ADAMIXTURE: Adaptive First-Order Optimization for Biobank-Scale Genetic Clustering

Joan Saurina-i-Ricos [1,2], **Daniel Mas Montserrat**[1], **Alexander G. Ioannidis**[1,2,*]

[1]*Department of Biomedical Data Science, Stanford University, 300 Pasteur Dr, Palo Alto, CA 94305, USA*
[2]*Genomics Institute, University of California, Santa Cruz, 2300 Delaware Ave, Santa Cruz, CA 95060, USA*

[*]Corresponding author: Alexander G. Ioannidis (ioannidis@stanford.edu)

Abstract

Motivation: Estimating genetic clusters from sequencing data is a fundamental task in population and medical genetics, enabling demographic inference and adjustment for population structure in association studies. ADMIXTURE, a widely used model-based clustering method, employs an accelerated Expectation–Maximization (EM) algorithm to infer population parameters; however, its computational demands scale poorly, limiting its usefulness for modern biobank-sized datasets. While recent EM acceleration strategies employing second-order quasi-Newton schemes preserve accuracy, they remain computationally intensive. Conversely, EM-free approaches that prioritize speed often compromise solution quality.

Results: We introduce ADAMIXTURE, a novel optimization framework that integrates the EM algorithm with Adaptive Moment Estimation (Adam). Unlike traditional acceleration methods, ADAMIXTURE utilizes first-order gradients with adaptive learning rates derived from raw and squared moments to approximate curvature information, bypassing the computational overhead of Hessian approximations. This approach surpasses the convergence efficiency of second-order methods while maintaining the low computational complexity of first-order updates. Across simulated and large-scale empirical datasets, ADAMIXTURE demonstrates substantial reductions in wall-clock runtime and enhanced scalability compared to state-of-the-art methods, all while maintaining comparable or improved inference accuracy.

Availability: The source code of ADAMIXTURE is available via: `https://github.com/ai-sandbox/adamixture`.

1 Introduction

The rapid expansion of population-scale biobanks and sequencing initiatives has revolutionized the genomic prediction of human traits and disease risk. However, the predictive accuracy of these models is often linked to genetic ancestry, rendering the precise characterization of population structure a prerequisite for robust analysis [12]. For decades, unsupervised ancestry inference has served as a cornerstone for correcting stratification in genome-wide association studies and reconstructing demographic histories. Yet, as datasets scale to millions of individuals and genomic variants [5], foundational algorithms increasingly face computationally prohibitive bottlenecks. Overcoming these scalability constraints is paramount for analyzing diverse biobanks and ensuring that the promise of precision medicine is equitably realized across all populations.

The standard computational framework for this task, established by `STRUCTURE` [14], models individuals as mosaics of fractional ancestries derived from K ancestral populations. These populations are inferred via unsupervised learning in a high-dimensional space defined by variant frequencies, allowing for the characterization of continuous genetic variation without reliance on subjective labels. The model typically operates on single nucleotide polymorphisms (SNPs), exploiting the fact that allele frequency distributions diverge among populations due to distinct demographic processes, including migration, genetic drift, and founder effects.

While the Bayesian framework introduced by `STRUCTURE` was seminal, its reliance on Markov Chain Monte Carlo (MCMC) sampling imposed severe computational costs on large datasets. To mitigate this, `ADMIXTURE` [1, 2] reformulated the problem within a maximum likelihood framework, employing block relaxation and Expectation-Maximization (EM) algorithms to accelerate inference. This shift significantly extended the applicability of ancestry estimation. Nevertheless, with the exponential growth of large-scale genomic cohorts, the computational demands of `ADMIXTURE` have likewise become prohibitive, rendering it intractable for biobank-scale analysis.

In response, recent years have seen the emergence of diverse acceleration strategies, broadly categorized into likelihood-free and likelihood-based approaches. Likelihood-free methods, such as the alternating least squares (ALS) scheme in `SCOPE` [6], bypass explicit probabilistic modeling to achieve rapid estimation. Conversely, likelihood-based accelerators focus on optimizing the EM procedure itself; for instance, `fastmixture` [15] combines mini-batch EM optimization with ALS-driven initialization. Concurrently, deep learning paradigms have been adapted to this domain, notably with `Neural Admixture` [8], which recasts ancestry inference as an autoencoder task. While these innovations have improved scalability, they often introduce significant trade-offs between computational speed and statistical rigor.

Addressing this dichotomy, we introduce **ADAMIXTURE**, a novel optimization framework that accelerates the EM algorithm by integrating it with Adaptive Moment Estimation (Adam) [10]. Unlike prior acceleration attempts, ADAMIXTURE leverages first-order *pseudo-gradients* with adaptive learning rates to drive robust parameter updates within a single EM step. This approach effectively approximates curvature information without the overhead of second-order methods, dramatically enhancing convergence speed while strictly preserving likelihood maximization. The result is a powerful, accessible tool that achieves state-of-the-art performance on biobank-scale datasets.

2 Subjects and methods

2.1 Model formulation

The framework we use was first introduced in the `STRUCTURE` model, which describes a probabilistic relationship between an observed genotype matrix, G, and two latent parameter matrices: the individual admixture proportions, Q, and the ancestral allele frequencies, F.

It takes as input the $M \times N$ genotype matrix G, where each entry G_{ji} contains the count of alternate alleles $\{0, 1, 2\}$ for SNP j in individual i, across M SNPs and N individuals. The two parameter matrices are:

- **Admixture Proportions (Q):** A $K \times N$ matrix where q_{ki} is the proportion of individual i's ancestry from ancestral population k, such that $0 \leq q_{ki} \leq 1$ and $\sum_{k=1}^{K} q_{ki} = 1$.
- **Ancestral Allele Frequencies (F):** An $M \times K$ matrix where f_{jk} is the frequency of the alternate allele at SNP j in ancestral population k, such that $0 \leq f_{jk} \leq 1$.

The model connects these by defining an expected individual-specific allele frequency $p_{ji} = \sum_{k=1}^{K} f_{jk} q_{ki}$, which represents the weighted average of the ancestral frequencies. The observed genotype G_{ji} is then modeled as a random draw from a binomial distribution, $G_{ji} \sim \text{Binomial}(2, p_{ji})$. The primary goal is to infer the latent matrices Q and F from the observed genotypes G.

