Etude pangénomique des effets cis haplotypiques sur l'expression des gènes dans les monocytes humains

Sophie Garnier[1,2], Vinh Truong[1,2], The Cardiogenics Consortium, Tanja Zeller[3], Stefan Blankenberg[3], Willem H. Ouwehand[4,5], François Cambien[1,2], Alison H. Goodall[6,7], David-Alexandre Trégouët[1,2]

(1) INSERM UMR_S 937, Paris, France.
(2) ICAN Institute for Cardiometabolism And Nutrition, Pierre and Marie Curie University (UPMC, Paris 6), Paris, France.
(3) Department of General and Interventional Cardiology, University Heart Center Hamburg, Hamburg, Germany.
(4) Human Genetics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom.
(5) Department of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge, United Kingdom
(6) Department of Cardiovascular Sciences, University of Leicester, Leicester, United Kingdom.
(7) National Institute for Health Research Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, United Kingdom.

Overview:

This project aims at assessing whether gene expression variability could be influenced by the presence of more than one proximal (eg cis-acting) single nucleotide polymorphisms (SNPs).
In contrast to most previous studies looking at single SNP associations, we focused on the influence of several SNPs through additive or more complex haplotype combinations and undertook a systematic search of cis acting haplotypes expression quantitative trait loci (eQTL) on the entire human genome.
This strategy required the analysis of more than 2 billion haplotypic combinations. Our local cluster was not suitable to run this project in a timely manner and large computing resources were needed. We therefore took advantage of the power of the European grid (EGI) to run all the statistical analyses. Our results demonstrated that the monocyte expressions of a substantial proportion of the human genes were under the influence of multiple cis-acting SNPs. About 75% of the detected genetic effects were related to independent additive SNP effects and the last quarter due to more complex haplotype effects. Of note, twenty-four of the genes identified to be affected by multiple cis eSNPs have been previously reported to reside at disease-associated loci. This could suggest that such multiple locus-specific genetic effects could contribute to the susceptibility to human diseases.

Enjeux scientifiques, besoin en calcul, stockage et visualisation :

De nombreux travaux récents ont montré que le niveau d'expression d'un gène peut être influencé par des polymorphismes génétiques ("SNPs" pour Single Nucleotide Polymorphisms) situés à sa proximité ("cis eSNPs"), certains de ces polymorphismes pouvant également être associés à un phénotype comme un risque de maladie ou la variabilité de traits quantitatifs [1-6].

La majorité des études portant sur ces effets cis ne se base cependant que sur l'influence d'un seul SNP sur le gène alors que de manière générale de nombreux SNPs situés en cis peuvent être impliqués dans la variabilité de son expression. Les effets de chaque SNP sont parfois indépendants et dans ce cas là, une approche SNP à SNP est pertinente. Ce n'est cependant pas toujours le cas. Non seulement, il faut pouvoir distinguer les vrais effets des SNPs des effets qui seraient dus à un phénomène de dépendance entre deux SNPs (déséquilibre de liaison) mais il faut aussi pouvoir dissocier, lorsque les SNPs ne sont pas indépendants, les effets additifs d'effets plus complexes (parfois dus à l'influence d'un haplotype particulier). Dans ce cas, seules les études considérant plusieurs SNPs, les études haplotypiques [7], sont pertinentes.

Dans ce projet, nous avons utilisé une approche dite haplotypique, qui permet d'analyser l'effet d'un groupe de SNPs sur l'expression d'un gène et, pour la première fois, cette étude a été menée de manière systématique sur l'ensemble du génome humain. Pour cela, nous avons utilisé les données de l'étude *Cardiogenics ([11])*, un projet européen dans lequel 758 individus ont été génotypés pour environ 350 000 SNPs et pour lequel l'expression des gènes a également été mesurée. Le tableau 1 donne les principales caractéristiques de cette étude. Notons que les sondes sont utilisées pour mesurer le niveau d'expression d'un gène. Le nombre de sondes est plus élevé que celui des gènes car l'expression d'un gène peut être mesurée par plusieurs sondes.

