SUPPLEMENTARY TEXT

Comparing temporal and divergence models when population size is not constant

Taking mutations into account, F_{ST} can be expressed in terms of probabilities of identity by descent (IBD) such as $F_{ST} = (f_w - f_b)/(1 - f_b)$. Here f_b is the probability of IBD for lineages picked from different populations and f_w is the probability of IBD for lineages picked from the same population (averaged over the different populations). For instance, if f_1 and f_{w2} are the probabilities of IBD in two different groups 1 and 2

$$F_{ST}=\frac{0.5(f_1+f_2)-f_b}{1-f_b}$$

The parametrization of the two different models ('temporal' and 'divergence') that we have considered in the main text in order to allow for different population sizes is described in Figure S1. It is useful to define $a(t_i,c_i)$: the probability of neither a mutation- nor a coalescence-event in two lineages during a time period t_i during which the population size was Nc_i. (i='1', '2' or 'A')

$$\alpha(t_i,c_i)=\left((1-\mu)^{t_i}\right)^2\left(1-(2Nc_i)^{-1}\right)^{t_i}\approx e^{-T_i(1+\theta c_i)/c_i}$$

where $T_i=t_i/2N$ and $\theta=4N\mu$. Probabilities of IBD (identity by descent) in the split model are

$$f_b=(1-\mu)^{t_1}(1-\mu)^{t_2}\frac{(2Nc_A)^{-1}}{(2Nc_A)^{-1}+2\mu}\approx\frac{e^{-\theta(T_1+T_2)/2}}{1+\theta c_A}$$

$$f_1=\frac{1-\alpha(t_1,c_1)}{1+\theta c_1}+\frac{\alpha(t_1,c_1)}{1+\theta c_A}$$

and

$$f_2=\frac{1-\alpha(t_2,c_2)}{1+\theta c_2}+\frac{\alpha(t_2,c_2)}{1+\theta c_A}.$$

Probabilities of IBD in the temporal model are

$$f_b=(1-\mu)^{t_1}(1-\mu)^{t_2}\frac{(2Nc_A)^{-1}}{(2Nc_A)^{-1}+2\mu}\approx\frac{e^{-\theta(T_1+T_2)/2}}{1+\theta c_A}$$

$$f_1=\frac{1-\alpha(t_1,c_1)}{1+\theta c_1}+\alpha(t_1,c_1)\frac{1-\alpha(t_2,c_2)}{1+\theta c_2}+\frac{\alpha(t_1,c_1)\alpha(t_2,c_2)}{1+\theta c_A}$$

$$f_2=\frac{1}{1+\theta c_A}$$

where $T_1=t_1/2N$ and $T_2=t_2/2N$ and t_1 and t_2 are the times (in generations) to the split of the two populations.

It may seem reasonable that if the t_i:s and the c_i:s are the same in the split and temporal models, then F_{ST} will be the same as well. This is however not the case unless $c_2 = c_A$. Looking at the probabilities of IBD we see that if the t_i:s and the c_i:s are the same, then f_b is always equal in the two models. Thus, if $f_1 + f_2$ equal each other in the two models then F_{ST} will be the same (equation 1).

In the split model:

$$f_1+f_2=\frac{1-\alpha(t_1,c_1)}{1+\theta c_1}+\frac{1-\alpha(t_2,c_2)}{1+\theta c_2}+\frac{\alpha(t_1,c_1)+\alpha(t_2,c_2)}{1+\theta c_A}$$

In the temporal model:

$$f_1+f_2=\frac{1-\alpha(t_1,c_1)}{1+\theta c_1}+\alpha(t_1,c_1)\frac{1-\alpha(t_2,c_2)}{1+\theta c_2}+\frac{1+\alpha(t_1,c_1)\alpha(t_2,c_2)}{1+\theta c_A}.$$

Setting these equations equal to each other leads to $c_2 = c_A$ (for $t_1 > 0$ and $t_2 > 0$, and f_1+f_2 will also be the same in the two models if at least one of t_1 or t_2 is zero).