2.2 Parameter estimation

The parameters of the model, Q and F, are typically estimated by maximizing the likelihood of the observed genotype data G. This is equivalent to minimizing the negative log-likelihood function. Assuming each genotype G_{ji} is an independent draw from a binomial distribution, the negative log-likelihood is given by:

$$\mathcal{L}(Q, F) = -\sum_{j=1}^{M} \sum_{i=1}^{N} \left[G_{ji} \log(p_{ji}) + (2 - G_{ji}) \log(1 - p_{ji}) \right] \quad (1)$$

This minimization problem is subject to constraints imposed by the parameter definitions. The resulting optimization problem is:

$$\begin{aligned} \underset{Q,F}{\text{minimize}} \quad & \mathcal{L}(Q, F) \\ \text{subject to} \quad & 0 \leq f_{jk} \leq 1, \quad \forall j, k \\ & \sum_{k=1}^{K} q_{ki} = 1, \quad \forall i \\ & q_{ki} \geq 0, \quad \forall k, i \end{aligned} \quad (2)$$

2.3 Expectation-Maximization

We employ the Expectation-Maximization (EM) procedure originally introduced in `FRAPPÉ` [17] and later adopted by `ADMIXTURE`, which iteratively updates the ancestry proportions Q and allele frequencies F.

In the Expectation step, we compute the posterior probabilities of latent ancestry assignments for each individual i, SNP j, and ancestral population k:

$$a_{jik}^{(t)} = \frac{q_{ki}^{(t)} f_{jk}^{(t)}}{\sum_{m=1}^{K} q_{mi}^{(t)} f_{jm}^{(t)}} \qquad b_{jik}^{(t)} = \frac{q_{ki}^{(t)} (1 - f_{jk}^{(t)})}{\sum_{m=1}^{K} q_{mi}^{(t)} (1 - f_{jm}^{(t)})} \quad (3)$$

In the Maximization step, the model parameters are updated by maximizing the expected complete-data log-likelihood with respect to the current posterior assignments:

$$\begin{aligned} f_{jk}^{(t+1)} &= \frac{\sum_{i=1}^{N} G_{ji}\, a_{jik}^{(t)}}{\sum_{i=1}^{N} \left[G_{ji}\, a_{jik}^{(t)} + (2 - G_{ji})\, b_{jik}^{(t)} \right]} \\ q_{ki}^{(t+1)} &= \frac{1}{2M} \sum_{j=1}^{M} \left[G_{ji}\, a_{jik}^{(t)} + (2 - G_{ji})\, b_{jik}^{(t)} \right] \end{aligned} \quad (4)$$

This procedure alternates between computing expected ancestry assignments and updating parameter estimates until convergence.

2.4 Initialization

The initialization of ancestry proportions (Q) and ancestral allele frequencies (F) plays a critical role in ensuring the stability and convergence speed of EM-based admixture models. Naive random initialization frequently leads to slow convergence or entrapment in suboptimal local maxima, a problem that is exacerbated when acceleration methods are employed.

While `Neural Admixture` [8] employs soft clustering on PCA-projected data to provide informative initialization points, it struggles to scale to large numbers of ancestral populations (K) or very large sample sizes. To address this, we adopt and refine the scalable spectral initialization strategy employed by `SCOPE` [6] and `fastmixture` [15]. Specifically, we implement Randomized Singular Value Decomposition (SVD) with Dynamic Shifts [9], a technique that utilizes dynamically shifted power iterations to refine the approximation of leading singular vectors. This is particularly effective for handling the slow spectral decay characteristic of genetic data.

The resulting low-rank structure yields an approximation of the centered genotype matrix, $\hat{H}$, which serves as the target for an Alternating Least Squares (ALS) factorization. This procedure jointly estimates the initial Q and F matrices prior to EM optimization. To ensure numerical stability during the ALS updates, particularly in cases of rank deficiency or ill-conditioning, we incorporate Tikhonov regularization when computing the matrix inverses. Additionally, following each update step, we explicitly project the parameters Q and F onto the probability simplex to strictly enforce the constraints defined in Eq. (2).

Algorithm 1 ADAMIXTURE Initialization

Require: $G \in \mathbb{R}^{M \times N}, K, \delta, \lambda, t_{\text{max}}$

$G_C \leftarrow \text{center}(G)$
$U_{1:K}, S_{1:K}, V_{1:K} \leftarrow \text{randomizedSVD}(G_C, K)$
$\hat{H} \leftarrow \frac{1}{2} U_{1:K} S_{1:K} V_{1:K}^\top$
$F \sim \mathcal{U}(0,1)^{M \times K}$
$Q_{\text{prev}} \leftarrow 0,\ \varepsilon_{\text{best}} \leftarrow \infty,\ t_{\text{stall}} \leftarrow 0$
HighCorr $\leftarrow$ False
$F_{\text{best}} \leftarrow F,\ Q_{\text{best}} \leftarrow 0$
repeat
  $Q \leftarrow \hat{H} F (F^\top F + \lambda I_K)^{-1}$
  $F \leftarrow \hat{H}^\top Q (Q^\top Q + \lambda I_K)^{-1}$
  $e \leftarrow \text{RMSE}(Q, Q_{\text{prev}})$
  $G_F \leftarrow F^\top F$
  $v \leftarrow \sqrt{\text{diag}(G_F)}$
  $C \leftarrow G_F \oslash (v v^\top)$
  if $\neg$HighCorr $\wedge \max_{i \neq j} |C_{ij}| > 0.95$ **then**
    HighCorr $\leftarrow$ True
  end if
  if HighCorr **then**
    if $e < \varepsilon_{\text{best}}$ **then**
      $\varepsilon_{\text{best}} \leftarrow \infty,\ t_{\text{stall}} \leftarrow 0$
      $Q_{\text{best}} \leftarrow Q,\ F_{\text{best}} \leftarrow F$
    else
      $t_{\text{stall}} \leftarrow t_{\text{stall}} + 1$
    end if
    if $t_{\text{stall}} \geq t_{\text{max}}$ **then**
      $Q \leftarrow Q_{\text{best}},\ F \leftarrow F_{\text{best}}$
      break
    end if
  end if
  $Q_{\text{prev}} \leftarrow Q$
until $e < \delta$
return Q, F

Beyond these algebraic constraints, we introduce a supplementary stabilization mechanism to handle model over-specification, occurring when the specified K exceeds the data's true latent dimensionality. In such regimes, the columns of F frequently exhibit high collinearity, leading to ill-conditioned updates and numerical instability. To mitigate this, our algorithm monitors the pairwise correlations between ancestral frequency vectors. Upon detecting significant collinearity, the initialization shifts to a robust termination criterion: rather than enforcing strict convergence tolerances—which can induce oscillatory behavior—the procedure tracks the reconstruction error and terminates after a fixed interval of stagnation, returning the optimal solution found. This ensures algorithm robustness even in degenerate, over-parameterized scenarios. Algorithm 1 shows a detailed description of the initialization steps, with the blue-colored steps highlighting the novel procedures incorporated compared to the basic randomized SVD + ALS approach used in `SCOPE` and `fastmixture`.

