PNAS
www.pnas.org

Supplementary Information for

Genomic evidence for shared common ancestry of East African hunting-gathering populations and insights into local adaptation

Authors: Laura Scheinfeldt, Sameer Soi, Charla Lambert, Wen-Ya Ko, Aoua Coulibaly, Alessia Ranciaro, Simon Thompson, Jibril Hirbo, William Beggs, Muntaser Ibrahim, Thomas Nyambo, Sabah Omar, Dawit Woldemeskel, Gurja Belay, Alain Froment, Junhyong Kim, Sarah Tishkoff

Sarah Tishkoff
Email: tishkoff@pennmedicine.upenn.edu

This PDF file includes:

Supplementary text
References for SI reference citations
Figs. S1 to S19
Tables S1 to S10
Captions for dataset tables S1 to S3

Other supplementary materials for this manuscript include the following:

Dataset Tables S1 to S3

www.pnas.org/cgi/doi/10.1073/pnas.1817678116

Supplementary Information Text

Details of STRUCTURE analysis using modern and inferred ancestral allele frequencies

A description of patterns of substructure at K = 9 is described in the main text. We also examined patterns at lower values of *K*. At *K*=2, AAC 1 (blue) represents Saharan ancestry and AAC 2 (green), sub-Saharan ancestry. At *K*=3, AAC 3 (orange) is most commonly found in individuals that speak languages belonging to the NC language family. We compared the distribution of AAC 3 in individuals speaking NC languages to individuals speaking AA, NS, and Khoisan languages using the Wilcoxon rank-sum test and found individuals speaking NC languages to have a significantly greater distribution of AAC 3 when compared to individuals not speaking NC languages (W=126,094; *p*-value $<1.0\text{x}10^{-16}$).

At *K*=4 we observe an AAC (light blue) that is found at highest frequency in Sabue individuals, 83.2% on average; we compared the distribution of AAC 4 in Sabue individuals to all other ethno-linguistic groupings and combined *p*-values with Fisher's method, testing for an excess of low *p*-values. We found a statistically significant enrichment of low *p*-values (χ^2=1,731.3, *d.f.*=100; *p*-value$<1.0\text{x}10^{-16}$) indicating the specificity of this AAC to the Sabue. This AAC is also found at high frequency in Hadza individuals (60.2% on average), which was significantly greater in all pair-wise comparisons with other ethnicities save the Sabue, Aari (AA), Hamer (AA) and Pokot (NS) after adjusting *p*-values for FDR (1). Furthermore, we found a statistically significant excess of low *p*-values in the Hadza comparisons, indicating that the Hadza share a significant proportion of the AAC enriched in the Sabue (χ^2=1,274.4, *d.f.*=100; *p*-value $<1.0\text{x}10^{-16}$). The Aari, Hamer and Pokot populations are located relatively close to the Sabue geographically (224-660 km); the Shabo language is hypothesized to have borrowed linguistic features or descended from the languages spoken by the ancestors of these populations (2). In contrast, the Hadza live more than 1,250 km away and have no previously documented cultural connection, linguistic or otherwise, with the Sabue.

At *K*=5, an AAC (yellow) is inferred that is at 59.5% frequency in Hadza individuals on average; when compared to other populations the distribution of this AAC was significantly greater in the Hadza than all other populations (χ^2=1,799.0, *d.f.*=100; p-value $<1.0\text{x}10^{-16}$). The levels of AAC 5 amongst San individuals (39.5% on average) was found to be significantly lower in other populations, save the Hadza (χ^2=1,692.8, *d.f.*=100; *p*-value $<1.0\text{x}10^{-16}$), which is also of note because of the geographic distance between the Hadza and San and the contentious classification of their respective languages into the same language family based largely on sharing a dental click phoneme (3).

At *K*=6, the genetic structure distinguished in the haplotypes cluster and genotype modes are discordant. At *K*=7, both types of markers produce consistent STRUCTURE runs, with two AACs emerging—one AAC (dark green) is common in RHG populations and one AAC (red) is common in NS-and AA-speaking individuals. The AAC (dark green) is found at highest frequencies in the Biaka, Baka, Bakola, and Bedzan (Western RHG) than in other populations (χ^2=1731.3, *d.f.*=92; *p*-value <1.0x10^{-16}). The Mbuti (Eastern RHG) have the second highest frequency of this AAC (χ^2=1475.5, d.f.=92; *p*-value < 1.0x10^{-16}). The distinct ancestry of the Mbuti is characterized by greater San ancestry than the Western RHG and less NC ancestry: the second highest frequency AAC in the Mbuti is found at high frequency in San individuals (28.2% frequency on average); in contrast, this AAC is found in the Western Pygmy at only 5.3% frequency on average (*W*=1,470, *p*-value= 2.6x10^{-10}). While the NC AAC is found amongst the Mbuti, 9.2% frequency on average, the proportion of NC ancestry is significantly higher in the Western Pygmy individuals, 35.5% on average (*W*=1,470, p= 2.6x10^{-10}). At *K*=8, an AAC (pink) emerges that is at highest frequency in Dahalo individuals when compared to other populations (χ^2=1,578.5*, d.f.*=100; *p*-value <1.0x10^{-16}).

We also examined the genetic relationship between AACs based on the inferred ancestral allele frequencies themselves. PCA of the ancestral allele frequencies for each AAC show clustering of EHG, consistent with a signal of shared ancestry amongst these groups. The first four principal components of the ancestral allele frequency matrix (*K*=9) account for 26.56%, 16.28%, 13.25%, and 12.34% of the variance, respectively. Results are shown for PC 1 to 3, which in total explain 56% of the variance in the data shown in **Fig. S13**, where the AAC's are represented by color: San (light green), Western Pygmy (dark green), Hadza (yellow), Dahalo (dark purple) and Sabue (light purple). The results of the PCA of ancestral allele frequencies inferred by STRUCTURE (AAC-PCA) broadly reflect the results obtained by PCA of individuals' genotypes on PC 1: the San-specific AAC (green) is at one extreme of AAC-PC 1 while the Mozabite-specific is at the other end. In addition, the Western Pygmy-specific AAC is closest to the San-specific AAC while AACs specific to other sub-Saharan groups cluster around the center of AAC-PC 1 at 0. For AAC-PC 2, the relationship with genotype PC 2 is less clear; the NC-specific, NS-specific, and Sabue-specific AACs cluster together at one end of AAC-PC 2 whereas on the genotype PC 2, NS-speaking and NC-speaking but not Sabue individuals are at the extreme end of genotype PC 2. This difference may reflect the evidence for recent common ancestry between the Sabue and NS-speaking populations previously hypothesized (2). Furthermore, on PC 3, four of the HG AACs are closest to each other on the negative-end of the axis: Hadza, San, Sabue, and Dahalo. Thus, while the AACs indicate that genetic drift has differentiated allele frequencies amongst these HG groups, a comparison of the inferred ancestral allele frequencies support a unique common ancestry for them.

Supplemental Identity-by-descent

Tracts of identity-by-descent (IBD) between pairs of haploid genomes, phased using BEAGLE, were obtained using the GERMLINE v2.2 software with "bits" set to 128, "haploid" set to true, "err_hom" set to 3 "-err_het" set to 2 and "-min_m" set to 2154. In total, 797,760 IBD tracts were identified between all pairs of individuals in our sample. Tracts of IBD identified between haploid chromosomes belonging to the same individual were categorized as runs-of-homozygosity (ROH). For each population, we visualized both the mean of the cumulative tract lengths i.e. the sum of all tract lengths between every possible pair of samples within the same ethno-linguistic affiliation, as well as the number of tracts between every single possible pair of individuals of the same ethno-linguistic affiliation.

To compare the number of tracts and their lengths shared within and between populations, we used the model described by Huff et al. (5). The authors developed this model to identify cryptic relatives—up to 12th degree cousins—in population samples such as the CEU in the HapMap sample. Briefly, in this model, the distribution of the number of IBD tracts is assumed to be Poisson with mean equal to the number of generations since diverging from their last common ancestor, censored below a given cutoff. The distribution of tract lengths is assumed to be exponentially distributed with mean inversely proportionally to the number of generations since divergence from the last common ancestor. However, here we use this estimate of the number of generations since two individuals share an ancestor as an IBD-based distance, slightly modified, for summarizing the extent of IBD sharing in a set of individuals. Namely, we calculated a statistic measuring the distance for a given set of IBD tracts-

$$t = \arg\max_{t} e^{-\frac{(r\tau+c)e^{-Ct/100}}{2^{t-1}}} \left[\frac{(rt+c)e^{-Ct/100}}{2^{t-1}} \right]^{|S|} \prod_{i \in S} \frac{100}{t} e^{t(C-i)/100}$$

where S represents the set of IBD tracts greater than C cM in size among a group of individuals, $|S|$ is the cardinality of set S, and the quantity r is the expected number of recombination events in a set of individuals, 35 per individual, and c is the number of chromosomes in a sample of individuals *i.e.* the number of individuals multiplied by 22(5). The estimates were obtained by maximizing the objective function using the R function *optim()*. Unlike Huff *et al.* (5)we did not account for background IBD present in a population and, thus, use the estimated quantity t_{IBD}solely as an IBD-based distance measure between the set of tracts from a chosen set of individuals. We inferred population-specific values of t_{IBD}based on all IBD tracts and ROH tracts within a given population whereas Huff *et al.* calculated t based on the set of IBD tracts identified between a pair of individuals or "within individuals" *i.e.* ROH. We calculated standard errors by jackknifing individuals.

In addition to examining IBD between individuals of the same population, we also examined IBD sharing between individuals of different populations to detect populations sharing. Similar to the within population case, we counted the number of IBD tracts and size of those tracks and estimated the quantity t_{IBD} except in this case the tracts were limited to those identified between individuals of two different populations i and j. We devised a between-population IBD statistic based on t_{IBD} analogous to F_{ST}:

$F_{IBD} = 1 - \frac{\bar{t}^{0}_{IBD}}{\bar{t}_{IBD}}$, where $\bar{t}^{0}_{IBD}$is the average of $\bar{t}^{0}_{IBD}$within a population and $\bar{t}_{IBD}$ is calculated pooling populations together. We created a distance matrix using this quantity and re-constructed a population tree using the NJ algorithm. Broadly, populations in this IBD tree cluster by language family and geography, consistent with the F_{ST}–based tree.

Correlations between among genetics, geography and culture

Several studies of world-wide human diversity as well as studies within Africa have demonstrated that population structure between pairs of populations increases with geographic distance (6). We tested for and found a statistically significant association between Great-Circle (GC) distances, which were log-transformed, and genetic distances, *g*, between all pairs of populations with ≥10 sampled individuals using the Mantel test with 10,000 permutations (M=0.400, p-value=$1.0\text{x}10^{-4}$) implemented in the *ade4* package for *R* (7). We also hypothesized that language and subsistence may impact the distribution of genetic variation as well. To investigate the relationship between linguistic and genetic distance, we created a language "distance matrix" where pairs of populations speaking the same language were coded as 0, and 1, otherwise. We then retained the residuals from a logistic regression of language distance with log-transformed GC distances for the Mantel test with genetic distances (*i.e.* a partial Mantel test), which was statistically significant (M=0.2329, p-value=$1.0\text{x}10^{-4}$). We used a similar procedure for a partial Mantel test with the subsistence distance matrix wherein pairs of populations practicing the same subsistence strategy were coded as 0, and 1, otherwise. As with language, we found subsistence strategy and genetic distance to be positively correlated (partial Mantel test: M=0.08190, p-value=0.00849).

Inferring Demographic History using Approximate Bayesian Computation (ABC) modeling

Motivated by the problem of inferring divergence times between EHG populations, who have experienced gene flow from neighboring populations and have different effective population sizes, using SNP data based on ascertained SNP arrays, we utilized ABC, constructing summary statistics from patterns of allele frequency differences, LD decay, and admixture LD decay in order to infer parameters such as time of divergence and migration.

Simulation framework

We used Hudson's *ms* program to simulate coalescent ARGs given a set of demographic parameters: effective population sizes, divergence times, and migration rates between populations. For these coalescent simulations we set $\theta = 4N_e\mu$, the mutation rate was chosen to be 1.1×10^{-8} per generation per base pair, based on a recent whole-genome sequencing study of Hutterite parent-offspring trios (8). For the parameter $\rho = N_e r$, we allowed for variation in recombination rate *r*, which has been shown to have an influence on observed summary statistics such as LD decay (9). For this we used the sex-averaged recombination map inferred by Kong *et al* (10). However, the SNP array used by the authors had a lower density (approximately 5/6th) than the one we used in our analyses; therefore, we fit a smoothing spline regression to the cumulative recombination distance between the SNPs on the Kong array to infer the recombination rate between SNPs not present on their array. We pre-selected randomly sampled 50-kbp regions from the Kong map and matched their average recombination rates in our simulations.

The demographic model we developed for EHG evolution in the context of agriculturalist and pastoralist expansion is shown in **Fig. S8** Parameters of the coalescent simulations were either randomly generated (from a prior distribution) if unknown or pre-determined based on previous studies (11, 12). The earliest event was the introduction of a population split 4,500 generations from the present (9, 13), corresponding to emergence of population structure in Africa. This population represents the ancestral population leading to modern agriculturalist and pastoralist populations (A/P). The second event was the divergence of a non-African representative population (used for ascertainment) 3,500 generations in the past (9, 13, 14). The simulated out-of-Africa event is accompanied by a population bottleneck (10X reduction in population size from 10,000 to 1,000 individuals) that lasts for 100 generations and is then followed by an exponential population growth to a current population size of 100,000 individuals (N_{OA}) (9, 15). The A/P population grew exponentially from 10,000 to 100,000 individuals (N_{OA}) at 250 generations in the past. In addition, the beginning of the A/P expansion is also accompanied by the onset of gene flow into HG population 1 and 2, parameterized by migration rates M_1 and M_2, respectively (16, 17). The expansion and accompanying gene flow of the A/P population, is intended to be analogous to the population expansions that occurred at the advent of the Neolithic revolution in Africa, which occurred 5-3 kya (18, 19). Thus, the time at which migration from the A/P population into the HG populations commenced (T_{mig}) was drawn from a uniform prior between 100 and 200 generations in the past. In addition, effective population sizes for the HG populations (N_1 and N_2) were drawn from a uniform prior between 10,000 and 30,000 individuals; the ancestral population size of the HG (N_{HG}) was drawn from a uniform distribution on the interval 20,000 to 100,000 individuals; the ancestral population size (N_{anc}) was also drawn from a uniform distribution between 20,000 and 100,000 individuals. Finally, the divergence time of the EHG populations (T_{HG}) was drawn from a wide uniform prior between 200 and 3,000 generations before present.