The reason why c_2 has to be equal to c_A can be traced to the fact that while, in the temporal model, the probability of IBD in population 2 only depends on c_A, it depends on both c_2 and c_A in the split model. Another way to illustrate this is to consider what happens in the limit when t_1 and t_2 goes to infinity. Then $f_b = 0$ and F_{ST} is the mean of f_1 and f_2. While $f_1 = 1/(1+ \theta c_1)$ in both models, $f_2 = 1/(1+ \theta c_2)$ in the split model but $f_2 = 1/(1+ \theta c_A)$ in the temporal model.

Given the split model in Figure S1 one may further ask if we can replace c_1 and c_2

with a single value in the temporal model (so that the population size is constant during $t_1 + t_2$). In other words (since f_b remains unchanged) solve for x in

$$\frac{1-\alpha(t_1,c_1)}{1+\theta c_1}+\frac{1-\alpha(t_2,c_2)}{1+\theta c_2}+\frac{\alpha(t_1,c_1)+\alpha(t_2,c_2)}{1+\theta c_A}=\frac{1-\alpha(t_1+t_2,x)}{1+\theta x}+\frac{1+\alpha(t_1+t_2,x)}{1+\theta c_A}.$$

Here x will be heavily dependent on θ and does not always have a solution (for instance when $T_1 = T_2 = c_1 = c_2 = 1$, $c_A = 0.1$ and $\theta = 10$ corresponding to an ancestral bottleneck).

F_{ST} under a simple island model with temporal structure

Another model frequently employed in population genetics is the island model where two or more populations are separate but exchange migrants. Here we study the simplest such model with two populations of equal size, N, and a symmetric migration rate between them. We add temporal structure to this set-up so that one population is sampled t generations before the other population. In this way, the probability of IBD for two lineages picked from the same population is independent of population ($f_1=f_2=f_w$) while the probability of IBD for two lineages picked from different populations (f_b) will depend on the time separation of the samples. If f_{b0} denotes the probability of IBD for lineages picked from different populations with no time separation between them and $g(t)$ is the probability for a lineage to be in the same population (as it started) in after t generations, then

$$F_{ST}=\frac{0.5(f_1+f_2)-f_b}{1-f_b}=\frac{f_w-(1-\mu)^t(g(t)f_{b0}+(1-g(t))f_w)}{1-(1-\mu)^t(g(t)f_{b0}+(1-g(t))f_w)}$$

since for two lineages to be ibd when they are picked at different time points from different populations - the younger lineage can not mutate during the t generations and when it reaches the older time point, it will with probability 1-$g(t)$ be in the same population as the other lineage.

Since the migration rate (the probability to change population one generation back in time), m, is symmetric the probability to remain in the same population is the probability of an even number of migration events during the t generations. Assuming t is even (the approximation is obviously the same also if t is an odd number of generations),

$$g(t)=\sum_{i=0}^{t/2}\binom{t}{2i}m^{2i}(1-m)^{t-2i}\approx e^{-MT}\sum_{i=0}^{NT}\frac{(MT)^{2i}}{(2i)!}\approx e^{-MT}0.5\left(e^{MT}+e^{-MT}\right)=0.5\left(1+e^{-2MT}\right)$$

where $M=2Nm$, $T=t/2N$ (as above), the first approximation comes from the Poisson approximation of the binomial and the second from the power series representation of the exponential function. Since $(1-\mu)^t$ is well approximated by $\exp(-\theta T/2)$ and ($\theta=4N\mu$) we get

$$F_{ST}=\frac{2f_w-e^{-\theta T/2}(f_{b0}+f_w)-e^{-(\theta+4M)T/2}(f_{b0}-f_w)}{2-e^{-\theta T/2}(f_{b0}+f_w)-e^{-(\theta+4M)T/2}(f_{b0}-f_w)}. \quad \text{(i)}$$