2.5 Adam-EM

Standard Expectation-Maximization (EM) algorithms exhibit linear convergence rates, which become prohibitively slow on biobank-scale datasets. While acceleration strategies based on second-order Quasi-Newton schemes [18]—such as those employed by `ADMIXTURE` and `fastmixture`—improve convergence, they typically necessitate two EM evaluations per parameter update to approximate the Hessian. This effectively doubles the computational cost per iteration, creating a bottleneck in high-dimensional settings.

Algorithm 2 Adam-EM

Require: $\theta_0, \alpha, \beta_1, \beta_2, \epsilon$

$m \leftarrow 0, \quad v \leftarrow 0, \quad t \leftarrow 0$
while not converged **do**
  $\theta_{EM} \leftarrow \text{EM}(\theta_t)$ ▷ Compute standard EM proposal
  $g_t \leftarrow \theta_{EM} - \theta_t$ ▷ Derive pseudo-gradient
  $m \leftarrow \beta_1 m + (1 - \beta_1) g_t$
  $v \leftarrow \beta_2 v + (1 - \beta_2) g_t^2$
  $\hat{m} \leftarrow m / (1 - \beta_1^{t+1})$
  $\hat{v} \leftarrow v / (1 - \beta_2^{t+1})$
  $\theta_{t+1} \leftarrow \theta_t + \alpha \frac{\hat{m}}{\sqrt{\hat{v}} + \epsilon}$ ▷ Apply adaptive update
  $t \leftarrow t + 1$
end while
return θ_t

To overcome this limitation, we introduce **Adam-EM**, a hybrid optimization framework that integrates the stability of EM with the adaptive dynamics of the Adam optimizer. Unlike second-order approaches, Adam-EM operates strictly with first-order information. We interpret the parameter shift induced by a standard EM step, $\Delta\theta = \theta_{EM} - \theta_t$, as a *pseudo-gradient*—a proxy for the steepest ascent direction on the likelihood manifold. By ingesting this vector into the Adam update rule, we leverage adaptive moment estimation to normalize updates across parameters with varying sensitivities. This effectively approximates curvature information and injects momentum without the computational overhead of Hessian construction. The detailed pseudocode for this general optimization strategy is presented in Algorithm 2.

Beyond ancestry inference, we propose Adam-EM as a general acceleration strategy for the broad class of latent variable models where EM updates are computationally expensive and exact gradients are intractable—such as probabilistic matrix factorization and topic modeling. The underlying mechanism, which decouples the search direction (provided by EM) from the step size dynamics (managed by Adam), suggests that similar performance gains can be expected in other settings characterized by high dimensionality and ill-conditioned parameter spaces. In these regimes, Adam-EM offers a scalable alternative that bridges the gap between the stability of first-order methods and the convergence speed of second-order schemes.

2.6 Optimization

We apply the Adam-EM acceleration scheme to simultaneously optimize the admixture proportions Q and the ancestral allele frequencies F. In each iteration, we utilize the standard EM update rules derived in Eqs. (3)–(4) to compute the pseudo-gradients, following the logic described in Algorithm 2. Subsequently, after applying the adaptive step, the parameters are explicitly projected onto the probability simplex to strictly satisfy the constraints formulated in Eq. (2). To strictly guarantee numerical stability, we incorporate a dynamic learning rate schedule that monitors the negative log-likelihood objective function $\mathcal{L}(Q, F)$

defined in Eq. (1). Since our goal is to minimize this objective, an increase in $\mathcal{L}$ signals that the step size is too aggressive. If this occurs, the learning rate α is immediately decayed by a factor $\gamma \in (0, 1)$ to regain stability and ensure monotonic convergence. The complete optimization procedure, including this stabilization strategy, is detailed in Algorithm 3.

Algorithm 3 ADAMIXTURE

Require: $G \in \mathbb{R}^{M \times N}, K, \alpha, \gamma, \beta_1, \beta_2, \epsilon, Q, F$

$m_Q = v_Q = m_F = v_F \leftarrow 0$
$t \leftarrow 0$
$\mathcal{L}_{\text{prev}} \leftarrow \infty$
while not converged **do**
  $Q_1, F_1 \leftarrow \text{EM}(G, Q, F)$
  $\nabla_Q \leftarrow Q_1 - Q, \quad \nabla_F \leftarrow F_1 - F$
  for $\theta \in \{Q, F\}$ **do**
    $m_\theta \leftarrow \beta_1 m_\theta + (1 - \beta_1)\nabla_\theta$
    $v_\theta \leftarrow \beta_2 v_\theta + (1 - \beta_2)(\nabla_\theta)^2$
    $\hat{m}_\theta \leftarrow m_\theta / (1 - \beta_1^{t+1})$
    $\hat{v}_\theta \leftarrow v_\theta / (1 - \beta_2^{t+1})$
    $\theta \leftarrow \theta + \alpha \frac{\hat{m}_\theta}{\sqrt{\hat{v}_\theta} + \epsilon}$
  end for
  $\mathcal{L}_{\text{curr}} \leftarrow \mathcal{L}(Q, F)$
  if $\mathcal{L}_{\text{curr}} > \mathcal{L}_{\text{prev}}$ **then**
    $\alpha \leftarrow \alpha \cdot \gamma$
  end if
  $\mathcal{L}_{\text{prev}} \leftarrow \mathcal{L}_{\text{curr}}$
  $t \leftarrow t + 1$
end while
return Q, F

3 Benchmarking

3.1 Assessment of results

We assess method performance through a dual strategy encompassing quantitative goodness-of-fit metrics and qualitative inspection of population structure. Given the unsupervised paradigm of ancestry inference, where absolute ground truth is absent in empirical datasets, we rely on the negative log-likelihood from Eq. (1) as the definitive objective function. This metric quantifies the statistical fidelity of the parameter estimates and the model's capacity to explain the observed genetic variation.

