AEon: A global genetic ancestry estimation tool

Naomi M Warren[1] and Mark Pinese[1,2]

[1]Children's Cancer Institute, Lowy Cancer Research Centre, UNSW Sydney, Sydney, NSW, Australia
[2]School of Clinical Medicine, UNSW Sydney, Sydney, NSW, Australia

Abstract

Genetic ancestry is a significant confounding factor in genetic association studies. It is therefore essential to estimate and account for ancestry if we hope to translate genomic research into equitable clinical utility for ethnically diverse populations. Despite this importance, accurate ancestry information is difficult to obtain, and existing methods to estimate ancestry are not designed for modern sequencing data. This methodological gap hinders the integration of genetic ancestry information in modern research, and limits progress in finding the genetic determinants of disease in diverse populations. To address this gap we present AEon, a probabilistic model-based global ancestry estimation tool, ready for use on modern genomic data. AEon predicts fractional population membership, accounting for possible admixture. Unlike previous global ancestry estimation tools such as EIGENSTRAT and ADMIXTURE, AEon takes input directly from a VCF/BCF, includes default training data based on the 26 reference populations of the 1000 Genomes Project, and produces visualisation aids and diagnostics to complement data output. AEon's turnkey design significantly reduces the time taken to estimate ancestry from VCFs, without compromising on accuracy compared to ADMIXTURE analysis. AEon is available at github.com/GenomicRisk/aeon.

Introduction

Genetic ancestry is a significant confounding factor in genetic association studies. Current association studies are heavily biased toward European populations [1,2], limiting the application of study findings [1,2,3]. Variation in linkage disequilibrium structure between populations can disrupt linkage between causal and marker alleles [3,4], and polygenic or indirect biomarkers can be confounded by uncharacterised modifiers with enriched allele frequency in specific population groups [3,5,6]. It is therefore essential to estimate and account for genetic ancestry if we hope to translate genomic research into equitable clinical utility for ancestrally diverse populations.

Unfortunately, ancestry information is not routinely incorporated in genomic analyses because it is difficult to reliably obtain. Genetic ancestry can be conflated with ethno-cultural identity, which raises the concern that either or both might be misrepresented by a simple "ancestry" survey. This demonstrates the need for a tool that can estimate genetic

ancestry from genome sequencing data itself, but existing tools for this purpose are not simple to access or use.

A major class of ancestry estimation (AE) tools implement global ancestry estimation, which estimates the proportions of a query genome that can be attributed to distinct ancestral populations. Global AE can be used to assess the appropriateness of any reference databases, allele frequencies or risk scores used to analyse a sample. Such information would provide important caveats to any findings and guide downstream curation.

However, current global AE software is limited. Existing global AE tools were either not designed for modern sequencing data and hence require significant manual processing overhead for contemporary use (e.g. ADMIXTURE [7]), or were developed within a specific framework and as a result rely on complex external infrastructure (e.g. gnomAD's AE tool on Hail [8]). More recent tools (e.g. SNVstory [9]) only classify individuals to a single ancestral population, without the capacity for mixture modelling, and so are only appropriate for contexts where admixture is unlikely – an unrealistic expectation for modern clinical cohorts. Without a tool that is admixture-aware, easily accessible, quick to run and interpret, and compatible with mainstream data formats, integrating ancestry information into genomic studies remains a difficult task.

To address this gap we present AEon, a probabilistic model-based global AE tool, ready for use on modern genomic data. AEon predicts fractional population membership of input samples given allele frequency data from known populations, accounting for possible admixture. It is simple to install and run, implemented in Python, and utilises existing mature libraries to process input and perform statistical modelling. AEon takes input directly from a VCF/BCF and outputs estimates in CSV format, enabling easy integration with existing software and pipelines. The tool also includes a default population reference file to provide turnkey ancestry estimation with no further configuration. However, the option to modify or replace this reference file is available should users wish to tailor the tool for their specific input data and questions of interest.

AEon removes the barriers preventing genetic ancestry estimation from being widely integrated into genome analysis pipelines, thereby enabling and accelerating ancestry-informed genetic research. Here we present a high-level overview of AEon, describe its performance and application to modern genomic sequencing data relative to the traditional ADMIXTURE tool, and make the tool and its databases available for use.