To get f_w and f_{b0} we solve the recursions

$$f_w=(1-\mu)^2[((1-m)^2+m^2)((1-(2\mathrm{N})^{-1})f_w+(2\mathrm{N})^{-1})+2m(1-m)f_{b0}]$$

$$f_{b0}=(1-\mu)^2[2m(1-m)((1-(2\mathrm{N})^{-1})f_w+(2\mathrm{N})^{-1})+((1-m)^2+m^2)f_{b0}]$$

which arise from considering what happens going one generation back in time. The solution to this is

$$f_w=\frac{(1-\mu)^2}{2N}\left(1-(1-\mu)^2-2m(1-m)+2m(1-m)(1-\mu)^2\right)/D$$

$$f_{b0}=\frac{(1-\mu)^2}{2N}2m(1-m)/D$$

where

$$D=1-(2-1/2N)(1-2m(1-m))(1-\mu)^2+(1-1/2N)(1-4m(1-m))(1-\mu)^4$$

Expressing things in terms of θ, M and T

$$f_w=\frac{(2N)^{-2}(\theta+2M+O(1/N))}{D}$$

$$f_{b0}=\frac{(2N)^{-2}(2M+O(1/N))}{D}$$

$$D=(2N)^{-2}(\theta^2+\theta+2M+4M\theta+O(1/N))$$

Inserting f_w and f_{b0} into equation (i) we get

$$F_{ST}=\frac{(\theta+4M)(1-e^{-\theta T/2})+\theta(1+e^{-(\theta+4M)T/2})}{(\theta+4M)(2\theta+1-e^{-\theta T/2})+\theta(1+e^{-(\theta+4M)T/2})}.$$ Note that

$$T\to 0\Rightarrow F_{ST}\to\frac{1}{1+\theta+4M}$$

which is the classic result for an island model without temporal structure;

$$M\to 0\Rightarrow F_{ST}\to\frac{1}{1+\theta}$$

which also makes sense since in this case $f_b=0$ so that $F_{ST}=f_w$ *and* $f_w \approx (1/2N)/(2\mu+1/2N)$;

$$M\to\infty \Rightarrow F_{ST}\to\frac{1-e^{-\theta T/2}}{2\theta+1-e^{-\theta T/2}}$$

which is the same as a pure temporal model with a constant population of double size (so that the scaled mutation rate and scaled migration rate is larger by a factor 2 while the scaled time is half as large).

Conditioning on the ancestral population being polymorphic (Nei's expectation on F_{ST} of divergence time)

An alternative expression for F_{ST} is in terms of variance in allele frequency:

$$F_{ST}=\frac{E\left[(X_1-X_2)^2\right]}{E\left[X_1(1-X_2)+X_2(1-X_1)\right]},$$

where X_1 and X_2 are the allele frequencies in populations 1 and 2 respectively. By conditioning on the ancestral population being polymorphic, we can rely on the expression by [28]:

$$E\left[(X_t-p)^2 \mid X_0=p\right]=p(1-p)\left(1-\left(1-\frac{1}{2\mathrm{N}}\right)^t\right)\approx p(1-p)\left(1-e^{-T}\right)$$

where X_t is the allele frequency after t generations (forward in time), X_0 is the initial allele frequency and $T=t/2N$.

Split model

If the allele frequency in the ancestral population in the split model is designated by Z and we condition on $0<Z<1$:

$$E\left[(X_1-X_2)^2 \mid 0<Z<1\right]=\int_0^1 E\left[(X_1-p-(X_2-p))^2 \mid Z=p\right]f_Z(p \mid 0<Z<1)dp$$

$$=\int_0^1 E\left[(X_1-p)^2+(X_2-p)^2-2((X_1-p)(X_2-p)) \mid Z=p\right]f_Z(p \mid 0<Z<1)dp$$

$$=\left(\left(1-e^{-T_1}\right)+\left(1-e^{-T_2}\right)\right)\int_0^1 p(1-p)f_Z(p \mid 0<Z<1)dp$$

