

Identification d'interactions Gène×Environnement par méta-analyse en génétique des plantes.

Contexte

En génétique d'association on collecte sur une série d'individus (panel) des informations phénotypiques (eg leur taille ou une réponse physiologique) et des informations génotypiques (ie on "séquence" l'individu en différentes positions le long du génome - appelées marqueurs - pour savoir quels sont les allèles présents à ces positions). On analyse ensuite les données de la façon suivante : pour chaque marqueur on teste si la présence des différents allèles influe sur le phénotype (variable réponse). A l'issue de l'analyse, on récupère autant de probabilités critiques qu'il y avait de marqueurs (généralement quelques centaines de milliers / millions). Il existe pour ce type d'analyse des outils dédiés qui permettent l'identification des marqueurs influant le caractère en quelques minutes [Lippert et al., 2011, Zhou and Stephens, 2012]. Cette stratégie a déjà permis l'identification de régions géniques (appelées QTL) impliquées dans des maladies à forte prévalence comme le cancer chez l'humain, ou à l'identification de gènes influant sur la résistance au stress hydrique ou aux parasites chez les plantes.

Il existe cependant des limites à l'approche précédemment décrite. La première réside dans la capacité de détection de QTLs dont les effets sont faibles : augmenter la puissance de détection nécessiterait d'augmenter les tailles des panels, ce qui n'est pas réalisable en pratique pour des raisons économiques (coûts trop élevés) ou logistiques (impossible de recruter un grand nombre de patients dans un même hôpital, de cultiver différentes variétés d'une même espèce dans un seul lieu, etc.). Par ailleurs l'inclusion dans une même étude de mesures individuelles collectées dans des conditions hétérogènes (différents hôpitaux, différentes parcelles expérimentales) résulte en une modélisation statistique plus complexe, ce qui se traduit in fine par des coûts algorithmiques associés à l'analyse qui s'avèrent être prohibitifs.

Une stratégie a récemment émergé en génétique humaine, consistant à combiner les résultats issus de plusieurs analyses réalisées sur des panels différents. On dispose donc maintenant pour chaque marqueur d'une série de probabilités critiques (une par panel) qu'il s'agit de combiner pour identifier les marqueurs associés au phénotype. Cette stratégie, appelée méta-analyse, a fait l'objet de nombreux développements et applications en génétique humaine ces dernières années [Al Olama et al., 2014, Houlston et al., 2010, Pharoah et al., 2013]. Plus récemment cette stratégie a aussi été considérée avec succès en génétique des plantes [Battenfield et al., 2018, Zhao et al., 2019], mais il n'existe actuellement aucune méthode adaptée pour le contexte particulier de l'application des méthodes de méta-analyse dans le contexte d'identification d'interactions gène×environnement.

Objectif du stage

On souhaite donc adapter les stratégies usuelles de méta-analyse pour prendre en compte les spécificités des études portant sur un même panel d'individus (ici de lignées) évalué dans un ensemble d'environnements différents. Le stage consistera à développer une stratégie de méta-analyse prenant en compte la similarité entre panels étudiés et à évaluer les performances de la stratégie proposée (en terme de puissance de détection des marqueurs influant sur le caractère étudié, et de contrôle de faux positifs) à l'aide d'une étude de simulations. La procédure d'analyse devra être implémentée de manière efficace pour être appliquée à de gros jeux de données (panels d'individus génotypés pour $\approx 10^6$ marqueurs, évalués dans une vingtaine d'environnements). Dans un deuxième temps la méthodologie sera appliquée à plusieurs jeux de données publiques portant sur une espèce modèle (*Arabidopsis*) et deux espèces d'intérêt économique (maïs et blé).

Compétences recherchées

Bonne maîtrise du modèle linéaire.
Programmation en R, notions de Rmarkdown.
Intérêt pour les applications statistiques en génétique.

Unité d'accueil, ressources mises à disposition

Le stage se déroulera dans l'UMR de statistique d'AgroParisTech (Paris 5ème) et/ou dans l'UMR de Génétique végétale du Moulon (Gif-sur-Yvette), et sera encadré par Tristan Mary-Huard (chargé de recherche INRA en statistique). Le stagiaire disposera d'un ordinateur personnel et pourra utiliser les ressources informatiques (serveurs + cluster de calcul) de l'unité. Le stagiaire percevra la gratification INRA. La durée du stage (entre 5 et 6 mois) et la date de commencement peuvent être adaptées en fonction des contraintes du candidat. Le stage pourra potentiellement être suivi par une thèse.

Contact

Tristan Mary-Huard, UMR 518 AgroParisTech / INRA MIA, maryhuar@agroparistech.fr

Références

- [Al Olama et al., 2014] Al Olama, A. A., Kote-Jarai, Z., Berndt, S. I., Conti, D. V., Schumacher, F., Han, Y., Benlloch, S., Hazelett, D. J., Wang, Z., Saunders, E., et al. (2014). A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nature genetics*, 46(10) :1103.
- [Battenfield et al., 2018] Battenfield, S. D., Sheridan, J. L., Silva, L. D., Miclaus, K. J., Dreisigacker, S., Wolfinger, R. D., Peña, R. J., Singh, R. P., Jackson, E. W., Fritz, A. K., et al. (2018). Breeding-assisted genomics : Applying meta-gwas for milling and baking quality in CIMMYT wheat breeding program. *PLoS one*, 13(11) :e0204757.

- [Houlston et al., 2010] Houlston, R. S., Cheadle, J., Dobbins, S. E., Tenesa, A., Jones, A. M., Howarth, K., Spain, S. L., Broderick, P., Domingo, E., Farrington, S., et al. (2010). Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26. 2, 12q13. 13 and 20q13. 33. *Nature genetics*, 42(11) :973.
- [Lippert et al., 2011] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10) :833.
- [Pharoah et al., 2013] Pharoah, P. D., Tsai, Y.-Y., Ramus, S. J., Phelan, C. M., Goode, E. L., Lawrenson, K., Buckley, M., Fridley, B. L., Tyrer, J. P., Shen, H., et al. (2013). Gwas meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nature genetics*, 45(4) :362.
- [Zhao et al., 2019] Zhao, J., Sauvage, C., Zhao, J., Bitton, F., Bauchet, G., Liu, D., Huang, S., Tieman, D. M., Klee, H. J., and Causse, M. (2019). Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. *Nature communications*, 10(1) :1534.
- [Zhou and Stephens, 2012] Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7) :821.