bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fast principal components analysis reveals independent evolution of ADH1B gene in Europe and East Asia

Kevin J. Galinsky[1,2], Gaurav Bhatia[2,3], Po-Ru Loh[2,3], Stoyan Georgiev[4], Sayan Mukherjee[5], Nick J. Patterson[3,*], Alkes L. Price[1,2,3,*]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, USA. [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, USA. [3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, USA. [4]Stanford University School of Medicine, Palo Alto, CA. [5]Departments of Statistical Science, Computer Science, and Mathematics, Duke University, Durham, NC.

Correspondence should be addressed to K.J.G. (galinsky@fas.harvard.edu), N.J.P. (nickp@broadinstitute.org) or A.L.P. (aprice@hsph.harvard.edu). *These authors contributed equally to this work.

Principal components analysis (PCA) is a widely used tool for inferring population structure and correcting confounding in genetic data[1–8]. We introduce a new algorithm, FastPCA, that leverages recent advances in random matrix theory[9–11] to accurately approximate top PCs while reducing time and memory cost from quadratic to linear in the number of individuals, a computational improvement of many orders of magnitude. We apply FastPCA to a cohort of 54,734 European Americans, identifying 5 distinct subpopulations spanning the top 4 PCs. Using a new test for natural selection based on population differentiation along these PCs, we replicate previously known selected loci and identify three new signals of selection, including selection in Europeans at the ADH1B gene. The coding variant rs1229984 has previously been associated to alcoholism[12–14] and shown to be under selection in East Asians[13,15–17]; we show that it is a rare example of independent evolution on two continents[18,19].

The FastPCA method generalizes the method of power iteration[20], a technique to estimate the largest eigenvalue and corresponding eigenvector of a matrix. A random vector is repeatedly multiplied by a target matrix and normalized. Thus, it is projected onto all the eigenvectors of the matrix and then scaled by their corresponding eigenvalues. The projection along the eigenvector with the largest eigenvalue grows fasters than the rest, and the product converges to this eigenvector. The method of power iteration can be combined with the Gram-Schmidt orthogonalization process to produce an orthonormal basis of the top eigenvectors, by repeating this process and orthogonalizing subsequent vectors against previous estimated eigenvectors[20]. In genetic data sets, it is of interest to estimate the top eigenvectors of a genetic relationship matrix (GRM) between individuals[1,2]. However, this matrix requires time $O(MN^2)$ to compute (where M is the #SNPs and N is the #individuals) and time $O(N^3)$ to decompose, a time cost that may be prohibitive in large data sets. Instead, FastPCA uses a block-Lanczos process to construct an accurate estimate for the top PCs; accuracy is improved by estimating additional PCs and using them to create a low-rank approximation of the genotype matrix[9–11]. Singular value decomposition is then applied to the low-rank genotype matrix approximation to approximate the top eigenvectors of the GRM (see Online Methods), reducing time cost and memory usage to $O(MN)$ – much more tractable than other methods (see below). In addition, we generalize a previous selection statistic developed for discrete subpopulations[21] to detect unusual allele frequency differences along inferred PCs. This is based on the fact that the squared correlation of each SNP to a PC, rescaled to account for genetic drift, follows a chi-square (1 d.o.f.) distribution under the null hypothesis of no selection. We have released open-source software implementing the methods (see Web Resources).

We used simulated data to compare the running time and memory usage of FastPCA to three previous methods: smartpca[1,2], PLINK2-pca[22], and flashpca[23] (see Web Resources). We simulated genotype data

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

from six populations with a star-shaped phylogeny using 100k SNPs (typical for real data after LD-pruning) and up to 100k individuals (see Online Methods). For each run, running time was capped at 100 hours and memory usage was capped at 40GB. The running time and memory usage of FastPCA scaled linearly with simulated dataset size (Figure 1), compared with quadratically or cubically for other methods. The computation became intractable at 50k-70k individuals for smartpca, PLINK2-pca and flashpca. The largest dataset, with 100k SNPs and 100k individuals, required only 56 minutes and 3.2GB of memory with FastPCA (Supplementary Table 1). Thus, FastPCA enables rapid principal components analysis without specialized computing facilities.

We next assessed the accuracy of FastPCA, using PLINK2-pca[22] as a benchmark. We used the same simulation framework as before, with 10k individuals (1,667k individuals per population) and 50k SNPs. We varied the divergence between populations, as quantified by F_{ST}[24]. We assessed accuracy using the Mean of Explained Variances (MEV) of the 5 population structure PCs (see Online Methods). We determined that the results of FastPCA and PLINK-pca were virtually identical (Figure 2). This indicates that FastPCA performs comparably to standard PCA algorithms while running much faster.

We ran FastPCA on the GERA cohort (see Web Resources), a large European American dataset containing 54,734 individuals and 162,335 SNPs after QC filtering and LD-pruning (see Online Methods). This computation took 57 minutes and 2.6GB of RAM. PC1 and PC2 separated individuals along the canonical Northwest European (NW), Southeast European (SE) and Ashkenazi Jewish (AJ) axes[25], as indicated by labeling the individuals by predicted fractional ancestry from SNPweights[26] (Figure 3). PC3 and PC4 detected additional population structure within the NW population.

To further investigate this subtle structure, we projected POPRES individuals from throughout Europe[27] onto these PCs[2] (see Online Methods). This analysis recapitulated the position of SE populations via the placement of the Italian individuals, and determined that PC3 and PC4 separate the NW individuals into Irish (IR), Eastern European (EE) and Northern European (NE) populations (Figure 4). This visual subpopulation clustering was confirmed via k-means clustering on the top 4 PCs, which consistently grouped the AJ, SE, NE, IR and EE populations separately (Supplementary Figure 1).

Population differentiation between closely related populations can be valuable in detecting signals of natural selection[21,25,28,29]. We generalized a previous method for detecting selection across discrete subpopulations[21] to detect unusual allele frequency differences along inferred PCs by analyzing the squared correlations of the genotypes at each SNP to a PC. These squared correlations, rescaled to account for population differences due to genetic drift, follow a chi-square (1 d.o.f.) distribution under the null hypothesis of no selection (see Online Methods), as confirmed by simulations (Supplementary Figure 2, Supplementary Table 2). Using the PCs computed on the 162,335 LD-pruned SNPs, we calculated these selection statistics for 608,981 non-LD-pruned SNPs (see Online Methods). The resulting Manhattan plots for PCs 1-4 are displayed in Figure 5 (QQ plots are displayed in Supplementary Table 3). Analyses of PCs 5-10 indicated that these PCs do not represent true population structure (Supplementary Figure 4), but are either dominated by a small number of long-range LD loci[30–32] or are correlated with the missing genotyping rate in individuals.

Genome-wide significant signals (listed in Table 1) included several known selection regions[33–37] and novel signals at ADH1B, IGFBP3 and IGH (see below). Suggestive signals were observed at additional known selection regions[36,38] (Supplementary Table 3). After removing the regions in Table 1, rerunning FastPCA and recalculating selection statistics, all of these regions remained significant except for a chromosomal inversion on chromosome 8[30,31] (Supplementary Figure 5, Supplementary Table 4). Thus, the remaining regions are not due to PC artifacts caused by SNPs inside these regions. Detecting subtle signals of selection benefitted from the large sample size, as subsampling the GERA data set at smaller

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sample sizes and recomputing PCs and selection statistics generally led to less significant signals (Supplementary Table 5).

