« Exploration statistique autour des variants génétiques rares »

Lieu : Laboratoire de Mathématiques Jean Leray et Institut du thorax, Université de Nantes

Durée / Période : 5 à 6 mois (à positionner sur la période Mars-Septembre 2013)

Encadrants/contacts :

- Lise Bellanger, email : lise.bellanger@univ-nantes.fr
  Laboratoire de Mathématiques Jean Leray (LMJ) UMR CNRS 6629 et EA 4275, "Biostatistique, Recherche Clinique et Mesures Subjectives en Santé", Université de Nantes
- Christian Dina, email : christian.dina@univ-nantes.fr
- Matilde Karakachoff, email : matilde.karakachoff@univ-nantes.fr
  Epidémiologie Génétique Institut du thorax Inserm UMR 1087 / CNRS UMR 6291 Nantes

Indemnité de stage : 1/3 smic

Résumé

Ce stage sera appliqué au domaine de la génomique et portera sur l'identification, à partir de méthodes statistiques, de variants génétiques rares associée à une pathologie donnée.

Sujet détaillé

L'implication de variants génétiques fréquents dans les pathologies humaines a été confirmée par un grand nombre d'études d'association génome entier. Il apparaît cependant qu'une partie importante de l'héritabilité soit due à la présence de variants rares, qui auraient des effets génétiques plus forts. Tester l'association de ces variants est problématique du fait de leur faible fréquence dans la population. En effet, malgré un effet potentiellement important, le petit nombre d'observations (due à une fréquence basse) ainsi que le grand nombre de variants rares (et donc le grand nombre de corrections statistiques nécessaires – *problèmes des tests multiples*) ne permet pas d'obtenir une puissance suffisante pour discriminer entre vrais et faux positifs.

Un grand nombre de méthodes ont été développées afin de combiner les variants rares en des tests dits poolés. Les variants sont d'abord définis comme pertinents (par exemple en ne retenant que les polymorphismes non-synonymes). Ensuite, les variants pertinents sont combinés (par exemple en créant des catégories de porteurs d'au moins un allèle rare sur un ensemble M de polymorphismes génétiques) et le nouveau score est testé par rapport au phénotype. Le défi principal réside dans la définition des variants génétiques pertinents mais également dans la façon de les regrouper et de les combiner.

Alternativement, il est possible de développer des tests basés sur la similarité entre individus (U-statistiques et kernel). Ces tests permettent une plus grande flexibilité d'analyse, surtout lorsque les effets des allèles rares ne sont pas unidirectionnels.

L'objectif de ce stage est de développer une méthode pour tester l'association entre un certain phénotype et la présence des variants rares, en tenant compte aussi des variables confondantes démographiques, notamment l'origine géographique des individus. Les données à disposition pour le développement de ce stage consistent en un ensemble de génotypes des cas (atteints d'une certaine maladie) et des contrôles (non atteints) recrutés dans le cadre du projet RAC, un projet de recherche spécifique à l'Institut du thorax. Le génotype d'un individu est représenté par une séquence de marqueurs (SNPs - Single Nucleotide Polymorphism), à partir de ceci il est possible de calculer les variations génétiques entre les groupes de cas et de témoins et ainsi d'estimer l'existence d'une association entre la variation génétique et la pathologie. Parallèlement il faudra aussi travailler à la simulation d'une base des données de génotypes d'une population spécifique, en tenant compte d'une certaine structure génétique et démographique. Les données simulées seront utilisées pour la mise en place et vérification de la méthode.

Les activités du stage peuvent ainsi être résumées:

- Etat de l'art :
  - revue des différentes méthodologies proposées (tests poolés et U-statistiques appliquées aux études d'associations génétiques) ;
  - revue des programmes informatiques disponibles.
- Développement d'une nouvelle méthode de mesure de ressemblance entre individus en utilisant les U-statistiques, une statistique pour tester l'association entre un set de marqueurs SNP et le phénotype, tout en

ajustant pour les covariables, pour tenir compte de la stratification géographique de la population. Cette deuxième étape prévoit

- la simulation des données de génotypes pour le développement de la méthodologie et la comparaison des puissances statistiques entre la méthode mise en place et d'autres méthodes disponibles en littérature (référence à titre d'exemple Neale, 2011)
- l'application aux données réelles : données de génotypage d'une étude cas/control pour la recherche de variants génétiques associées au Rétrécissement aortique calcifié (RAC) (référence à titre d'exemple Köttgen, 2012)

Profil recherché

Cette proposition de stage s'adresse à un étudiant de niveau M2 compétent en statistique et en programmation informatique et motivé par les applications biologiques. Cet étudiant travaillera dans le cadre d'une collaboration interdisciplinaire sous la direction de Lise Bellanger (statistique - Laboratoire de Mathématiques Jean Leray (LMJ) UMR CNRS 6629 et chercheur associé à l'EA 4275, "Biostatistique, Recherche Clinique et Mesures Subjectives en Santé", Université de Nantes), Christian Dina et Matilde Karakachoff (Epidémiologie Génétique Institut du thorax Inserm UMR 1087 / CNRS UMR 6291 Nantes).