The SNP ascertainment scheme approximation was applied to each simulated 50 kbp region, before calculating summary statistics. The scheme entailed removing all SNPs with an allele frequency below 5% frequency in the non-African-representative sample from the African-representative populations. Next, pair-wise r^2 was calculated as described above for all SNPs in the non-African population sample; for pairs of SNPs with an $r^2 > 0.7$, one SNP was randomly chosen to be removed from the African-representative sample. Finally, for simulations compared to actual data, loci were randomly sampled in the "African" samples to simulate the density of SNPs observed on the array. This number was generated by first sampling from a log-normal distribution with the same mean and standard-deviation as the log of the number of SNPs in the pre-selected 50-kbp segments on the Illumina 1M-Duo. If the random number was less than a cutoff, a minimum number of SNPs (eight) was selected; otherwise the random number was rounded and that number of SNPs was randomly selected from the region.

Summary statistics

The parameters of our demographic model for the evolution of HG populations included divergence times, current effective population sizes, ancestral effective population sizes, and rates of migration. The f_2 statistic summarizes the deviation of allele frequencies between a pair of populations (20). The f_2 statistic measures the extent of genetic drift between two populations; drift itself is a function of time of divergence and effective population size. For each site f_2 statistics were calculated; the logs of the mean and variance of all f_2 statistics between the EHG and each EHG with the A/P population were used as summary statistics for each simulation.

The shape of the LD decay curve contains information regarding effective population sizes(21). Instead of relying on an analytical relationship between LD and N_e derived from a simple model, the relationship between effective population size and the shape of the curve was empirically explored. A simple, flexible approach for fitting curves is to use polynomial regression. In this case, a 4th degree orthogonal polynomial was fit with to $E[r^2]$ as a function of pairwise distance between SNPs (in kbp) for a population. However, we found estimates $E[r^2]$ to be subject to noise due to sampling variance, especially at SNPs farther apart for which fewer observations were available. Therefore, we fit the polynomial regression to a smoothed LD decay curve, which we estimated through bootstrap aggregated (bagged) local linear regression. The resulting coefficients of the polynomial regression were used as the summary statistics relevant to N_e. Analogously, the coefficients for a 4th degree polynomial regression fit to the admixture LD decay $E[a]$ as a function of pairwise distance between SNPs were used as summary statistics sensitive to migration rate (with smoothing first applied as well). This is in contrast to the approach of Moorjani and Loh (20, 22) who assume a simple model for the change of ALD as a function of admixture rate and time since admixture. In addition, their model assumes a single pulse of admixture as it is a more analytically tractable model (more recently extended to two pulses (23, 24). In contrast, we allow for a continuous migration rate, which is more plausible for East African and Pygmy hunting-gathering populations who have likely co-existed with neighboring agriculturalist and pastoralist populations for extended periods of time (25-29).

Demographic inference

Following Fearnhead and Prangle (30), we used these summary statistics in a pilot stage for constructing lower dimensional summary statistics that corresponded to estimates of the means of the marginal posteriors for each parameter. Thus, for each demographic scenario we simulated B_1 data sets from a prior distribution on the parameters; summary statistics $S(X_{sim})$ were calculated for each simulation.

The gradient boosting machine method (31, 32), as implemented in the R package *gbm* (33), was used to estimate the posterior mean of the summary statistics for each parameter, denoted $\hat{G}_p(S(X_{sim}))$, where p indicates the parameter of interest. Before fitting the GBM, we normalized the parameters, which were each drawn from a uniform distribution, to the range [0.001, 0.999] and then applied the inverse-CDF of the Gaussian, such that the parameters were approximately normally distributed with mean 0 and standard deviation 1. At each iteration of the GBM, we used a random sample of 50% of the data for training a decision tree. To capture interactions between summary statistics, decision trees were allowed to include up to 3 summary statistics; however, to prevent over-fitting only regression trees including at least 50 observations were included; a learning rate of 0.01 was used for regularization purposes. We fit up to 2,000 regression trees. Cross-validation was used to pick the total number of trees with the best out of sample performance.

We sought to ensure that the GBM learned a function approximation invariant to the order in which populations were analyzed *e.g.* predictions for migration rates from the Iraqw into the Hadza and Sandawe should be the same whether the first set of ALD statistics is calculated with the Hadza and Iraqw or Sandawe and Iraqw and *vice versa*. Thus, instead of training separate GBM's for migration rate 1 and 2 we trained a single GBM for each of these pairs of parameters. This was accomplished by only retaining relevant summary statistics for each parameter *e.g.* for migration rate between population 1 and the A/P population we used f_2 between the two HG populations, f_2 between the HG population 1 and the A/P population, coefficients for LD decay in population 1, and coefficients for LD decay in population 1; we then concatenated the corresponding statistics for migration rate into population 2 (*e.g.* f_2 between the HG population 2) to the matrix of statistics for migration rate into population 1, giving us a total of $2B_1$ rows. This process was also employed for effective population size in the HG populations.

For the other parameters that may have had an effect on variation in both extant populations (ancestral population sizes and time of divergence), we trained a GBM that would learn the relationship between the summary statistics and parameters regardless of the order in which the HG populations were analyzed (*e.g.* the divergence time estimated for the Hadza and Sandawe should be the same whether population 1 were assigned to be the Hadza or Sandawe). Thus, the matrix of summary statistics was modified by concatenating a version of the matrix where population specific variables were swapped. For example, in the original matrix the coefficients for LD decay in population 1 were columns j through j+4 and in the coefficients for LD decay in population 2 were columns j+5 through j+9; in the second version of the matrix that was concatenated to the original matrix, these were swapped such that the coefficients for LD decay in population 2 were columns j through j+4 and in the coefficients for LD decay in population 1 were columns j+5 through j+9. In this way, we minimized the effect of the order in which populations were analyzed as *a priori* this has no inherent relevance to the model.

In the second stage B_2 data sets were simulated from the same prior distributions. The original set of summary statistics $S(X_{sim})$ were calculated and the posterior means were predicted using the $\hat{G}_p(S(X_{sim}))$ predicted in the first stage. Thus the algorithm can be given as follows-

1) Stage 1: semi-automatic summary statistic generation

    a. Draw a parameter from prior distribution on θ

    b. Generate B_1 data sets from coalescent simulator

    c. Create genotype data by randomly pairing genotypes

d. Ascertain SNPs in “outgroup” population; retain only those in other populations

e. Calculate summary statistics of simulated $S(X_{sim})$ data

f. Repeat steps 1.a—1.e B_1 times

g. Train GBM on these data: $\hat{G}_p(S(X_{sim}))$

2) Stage 2: Rejection-based posterior inference

a. Repeat steps 1.a—1.e B_2 times

b. Calculate distance between simulated $\hat{G}_p(S(X_{sim}))$ and observed statistics $\hat{G}_p(S(X_{obs}))$

c. Accept a fraction of the B_2 parameters closest to the observed summary statistics, as selected by cross-validation

d. Utilize local linear regression to estimate posterior density (ABC with regression adjustment)

The *abc* library (34) in the *R* environment was used to sample the posterior distribution as described in stage 2 of the algorithm above.

Performance of ABC estimates on simulated data

To test the performance of f_2 and ALD-based summary statistics for inferring the rate and time of admixture while accounting for ascertainment, we simulated a demographic scenario with three populations with 20 individuals each. Values for migration rates ($4N_em$) from population 2 to population 3 were sampled from a uniform prior from 0.1 to 500. The time at which migration initiated varied from 200 to 3000 generations in the past. Furthermore, the current N_e for population 3 varied from 10,000 to 30,000 individuals; the ancestral effective population size values were drawn from a uniform distribution between 20,000 and 100,000. A total of B_1=1,000 simulations were run in the first stage and B_2=1,000 simulations in the second stage.

The quality of the semi-automatic summary statistics was evaluated by comparing the fitted values to the actual values of the parameters (**Figures S14-S17**). We observed good performance from the trained GBMs on simulated data, with the exception of N_1, N_2, and

N_{HG} parameters. However, our primary motivation was to understand divergence time of the EHG; thus, we left the issue of GBM fit for these parameters for future study.

We selected the ABC tolerance, *i.e.* the fraction of parameter values to keep, by leave-one-out cross-validation of divergence time. In each iteration a simulated observation was omitted and its parameter value was inferred using regression-adjusted ABC at a set of increase tolerance levels (1 to 25th percentiles); this procedure was repeated for 20 simulated observations. The mean squared error was calculated for each tolerance level across the 20 chosen points and the tolerance level (20th percentile) with the lowest MSE (144848.7) was chosen for ABC inference on the whole data set. The performance of the CV analysis is shown in **Figure S18**.

In addition, the coverage probability (*i.e.* the probability for the 95% credible for each parameter) was calculated and confirmed, demonstrating that the true parameter was reliably within the credible interval: M_1 (94.5%), M_2 (95.8%), T_{HG} (94.6%), T_{mig} (95.2%), N_1 (93.8%), N_2 (92.4%), N_{HG} (92.3%), N_{anc} (95.9%). An example of a parameter falling into the 95% CI is shown (**Figure S19**).

Application to African SNP data

As we observed strong evidence for common ancestry amongst a set of East African hunting-gathering populations we sought to date the divergence of these populations while accounting for changes in effective population size, gene flow and ascertainment bias. We used ABC to calculate the divergence time between the EHG populations (Hadza, Sandawe, Sabue, and Dahalo). For each possible EHG pair, we included one of three A/P populations—Yoruba (NC), Dinka (NS), and Iraqw (AA)—as a source of gene flow. Thus, we inferred divergence time for 18 population combinations in total. However, for certain combinations we had greater *a priori* belief that the A/P population had contributed gene flow in the past; this was supported with inspection of STRUCTURE results (**Fig. S4)**. For each combination, we calculated the f_2, LD, and ALD summary statistics.

The maximum *a posteriori* (MAP) estimate and 95% credible interval for the divergence time are shown in **Fig. 4** and discussed in the main text. We also examined the posterior estimates for other parameters (**Tables S4-S10**). Notably we observed migration rates consistent with *a priori* knowledge about populations as well as with STRUCTURE results. Namely, we find high levels of gene flow from the Iraqw into Hadza and Sandawe populations, which has been observed in prior genetic studies (35) and also is supported by the fact that the Iraqw live in close proximity to both these EHG populations (36). Our estimates of ancestral Ne in African populations are relatively large ($\sim 1.5 \times 10^5$ across population pairs) compared to other estimates (15).

References

1. Benjamini Y & Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*:1165-1188.
2. Ehret C (1992) Do Krongo and Shabo belong in Nilo-Saharan. *Proceedings of the Fifth Nilo-Saharan Linguistics Colloquium, Nice*, pp 169-193.
3. Guldemann T & Stoneking M (2008) A Historical Appraisal of Clicks: A Linguistic and Genetic Population Perspective. *Annual Review of Anthropology,* Annual Review of Anthropology, (Annual Reviews, Palo Alto), Vol 37, pp 93-109.
4. Felsenstein J (1989) Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164-166.
5. Huff CD*, et al.* (2011) Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome research* 21(5):768-774.
6. Ramachandran S*, et al.* (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* 102(44):15942-15947.
7. Chessel D, Dufour A, & Thioulouse J (2004) The ade4 package. *R news*.
8. Campbell CD*, et al.* (2012) Estimating the human mutation rate using autozygosity in a founder population. *Nature genetics* 44(11):1277-1281.
9. Schaffner SF*, et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome research* 15(11):1576-1583.
10. Kong A*, et al.* (2010) Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319):1099-1103.
11. Stringer CB, Grün R, Schwarcz H, & Goldberg P (1989) ESR dates for the hominid burial site of Es Skhul in Israel. *Nature* 338(6218):756-758.
12. Emery LS, Felsenstein J, & Akey JM (2010) Estimators of the human effective sex ratio detect sex biases on different timescales. *The American Journal of Human Genetics* 87(6):848-856.
13. Gronau I, Hubisz MJ, Gulko B, Danko CG, & Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics* 43(10):1031-1034.
14. Gutenkunst RN, Hernandez RD, Williamson SH, & Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics* 5(10):e1000695.
15. Li H & Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature* 475(7357):493-496.
16. de Filippo C*, et al.* (2011) Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. *Mol Biol Evol* 28(3):1255-1269.
17. Patin E*, et al.* (2009) Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet* 5(4):e1000448.
18. Philipson D (1975) The Chronology of the Iron Age in Bantu Africa. *The Journal of African History* 16(3):321-342.
19. Ehret C (1967) Cattle-Keeping and Milking in Eastern and Southern African History: The Linguistic Evidence *The Journal of African History* 8(1):1-17.
20. Loh P-R*, et al.* (2013) Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193(4):1233-1254.
21. Tenesa A*, et al.* (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome research* 17(4):520-526.
22. Moorjani P*, et al.* (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet* 7(4):e1001373.

23. Henn BM*, et al.* (2012) Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS genetics* 8(1):e1002397.
24. Pickrell JK*, et al.* (2014) Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences* 111(7):2632-2637.
25. Stiles D (1981) Hunters of the northern East African coast: Origins and historical processes. *Africa* 51(04):848-862.
26. Headland TN*, et al.* (1989) Hunter-Gatherers and Their Neighbors from Prehistory to the Present [and Comments and Replies]. *Current Anthropology*:43-66.
27. Stiles D (1992) The hunter-gatherer'revisionist'debate. *Anthropology Today* 8(2):13-17.
28. Kusimba SB (2005) What is a hunter-gatherer? Variation in the archaeological record of eastern and southern Africa. *Journal of Archaeological Research* 13(4):337-366.
29. Bahuchet S (2006) Languages of African rainforest "Pygmy" hunter-gatherers: Lanugage shifts without cultural admixture. *Hunter-gatherers and linguistic history: A global perspective*, eds Güldemann T, McConvell P, & Rhodes R (Cambridge University Press, Cambridge).
30. Fearnhead P & Prangle D (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74(3):419-474.
31. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics*:1189-1232.
32. Friedman JH (2002) Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38(4):367-378.
33. Ridgeway G (2007) Generalized Boosted Models: A guide to the gbm package. *Update* 1(1).
34. Csilléry K, Blum MG, Gaggiotti OE, & François O (2010) Approximate Bayesian computation (ABC) in practice. *Trends in ecology & evolution* 25(7):410-418.
35. Tishkoff SA*, et al.* (2009) The genetic structure and history of Africans and African Americans. *Science (New York, N.Y.)* 324(5930):1035-1044.
36. Tishkoff SA*, et al.* (2009) The genetic structure and history of Africans and African Americans. *Science* 324(5930):1035-1044.

Figure S1. Principal components analysis of individuals' genotypes. A) The percentage of variance explained by the first ten principal components is shown. B) On PC 4, San individuals are on one end of the axis and WRHG individuals are on the other. C) On PC 5, Hadza individuals are on one end of the axis and San individuals are on the other. D) On PC 6, Mbuti individuals are on one end of the axis and WRHG individuals are on the other. E) On PC 7, Sabue individuals are on one end of the axis and AA and NC individuals are on the other. F) On PC 8, Mozabite and Sabue individuals are on one end of the axis and Fulani individuals are on the other. G) On PC 9, Dahalo individuals are on one end of the axis and Elmolo individuals are on the other. H) Displays the proportion of variance explained by each PC.

