BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–2

PSIKO2: a fast and versatile tool to infer population stratification on various levels in GWAS

Andrei-Alin Popescu [1]*and Katharina T. Huber [1]

[1]School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Genome-Wide Association Studies are an invaluable tool for identifying genotypic loci linked with agriculturally important traits or certain diseases. The signal on which such studies rely upon can however be obscured by population stratification making it necessary to account for it in some way. Population stratification is dependent on when admixture happend and thus can occur at various levels. To aid in its inference at the genome-level, we recently introduced PSIKO and comparison with leading methods indicate that it has attractive properties. However uptil now it could not be used for local ancestry inference (LAI) which is preferable in cases of recent admixture as the genome level tends to be too coarse to properly account for processes acting on small segments of a genome. To also bring the powerful ideas underpinning PSIKO to bear in such studies, we extended it to PSIKO2 which we introduce here.

Availability: Source code, binaries, and user manual are freely available at `https://www.uea.ac.uk/computing/psiko`.

Contact: Katharina.Huber@cmp.uea.ac.uk,
Andrei-Alin.Popescu@uea.ac.uk

1 INTRODUCTION

A major confounding factor in Genome-Wide Association Studies (GWAS) is population stratification, that is, reproductive isolation of a sampled population. A powerful way to account for it is to assume that the genotype of each individual (generally called an accession and represented in terms of a Single Nucleotide Polymorphism (SNP) sequence) in a study is an admixture of genotypes of $K \geq 2$ (generally unknown) founder (populations). This admixture can then be expressed in terms of a dataset's principal components (PCs) or its population stratification matrix (i. e. its Q-matrix) which indicates for each accession of a study the proportion of its genotype that came from each of the K founders. Contrary to leading tools such as EIGENSTRAT (Price *et al.*, 2006) which only infers a dataset's PCs and STRUCTURE (Pritchard *et al.*, 2000) (and its extension to FASTSTRUCTURE (Raj *et al.*, 2014)), ADMIXTURE (Alexander *et al.*, 2009), and SNMF (Fricho *et al.*, 2014) which only infer a dataset's Q-matrix, PSIKO (Popescu *et al.*, 2014) is able to infer both. Furthermore, comparison of PSIKO against competing methods suggest that whilst the quality of its Q-matrices is on par with those generated by them, PSIKO has better scaling properties. However, uptil now PSIKO could not be used for local ancestry inference (LAI) which is important for applicatons ranging from human population studies to identification of disease causative loci (Brisbin *et al.*, 2012). Furthermore it could only be used on a LINUX platform and the efficiency of its PCA-step was not benchmarked. PSIKO2 rectifies these drawbacks.

*to whom correspondence should be addressed

2 FEATURES

PSIKO combines linear-kernel PCA with least-squares optimisation to quickly infer the PCs, number of founders, and Q-matrix of a dataset. Its successor PSIKO2 signigficantly extends it by also allowing for LAI and usage of PSIKO within a Mac environment.

2.1 PCs and number K of founders of a dataset

To obtain a dataset's PCs and thus estimates for its K, we perform a PCA-analysis. Rather than using standard PCA, we employ linear-kernel PCA (see e. g. Murphy (2012)) due to its good scaling properties in terms of the number of variables (SNPs in our case).

2.2 Q-matrix inference

We return the Q-matrix for the PCA-reduced dataset as Q-matrix for a given dataset. To obtain that matrix, we combine properties of PCA relating to simplices observed in e. g. Ma and Amos (2012) and Patterson *et al.* (2006) with an iterative least squares approach.

2.3 Local Ancestry Inference

We combine a sliding window approach and information about founder genotypes to map, for each individual of a dataset, each such window to one of the K founders. The window size is chosen by the user and the mapping is closely related to the one used by PCADMIX (Brisbin *et al.*, 2012). Contrary to PCADMIX which requires information about founder genotypes as input and thus cannot be used in its absence, this input is optional for PSIKO2. For datasets where this information is not available we infer it from the estimate of the Q-matrix it found for it (see Supplement for details).

3 IMPLEMENTATION AND USAGE

Released under a GPL license, PSIKO2 is command-line based and takes as input a genotype matrix in the form of the widely used .geno file format (Price *et al.*, 2006). It is written in C++ and comes with directly linked binary executable files that should work on all

modern Linux platforms/Mac environments. Alternatively, the user may compile the program himself if all required libraries are present on their system (see user manual for details). Its output can be used to either inform a study in terms of a dataset's local ancestry, PCs, K, and Q-matrix (which can then be graphically represented in terms of a barplot using R (R Core Team, 2014)) or as input to approaches such as STRUCTURE (in the form of e. g. a starting point for estimation of that number), or packages for association mapping such as TASSEL (Bradbury *et al.*, 2007), BOLT-LMM (Loh *et al.*, 2015), and FAST-LMM (Lippert *et al.*, 2011).