Complementing statistical fit, we benchmark computational efficiency by recording wall-clock runtime across varying dataset dimensions to evaluate scalability, and qualitatively validate the results by visualizing the estimated admixture proportions (Q) using *pong* [3], enabling a direct comparative assessment of the biological interpretability, coherence, and stability of the inferred population structures across algorithms. The results from the experiments conducted on both empirical and simulated datasets are publicly available on figshare [16].

3.2 Empirical datasets

To evaluate ancestry inference under realistic conditions, we selected four genomic cohorts representing distinct regimes of sample size, SNP density, and population structure. This diverse set provides a robust framework for assessing both scalability and accuracy, capturing patterns from fine-scale stratification to global population divergence.

The **UK Biobank (UKB)** [5] serves as our primary benchmark for biobank-scale inference. For the central accuracy and runtime comparisons, we curated a standard subset comprising 100,000 individuals and 100,000 SNPs, obtained via random subsampling and minor allele frequency (MAF) filtering. Additionally, to systematically evaluate algorithmic scalability across different data dimensions, we generated a spectrum of auxiliary subsets with sample sizes ranging from 50,000 to 500,000 individuals and SNP counts from 100,000 to 500,000. This cohort, predominantly of European ancestry, presents a challenge of scale and fine-structure resolution, representative of modern Genome-Wide Association Study (GWAS) contexts.

The **Domestic Dog** dataset [13] includes genotype data from 1,355 individuals spanning 166 breeds (150,131 SNPs). Originating from diverse lineages across all inhabited continents, this dataset exhibits strong genetic drift and sharp population boundaries due to breed isolation. It serves as a rigorous test for model robustness and adaptability, particularly when inferring high numbers of ancestral populations (K).

The **Human Genome Diversity Project (HGDP)** and **1000 Genomes Project** merged dataset comprises genotype data from 4,091 individuals representing a wide range of globally distributed populations. After applying minor allele frequency (MAF) filtering, the resulting dataset contains approximately 500,000 single-nucleotide polymorphisms (SNPs) spanning all autosomal chromosomes (1–22), with sex chromosomes excluded. In contrast to the UK Biobank cohort, this combined dataset captures substantial worldwide genetic diversity and pronounced population structure, making it a complementary benchmark for evaluating model performance under complex, multi-ancestry settings [7, 11].

3.3 Simulated datasets

To ensure reproducibility and enable direct comparison with state-of-the-art acceleration methods, we employ the standardized simulation benchmarks introduced by `fastmixture` [15]. These datasets were generated using coalescent simulations assuming a constant effective population size of 10,000 to model neutral evolution.

We analyze four distinct demographic scenarios (A–D). **Scenarios A**, **B**, and **C** model discrete populations under constant size assumptions, while **Scenario D** incorporates a complex American-Admixture event [4]. The cohorts consist of 1,000 individuals for Scenarios A, B, and D, and 1,600 individuals for Scenario C. Post-processing retained variants with MAF > 0.05, resulting in datasets ranging from 500,000 to 700,000 SNPs.

3.4 Experimental setup

All experiments were conducted on the UCSC Genomics Institute Phoenix cluster. Each computational run was allocated 64 logical cores on an AMD EPYC 7713 processor. As `Neural Admixture` is the only method capable of leveraging hardware acceleration, it was provisioned with a single NVIDIA A100 GPU (80 GB VRAM); all other methods were restricted to CPU execution.

Table 1: **Log-likelihood measures on Simulated Datasets.** Each dataset was run with the specified number of clusters: Dataset A ($K = 3$), B ($K = 4$), C ($K = 5$), D ($K = 3$). For each method, the mean of 3 runs is reported, with the standard deviation shown in parentheses.

Method	A	B	C	D
ADAMIXTURE	**-637,729,492.5 (0.1)**	**-629,360,598.5 (0.3)**	**-1,006,294,344.6 (0.2)**	-429,286,152.9 (0.4)
fastmixture	-637,729,494.3 (0.6)	-629,360,599.1 (0.3)	-1,006,294,347.8 (0.6)	**-429,286,152.4 (0.05)**
Neural Admixture	-640,179,649.8 (3.4×10^4)	-633,931,492.0 (3.2×10^4)	-1,013,856,785.5 (5.2×10^4)	-433,961,974.7 (6.6×10^5)
ADMIXTURE	-637,729,492.5 (0.5)	-629,360,598.7 (1.8)	-1,006,294,346.7 (2.0)	-429,286,157.3 (3.8)
SCOPE	-637,846,219.5 (7.2)	-629,538,194.4 (37.1)	-1,006,604,611.6 (251.3)	-429,584,916.5 (25.2)

We benchmark the performance of ADAMIXTURE against the following state-of-the-art methods, all executed using default hyperparameters:

- **`ADMIXTURE (1.4.1)`**: A maximum-likelihood method based on block relaxation: `http://dalexander.github.io/admixture/download.html`
- **`Neural Admixture (1.4.4)`**: A deep learning autoencoder-based approach: `https://github.com/AI-sandbox/neural-admixture`
- **`fastmixture (1.2.0)`**: An accelerated EM implementation using quasi-Newton methods: `https://github.com/Rosemeis/fastmixture`
- **`SCOPE`**: A likelihood-free method based on alternating least squares: `https://github.com/sriramlab/SCOPE`

4 Results

4.1 Performance on Simulated Benchmarks

Quantitative benchmarking on simulated data confirms that ADAMIXTURE consistently recovers the global maximum likelihood solution, exhibiting numerical precision and stability strictly comparable to, or exceeding, the exact EM implementations of `ADMIXTURE` and `fastmixture`. While all three likelihood-based methods converge to the same global mode, ADAMIXTURE achieves the highest absolute log-likelihood values across Datasets A, B, and C, effectively outperforming the baselines in convergence tightness (Table 1).

In contrast, the approximate methods fail to reach this optimality. `SCOPE` yields significantly suboptimal fits, consistently falling short of the likelihood quality obtained by EM-based approaches. Similarly, `Neural Admixture` displays marked instability and high variance across replicates, failing to provide reliable parameter estimates.

4.2 Performance on Empirical Benchmarks

Evaluation on large-scale empirical cohorts further validates the efficacy of our framework. As detailed in Table 2, ADAMIXTURE attains the maximum log-likelihood across all evaluated datasets—UK Biobank, Dogs, and HGDP. Crucially, our method demonstrates superior numerical stability, maintaining negligible standard deviations across replicates, whereas competing accelerated methods frequently exhibit substantial variance. Across all cohorts, `Neural Admixture` and `SCOPE` systematically fail to converge to the optimal mode, resulting in model fits that differ by orders of magnitude in likelihood compared to exact methods.

