Evaluation of Model Fit of Inferred Admixture Proportions

Genís Garcia-Erill[1], Anders Albrechsten[1]

[1] Department of Biology, Section for Computational and RNA Biology, University of Copenhagen, 2200 Copenhagen N, Denmark

July 19, 2019

Abstract

Model based methods for genetic clustering of individuals such as those implemented in *structure* or ADMIXTURE allow to infer individual ancestries and study population structure. The underlying model makes several assumptions about the demographic history that shaped the analyzed genetic data. One assumption is that all individuals are a result of K ancestral homogeneous populations that are all represented well in the data while another assumption is that no drift happened after the admixture event. The histories of many real world populations do not conform to that model, and in that case taking the inferred admixture proportions at face value might be misleading. We propose a method to evaluate the fit of admixture models based on calculating the genotypes predicted by the admixture model, and obtaining the residuals as the difference between the true and predicted genotypes. The correlation of residuals between pairs of individuals can then be used as a measure of model fit. When the model assumptions are not violated and the inferred admixture proportions are accurate then the residuals from a pair of individuals are not correlated. In case of a bad fit, individuals with similar histories have a positive correlation of their residuals. Using simulated and real data, we show how the method is able to detect a bad fit of inferred admixture proportions due to using an insufficient number of clusters K or to demographic histories that deviate significantly from the admixture model assumptions, such as admixture from ghost populations, drift after admixture events and non-discrete ancestral homogeneous populations.

1 Introduction

Characterization of population structure is a key step in population genetics. It is essential for understanding the population histories and is important to take into account when analysing genetic data for example in genetic association studies [1]. Model-based clustering of individuals is a widely used approach to do so, with several programs available [2, 3, 4, 5]. Each software has different inference procedures and other variations, but they all share a common likelihood model, which we will refer to as *admixture model*. Under this model, each allele is an independent sample from a binomial distribution, with probability equal to an *individual allele frequency* specific to each individual and loci. From a predefined number of clusters (K) individual frequencies are characterized by the individual admixture proportions (Q) and the ancestral population allele frequencies (F).

This model implies several assumptions about the demographic history behind the analyzed individuals. First, it assumes that all individuals derive from K homogeneous ancestral populations, all of them in Hardy-Weinberg equilibrium (HWE) and well represented in the data set. Moreover, individuals are assumed to be unrelated and no linkage disequilibrium (LD) among loci is expected (with some exceptions [6]). In cases of admixture, i.e.

when individuals are assigned to more than one cluster, the ancestral and present population frequencies are assumed to be the same, corresponding to a recent admixture event where the effect of genetic drift is negligible. At the same time, the source ancestral population of the two alleles of an individual at the same loci are treated as independent, which doesn't account for the presence of recent hybridization.

Choosing the most appropriate value of K to model the data is a typical worry when applying an admixture analysis. Several model selection procedures have been proposed [2, 7, 8], but none has lead to an unambiguous general solution to the problem [9, 10]. Moreover, many population histories do not conform to the model. In that case there does not really exist anything similar to a correct K, and the application of an admixture analysis can lead to misleading conclusions. A recent paper [11] showed and discussed several cases of problematic admixture analysis, and introduced badMIXTURE, a tool to evaluate the model fit of admixture results. The method requires applying CHROMOPAINTER [12] and comparing its results with what has been obtained from the admixture analysis. CHROMOPAINTER is based on patterns of phased haplotype sharing, and therefore requires the presence of LD in the data and the existence of linkage maps for the analyzed organism. badMIXTURE can still be applied to unlinked loci, but its resolution decreases sensibly in that case, and it requires a certain amount of markers to detect some signal [11]. Its requirements are therefore out of reach to many kinds of data to which admixture analyses are commonly applied.

We propose an alternative method to asses the admixture model fit, explicitly based on evaluating to what extent the estimated individual allele frequencies are capturing the individual allele frequencies, and allowing in most cases for an assessment of the model fit at the individual level. It can easily be applied to any situation where an admixture analysis is used, including when working with low-depth next generation sequencing (NGS) data, where the admixture analysis should be done from genotype likelihoods to avoid the bias introduced by calling genotypes [5]. After introducing the method, we show on simulated and real data how it can detect a bad admixture model fit due to using an insufficient number of ancestral clusters K, to demographic histories that deviate from the model assumptions or to the presence of related individuals.

2 Methods

2.1 Correlation of residuals as a measure of admixture fit

For N individuals, M sites and K ancestral populations, we have an $N \times M$ genotype matrix G, and we have estimated with an admixture analysis an $N \times K$ admixture proportions matrix $\hat{Q}$ and an $M \times K$ ancestral frequencies matrix $\hat{F}$. We assume for convenience that genotypes indicate the number of minor alleles individual i carries at site j; so $g_{ij} \in \{0, 1, 2\}$. Under the admixture model, each genotype g_{ij} is a sample from a binomial distribution with parameters $n = 2$ and $p = \hat{\pi}_{ij}$, where $\hat{\pi}_{ij}$ is the estimated individual allele frequency of individual i at site j, and comes from an $M \times N$ matrix $\hat{\Pi}$ where each entry is

$$\hat{\pi}_{ij} = \sum_{k=1}^{K} \hat{f}_{jk} \hat{q}_{ik}. \tag{1}$$

Here $\hat{f}_{jk}$ and $\hat{q}_{ik}$ are entries in the $\hat{F}$ and $\hat{Q}$ matrices. Let $\hat{\Pi}_i$ be a vector containing the

estimated individual allele frequencies of all sites for individual i. We treat each of the true genotypes as a realization from a binomial distribution with probability given by unknown true individual frequencies Π, and $n = 2$. The expected true genotype will be $\mathbb{E}[g_{ij}] = 2\pi_{ij}$. As an evaluation of the admixture model fit, we aim to test, for each individual i, if $\hat{\Pi}_i$ are unbiased estimates of the true Π_i.

We define the predicted genotypes $\hat{G}$, an $N \times M$ matrix containing the expectation of the genotypes conditional on the estimated individual frequencies

$$\hat{G} = \mathbb{E}[G|\hat{\Pi}] = 2\hat{\Pi}, \tag{2}$$

and R, another $N \times M$ residual matrix containing the difference between the true and the predicted genotypes

$$R = G - \hat{G}. \tag{3}$$

R_i is a vector containing the residuals of all M sites from individual i. If there is a good admixture model fit, i.e. $\mathbb{E}[\hat{\pi}_{ij}] = \pi_{ij}$, the expected residual $\mathbb{E}[r_{ij}] = 0$ for every site. But with a bad model fit, the expected estimated frequency will deviate from the true frequency by some quantity $\mathbb{E}[\hat{\pi}_{ij}] = \pi_{ij} - \delta_{ij}$; and this will be present in the expected residuals

$$\mathbb{E}[r_{ij}] = 2\pi_{ij} - 2(\pi_{ij} - \delta_{ij}) = 2\delta_{ij}. \tag{4}$$

This deviation will be specific to every site; but individuals from the same population, with similar histories that not captured by the admixture model, will tend to share the same specific deviations at all sites. As a measure of admixture model fit, we use the correlation of residuals between pairs of individuals, e.g. between individuals 1 and 2,

$$\rho_{12} = Cor(R_1, R_2) = \frac{\sum_{j=1}^{M}(r_{1j} - \overline{r}_1)(r_{2j} - \overline{r}_2)}{\sqrt{\sum_{j=1}^{M}(r_{1j} - \overline{r}_1)^2}\sqrt{\sum_{j}^{M}(r_{2j} - \overline{r}_2)^2}}. \tag{5}$$

If individual 1 and 2 are from the same population, under a bad fit the sharing of a systematic error in their individual frequencies will result in a positive correlation of their residuals. When we have a good model fit, the residuals should only be a result from the binomial variance, and they should be independent between any pair of individuals. However, this is not the case when working with a finite sample size. There is a bias that arrives from the estimation of the ancestral frequencies $\hat{f}_{jk}$. Because $\hat{f}_{jk}$ is estimated using the genotypes of all individuals with some admixture proportion from population k, this will introduce a bias in the correlation of the residuals between individuals that share ancestry from one or more populations. For example if individuals 1 and 2 are non-admixed individuals assigned to the same k population, the expected correlation under a good model fit is

$$\mathbb{E}[\rho_{12}] = -\frac{1}{N_k - 1}, \tag{6}$$

where N_k is the number of individuals assigned to ancestral population k, assuming that all individuals with ancestry from that population are not admixed (See S1 for details).

2.1.1 Frequency correction

We can make the residuals between individuals 1 and 2 independent under a good model fit, by removing the contribution of individual 1 from the frequency used for the calculation of individual 2's residuals. First assuming for simplicity that the genome of every individual comes from a single ancestral population k, for a site j

$$\tilde{\pi}_{2j,-1} = \tilde{f}_{jk,-1} = \frac{\hat{f}_{jk} - g_{1j}}{2(N_k - 1)} = \frac{1}{2(N_k - 1)} \sum_{i=2}^{N_k} g_{ij} \tag{7}$$

In this case, when $\hat{f}_{jk}$ is an unbiased estimate of the true frequency f_{jk}, $r_{1j} = g_{1j} - 2\hat{\pi}_{1j}$ and $r_{2j} = g_{2j} - 2\tilde{\pi}_{2j,-1}$ are independent and we expect a zero correlation of their residuals (proof in S2).

