Inferring Population Structure and Admixture Proportions in Low Depth NGS Data

Jonas Meisner & Anders Albrechtsen

Supplementary Material

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 base quality score $base_Q$ is usually in Phred scale such that the probability of an error in the observed base call is given by $\epsilon = 10^{\frac{-base_Q}{10}}$. 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 [1, 2]. The genotype likelihood for individual i in site s can be defined as follows for a multi-allelic case derived from the approach in [3]:

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

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.

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) [4] as follows by assuming independence between all n individuals:

$$\mathcal{L}(p_s) = P(\mathbf{X}_s \,|\, p_s) \propto \prod_{i=1}^{n} P(X_{is} \,|\, p_s). \tag{2}$$

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 possible genotypes. Thus for individual i in site s, $P(X_{is} \,|\, p_s)$ can now be defined as:

$$P(X_{is}\,|\,p_s) = \sum_{g=0}^{2} P(X_{is}\,|\,G=g)P(G=g\,|\,p_s)\,, \tag{3}$$

where $P(X_{is}\,|\,G_{is}=g)$ is the genotype likelihood and $P(G_{is}=g\,|\,p_s)$ is the genotype probability. By assuming Hardy-Weinberg equilibrium (HWE) in the whole sample, the 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:

$$\hat{p}_s^{(\mathrm{ML})} = \underset{p_s}{\operatorname{argmax}} \prod_{i=1}^{n} P(X_{is}\,|\,p_s)\,. \tag{4}$$

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:

$$P(G_{is}=g\,|\,X_{is},p_s) = \frac{P(X_{is}\,|\,G_{is}=g)P(G_{is}=g\,|\,p_s)}{\sum_{g'=0}^{2} P(X_{is}\,|\,G_{is}=g')P(G_{is}=g'\,|\,p_s)}\,. \tag{5}$$

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

$$\mathbb{E}[G_{is}\,|\,X_{is},p_s] = \sum_{g=0}^{2} gP(G_{is}=g\,|\,X_{is},p_s)\,. \tag{6}$$

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 ($\mathbb{E}[G]=2p$):

$$\hat{p}_s^{(t+1)} = \frac{\sum_{i=1}^{m} \mathbb{E}[G\,|\,X_{is},\hat{p}_s^{(t)}]}{2n}\,. \tag{7}$$

PCA - Sequencing depth

Simulated dataset

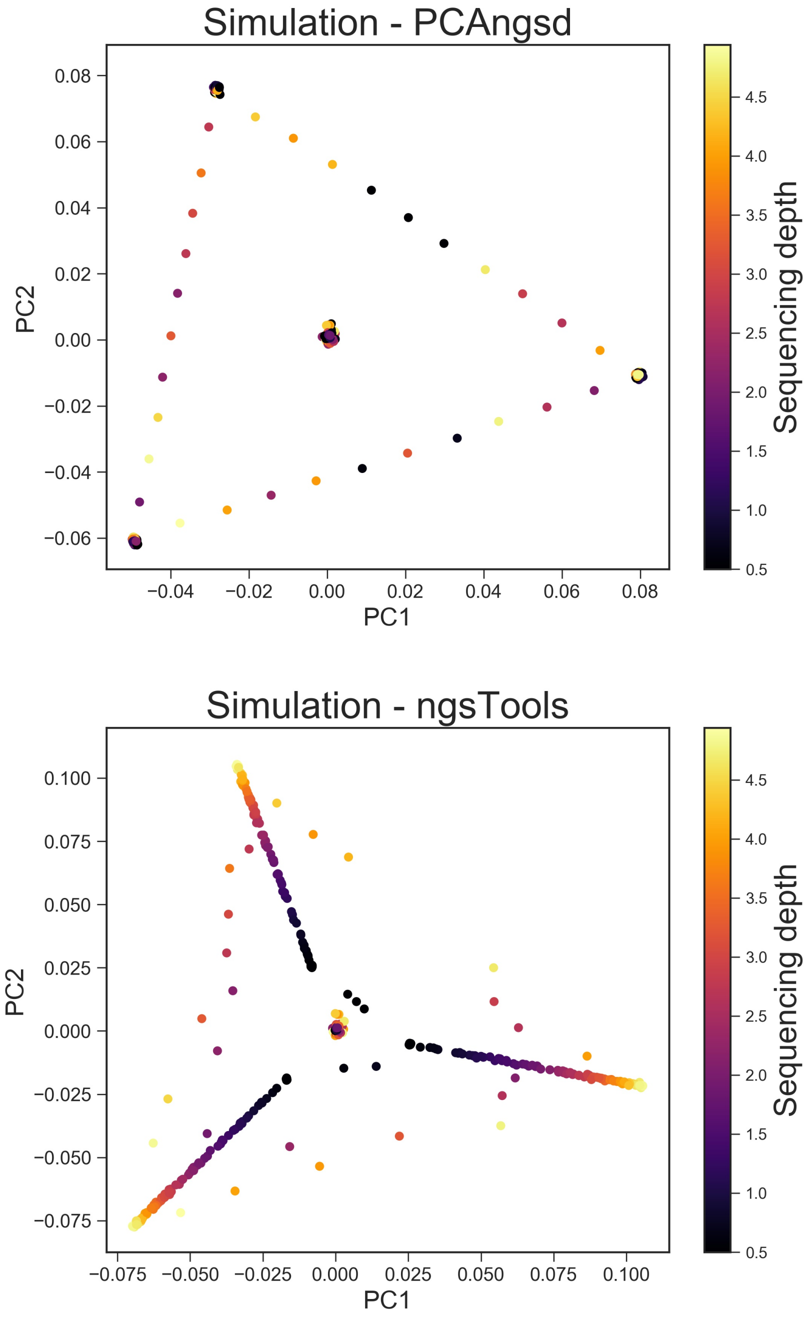


Figure S1: PCA plots of the simulated dataset as in Figure 1 but with individuals colored by their individual sampled sequencing depth. The upper PCA plot is of PCAngsd and the bottom is of the ngsTools model.

1000 Genomes dataset

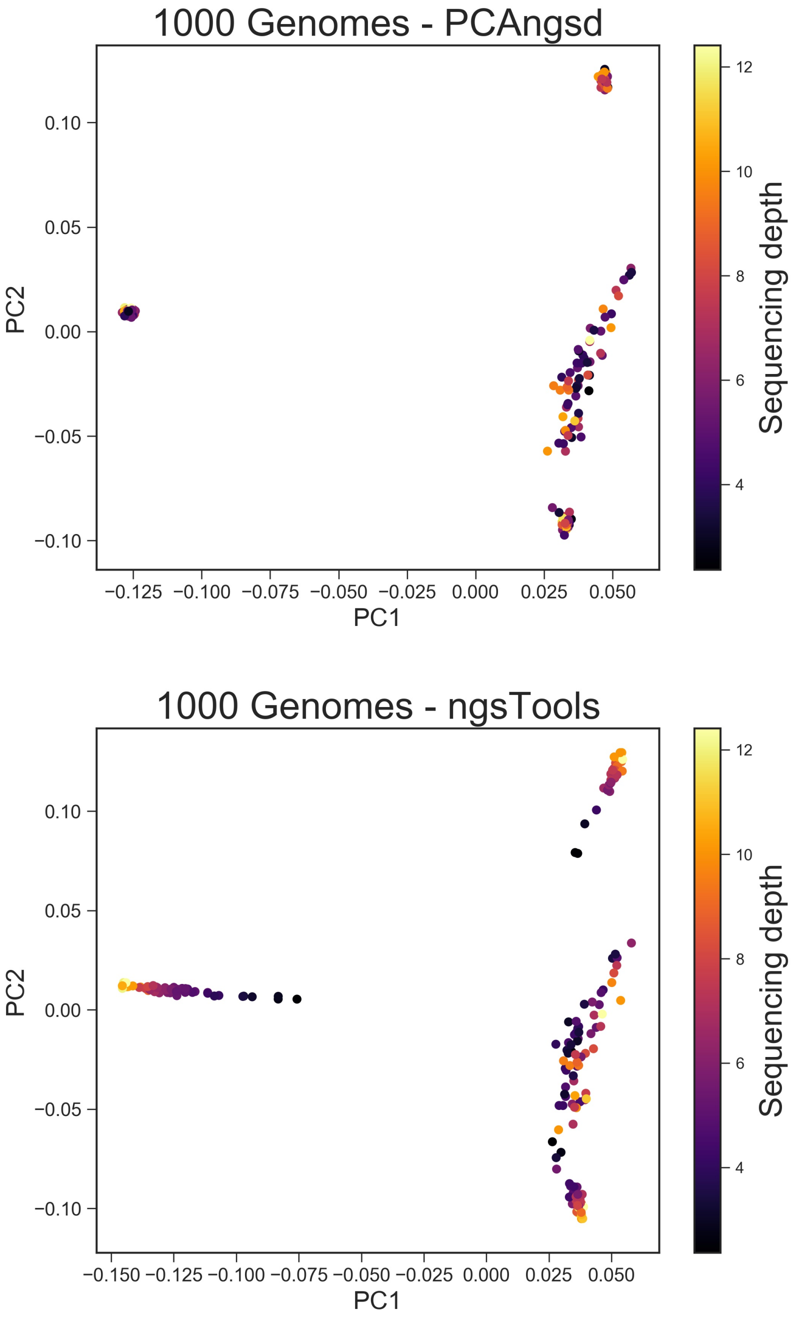


Figure S2: PCA plots of the 1000 Genomes dataset as in Figure 3 but with individuals colored by their individual sequencing depth. The upper PCA plot is of PCAngsd and the bottom is of the ngsTools model. The sequencing depths are estimated in ANGSD [5].

Waterbuck dataset

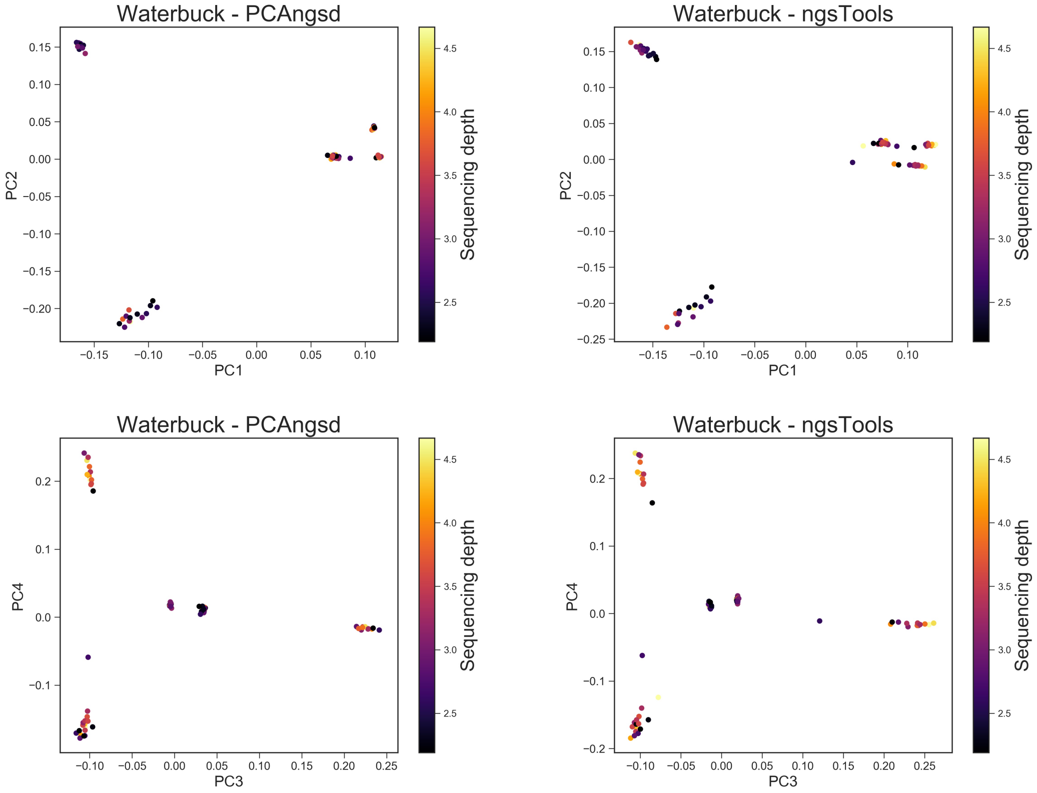
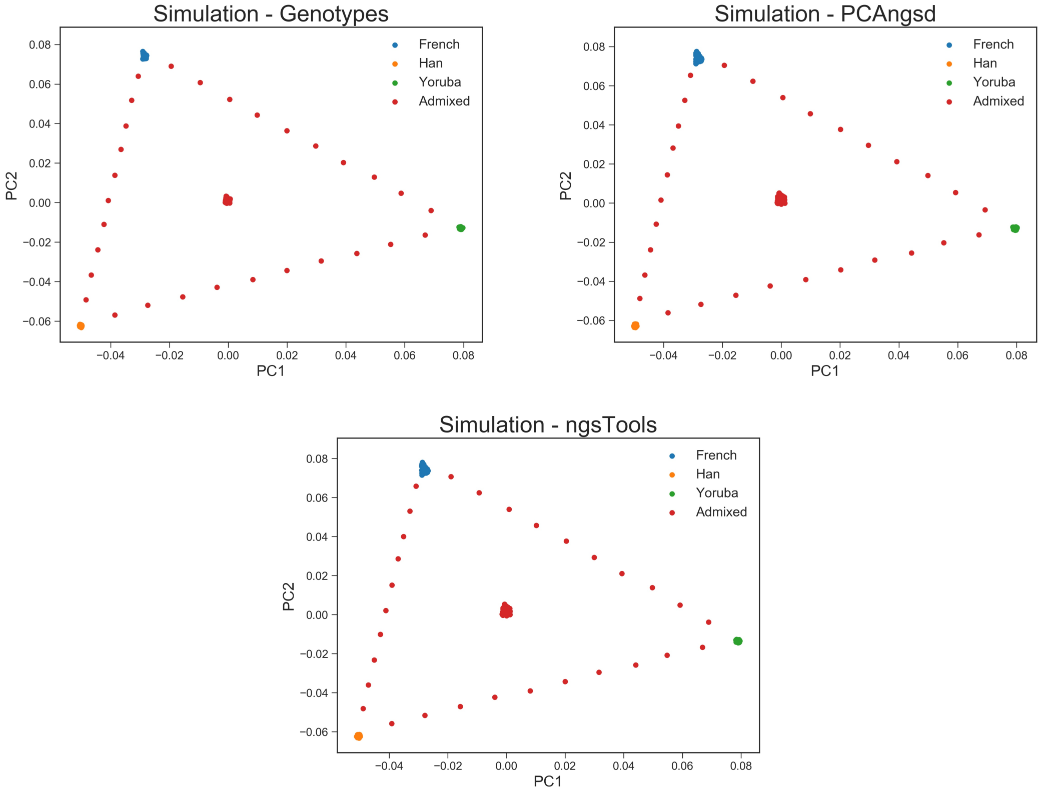


Figure S3: PCA plots of the waterbuck dataset as in Figure 6 but with individuals colored by their individual sequencing depth. The PCA plots of the left column are of PCAngsd and the plots of the right column are of the ngsTools model. The sequencing depths are estimated in ANGSD [5].

Simulated dataset - Equal sequencing depth (2.5X)


Simulation - Genotypes

Simulation - PCAngsd

Simulation - ngsTools

Figure S4: PCA plots of a simulated dataset consisting of 380 individuals and 0.4 million variable sites. Each individual has been simulated with a sequencing depth of 2.5X in the same way as described in the Material and Methods section. The upper left plot shows the PCA performed on the known genotypes using equation 2, the upper right plot shows the PCA performed using PCAngsd and the bottom plot is the PCA performed using the ngsTools model. Procrustes analyses showed a RMSD value of 0.000508 for PCAngsd and 0.000721 for the ngsTools model when compared to the inferred principal components of the genotypes.

Downsampling 1000 Genomes dataset

The 1000 Genomes dataset has been downsampled at different rates S using ANGSD [5] to test the robustness of PCAngsd. S represents the fraction of sequencing reads kept in the estimation of genotype likelihoods, and here $S = 0.25, 0.10, 0.05, 0.01$ have been tested. The same filters have been used to call SNPs and generate genotype likelihoods for all four downsampling rates in ANGSD, which will have an effect when the downsampling rate becomes small as seen below for the number of variable sites evaluated. The filters applied can be seen in the command-line example for ANGSD below. The four different downsampling rates yield the following sequencing depths:

- $S = 0.25$: $0.42 - 3.23$X (7.5 million sites)
- $S = 0.10$: $0.18 - 1.38$X (6.5 million sites)
- $S = 0.05$: $0.11 - 0.79$X (3.4 million sites)
- $S = 0.01$: $0.23 - 3.30$X (2412 sites)

PCA plots

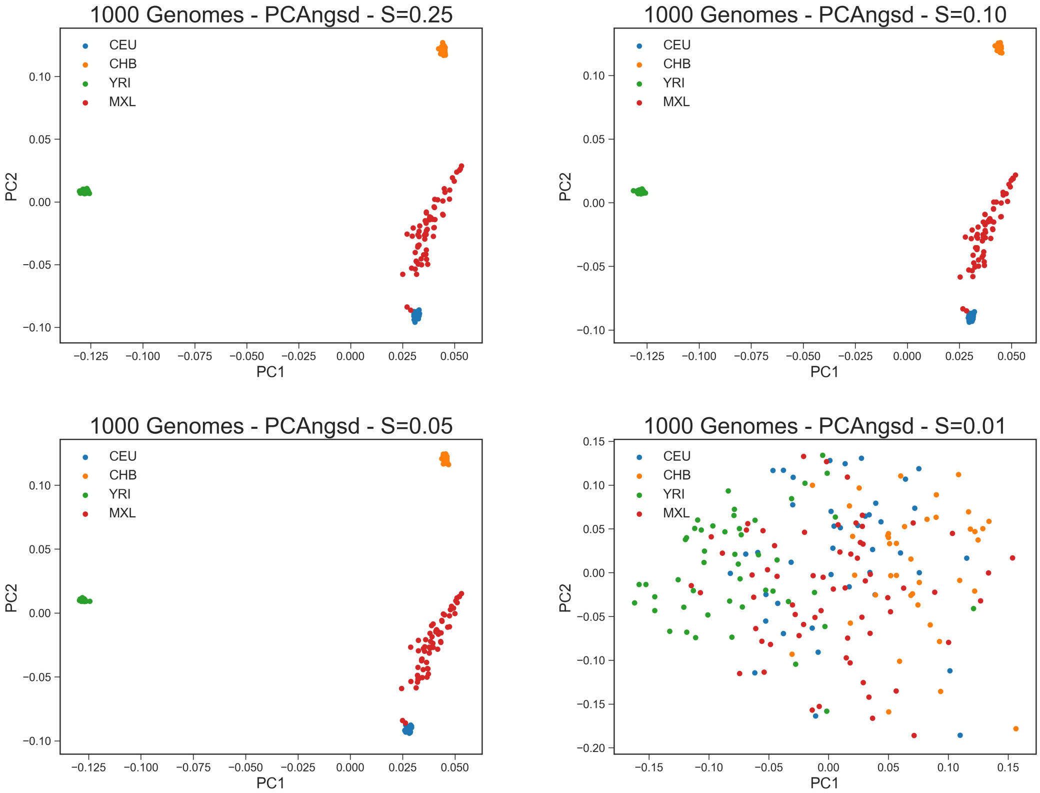
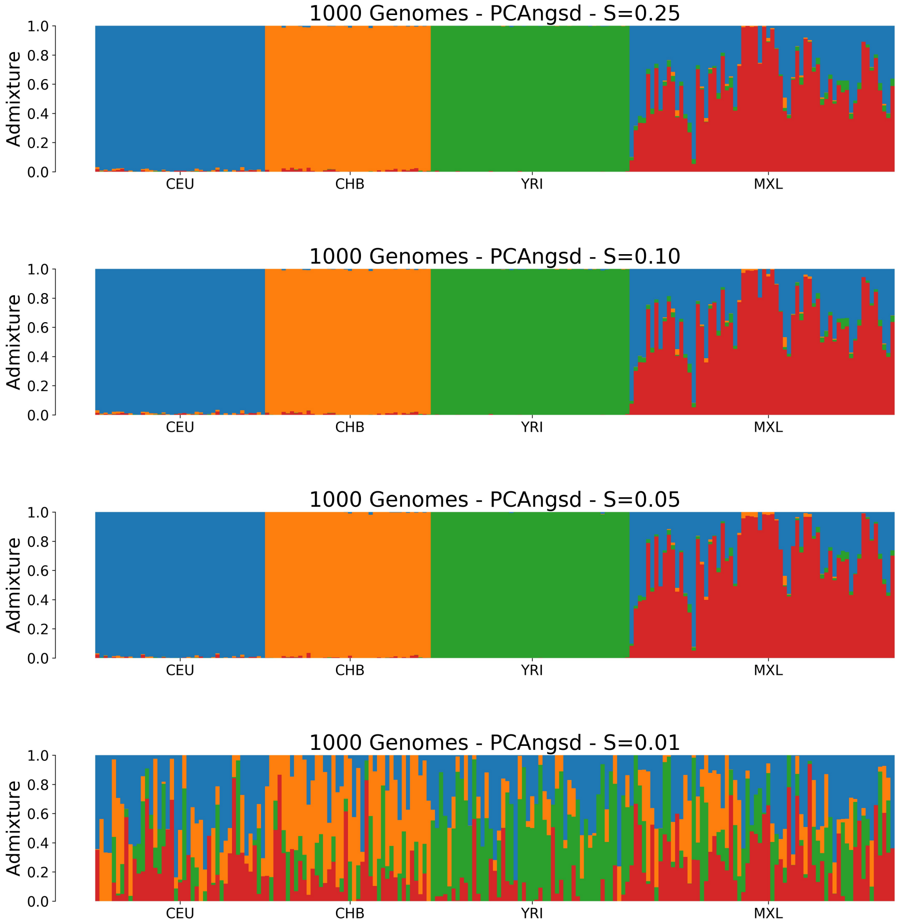


Figure S5: PCA plots of the downsampled 1000 Genomes datasets performed using PCAngsd. The upper left plot is performed with a downsampling rate of $S = 0.25$, the upper right plot is performed with a downsampling rate of 0.10, the bottom left plot is performed with a downsampling rate of 0.05 and the bottom right is performed with a downsampling rate of 0.01. Procrustes analyses reported RMSD values of 0.00292, 0.00356, 0.00399 and 0.0827 for the four plots, respectively, in comparison to the PCA performed on the reliable genotypes using equation 2.

Admixture proportions



1000 Genomes - PCAngsd - S=0.25

1000 Genomes - PCAngsd - S=0.10

1000 Genomes - PCAngsd - S=0.05

1000 Genomes - PCAngsd - S=0.01

Figure S6: Admixture plots for $K = 4$ of the downsampled versions of the 1000 Genomes dataset performed using PCAngsd. The upper plot shows the admixture plot with a downsampling rate of $S = 0.25$ and $\alpha = 815$, the second plot shows the admixture plot with a downsampling rate of 0.10 and $\alpha = 665$, the third plot is the admixture plot with a downsampling rate of 0.05 and $\alpha = 555$ and the bottom plot is the admixture plot with a downsampling rate of 0.01 and $\alpha = 1.5$. The RMSD values are 0.0173, 0.0331, 0.0530 and 0.347 for the four plots, respectively, in comparison to the admixture proportions estimated using ADMIXTUE [6] from the reliable genotypes.

Naive genotype calling

Genotypes have naively been called from genotype likelihoods with a uniform prior such that the calling is based on the highest genotype likelihood. Skotte et al. [7] showed that genotypes called from the genotype likelihoods performed better than genotypes called from posterior genotype probabilities using the population allele frequencies as prior for inferring population structure in low depth scenarios for samples of diverse ancestry. Here we show the population structure inferences from the called genotypes in the simulated and 1000 Genomes datasets including a downsampled scenario for $S = 0.05$ as seen above.

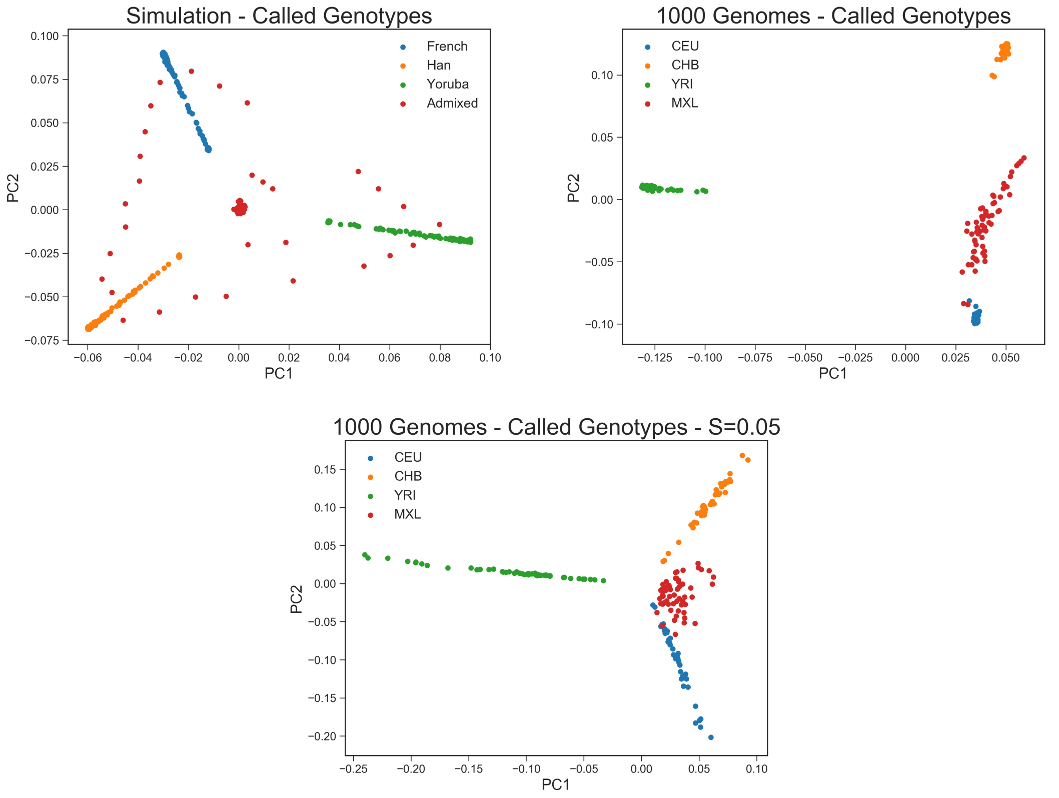


Figure S7: PCA plots performed on the called genotypes from genotype likelihoods with a uniform prior. The upper left plot is performed on the simulated dataset, the upper right plot is performed on the 1000 Genomes dataset and the bottom plot is performed on the downsampled 1000 Genomes dataset using $S = 0.05$. Procrustes analyses showed RMSD values of 0.0123, 0.00310 and 0.0296, respectively, in comparison to the PCA performed on the known genotypes of both datasets (compared to RMSD values of 0.00121, 0.00182 and 0.00399 using PCAngsd).

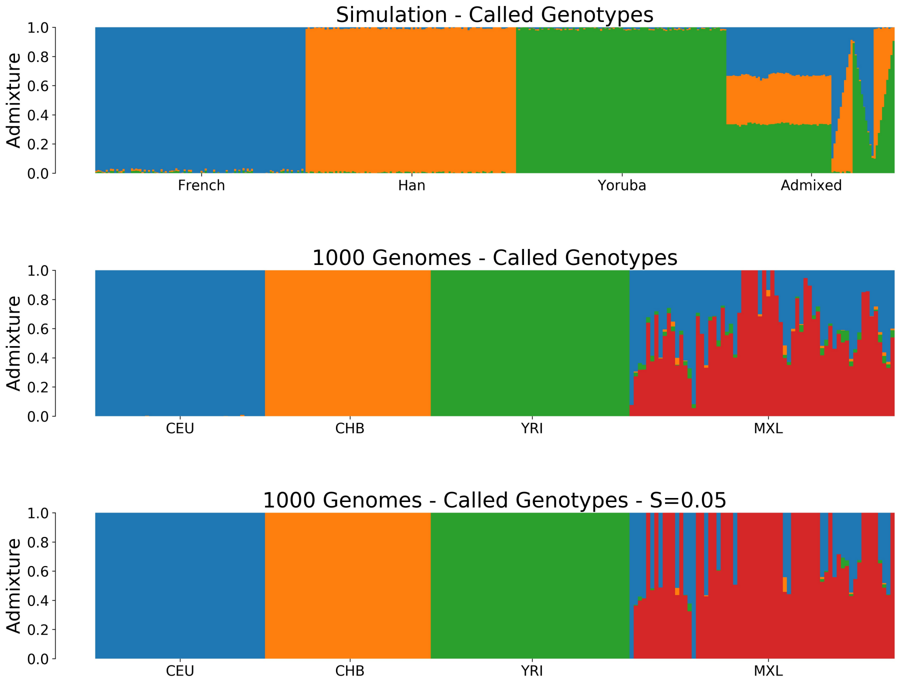


Figure S8: Admixture plots estimated using ADMIXTURE [6] on the called genotypes from genotype likelihoods with a uniform prior. The top plot is estimated from the simulated dataset, the middle plot is estimated from the 1000 Genomes dataset and the bottom plot is performed on the downsampled 1000 Genomes dataset using $S = 0.05$. The RMSD values are 0.00995, 0.00865 and 0.0994 for the three plots, respectively, in comparison to the admixture proportions estimated from the genotypes of both datasets (compared to RMSD values of 0.00476, 0.0108 and 0.0530 using PCAngsd).

Individual allele frequencies

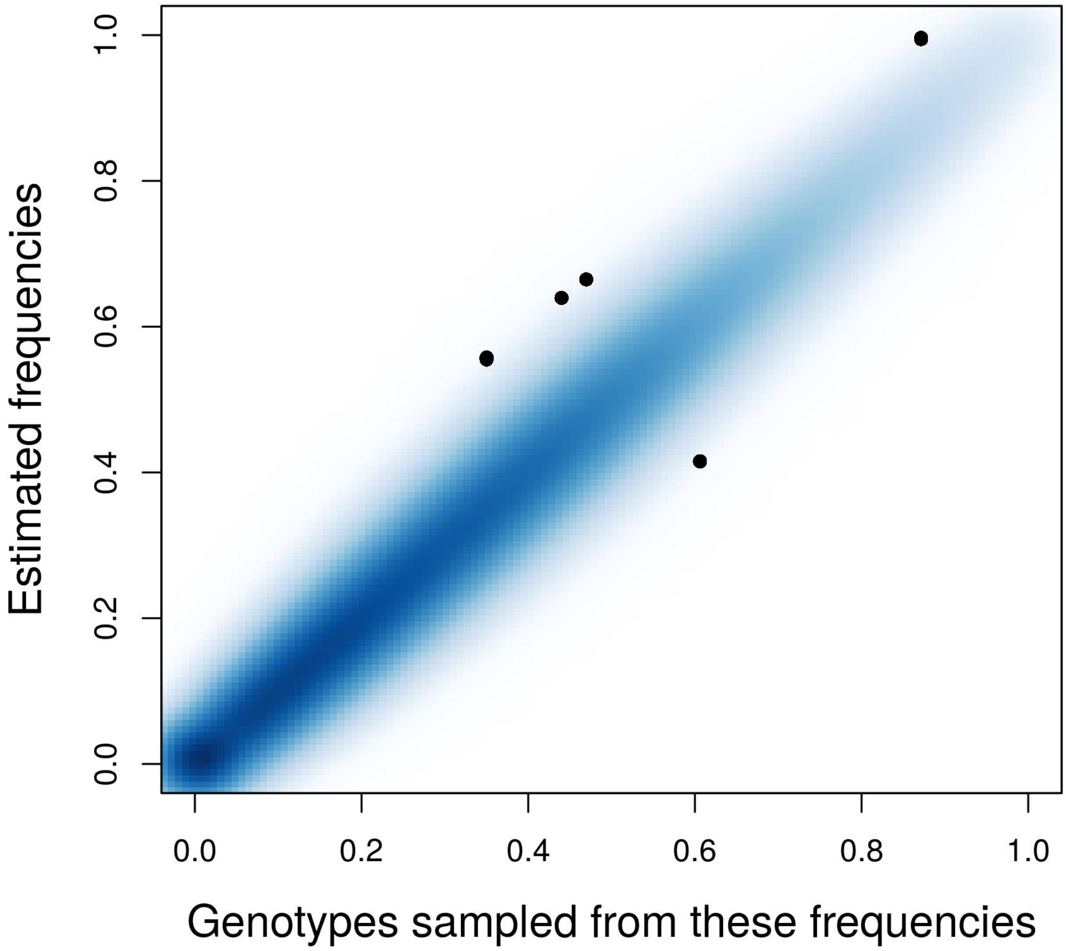


Figure S9: Smooth scatterplot of the frequencies from which the simulated genotypes have been sampled from against the individual allele frequencies estimated using PCAngsd. Outliers are highlighted with black dots.

NMF α parameter

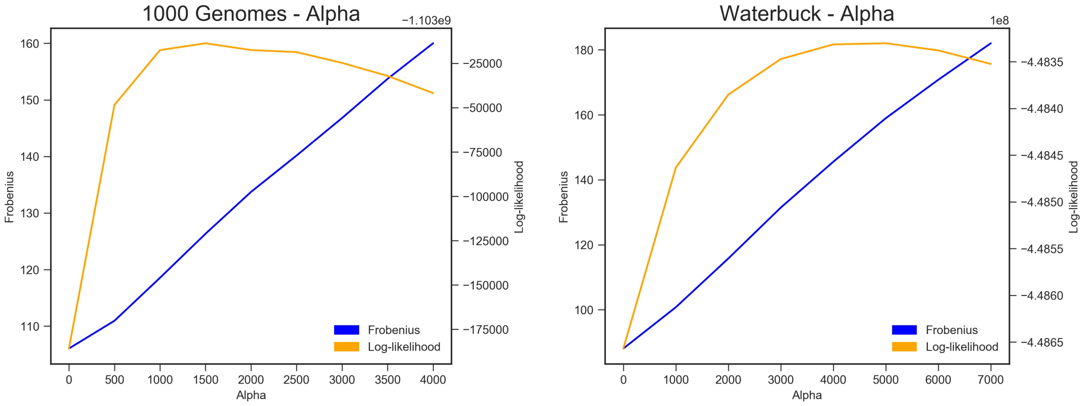


Figure S10: Combined plots of the Frobenius error and likelihood measure obtained using different α values in the estimation of admixture proportions for the real datasets. The left figure shows the plot for the 1000 Genomes dataset with an optimal $\alpha = 1500$ in terms of maximizing the likelihood measure. The right figure shows the same for the waterbuck dataset with an optimal $\alpha = 5000$. $B = 5$ was used in both cases.

Command-line examples

ANGSD

Call SNPs and estimate genotype likelihoods for the low coverage 1000 Genomes dataset using ANGSD [5].

```
./angsd -bam 1000g_bamlist -GL 1 -out 1000g_GL -doGlf 2 -doMajorMinor 1
-doMaf 2 -minMaf 0.05 -SNP_pval 1e-6 -minQ 20 -minMapQ 30 -skipTriallelic 1
-minInd 50 -rf chrFile -doDepth 1 -doCounts 1 -P 20
```

PCAngsd

Perform PCA and estimate admixture proportions for the genotype likelihoods of the low coverage 1000 Genomes dataset using PCAngsd.

```
python pcangsd.py -beagle 1000g_GL.beagle.gz -o 1000g_pcangsd -threads 20
-admix -admix_alpha 1500
```

References

1. 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).
2. 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).
3. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–1303 (2010).
4. Kim, S. Y. *et al.* Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC bioinformatics* **12,** 231 (2011).

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

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

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