterminer si l'environnement et la société actuelle (nourriture plus grasse par exemple) ont un impact au niveau de notre génotype et chercher à comprendre à quel point et quelle vitesse le pool génétique de nos populations change face à l'impacte de la modernité de celles-ci.

METHODE



Détecter une pression de sélection par Hardy-Weinberg (HW): Selon le principe d'Hardy-Weinberg, dans une population idéale, les fréquences alléliques sont à l'équilibre au fil des générations. De plus, une population est considérée idéale si elle répond à aux critères suivants:

- Population de grande taille
- Espèce diploïde avec reproduction sexuée
- Panmixie
- Pas de migration
- Pas de pression de sélection

Ainsi, si certains SNPs ne sont pas à l'équilibre HW, cela signifie qu'ils ne remplissent pas une ou plusieurs de ces conditions. L'objectif étant de détecter une pression de sélection, les autres causes doivent être contrôlées au préalable. Les deux premières conditions (taille de la population et diploïdie) sont vérifiées. La panmixie ne peut pas être vérifiée. En effet, une association en fonction des phénotypes (et génotype) est probable. La migration a été vérifiée (données sur l'origine).

Vérification du génotypage: Afin de vérifier la qualité du génotypage, deux algorithmes ont été utilisés: Espérance-maximisation (EM) et K-means. Ces algorithmes vont vérifier que les SNPs génotypés sont bien répartis en trois groupes distincts. Les résultats suivants ont été observés:

- EM -> 10% des SNPs sont validés
- K-means -> 80% sont validés

Après un contrôle visuel de plusieurs génotypages sélectionnés aléatoirement, il s'est avéré que K-means est plus pertinent.

