Peer Community Journal

Section: Evolutionary Biology

Research article

Published
2024-12-12

Cite as
Cindy G. Santander, Alba Refoyo Martinez and Jonas Meisner (2024) *Faster model-based estimation of ancestry proportions*, Peer Community Journal, 4: e115.

Correspondence
jonas.meisner@sund.ku.dk

Peer-review
Peer reviewed and recommended by PCI Evolutionary Biology,
https://doi.org/10.24072/pci.evolbiol.100838

Faster model-based estimation of ancestry proportions

Cindy G. Santander[1], Alba Refoyo Martinez[2], and Jonas Meisner[3,4]

Volume **4** (2024), article e115

https://doi.org/10.24072/pcjournal.503

Abstract

Ancestry estimation from genotype data in unrelated individuals has become an essential tool in population and medical genetics to understand demographic population histories and to model or correct for population structure. The *ADMIXTURE* software is a widely used model-based approach to account for population stratification, however, it struggles with convergence issues and does not scale to modern human datasets or the large number of variants in whole-genome sequencing data. Likelihood-free approaches optimize a least square objective and have gained popularity in recent years due to their scalability. However, this comes at the cost of accuracy in the ancestry estimates in more complex admixture scenarios. We present a new model-based approach, *fastmixture*, which adopts aspects from likelihood-free approaches for parameter initialization, followed by a mini-batch expectation-maximization procedure to model the standard likelihood. In a simulation study, we demonstrate that the model-based approaches of *fastmixture* and *ADMIXTURE* are significantly more accurate than recent and likelihood-free approaches. We further show that *fastmixture* runs approximately 30× faster than *ADMIXTURE* on both simulated and empirical data from the 1000 Genomes Project such that our model-based approach scales to much larger sample sizes than previously possible.

[1] Department of Biology, University of Copenhagen, Denmark, [2] Center for Health Data Science, University of Copenhagen, Denmark, [3] Mental Health Centre Copenhagen, Copenhagen University Hospital, Denmark, [4] Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Denmark

e-ISSN 2804-3871

Introduction

For the past two decades, unsupervised ancestry estimation has been a crucial element in studies on human evolutionary genetics and genome-wide association studies (Marchini et al., 2004; Novembre et al., 2008). Estimating global ancestry proportions in unrelated individuals and corresponding ancestral allele frequencies has been a classic way of correcting for population structure and understanding demographic processes that have shaped the evolutionary history of modern populations. Ancestry estimation from genotype data emerged with the model-based clustering approach proposed with STRUCTURE (Pritchard et al., 2000), where individuals are proportionally assigned to an assumed number of ancestral populations. They modeled the probability of the observed genotype data given ancestry proportions and ancestral allele frequencies using a Bayesian approach. Due to scalability issues, the Bayesian approach was later replaced by maximum likelihood models, which were optimized using expectation-maximization (EM) and block relaxation algorithms, which include the widely used software ADMIXTURE (Alexander et al., 2009; Tang et al., 2005).

In the era of big data, where ever-growing cohorts contain thousands of individuals with genotype data for millions of genetic variants, classic state-of-the-art model-based approaches, such as ADMIXTURE and STRUCTURE, fail to scale due to computational intractability. Over the past decade, there have been multiple attempts to scale unsupervised ancestry estimation. These efforts have primarily been rooted in likelihood-free approaches, with a few exceptions that have attempted to scale the standard model-based approach. Model-based and likelihood-free approaches have been shown to be connected within the framework of matrix factorization, where both approximate an observed genotype matrix by lower rank matrices under different assumptions and constraints (Engelhardt and Stephens, 2010). In likelihood-free approaches, it is common to optimize a different least square objective using either non-negative matrix factorization (NMF) (Frichot et al., 2014; Meisner and Albrechtsen, 2018) or alternating least square (ALS) (Cabreros and Storey, 2019; Chiu et al., 2022). The SCOPE software (Chiu et al., 2022) has gained increased popularity due to its efficient implementation of an ALS approach, which is well-suited for biobank-scale datasets. Meanwhile for the model-based approaches, a stochastic variational inference algorithm (Gopalan et al., 2016) and a neural network autoencoder (Dominguez Mantes et al., 2023) have also been proposed in recent times.

As large-scale cohorts begin to expand and include more cosmopolitan representations of individuals around the globe, obtaining accurate ancestry estimates has become a crucial task of modern genomics. This is essential to properly correct for population structure in genome-wide association studies and mitigate the bias in the over-representation of European-descent samples in current cohorts due to the ever-increasing focus on genomics for population health applications (Martin et al., 2019; Wang et al., 2022). While convergence within a reasonable time frame is highly desirable for these extensive datasets, recent attempts to improve scalability most often compromise accuracy in the ancestry estimation, as demonstrated here. The field still lacks a method that scales to larger datasets while being as accurate as the classic state-of-the-art approaches.

We introduce `fastmixture`, a novel model-based method for estimating ancestry proportions and ancestral allele frequencies in unrelated individuals. We leverage randomized singular value decomposition (SVD) for initializing the ancestry proportions and ancestral allele frequencies, followed by a mini-batch accelerated scheme to speed up the convergence of the EM algorithm. We demonstrate in an extensive simulation study and on real data that `fastmixture` significantly outperforms the original `ADMIXTURE` software in terms of speed while maintaining higher accuracy compared to recently developed approaches for ancestry estimation.

Material and methods

We define a diallelic genotype matrix of N individuals and M variants or single-nucleotide polymorphisms (SNPs) as $\mathbf{G} \in \{0,1,2\}^{N\times M}$, which corresponds to the minor allele counts. We describe the estimation of ancestry proportions and ancestral allele frequencies with K ancestral sources as a low-rank matrix factorization problem such that $\mathbf{G} \approx 2\mathbf{QP}^T$, with $\mathbf{Q} \in [0,1]^{N\times K}$ and the constraint $\sum_{k=1}^{K} q_{ik} = 1$, for $i = 1,\dots,N$, and $\mathbf{P} \in [0,1]^{M\times K}$. The matrix factorization problem can also be interpreted as the estimation of individual allele frequencies from the genotype matrix, $\mathbf{H} = \mathbf{QP}^T$. Here, $h_{ij} = \sum_k^K q_{ik}p_{jk}$ is the individual allele frequency of individual i at variant j assuming K ancestral sources. The individual allele frequency will correspond to the underlying parameter in a binomial sampling process of a genotype conditioned on population structure.