Nombre d'individus	758
Nombre de SNPs	346 749
Nombre de gènes concernés	15 426
Nombre de sondes	19 805
Nombre d'analyses	2 097 693 183

Tableau 1 : Caractéristiques du projet Cardiogenics

Une analyse haplotypique pour un gène donné nécessite de prendre en compte toutes les combinaisons possibles entre les SNPs et d'étudier leurs effets sur l'expression du gène. En pratique, nous avons limité à 4 le nombre maximum de SNPs par haplotype en considérant des fenêtres de 30 SNPs. Par exemple, un ensemble de 70 SNPs engendre avec une telle stratégie 195 530 combinaisons.

Pour ce projet, nous avons analyser plus de 2 milliards de combinaisons. L'analyse de l'expression d'un gène pour une sonde dure de quelques secondes à plusieurs minutes selon le nombre de SNPs à étudier au sein du gène. En considérant qu'elle dure 3 secondes, il aurait alors fallu plus de 10 ans pour finir le projet sur un seul CPU. L'ensemble des analyses étant indépendantes entre elles, nous avons commencé à utiliser un cluster en exploitant l'ensemble des ressources (une vingtaine de CPUs). Mais cette solution ne nous permettait pas de finir le projet dans un temps raisonnable. Nous avons ainsi décidé d'utiliser une grille de calcul pour terminer nos expériences en quelques mois.

Développements, utilisation des infrastructures :

L'ensemble des expériences a été conduit sur la grille de calcul européenne EGI (European Grid Infrastructure) avec *biomed* comme organisation virtuelle. L'utilisation de la grille nous a permis d'accéder à la puissance de calcul nécessaire pour terminer le projet en moins d'un an. L'ensemble de nos expériences a finalement duré environ 8 mois.

L'ensemble des tests d'association a été réalisé avec le programme *combinHaplo [7].* Nous pouvons noter que ce logiciel est intégré à *GridHaplo* avec l'interface *Easy-gLite [10]*, qui a été conçu pour lancer ces calculs sur une grille de calcul. Cet outil a déjà été utilisé avec succès pour faire une recherche d'association sur le génome entier entre un ensemble de SNPs et la maladie d'Alzheimer [9] . Contrairement à cette précédente étude, notre projet s'intéresse à plusieurs phénotypes (l'expression des gènes) et aux SNPs se trouvant à proximité de chaque gène. *GridHaplo* nous a semblé moins adapté pour notre projet et avons donc développé nos propres scripts en Python que l'on va décrire succinctement.

Nous avons préalablement générer des fichiers pour chacune des sondes qui contiennent le niveau d'expression mesuré par la sonde ainsi que les allèles des SNPs se trouvant à proximité du gène. *combinHaplo* génère une liste indicée de combinaisons de façon déterministe et peut exécuter les tests d'association pour un sous-ensemble de combinaisons en spécifiant uniquement les indices correspondantes. Nous avons exploité cette caractéristique pour spécifier des blocs de combinaisons à tester.

Le script principal permet de faire exécuter une série de tests sur la grille de calcul. Elle exécute les étapes suivantes:

- sélectionner une sonde
- définir les blocs de combinaisons à tester
- spécifier les tâches à soumettre à la grille de calcul (génération automatique des fichiers jdl)
- soumettre les tâches à la grille

Une tâche correspond donc à un ensemble de tests d'association à exécuter. Pour ce projet, nous avons considéré au maximum 4500 tests par tâche. Ce nombre a été fixé empiriquement pour que la tâche puisse s'exécuter pendant la durée du proxy (c'est à dire 24h par défaut). Cependant, cela a impliqué un nombre conséquent de tâches à gérer sur la grille. Nous avons donc été obligés de les soumettre automatiquement et à espace régulier.

Un certain nombre de tâches n'ayant pas pu s'exécuter correctement, il nous a fallu concevoir des scripts permettant de vérifier l'intégrité et l'intégralité des fichiers résultats, de spécifier les blocs de combinaisons (et la probe associée) qui ont posé problème et de soumettre ces nouvelles tâches. Les fichiers résultats correctement générés ont ensuite été transférés

sur le cluster local pour traiter les données.