To extend the frequency correction to the case with admixed individuals, we need to consider that the number of alleles contributed by each individual to the calculation of the ancestral frequency will not be an integer. Consequently, also the sample size N_k used to obtain f_k will not be an integer. Individual's i contribution to the calculation of the allele frequency of site j in population k will depend on the probability that one of its alleles $\mathbb{A} \in \{0, 1\}$ comes from ancestral population k. We denote the ancestry of the allele $\mathbb{A}$ of individual i for site j as $A_{ij} = k$ with probability

$$P(A_{ij} = k|\hat{F}, \hat{Q}, \mathbb{A} = 1) = \frac{P(\mathbb{A} = 1|A_{ij} = k, \hat{F}, \hat{Q})P(A_{ij} = k|\hat{F}, \hat{Q})}{\sum_{k'} P(\mathbb{A} = 1|A_{ij} = k', \hat{F}, \hat{Q})P(A_{ij} = k|\hat{F}, \hat{Q})} = \frac{f_{jk}q_{ik}}{\sum_{k'}^{K} f_{jk'}q_{ik'}}, \tag{8}$$

and

$$P(A_{ij} = k|\hat{F}, \hat{Q}, \mathbb{A} = 0) = \frac{P(\mathbb{A} = 0|A_{ij}^1 = k, \hat{F}, \hat{Q})P(A_{ij} = k|\hat{F}, \hat{Q})}{\sum_{k'} P(\mathbb{A} = 0|A_{ij} = k', \hat{F}, \hat{Q})P(A_{ij} = k|\hat{F}, \hat{Q})} = \frac{(1 - f_{jk})q_{ik}}{\sum_{k'}^{K}(1 - f_{jk'})q_{ik'}}. \tag{9}$$

The contribution of each individual to the frequency calculation is given by the number of minor and major alleles it carries, given by its genotype, times the probabilities calculated with equations (8) and (9). The estimated ancestral frequencies can be written as

$$\hat{f}_{jk} = \frac{\sum_i^N P(A_{ij}^1 = k|\hat{F}, \hat{Q}, \mathbb{A} = 1)g_{ij}}{\sum_i^N \Big(P(A_{ij}^1 = k|\hat{F}, \hat{Q}, \mathbb{A} = 1)g_{ij} + P(A_{ij}^1 = k|\hat{F}, \hat{Q}, \mathbb{A} = 0)(2 - g_{ij})\Big)}. \tag{10}$$

Then we could obtain a new frequency estimate without an individual's contribution by summing all the other individuals. For example the frequency estimate without the contribution of individual 1 is

$$\tilde{f}_{jk,-1} = \frac{\sum_{i=2}^{N} P(A_{ij} = k|\hat{F}, \hat{Q}, \mathbb{A} = 1) g_{ij}}{\sum_{i=2}^{N} \Big(P(A_{ij} = k|\hat{F}, \hat{Q}, \mathbb{A} = 1) g_{ij} + P(A_{ij} = k|\hat{F}, \hat{Q}, \mathbb{A} = 0)(2 - g_{ij}) \Big)}. \tag{11}$$

Because we used the initial frequencies to obtain the ancestral probabilities with equations (8) and (9), the new estimate will still contain some contribution of individual 1, and therefore the correlation of residuals will still be biased. To obtain an estimate of the unbiased frequencies, we used an expectation maximization (EM) iterative algorithm. At each iteration, the ancestral probabilities calculated with equations (8) and (9) are calculated using the frequency estimated in the previous iteration with equation (11). We found that 5 iterations were enough at every test to obtain an unbiased correlation under a good model fit. Then individual 2's individual allele frequencies $\tilde{\Pi}_{2,-1}$ are given by

$$\tilde{\pi}_{2j,-1} = \sum_{k=1}^{K} \tilde{f}_{jk,-1} \hat{q}_{1k} \tag{12}$$

And these frequencies are used to obtain individual's 2 residuals used to calculate the correlation with individual's 1 residuals

$$R_{2,-1} = G_1 - \tilde{\Pi}_{2,-1}. \tag{13}$$

The residuals of individual 2 are calculated from the frequencies from the admixture model (3). In case of a good admixture model fit,

$$\mathbb{E}[\rho_{12}] = \mathbb{E}[Cor(R_1, R_{2,-1})] = 0. \tag{14}$$

2.2 Method with NGS data

When working with NGS data, we work with genotype likelihoods instead of called genotypes. Genotype likelihoods give the probability of observing the sequencing data X given that the true genotype is g_{ij}; $P(X_{ij}|G_{ij} = g_{ij})$. From the genotype likelihoods, we can use Bayes theorem to obtain the probability of each genotype given the data

$$P(G_{ij} = g_{ij}|X_{ij}) = \frac{P(X_{ij}|G_{ij} = g_{ij}) P(G_{ij} = g_{ij}|\hat{\pi}_{ij})}{\sum_{g' \in (0,1,2)} P(X_{ij}|G_{ij} = g') P(G_{ij} = g'|\hat{\pi}_{ij})}. \tag{15}$$

We use the individual frequencies obtained from the admixture results as prior, and assume Hardy-Weinberg equilibrium

$$P(G_{ij} = g_{ij}|\hat{\pi}_{ij}) = \begin{cases} (1 - \hat{\pi}_{ij})^2 & if \quad g_{ij} = 0 \\ 2(1 - \hat{\pi}_{ij})\hat{\pi}_{ij} & if \quad g_{ij} = 1 \\ \hat{\pi}_{ij}^2 & if \quad g_{ij} = 2 \end{cases}. \tag{16}$$

Considering all three possible genotypes, we calculate the expected genotype given the sequencing data

$$\mathbb{E}[G_{ij}|X_{ij}] = \sum_{g' \in (0,1,2)} g' P(G_{ij} = g'|X_{ij}) = P(G_{ij} = 1|X_{ij}) + 2P(G_{ij} = 2|X_{ij}). \tag{17}$$

This expected genotype is used instead of the true known genotype in equations (10), (11) and (13). Moreover, because we used the estimated individual frequencies as prior, we need to update $\mathbb{E}[G|X]$ at each iteration of the EM algorithm during the frequency correction, using the individual frequency estimates of the previous step in equation (16).

2.3 Simulations

2.3.1 Genotypes

To test the method, we simulated genotype data and genotype likelihoods for two different scenarios. The two scenarios were generated from allele frequency data of the Human Genome Diversity Project (HGDP) [13]. In Scenario 1 we used an European (French) an East Asian (Han Chinese) and an African (Yoruba) populations, and furthermore we created an admixed population with admixture proportions $(Q_{Fr}, Q_{Han}, Q_{Yor}) = (0.7, 0.3, 0)$. In scenario 2 we included also the French, Han and Yoruba populations, and an admixed population that in this case has ancestry from a Native American population (Zapotec), with $(Q_{Fr}, Q_{Han}, Q_{Yor}, Q_{Zap}) = (0.7, 0, 0, 0.3)$. For each scenario we simulated genotypes of a total 80 individuals from around 0.5 million sites. In both of them there were 20 individuals from the French, the Han and the Yoruba populations, and 20 from the admixed population. The genotypes were generated by sampling from a binomial distribution using the individual frequencies as given by (1) (which in non-admixed individuals reduce to the population frequencies). All sites were in HWE and without LD between them. A minimum allele frequency (MAF) filter was applied using PLINK v.1.07 [14]; around 0.4 million sites were left in each scenario. Admixture proportions and ancestral population frequencies were estimated for both scenarios using ADMIXTURE [4] with $K = 3$.

2.3.2 Genotype likelihoods

For each of these genotype datasets, we also simulated low-coverage data by sampling reads with a mean depth of $3X$, assuming a Poisson distribution. Sequencing errors were included by having a probability $e = 0.01$ of sampling the wrong base. Genotype likelihoods were then calculated from the sampled reads. A more detailed description of the simulation protocol used can be found in [15]. Admixture proportions and ancestral allele frequencies were estimated from the genotype likelihoods using NGSadmix [5] with default settings.

2.4 Badmixture simulations

We obtained the freely available data of the three simulated scenarios used in [11] as PLINK files. The simulations comprise a total of 13 populations that aim at mimicking human population history, originally performed as part of a study on the origin of the

Ethiopian Ari populations [16]. Each of the three scenarios differ in the relationships between 4 populations that correspond to Ethiopian-like populations, in which we and the previous studies focused and that we label Pop1, Pop2, Pop3 and Pop4. The whole dataset contains a total of 795 individuals in the scenarios labelled as Bottleneck and Ghost Admixture, and 785 in the Recent Admixture. Regarding the four populations of interest, the Bottleneck and Ghost Admxiture scenario contain 15 individuals in Pop1, 25 in Pop2, 100 in Pop3 and 25 in Pop4, and Recent Admixture there are 35 individuals in Pop1, 25 in Pop2, 70 in Pop3 and 25 in Pop4. We used PLINK v.1.07 variants with MAF below 0.05 and removed LD by removing one of each pair of sites with an r^2 above 0.1 within windows of 100 kb; about 40000 sites were left after filtering in each scenario. For each of the scenarios, we ran ADMIXTURE in the whole dataset assuming $K = 13$, and calculated the correlation of residuals between individuals from Pop1, Pop2, Pop3 and Pop4.

2.5 1000 Genomes Data

2.5.1 Genotypes

We also tested the method on real human data from the 1000 Genomes Project Consortium (1000G) [17]. We used 435 individuals from 5 populations: 108 from Yoruba in Ibadan, Nigeria (YRI), 61 from African Ancestry in Southwest US (ASW), 99 from Utah residents with Northern and Western European ancestry (CEU), 63 from Mexican Ancestry in Los Angeles, California (MXL) and 103 from Han Chinese in Beijing, China (CHB). We used autosomal sites from the Human Origins panel [18], with 407441 sites left after filtering sites with minor allele frequency (MAF) lower than 0.05. We inferred relatedness between all pairs of individuals in the dataset using relateAdmix [19], that is based on inferring the fraction of sites in which a pair of individuals have 1 (k_1) or 2 (k_2) alleles identical by descent (IBD). Kinship coefficients can then be calculated as $\theta = \frac{k_1/2+k_2}{2}$. We used the admixture results with $K = 4$ to control for the effect of admixture in relatedness inference.

2.5.2 Genotype likelihoods

To test the NGS version of the method, we used low-coverage sequencing data from the same 1000G populations, individuals and sites used for the genotypes version. We calculated genotypes likelihoods using ANGSD [20] applying the model described in [21], and filtering bases with sequence quality below 30 and reads with mapping quality below 20, excluding sites without data in more than 50 individuals, with a SNP significance threshold of $1e^{-6}$ and excluding variants with MAF below 0.05. A total of 408240 SNPs were left after filtering. The mean sequencing depth per individual ranged from $2.6X$ to $15.8X$, with a median of $6.8X$. We ran NGSadmix with $K = 3$ and $K = 4$, and calculated the correlation of residuals in each case.

3 Results

3.1 Simulations

We simulated two similar scenarios, that differ in the demographic history of the admixed populations. In Scenario 1 there is admixture between the French and Han populations,

while in Scenario 2 it is the unsampled Zapotec population that admixes with the French. It is therefore a case of admixture from a ghost population.

Both scenarios result in nearly identical estimated admixture proportions. The individuals from the admixed populations are inferred as having ancestry from the French and the Han populations (Figures 1 and 2). This results are accurately capturing the simulated model in Scenario 1, and we obtain uncorrelated residuals between all individuals (Fogire 1). In Scenario 2 the inferred admixture model is not a good fit to the data. Because the individual frequencies are here calculated using the Han frequencies instead of the Zapotec, there is a systematic error in the predicted genotypes, so the residuals of admixed individuals are positively correlated (Figure 2). We also applied the method to genotype likelihoods obtained from simulated sequence data from the same datasets with a mean depth of $3X$, resulting in identical results (Supplementary Figures S1 and S2).

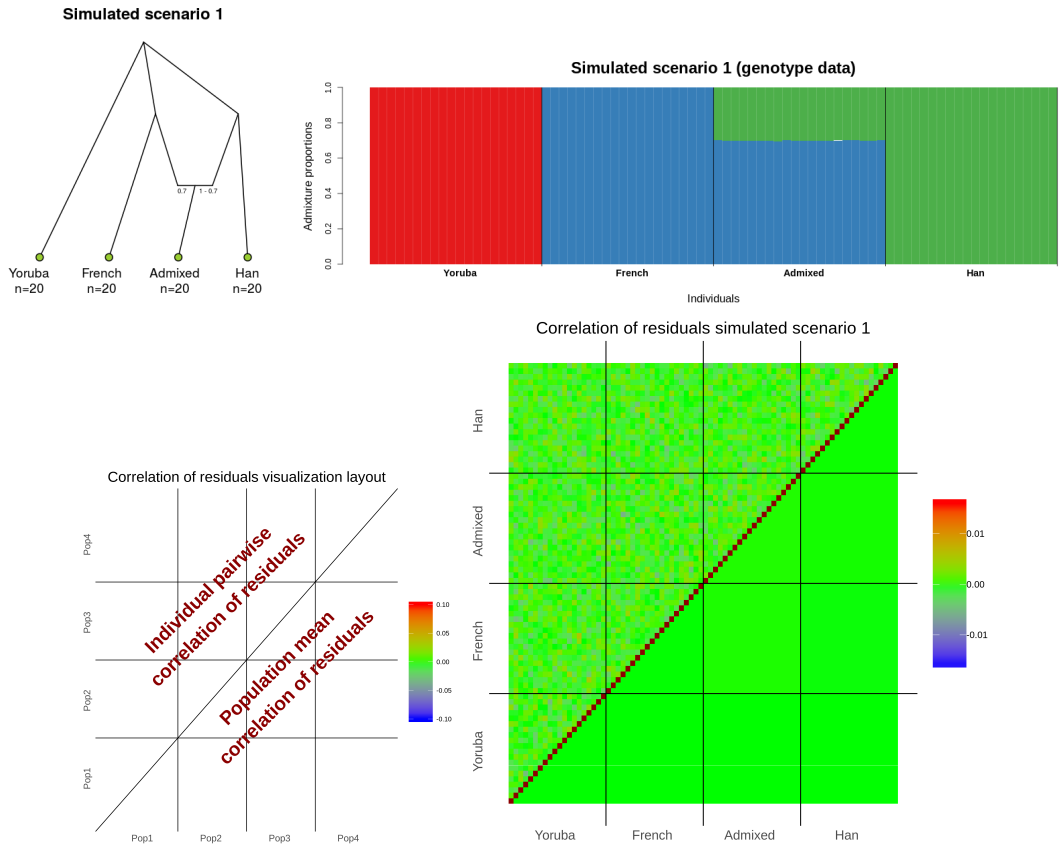

Figure 1: Population tree of the simulated demographic history in Scenario 1 (upper left panel), inferred admixture proportions assuming $K = 3$ from genotype data using ADMIXTURE (upper right panel), diagram with the placement of data in the heatmap used for visualizing the correlation of residuals (bottom left panel), and evaluation of model fit of the admixture results as the correlation of residuals (bottom right panel). Residuals are independent between all individuals, indicating a good admixture model fit.

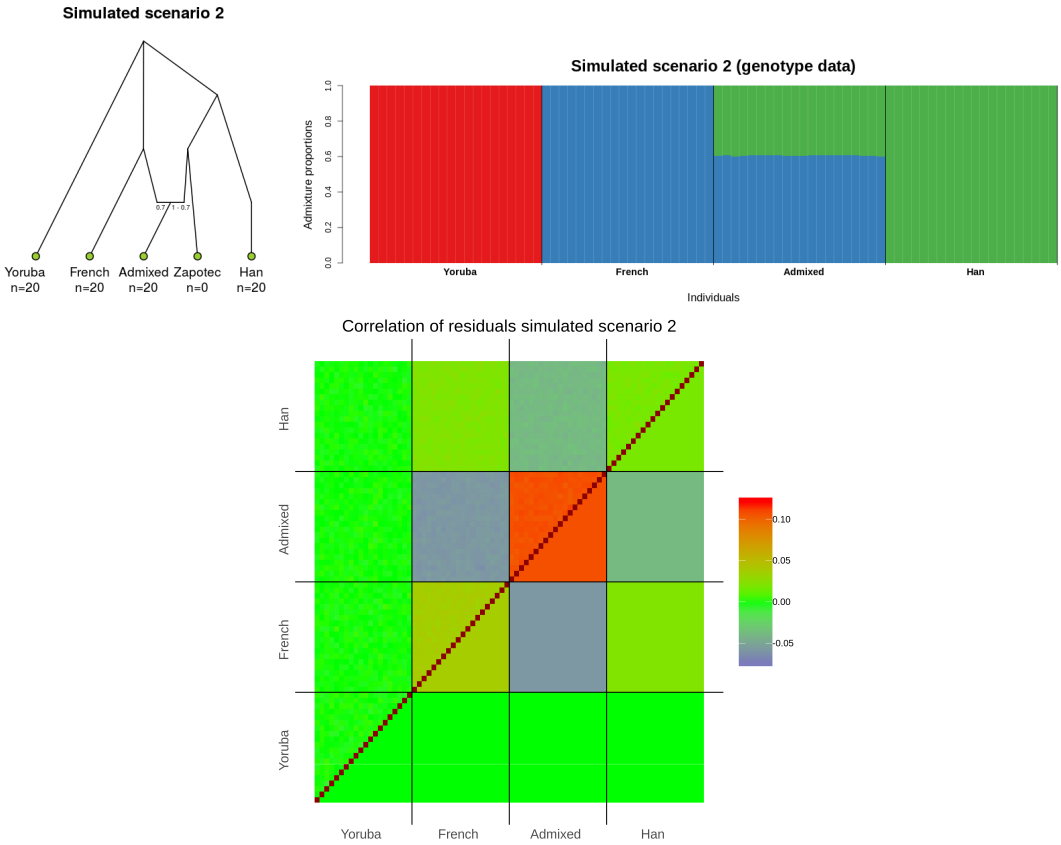

Figure 2: Population tree of the simulated demographic history in Scenario 2 (upper left panel), inferred admixture proportions assuming $K = 3$ from genotype data using ADMIXTURE (upper right panel) and evaluation of model fit of the admixture results as the correlation of residuals (bottom panel). A positive correlation within the admixed population, and to a lesser extend in the two source populations, indicate a bad fit of the inferred admixture proportions.

3.2 BadMIXTURE simulations

The simulations used in the BadMIXTURE paper comprise three scenarios with different demographic histories that however when an admixture analysis is applied result in nearly identical inferred admixture proportions (Figure 3). After running admixture analysis in the same datasets, we calculated the correlation of residuals for each scenario. In the Bottleneck scenario, a recent bottleneck in Pop1 results in Pop2 being modelled as admixed despite the absence of any gene flow. We can detect this bad model fit as a positive correlation of residuals between individuals from Pop2 (Figure 3). In the Ghost admixture scenario, Pop2 has received admixture from an unsampled population. Again a positive correlation of residuals shows that the admixture results are not an accurate representation of its demographic history. In the Recent admixture scenario, Pop2 is the result of an admixture event between Pop1, Pop2 and Pop3 in proportions similar to the inferred ones. In this case we find uncorrelated residuals between individuals from Pop2 (Figure 3). In this scenario there are two groups of 15 and 25 individuals from Pop3 that have correlated residuals between them, suggesting the presence of population substructure. The signal was attenuated after removing variants in LD from the dataset (Figure 3, compared to Supplementary Figure S3). Because nothing in their reported demographic history suggests that Pop3 should have any kind of substructure, this signal is most likely due to a simulation artifact.

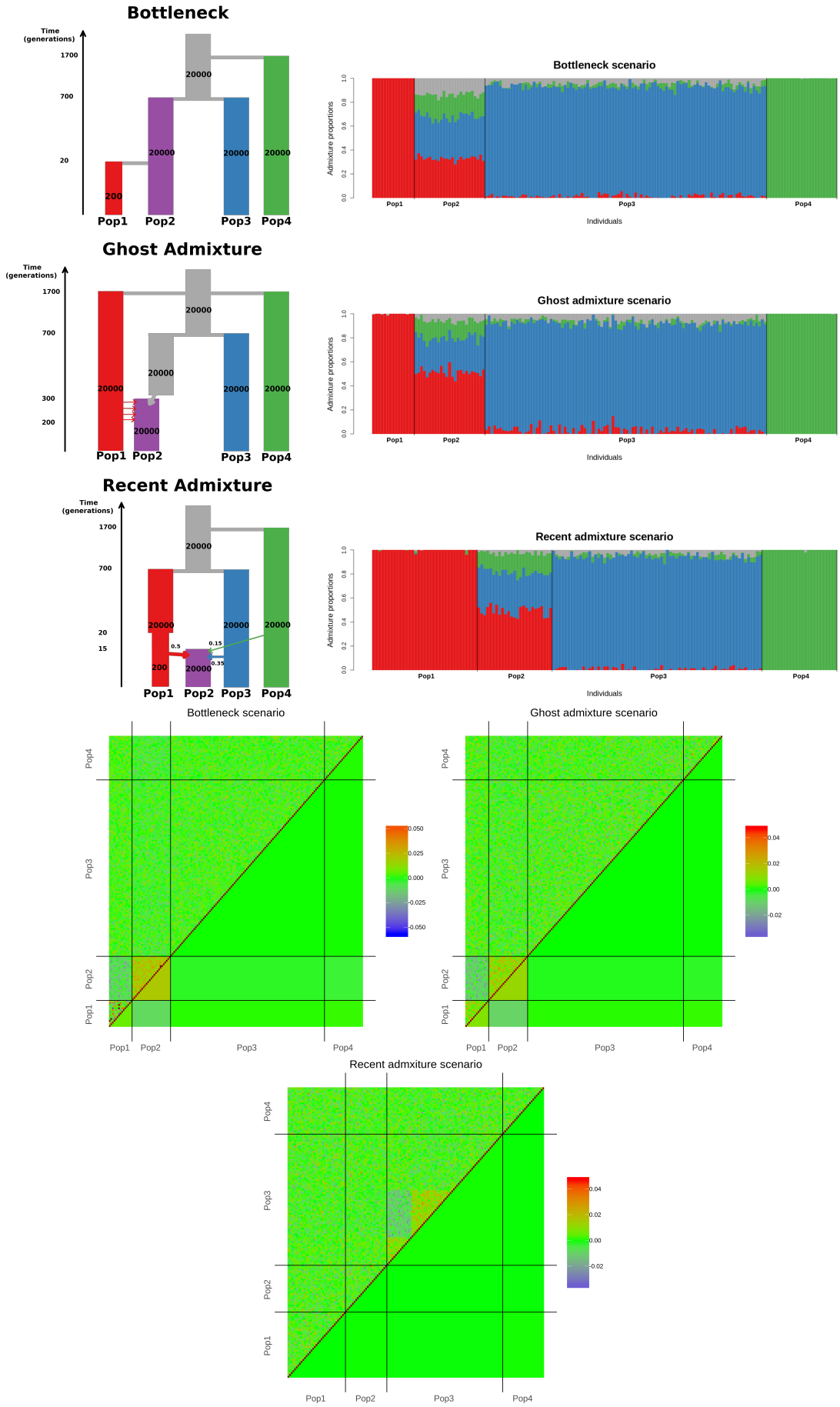

Figure 3: Demographic models of each of the three simulated scenarios from the badMIXTURE paper, including only the four populations of interest (three upper left panels, time and population sizes are not scaled), with the corresponding inferred admixture proportions assuming $K = 11$, colouring only the three relevant ancestral populations, and pooling together the rest as grey (three upper right panels). Evaluation of the estimated admixture proportions for each scenario as the correlation of residuals.

3.3 1000 Genomes

We performed admixture analyses to a subset of 5 populations (YRI, ASW, CEU, MXL and CHB, see methods) from the 1000 Genomes project, assuming 3 and 4 ancestral clusters. In the case with $K = 3$, the inferred clusters correspond to African, European and East Asian ancestral populations. The Native American ancestry in the MXL population is modelled as East Asian ancestry, which results in a positive correlation of the residuals between individuals from that populations, with magnitude varying depending on the amount of Native American ancestry the individuals have (Figure 4). Moreover, there are a few individuals from the ASW population with Native American ancestry, who have also a positive correlation with MXL individuals. When doing the analysis with $K = 4$, a cluster corresponding to the Native American ancestry is added. In this case, individuals with Native American ancestry have uncorrelated residuals (Figure 4). In none of the population there is a visible mean correlation when using a normal scale. However, when increasing the resolution we find a weak signal in the ASW population (Supplementary Figure S4), suggesting that the YRI and CEU populations are not a perfect proxy for the ancestry of African Americans.

There are a few individuals that show a highly positive correlation of their residuals with the same magnitude at both $K = 3$ and $K = 4$. The majority of these cases, and also the ones with the strongest signal, are from the ASW population (Figure 4). We inferred relatedness using relateAdmix, and found five individuals from the ASW population with a kinship coefficient of 0.25 and one with a kinship coefficient of 0.12. These six pairs were the same that had the highest correlation of residuals (Figure 5).

We ran the analyses using both genotype data and sequencing data, obtaining in both cases identical inferred admixture proportions and correlation of residuals (Supplementary Figure S5).

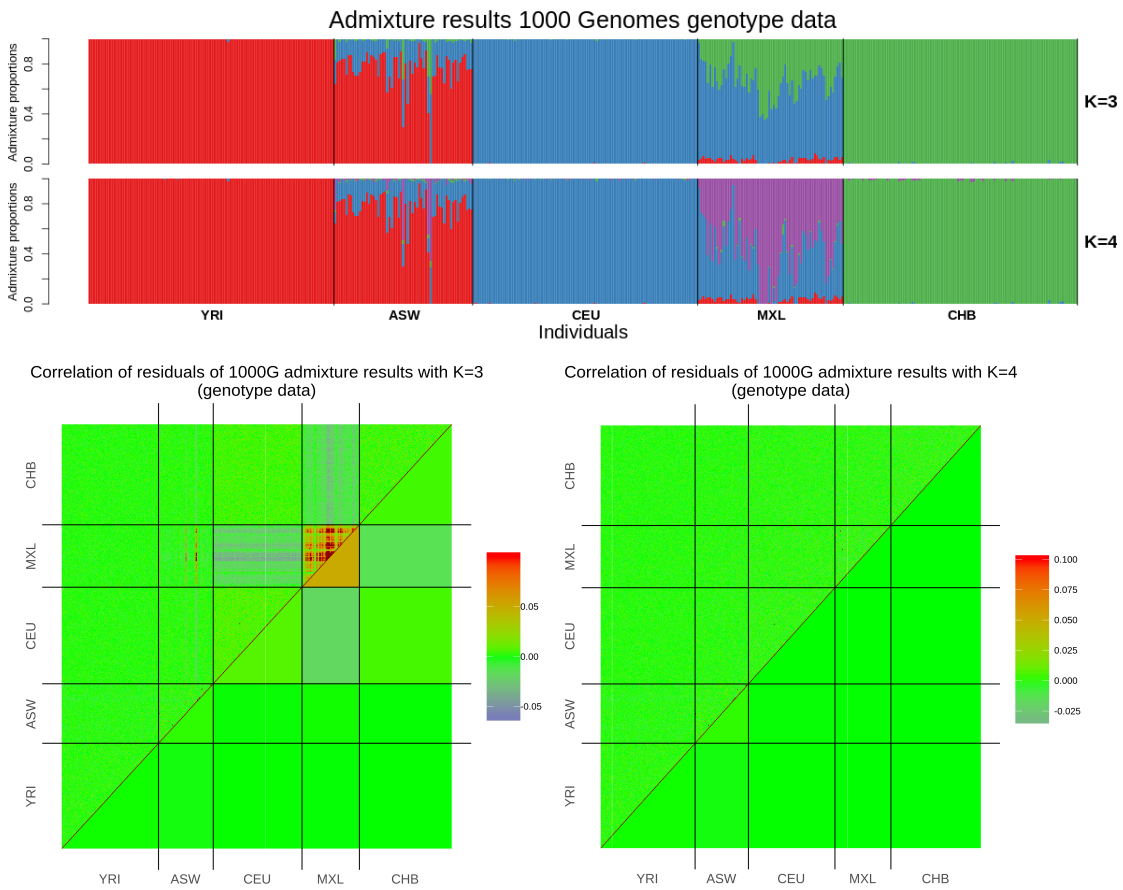

Figure 4: Inferred admixture proportions of the genotype data from the 1000 Genomes dataset, assuming $K = 3$ and $K = 4$ (upper panel). Evaluation of the estimated admixture proportions as correlation of residuals in each case, showing that four ancestral populations are needed to accurately model the data (bottom right panel, in contrast to bottom left panel). Correlation values higher than 0.1 are plotted as dark red.

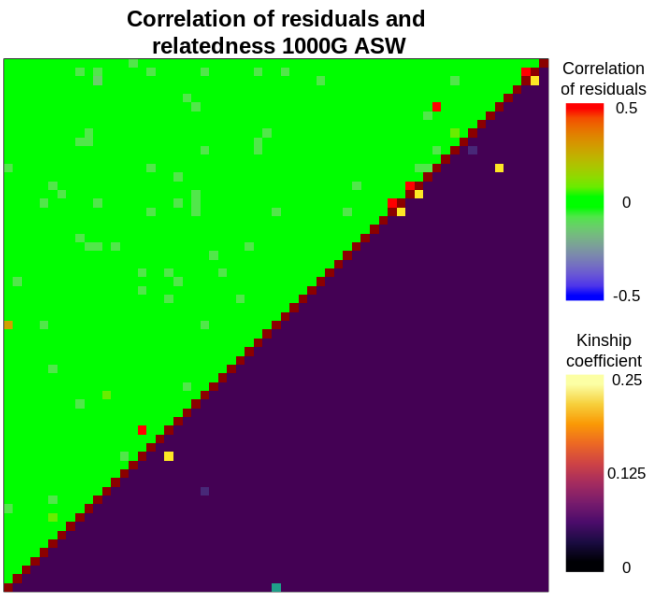

Figure 5: Correlation of residuals between the ASW individuals from the 1000 Genomes dataset with admixture proportions inferred assuming $K = 4$ (above diagonal) and kinship coefficients between the same individuals (below diagonal). The pairs with highest correlation show also the highest kinship coefficient.

3.4 Implementation

The method presented here has been implemented as evalAdmix in C++, and can be run multithreaded. The time complexity is $\mathcal{O}(N^2MK)$, It is quadratic with respect to the number of individuals since it has to estimate ancestral allele frequencies N times in contrast which ADMIXTURE which is linear, $\mathcal{O}(NMK)$. However, it is still fast enough to run on thousands of samples (Supplementary Figure S6). The implementation is freely available at `https://github.com/GenisGE/evalAdmix`.

4 Discussion

We have introduced a method to evaluate the model fit of admixture proportions, and we have shown how it can detect several cases of violations of the model assumptions, such as assuming a too low number of ancestral clusters, drift between recently diverged populations, admixture from ghost populations and the presence of related individuals in the sample. In each case, the magnitude of the correlation varies. Above what threshold can it be considered that there is a bad model fit is a question that needs addressing. However, we do not think it possible to give a general binary answer to that question. Real data will never have a perfect fit to the admixture model assumptions, so the question becomes when is the model fit good enough. Our analyses with the 1000G data provide a relevant example. While with $K = 3$ it is clear that we are missing an ancestral cluster to model the Native American ancestry, at $K = 4$ we can still detect a certain correlation in the mean within the ASW population, probably reflecting that there is more diversity in the Afrian and European admixture sources of the African Americans than what is captured using only the YRI and CEU populations as sources [22]. However, the correlation is very low and we need to reduce the scale of the visualization to a level in which the signal of

individual correlations becomes mostly noise. For many purposes, we could consider that this signal is negligible enough so that the results with four ancestral clusters can be seen a good approximation.

Increasing the number of ancestral clusters K means using more parameters in the calculation of the individual allele frequencies, which will always lead to a decrease in the total mean correlation of residuals. What our method can provide therefore is a lower bound in the number of ancestral populations needed to model the data, as the smallest value of K from which the resulting correlation is good enough. In contrast, most procedures proposed to select K, such as *structure*'s model evidence [2], ADMIXTURE's cross-validation error [23] or the ΔK statistic [7], focus on avoiding overfitting and consequently give an upper bound to K, which can lead to underestimating the amount of population substructure present in the data [10]. The presence of population substructure within clusters would be detected by the correlation of residuals as long as there are at least 2 individuals from the subpopulation. This can also help to identify biases in ancestral cluster assignments by uneven sample sizes, that can cause highly diverged populations not to be assigned their own cluster if too few individuals are present in the sample [24]. Another potential source of bias in admixture analysis is the presence of related individuals [25]. Relatedness leads to non-independence of genotypes between individuals, which as we have shown can also be detected by the correlation of residuals. The five ASW individual pairs with the highest correlation had been previously identified to be parent-offspring, and the sixth pair with a moderate correlation to be aunt/niece [26]. It is worth noting that even though the admixture model is violated it does not change the interpretation of the results of the analysis. In this case there is enough African and European ancestry in the analysis that the few pairs of related individuals do not create sub clusters or otherwise interfere with the inferred admixture proportions.

An excessive focus on the choice of K can lead to the implicit assumption that the demographic history of the sample can be represented as a recent mixture of divergent ancestral populations. In most cases, this will not be true. However, the admixture proportions $\hat{Q}$ will in any case be used by the model to describe variation in the dataset [11]. Therefore, the values in $\hat{Q}$ that we call admixture proportions will not always be an estimate of the proportion of each individual's genome that derives from different source populations. A good fit of the data to the admixture model is needed in order for $\hat{Q}$ to be taken as representing actual admixture proportions.

The method that we propose only requires the input and output used for an admixture analysis, which makes it very simple to integrate them. The runtime migh in some cases be a problem, since the need to correct the frequencies makes the complexity to increase quadratically with the number of individuals. In most cases it is still reasonably fast, specially if the program can be ran using several CPUs. However, if many thousands of individuals are included in the analysis and the chosen K is high then it can become a problem. For this cases, we have implemented the option to calculate the correlation only between a selected set of individuals of interest. Additionally, a proportion of the total sites can be specified to be randomly selected, discarding the rest. This decreases linearly the runtime, although if too few sites are left the resolution will also decrease.

We believe the tool we propose can be of great utility in the interpretation of admixture analyses results. It can provide additional guidance in the persistent problem of choosing an appropriate value of K and allows for an assessment of the model fit at the individual level. It is in this sense very similar to badMIXTURE [11], but its lower requirements and ease of use make it an attractive alternative that can be applied in a broader range of situations.

Acknowledgement

We would like to thank Daniel Falush for the idea for this project which stems from conversations with him at the 2018 Workshop on Population and Speciation Genomics, Cesky Krumlov. We would also like to thank Jonas Meisner for his help in performing the simulations, and Emil Jørsboe for his help in performing the simulations and in solving problems with the implementation. This work was supported by a grant from the Lundbeck Foundation (R215-2015-4174).

References

[1] D. J. Lawson et al. "Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity?" In: *Hum. Genet.* (2019).

[2] J. K. Pritchard, M. Stephens, and P. Donnelly. "Inference of population structure using multilocus genotype data". In: *Genetics* 155.2 (2000), pp. 945–959.

[3] H. Tang et al. "Estimation of individual admixture: analytical and study design considerations". In: *Genet. Epidemiol.* 28.4 (2005), pp. 289–301.

[4] D. H. Alexander, J. Novembre, and K. Lange. "Fast model-based estimation of ancestry in unrelated individuals". In: *Genome Res.* 19.9 (2009), pp. 1655–1664.

[5] L. Skotte, T. S. Korneliussen, and A. Albrechtsen. "Estimating individual admixture proportions from next generation sequencing data". In: *Genetics* 195.3 (2013), pp. 693–702.

[6] D. Falush, M. Stephens, and J. K. Pritchard. "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies". In: *Genetics* 164.4 (2003), pp. 1567–1587.

[7] G. Evanno, S. Regnaut, and J. Goudet. "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study". In: *Mol. Ecol.* 14.8 (2005), pp. 2611–2620.

[8] R. Verity and R. A. Nichols. "Estimating the Number of Subpopulations (K) in Structured Populations". In: *Genetics* 203.4 (2016), pp. 1827–1839.

[9] J. Novembre. "Pritchard, Stephens, and Donnelly on Population Structure". In: *Genetics* 204.2 (2016), pp. 391–393.

[10] J. K. Janes et al. "The K = 2 conundrum". In: *Mol. Ecol.* 26.14 (2017), pp. 3594–3602.

[11] D. J. Lawson, L. van Dorp, and D. Falush. "A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots". In: *Nat Commun* 9.1 (2018), p. 3258.

[12] D. J. Lawson et al. "Inference of population structure using dense haplotype data". In: *PLoS Genet.* 8.1 (2012), e1002453.

[13] H. M. Cann et al. "A human genome diversity cell line panel". In: *Science* 296.5566 (2002), pp. 261–262.

[14] S. Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *Am. J. Hum. Genet.* 81.3 (2007), pp. 559–575.

[15] J. Meisner and A. Albrechtsen. "Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data". In: *Genetics* 210.2 (2018), pp. 719–731.

[16] L. van Dorp et al. "Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference". In: *PLoS Genet.* 11.8 (2015), e1005397.

[17] A. Auton et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74.

[18] I. Lazaridis et al. "Ancient human genomes suggest three ancestral populations for present-day Europeans". In: *Nature* 513.7518 (2014), pp. 409–413.

[19] I. Moltke and A. Albrechtsen. "RelateAdmix: a software tool for estimating relatedness between admixed individuals". In: *Bioinformatics* 30.7 (2014), pp. 1027–1028.

[20] T. S. Korneliussen, A. Albrechtsen, and R. Nielsen. "ANGSD: Analysis of Next Generation Sequencing Data". In: *BMC Bioinformatics* 15 (2014), p. 356.

[21] A. McKenna et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". In: *Genome Res.* 20.9 (2010), pp. 1297–1303.

[22] S. A. Tishkoff et al. "The genetic structure and history of Africans and African Americans". In: *Science* 324.5930 (2009), pp. 1035–1044.

[23] D. H. Alexander and K. Lange. "Enhancements to the ADMIXTURE algorithm for individual ancestry estimation". In: *BMC Bioinformatics* 12 (2011), p. 246.

[24] S. J. Puechmaille. "The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem". In: *Mol Ecol Resour* 16.3 (2016), pp. 608–627.

[25] E. C. Anderson and K. K. Dunham. "The influence of family groups on inferences made with the program Structure". In: *Mol Ecol Resour* 8.6 (2008), pp. 1219–1229.

[26] S. Gazal et al. "High level of inbreeding in final phase of 1000 Genomes Project". In: *Sci Rep* 5 (2015), p. 17453.

Supplementary Material

S1 Expected correlation using sample frequencies

We assume that we have N diploid genotypes that are sampled from the same panmictic population k with minor allele frequency f_k. Genotypes are the minor allele count, so $g \in \{0, 1, 2\}$ and the expected genotype is $\mathbb{E}[g] = 2f$. We use the sampled genotypes to calculate the sample frequency $\hat{f}_k = \frac{1}{2N}\sum_i^N g_i$; with $\mathbb{E}[\hat{f}_k] = f_k$.

The residual r is the difference between the true sampled genotype and the genotype predicted from the estimated sample frequency; so individual 1's residual is

$$r_1 = g_1 - \mathbb{E}[G|\hat{f}_k] = g_1 - \frac{1}{N}\sum_{i=1}^{N} g_i. \tag{1}$$

What we are interested in is the correlation between the residuals of pairs of individuals. Taking individuals 1 and 2, since for any i $\mathbb{E}[r_i] = 0$, we can write the expected correlation $Cor(r_1, r_2) = \rho_{12}$ as

$$\mathbb{E}[\rho_{12}] = \frac{\mathbb{E}[r_1 r_2]}{\sqrt{\mathbb{E}[r_1 r_1]}\sqrt{\mathbb{E}[r_2 r_2]}} = \frac{\mathbb{E}[(g_1 - \frac{1}{N}\sum_i^N g_i)(g_2 - \frac{1}{N}\sum_i^N g_i)]}{\sqrt{\mathbb{E}[(g_1 - \frac{1}{N}\sum_i^N g_i)(g_1 - \frac{1}{N}\sum_i^N g_i)]}\sqrt{\mathbb{E}[(g_2 - \frac{1}{N}\sum_i^N g_i)(g_2 - \frac{1}{N}\sum_i^N g_i)]}}$$
$$= \frac{\mathbb{E}[g_1 g_2] - \mathbb{E}[g_1 \frac{1}{N}\sum_i^N g_i] - \mathbb{E}[g_2 \frac{1}{N}\sum_i^N g_i] + \mathbb{E}[(\frac{1}{N}\sum_i^N g_i)(\frac{1}{N}\sum_i^N g_i])}{\sqrt{\mathbb{E}[g_1 g_1] - 2\mathbb{E}[g_1 \frac{1}{N}\sum_k^N g_i]\mathbb{E}[(\frac{1}{N}\sum_i^N g_i)(\frac{1}{N}\sum_i^N g_i])}\sqrt{\mathbb{E}[g_2 g_2] - 2\mathbb{E}[g_2 \frac{1}{N}\sum_i^N g_i]\mathbb{E}[(\frac{1}{N}\sum_i^N g_i)(\frac{1}{N}\sum_i^N g_i])}}. \tag{2}$$

We can then expand the expectations of products of sums over genotypes such that

$$\mathbb{E}[g_1 \frac{1}{N}\sum_i^N g_i] = \frac{1}{N}\mathbb{E}[g_1 g_1 + \sum_{j=2}^N g_1 g_j] = \frac{1}{N}(\mathbb{E}[g_1 g_1] + \sum_{j=2}^N \mathbb{E}[g_1 g_j]) \tag{3}$$

and

$$\mathbb{E}[\frac{1}{N}\sum_i^N g_i \frac{1}{N}\sum_i^N g_i] = \frac{1}{N^2}\mathbb{E}[\sum_i^N g_i g_i + \sum_i^N \sum_{j\neq i}^N g_i g_j] = \frac{1}{N^2}(\sum_i^N \mathbb{E}[g_i g_j] + \sum_i^N \sum_{j\neq i}^N \mathbb{E}[g_i g_j]). \tag{4}$$

When assuming that genotypes are drawn from the same population, we can assume them to be IID, and therefore

$$\begin{aligned} \mathbb{E}[g_i] = \mathbb{E}[g_j] = \mathbb{E}[g], \\ \mathbb{E}[g_i g_i] = \mathbb{E}[g_j g_j] = \mathbb{E}[gg], \\ \mathbb{E}[g_i g_j] = \mathbb{E}[g_i]\mathbb{E}[g_j] = \mathbb{E}[g]\mathbb{E}[g], \end{aligned} \tag{5}$$

where g is any genotype of an individual from the same population. When applying this to expansion (3), we can develop it further and denote its value as d

$$\frac{1}{N}(\mathbb{E}[g_1 g_1] + \sum_{j=2}^N \mathbb{E}[g_1 g_j]) = \frac{1}{N}(\mathbb{E}[gg] + (N-1)\mathbb{E}[g]\mathbb{E}[g]) = \frac{1}{N}(\mathbb{E}[gg] - \mathbb{E}[g]\mathbb{E}[g]) + \mathbb{E}[g]\mathbb{E}[g] = d. \tag{6}$$

Similarly (4) will have the same value d

$$\begin{aligned} \frac{1}{N^2}(\sum_i^N \mathbb{E}[g_i g_i] + \sum_i^N \sum_{j\neq i}^N + \mathbb{E}[g_i g_j]) &= \frac{1}{N^2}(N\mathbb{E}[gg] + N(N-1)\mathbb{E}[g]\mathbb{E}[g]) \\ &= \frac{1}{N}(\mathbb{E}[gg] - \mathbb{E}[g]\mathbb{E}[g]) + \mathbb{E}[g]\mathbb{E}[g] = d. \end{aligned} \tag{7}$$

Going back to the correlation, and initially denoting expectations of products of sums by $d = \frac{1}{N}(\mathbb{E}[gg] - \mathbb{E}[g]\mathbb{E}[g]) + \mathbb{E}[g]\mathbb{E}[g]$:

$$\begin{aligned}
\mathbb{E}[\rho_{12}] &= \frac{\mathbb{E}[g]\mathbb{E}[g] - d - d + d}{(\sqrt{\mathbb{E}[gg] - 2d + d})^2} = \frac{\mathbb{E}[g]\mathbb{E}[g] - d}{\mathbb{E}[gg] - d} \\
&= \frac{\mathbb{E}[g]\mathbb{E}[g] - \frac{1}{N}(\mathbb{E}[gg] - \mathbb{E}[g]\mathbb{E}[g]) - \mathbb{E}[g]\mathbb{E}[g]}{\mathbb{E}[gg] - \frac{1}{N}(\mathbb{E}[gg] - \mathbb{E}[g]\mathbb{E}[g]) - \mathbb{E}[g]\mathbb{E}[g]} \\
&= \frac{-\frac{1}{N}(\mathbb{E}[gg] - \mathbb{E}[g]\mathbb{E}[g])}{\mathbb{E}[gg] - \mathbb{E}[g]\mathbb{E}[g] - \frac{1}{N}(\mathbb{E}[gg] - \mathbb{E}[g]\mathbb{E}[g])} \\
&= -\frac{\frac{1}{N}}{1 - \frac{1}{N}} = -\frac{1}{N-1}.
\end{aligned} \tag{8}$$

Thus for unadmixed individuals in the same population we expect a negative correlation of the residuals.

S2 Expected correlation with frequency correction

We can make the residuals within one discrete population independent if we remove the contribution of one of the individuals from the frequency used in the calculation of the other individual's residual. So to calculate the correlation between individuals 1 and 2, we can use for individual 1 the sample frequency $\hat{f}_k = \frac{1}{2N}\sum_i^N g_i$, while for individual 2 we use

$$\tilde{f}_{k,-1} = \frac{1}{2(N-1)}\sum_{i=2}^{N} g_i. \tag{9}$$

The residuals are then

$$\begin{aligned}
r_1 &= g_1 - 2\hat{f}_k = g_1 - \frac{1}{N}\sum_{i}^{N} g_i, \\
r_{2,-1} &= g_2 - 2\tilde{f}_{k,-1} = g_2 - \frac{1}{N-1}\sum_{i=2}^{N-1} g_i,
\end{aligned} \tag{10}$$

and the expected correlation $Cor(r_1, r_2) = \rho_{12}$ is

$$\begin{aligned}
\mathbb{E}[\rho_{12}] &= \frac{\mathbb{E}[r_1 r_2]}{\sqrt{\mathbb{E}[r_1 r_2]}\sqrt{\mathbb{E}[r_2 r_2]}} = \frac{\mathbb{E}[g_1 - \frac{1}{N}\sum_i^N g_i)(g_2 - \frac{1}{N-1}\sum_{i=2}^N g_i)]}{\sqrt{\mathbb{E}[g_1 - \frac{1}{N}\sum_i^N g_i)(g_1 - \frac{1}{N}\sum_i^N g_i)]}\sqrt{\mathbb{E}[(g_2 - \frac{1}{N-1}\sum_{i=2}^N g_i)(g_2 - \frac{1}{N-1}\sum_{i=2}^N g_i)]}} \\
&= \frac{\mathbb{E}[g_1 g_2] - \mathbb{E}[g_1 \frac{1}{N-1}\sum_{i=2}^N g_i] - \mathbb{E}[g_2 \frac{1}{N}\sum_i^N g_i] + \mathbb{E}[(\frac{1}{N}\sum_i^N g_i)(\frac{1}{N-1}\sum_{i=2}^N g_i)]}{\sqrt{\mathbb{E}[g_1 g_1] - 2\mathbb{E}[g_1 \frac{1}{N}\sum_i^N g_i]\mathbb{E}[(\frac{1}{N}\sum_i^N g_i)(\frac{1}{N}\sum_i^N g_i)]}\sqrt{\mathbb{E}[g_2 g_2] - 2\mathbb{E}[g_2 \frac{1}{N-1}\sum_{i=2}^N g_i]\mathbb{E}[(\frac{1}{N-1}\sum_{i=2}^N g)(\frac{1}{N-1}\sum_{i=2}^N g_i)]}}.
\end{aligned} \tag{11}$$

To prove that in this case $\mathbb{E}[\rho_{12}] = 0$, we only need to show that the expected value of the numerator is 0, since the denominator is the product of two variances, which by definition will have a positive value. There are three expectations of products of sums over genotypes to expand, one of them already given by equation 3, which we denoted as d,

$$\mathbb{E}[g_2 \frac{1}{N} \sum_{i}^{N} g_i] = \frac{1}{N}(\mathbb{E}[gg] - \mathbb{E}[g]\mathbb{E}[g]) + \mathbb{E}[g]\mathbb{E}[g] = d. \tag{12}$$

Then we have two new expansions

$$\mathbb{E}[g_1 \frac{1}{N-1} \sum_{i=2}^{N} g_i] = \frac{1}{N-1} \sum_{i=2}^{N} \mathbb{E}[g_1 g_j] = \frac{1}{N-1}(N-1)\mathbb{E}[g]\mathbb{E}[g] = \mathbb{E}[g]\mathbb{E}[g], \tag{13}$$

and

$$\begin{aligned} \mathbb{E}[(\frac{1}{N} \sum_{i}^{N} g_i)(\frac{1}{N-1} \sum_{i=2}^{N} g_i)] &= \frac{1}{N(N-1)}(\sum_{i=2}^{N} \mathbb{E}[g_i g_i] + \sum_{i=2}^{N} \sum_{j=1;j\neq i}^{N} \mathbb{E}[g_i]\mathbb{E}[g_j]) \\ &= \frac{1}{N}(\mathbb{E}[gg] + (N-1)\mathbb{E}[g]\mathbb{E}[g]) = \frac{1}{N}(\mathbb{E}[gg] - \mathbb{E}[g]\mathbb{E}[g]) + \mathbb{E}[g]\mathbb{E}[g] = d \end{aligned} \tag{14}$$

We denote the non zero denominator as A, and by substituting with what we have shown in equations 12, 13 and 14, we find that the expected correlation is

$$\begin{aligned} \mathbb{E}[\rho_{12}] = \frac{\mathbb{E}[g_1 g_2] - \mathbb{E}[g_1 \frac{1}{N-1} \sum_{i=2}^{N} g_i] - \mathbb{E}[g_2 \frac{1}{N} \sum_{i}^{N} g_i] + \mathbb{E}[(\frac{1}{N} \sum_{i}^{N} g_i)(\frac{1}{N-1} \sum_{i=2}^{N} g_i)]}{A} \\ = \frac{\mathbb{E}[g]\mathbb{E}[g] - d - \mathbb{E}[g]\mathbb{E}[g] + d}{A} = 0. \end{aligned} \tag{15}$$

Thus if we remove the contribution of individual 1 from individual's 2 frequency, their expected residuals will be independent.

S3 Supplementary Figures

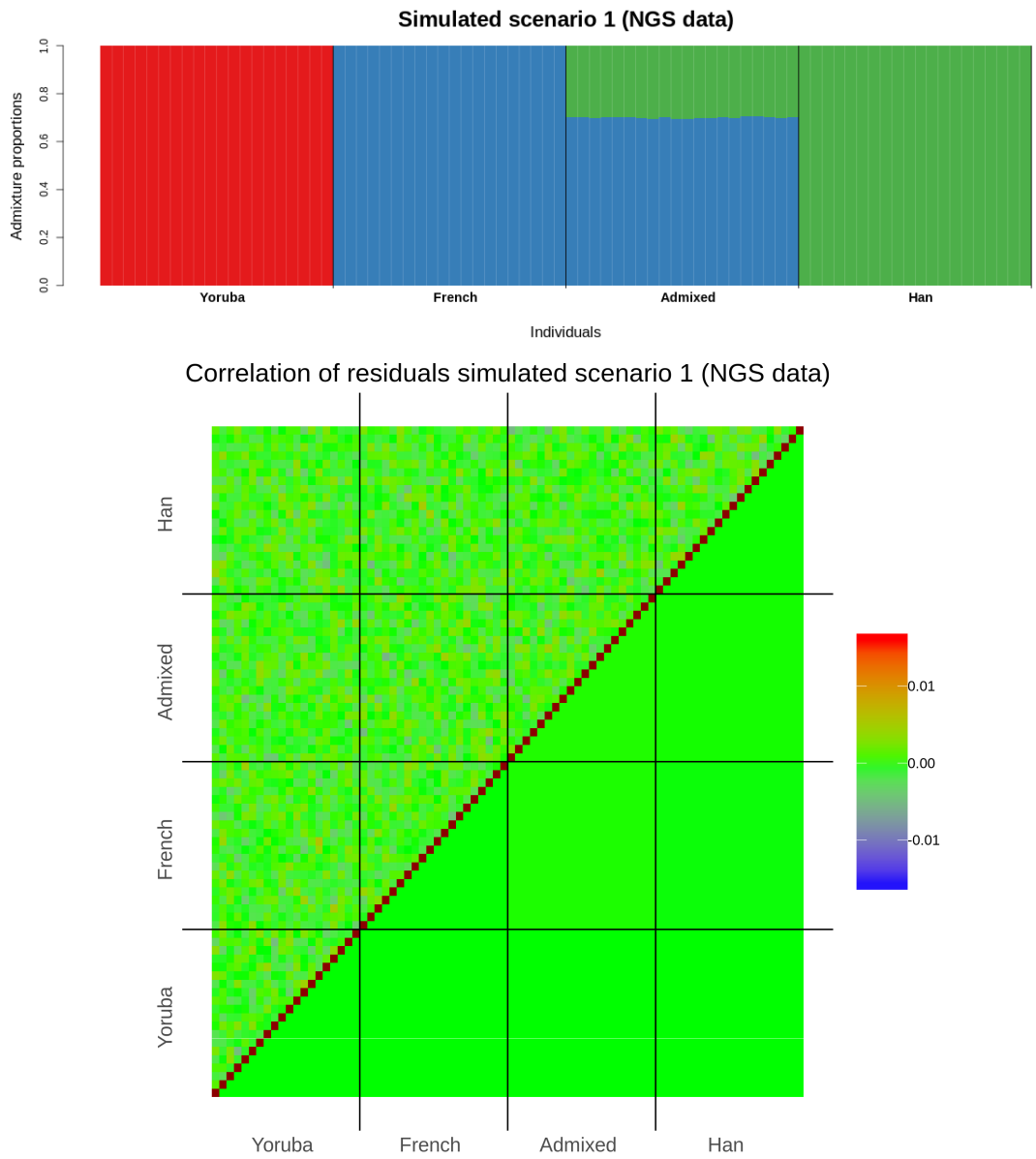

Figure S1: Admixture proportions inferred with NGSadmix assuming $K = 3$ from genotype likelihood data for simulated Scenario 1 (upper panel) and evaluation of model fit as correlation of residuals (bottom panel.

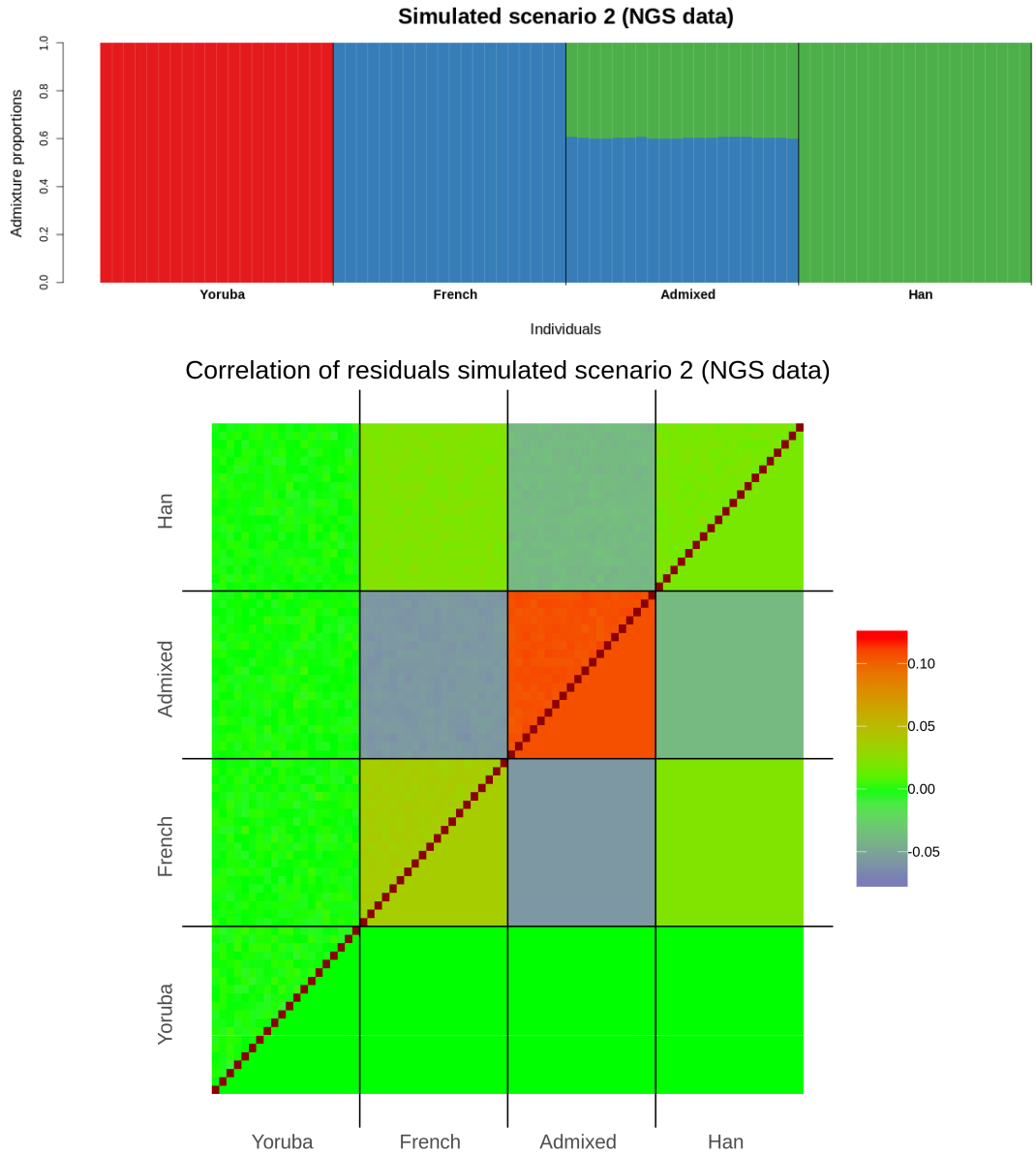

Figure S2: Admixture proportions inferred with NGSadmix assuming $K = 3$ from genotype likelihood data for simulated Scenario 2 (upper panel) and evaluation of model fit as correlation of residuals (bottom panel

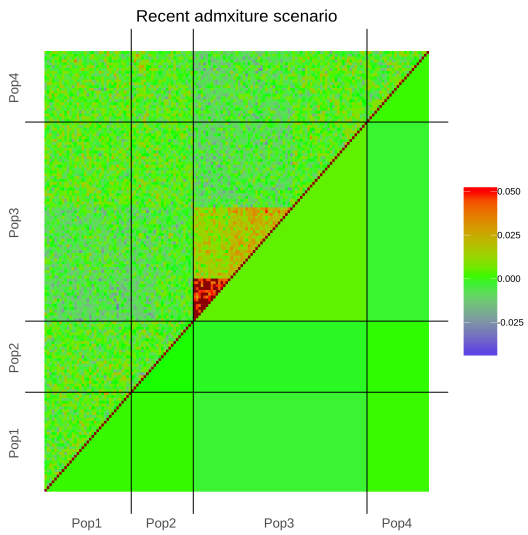

Figure S3: Evaluation of the Recent Admixture simulated scenario as correlation of residuals, without filtering out variants in LD, which results in an increased correlation between the subgroups of 15 and 25 individuals from Pop3. Correlation values higher than 0.1 are plotted as dark red.

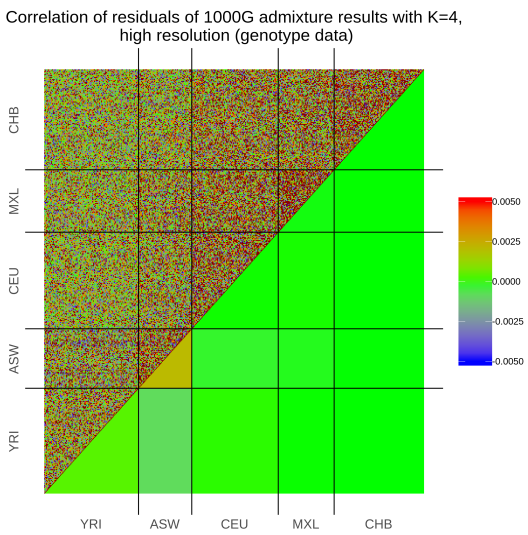

Figure S4: Evaluation of the inferred admixture proportions of the 1000G genotype data, assuming $K = 4$ and reducing the scale of the heatmap to highlight the presence of a positive correlation in the mean within ASW individuals. Correlation values higher than 0.1 are plotted as dark red.

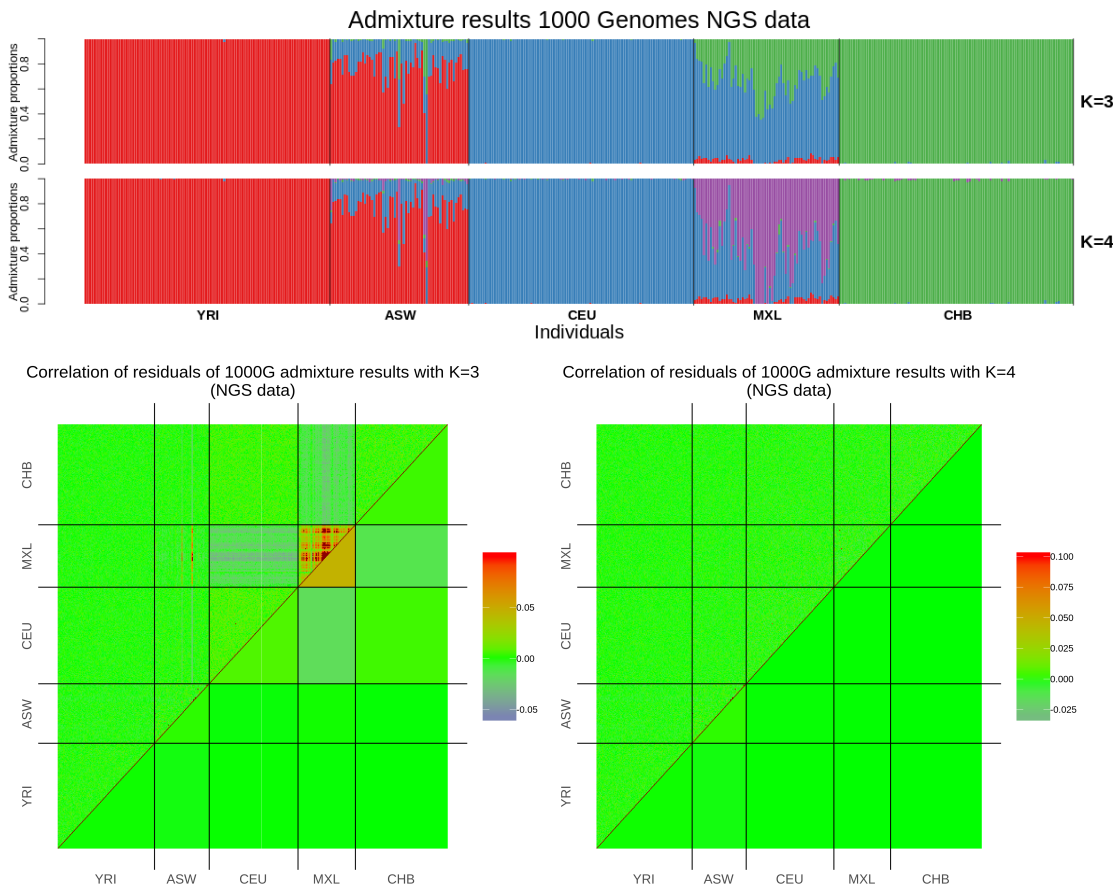

Figure S5: Inferred admixture proportions of the genotype likelihood data from the 1000G dataset, assuming $K = 3$ and $K = 4$ (upper panel). Evaluation of the estimated admixture proportions as correlation of residuals in each case, showing that four ancestral populations are needed to accurately model the data (bottom right panel, in contrast to bottom left panel). Correlation values higher than 0.1 are plotted as dark red.

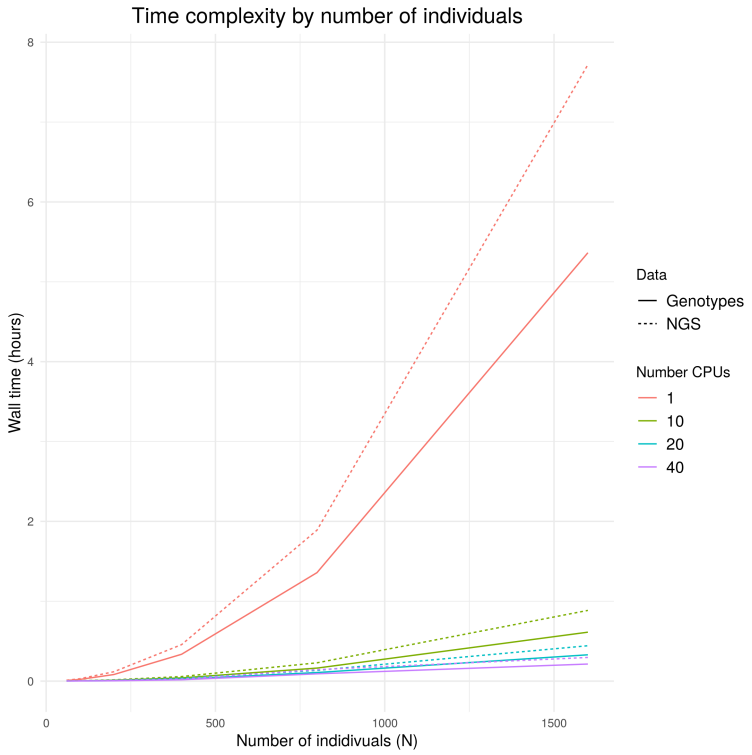

Figure S6: Total runtime of the genotype and NGS version of the program, with different number of individuals (N), and using 1, 10, 20 or 40 threads. In all cases the number of sites (M) is 100000 and the number of ancestral populations (K) is 6.