We identified a genome-wide significant signal of selection at rs1229984, a coding SNP (Arg47His) in the ADH1B alcohol dehydrogenase gene (Table 1). The derived allele has been shown to have a protective effect on alcoholism[39] and to produce an REHH signal[40] in East Asians[16], but was not previously known to be under selection in Europeans. (Previous studies noted the higher frequency of the derived T allele in western Asia compared to Europe, but indicated that selection or random drift were both plausible explanations[41,42].) We examined the allele frequency of the derived T allele in the five subpopulations: AJ, SE, NE, IR and EE (Supplementary Table 6). We observed derived allele frequencies (DAF) of 0.21 in AJ, 0.10 in SE, and 0.05 or lower in other subpopulations, consistent with the higher frequency of the derived allele in western Asia. A comparison of NE to the remaining subpopulations using the discrete subpopulation selection statistic[21] also produced a genome-wide significant signal after correcting for all hypotheses tested (Supplementary Table 7); this is not an independent experiment, but indicates that this finding is not due to assay artifacts affecting PCs.

To further understand the selection at this locus, we examined the allele frequency of rs1229984 in 1000 Genomes project[43] populations (see Web Resources), along with the allele frequency of the regulatory SNP rs3811801 that may also have been a target of selection in Asian populations[13]. The haplotype carrying the derived allele at rs3811801 (and corresponding haplotype H7) was absent in populations outside of East Asia (Supplementary Table 8). This indicates that if natural selection acted on this SNP in Asian populations, selection acted independently at this locus in Europeans. One possible explanation for these findings is that rs1229984 is an older SNP under selection in Europeans, while rs3811801 is a newer SNP under strong selection in Asian populations leading to the common haplotype found in those populations.

The IGFBP3 insulin-like growth factor-binding protein gene had two SNPs reaching genome-wide significance. Genetic variation in IGFBP3 is associated with increased risk of breast cancer[44] and is also associated with pulse pressure[45], blood pressure and hypertension[46]. The IGH immunoglobulin heavy locus had one genome-wide-significant SNP and two suggestive SNPs with p-value $< 10^{-6}$. Genetic variation in IGH is associated with multiple sclerosis[47] The IGFBP3 and IGH SNPs each had substantially higher minor allele frequencies in Eastern Europeans (Supplementary Table 6), but were not genome-wide significant under the discrete subpopulation selection statistic[21] (Supplementary Tables 9-10), but the existence of multiple SNPs at each of these loci with $p < 10^{-6}$ for the PC-based selection statistic suggests that these findings are not the result of assay artifacts.

We have presented FastPCA, a computationally efficient (linear-time and linear-memory) algorithm for accurately estimating top PCs. Although mixed model association methods are increasingly appealing for conducting genetic association studies[48,49], we anticipate that PCA will continue to prove useful in population genetic studies, in characterizing population stratification when present in association studies, in supplementing mixed model association methods by including PCs as fixed effects in studies with extreme stratification, and in correcting for stratification in analyses of components of heritability[50,51]. We have also presented a new method to detect selection along top PCs in datasets with subtle population structure. This method can detect selection at genome-wide significance, an important consideration in genome-wide selection scans. In particular, we detected genome-wide significant evidence of selection in Europeans at the ADH1B locus, which was previously reported to be under selection in east Asian populations[13,15–17] using REHH[40] (which can only detect relatively recent signals and does not work on standing variation[52]) – and at the disease-associated IGFBP3 and IGH loci.

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We note that our work has several limitations. First, top PCs do not always reflect population structure, but may instead reflect assay artifacts[53] or regions of long-range LD[31]; however, PCs 1-4 in GERA data reflect true population structure and not assay artifacts. Second, common variation may not provide a complete description of population structure, which may be different for rare variants[54]; we note that based on analysis of real sequencing data with known structure, we recommend that LD-pruning and removal of singletons (but not all rare variants) be applied in data sets with pervasive LD and large numbers of rare variants (see Supplementary Note). Third, our selection statistic is only capable of detecting that selection occurred, but not when or where it; indeed, top PCs may not perfectly represent the geographic regions in which selection occurred. Despite these limitations, we anticipate that the methods introduced here will prove valuable in analyzing the very large data sets of the future.

Web Resources

EIGENSOFT version 6.0.1, including open-source implementation of FastPCA and smartpca: http://www.hsph.harvard.edu/alkes-price/software/

PLINK2: https://www.cog-genomics.org/plink2/

flashpca: https://github.com/gabraham/flashpca

GERA cohort: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000674.v1.p1

1000 Genomes: http://www.1000genomes.org/

Acknowledgements

We are grateful to D. Reich for helpful discussions and S. Pollack for assistance with FastPCA software. This research was funded by NIH grant R01 HG006399. SM is funded by NSF grants DMS-1209155 and DMS-1418261.

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
2. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genet* **2,** e190 (2006).
3. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40,** 646–649 (2008).
4. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).
5. McVean, G. A Genealogical Interpretation of Principal Components Analysis. *PLoS Genet.* **5,** e1000686 (2009).
6. Schlebusch, C. M. *et al.* Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science* **338,** 374–379 (2012).
7. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
8. Moreno-Estrada, A. *et al.* The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* **344,** 1280–1285 (2014).
9. Rokhlin, V., Szlam, A. & Tygert, M. A Randomized Algorithm for Principal Component Analysis. *SIAM J. Matrix Anal. Appl.* **31,** 1100–1124 (2009).
10. Halko, N., Martinsson, P., Shkolnisky, Y. & Tygert, M. An Algorithm for the Principal Component Analysis of Large Data Sets. *SIAM J. Sci. Comput.* **33,** 2580–2594 (2011).
11. Halko, N., Martinsson, P. & Tropp, J. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Rev.* **53,** 217–288 (2011).
12. Whitfield, J. B. Alcohol Dehydrogenase and Alcohol Dependence: Variation in Genotype-Associated Risk between Populations. *Am. J. Hum. Genet.* **71,** 1247–1250 (2002).
13. Li, H. *et al.* Diversification of the ADH1B Gene during Expansion of Modern Humans. *Ann. Hum. Genet.* **75,** 497–507 (2011).
14. Edenberg, H. J. & Foroud, T. Genetics and alcoholism. *Nat. Rev. Gastroenterol. Hepatol.* **10,** 487–494 (2013).
15. Osier, M. V. *et al.* A Global Perspective on Genetic Variation at the ADH Genes Reveals Unusual Patterns of Linkage Disequilibrium and Diversity. *Am. J. Hum. Genet.* **71,** 84–99 (2002).
16. Han, Y. *et al.* Evidence of positive selection on a class I ADH locus. *Am. J. Hum. Genet.* **80,** 441–456 (2007).
17. Peter, B. M., Huerta-Sanchez, E. & Nielsen, R. Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLoS Genet* **8,** e1003011 (2012).
18. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39,** 31–40 (2007).
19. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39,** 1256–1260 (2007).
20. Golub, G. H. & Van Loan, C. F. *Matrix Computations*. (Johns Hopkins University Press, 1996).
21. Bhatia, G. *et al.* Genome-wide Comparison of African-Ancestry Populations from CARe and Other Cohorts Reveals Signals of Natural Selection. *Am. J. Hum. Genet.* **89,** 368–381 (2011).
22. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
23. Abraham, G. & Inouye, M. Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLoS ONE* **9,** e93766 (2014).
24. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23,** 1514–1521 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

