Inferring population structure in biobank-scale genomic data

Authors

Alec M. Chiu, Erin K. Molloy, Zilong Tan, Ameet Talwalkar, Sriram Sankararaman

Correspondence

sriram@cs.ucla.edu

Chiu et al., 2022, The American Journal of Human Genetics *109*, 727–737
April 7, 2022
https://doi.org/10.1016/j.ajhg.2022.02.015

CellPress

Inferring population structure in biobank-scale genomic data

Alec M. Chiu,[1] Erin K. Molloy,[2,3] Zilong Tan,[4] Ameet Talwalkar,[5] and Sriram Sankararaman[1,2,6,7,*]

Summary

Inferring the structure of human populations from genetic variation data is a key task in population and medical genomic studies. Although a number of methods for population structure inference have been proposed, current methods are impractical to run on biobank-scale genomic datasets containing millions of individuals and genetic variants. We introduce SCOPE, a method for population structure inference that is orders of magnitude faster than existing methods while achieving comparable accuracy. SCOPE infers population structure in about a day on a dataset containing one million individuals and variants as well as on the UK Biobank dataset containing 488,363 individuals and 569,346 variants. Furthermore, SCOPE can leverage allele frequencies from previous studies to improve the interpretability of population structure estimates.

Introduction

Inference of population structure is a central problem in human genetics with applications that range from fine-grained understanding of human history[1] to correcting for population stratification in genome-wide association studies (GWASs).[2] Approaches to population structure inference[3–8] typically formalize the problem as one of estimating admixture proportions of each individual and ancestral population allele frequencies given genetic variation data.

The growth of repositories of genetic variation data over large numbers of individuals has opened up the possibility of inferring population structure at increasingly finer resolution.[9,10] For instance, the UK Biobank[9] contains genotype data from approximately half a million British individuals across millions of SNPs. This development has necessitated methods that can be applied to large-scale datasets with reasonable runtime and memory requirements. Existing methods, however, do not scale to these datasets. Thus, we have developed SCOPE (scalable population structure inference)—a scalable method capable of inferring population structure on biobank-scale data.

SCOPE utilizes a previously proposed likelihood-free framework[8] that involves estimation of the individual allele frequency (IAF) matrix through a statistical technique known as latent subspace estimation (LSE)[11] followed by a decomposition of the estimated IAF matrix into ancestral allele frequencies and admixture proportions. SCOPE uses two ideas to substantially improve the scalability of this approach. First, SCOPE uses randomized eigendecomposition[12] to efficiently estimate the latent subspace. Specifically, SCOPE avoids the need to form matrices that are expensive to compute on or require substantial memory and instead works directly with the input genotype matrix. Second, SCOPE leverages the insight that the resulting method involves repeated multiplications of the genotype matrix and uses the Mailman algorithm for fast multiplication of the genotype matrix.[13]

We benchmarked the accuracy and efficiency of SCOPE on simulated and real datasets. In simulations, SCOPE obtains accuracy comparable to existing methods while being up to 1,800 times faster. Relative to the previous state-of-the-art scalable method (TeraStructure[7]), SCOPE is three to 144 times faster. SCOPE can estimate population structure in about a day for a simulated dataset consisting of one million individuals and SNPs for six latent populations, whereas TeraStructure is extrapolated to require approximately 20 days on this same dataset. We additionally used SCOPE to infer continental ancestry proportions (four ancestry groups) on the UK Biobank dataset (488,363 individuals and 569,346 SNPs) in about a day. We find that the inferred continental ancestry proportions are highly concordant with self-reported race and ethnicity (SIRE).

SCOPE additionally can be applied in a supervised setting. Given allele frequencies from reference populations,[14,15] SCOPE can estimate admixture proportions corresponding to the reference populations to enable greater interpretability.

Subjects and methods

The structure/admixture model

The structure/admixture model links the $m \times n$ genotype matrix $\boldsymbol{X}$ (where rows refer to single nucleotide polymorphisms [SNPs] and

[1]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, USA; [2]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90095, USA; [3]Institute for Advanced Computer Studies, University of Maryland, College Park, College Park, MD 20742, USA; [4]Facebook, Inc., Menlo Park, CA 94025, USA; [5]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA; [6]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA; [7]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
*Correspondence: sriram@cs.ucla.edu
https://doi.org/10.1016/j.ajhg.2022.02.015.

columns refer to individual diploid genotypes, $x_{ij} \in \{0, 1, 2\}, i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}$) to the $m \times n$ individual allele frequency (IAF) matrix $\boldsymbol{F}$, $m \times k$ ancestral population allele frequencies $\boldsymbol{P}$, and the $k \times n$ individual admixture proportions $\boldsymbol{Q}$ (also termed the global ancestry of an individual). Here, m denotes the number of SNPs, n denotes the number of individuals, and k denotes the number of latent populations. The IAF matrix, ancestral allele frequencies, and admixture proportions are mathematically related as $\boldsymbol{F} = \boldsymbol{PQ}$. Furthermore, there are constraints on $\boldsymbol{P}$ and $\boldsymbol{Q}$. Each element of $\boldsymbol{P}$ is constrained to lie between 0 and 1 ($0 \leq p_{il} \leq 1, i \in \{1, \ldots, m\}, l \in \{1, \ldots, k\}$). Each element of $\boldsymbol{Q}$ is non-negative ($q_{lj} \geq 0, l \in \{1, \ldots, k\}, j \in \{1, \ldots, n\}$) and the admixture proportion of each individual must sum to one ($\sum_l q_{lj} = 1$). Finally, each entry of the genotype matrix is an independent draw from the corresponding entry of the IAF matrix F as: $x_{ij} | f_{ij} \sim \text{Binomial}(2, f_{ij})$. The goal of population structure inference under the structure/admixture model is to estimate $\boldsymbol{P}$ and $\boldsymbol{Q}$ given $\boldsymbol{X}$.

SCOPE

For scalable inference, SCOPE uses as its starting point a likelihood-free estimator of population structure previously proposed in ALStructure.[8] This estimator has two major steps: latent subspace estimation (LSE) and alternating least-squares (ALS). LSE attempts to estimate the subspace spanned by the rows of $\boldsymbol{Q}$[11] by computing a low-rank approximation to the matrix $\boldsymbol{G} = (1/m)\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{D}$ where each entry d_j of the $n \times n$ diagonal matrix $\boldsymbol{D}$ is obtained as $d_j = (1/m)\sum_{i=1}^{m} 2x_{ij} - x_{ij}^2$. The latent subspace of $\boldsymbol{Q}$ is estimated as the span of the top k eigenvectors of $\boldsymbol{G}$: $v_1, \ldots, v_k$. After obtaining the top k eigenvectors $\boldsymbol{V} = [v_1, \ldots, v_k]$, ALStructure projects the data $\boldsymbol{X}$ onto $\boldsymbol{V}$ to obtain an estimate of $\boldsymbol{F}$: $\widehat{\boldsymbol{F}} = (1/2)\boldsymbol{XVV}^T$. It then uses truncated alternating least-squares (ALS) to factorize the estimate, $\widehat{\boldsymbol{F}}$, into estimates of $\boldsymbol{P}$ and $\boldsymbol{Q}$: $\widehat{\boldsymbol{F}} = \widehat{\boldsymbol{P}}\widehat{\boldsymbol{Q}}$. $\widehat{\boldsymbol{Q}}$ are the estimates of the individual admixture proportions.