Results

Existing global genetic ancestry estimation tools require significant hands-on time to apply on modern sequencing data, limiting their uptake. AEon streamlines global ancestry estimation by running directly from VCF files. This decreased complexity reduces analysis burden and eliminates error-prone data conversion steps. Here we demonstrate that this

simplicity does not come at the expense of accuracy, and that AEon yields accurate ancestry estimates in single-population and admixed samples, competitive with gold-standard global approaches such as ADMXITURE.

AEon identifies the most significant ancestry in an individual

We first evaluated AEon's ability to correctly infer the largest contributing ancestry in a sample. We applied AEon to 30 samples from the 1000 Genomes project, and measured the concordance between AEon's largest estimated ancestry fraction (EAF) and the sample's labelled population.

AEon was highly accurate in identifying the primary population contributing to each sample, correctly identifying the major contributing population in 27 out of 30 cases where 1,000 loci were used for inference, and every case (30/30) where 2,000 loci or more were used. With increasing locus count the prediction became more confident, with AEon assigning fractions approaching 1 to the primary population (Figure S1a).

In some populations, AEon consistently assigned low EAF to other ancestry groups, regardless of the number of loci supplied. For example, African Ancestry in SW USA (ASW) samples consistently had a small fraction of ancestry assigned to other African populations; and Puerto Rican in Puerto Rico (PUR) and Mexican Ancestry in Los Angeles (MXL) samples consistently showed minor fractions of European ancestry, even with 100,000 loci (Figure S1b-d). Notably, this spread EAF pattern was predominantly observed in populations which are known to harbour genetically admixed individuals [10], suggesting that AEon may be detecting latent low-level admixture in these cases. Given the potential for AEon to detect and quantify low levels of admixture, we then explicitly investigated AEon's ability to accurately estimate EAF in known admixed individuals.

AEon accurately estimates ancestry of admixed individuals

To evaluate AEon's ability to estimate known admixture, we simulated offspring from random pairings of individuals from the 1000 Genomes CEU and STU populations. This procedure results in offspring with a fixed 0.5:0.5 CEU:STU admixture fraction. AEon accurately predicted this fractional composition when run with 10,000 loci, with estimates differing from the true label vector by an L2-distance of only 0.035 at a superpopulation level and 0.086 at a population level. Additional loci did not provide significant benefit at a superpopulation level, but some improvement was seen at a population level when 50,000 loci or more were used for prediction (L2-distance of 0.047 at 50,000 loci and 0.038 at 100,000 loci) (Figure 1a).

These results establish AEon's ability to estimate ancestry composition in admixed samples and supports the use of 10,000 loci or more for this application. To further characterise

AEon's performance, we compared its estimates and computational performance to the gold-standard ADMIXTURE method.