25. Price, A. L. *et al.* The Impact of Divergence Time on the Nature of Population Structure: An Example from Iceland. *PLoS Genet* **5,** e1000505 (2009).
26. Chen, C.-Y. *et al.* Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29,** 1399–1406 (2013).
27. Nelson, M. R. *et al.* The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. *Am. J. Hum. Genet.* **83,** 347–358 (2008).
28. Xu, S. *et al.* Genomic Dissection of Population Substructure of Han Chinese and Its Implication in Association Studies. *Am. J. Hum. Genet.* **85,** 762–774 (2009).
29. Yi, X. *et al.* Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* **329,** 75–78 (2010).
30. Fellay, J. *et al.* A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. *Science* **317,** 944–947 (2007).
31. Tian, C. *et al.* Analysis and Application of European Genetic Substructure Using 300 K SNP Information. *PLoS Genet* **4,** e4 (2008).
32. Zou, F., Lee, S., Knowles, M. R. & Wright, F. A. Quantification of Population Structure Using Correlated SNPs by Shrinkage Principal Components. *Hum. Hered.* **70,** 9–22 (2010).
33. Bersaglieri, T. *et al.* Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *Am. J. Hum. Genet.* **74,** 1111–1120 (2004).
34. De Bakker, P. I. W. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38,** 1166–1172 (2006).
35. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol* **4,** e72 (2006).
36. Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
37. Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19,** 826–837 (2009).
38. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449,** 913–918 (2007).
39. Dick, D. M. & Foroud, T. Candidate Genes for Alcohol Dependence: A Review of Genetic Evidence From Human Studies. *Alcohol. Clin. Exp. Res.* **27,** 868–879 (2003).
40. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
41. Li, H. *et al.* Geographically Separate Increases in the Frequency of the Derived ADH1B*47His Allele in Eastern and Western Asia. *Am. J. Hum. Genet.* **81,** 842–846 (2007).
42. Treutlein, J., Frank, J., Kiefer, F. & Rietschel, M. ADH1B Arg48His Allele Frequency Map: Filling in the Gap for Central Europe. *Biol. Psychiatry* **75,** e15 (2014).
43. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
44. Al-Zahrani, A. et al. IGF1 and IGFBP3 tagging polymorphisms are associated with circulating levels of IGF1, IGFBP3 and risk of breast cancer. *Hum. Mol. Genet.* **15**, 1–10 (2006).
45. Ganesh, S. K. *et al.* Effects of Long-Term Averaging of Quantitative Blood Pressure Traits on the Detection of Genetic Associations. *Am. J. Hum. Genet.* **95,** 49–65 (2014).
46. Zhu, X. *et al.* Meta-analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension. *Am. J. Hum. Genet.* **96,** 21–36 (2015).
47. Buck, D. *et al.* Genetic variants in the immunoglobulin heavy chain locus are associated with the IgG index in multiple sclerosis. *Ann. Neurol.* **73,** 86–94 (2013).
48. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46,** 100–106 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

49. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47,** 284–290 (2015).
50. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–569 (2010).
51. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43,** 519–525 (2011).
52. Novembre, J. & Di Rienzo, A. Spatial patterns of variation due to natural selection in humans. *Nat. Rev. Genet.* **10,** 745–755 (2009).
53. Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37,** 1243–1246 (2005).
54. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44,** 243–246 (2012).
55. Balding, D. J. & Nichols, R. A. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96,** 3–12 (1995).
56. Galassi, M. *et al. GNU Scientific Library Reference Manual*. (Network Theory Limited, 2009).
57. Wang, Q., Zhang, X., Zhang, Y. & Yi, Q. AUGEM: Automatically Generate High Performance Dense Linear Algebra Kernels on x86 CPUs. in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* 25:1–25:12 (ACM, 2013). doi:10.1145/2503210.2503219
58. Hoffmann, T. J. *et al.* Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98,** 79–89 (2011).
59. Billingsley, P. *Probability and Measure*. (Wiley-Interscience, 1995).

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Online Methods

Description of FastPCA method

We are given an input $M \times N$ genotype matrix $\boldsymbol{X}$, where M is the number of SNPs and N is the number of individuals (e.g. each row is a SNP, each column is a sample). Each entry in this matrix takes its values from $\{0,1,2\}$ indicating the count of variant alleles for a sample at a SNP. From this matrix we can generate the normalized $M \times N$ genomic matrix $\boldsymbol{Y} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \boldsymbol{y}_M^T)^T$ where each row $\boldsymbol{y}_i$ has approximately mean 0 and variance 1 for SNPs in Hardy-Weinberg equilibrium.

$$\begin{aligned} \hat{p}_i &= \frac{\sum_{j=1}^{N} x_{ij}}{2N_i} = \frac{\boldsymbol{x}_i \boldsymbol{1}}{2\boldsymbol{1}^T \boldsymbol{1}} \\ y_{ij} &= \frac{x_{ij} - 2\hat{p}_i}{2\hat{p}_i(1-\hat{p}_i)} \\ \boldsymbol{y}_i &= (y_{i1}, y_{i2}, \dots, y_{iN}) = \frac{\boldsymbol{x}_i - 2\hat{p}_i \boldsymbol{1}^T}{2\hat{p}_i(1-\hat{p}_i)} \end{aligned} \tag{1}$$

Here, $\boldsymbol{x}_i$ is the row vector of genotypes for SNP i and $\boldsymbol{y}_i$ is the normalized row vector. x_{ij} and y_{ij} are the genotype/normalized genotype at SNP i for sample j. N_i is the number of valid genotypes at SNP i. All this is used to calculate $\hat{p}_i$, the sample allele frequency for SNP i, which is used to normalize the genotypes. In practice, the genotype matrix is normalized through the use of a lookup table mapping from genotypes (stored as 0, 1 or 2 copies of the alternate allele, or missing data) to normalized genotypes (using the above formula, with missing data having a normalized value of 0).

We are seeking the top K PCs for the normalized genomic matrix $\boldsymbol{Y}$. Traditional PCA algorithms compute the PCs by performing the eigendecomposition of the genetic relationship matrix ($GRM = \boldsymbol{Y}^T\boldsymbol{Y}/M$), a costly procedure which returns all the principal components. FastPCA speeds this process up by only approximating the top K PCs.

FastPCA is seeded with a random $N \times L$ matrix $\boldsymbol{G}_0$ composed of values drawn from a standard Gaussian distribution. L affects the accuracy of the result and L should be greater than K. For $K = 10$, $L = 20$ is a good choice. Then, for I iterations, $\boldsymbol{H}_i = \boldsymbol{Y} \times \boldsymbol{G}_i$ and $\boldsymbol{G}'_{i+1} = \boldsymbol{Y}^T \times \boldsymbol{H}_i$. $\boldsymbol{G}_{i+1}$ is found by taking the QR-decomposition of $\boldsymbol{G}'_{i+1}$ where $\boldsymbol{G}'_{i+1} = \boldsymbol{G}_{i+1}\boldsymbol{R}$. This step normalizes $\boldsymbol{G}_i$ to prevent rounding errors during the computation.

