Scalable probabilistic PCA for large-scale genetic variation data

Aman Agrawal †1, Alec M. Chiu †2, Minh Le[3], Eran Halperin[2,3,4,5,6], and Sriram Sankararaman *2,3,4,6

[1]Department of Computer Science, Indian Institute of Technology, Delhi, India
[2]Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA USA
[3]Department of Computer Science, University of California, Los Angeles, CA USA
[4]Department of Human Genetics, University of California, Los Angeles, CA USA
[5]Department of Anesthesiology and Perioperative Medicine, University of California, Los Angeles, CA USA
[6]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA USA

Abstract

Principal component analysis (PCA) is a key tool for understanding population structure and controlling for population stratification in genome-wide association studies (GWAS). With the advent of large-scale datasets of genetic variation, there is a need for methods that can compute principal components (PCs) with scalable computational and memory requirements. We present ProPCA, a highly scalable method based on a probabilistic generative model, which computes the top PCs on genetic variation data efficiently. We applied ProPCA to compute the top five PCs on genotype data from the UK Biobank, consisting of 488,363 individuals and 146,671 SNPs, in less than thirty minutes. Leveraging the population structure inferred by ProPCA within the White British individuals in the UK Biobank, we scanned for SNPs that are not well-explained by the PCs to identify several novel genome-wide signals of recent putative selection including missense mutations in RPGRIP1L and TLR4.

†Equal contribution

*email: sriram@cs.ucla.edu

Author Summary

Principal component analysis is a commonly used technique for understanding population structure and genetic variation. With the advent of large-scale datasets that contain the genetic information of hundreds of thousands of individuals, there is a need for methods that can compute principal components (PCs) with scalable computational and memory requirements. In this study, we present ProPCA, a highly scalable statistical method to compute genetic PCs efficiently. We systematically evaluate the accuracy and robustness of our method on large-scale simulated data and apply it to the UK Biobank. Leveraging the population structure inferred by ProPCA within the White British individuals in the UK Biobank, we identify several novel signals of putative recent selection.

Introduction

Inference of population structure is a key step in population genetic analyses [1] with applications that include understanding genetic ancestry [2–4] and controlling for confounding in genome-wide association studies (GWAS) [5]. While several methods have been proposed to infer population structure (e.g., [6–10]), principal component analysis (PCA) is one of the most widely used [11, 6]. Unfortunately, the naive approach for estimating principal components (PCs) by computing a full singular value decomposition (SVD) scales quadratically with sample size (for datasets where the number of SNPs is larger than sample size), resulting in runtimes unsuitable for large data sets.

In light of these challenges, several solutions have been proposed for the efficient computation of PCs. One approach taken by two recent scalable implementations (FastPCA [12] and FlashPCA2 [13]) takes advantage of the fact that typical applications of PCA in genetics only require computing a small number of top PCs; *e.g.* GWAS typically use 5-20 PCs to correct for stratification [14]. An alternative approach for efficient computation of PCs takes advantage of the parallel computation infrastructure of the cloud [15]. However, the cost of cloud usage is roughly proportional to the number of CPU hours used by these algorithms, making them cost-prohibitive. Finally, these scalable implementations lack a full probabilistic model, making them challenging to extend to settings with missing genotypes or linkage disequilibrium (LD) between SNPs.

In this work, we describe ProPCA, a scalable method to compute the top PCs on genotype data. ProPCA is based on a previously proposed probabilistic model [16, 17], of which PCA is a special case. While PCA treats the PCs and the PC scores as fixed parameters, probabilistic PCA imposes a prior on the PC scores. This formulation leads to an iterative Expectation Maximization (EM) algorithm for computing the PCs. ProPCA leverages the structure of genotype data so that each iteration of the EM algorithm can be computed in time that scales sub-linear in the number of individuals or SNPs. The EM algorithm requires only a small number of iterations to obtain accurate estimates of the PCs resulting in a highly scalable algorithm.

In both simulated and real data, ProPCA is able to accurately infer the top PCs while scaling favorably with increasing sample size. We applied ProPCA to compute the top five PCs on genotype data from the UK Biobank, consisting of 488,363 individuals and 146,671 SNPs, in less than thirty minutes. Leveraging the population structure inferred by ProPCA within the White British individuals in the UK Biobank [18], we scanned for SNPs that are not well-modeled by the top PCs to identify several novel genome-wide signals of recent positive selection. Our scan recovers sixteen loci that are highly differentiated across the top five PCs that are likely signals of recent selection. While these loci include previously reported targets of selection [12], the larger sample size that we analyze here allows us to identify eleven novel signals including a missense mutation in RPGRIP1L ($p = 2.09 \times 10^{-9}$) and another in TLR4 ($p = 7.60 \times 10^{-12}$).

A number of algorithms that analyze genotype data, including methods for heritability estimation and association testing, can be modeled as iterative procedures where the core computational operation is similar to that solved by ProPCA. Thus, the algorithm that we employ in this work can potentially lead to highly scalable algorithms for a broad set of population genetic analyses.

Results

Accuracy We first assessed the accuracy of ProPCA using the simulation framework described in the Methods. We generated datasets containing $50,000$ SNPs and $10,000$ individuals across q populations, where q was chosen to be 5 and 10. The populations were simulated with varying levels of population differentiation that are typical of present-day human populations (values of F_{st} ranging from 0.001 to 0.01) and were small enough so that we could compute the full SVD thereby allowing us to estimate the accuracy of the PCs computed by ProPCA. To measure accuracy, we computed the mean of explained variances (MEV), a measure of the overlap between the subspaces spanned by the PCs estimated by ProPCA compared to the PCs computed using a full SVD (Methods). ProPCA estimates highly accurate PCs (values of MEV close to 1) across the range of parameters (Table 1).

Runtime We assessed the scalability of ProPCA with increasing sample size (Methods). We simulated genotypes from five populations containing $100,000$ SNPs and sample sizes varying from $10,000$ to $1,000,000$ with $F_{st} = 0.10$.

