Supplementary Method:
Site Frequency Spectrum for Variable Population Size

In this note we determine the site frequency spectrum (SFS) for a variable population size model, and apply the method to investigate a specific demographic model for three subspecies of chimpanzees. The note falls in three parts: In Section 1 we develop the general theory for obtaining the SFS for a variable population size model. The main problem is to determine the mean time between coalescence events, and in Section 2 we describe a Monte Carlo procedure for obtaining these mean coalescence times. Finally in Section 3 we consider the demographic model and corresponding SFS for three chimpanzee subspecies.

1 Simulating waiting times between coalescence events

Consider a tree with n leaves and denote by ℓ the number of branches that are present in the tree at a given time point. We call ℓ $(2 \leq \ell \leq n)$ the level of the tree.

The probability $P_{i|\ell}$ that a mutation at level ℓ is present in i samples (the probability that a branch at level ℓ is ancestral to i leaves) is given by (see e.g. Durrett, 2008, equation (2.1) page 54)

$$P_{i|\ell} = \frac{\binom{n-i-1}{\ell-2}}{\binom{n-1}{\ell-1}}, \quad i = 1, \ldots, n-\ell. \tag{1}$$

Let ξ_i $(1 \leq i \leq n-1)$ be the number of sites with i derived alleles and $(n-i)$ ancestral alleles. Conditional on the times $(T_2, \ldots, T_n)$ between coalescence events we have

$$\mathrm{E}\big[\xi_i | (T_2, \ldots, T_n)\big] = \sum_{\ell=2}^{n-i+1} \ell \frac{\theta}{2} T_\ell P_{i|\ell},$$

where $P_{i|\ell}$ is defined above and θ is the scaled mutation rate. We get

$$\mathrm{E}[\xi_i] = \theta \sum_{\ell=2}^{n-i+1} \frac{\ell}{2} \frac{\binom{n-i-1}{\ell-2}}{\binom{n-1}{\ell-1}} \mathrm{E}\big[T_\ell\big],$$

so all we need in order to determine the site frequency count for a variable population size model is to determine $\mathrm{E}[T_\ell]$ $(2 \leq \ell \leq n)$.

Using a time change we can easily *simulate* the times between coalescence events in a variable population size model (e.g. Tavare, 2004, Section 2.4, Algorithm 2.1). We therefore use a Monte Carlo estimate to determine the mean between coalescent times.

2 Simulating waiting times between coalescence events

For convenience we re-state Algorithm 2.1 in Tavare (2004) below.

Algorithm: Simulating times between coalescence events.

Input: Sample size n and variable population size with integrated intensity function Λ.
Output: Sample of times $T_2, \ldots, T_n$ between coalescence events.
Step 1: Generate $t'_j = -2\log(U_j)/(j(j-1)),\ j = 2, \ldots, n$.
Step 2: Form $s'_j = t'_j + \ldots + t'_n\ j = 2, \ldots, n$.
Step 3: Compute $t_n = \Lambda^{-1}(s'_n)$, $t_j = \Lambda^{-1}(s'_j) - \Lambda^{-1}(s'_{j+1}),\ j = n-1, \ldots, 2$.
Step 4: Return $T_j = t_j,\ j = 2, \ldots, n$.

Here the integrated intensity function Λ is defined as follows. Let N_0 be the scaling factor such that the relative population size function in scaled time t (generation tN_0 ago) is given by

$$f(t) = N(tN_0)/N_0,\ t \geq 0.$$

We then define the intensity function

$$\lambda(u) = \frac{1}{f(u)},\ u \geq 0,$$

and the integrated intensity function

$$\Lambda(t) = \int_0^t \lambda(u)du,\ t \geq 0.$$

For the chimpanzee data we are particularly interested in variable population size models where the population size is constant for two or three epochs.

2.1 Two epochs of constant size

Consider the demographic scenario depicted in Figure 1. Present population size is N, and the population size stays at N until time aN in the past. From time aN and further back in time the population size is a fraction α of the present population size.