After the iterative step completes, the singular value decomposition of matrix $\boldsymbol{H} = (\boldsymbol{H}_0|\boldsymbol{H}_1| \dots |\boldsymbol{H}_I)$ is taken: $\boldsymbol{H} = \boldsymbol{U}_H \boldsymbol{\Sigma}_{\mathrm{H}} \boldsymbol{V}_H^T$. $\boldsymbol{U}_H$ is a low-rank approximation to the column-space of $\boldsymbol{Y}$, where $\boldsymbol{Y} \approx \boldsymbol{U}_H \boldsymbol{U}_H^T \boldsymbol{Y}$. The SVD of $\boldsymbol{T} = \boldsymbol{U}_H^T \boldsymbol{Y} = \boldsymbol{U}_T \boldsymbol{\Sigma}_T \boldsymbol{V}_T^T$ can be computed efficiently and approximates the SVD of $\boldsymbol{Y}$ since $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \approx \boldsymbol{U}_H \boldsymbol{T} = \boldsymbol{U}_H \boldsymbol{U}_T \boldsymbol{\Sigma}_T \boldsymbol{V}_T^T$. For the PCA, we are only interested in the left K columns of $\boldsymbol{V}_T$ and the first K entries along the diagonal of $\boldsymbol{\Sigma}_T$.

Simulation framework

Simulated genotypes at a particular SNP were generated for multiple populations separated by a given fixation index (F_{ST}) by first generating an ancestral population allele frequency p from a $Uniform(0.05,0.95)$ distribution, and then generating individual population frequencies from a truncated $Beta\left(p \times \frac{1-F_{ST}}{F_{ST}}, (1-p) \times \frac{1-F_{ST}}{F_{ST}}\right)$ distribution, where allele frequencies outside of $[0.01,0.99]$ are discarded[21,55]. This was to facilitate generation of more complicated population structures; a descendent population frequency could be plugged into the above equation to generate additional population frequencies separated by a different F_{ST}. When the minor allele frequency approached 0, the method to generate the beta random variate would crash. Once a population allele

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

frequency p_i was established, N_i individual genotypes would be generated from a $Binomial(2, p_i)$ distribution.

To assess running time, the simulated datasets had $F_{ST} = 0.01$, $M = 100k$ SNPs and $N \approx \{1k, 1.5k, 2k, 3k, 5k, 7k, 10k, 15k, 20k, 30k, 50k, 70k, 100k\}$ individuals (since there were 6 populations, $N_i = round\left(\frac{N}{6}\right)$). Throughout this paper we report CPU time, but due to multithreading present in the GSL[56] and OpenBLAS[57] libraries run time was about 60% of CPU time. Accuracy was assessed using $M = 50k$ SNPs and $N \approx 10k$ individuals at $F_{ST} = \{0.001, 0.002, \ldots, 0.010\}$.

Assessing accuracy

Accuracy was assessed via the Mean or Explained Variances (MEV) of eigenvectors. Two different sets of K N-dimensional principal components each produce column space. A metric for the performance of a PCA algorithm against some baseline is to see how much the column overlap. This is done by projecting the eigenvectors of one subspace onto the other and finding the mean lengths of the projected eigenvectors. If we have a reference set of PCs $(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K)$ against which we wish evaluate the performance a set of computed PCs $(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_K)$, then the performance calculation becomes:

$$MPL = K^{-1} \sum_{j=1}^{K} \sqrt{\sum_{j=1}^{K} \left(\boldsymbol{v}_k \cdot \boldsymbol{u}_j\right)^2} = K^{-1} \sum_{j=1}^{K} \|\boldsymbol{U}^T \boldsymbol{v}_k\| \qquad (2)$$

Here, $\boldsymbol{U}$ is a matrix whose column vectors are the PCs which we are testing. The test matrix can either be the result of another computation or the truth for a simulated sample. K eigenvectors can describe the population structure in a dataset with $K + 1$ populations. They can be constructed by first creating a vector $\boldsymbol{v}_k^* = (v_{k,1}^*, v_{k,2}^*, \ldots v_{k,N}^*)$ where $v_{k,j}^* = 1$ if individual j is in population k and 0 otherwise. The set of eigenvectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K\}$ are constructed by taking K of these vectors, normalizing them to have mean 0, and scaling/orthogonalizing them via the Gram-Schmidt process.

GERA data set

The GERA dataset comprises 670,176 SNPs and 62,318 individuals of European descent from Northern California[58]. Individuals were filtered to remove those with missing sex information, individuals related according to the provided pedigree data or with observed genomic relatedness greater than 0.05 in the GRM[22] and individuals with less than 90% European ancestry as predicted by SNPweights[26] using a worldwide dataset containing European, African, Asian and Native American ancestry. After filtering, 54,734 individuals remained.

SNPs were initially filtered to remove non-autosomal SNPs, SNPs with minor allele frequency less than 1%, and SNPs with >1% missing data, leaving 608,981 SNPs. The second stage of filtering removed SNPs that failed PLINK's Hardy-Weinberg Equilibrium test[22] with $p < 10^{-6}$, and performed LD-pruning using PLINK. Due to regions of long-range LD, LD persisted even after one filtering run. Multiple rounds of LD filtering were performed using an r^2 cutoff of 0.2 until additional rounds of LD filtering did not remove additional SNPs, leaving 162,335 SNPs. Selection statistics (see below) were computed on the set of 608,981 SNPs, prior to H-W filtering and LD-pruning. We note that many of the SNPs producing signals of selection generated significant H-W *p*-values (e.g. H-W $p = 1.37 \times 10^{-79}$ for LCT SNP rs6754311), which is an expected consequence of unusual population differentiation.

SNPweights[26] was used to predict fractional Northwest European, Southeast European, and Ashkenazi Jewish ancestry for each individual. In Figure 3, percentage ancestry in each of these three populations

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was mapped to an integer in $[0,255]$, which was then used for the RGB color value for that sample, so a NW sample would appear red, SE would appear green and AJ would appear blue.

PC Projection

POPRES[27] individuals were projected onto these PCs. The left singular vectors ($\boldsymbol{U}$) were generated by multiplying normalized genotypes for all SNPs in GERA ($\boldsymbol{Y}_{GERA}$) by the PCs ($\boldsymbol{V}$) and scaling by the singular values ($\boldsymbol{\Sigma}$), the number of SNPs used to calculate the PCs (M) and the number of SNPs used for projection (M_{GERA}): $\boldsymbol{U} = \boldsymbol{Y}_{GERA}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}M/M_{GERA}$. Projected PCs were then calculated by multiplying the corresponding set of SNPs in POPRES by these singular vectors and scaling again by the singular values: $\boldsymbol{V}_{POPRES} = \boldsymbol{Y}_{POPRES}^{T}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}$. The projected individuals were overlaid on the PCA plot of GERA individuals and colored according to population membership and consistently with population assignment from SNPweights[26].

Selection statistic

Previous work[21] shows that for a SNP i genotyped in two populations, the difference in allele frequency estimates approximately follows a normal distribution.

$$D_i = \hat{p}_{i1} - \hat{p}_{i2} \sim N\left[0, \hat{p}_i(1-\hat{p}_i)\left(2F_{ST} + \frac{1}{2N_1} + \frac{1}{2N_2}\right)\right] = N[0, \sigma_D^2] \tag{3}$$

Here, $\hat{p}_{iq}$ is the allele frequency estimate of SNP i for a sample of size N_q from population q and F_{ST} is the measure of differentiation between the two populations. Our goal is to extend this formula to individuals with fractional ancestries, and then to continuous-valued PCs.