We compared the wall-clock time for running ProPCA, the SVD implementation in PLINK (PLINK_SVD [19]), FastPCA [12] and FlashPCA2 [13]. The SVD implementation in PLINK could not run in reasonable time on datasets exceeding $70,000$ individuals (Figure 1a). While FastPCA, FlashPCA2 and ProPCA all scale with sample size, ProPCA is about four times faster than FastPCA and twice as fast as FlashPCA2 (Figure 1b) . ProPCA computes PCs in about 30 minutes even on the largest data containing a million individuals and $100,000$ SNPs.

Since ProPCA, FastPCA, and FlashPCA2 are all based on iterative algorithms, their runtimes depend on details of convergence criterion. We performed an additional experiment to compare the runtime of ProPCA and FastPCA (for which we could instrument the source code) for a single iteration and found ProPCA to be three to four times faster than FastPCA across the range of sample sizes (Figure S1). Measuring the accuracy of the PCs (MEV) as a function of runtime (on datasets with a range of F_{st} containing $50,000$ SNPs and $10,000$ individuals so that we could compare the estimated PCs to exact PCs), ProPCA attains a given MEV in about half the time as FastPCA (Figure 2). We were unable to include FlashPCA2 in these comparisons as we were unable to instrument the source code to control the number of iterations.

Application to real genotype data We applied ProPCA to genotype data from Phase 1 of the 1000 Genomes project [20]. On a dataset of 1092 individuals and $442,350$ SNPs, ProPCA computes the top five PCs with an MEV of 0.968 producing PCs that are qualitatively indistinguishable from running a full SVD (Figure S4, Table S1). We also applied ProPCA to genotype data from the UK Biobank [18] consisting of $488,363$ individuals and $146,671$ SNPs after QC. ProPCA can compute the top five PCs in about 27 minutes and the resulting PCs reflect population structure within the UK Biobank, consistent with previous studies [18] (Figure 3a).

Application to scans for selection We developed a statistical test to search for SNPs that are not well-modeled by the ProPCA model as a means of discovering signals of natural selection

(Methods).This statistic relies on the observation that a SNP evolving under positive selection is expected to exhibit differentiation in the frequencies of its alleles that is extreme compared to a typical SNP that is evolving neutrally [21]. Since deviations from the ProPCA model can occur due to reasons unrelated to selection, we filtered out SNPs with high rates of missingness, low minor allele frequency (MAF), and presence in regions of long-range LD [22](Methods). We ran ProPCA to infer the top five PCs on $276,736$ unrelated White British samples and the UK Biobank SNP set consisting of $146,671$ SNPs obtained by further removing SNPs in high LD (Figure S5).

The Pearson correlation coefficient between birth location coordinates and the PC score for each individual reveals that the estimated PCs capture geographic structure within the UK (Figure 3b, Figure S6, Table S2). We used these PCs to perform a selection scan on a larger set of $516,140$ SNPs and we report SNPs that are genome-wide significant after accounting for the number of SNPs as well as PCs tested (p-value $< \frac{0.05}{6\times 516,140}$; we use 6 to account for the additional combined test statistic that we describe later). We ensured that the selection statistic for each PC was well-calibrated against a χ^2_1 distribution (Fig. S7) and genomic inflation (λ_{GC}) values for each of the PCs showed no substantial inflation (Table S3). While our statistic is closely related to a previously proposed statistic to detect selection on PCs [12], we found that our proposed statistic is better calibrated (Table S3).

Our scan revealed a total of 59 SNPs that were genome-wide significant (Table S4). Clustering these signals into 1 Mb windows centered around the most significant SNP for each PC, we obtained twelve non-overlapping loci that contain putative targets of selection (Figure 4, Table S5). These twelve loci include five that were previously reported to be signals of selection in the UK with genome-wide significance: *LCT* (rs7570971 with $p = 8.51 \times 10^{-16}$), *TLR1* (rs5743614, $p = 5.65 \times 10^{-25}$), *IRF4* (rs62389423, $p = 8.80 \times 10^{-42}$), *HLA* (rs9267817, $p = \times 6.17 \times 10^{-9}$), and *FUT2* (rs492602, $p = 7.02 \times 10^{-10}$) [12]. The larger sample size that we analyze here also reveals novel signals at additional loci. Four of the twelve signals were previously suggested to be signals of selection but were not genome-wide significant: *HERC2* (rs12913832, $p = 5.21\times 10^{-10}$), *RPGRIP1L* (rs61747071, $p = 2.09 \times 10^{-9}$), *SKI* (rs79907870, $p = 2.58 \times 10^{-9}$), *rs77635680* ($p = 2.22 \times 10^{-10}$) [12] while the remaining three loci: *HERC6* (rs112873858, $p = 2.68 \times 10^{-11}$), *rs6670894* ($p = 4.98 \times 10^{-9}$), and *rs12380860* ($p = 8.62 \times 10^{-9}$) appear to be previously unreported (Section S7).

To validate our findings, we utilized birth location coordinates for each individual and assigned them to geographical regions in the UK as defined in the Nomenclature of Territorial Units for Statistics level 3 (NUTS3) classification. We performed a test of association between the allele frequency of the top SNP in each of our novel loci with geographical regions (Table S9) and confirmed that SNPs identified in our selection scan show differences in allele frequencies across specific geographical regions (Table S9).

One of the novel genome-wide significant loci is *RPGRIP1L*. *RPGRIP1L* is a highly conserved gene that encodes a protein that localizes at primary cilia and is important in development [23]. Mutations in this gene have been implicated with neurological disorders such as Joubert syndrome and Meckel syndrome [24], conditions that sometimes also result in additional symptoms such as eye diseases and kidney disorders [25]. The SNP with the most significant p-value in our scan in *RPGRIP1L*, rs61747071, is a missense loss-of-function mutation A229T that has been shown to lead to photoreceptor loss in ciliopathies [26].

We created an additional variant of our selection statistic which tests for SNPs that are not well-modeled by a linear combination of the first five PCs by summing the per-PC χ^2_1 statistics resulting in a new chi-squared statistic with five degrees of freedom. Combining signals across PCs has been previously shown to boost power in association testing [27]. We verified that the resulting combined statistic is also calibrated (Fig S7 and Table S3). Under this combined statistic, we recover majority of the loci found on each individual PC, but we also discover four additional novel loci: *AMPH* (rs118079376, $p = 2.64 \times 10^{-10}$), *TLR4* (rs4986790, $p = 7.60 \times 10^{-12}$), *rs9856661* ($p = 6.46 \times 10^{-9}$), and *rs116352364* ($p = 5.24 \times 10^{-11}$) (Table S8).

