Projet GENOM 2023-2024
Données GWAS et variations phénotypiques

Sophie Garnier : sophie.garnier@sorbonne-universite.fr

INTRODUCTION

L'étude GHS (Gutenberg Health Study) est un projet collaboratif, initié en 2006, entre l'université Johannes-Gutenberg de Mayence (Allemagne) et l'INSERM. C'est une étude prospective dont le but est d'identifier les facteurs de risques de nombreuses maladies dont l'athérosclérose. A l'origine de bon nombre de maladies cardiovasculaires, l'athérosclérose se caractérise par le dépôt d'une plaque de lipides (athérome) sur la paroi des artères, entraînant par la suite sa lésion (sclérose).

3300 individus de la cohorte GHS, recrutés dans la population générale allemande et donc asymptomatiques, ont donc été testés, après avoir donné leur consentement écrit, pour plusieurs phénotypes cardiovasculaires et génotypés avec la puce *Affymetrix6.0*. GHS est une étude prospective, car les individus recrutés, initialement exempts de tout phénotype morbide, sont suivis régulièrement afin de détecter l'apparition éventuelle de ces derniers. Chez les individus développant de tels phénotypes on peut, *a posteriori*, étudier conjointement les phénotypes cardiovasculaires développés et les génotypes afin d'identifier des facteurs de risques aux maladies cardiovasculaires.

PROJET A

Le but de ce projet bio-informatique est de développer les scripts et outils permettant de filtrer les données de génotypage de la puce *Affymetrix6.0*, de tester l'homogénéité génétique entre les individus afin d'éliminer les « outliers », et enfin de réaliser l'étude d'association sur les variants et les individus testés pour deux phénotypes intermédiaires de l'athérosclérose, le taux de HDL-Cholestérol et le taux d'Apolipoprotéine A1, ApoA1. Les résultats bruts de ces diverses analyses pourront ensuite être affinés ou ajustés sur différents critères (âge, sexe, indice de masse corporel …), représentés graphiquement et interprétés. Les résultats devront ensuite être présentés oralement.

PROJET B

Ce 2ème projet consiste, à partir du même jeu de données de réaliser une analyse d'imputation. Les 1ères étapes, communes avec le premier projet, seront de nettoyer le jeu de données et de vérifier l'homogénéité génétique entre les individus afin d'éliminer les « outliers ». Par la suite, l'imputation proprement dite sera réalisée afin de récupérer les données génotypiques pour les marqueurs non présents sur la puce afin d'accroître l'informativité de l'analyse. Cette imputation se déroule en deux étapes, le phasage et l'imputation proprement dite. Suite à cette imputation, vous pourrez, si vous en avez le temps réaliser l'étude d'association sur les variants et les individus testés pour les deux phénotypes intermédiaires de l'athérosclérose, le taux de HDL-Cholestérol et le taux d'Apolipoprotéine A1, ApoA1.

DONNEES

- Données de génotypage pour les chromosomes 2, 5, 13 et 16, soit 190857 marqueurs pour 3300 individus, dans des fichiers .bim, .fam et .bed (format plink).
    - Les fichiers .bim donnent les informations sur les SNPs, leur identification, leur localisation et leurs allèles

Exemple de fichier bim					
Chr	Identifiant du SNP	Position en cM	Position en pb	Allèle 1	Allèle 2
9	rs10751931	0	49949	C	T
8	rs11252127	0	52087	A	C

    - Les fichiers .fam donnent les identifiants et le statut des individus. Ce sont les 6 premières colonnes d'un fichier.ped au format plink (identifiant de l'individu, identifiant de la famille, identifiant du père, de la mère, le sexe et statut)

Exemple de fichier fam					
Identifiant familial	Identifiant individuel	Père	Mère	Sexe	Statut
FAM1	NA06991	0	0	1	1
0	NA06993	0	0	2	1
0	NA06994	0	0	1	2

    - Les fichiers .bed contiennent les données génotypiques encodées en binaire. Il n'est pas possible de lire ces fichiers, mais on peut, en utilisant le logiciel plink, revenir à des data non encodées et donc lisibles. En ce cas, les fichiers bed, bim et fam seront recodés en fichiers .map et .ped. Les fichiers .map donnent les informations sur les SNPs, leur identification, leur localisation …, tandis que les fichiers .ped donnent pour un individu son identifiant, celui de sa famille, de son père, de sa mère, son sexe (1 pour un homme et 2 pour une femme) et son statut (1 pour un individu sain, 2 pour un individu atteint) suivi de ses génotypes pour tous les marqueurs testés du chromosome. Les données manquantes sont codées par 0.

Les données de génotypage sont également disponibles dans des fichiers au format .map et .ped. Les fichiers .map donnent les informations sur les SNPs, leur identification, leur localisation …, tandis que les fichiers .ped donnent les identifiants, statuts et génotypes des individus.

