pcadapt: an R package for performing genome scans for selection based on principal component analysis

Eric Bazin*[1], Michael Gb Blum[2], and Keurcien Luu[2]

[1]Laboratoire d'écologie alpine (LECA) – CNRS : UMR5553, Université Joseph Fourier - Grenoble I, Université de Savoie – bat. D - Biologie 2233 Rue de la piscine - BP 53 38041 GRENOBLE CEDEX 9, France

[2]Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG) – CNRS : UMR5525, Université Joseph Fourier - Grenoble I – Domaine de la Merci - 38706 La Tronche, France

Abstract

We introduce the R package pcadapt that provides list of candidate genes under selection based on population genomic data. The package is fast and can handle large-scale data generated with next-generation technologies. It works at the individual scale and can handle admixed individuals because it does not assume that individuals should be grouped into populations. It returns a list of P-values, which provides the opportunity to control for the false discovery rate. The statistical method implemented in pcadapt assumes that markers that are excessively related with population structure, as ascertained with principal component analysis, are candidates for local adaptation. The package computes vectors that measure associations between genetic markers and principal components. For outlier detection, a vector of associations is then transformed into a test statistic using Mahalanobis distance. Using simulated data, we compared the false discovery rate and statistical power of pcadapt to the ones obtained with Flk, Outlflank and BayeScan. For data simulated under an island model, we find that all software provide comparable results. However, in a model of divergence between populations, we find that BayeScan is too liberal, Outflank and BayeScan are too conservative, and pcadapt provide intermediate results. In terms of running time, we find that pcadapt is the fastest software of all included in the comparison. Because pcadapt can handle molecular data generated with next sequencing technologies, we anticipate that it will be a valuable tool for modern analysis in molecular ecology.

Keywords: genome scan, SNP, PCA

*Speaker