and

$$E[X_1(1-X_2)+X_2(1-X_1)\,|\,0<Z<1]=2\int_0^1 p(1-p)f_Z(p\,|\,0<Z<1)dp$$

In both of these derivations we have used the independence of X_1 and X_2 given $Z = p$ so that $E[X_1X_2|Z=p]=p^2$. We have

$$F_{ST}=\frac{\left(1-\mathrm{e}^{-T_1}\right)+\left(1-\mathrm{e}^{-T_2}\right)}{2}$$

which is equal to $1\text{-}e^{-T/2}$ for $T_1 = T_2 = T/2$.

Temporal model

In the temporal model conditioning on $0 < X_2 < 1$, (the older population being polymorphic), we have:

$$E\left[(X_1-X_2)^2\,|\,0<X_2<1\right]=\int_0^1 E\left[(X_1-p)^2\,|\,X_2=p\right]f_{X_2}(p\,|\,0<X_2<1)dp$$

$$=\left(1-e^{-T}\right)\int_0^1 p(1-p)f_{X_2}(p\,|\,0<X_2<1)dp$$

$$E[X_1(1-X_2)+X_2(1-X_1)\,|\,0<X_2<1]=2\int_0^1 p(1-p)f_{X_2}(p\,|\,0<X_2<1)dp$$

so that

$$F_{ST}=\frac{\left(1-\mathrm{e}^{-T}\right)}{2}.$$

A comment

Note that the frequency density in the integrals (both in the split model and the temporal model) could be replaced by any frequency density (since the integral cancel in the ratio) just as long as the integral is not evaluated at zero. We could for instance replace the conditioning on a polymorphic state ($0<X<1$) by a specific ascertainment scheme. See also Bhatia, et al (2013) for an excellent review and discussion of this (and similar) definitions of F_{ST}.

REFERENCES

Bhatia G, Patterson N, Sankararaman S, Price AL 2013. Estimating and interpreting F_{ST}: the impact of rare variants. Genome Res 23: 1514–1521.

SUPPLEMENTARY FIGURES

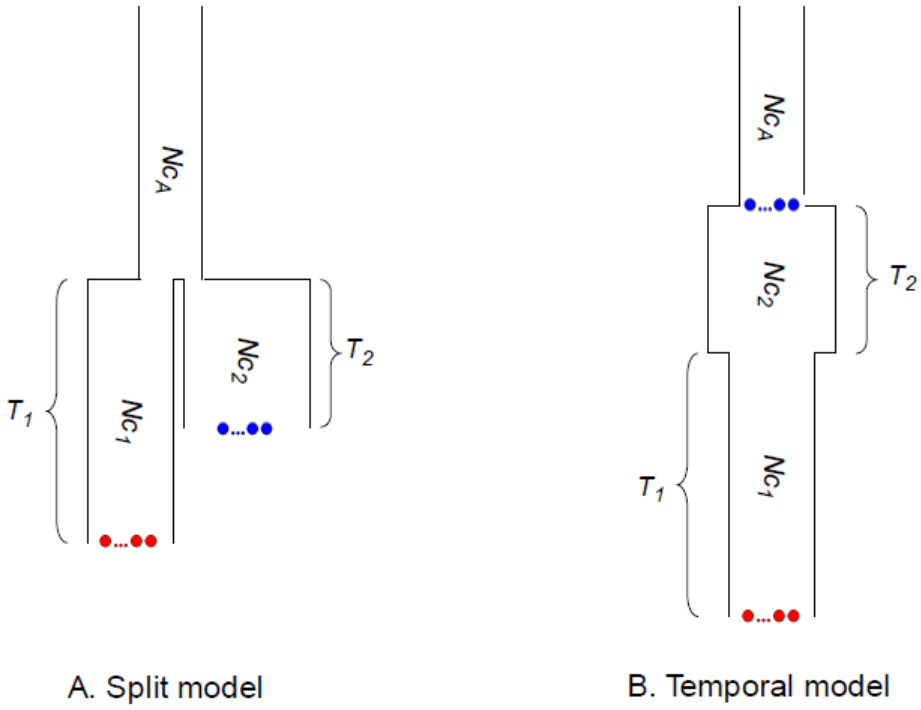

Figure S1. Illustration of the two models considered in terms of expected F_{ST} under variable population size.

