nature computational science

Article

https://doi.org/10.1038/s43588-023-00482-7

Neural ADMIXTURE for rapid genomic clustering

Received: 25 February 2022

Accepted: 6 June 2023

Published online: 6 July 2023

Check for updates

Albert Dominguez Mantes[1,2,3], Daniel Mas Montserrat[1], Carlos D. Bustamante[4], Xavier Giró-i-Nieto[2] & Alexander G. Ioannidis[1,5]

Characterizing the genetic structure of large cohorts has become increasingly important as genetic studies extend to massive, increasingly diverse biobanks. Popular methods decompose individual genomes into fractional cluster assignments with each cluster representing a vector of DNA variant frequencies. However, with rapidly increasing biobank sizes, these methods have become computationally intractable. Here we present Neural ADMIXTURE, a neural network autoencoder that follows the same modeling assumptions as the current standard algorithm, ADMIXTURE, while reducing the compute time by orders of magnitude surpassing even the fastest alternatives. One month of continuous compute using ADMIXTURE can be reduced to just hours with Neural ADMIXTURE. A multi-head approach allows Neural ADMIXTURE to offer even further acceleration by computing multiple cluster numbers in a single run. Furthermore, the models can be stored, allowing cluster assignment to be performed on new data in linear time without needing to share the training samples.

The rapid growth in sequenced human genomes and the proliferation of population-scale biobanks have enabled the creation of increasingly accurate models to predict traits and disease risk using an individual's genome. However, different predictive models can be required depending on an individual's genetic ancestry, and this necessitates accurately characterizing genetic cluster composition at the individual level[1]. Such characterization is also an essential part of most modern population genetics studies and national biobanking efforts[2,3]. However, many existing algorithms for this task struggle with next-generation sequencing datasets, where both the number of samples and the number of sequenced positions along the genome are much greater than earlier case–control genotyping studies. Scalable algorithms to characterize the population structure of genetic sequences are especially important for more diverse biobanks, themselves needed to correct the extreme imbalance towards European-descent samples in existing studies in order to avoid a new divide in healthcare arising through omitting most of the world's population from precision health research[4].

A common approach for characterizing the population structure within a genetic dataset is to describe each sample as a set of fractional assignments to each cluster. These clusters are centroids found via an unsupervised algorithm in a space spanning the frequencies of each variant. By avoiding the culture-specific labels and subjective constructs (for example, ethnicity) of supervised classification methods[5], these unsupervised approaches can better reflect the spectrum of genetic structure across samples. Generally, the input variants are the individual's sequence of single nucleotide polymorphisms (SNPs), that is, single positions along the genome known to vary between individuals. Smaller datasets of less numerous variants, such as microsatellites, have also been used. There are millions of SNPs in the human genome and most are biallelic (two variants) permitting a binary encoding. For instance, zero could be used to encode the most common (or reference) variant at an SNP position on the genome and one to encode the minority (or alternate) variant. The frequency distribution of these variants will vary between populations due to differing histories: founder events, migration, isolation, and drift.

[1]Department of Biomedical Data Science, Stanford Medical School, Stanford, CA, US. [2]Signal Theory and Communications Department, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain. [3]School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Vaud, Switzerland. [4]Galatea Bio, Hialeah, FL, US. [5]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, US. e-mail: ioannidis@stanford.edu

https://doi.org/10.1038/s43588-023-00482-7

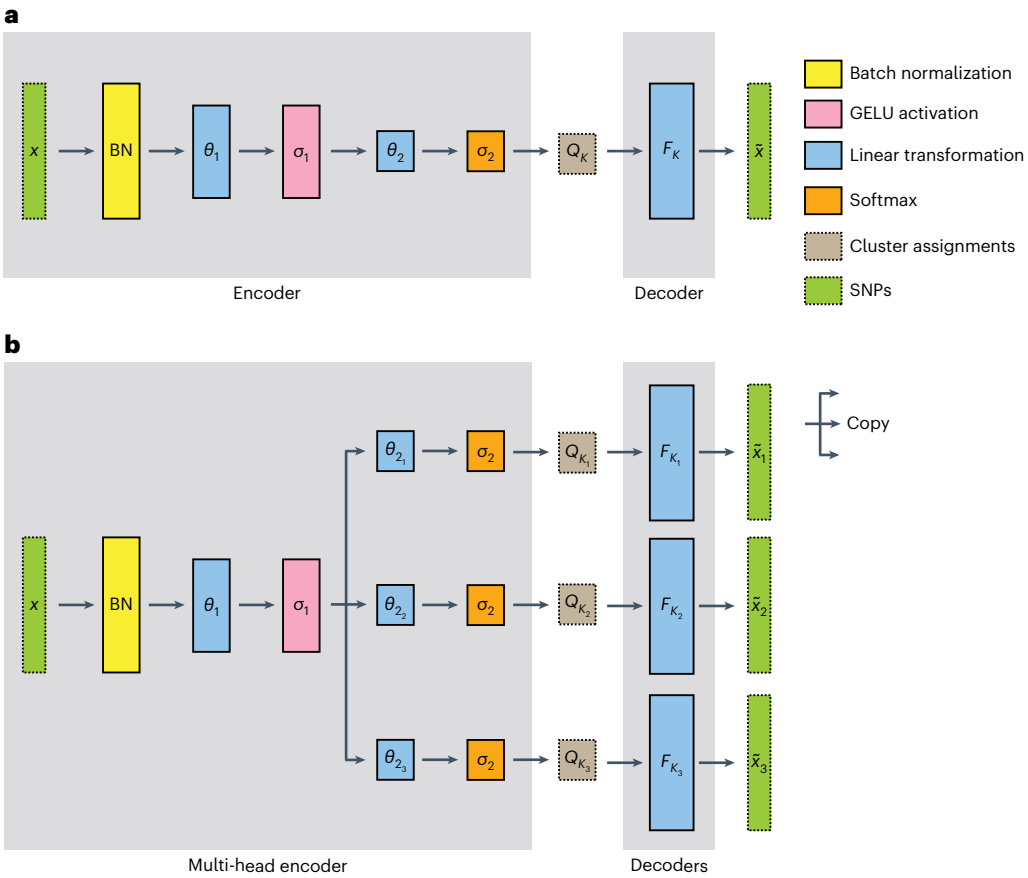


Fig. 1 | Neural ADMIXTURE model architecture. a, Single-head architecture. The input sequence (x) is projected into 64 dimensions using a linear layer (θ_1) and processed by a GELU non-linearity (σ_1). The cluster assignment estimates Q are computed by feeding the 64-dimensional sequence to a K-neuron layer (parametrized by θ_2) activated with a softmax (σ_2). Finally, the decoder outputs a reconstruction of the input ($\tilde{x}$) using a linear layer with weights F. Note that the decoder is restricted to this linear architecture to ensure interpretability. **b**, Simple multi-head example with $H = 3$. The 64-dimensional hidden vector is copied and processed independently by different sets of weights (θ_{2_h}), which yield vectors of different dimensions, corresponding to the different K values. Each different Q_{K_h} matrix is processed independently by different decoder matrices F_{K_h} yielding H different reconstructions. All parameters are optimized jointly in an end-to-end fashion.

We present an autoencoder that expands on the clustering method for genomes: ADMIXTURE[6,7]. ADMIXTURE was developed as a computationally efficient alternative to STRUCTURE[8], and we take this pursuit of efficiency now to the next generation of datasets. Our proposed method, Neural ADMIXTURE, follows the same modeling assumptions as ADMIXTURE, but reframes the task as a neural-network-based autoencoder, providing faster computational times, both on graphics and central graphics units (GPUs and on CPUs), while maintaining high-quality assignments.

Results

Model overview

Neural ADMIXTURE (Fig. 1a) is an interpretable autoencoder with two main components: (1) an encoder, composed of two linear layers with a Gaussian error linear unit (GELU) activation[9] in-between, then a softmax activation, which projects a genotype sequence onto a vector representing fractional ancestry assignments for each individual (Q); and (2) a decoder, which is a single linear layer whose weights are restricted to lie between 0 and 1, leading to an interpretable projection matrix that learns the cluster centroids, or equivalently, the average variant frequency at each site for each population (F). Additionally, we introduce Multi-head Neural ADMIXTURE (Fig. 1b), which includes multiple decoders in a single network to obtain results analogous to training ADMIXTURE repeatedly for different numbers of clusters, but needing only a single training for all numbers of clusters desired.

Neural ADMIXTURE was trained with a standard binary cross-entropy, leading to an equivalence with the traditional ADMIXTURE model's objective function (Methods). Two initialization techniques, one based on principal component analysis[10–12] and the other on archetypal analysis[13], were used as an alternative to common network initializations to speed up the training process and improve results (Supplementary section 'Decoder initialization'). Furthermore, two mechanisms are available to incorporate prior knowledge about the amount of admixture in a dataset by controlling the softness of the cluster assignments: applying L2 regularization during training (Methods) and softmax tempering (Supplementary section 'Softmax tempering'). Both single-head and multi-head approaches can be adapted to a supervised version that performs regular classification given known training labels (Supplementary section 'Supervised training'). The proposed method is fully compatible with the original ADMIXTURE framework, allowing the use of ADMIXTURE results as an initialization for Neural ADMIXTURE parameters (Supplementary section 'Pretrained mode'), and vice versa. We performed an in-depth evaluation of the proposed method and compared it with competing approaches across multiple datasets, including using simulations from a variety of systems[14–17] and using samples from large-scale, real-world biobanks (Methods, Supplementary Table 1, Supplementary Table 2, and Supplementary section 'Dataset description').