Choosing $N_0 = N$ we get

$$f(x) = \begin{cases} 1 & 0 \leq x < a \\ \alpha & x \geq a \end{cases}$$

and

$$\Lambda^{-1}(u) = \begin{cases} u & 0 \leq u < a \\ a + \alpha(u - a) & u \geq a. \end{cases}$$

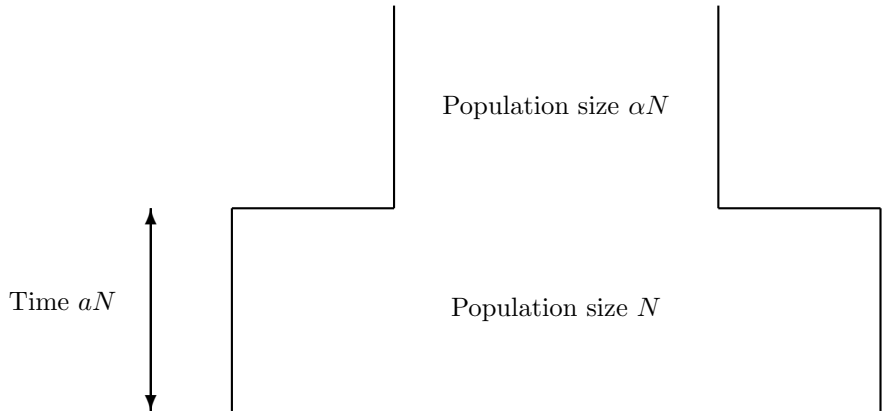


Figure 1: Backwards in time the population size is N in the first epoch and αN in the second epoch. The chainge in population size happens at time a in the past.

2.2 Three epochs of constant size

With three epochs of constant size the relative size function becomes

$$f(x) = \begin{cases} 1 & 0 \leq x < a \\ \alpha & a \leq x < b \\ \beta & x \geq b \end{cases}$$

and

$$\Lambda^{-1}(u) = \begin{cases} u & 0 \leq u < a \\ a + \alpha(u-a) & a \leq u < a + \frac{b-a}{\alpha} \\ b + \beta\big[(u-a) - \frac{(b-a)}{\alpha}\big] & u \geq a + \frac{b-a}{\alpha}. \end{cases}$$

3 Demographic model for chimpanzees

In Figure 2 we show the observed and estimated site frequency spectra for the three subspecies of chimpanzees. The estimated site frequency spectra are based on the demographic model obtained from the ABC procedure. In Figure 3 we show the observed and estimated folded site frequency spectra from the central, eastern and western chimpanzees.

References

Tavare, S. (2004). *Ancestral inference in population genetics.* Lectures on Probability Theory and Statistics. Volume 1837, 2004, pp 1-188. Springer, Berlin Heidelberg.