A naive approach to compute the top k eigenvectors of G would involve first forming the matrix $\boldsymbol{G}$ and then computing its top k eigenvectors, which would require $\mathcal{O}(n^2m + n^2k)$ (if a full SVD is performed, this step would require $\mathcal{O}(\min(n, m)nm)$). To perform scalable LSE, SCOPE uses techniques from randomized linear algebra,[12] specifically the implicitly restarted Arnoldi method,[16] to obtain the top k eigenvectors. This step involves repeatedly multiplying estimates of the eigenvectors $v_l : l \in \{1, \ldots, k\}$ with the genotype matrix: $((1/m)\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{D})v_l = (1/m)((\boldsymbol{X}v_l)^T\boldsymbol{X})^T - \boldsymbol{D}v_l$ and can be performed without explicitly forming the matrix $\boldsymbol{G}$. Instead, this approach requires repeatedly computing $w_l \equiv \boldsymbol{X}v_l$, $w_l^T\boldsymbol{X}$, and $\boldsymbol{D}v_l$, which can be computed in $\mathcal{O}(nmk)$ time. We use the C++ Spectra library (web resources) to implement these computations in SCOPE.

To efficiently compute $\widehat{\boldsymbol{P}}$ and $\widehat{\boldsymbol{Q}}$ with truncated ALS, we randomly initialized the matrix $\widehat{\boldsymbol{P}}$ with all values between 0 and 1 ($0 \leq \widehat{p}_{il} \leq 1$). We iteratively solve for estimates of $\boldsymbol{P}$ and $\boldsymbol{Q}$, projecting the estimates onto the constraint space until convergence:

$$\widehat{\boldsymbol{Q}} = \frac{1}{2}(\widehat{\boldsymbol{P}}^T\widehat{\boldsymbol{P}})^{-1}\widehat{\boldsymbol{P}}^T\boldsymbol{XVV}^T$$

$$\widehat{\boldsymbol{P}} = \frac{1}{2}\boldsymbol{XVV}^T\widehat{\boldsymbol{Q}}(\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^T)^{-1}.$$

All values in $\widehat{\boldsymbol{P}}$ are truncated to be between 0 and 1 while $\widehat{\boldsymbol{Q}}$ is projected onto the appropriate simplex. Each step of the ALS algorithm has runtime $\mathcal{O}(nmk)$. We note here that we never store $\widehat{\mathbf{F}}$ but instead compute it implicitly per iteration. This allows us to reduce the memory footprint of SCOPE, as $\widehat{\mathbf{F}}$ is a continuous, real-valued matrix with the same dimensions as the genotype matrix. It is not feasible for most computers to be able to store this in memory. For instance, to store our larger UK Biobank dataset (488,363 individuals and 569,346 SNPs), one is estimated to require around 2,072 GB of memory.

Each of the computations in SCOPE requires multiplying a genotype matrix with entries consisting of only 0, 1, and 2 for diploid genotype. These operations can be efficiently performed with the Mailman algorithm,[13] which provides computational savings when there are repeated multiplications involving a matrix with a finite alphabet. We utilize the Mailman algorithm in computations involving the genotype matrix in both LSE and ALS so that the final time complexity of SCOPE is $\mathcal{O}(nmk/\max(\log_3 n, \log_3 m))$.

Supervised population structure inference

SCOPE can utilize allele frequencies from reference populations to infer corresponding admixture proportions. In this scenario, we assume $\widehat{\boldsymbol{P}}$, the population allele frequencies, are known. As a result, one only needs to compute $\widehat{\boldsymbol{Q}}$ by using the supplied $\widehat{\boldsymbol{P}}$. This allows the admixture proportions corresponding to the reference populations to be inferred in a single step of ALS once the LSE step is completed.

Permutation matching of inferred results

The output of population structure inference methods can result in output that is permuted even between different runs of the same method. It is critical to correctly match latent populations between methods and runs in order to properly assess results. To perform permutation matching, we employed a strategy similar to that of Behr et al.[17] This permutation matching problem is better known as the assignment problem, which can be solved efficiently with linear programming. We first construct a score matrix by using the distance metric created in Behr et al.[17] The optimal permutation match can then be found by optimizing the total score from assignments through linear programming. We utilize the *lpSolve* (web resources) package in R to solve the linear program.

PSD model simulations

We perform simulations under the STRUCTURE or Pritchard-Stephens-Donnelly (PSD) model.[3] In the PSD model, priors are placed on $\boldsymbol{P}$ and $\boldsymbol{Q}$:

$$p_{il} \overset{iid}{\sim} \text{Beta}\left(\frac{1 - F_{ST}}{F_{ST}}p_A, \frac{1 - F_{ST}}{F_{ST}}(1 - p_A)\right), i \in \{1, \ldots, m\}, l \in \{1, \ldots, k\}$$

$$q_{:j} \overset{iid}{\sim} \text{Dirichlet}(\alpha \mathbf{1}_K), j \in \{1, \ldots, n\}.$$

The allele frequencies p_{il} are drawn from the Balding-Nichols model,[18] which is a beta distribution parametrized by the fixation index (F_{ST}) and an initial allele frequency (p_A). For our simulations, we calculated F_{ST} and p_A from our real datasets. Admixture proportions $\boldsymbol{q}_{:j}$ are drawn at random from a Dirichlet distribution. We take the product of the two matrices to form the IAF matrix, $\boldsymbol{F} = \boldsymbol{PQ}$, and draw each genotype from a binomial distribution parametrized by entries of $\boldsymbol{F}$: $x_{ij} \sim \text{Binomial}(2, f_{ij})$.

Spatial model simulations

We also perform simulations under a spatial model similar to that in Ochoa and Storey.[19] In the spatial model, allele frequencies p_{il} are drawn as in the PSD model, but the admixture proportions, $\boldsymbol{q}$, are drawn from a 1D geography.

$$z \equiv (1, \ldots, k)$$

$$y_j \overset{\widetilde{iid}}{\sim} \text{Uniform}(0, k+1)$$

$$q_{lj} = \frac{f_{z_l}(y_j)}{\sum_{l=1}^{k} f_{z_l}(y_j)}$$

Populations are placed at integer values on a line. We get the resulting population position vector, $z \equiv (1, \ldots, k)$. Each individual has a position, y_j drawn from a uniform distribution between 0 and $k+1$. Proportions for each population are generated via a normal distribution, where f_{z_l} denotes the normal density function with z_l ($l \in \{1, \ldots, k\}$) as the mean and σ^2 as variance. The resulting vector of proportions is then normalized to satisfy the constraints on $\boldsymbol{Q}$. We used $\sigma^2 = 4$ for our simulations.