https://doi.org/10.1038/s43588-023-00482-7

Table 1 | Performance comparison of several global ancestry inference algorithms

Dataset	Algorithm	$\Delta(Q, Q_{GT})$	RMSE(Q, Q_{GT})	RMSE(F, F_{GT})	Runtime (CPU)	Runtime (GPU)
All-Chms	ADMIXTURE	0.042	0.153	0.062	>1day	–
	AlStructure	0.064	0.159	0.032	06:04:28	–
	TeraStructure	0.033	0.133	–	02:12:46	–
	HaploNet	0.026	0.114	–	–	03:17:00
	Neural ADMIXTURE	**0.025**	**0.108**	**0.011**	**00:11:21**	**00:01:32**
Chm-22	ADMIXTURE	0.048	0.161	0.068	02:56:29	–
	fastSTRUCTURE	0.055	0.162	–	03:31:00	–
	AlStructure	0.116	0.256	0.068	00:46:49	–
	TeraStructure	0.050	0.170	–	00:43:48	–
	HaploNet	0.053	0.170	–	–	01:09:29
	Neural ADMIXTURE	**0.033**	**0.140**	**0.016**	**00:05:46**	**00:00:45**
Chm-22-Sim	ADMIXTURE	0.046	0.197	0.067	09:48:18	–
	fastSTRUCTURE	0.069	0.237	–	>1day	–
	AlStructure	0.126	0.286	0.076	02:51:36	–
	TeraStructure	0.040	0.175	–	06:37:14	–
	HaploNet	0.026	0.113	–	–	02:07:54
	Neural ADMIXTURE	**0.011**	**0.070**	**6.02×10^{-3}**	**00:20:41**	**00:01:34**
PAB	ADMIXTURE	**1.44×10^{-4}**	**0.010**	**5.97×10^{-3}**	03:31:01	–
	AlStructure	1.45×10^{-3}	0.026	7.83×10^{-3}	05:10:42	–
	TeraStructure	1.97×10^{-4}	0.012	–	01:13:38	–
	HaploNet	0.039	0.248	–	–	02:37:09
	Neural ADMIXTURE	4.34×10^{-3}	0.055	7.01×10^{-3}	**00:14:27**	**00:01:48**
Synthetic	ADMIXTURE	**1.37×10^{-4}**	**0.011**	**0.028**	00:08:06	–
	AlStructure	2.74×10^{-4}	0.014	0.030	00:03:07	–
	TeraStructure	1.13×10^{-3}	0.032	–	00:03:28	–
	HaploNet	0.022	0.123	–	–	00:04:04
	Neural ADMIXTURE	8.60×10^{-4}	0.030	**0.028**	**00:01:25**	**00:00:12**

Metrics reported from the training data. Root mean squared error (RMSE) (F, F_{GT}), as defined in the Methods section, for fastSTRUCTURE, TeraStructure, and HaploNet was not computed because the first two lack an allele frequency matrix and the third lacks interpretability. HaploNet was not run on CPU because its resource and time requirements exceed system capabilities. Runtime format is HH:MM:SS and denotes wall-clock time. A runtime longer than a day denotes that the algorithm could not finish on the described hardware within 24 h, requiring it to be run on alternative hardware for longer. The best performing method for a given metric is highlighted in bold.

Single-head and multi-head results

Neural ADMIXTURE is systematically faster than alternative algorithms, both on CPU and GPU (Table 1, Supplementary Fig. 1). This speedup is further enhanced when using the Multi-head Neural ADMIXTURE architecture, which can perform clusterings for different K values simultaneously. For example, in the All-Chms dataset, we observed that Neural ADMIXTURE trained in less than 2 min, whereas ADMIXTURE required more than a day. Neural ADMIXTURE performs at least as well as existing algorithms on both predicting the ancestry assignments (Q) and the allele frequencies (F). On average, Neural ADMIXTURE's Q estimates appear to be more similar to the matrix of known labels than the Q estimates from previous methods (Extended Data Fig. 1).

Table 2 shows the accuracy and time performance of ADMIXTURE and Neural ADMIXTURE on the test data for three different datasets. Both ADMIXTURE and Neural ADMIXTURE are able to generalize and produce consistent assignments on unseen data. However, Neural ADMIXTURE is much faster than ADMIXTURE on both CPU and GPU, because ADMIXTURE must optimize the objective with a fixed F to find Q for unseen data, whereas Neural ADMIXTURE directly learns a function that estimates Q. We note that inference on GPU is extremely fast (generally less than a second for a forward pass); the computational bottleneck comes simply from reading and processing of the data, which could be further addressed.

We visualized the Q estimates of ADMIXTURE and Neural ADMIXTURE on the Chm-22-Sim dataset using pong[18] (Fig. 2a–d). The SNP frequencies (the entries in the F matrix) from both models can be observed as projections onto the first two principal components of the training data (Fig. 2e). Neural ADMIXTURE provides harder cluster predictions, with many samples being assigned only to a single population, whereas ADMIXTURE provides softer cluster predictions with partial assignments to multiple clusters. On this dataset, ADMIXTURE does not assign different clusters to Native Americans (AMR) and East Asians (EAS); instead, it partitions Africans (AFR) into two different ancestry clusters (Fig. 2a,b). Neural ADMIXTURE, however, does split AMR and EAS populations (Fig. 2c–e). Depictions of the cluster assignments (Q) of all algorithms on several datasets can be found in Supplementary Figs. 2–5.

We applied Neural ADMIXTURE, trained on Chm-22-Sim, to admixed populations that were not present in the training data: Mexican Ancestry in Los Angeles, California (MXL, 118), and Puerto Ricans in Puerto Rico (PUR, 104) (Fig. 2f).

We evaluated Multi-head Neural ADMIXTURE with Chm-22-Sim (Extended Data Fig. 2) and showed that as the number of clusters increases, each population group gets assigned its own cluster.

https://doi.org/10.1038/s43588-023-00482-7

Table 2 | Performance comparison of ADMIXTURE and Neural ADMIXTURE on test data

Dataset	Algorithm	$\Delta(Q, Q_{GT})$	RMSE(Q, Q_{GT})	Runtime (CPU)	Runtime (GPU)
Chm-22	ADMIXTURE	0.056	0.171	00:06:40	–
	Neural ADMIXTURE	**0.043**	**0.146**	**00:00:14**	**00:00:07**
Chm-22-Sim	ADMIXTURE	0.060	0.206	00:18:00	–
	Neural ADMIXTURE	**0.025**	**0.110**	**00:00:25**	**00:00:07**
PAB	ADMIXTURE	$\mathbf{4.29\times10^{-3}}$	**0.045**	00:10:26	–
	Neural ADMIXTURE	5.68×10^{-3}	0.062	**00:00:23**	**00:00:10**

ADMIXTURE results were computed using the Projection analysis mode, which reuses the *F* matrix computed during the fitting stage using the training data. Neural ADMIXTURE results were computed by simply feeding the sequences to the trained encoder, hence the extremely fast execution time. AlStructure, TeraStructure, and HaploNet lack the ability to compute ancestry assignments on data they were not trained on and so are not taken into account. Runtime format is HH:MM:SS and denotes wall-clock time. The best performing method for a given metric is highlighted in bold.