Likelihood model

We estimate ancestry proportions and ancestral allele frequencies by maximizing the likelihood model introduced in `ADMIXTURE` (Alexander et al., 2009). The log-likelihood model is defined as follows, assuming independence for individuals and variants:

$$(1) \qquad L(\mathbf{Q},\mathbf{P}) = \sum_i^N \sum_j^M g_{ij}\log(h_{ij}) + (2-g_{ij})\log(1-h_{ij})\,.$$

We utilize the expectation-maximization (EM) algorithm of `frappe` (Tang et al., 2005) and `ADMIXTURE` (Alexander et al., 2009) to maximize the log-likelihood, where the EM updates at iteration t for entries of $\mathbf{Q}$ and $\mathbf{P}$ are defined as follows, respectively:

$$(2) \qquad q_{ik}^{(t+1)} = \frac{q_{ik}^{(t)}}{2M}\sum_{j=1}^{M}\left(\frac{g_{ij}p_{jk}^{(t)}}{h_{ij}^{(t)}} + \frac{(2-g_{ij})(1-p_{jk}^{(t)})}{1-h_{ij}^{(t)}}\right),$$

$$(3) \qquad p_{jk}^{(t+1)} = \left(p_{jk}^{(t)}\sum_{i=1}^{N}\frac{g_{ij}q_{ik}^{(t)}}{h_{ij}^{(t)}}\right)\Bigg/\left(\sum_{i=1}^{N}\frac{g_{ij}q_{ik}^{(t)}p_{jk}^{(t)}}{h_{ij}^{(t)}} + \frac{(2-g_{ij})q_{ik}^{(t)}(1-p_{jk}^{(t)})}{1-h_{ij}^{(t)}}\right).$$

The EM algorithm is notorious for its slow convergence rate and to expedite this process, we employ a quasi-Newton (QN) acceleration scheme (Zhou et al., 2011). The QN acceleration scheme combines multiple EM updates into a larger jump in parameter space at the expense of an increased computational cost per iteration. More details on the acceleration scheme can be found in the supplementary material and Algorithm S1. The convergence criteria of the EM algorithm is defined by the difference in log-likelihood between two successive iterations:

$$(4) \qquad L(\mathbf{Q}^{(t+1)},\mathbf{P}^{(t+1)}) - L(\mathbf{Q}^{(t)},\mathbf{P}^{(t)}) < \epsilon\,,$$

with ϵ being a user-defined threshold.

SVD initialization

Multiple approaches for inferring population structure are connected under the umbrella of matrix factorization, each incorporating different conditions and constraints (Engelhardt and Stephens, 2010). Therefore, we leverage the efficiency and speed of likelihood-free approaches to provide a better initialization of $\mathbf{Q}$ and $\mathbf{P}$ to aid the convergence rate of the EM algorithm in comparison to a standard random initialization.

We initialize $\mathbf{Q}$ and $\mathbf{P}$ using individual allele frequencies estimated from randomized singular value decomposition (SVD) performed on the genotype matrix, combined with an alternating least squares (ALS) approach. SVD is a widely used dimensionality reduction approach in population genetics, which infers continuous structure by extracting axes of genetic variation. The randomized SVD is defined as $\mathbf{G}_C \approx \mathbf{U}_{[1:D]}\mathbf{S}_{[1:D]}\mathbf{V}_{[1:D]}^T$, which extracts the top D singular values and singular vectors, with $\mathbf{G}_C$ being the centered genotype matrix. The individual allele frequency of individual i at variant j is then approximated using the SVD as $\hat{h}_{ij} = f_j + \frac{1}{2}\sum_{d=1}^{D} u_{id}s_d v_{jd}$, where f_j is the minor allele frequency of variant j. We use $D = K - 1$, as the top $K - 1$ singular vectors will capture the population structure of K distinct populations (Patterson et al., 2006). By initializing $\mathbf{P}$ randomly, we can then factorize the individual allele frequency matrix, $\hat{\mathbf{H}}$, using an ALS approach to iteratively estimate both $\mathbf{Q}$ and $\mathbf{P}$, which minimizes the following least square objective:

$$\min_{\mathbf{Q},\mathbf{P}} \|\hat{\mathbf{H}} - \mathbf{Q}\mathbf{P}^T\|_F^2 \,, \tag{5}$$

where $\|.\|_F$ is the Frobenius norm. The concept is similar to `ALStructure` (Cabreros and Storey, 2019) and `SCOPE` (Chiu et al., 2022), which instead rely on latent subspace estimation for approximating the individual allele frequencies. The convergence criteria of the ALS approach is defined by the root mean square error (RMSE) of $\mathbf{Q}$ matrices between two successive iterations:

$$\sqrt{\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(q_{ik}^{(t+1)} - q_{ik}^{(t)}\right)^2} < \delta \,, \tag{6}$$

with δ being a user-defined threshold. The initialization of $\mathbf{Q}$ and $\mathbf{P}$ is fully described in Algorithm 1.

Algorithm 1 `fastmixture` initialization

```
Given: $\mathbf{G}, \mathbf{f}, D, K$
$\mathbf{G}_C \leftarrow$ center($\mathbf{G}, \mathbf{f}$)                          ▷ Centered genotype matrix
$\mathbf{U}_{[1:D]}, \mathbf{S}_{[1:D]}, \mathbf{V}_{[1:D]} \leftarrow$ randomizedSVD($\mathbf{G}_C$, $D$)        ▷ Randomized SVD
$\hat{\mathbf{H}} = \frac{1}{2}\mathbf{U}_{[1:D]}\mathbf{S}_{[1:D]}\mathbf{V}_{[1:D]}^T + \mathbf{f}$        ▷ Individual allele frequencies based on SVD
$\mathbf{P} \in [0,1]^{M \times K}$ initialized randomly
while not converged do                  ▷ Alternating least squares for solving equation 5
    $\mathbf{Q} \leftarrow \hat{\mathbf{H}}\mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}$
    $\mathbf{P} \leftarrow \hat{\mathbf{H}}^T\mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}$
    if rmse($\mathbf{Q}, \mathbf{Q}_{prev}$) $< \delta$ then              ▷ Convergence criteria for ALS
        break
    end if
    $\mathbf{Q}_{prev} \leftarrow \mathbf{Q}$
end while
Return: $\mathbf{Q}, \mathbf{P}$
```