Ce stage sera susceptible d'être poursuivi dans le cadre d'une thèse en Statistique Appliquée à l'Epidémiologie Génétique à l'Institut du Thorax, co-encadrée par Lise Bellanger et Christian Dina (financement obtenu).

Bibliographie utile

Pooled association tests for rare variant identification

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. PLoS Genet. 2011 Mar;7(3): e1001322. doi: 10.1371/journal.pgen.1001322. Epub 2011 Mar 3. PubMed PMID: 21408211; PubMed Central PMCID: PMC3048375.

Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012 Sep;13(4):762-75. doi: 10.1093/biostatistics/kxs014. Epub 2012 Jun 14. PubMed PMID: 22699862; PubMed Central PMCID: PMC3440237.

Wang Y, Chen YH, Yang Q. Joint rare variant association test of the average and individual effects for sequencing studies. PLoS One. 2012;7(3):e32485. doi: 10.1371/journal.pone.0032485. Epub 2012 Mar 16. Erratum in: PLoS One. 2012;7(6). doi: 10.1371/annotation/9a117865-2fb1-4356-a0ec-fda016d471a1. PubMed PMID: 22468164; PubMed Central PMCID: PMC3309869.

Wan-Yu L**,** Boshao Z**,** Nengjun Y**,** Guimin G**,** Nianjun L. Evaluation of pooled association tests for rare variant identification. *BMC* Proceedings 2011, 5(Suppl 9):S118 doi: 10.1186/1753-6561-5-S9-S118

Köttgen A, Yang Q, Shimmin LC, Tin A, Schaeffer C, Coresh J, Liu X, Rampoldi L, Hwang SJ, Boerwinkle E, Hixson JE, Kao WH, Fox CS. Association of estimated glomerular filtration rate and urinary uromodulin concentrations with rare variants identified by UMOD gene region sequencing. PLoS One. 2012;7(5):e38311. doi: 10.1371/journal.pone.0038311. Epub 2012 May 31. PubMed PMID: 22693617; PubMed Central PMCID: PMC3365030.

U-stats

Wei Z, Li M, Rebbeck T, Li H. U-statistics-based tests for multiple genes in genetic association studies. Ann Hum Genet. 2008 Nov;72(Pt 6):821-33. doi: 10.1111/j.1469-1809.2008.00473.x. Epub 2008 Aug 6. PubMed PMID: 18691161; PubMed Central PMCID: PMC2574710.

Li M, Fu W, Lu Q. An aggregating U-Test for a genetic association study of quantitative traits. BMC Proc. 2011 Nov 29;5 Suppl 9:S23. doi: 10.1186/1753-6561-5-S9-S23. PubMed PMID: 22373246; PubMed Central PMCID: PMC3287858.

Li H. U-statistics in genetic association studies. Hum Genet. 2012 Sep;131(9):1395-401. doi: 10.1007/s00439-012-1178-y. Epub 2012 May 20. Review. PubMed PMID: 22610525; PubMed Central PMCID: PMC3419299.

Genome Wide Association Studies (GWAS)

Fox ER, Musani SK, Barbalic M, Lin H, Yu B, Ogunyankin KO, Smith NL, Kutlar A, Glazer NL, Post WS, Paltoo DN, Dries DL, Farlow DN, Duarte CW, Kardia SL, Meyers KJ, Sun YV, Arnett DK, Patki AA, Sha J, Cui X, Samdarshi TE, Penman AD, Bibbins-Domingo K, Buzková P, Benjamin EJ, Bluemke DA, Morrison AC, Heiss G, Carr JJ, Tracy RP, Mosley TH, Taylor HA, Psaty BM, Heckbert SR, Cappola TP, Vasan RS. Genome-Wide Association Study of Cardiac Structure and Systolic Function in African Americans: The Candidate Gene Association Resource (CARe) Study. Circ Cardiovasc Genet. 2012 Dec 28. [Epub ahead of print] PubMed PMID: 23275298.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007 Feb 22;445(7130):881-5. Epub 2007 Feb 11. PubMed PMID: 17293876.