Performing Highly Efficient Genome Scans for Local Adaptation with R Package pcadapt Version 4

Florian Privé,*[,1,2] Keurcien Luu,[2] Bjarni J. Vilhjálmsson,[1] and Michael G.B. Blum[2,3]

[1]National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark

[2]University of Grenoble Alpes, Laboratoire TIMC-IMAG, UMR 5525, La Tronche, France

[3]OWKIN France, Paris, France

***Corresponding author:** E-mail: florian.prive.21@gmail.com.

Associate editor: Michael Rosenberg

Abstract

R package pcadapt is a user-friendly R package for performing genome scans for local adaptation. Here, we present version 4 of pcadapt which substantially improves computational efficiency while providing similar results. This improvement is made possible by using a different format for storing genotypes and a different algorithm for computing principal components of the genotype matrix, which is the most computationally demanding step in method pcadapt. These changes are seamlessly integrated into the existing pcadapt package, and users will experience a large reduction in computation time (by a factor of 20–60 in our analyses) as compared with previous versions.

Key words: **R package, local adaptation, efficient.**

We developed the R package pcadapt as a method to detect signs of local adaptation in genetic data (Luu et al. 2017). There are two main functions in the package: read.pcadapt which makes sure the data are in the right format and pcadapt which performs computations. The pcadapt method first computes the Principal Component Analysis (PCA) of a scaled genotype matrix. It then regresses all variants onto the resulting PCs to get a matrix of *Z*-scores (i.e., one *Z*-score for each variant and each PC). Then, it computes robust Mahalanobis distances of these *Z*-scores to integrate all PCA dimensions in one multivariate distance for each variant (Luu et al. 2017). These distances approximately follow a chi-squared distribution, which enables derivation of one *P*-value for each genetic variant. In essence, the pcadapt method tests how much each variant is associated with population structure, assuming that outlier variants are indicative of local adaptation.

Previous versions of package pcadapt used format "pcadapt", which is a text file of characters separated by spaces where each line stores all samples' genotypes for one variant (0, 1, 2, and 9 for missing values). It was also possible to convert from "ped" and "vcf" files to format "pcadapt". In pcadapt v4, the preferred format is now the PLINK "bed" format (Purcell et al. 2007). Format "bed" is very compact; it stores each genotype using only 2 bits, making it eight times smaller than the corresponding "pcadapt" file. Moreover, format "bed" can be memory-mapped to be used in both R and C++ almost as a standard R(cpp) matrix; see, for example, R package BEDMatrix that provides matrix-like accessors to "bed" files (Grueneberg and de los Campos 2019). Format "bed" is also advantageous because the widely used software PLINK can be used to convert "ped" and "vcf" files to "bed" files, and to perform quality control (Chang et al. 2015). For an existing "pcadapt" file, function read.pcadapt now creates a new file with extension ".pcadapt.bed" to be used by the main function pcadapt. Updating to version 4 is seamless for the user.

Previous versions of pcadapt computed the eigen decomposition of the Genetic Relationship Matrix (GRM). In pcadapt v4, we use the implicitly restarted Arnoldi method (IRAM), which is both fast and accurate for computing first PCs (Lehoucq and Sorensen 1996; Abraham et al. 2017; Privé et al. 2018). To compare the performance of the newest version of R package pcadapt (v4.1.0 here) with previously published versions (v3.0.4 here), we use publicly available data on 4,342 domestic dogs genotyped at 144,474 variants after quality control (Hayward et al. 2016). Function pcadapt v3.0.4 takes 2,111 s (35 min) to complete for this data. The bulk of this time is spent computing the GRM ($O(N^2P)$, where N is the number of samples and P is the number of variants). With pcadapt v4.1.0, it takes only 35, 60, and 102 s to run for $K = 5$, 10, and 20 PCs, respectively (fig. 1, $\sim O\,(NPK)$). This represents a 60-, 35-, and 20-folds improvement in computation time.

Linkage disequilibrium (LD) can confound genome scans in admixed populations (Price et al. 2008; Abdellaoui et al. 2013; Galinsky et al. 2016). In previous versions of pcadapt, the only way to deal with this problem was to reduce K, the number of principal components (PCs) used by pcadapt, to include PCs capturing population structure only. In version 4, we have added an option for performing LD clumping that removes variants in LD. This ensures that more PCs capture population structure instead of LD structure (Privé et al. 2018).