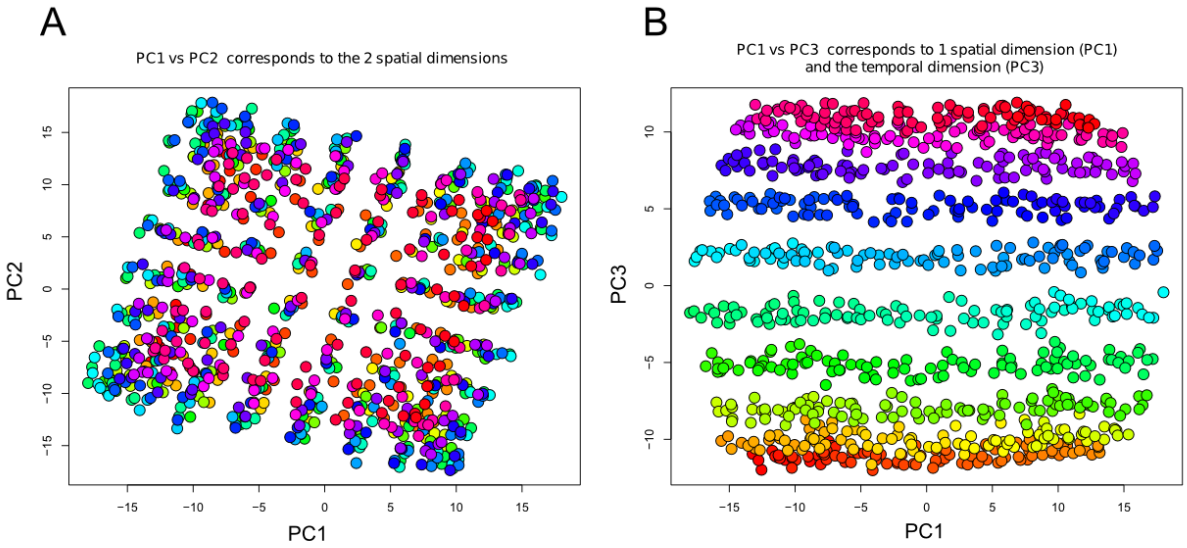

Figure S2. Principal component analysis under an isolation-by-distance model with temporal sampling. C) PC1 and PC2 recapitulate the two spatial (Procrustes correlation: 0.985, $P << 10^{-5}$). D) PC1 and PC3 recapitulate one spatial dimension and the temporal dimension of the model. As expected when migration rates are spatially homogeneous, the first two PCs explained similar fractions of the total variance (~2.2% in both cases). PC3, reflecting the temporal dimension, explained less of the total variance (1.6%), but this will depend on the time between samples and the spatial structure of the genetic model.