Assessment of results

We assess our results by using two metrics: average Jensen-Shannon divergence (JSD) and average root-mean-square error (RMSE). We calculate the metrics between the true global ancestry proportions, $\boldsymbol{Q}$, and the estimates, $\widehat{\boldsymbol{Q}}$, after $\widehat{\boldsymbol{Q}}$ has been permutation matched to the true proportions.

$$\text{RMSE}(\boldsymbol{Q}, \widehat{\boldsymbol{Q}}) = \frac{1}{\sqrt{nk}} \|\boldsymbol{Q} - \widehat{\boldsymbol{Q}}\|_F$$

$$\text{JSD}(\boldsymbol{Q}, \widehat{\boldsymbol{Q}}) = \frac{1}{2}\left[\text{KL}\left(\boldsymbol{Q}, \frac{1}{2}[\boldsymbol{Q} + \widehat{\boldsymbol{Q}}]\right) + \text{KL}\left(\widehat{\boldsymbol{Q}}, \frac{1}{2}[\boldsymbol{Q} + \widehat{\boldsymbol{Q}}]\right)\right]$$

$\|\cdot\|_F$ represents the Frobenius norm. KL is the Kullback-Leibler divergence, which is defined as:

$$\text{KL}(\boldsymbol{Q}, \widehat{\boldsymbol{Q}}) = \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{k} q_{lj} \log\left(\frac{q_{lj}}{\widehat{q}_{lj}}\right).$$

In the JSD calculations, we replace values of 0 in $\boldsymbol{Q}$ or $\widehat{\boldsymbol{Q}}$ with 1×10^{-9} to avoid numerical issues.

Datasets

We use the 1000 Genomes Project (TGP),[14,15] Human Origins (HO),[20] Human Genome Diversity Project (HGDP),[21,22] and the UK Biobank (UKB)[9] in this study. The HGDP dataset is the complete Stanford HGDP SNP genotyping data filtered to only include individuals in the H952 set,[23] greater than 95% genotyping rate, and greater than 1% minor allele frequency (MAF), resulting in 940 individuals and 642,951 SNPs. The TGP dataset is the 2012-01-31 Omni Platform genotypes filtered to only include unrelated individuals, greater than 95% genotyping rate, and greater than 1% MAF, resulting in 1,718 individuals and 1,854,622 SNPs. The HO dataset was filtered for human-only samples, greater than 99% genotyping rate, and greater than 5% MAF, resulting in 1,931 individuals and 385,089 SNPs. For the UK Biobank, we filtered the UK Biobank Axiom Array genotypes for greater than 1% MAF, long-range linkage disequilibrium (LD), and pairwise LD pruning in 50 kilobase windows, 80 variant step size, and an r^2 threshold of 0.1, resulting in 488,363 individuals and 568,346 SNPs. This is similar to the UK Biobank manuscript's first round of quality control for principal-component analysis (PCA)[9] with the differences of using all individuals and no genotype filter. We also use the UK Biobank's final set of PCA SNPs,[9] which consists of 147,604 SNPs, to explore higher number of latent populations. We calculate metrics such as F_{ST} from the provided population and superpopulation labels provided by each dataset. To perform our supervised analyses, we use the common SNPs between the datasets involved. All genotype processing was performed with PLINK.[24] Links to the publicly available datasets as well as scripts to apply our preprocessing are available in the code repository for SCOPE.

Visualization of results

We visualize our inferred admixture proportions as stacked bar plots. We permutation matched estimates from all methods to enable easy comparison. For our PSD simulations, we performed hierarchical clustering with complete linkage on a Euclidean distance matrix calculated from the true admixture proportion matrix ($\boldsymbol{Q}$) to obtain the order of samples. For our spatial simulations, we sorted by decreasing membership of the first population. For our real datasets, we perform the same hierarchical clustering strategy used for our PSD simulations but use the estimates from ADMIXTURE ($\widehat{\boldsymbol{Q}}$) in place of the true admixture proportions. For the HGDP, TGP, and UK Biobank, we first took the average proportions for each SIRE group and performed hierarchical clustering on the averages to determine the order of the SIRE groups. We then performed hierarchical clustering within each SIRE group to determine the order of individuals within groups. For large datasets, we utilized *genieclust*,[25] a scalable method for hierarchical clustering.

Benchmarking

We compared SCOPE to ADMIXTURE v1.3.0,[5] fastSTRUCTURE,[6] TeraStructure,[7] ALStructure v0.1.0,[8] and sNMF v1.2.[26]

ADMIXTURE computes maximum-likelihood estimates while TeraStructure and fastSTRUCTURE compute approximate posterior estimates in a Bayesian model with variational inference. ALStructure, the framework that SCOPE builds upon, utilizes a two-stage strategy of first performing dimensionality reduction (latent subspace estimation) followed by matrix factorization (alternating least-squares).

Each method was run with eight threads with the exception of fastSTRUCTURE and ALStructure, which do not have multi-threaded implementations. Default parameters were used. TeraStructure has an additional "rfreq" parameter, which was set to 10% of the number of SNPs as recommended by its authors. For SCOPE, we used convergence criteria of either 1,000 iterations of the ALS algorithm or a change between iterations less than 1×10^{-5}, which we calculate as the RMSE between the estimated admixture matrices between two iterations. All experiments were performed on a server with two AMD EPYC 7501 32-Core Processors and 1 terabyte of RAM.

Results

Accuracy

We assessed the accuracy of SCOPE by using simulations under the Pritchard-Stephens-Donnelly (PSD) model[3] to study accuracy under a standard population genetics

