tess3r : un package R pour l'estimation de la structure génétique des populations spatialisées

K. Caye[a], **O. Michel**[b] et **O. François**[c]

[a]Université de Grenoble-Alpes
Laboratoire TIMC-IMAG, UMR CNRS 5525
38042 Grenoble Cedex
kevin.caye@imag.fr

[b]Université de Grenoble-Alpes
Laboratoire GIPSA-lab, UMR CNRS 5216
38042 Grenoble Cedex
olivier.michel@gipsa-lab.grenoble-inp.fr

[c]Université de Grenoble-Alpes
Laboratoire TIMC-IMAG, UMR CNRS 5525
38042 Grenoble Cedex
olivier.francois@imag.fr

Mots clefs : Génétique des populations, Analyse de structure, Balayage pangénomique.

Une étape importante lors de l'étude de données de grande dimension est la recherche d'une structure de faible dimension mettant en évidence les principales caractéristiques des données. Cette étape est très importante en génétique des populations. En particulier elle est utile pour détecter les signatures laissées par l'histoire démographique de la population étudiée. De nombreuses méthodes ont été développées pour estimer la structure génétique des populations dans des échantillons où les individus sont issus du métissage de plusieurs sous-populations [2]. La plupart des jeux de données issus d'espèces naturelles présentent une forte autocorrélation spatiale. Quand elle est disponible il est donc intéressant d'utiliser l'information spatiale pour estimer la structure de population. Cela permet de représenter la répartition spatiale des sous-populations de manière continue. De plus, on sait que l'adaptation des individus à leur milieu induit des différences de fréquence allélique entre les différentes sous-populations. Il est possible d'utiliser l'estimation de la structure de population pour détecter des gènes potentiellement sous sélection environnementale. Enfin il est important de noter que la taille des jeux données récoltés par les généticiens des populations à largement explosée ces dernières années. Il est donc nécessaire que les outils d'analyse s'adaptent à cette tendance.

Le package tess3r permet d'estimer les coefficients de métissage individuels et les fréquences alléliques de chaque sous-population à partir de la matrice de génotype et des coordonnées spatiales des individus. La méthode repose sur un problème de factorisation de matrice régularisé par un graphe afin de garantir la continuité spatiale de l'estimation des coefficients de métissage individuels [1]. Cette méthode permet d'avoir un temps de calcul de l'ordre de l'heure sur des matrices allant jusqu'à 10^3 individus et 10^6 gènes. Le package propose des fonctions pour projeter sur une carte les coefficients de métissage ainsi calculés. À partir de l'estimation des fréquences alléliques de chaque sous-population et des coefficients de métissage individuels, le package tess3r calcule une statistique de test de type F_{ST} pour chaque gène. Cette statistique peut être utilisée pour effectuer un balayage pangénomique dans le but de détecter des gènes potentiellement responsables d'une adaptation à l'environnement. En résumé, l'intérêt du pack-

age tess3r est de permettre une estimation efficace de la structure des populations spatialisées et d'une statistique F_{ST} en s'intégrant dans l'environnement de travail R. L'utilisateur peut alors profiter des autres packages de R pour la visualisation des résultats ou encore la gestion des formats de données.

Références

[1] Caye, C., Deist, T. M., Martins, H., Michel, O., Francois, O. (2016). TESS3: Fast Inference of Spatial Population Structure and Genome Scans for Selection. *Molecular Ecology Resources*, **16**(2), 540–548

[2] Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., Francois, O. (2014). Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics*, **196**(4), 973–983