Le format habituel d'un fichier .ped se constitue de 6 colonnes standards (identifiant de l'individu, identifiant de la famille, identifiant du père, de la mère, sexe et statut) suivi des génotypes de l'individu pour tous les marqueurs testés du chromosome. Les données manquantes sont codées par 0, pour le sexe 1 correspond à un homme et 2 à une femme, pour le statut 1 à un individu sain, 2 à un individu atteint.

- Données phénotypiques, ici pour ApoA1 et HDL_Cholestérol, dans le fichier phenotype_plink.txt. Ce fichier comprend une colonne identifiant et les données phénotypiques pour le taux d'APOA1 et d'HDL_cholestérol. Selon les logiciels utilisés ou les programmes codés, le format de fichier peut varier. Il faudra donc éventuellement transformer ce fichier brut afin qu'il soit correctement reconnu.
- Pour le projet A notamment, une fois l'analyse d'association brute effectuée, et en fonction des orientations que vous voudrez donner à votre projet, d'autres données pourront être utilisées comme
    - Données de covariables (âge, sexe, indice de masse corporel ou BMI) pour des études avec ajustement diverses co-variables dans le fichier covar_plink.txt
    - Données d'expression pour les probes des régions chromosomiques mises en évidence par la GWAS …
- Pour le projet B, le phasage nécessite un panel de référence provenant de la dernière version du projet 1000 Genomes (http://csg.sph.umich.edu/abecasis/mach/ download/ 1000G.Phase3.v5.html).

Logiciels d'analyse

- Plink[1] : Traitement des données génotypées et analyse d'association sur ces données (http://zzz.bwh.harvard.edu/plink/ dataman.shtml#recode)
- Mach1[2,3] ou ShapeIT[4] ou … : Phasage, étape 1 de l'imputation
- Minimac2[5,6] ou Impute2[7] ou … : Imputation
- Mach2qtl[2,3] : analyse d'association sur données imputées
- R (graphiques …)
- Tout autre logiciel qui aurait votre préférence

Langage de programmation

Les étudiants restent libres de programmer dans le langage qui leur convient.

ETAPES

Le calendrier de travail proposé ci-après n'est qu'une suggestion. Certains d'entre vous iront peut être plus vite … ou moins vite, la totalité du projet ne sera peut être pas réalisé … l'important est ici que vous initiiez aux études d'association et aux étapes de prétraitements et d'ajustements nécessaire à l'analyse correcte des données tout en programmant par vous-même les outils qui vous seront nécessaires.

1ères séances, préparation des données pour l'analyse d'association ou l'imputation

Données de génotypage SNP

Les données de génotypage, fournies dans les fichiers .ped, sont les données brutes obtenues après le scan de la puce, sans qu'aucun filtre n'ait été appliqué. Pour réaliser une étude d'association, il convient tout d'abord de « nettoyer » ces données en les filtrant sur différents critères. C'est ce que l'on appelle le contrôle qualité (QC).

A vous de réfléchir aux critères de validité des data qu'il vous semble important de respecter puis de les présenter à un des enseignants qui les validera ou non. Une fois ces critères établis, il conviendra de réaliser le/les programme(s) permettant de filtrer les données comme vous l'avez décidé. A la fin de cette étape un premier fichier contenant la liste des SNPs à exclure de l'analyse peut être généré. En soustrayant ces SNPs aux fichiers initiaux on obtiendra alors des fichiers « propres » pour les génotypes, nommés par exemple chrx_QC.map ou .ped.

Que pensez-vous de la qualité des données ?

Du nombre de marqueurs génotypés et finalement analysés? …

Individus à inclure dans l'analyse

Une fois les fichiers de génotypage nettoyées, une autre étape de prétraitement des données consiste en l'étude de l'homogénéité de la population étudiée.

Afin, entre autre, d'éviter les biais de stratification (association fallacieuse due à des différences de fréquences alléliques chez les individus du fait de différentes provenances ethniques et non pas causées par des différences de statut des individus) il convient par exemple de vérifier que les individus testés sont globalement ressemblant en ce qui concerne leurs marqueurs étudiés c'est-à-dire qu'ils ont la même provenance ethnique, ici européenne.

Ceci peut, par exemple, être réalisé en lançant des matrices de distance sur chacun des 4 chromosomes pour comparer les individus deux à deux et repérer ceux qui s'écartent de la moyenne, les « outliers ».

A la fin de cette étape un fichier des individus à exclure peut être généré (indtoremove.txt). Ce fichier comportera, par exemple, les identifiants individuels et familiaux des individus à extraire sur deux colonnes avec des en-têtes ID et FID. En soustrayant ces individus aux fichiers initiaux on obtiendra des fichiers « propres » pour les individus.