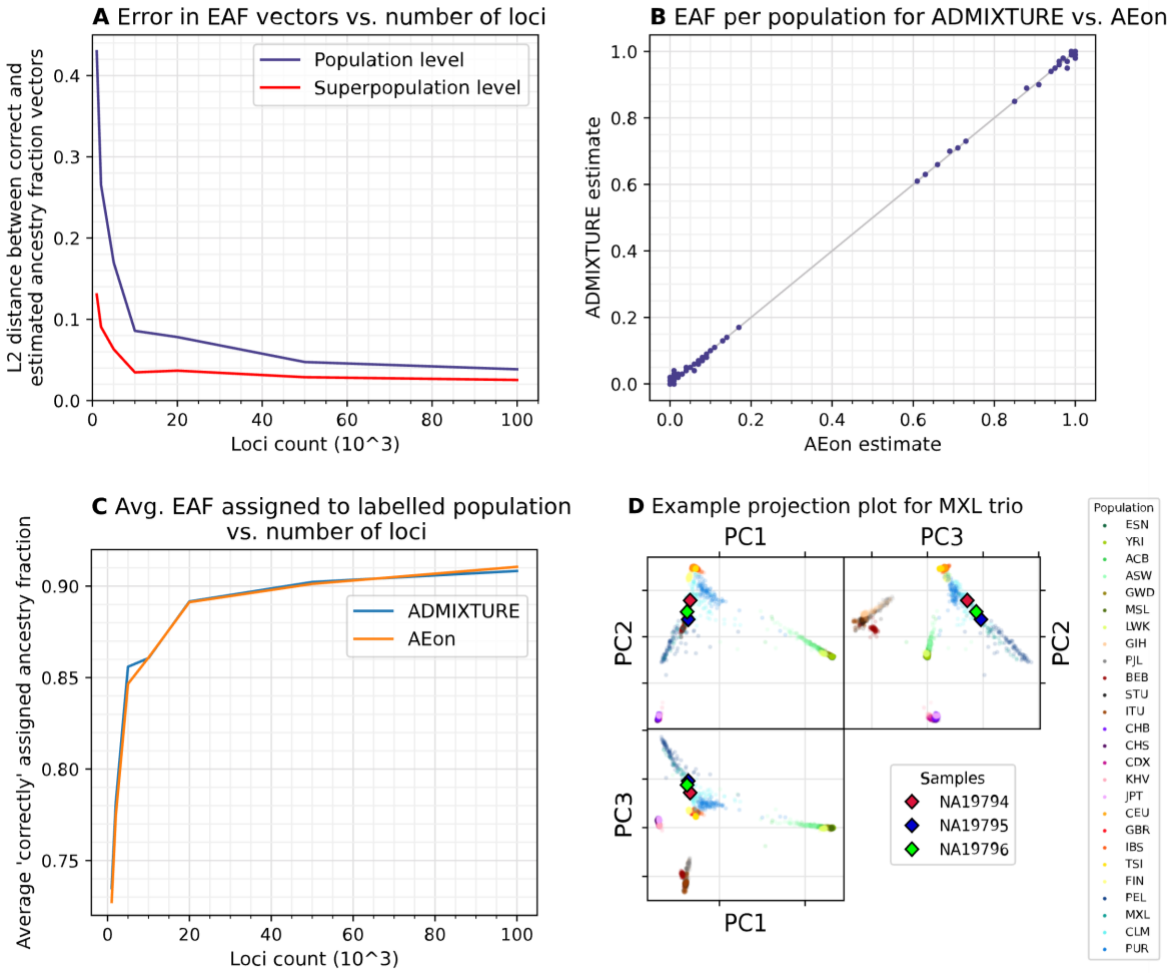

Figure 1. High accuracy of population composition can be achieved at both the superpopulation and population level with sufficient loci, with estimates reaching optimal accuracy from approximately 10,000 loci **(a)**. EAFs are nearly identical between ADMIXTURE and AEon, lying along the line y=x. Each point represents the EAF assigned to a population for a given sample, with 26 points per sample **(b)**. AEon and ADMIXTURE assign very similar fractions to a sample's labelled population when given the same loci on which to base the prediction, with both increasing in accuracy as more data is provided **(c)**. Samples can be visualised in projection space compared to the 1000 Genomes populations as an orthogonal aid **(d)**. African reference populations are in greens, South Asian in grey/browns, East Asian in purples, European in oranges and American in blues.

AEon matches ADMIXTURE's accuracy, with enhanced performance

AEon ancestry estimates were equivalent to those from ADMIXTURE. Across 30 1000 Genomes Project samples, AEon and ADMIXTURE assigned nearly identical ancestry fractions to each population, with a mean difference of 0.00 and maximum difference of 0.03, when considering all 128,907 loci (Figure 1b). This equivalence in accuracy was independent of the number of loci used to inform ancestry prediction (Figure 1c).

Although AEon demonstrated equivalent accuracy to ADMIXTURE, it was more computationally efficient, particularly for large locus counts. AEon ancestry estimation was significantly faster than ADMIXTURE in wallclock time, both as a function of number of loci and the number of samples (Figure 2a,b). This is in large part due to the costly format conversion step required for ADMIXTURE (Table S1), as well as AEon's more efficient utilisation of computational resources (Figures 2c,S2). Although AEon did require more system memory for its computation, it had superior memory scaling to ADMIXTURE, and the RAM requirements for either tool are unlikely to be limiting on modern hardware (Figure 2d).