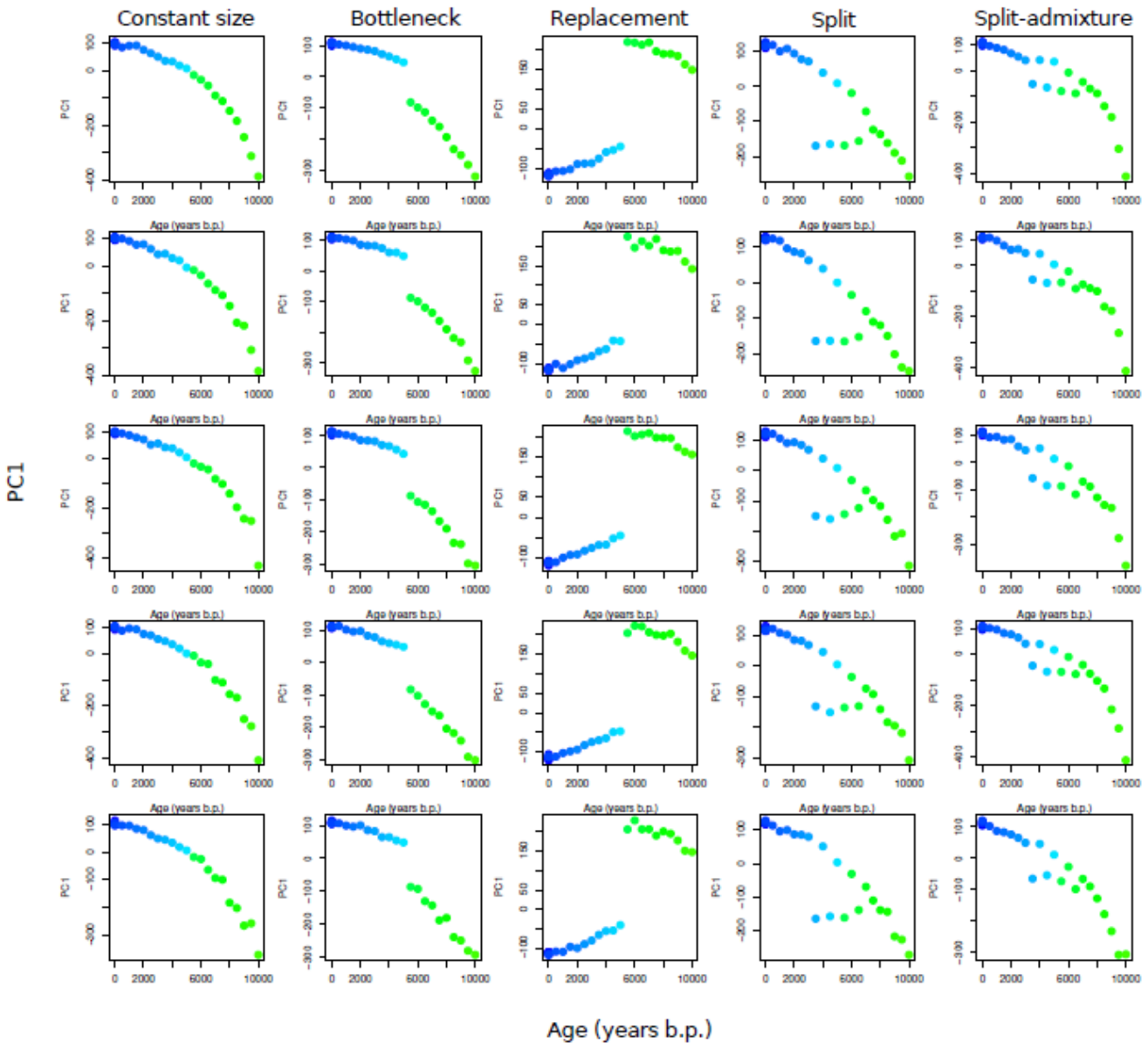

Figure S3. PCA of 5 replicates of each of the models in Figure 5 and 7.