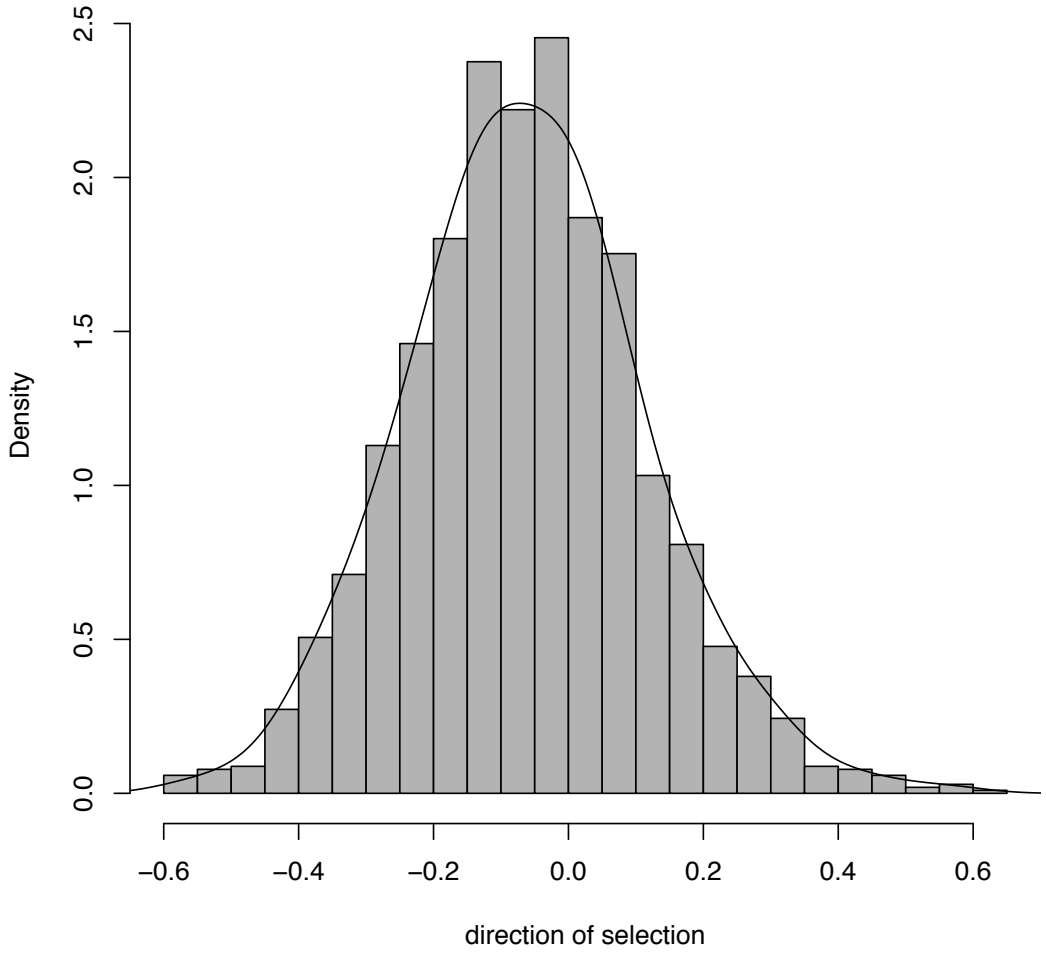
Durrett, R. (2008). *Probability Models for DNA Sequence Evolution.* 2nd edition. Springer, New York.

List of supplementary tables & figures

Supplementary Table 1 Geographic origin and depth of sequencing of individuals analyzed.