4 TESTING AND PERFORMANCE MEASURE

To ensure the correctness of PSIKO2's predecessor regarding K and Q-matrix estimation, we rigorously tested it in (Popescu *et al.*, 2014) on both real biological and simulated data. This testing did not include scrutinizing the efficiency of our implementation of linear-kernel PCA. We rectify this here by comparing PSIKO2 against an implementation of that strategy in the freely available SKLEARN software (Pedregosa *et al.*, 2011). Furthermore, to better understand PSIKO2's performance with regards to LAI, we used an approach similar to the one described in (Brisbin *et al.*, 2012) to compare it against PCADMIX. See Supplement for details for both.

4.1 Linear-kernel PCA

Key to PSIKO's improved performance in (Popescu *et al.*, 2014) with regards to large (in the number of SNPs) datasets common with NGS data, is a compact representation of genotype data. One of the reasons for this is owing to the fact that SNPs can only attain one of three possible values (i. e. 0, 1 or 2) allowing PSIKO to efficiently store them using a bitwise implementation. In the context of linear-kernel PCA, this also allows for quick dot product computation which PSIKO requires for finding a dataset's PCs (see Supplement). Using the R^2 correlation coefficient between the PCA-reduced datasets generated by the two kernel-PCA methods (where we generated the input datasets as described in (Popescu *et al.*, 2014)) as asessment meassure, we found R^2 to be larger that 0.999 for all simulated datasets indicating that both methods produce, to all extends and purposes, identical output with differences attributable to floating point arithmetic precision errors. However PSIKO's runtime was a fraction of that of SKLEARN (see Table 1), with SKLEARN running out of memory for sequences of length 2.5M.

	100K	250K	1M	2.5M
PSIKO	4.0/36	10.5/36	35/833	75/1192
SKLEARN	71.4/2436	337.9/10284	1415/39891	NA / >63000

Table 1. Runtime comparison (in seconds)/memory consumption (in megabites) of PSIKO and SKLEARN.

4.2 LAI algorithm

To assess PSIKO2's suitability for LAI, we used simulated datasets which we generated by combining the dataset provided by PCADMIX with the methodology described in (Brisbin *et al.*, 2012). We considered two main scenarios. In the first we provided PSIKO2 with founder (genotypes) and thus our results are directly comparable with those reported for PCADMIX in (Brisbin *et al.*, 2012). In the second, we withheld that information rendering PCADMIX inapplicable as it requires that information as input. Using the Q-matrix estimated by PSIKO2 and taking as proxy for the founders all accessions which had more than 91% of their genome originating from the same population, this dataset did not pose a problem for PSIKO2 .

In both cases, the performance of PSIKO2 was notable with it correctly reporting within, less than a second, the ancestry of 91.2%/91.1% (first/second scenario) of the loci under consideration for the input dataset. This is of the same quality as the results that PCADMIX obtained for a dataset with similarly diverged founders.

In summary, we believe PSIKO2 to hold great promise for population stratification correction on various genomic levels.

Acknowledgment A. -A. P. thanks the Norwich Research Park for support. Both authors thank the referees for helpful comments. The research presented was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia.

This is a pre-copyedited, author-produced PDF of an article accepted for publication in Bioinfromatics following peer review. The version of record [insert complete citation information here] is available online at: xxxxxxx [insert URL that the author will receive upon publication here].

REFERENCES

Alexander, D.H. et al (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research,* **19** (9), 1655–1664.

Bradbury, P.J. et al (2007) Tassel: software for association mapping of complex traits in diverse samples. *Bioinformatics,* **23** (19), 2633–2635.

Brisbin, A. et al (2012) Pcadmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human Biology,* **84** (4), 343.

Fricho, E. et al (2014) Fast inference of admixture coefficients using sparse non-negative matrix factorization algorithms. *Genetics,* **196** (4), 973–983.

Lippert, C. et al (2011) FaST linear mixed models for genome-wide association studies. *Nature Methods,* **8**, 833–835.

Loh, P.R. et al (2015) Efficient Baysian mixed-model analysis increases association power in large cohorts. *Nature Genetics,* **47**, 284–290.

Ma, J. and Amos, C.I. (2012) Principal components analysis of population admixture. *PLoS ONE,* **7** (7), e40115.

Murphy, K.P. (2012) *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, Massachusetts.

Patterson, N. et al (2006) Population structure and Eigenanalysis. *PLoS Genet,* **2** (12).

Pedregosa, F. et al (2011) Scikit-learn: machine learning in Python. *Journal of Machine Learning Research,* **12**, 2825–2830.

Popescu, A.A. et al (2014) A novel and fast approach for population structure inference using kernel-pca and optimisation (PSIKO). *Genetics,* **198** (4), 1421–1431.

Price, A.L. et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics,* **38** (8), 904–909.

Pritchard, J.K. et al (2000) Inference of population structure using multilocus genotype data. *Genetics,* **155** (2), 945–959.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

Raj, A. et al (2014) fastSTRUCTURE: variational inference of population structure in large SNP datasets. *Genetics,* **197** (2), 573–589.