While `fastmixture` and classical `ADMIXTURE` generally converge toward near-optimal solutions, ADAMIXTURE consistently outperforms both methods in terms of final precision and convergence stability. The practical impact of this improved robustness is illustrated in Figure 1. Qualitative inspection shows that the population structure inferred by ADAMIXTURE is visually indistinguishable from that recovered by `fastmixture` and classical `ADMIXTURE`, demonstrating that the proposed acceleration strategy preserves fine-scale population structure without sacrificing resolution relative to the other two methods.

Conversely, likelihood-free accelerators introduce significant artifacts. `Neural Admixture` produces over-smoothed estimates, assigning nearly identical ancestry proportions to distinct sub-populations, likely due to mode collapse or aggressive regularization. `SCOPE` yields noisy, inconsistent cluster assignments that obscure underlying genetic signals. Notably, the solutions produced by `SCOPE` closely mirror those obtained using only our spectral initialization step without EM refinement. This suggests that `SCOPE`'s ALS-based strategy may be stagnating at a superficial local minimum equivalent to a high-quality linear initialization, which explains its rapid runtime but justifies its substantially reduced solution quality.

Table 2: **Log-likelihood measures on Empirical Datasets.** Each dataset was run with the specified number of clusters: UKB ($K = 5$), DOGS ($K = 20$), HGDP ($K = 8$). For each method, the mean of 3 runs is reported, with the standard deviation shown in parentheses.

Method	UKB	Dogs	HGDP
ADAMIXTURE	**-5,324,953,197.1 (20.9)**	**-194,656,413.9 (1.4)**	**-1,131,620,425.7 (1.7)**
fastmixture	-5,325,159,266.5 (290,581.1)	-194,656,417.1 (0.5)	-1,131,620,448.5 (5.5)
Neural Admixture	-5,335,153,789.3 (231,021.0)	-209,223,075.7 (5,696.8)	-1,152,976,562.6 (45,148.3)
ADMIXTURE	-5,324,953,304.6 (35.8)	-194,769,680.1 (50,278.7)	-1,131,908,125.6 (406,377.6)
SCOPE	-5,340,146,107.4 (3,018,310.2)	-195,985,486.5 (48.7)	-1,134,191,014.5 (2,669.4)

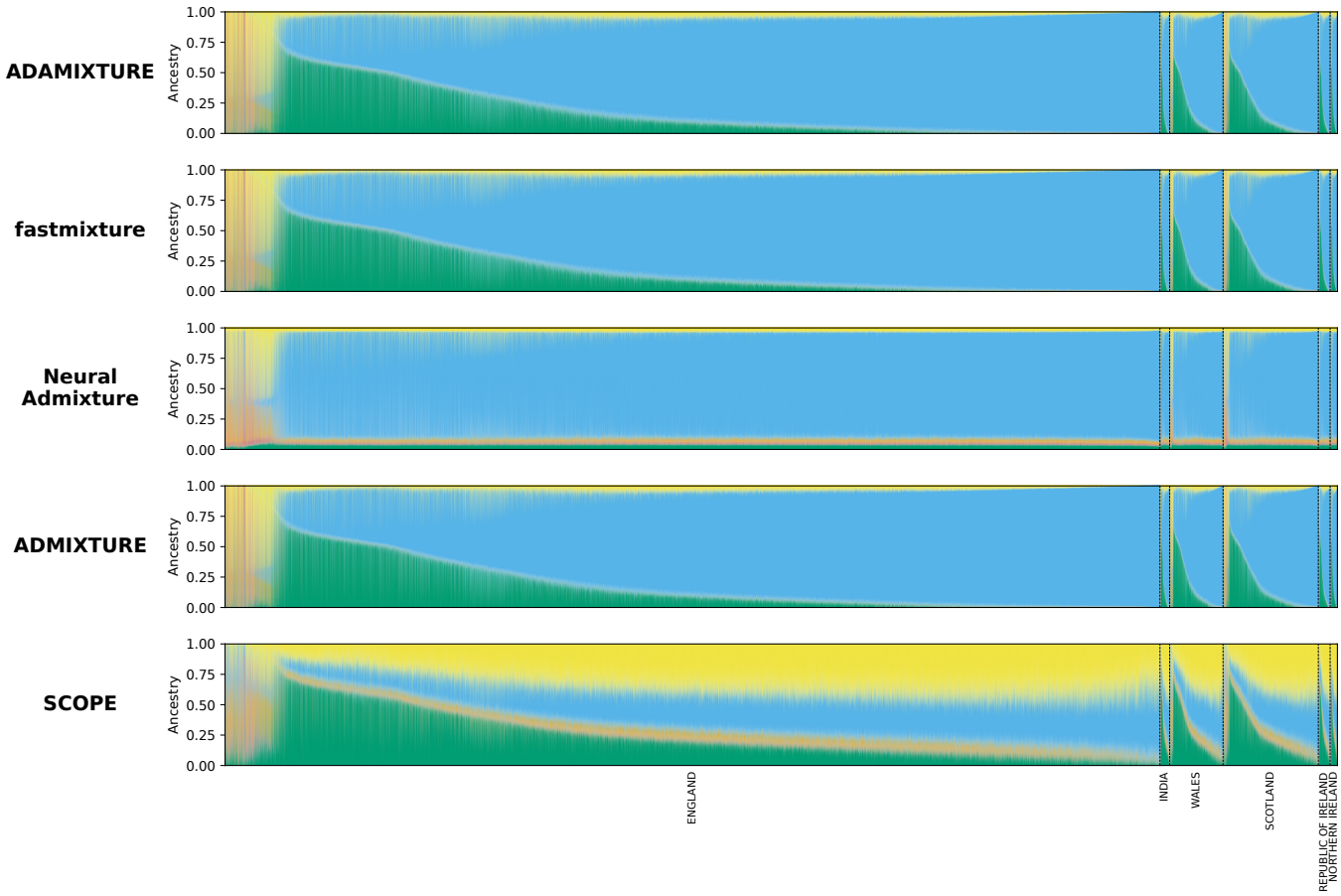

Figure 1: **Visualization of the estimated ancestry proportions for the UK Biobank dataset ($K = 5$) obtained using the different benchmarked methods.** Labels correspond to self-reported location of birth. Only regions with a sufficient number of individuals to ensure clear visualization are shown; these include England, India, Wales, Scotland, the Republic of Ireland, and Northern Ireland.