Mini-batch optimization

To further accelerate the convergence rate of the EM algorithm, we introduce simple mini-batch updates inspired by stochastic gradient descent and stochastic EM algorithms (Ruder, 2016). In each iteration, we randomly split the M variants into B batches, and perform QN accelerated EM updates sequentially in each of the B batches. The strategy of sub-sampling variants was also explored in `TeraStructure` (Gopalan et al., 2016). This results in the entries of $\mathbf{Q}$ being updated B times across the batches, while the entries of $\mathbf{P}$ are still only updated once per cycle. Following the mini-batch updates, full QN accelerated EM updates are applied to stabilize the parameters in every iteration. Given the stochastic nature of our mini-batch updates, we halve the number of batches B every time the log-likelihood estimate (Equation 1) fluctuates between iterations. The algorithm will therefore gradually converge towards a standard QN accelerated algorithm for $B \to 1$. Our proposed mini-batch setting resembles a mini-batch gradient descent approach, where the batch size is increased over time. For a detailed description of the `fastmixture` method, please refer to Algorithm 2.

Algorithm 2 `fastmixture` estimation

```
Given: $\mathbf{G}, \mathbf{f}, K, B$
$\mathbf{Q}, \mathbf{P} \leftarrow$ initialization($\mathbf{G}, \mathbf{f}, K$)            ▷ Algorithm 1: SVD initialization
$L_{prev} \leftarrow L(\mathbf{Q}, \mathbf{P})$                                    ▷ Log-likelihood (Equation 1)
while not converged do
    Randomly split $M$ variants into $B$ batches
    for $b = 1, \dots, B$ do
        $\mathbf{Q}, \mathbf{P}_{[b]} \leftarrow \text{QN}(\mathbf{G}_{[b]}, \mathbf{Q}, \mathbf{P}_{[b]})$       ▷ Mini-batch updates (Algorithm S1)
    end for
    $\mathbf{Q}, \mathbf{P} \leftarrow \text{QN}(\mathbf{G}, \mathbf{Q}, \mathbf{P})$                        ▷ Full updates (Algorithm S1)
    if $L(\mathbf{Q}, \mathbf{P}) < L_{prev}$ then
        $B = B/2$                                               ▷ Halve the number of batches
    end if
    if $B = 1$ and $L(\mathbf{Q}, \mathbf{P}) - L_{prev} < \epsilon$ then            ▷ Convergence criteria for EM
        break
    end if
    $L_{prev} \leftarrow L(\mathbf{Q}, \mathbf{P})$
end while
Return: $\mathbf{Q}, \mathbf{P}$
```

Implementation details

The `fastmixture` software is implemented as a multithreaded command-line tool written in Python and Cython, which takes binary `PLINK` (Chang et al., 2015) files as input. It utilizes the NumPy library (Harris et al., 2020) for efficient array manipulation. We assume that the user has performed standard quality control and preprocessing, (e.g., variant filtering based on a minor allele frequency threshold and to only include unrelated samples). The genotype matrix is stored in an 8-bit integer format. For Algorithm 1, we read the centered genotypes in chunks to reduce the memory consumption and perform randomized SVD as introduced in `PCAone` (Li et al., 2023). This approach minimizes the memory footprint of `fastmixture`, which is primarily dominated by the genotype matrices of $MN + 8CN$ bytes, with C being the chunk size of variants used in the randomized SVD. This allows our method to handle large-scale datasets effectively. The individual allele frequency matrix obtained in the randomized SVD, $\hat{\mathbf{H}}$, is computed implicitly through the singular matrices in the ALS approach to further reduce the memory footprint. Throughout the study, we use a starting point of $B = 32$ mini-batches to speed up the convergence of the EM algorithm, which works well across all the tested scenarios. We set the convergence criteria of the ALS approach and the EM algorithm to $\delta = 1.0 \times 10^{-4}$ and $\epsilon = 0.5$, respectively. Given the computational expense of the log-likelihood estimation step (Equation 1), which requires a full pass over the data, we only evaluate it every fifth iteration along with convergence checks to reduce the number of computations. If the entries of $\mathbf{Q}$ and $\mathbf{P}$ are out of domain during the ALS or accelerated EM updates, we simply map them back to their domain through truncation and projection procedures.

Simulations

We simulate genotypes using the `msprime` (Baumdicker et al., 2022) backwards-in-time coalescent model and infer true ancestral tracts using `tspop` (Tsambos et al., 2023). To evaluate the capabilities of `fastmixture`, we assume different demographic models in four different scenarios, all featuring a single or multiple admixture events where the true ancestry proportions are known. In each scenario, we simulate a genetic segment of 250 Mb using a constant recombination rate of 1.28×10^{-8} (The International HapMap Consortium, 2007) and a mutation rate of 2.36×10^{-8} (Gravel et al., 2011). We visualize the different demographic models in Figure 1A, S1A, S2A, and S3A. A census event precedes the first admixture event in all simulations to track the true ancestry of the inherited segments in the sampled individuals. The lengths of the segments each individual has inherited from each source population are then aggregated to estimate the ground truth ancestry proportions using `tspop`. Scenarios A, B, and C are constructed from simple demographic models, where we assume a constant population size of 10,000 for all simulated populations, while Scenario D extends the out-of-Africa model (Gravel et al., 2011) with an additional admixture event (American-Admixture) (Browning et al., 2018). In Scenario A, B, and D, we sample 1,000 individuals, while in the more complex Scenario C, we sample 1,600 individuals. We perform standard filtering on minor allele frequencies at a threshold of 0.05, resulting in datasets consisting of 689,563 SNPs, 687,107 SNPs, 685,592 SNPs, and 500,114 SNPs for Scenario A, B, C, and D, respectively. An overview of the simulated datasets is provided in the supplementary material.