Outils, le cas échéant complémentarité des ressources, difficultés rencontrées :

Notre approche consiste à soumettre un grand nombre de jobs en automatisant le plus possible les différentes étapes. Une attention particulière a été faite pour ne pas noyer un élément de calcul sous un grand nombre de tâches. A chaque soumission, nous avons uniquement regardé les CE qui n'ont pas de tâches en attente et qui ont plus de 10 processeurs disponibles. Cette stratégie a permis d'exécuter un grand nombre de tâches soumises.

De plus, pour des raisons diverses, des éléments de calcul (CE) de la grille n'arrivaient pas à exécuter les tâches demandées. L'une des causes venait du système d'exploitation utilisé par les CE. Nous avons en effet utilisé une version compilée sous linux de *combin_haplo.* Malheureusement, nous n'avons pas identifié toutes les causes de ces échecs. Pour contourner ce problème, nous avons utilisé une solution simple : nous avons spécifié une liste noire de CE. Nous avons ainsi écarté 54 CE.

De plus, la gestion des proxys n'a pas été facile. Contrairement à d'autres travaux, nous n'avions pas réussi à utiliser les proxys de longues durées ou leurs renouvellements automatiques. En effet, de telles fonctionnalités peuvent représenter pour certaines VO des faiblesses au niveau de la sécurité de la grille. C'était, semble-il , le cas pour la VO *biomed* lorsque nous avons exploité la grille. Les proxys ont dû être générés régulièrement à la main. Cette contrainte a particulièrement compliqué la gestion et l'exécution des différentes tâches.

Résultats scientifiques :

Nous avons réalisé une approche haplotypique sur les données de *Cardiogenics* à la recherche de groupe de SNPs pouvant influer sur l'expression d'un gène à proximité. Comme nous nous intéressons à des effets dus à de multiples SNPs, nous avons fait une analyse SNP à SNP et sélectionné les sondes et les SNPs dont l'effet haplotypique est meilleur que l'effet d'un seul SNP. Cette phase a permis de retenir 13,4 % des sondes (2650). Une phase de réplication a permis de confirmer une partie de ces résultats (une centaine de sondes).

Au final, notre étude montre que l'expression d'un grand nombre de gènes peut être influencée par plusieurs SNPs à travers soit des effets additifs soit des effets plus complexes. Il est donc important de prendre en compte ce type de phénomène avant de se lancer dans des expériences expressions fonctionnelles.

Perspectives :

Une partie des sondes ,dont les expressions ont été trouvées influencées par plusieurs SNPs dans notre projet, ont été rapportées dans la littérature comme associées à des risques de maladie ou à des traits quantitatifs [8]. Ces résultats suggèrent donc que des effets haplotypiques pourraient également contribuer au risque de maladie.

Nous nous sommes limités à des haplotypes comportant au plus 4 SNPs or il peut exister des effets cis dus à des haplotypes de taille plus importante. Une recherche systématique paraît cependant difficile, même avec l'utilisation d'une grille de calcul. En effet, le nombre de combinaisons augmente exponentiellement avec le nombre de SNPs. Cela dit, une recherche plus localisée sur des gènes candidats pourrait être envisagée.

Références :

[1] Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39: 1208-1216.
[2] Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. Nat Genet 39: 1202-1207.
[3] Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217-1224.
[4] Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452: 423-428.
[5] Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6: e107.
[6] Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, et al. (2010) Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. PLoS One 5: e10693.
[7] Tregouet DA, Konig IR, Erdmann J, Munteanu A, Braund PS, et al. (2009) Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nat Genet 41: 283-285
[8] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.
[9] Lambert J-C, Grenier-Boley B, et al (2012) Genome-wide haplotype association study identifies the FRMD4A gene as a risk locus for Alzheimer's disease, Molecular Psychiatry
[10] http://genecanvas.ecgene.net/downloads.php?cat_id=1
[11] http://www.cardiogenics.eu