First, consider the case with two discrete subpopulations. Rather than treating the subpopulations separately, we define a vector $\boldsymbol{\alpha}$ where α_j indicates the ancestry in population 1 (e.g. $\alpha_j = 1$ if sample j is in population 1 and 0 if sample j is in population 2). D_i can be rewritten as:

$$\hat{p}_1 = \frac{\boldsymbol{x}_i\boldsymbol{\alpha}}{2\boldsymbol{1}^T\boldsymbol{\alpha}}, \qquad \hat{p}_2 = \frac{\boldsymbol{x}_i(\boldsymbol{1}-\boldsymbol{\alpha})}{2\boldsymbol{1}^T(\boldsymbol{1}-\boldsymbol{\alpha})}, \qquad D_i = \frac{\boldsymbol{x}_i\boldsymbol{\alpha}}{2\boldsymbol{1}^T\boldsymbol{\alpha}} - \frac{\boldsymbol{x}_i(\boldsymbol{1}-\boldsymbol{\alpha})}{2\boldsymbol{1}^T(\boldsymbol{1}-\boldsymbol{\alpha})} \tag{4}$$

If we run PCA on this sample, we would ideally get an eigenvector $\boldsymbol{v}$ that has value v_1 for individuals in population 1 and $-v_2$ for individuals in population 2, where (since $\boldsymbol{v}^T\boldsymbol{1} = 0$, $\boldsymbol{v}^T\boldsymbol{v} = 1$)

$$v_q = \frac{1}{N_q}\sqrt{\frac{N_1N_2}{N}} \tag{5}$$

In this case, D_i can be rewritten as:

$$D_i = \frac{1}{2}\sqrt{\frac{N_1N_2}{N}}\boldsymbol{x}_i\boldsymbol{v} \tag{6}$$

In the limiting case where F_{ST} approaches 0, the statistic becomes:

$$\frac{D_i^2}{\sigma^2} = \frac{\frac{1}{4}\frac{N_1N_2}{N}\boldsymbol{x}_i\boldsymbol{v}}{\hat{p}_i(1-\hat{p}_i)\left(\frac{1}{2N_1}+\frac{1}{2N_2}\right)} = \left[\left(\frac{\boldsymbol{x}_i - 2\hat{p}_i\boldsymbol{1}^T}{2\hat{p}_i(1-\hat{p}_i)}\right)\boldsymbol{v}\right]^2 = [\boldsymbol{y}_i\boldsymbol{v}]^2 \tag{7}$$

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Thus, the square of the SNP weight follows a chi-square 1-d.o.f. distribution in the case where $F_{ST} \to 0$. In the case where $F_{ST} \neq 0$, then the scaling parameter has to be changed, but D_i still follows a normal distribution.

In the case with fractional ancestry ($\alpha_j \in [0,1]$), $\hat{p}_1$, $\hat{p}_2$ and D_i can still be estimated using equation (4). The individual $\hat{p}_q$s will still asymptotically follow a normal distribution (because of the Lyapunov central limit theorem[59]), but will be correlated due to individuals with fractional ancestry contributing to both estimates. Thus, D_i will still follow a normal distribution, but the variance of equation (3) will not hold.

Now consider the case where we do not have fractional ancestries, but rather an eigenvector that separates individuals along some axis of variation. We can treat the eigenvector as a linear transformation of the ancestry vector:

$$\boldsymbol{\alpha} = \beta_0 + \beta_1 \boldsymbol{v} \tag{8}$$

Substituting these values into (4), we find:

$$D_i = \frac{\beta_1}{2N\beta_0(1-\beta_0)} \boldsymbol{x}_i \boldsymbol{v} \propto \boldsymbol{y}_i \boldsymbol{v} \tag{9}$$

Thus, our new selection statistic D_i is based on the dot product of the normalized genotypes and the eigenvector. Since the variance of D_i is not known, it will need to be rescaled in order to follow a $N(0,1^2)$ distribution.

(1) If we are operating on the same set of SNPs that we used for PCA, then the rescaling of $\boldsymbol{y}_i \boldsymbol{v}$ is straightforward. Because PCA is the same as SVD, we see that:

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \\ \boldsymbol{U} &= \boldsymbol{Y}\boldsymbol{V}\boldsymbol{\Sigma}^{-1} \end{aligned} \tag{10}$$

Here, $\boldsymbol{V}$ contains the right singular vectors which are equivalent to the PCs, $\boldsymbol{U}$ contains the left singular vectors which are rescaled SNP weights and $\boldsymbol{\Sigma}$ contains the singular values which are the square roots of the eigenvalues of the GRM. $\boldsymbol{V}$ and $\boldsymbol{U}$ are unitary, so the columns of $\boldsymbol{U}$ are guaranteed to have a norm of 1. Multiplying $\boldsymbol{U}$ by $\sqrt{M}$ will then produce a properly normalized vector of differences $\boldsymbol{D} = (D_1, D_2, \dots, D_M)^T$. In other words:

$$\frac{\sqrt{M}}{\Sigma_k} \boldsymbol{y}_i \boldsymbol{v}_k \sim N(0,1) \tag{11}$$

In the case where we are calculating PCs on a different set of SNPs than the one for which we are calculating weights, then the above property is not guaranteed to hold. In this case, (2) a properly normalized $\boldsymbol{D}$ can be obtained by scaling $\boldsymbol{Y}\boldsymbol{V}$ so that it has norm M, i.e. scaling $\boldsymbol{y}_i \boldsymbol{v}$ so it has variance 1. This is the approach used in all of our analyses. When rescaling the weights in GERA using equation (11), the variances for PCs 1-4 were 1.03-1.07, while the variances for PCs 5-10 ranged 0.93-8.12.