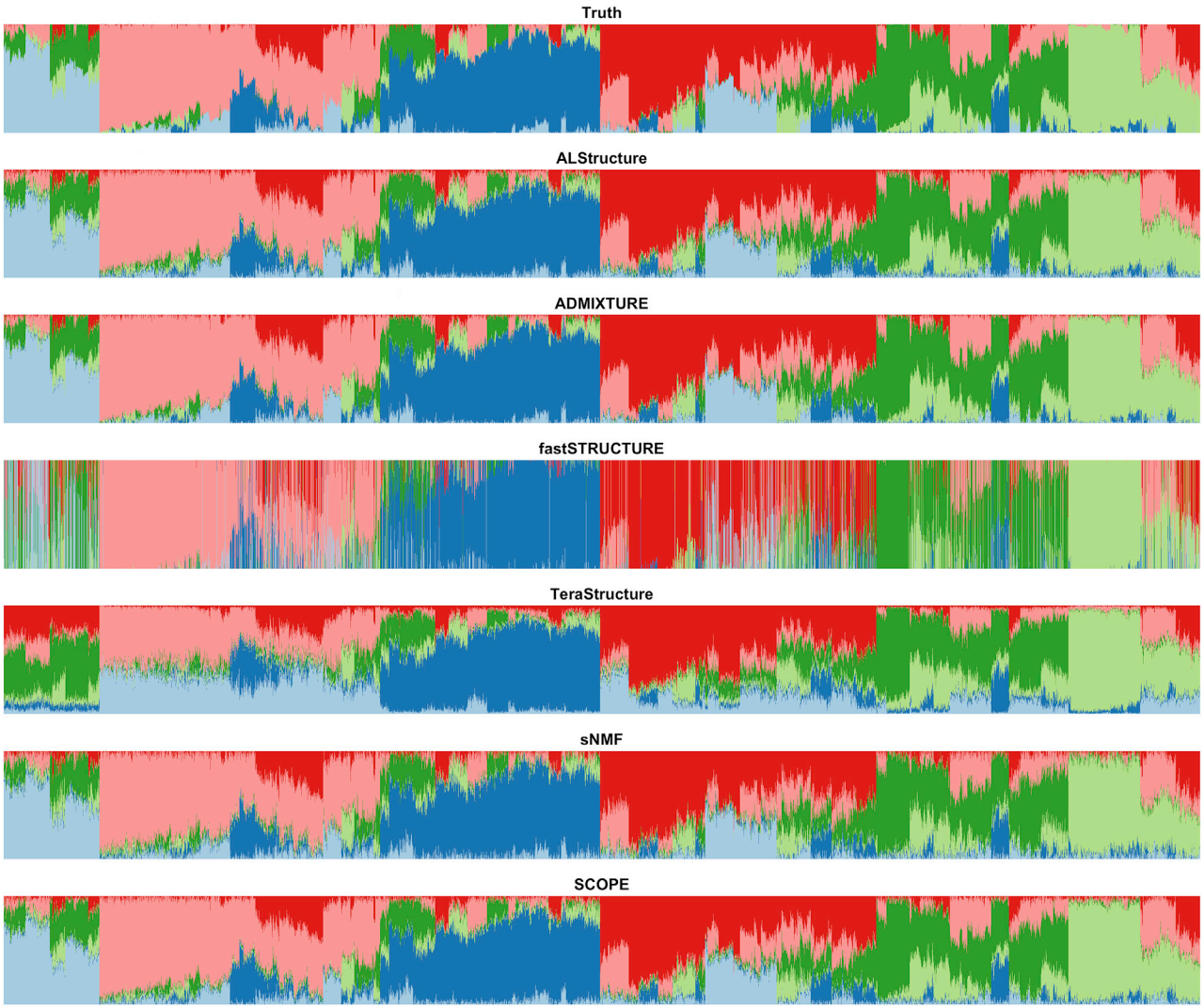

Figure 1. Population structure inference for simulations under PSD model generated with 1000 Genomes Phase 3 data
PSD model parameters were drawn from TGP data to generate a simulation dataset with 10,000 samples and 10,000 SNPs. The true admixture proportions and resulting inferred admixture proportions from each method are shown. Colors and order of samples are matched between each method to the truth.

model and a basic model of spatial structure[19] to study the robustness of SCOPE and other methods in the presence of model violations. We simulated several independent datasets by using parameters calculated from two real datasets: the 1000 Genomes Project (TGP)[15] and the Human Genome Diversity Project (HGDP)[27] (see "benchmarking" sections of subjects and methods). It is important to note that each simulation dataset was created independently of the others and they are not subsets of the largest dataset. Thus, performance should only be compared between methods run on the same dataset.

Under the PSD model, which matches the assumptions of the methods tested, ADMIXTURE is the most accurate followed by SCOPE and ALStructure (Figures 1, S1, S2, S3, and S4). Among the scalable methods, TeraStructure and SCOPE, SCOPE tends to be more accurate in terms of both Jensen-Shannon divergence (JSD) (Table 1) and root-mean-square error (RMSE) (Table 2). We also assessed accuracy under a spatial model, which violates the assumptions of the PSD model by inducing a spatial relationship between the admixture proportions (Figures 2, S5, S6, and S7). Under this scenario, SCOPE, ALStructure, and sNMF are typically the most accurate (Tables 1 and 2).

We also observe similar trends when calculating Kullback-Leibler (KL) divergence (Tables S1 and S2) but opt to use JSD as a primary accuracy measurement because of the asymmetric nature of KL divergence, which changes depending on the order of inputs. We also assessed whether SCOPE can consistently arrive at similar solutions across runs regardless of the stochastic approximations used in SCOPE's algorithm. We ran five replicates of SCOPE from 2–40 inferred populations on a HGDP PSD simulation (Figure S8A), TGP PSD simulation (Figure S8B), HGDP dataset (Figure S8C), and HO dataset (Figure S8D). We observe in our simulated datasets that SCOPE is consistent across both JSD and RMSE between solutions up to the

Table 1. Jensen-Shannon divergence measurements for methods on simulated data

Dataset type	Base dataset	k	n	m	ADMIXTURE	fastStructure	TeraStructure	ALStructure	sNMF	SCOPE
PSD	HGDP	6	10,000	10,000	2.4*	6.3	13.7	3.6	2.4*	3.6
PSD	TGP	6	10,000	10,000	0.8*	11.3	8.8	1.9	2.4	1.9
PSD	TGP	6	10,000	1,000,000	0.03*	8.1	0.2	–	–	0.2
PSD	TGP	6	100,000	1,000,000	–	–	0.3	–	–	0.2*
PSD	TGP	6	1,000,000	1,000,000	–	–	–	–	–	0.2*
Spatial	HGDP	6	10,000	10,000	6.5	33.9	5.7	2.1*	2.3	2.6
Spatial	TGP	6	10,000	10,000	6.8	31.1	3.4	2.4*	4.0	3.3
Spatial	TGP	10	10,000	100,000	12.4	34.7	6.3	8.1	5.7	5.6*
Spatial	TGP	10	10,000	1,000,000	–	–	10.0	–	–	8.2*

Jensen-Shannon divergence (JSD) was computed against the ground truth admixture proportions for each simulation. Values are displayed as percentages rounded to one decimal place. Estimated proportions of 0 were set to 1×10^{-9} (see subjects and methods). A dash denotes that the method was not run because of projected time or memory usage. Values with an asterisk denote the best value for each dataset.

simulated number of populations. Both accuracy measures decrease when inferred more populations than simulated. For the HGDP and HO datasets, we observed that SCOPE is mostly consistent even up to 40 inferred populations. On occasion, we see slight inconsistency, but this is largely because one replicate differed from the other (Figure S9).

Runtime and memory

Using simulated and real datasets, we compared the runtime of SCOPE to ADMIXTURE, fastStructure, TeraStructure, sNMF, and ALStructure (Table 3). Not all of the compared methods could be run on all datasets within practical constraints of time and memory. On the largest PSD datasets that each method could be run on, SCOPE is over 150 times faster than ADMIXTURE (10,000 individuals by 1 million SNPs), over 500 times faster than fastStructure (10,000 individuals by 1 million SNPs), about 100 times faster than ALStructure (10,000 individuals by 100,000 SNPs), over 110 times faster than TeraStructure (100,000 individuals by 1 million SNPs), and as fast as sNMF (10,000 individuals by 10,000 SNPs). SCOPE is also capable of running on a dataset containing one million SNPs and individuals in just over 24 h (≈ 1 day), whereas TeraStructure is extrapolated to require about 500 h (≈ 20 days) on the basis of times reported in its manuscript[7] as well as our experiments (see "benchmarking" sections of subjects and methods).

The runtime of all methods increases under the spatial model. In this scenario, SCOPE is over 1,800 times faster than ADMIXTURE (10,000 individuals by 100,000 SNPs), about 210 times faster than fastStructure (10,000 individuals by 100,000 SNPs), over 155 times faster than ALStructure (10,000 individuals by 100,000 SNPs), about nine times faster than TeraStructure (10,000 individuals by 1 million SNPs), and four times faster than sNMF (10,000 individuals by 100,000 SNPs) on the largest dataset each method could be run on. Over all of the datasets, SCOPE is up to 1,800 times faster than existing methods and three to 144 times faster than TeraStructure. Furthermore, SCOPE scales linearly with the number of latent populations inferred (Figure S10). Additional threads can also be used by SCOPE to speed up

Table 2. Root-mean-square error measurements for methods on simulated data

Dataset type	Base dataset	k	n	m	ADMIXTURE	fastStructure	TeraStructure	ALStructure	sNMF	SCOPE
PSD	HGDP	6	10,000	10,000	4.0*	10.3	16.6	5.6	4.1	5.6
PSD	TGP	6	10,000	10,000	1.8*	15.9	13.7	3.2	4.1	3.2
PSD	TGP	6	10,000	1,000,000	0.2*	12.4	0.9	–	–	0.3
PSD	TGP	6	100,000	1,000,000	–	–	1.0	–	–	0.4*
PSD	TGP	6	1,000,000	1,000,000	–	–	–	–	–	0.5*
Spatial	HGDP	6	10,000	10,000	11.9	31.1	10.2	5.7*	5.7*	6.5
Spatial	TGP	6	10,000	10,000	12.5	29.1	6.8*	7.5	9.4	7.3
Spatial	TGP	10	10,000	100,000	10.8	22.8	8.8	8.5	6.7*	6.7*
Spatial	TGP	10	10,000	1,000,000	–	–	6.6*	–	–	7.2

Root-mean-square error (RMSE) was computed against the ground truth admixture proportions for each simulation. RMSE is displayed in percentage and rounded to the first decimal place. A dash denotes that the method was not run due to projected time or memory usage. Values with an asterisk denote the best value for each dataset.