Séances 3 et 4, analyse d'association ou imputation

Générations des fichiers « propres »

Une fois à votre disposition les fichiers corrigés pour les marqueurs et la liste des individus à exclure (chrx_QC et toremove), il convient de générer de nouveau fichiers « propres » .ped et .map pour les futures analyses d'association. Les fichiers initiaux sont à conserver pour d'éventuelles comparaisons de résultats avec les résultats des fichiers « nettoyés ».

Projet A

Analyse d'association

Une fois les jeux de données filtrés produits, les études d'association peuvent ensuite être lancées via le logiciel PLINK ou tout autre logiciel qui aurait votre préférence et les résultats présentés sous forme de graphique et/ ou de tableaux (Manhattan Plot sous R pour les résultats d'association …)

Analyse des résultats

Les résultats des analyses d'association sont des résultats bruts qu'il convient par la suite de raffiner, d'interpréter afin d'optimiser la présentation des résultats.

A vous de réfléchir aux meilleures façons d'analyser ces résultats en particuliers quels sont les « meilleurs » SNPs qui ressortent de l'analyse ? Sous quels critères pensez-vous qu'ils sont les meilleurs … Vous pouvez vous baser entre autres sur différentes représentations :

Manhattan Plot, QQ-plot, SnapShot, Haploview qui donne des informations sur le déséquilibre de liaison …

Projet B

Imputation

Une fois les jeux de données filtrés produits, le phasage et l'imputation des données peut être lancée (logiciels Mach1/minimac3 ou SHAPEIT/IMPUTE2 ou …) en utilisant le panel de référence europen du 1000Génomes pour les chromosomes 2, 5, 13 et 16.

Analyse d'association

Une fois les données imputées disponibles, si vous en avez le temps, les études d'association peuvent être lancées *via* le logiciel Mach2qtl puis les résultats présentés sous forme de graphique et/ ou de tableaux (Manhattan Plot sous R pour les résultats d'association …)

Analyse des résultats

Les résultats des analyses d'association sont des résultats bruts qu'il convient par la suite de raffiner, d'interpréter afin d'optimiser la présentation des résultats.

A vous de réfléchir aux meilleures façons d'analyser ces résultats en particuliers quels sont les « meilleurs » SNPs qui ressortent de l'analyse ? Sous quels critères pensez-vous qu'ils sont les meilleurs … Vous pouvez vous baser entre autre sur

- La représentation du Manhattan Plot
- QQ-plot
- SnapShot
- Haploview, qui donne accès à des informations sur le déséquilibre de liaison …

Séances suivantes

Optimisation des résultats en ajustant les données sur une ou plusieurs covariables (phénotype, âge, sexe, statut …), en excluant les meilleurs marqueurs de l'analyse ….

Si des régions chromosomiques sont mises en évidences on peut également s'interroger sur les gènes présents au sein de cette/ces région(s) (fonction, déséquilibre de liaison …), sur l'effet des marqueurs sur le niveau d'expression des gènes, sur l'existence d'une association haplotypique … Il peut également être intéressant d'étudier l'impact du contrôle qualité sur les résultats de l'analyse …

A vous de lancer autant d'analyses qu'il vous semblera nécessaire en ajustant sur les variables qui vous semblent pertinentes et de comparer les résultats de ces diverses analyses.
En fonction du temps dont vous disposez, et de vos envies !, vous pouvez vous partager le travail au sein du groupe et orienter chacun votre travail dans la voie qui vous convient
En fonction de l'avancée de votre projet il va peut être temps de commencer à mettre vos résultats au propre afin de nous les présenter ! N'oubliez pas d'interpréter les résultats, les statistiques, de produire des graphiques …

Suggestions de références bibliographiques

1. *Purcell S, et al. PLINK a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81:559–575*
2. *Li Y, et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34:816-834*
3. *Li Y, et al. Genotype Imputation. Annu Rev Genomics Hum Genet. 2009; 10:387-406*
4. *O. Delaneau, J. Marchini, JF. Zagury (2012) A linear complexity phasing method for thousands of genomes. Nat Methods. 9(2):179-81. doi: 10.1038/nmeth.1785*
5. *Fuchsberger C, et al. minimac2: faster genotype imputation. Bioinformatics 2015; 31, Issue 5 :782–784*
6. *Howie B, et al. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 2012; 44:955–959*
7. *Marchini J, et al. A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics. 2007; 39: 906-913*
8. *Laurie CC et al.Cathy C. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol. 2010 Sep; 34(6): 591–602.*
9. *Marees AT, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018 Jun;27(2)*
10. *https://bioconductor.org/packages/release/bioc/manuals/GWASTools/man/GWASTools.pdf*

Sites utiles

https://zzz.bwh.harvard.edu/plink/
https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html