Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genetic and population analysis

IPCAPS: an R package for iterative pruning to capture population structure

Kridsadakorn Chaichoompu[1*], Fentaw Abegaz Yazew[1,2], Sissades Tongsima[3], Philip James Shaw[4], Anavaj Sakuntabhai[5,6], Luísa Pereira[7,8], and Kristel Van Steen[1,2*]

[1]GIGA-Medical Genomics, University of Liege, Avenue de l'Hôpital 11, 4000 Liege, Belgium, [2]WELBIO (Walloon Excellence in Lifesciences and Biotechnology) [3]Genome Technology Research Unit, National Center for Genetic Engineering and Biotechnology, 113 Thailand Science Park, Phahonyothin Road, Khlong Neung, Khlong Luang, Pathum Thani 12120, Thailand, [4]Medical Molecular Biology Research Unit, National Center for Genetic Engineering and Biotechnology, 113 Thailand Science Park, Phahonyothin Road, Khlong Neung, Khlong Luang, Pathum Thani 12120, Thailand, [5]Functional Genetics of Infectious Diseases Unit, Institut Pasteur, 25-28, rue du Docteur Roux, 75015 Paris, France [6]Centre National de la Recherche Scientifique, URA3012, Paris, France [7]Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Rua Alfredo Allen, 208 | 4200-135 Porto, Portugal and [8]Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Rua Júlio Amaral de Carvalho, 45 | 4200-135 Porto, Portugal

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Resolving population genetic structure is challenging, especially when dealing with closely related populations. Although Principal Component Analysis (PCA)-based methods and genomic variation with single nucleotide polymorphisms (SNPs) are widely used to describe shared genetic ancestry, improvements can be made targeting fine-level population structure. This work presents an R package called *IPCAPS*, which uses SNP information for resolving possibly fine-level population structure. The *IPCAPS* routines are built on the iterative pruning Principal Component Analysis (*ipPCA*) framework to systematically assign individuals to genetically similar subgroups. Our tool is able to detect and eliminate outliers in each iteration to avoid misclassification. It can be extended to detect subtle subgrouping in patients as well. In addition, *IPCAPS* supports different measurement scales for variables used to identify substructure. Hence, panels of gene expression and methylation data can be accommodated.

Availability and implementation: The R package *IPCAPS* is required R version >= 3.0.0 and is open source under the GPL >= 2 license. It is freely available from bio3.giga.ulg.ac.be/ipcaps.

Contact: kridsadakorn.chaichoompu@ulg.ac.be and kristel.vansteen@ulg.ac.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Single Nucleotide Polymorphisms (SNPs) can be used to identify population substructure, but resolving complex substructures remains challenging (Neuditschko *et al.*, 2012). Owing to the relatively low information load carried by single SNPs, usually thousands of them are needed to generate sufficient power for effective resolution of population strata due to shared genetic ancestry (Price *et al.*, 2006). Moreover in practice with high-density genome-wide SNP datasets, linkage disequilibrium (LD) and haplotype patterns are likely to exist, which can be

exploited for the inference of population structure (Lawson *et al.*, 2012). On the one hand, exploiting haplotype patterns is potentially informative, but comes with a high computational burden. On the other hand, although removing LD by pruning strategies can eliminate some spurious substructure patterns (Price *et al.*, 2006), it may limit our ability to identify subtle subgroupings.

The identification of substructure in in a genome wide association study sample of healthy controls or patients is a clustering problem. Conventional population structure analyses use Bayesian statistics to show relationships amongst individuals in terms of their so-called admixture profiles, where individuals can be clustered by using ratios of ancestral components, see also (Corander *et al.*, 2008). The iterative pruning Principal Component Analysis (*ipPCA*) approach differs from this paradigm as it assigns individuals to subpopulations without making assumptions of population ancestry (Intarapanich *et al.*, 2009). At the heart of *ipPCA* lies performing PCA with genotype data, similar to EIGENSTRAT (Price *et al.*, 2006). If substructure exists in a principal component (PC) space (ascertained using, for instance, Tracy-Widom statistics (Intarapanich *et al.*, 2009), or the EigenDev heuristic (Limpiti *et al.*, 2011)), individuals are assigned into one of two clusters using a 2-means algorithm for which cluster centers are initialized with a fuzzy c-means algorithm. The test for substructure and clustering is performed iteratively on nested datasets until no further substructure is detected, i.e. until a stopping criterion is satisfied. The software developed to perform *ipPCA* has some shortcomings though. Notably, it is limited to a MATLAB environment, which is not freely available. Also, outliers can severely disturb the clustering analysis. These limitations are addressed in *IPCAPS*, which improves the power of fine-level population structure, while appropriately identifying and handling outliers.