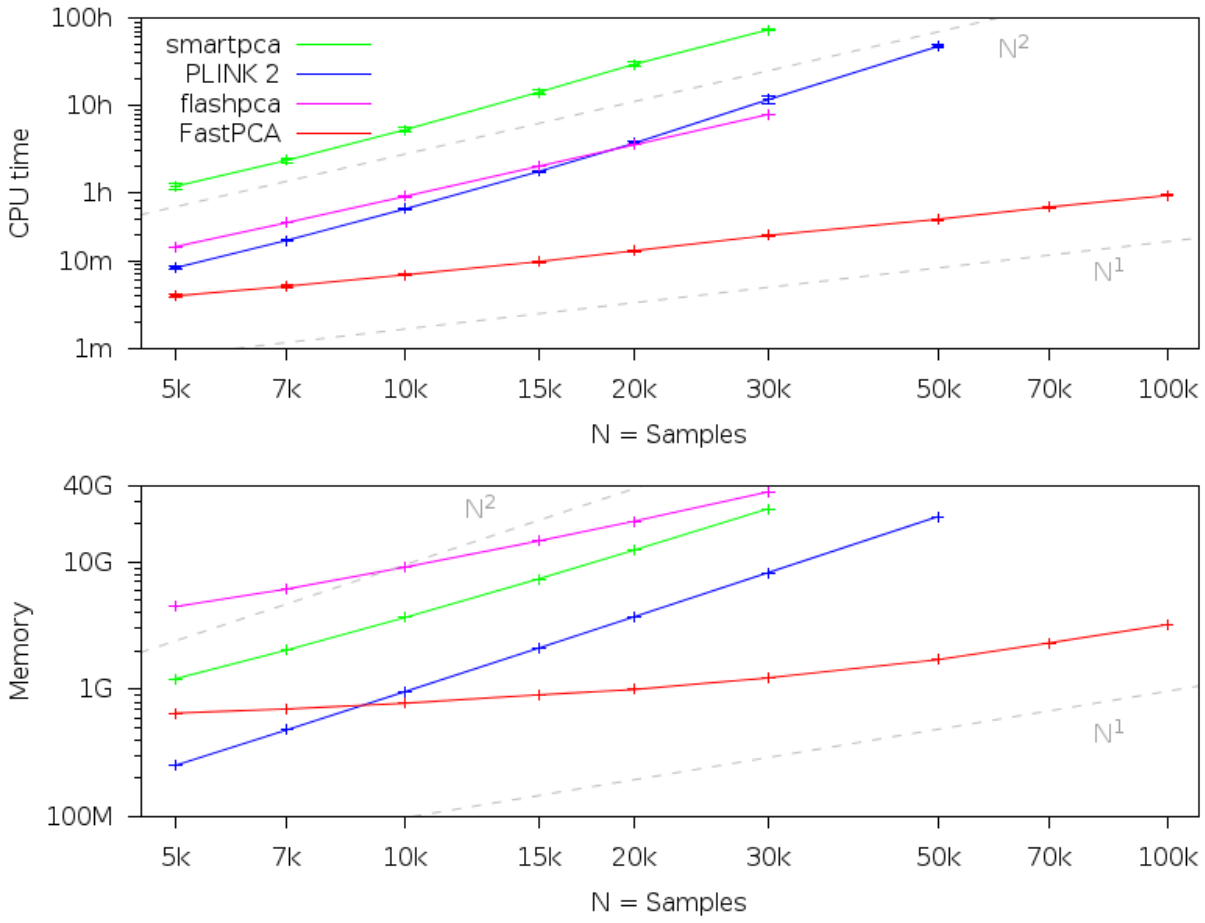
One assumption underlying the statistic is that the true minor allele frequency is not extremely small, otherwise the assumption of normality will not hold[21]. For this reason, the selection statistic was only computed for those SNPs containing minor allele frequency greater than 1%.

bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

Figure 1. Running time and memory requirements of FastPCA and other methods.

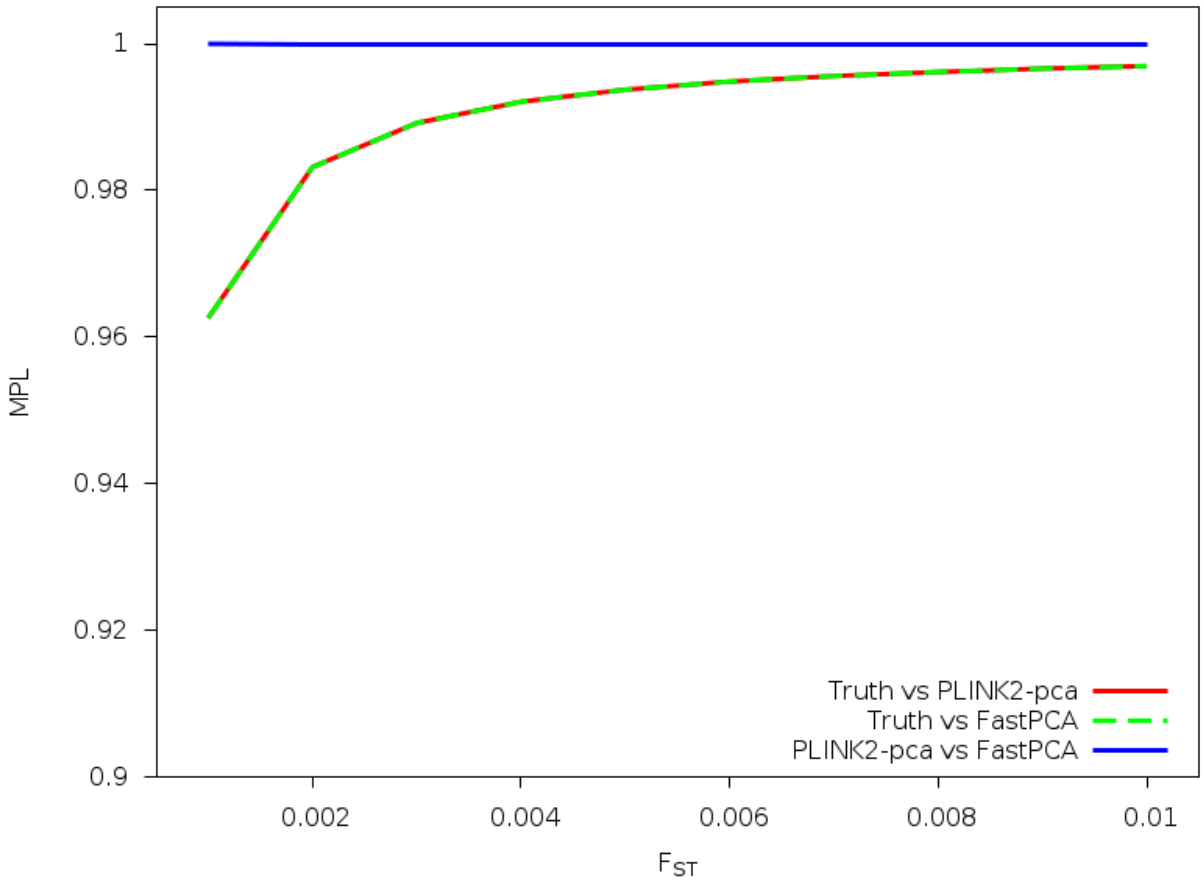
The CPU time and memory usage of FastPCA scale linearly with the number of individuals. On the other hand, smartpca and PLINK2-pca scale between quadratically and cubically, depending on whether computing the GRM (quadratic) or the eigendecomposition (cubic) is the rate-limiting step. The running time of flashpca scales quadratically (because it computes the GRM), but its memory usage scales linearly because it stores the normalized genotype matrix in memory. With 50k individuals, smartpca exceeded the time constraint (100 hours) and flashpca exceeded the memory constraint (40GB). With 70k individuals, PLINK2-pca exceeded the memory constraint (40GB). Run times are based on one core of a 2.26-GHz Intel Zeon L5640 processor; we caution that run time comparisons may vary by a small constant factor as a function of the computing environment. Numerical data are provided in Supplementary Table 1.



bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2. Accuracy of FastPCA and PLINK2-pca.