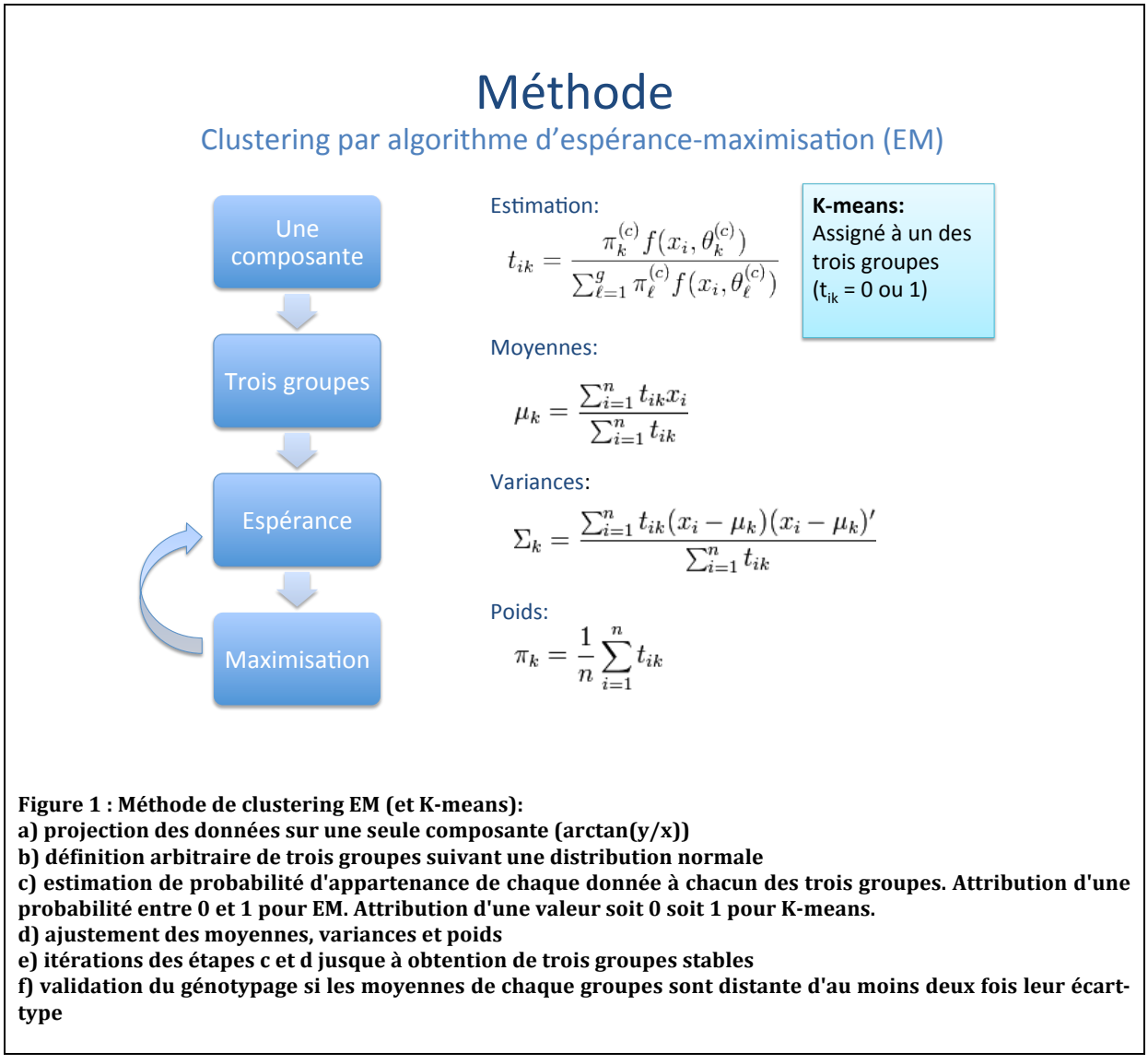


Figure 1 : Méthode de clustering EM (et K-means):
a) projection des données sur une seule composante (arctan(y/x))
b) définition arbitraire de trois groupes suivant une distribution normale
c) estimation de probabilité d'appartenance de chaque donnée à chacun des trois groupes. Attribution d'une probabilité entre 0 et 1 pour EM. Attribution d'une valeur soit 0 soit 1 pour K-means.
d) ajustement des moyennes, variances et poids
e) itérations des étapes c et d jusque à obtention de trois groupes stables
f) validation du génotypage si les moyennes de chaque groupes sont distante d'au moins deux fois leur écart-type

Variation due à l'immigration : Le but de cette sous-partie du projet est de déterminer parmi les SNPs ne se trouvant pas à l'équilibre de Hardy-Weinberg, ceux qui sont soumis à une pression de migration due à l'origine de ces personnes. Pour ce faire, l'analyse de composantes principales a été utilisée dans le but

de déterminer les deux principales composantes (PC1&PC2) représentant la variation génétique maximale.

Depuis la publication d'une étude[1] en 2008, il est établi que la variation génétique maximale (représenté par PC1 & PC2) est associée à la distance géographique des personnes.

Parmi une liste de SNPs donnés n'étant pas à l'équilibre de H-W et étant bien génotypés, il a été sélectionné avec un seuil de 10 % les SNPs avec la variation maximale sur PC1 et PC2. Cette variation représentant l'origine des personnes, ces 10 % ont été enlevés de la liste. Le 90 % restant a été transmis par la suite au groupe s'occupant de la corrélation avec la date de naissance, dans le but de voir quel(s) SNP(s) pourrai(en)t être actuellement soumis à une pression de sélection.

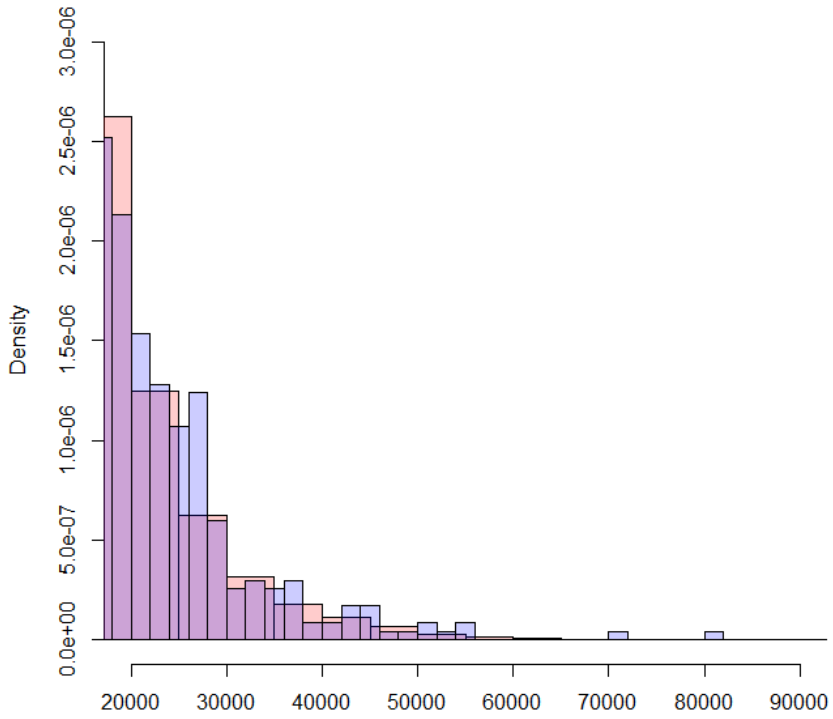


Figure 2; Densité en fonction des valeurs absolues des composantes de pc1. La couleur rose représente l'ensemble des SNP alors que la couleur bleue représente la liste donnée. Les parties en violet montre la superposition des deux cités plus haut. 1.3% des données totales sont montrées.