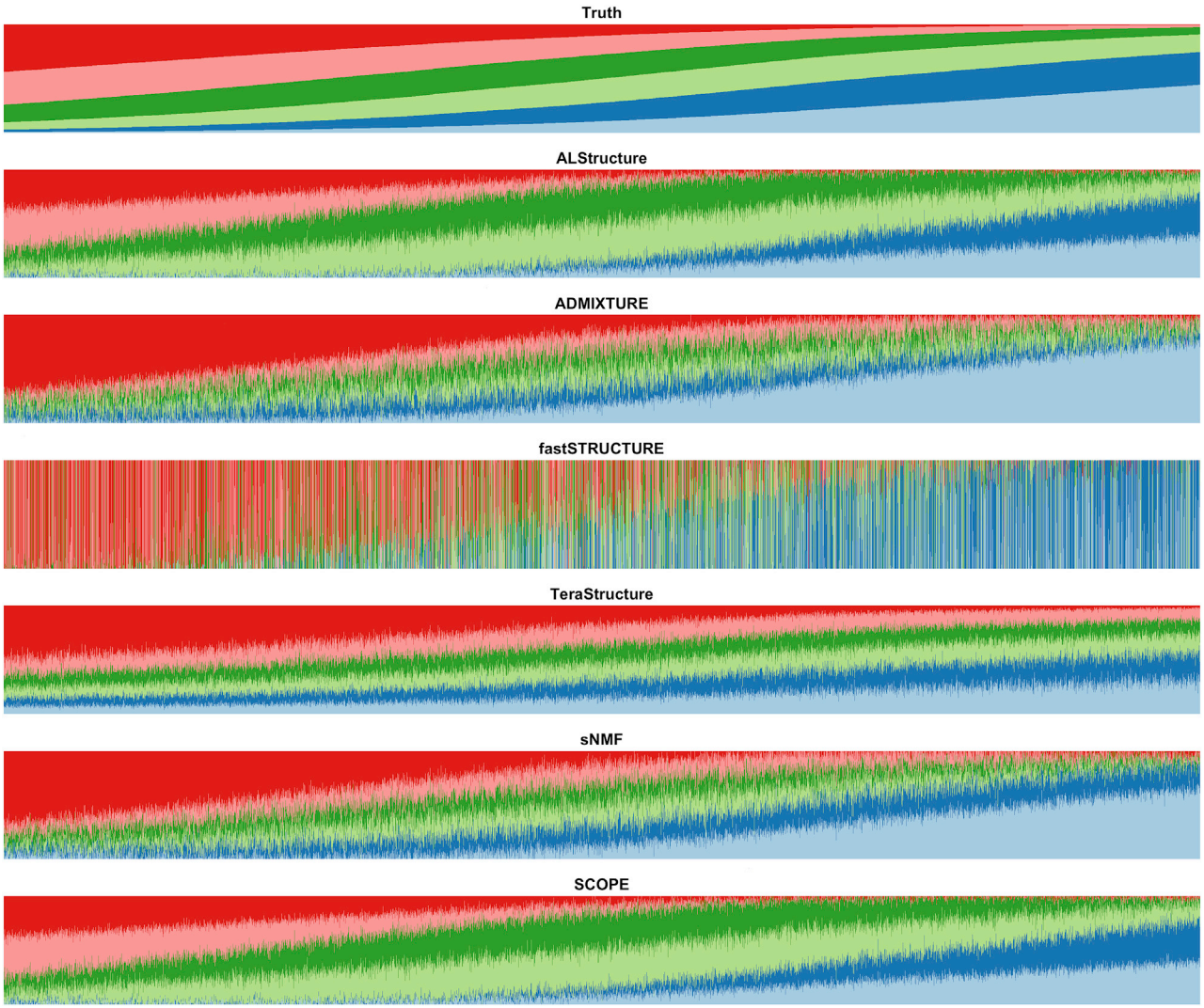

Figure 2. Population structure inference for simulations under a spatial model generated with 1000 Genomes Phase 3 data
Model parameters were drawn from TGP data to generate a simulation dataset with 10,000 samples and 10,000 SNPs under a spatial model (see subjects and methods). The true admixture proportions and resulting inferred admixture proportions from each method are shown. Colors and order of samples are matched between each method to the truth.

runtime up until a fundamental I/O bound is reached (Figure S11).

SCOPE has a reasonable memory footprint: for large datasets for which only TeraStructure and SCOPE were feasible, SCOPE uses slightly less memory than TeraStructure. The memory usage of SCOPE also scales linearly in the size of genotype matrix (i.e., the number of individuals times the number of SNPs) (Table S3). SCOPE requires less than 250 GB for the UK Biobank dataset (488,363 individuals and 569,346 SNPs) and 750 GB for the dataset consisting of one million individuals and SNPs. When using smaller SNP sets such as the UK Biobank's PCA set (147,604 SNPs), SCOPE uses about 60 GB of memory (488,363 individuals and 147,604 SNPs).

Accuracy of supervised analysis

Out of the methods tested, only SCOPE and ADMIXTURE are able to use supplied allele frequencies to perform population structure inference in a supervised fashion (Tables 4 and S4). In the PSD model simulations, we observe a small improvement to both RMSE and JSD relative to unsupervised population structure inference (Figures S12, S13, S14, S15, and S16). Under the spatial model simulations, the use of supervision obtains much greater accuracy compared to unsupervised inference (Figures 3, S17, S18, and S19).

Application to real genotype data

We applied SCOPE to several real, genomic datasets: TGP (1,718 individuals and 1,184,622 SNPs) with eight latent populations ($k = 8$) (Figure S20), HGDP (940 individuals and 642,951 SNPs) with ten populations ($k = 10$) (Figure S21), Human Origins (HO) (1,931 individuals and 385,089 SNPs)[20] with 14 populations ($k = 14$) (Figure S22), the UK Biobank (488,363 individuals and 569,346 SNPs) with four populations ($k = 4$) (Figure 4), and the UK Biobank

