François O, Frichot E (2015). LEA : Un package R pour la génomique du paysage. *Quatrièmes Rencontres R*, Grenoble 24-26 juin 2015.

LEA : Un package R pour la génomique du paysage

O. François **E. Frichot**

Université Grenoble-Alpes
Laboratoire TIMC-IMAG, UMR CNRS 5525
38042 Grenoble cedex
olivier.francois@imag.fr

Mots clefs : Génétique du paysage, Génomique des populations, Analyse de structure, Etude d'association écologique.

Comprendre les bases moléculaires de l'adaptation est une étape fondamentale en biologie de l'évolution, en écologie moléculaire, ou en biologie de la conservation [1]. Depuis Darwin, nous savons que l'adaptation locale induit des changements évolutifs dans les populations par le biais de la sélection naturelle. En génomique du paysage, les signatures moléculaires de l'adaptation locale peuvent être détectées en identifiant les fréquences d'allèles présentant une association significative avec certains gradients écologiques. De telles études d'association, dites pan-génomiques, proposent de cribler des millions de génotypes individuels. Dans ces méthodes, les facteurs écologiques d'intérêt englobent par exemple des variables climatiques, des descripteurs de l'habitat tels que l'altitude, ou la densité de pathogènes et varient spatialement.

Le programme `LEA` cherche à faciliter les études portant sur la génétique du paysage et les études d'association écologique. Le programme permet tout d'abord d'effectuer des analyses de la structure génétique d'une population. De plus, il permet d'effectuer des balayages génomiques dans le but d'identifier les gènes potentiellement impliqués dans la réponse adaptative des organismes à l'environnement. Le package `LEA` incorpore des méthodes statistiques pour l'estimation des coefficients d'ascendance individuelle à partir de matrices de génotypes, et des méthodes pour l'identification de polymorphismes génétiques présentant une corrélation significative avec des variables environnementales. Le package `LEA` est principalement construit à partir de programmes en langage C optimisés pour traiter de très gros volumes de données génomiques.

Les outils bioinformatiques mettant en œuvre des tests d'association écologique s'appuient sur des modèles de régression linéaire généralisée ou des modèles linéaires mixtes. Ces derniers modélisent les facteurs de confusion générés par les patrons d'isolement par la distance, la structure de la population ou le contexte génomique, qui induisent de nombreux faux positifs. Les méthodes existantes ont l'inconvénient de nécessiter des pré-traitements complexes pour analyser de la structure de la population, ainsi que des post-traitements pour contrôler le taux de fausse découverte ou visualiser les résultats. Le programme `LEA` permet aux utilisateurs d'effectuer les traitements mentionnés ci-dessus à partir de la ligne de commande de `R` en utilisant une interface unique [2]. Le package optimise la vitesse algorithmique et l'allocation de mémoire tout en préservant la flexibilité des analyses statistiques utilisant `R`. Les fonctions traitent des données génomiques massives à partir de la ligne de commande de `R` sans charger la mémoire du programme.

Dans cet exposé, nous présentons un cadre intégré pour les analyses de structure génétique d'une population et pour les études d'association écologique. Nous décrivons le programme `LEA` qui fournit des méthodes de correction pour les tests multiples et génère des sorties graphiques pour les résultats. Plus spécifiquement, `LEA` est une boîte à outils contenant des méthodes d'estimation de coefficients d'ascendance individuelle par analyse en composantes principales (ACP) ou par factorisation de matrice non-negative (`snmf`, [3]). `LEA` contient de plus des méthodes corrélatives basées sur des modèles mixtes à facteurs latents (LFMM, `lfmm`, [4]). Les modèles de type LFMM permettent la correction des biais dûs à la structure des populations ainsi qu'à d'autres facteurs de confusion non-observables.

En conclusion, l'intérêt majeur du package `LEA` est de permettre à ses utilisateurs d'effectuer des analyses informatiques pour des jeux de données massives, tout en bénéficiant des méthodes statistiques et de visualisation disponibles à partir de `R`. Le détail des options de `LEA` sont disponibles dans des fichiers de documentation en ligne et dans un tutoriel Web. `LEA` peut être installé à partir du site `http://www.bioconductor.org`.

Références

[1] Schoville, S. D., Bonin, A., François, O., Lobreaux, S., Melodelima, C., Manel, S. (2012). Adaptive genetic variation on the landscape: methods and cases.*Annual Review of Ecology, Evolution, and Systematics*, 43, 23-43.
[2] Frichot, E., François, O. (2015). LEA: An R package for landscape and ecological association studies. To appear in *Methods in Ecology and Evolution*.
[3] Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., François, O. (2014). Fast and efficient estimation of individual ancestry coefficients.*Genetics*, 196(4), 973-983.
[4] Frichot, E., Schoville, S. D., Bouchard, G., François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, 30(7), 1687-1699.