Inferring Population Structure and Admixture Proportions in Low Depth Next-Generation Sequencing Data

Jonas Meisner & Anders Albrechtsen

16th April 2018

Abstract

We here present two new methods for inferring population structure and admixture proportions in low depth next-generation sequencing (NGS) data. Inference of population structure is essential in both population genetics and association studies and is often performed using principal component analysis (PCA) or clustering-based approaches. NGS methods provide large amounts of genetic data but are associated with statistical uncertainty for especially low depth sequencing data. Probabilistic methods have therefore been employed to account for this uncertainty by working directly on genotype likelihoods of the unobserved genotypes. We propose a new method for inferring population structure through principal component analysis based on an iterative approach of estimating individual allele frequencies, and demonstrate a greatly improved accuracy in samples with low and variable sequencing depth for both simulated and real datasets. At last, we use the estimated individual allele frequencies in a new fast non-negative matrix factorization method to estimate admixture proportions. Both methods have been implemented in the PCAngsd framework available at `http://www.popgen.dk/software/`.

1 Introduction

Population genetic studies often consist of individuals of diverse ancestries, and inference of population structure therefore plays an important role in population genetics and association studies. Population stratification can act as a confounding factor in association studies as it can lead to spurious associations [1]. Principal component analysis (PCA) was first introduced to genetic data in Menozzi et al. (1978) [2] to produce synthetic maps in an exploratory analysis of genetic variation. PCA is now a common tool in population genetic studies, where its dimension reduction properties can be used to visualize genetic data by summarizing the genetic variation through principal components [3] as well as to be used to infer population structure to correct for population stratification in association studies, investigating demographic history [4–6] and performing genome selection scans [7–9]. PCA is an appealing approach to infer population structure as the aim is not to classify the individuals into discrete populations, however instead describe continuous axes of genetic variation such that heterogeneous populations and admixed individuals can be better represented [4]. Another successful approach in modeling complex population structure has been to estimate admixture proportions based on clustering-based methods [10–13], such as the popular software ADMIXTURE, which have also been used for correction of population stratification in association studies [14].

Next-generation sequencing (NGS) methods [15] produce a large amount of reliable DNA sequencing data at low cost and are commonly used in population genetic studies [16]. Many NGS studies are based on medium (<15X) and low (<5X) depth data due the demand for large sample sizes as seen in large-scale sequencing studies, e.g. 1000 Genomes Project Consortium [17, 18]. However, the use of medium and especially low depth sequencing data introduces challenges rooted in the statistical uncertainty induced when calling SNPs and genotypes in these scenarios. The high error rates associated with NGS methods are usually caused by several factors such as sampling, alignment and sequencing errors [16]. The statistical uncertainty increases for low depth samples due to the increased difficulty of distinguishing between a variable site and a sequencing error with the information provided. Chromosomes are also sampled with replacement in the sequencing process and both alleles may therefore not have been sampled for a heterozygous individual in low depth scenarios. Homozygous genotypes may also be wrongly inferred as heterozygous due to sequencing errors. Thus, genotype calling will associate individuals with a statistical uncertainty which should be taken into account [16, 19].

To overcome these problems related to NGS data and genotype calling, probabilistic methods have been developed to take use of genotype likelihoods in combination with external information for various population genetic parameters [5, 13, 16, 20–23], such that posterior genotype probabilities can be used to model the related uncertainty. Genotype likelihoods can be estimated to incorporate errors of the sequencing process such as the base quality scores as well as the allele sampling [24]. These posterior genotype probabilities have also been used to call genotypes with a higher accuracy than previous methods for low depth NGS data [16, 19].

We present two new methods for low depth NGS data using genotype likelihoods to model complex population structure that connect the results of PCA with the admixture proportions of the clustering-based methods. A method has been developed to perform PCA in an iterative approach of estimating individual allele frequencies to compute a covariance matrix, while another method uses the estimated individual allele frequencies in an accelerated non-negative matrix factorization (NMF) approach to estimate admixture proportions. The performances of the two methods are assessed on both simulated and real datasets in regards to existing methods for both low depth NGS and genotype data. The methods have been implemented in a framework called PCAngsd (Principal Component Analysis of Next-Generation Sequencing Data).

2 Methods

We will analyze NGS data of m diploid individuals across n variable sites. These sites will either be known or called single-nucleotide polymorphisms (SNPs), which are assumed to be diallelic such that the major and minor allele of each SNP have been inferred. This can either be done from the sequencing reads [20] or from the genotype likelihoods [21] and only three different genotypes will be possible. Thus, we assume that a genotype G can be seen as a Binomial random variable with realizations 0, 1 and 2 that represent the number of copies of the minor allele in a site for a given individual in the absence of population structure. The expectation and variance of G can therefore be defined as $\mathbb{E}[G] = 2p$ and $\text{Var}[G] = 2p(1-p)$ with p representing the allele frequency of a population, which we also refer to as population allele frequency.

However, genotypes are not observed in NGS data and we will instead work on genotype likelihoods that also include information of the sequencing process. The genotype likelihoods are the probability of the observed sequencing data X given the three different possible genotypes, $P(X \mid G = g)$, for $g = 0, 1, 2$. One method to compute the genotype likelihoods from sequencing reads is described in the supplementary material based on the model in McKenna et al. (2010) [24].

External information can be incorporated to define posterior genotype probabilities using Bayes' theorem in combination with the genotype likelihoods [19]. The population allele frequency is often used as information in the prior genotype probability $P(G_{is} \mid p_s)$, for an individual i in site s [5, 16, 20, 22]. Assuming the population is in Hardy-Weinberg Equilibrium (HWE) for a site s, the population allele frequency is used to define the prior genotype probability such that $P(G_{is} = 0 \mid p_s) = (1 - p_s)^2$, $P(G_{is} = 1 \mid p_s) = 2p_s(1 - p_s)$ and $P(G_{is} = 2 \mid p_s) = p_s^2$ for the three different possible genotypes. Using the estimated population allele frequency $\hat{p}_s$ for computing the posterior genotype probability, $P(G_{is} = g \mid X_{is}, \hat{p}_s)$, such as defined in Kim et al. (2011) [20], is given as follows for individual i in site s:

$$P(G_{is} = g \mid X_{is}, \hat{p}_s) = \frac{P(X_{is} \mid G_{is} = g) P(G_{is} = g \mid \hat{p}_s)}{\sum_{g'=0}^{2} P(X_{is} \mid G_{is} = g') P(G_{is} = g' \mid \hat{p}_s)}. \tag{1}$$

2.1 PCA

The standard way of performing PCA in population genetics and using it to infer population structure is based on the method defined in Patterson et al. (2006) [4]. For a genotype matrix $\mathbf{G}$ of m individuals and n variable sites, the $m \times m$ covariance matrix $\mathbf{C}$, also known as the genetic relationship matrix (GRM), is computed as follows for two individuals i and j:

$$c_{ij} = \frac{1}{n} \sum_{s=1}^{n} \frac{(g_{is} - 2\hat{p}_s)(g_{js} - 2\hat{p}_s)}{2\hat{p}_s(1 - \hat{p}_s)}. \tag{2}$$

Here g_{is} is the observed genotype for individual i in site s to distinguish it from G defined above for unobserved genotypes, and $\hat{p}$ is the estimated population allele frequency. The principal components are then computed by performing an eigendecomposition of the covariance matrix, where $\mathbf{C} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T$ with $\mathbf{V}$ being the matrix of eigenvectors and $\boldsymbol{\Sigma}$ the diagonal matrix of eigenvalues. Principal components and eigenvectors will be used interchangeably throughout this study. The top principal components capture most of the population structure as they represent axes of genetic variation in the dataset [4].

This method has been extended to NGS data in Fumagalli et al. (2013) [5, 25] using the probabilistic framework described above, by summing over the joint posterior genotype probabilities for the two individuals under the assumption of HWE in the whole sample. The method has been implemented in the ngsTools framework [26]. The covariance matrix is estimated as follows for NGS data using only known variable sites for two individuals i and j:

$$c_{ij} = \frac{1}{n} \sum_{s=1}^{n} \frac{\sum_{g_i=0}^{2} \sum_{g_j=0}^{2} (g_i - 2\hat{p}_s)(g_j - 2\hat{p}_s) P(G_{is} = g_i, G_{js} = g_j \mid X_{is}, X_{js}, \hat{p}_s)}{2\hat{p}_s(1 - \hat{p}_s)}. \tag{3}$$

Fumagalli et al. (2013) splits up the joint posterior probability $P(G_{is} = g_i, G_{js} = g_j \mid X_{is}, X_{js}, \hat{p}_s)$ into $P(G_{is} = g_i \mid X_{is}, \hat{p}_s)P(G_{js} = g_j \mid X_{js}, \hat{p}_s)$ for $i \neq j$ by assuming conditional independence between individuals given the estimated population allele frequencies. The non-diagonal entries in the covariance matrix are now directly estimated from the posterior expectations of the genotype instead of the observed genotypes as described in Patterson et al. (2006). The original method by Fumagalli et al. (2013) weighs each site by its probability of being a variable site such that SNP calling is not needed prior to the covariance matrix estimation. This is not taking into account in this study as we are using called variable sites to infer population structure. The population allele frequencies are estimated from the genotype likelihoods using an expectation maximization (EM) algorithm [20] as described in the supplementary material.

The problem with this approach is that the assumption of conditional independence between individuals given the population allele frequency is only valid when there is no population structure. Here we propose a novel approach of estimating the covariance matrix using iteratively estimated individual allele frequencies to update the prior information of the posterior genotype probability. Thereby conditioning on the individual allele frequencies as in the clustering-based approaches.

2.1.1 Individual allele frequencies

A model for estimating individual allele frequencies based on population structure was introduced by Pritchard (2000) [10] as later described in equation 14. Hao et al. (2015) [8] proposed a different model for estimating individual allele frequencies $\mathbf{\Pi}$ using the information in the principal components instead of having an assumption of K ancestral populations. The model is defined as follows,

$$\mathbf{\Pi} = \mathbf{S}\mathbf{A}, \tag{4}$$

where $\mathbf{S}$ represents the population structure such that $\mathbf{A}$ represents the mapping of the population structure $\mathbf{S}$ in the allele frequencies. Hao et al. estimate the individual allele frequencies through a singular value decomposition (SVD) method, where the genotypes are reconstructed using only the top D principal components such that they are modeled by population structure. A similar approach has been proposed in Conomos et al. (2016) [27] where the inferred principal components are used to estimate individual allele frequencies through a simple linear regression model. However, due to working on NGS data and not knowing the genotypes, we are extending their method to NGS data by using the posterior expectations of the genotypes, referred to as genotype dosages, instead of genotypes. Thus we will be using,

$$\mathbb{E}[G_{is} \mid X_{is}, \hat{p}_s] = \sum_{g=0}^{2} g\, P(G_{is} = g \mid X_{is}, \hat{p}_s), \tag{5}$$

for individual i in site s.

The individual allele frequencies are estimated by performing SVD on the centered genotype dosages and reconstructing them using only the top D principal components. In this way the centered genotype dosages are modeled by population structure, which is represented through the top principal components explaining most of the genetic variance in the dataset. $2\hat{\mathbf{p}}$ is then added to the reconstruction and scaled by $\frac{1}{2}$ based on our

Binomial distribution assumption of G_{is}, for $i = 1, \ldots, m$ and $s = 1, \ldots, n$, to produce the individual allele frequencies. Since a SVD is a real valued method, we will have to truncate the estimated individual allele frequencies in order to constrain them in the range $[0, 1]$. However, Hao et al. showed that the resulting estimates were still very accurate considering this limitation. For ease of notation, let $\mathbf{E}$ be the $m \times n$ matrix of genotype dosages, $e_{is} = \mathbb{E}[G_{is} \,|\, X_{is}, \hat{p}_s]$, for $i = 1, \ldots, m$ and $s = 1, \ldots, n$. The following steps for estimating the individual allele frequencies are adopted from the SVD based algorithm of Hao et al. (2015) [8]:

Algorithm 1: SVD based method for estimating individual allele frequencies.

1. The centered genotype dosages are constructed as $\mathbf{E}^{(C)} = \mathbf{E} - 2\hat{\mathbf{p}}$.
2. Perform SVD on the centered genotype dosages, $\mathbf{E}^{(C)} = \mathbf{W\Delta U}^T$, where $\mathbf{W}$ will represent population structure similarly to $\mathbf{V}$.
3. Define $\mathbf{E}_D^{(C)}$ to be the prediction of the centered genotype dosages using only the top D principal components, $\mathbf{E}_D^{(C)} = \mathbf{W}_{1:D}\mathbf{\Delta}_{1:D}\mathbf{U}_{1:D}^T$.
4. Estimate $\hat{\mathbf{\Pi}}$ by adding $2\hat{\mathbf{p}}$ to $\mathbf{E}_D^{(C)}$ row-wise and scaling with $\frac{1}{2}$, based on $\hat{\pi}_{is} \approx \frac{1}{2}\mathbb{E}[G_{is}]$.

For matrix notations define $\hat{\mathbf{S}} = [\mathbf{1}, \mathbf{W}_1, \ldots, \mathbf{W}_D]$ and $\hat{\mathbf{A}}^T = \frac{1}{2}\,[2\hat{\mathbf{p}}, \mathbf{U}_1\delta_1, \ldots, \mathbf{U}_D\delta_D]$, all representing column vectors, such that equation 4 can be approximated as $\hat{\mathbf{\Pi}} = \hat{\mathbf{S}}\hat{\mathbf{A}}$. Finally, $\hat{\mathbf{\Pi}}$ is truncated in order for for allele frequency estimates to be in range $[0, 1]$ based on a small value γ such that,

$$\hat{\pi}_{is} = \begin{cases} \gamma & \text{if } \hat{\pi}_{is} \leq \gamma \\ \hat{\pi}_{is} & \text{if } \gamma \leq \hat{\pi}_{is} \leq 1 - \gamma \\ 1 - \gamma & \text{if } \hat{\pi}_{is} \geq 1 - \gamma. \end{cases} \tag{6}$$

We now incorporate the individual allele frequencies into the estimation of the posterior genotype probabilities. The estimated individual allele frequencies are used as updated prior information instead of the population allele frequencies in the estimation of the prior genotype probabilities. The individual allele frequencies, including information of population structure, will then able to provide a better estimate of the underlying Binomial distribution that genotypes of each individual have been assumed sampled from. Thus, the posterior genotype probabilities are estimated as follows for individual i in site s:

$$P(G_{is} = g \,|\, X_{is}, \hat{\pi}_{is}) = \frac{P(X_{is} \,|\, G_{is} = g)P(G_{is} = g \,|\, \hat{\pi}_{is})}{\sum_{g'=0}^{2} P(X_{is} \,|\, G_{is} = g')P(G_{is} = g' \,|\, \hat{\pi}_{is})}. \tag{7}$$

Each individual are now seen as a single population using the individual allele frequencies as prior information. The prior genotype probability are estimated by assuming HWE such that, $P(G = 0 \,|\, \pi_{is}) = (1 - \pi_{is})^2$, $P(G = 1 \,|\, \hat{\pi}_{is}) = 2(1 - \hat{\pi}_{is})\hat{\pi}_{is}$ and $P(G = 2 \,|\, \hat{\pi}_{is}) = \hat{\pi}_{is}^2$. An updated definition of the posterior expectations of the genotypes are then given as:

$$\mathbb{E}[G \,|\, X_{is}, \hat{\pi}_{is}] = \sum_{g=0}^{2} g\, P(G = g \,|\, X_{is}, \hat{\pi}_{is}). \tag{8}$$

This procedure of updating the prior information can be iterated to estimate new individual allele frequencies on the basis of an updated population structure. Therefore, we propose the following algorithm for an iterative procedure of estimating the individual allele frequencies.

Algorithm 2: Iterative estimation of individual allele frequencies.

1. Estimate population allele frequencies $\hat{\mathbf{p}}$ from genotype likelihoods (See supplementary materials).
2. Estimate posterior genotype probabilities and genotype dosages $\mathbf{E}$ based on genotype likelihoods and $\hat{\mathbf{p}}$.
3. Estimate $\hat{\mathbf{\Pi}}$ using SVD based method on $\mathbf{E}$ as described in Algorithm 1.
4. Estimate posterior genotype probabilities and genotype dosages $\mathbf{E}$ using updated prior information, $\hat{\mathbf{\Pi}}$.
5. Repeat step 3 and 4 until individual allele frequencies have converged.

Convergence of our iterative method is defined as when the root-mean-square deviation (RMSD) of the estimated individual allele frequencies of two successive iterations are smaller than a value δ (5.0×10^{-5}). The RMSD of iteration $t+1$ is defined as,

$$\text{RMSD} = \sqrt{\frac{1}{mn} \sum_{i=1}^{m} \sum_{s=1}^{n} \left(\hat{\pi}_{is}^{(t+1)} - \hat{\pi}_{is}^{(t)} \right)^2}. \tag{9}$$

2.1.2 Covariance matrix

We now use the final set of individual allele frequencies to estimate an updated covariance matrix in a similar model proposed by Fumagalli et al. (2013), but with the individual allele frequencies incorporated into the joint posterior probability of equation 3. The entries of the covariance matrix $\mathbf{C}$ are therefore defined as follow for individuals i and j:

$$c_{ij} = \frac{1}{n} \sum_{s=1}^{n} \frac{\sum_{g_i=0}^{2} \sum_{g_j=0}^{2} (g_i - 2\hat{p}_s)(g_j - 2\hat{p}_s) P(G_i = g_i, G_j = g_j \,|\, X_{is}, X_{js}, \hat{\pi}_{is}, \hat{\pi}_{js})}{2\hat{p}_s(1 - \hat{p}_s)}. \tag{10}$$

For $i \neq j$, the joint posterior probability can be computed as $P(G_i = g_i, \,|\, X_{is}, \hat{\pi}_{is}) P(G_j = g_j, \,|\, X_{js}, \hat{\pi}_{js})$, since the two terms are conditionally independent given the individual allele frequencies in contrary to the assumption made in the model of Fumagalli et al. (2013) using population allele frequencies. The above equation can be expressed in terms of the genotype dosages for ease of notation and computation:

$$\begin{aligned} C_{ij} &= \frac{1}{n}\sum_{s=1}^{n} \frac{\sum_{g_i=0}^{2}\sum_{g_j=0}^{2}(g_i - 2\hat{p}_s)(g_j - 2\hat{p}_s)P(G_i = g_i \,|\, X_{is}, \hat{\pi}_{is})P(G_j = g_j \,|\, X_{js}, \hat{\pi}_{js})}{2\hat{p}_s(1-\hat{p}_s)} \\ &= \frac{1}{n}\sum_{s=1}^{n} \frac{(\mathbb{E}[G_i \,|\, X_{is}, \hat{\pi}_{is}] - 2\hat{p}_s)(\mathbb{E}[G_j \,|\, X_{js}, \hat{\pi}_{js}] - 2\hat{p}_s)}{2\hat{p}_s(1-\hat{p}_s)} . \end{aligned} \tag{11}$$

However for $i = j$ (diagonal of the covariance matrix), the joint posterior probability is simplified to $P(G_i \,|\, X_{is}, \hat{\pi}_{is})$ such that the estimation of the diagonal covariance entries is given as:

$$C_{ii} = \frac{1}{n}\sum_{s=1}^{n} \frac{\sum_{g_i=0}^{2}(g_i - 2\hat{p}_s)^2 P(G_i = g_i \,|\, X_{is}, \hat{\pi}_{is})}{2\hat{p}_s(1-\hat{p}_s)}. \tag{12}$$

An eigendecomposition of the updated estimated covariance matrix is then performed to obtain the principal components as described earlier, $\mathbf{C} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$. Note that $\mathbf{V}$ and $\mathbf{W}$ are not the same even though both represent population structure through axes of genetic variation in the dataset.

2.1.3 Number of principal components

It can be hard to determine the optimal number of significant principal components that represent population structure. In our method, we are using Velicier's minimum average partial (MAP) test as proposed by Shriner (2011) [28] to automatically detect the number of top principal components D used for estimating the individual allele frequencies. Shriner showed that the test based on a Tracy-Widom distribution, described in Patterson et. al (2006)[4], systematically overestimates the number of significant principal components and even performs worse for datasets including admixed individuals. However, in order to be able to perform the MAP test and detect the optimal D, an initial covariance matrix is estimated based on the model in equation 3.

The MAP test is performed on the estimated initial covariance matrix $\mathbf{C}$ for NGS data as an approximation of a Pearsson correlation matrix used by Shriner. Using the notation of Shriner, $\mathbf{C}_d^*$ is defined as the matrix of partial correlations after having partialed out the first d principal components. Velicer (1976) [29] proposed the summary statistic $f_d = \sum_{i=1,i\neq j}^{m}\sum_{j=1}^{m}\frac{(\mathbf{C}_{d,ij}^*)^2}{m(m-1)}$, where $C_{d,ij}^*$ represents the entry in $\mathbf{C}_d^*$ for individuals i and j. Thus, the test statistic f_d represents the average squared correlation after partialing out the top d principal components. The number of top principal components that represent population structure is then chosen as $\text{argmin}_d\, f_d$, for $d = 0, \ldots, m-1$. We have used the same implementation of the MAP test as Shriner (2011) [28].

The MAP test and the preceding estimation of the initial covariance matrix can be avoided by having prior knowledge of an optimal D for the dataset being analyzed such that D is manually selected.

2.1.4 Genotype calling

As previously shown in [5, 16], genotypes can be called from posterior genotype probabilities to achieve higher accuracy in low depth NGS scenarios. We can adapt this concept to our posterior genotype probabilities based on individual allele frequencies, such that genotypes can be called at a higher accuracy in structured populations from low depth NGS data. The genotype for individual i in site s is called as follows:

$$\hat{g}_{is} = \operatorname*{argmax}_{g} P(G_{is} = g \,|\, X_{is}, \pi_{is}), \text{ for } g = 0, 1, 2. \tag{13}$$

2.2 Admixture proportions

Based on the likelihood model defined by Pritchard et al. (2000) [10], individual allele frequencies $\mathbf{\Pi}$ can be estimated using admixture proportions $\mathbf{Q}$ and population-specific allele frequencies $\mathbf{F}$ [12], such that:

$$\pi_{is} = \sum_{k=1}^{K} q_{ik} f_{sk}, \tag{14}$$

for an individual i in a variable site s. This is based on an assumption of K ancestral populations where $\sum_{k=1}^{K} q_{ik} = 1$ and $0 \leq q, f \leq 1 \;\forall\; q, f \in (\mathbf{Q}, \mathbf{F})$. However, $\mathbf{Q}$ and $\mathbf{F}$ must be inferred in order to estimate the individual allele frequencies, where as K is assumed to be known. One probabilistic approach for inferring population structure through admixture proportions in low depth NGS data has been implemented in the NGSadmix software by Skotte et al. (2013) [13]. Here both parameters are estimated jointly in an EM algorithm using the genotype likelihoods.

In our case, we have already estimated the individual allele frequencies based on our iterative procedure using PCA described above. K can be chosen as $D+1$, since it would explain the number of distinct ancestral population from which the individual allele frequencies have been estimated from. There is however no direct interpretation between principal components and admixture proportions [12]. Therefore, we propose an approach based on non-negative matrix factorization (NMF) to infer $\mathbf{Q}$ and $\mathbf{F}$ using only our estimated individual allele frequencies as information for low depth NGS data. NMF has previously been applied directly on genotype data to infer population structure and admixture proportions by Frichot et al. (2014) [30], where their method showed comparable accuracy and faster run-time in comparison to ADMIXTURE by Alexander et al. (2009) [12]. NMF has also been well applied in gene expression studies [31].

NMF is a dimension reduction and factor analysis method for finding a low-rank approximation of a matrix, which is similar to PCA, but NMF is constrained to find non-negative low-rank matrices. For an non-negative matrix $\mathbf{\Pi} \in \mathbb{R}_{+}^{M \times N}$, the goal of NMF is to find an approximation of $\mathbf{\Pi}$ based on two non-negative factor matrices $\mathbf{Q} \in \mathbb{R}_{+}^{m \times K}$ and $\mathbf{F} \in \mathbb{R}_{+}^{n \times K}$, such that:

$$\mathbf{\Pi} \approx \mathbf{Q}\mathbf{F}^{T}. \tag{15}$$

$\mathbf{Q}$ will consist of columns of non-negative basis vectors such that linear combinations of these approximates $\mathbf{\Pi}$ through $\mathbf{F}$. Thus based on the non-negative nature of our parameters, we can apply the ideas of NMF to infer admixture proportions and population-specific allele frequencies from the the individual allele frequencies. We use a combination of recent research in NMF to minimize the following least squares with an added sparseness constraint on $\mathbf{Q}$:

$$\min_{\mathbf{Q},\mathbf{F}} \left\| \hat{\mathbf{\Pi}} - \mathbf{Q}\mathbf{F}^T \right\|_F^2 + \alpha \sum_{i=1}^{m} \sum_{k=1}^{K} |q_{ik}|, \tag{16}$$

for $Q \geq 0$, $F \geq 0$ and $\alpha \geq 0$. Here $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |x_{ij}|^2}$ is the Frobenius norm of a matrix $\mathbf{X}$ and α is the regularization parameter controlling the sparseness enforced.

Lee and Seung (1999, 2001) [32, 33] proposed an multiplicative update (MU) algorithm to solve the standard NMF problem without the sparseness constraint included above. Their update rules can be seen as conservative steps for the two factor matrices in a gradient descent optimization problem, which ensure that the non-negative constraint holds for each update. MU and its relation to gradient descent is described in the supplementary material. Hoyer (2002) [34] extended the MU to incorporate a sparseness constraint as described in equation 16 for $\mathbf{Q}$. For $\alpha > 0$, the regularization parameter is used to reduce noise, especially induced by the uncertainty of low depth NGS data, in the estimated admixture proportions by enforcing sparseness in the solution.

The Euclidean cost (16) is guaranteed not to increase for each update of a factor matrix and MU converges towards a stationary using a small modification by Gillis and Glineur (2008, 2011) [35, 36]. Here the entries of a factor matrix are forced to be greater than a small value γ (1.0×10^{-4}) after each update. The update rules for $\mathbf{F}$ and $\mathbf{Q}$, with α included, are therefore defined as follows in iteration t:

$$\hat{\mathbf{F}}^{(t+1)} = \max\left(\gamma, \hat{\mathbf{F}}^{(t)} \otimes \frac{\hat{\mathbf{\Pi}}^T \hat{\mathbf{Q}}^{(t)}}{\hat{\mathbf{F}}^{(t)} \hat{\mathbf{Q}}^{(t)\,T} \hat{\mathbf{Q}}^{(t)}}\right). \tag{17}$$

$$\hat{\mathbf{Q}}^{(t+1)} = \max\left(\gamma, \hat{\mathbf{Q}}^{(t)} \otimes \frac{\hat{\mathbf{\Pi}} \hat{\mathbf{F}}^{(t+1)}}{\hat{\mathbf{Q}}^{(t)} \hat{\mathbf{F}}^{(t+1)\,T} \hat{\mathbf{F}}^{(t+1)} + \alpha}\right), \tag{18}$$

Here $\otimes$ represents element-wise multiplication while the division operator and max function are element-wise as well. However, MU has been shown to have a slow convergence rate, especially for dense matrices, and our approach is therefore to accelerate this procedure by combining two different techniques.

Gillis and Glineur (2011) [36] proposed an acceleration scheme where a factor matrix can be updated a fixed number of times at a lower computational cost while keeping the other factor matrix fixed without losing convergence properties. In this way, they showed an improvement in the convergence rate of MU.

Another approach for improving the convergence rate of MU in NMF has been proposed by Serizel et al. (2016) [37] using an algorithm based on asymmetric stochastic gradient descent, called ASG-MU. ASG-MU works by shuffling the columns of $\mathbf{\Pi}$ and splitting the column indices into B equally sized batches. The shuffling of the columns in $\hat{\mathbf{\Pi}}$ is performed to approximate equal variability across all the batches. The following batch update procedure is then iterated in the ASG-MU algorithm. The order of the batches is randomly permuted B_{rand} and each batch $b \in B_{\text{rand}}$ is used to update $\hat{\mathbf{F}}_b$ and $\hat{\mathbf{Q}}$ sequentially. Here $\hat{\mathbf{F}}_b$ is the subset of columns for batch $b \in \hat{\mathbf{F}}$. Thus, a full update of $\hat{\mathbf{F}}$ has only occurred after looping through all B batches, while $\hat{\mathbf{Q}}$ will be updated for every single batch. The update rules for $\hat{\mathbf{F}}$ and $\hat{\mathbf{Q}}$ are then defined as follows for batch b in iteration t:

$$\hat{\mathbf{F}}_b^{(t+1)} = \max\left(\gamma, \hat{\mathbf{F}}_b^{(t)} \otimes \frac{\hat{\mathbf{\Pi}}_b^T \hat{\mathbf{Q}}^{(t,b')}}{\hat{\mathbf{F}}_b^{(t)} \hat{\mathbf{Q}}^{(t,b')\,T} \hat{\mathbf{Q}}^{(t,b')}}\right), \tag{19}$$

$$\hat{\mathbf{Q}}^{(t,b)} = \max\left(\gamma, \hat{\mathbf{Q}}^{(t,b')} \otimes \frac{\hat{\mathbf{\Pi}}_b \hat{\mathbf{F}}_b^{(t)}}{\hat{\mathbf{Q}}^{(t,b')} \hat{\mathbf{F}}_b^{(t)\,T} \hat{\mathbf{F}}_b^{(t)} + \alpha}\right), \tag{20}$$

where b' represents the previous batch used to update $\hat{\mathbf{Q}}$. Note that t will only increase for $\hat{\mathbf{Q}}$ when all B batches has been looped through.

We can then extend the ASG-MU update procedure to integrate the accelerated update scheme from Gillis and Glineur (2011) [36] in each factor matrix update. The idea of introducing an acceleration scheme for MU in a stochastic gradient descent approach has also been described in Kasai (2017) [38]. However, further modifications are needed for our procedure as we need to satisfy $\sum_{k=1}^{K} q_{ik} = 1$, for $i = 1, \ldots, m$, as well as having all entries of the factor matrices in range $[\gamma, 1-\gamma]$. The rows of $\hat{\mathbf{Q}}$ are therefore normalized after each update and the entries of both $\hat{\mathbf{Q}}$ and $\hat{\mathbf{F}}$ truncated. The normalization of the $\hat{\mathbf{Q}}$ will also ensure that the NMF algorithm finds a unique solution.

We propose the following algorithm for combining the two acceleration approaches to estimate admixture proportions and population-specific allele frequencies from low depth NGS data, using only the estimated individual allele frequencies:

```
Algorithm 3: Estimation of admixture model parameters based on NMF.
  1. Initiate $\hat{\mathbf{Q}}$ and $\hat{\mathbf{F}}$ randomly with entries in range $[\gamma, 1-\gamma]$.
  2. Normalize rows of $\hat{\mathbf{Q}}$ to sum to one.
  3. Let $\hat{\mathbf{\Pi}}^*$ be $\hat{\mathbf{\Pi}}$ after column shuffling and let $B$ be the set of batches.
  4. Randomly permute batches in $B$, and for each $b \in B$:
      (a) Update $\hat{\mathbf{F}}_b$ using $\hat{\mathbf{Q}}$ and $\hat{\mathbf{\Pi}}_b^*$ in equation 19 with acceleration scheme.
      (b) Truncate entries of $\hat{\mathbf{F}}_b$ in range $[\gamma, 1-\gamma]$.
      (c) Update $\hat{\mathbf{Q}}$ using $\hat{\mathbf{F}}_b$ and $\mathbf{\Pi}_b^*$ in equation 20 with acceleration scheme.
      (d) Truncate entries of $\hat{\mathbf{Q}}$ in range $[\gamma, 1-\gamma]$.
      (e) Normalize rows of $\hat{\mathbf{Q}}$ to sum to one.
  5. Repeat from step 3 until admixture proportions have converged.
  6. Reshuffle columns of $\hat{\mathbf{F}}$ for column indices to match the originals of $\hat{\mathbf{\Pi}}$.
```

Convergence in the estimation of admixture proportions is defined as when the RMSD of estimated admixture proportions of two successive iterations are smaller than a value ϕ (5.0×10^{-5}). The RMSD of iteration $t+1$ is defined as,

$$\text{RMSD} = \sqrt{\frac{1}{mK} \sum_{i=1}^{m} \sum_{k=1}^{K} (\hat{q}_{ik}^{(t+1)} - \hat{q}_{ik}^{(t)})^2}. \tag{21}$$

The α parameter enforcing sparseness in the estimated solution of Q is arbitrarily specified, however the use of the likelihood measure in the NGSdamix [13] model can be used to determine a proper α parameter fitting the dataset. The likelihood measure is defined as:

$$\mathcal{L}(\hat{\mathbf{Q}}, \hat{\mathbf{F}}) = \prod_{i=1}^{m} \prod_{s=1}^{n} \sum_{g=0}^{2} P(X_{is} \,|\, G_{is} = g) P(G_{is} = g \,|\, \hat{\pi}_{is}) \tag{22}$$

where $\hat{\pi}_{is} = \sum_{k=1}^{K} \hat{q}_{ik} \hat{f}_{sk}$. Based on the fast estimation of admixture proportions using our NMF algorithm, a set of α values can be tested and measured sequentially using the likelihood measure. This can be performed without sacrificing significant run-time compared to NGSadmix due to already having estimated the individual allele frequencies for a particular K.

2.3 Implementation

Both presented methods have been implemented in a Python framework named PCAngsd (Principal Component Analysis of Next Generation Sequencing Data). The framework is freely available at `http://www.popgen.dk/software/`.

The memory requirements for using PCAngsd is $\mathcal{O}(mn)$ as the genotype likelihoods need to be stored in memory, and the most computational expensive step is the estimation of individual allele frequencies and covariance matrix ($\mathcal{O}(m^2 n)$). However, a fast SVD method for only computing the top D eigenvectors, implemented in the Scipy library [39] using ARPACK [40] as an eigensolver, has been used to speed up the estimations for the individual allele frequencies. PCAngsd is multithreaded as well to take advantage of several cores and the backbone of the framework is based on Numpy [41] data structures using the Numba [42] library to speed up bottlenecks with just-in-time (JIT) compilation.

3 Data

3.1 Simple simulation of genotypes and sequencing data

Low depth NGS data has been simulated as genotype likelihoods to test the capabilities of our two presented methods. Allele frequencies of the reference panel of the Human Genome Diversity Project (HGDP) [43] have been used to generate a total of 380 individuals from three distinct populations (French, Han Chinese, Yoruba) including admixed individuals in approximately 0.4 million SNPs across all autosomes. As the allele frequencies are known for each population, the genotypes of each individual can be sampled from a Binomial distribution for each diallelic SNP, using the population-specific allele frequency or an admixed allele frequency as parameter. No LD has been simulated. The genotypes are therefore known and are used in the evaluation of our methods in our low depth scenarios. The number of reads in each SNP are sampled from a Poisson distribution with a mean parameter resembling the average sequencing depth of the individual, and the genotype is used to sample the number of derived alleles from a Binomial distribution using the sampled depth as parameter. The sequencing depth of each individual is sampled uniformly random from a range of $[0.5, 5]$. Sequencing errors are incorporated by sampling each read with a probability $\epsilon = 0.01$ of being wrong. The genotype likelihoods are then finally generated from the probability mass function of a Binomial distribution using the sampled parameters and ϵ. This approach of genotype likelihood simulation has previously been used in [13, 20, 22].

A complex admixture scenario has been constructed to test the capabilities of our methods. 100 individuals have been sampled directly from each of the population specific allele frequencies (non-admixed), while 50 individuals have been sampled to have equal ancestry from each of the three distinct populations (three-way admixture). At last, 30 individuals have been sampled from a gradient of ancestry between all pairs of the ancestral populations (two-way admixture).

3.2 1000 Genomes

We also analyze human low coverage NGS data of 193 individuals from the 1000 Genomes Project Consortium [17, 18]. The individuals are from four different populations consisting of 41 from CEU (Utah residents with Northern and Western European ancestry), 40 from CHB (Han Chinese in Beijing), 48 from YRI (Yoruba in Ibadan) and 64 individuals from MXL (Mexican ancestry in Los Angeles) representing an admixed scenario of European and Native American ancestry. The individuals from the low coverage datasets used here have a varying sequencing depth from $3 - 14$X after filtering. An advantage of using the 1000 Genomes Project data is that reliable genotypes of the individuals in the low coverage sequencing dataset are available, such that we can use them for validation purposes.

SNP calling and estimation of genotype likelihoods of the 1000 Genomes dataset has been performed in ANGSD [21] using simple read quality filters. A significance threshold of 1.0×10^{-6} has been used for SNP calling alongside a MAF filter of 0.05 removing rare variants. The number of SNPs is also thinned by removing every eighth SNP in order to reduce the dataset and reduce the effect of LD patterns. A total number of 1 million variable sites across all autosomes have been used in the analyses. The full ANGSD command used to generate the genotype likelihoods is provided in the supplementary material.

3.3 Waterbuck

Lastly, an animal dataset (non-model organism) has also been included in our study. A reduced low depth NGS dataset of the waterbuck (*Kobus ellipsiprymnus*) originating from Pedersen et al. (2018, unpublished) [44] has been analyzed. The dataset consists of 73 samples that have been sampled at 5 different sites in Africa with a varying sequencing depth from $2.2 - 4.7$X. The dataset has been reduced to only include sampling sites with more than 10 samples such that the inferred axes of genetic variation will reflect true population structure. As performed for the 1000 Genomes dataset, genotype likelihoods has been estimated in ANGSD with the same SNP and MAF filters. A total number of 10 million SNPs across the autosomes of the waterbuck is analyzed in this study.

4 Results

For the simulated and 1000 Genomes datasets, results estimated in PCAngsd on low depth NGS data are evaluated against the results estimated from reliable genotype data. The model in Patterson et al. (2006) [4] (equation 2) is used to perform PCA, while ADMIXTURE [12] is used to estimate admixture proportions on the genotype datasets. The performance of PCAngsd is also compared to existing NGS methods in performing

PCA, the ngsTools [26] model (equation 3), and estimating admixture proportions, NGSadmix [13], that are both based on probabilistic frameworks using genotype likelihoods as well. In all the following cases of admixture plots estimated by PCAngsd, α has been selected by choosing the one maximizing the likelihood measure described above (equation 22).

RMSD is used to evaluate the performances of both NGS methods for estimating admixture proportions in terms of accuracy:

$$\text{RMSD} = \sqrt{\frac{1}{mK}\sum_{i=1}^{m}\sum_{k=1}^{K}(\hat{q}_{ik} - q_{ik})^2}, \tag{23}$$

where q_{ik} and $\hat{q}_{ik}$ represents the estimated admixture proportion for individual i in ancestral population k from known genotypes and NGS data, respectively. The accuracy of the estimated PCA plots of both NGS methods are evaluated with a Procrustes analysis [5, 45] producing a residual sum-of-squares (RSS) value using the estimated PCA plot of the known genotypes for the simulated and 1000 Genomes datasets.

As well as measuring the accuracy of the presented methods, we also evaluate the number of ancestral populations K chosen using residual matrices based on genotype dosages and individual allele frequencies. The residual matrix $\mathbf{R}$ will be defined as follows for individual i in site s:

$$r_{is} = 2\hat{\pi}_{is} - \mathbb{E}[G_{is} \,|\, X_{is}, \hat{\pi}_{is}] \tag{24}$$

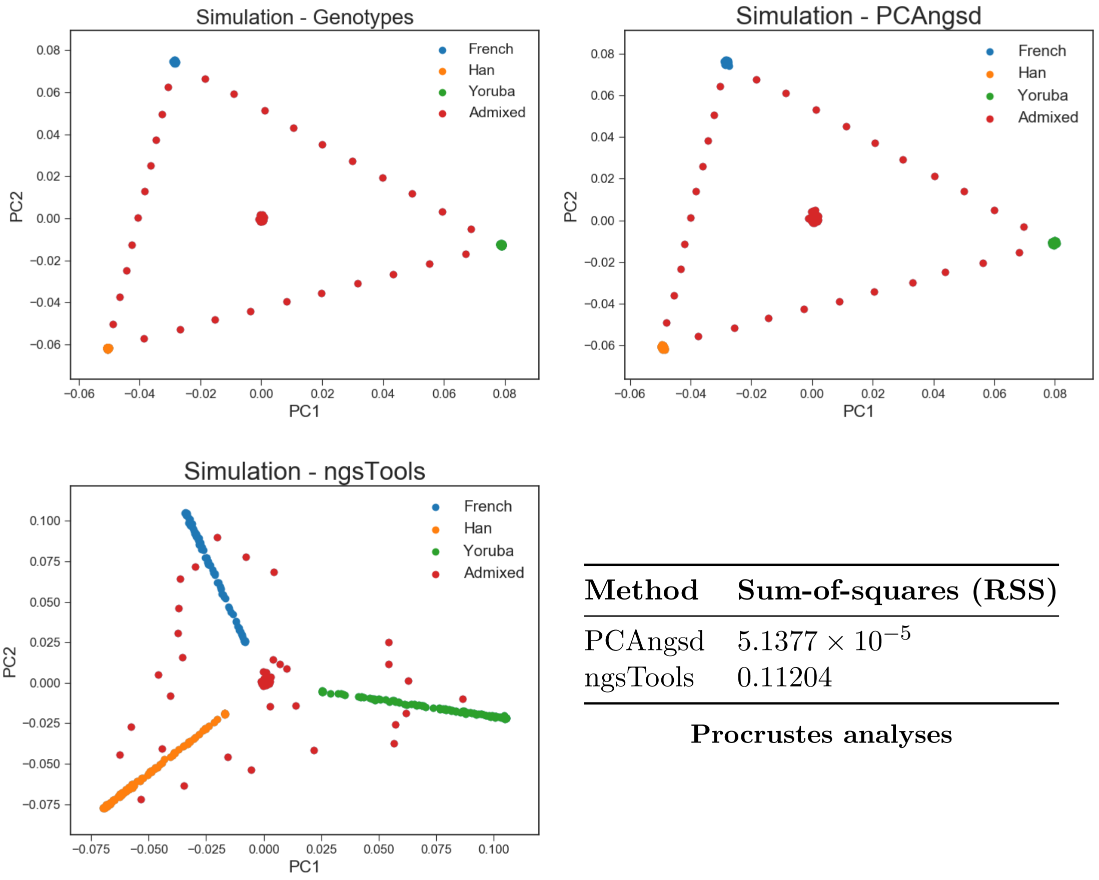
The correlation matrix is then computed from $\mathbf{R}$. Therefore, if the number of assumed ancestral populations K is not representative of the dataset, then we would see a positive correlation in the residuals between the individuals within a population, as K is not sufficient to model the individual allele frequencies. A plot of the correlation matrix can therefore serve as a verification of the chosen K as well as the inferred number of eigenvectors in the MAP test ($D = K - 1$).

All tests in this study have been performed server-side using 32 threads (Intel® Xeon® CPU E5-2690) for both PCAngsd and NGSadmix.

4.1 Simulation

The results of performing PCA on the simulated dataset are displayed in Figure 1. The MAP test reported 2 significant principal components which was also expected for individuals simulated from three distinct populations. The inferred principal components clearly shows the importance of taking the estimated individual allele frequencies into account in the probabilistic framework. Here PCAngsd is able to infer the population structure of individuals from distinct populations and admixed individuals nicely as also seen by the Procrustes analysis with a RSS value of 5.14×10^{-5}. There is clear bias in the results of the ngsTools model where the patterns are representing sequencing depth instead of population structure as made apparent in Figure 9. The individuals are acting as a gradient towards the origin due to their varying sequencing depth. The biased performance of ngsTools is also reflected in the Procrustes analysis with an estimated RSS value of 0.112.

The estimated admixture proportions for the simulated dataset are displayed in Figure 2. PCAngsd estimates the admixture proportions well with a RMSD of 0.00476 compared



Method	Sum-of-squares (RSS)
PCAngsd	5.1377×10^{-5}
ngsTools	0.11204

Procrustes analyses

Figure 1: PCA using different methods of the top 2 principal components in the simulated dataset consisting of 380 individuals and 0.4 million variable sites. The top left plot shows the PCA performed on the sampled genotypes using the model described by Patterson et al. (2006) (equation 2). The top right plot shows the PCA performed by PCAngsd, and the bottom plot left displays the PCA performed by the ngsTools model (equation 3). The accuracy of the PCA plots of the NGS methods are summarized in the table in the bottom right based on Procrustes analysis.

to the ADMIXTURE estimates of the known genotypes, but is however outperformed by NGSadmix with a RMSD of 0.00184. For the 380 individuals and 0.4 million SNPs using $K = 3$, PCAngsd had an average run-time of only 3.5 minutes while NGSadmix had an average run-time of 7.9 minutes.

4.2 1000 Genomes

The methods of PCAngsd have also been applied to the 1000 Genomes dataset. The MAP test indicated evidence of 3 significant principal components meaning that the Native American ancestry explains enough genetic variance in the dataset to get an axis of its own. The results of the PCA are displayed in Figure 3. As was also seen for the simulated dataset, PCAngsd is able to cluster all individuals almost perfectly, while the ngsTools model is only able to capture some of the same population structure patterns. Its results are still biased by the variable sequencing depth as seen as well in Figure 10. The RSS

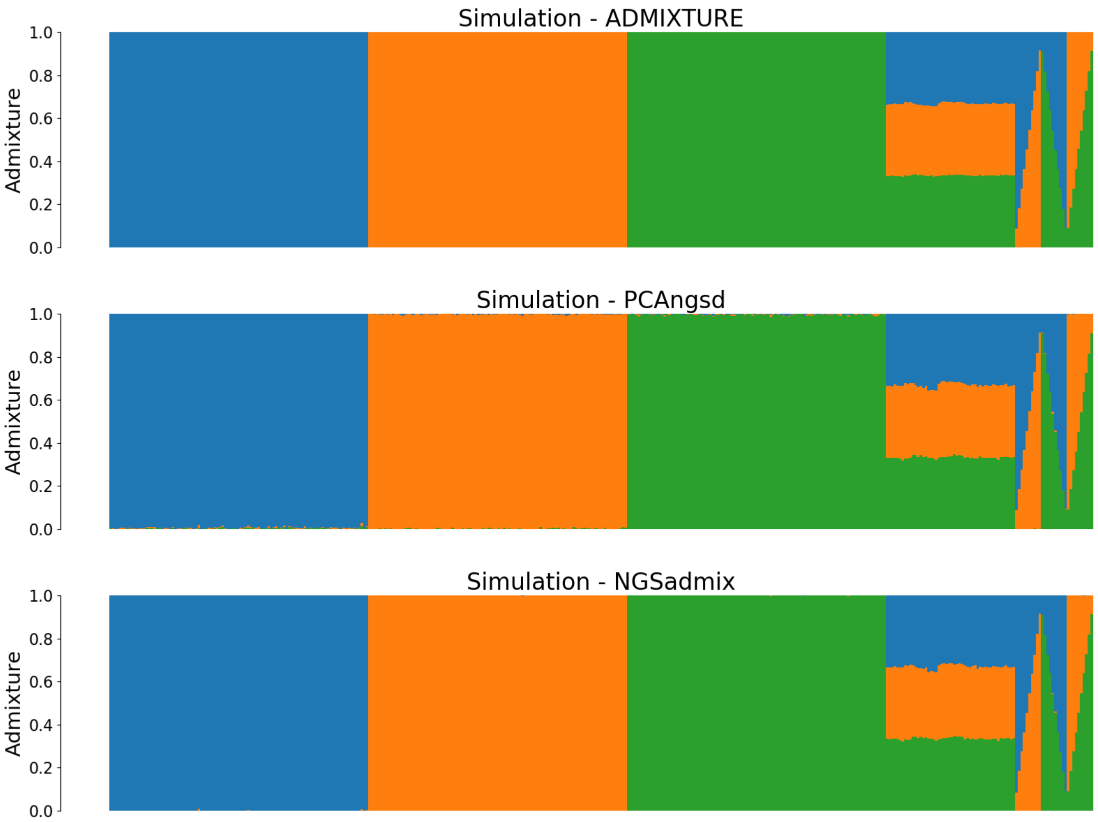


Figure 2: Admixture plots for $K = 3$ of the simulated dataset. The first plot is the admixture proportions estimated in ADMIXTURE [12] using the known genotypes, which represents the ground-truth in our simulation studies. The second plot shows admixture proportions estimated using PCAngsd with parameters $\alpha = 0$ and $B = 5$ and the bottom plot using NGSadmix [13].

values of the Procrustes analyses verify the observations, where PCAngsd has a RSS value of 0.000575 compared to ngsTools with a RSS value of 0.00814.

The admixture plots are displayed in Figure 4. PCAngsd is not able to outperform NGSadmix in terms of accuracy, however it is still able to estimate a very similar result. PCAngsd has some issues with noise in its estimation but is however able to reduce it with the use of the sparseness parameter α. The likelihood measure in equation 22 has been used to easily find an optimal α as seen in Figure 12. PCAngsd estimates the admixture proportions with a RMSD of 0.0121 compared to NGSadmix with a RMSD of 0.0108. The average run-time for 193 individuals and 1 million SNPs using $K = 4$ was 3.6 minutes for PCAngsd and 14.9 minutes for NGSadmix, making PCAngsd more than 4.1x faster than NGSadmix.

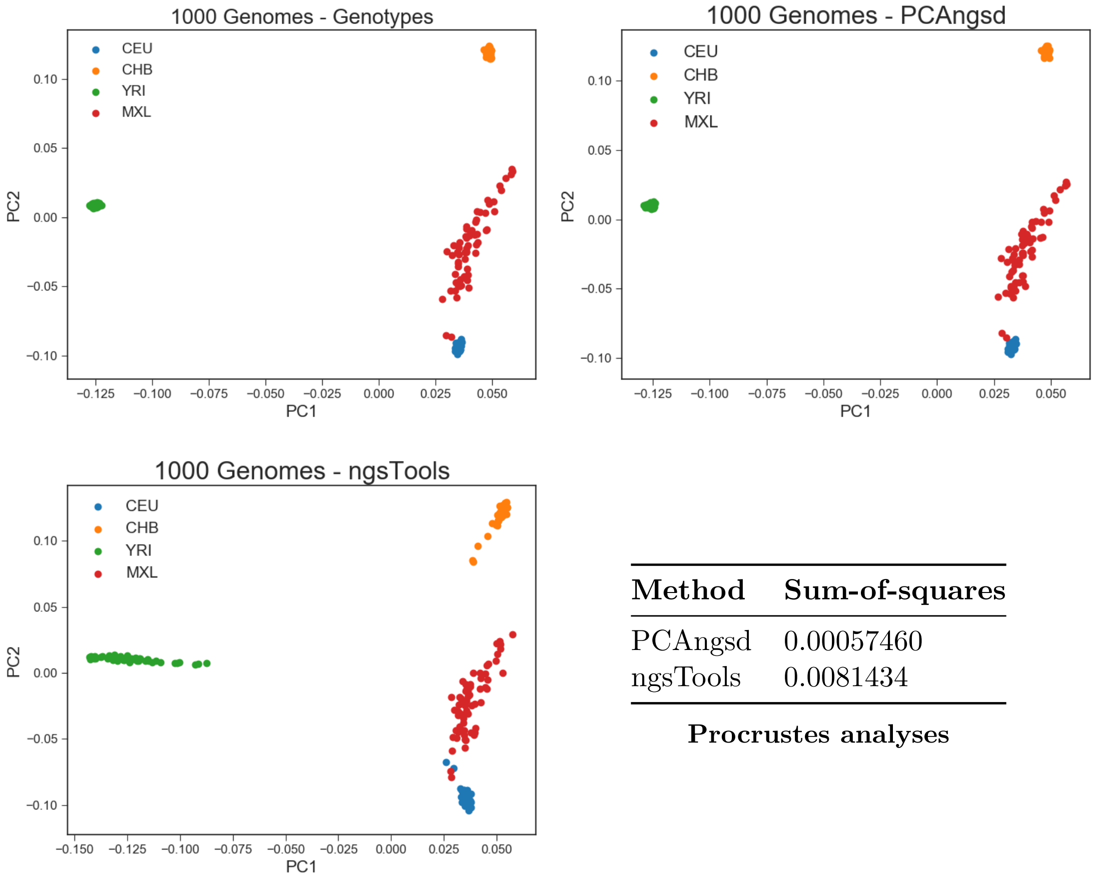
We have computed correlation matrices based on the residuals (equation 24) for the 1000 Genomes dataset in Figure 5 using $K = 3, 4$. Here we show the difference between the assumption of 3 or 4 ancestral populations when estimating admixture proportions. It is clearly seen that the assumption of only 3 ancestral populations is not enough to fully explain the population structure in the sample as the residuals are positively correlated for the individuals with Mexican ancestry. For $K = 4$, these individuals can be modeled more accurately as seen in the bottom right corner of both plots. These results show that the individuals with Mexican ancestry can not only be modeled by European and Asian ancestry but would need the presence of assumed Native American ancestry as well.

4.3 Waterbuck

Lastly, we have analyzed the waterbuck dataset. The MAP test reported 4 significant principal components for explaining the genetic variation in the dataset which also fits with having 5 distinct waterbuck sampling sites. The PCA plots are visualized in Figure 6, where the top 4 principal components have been plotted for each method. Once again, PCAngsd is able to cluster the populations much better than the ngsTools model, however the effect is not as apparent as for the other datasets. This is very likely due to the low number of individuals in each population which means that the principal components and individual allele frequencies can not be as well described.

The bias, which affects the estimation of the individual allele frequencies, will of course also affect the admixture plots seen in Figure 7, where additional noise is hard to remove without also affecting the true ancestry signals. Still, PCAngsd is capturing the same ancestry signals as NGSadmix with the use of the sparseness parameter and the RMSD between the estimates of the two methods is merely 0.00927. It is worth noting that an admixed individual of Ugalla and QENP is captured in both PCA and admixture estimation of PCAngsd as also verified by the NGSadmix method. The difference in run-times for the waterbuck dataset of 73 samples and 10 million SNPs using $K = 5$, where PCAngsd had an average run-time of 19 minutes while NGSadmix had an average run-time of 3.2 hours, thus making PCAngsd more than 10x faster.

As for the 1000 Genomes dataset, we have computed correlation matrices of the residuals for $K = 4, 5$ in the waterbuck dataset. The results can be seen in Figure 8. The plots enforces the evidence of 5 distinct populations ($K = 5$), as inferred by the MAP test, due to the positive correlation of residuals seen in the bottom right corner for $K = 4$. A negative correlation can be seen between the individuals within the same population, as deviations from the population-specific mean will become much more apparent for a low number of individuals using low depth NGS data.



Method	Sum-of-squares
PCAngsd	0.00057460
ngsTools	0.0081434

Procrustes analyses

Figure 3: PCA plots of the top 2 principal components for the 1000 Genomes dataset with 193 individuals and 1 million variable sites. The top left PCA plot is based on the known genotypes of the variable sites used in the low depth NGS data, top right PCA is performed by PCAngsd and the bottom left PCA is performed by the ngsTools model. The accuracy of the PCA plots of the NGS methods are summarized in the table in the bottom right based on Procrustes analysis.

Dataset	m	n	**PCAngsd**	**NGSadmix**	**Depth**
Simulated	380	0.4 million	3.5 min	7.9 min	0.5 – 5X
1000 Genomes	193	1 million	3.6 min	14.9 min	3 – 14X
Waterbuck	73	10 million	19 min	192 min (3.2 hours)	2.2 – 4.7X

Table 1: Average run-times of 10 initializations for both PCAngsd and NGSadmix. The run-times reported for PCAngsd include both estimation of covariance matrix, individual allele frequencies and admixture proportions. All tests have been performed server-side using 32 threads.

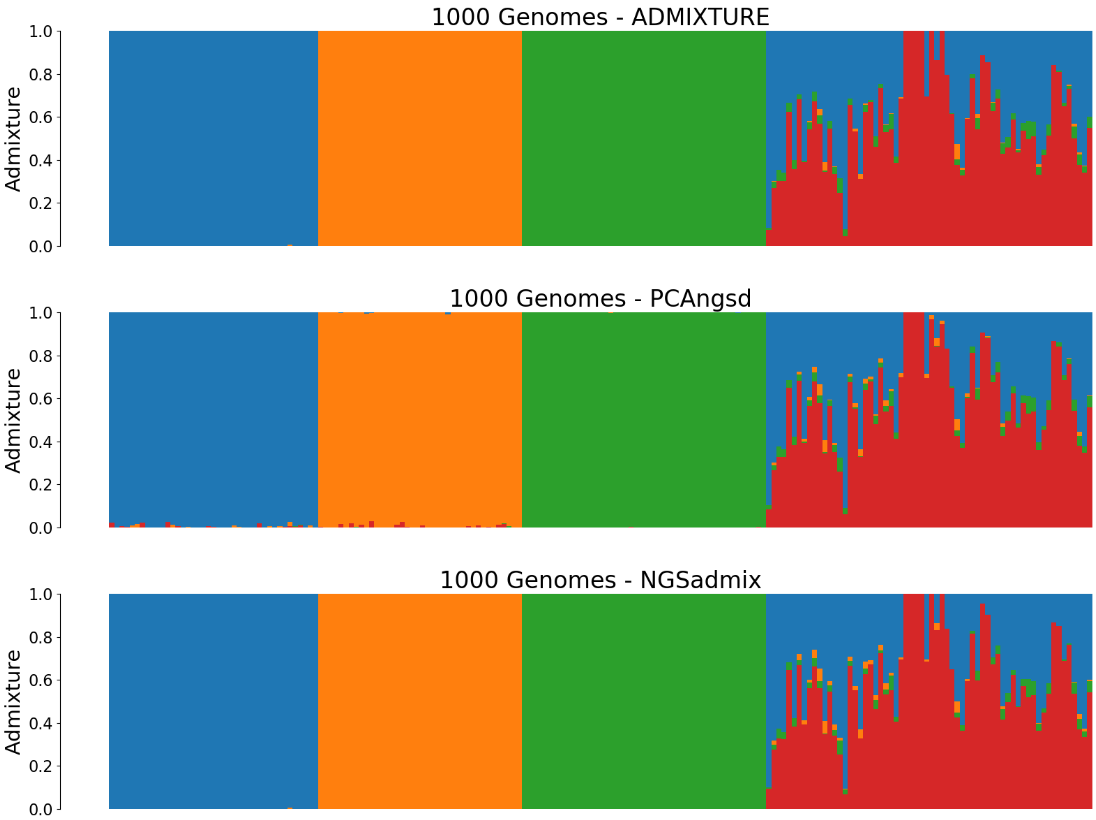


Figure 4: Admixture plots for $K = 4$ of the 1000 Genomes dataset. The first plot is the admixture proportions estimated in ADMIXTURE [12] using the known genotypes, the second plot shows admixture proportions estimated in PCAngsd with parameters $\alpha = 250$ and $B = 5$ and the last plot is the admixture proportions estimated in NGSadmix [13].

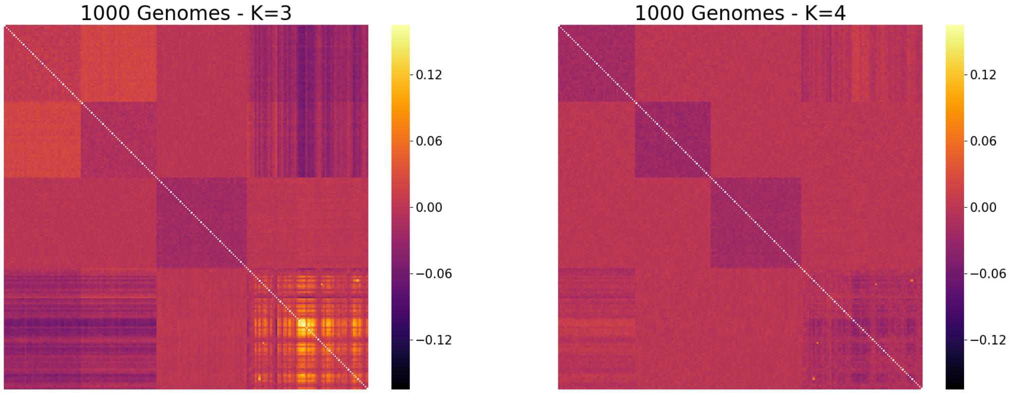


Figure 5: Correlation matrices of the residuals based on equation 24 assuming $K = 3, 4$. A positive correlation between the residuals of two individuals indicate that the number of assumed ancestral population is not sufficient to describe the population structure of the dataset.

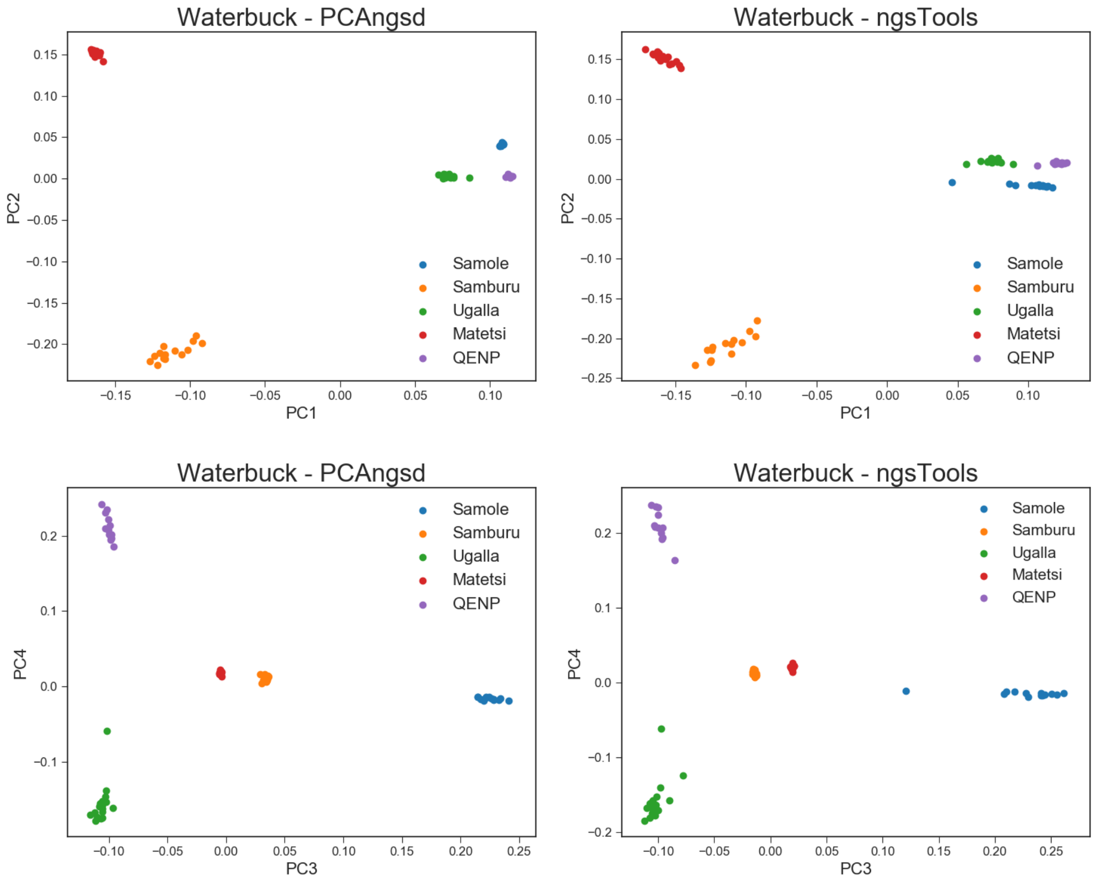


Figure 6: PCA plots of the top 4 principal components for the waterbuck dataset with 73 individuals and 10 million variable sites. The first row displays the plotting of the first and second principal components for PCAngsd and the ngsTools model, respectively, while the second row displays the plotting of the third and fourth principal components.

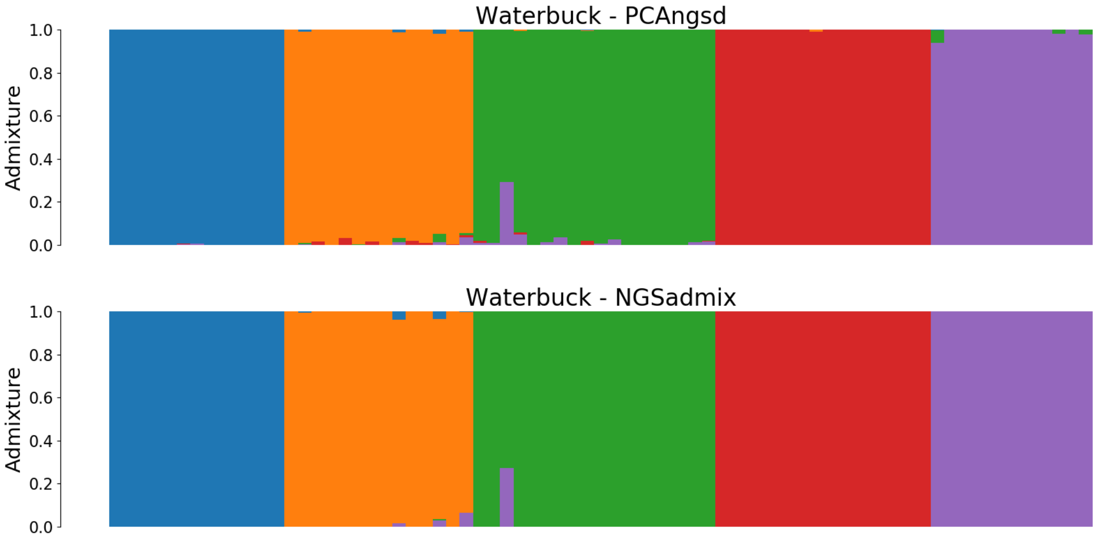


Figure 7: Admixture plots for $K = 5$ of the waterbuck dataset. The first plot is the admixture proportions estimated in PCAngsd with parameters $\alpha = 5000$ and $B = 5$ and the second plot shows the admixture proportions estimated in NGSadmix [13].

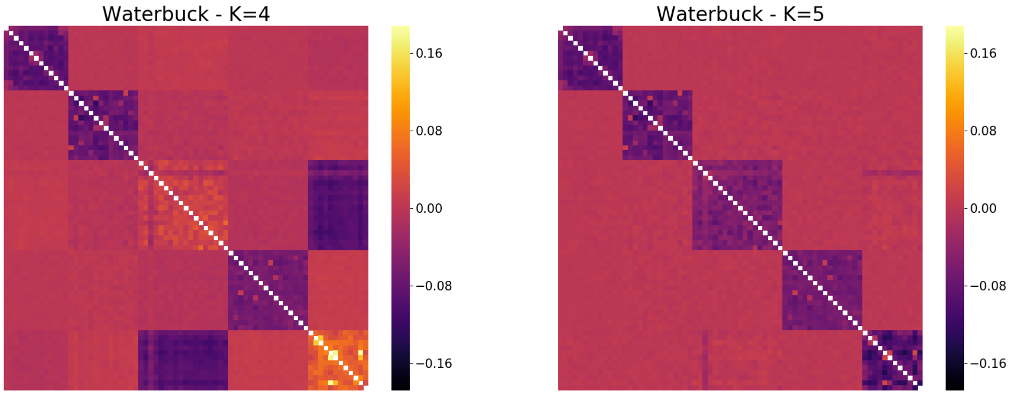


Figure 8: Correlation matrices of the residuals for the waterbuck dataset assuming $K = 4, 5$. A positive correlation between the residuals of two individuals indicate that the number of assumed ancestral population is not sufficient to describe the population structure of the dataset.

5 Discussion

We have presented two new methods for inferring population structure and admixture proportions in low depth NGS data and both methods have been implemented in a framework named PCAngsd. We have developed a probabilistic framework using genotype likelihoods to iteratively estimate individual allele frequencies in which we have connected principal components to admixture proportions such that we are able to infer and estimate both in a very fast approach.

Based on the results when inferring population structure using PCA, it is clear that the increased uncertainty of low depth sequencing data biases the clustering of populations using the ngsTools model. Contrary to PCAngsd, population structure is not taking into account when using the posterior genotype probabilities to estimate the covariance matrix. The ngsTools model uses population allele frequencies as prior information for all individuals such that individuals are assumed to be sampled from a homogeneous population. This assumption is of course violated when individuals are sampled from structured populations with diverge ancestries. Missing data is therefore modeled by population allele frequencies that resemble an average across the entire sample. As an effect of this, the low depth individuals are modeled by sequencing depth instead of population structure. These results may lead to misinterpretations of population structure or admixture only due to low and variable sequencing depth. However, PCAngsd is able to overcome the observed bias of low and variable sequencing depth by using individual allele frequencies as prior information, which leads to more accurate results in all datasets of the study, as missing data is modeled by inferred population structure. The assumption of conditional independence between individuals in the estimation of the covariance matrix (equation 11) also holds for structured populations by using the estimated individual allele frequencies.

The number of significant eigenvectors used in the estimation of individual allele frequencies is determined by the MAP test. The MAP test is performed on the covariance matrix estimated from the ngsTools model, which we have shown to be biased by low and variable sequencing depth. Thus in cases of complex population structure and low and variable sequencing depth, it is possible that the MAP test will not find a suitable number of significant eigenvectors to represent the genetic variation of the dataset. It could therefore be more relevant to use prior information regarding the number of eigenvectors needed for the dataset instead. However for each of the cases presented in this study, the MAP test inferred the expected number of significant eigenvectors to describe the population structure.

PCAngsd is able to approximate the results of NGSadmix to a high degree when estimating admixture proportions using solely the estimated individual allele frequencies. However, PCAngsd is not able to outperform NGSadmix in terms of accuracy, but it is however able to capture the exact same ancestry patterns as the clustering-based methods in a much faster approach, as shown by the run-times of each method. Another advantage of PCAngsd is that the estimated individual allele frequencies are only needed to be computed once for a specific K, thus multiple different α's and random seeds can be tested in the same run for an even greater speed advantage over NGSadmix, since the iterative estimation of individual allele frequencies is the most computational expensive step in PCAngsd. PCAngsd is therefore an appealing alternative for estimating admixture proportions for low depth NGS data as convergence and run-time can be a problem for a large number of parameters in NGSadmix [13]. PCAngsd was only seen to converge to a single solution for all our practical tests. We recommend to have at least 100000 SNPs in each batch to reduce the probability of having an unfortunate split and shuffling of the variable sites, and thus ensuring approximately equal variability across the batches.

Both methods of the PCAngsd framework rely on an representative estimation of individual allele frequencies which are modeled using the inferred principal components of the SVD on the genotype dosages. The number of individuals representing each population or subpopulation is essential for inferring principal components that describe true population structure as each individual will contribute to the construction of these axes of genetic variation. This particular effect can be seen in the PCA results of the waterbuck dataset where the populations are only described by a low number of individuals such that some of the clusters are not so well defined as for the other datasets. The admixture proportions estimated from the waterbuck dataset are therefore affected as well which can be seen by the additional noise in the admixture plots.

The PCAngsd framework might be able to push the lower boundaries of sequencing depth required to perform population genetic analyses using NGS data of large-scale genetic studies. PCAngsd demonstrates an efficient approach to be able to deal with merged datasets with various sequencing depths as well. The estimated individual allele frequencies contain a lot of information regarding population structure and can open up for the development and extension of population genetic models based on a similar probabilistic framework to naturally correct for population structure in order to obtain more accurate estimates in heterogeneous populations.

6 Supplementary Material

6.1 Genotype likelihoods

Genotype likelihoods are the probability of the observed sequencing data given the unobserved genotypes. They can be computed from next-generation sequencing (NGS) data using the uncertainty of each base from the raw quality scores of sequencing machines. The quality scores Q are usually in Phred scale such that the probability of an error in the observed base call is given by $\epsilon = 10^{\frac{-Q}{10}}$. The probability of observing a base b of read r in a site s can be seen as the likelihood of the given allele. For having L reads covering s and assuming independence between the reads (and the error probabilities), the genotype likelihood can be computed by the product of the allelic likelihoods for the site [16, 19]. The genotype likelihood for individual i in site s can be defined as follows for a multi-allelic case derived from the approach in [24]:

$$P(X_{is} \,|\, G = A_1A_2) \propto \prod_{r=1}^{L} \left(\frac{P(b_r^{(i)} \,|\, A_1)}{2} + \frac{P(b_r^{(i)} \,|\, A_2)}{2} \right) . \tag{25}$$

Here X_{is} is the sequencing data, $P(b \,|\, A) = 1 - \epsilon$, for $b = A$, and $P(b \,|\, A) = \frac{\epsilon}{3}$, for $b \neq A$, with ϵ being the probability of error in the observed base call. This is for an arbitrary genotype A_1A_2.

6.2 Population allele frequencies

The population allele frequencies $\mathbf{p}$ can be estimated from NGS data using an Expectation Maximization (EM) algorithm to compute the maximum likelihood estimator for each site. The likelihood function of $\mathbf{p}$ in a site s is defined in Kim et al. (2011) [20] as follows by assuming independence between all m individuals:

$$\mathcal{L}(p_s) \propto P(\mathbf{X}_s \,|\, p_s) = \prod_{i=1}^{m} P(X_{is} \,|\, p_s) \,. \tag{26}$$

Here $\mathbf{X}_s$ is the observed sequencing data in site s. Since the genotype is not observed for NGS data, a latent variable G is introduced by taking the sum over the realizations of the genotype. Thus for individual i in site s, $P(X_{is} \,|\, p_s)$ can now be defined as:

$$P(X_{is} \,|\, p_s) = \sum_{g=0}^{2} P(X_{is} \,|\, G = g) P(G = g \,|\, p_s) \,, \tag{27}$$

where $P(X_{is} \,|\, G_{is} = g)$ is the genotype likelihood and $P(G_{is} = g \,|\, p_s)$ is the prior genotype probability. By assuming Hardy-Weinberg equilibrium (HWE) in the whole sample, the prior genotype probabilities are estimated as $P(G_{is} = 0 \,|\, p_s) = (1 - p_s)^2$, $P(G_{is} = 1 \,|\, p_s) = 2p_s(1-p_s)$ and $P(G_{is} = 2 \,|\, p_s) = p_s^2$. The maximum likelihood estimator of p_s is then defined as follows:

$$\hat{p}_s^{(\mathrm{ML})} = \underset{p_s}{\operatorname{argmax}} \prod_{i=1}^{m} P(X_{is} \,|\, p_s) \,. \tag{28}$$

The maximum likelihood solution is found by estimating the mean posterior expectations of the latent variable G iteratively for all individuals. The posterior genotype probability for individual i in site s is given as:

$$P(G_{is} = g \,|\, X_{is}, p_s) = \frac{P(X_{is} \,|\, G_{is} = g)P(G_{is} = g \,|\, p_s)}{\sum_{g'=0}^{2} P(X_{is} \,|\, G_{is} = g')P(G_{is} = g' \,|\, p_s)} \,. \tag{29}$$

And the posterior expectation of the genotype is then given as:

$$\mathbb{E}[G_{is} \,|\, X_{is}, p_s] = \sum_{g=0}^{2} g P(G_{is} = g \,|\, X_{is}, p_s) \,. \tag{30}$$

Now the update step for iteration $t+1$ in the EM algorithm can be defined as the mean of the posterior expectations of the genotype. The population allele frequency for each site is then obtained by scaling with 2 based on an assumption of G being Binomial distributed ($\mathbb{E}[G] = 2p$):

$$\hat{p}_s^{(t+1)} = \frac{\sum_{i=1}^{m} \mathbb{E}[G \,|\, X_{is}, \hat{p}_s^{(t)}]}{2m} \,. \tag{31}$$

6.3 Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is a dimension reduction and factor analysis method. Given a matrix $\mathbf{\Pi} \in \mathbb{R}_+^{m \times n}$, NMF finds two factor matrices $\mathbf{Q} \in \mathbb{R}_+^{m \times K}$ and $\mathbf{F} \in \mathbb{R}_+^{n \times K}$, such that $\mathbf{\Pi} \approx \mathbf{Q}\mathbf{F}^T$. The Euclidean cost J is usually used as an objective function in NMF and can be minimized as an optimization problem with respect to $\mathbf{Q}$ and $\mathbf{F}$:

$$J(\mathbf{Q}, \mathbf{F}) = \frac{1}{2} \left\| \mathbf{\Pi} - \mathbf{Q}\mathbf{F}^T \right\|_F^2 = \frac{1}{2} \sum_{i=1}^{m} \sum_{s=1}^{n} \left| \pi_{is} - (\mathbf{Q}\mathbf{F}^T)_{is} \right|^2 . \tag{32}$$

Here $\|\mathbf{X}\|_F$ is the Frobenius norm of a given matrix $\mathbf{X}$. The gradient matrices of J with respect to $\mathbf{Q}$ and $\mathbf{F}$, respectively, can be defined as:

$$\nabla_{\mathbf{Q}} J(\mathbf{Q}, \mathbf{F}) = \mathbf{\Pi}\mathbf{F} - \mathbf{Q}\mathbf{F}^T\mathbf{F}, \tag{33}$$

$$\nabla_{\mathbf{F}} J(\mathbf{Q}, \mathbf{F}) = \mathbf{\Pi}^T\mathbf{F} - \mathbf{F}\mathbf{Q}^T\mathbf{Q}, \tag{34}$$

which we are using in equation in equation 36 and 37. Lee and Seung (1999, 2001) [32, 33] presented multiplicative update rules to minimize the cost J with respect to $\mathbf{Q}$ and $\mathbf{F}$, which have been defined element-wise for $\mathbf{Q}$ and $\mathbf{F}$ as follows:

$$q_{ik} = q_{ik} \frac{(\mathbf{\Pi}\mathbf{F})_{ik}}{(\mathbf{Q}\mathbf{F}^T\mathbf{F})_{ik}}, \; f_{sk} = f_{sk} \frac{(\mathbf{\Pi}^T\mathbf{Q})_{sk}}{(\mathbf{F}\mathbf{Q}^T\mathbf{Q})_{sk}}, \tag{35}$$

for $i = 1, \ldots, m$, $s = 1, \ldots, n$ and $k = 1, \ldots, K$. The multiplicative update rules can be contrasted to gradient descent by displaying an additive update to minimize J using the gradients described above:

$$q_{ik} = q_{ik} + \eta_{ik}^{(q)} (\nabla_{\mathbf{Q}} J(\mathbf{Q}, \mathbf{F}))_{ik} \tag{36}$$

$$f_{sk} = f_{sk} + \eta_{sk}^{(f)} (\nabla_{\mathbf{F}} J(\mathbf{Q}, \mathbf{F}))_{sk} \tag{37}$$

Then by setting the steplengths (η) to small positive numbers, the additive updates would be equivalent to a standard gradient descent approach that decreases the Euclidean cost J. However by setting $\eta_{ik}^{(q)} = \frac{q_{ik}}{(\mathbf{Q}\mathbf{F}^T\mathbf{F})_{ik}}$ and $\eta_{ik}^{(f)} = \frac{f_{sk}}{(\mathbf{F}\mathbf{Q}^T\mathbf{Q})_{sk}}$ is equal to the multiplicative update rules (equation 35), where the steps are conservative enough to ensure non-negative entries for each update [33, 46].

6.4 PCA - Sequencing depth

6.4.1 Simulation

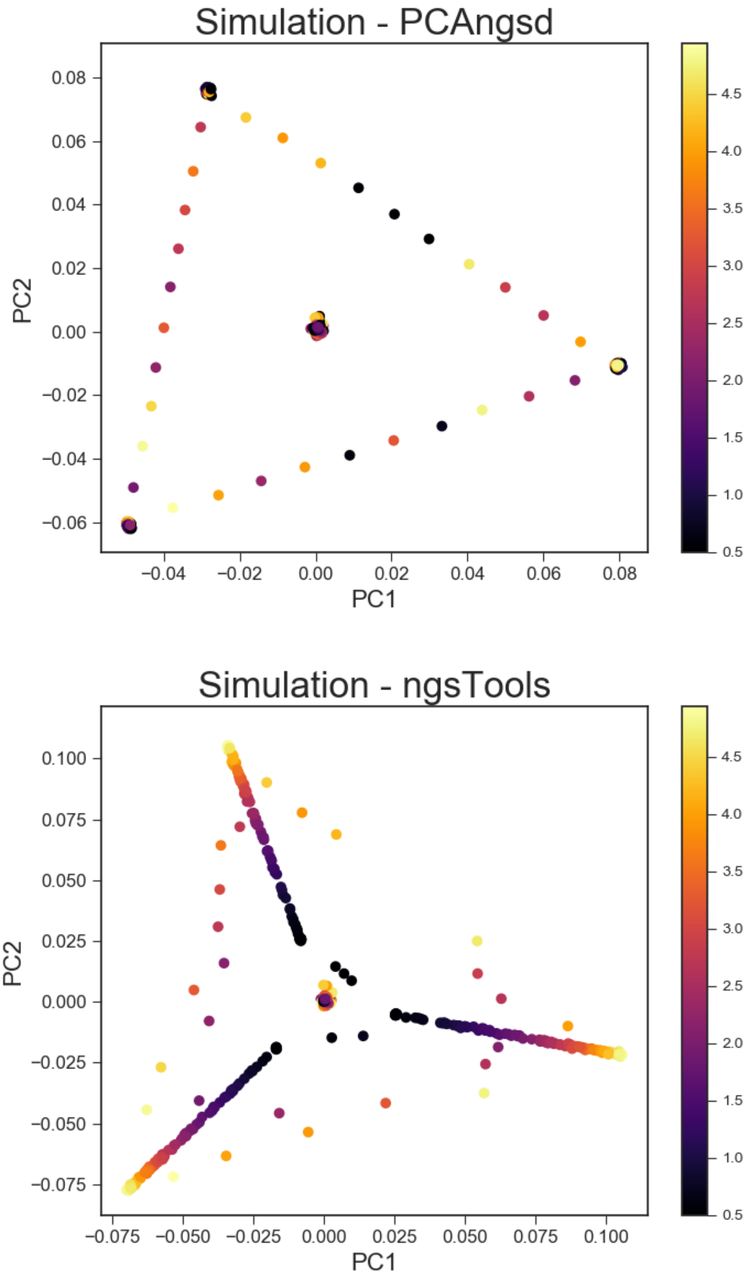


Figure 9: PCA plots of the simulated dataset with individuals colored by their individual sampled sequencing depth. The upper PCA plot is of PCAngsd and the bottom is of the ngsTools model.

6.4.2 1000 Genomes

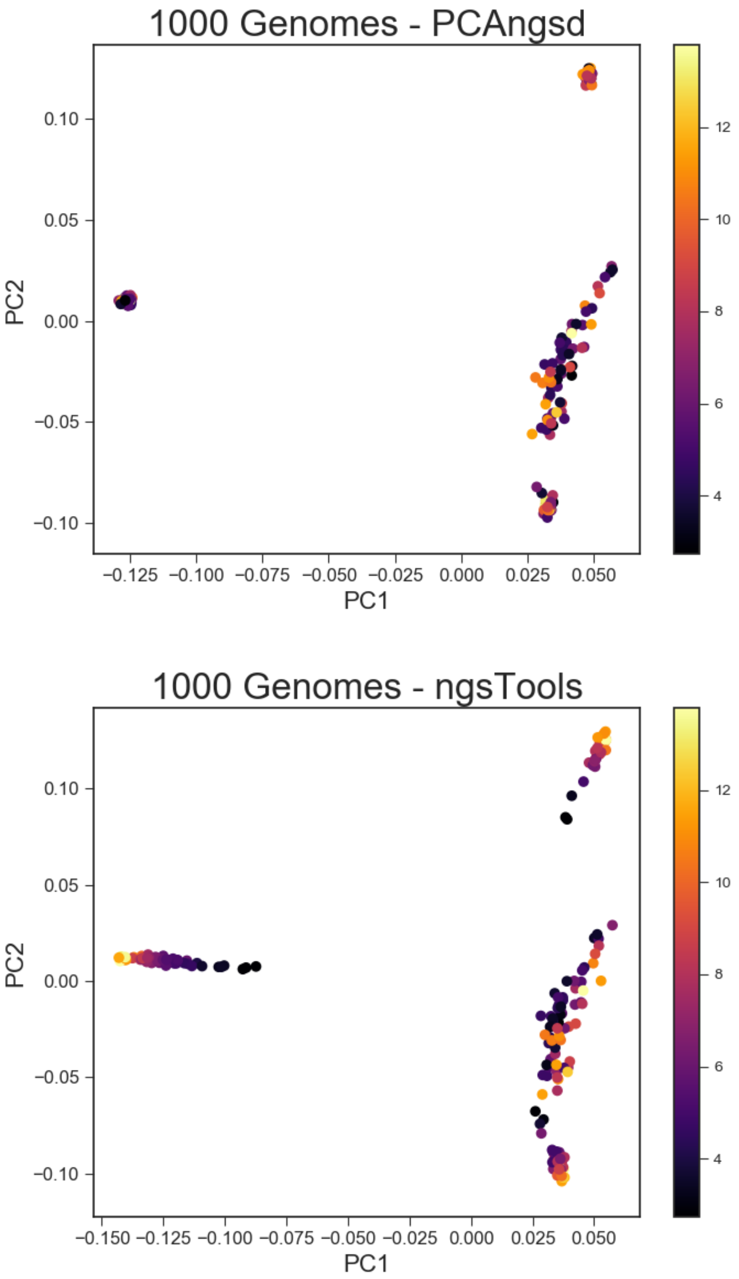


Figure 10: PCA plots of the 1000 Genomes dataset with individuals colored by their individual sequencing depth. The upper PCA plot is of PCAngsd and the bottom is of the ngsTools model. The sequencing depths are estimated in ANGSD [21].

6.4.3 Waterbuck

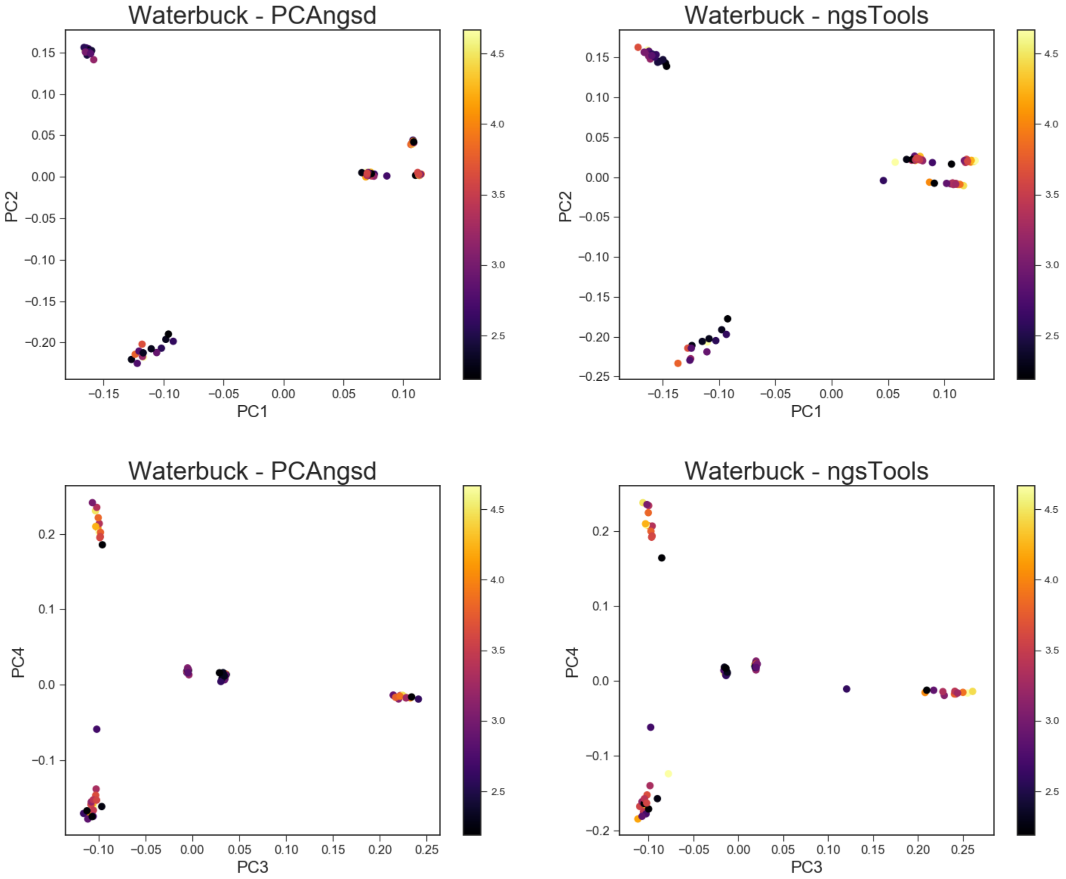


Figure 11: PCA plots of the waterbuck dataset with individuals colored by their individual sequencing depth. The PCA plots of the left column are of PCAngsd and the plots of the right column are of the ngsTools model. The sequencing depths are estimated in ANGSD [21].

6.5 NMF α parameter

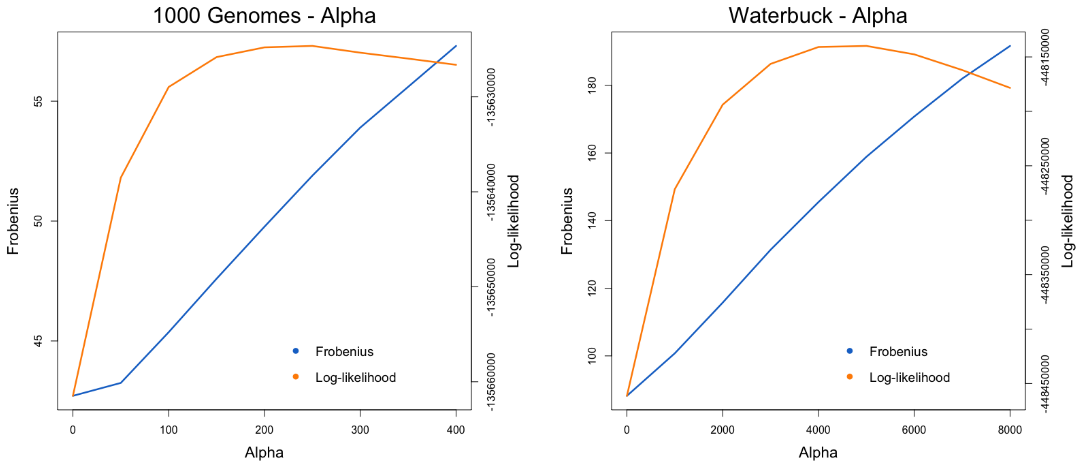


Figure 12: Combined plots of the Frobenius error and likelihood measure obtained using different α values in the estimation of admixture proportions for the real datasets. The left figure shows the plot for the 1000 Genomes dataset with an optimal $\alpha = 250$ in terms of maximizing the likelihood measure. The right figure shows the same for the waterbuck dataset with an optimal $\alpha = 5000$. $B = 5$ was used in both cases.

6.6 ANGSD

Genotype likelihoods and SNP calling

```
./angsd -bam ngslist -GL 1 -out ngsGL -doGlf 2 -doMajorMinor 1 -doMaf 2 -minMaf 0.05 -SNP_pval 1e-6 -minInd 170 -rf chrFile -doDepth 1 -doCounts 1 -P 20
```

References

1. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nature genetics* **36,** 512–517 (2004).
2. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201,** 786–792 (1978).
3. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics* **40,** 646–649 (2008).
4. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS genet* **2,** e190 (2006).
5. Fumagalli, M. *et al.* Quantifying population genetic differentiation from next-generation sequencing data. *Genetics* **195,** 979–992 (2013).
6. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38,** 904 (2006).
7. Galinsky, K. J. *et al.* Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics* **98,** 456–472 (2016).
8. Hao, W., Song, M. & Storey, J. D. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics* **32,** 713–721 (2015).
9. Luu, K., Bazin, E. & Blum, M. G. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources* **17,** 67–77 (2017).
10. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).

11. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology* **28,** 289–301 (2005).

12. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19,** 1655–1664 (2009).

13. Skotte, L., Korneliussen, T. S. & Albrechtsen, A. Estimating individual admixture proportions from next generation sequencing data. *Genetics* **195,** 693–702 (2013).

14. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11,** 459–463 (2010).

15. Metzker, M. L. Sequencing technologies–the next generation. *Nature reviews. Genetics* **11,** 31 (2010).

16. Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y. & Wang, J. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PloS one* **7,** e37558 (2012).

17. Consortium, 1. G. P. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

18. Consortium, 1. G. P. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

19. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12,** 443–451 (2011).

20. Kim, S. Y. *et al.* Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC bioinformatics* **12,** 231 (2011).

21. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC bioinformatics* **15,** 356 (2014).

22. Vieira, F. G., Fumagalli, M., Albrechtsen, A. & Nielsen, R. Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. *Genome research* **23,** 1852–1861 (2013).

23. Kousathanas, A. *et al.* Inferring heterozygosity from ancient and low coverage genomes. *Genetics* **205,** 317–332 (2017).

24. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–1303 (2010).

25. Skotte, L., Korneliussen, T. S. & Albrechtsen, A. Association Testing for Next-Generation Sequencing Data Using Score Statistics. *Genetic epidemiology* **36,** 430–437 (2012).

26. Fumagalli, M., Vieira, F. G., Linderoth, T. & Nielsen, R. ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics* **30,** 1486–1487 (2014).

27. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics* **98,** 127–148 (2016).

28. Shriner, D. Investigating population stratification and admixture using eigenanalysis of dense genotypes. *Heredity* **107,** 413–420 (2011).

29. Velicer, W. F. Determining the number of components from the matrix of partial correlations. *Psychometrika* **41,** 321–327 (1976).

30. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & François, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196,** 973–983 (2014).

31. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences* **101,** 4164–4169 (2004).

32. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791 (1999).

33. Lee, D. D. & Seung, H. S. *Algorithms for non-negative matrix factorization* in *Advances in neural information processing systems* (2001), 556–562.

34. Hoyer, P. O. *Non-negative sparse coding* in *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on* (2002), 557–565.

35. Gillis, N. & Glineur, F. Nonnegative factorization and the maximum edge biclique problem. *arXiv preprint arXiv:0810.4225* (2008).

36. Gillis, N. & Glineur, F. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural computation* **24,** 1085–1105 (2012).

37. Serizel, R., Essid, S. & Richard, G. *Mini-batch stochastic approaches for accelerated multiplicative updates in nonnegative matrix factorisation with beta-divergence* in *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on* (2016), 1–6.

38. Kasai, H. Stochastic variance reduced multiplicative update for nonnegative matrix factorization. *arXiv preprint arXiv:1710.10781* (2017).

39. Jones, E., Oliphant, T. & Peterson, P. {SciPy}: open source scientific tools for {Python} (2014).

40. Lehoucq, R. B., Sorensen, D. C. & Yang, C. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods* (Siam, 1998).

41. Walt, S. v. d., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering* **13,** 22–30 (2011).

42. Lam, S. K., Pitrou, A. & Seibert, S. *Numba: A llvm-based python jit compiler* in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* (2015), 7.

43. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296,** 261–262 (2002).

44. Pedersen, C.-E. T. *et al. Genomic patterns of subspecies divergence in the African Waterbuck (Kobus ellipsiprymnus), UNPUBLISHED* 2018.

45. Wang, C. *et al.* Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Statistical applications in genetics and molecular biology* **9** (2010).

46. Kim, J., He, Y. & Park, H. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization* **58,** 285–319 (2014).