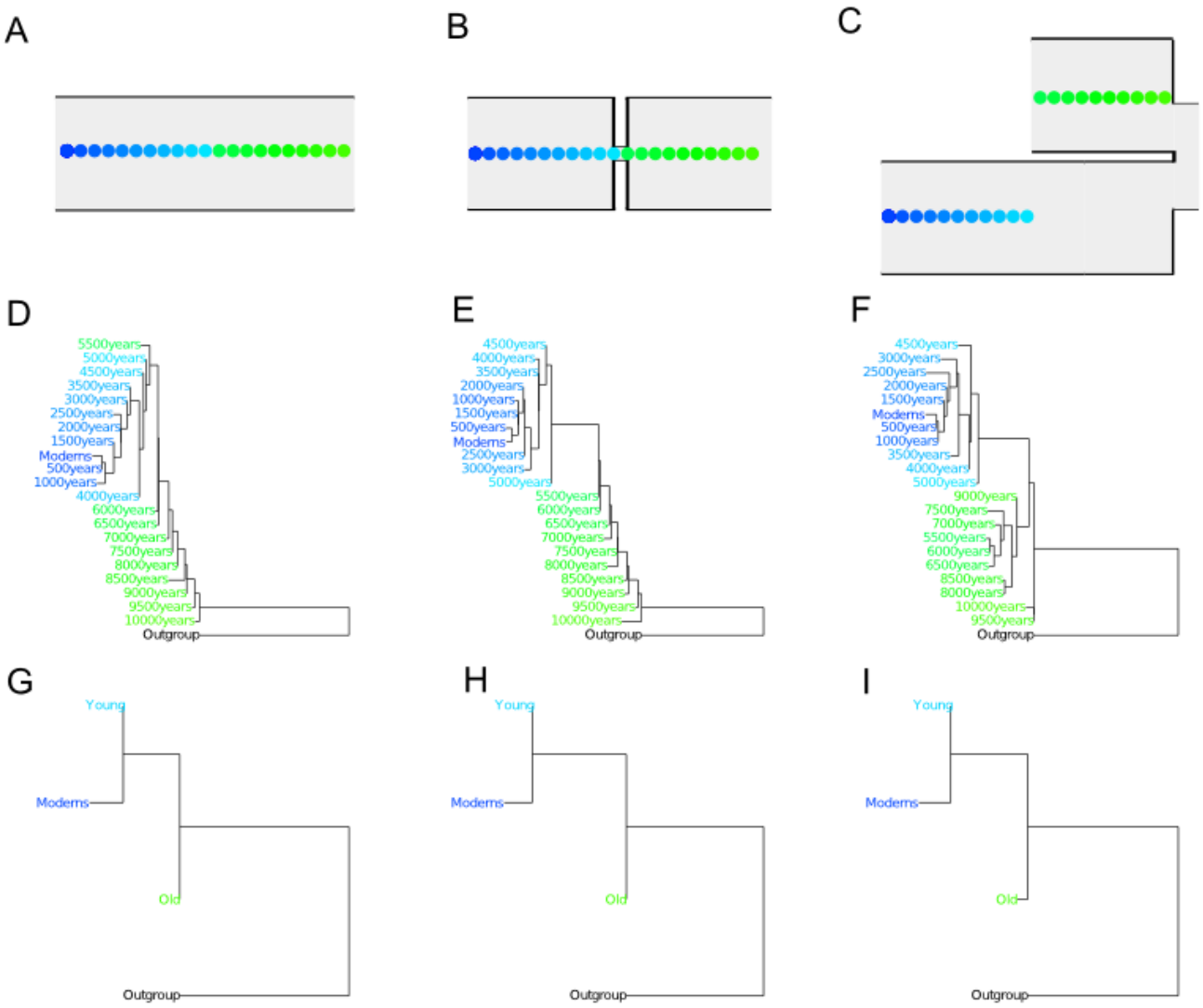

Figure S4. Treemix (version 1.11) analysis of temporal data from models displayed in Figure 5. The trees were reconstructed assuming no migration and using a block size of 1000 SNPs for estimating standard errors. The outgroup was constructed by taking the ancestral allele at each SNP. A) Constant-size model. B) Bottleneck model. C) Replacement model. D-F) Treemix analysis of models in A-C with only modern individuals pooled. G-I) Treemix analysis of models in A-C with three temporal groups. 'Moderns': 20 samples from time 0. 'Young': 10 samples from time 0-5000 years ago. 'Old': 10 samples from 5500 years ago to 10,000 years ago. The maximum likelihood trees recapitulate the patterns seen in PC analyses and can distinguish among models if a wide temporal range of samples is considered (D-F). Treemix cannot distinguish among models if the samples are grouped in the analysis.