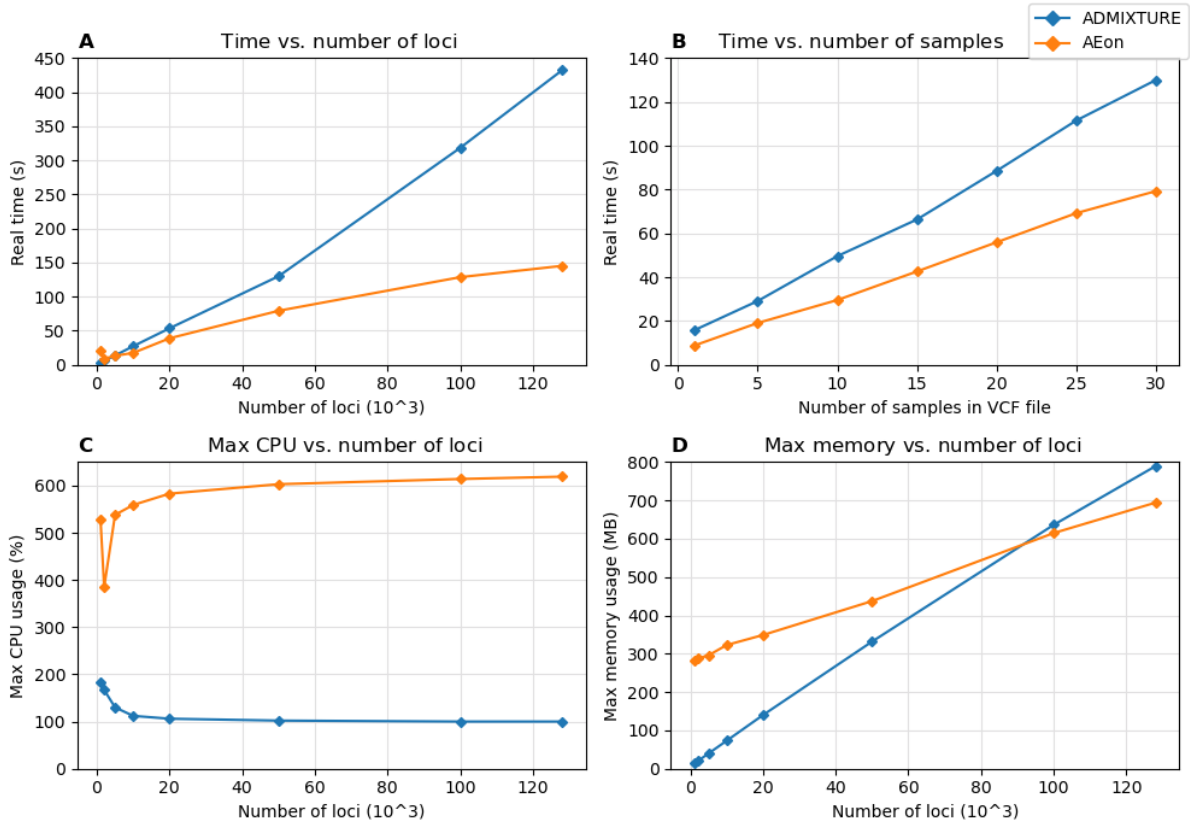

Figure 2. Comparison between the resources used by AEon (orange) and ADMIXTURE (blue). AEon is faster than ADMIXTURE, and its runtime scales better with increasing locus count **(a)** or sample count **(b)**. AEon automatically makes use of available CPU regardless of loci count **(c)**. AEon's memory usage also scales better with loci count **(d)**.

In summary, AEon simplifies the process of estimating ancestry, decreasing the time required to perform analysis both in terms of wallclock time and analyst effort. AEon's simplicity does not compromise accuracy, resulting in ancestry estimates that are of the same quality as the current standard tool ADMIXTURE.

AEon provides a visualisation aid to inspect samples

In addition to .csv output, AEon generates a plot displaying the location of input samples with respect to the 1000 Genomes reference populations in 3-dimensional projection space (Figure 1d). This plot can help situate samples within the wider reference population context.

Discussion

We have presented AEon, a tool to estimate global genetic ancestry from modern sequencing data. AEon has been engineered for ease-of-use and flexibility, minimising data handling overhead and the potential for errors. As an updated implementation of the classical global ancestry mixture model, AEon complements recent methods for ancestry classification, and addresses the usability gap left by current ancestry mixture modelling tools.

AEon's primary contribution is as an ancestry mixture modelling method that works natively on modern genotype data formats. Classical methods for this modelling, such as ADMIXTURE, were designed for genotyping microarray data. Such microarray data differ significantly from modern genotyping-by-sequencing (GBS) in both file format and interpretation, particularly with respect to the handling of ungenotyped loci and homozygous reference loci [11]. As a consequence, although it is possible to apply ADMIXTURE and related methods to modern GBS data in VCF format, this is complex, error-prone, and comes with significant manual and computational overhead (Figure 2). By working natively with VCF data, AEon circumvents these issues and provides a highly performant, lightweight, and convenient framework to estimate ancestry from genomic data.