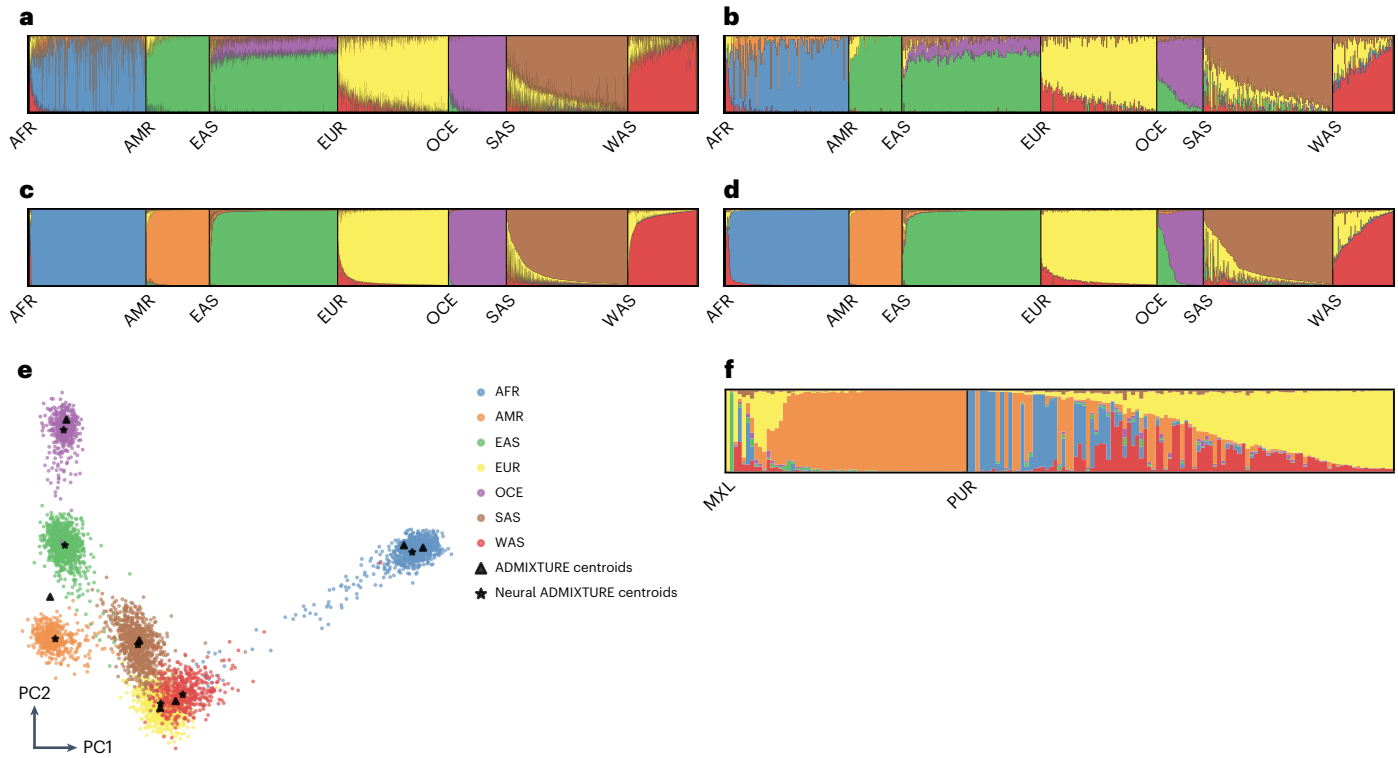


Fig. 2 | Visualization of several results of ADMIXTURE and Neural ADMIXTURE trained on the dataset Chm-22-Sim ($K = 7$). **a**, Q estimates of ADMIXTURE on training data. **b**, Q estimates of ADMIXTURE on test data. **c**, Q estimates of Neural ADMIXTURE on training data. **d**, Q estimates of Neural ADMIXTURE on test data. **e**, Two-dimensional principal component analysis (PCA) projection of the training data and the matrix F learnt by both ADMIXTURE and Neural ADMIXTURE, which correspond to the cluster centroids. The color of each individual in the PCA represents its ground truth regional label. **f**, Q estimates of Neural ADMIXTURE on admixed populations not present in the training data. Among the MXL samples, we observe mainly an orange AMR component with a red and yellow component (West Asians (WAS) and Europeans (EUR), respectively). These latter components likely originate from the immigration of Spanish, Morisco, and Sephardic Jewish individuals into Mexico during the colonial period. The PUR samples exhibit EUR, WAS, AMR, and AFR ancestry clusters. The additional AFR component is likely linked to the introduction of enslaved West Africans during the colonial period. In the barplots (used to visualize Q), each vertical bar represents an individual sample and bar color lengths represent the proportion of the sample's ancestry assigned to that colored cluster. OCE, Oceanians; SAS, South Asians.

Furthermore, we showed that Multi-head Neural ADMIXTURE can be successfully applied to closely related populations (Extended Data Fig. 3). Finally, we showed that the proposed method can be applied on real, admixed datasets (Extended Data Fig. 4).

UK Biobank computational analysis

To assess the clustering speed on a very large dataset, we ran Neural ADMIXTURE in its multi-head mode on the entire UK Biobank—a total of 488,377 samples—and using 147,604 SNPs subsetted to remove linkage disequilibrium (LD) by pruning the full set[19]. Neural ADMIXTURE was able to process the complete dataset within 11 h, providing results from $K = 2$ to $K = 6$, whereas ADMIXTURE would take about a month to do the same, given that it took 5.5 days to provide results for $K = 2$. Traditional techniques such as ADMIXTURE are thus too slow for such large biobanks, particularly because multiple additional runs with different parameters and subsets of data are generally needed in a study. Neural ADMIXTURE was trained without regularization ($\lambda = 0$, Methods) and using the PCK-means initialization (Supplementary Algorithm 1). During inference, the temperature was set to $\tau = \frac{3}{2}$ (Supplementary section 'Softmax tempering'). Figure 3 displays these cluster assignments for the UK Biobank genomes. We group the individuals by their reported country of birth; those with missing or non-existent country-of-birth labels were excluded from the plots.

https://doi.org/10.1038/s43588-023-00482-7

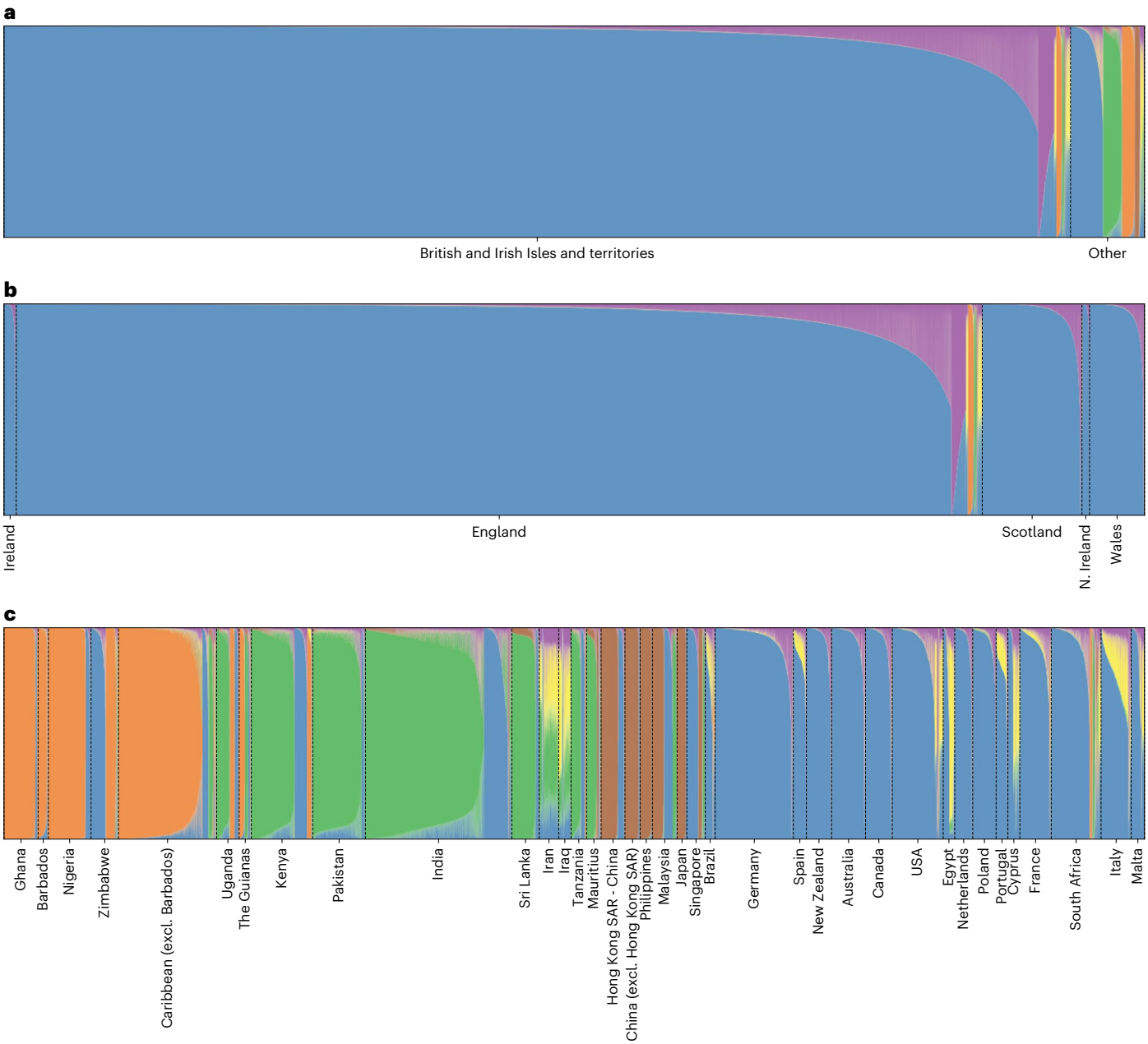


Fig. 3 | *Q* fractional genetic cluster estimates across the entire UK Biobank dataset (N = 488,377) obtained using Multi-head Neural ADMIXTURE (K = 6 displayed). Although results are only displayed for $K = 6$, the multi-head architecture was trained for $K = 2$ to $K = 6$ simultaneously in approximately 11 h. In the barplots (used to visualize Q), each vertical bar represents an individual sample and stacked bar color heights represent the proportion of the sample's ancestry assigned to that colored genetic cluster. Since they result from unsupervised clustering, interpretation of the cluster colors is left open. **a**, Q estimates of all the samples. Although many samples are clustered together (blue cluster, representing a northern European/British ancestry component), other clusters emerge reflecting the diverse modern populations now living within the United Kingdom. **b**, Q estimates of individuals born in the British and Irish Isles and territories. Samples from Gibraltar and the Channel Islands are excluded as they contain a very small number of individuals. **c**, Q estimates for individuals born outside of the British and Irish Isles are labeled by their country or region of birth, showcasing clusters representing Africans, East Asians, South Asians, Northern Europeans, and West Asians (sharing a cluster in part with Southern Europeans). Despite the large ancestry imbalance, Neural ADMIXTURE characterizes the globally diverse genetic variation found in the UK Biobank. Many UK residents born in other countries appear to have northern European (British) ancestry. These likely represent children born abroad to British parents, who later repatriated. We also note a sizeable South-Asian-like genetic ancestry cluster seen in many individuals born in East Africa. This likely stems from the decolonization era exodus out of East Africa of South Asians, who had settled there during the British Empire. The predicted cluster assignments for $K = 2$ to $K = 6$ for individuals born outside of the British and Irish Isles can be found in Extended Data Fig. 5.

Scalability analysis

To assess the scalability of different methods, we simulated multiple datasets with various numbers of variants and samples using the software reported previously[17]. The datasets consist of combinations of $N \in \{1{,}000, 5{,}000, 10{,}000, 20{,}000, 50{,}000\}$ and $M \in \{1{,}000, 10{,}000, 50{,}000, 100{,}000\}$, where N and M are the number of samples and SNPs, respectively.