Figure S2. Regression analysis of PC 1 and 2 with geography and language. A) We fit a linear model relating latitude, longitude as well as different language family and subsistence strategies, as categorical variables, to the projection of indviduals' genotypes onto PC1; the fit of the data was compared to the original data and fits the data well (R2 =0.86; p-value <1.0x10-16). B) We repeated the process for PC 2; however, the linear fit appears to be weaker (R2 =0.56; p- value <1.0x10-16).

Figure S3. The distribution of EHG and Dinka individuals in PC space. A) Euclidean distances were calculated between each EHG individuals using the first two PC's as a two- dimensional space; in addition, Euclidean distances were calculated between every possible pair of an EHG individual with a non-EHG individual. These two distributions are juxtaposed and a statistical test bears out that the EHG individuals are closer to each other than to other individuals in the first two PC's. B) Absolute distances of projection of individuals' genotypes onto PC 3 between all possible pairs of Sabue, Dinka and Hadza individuals compared to all possible pairs of individuals in these three populations with individuals not in these three populations show that Sabue, Dinka and Hadza individuals are closer to each other on PC 3.

Figure S4. STRUCTURE results for K=2 to K=9 for haplotype clusters at 20,000 unlinked loci. Haplotype clusters identified from BEAGLE were plotted using DISTRUCT; the patterns are largely concordant across K's. At K=6 and 7 AAC's enriched in the Mbuti and NS populations appear, respectively, which is in contrast to the genotype-based results wherein for the AAC's at K=6 and 7 the NS and Mbuti populations, respectively, are enriched.

Figure S5. Data likelihood of STRUCTURE at K=2 to K=14 across replicate runs. A)

STRUCTURE analysis of genotypes at 20,000 unlinked loci shows data likelihood increasing as a function of K; however, the variance of the data likelihood increases dramatically at K=10 B) STRUCTURE analysis of haplotype clusters at 20,000 unlinked loci showed fit increasing as a function of K; as with analysis of genotypes, the variance of the data likelihood increases dramatically at K=10.

Figure S6. Linkage disequilibrium (average r2) decay in populations. The decay of average r2 (y-axis) over genetic distance, measured in kb (x-axis) is displayed.

Figure S7. Inference of demographic history based on linkage disequilibrium and haplotype sharing. A) Effective population size was estimated from LD decay; EHG populations are indicated with asterisks. **B**) The relationship between mean cumulative ROH and IBD.

Figure S8. EHG Demographic Model. The model employed for EHG demographic history in the context of agriculturalist and pastoralist expansion includes the following parameters: effective population sizes, population splits, and migration.

Figure S9. PCA of population samples used in neutrality testing. PC1 is plotted along the X- axis, and PC2 is plotted along the Y-axis.

Figure S10. Distribution of iHS candidate loci across population groupings. The number of candidate loci identified in the top 0.1% iHS results that are present in a given number of population groupings are displayed on the Y-axis. The X-axis displays each population grouping number category.

Figure S11. Distribution of D candidate loci across population groupings. The number of candidate loci identified in the top 0.1% D results that are present in a given number of population groupings are displayed on the Y-axis. The X-axis displays each population grouping number category.

Figure S12. Distribution of XP-CLR candidate loci across population groupings. The number of candidate loci identified in the top 0.1% XP-CLR results that are present in a given number of population groupings are displayed on the Y-axis. The X-axis displays each population grouping number category.

Figure S13. PCA of AAC (K=9) inferred allele frequencies. The AAC's are projected onto axes obtained from the PCA; the AACs are denoted by which population is most

enriched for a given AAC. A) On the PC 1 axis, which explains 26.56% of the variance in AAC allele frequencies, we observe a distribution of population-specific AACs consistent with genotypic PC 1, wherein San and Mozabite populations are at opposite ends of the axis. In contrast, AAC PC axis 2, which explains 16.28% of the variance, is not so clearly correlated with genotypic PC 2. B) AAC PC 3, explaining 13.25% of the data variance, also shows a weak correspondence with genotypic PC 3; the AAC-specific to the WRHG is drawn out most on one end of the axis while the Hadza are on the other end, with the Sabue closest. C) AAC PC4, explaining 12.2% of the data variance. D) AAC PC5, explaining 9% of the variance.

Figure S14. Transformed (inverse Gaussian CDF) simulated and estimated Nanc (left) and NHG (right) parameter values; the 45 degree red line represents perfect correlation.

Figure S15. Transformed (inverse Gaussian CDF) simulated and estimated M1 (left) and M2 (right) parameter values; the 45 degree red line represents perfect correlation.

Figure S16. Transformed (inverse Gaussian CDF) simulated and estimated N1 (left) and N2 (right) parameter values.

Figure S17. Transformed (inverse Gaussian CDF) simulated and estimated divergence time (left) and time of migration (right) parameter values; the 45 degree red line represents perfect correlation.

Figure S18. Cross-validation error for divergence time shown at varying tolerance levels (yellow to red is lowest to highest i.e. 1st percentile to 25th percentile; the black line represents perfect correlation).

Figure S19. Example posterior distribution inferred for Nanc where the true, simulated parameter value is within the 95% credible interval. The dotted line reflects the prior; the black line is the unadjusted, standard ABC posterior distribution and the red line is the regression adjusted ABC posterior.

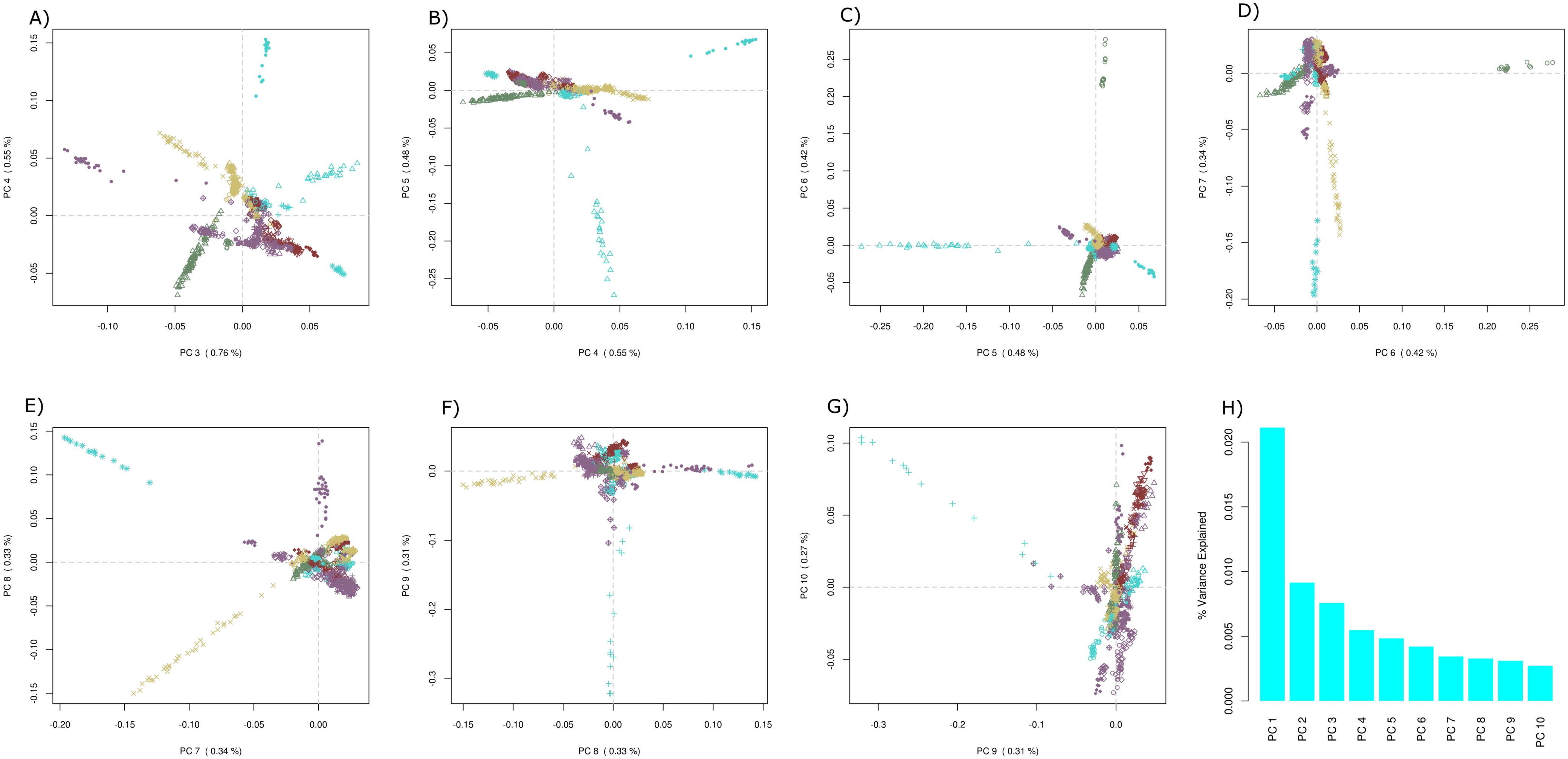
A)
PC 3 (0.76 %)
PC 4 (0.55 %)
-0.10
-0.05
0.00
0.05
0.15
0.10
0.05
0.00
-0.05
B)
PC 4 (0.55 %)
PC 5 (0.48 %)
-0.05
0.00
0.05
0.10
0.15
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
C)
PC 5 (0.48 %)
PC 6 (0.42 %)
-0.25
-0.20
-0.15
-0.10
-0.05
0.00
0.05
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
D)
PC 6 (0.42 %)
PC 7 (0.34 %)
-0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.00
-0.05
-0.10
-0.15
-0.20
E)
PC 7 (0.34 %)
PC 8 (0.33 %)
-0.20
-0.15
-0.10
-0.05
0.00
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
F)
PC 8 (0.33 %)
PC 9 (0.31 %)
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15
0.0
-0.1
-0.2
-0.3
G)
PC 9 (0.31 %)
PC 10 (0.27 %)
-0.3
-0.2
-0.1
0.0
0.10
0.05
0.00
-0.05
H)
% Variance Explained
0.020
0.015
0.010
0.005
0.000
PC 1
PC 2
PC 3
PC 4
PC 5
PC 6
PC 7
PC 8
PC 9
PC 10

Fig. S1

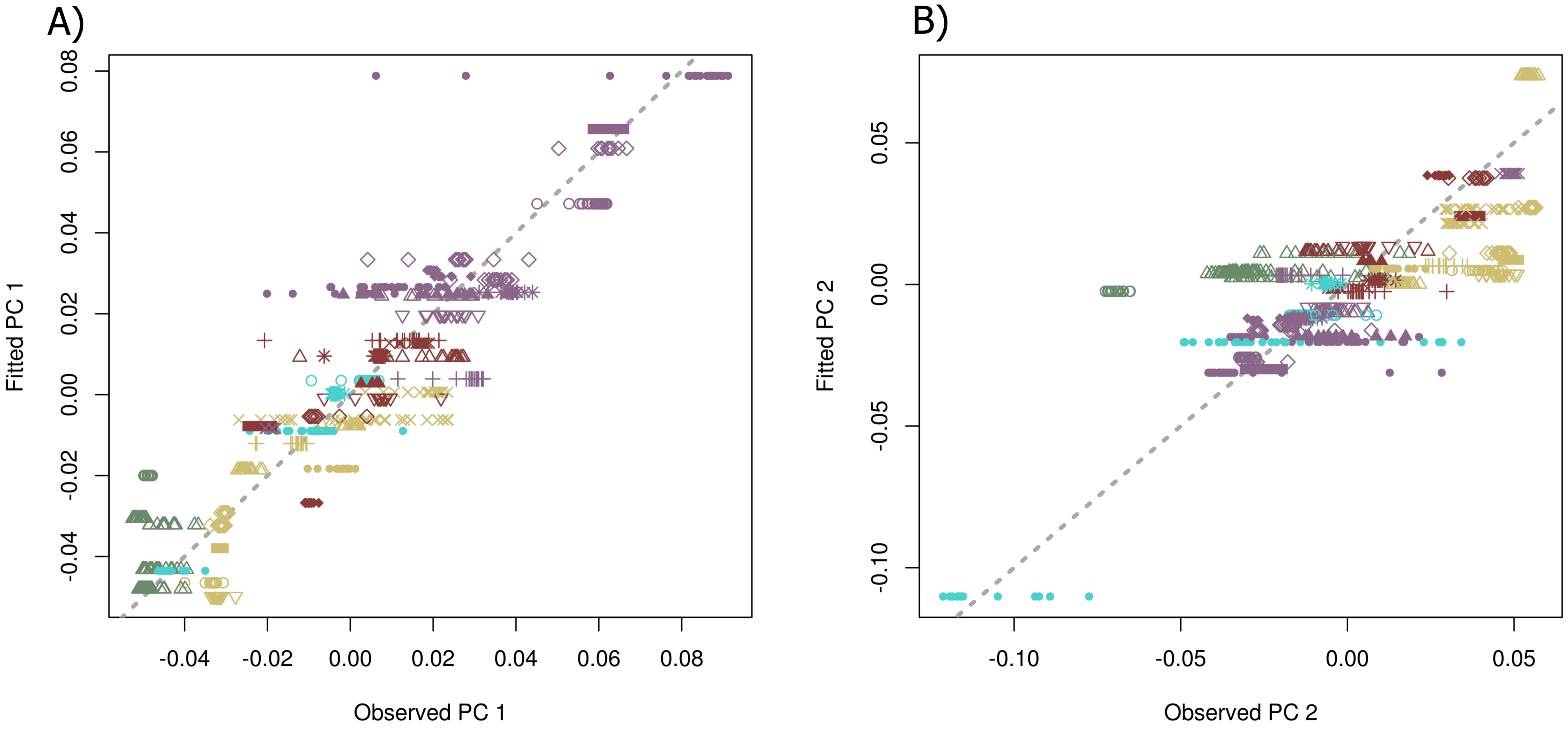
A)
Fitted PC 1
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.04
-0.02
0.00
0.02
0.04
0.06
0.08
Observed PC 1
B)
Fitted PC 2
0.05
0.00
-0.05
-0.10
-0.10
-0.05
0.00
0.05
Observed PC 2