A key advantage of AEon is its easy and flexible setup. AEon has straightforward single-step execution, lightweight dependencies, and comes with a default ancestry reference database. The inclusion of a reference ancestry database is a notable benefit of AEon, enabling new users to generate accurate ancestry estimates with minimal effort. This simplicity of AEon does not come at the expense of flexibility, and AEon accepts custom ancestry databases if required. To our knowledge, this flexibility is unique, as existing tools either require the user to generate a reference database first (eg ADMIXTURE), or do not allow for the use of a custom ancestry reference (eg SNVStory [9]). With our findings that potential reference database mismatch or incompleteness can have a significant impact on results (Figure S1), this flexibility offered by AEon is a substantial benefit. Ancestry derived entirely from non-reference populations cannot be captured by current tools, and as such the flexibility to accommodate new data as it becomes available is essential for continued equitable use.

As a modern implementation of the classical ancestry mixture model, AEon complements other new tools. For example, the SNVStory method, and the ancestry estimation module of Hail [8], also aim to ascertain ancestry from modern GBS data. However, these methods are designed as classification tools, which assign individuals to a single ancestry. For populations with low admixture this is an appropriate methodology, however it is unsuitable for characterising admixed individuals. AEon, like the classical ADMIXTURE and STRUCTURE methods, natively fits ancestral mixtures and so naturally and accurately handles admixture, which is commonly encountered in modern genomic cohorts. AEon also provides a PCA-based visualisation of the results for further assessment of reference fit and suitability.

Overall AEon is an effective tool for estimating global genetic ancestry from WGS data, accurately determining the fractional ancestral populations contributing to an individual. Built on python to run with a single command and including default reference data, it allows meaningful information to be extracted from genotype data immediately upon installation, while also providing input flexibility that enables users to include the most relevant data for their analyses.

Implementation

Approach

AEon estimates population membership for individuals based on their genotype calls and a population-specific allele frequency reference.

Briefly, AEon employs `pysam` to convert input genotypes in VCF/BCF format to dosages of pre-specified ancestry-informative alleles. These dosages are then fit to an ADMIXTURE-like model by maximum likelihood using `pyro` [12], to yield a per-individual population mixture estimate. This fit requires a reference table giving frequencies of each allele in each population of interest; we supply a default table for GRCh38 references across the 1000 Genomes Project populations, but AEon also supports user-supplied tables.

As output AEon produces per-individual population mixtures in CSV format, as well as diagnostic data and plots to verify fit and reference matrix suitability.

Statistical Model

AEon models an individual as a mixture of prespecified ancestral populations. Specifically, denote the count of non-reference alleles of an individual as $G_j \in \{0, 1, 2\}$, for a biallelic locus j. Under the gene pool model, we suppose each G_j is drawn independently with probabilities $\left[\left(1 - r_j\right)^2, 2r_j\left(1 - r_j\right), r_j^2\right]$ for $G_j = [0, 1, 2]$, where r_j is the non-reference allele frequency for the gene pool from which the individual was sampled. We model the

gene pool frequency for this individual r_j as a mixture of ancestral frequencies, $r_j = \sum_i f_{ij} p_i$ where $p_i \in [0,1]$ is the fraction of ancestral population i contributing to this individual, and f_{ij} is the non-reference allele frequency at locus j in ancestral population i, $\sum_i p_i = 1$. Given prespecified f_{ij} and supplied individual G_j from genotyping, AEon estimates the full vector of population memberships p.

Generating the Population Reference

AEon includes a default reference table of ancestry-informative allele frequencies from the 2,504 unrelated individuals in the 1000 Genomes (G1K) Project Phase 3 dataset [13], sampled from 26 populations grouped into 5 superpopulations. To generate this table, we filtered an initial list of 133,872 population-informative alleles [14] to remove alternate alleles absent from GRCh38 and loci missing from the 1000 Genomes call set, to yield a final table of allele frequencies for 128,097 alternate alleles across 26 populations.