Table 3. Runtimes and fold-speedups of methods on simulations and real datasets

Dataset type	Base dataset	k	n	m	ADMIXTURE	fastStructure	TeraStructure	ALStructure	sNMF	SCOPE
PSD	HGDP	6	10,000	10,000	0:14 (48)	3:44 (746)	0:11 (36)	0:30 (101)	< 1 min* (1)	< 1 min*
PSD	TGP	6	10,000	10,000	0:17 (206)	1:22 (987)	0:12 (144)	0:23 (271)	< 1 min* (1)	< 1 min*
PSD	TGP	6	10,000	1,000,000	35:12 (156)	114:51 (509)	20:31 (91)	–	–	0:14*
PSD	TGP	6	100,000	1,000,000	–	–	237:02 (113)	–	–	2:06*
PSD	TGP	6	1,000,000	1,000,000	–	–	–	–	–	24:37*
Spatial	HGDP	6	10,000	10,000	5:52 (440)	4:06 (308)	0:03 (3)	1:39 (124)	< 1 min* (1)	< 1 min*
Spatial	TGP	6	10,000	10,000	3:11 (239)	3:19 (249)	0:07 (9)	1:55 (144)	~ 1 min* (1)	< 1 min*
Spatial	TGP	10	10,000	100,000	284:47 (1,808)	33:03 (210)	4:29 (28)	24:51 (158)	0:33 (4)	0:09*
Spatial	TGP	10	10,000	1,000,000	–	–	15:22 (9)	–	–	1:47*
Real	HGDP	10	940	642,951	4:24 (31)	4:39 (33)	0:40 (5)	0:55 (7)	0:16 (2)	0:08*
Real	HO	14	1,931	385,089	13:28 (122)	24:49 (224)	1:37 (15)	2:11 (20)	0:30 (4)	0:07*
Real	TGP	8	1,718	1,854,622	31:33 (33)	8:53 (9)	4:20 (5)	11:16 (12)	–	0:57*
Real	UKB	4	488,363	569,346	–	–	–	–	–	25:57*
Real	UKB	20	488,363	147,604	–	–	–	–	–	23:42*
Real	UKB	40	488,363	147,604	–	–	–	–	–	51:25*

ADMIXTURE, TeraStructure, sNMF, and SCOPE were run with eight threads. ALStructure and fastStructure were run on a single thread because of their lack of multithreading implementations. Default parameters were used. TeraStructure's "–rfreq" parameter was set to 10% of the number of SNPs. Times are rounded to the nearest minute and displayed in h:min. The fold-speedup (runtime of method in seconds divided by runtime of SCOPE in seconds) achieved by SCOPE is denoted with each time in parentheses and rounded to the nearest integer. Values with an asterisk denote the best value for each dataset. Runtimes for SCOPE under one minute are denoted as "< 1 min." A dash denotes that the method was not run because of projected time or memory usage.

(488,363 individuals and 147,604 SNPs) with 20 populations ($k = 20$) (Figure S24) and 40 populations ($k = 40$) (Figure S25) (see subjects and methods for quality control). We chose the number of latent populations to be consistent with previous studies on these datasets.[7,8] For the UK Biobank analysis, we chose four latent populations to infer continental ancestry groups for the larger SNP set and 20 and 40 latent populations to explore SCOPE's ability to infer larger numbers of latent populations on real data. In terms of runtime and memory, we continued to observe trends consistent with our simulations where SCOPE is orders of magnitude faster than other methods while consuming reasonable amounts of memory (Tables 3 and S3). We note that the runtime for inference on the larger UK Biobank dataset is about the same as the runtime for our 1 million individual and SNP simulation despite the fact that the UK Biobank dataset is approximately a quarter of its size, consistent with the increase in runtimes with model deviations as seen in the context of spatial simulations.

Table 4. Accuracy of supervised population structure inference with supplied allele frequencies on simulations

					Supervised		Unsupervised	
Dataset type	Base dataset	k	n	m	RMSE	JSD	RMSE	JSD
PSD	HGDP	6	10,000	10,000	2.9*	1.5*	5.6	3.6
PSD	TGP	6	10,000	10,000	2.0*	0.9*	3.2	1.9
PSD	TGP	6	10,000	1,000,000	0.2*	0.1*	0.3	0.2
PSD	TGP	6	100,000	1,000,000	0.2*	0.1*	0.4	0.2
PSD	TGP	6	1,000,000	1,000,000	0.2*	0.1*	0.5	0.2
Spatial	HGDP	6	10,000	10,000	2.4*	0.6*	6.5	2.6
Spatial	TGP	6	10,000	10,000	1.7*	0.3*	7.3	3.3
Spatial	TGP	10	10,000	100,000	0.6*	0.3*	6.7	5.6
Spatial	TGP	10	10,000	1,000,000	0.3*	0.1*	8.2	7.2

True allele frequencies were supplied to SCOPE to use in supervised population structure inference. Root-mean-square error (RMSE) and Jensen-Shannon divergence (JSD) were computed against the true admixture proportions. Estimated proportions of 0 were set to 1×10^{-9} for JSD calculations (see subjects and methods). Values are displayed in percentages and rounded to the first decimal place. Values with an asterisk denote the best value for each dataset.

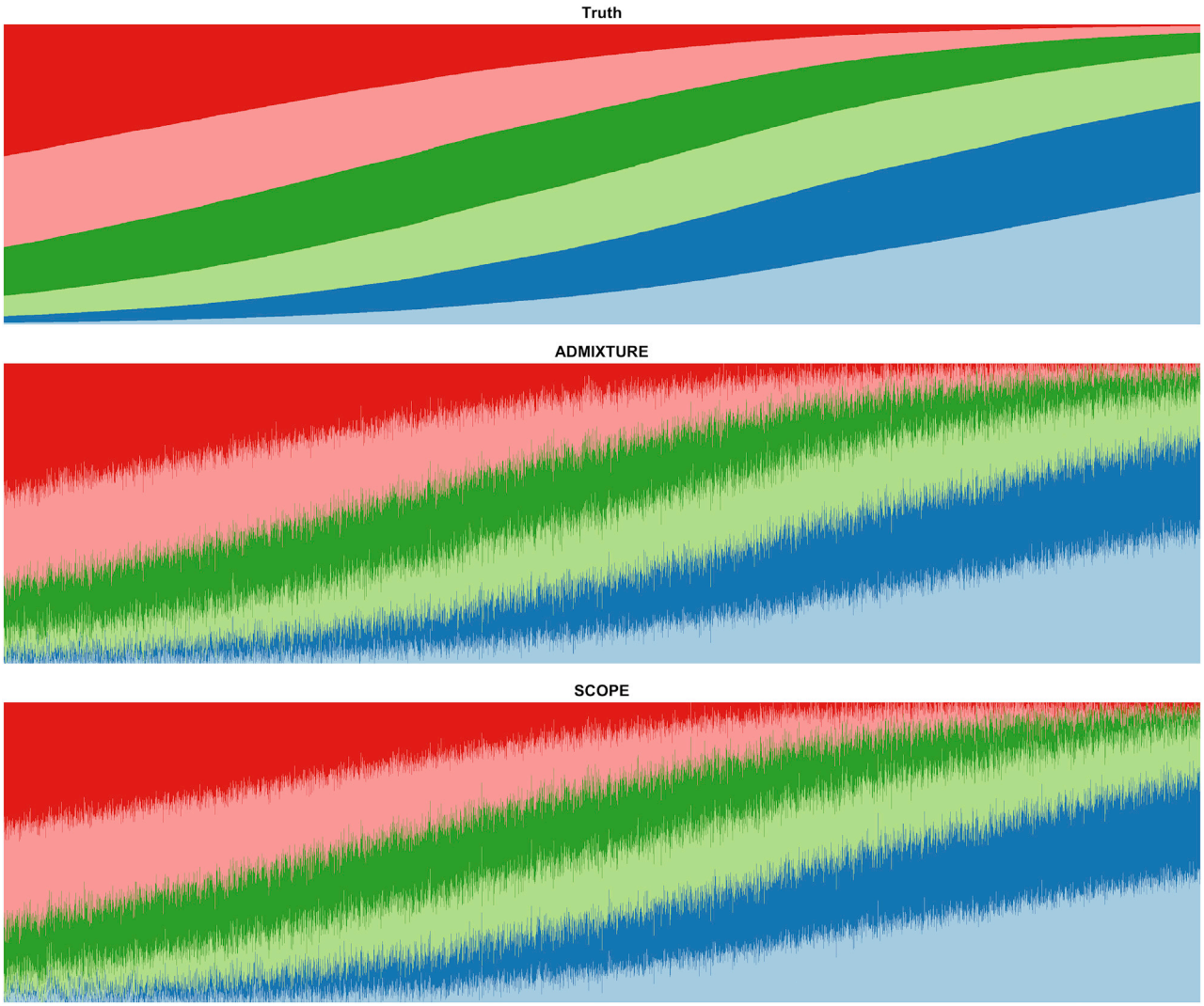

Figure 3. Supervised population structure inference for simulations under a spatial model generated with 1000 Genomes Phase 3 data