We compared the training times of ADMIXTURE, AlStructure, TeraStructure, and Neural ADMIXTURE, both on CPU and GPU, across different dataset sizes (Fig. 4). Neural ADMIXTURE is consistently

https://doi.org/10.1038/s43588-023-00482-7

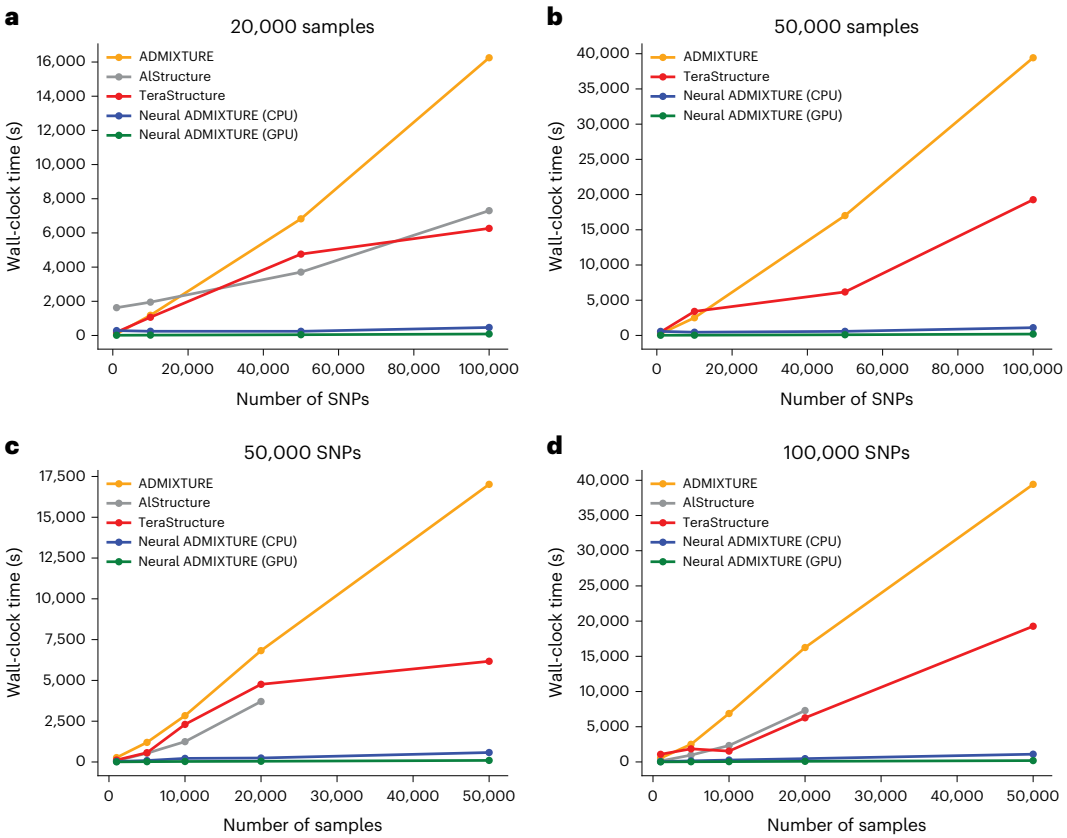


Fig. 4 | Evolution of execution time when increasing number of samples and number of variants. Neural ADMIXTURE has clearly faster execution times than the other benchmarked methods on both CPU and GPU. AlStructure results are not reported on the 50,000 samples because this method has prohibitively slow execution times.

faster than the alternatives. Moreover, Neural ADMIXTURE accelerates substantially using GPUs in contrast to the other methods. The hyperparameters used are described in Supplementary Table 3.

Discussion

Many unsupervised clustering methods for genotype sequences have been introduced[8,20–25] including the most commonly used, ADMIXTURE[6,7]. These methods, which resemble a non-negative matrix factorization, decompose each input sequence into a set of cluster assignments and compute a centroid for each cluster. The cluster assignments give the proportion of each genetic ancestry cluster for an individual, whereas the cluster centroids give the SNP variant frequencies at each genetic position corresponding to each cluster. As a diploid organism, most humans have a paternal and maternal copy of each non-sex chromosome. Therefore, for a given individual at each genomic position, we have the possibility of four different combinations of biallelic SNPs (0/0, 0/1, 1/0, 1/1). It is common practice to sum both maternal and paternal variants, obtaining a count sequence n_{ij}. In this scenario, an individual i has $n_{ij} \in \{0, 1, 2\}$ copies of the minority SNP j. ADMIXTURE models each individual's count sequence, given a fixed number of population groups K, as $n_{ij} \sim \text{Bin}(2, p_{ij})$, where $p_{ij} = \sum_k q_{ik} f_{kj}$, with q_{ik} denoting the fraction of population k assigned to i, and f_{kj} denoting the frequency of SNPs with a value of '1' j in population k. ADMIXTURE applies block relaxation to find the parameters Q and F that minimize the negative log-likelihood function shown in equation (1). The value of K (number of clusters) is typically chosen by using an ad hoc cross-validation procedure[7], necessitating runs across a range of values.

The block relaxation optimization in ADMIXTURE runs much faster than other approaches used by its main competitors, namely FRAPPE[21] and STRUCTURE[8]. Although it can be run in multi-threading mode, greatly boosting the execution time, it is insufficient when dealing with either a large number of samples or a large number of SNPs. Here we instead use neural networks, whose architectures have begun to be explored for several other genetic structure tasks including haplotype segmentation, dimensionality reduction, and classification[26–35] (Supplementary section 'Related work').

An important caveat when using soft-clustering techniques, such as Neural ADMIXTURE or ADMIXTURE, is that these techniques follow a modeling assumption that there are some 'prototype' populations and that each individual can be placed within the convex hull of such prototypes. Note that this model might not reflect the underlying structure of real-world populations particularly when independent genetic drift has occurred in each population following admixture events. This limitation is particularly acute in the case of ancient admixture events, and in such cases, other complementary techniques should also be used. Future experiments to quantify these effects using simulations would be valuable. Combining unsupervised clustering with tree-based methods to account for this drift would also be a useful direction. This could complement the progress being made in ancestral recombination graphs.

Although the computational times of Neural ADMIXTURE enable practitioners to obtain rapid results with multiple hyperparameters and different values of K, properly selecting the best results still involves a subjective element, and additional experiments and new quantitative measures are needed. Further, unsupervised clustering methods, and more generally dimensionality-reduction techniques, are affected by sampling imbalances between population groups, which can alter population structure detection and prioritization[36,37]. Additionally, even if structure is not present within the data, these techniques can indicate otherwise[38,39].

Methods

Single-head Neural ADMIXTURE

As described in the Discussion, the existing ADMIXTURE algorithm minimizes the negative log-likelihood:

$$\begin{aligned} \min_{Q,F} \quad & \mathcal{L}_{\mathrm{C}}(Q,F) = -\sum_{i,j} n_{ij} \log\left(\sum_k q_{ik} f_{kj}\right) + (2-n_{ij}) \log\left(1-\sum_k q_{ik} f_{kj}\right) \\ \text{subject to} \quad & 0 \le f_{kj} \le 1 \\ & \sum_k q_{ik} = 1 \\ & q_{ik} \ge 0 \end{aligned} \tag{1}$$

with $Q = (q_{ik})$ and $F = (f_{kj})$.

This can be formulated as a non-negative matrix factorization problem. Let X denote the training samples, where the features are the alternate allele normalized counts per position and the jth SNP of the ith individual is represented as $x_{ij} = \frac{n_{ij}}{2} \in \{0, 0.5, 1\}$. Then, $X \approx QF$, where Q is the assignments, F is the alternate allele frequencies per SNP and population, and the negative log-likelihood in equation (1) is a distance between X and QF. This can be translated into a neural network as an autoencoder with $Q = \Psi(X)$ being the bottleneck computed by the encoder function Ψ and F being the decoder weights themselves (Fig. 1a). Because Q is estimated at every forward pass and not learnt as a whole for the training data, to retrieve Q assignments on previously unseen data, we can perform a simple forward pass instead of running the optimization process fixing F, unlike with ADMIXTURE.

Note that the restrictions in the optimization problem (equation (1)) impose restrictions in the architecture. Those relating to Q ($\sum_k q_{ik} = 1$ and $q_{ik} \ge 0$) can be enforced by applying a softmax activation at the encoder output, making the bottleneck equivalent to the cluster assignments. Although the decoder restriction ($0 \le f_{kj} \le 1$) could be enforced by applying the sigmoid function to the decoder weights, we found that it suffices to project the weights of the decoder to the interval $[0, 1]$ after every optimization step, one of the most common forms of projected gradient descent[40].

The decoder must be linear and cannot be followed by a non-linearity, as this would break the interpretability of the F matrix; the equivalence between the decoder weights and cluster centroids would be lost. On the other hand, the encoder architecture is free from constraints, and it may be composed of several layers. The proposed architecture includes a 64-dimensional, non-linear layer with a GELU activation before the bottleneck and batch normalization acting directly on the input. The latter re-scales the data to have zero mean and unit variance. Since the mean for each SNP is its frequency p, and the standard deviation σ is $\sqrt{p(1-p)}$, the $\{0, 1\}$ input gets encoded as $\left\{-\sqrt{\frac{p}{1-p}}, \sqrt{\frac{1-p}{p}}\right\}$, thereby supplying more explicitly the information of the allele frequencies to the network.

The ADMIXTURE model does not precisely reconstruct the input data as a regular autoencoder would do, because the input SNP genotype sequences, $n_{ij} \in \{0, 1, 2\}$, and the reconstructions, $p_{ij} \in [0, 1]$, do not have matching ranges. This can easily be remedied by dividing the genotype counts by two, so that the input data are $x_{ij} = \frac{n_{ij}}{2} \in \{0, 0.5, 1\}$. Moreover, instead of minimizing $\mathcal{L}_{\mathrm{C}}$ (equation (1)), we propose minimizing the binary cross-entropy instead, using a penalty term on the Frobenius norm of the encoder weights, θ:

$$\mathcal{L}_{\mathrm{N}}(Q,F) = -\sum_{i,j} x_{ij} \log\left(\sum_k q_{ik} f_{kj}\right) + (1-x_{ij}) \log\left(1-\sum_k q_{ik} f_{kj}\right) + \lambda \|\theta\|_F^2. \tag{2}$$

This regularization term avoids hard assignments in the bottleneck, which helps during the training process and reduces overfitting. In equation (3) we show that the proposed optimization problem and the ADMIXTURE one are equivalent (excluding the regularization term) by using equations (1) and (2):

$$\begin{aligned} \mathcal{L}_{\mathrm{N}}^{\lambda=0}(Q,F) &= -\sum_{i,j} x_{ij} \log\left(\sum_k q_{ik} f_{kj}\right) + (1-x_{ij}) \log\left(1-\sum_k q_{ik} f_{kj}\right) \\ &= -\sum_{i,j} \tfrac{n_{ij}}{2} \log\left(\sum_k q_{ik} f_{kj}\right) + \left(1-\tfrac{n_{ij}}{2}\right) \log\left(1-\sum_k q_{ik} f_{kj}\right) = \\ &= -\tfrac{1}{2} \sum_{i,j} n_{ij} \log\left(\sum_k q_{ik} f_{kj}\right) + (2-n_{ij}) \log\left(1-\sum_k q_{ik} f_{kj}\right) = \\ &= \tfrac{1}{2} \mathcal{L}_{\mathrm{C}}(Q,F). \end{aligned} \tag{3}$$

A perfect reconstruction can of course be obtained by setting the number of clusters (K) equal to the number of training samples or to the dimension of the input (number of SNPs). However, the bottleneck should ideally capture elementary information about the population structure of the given sequences; therefore, we make use of low-dimensional bottlenecks.