4.3 Runtime and Accuracy Trade-off

The runtime efficiency of ADAMIXTURE was systematically benchmarked against competing methods across all datasets, with results summarized in Figure 2. The most dramatic acceleration is observed relative to the classical `ADMIXTURE` algorithm; on the largest dataset (UK Biobank), `ADMIXTURE` necessitated over 57 hours to converge, whereas our method completed the task in approximately 53 minutes. This corresponds to a speedup factor exceeding $60\times$, effectively transforming the analysis of biobank-scale data from a multi-day computationally intensive task into a sub-hour operation. Across all evaluated cohorts, ADAMIXTURE consistently outperforms other accelerated approaches, achieving an average speedup of approximately $3.5\times$ over `fastmixture` and $4\times$ over `Neural Admixture`.

A nuanced performance profile is observed for `SCOPE`. While it exhibits the lowest absolute wall-clock times on datasets with moderate feature dimensionality ($< 500,000$ SNPs; specifically D, UKB, Dogs, and HGDP), ADAMIXTURE proves more efficient on high-dimensional datasets (A, B, C). Crucially, as established by the accuracy analysis, `SCOPE`'s speed advantage in low-complexity regimes is obtained at the direct expense of solution quality.

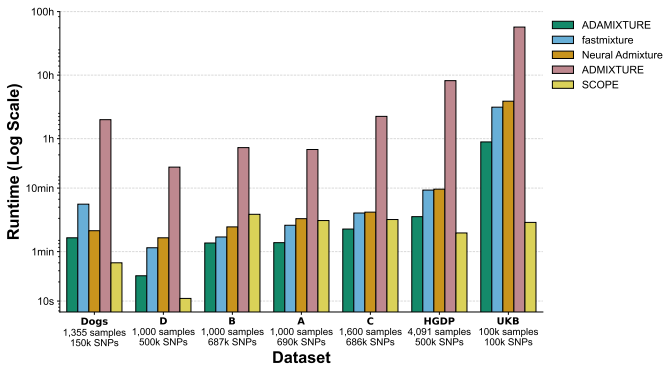

Figure 2: **Comparison of runtime performance across synthetic (A-D) and real-world (UKB, Dogs, HGDP) datasets.** The y-axis is shown on a logarithmic scale to accommodate the variance in execution times. ADAMIXTURE (green) demonstrates superior scalability among methods that achieve optimal solutions, particularly on the massive UK Biobank dataset. Results represent the mean of 3 runs.

To synthesize the global performance landscape, Figure 3 presents the aggregate Mean Normalized Log-Likelihood against Mean Normalized Runtime. In this coordinate space, the theoretical ideal lies at the origin of the efficiency frontier (runtime ≈ 0, normalized likelihood ≈ 1).

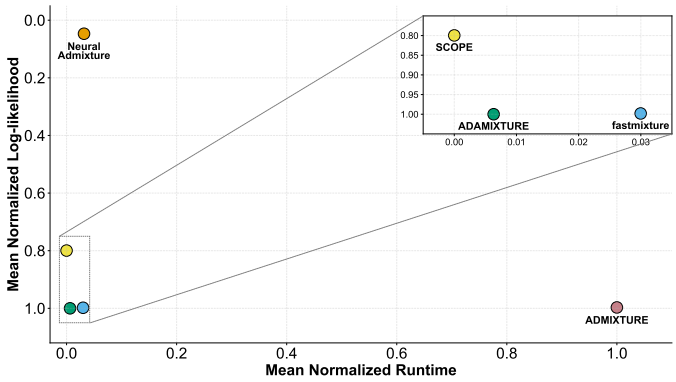

Figure 3: **Global trade-off between solution quality and computational cost.** The scatter plot aggregates performance across all datasets, displaying Mean Normalized Runtime (x-axis, lower is better) versus Mean Normalized Log-likelihood (y-axis, inverted scale where 1.0 is optimal). The inset zooms into the high-accuracy region. ADAMIXTURE (green) uniquely occupies the ideal position in the Pareto frontier, achieving the optimal likelihoods (1.0) of standard `ADMIXTURE` while maintaining runtime efficiency comparable to `SCOPE`.

The plot highlights the structural limitations of existing fast approaches. `Neural Admixture`, despite achieving low runtimes, suffers from severe degradation in solution quality, appearing far from the optimal frontier with a mean normalized likelihood close to 0. Similarly, while `SCOPE` is depicted in the inset as a highly efficient method, it fails to reach the global optimum, plateauing at a normalized likelihood of approximately 0.80. This confirms that its computational speed is achieved by sacrificing the statistical rigor required for accurate population structure inference.

In contrast, ADAMIXTURE and `fastmixture` are the only accelerated methods that match the solution quality of the gold-standard `ADMIXTURE`, both achieving a mean normalized likelihood of 1.0. However, ADAMIXTURE is distinctly superior in efficiency, occupying a strictly dominant position relative to `fastmixture` on the runtime axis. Consequently, ADAMIXTURE represents the optimal balance in the current landscape of admixture inference, providing the accuracy of likelihood-based solvers with the computational speed typically reserved for heuristic approximations.

4.4 Scalability Analysis

To ensure a rigorous assessment of computational efficiency, we restrict our scalability benchmarks to ADAMIXTURE and `fastmixture`. As demonstrated in Figure 3, heuristic approximations such as `SCOPE` and `Neural Admixture` fail to converge to the global maximum likelihood mode. Consequently, including them in a direct efficiency comparison would be methodologically flawed, as their lower runtimes derive primarily from premature termination at suboptimal local optima rather than algorithmic superiority. We therefore focus on the two methods capable of reliably recovering the true latent structure.

We first evaluate the impact of model complexity (K) on runtime using a fixed subset of the UK Biobank. As shown in Figure 4, ADAMIXTURE exhibits a significantly flatter growth curve compared to `fastmixture`. This efficiency gain is intrinsic to our Adam-EM formulation: by avoiding the dual-step EM evaluation required for the quasi-Newton updates in `fastmixture`, we effectively halve the computational burden associated with estimating the population matrices (F and Q) at each iteration, a saving that becomes increasingly critical as K grows.