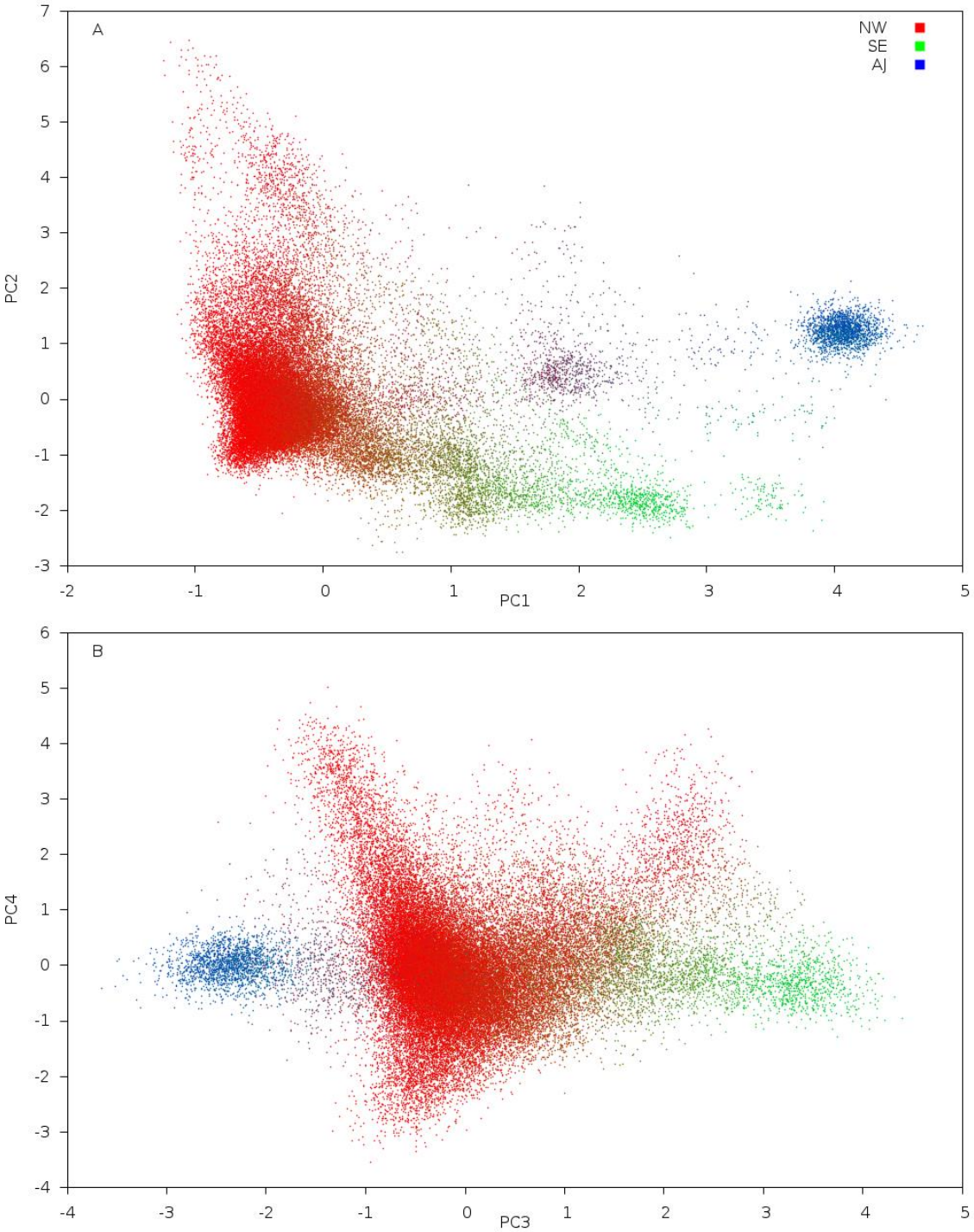
FastPCA and PLINK2-pca were run on simulated populations of varying divergence. The simulated data comprised 50k SNPs and 10k total individuals from six subpopulations derived from a single ancestral population. PCs computed by PLINK2-pca and FastPCA were compared to the true population PCs and to each other using the Mean of Explained Variances (MEV) metric (see text). FastPCA explained the same amount of true population variance as PLINK2-pca in all experiments, and the methods output nearly identical PCs (MEV>0.999).



bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

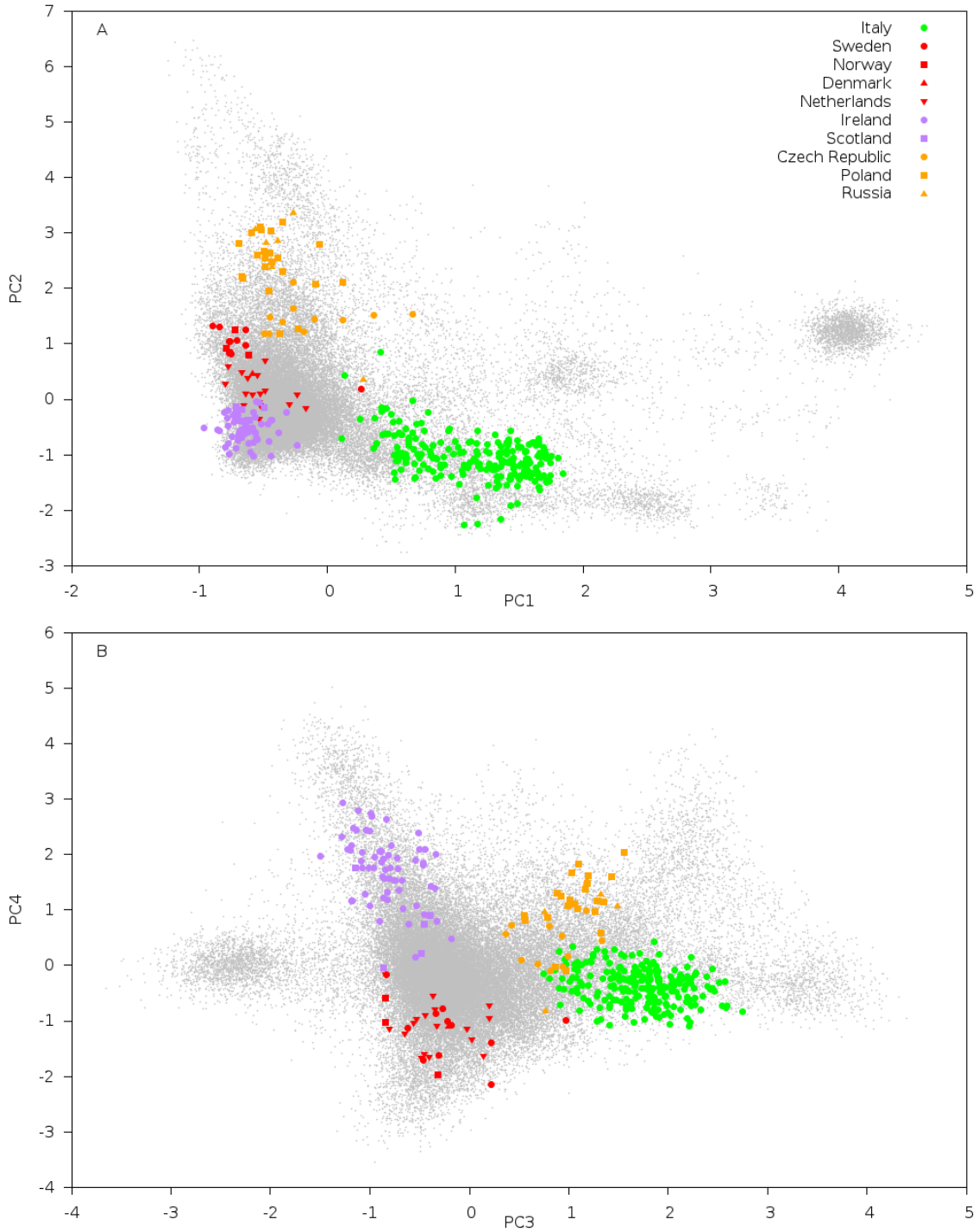
Figure 3. PCA results on GERA data set.