Fig. S2

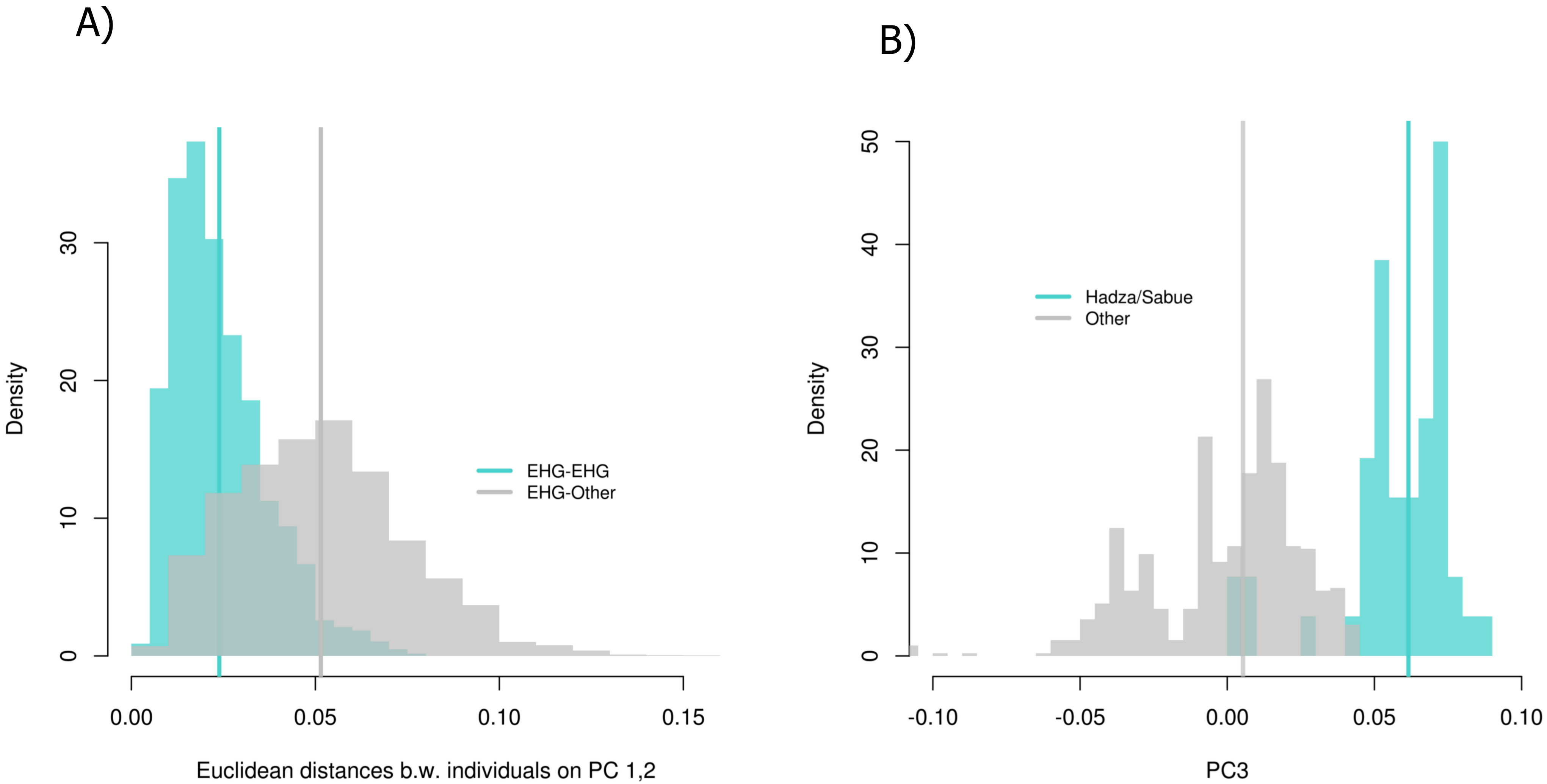
A)
Density
0
10
20
30
0.00
0.05
0.10
0.15
Euclidean distances b.w. individuals on PC 1,2
EHG-EHG
EHG-Other
B)
Density
0
10
20
30
40
50
-0.10
-0.05
0.00
0.05
0.10
PC3
Hadza/Sabue
Other

Fig. S3

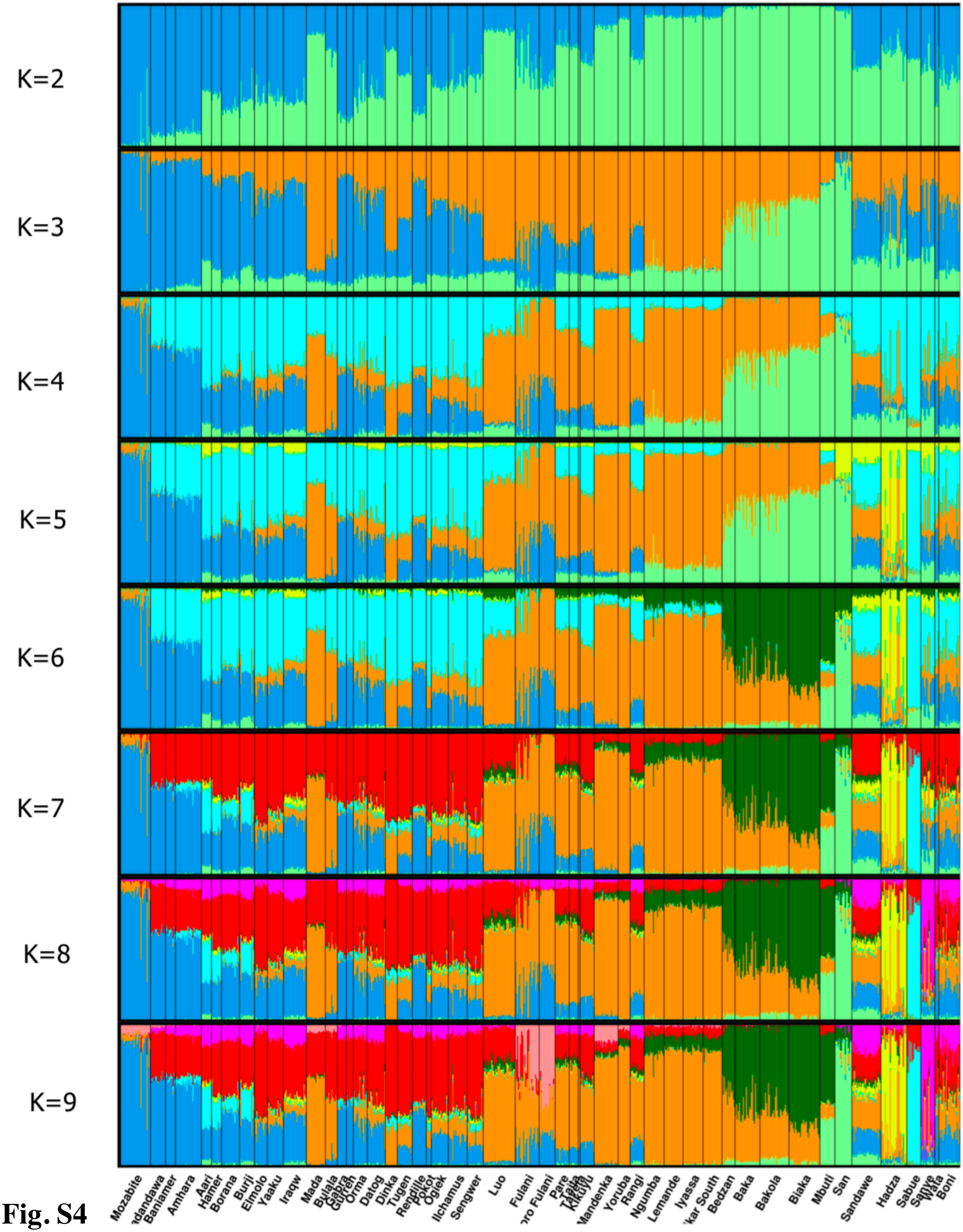
K=2
K=3
K=4
K=5
K=6
K=7
K=8
K=9
Mozabite
adandawa
Baniamer
Amhara
Aari
Hamer
Borana
Burji
Elmolo
Yaaku
Iraqw
Mada
Bulala
Gabra
Gurreh
Orma
Datog
Dinka
Tugen
Rendille
Pokot
Ogiek
Ilchamus
Sengwer
Luo
Fulani
oro Fulani
Pare
Taita
Tagota
Kikuyu
Mandenka
Yoruba
Rangi
Ngumba
Lemande
Iyassa
kar South
Bedzan
Baka
Bakola
Biaka
Mbuti
San
Sandawe
Hadza
Sabue
Sanye
Wata
Boni

Fig. S4

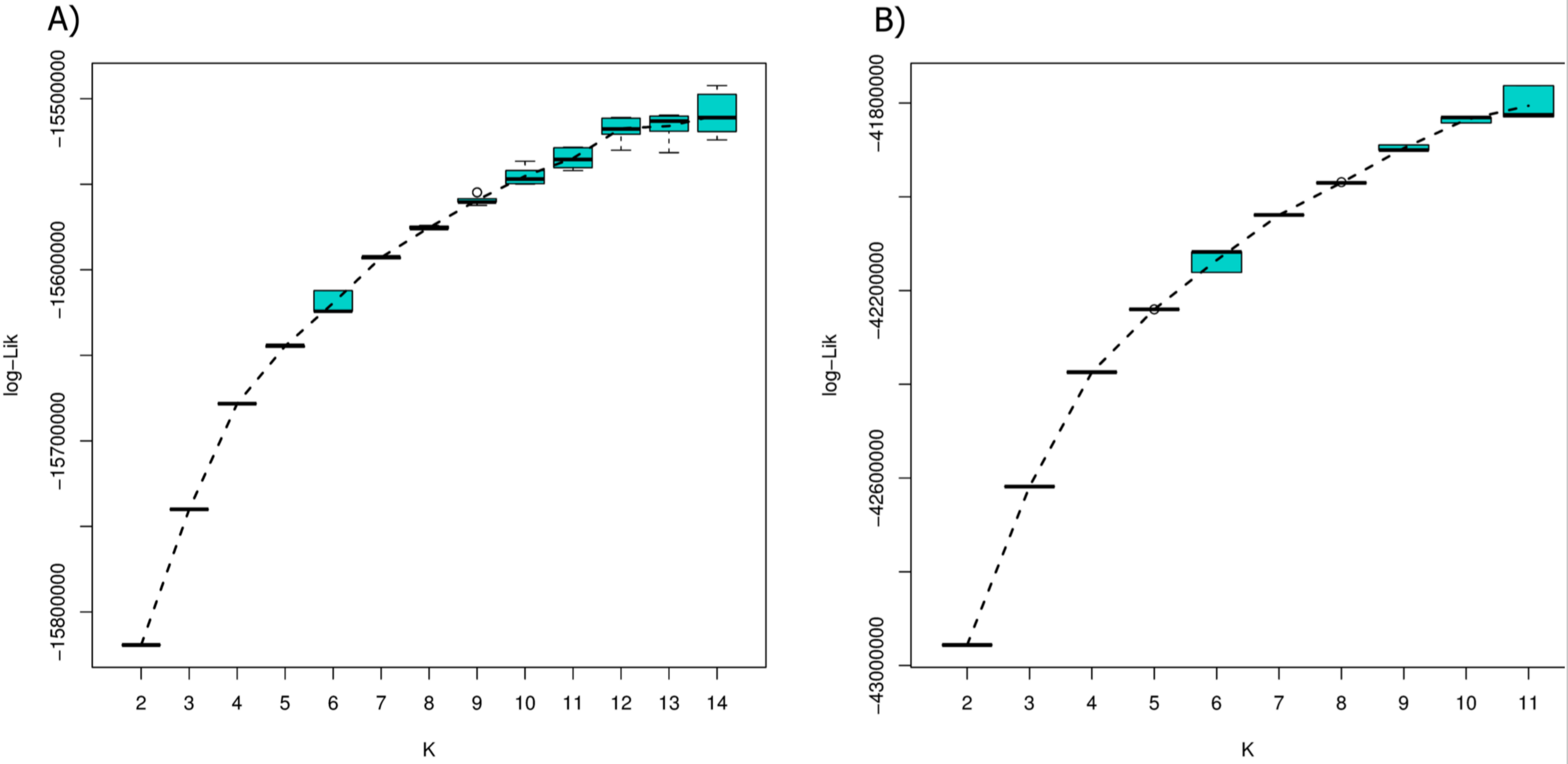
A)
−15500000
−15600000
−15700000
−15800000
log−Lik
2
3
4
5
6
7
8
9
10
11
12
13
14
K
B)
−41800000
−42200000
−42600000
−43000000
log−Lik
2
3
4
5
6
7
8
9
10
11
K

Fig. S5

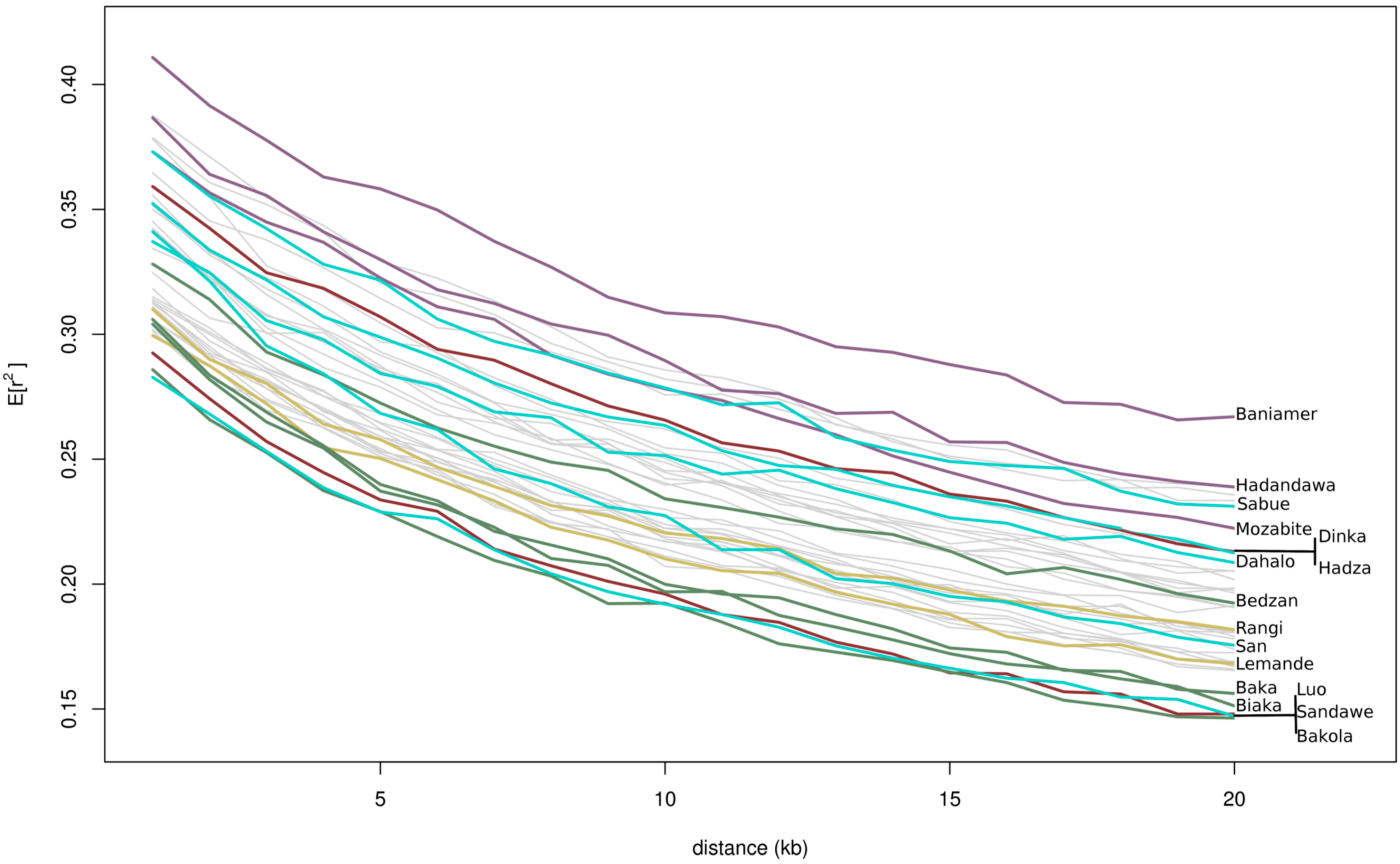
E[r^2]
0.15
0.20
0.25
0.30
0.35
0.40
5
10
15
20
distance (kb)
Baniamer
Hadandawa
Sabue
Mozabite
Dinka
Dahalo
Hadza
Bedzan
Rangi
San
Lemande
Baka
Luo
Biaka
Sandawe
Bakola