TLR4 is a member of the toll-like receptor family. The TLR gene family is known to play a fundamental role in pathogen recognition and activation of innate immunity, but TLR4 in particular is involved with proinflammatory cytokines and has a pro-carcinogenic function [28]. The SNP with the most significant p-value at our $TLR4$ locus is rs4986790, a missense D299G mutation and D259G mutation on two different transcripts for the TLR4 gene. The D299G mutation is of particular interest as this mutation is strongly correlated with increased infection by *Plasmodium falciparum*, a parasite that causes malaria [29, 30]. Additional details on the signals of selection can be found in the Supplementary Information (SI Section S7).

Discussion

We have presented, ProPCA, a scalable method for PCA on genotype data that relies on performing inference in a probabilistic model. Inference in this model consists of an iterative procedure that uses a fast matrix-vector multiplication algorithm. We have demonstrated its accuracy and efficiency across diverse settings. Further, we have demonstrated that ProPCA can accurately estimate population structure within the UK Biobank dataset and how this structure can be leveraged to identify targets of recent putative selection.

The algorithm that we employ here to accelerate the EM updates is of independent interest. Beyond PCA, several algorithms that operate on genotype data perform repeated matrix-vector multiplication on the matrix of genotypes. For example, association tests and permutation tests, can be formulated as computing a matrix-vector product where the matrix is the genotype matrix while the vector consists of phenotype measurements. The idea that SVD computations can leverage fast matrix-vector multiplication operations to obtain computational efficiency is well known in the numerical linear algebra literature [31]. Indeed, the algorithms [31] implemented in FastPCA [12] as well as FlashPCA2 [13] can also utilize these ideas to gain additional computational efficiency. Alternate approaches to improve matrix-vector multiplication in the genetics setting include approaches that rely on sparsity of the genotype matrix. It is important to note that the speedup obtained from the Mailman algorithm does not rely explcitly on sparsity and could be applied even to dense matrices. It would be of interest to contrast these approaches and to investigate the potential to combine to leverage sparsity as well as the discrete nature of the genotype matrix.

The probabilistic formulation underlying ProPCA allows the algorithm to be generalized in several directions. One direction is the application of PCA in the presence of missing data that often arises when analyzing multiple datasets. We have explored an extension of the ProPCA model to this setting (SI Section S5). While this approach is promising, a limitation of the use of the Mailman algorithm within ProPCA is the requirement of discrete genotypes, which prevents ProPCA from being directly applied to dosages. Another potential future direction is in modeling linkage disequilibrium and in incorporating rare variants which have the potential to reveal structure that is not apparent from the analysis of common SNPs [32, 33]. Current applications of PCA remove correlated SNPs and singletons though this has been shown to discard information [12]. One possible way to incorporate LD would leverage the connection between haplotype copying models [34] and the multivariate normal model of PCA [35], or by a whitening transformation [4]. Further, the observation model can also be modified to account for the discrete nature of genotypes [3, 36]. A number of non-linear dimensionality reduction methods have been recently proposed [37, 38]. A comparison of these methods to ProPCA (in terms of statistical structure that the methods aim to detect, robustness and ability to handle missing data, as well computational scalability) would be of great interest. Finally, leveraging fine-scale population structure inferred from large-scale data to study recent positive selection in human history is an important direction for future work. The challenge is to design realistic statistical models of population structure while enabling inference at scale.

ProPCA is available at https://github.com/sriramlab/ProPCA.

Material and Methods

Principal Components Analysis (PCA) We observe genotypes from n individuals at m SNPs. The genotype vector for individual i is a length m vector denoted by $\boldsymbol{g}_i \in \{0,1,2\}^m$. The j^{th} entry of $\boldsymbol{g}_i$ denotes the number of minor allele carried by individual i at SNP j. Let $\boldsymbol{G}$ be the $m \times n$ genotype matrix where $\boldsymbol{G} = [\boldsymbol{g}_1 \dots \boldsymbol{g}_n]$. Let $\boldsymbol{Y}$ denote the matrix of standardized genotypes obtained by centering and rescaling each row of the genotype matrix $\boldsymbol{G}$ so that $\sum_j y_{i,j} = 0$ and $\sum_j y_{i,j}^2 = 1$ for all $i \in \{1, \dots, m\}$.

Principal components analysis (PCA) [11] attempts to find a low-dimensional linear transformation of the data that maximizes the projected variance or, equivalently, minimizes the reconstruction error. Given the $m \times n$ matrix $\boldsymbol{Y}$ of standardized genotypes and a target dimension k, PCA attempts to find a $m \times k$ matrix with orthonormal columns $\boldsymbol{W}$ and $n \times k$ matrix $\boldsymbol{Z}$ that minimizes the reconstruction error: $\|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{Z}^{\mathrm{T}}\|_F$ where $\|A\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$ is the Frobenius norm of the matrix $\boldsymbol{A}$. To solve the PCA problem, we perform a singular-value decomposition (SVD) of the standardized genotype matrix $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ and set $\widehat{\boldsymbol{W}} = \boldsymbol{U}_K$, where $\boldsymbol{U}_K$ is a $m \times k$ matrix containing the k columns of $\boldsymbol{U}$ corresponding to the k largest singular vectors of $\boldsymbol{Y}$.

Probabilistic PCA PCA can be viewed as a limiting case of the probabilistic PCA model [10, 16, 17]. Probabilistic PCA models the observed data $\boldsymbol{y}_i \in \mathbb{R}^m, i \in \{1 \dots, n\}$ as a linear transformation of a k-dimensional latent random variable $\boldsymbol{x}_i$ $(k \leq m)$ with additive Gaussian noise. Denoting the linear transformation by the $m \times k$ matrix $\boldsymbol{C}$, and the (m-dimensional) noise by $\boldsymbol{\epsilon}_i$ (with isotropic covariance matrix $\sigma^2 \boldsymbol{I}_m$), the generative model can be written as

$$\begin{aligned} \boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\epsilon}_i &= \boldsymbol{C}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{x}_i &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k) \\ \boldsymbol{\epsilon}_i &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m) \end{aligned} \tag{1}$$

The maximum likelihood estimate of the matrix $\boldsymbol{C}$ in this model has been shown to span the k-dimensional principal subspace of the data $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots \boldsymbol{y}_n]$ [39].

EM algorithm for PCA Since probabilistic PCA is a probabilistic model endowed with latent variables, the EM algorithm presents a natural approach to compute the maximum likelihood estimates of the model parameters $(\boldsymbol{C}, \sigma^2)$ [16, 17]. The EM algorithm for learning the principal components can be derived as a special case of the EM algorithm for the probabilistic PCA model where the variance of the observation noise σ^2 tends to zero leading to these updates:

$$\text{E-step :} \quad \boldsymbol{X} = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{Y} \tag{2}$$

$$\text{M-step :} \quad \boldsymbol{C} = \boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1} \tag{3}$$

Here $\boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_n]$ is a $k \times n$ matrix and $\boldsymbol{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_n]$ is a $m \times n$ matrix. Noting that all matrix inversions require inverting a $k \times k$ matrix, the computational complexity of the E-step is

$\mathcal{O}(k^2m+k^3+k^2m+mnk)$ while the computational complexity of the M-step is $\mathcal{O}(k^2n+k^3+k^2n+mnk)$. For small k and large m, n, the per-iteration runtime complexity is $\mathcal{O}(mnk)$. Thus, the EM algorithm provides a computationally efficient estimator of the top k PCs when the number of PCs to be estimated is small.

Sub-linear time EM The key bottleneck in the EM algorithm is the multiplication of the matrix $\boldsymbol{Y}$ with matrices $\boldsymbol{E} = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T$ and $\boldsymbol{M} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$.

The vectors representing the sample mean and standard deviation of the genotypes at each SNP are denoted $\overline{\boldsymbol{g}}$ and $\boldsymbol{s}$. Assuming no entry in $\boldsymbol{s}$ is zero (we remove SNPs that have no variation across samples), the matrix of standardized genotypes $\boldsymbol{Y}$ can be written as:

$$\boldsymbol{Y} = diag(\boldsymbol{s})^{-1}\boldsymbol{G} - \boldsymbol{\rho}\mathbf{1}_n^{\mathrm{T}}$$

Here $diag(\boldsymbol{x})$ is an operator that constructs a diagonal matrix with the entries of $\boldsymbol{x}$ along its diagonals, $\mathbf{1}_n$ is a length n vector with each entry equal to one, and $\boldsymbol{\rho}$ is a length m vector with $\rho_j = \frac{\overline{g}_j}{s_j}, j \in \{1, \ldots, m\}$.

The EM updates can be written as:

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{E}\boldsymbol{Y} = \boldsymbol{E}\,\mathsf{diag}\left(\boldsymbol{s}\right)^{-1}\boldsymbol{G} - \boldsymbol{E}\boldsymbol{\rho}\mathbf{1}_n^{\mathrm{T}} \\ &= \widetilde{\boldsymbol{E}}\boldsymbol{G} - \boldsymbol{E}\boldsymbol{\rho}\mathbf{1}_n^{\mathrm{T}} \qquad (4) \\ \boldsymbol{C} &= \boldsymbol{Y}\boldsymbol{M} = diag(\boldsymbol{s})^{-1}\boldsymbol{G}\boldsymbol{M} - \boldsymbol{\rho}\mathbf{1}_n^{\mathrm{T}}\boldsymbol{M} \qquad (5) \end{aligned}$$

Here $\widetilde{\boldsymbol{E}}$ can be computed in time $\mathcal{O}(km)$ while $\boldsymbol{E}\boldsymbol{\rho}\mathbf{1}_n^{\mathrm{T}}$ and $\boldsymbol{\rho}\mathbf{1}_n^{\mathrm{T}}\boldsymbol{M}$ can be computed in time $\mathcal{O}(nk+mk)$.

The key bottleneck in the E-step is the multiplication of the genotype matrix $\boldsymbol{G}$ by each of the k rows of the matrix $\widetilde{\boldsymbol{E}}$ and in the M-step, multiplication of $\boldsymbol{G}$ by each of the k columns of the matrix $\boldsymbol{M}$ respectively. Leveraging the fact that each element of the genotype matrix $\boldsymbol{G}$ takes values in the set $\{0, 1, 2\}$, we can improve the complexity of these multiplication operations from $\mathcal{O}(nmk)$ to $\mathcal{O}(\frac{nmk}{\max(\log_3 n, \log_3 m)})$ by extending the Mailman Algorithm [40]. For additional implementation details, see SI Section S1.

The Mailman algorithm In the M-step, we need to compute $\boldsymbol{c} = \boldsymbol{A}\boldsymbol{b}$ for an arbitrary real-valued vector $\boldsymbol{b}$ and a $m \times n$ matrix $\boldsymbol{A}$ whose entries take values in $\{0, 1, 2\}$. We assume that $m = \lceil \log_3(n) \rceil$. Naive matrix-vector multiplication takes $\mathcal{O}(\lceil \log_3(n) \rceil n)$ time.

The Mailman algorithm decomposes $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{U}_n\boldsymbol{P}$. Here $\boldsymbol{U}_n$ is the $m \times r$ matrix whose columns containing all $r = 3^m$ possible vectors over $\{0, 1, 2\}$ of length m. We set an entry $P_{i,j}$ to 1 if column j of $\boldsymbol{A}$ matches column i of U_n: $A^{(j)} = U_n^{(i)}$. The decomposition of any matrix $\boldsymbol{A}$ into $\boldsymbol{U}_n$ and $\boldsymbol{P}$ can be done in $\mathcal{O}(nm)$ time. Given this decomposition, the desired product $\boldsymbol{c}$ is computed in two steps, each of which has $\mathcal{O}(n)$ time complexity [40]:

$$\boldsymbol{d} = \boldsymbol{P}\boldsymbol{b}, \quad \boldsymbol{c} = \boldsymbol{U}_n\boldsymbol{d}$$

The Mailman algorithm provides computational savings in a setting where the cost of computing the decomposition of $\boldsymbol{A}$ are offset by the gains in repeated multiplication involving $\boldsymbol{A}$.

Similarly, in the E-step, we need to compute $\boldsymbol{f}^{\mathrm{T}}\boldsymbol{A}$ in $\mathcal{O}(\lceil \log_3(n) \rceil n)$ time by computing $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{f}$ and computing a decomposition of $\boldsymbol{A}^{\mathrm{T}}$. A drawback of this approach is the need to store both decompositions that would double the memory requirements of the algorithm. Instead, we propose a novel variant of the Mailman algorithm that can compute $\boldsymbol{f}^{\mathrm{T}}\boldsymbol{A}$ in $\mathcal{O}(\lceil \log_3(n) \rceil n)$ time using the same decomposition as $\boldsymbol{A}$ (SI Section S2).

Additional details on efficient implementation of the EM and Mailman algorithms can be found in SI Section S1.

Simulations We simulated genotypes at m independent SNPs across n individuals in which a single ancestral population diverged into q sub-populations with drift proportional to the F_{st}, a measure of population differentiation. The allele frequency at SNP $f_{j,0}, j \in \{1, \ldots, m\}$ in the ancestral population was sampled from a uniform distribution such that $f_{j,0} \overset{iid}{\sim} Unif(0.05, 0.95)$. Allele frequencies in each of the l subpopulations were generated by simulating neutral drift from the ancestral allele frequency, $f_{j,l} \sim \mathcal{N}(f_{j,0}, f_{j,0}(1 - f_{j,0})F_{st}), l \in \{1, \ldots, q\}$ and were set to 0 or 1 if they fell outside the interval $[0, 1]$. The genotypes of an individual in population l at SNP j was sampled from a $Binomial(2, f_{j,l})$ distribution.

Benchmarking To compare estimated PCs to reference PCs, we computed the mean of explained variance (MEV) – a measure of the overlap between the subspaces spanned by the two sets of PCs. Two different sets of K principal components each produce a K-dimensional column space. A metric for the performance of a PCA algorithm against some baseline is to see how much the column spaces 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 $(v_1, v_2, ..., v_k)$ against which we wish to evaluate the performance of a set of estimated PCs $(u_1, u_2, ..., u_k)$, $MEV = \frac{1}{k}\sum_{i=1}^{k}\sqrt{\sum_{j=1}^{k}(\mathbf{v_i} \cdot \mathbf{u_j})^2} = \frac{1}{k}\sum_{i=1}^{k}\|\mathbf{U^T v_i}\|$ where $\mathbf{U}$ is a matrix whose column vectors are the PCs which we are testing.

In practice, when attempting to compute the top k PCs, ProPCA was found to converge faster by computing l PCs for $l > k$ PCs and retaining the top k PCs. We set $l = 2k$ in our experiments. While ProPCA could be run to convergence, we found that running it for k iterations already gave accurate results across the range of parameters considered. Our empirical results are consistent with our theoretical result that the EM algorithm converges exponentially fast in the spectral norm of the error matrix [31, 41] (SI Section S3).

We compared ProPCA to the current state-of-the-art methods for computing PCs from genotype data: the SVD implementation in PLINK (PLINK_SVD [19]), FastPCA [12] and FlashPCA2 [13]. PLINK_SVD refers to an exact computation of PCs using the full Singular Value Decomposition as implemented in the PLINK package (PLINK_SVD). FastPCA [12] is an implementation of the randomized subspace iteration method [31] while FlashPCA2 [13] is an implementation of the implicitly restarted Arnoldi method [42]. We used default parameters for all methods. All experiments were performed on a Intel(R) Xeon(R) CPU 2.10GHz server with 128 GB RAM, restricted to a single core, capped to a maximum runtime of 100 hours and a maximum memory of 64 GB.

Selection scan The White British cohort was identified by the UK Biobank as participants who self-identified as 'British' within the broader-level group 'White' while having similar ancestral background [18]. For our selection scan, we further filtered the $409,634$ individuals in the White British subset to obtain an unrelated White British subset by removing individuals with one other related individual in the data set (individuals with kinship coefficients greater than 0.0625 (third-degree relatedness) to any other individual as determined by KING [43]). After removing these individuals, we obtained an unrelated White British subset containing $276,736$ individuals.

We inferred the top five PCs using ProPCA on all $276,736$ unrelated White British individuals and a filtered SNP set containing $146,671$ SNPs (UK Biobank SNP set). SNPs in the UK Biobank SNP set consist of SNPs on the UK Biobank Axiom array from which SNPs were removed if they have missing rates greater than 1.5%, minor allele frequencies (MAF) less than 1%, or if they were in regions of long-range linkage disequilibrium. The remaining SNPs were then pruned for pairwise r^2 less than 0.1 using windows of 1000 base pairs (bp) and a step-size of 80 bp.

We developed a selection statistic to search for SNPs whose variation is not well-explained by the ProPCA model (closely related to the selection statistic proposed in [12]). Under the probabilistic PCA model, the normalized genotype matrix is modeled by a low rank approximation and Gaussian noise, $\mathbf{Y} = \mathbf{CX} + \boldsymbol{\epsilon}$. Given our low rank approximation of the genotype matrix, $\hat{\mathbf{Y}} = \mathbf{CX}$, we have the residual : $\mathbf{Y} - \hat{\mathbf{Y}} = \boldsymbol{\epsilon}$. For a SNP j, the Gaussian noise, $\boldsymbol{\epsilon_j} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Projecting this residual onto a PC results in a univariate Gaussian with zero mean and constant variance across SNPs. This variance can be estimated as the sample variance $\hat{\sigma}^2$ of the resulting statistics across SNPs. In summary, we propose the statistic: $\frac{((\mathbf{y}_j - \hat{\mathbf{y}}_j)^T \mathbf{x}_k)^2}{\hat{\sigma}^2} \sim \chi^2_1$ for SNP j, given the k-th PC. The projection of the residual onto a PC allows the signal of selection to be interpreted in the context of deviations from ancestry captured by the specific PC.

Furthermore, a variant of this statistic, which we call the combined statistic, can be generated from the selection statistics computed on each individual PC using the observation that the resulting chi-squared statistics are independent of each other. This allows us to create an additional statistic by summing the individual PC statistics to create a combined statistic that follows a chi-squared distribution with additional degrees of freedom for each PC used.

Using the results from the PCA on the UK Biobank SNP set, we performed our selection scan on a different set of $516,140$ SNPs. We generated this set of SNPs by removing SNPs that were multi-allelic, had genotyping rates less than 99%, had minor allele frequencies less than 1%, and were not in Hardy-Weinberg equilibrium ($p < 10^{-6}$).

We performed an allele frequency test for each novel SNP using the Nomenclature of Territorial Units for Statistics level 3 (NUTS3) classification of regions for the UK. The NUTS3 classification defines non-overlapping borders for each region in the UK, allowing us to uniquely map each individual to a region in the UK using their birth location coordinates by checking which NUTS3 regions they fell into. For each of our novel loci, we then performed an two-tailed Z-test between each region's allele frequency against all other regions. We corrected for multiple testing using the Bonferroni correction.

Supplemental Data

Supplemental data includes 7 figures and 8 tables. 2 of 8 of the tables can be found in the accompanying Excel file.

Acknowledgements

We would like to thank Bogdan Pasaniuc, members of his lab, and other members of the Sankararaman lab for advice and comments on this project. This research was conducted using the UK Biobank Resource under application 33127.

References

1. John Novembre and Sohini Ramachandran. Perspectives on human population structure at the cusp of the sequencing era. *Annual review of genomics and human genetics*, 12:245–274, 2011.

2. J Novembre, T Johnson, K Bryc, Z Kutalik, AR Boyko, A Auton, A Indap, KS King, S Bergmann, MR Nelson, M Stephens, and Bustamante CD. Genes mirror geography within europe. *Nature*, 456(7219):274, 2008.

3. Wen-Yun Yang, John Novembre, Eleazar Eskin, and Eran Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nature genetics*, 44(6):725–731, 2012.

4. Yael Baran, Inés Quintela, Ángel Carracedo, Bogdan Pasaniuc, and Eran Halperin. Enhanced localization of genetic samples through linkage-disequilibrium correction. *The American Journal of Human Genetics*, 92(6):882–894, 2013.

5. Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature reviews. Genetics*, 11(7):459, 2010.

6. Nick Patterson, Alkes L. Price, and David Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190+, December 2006.

7. C. L. Hanis, R. Chakraborty, R. E. Ferrell, and W. J. Schull. Individual admixture estimates: disease associations and individual risk of diabetes and gallbladder disease among mexican-americans in starr county, texas. *Am J Phys Anthropol*, 70(4):433–441, August 1986.

8. J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

9. Chibiao Chen, Eric Durand, Florence Forbes, and Olivier François. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Resources*, 7(5):747–756, 2007.

10. Barbara E Engelhardt and Matthew Stephens. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS genetics*, 6(9):e1001117, 2010.

11. Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

12. Kevin J Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J Patterson, and Alkes L Price. Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *The American Journal of Human Genetics*, 98(3):456–472, 2016.

13. Gad Abraham, Yixuan Qiu, and Michael Inouye. Flashpca2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics*, 2017.

14. A. Price, N. Patterson, R. Plenge, M. Weinblatt, N. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.

15. Oriol Canela-Xandri, Andy Law, Alan Gray, John A Woolliams, and Albert Tenesa. A new tool called dissect for analysing large genomic data sets using a big data approach. *Nature communications*, 6:10162, 2015.

16. Sam T Roweis. Em algorithms for pca and spca. In *Advances in neural information processing systems*, pages 626–632, 1998.

17. Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

18. Bycroft C et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562:203–209, 2018.

19. Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

20. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov 2012.

21. Mark D Shriver, Giulia C Kennedy, Esteban J Parra, Heather A Lawson, Vibhor Sonpar, Jing Huang, Joshua M Akey, and Keith W Jones. The genomic distribution of population substructure in four populations using 8,525 autosomal snps. *Human genomics*, 1(4):274, 2004.

22. Chao Tian, Robert M Plenge, Michael Ransom, Annette Lee, Pablo Villoslada, Carlo Selmi, Lars Klareskog, Ann E Pulver, Lihong Qi, Peter K Gregersen, et al. Analysis and application of european genetic substructure using 300 k snp information. *PLoS genetics*, 4(1):e4, 2008.

23. Gerhardt C Wiegering A, Ruther U. The ciliary protein rpgrip1l in development and disease. *Dev Biol*, 442(1):60–68, 2018.

24. Marion Delous et al. The ciliary gene rpgrip1l is mutated in cerebello-oculo-renal syndrome (joubert syndrome type b) and meckel syndrome. *Nature Genetics*, 39:875–881, 2007.

25. Oliver Devuyst and Veronique J. Arnould. Mutations in rpgrip1l : extending the clinical spectrum of ciliopathies. *Nephrology Dialysis Transplantation*, 23(5):15001503, 2008.

26. Khanna H et al. A common allele in rpgrip1l is a modifier of retinal degeneration in ciliopathies. *Nature Genetics*, 41(6):739–45, 2009.

27. Aschard et al. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *AJHG*, 94(5):662–76, 2014.

28. Korneev et al. Tlr-signaling and proinflammatory cytokines as drivers of tumorigenesis. *Cytokine*, 89:127–135, 2017.

29. Mockenhaupt et al. Toll-like receptor (tlr) polymorphisms in african children: Common tlr-4 variants predispose to severe malaria. *PNAS*, 103(1):177182, 2006.

30. Van der Graaf et al. Toll-like receptor 4 asp299gly/thr399ile polymorphisms are a risk factor for candida bloodstream infection. *European Cytokine Network*, 17(1):2934, 2006.

31. Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

32. Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, 44(3):243, 2012.

33. Garrett Hellenthal, Adam Auton, and Daniel Falush. Inferring human colonization history using a copying model. *PLoS Genet*, 4(5):e1000078, 05 2008.

34. Na Li and Matthew Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165(4):2213–2233, 2003.

35. Xiaoquan Wen and Matthew Stephens. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, 4(3):1158, 2010.

36. Andrew I Schein, Lawrence K Saul, and Lyle H Ungar. A generalized linear model for principal component analysis of binary data. In *AISTATS*, volume 3, page 10, 2003.

37. Li et al. Application of t-sne to human genetic data. *J Bioinform Comput Biol.*, 15(4):1750017, 2017.

38. Becht et al. Dimensionality reduction for visualizing single-cell data using umap. *Nat Biotechnol.*, 37:38–44, 2019.

39. Theodore W Anderson and Herman Rubin. Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pages 111–150, 1956.

40. Edo Liberty and Steven W Zucker. The mailman algorithm: A note on matrix–vector multiplication. *Information Processing Letters*, 109(3):179–182, 2009.

41. Arthur Szlam, Andrew Tulloch, and Mark Tygert. Accurate low-rank approximations via a few iterations of alternating least squares. *SIAM Journal on Matrix Analysis and Applications*, 38(2):425–433, 2017.

42. Richard B Lehoucq and Danny C Sorensen. Deflation techniques for an implicitly restarted arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, 17(4):789–821, 1996.

43. Manichaikul et al. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):28672873, 2010.

Figure 1: **ProPCA is computationally efficient**: Comparison of runtimes over simulated genotype data containing 100, 000 SNPs, six subpopulations, $F_{st} = 0.10$ and individuals varying from 10, 000 to 1, 000, 00. We report the mean and standard deviation over ten trials. Figures 1a and 1b display the total runtime (capped to a maximum of 100 hours and a maximum memory of 64 GB). Figure 1b compares the runtimes of all algorithms excluding PLINK_SVD which could only run successfully up to a sample size of 70, 000.

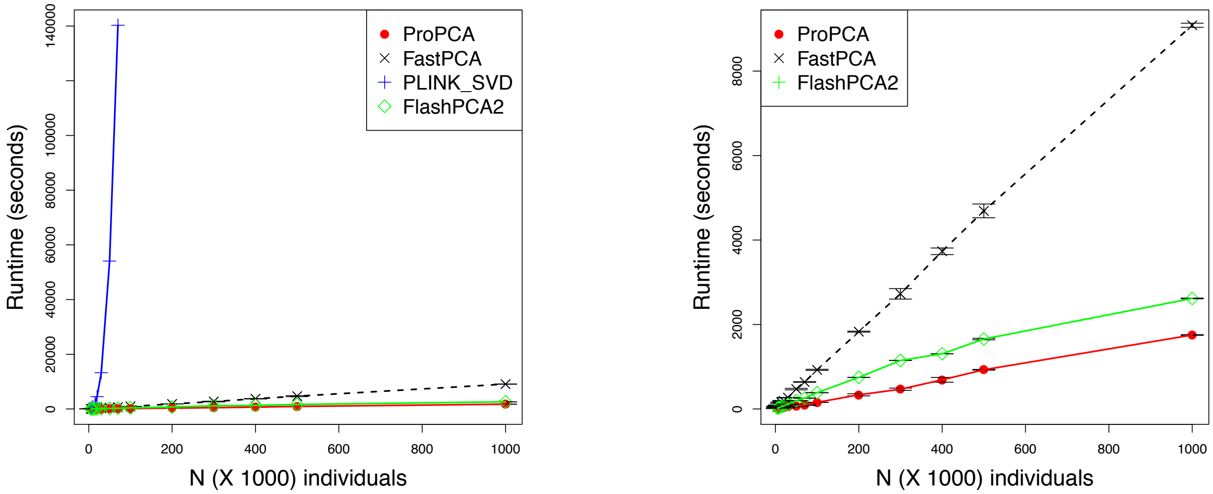
Figure 2: **ProPCA is computationally efficient relative to FastPCA**: We compute the total time taken to estimate the top five principal components as a function of a measure of accuracy (MEV) for ProPCA compared to FastPCA. We performed these comparisons on simulated genotype data containing 50, 000 SNPs, 10, 000 individuals, six subpopulations, and $F_{st} \in \{0.001, 0.005, 0.10\}$. We were unable to leverage the source code of FlashPCA2 to include in these comparisons.

Figure 3: **Principal components uncover population and geographic structure in the UK Biobank**: We used ProPCA to compute PCs on the UK Biobank data. Figure 3a shows the first two principal components to reveal population structure. Figure 3b shows geographic structure by plotting the score of 276, 736 unrelated White British individuals on the first principal component on their birth location coordinates.

Figure 4: **Selection scan for the first five principal components in the white British individuals in the UK Biobank**: A Manhattan plot with the $-\log_{10} p$ values associated with the test of selection displayed for the first five principal components for the unrelated White British subset of the UK Biobank. The red line represents the Bonferroni adjusted significance level ($\alpha = 0.05$). Significant loci are labeled. Signals above $-log_{10}(p) = 18$ were capped at this value for better visualization.

F_{st}	MEV	
	$K = 5$	$K = 10$
0.001	0.987	1.000
0.002	0.999	1.000
0.003	0.999	1.000
0.004	0.999	1.000
0.005	1.000	1.000
0.006	1.000	1.000
0.007	1.000	1.000
0.008	1.000	1.000
0.009	1.000	1.000
0.010	1.000	1.000

Table 1: **ProPCA accurately estimates principal components**: The principal components computed by ProPCA are compared to the PCs obtained from a full SVD on a genotype dataset containing 50, 000 SNPs and 10, 000 individuals. Accuracy was measured by the mean of explained variance (MEV) which measures the overlap between the set of PCs inferred from ProPCA and those from SVD across values of $F_{st} \in \{0.001, \ldots, 0.01\}$. We report MEV for $K = 5$ using 5 populations as well as for $K = 10$ PCs using 10 populations.



(a)

(b)

Figure 1: **ProPCA is computationally efficient**: Comparison of runtimes over simulated genotype data containing 100, 000 SNPs, six subpopulations, $F_{st} = 0.10$ and individuals varying from 10, 000 to 1, 000, 00. We report the mean and standard deviation over ten trials. Figures 1a and 1b display the total runtime (capped to a maximum of 100 hours and a maximum memory of 64 GB). Figure 1b compares the runtimes of all algorithms excluding PLINK_SVD which could only run successfully up to a sample size of 70, 000.

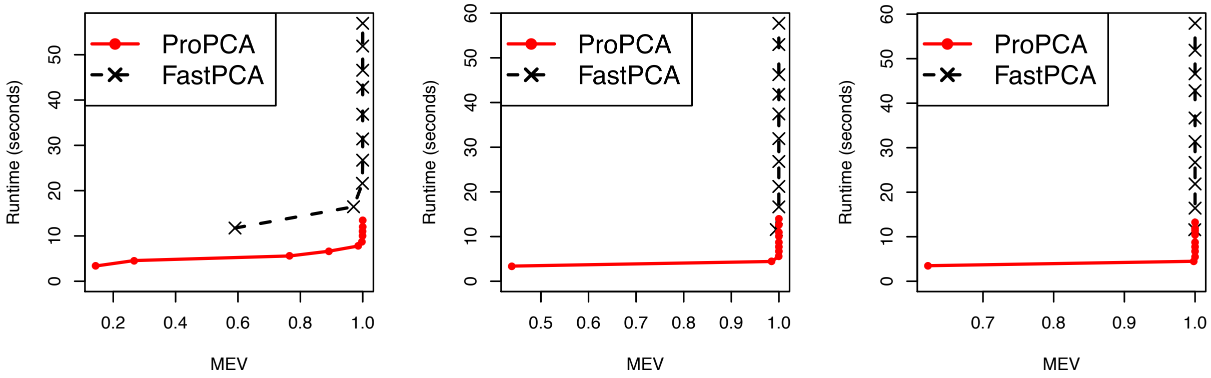
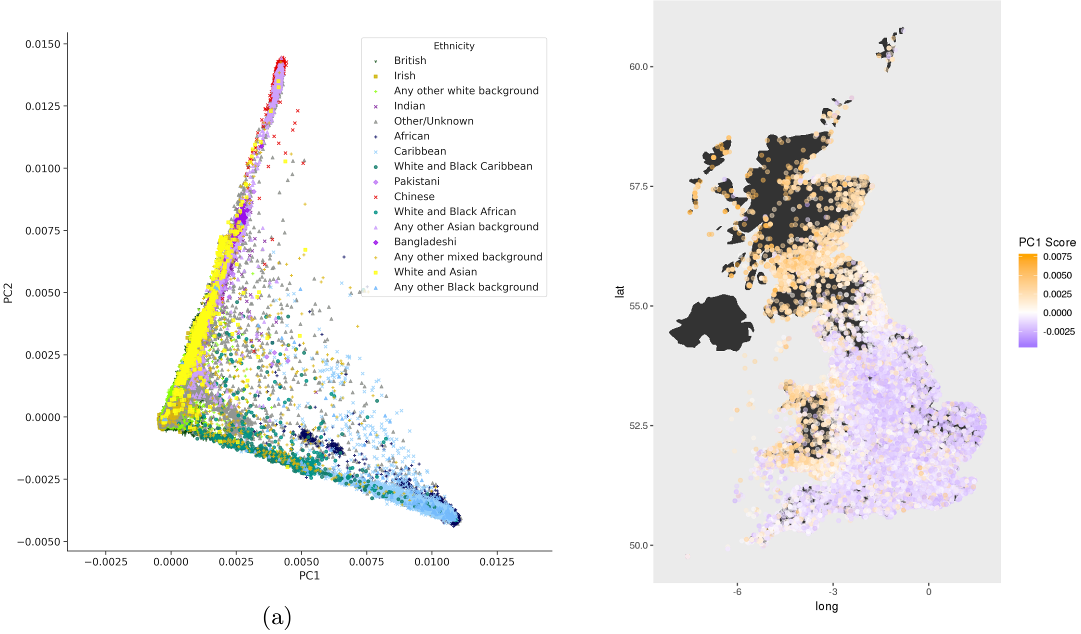


Figure 2: **ProPCA is computationally efficient relative to FastPCA**: We compute the total time taken to estimate the top five principal components as a function of a measure of accuracy (MEV) for ProPCA compared to FastPCA. We performed these comparisons on simulated genotype data containing 50, 000 SNPs, 10, 000 individuals, six subpopulations, and $F_{st} \in \{0.001, 0.005, 0.10\}$. We were unable to leverage the source code of FlashPCA2 to include in these comparisons.



(a)

(b)

Figure 3: **Principal components uncover population and geographic structure in the UK Biobank**: We used ProPCA to compute PCs on the UK Biobank data. Figure 3a shows the first two principal components to reveal population structure. Figure 3b shows geographic structure by plotting the score of 276, 736 unrelated White British individuals on the first principal component on their birth location coordinates.

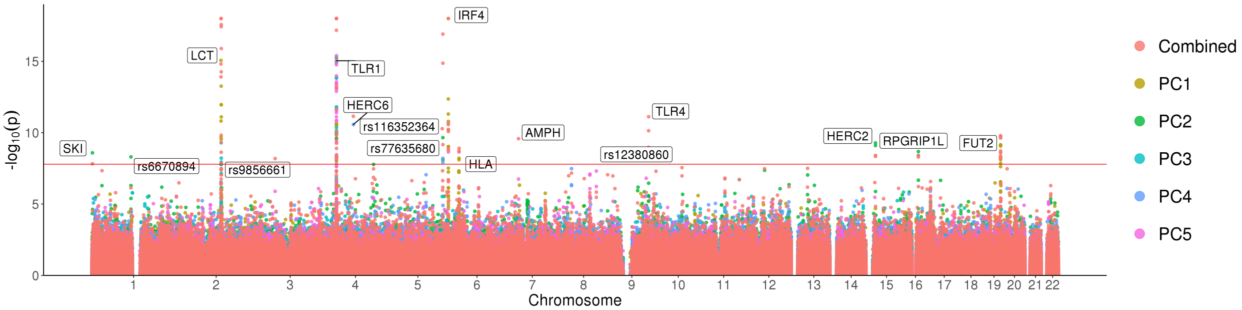


Figure 4: **Selection scan for the first five principal components in the white British individuals in the UK Biobank**: A Manhattan plot with the $-\log_{10} p$ values associated with the test of selection displayed for the first five principal components for the unrelated White British subset of the UK Biobank. The red line represents the Bonferroni adjusted significance level ($\alpha = 0.05$). Significant loci are labeled. Signals above $-log_{10}(p) = 18$ were capped at this value for better visualization.