Overall, version 4 of R package pcadapt is much more efficient in terms of disk space, memory, and time requirements compared with its previous versions. This enables the analysis of large genotype data sets using a personal laptop.

Brief Communication

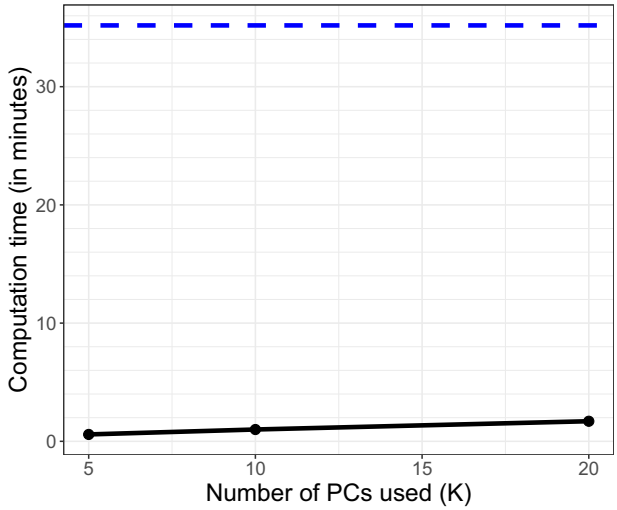

FIG. 1. Computation times of pcadapt as a function of the number of PCs used. Black points represent timings with pcadapt v4, and the dashed blue line represents the timing with v3, which is independent of the number of PCs used.

Moreover, using the PLINK "bed" format instead of format "pcadapt" makes pcadapt easier to use along with PLINK, which can be used for file conversion and quality control. Therefore, pcadapt v4 supports all formats that can be converted to the "bed" format, such as "vcf" and "ped". As for existing "pcadapt" files, they are seamlessly converted to format "bed" when using function read.pcadapt, which makes the new version backward compatible. Most users analyzing small data sets will likely not even notice the changes in version 4. Moreover, results are expected to be very similar (see supplementary note, Supplementary Material online). However, as sample sizes continue to grow, the importance of computational efficiency and robust methods also grows, and this is exactly what we address in version 4 of pcadapt.

Software and Code Availability

R package pcadapt is available on CRAN. It also has a GitHub repository where you can open issues (https://github.com/bcm-uga/pcadapt/issues). A tutorial on using pcadapt to detect local adaptation is available at https://bcm-uga.github.io/pcadapt/articles/pcadapt.html. The code used in this paper is available at https://github.com/bcm-uga/pcadapt/tree/master/new-paper/code.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

F.P. and B.J.V. are supported by the Danish National Research Foundation (Niels Bohr Professorship to John McGrath). The authors thank Katherine Musliner for her feedback on text.

References

Abdellaoui A, Hottenga J-J, De Knijff P, Nivard MG, Xiao X, Scheet P, Brooks A, Ehli EA, Hu Y, Davies GE, et al. 2013. Population structure, migration, and diversifying selection in the netherlands. *Eur J Hum Genet*. 21(11):1277–1285.

Abraham G, Qiu Y, Inouye M. 2017. FlashPCA2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics* 33(17):2776–2778.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4(1):7.

Galinsky KJ, Bhatia G, Loh P-R, Georgiev S, Mukherjee S, Patterson NJ, Price AL. 2016. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am J Hum Genet*. 98(3):456–472.

Grueneberg A, de los Campos G. 2019. BGData—a suite of R packages for genomic analysis with big data. *G3 (Bethesda)* 9(5):1377–1383.

Hayward JJ, Castelhano MG, Oliveira KC, Corey E, Balkman C, Baxter TL, Casal ML, Center SA, Fang M, Garrison SJ, et al. 2016. Complex disease and phenotype mapping in the domestic dog. *Nat Commun*. 7(1):10460.

Lehoucq RB, Sorensen DC. 1996. Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM J Matrix Anal Appl*. 17(4):789–821.

Luu K, Bazin E, Blum MG. 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol Ecol Resour*. 17(1):67–77.

Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, Ge D, Rotter JI, Torres E, Taylor KD, et al. 2008. Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet*. 83(1):132–135.

Privé F, Aschard H, Ziyatdinov A, Blum MGB. 2018. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* 34(16):2781–2787.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 81(3):559–575.