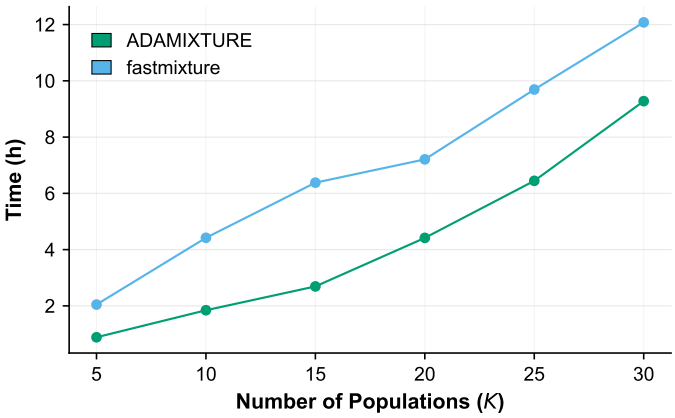

Figure 4: **Runtime scalability with respect to model complexity (K).** Runtime comparison between ADAMIXTURE (green) and `fastmixture` (blue) on a fixed subset of the UK Biobank dataset ($100,000$ samples, $100,000$ SNPs). While both methods exhibit linear scaling with K, ADAMIXTURE demonstrates a significantly lower slope, indicating superior efficiency in high-complexity regimes.

Finally, we analyze scalability with respect to data volume. Figure 5 illustrates runtime performance when varying sample size (N) and number of SNPs (M) with fixed number of populations ($K = 5$). ADAMIXTURE consistently outperforms `fastmixture` across all tested configurations. Crucially, the performance gap widens as the number of samples increases—the limiting factor in modern biobank analyses. This diverging trend highlights ADAMIXTURE's suitability for massive cohorts where N and M reach the scale of hundreds of thousands, offering a linear-time solution that does not compromise on statistical rigor.

5 Discussion

The results presented in this study demonstrate that the integration of first-order adaptive optimization with the Expectation-Maximization framework offers an effective solution to the scalability bottlenecks inherent in classical genetic clustering and ancestry inference. A central finding of our work is the redundancy of performing second-order updates in high-dimensional genomic settings. While traditional acceleration strategies, such as the quasi-Newton schemes employed by `fastmixture` and `ADMIXTURE`, necessitate two EM evaluations per parameter update to approximate the curvature of the likelihood surface, our analysis suggests that this computational expense is unnecessary. ADAMIXTURE effectively navigates the parameter space using pseudo-gradients derived from a single EM step. By coupling adaptive moment estimation with a dynamic learning rate schedule based on likelihood monitoring, our approach achieves robust convergence without the overhead of secant approximations. The dynamic schedule acts as a stabilizer, ensuring that

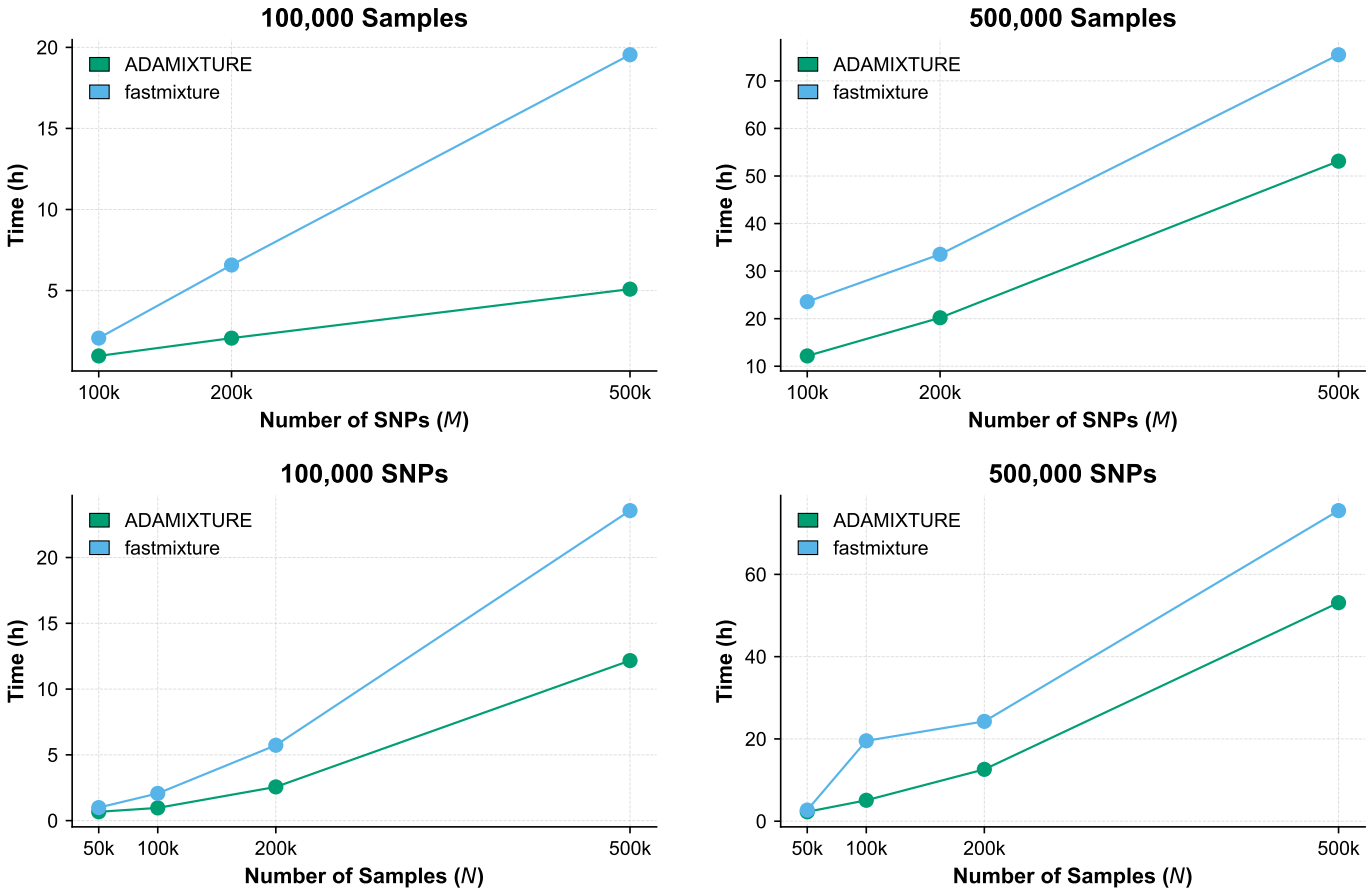

Figure 5: **Scalability with respect to dataset dimensions.** The plot illustrates runtime growth on the UK Biobank dataset as a function of the number of SNPs (M) (top panels) and sample size (N) (bottom panels), with the number of populations fixed at $K = 5$. ADAMIXTURE (green) consistently outperforms `fastmixture` (blue) across all test cases, demonstrating robust scalability for biobank-scale cohorts.

the aggressive step sizes characteristic of Adam do not lead to divergence, thereby rendering the transition to complex quasi-Newton refinements obsolete.