Testing on 1000 Genomes populations

Ancestry estimation with AEon was tested on WGS data from the 1000 Genomes Project. Two trios from each of the five superpopulations labelled in the G1K database were selected as test samples (Table S2), for a total of 30 individuals. Genotype calls for these 30 individuals was provided to AEon as an indexed BCF using default flags plus the flag `--no_visualisation`. A random subset of 1,000, 2,000, 5,000, 10,000, 20,000, 50,000 or 100,000 loci from the 128,907 default reference alleles were provided in the reference allele frequency file to compare performance with varying amounts of information, as well as a run with the complete set of 128,907 loci. The estimated ancestry fractions provided as output from these tests were then compared against the labelled population assigned by the G1K database for each sample.

Testing on Simulated Admixed Individuals

Since all individuals from the G1K database are intended to be representative of specific populations, we simulated admixture to test AEon's performance on individuals with genetic ancestry from multiple population groups. Mixtures of CEPH European Ancestry in Utah (CEU) and Sri Lankan Tamil in the UK (STU) samples were generated by treating all loci as unlinked and randomly passing on one allele from each parent to the simulated child. This was repeated to produce 25 simulated individuals with a mix of 50CEU:50STU ancestry. These simulated individuals were run through AEon as above using eight randomly sampled subsets of loci for each size from 1,000 to 100,000 loci, and average predictions were calculated.

Given that these individuals did not belong to a sole population group, their estimated ancestry fractions from each trial were compared against the G1K labelled fractions using

L2-distance between ancestry fraction vectors. The 'correct' ancestry fraction vector had a value of 0.5 at the positions corresponding to CEU and STU and 0 elsewhere.

Comparison to ADMIXTURE

Having validated AEon's results against the labelled populations, we also compared its performance to the existing ancestry estimation tool ADMIXTURE.

To run ADMIXTURE, a pre-processing script was created in bash to convert the initial BCF into the required format (Supplementary Material). In addition, ADMIXTURE required an initial run on the reference set, after which ancestry estimation for the test samples was performed using ADMIXTURE's projection mode. Max memory usage, max CPU and real time spent were recorded for both the pre-processing step and the predictive projection-mode ADMIXTURE run using the same selection of loci (between 1,000 and 100,000) as the AEon trials. We compared these resource metrics between tools, as well as the per-sample population mixtures predicted by each tool, on a MacBook Pro machine (Apple M2 Pro chip with 10-core CPU (6 performance cores, 4 efficiency cores), 16GB memory).

Availability and requirements

AEon is available at:

- GitHub: github.com/GenomicRisk/aeon
- Docker: https://hub.docker.com/r/naomiwren/aeon

Requirements:

Installation from the GitHub repository requires a working version of Python (tested using Python 3.10) and pip. Use of a virtual environment (e.g. conda) is also recommended.

References

1. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538:161–4.

2. Wong KHY, Ma W, Wei C-Y, Yeh E-C, Lin W-J, Wang EHF, et al. Towards a reference genome that captures global genetic diversity. Nat Commun. 2020;11:5482.

3. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177:26–31.

4. Galinsky KJ, Reshef YA, Finucane HK, Loh P, Zaitlen N, Patterson NJ, et al. Estimating cross-population genetic correlations of causal effect sizes. Genet Epidemiol. 2019;43:180–8.

5. Tahmasebi H, Trajcevski K, Higgins V, Adeli K. Influence of ethnicity on population reference values for biochemical markers. Crit Rev Clin Lab Sci. 2018;55:359–75.

6. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51:584–91.

7. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.

8. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.

9. Bollas AE, Rajkovic A, Ceyhan D, Gaither JB, Mardis ER, White P. SNVstory: inferring genetic ancestry from genome sequencing data. BMC Bioinformatics. 2024;25:76.

10. Salzano FM, Sans M. Interethnic admixture and the evolution of Latin American populations. Genet Mol Biol. 2014;37:151–70.

11. Darrier B, Russell J, Milner SG, Hedley PE, Shaw PD, Macaulay M, et al. A Comparison of Mainstream Genotyping Platforms for the Evaluation and Use of Barley Genetic Resources. Front Plant Sci. 2019;10.

12. Bingham E, Chen JP, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, et al. Pyro: Deep Universal Probabilistic Programming. Journal of Machine Learning Research. 2019;20:1–6.

13. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. 2022;185:3426-3440.e19.

14. Pinese M, Lacaze P, Rath EM, Stone A, Brion M-J, Ameur A, et al. The Medical Genome Reference Bank contains whole genome and phenotype data of 2570 healthy elderly. Nat Commun. 2020;11:435.

15. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517:327–32.