Fig. S6

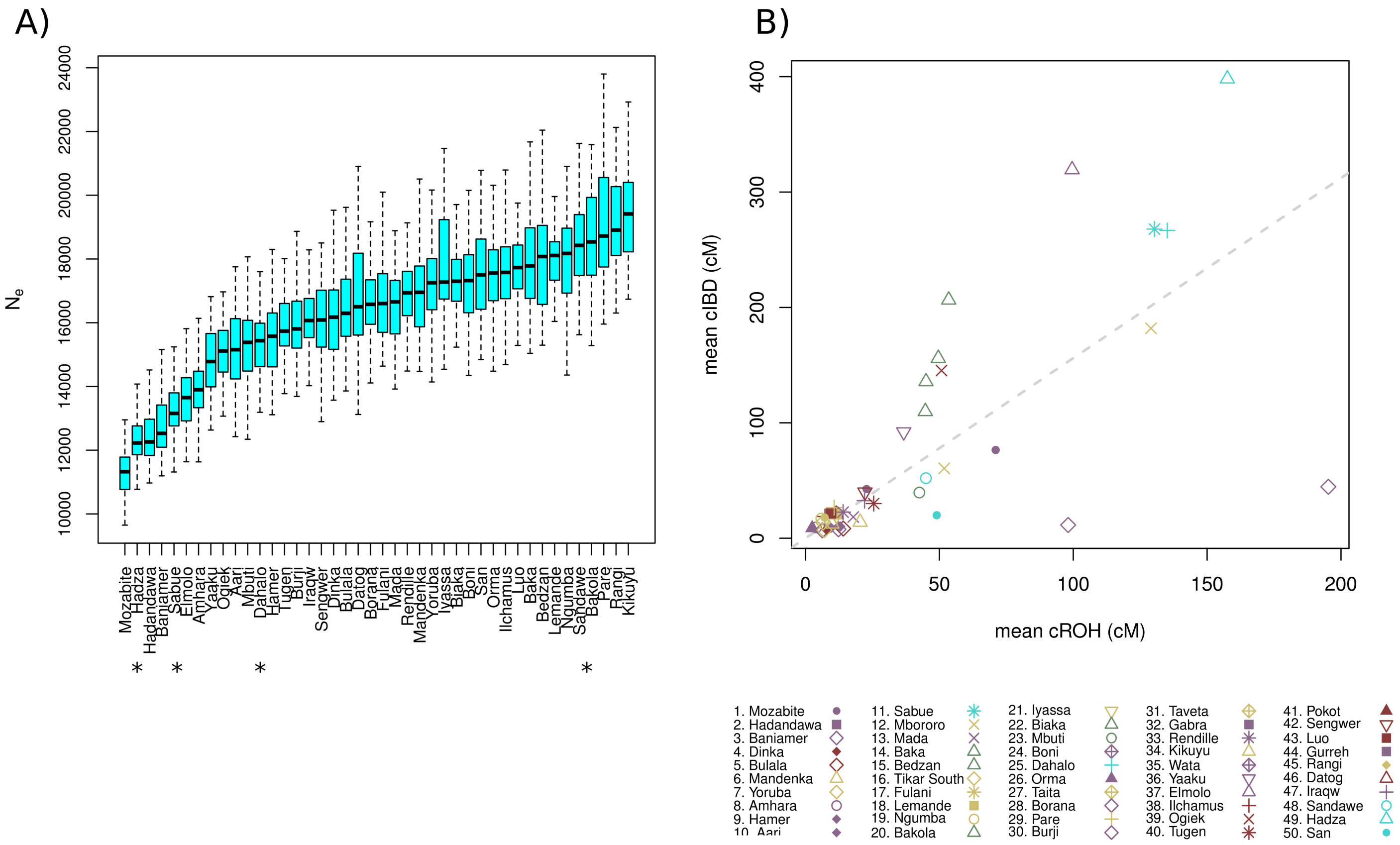
A)
B)
N_e
10000
12000
14000
16000
18000
20000
22000
24000
Mozabite
Hadza
Hadandawa
Baniamer
Sabue
Elmolo
Amhara
Yaaku
Ogiek
Aari
Mbuti
Dahalo
Hamer
Tugen
Burji
Iraqw
Sengwer
Dinka
Bulala
Datog
Borana
Fulani
Mada
Rendille
Mandenka
Yoruba
Iyassa
Biaka
Boni
San
Orma
Ilchamus
Luo
Baka
Bedzan
Lemande
Ngumba
Sandawe
Bakola
Pare
Rangi
Kikuyu
*
*
*
*
mean cIBD (cM)
0
100
200
300
400
mean cROH (cM)
0
50
100
150
200
1. Mozabite ●
2. Hadandawa ■
3. Baniamer ◇
4. Dinka ◆
5. Bulala ◇
6. Mandenka △
7. Yoruba ◇
8. Amhara ○
9. Hamer ◆
10. Aari ◆
11. Sabue ✳
12. Mbororo ×
13. Mada ×
14. Baka △
15. Bedzan △
16. Tikar South ◇
17. Fulani ✳
18. Lemande ■
19. Ngumba ○
20. Bakola △
21. Iyassa ▽
22. Biaka △
23. Mbuti ○
24. Boni ⊕
25. Dahalo +
26. Orma ▲
27. Taita ⊕
28. Borana ◇
29. Pare +
30. Burji ◇
31. Taveta ⊕
32. Gabra ■
33. Rendille ✳
34. Kikuyu △
35. Wata ⊕
36. Yaaku ▽
37. Elmolo △
38. Ilchamus +
39. Ogiek ×
40. Tugen ✳
41. Pokot ▲
42. Sengwer ▽
43. Luo ■
44. Gurreh ■
45. Rangi ◆
46. Datog △
47. Iraqw +
48. Sandawe ○
49. Hadza △
50. San ●

Fig. S7

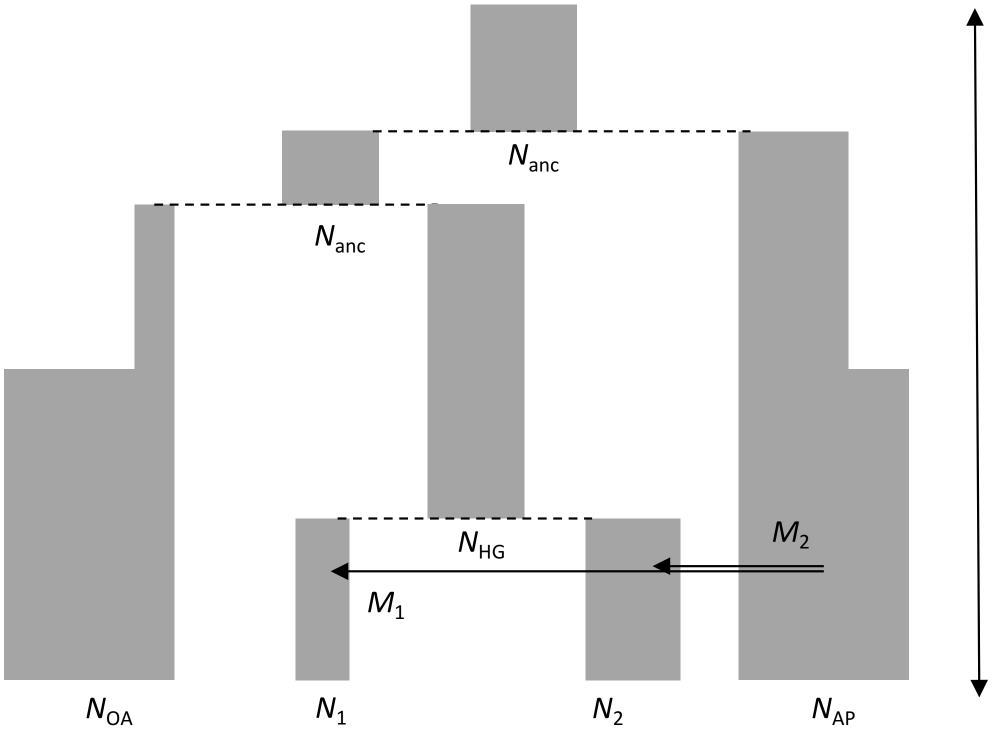

Fig. S8

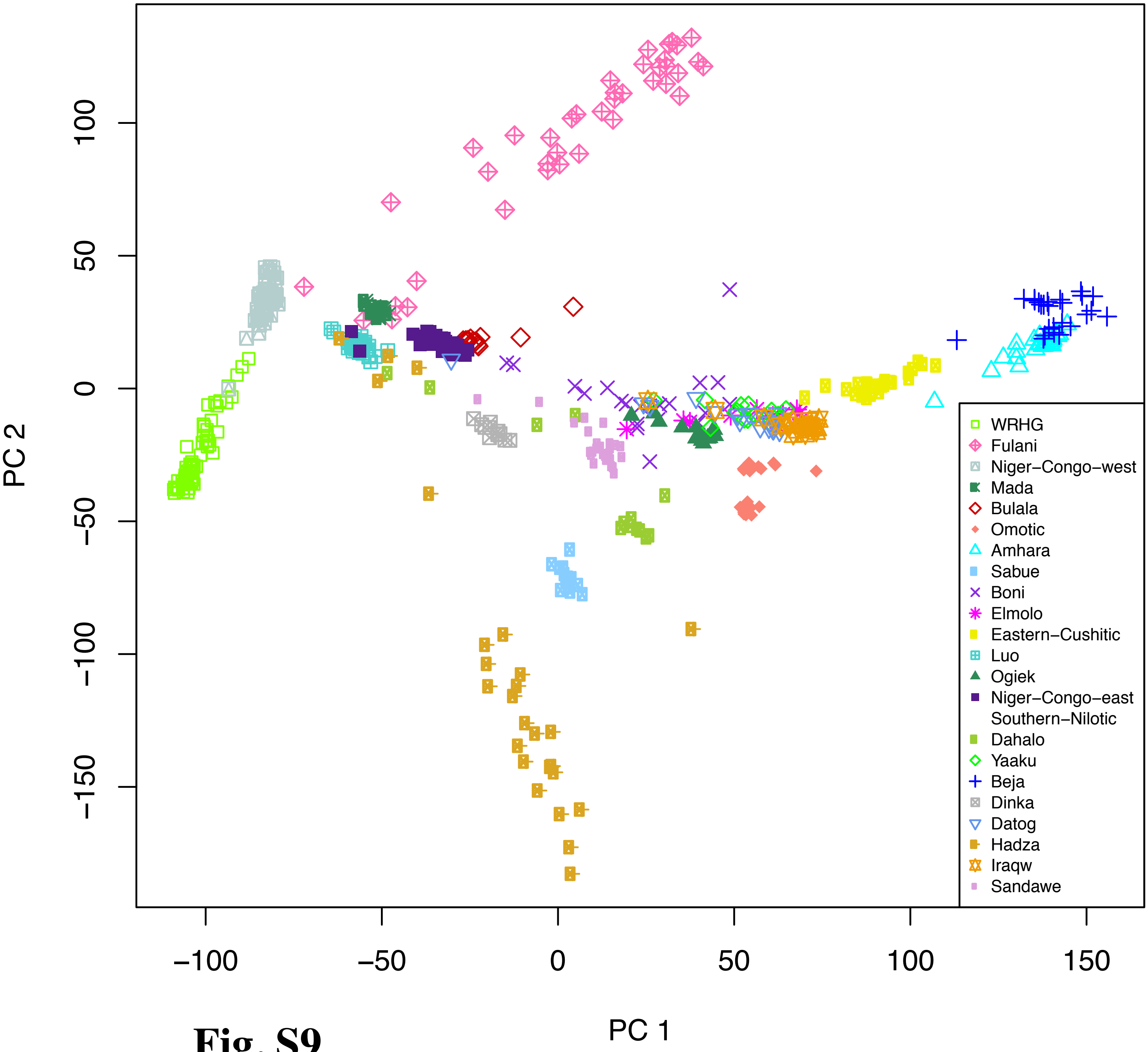
PC 1
PC 2
100
50
0
−50
−100
−150
−100
−50
0
50
100
150
WRHG
Fulani
Niger−Congo−west
Mada
Bulala
Omotic
Amhara
Sabue
Boni
Elmolo
Eastern−Cushitic
Luo
Ogiek
Niger−Congo−east
Southern−Nilotic
Dahalo
Yaaku
Beja
Dinka
Datog
Hadza
Iraqw
Sandawe