La figure 1 a montré clairement de la variance entre les deux groupes. Il a été enlevé comme mentionné les 10 % les plus extrèmes (représenté sur la partie droite de la figure 1). Sur la base visuelle de cette histogramme, un test de Levene a été effectué. Ce test permet de comparer la variance entre deux groupes (Tous les SNP + liste donnée dans notre cas). La p-valeur retournée étant largement significative, il a été conclu que les SNPs dont la variance était grande, étaient probablement sous pression par la migration.

Corrélation avec l'âge: L'objectif est d'analyser s'il y a une corrélation entre les génotypes des individus de Colaus et leur date de naissance. Cette corrélation indiquerait un changement de distribution des fréquences des génotypes dans la population au cours des générations. Ceci indiquerait une éventuelle pression de sélection actuellement active qui pourrait expliquer pourquoi les SNP en questions ne sont pas à l'équilibre de Hardy-Weinberg.

Nous avons utilisé la régression linéaire, qui permet de créer un modèle linéaire des valeurs attendues, selon l'hypothèse nulle qu'il n'y a pas de corrélation entre les génotypes et l'âge. Ceci permet ensuite de déterminer si nos données observées sont significativement différentes des valeurs attendues.

Un QQ-plot, permet de visualiser si les p-valeurs obtenues suite à cette régression, sont plus petites que les p-valeurs attendues.

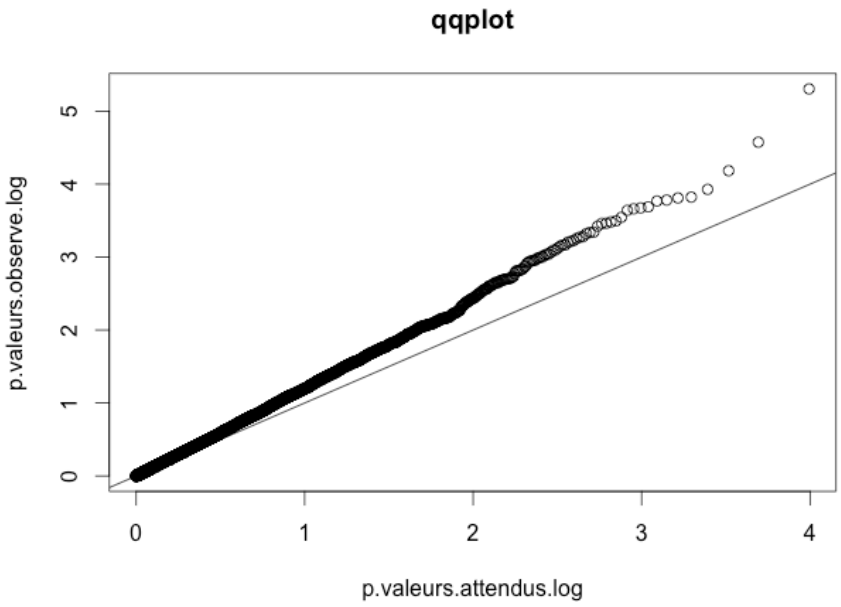


Figure 3: QQ-plot des p-valeurs observées en fonction des p-valeurs attendues

Résultats: Le QQ-plot nous indique que nous avons potentiellement quelque chose de significatif puisque les points extrêmes, correspondant à des petites p-valeurs, s'éloignent en dessus de la droite des valeurs attendues, montrant que ces valeurs sont plus petites que attendues.

Après avoir corrigé les p-valeurs obtenues pour des tests multiples, un seul SNP est significatif avec un seuil de 5% (p-valeurs=0.0488). Ce SNP se trouve dans le génome à proximité de trois gènes différents : SLC1A1 (code pour un transporteur de glutamate), SPATA6L (impliqué dans la spermatogenèse) et GLIS3 (protéine à 5 domaines doigt de Zinc, impliqué dans régulation de transcription et des mutations dans ce gène ont été associées au diabète).

Si ce SNP a une influence sur ces gènes, il est possible que cela indique qu'il y a une pression de sélection qui s'exerce sur cette région du génome.

Conclusion

Des données de CoLaus nous avons déterminé qu'un certain nombre des SNP n'étaient pas à l'équilibre d'Hardy-Weinberg. De ceux-ci certains étaient mal génotypés, d'autres était liés à la migration et un SNP indiquait une éventuelle pression de sélection.

Une suite possible de ce projet serrait d'effectuer la même analyse sur une autre population afin de déterminer si les résultats sont replicables. De plus, il serait intéressant de déterminer si le dernier SNP a réellement un effet sur les gènes aux alentours et s'il est lié à certains phénomènes actuels tels que baisse de fertilité masculine, diabète ou autres.

References

[1] Genes mirror geography within Europe. Nature 456, 98-101 (6 November 2008)