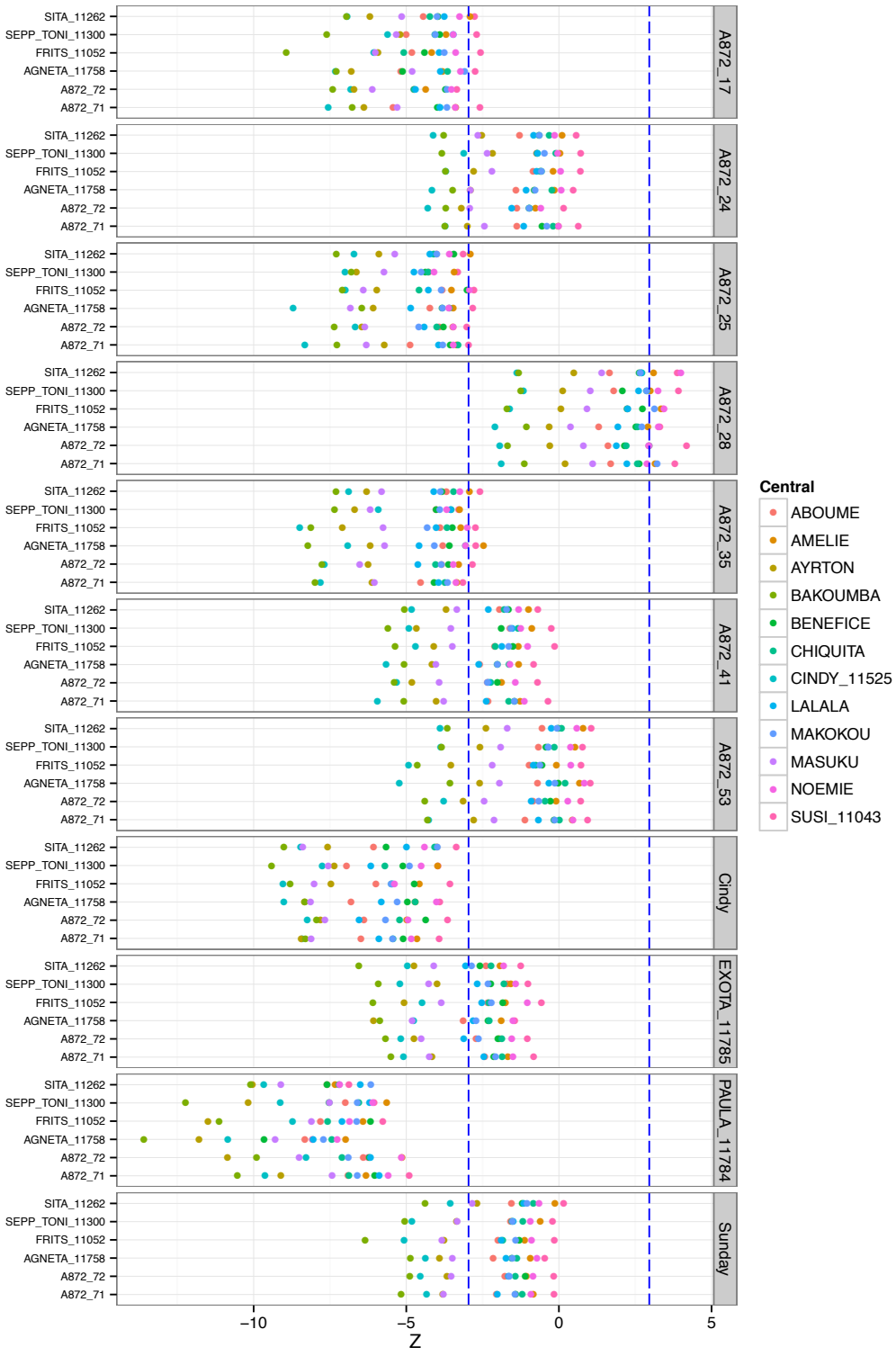
Individual ID	Sex	Origin	Subspecies	Avg. Seq. depth
Yoyo	F	Democratic Republic of Congo	*Pan troglodytes schweinfurthii*	*32.9*
Umugenzi	M	Democratic Republic of Congo	*Pan troglodytes schweinfurthii*	*32.2*
Ikuru	F	Democratic Republic of Congo	*Pan troglodytes schweinfurthii*	*38.2*
Tumbo	M	Kibale (Uganda)	*Pan troglodytes schweinfurthii*	*32.2*
Natasha	F	Kibale (Uganda)	*Pan troglodytes schweinfurthii*	*35.5*
Nkuumwa	F	Uganda	*Pan troglodytes schweinfurthii*	*38.9*
Pasa	F	Democratic Republic of Congo	*Pan troglodytes schweinfurthii*	*33.1*
Cindy	F	Budongo (Uganda)	*Pan troglodytes schweinfurthii*	*28.8*
Sunday	M	Democratic Republic of Congo	*Pan troglodytes schweinfurthii*	*33.4*
Exota (11785)	F	Wild caught	*Pan troglodytes schweinfurthii*	*36.7*
Paula 11784)	F	Wild caught	*Pan troglodytes schweinfurthii*	*46.2*
Susi (11043)	F	Wild caught	*Pan troglodytes troglodytes*	*36.6*
Cindy (11525)	F	Wild caught	*Pan troglodytes troglodytes*	*46.2*
Aboume	M	Gabon	*Pan troglodytes troglodytes*	
Amelie	F	Gabon	*Pan troglodytes troglodytes*	*35.9*
Ayrton	M	Moanda (Gabon)	*Pan troglodytes troglodytes*	*37.3*
Bakoumba	M	Gabon	*Pan troglodytes troglodytes*	*37.6*
Benefice	F	Makokou (Gabon)	*Pan troglodytes troglodytes*	*40.1*
Chiquita	F	Gabon	*Pan troglodytes troglodytes*	*37.7*
Lalala	F	Libreville (Gabon)	*Pan troglodytes troglodytes*	*29.8*
Makokou	F	Ogooué Ivindo (Gabon)	*Pan troglodytes troglodytes*	*38.8*
Masuku	F	Haut Ogooué (Gabon)	*Pan troglodytes troglodytes*	*34.6*
Noemie	F	Equatorial Guinea	*Pan troglodytes troglodytes*	*34.4*
Sita (11262)	F	1.gen of wild caught parents from Liberia	*Pan troglodytes verus*	*37.4*
Sepp-Toni (11300)	M	Wild Caught Liberia	*Pan troglodytes verus*	*33.5*
Olga (12314)	F	1.generation individual of wild caught parents from West Africa.	*Pan troglodytes verus*	*38.4*
Moreno (12341)	M	Wild caught West Africa	*Pan troglodytes verus*	*32.7*
Agneta (11758)	F	1.generation individual of wild caught parents from West Africa.	*Pan troglodytes verus*	*34.2*
Frits (11052)	M	Wild caught Sierra Leone	*Pan troglodytes verus*	*35.1*

