Bioinformatics, 33(17), 2017, 2776–2778
doi: 10.1093/bioinformatics/btx299
Advance Access Publication Date: 5 May 2017
Applications Note

OXFORD

Genetics and population analysis

FlashPCA2: principal component analysis of Biobank-scale genotype datasets

Gad Abraham[1,2,*], Yixuan Qiu[3] and Michael Inouye[1,2]

[1]Centre for Systems Genomics, School of BioSciences, [2]Department of Pathology, University of Melbourne, Parkville, VIC 3010, Australia and [3]Department of Statistics, Purdue University, West Lafayette, IN 47907-2066, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on January 18, 2017; revised on April 19, 2017; editorial decision on May 1, 2017; accepted on May 4, 2017

Abstract

Motivation: Principal component analysis (PCA) is a crucial step in quality control of genomic data and a common approach for understanding population genetic structure. With the advent of large genotyping studies involving hundreds of thousands of individuals, standard approaches are no longer feasible. However, when the full decomposition is not required, substantial computational savings can be made.

Results: We present FlashPCA2, a tool that can perform partial PCA on 1 million individuals faster than competing approaches, while requiring substantially less memory.

Availability and implementation: https://github.com/gabraham/flashpca.

Contact: gad.abraham@unimelb.edu.au

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

1 Introduction

Principal component analysis (PCA) of genotypes is an established approach for detecting and adjusting for population stratification and technical artefact in genome-wide association studies and similar genomic analyses (Galinsky *et al.*, 2016; Novembre and Stephens, 2008; Patterson *et al.*, 2006; Price *et al.*, 2010). The widely-used smartpca (EIGENSOFT) implementation has proven useful but it relies on two computationally expensive steps: (i) computing the genetic relatedness matrix (GRM) $\frac{1}{m}\mathbf{X}\mathbf{X}^{\mathrm{T}}$ ($\mathbf{X}$ is the $n \times m$ matrix of standardised genotypes for n individuals and m single nucleotide polymorphisms, SNPs), and (ii) eigen-decomposition of the GRM. Although this approach is effective for relatively small datasets (up to several thousand individuals), it becomes infeasible both in terms of memory requirements and computation time for larger datasets ($\mathcal{O}(mn^2)$ and $\mathcal{O}(n^3)$, respectively).

Since most genomic analyses involving PCA only make use of the top 10–20 or so principal components, alternative approaches that perform a partial decomposition have been proposed, including FlashPCA1 (Abraham and Inouye, 2014) and FastPCA (Galinsky *et al.*, 2016). These tools have enabled analyses of far-larger datasets than would be practical otherwise. However, as shown below, these algorithms may not always converge rapidly to the solution or have substantial memory requirements. These shortcomings will be particularly challenging when analysing large datasets that are now becoming available, such as the UK Biobank (Sudlow *et al.*, 2015) ($n = 500\,000$ individuals) or the Precision Medicine Initiative (Collins and Varmus, 2015), which intends to genotype 1 million individuals in the coming years.

Here we present FlashPCA2, which outperforms existing tools in terms of computation time on large datasets ($n = 1\,000\,000$ individuals and 100 000 SNPs or more), while utilising bounded memory and maintaining high accuracy for the top eigenvalues/eigenvectors. FlashPCA2 is implemented in C++ (based on the Eigen numerical library, http://eigen.tuxfamily.org), and relies on the Implicitly Restarted Arnoldi Method as implemented in the C++ library Spectra (https://spectralib.org).

2 Materials and Methods

Key to performance is the fact that the Arnoldi iterations only rely on matrix-vector multiplication with the genotype matrix $\mathbf{X}$. FlashPCA2 employs a blockwise approach whereby a suitably sized subset of the matrix is loaded into memory at one time. Other

