Recherche d'associations haplotypiques dans le cadre de la maladie d'Alzheimer

Jean-Charles Lambert, Benjamin Grenier-Boley

Recherche d'associations haplotypiques dans le cadre de la maladie d'Alzheimer

Jean-Charles LAMBERT(1), Benjamin GRENIER-BOLEY(2)
(1) jean-charles.lambert@pasteur-lille.fr, INSERM U744, Institut Pasteur de Lille, Université Lille-Nord de France.
(2) benjamin.grenier-boley@pasteur-lille.fr, INSERM U744, Institut Pasteur de Lille, Université Lille-Nord de France.

Overview :

Following a three-step genome wide haplotype association study contrasting different populations from seven European countries and totalling 9,938 Alzheimer's Disease (AD) patients and 15,745 controls, we identified a rare risk haplotype associated with AD risk (odds ratio=1.68 (95% Confidence Interval, [1.44-1.99], $P=1x10^{-10}$). This haplotype maps to the *FRMD4A* gene, a locus that had not been detected by previous genome-wide association studies relying on the univariate analysis of single nucleotide polymorphism analysis.

Enjeux scientifiques, besoin de la grille :

La maladie d'Alzheimer est une maladie neurodégénérative qui entraîne la perte progressive et irréversible des fonctions cognitives touchant environ 860000 personnes en France (2007). Dans la majorité des cas (99%), elle se déclare après 60 ans et est multi-factorielle, résultant de l'interaction de facteurs environnementaux (20-40%) et génétiques (60-80%). Le premier gène lié à la pathologie a été identifié en 1993 et concerne l'allèle ε4 du gène APOE [1]. Il a fallu attendre ces deux dernières années et les progrès technologiques dans le domaine de la génétique pour identifier de nouveaux gènes grâce notamment à des études Genome-wide Association studies (GWAs) ou bien à des méta-analyses de plusieurs de ces études [2,3,4,5,6]. Ces études GWAs permettent de comparer les différences de distribution des allèles entre les cas et les témoins à certains endroits du génome correspondant à des mutations ponctuelles (ou SNP pour Single Nucleotide Polyphormism).
Afin de trouver de nouveaux gènes liés au risque de développer la maladie, une approche Genome-wide Haplotype Association Study (GWHAs) a été effectuée. Cette approche, qui a déjà permis d'identifier avec succès de nouveaux gènes dans une autre pathologie [7], analyse non plus un SNP à la fois mais une combinaison de plusieurs SNPs (ou haplotype). Cette approche augmente donc grandement le nombre de combinaisons à tester (environ 37000000 contre 600000 pour une GWA) mais permet d'identifier des gènes potentiels indétectables par les méthodes classiques de GWAs ou de méta-analyses. Une estimation du temps de calcul nécessaire a mis en évidence une durée totale comprise entre 17 et 95 ans sur un seul CPU pour effectuer tous les calculs, rendant de ce fait l'utilisation d'une grille de calcul obligatoire.

Développements, déploiement sur la grille :

Afin de tester les combinaisons haplotypiques par cette approche, un logiciel open source, nommé combin_haplo, a été développé par David Tregouët [7,8]. De plus, afin de pouvoir utiliser celui-ci simplement sur une grille de calcul, il a aussi développé Gridhaplo couplé avec une interface Easy-gLite. Ce premier permet, en plus de réaliser les tests d'association haplotypique, de gérer facilement ses données pour pouvoir lancer ses calculs sur une grille (découpage des données, tri des résultats, outils de visualisation etc.). Quant à l'interface, elle permet de lancer ses tâches et d'assurer un

audit de façon facilitée sur des grilles utilisant gLite comme middleware (soumission des tâches, resoumission, récupération des sorties etc.).
Ces outils ayant déjà été testés et utilisés avec succès sur la grille EGI [7], aucun développement additionnel n'a été réalisé et le couple « Gridhaplo/Easy-gLite » a été utilisé tel quel.

Outils, difficultés rencontrées :
Les difficultés rencontrées ont été liées au nombre conséquent de tâches qu'il a fallu lancer sur la grille. En effet, afin de profiter au mieux de la puissance de la grille de calcul, chaque tâche était responsable d'un certain nombre de combinaisons à effectuer pour une durée d'environ quelques heures de calcul. Ceci a été possible car chaque combinaison haplotypique à tester est indépendante. L'inconvénient a été que le nombre de tâche à lancer a été conséquent, même si nos données ont été découpées au préalable par morceau de chromosome. L'interface ne gérant pas les « collections de jobs », chaque tâche a été lancée à la suite sur la grille prenant, à chaque fois, beaucoup de temps.
De plus, même si l'interface a aidé à la gestion des tâches, le nombre important de tâches ayant échouées pour diverses raisons et nécessitant une resoumission, a été difficile à gérer.
La durée totale mis pour réaliser tous les calculs a été d'environ 5 à 6 mois.