Fig. S9

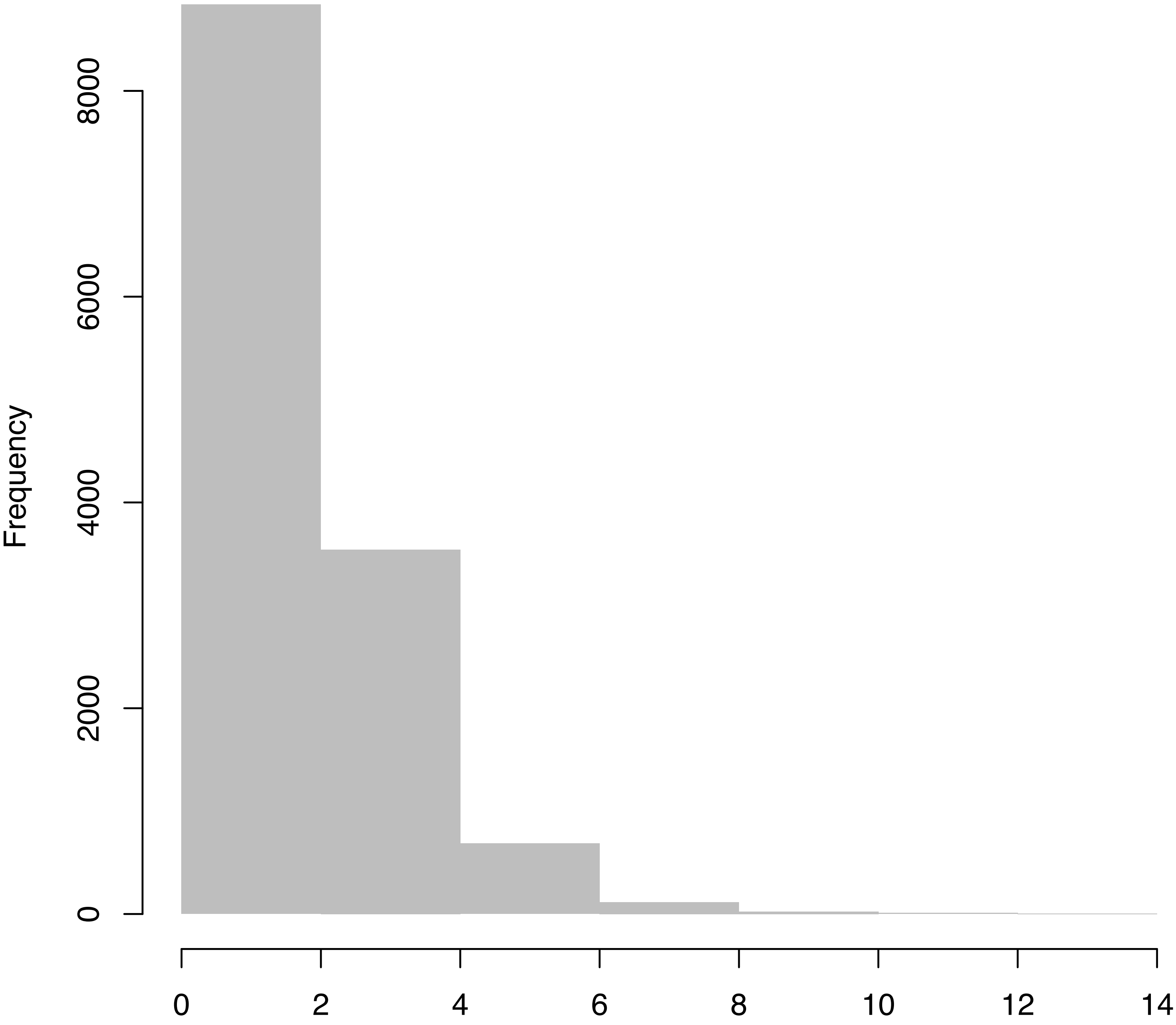

Fig. S10 # of pop groupings a top D candidate gene is identified in

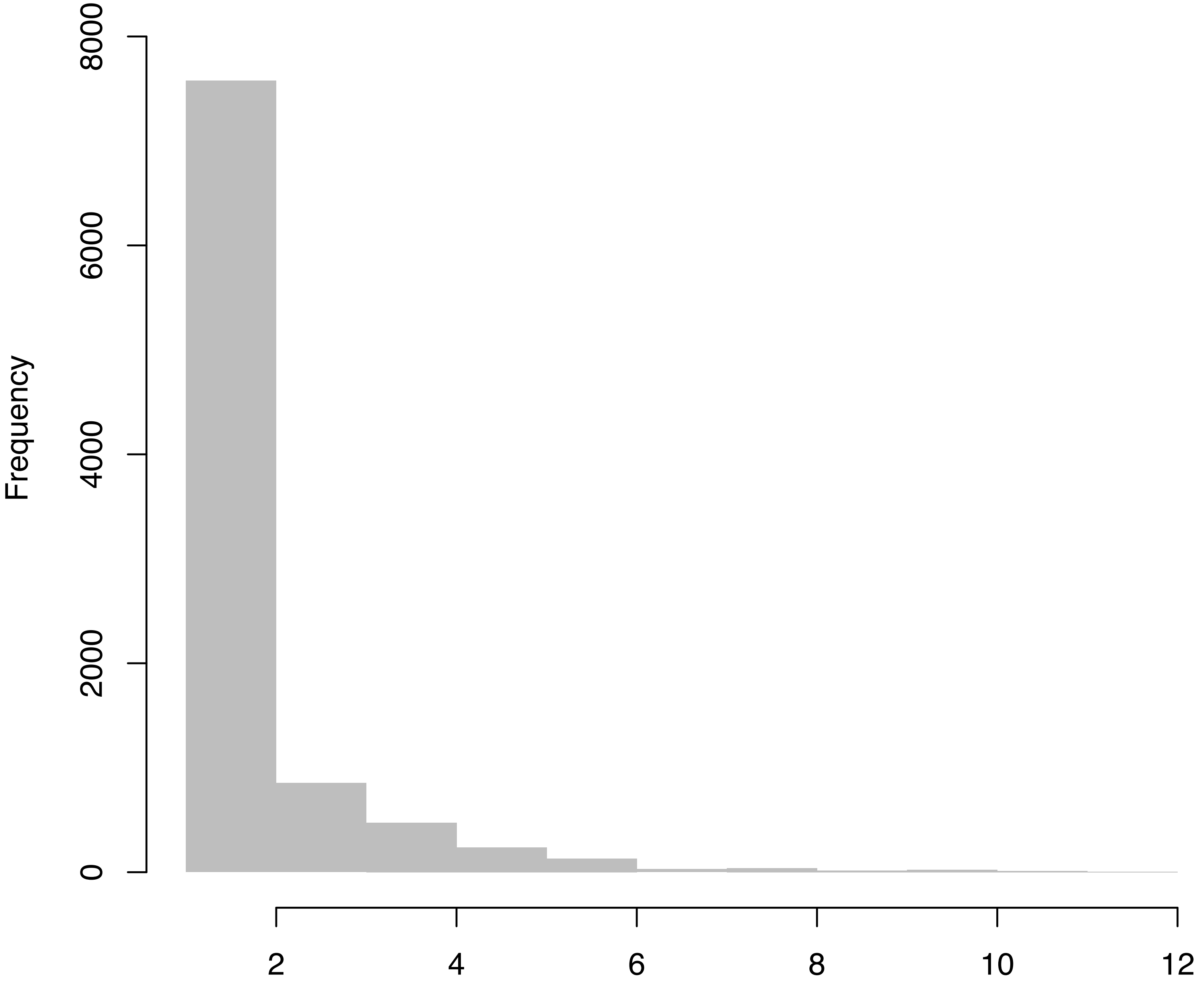
8000
6000
4000
2000
0
Frequency
2
4
6
8
10
12
# of pop groupings a top iHS candidate gene is identified in

Fig. S11

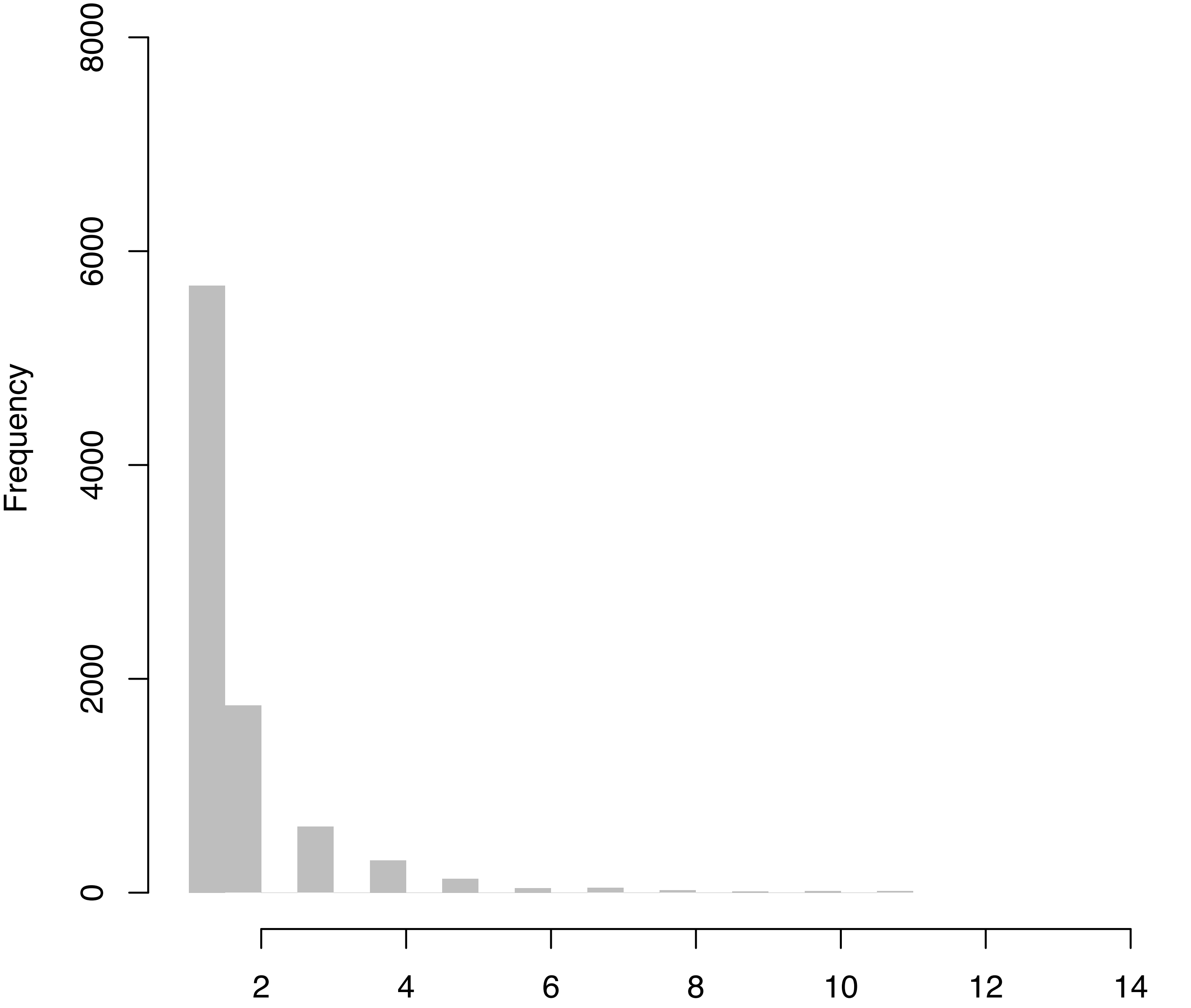
Frequency
0
2000
4000
6000
8000
2
4
6
8
10
12
14
# of pop groupings a top 0.1% XP−CLR candidate gene is identified in

Fig. S12

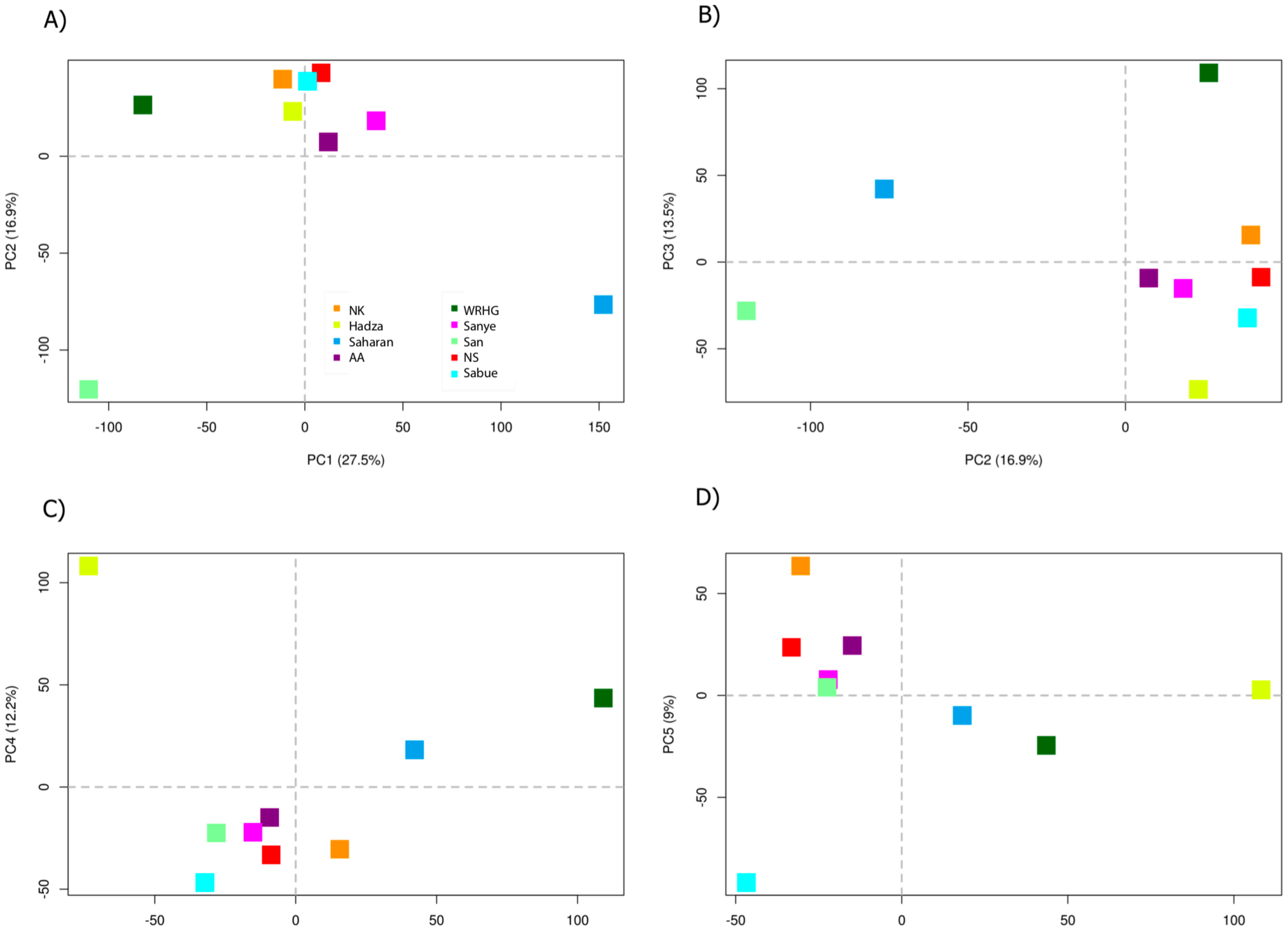
A)
B)
C)
D)
PC1 (27.5%)
PC2 (16.9%)
PC3 (13.5%)
PC4 (12.2%)
PC5 (9%)
NK
Hadza
Saharan
AA
WRHG
Sanye
San
NS
Sabue
-100
-50
0
50
100
150