Multi-head Neural ADMIXTURE

In ADMIXTURE, cross-validation must be performed to choose the number of population clusters (K), unless specific prior information about the number of population ancestries is known. Furthermore, in many applications, practitioners desire to observe how cluster assignments change as the number of clusters increases. As the number of both sequenced individuals and variants increases, the feasible number of different cluster numbers that can be run for cross-validation rapidly decreases due to the additional computational cost. As a solution, Multi-head Neural ADMIXTURE allows all cluster numbers to be run simultaneously by taking advantage of the 64-dimensional latent representation computed by the encoder. This shared representation is jointly learnt for the different values of K, $\{K_1, \ldots, K_H\}$.

Figure 1b shows how the shared representation is split into H different heads in the multi-head architecture. The ith head consists of a non-linear projection to a K_i-dimensional vector, which corresponds to an assignment that assumes there are K_i different genetic clusters in the data. Although every head could be concatenated and fed through a decoder, this would cause the decoder weights F to not be interpretable. Therefore, every head needs to have its own decoder and, thus, H different reconstructions of the input are retrieved.

As we have H reconstructions, we will now have H different loss values. We can train this architecture by minimizing equation (4):

$$\mathcal{L}_{\mathrm{MNA}}(Q_{K_{1,\ldots,H}}, F_{K_{1,\ldots,H}}) = \sum_{h=1}^{H} \mathcal{L}_{\mathrm{N}}\left(Q_{K_h}, F_{K_h}\right), \tag{4}$$

where Q_{K_h} and F_{K_h} are, respectively, the cluster assignments and the SNP frequencies per population for the hth head. The restrictions of the ADMIXTURE optimization problem (equation (1)) must be satisfied by Q_{K_h} and F_{K_h} $\forall h \in \{1, \ldots, H\}$.

The multi-head architecture allows computation of H different cluster assignments corresponding to H different values for K, efficiently, in a single forward pass. Results can then be quantitatively and qualitatively analyzed by the practitioner to decide which value of K is the most suitable for the data.

Evaluation setup

Let N denote the number of samples and M the number of variants (SNPs). To assess the performance of the Q estimates, we match the assignments with the known labels and report the RMSE between them,

$$\mathrm{RMSE}(Q, Q_{\mathrm{GT}}) = \frac{1}{\sqrt{NK}} \|Q - Q_{\mathrm{GT}}\|_F \tag{5}$$

and the RMSE between the known allele frequencies (F_{GT}) and the estimated frequencies (F),

$$\mathrm{RMSE}(F, F_{\mathrm{GT}}) = \frac{1}{\sqrt{KM}} ||F - F_{\mathrm{GT}}||_F \quad (6)$$

We also use a new metric, Δ, defined as

$$\Delta(Q, Q_{\mathrm{GT}}) = \frac{1}{N^2} \left\| QQ^{\mathrm{T}} - Q_{\mathrm{GT}} Q_{\mathrm{GT}}^{\mathrm{T}} \right\|_F^2, \quad (7)$$

which is equivalent to the mean squared difference between the covariance matrices of the estimated and the target populations. In case the Q estimates completely agree with Q_{GT} (up to permutation), Δ will be zero. The larger the disagreement, the higher the value of Δ. We are interested in these metrics, as they are more easily interpreted than the loss function value itself. We are aware that these pseudo-supervised metrics, when applied to datasets simulated from real individuals, do not yield the true quality of the predictions of the models, since the biogeographic labels assigned to the real sequences used to simulate datasets might not reflect the true genomics clusters and variation within the populations. To further investigate this issue, we also used fully simulated population clusters to evaluate the methods.

Dataset preparation. For reproducibility we have used a comprehensive set of publicly available, labeled human whole-genome sequences from diverse populations across the world, combining the 1000 Genomes Project[41], the Simons Genome Diversity Project[42], and the Human Genome Diversity Project[43], as well as data simulated from these samples using PyAdmix[14] and data simulated de novo using the Balding–Nichols Pritchard–Stephens–Donnely model[8,23]. The populations within the combined real datasets can be found in Supplementary Table 2. Each subpopulation is aggregated into a continental-level label according to its geographical location (Supplementary section 'Dataset description'). Additionally, we used the entire UK Biobank genotype dataset.

Benchmarking setup. We compared Neural ADMIXTURE computational time and clustering quality with ADMIXTURE, fastSTRUCTURE[24], AlStructure[22], and TeraStructure[23]. fastSTRUCTURE assumes the STRUCTURE model but uses accelerated variational methods instead of MCMC, yielding speedups of more than two orders of magnitude against STRUCTURE. TeraStructure iteratively computes Q and F while avoiding a high computational load by subsampling SNPs at every iteration, which makes the algorithm faster. AlStructure first estimates a low-dimensional linear subspace of the admixture components and then searches for a model in the latter subspace that satisfies the modeling constraints, yielding a fast alternative to the iterative or maximum likelihood schemes followed by most algorithms. Furthermore, we also compared against HaploNet[26], a variational autoencoder that maps parts of the sequence (windows) to a low-dimensional latent space, on which clustering is then performed using Gaussian mixture priors. Although the global structure of the data is preserved in the low-dimensional space, direct interpretability of the allele frequencies (available in Neural ADMIXTURE) is not preserved.

All models were optimized using 16 threads on an AMD EPYC 7742 (x86_64) processor, which consists of 64 cores and 512 GB of RAM. We restricted the number of threads to 16 despite the fact that more cores are available to run several executions in parallel. To assess GPU performance of Neural ADMIXTURE, all networks were trained on an NVIDIA Tesla V100 SXM2 of 32 GB. The same GPUs were used to run inference on the trained models.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The samples used in the 'Experiments' section were compiled from public datasets: 1000 Genomes Project (https://www.internationalgenome.org/data/)[41], the Simons Genome Diversity Project (https://www.simonsfoundation.org/simons-genome-diversity-project/)[42], and the Human Genome Diversity Project (https://www.internationalgenome.org/data-portal/data-collection/hgdp)[43]. The compiled datasets (All-Chms, Chm-22 and Chm-22-Sim) are available on figshare[44]. The UK Biobank has approval from the North West Multi-centre Research Ethics Committee as a Research Tissue Bank. This dataset is available to researchers through an open application via https://www.ukbiobank.ac.uk/register-apply/. The entire dataset of genotypes available to download from the UK Biobank portal were used. Source data are provided with this paper.

Code availability

The software is available as an installable package in the PyPi repository under the name 'neural-admixture'. The source code can be found at https://github.com/ai-sandbox/neural-admixture ref. 45.

References

1. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
2. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
3. Privé, F. Using the UK Biobank as a global reference of worldwide populations: application to measuring ancestry diversity from GWAS summary statistics. *Bioinformatics* **38**, 3477–3480 (2022).
4. Morales, J. et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 1–10 (2018).
5. Mathieson, I. & Scally, A. What is ancestry? *PLoS Genet.* **16**, e1008624 (2020).
6. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
7. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinform.* **12**, 246 (2011).
8. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
9. Hendrycks, D. & Gimpel, K. Gaussian error linear units (GELUs). Preprint at https://doi.org/10.48550/arXiv.1606.08415 (2020).
10. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
11. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
12. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
13. Cutler, A. & Breiman, L. Archetypal analysis. *Technometrics* **36**, 338–347 (1994).
14. Kumar, A., Montserrat, D. M., Bustamante, C. & Ioannidis, A. XGMix: local-ancestry inference with stacked XGBoost. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.21.053876 (2020).
15. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
16. Karavani, E. et al. Screening human embryos for polygenic traits has limited utility. *Cell* **179**, 1424–1435.e8 (2019).
17. Chiu, A., Molloy, E., Tan, Z., Talwalkar, A. & Sankararaman, S. Inferring population structure in biobank-scale genomic data. *Am. J. Hum. Genet.* **109**, 727–737 (2022).
18. Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P. & Ramachandran, S. Pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32**, 2817–2823 (2016).

19. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
20. Bradburd, G. S., Coop, G. M. & Ralph, P. L. Inferring continuous and discrete population genetic structure across space. *Genetics* **210**, 33–52 (2018).
21. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**, 289–301 (2005).
22. Cabreros, I. & Storey, J. D. A likelihood-free estimator of population structure bridging admixture models and principal components analysis. *Genetics* **212**, 1009–1029 (2019).
23. Gopalan, P., Hao, W., Blei, D. & Storey, J. Scaling probabilistic models of genetic variation to millions of humans. *Nat. Genet.* **48**, 1587–1590 (2016).
24. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
25. Gimbernat-Mayol, J., Dominguez Mantes, A., Bustamante, C. D., Mas Montserrat, D. & Ioannidis, A. G. Archetypal analysis for population genetics. *PLoS Comput. Biol.* **18**, e1010301 (2022).
26. Meisner, J. & Albrechtsen, A. Haplotype and population structure inference using neural networks in whole-genome sequencing data. *Genome Res.* **32**, 1542–1552 (2022).
27. Joo, W., Lee, W., Park, S. & Moon, I.-C. Dirichlet variational autoencoder. *Pattern Recognit.* **107**, 107514 (2020).
28. Keller, S. M., Samarin, M., Torres, F. A., Wieser, M. & Roth, V. Learning extremal representations with deep archetypal analysis. *Int. J. Comput. Vis.* **129**, 805–820 (2021).
29. Ausmees, K. & Nettelblad, C. A deep learning framework for characterization of genotype data. *G3* **12**, jkac020 (2022).
30. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* **36**, 3418–3421 (2020).
31. Battey, C., Coffing, G. C. & Kern, A. D. Visualizing population structure with variational autoencoders. *G3* **11**, jkaa036 (2021).
32. Montserrat, D. M., Bustamante, C. & Ioannidis, A. LAI-Net: local-ancestry inference with neural networks. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing* 1314–1318 (IEEE, 2020).
33. Oriol Sabat, B., Mas Montserrat, D., Giro-i Nieto, X. & Ioannidis, A. G. SALAI-Net: species-agnostic local ancestry inference network. *Bioinformatics* **38**, ii27–ii33 (2022).
34. Romero, A. et al. Diet networks: thin parameters for fat genomics. In *5th International Conference on Learning Representations* (OpenReview.net, 2017).
35. Battey, C. J., Ralph, P. L. & Kern, A. D. Predicting geographic location from genetic variation with deep neural networks. *eLife* **9**, e54507 (2020).
36. Toyama, K. S., Crochet, P.-A. & Leblois, R. Sampling schemes and drift can bias admixture proportions inferred by structure. *Mol. Ecol. Resour.* **20**, 1769–1785 (2020).
37. Elhaik, E. Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Sci. Rep.* **12**, 14683 (2022).
38. Chari, T., Banerjee, J. & Pachter, L. The specious art of single-cell genomics. Preprint at *bioRxiv* https://doi.org/10.1101/2021.08.25.457696 (2021).
39. Montserrat, D. M. & Ioannidis, A. G. Adversarial attacks on genotype sequences. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2023).
40. Lin, C.-J. Projected gradient methods for nonnegative matrix factorization. *Neural Comput.* **19**, 2756–2779 (2007).
41. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
42. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
43. Bergström, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
44. Dominguez Mantes, A. et al. Neural ADMIXTURE - datasets. *figshare* https://doi.org/10.6084/m9.figshare.19387538.v1 (2022).
45. Dominguez Mantes, A., Ioannidis, A. G. & Montserrat, D. M. AI-sandbox/neural-admixture: stable release. *Zenodo* https://doi.org/10.5281/zenodo.7938892 (2023).

Acknowledgements

This work was partially supported by a grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI), NIH grants 7U01HG009080 and R01HG010140, and project PID2020-117142GB-I00 funded by MCIN/AEI /10.13039/501100011033. This research was conducted using the UK Biobank Resource under Application Number 89006.

Author contributions

A.G.I. and D.M.M. designed the research. A.D.M. performed the research and wrote the software. A.D.M., D.M.M., X.G.N., and A.G.I. interpreted the results. C.D.B. contributed data. A.D.M., D.M.M., and A.G.I. wrote the manuscript.

Competing interests

C.D.B. is the chief executive officer of Galatea Bio, and A.G.I. also holds shares. The remaining authors declare no competing interests.

Additional information

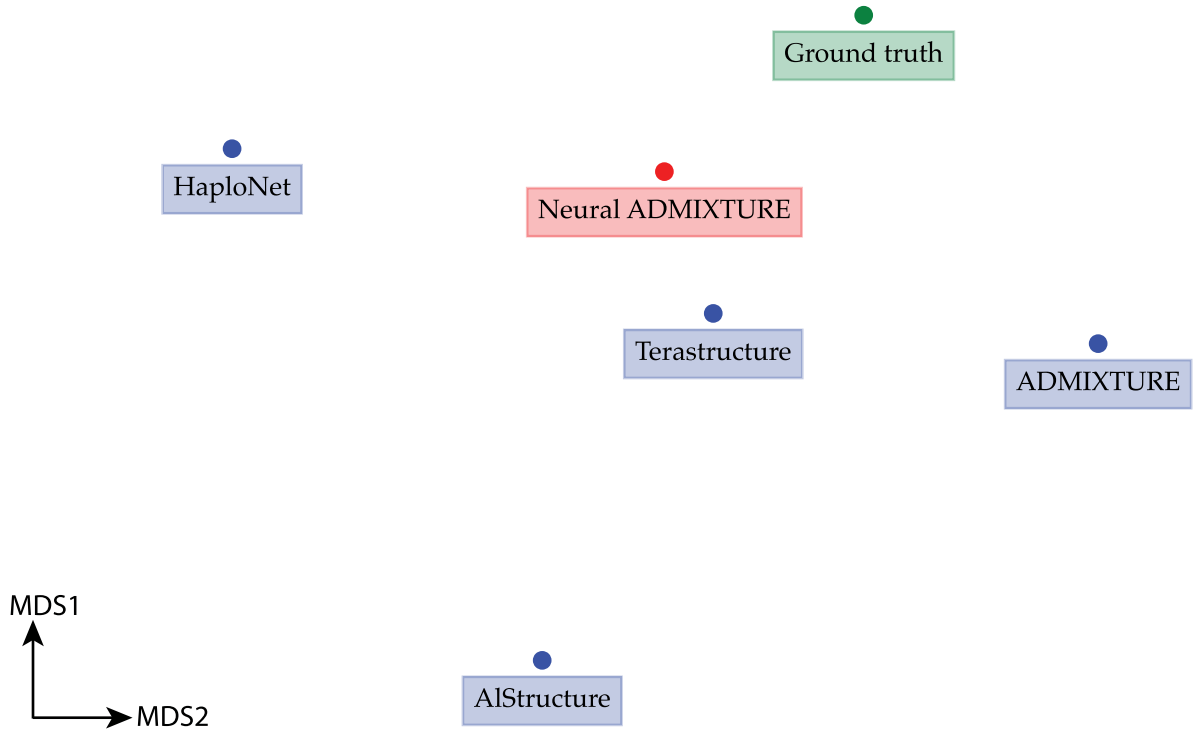
Extended data is available for this paper at https://doi.org/10.1038/s43588-023-00482-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-023-00482-7.

Correspondence and requests for materials should be addressed to Alexander G. Ioannidis.

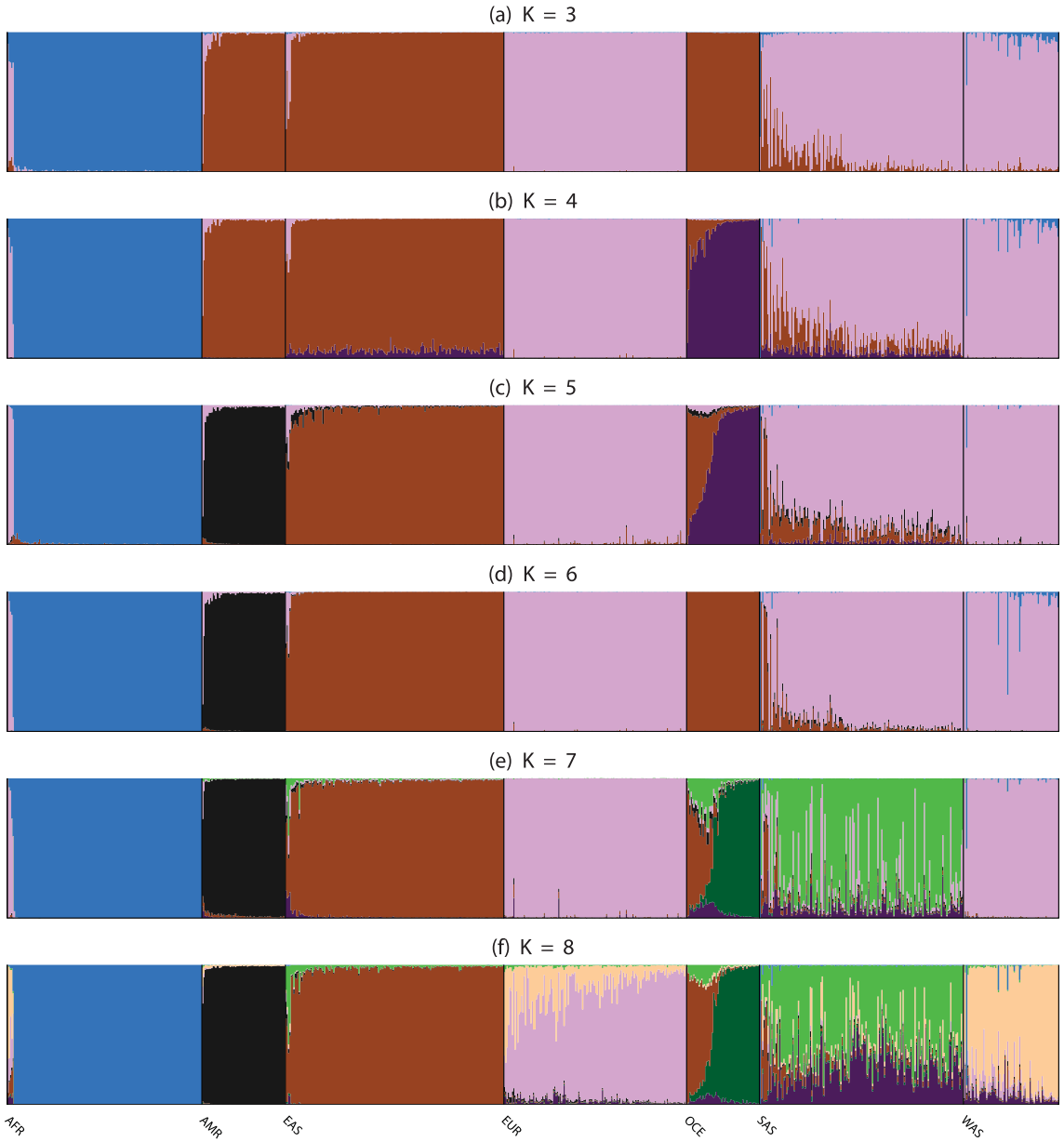
Peer review information *Nature Computational Science* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Ananya Rastogi, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