Supplementary Data

Tables

Table S1. Real time for ADMIXTURE processing of a 30 sample file

# loci	**Time (s)**		
	Conversion step	Prediction step	Total
1,000	2.64	1.26	3.90
2,000	4.24	1.93	6.17
5,000	9.09	4.60	13.69
10,000	17.24	10.13	27.37
20,000	33.46	20.16	53.62
50,000	81.99	47.99	129.98
100,000	163.40	154.99	318.39
128,097	212.39	220.08	432.47

Table S2. Selected test samples from the 1000 Genomes Project

Superpopulation	Population Code	Population Name	SampleID
East Asian	CHS	Han Chinese South	HG00403
			HG00404
			HG00405
	KHV	Kinh in Ho Chi Minh City	HG02015
			HG02016
			HG02017
South Asian	PJL	Punjabi in Lahore	HG02490
			HG02491
			HG02492
	STU	Sri Lankan Tamil in UK	HG03943
			HG03992
			HG04036
European	IBS	Iberian in Spain	HG02238
			HG02239
			HG02240
	CEU	CEPH (Utah residents with Northwest European Ancestry)	NA06984
			NA06989
			NA12329
American	MXL	Mexican Ancestry in Los Angeles	NA19794
			NA19795
			NA19796

	PUR	Puerto Rican	HG00551
			HG00552
			HG01075
African	YRI	Yoruba in Ibadan	NA19247
			NA19248
			NA19249
	ASW	African Ancestry in Southwest US	NA20356
			NA20357
			NA20358