2 DESCRIPTION

The *IPCAPS* package implements unsupervised strategies that facilitate the detection of fine-level structure in samples, extracted from informative genetic markers. For general populations, information regarding substructure may come directly from SNPs. For patient samples, general population structure should first be removed via regressing out ancestry informative markers prior to clustering. The latter is incorporated in *IPCAPS*. Currently, *IPCAPS* accepts three data input formats: text, PLINK binary (BED, BIM, FAM), and RData (more details in Table S1). In the sequel, we will assume the availability of a sufficiently large SNP panel that is called on a collection of population samples.

Prior to clustering with *IPCAPS*, adequate data quality control (QC) steps need to be taken. These are not supported by *IPCAPS* itself but can easily be performed in PLINK (1.9) (Purcell and Chang). Suggested PLINK parameters include: restrict to founders (*--filter-founders*), select chromosome 1-22 (*--not-chr 0,x,y,xy,mt*), perform LD pruning (*--indep-pairwise 50 5 0.2*), test for Hardy–Weinberg equilibrium (*--hwe 0.001*), use call rate at least 95% (*--mind 0.05*), filter out missing SNP above 2% (*--geno 0.02*), and remove low minimum allele frequency (*--maf 0.05*). The remaining missing genotype values are SNP-wise imputed by medians.

Rather than performing 2-means clustering in PCA-space, at each iteration, *IPCAPS* clustering potentially involves the consecutive application of 2 clustering modules. The first, which we call rubikClust, is applied in the 3-dimensional space determined by the 3 first principal components (axes) at an iteration step. It involves applying rotations in 3D by consecutively performing rotations around PC1, PC2, PC3, and may provide >2 clusters (more details in Fig. S2). Notably, this approach also allows for a rapid identification of outliers. When samples cannot be divided into 2 groups in this way, the existing R function *mixmod* (package *Rmixmod*) is used for latent subgroup detection. In particular, earlier computed PCs (untransformed) at a particular iteration are subjected to multivariate Gaussian mixture modeling and Clustering EM (CEM) estimation (Lebret *et al.*, 2015), allowing for up to three clusters at each iteration. The iterative loop of *IPCAPS* can be terminated automatically by calling one of three possible stopping criteria: the number of subgroups is lower than a minimum, the fixation index (F_{ST}) is lower than a threshold, and *EigenFit* is lower than a pre-specified cutoff. The *EigenFit* criterion is defined by the differences between the logarithms of consecutive eigenvalues, sorted from high to low.

All *IPCAPS* results are saved in a single directory including textual information about cluster allocations, and visual information such as PC plots and hierarchical trees of group membership. Due to memory restrictions in R, large datasets (i.e., large number of subjects) may need to be split in multiple files and loaded into computer memory via the *IPCAPS* option *files*, after which they are internally merged again for iterative PCA. Extra attention is paid on efficient PC calculation (Clayton, 2015), also relying on the R package *rARPACK*. To further improve computational efficiency, *IPCAPS* supports multithreads for parallel analyses, relying on the R package *doMC*. The R package I*PCAPS* provides one synthetic dataset and 5 functions; *ipcaps*, *cal.PC.linear*, *fst.hudson*, *fst.each.snp.hudson*, and *plot.3views*. The R package reference manual is provided as supplementary File S3.