1000 Genomes Project

We also evaluate our `fastmixture` software in the phase 3 release of the 1000 Genomes Project (1KGP) (The 1000 Genomes Project Consortium, 2015). The dataset consists of genotype data of 2,504 individuals from 26 populations across the world, assigned to five superpopulations: AFR (African ancestry), AMR (American ancestry), EAS (East Asian ancestry), EUR (European ancestry), SAS (South Asian ancestry). We keep diallelic SNPs with a minor allele frequency greater than the standard threshold of 0.05, resulting in a total of 6,864,700 SNPs. Due to computational complexity and runtime considerations for comparative analyses, we also construct a downsampled dataset, which we refer to as “1KGP Down". This dataset is obtained by random downsampling or thinning the number of SNPs from the full dataset by a factor of ten, such that the downsampled dataset consists of 686,470 SNPs.

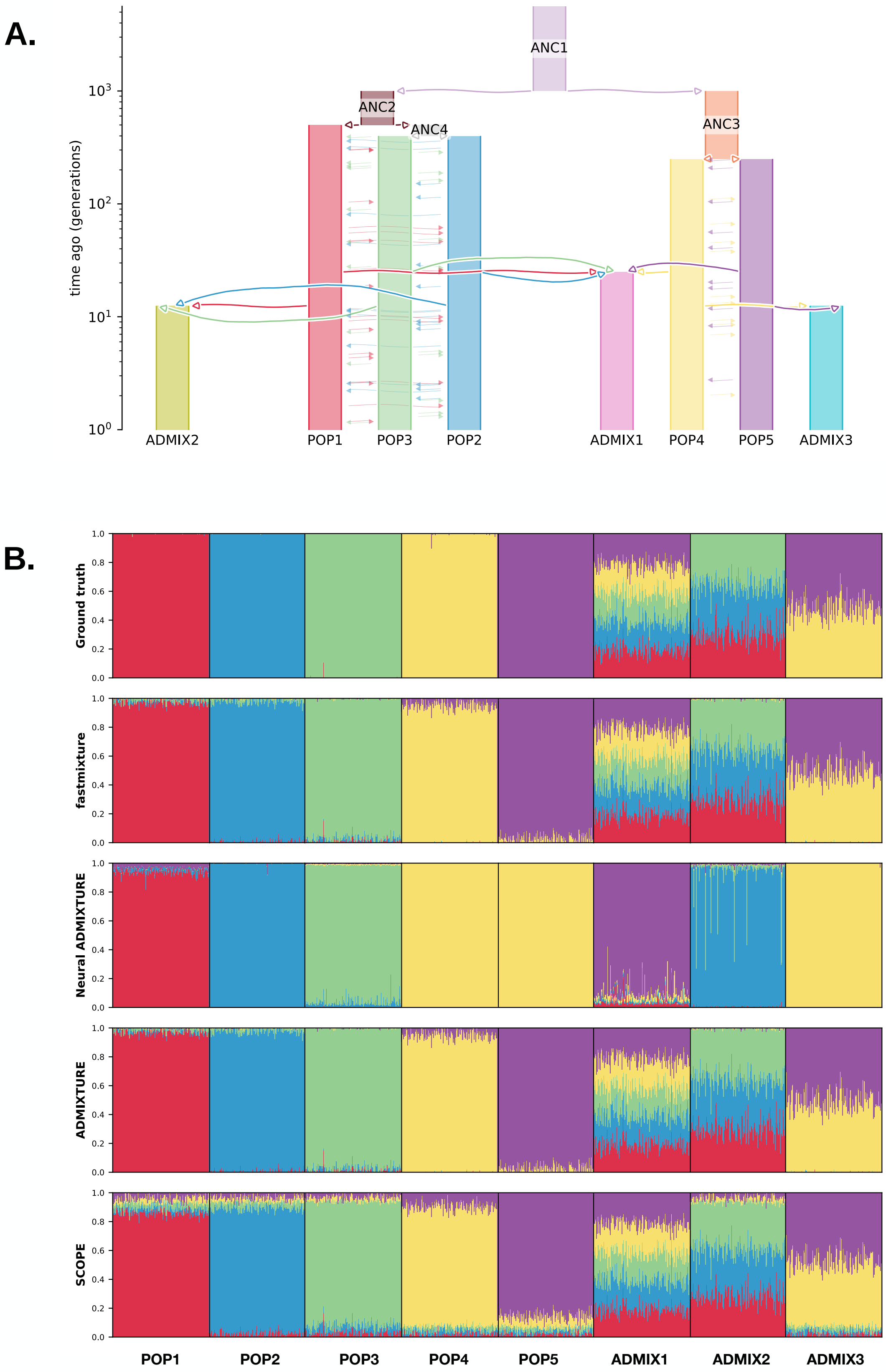

Figure 1 – **A.** Demographic model of *Scenario C* with 200 individuals sampled from each of the eight populations. **B.** Admixture plots of the ancestry proportions in the simulated individuals for $K = 5$ with the ground truth at the top followed by the different software using their run with the highest log-likelihood.

Results

We evaluate and compare the performance of fastmixture (v0.93) against widely used software for estimating ancestry proportions: ADMIXTURE (v1.3.0) (Alexander et al., 2009), Neural ADMIXTURE (v1.4.1) (Dominguez Mantes et al., 2023) and SCOPE (Chiu et al., 2022). All software are executed with their default parameters. ADMIXTURE and Neural ADMIXTURE are model-based approaches, like ours. ADMIXTURE maximizes the likelihood model using a block relaxation method combined with a similar quasi-Newton acceleration scheme to fastmixture, while Neural ADMIXTURE employs a neural network autoencoder approach. On the other hand, SCOPE is a likelihood-free approach that estimates the ancestry proportions and ancestral allele frequencies using an ALS method, differing from the maximum likelihood approaches of the other tools. In our simulation study, we assess these tools based on the root mean square error (RMSE) (Table 1) and the Jenson-Shannon divergence (JSD) (Table S2) between the estimated ancestry proportions and the ground truth ancestry proportions. We also report the log-likelihood estimates (Equation 1) (Table S3) for both the simulated and the empirical datasets of the 1000 Genomes Project. The different measures of assessment in the simulation study are further detailed in the supplementary material.

Additionally, we compare the computational runtimes of fastmixture to the three other software (Figure 3 and Table S1). Notably, ADMIXTURE exhibits significant scalability issues, being approximately 30 times slower than fastmixture across all evaluated datasets. For instance, ADMIXTURE needs more than 40 hours to complete a single run for $K = 5$ on the full 1KGP dataset. In contrast, fastmixture has comparable runtimes to the two other faster approaches in the simulation study. SCOPE is the fastest of all evaluated approaches, showcasing the appealing choice of optimizing the least square objective, a step also used for parameter initialization in fastmixture.

Ancestry estimation in simulation studies

Under the simple demographic model of Scenario A (Figure S1), ADMIXTURE and fastmixture were the two most accurate software followed by SCOPE. When only considering the scalable methods, fastmixture clearly outperformed SCOPE in terms of RMSE (Table 1) and JSD (Table S2), where SCOPE appeared to produce highly noisy ancestry proportions in individuals of the unadmixed populations. On all accounts, Neural ADMIXTURE severely underperformed in estimating the ancestry proportions of the admixed individuals. Notably when admixture was introduced from more than three sources, as in Scenario B (Figure S2), Neural ADMIXTURE failed to detect POP4 as a separate population from POP3 and modeled the ADMIX population incorrectly as a separate unadmixed population. For Scenario B, both ADMIXTURE and fastmixture were again more accurate than SCOPE, while fastmixture was $\sim$34× faster than ADMIXTURE (Figure 3 and Table S1). The noise introduced by SCOPE only increased for $K = 4$ in comparison to $K = 3$ in the simpler Scenario A.

We further evaluated the different software in a more complex simulation scenario, Scenario C, which includes five ancestral sources ($K = 5$) with symmetric migration patterns and three admixed populations (Figure 1). Consistent with results from the simpler scenarios, fastmixture and ADMIXTURE outperformed the two other approaches, with fastmixture being $\sim$28× faster than ADMIXTURE. Due to the increased complexity of the simulation scenario, SCOPE exhibited an even greater increase of noise in its ancestry estimates, while Neural ADMIXTURE again failed

Table 1 – Root mean square error (RMSE) measures for estimated ancestry proportions in the four different simulation scenarios for the evaluated methods against the ground truth. The mean across five different runs is reported with the corresponding standard deviation in parenthesis.

Scenario	`fastmixture`	`Neural ADMIXTURE`	`ADMIXTURE`	`SCOPE`
A	0.0364 (1.0e-4)	0.1289 (0.0200)	0.0367 (3.1e-7)	0.0677 (7.8e-5)
B	0.0278 (4.5e-5)	0.3690 (0.0022)	0.0279 (2.2e-7)	0.0661 (8.2e-5)
C	0.0255 (8.7e-6)	0.3094 (0.0044)	0.0256 (1.9e-7)	0.0658 (3.2e-5)
D	0.0085 (4.3e-7)	0.1010 (0.0372)	0.0085 (4.5e-8)	0.0186 (3.3e-6)

to detect one of the unadmixed population sources and modeled two admixed populations as ancestral sources. Examining the accuracy of the ancestry estimates in each of the eight populations, we observed that `fastmixture` and `ADMIXTURE` performed similarly across the unadmixed and admixed populations, whereas `SCOPE` inferred more accurate ancestry estimates in the admixed populations in comparison to the unadmixed populations (Table S4 and S5).

Using the American-Admixture demographic model in Scenario D (Figure S3), we still consistently observed that `ADMIXTURE` and `fastmixture` perform similarly in terms of accuracy and log-likelihood, outputting results closest to the ground truth (Table 1 and Table S2), with `fastmixture` being $\sim$28× faster than `ADMIXTURE`. While across all scenarios `SCOPE` exhibited the fastest runtimes, it also consistently underperformed in accuracy as measured with RMSE and JSD, however, markedly better than `Neural ADMIXTURE`.

Testing hyperparameters. The number of initial batches in `fastmixture`, used for its mini-batch optimization, is a hyperparameter. We tested the effect of changing the number of mini-batches in the more complex simulation scenario, Scenario C, having multiple admixture events and five source populations. We utilized $B = \{8, 16, 32, 64, 128\}$, including the default choice of $B = 32$, and reported the computational runtimes, log-likelihoods, RMSE, and JSD measures. Our results showed that `fastmixture` was robust to changes in B, as all evaluated choices consistently captured the same solutions with highly comparable assessment measures (Figure S4 and Table S6). Based on these findings, we conclude that $B = 32$ was an optimal choice, balancing both fast runtimes and highly accurate ancestry estimations.

We further evaluated the effectiveness of our SVD initialization by comparing it to random parameter initialization inside the `fastmixture` framework for Scenario C. We reported computational runtimes, log-likelihoods, RMSE, and JSD measures in Table S7, where the two initializations performed similarly but the SVD initialization approximately halves the runtime on average in comparison to having a random initialization. Therefore, our observed runtime gains relative to `ADMIXTURE` could largely be attributed to our proposed mini-batch optimization.

Robustness to model misspecification. For most scenarios in ancestry estimation, the true number of ancestral sources is rarely known. We, therefore, tested and compared all software and their capabilities to deal with model misspecifications related to the number of ancestral sources used for Scenario C, which had a ground truth of $K = 5$. Here we would expect the ancestry estimations to capture older events in the demographic model for $K < 5$. The results comparing the software for $K = \{2, 3, 4\}$ are displayed in Figures S6, S7 and S8, respectively, and their corresponding log-likelihoods are reported in Table S8. We note that `ADMIXTURE` only found the optimal solution in four out of five runs, thus showcasing its vulnerabilities due to random parameter initialization and a standard optimization approach. Due to the increased complexity

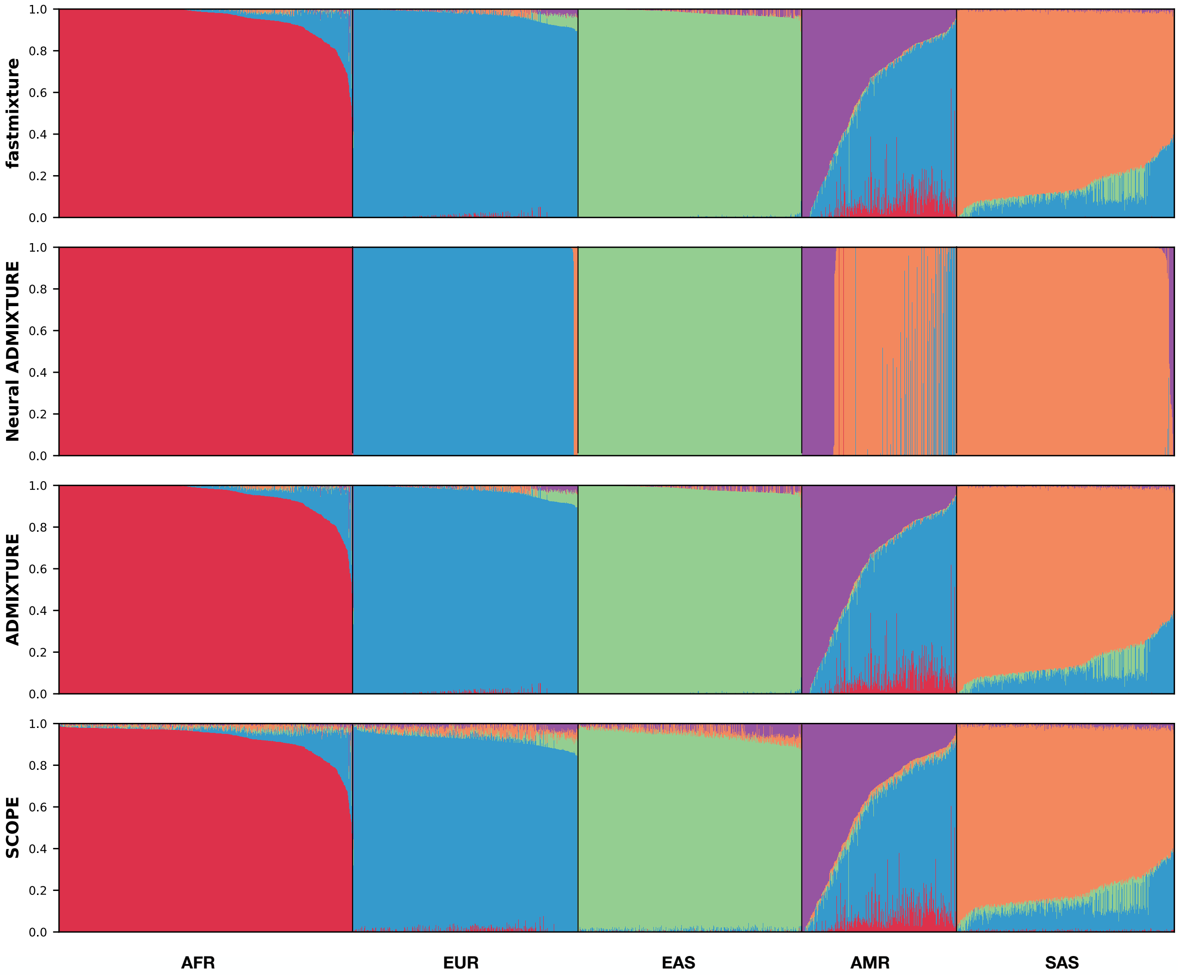

Figure 2 – Admixture plots of the estimated ancestry proportions for $K = 5$ in the full version of the 1000 Genomes Project dataset (1KGP) by the different software using their run with the highest log-likelihood. AFR: African, EUR: European, EAS: East Asian, AMR: American, and SAS: South Asian ancestry.

of the simulation scenario, SCOPE exhibited an even further increased noise level in its ancestry estimates across all three values of K.

Ancestry estimation in the 1000 Genomes Project

In addition to our simulation study, we also applied fastmixture to empirical data of the 1000 Genomes Project (1KGP), using both a downsampled version and the full set of variants. Specifically, the downsampled version was the only way to properly assess the performance of ADMIXTURE over multiple runs due to scalability issues (Figure S9). Here we observed, as in the simulation study, that fastmixture and ADMIXTURE performed comparably and the two methods achieved the highest log-likelihoods, followed by SCOPE and then Neural ADMIXTURE (Table S3), even though Neural ADMIXTURE maximizes the log-likelihood in its optimization approach.

For the full 1KGP dataset, we only performed a single run for ADMIXTURE due to its excessive computational runtime of $>$ 40 hours, in comparison to the other software with runtimes of $<$ 2 hours (Figure 3 and Table S1). Here fastmixture was ~30× faster on average than the ADMIXTURE run. We demonstrated again that fastmixture was the best performing method in terms of achieving the highest log-likelihood (Figure 2). Strikingly, Neural ADMIXTURE failed to

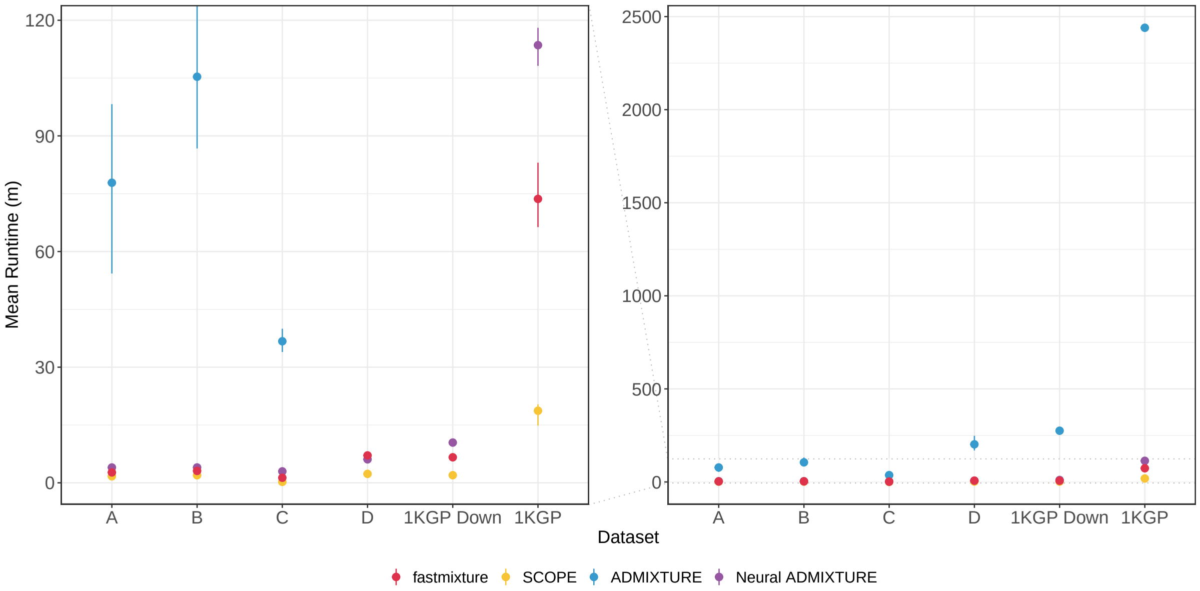

Figure 3 – Runtime (in minutes) for the different software measured across five runs in the four simulated scenarios (A, B, C, and D), and in the two 1000 Genomes Project datasets: one including all individuals (1KGP), and the downsampled dataset (1KGP Down). The intervals display the full range of runtimes across the six runs. On the left, a zoomed-in version highlights the runtime differences between the software with more similar times. The results are also described in Table S1. All analyses have been performed on a cluster with an Intel(R) Xeon(R) Gold 6152 CPU using 64 threads.

accurately distinguish ancestral contributions in the AMR and SAS super-populations in both the full dataset and in 1KGP Down. Moreover, we saw a correlation of $r^2 = 0.72$ between the estimated ancestry proportions in the downsampled and the full dataset for `Neural ADMIXTURE`, indicating poor robustness. This is in strong contrast to `fastmixture`, `ADMIXTURE` and `SCOPE` where the results between the full dataset and the downsampled one were highly correlated ($r^2 \approx 1$). We observed the same pattern of increased noise in the estimated ancestry proportions of the unadmixed individuals from `SCOPE`, which was consistent across the two 1KGP datasets.

Discussion

We have presented, `fastmixture`, our novel method and software for ancestry estimation in genotype data using our extended model-based approach. We demonstrate that our approach performs comparably to `ADMIXTURE` while being $\sim$30× faster on average across all evaluated datasets. Among the four methods assessed, `ADMIXTURE` and `fastmixture` stand out as the most accurate approaches as shown in our simulation study. In general, `fastmixture` estimates fast and accurate solutions that are robust to changes in model parameters and hyperparameters, such as initialization and number of initial mini-batches, which can be attributed to both its SVD initialization and accelerated mini-batch optimization procedure. It is well known that `ADMIXTURE` has scalability issues for large sample sizes and whole-genome sequencing data, which we also demonstrate using the 1KGP datasets. The results of `ADMIXTURE` are only based on a single seed in the full 1KGP SNP set due to a prohibitively excessive computational runtime of more than 40 hours, in contrast to $\sim$74 minutes using `fastmixture`. We have therefore evaluated all software on a downsampled version as well, with 10× less SNPs, which is a common procedure in population genetic studies.

fastmixture has runtimes comparable to the recently introduced Neural ADMIXTURE software, which employs an autoencoder framework for speeding up the ancestry estimation. Note that the testing of the software has exclusively been conducted in a CPU-based setup. However, Neural ADMIXTURE struggles across all evaluated datasets and consistently performs the poorest among the evaluated methods, as it can only manage to model unadmixed populations in simpler demographic models. In Scenario B and C, Neural ADMIXTURE fails to detect unadmixed populations and models admixed populations incorrectly as ancestral sources. SCOPE emerges as the fastest approach across all evaluated datasets, demonstrating excellent scalability. However, its optimization of a simpler least squares objective compromises its ability to accurately estimate ancestry proportions, where it can be difficult to distinguish real admixture signals from noise.

When optimizing a different objective, it is expected that the log-likelihood estimates for SCOPE would be lower compared to model-based approaches. However, based on the RMSE and JSD measures against the ground truth in the simulation study, our results showcase that the least square objective used in likelihood-free approaches, such as SCOPE, is not an optimal choice in comparison to the likelihood model. Our findings suggest that the added noise in the ancestry proportions estimated in SCOPE are likely to increase further in scenarios with a larger K or more complex demographic models, as demonstrated in Scenario C. This limits the utility of SCOPE in association studies and precision medicine. Furthermore, we observed a general trend of major optimization issues for the Neural ADMIXTURE software across all evaluated scenarios. A critical issue appears to be in their convergence evaluation, where log-likelihood estimates are normalized across individuals and variants in their mini-batch training setup, causing premature convergence. This premature convergence negatively impacts their performance, despite Neural ADMIXTURE erroneously reporting faster runtimes than it would achieve at optimal solutions.

While our fastmixture software does not entirely solve the scalability issues of model-based approaches, it represents a significant step by enabling researchers to estimate accurate ancestry proportions for much larger sample sizes and whole-genome sequencing data. It will also facilitate a more feasible exploration of increased numbers of ancestral sources. We anticipate that fastmixture will be the preferred alternative to ADMIXTURE in future population genetic studies, and it will also allow researchers to correct for population structure in genome-wide association studies of moderate sample sizes, leveraging our accurate estimates of ancestry proportions.

Acknowledgements

Preprint version 3 of this article has been peer-reviewed and recommended by Peer Community In Evolutionary Biology (https://doi.org/10.24072/pci.evolbiol.100838; Fumagalli, 2024).

Fundings

J.M. is supported by a grant from the Lundbeck Foundation (R380-2021-1225).

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Data, script, code, and supplementary information availability

The study uses genotype data from the phase 3 release of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015), publicly available at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. The results (trees and ancestry estimations) and processed genotype data files used in the study are available on Zenodo (https://doi.org/10.5281/zenodo.12683371; Meisner et al., 2024a) for reproducibility.

The fastmixture software is open-source and freely available on GitHub (https://github.com/Rosemeis/fastmixture). Scripts for reproducing the entire simulation study are available on Zenodo (https://doi.org/10.5281/zenodo.12683371; Meisner et al., 2024a).

Supplementary information is available online (https://doi.org/10.5281/zenodo.14288116; Meisner et al., 2024b)

References

Alexander DH, Novembre J, Lange K (2009). *Fast model-based estimation of ancestry in unrelated individuals*. *Genome research* **19**, 1655–1664. https://doi.org/10.1101/gr.094052.109.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, et al. (2022). *Efficient ancestry and mutation simulation with msprime 1.0*. *Genetics* **220**, iyab229. https://doi.org/10.1093/genetics/iyab229.

Browning SR, Browning BL, Daviglus ML, Durazo-Arvizu RA, Schneiderman N, Kaplan RC, Laurie CC (2018). *Ancestry-specific recent effective population size in the Americas*. *PLoS genetics* **14**, e1007385. https://doi.org/10.1371/journal.pgen.1007385.

Cabreros I, Storey JD (2019). *A likelihood-free estimator of population structure bridging admixture models and principal components analysis*. *Genetics* **212**, 1009–1029. https://doi.org/10.1534/genetics.119.302159.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015). *Second-generation PLINK: rising to the challenge of larger and richer datasets*. *Gigascience* **4**, s13742–015. https://doi.org/10.1186/s13742-015-0047-8.

Chiu AM, Molloy EK, Tan Z, Talwalkar A, Sankararaman S (2022). *Inferring population structure in biobank-scale genomic data*. *The American Journal of Human Genetics* **109**, 727–737. https://doi.org/10.1016/j.ajhg.2022.02.015.

Dominguez Mantes A, Mas Montserrat D, Bustamante CD, Giró-i-Nieto X, Ioannidis AG (2023). *Neural ADMIXTURE for rapid genomic clustering*. *Nature Computational Science*, 1–9. https://doi.org/10.1038/s43588-023-00482-7.

Engelhardt BE, Stephens M (2010). *Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis*. *PLoS genetics* **6**, e1001117. https://doi.org/10.1371/journal.pgen.1001117.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014). *Fast and efficient estimation of individual ancestry coefficients*. *Genetics* **196**, 973–983. https://doi.org/10.1534/genetics.113.160572.

Fumagalli M (2024). **fastmixture* generates fast and accurate estimates of global ancestry proportions and ancestral allele frequencies*. *Peer Community in Evolutionary Biology* **1**, 100838. https://doi.org/10.24072/pci.evolbiol.100838.

Gopalan P, Hao W, Blei DM, Storey JD (2016). *Scaling probabilistic models of genetic variation to millions of humans*. *Nature genetics* **48**, 1587–1590. https://doi.org/10.1038/ng.3710.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Project 1G, Bustamante CD, et al. (2011). *Demographic history and rare allele sharing among human populations*. *Proceedings of the National Academy of Sciences* **108**, 11983–11988. https://doi.org/10.1073/pnas.1019276108.

Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. (2020). *Array programming with NumPy*. *Nature* **585**, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Li Z, Meisner J, Albrechtsen A (2023). *Fast and accurate out-of-core PCA framework for large scale biobank data*. *Genome Research* **33**, 1599–1608. https://doi.org/10.1101/gr.277525.122.

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004). *The effects of human population structure on large genetic association studies*. *Nature genetics* **36**, 512–517. https://doi.org/10.1038/ng1337.

Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ (2019). *Clinical use of current polygenic risk scores may exacerbate health disparities*. *Nature genetics* **51**, 584–591. https://doi.org/10.1038/s41588-019-0379-x.

Meisner J, Albrechtsen A (2018). *Inferring population structure and admixture proportions in low-depth NGS data*. *Genetics* **210**, 719–731. https://doi.org/10.1534/genetics.118.301336.

Meisner J, Santander C, Refoyo-Martinez A (2024a). *Supplemental data for reproducing "Faster model-based estimation of ancestry proportions"*. Zenodo. Version 0.93.4. https://doi.org/10.5281/zenodo.14106454.

Meisner J, Santander C, Refoyo-Martinez A (2024b). *Supplemental Information: Faster model-based estimation of ancestry proportions*. Zenodo. https://doi.org/10.5281/zenodo.14288116.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. (2008). *Genes mirror geography within Europe*. *Nature* **456**, 98–101. https://doi.org/10.1038/nature07331.

Patterson N, Price AL, Reich D (2006). *Population structure and eigenanalysis*. *PLoS genetics* **2**, e190. https://doi.org/10.1371/journal.pgen.0020190.

Pritchard JK, Stephens M, Donnelly P (2000). *Inference of population structure using multilocus genotype data*. *Genetics* **155**, 945–959. https://doi.org/10.1093/genetics/155.2.945.

Ruder S (2016). *An overview of gradient descent optimization algorithms*. *arXiv:1609.04747*. https://doi.org/10.48550/arXiv.1609.04747.

Tang H, Peng J, Wang P, Risch NJ (2005). *Estimation of individual admixture: analytical and study design considerations*. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **28**, 289–301. https://doi.org/10.1002/gepi.20064.

The 1000 Genomes Project Consortium (2015). *A global reference for human genetic variation*. *Nature* **526**, 68. https://doi.org/10.1038/nature15393.

The International HapMap Consortium (2007). *A second generation human haplotype map of over 3.1 million SNPs*. *Nature* **449**, 851. https://doi.org/10.1038/nature06258.

Tsambos G, Kelleher J, Ralph P, Leslie S, Vukcevic D (2023). *link-ancestors: fast simulation of local ancestry with tree sequence software*. *Bioinformatics Advances* **3**, vbad163. https://doi.org/10.1093/bioadv/vbad163.

Wang Y, Tsuo K, Kanai M, Neale BM, Martin AR (2022). *Challenges and opportunities for developing more generalizable polygenic risk scores*. *Annual review of biomedical data science* **5**, 293–320. https://doi.org/10.1146/annurev-biodatasci-111721-074830.

Zhou H, Alexander D, Lange K (2011). *A quasi-Newton acceleration for high-dimensional optimization algorithms*. *Statistics and computing* **21**, 261–273. https://doi.org/10.1007/s11222-009-9166-3.