Fig. S13

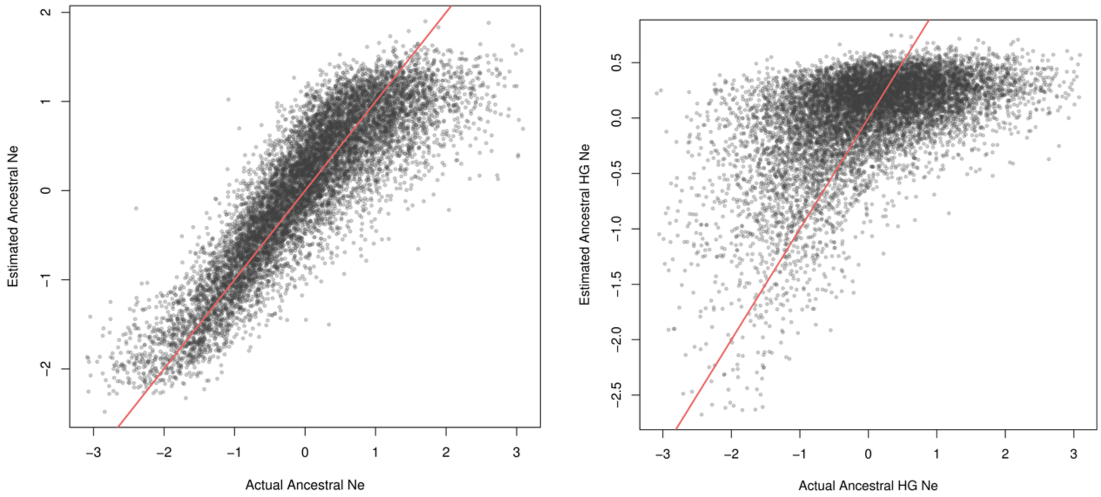

Fig. S14 Transformed (inverse Gaussian CDF) simulated and estimated N_{anc} (left) and N_{HG} (right) parameter values; the 45 degree red line represents perfect correlation

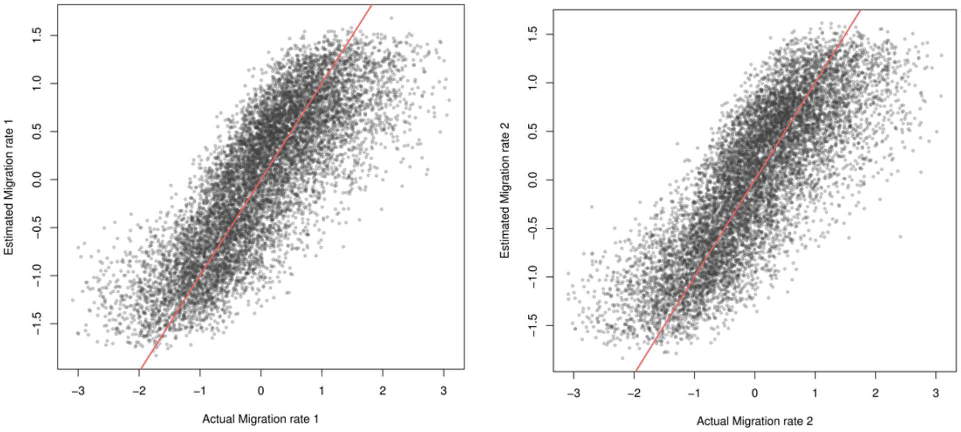

Fig. S15 Transformed (inverse Gaussian CDF) simulated and estimated M_1 (left) and M_2 (right) parameter values; the 45 degree red line represents perfect correlation

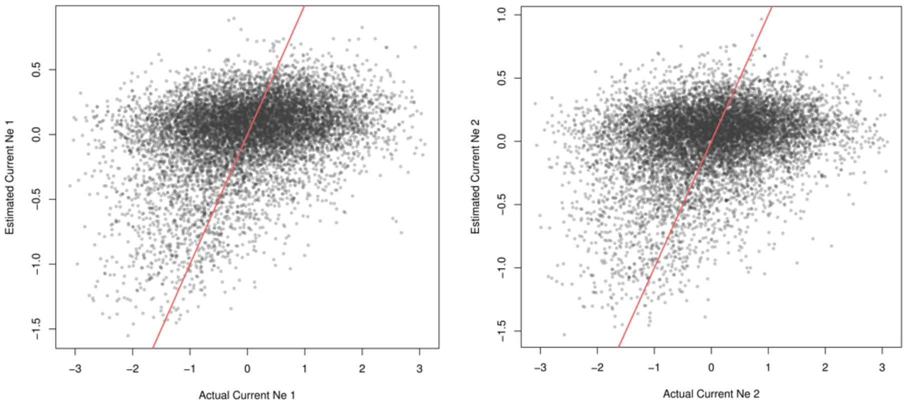

Fig. S16 Transformed (inverse Gaussian CDF) simulated and estimated N_1 (left) and N_2 (right) parameter values

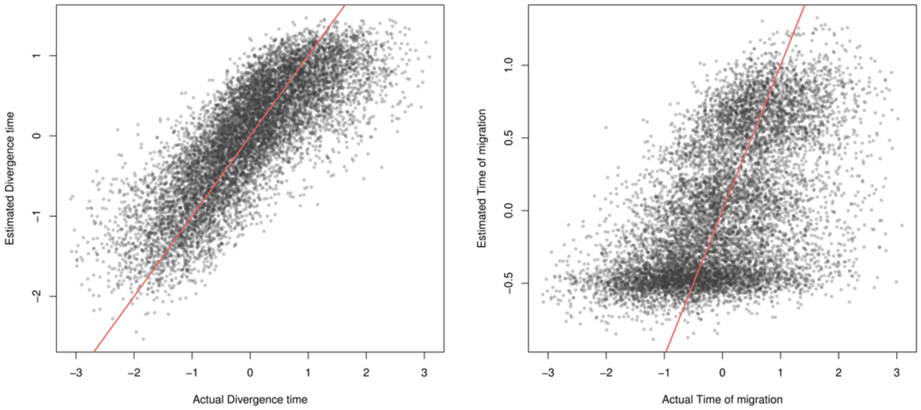

Fig. S17 Transformed (inverse Gaussian CDF) simulated and estimated divergence time (left) and time of migration (right) parameter values; the 45 degree red line represents perfect correlation

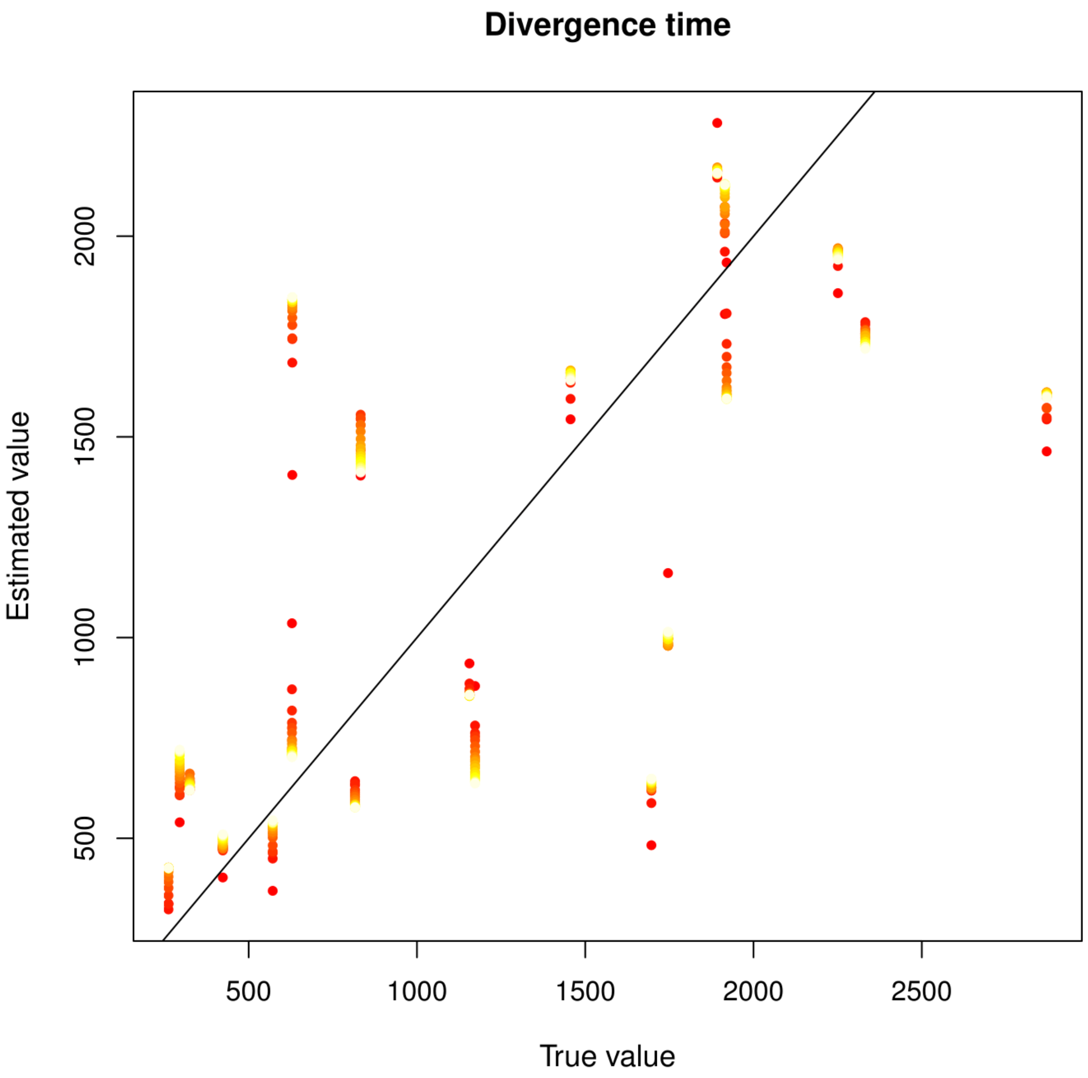

Fig. S18 Cross-validation error for divergence time shown at varying tolerance levels (yellow to red is lowest to highest *i.e.* 1st percentile to 25th percentile; the black line represents perfect correlation).

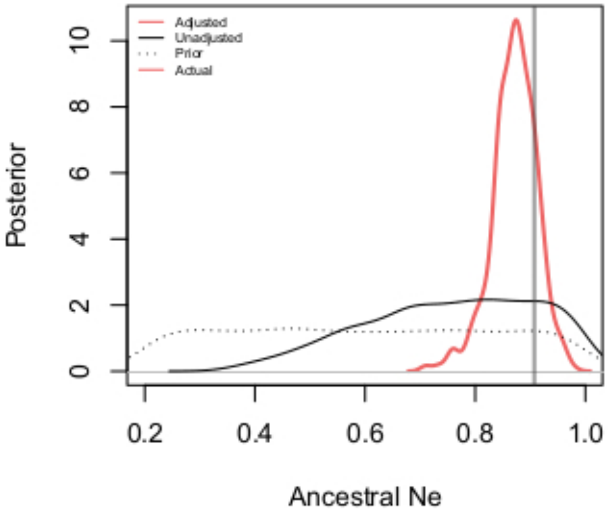

Fig. S19 An example posterior distribution inferred for N_{anc} where the true, simulated parameter value is within the 95% credible interval. The dotted line reflects the prior; the black line is the unadjusted, standard ABC posterior distribution and the red line is the regression adjusted ABC posterior.

Table S1. **The self-identified ethnicity, linguistic affiliation, country, and sample size for each population analyzed in the study.** Populations samples obtained from the publicly available resources are indicated with an asterisk. Populations referred to as East African Hunting-Gathering in the text (EHG) are highlighted in bold.

Ethnicity	**Country**	**Language**	**Historical Subsistence**	**Sample size**
Datog	Tanzania	Nilo-Saharan	Pastoralist	18
Dinka	Sudan	Nilo-Saharan	Pastoralist	12
Ilchamus	Kenya	Nilo-Saharan	Pastoralist	19
Tugen	Kenya	Nilo-Saharan	Pastoralist	15
Luo	Kenya	Nilo-Saharan	Pastoralist	32
Pokot	Kenya	Nilo-Saharan	Pastoralist	5
Ogiek	Kenya	Nilo-Saharan	Hunter-Gatherer	17
Sengwer	Kenya	Nilo-Saharan	Hunter-Gatherer	16
Sabue	Ethiopia	Nilo-Saharan	Hunter-Gatherer	14
Bulala	Chad	Nilo-Saharan	Agriculturalist	12
Fulani	Cameroon	Niger-Kordofanian	Pastoralist	40
Rangi	Tanzania	Niger-Kordofanian	Agriculturalist	14
Pare	Kenya	Niger-Kordofanian	Agriculturalist	14
Mbuti*	DRC	Niger-Kordofanian	Hunter-Gatherer	15
Biaka*	CAR	Niger-Kordofanian	Hunter-Gatherer	31
Bedzan	Cameroon	Niger-Kordofanian	Hunter-Gatherer	13
Baka	Cameroon	Niger-Kordofanian	Hunter-Gatherer	25
Bakola	Cameroon	Niger-Kordofanian	Hunter-Gatherer	29

Taveta	Kenya	Niger-Kordofanian	Agropastoralist	2
Taita	Kenya	Niger-Kordofanian	Agropastoralist	9
Iyassa	Cameroon	Niger-Kordofanian	Agriculturalist	20
Mandenka*	Senegal	Niger-Kordofanian	Agriculturalist	24
Yoruba	Nigeria	Niger-Kordofanian	Agriculturalist	12
Kikuyu	Kenya	Niger-Kordofanian	Agriculturalist	14
Lemande	Cameroon	Niger-Kordofanian	Agriculturalist	19
Ngumba	Cameroon	Niger-Kordofanian	Agriculturalist	20
Tikar South	Cameroon	Niger-Kordofanian	Agriculturalist	19
Hadza	Tanzania	Khoisan	Hunter-Gatherer	26
Sandawe	Tanzania	Khoisan	Hunter-Gatherer	29
San*	Namibia	Khoisan	Hunter-Gatherer	17
Hadandawa	Sudan	Afro-Asiatic	Pastoralist	15
Baniamer	Sudan	Afro-Asiatic	Pastoralist	10
Borana	Kenya	Afro-Asiatic	Pastoralist	18
Gabra	Kenya	Afro-Asiatic	Pastoralist	9
Rendille	Kenya	Afro-Asiatic	Pastoralist	14
Orma	Kenya	Afro-Asiatic	Pastoralist	14
Gurreh	Kenya	Afro-Asiatic	Pastoralist	7
Aari	Ethiopia	Afro-Asiatic	Pastoralist	10
Hamer	Ethiopia	Afro-Asiatic	Pastoralist	10

Iraqw	Tanzania	Afro-Asiatic	Agriculturalist	23
Burji	Kenya	Afro-Asiatic	Agriculturalist	15
Elmolo	Kenya	Afro-Asiatic	Hunter-Gatherer	13
Yaaku	Kenya	Afro-Asiatic	Hunter-Gatherer	16
Boni	Kenya	Afro-Asiatic	Hunter-Gatherer	21
Wata	Kenya	Afro-Asiatic	Hunter-Gatherer	4
Dahalo	Kenya	Afro-Asiatic	Hunter-Gatherer	14
Amhara	Ethiopia	Afro-Asiatic	Agriculturalist	26
Mada	Cameroon	Afro-Asiatic	Agriculturalist	19
Mozabite*	Algeria	Afro-Asiatic	Agriculturalist	30