3 APPLICATION

We simulated genotype data for 10,000 independent SNPs and 760 individuals belonging to one of three populations (250 individuals each) and 10 outliers. The pairwise genetic distance between populations was set to F_{ST}=0.005 (Balding and Nichols, 1995). Ten outlying individuals were generated by replacing the 1st and the 2nd eigenvectors by extreme values, and then the SNP matrix was reconstructed using the singular value decomposition formula (Liu *et al.*, 2013). Two-dimensional PC plots of the first 3 PCs only reveals a separation between populations (with overlap) for PC2 versus PC3 (Fig. S4). However, application of *IPCAPS* on the simulated data and thus flexible use of PC information and clustering stopping rules as described before, could clearly identify sample substructure (Fig. S5). Non-outlying individuals were correctly assigned to their respective subgroups. In a real-life data application, we considered four populations of HAPMAP (CEU, YRI, CHB, and JPT). These populations have been considered before in the evaluation of non-linear PCA to detect fine substructure (Alanis-Lobato *et al.*, 2015). After data QC as described before, 132,873 SNPs and 395 individuals remained. Using classic PCA, visualizing data into two-dimensional space based on the first two PCs is not enough to fully describe substructures. Whereas non-linear PCA is able to provide a hierarchical visualization with only the first 2 PCs, as claimed by the authors (Alanis-Lobato *et al.*, 2015), including PC3 clearly improves the detection of substructure of four strata, but the authors do not give recommendations on how to select the optimal number of non-linear PCs (Fig. S6). The iterative approach adopted in *IPCAPS* can distinguish populations for which the internal substructure becomes increasingly finer: CEU, YRI, CHB, and JPT populations are well separated by *IPCAPS,* which also separates the genetically rather similar populations CHB and JPT, with only one misclassified subject (Fig. S7).

4 CONCLUSIONS

Fine-scale resolution of population substructure can be captured using independent SNPs once all redundancies are filtered. In this work, we have introduced a flexible R package to accomplish an unsupervised clustering without prior knowledge, in the search for strata of individuals with similar genetic profiles. The tool performs well in fine-scale and broad-scale resolution settings. The *IPCAPS* allows relatively easy extension to input data derived from transcriptome or epigenome experiments.

ACKNOWLEDGEMENTS

The authors thank Pongsakorn Wangkumhang, and Alisa Wilantho for helpful discussions. We also thank Chumpol Ngamphiw, Raphaël Philippart, and Alain Empain for critical help on computing clusters.

Funding: This work was supported by the Fonds de la Recherche Scientifique (FNRS PDR T.0180.13) [KC, KVS]; the Walloon Excellence in Lifesciences and Biotechnology (WELBIO) [FA, KVS]; the French National Research Agency (ANR GWIS-AM, ANR-11-BSV1-0027) [AS]; the National Science and Technology Development Agency 2011 NSTDA Chair grant [ST], and the Thailand Research Fund (RSA5780007) [PJS].

Conflict of Interest: none declared.

References

Alanis-Lobato,G. *et al.* (2015) Highlighting nonlinear patterns in population genetics datasets. *Sci. Rep.*, **5**, 8140.

Balding,D.J. and Nichols,R.A. (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Clayton,D. (2015) snpStats: SnpMatrix and XSnpMatrix classes and methods.

Corander,J. *et al.* (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, **9**, 539.

Intarapanich,A. *et al.* (2009) Iterative pruning PCA improves resolution of highly structured populations. *BMC Bioinformatics*, **10**, 382.

Lawson,D.J. *et al.* (2012) Inference of Population Structure using Dense Haplotype Data. *PLoS Genet.*, **8**, e1002453.

Lebret,R. *et al.* (2015) Rmixmod: The R Package of the Model-Based Unsupervised, Supervised, and Semi-Supervised Classification Mixmod Library. *J. Stat. Softw.*, **67**.

Limpiti,T. *et al.* (2011) Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and structure. *BMC Bioinformatics*, **12**, 255.

Liu,L. *et al.* (2013) Robust methods for population stratification in genome wide association studies. *BMC Bioinformatics*, **14**, 132.

Neuditschko,M. *et al.* (2012) NetView: A High-Definition Network-Visualization Approach to Detect Fine-Scale Population Structures from Genome-Wide Patterns of Variation. *PLoS ONE*, **7**, e48375.

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

Purcell,S. and Chang,C. PLINK 1.9. *BGI Cogn. Genomics*.