Model parameters were drawn from TGP data to generate a simulation dataset with 10,000 samples and 10,000 SNPs under a spatial model. Both methods were provided the true population allele frequencies as input. Colors and order of samples are matched between each method to the truth.

Because there is no ground truth to assess accuracy on these datasets, we used concordance between SIRE and inferred admixture proportions as a metric. We trained multinomial logistic regression models to predict continental ancestry for the TGP (five populations) and HGDP (seven populations) by using the inferred admixture proportions from each method (Table S5). We find that all methods perform similarly on both datasets. For the UK Biobank, SCOPE is able to obtain 88.27% accuracy when using labels provided by UK Biobank (22 labels) and 95.75% accuracy when ambiguous/heterogeneous labels (e.g., "other," "mixed") are removed and population labels are collapsed to continental groupings (eight labels). We did not perform this analysis for the HO dataset because several population labels only contained one sample.

We additionally assessed SCOPE's ability to infer finer population structure by using the British individuals in the UK Biobank. We trained ordinary least-squares models to predict the self-reported birth location GPS coordinate by using the inferred proportions from the different runs of SCOPE under different numbers of latent populations (four, 20, and 40 latent populations) (Table S6). Increasing the number of latent populations generally improves the prediction accuracy when measured through coefficient of determination (R^2). With four latent populations, the R^2 is 0.007 and 0.008 for latitude and longitude prediction, respectively. This increases to 0.2–0.3 and approximately 0.15 when increasing the number of latent populations to 20 and 40. We also examined the prediction accuracy in terms of residual distance (difference between predicted and reported location). The 95% quantile for the residual distances decreases from $\approx$ 334 km to $\approx$ 290 km when increasing the number of inferred populations from four to 20 or 40.

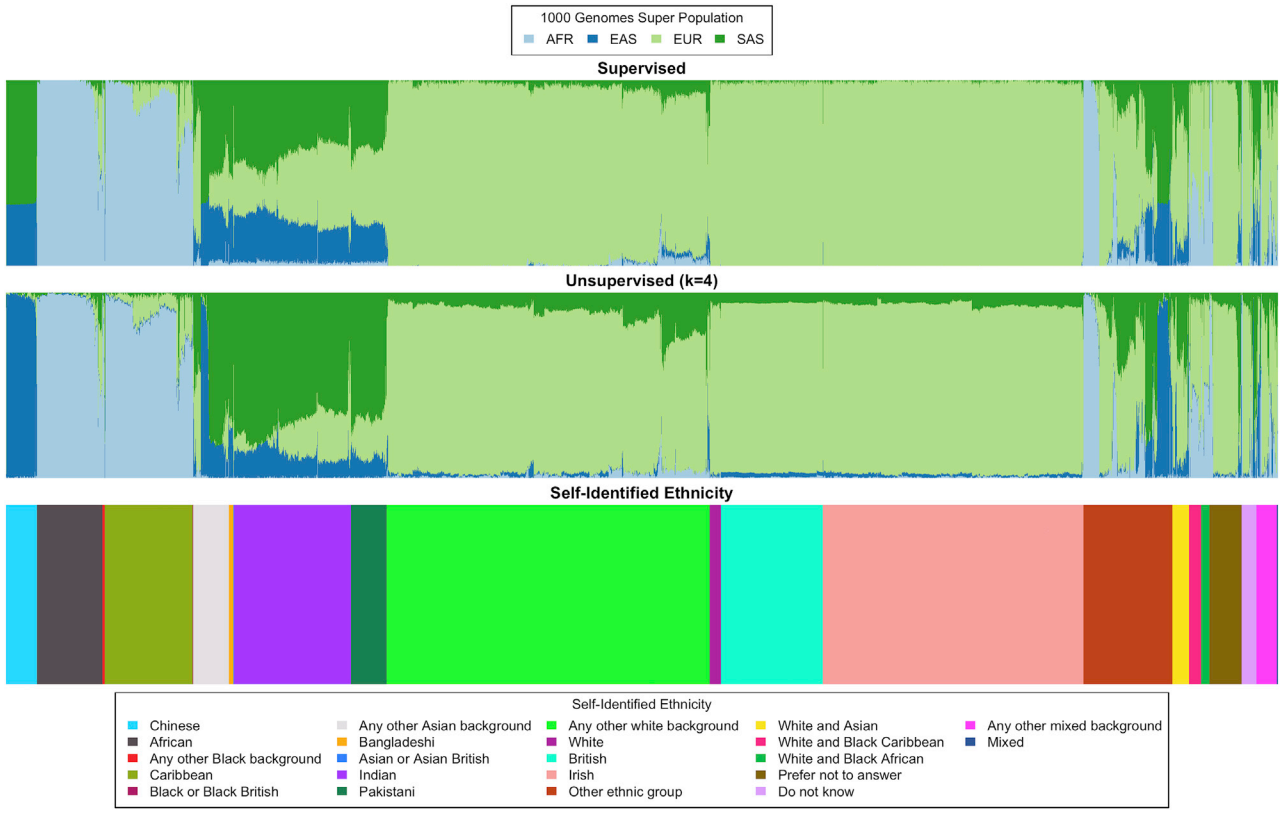

Figure 4. Continental ancestry inference on the UK Biobank
We ran population structure inference by using SCOPE on the UK Biobank (488,363 individuals and 569,346 SNPs) both supervised with 1000 Genomes Phase 3 allele frequencies (top) and unsupervised with four latent populations (middle). For reference, we plot the self-identified race/ethnicity (bottom). For visualization purposes, we reduced the number of self-identified British individuals to a random subset of 5,000 individuals. Colors and order of samples are matched between each row of the figure. The full figure without individuals removed can be found in Figure S23.

We also utilized the supervised mode of SCOPE by using known population allele frequencies from TGP superpopulations to infer continental ancestry for all individuals in the UK Biobank. We find that the supervised mode of SCOPE largely agreed with the unsupervised inference (Figures 4 and S23).

Discussion

We have presented SCOPE, a scalable method for inferring population structure from biobank-scale genomic data. We show that SCOPE remains accurate while being scalable in terms of runtime and memory requirements. SCOPE is also able to perform supervised analyses that leverage allele frequency estimates from previous studies to improve interpretability, runtime, and accuracy.