Supplementary Figure 1: Non-smoothed histogram of DoS values observed across autosomal windows in central chimpanzees.

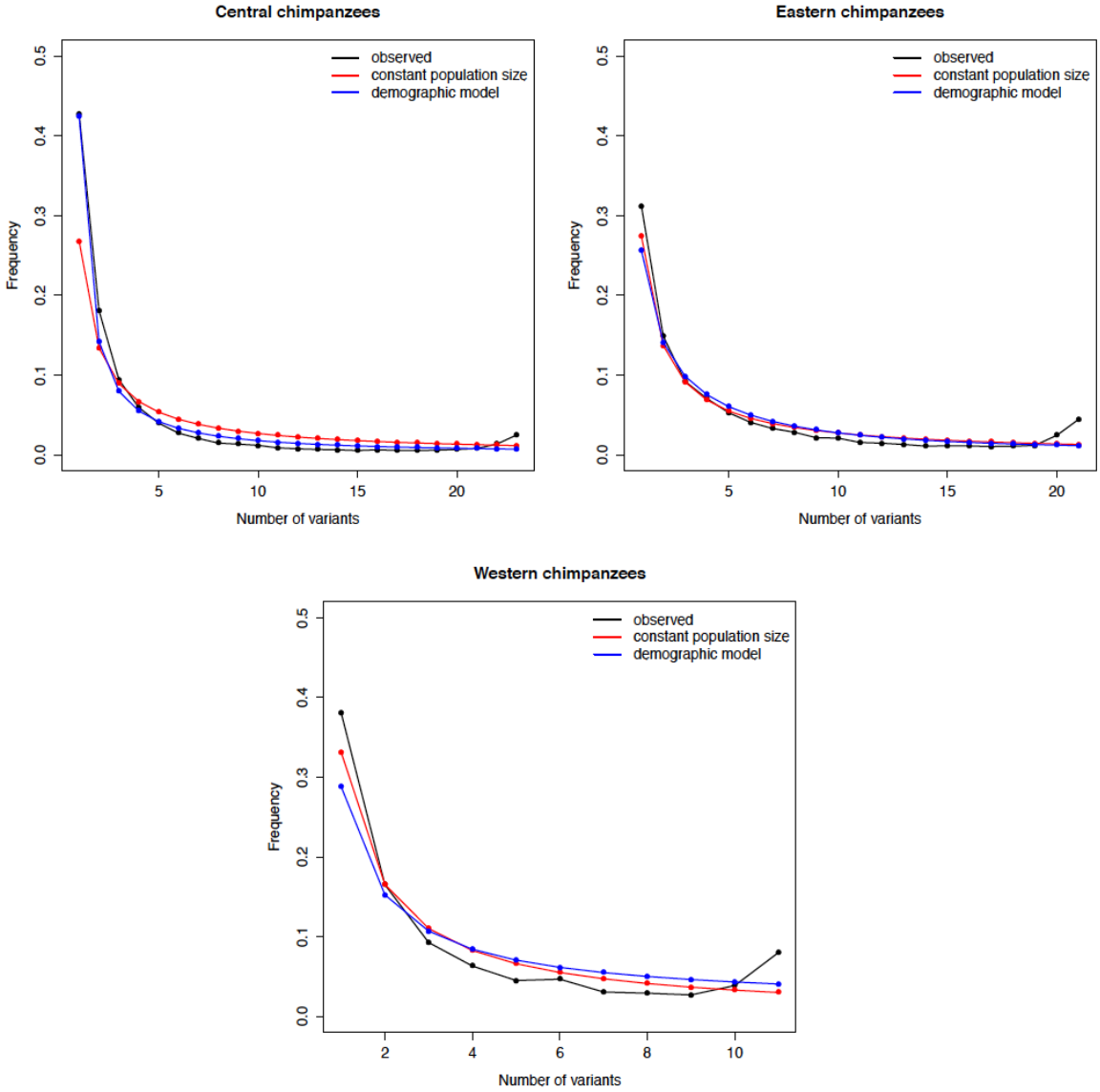


Supplementary Fig. 2 Distribution of D statistics for all three-way comparison of individuals

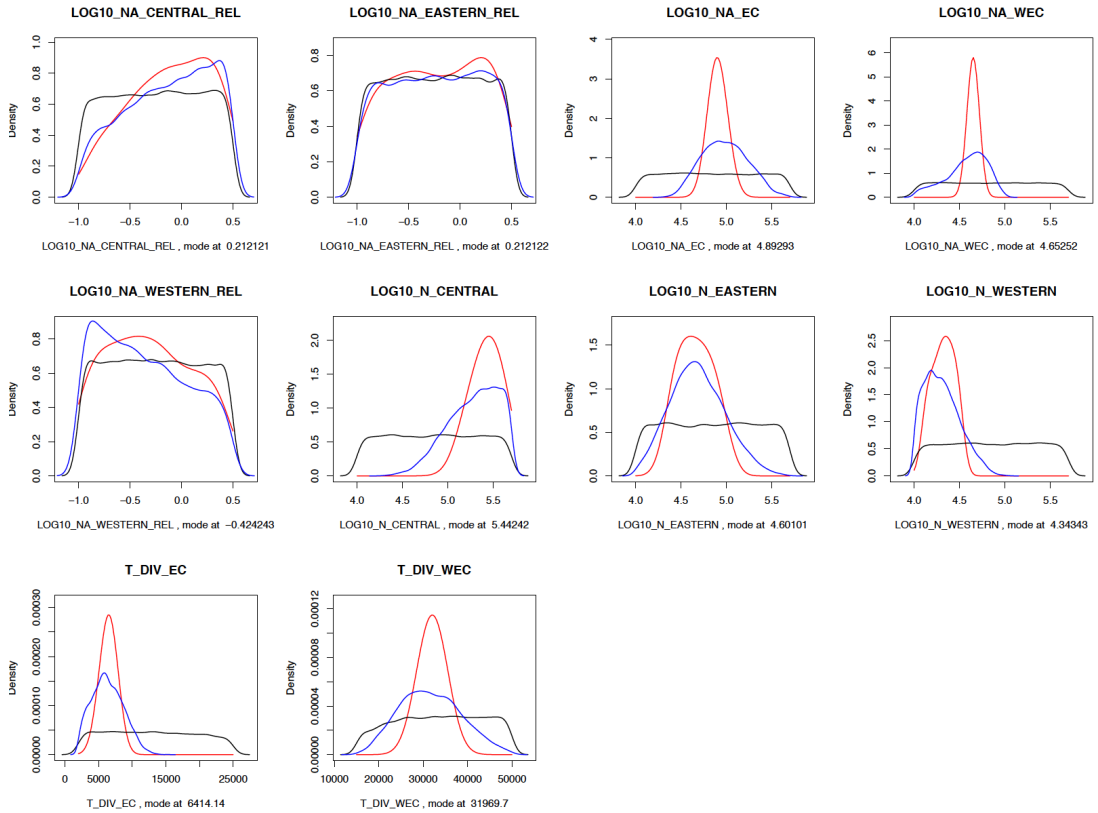
The plot shows all D statistics as Z scores with the significance threshold shown as vertical dashed lines. Each dot represents one three-way comparison. The different sub-plots correspond to different eastern chimpanzees, the different y-axis lines different western chimpanzees, and the different colors different central chimpanzees.