Figures

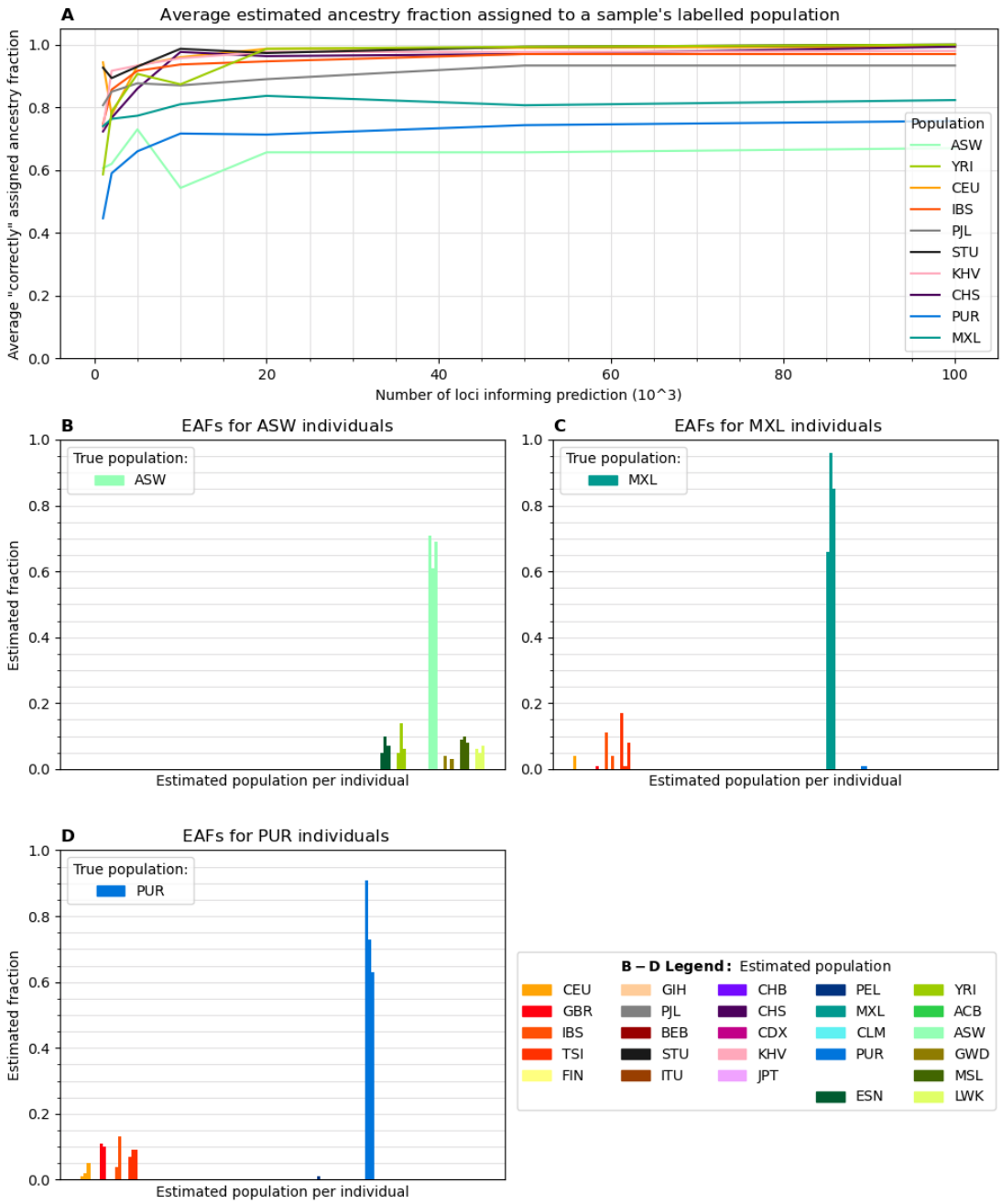

Figure S1. Increasing the number of loci used to inform prediction produces EAFs close to 1 for a sample's labelled population in most cases. The exceptions to this are samples from the African population ASW and the two American populations, PUR and MXL **(a)**. For ASW individuals, the remaining fraction is assigned to other African populations (shades of green) **(b)**. We note that African populations tend to have greater within-population genetic diversity than European populations [15] and therefore would be harder to characterise by representative samples, and the ASW population group had the smallest reference sample size with only 61 samples.
The remaining ancestry fraction assigned to MXL and PUR individuals outside their labelled population was to European populations (shades of orange) **(c-d)**. We consider this a reasonable estimate due to the historic movement of European peoples to this area [10].

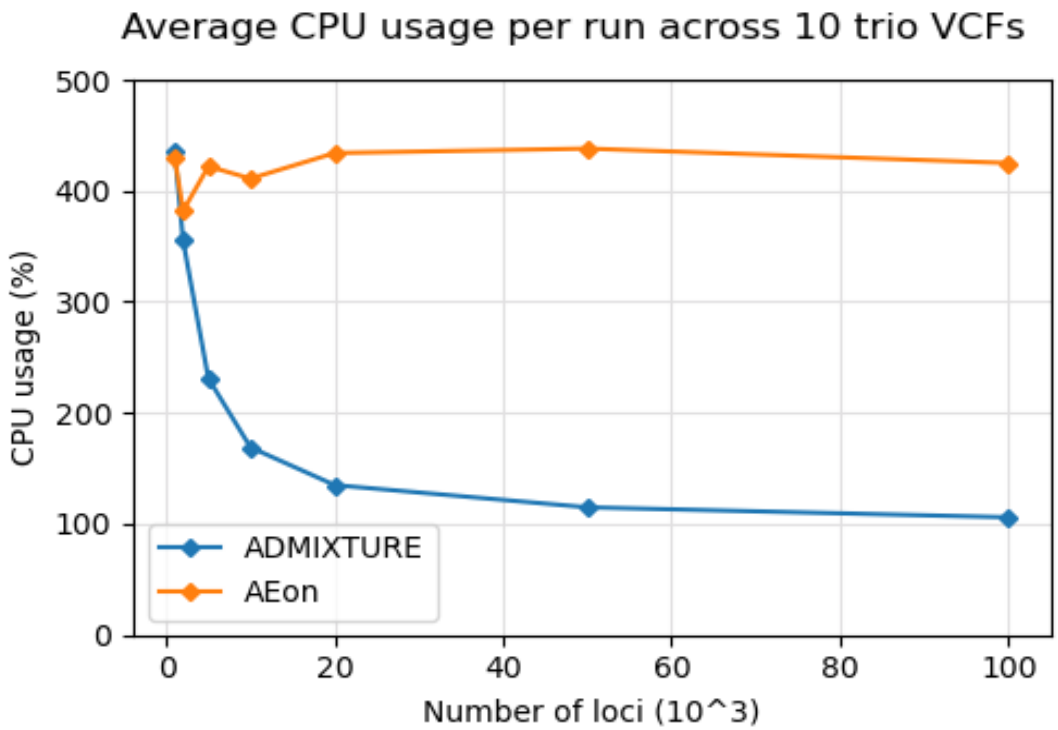

Figure S2. When run on files containing only three samples, ADMIXTURE was able to make use of the same amount of CPU as AEon with small number of loci.