Table S2. **Population grouping affiliations employed in neutrality tests.**

ethnolinguistic affiliation	N	country of residence	language family	population grouping	historical subsistence
Aari	10	Ethiopia	Afro-Asiatic	Omotic	agro-pastoral
Amhara	26	Ethiopia	Afro-Asiatic	Amhara	agricultural
Baka	25	Cameroon	Niger-Kordofanian	WRHG	hunting-gathering
Bakola	29	Cameroon	Niger-Kordofanian	WRHG	hunting-gathering
Baniamer	10	Sudan	Afro-Asiatic	Beja	pastoral
Bedzan	13	Cameroon	Niger-Kordofanian	WRHG	hunting-gathering
Boni	21	Kenya	Afro-Asiatic	Boni	hunting-gathering
Bulala	12	Chad	Nilo-Saharan	Bulala	agro-pastoral
Dahalo	13	Kenya	Afro-Asiatic	Dahalo	hunting-gathering
Datog	18	Tanzania	Nilo-Saharan	Datog	agro-pastoral
Dinka	12	Sudan	Nilo-Saharan	Dinka	pastoral
Elmolo	13	Kenya	Afro-Asiatic	Elmolo	hunting-gathering
Fulani	16	Cameroon	Niger-Kordofanian	Fulani	pastoral
Fulani	8	Nigeria	Niger-Kordofanian	Fulani	pastoral
Gabra	9	Kenya	Afro-Asiatic	Eastern-Cushitic	pastoral
Gurreh	7	Kenya	Afro-Asiatic	Eastern-Cushitic	pastoral
Hadandawa	15	Sudan	Afro-Asiatic	Beja	pastoral
Hadzabe	26	Tanzania	Khoisan	Hadza	hunting-gathering
Hamer	10	Ethiopia	Afro-Asiatic	Omotic	pastoral
Iraqw	23	Tanzania	Afro-Asiatic	Iraqw	agro-pastoral
Lemande	19	Cameroon	Niger-Kordofanian	Niger-Congo-west	agricultural
Luo	32	Kenya	Nilo-Saharan	Luo	pastoral
Mada	19	Cameroon	Afro-Asiatic	Mada	agricultural
Mbororo Fulani	16	Cameroon	Niger-Kordofanian	Fulani	pastoral
Ngumba	20	Cameroon	Niger-Kordofanian	Niger-Congo-west	agricultural
Ogiek	17	Kenya	Nilo-Saharan	Ogiek	hunting-gathering
Pare	14	Kenya	Niger-Kordofanian	Niger-Congo-east	agro-pastoral
Pokot	5	Kenya	Nilo-Saharan	Southern-Nilotic	pastoral
Rendille	14	Kenya	Afro-Asiatic	Eastern-Cushitic	pastoral
Sabue	14	Ethiopia	Nilo-Saharan	Sabue	hunting-gathering
Sandawe	29	Tanzania	Khoisan	Sandawe	hunting-gathering
Sengwer	16	Kenya	Nilo-Saharan	Southern-Nilotic	agricultural
Taita	9	Kenya	Niger-Kordofanian	Niger-Congo-east	agricultural
Taveta	2	Kenya	Niger-Kordofanian	Niger-Congo-east	agro-pastoral
Tikar South	19	Cameroon	Niger-Kordofanian	Niger-Congo-west	agricultural

Yaaku	16	Kenya	Afro-Asiatic	Yaaku	hunting-gathering
Yoruba	12	Nigeria	Niger-Kordofanian	Niger-Congo-west	agricultural

Table S3. **Pathway enrichment results.** Population grouping, test statistic, pathway, and pathway enrichment results are listed in a supplementary data file.

Statistic	Population Grouping	Pathway ID	# Observed	# Expected	P Value	Corr-ected P Value	Pathway Name
XP-CLR	Eastern-Cushitic	P05734	9	2.36	0.00	0.04	Synaptic vesicle trafficking
XP-CLR	Amhara	P00059	21	8.48	0.00	0.02	p53 pathway
XP-CLR	Amhara	P00043	12	3.79	0.00	0.05	Muscarinic acetylcholine receptor 2 and 4 signaling pathway
XP-CLR	Elmolo	P00031	20	7.80	0.00	0.02	Inflammation mediated by chemokine and cytokine signaling pathway
XP-CLR	Fulani	P00057	57	32.76	0.00	0.03	Wnt signaling pathway
XP-CLR	Mada	P00011	11	3.15	0.00	0.02	Blood coagulation
XP-CLR	Omotic	P00048	13	4.25	0.00	0.04	PI3 kinase pathway
XP-CLR	Sabue	P02766	2	0.09	0.00	0.04	Phenylethylamine degradation
XP-CLR	Sandawe	P00040	9	2.35	0.00	0.04	Metabotropic glutamate receptor group II pathway
iHS	Eastern-Cushitic	P00019	22	8.96	0.00	0.02	Endothelin signaling pathway
iHS	Beja	P02771	4	0.49	0.00	0.03	Pyrimidine Metabolism
iHS	Dinka	P04373	11	3.29	0.00	0.04	5HT1 type receptor mediated signaling pathway
iHS	Fulani	P00059	22	9.21	0.00	0.03	p53 pathway
iHS	Hadza	P00057	63	36.72	0.00	0.02	Wnt signaling pathway
iHS	Iraqw	P04395	10	2.51	0.00	0.03	Vasopressin synthesis
iHS	Mada	P04373	17	6.26	0.00	0.04	5HT1 type receptor mediated signaling pathway
iHS	Niger-Congo-east	P00013	8	1.68	0.00	0.02	Cell cycle
iHS	Sabue	P00057	28	12.59	0.00	0.02	Wnt signaling pathway
iHS	Yaaku	P00044	11	3.20	0.00	0.02	Nicotinic acetylcholine receptor signaling pathway
D	Amhara	P00049	21	8.23	0.00	0.02	Parkinson disease
D	Boni	P00057	76	46.46	0.00	0.03	Wnt signaling pathway
D	Boni	P00001	12	3.59	0.00	0.05	Adrenaline and noradrenaline biosynthesis
D	Eastern-Cushitic	P04378	29	12.22	0.00	0.02	Beta2 adrenergic receptor signaling pathway
D	Luo	P00004	54	29.98	0.00	0.03	Alzheimer disease-presenilin pathway
D	Ogiek	P00037	19	7.10	0.00	0.02	Ionotropic glutamate receptor pathway
D	Ogiek	P00011	15	5.19	0.00	0.04	Blood coagulation
D	Niger-Congo-east	P00060	15	4.77	0.00	0.02	Ubiquitin proteasome pathway
D	Southern-Nilotic	P00014	9	2.14	0.00	0.03	Cholesterol biosynthesis
D	Sabue	P00012	32	15.49	0.00	0.03	Cadherin signaling pathway
D	Sandawe	P04372	10	2.54	0.00	0.03	5-Hydroxytryptamine degredation
D	Dahalo	P00037	23	9.43	0.00	0.02	Ionotropic glutamate receptor pathway

Table S4. MAP estimates of ancestral Ne (N_{anc}) as well as upper and lower credible intervals in units of N_{OA} = 100,000.

Population Combination	**MAP**	**5th %ile**	**95th %ile**
Hadza:Sandawe:Iraqw	1.486	1.003	2.233
Hadza:Sandawe:Dinka	1.900	1.233	3.101
Hadza:Dahalo:Iraqw	1.634	1.073	2.767
Hadza:Dahalo:Dinka	1.950	1.355	2.969
Hadza:Sabue:Iraqw	1.445	1.008	2.182
Hadza:Sabue:Dinka	1.789	1.203	2.806
Sandawe:Dahalo:Iraqw	1.662	1.118	2.687
Sandawe:Dahalo:Dinka	1.994	1.414	2.978
Sandawe:Sabue:Iraqw	1.577	1.140	2.192
Sandawe:Sabue:Dinka	1.891	1.323	2.897
Dahalo:Sabue:Iraqw	1.513	1.081	2.246
Dahalo:Sabue:Dinka	1.903	1.363	2.795

Table S5. MAP estimates of ancestral HG Ne (N_{HG}) as well as upper and lower credible intervals in units of N_{OA} = 100,000.

Population Combination	**MAP**	**5th %ile**	**95th %ile**
Hadza:Sandawe:Iraqw	0.449	0.194	1.020
Hadza:Sandawe:Dinka	0.226	0.186	0.343
Hadza:Dahalo:Iraqw	0.379	0.195	0.866
Hadza:Dahalo:Dinka	0.213	0.129	0.562
Hadza:Sabue:Iraqw	0.845	0.365	0.964
Hadza:Sabue:Dinka	0.215	0.145	0.454
Sandawe:Dahalo:Iraqw	0.414	0.230	0.824
Sandawe:Dahalo:Dinka	0.216	0.136	0.515
Sandawe:Sabue:Iraqw	0.856	0.351	0.993
Sandawe:Sabue:Dinka	0.220	0.171	0.356
Dahalo:Sabue:Iraqw	0.507	0.255	0.976
Dahalo:Sabue:Dinka	0.239	0.133	0.684

Table S6. MAP estimates of N_1 (left-most population) as well as upper and lower credible intervals in units of N_{OA} = 100,000.

Population Combinations	**MAP**	**5th %ile**	**95th %ile**
Hadza:Sandawe:Iraqw	0.285	0.094	0.396
Hadza:Sandawe:Dinka	0.223	0.103	0.340
Hadza:Dahalo:Iraqw	0.158	0.106	0.339
Hadza:Dahalo:Dinka	0.128	0.090	0.273
Hadza:Sabue:Iraqw	0.146	0.090	0.271
Hadza:Sabue:Dinka	0.129	0.088	0.284
Sandawe:Dahalo:Iraqw	0.171	0.116	0.350
Sandawe:Dahalo:Dinka	0.130	0.091	0.280
Sandawe:Sabue:Iraqw	0.197	0.108	0.401
Sandawe:Sabue:Dinka	0.149	0.094	0.326
Dahalo:Sabue:Iraqw	0.134	0.093	0.223
Dahalo:Sabue:Dinka	0.133	0.095	0.278

Table S7. MAP estimates of N_2 (center population) as well as upper and lower credible intervals in units of N_{OA} = 100,000.

Population Combinations	MAP	5th %ile	95th %ile
Hadza:Sandawe:Iraqw	0.353	0.137	0.397
Hadza:Sandawe:Dinka	0.243	0.111	0.376
Hadza:Dahalo:Iraqw	0.127	0.091	0.244
Hadza:Dahalo:Dinka	0.150	0.096	0.283
Hadza:Sabue:Iraqw	0.145	0.100	0.352
Hadza:Sabue:Dinka	0.158	0.099	0.308
Sandawe:Dahalo:Iraqw	0.131	0.094	0.231
Sandawe:Dahalo:Dinka	0.152	0.097	0.291
Sandawe:Sabue:Iraqw	0.146	0.101	0.409
Sandawe:Sabue:Dinka	0.158	0.100	0.356
Dahalo:Sabue:Iraqw	0.159	0.102	0.384
Dahalo:Sabue:Dinka	0.142	0.094	0.275

Table S8. MAP estimates of M_1 (gene flow from the right-most population to the left-most population) as well as upper and lower credible intervals in units of migrants per generation.

Population Combinations	MAP	5th %ile	95th %ile
Hadza:Sandawe:Iraqw	559.255	214.011	1263.617
Hadza:Sandawe:Dinka	186.140	60.167	1221.276
Hadza:Dahalo:Iraqw	1184.150	535.102	2585.064
Hadza:Dahalo:Dinka	113.004	31.775	1238.339
Hadza:Sabue:Iraqw	1246.123	603.577	2694.996
Hadza:Sabue:Dinka	110.173	28.138	1442.733
Sandawe:Dahalo:Iraqw	2518.148	1392.942	3396.298
Sandawe:Dahalo:Dinka	126.396	34.408	1380.123
Sandawe:Sabue:Iraqw	1671.266	988.408	2404.591
Sandawe:Sabue:Dinka	137.682	43.718	1248.757
Dahalo:Sabue:Iraqw	142.927	41.543	1091.403
Dahalo:Sabue:Dinka	105.415	27.037	1179.507

Table S9. MAP estimates of M_2 (gene flow from the right-most population to the center population) as well as upper and lower credible intervals in units of migrants per generation.

Population Combinations	**MAP**	**5th %ile**	**95th %ile**
Hadza:Sandawe:Iraqw	1650.526	920.863	2184.846
Hadza:Sandawe:Dinka	269.567	71.699	1336.820
Hadza:Dahalo:Iraqw	214.254	51.649	887.214
Hadza:Dahalo:Dinka	129.716	35.287	972.618
Hadza:Sabue:Iraqw	280.501	78.015	1651.883
Hadza:Sabue:Dinka	121.714	32.282	1006.969
Sandawe:Dahalo:Iraqw	198.423	48.476	750.554
Sandawe:Dahalo:Dinka	125.961	34.044	917.488
Sandawe:Sabue:Iraqw	175.427	67.942	825.138
Sandawe:Sabue:Dinka	147.405	32.784	1081.539
Dahalo:Sabue:Iraqw	402.398	101.953	1715.500
Dahalo:Sabue:Dinka	116.637	32.033	921.510

Table S10. MAP estimates of the time of migration between A/P (right-most) population into HG populations (left-most and center) began in units of generations.

Population Combinations	MAP	5th %ile	95th %ile
Hadza:Sandawe:Iraqw	166.547	125.876	278.085
Hadza:Sandawe:Dinka	114.723	88.900	253.770
Hadza:Dahalo:Iraqw	128.562	101.407	264.067
Hadza:Dahalo:Dinka	115.268	94.155	245.100
Hadza:Sabue:Iraqw	136.172	103.746	291.389
Hadza:Sabue:Dinka	121.038	99.337	252.103
Sandawe:Dahalo:Iraqw	162.808	126.564	279.674
Sandawe:Dahalo:Dinka	121.401	99.538	250.259
Sandawe:Sabue:Iraqw	173.288	121.168	292.491
Sandawe:Sabue:Dinka	121.679	97.220	271.026
Dahalo:Sabue:Iraqw	118.602	95.515	261.357
Dahalo:Sabue:Dinka	115.968	95.699	235.353

Dataset S1. Top 0.1% results for the D test. Population grouping, chromosome, chromosome position (B37), SNP identifier (rsid), and genes within 100 kb are listed in a supplementary data file.

Dataset S2. Top 0.1% results for the iHS test. Population grouping, iHS test statistic, chromosome, chromosome position (B37), SNP identifier (rsid), and genes within 100 kb are listed in a supplementary data file.

Dataset S3. Top 0.1% results for the XP-CLR test. Population grouping, XP-CLR test statistic, chromosome, chromosome position (B37) of the start and end point of each tested region, and genes within 100 kb are listed in a supplementary data file.