Our evaluation across diverse benchmarks confirms that this algorithmic efficiency does not come at the cost of inference quality. In both simulated scenarios and large-scale empirical datasets such as the UK Biobank, our method consistently recovered the optimal likelihood solutions, matching the statistical fidelity of exact EM baselines while maintaining superior numerical stability. This stands in sharp contrast to recent likelihood-free or variational approaches (e.g., `SCOPE`, `Neural Admixture`), which, despite their speed, frequently yielded suboptimal fits and failed to capture fine-scale population structures. By strictly preserving the maximum-likelihood formulation, our approach ensures that the resulting ancestry estimates remain interpretable and robust, effectively resolving the trade-off between computational speed and statistical accuracy that has historically constrained the field.

Regarding scalability, our empirical analysis reveals that ADAMIXTURE offers distinct asymptotic advantages over existing accelerated methods. As demonstrated by our runtime benchmarks, the method exhibits a significantly flatter growth curve with respect to model complexity (K) compared to `fastmixture`. This efficiency is directly attributable to the decoupled nature of the Adam-EM update, which avoids the compounding overhead of the dual-step evaluations required by quasi-Newton methods. Furthermore, the linear scaling observed with respect to sample size (N) positions our framework as a uniquely viable tool for next-generation biobanks, where cohort sizes are projected to reach millions of individuals. As dataset dimensions expand, the performance gap between ADAMIXTURE and traditional solvers widens, highlighting the necessity of first-order adaptive methods for massive-scale inference.

Despite these advances, the current implementation remains bound by the inherent limitations of Central Processing Unit (CPU) architectures. While ADAMIXTURE significantly outperforms existing CPU-based methods—reducing runtimes from days to minutes for standard subsets—the analysis of full-scale cohorts (e.g., 500,000+ individuals) remains a non-trivial computational challenge, extrapolating to over twenty hours of continuous computation on standard hardware. The massive parallelism offered by Graphics Processing Units (GPUs) remains an underutilized resource for high-precision inference. Although `Neural Admixture` attempted to bridge this gap, our analysis indicates that it did not obtain optimal results in terms of likelihood maximization. This highlights a significant opportunity for future research: establishing a modeling framework—whether by adapting ADAMIXTURE to GPU kernels, re-engineering neural approaches, or developing novel hybrid architectures—that successfully reconciles massive hardware parallelism with rigorous statistical guarantees. Realizing this

potential could effectively compress the computational timeline from the current multi-hour standard to mere minutes, enabling real-time results even at biobank-scale.

In summary, ADAMIXTURE represents a pivotal advancement in the analysis of large-scale genomic data. By successfully coupling the rigorous statistical foundation of maximum-likelihood estimation with modern adaptive optimization and dynamic scheduling, we have effectively eliminated the historical necessity to compromise between computational feasibility and scientific precision. As biobank cohorts expand to millions of individuals, the ability to perform robust, interpretable, and scalable inference becomes a prerequisite for discovery. ADAMIXTURE fills this critical gap, providing the community with a high-performance engine to uncover the fine-scale population structures that underpin our understanding of human genetic variation, demographic history, and complex disease architecture.

6 Author contributions

J.S.R., A.G.I., and D.M.M. designed the research. J.S.R. wrote the software. J.S.R., D.M.M., and A.G.I. interpreted the results. J.S.R., D.M.M., and A.G.I. wrote the manuscript.

7 Acknowledgements

This work uses data provided by patients and collected by the NHS as part of their care and support. We thank the UK Biobank and its participants and the NHS. This research has been conducted using the UK Biobank Resource under Application Number 24983.

References

[1] D. H. Alexander and K. Lange. Enhancements to the admixture algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12:246, 2011.

[2] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

[3] A. A. Behr, K. Z. Liu, G. Liu-Fang, P. Nakka, and S. Ramachandran. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, 32(18): 2817–2823, 2016.

[4] S. R. Browning, B. L. Browning, M. L. Daviglus, R. A. Durazo-Arvizu, N. Schneiderman, R. C. Kaplan, and C. C. Laurie. Ancestry-specific recent effective population size in the americas. *PLoS Genetics*, 14(5):e1007385, 2018.

[5] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562:203–209, 2018.

[6] A. M. Chiu, E. K. Molloy, Z. Tan, A. Talwalkar, and S. Sankararaman. Inferring population structure in biobank-scale genomic data. *The American Journal of Human Genetics*, 109:727–737, April 2022.

[7] G. P. Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[8] A. Dominguez Mantes, D. Mas Montserrat, C. D. Bustamante, X. Giró-i Nieto, and A. G. Ioannidis. Neural admixture for rapid genomic clustering. *Nature Computational Science*, 2023.

[9] X. Feng, W. Yu, Y. Xie, and J. Tang. Algorithm 1043: Faster randomized svd with dynamic shifts. *ACM Transactions on Mathematical Software*, 50(2):1–27, 2024.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] J. Z. Li et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319: 1100–1104, 2008.

[12] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, 2019.

[13] H. G. Parker, D. L. Dreger, M. Rimbault, B. W. Davis, A. B. Mullen, G. Carpintero-Ramirez, and E. A. Ostrander. Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell Reports*, 19(4):697–708, 2017.

[14] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[15] C. G. Santander, A. Refoyo Martinez, and J. Meisner. Faster model-based estimation of ancestry proportions. *Peer Community Journal*, 4:e115, 2024.

[16] J. Saurina-i-Ricos, D. Mas Montserrat, and A. G. Ioannidis. Results for ADAMIXTURE: Ancestry proportions (Q) and allele frequencies (P) on simulated and public empirical datasets, 2026. URL https://doi.org/10.6084/m9.figshare.31065292.

[17] H. Tang, J. Peng, P. Wang, and N. J. Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301, 2005.

[18] H. Zhou, D. Alexander, and K. Lange. A quasi-Newton acceleration for high-dimensional optimization algorithms. *Optimization*, 58(5):655–672, 2009.