https://doi.org/10.1038/s43588-023-00482-7



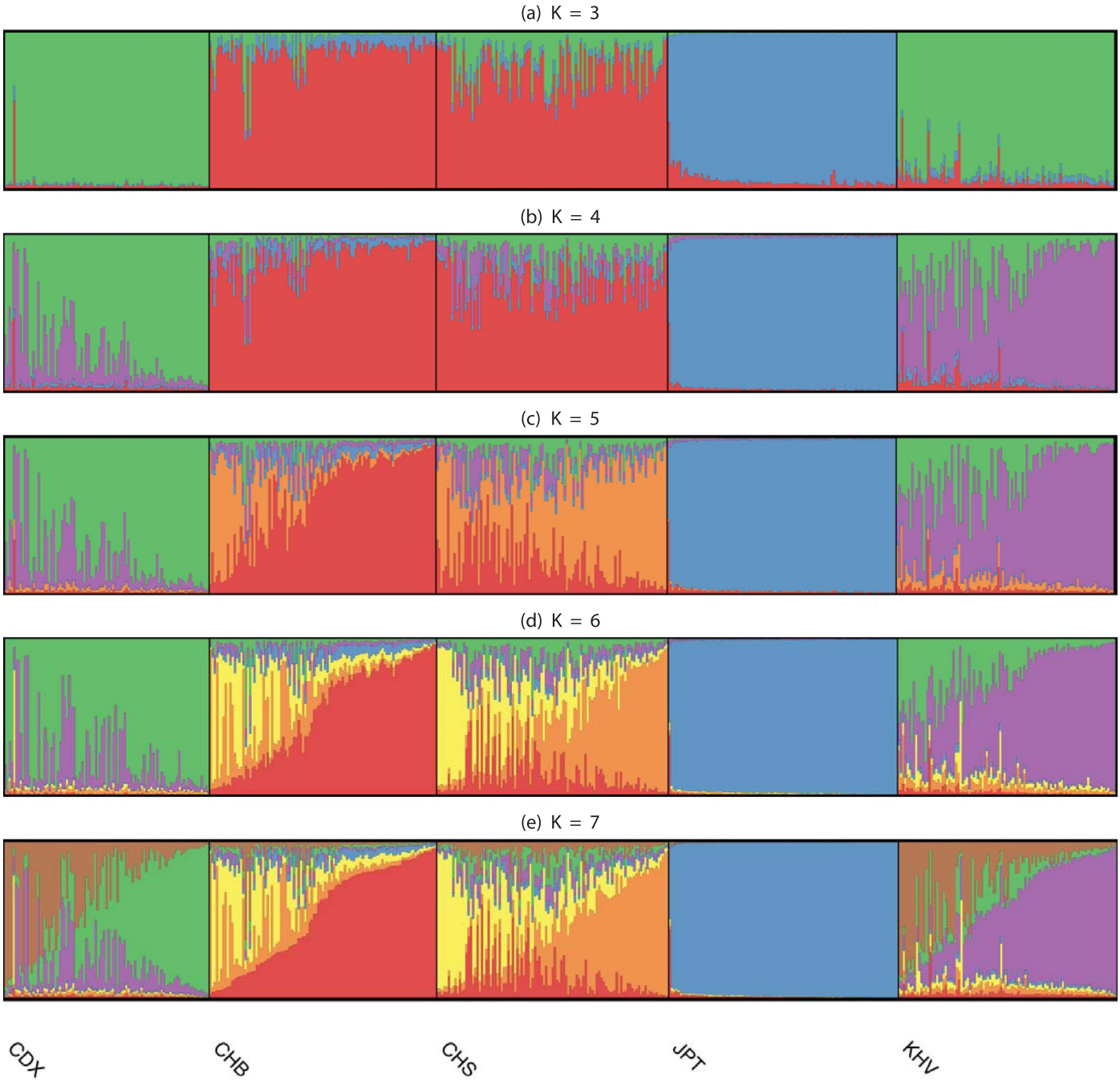
Extended Data Fig. 1 | 2D visualization of Q estimates using multidimensional scaling (MDS) Algorithms appearing closer in the MDS projection have more similar estimates than those farther away. In order to use MDS, a distance matrix of the Q results of different algorithms (including the ground truth matrix) has been computed by using the Frobenius norm between the different Q matrices. The average of the normalized distances has been taken across all datasets in order to retrieve a single distance matrix.

https://doi.org/10.1038/s43588-023-00482-7



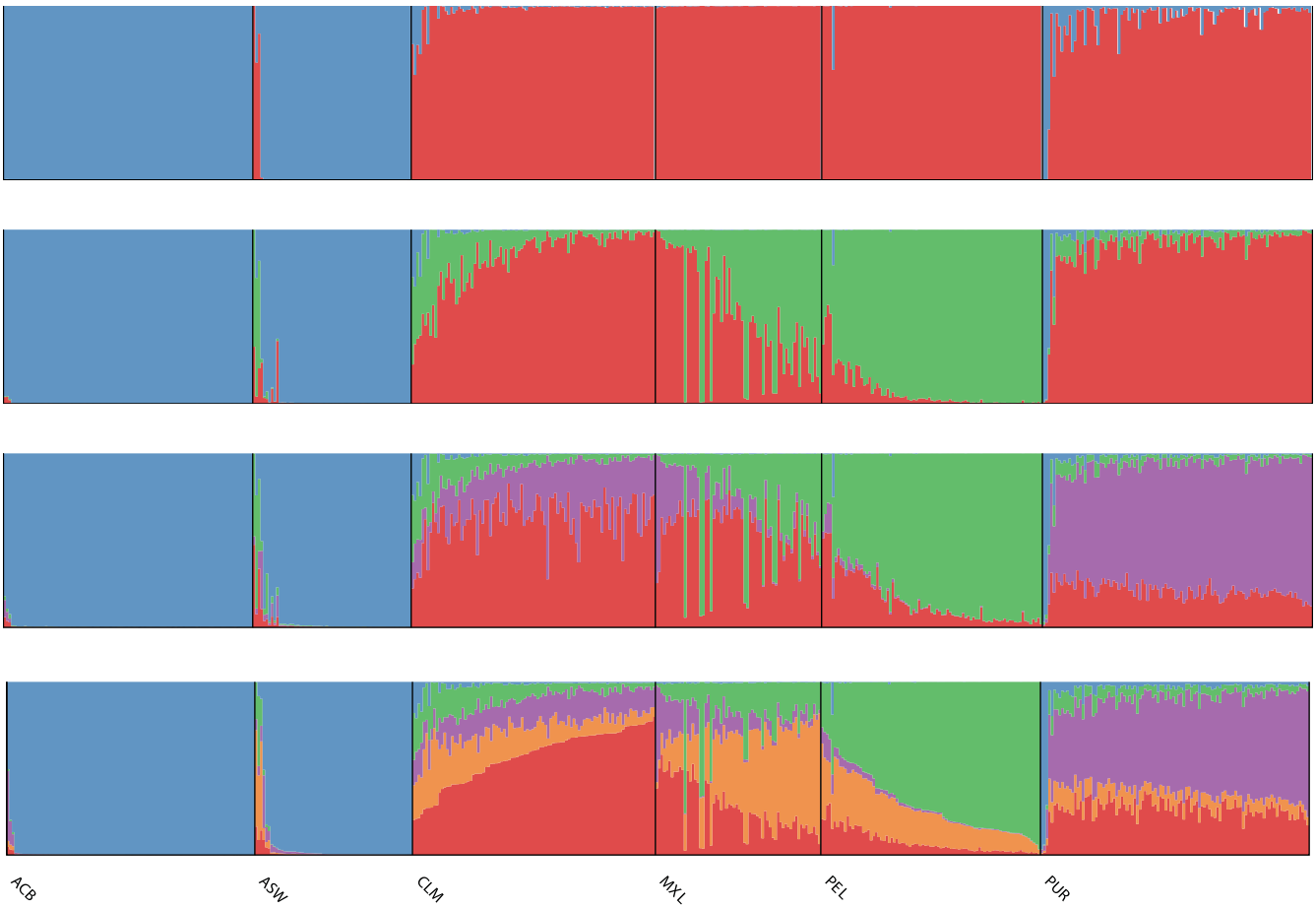
Extended Data Fig. 2 | Results from Multi-head Neural ADMIXTURE (K=3 to K=8) on the test set of Chm-22-Sim For K=3, European (EUR), West Asian (WAS) and South Asian (SAS) are combined within the same cluster, while American (AMR), Oceanian (OCE), and East Asian (EAS) are clustered together, and African (AFR) has its own cluster. These results reflect the genetic similarity between the respective groups due to their Out-of-Africa migration patterns and subsequent gene flow. After increasing to K=5, OCE obtains its own cluster, reflecting the ancient divergence from the others of that population consisting in our study of the Australo-Papuan groups-Native Australian (SGDP), Papuan Highlands (HGDP), Papuan Sepik (HGDP), Bougainville (HGDP), and Dusun (HGDP). As more clusters are incorporated, American (AMR) and EAS obtain their own clusters and OCE is divided between a component found predominantly in OCE and a component characteristic of EAS. The latter likely reflects the later migration of Austronesian speakers from East Asia out into the Pacific Islands, where they contributed their ancestry to the Oceanian inhabitants. A shared component between EUR, SAS and WAS is maintained, independent of the cluster number K. This could be linked to early farmer expansions out of West Asia and into both Europe and South Asia following the birth of agriculture, or to the much later expansion of the Indo-European language family across all of these regions. Other genetic exchanges between these neighboring regions doubtlessly played a role. With a sufficiently high number of clusters, a shared component between WAS and some AFR populations appears, perhaps reflecting North African gene flow.