SCOPE enables new analyses by improving the scalability of admixture proportion inference. The inclusion of more individuals and/or genomic sites allows more rare latent population structure to be discovered in addition to improving estimation of the true latent population frequencies. These are often the cases where scaling to biobank-level data becomes a necessity. Furthermore, many admixture tools are often used as an exploratory analysis being run with different numbers of latent populations (i.e., k). Being able to perform several runs quickly becomes important for initial analysis.

The use of SCOPE is not without limitations for real data analysis and interpretation. For instance, although larger non-trivial numbers of latent populations (k) such as 20 (Figure S24) and 40 (Figure S25) from the UK Biobank explored in this study increase our ability to dissect finer-scale population structure, they remain very difficult to interpret. Furthermore, when exploring these settings, care must be taken to curate a well-defined SNP set. For example, we see a decrease in prediction accuracy when moving from 20 to 40 latent populations in the UK Biobank. This may be attributed to the fact that the UK Biobank's PCA SNP set was curated to differentiate continental population structure rather than intracontinental structure. We also observed that SCOPE is consistent when inferring a large number of latent populations as exemplified by our replicate studies on the HGDP (Figure S8C) and HO (Figure S8D) datasets, which suggests there is more fine-scale population structure being detected and opens the question of what these latent populations may correspond to. While the ability to use

supervised analysis as we did for the UK Biobank can greatly improve interpretability, supervision with SCOPE largely depends on the accuracy of the reference dataset and frequencies used. Finally, there is still the open question of choosing the appropriate number of latent populations (k). Although SCOPE allows one to run several different values for k, we do not provide any criteria to choose a specific value of k. We defer deeper analysis of these questions for future studies.

The methodology used in SCOPE can also be extended in several ways. Several methods that perform structure inference on other genomic datasets[28,29] utilize semi-supervised approaches where there are both known and unknown populations. A possible approach for semi-supervision with SCOPE is to perform a multi-stage inference procedure where supervised inference is first applied and unsupervised inference is applied on the residual or unexplained structure. Most current methods, including SCOPE, ignore additional information within the data, such as correlation patterns (i.e., linkage disequilibrium [LD]). Some methods such as fineSTRUCTURE[30] can perform LD-aware population structure inference but are challenging to scale. The development of methods that can model LD while retaining scalability is a key step in advancing population structure inference.

Though not directly related to the admixture model, there are several approaches to finding broader forms of structure that are not explicitly in the form of admixture proportions. For instance, possible usage of non-linear dimensionality reduction techniques such as UMAP[31] could provide promising ways to extend beyond current methods, which solely utilize linear methods such as PCA. Other approaches to detecting fine-scale structure include using identity-by-descent (IBD)[32] or tree-based methods.[33] Finding ways to scalably bridge these different approaches with the admixture model is still an open question. Finally, extensions of the techniques used in SCOPE can be used to infer relevant structure in other domains such as metagenomics and single-cell transcriptomics.

Data and code availability

SCOPE can be found at https://github.com/sriramlab/SCOPE. Scripts for simulations, visualization, assessment, downloading of publicly available data, and real data filtering and additional code used in this study can be found at the repository as well. UK Biobank dataset is the only dataset used in this study that is not publicly available but can be obtained by application (https://www.ukbiobank.ac.uk/).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.02.015.

Acknowledgments

We would like to thank Bogdan Pasaniuc and members of the Pasaniuc and Sankararaman labs for advice and comments on this project. This research was conducted with the UK Biobank Resource under application 33127. This work was funded by NIH grants T32HG002536 (A.M.C.) and R35GM125055 (E.K.M., S.S.), an Alfred P. Sloan Research Fellowship (S.S.), and NSF grants DGE-1829071 (A.M.C.) and III-1705121 (A.T. and S.S.).

Declaration of interests

The authors declare no competing interests.

Received: November 8, 2021
Accepted: February 21, 2022
Published: March 16, 2022

Web resources

C++ Spectra library, https://spectralib.org/
lpSolve, https://CRAN.R-project.org/package=lpSolve

References

1. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.
2. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
3. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
4. Cheng, Jade Yu, Mailund, Thomas, and Nielsen, Rasmus (2017). Fast admixture analysis and population tree estimation for SNP and NGS data. Bioinformatics *33*, 2148–2155.
5. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
6. Raj, A., Stephens, M., and Pritchard, J.K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics *197*, 573–589.
7. Gopalan, P., Hao, W., Blei, D.M., and Storey, J.D. (2016). Scaling probabilistic models of genetic variation to millions of humans. Nat. Genet. *48*, 1587–1590.
8. Cabreros, I., and Storey, J.D. (2019). A likelihood-free estimator of population structure bridging admixture models and principal components analysis. Genetics *212*, 1009–1029.
9. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
10. Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol. *70*, 214–223.
11. Chen, X., and Storey, J.D. (2015). Consistent estimation of low dimensional latent structure in high-dimensional data. Preprint at arXiv, 1510.03497v1.
12. Halko, N., Martinsson, P.-G., and Joel, A. (2011). Tropp. Finding structure with randomness: Probabilistic algorithms

for constructing approximate matrix decompositions. SIAM Rev. *53*, 217–288.
13. Liberty, E., and Zucker, S.W. (2009). The mailman algorithm: A note on matrix–vector multiplication. Inf. Process. Lett. *109*, 179–182.
14. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
15. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A.; and 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
16. Abraham, G., Qiu, Y., and Inouye, M. (2017). FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics *33*, 2776–2778.
17. Behr, A.A., Liu, K.Z., Liu-Fang, G., Nakka, P., and Ramachandran, S. (2016). Pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics *32*, 2817–2823.
18. Balding, D.J., and Nichols, R.A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica *96*, 3–12.
19. Ochoa, A., and Storey, J.D. (2021). Estimating fst and kinship for arbitrary population structures. PLoS Genet. *17*, e1009241.
20. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413.
21. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.
22. Cavalli-Sforza, L.L. (2005). The Human Genome Diversity Project: past, present and future. Nat. Rev. Genet. *6*, 333–340.
23. Rosenberg, N.A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann. Hum. Genet. *70*, 841–847.
24. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
25. Gagolewski, M., Bartoszuk, M., and Cena, A. (2016). Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm. Inf. Sci. *363*, 8–23.
26. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. Genetics *196*, 973–983.
27. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science *319*, 1100–1104.
28. Shenhav, L., Thompson, M., Joseph, T.A., Briscoe, L., Furman, O., Bogumil, D., Mizrahi, I., Pe'er, I., and Halperin, E. (2019). FEAST: fast expectation-maximization for microbial source tracking. Nat. Methods *16*, 627–632.
29. Caggiano, C., Celona, B., Garton, F., Mefford, J., Black, B., Lomen-Hoerth, C., Dahl, A., and Zaitlen, N. (2020). Estimating the rate of cell type degeneration from epigenetic sequencing of cell-free dna. Preprint at bioRxiv. https://doi.org/10.1101/2020.01.15.907022.
30. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLOS Genet. *8*, 1–16.
31. Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., and Gravel, S. (2019). Umap reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLOS Genet. *15*, 1–24.
32. Nait Saada, J., Kalantzis, G., Shyr, D., Cooper, F., Robinson, M., Gusev, A., and Palamara, P.F. (2020). Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. Nat. Commun. *11*, 6130.
33. Kelleher, J., Wong, Y., Wohns, A.W., Fadil, C., Albers, P.K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. Nat. Genet. *51*, 1330–1338.