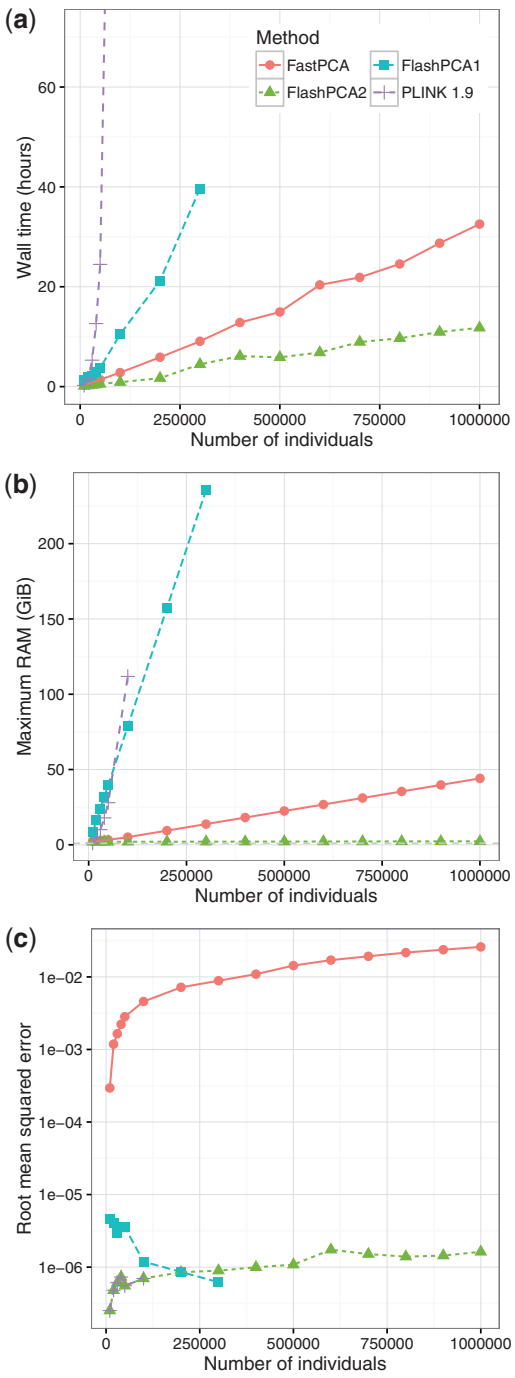

Fig. 1. Performance on the HAPGEN2 chromosome 1 simulated datasets (104 531 SNPs), as a function of sample size. **(a)** Wall time (hours, average over three runs), truncated at 72 h; **(b)** memory usage (GiB); **(c)** accuracy of the rank $K = 20$ decomposition (Equation 1). The wall time excluded LD-thinning in PLINK.

computational gains stem from precomputing the SNP-wise mean 2 p_j and standard deviation $\sqrt{2p_j(1-p_j)}$, where p_j is the minor allele frequency for the jth SNP, only once and performing a lookup of these values in subsequent passes. In addition, in our experiments (below) the Arnoldi method exhibited better convergence than the algorithm of FlashPCA1, which can get stuck in local minima before converging to the final estimates.

We used HAPGEN2 (Su *et al.*, 2011) with the 1000 Genomes 2009 CEU haplotypes (1000 Genomes Project Consortium, 2015) to simulate genotypes on chr1 (600 k SNPs) for up to 1 000 000 individuals. We used PLINK 1.9 (Chang *et al.*, 2015) to thin the SNPs by linkage-disequilibrium (LD) down to 104 531 SNPs, making the SNPs approximately independent (Patterson *et al.*, 2006). We first characterised the peformance of FlashPCA2 as a function of memory allocated to it (Supplementary Fig. S1). Best performance was obtained when the data could be loaded into RAM fully, however, this strategy is not practical for large datasets, and we chose to allow a total of 2GiB RAM as a compromise.

Next, we compared FlashPCA2 with FastPCA, FlashPCA1 and PLINK 1.9, examining wall run time, memory usage, and the decomposition error as a function of the n individuals and m SNPs using $K = 20$ dimensions for FlashPCA2 and FastPCA. The error was defined as

$$\mathrm{RMSE}_K = \left[\frac{1}{nK}\sum_{k=1}^{K}\left\|\frac{1}{m}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{u}_k - \mathbf{u}_k d_k^2\right\|_2^2\right]^{1/2}, \qquad (1)$$

where $||\cdot||_2$ is the Euclidean norm, $\mathbf{u}_k$ is the kth left singular vector and d_k is the kth singular value of $\mathbf{X}$.

As Figure 1 shows, FlashPCA2 was the fastest, followed by FastPCA (2.6× slower), FlashPCA1 (8.9× slower) and finally PLINK (316× slower) (see Supplementary Fig. S2 for run time as a function of the number of SNPs). Note that some FlashPCA1 and PLINK analyses could not be run due to the large memory requirements. FastPCA used up to 44GiB RAM for the largest analysis ($n =$ 1 000 000), whereas FlashPCA2 used only 2GiB RAM for all analyses. FlashPCA2 matched the accuracy of PLINK (full eigendecomposition), followed closely by FlashPCA1, whereas FastPCA had an RMSE three to four orders of magnitude higher (see Supplementary Fig. S3 for results of up to 100 000 individuals).

3 Conclusion

FlashPCA2 enables scalable and accurate PCA of large genotype datasets, using small amounts of memory (2GiB for 1 000 000 individuals and 100 000 SNPs in <12 h, single core), making it feasible to run such analyses on a standard personal computer, all within the R environment.

Funding

This work was supported by National Health and Medical Research Council (NHMRC) Early Career Fellowship (No. 1090462 to G.A.); NHMRC and Australian Heart Foundation Career Development Fellowship (No. 1061435 to M.I.).

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Abraham,G. and Inouye,M. (2014) Fast principal component analysis of large-scale genome-wide data. *PLoS One*, **9**, e93766.

Chang,C.C. *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.

Collins,F.S. and Varmus,H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.*, **372**, 793–795.

Galinsky,K.J. *et al*. (2016) Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet*., **98**, 456–472.

Novembre,J. and Stephens,M. (2008) Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet*., **40**, 646–649.

Patterson,N. *et al*. (2006) Population structure and eigenanalysis. *PLoS Genet*., **2**, e190.

Price,A.L. *et al*. (2010) New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet*., **11**, 459–463.

Su,Z. *et al*. (2011) HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27**, 2304–2305.

Sudlow,C. *et al*. (2015) UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*., **12**, e1001779.