Résultats scientifiques :
Cette première phase de calcul a été effectuée sur une population européenne, EADI1 (European Alzheimer's Initiative 1), regroupant 2025 cas et 5328 témoins. Trois des gènes connus comme associés au risque de développer la pathologie (APOE, CR1 et BIN1) ont été retrouvés par cette approche et ce, même si ce dernier gène était indétectable sur cette population en utilisant une GWA classique. Cela confirme l'intérêt de cette approche. Après exclusion de ces 3 loci, 153 régions potentielles d'intérêt ont été sélectionnées suivant des critères à priori de significativité.
Dans une seconde phase, nous avons cherché à répliquer ces régions dans une population indépendante : GERAD1 (Genetic and Environmental Risk in Alzheimer's Disease 1). Même si aucune région n'a passé le seuil de significativité après correction de Bonferroni, 2 régions ont montré des résultats cohérents : même meilleure combinaison haplotypique, même force et même direction d'association. La première région se situe sur le chromosome 6, locus 6p23 (Odd ratio (OR) : 1.53 ; Intervalle de confiance à 95% (95%CI) : [1.31-1.79] ; $P=8.1x10^{-8}$). La deuxième, elle, se situe sur le chromosome 10, locus 10p13 (OR : 1.76 ; 95%CI : [1.44-2.15] ; $P=2.3x10^{-8}$).
La troisième étape a consisté à génotyper les SNPs de ces 2 combinaisons haplotypiques dans 5 autre études cas-témoins : Allemagne, Belgique, Espagne, Finlande et Italie pour un total de 4740 cas et 4181 témoins. La première région n'a pas montré d'association avec le risque de développer la maladie (OR : 0.89 ; 95%CI : [0.74-1.08] ; P=0.23) et de ce fait n'a donc pas été répliquée. Par contre, la deuxième région a montré une association significative dans ces 5 populations (OR : 1.55 ; 95%CI : [1.19-2.00] ; $P=9.2x10^{-4}$) et dans une méta-analyse pour les 7 populations (OR : 1.68 ; 95%CI : [1.43-1.96] ; $P=1.1x10^{-10}$).
Cette région sur le chromosome 10, qui inclus le gène FRMD4A, n'a jamais été identifiée par les études GWAs et par méta-analyse comme région d'intérêt pour le risque de développer la maladie. Les fonctions de ce gène sont encore mal connues mais les connaissances actuelles suggèrent que ce gène pourrait être un candidat potentiel dans le risque de développer la maladie d'Alzheimer.

Perspectives :

Tous les calculs effectués par cette approche constituent une solide base de données qui servira dans de nombreux autres projets du laboratoire comme complément d'information. Même si la totalité de la méthode ne sera pas réitérée, il est possible que certaines régions d'intérêt soient recalculées en augmentant le nombre de SNPs afin d'avoir un vision plus fine des combinaisons haplotypiques et leur association avec la maladie.

Avec l'avancée des technologies de séquençage à haut-débit, les données que nous avons à traiter sont toujours de plus en plus nombreuses et volumineuses (ex : exomes complets), obligeant certaines de nos futures analyses à être effectuées sur une grille de calcul.

Enfin, grâce au projet 1000 Genomes [9], nous pouvons imputer de nombreux SNPs supplémentaires dans la population EADI, augmentant ainsi le nombre de SNPs à étudier de 600000 à plus de 11 millions. La phase d'imputation en elle-même est très consommatrice en ressources et l'analyse des données produites nécessitent aussi l'utilisation de grilles de calculs ou de clusters.

Références :

[1] Saunders AM *et al.* Association of alipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* 43:1467-72 (1993).

[2] Lambert J-C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* 41 : 1094-99 (2009).

[3] Harold D. *et al.* Genome-wide association study identidies variants at CLU and PICALM associated with Alzheimer's disease. *Nat. Genet.* 41 : 1088-93 (2009).

[4] Seshadri S. *et al.* Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA* 303 : 1832-40 (2010).

[5] Hollingworth P. *et al.* Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat. Genet.* 43:429-35 (2011).

[6] Naj AC. *et al.* Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet.* 43 : 436-41 (2011).

[7] Tregouët D-A. *et al.* Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat. Genet.* 41 : 283-85 (2009).

[8] Site : http://genecanvas.ecgene.net/downloads.php?cat_id=1

[9] 1000 Genomes Project Consortium. A map of human variation from population-scale sequencing. *Nature* 467 : 1061-73 (2010).