FastPCA and SNPweights[26] were run on the GERA cohort and the principal components from FastPCA were plotted. Individuals were colored by mapping Northwest European (NW), Southeast European (SE) and Ashkenazi Jewish (AJ) ancestry estimated by SNPweights to the red/green/blue color axes (see Online Methods). PC1 and PC2 separate the GERA cohort into northwest (NW), southeast (SE) and Ashkenazi Jewish (AJ) subpopulations. PC3 separates the AJ and SE individuals, while PC3 and PC4 further separates the NW European individuals.



bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

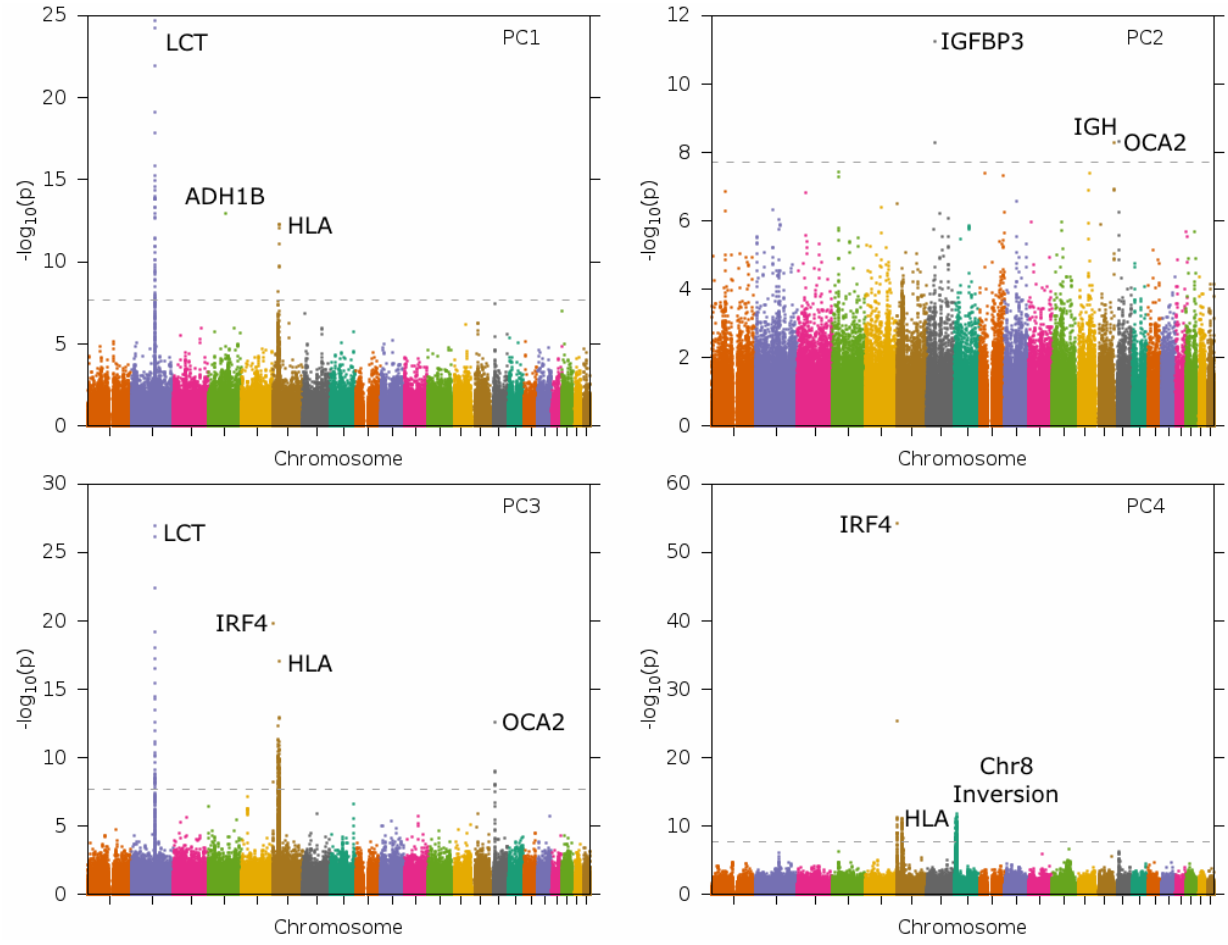
Figure 4. Separation of Irish, Eastern European and Northern European individuals in GERA data. We report results of projecting POPRES[27] individuals onto top PCs. The plot of PC3 vs PC4 shows that the Northwest European (NW) individuals are further separated into Irish and Eastern European and Northern European populations. Projected populations were colored based on correspondence to the ancestry assignment from SNPweights[26], except that Irish and Eastern European individuals were colored purple and orange, respectively, to indicate additional population structure.



bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 5. Signals of selection in the top PCs of GERA data.

We display Manhattan plots for selection statistics computed using each of the top 4 PCs. The grey line indicates the genome-wide significance threshold of 2.05 x 10^{-8} based on 2,435,924 hypotheses tested ($\alpha = 0.05$, 608,981 SNPs x 4 PCs).



bioRxiv preprint doi: https://doi.org/10.1101/018143; this version posted April 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tables

Table 1. Genome-wide significant signals of selection in GERA data.

We list regions with genome-wide significant ($\alpha = 0.05$, Bonferroni correction with 608,981 SNPs x 4 PCs = 2,435,924 hypotheses tested, $p < 2.05 \text{ x } 10^{-8}$) evidence of selection in the top 4 PCs. Loci that were not previously known to be under selection in Europeans are indicated in bold font. The chromosome 8 inversion signal is due to a PC artifact (see main text). Regions with suggestive evidence of selection ($10^{-6} < p < 2.05 \text{ x } 10^{-8}$) are listed in Supplementary Table 3.

Locus	Chromosome	Region (Mb)	PC	Best Hit	p-value
LCT[33]	2	134.8 – 137.6	1	rs6754311	2.15×10^{-25}
			3	rs4988235	1.15×10^{-27}
ADH1B	**4**	**100.5**	**1**	**rs1229984**	$\mathbf{1.26 \times 10^{-13}}$
IRF4[36,37]	6	0.3 – 0.5	3	rs12203592	1.76×10^{-20}
			4	rs12203592	5.52×10^{-55}
HLA[34]	6	30.8 – 32.9	1	rs382259	5.38×10^{-13}
			3	rs9268628	8.66×10^{-18}
			4	rs4394275	9.36×10^{-12}
IGFBP3	**7**	**45.3-45.9**	**2**	**rs150353309**	$\mathbf{5.82 \times 10^{-12}}$
Chr8 Inversion[31]	8	8.2 – 11.9	4	rs6984496	1.86×10^{-12}
IGH	**14**	**106.0-106.1**	**2**	**rs34614900**	$\mathbf{5.23 \times 10^{-9}}$
OCA2[35,37]	15	25.9 – 26.2	2	rs12916300	4.82×10^{-9}
			3	rs12916300	2.80×10^{-13}