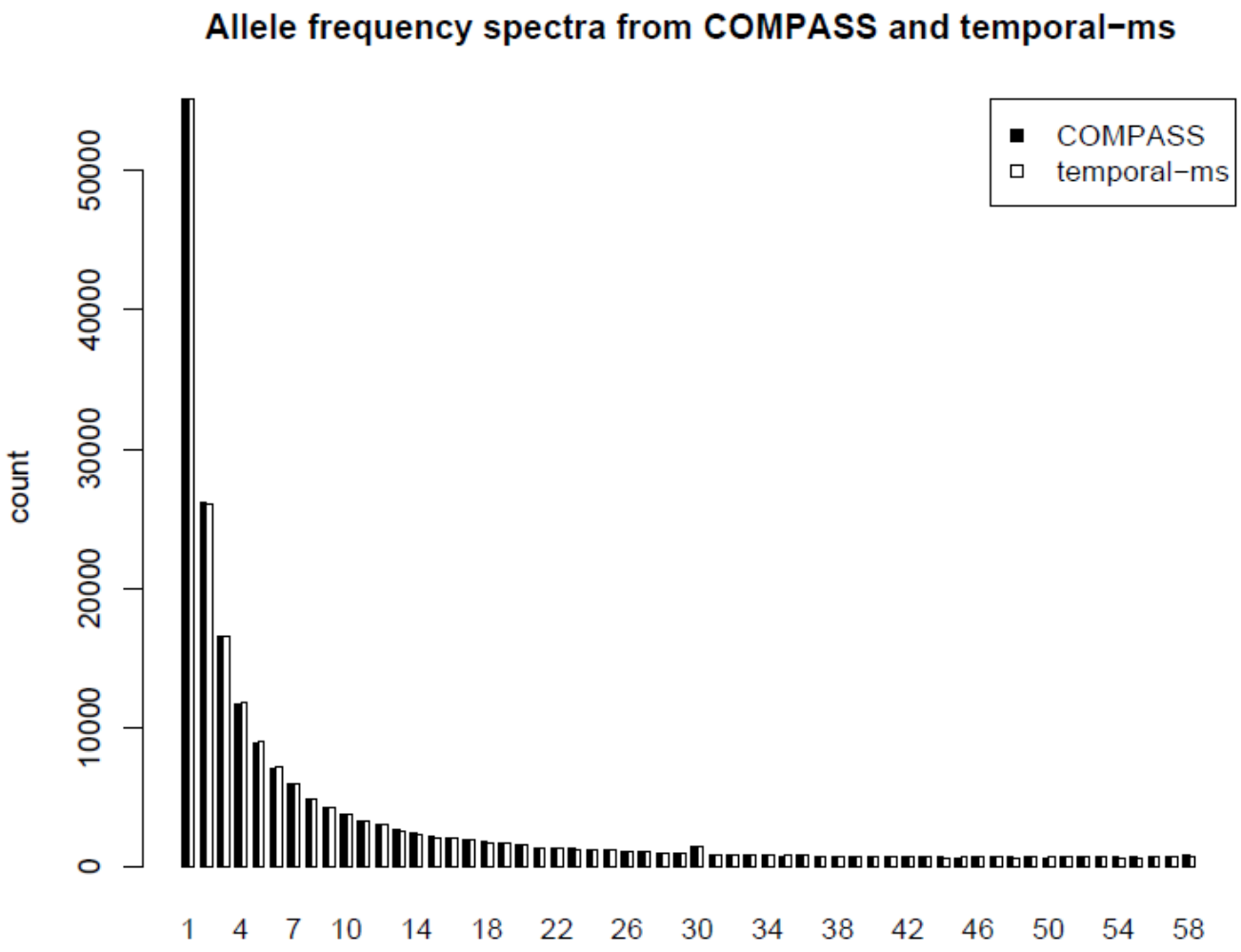

Figure S5. Site frequency spectra of data simulated as in the model in Figure 1A using COMPASS and the temporal algorithm added onto ms.