Supplementary Figure 3. Observed versus predicted synonymous autosomal SFS for each subspecies.

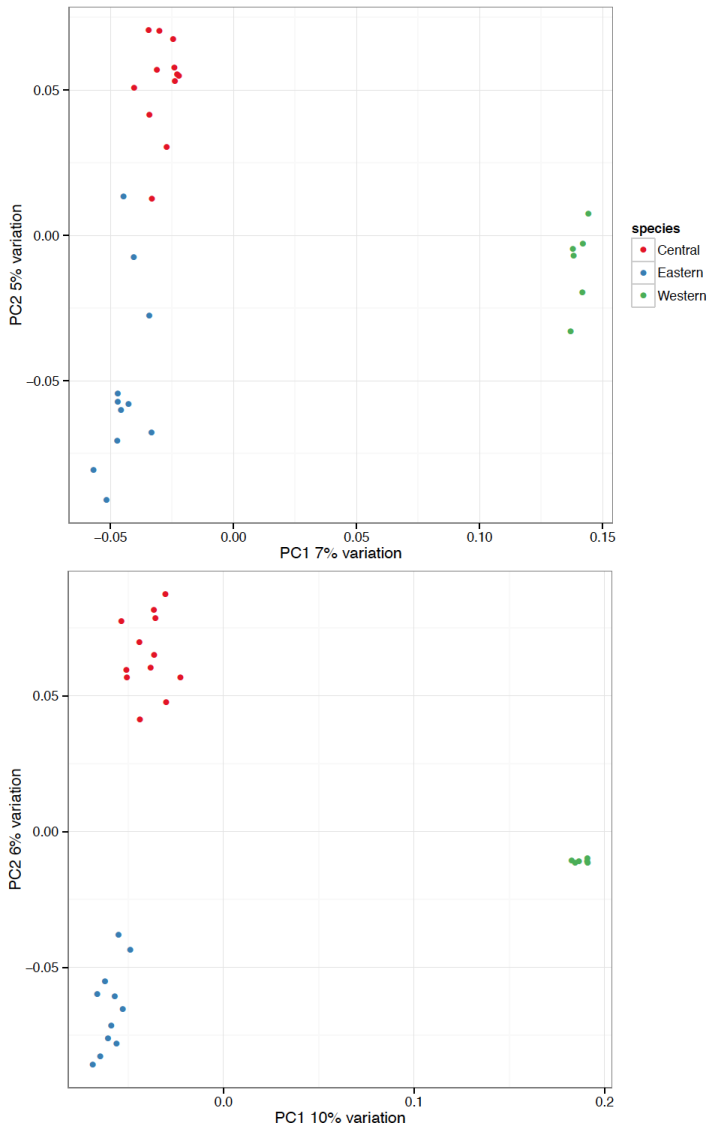


Predicted SFS under a constant population model (red) and the fitted demographic model (blue) reported in Figure 3.



Supplementary Figure 4

Prior (black) distribution and posterior distribution before (blue) and after post sampling adjustment (red) of parameters of the demographic model fitted to autosomal SNP data using our ABC procedure.



Supplementary Figure 5: comparison of PCA of indels versus SNPs.

To factor out the effect of sample size on the differentiation of populations revealed by PCA analysis of SNPs and indels, we down sampled the SNP data set to the size of the indel data set (n=2073) and redid the PCA analysis. The analysis still shows a smaller degree of population differentiation in the indels (TOP) than the SNP (BOTTOM). Note that to make the plots directly comparable the resulting PCA coordinates were normalized with the sum of standard deviations of the principal components.