https://doi.org/10.1038/s43588-023-00482-7



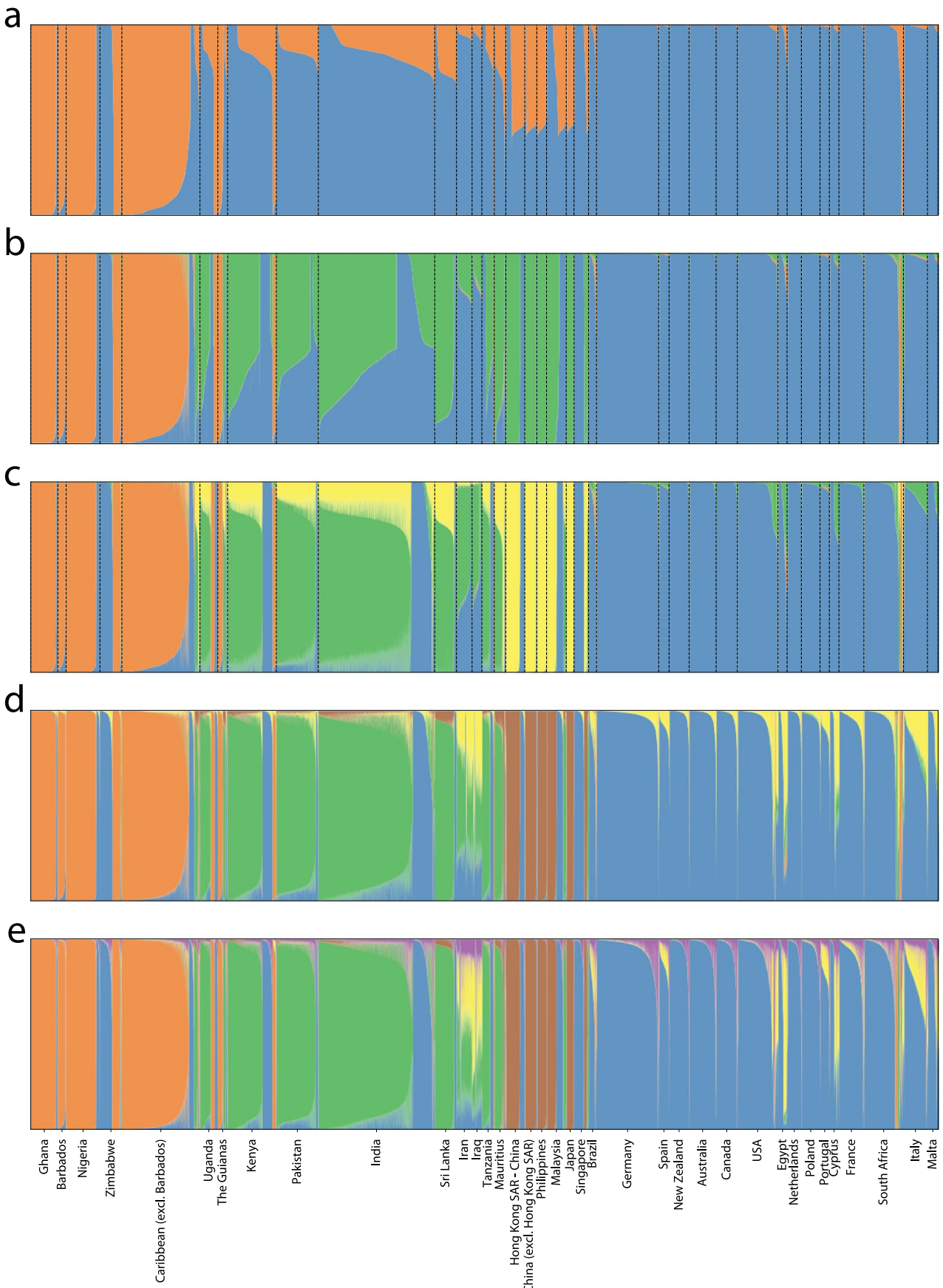
Extended Data Fig. 3 | Multi-head Neural ADMIXTURE results on a dataset consisting of closely related groups. To qualitatively assess the performance of Neural ADMIXTURE on related groups, we ran multi-head Neural ADMIXTURE on a subset of the dataset All-Chms containing 504 East Asian (EAS) individuals from neighboring regions. The self-reported ancestry of these individuals are Chinese Dai in Xishuangbanna, China (CDX, 93), Han Chinese in Beijing, China (CHB, 103), Han Chinese South (CHS, 105), Japanese in Tokyo, Japan (JPT, 104) and Kinh in Ho Chi Minh City, Vietnam (KHV, 99). The network was trained in its multi-head version from K=3 to K=7 using the PCK-Means initialization. The Japanese samples (JPT) are differentiated and clearly assigned their own cluster (blue), which is present only marginally in other populations. CDX (Chinese Dai) and KHV (Vietnamese Kinh) initially share the same cluster (K=3, green), reflecting their common Southeast Asian lineage, but are split into different groups at K=4 (purple and green). As expected CHB (Han Chinese in Beijing) and CHS (Han Chinese from South China) samples share the same cluster at first (red) and are only differentiated last (at K=5, red and orange). Further structure (yellow and brown) is seen within some populations at higher K.

https://doi.org/10.1038/s43588-023-00482-7



Extended Data Fig. 4 | Q estimates of multi-head Neural ADMIXTURE on a dataset consisting of only admixed samples. To assess performance of the model using real admixed samples, we have trained a multi-head Neural ADMIXTURE model (from K=2 to K=5) with samples whose self-reported ancestry are African Caribbean in Barbados (ACB, 96), African Ancestry in Southwest US (ASW, 61), Colombian in Medellin, Colombia (CLM, 94), Mexican Ancestry in Los Angeles, California (MXL, 64), Peruvian in Lima, Peru (PEL, 85) and Puerto Rican in Puerto Rico (PUR 104). The groups have been selected from the 1000 Genomes Project. The variants used (839629) are the same as in the dataset All-Chms. The network was trained using the PCK-Means initialization (Supplementary Text 'Decoder initialization'). At K=2, ACB and ASW are assigned predominantly to their own cluster, separating their mostly African origins from the remaining out-of-Africa components. When introducing the next new cluster (K=3), admixed individuals in CLM, MXL and PEL are assigned some fraction to it, differentiating an Indigenous American component in them from their European component. At K=4 the individuals in the PUR population are assigned some fraction of the new cluster, and this cluster is also present in small amounts in CLM and smaller amounts in some MXL. This component, which does not decrease the Indigenous American component fraction in the samples, likely represents an early colonial-era Spanish (European-ancestry) founder effect on the island of Puerto Rico perhaps reflecting the subsequent early colonial expansion from the Spanish Caribbean to coastal Colombia and Mexico. Structure in the European component appears at K=5.

https://doi.org/10.1038/s43588-023-00482-7



Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s43588-023-00482-7

Extended Data Fig. 5 | Cluster assignments computed by Neural ADMIXTURE for individuals born outside the British and Irish Isles in the UK Biobank training data. **(a)** K=2 **(b)** K=3 **(c)** K=4 **(d)** K=5 **(e)** K=6. Because the majority of the dataset is composed of individuals with white British ancestry, we only plot the cluster assignments of individuals that reported a country-of-birth outside British and Irish Isles. We can observe that K=2 approximately divides samples between European and non-European populations. With K=3 European, South-and-East Asian, and African ancestry clusters emerge. When K=4 a fine-grained clustering emerges dividing East and South Asian populations. K=5 adds a fifth cluster shared in common (with different proportions) between Southern European (Mediterranean) and West Asian (Near Eastern) populations. Finally, K=6 seems to introduce a cluster mostly present in Northern and Eastern European populations.

nature portfolio

Corresponding author(s): Alexander Ioannidis

Last updated by author(s): 05/11/2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	The software is available as an installable package in the PyPi repository under the name neural-admixture. The source code from this paper is available from the address https://github.com/ai-sandbox/neural-admixture listed in the paper. In addition version 1.7.1 of PyTorch, Python 3.8, pong 1.5, Admixture 1.3.0, Alstructure 0.1.0, Terastructure 1.0, Haplonet 0.2, and Faststructure 1.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets used in the Experiments section of this article have been compiled from the publicly available 1000 Genomes Project (https://www.internationalgenome.org/data/), the Simons Genome Diversity Project (https://www.simonsfoundation.org/simons-genome-diversity-project/), and the

Human Genome Diversity Project (https://www.internationalgenome.org/data-portal/data-collection/hgdp). Moreover, several compiled datasets used in the Experiments section of the article (All-Chms, Chm-22 and Chm-22-Sim) have been made available in figshare (https://doi.org/10.6084/m9.figshare.19387538.v1). Data from the UK Biobank study was also analyzed. This dataset is available to researchers through an open application via https://www.ukbiobank.ac.uk/register-apply/.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	No sex or gender based analyses were performed, as this study focuses only on computational performance on autosomes.
Reporting on race, ethnicity, or other socially relevant groupings	Grouping labels were sourced from the 1000 Genomes project data, from the Human Genome Diversity project data, or from the UK Biobank dataset, as collected in those studies. Additionally, labels based on geographical continent of origin are used for high level grouping.
Population characteristics	Genotypic information only was used.
Recruitment	No recruitment was conducted. All data originates from public databases as described above.
Ethics oversight	The UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB) approval. This is a publicly available dataset as described above.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was chosen based on existing public dataset size. That is, all samples available in the public datasets were used. These datasets were large enough to compute necessary scaling performance as demonstrated in the study.
Data exclusions	There were no exclusions.
Replication	There were no replicates.
Randomization	This is not relevant to our study as we are not looking for statistical associations; we are simply measuring computational performance.
Blinding	This is not relevant to our study as we are not looking for statistical associations; we are simply measuring computational performance.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging