PNAS
www.pnas.org

Supplementary Information for

Nonparametric coalescent inference of mutation spectrum history and demography

William S. DeWitt, Kameron Decker Harris, Aaron P. Ragsdale, Kelley Harris

wsdewitt@uw.edu (WSD)
harriske@uw.edu (KH)

This PDF file includes:

Supporting Information Text

1. SI Appendix

A. Proof of *Theorem* : the expected SFS given demographic and mutation intensity histories. Suppose n haplotypes are sampled in the present, and let random vector $\mathbf{T} = [T_2, \ldots, T_n]^\intercal$ denote the coalescent times measured retrospectively from the present, i.e. T_n is the most recent coalescent time, and T_2 is the TMRCA of the sample.

As in Section 3 of (1), we consider a marked Poisson process in which every mutation is assigned a random label drawn iid from the uniform distribution on $(0, 1)$. This is tantamount to the infinite sites assumption, with the unit interval representing the genome, and the random variate labels representing mutant sites. Further suppose that mutation intensity at time t (measured retrospectively from the present in units of Wright-Fisher generations) is a function of time $0 \leq \mu(t) < \infty$ (measured in mutations per genome per generation) applying equally to all lines in the coalescent tree. A given line in the coalescent tree then acquires mutations on a genomic subinterval $(p, p + dp) \subseteq (0, 1)$ at rate $\mu(t)dp$.

Let $\mathcal{E}_{dp,b}$ denote the event that a mutation present in $b \in \{1, 2, \ldots, n-1\}$ haplotypes in the sample occurred within a given genomic interval $(p, p + dp)$. Given the uniform labeling assumption, the probability of this event is independent of p, but the following argument can be generalized to allow the labelling distribution to be nonuniform over the unit interval without changing the result. Let I_k denote the kth intercoalescent time interval, i.e. $I_n = (0, T_n), I_{n-1} = (T_n, T_{n-1}), \ldots, I_2 = (T_3, T_2)$. Let $\mathcal{E}_{dp,b,k}$ denote the event that the mutation $\mathcal{E}_{dp,b}$ occurred during interval I_k (index k here is not to be confused for the k-mer size in the main text). For small dp and finite $\mu(t)$ we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{dp,b} \mid \mathbf{T}) &= \sum_{k=2}^{n} \mathbb{P}(\mathcal{E}_{dp,b,k} \mid \mathbf{T}) \\ &= \sum_{k=2}^{n} p_{n,k}(b) \left(k\, dp \int_{t \in I_k} \mu(t)dt + O\left((dp)^2\right) \right),\end{aligned}$$

where

$$p_{n,k}(b) \equiv \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} \qquad [1]$$

is the probability that a mutant that arose when there were k ancestral lines of n sampled haplotypes will be present in b of them (see (1), eqn. 1.9). The quantity in parentheses is the probability that a mutation arose during the kth intercoalescent interval in a genomic interval of size dp. Marginalizing $\mathbf{T}$ gives

$$\mathbb{P}(\mathcal{E}_{dp,b}) = dp \sum_{k=2}^{n} k p_{n,k}(b) \mathbb{E}_{\mathbf{T}} \left[\int_{t \in I_k} \mu(t)dt \right] + O\left((dp)^2\right).$$

For small dp, each genomic interval $(p, p + dp)$ contains zero or one mutations. Therefore, taking the limit $dp \to 0$ and integrating over the genome, the expected number of mutations subtending b haplotypes (i.e. the bth component of the SFS) is

$$\xi_b = \int_0^1 \mathbb{P}(\mathcal{E}_{dp,b}) = \sum_{k=2}^{n} k p_{n,k}(b) \mathbb{E}_{\mathbf{T}} \left[\int_{t \in I_k} \mu(t)dt \right]$$

We now substitute in the bounds of every intercoalescent interval $I_k = (T_{k+1}, T_k)$, giving

$$\begin{aligned}\xi_b &= \sum_{k=2}^{n} k p_{n,k}(b) \mathbb{E}_{T_k} \left[\int_0^{T_k} \mu(t)dt \right] - \sum_{k=2}^{n-1} k p_{n,k}(b) \mathbb{E}_{T_{k+1}} \left[\int_0^{T_{k+1}} \mu(t)dt \right] \\ &= \sum_{k=2}^{n} k p_{n,k}(b) \mathbb{E}_{T_k} \left[\int_0^{T_k} \mu(t)dt \right] - \sum_{k=3}^{n} (k-1) p_{n,k-1}(b) \mathbb{E}_{T_k} \left[\int_0^{T_k} \mu(t)dt \right] \\ &= \sum_{k=2}^{n} B_{b,k} \mathbb{E}_{T_k} \left[\int_0^{T_k} \mu(t)dt \right], \qquad [2]\end{aligned}$$

where

$$B_{b,k} \equiv \begin{cases} k p_{n,k}(b), & k = 2 \\ k p_{n,k}(b) - (k-1) p_{n,k-1}(b), & k > 2 \end{cases} \qquad [3]$$

are combinatorial terms.

Polanski et al. (2), eqns. 5-8, give the marginal density for the coalescent time T_k as

$$\pi_k(t_k) = \sum_{j=k}^{n} A_{k,j} q_j(t_k), \qquad [4]$$

for $k = 2, \ldots, n$, where $\mathbf{A}$ is an $(n-1) \times (n-1)$ matrix indexed from $2, \ldots, n$ with

$$A_{k,j} \equiv \begin{cases} 1, & k = j = n \\ 0, & j < k, \\ \frac{\prod_{l=k\neq j}^{n} \binom{l}{2}}{\prod_{l=k\neq j}^{n} \left(\binom{l}{2} - \binom{j}{2}\right)}, & \text{otherwise} \end{cases}$$

and

$$q_j(t) \equiv \frac{\binom{j}{2}}{\eta(t)} \exp\left[-\binom{j}{2} \int_0^t \frac{dt'}{\eta(t')}\right],$$

for $j = 2, \ldots, n$, and $\eta(t)$ is the haploid effective population size history. We assume that $0 < \eta(t) < \infty$. Note that $q_j(t)$ is the probability density of the time to the first coalescent event among any subset of j individuals in the present, with inhomogeneous Poisson intensity function $\binom{j}{2}/\eta(t)$.

The expectations in Eq. (2) can be expressed using Eq. (4) as

$$\begin{aligned} \mathbb{E}_{T_k}\left[\int_0^{T_k} \mu(t)dt\right] &= \int_0^\infty \pi_k(t_k) \int_0^{t_k} \mu(t)dt\, dt_k \\ &= \sum_{j=k}^{n} A_{k,j} \int_0^\infty q_j(t_k) \int_0^{t_k} \mu(t)dt\, dt_k \\ &= \sum_{j=k}^{n} A_{k,j} \int_0^\infty q_j(t_k) \int_0^\infty \mathbb{1}_{[0<t<t_k]}\mu(t)dt\, dt_k \\ &= \sum_{j=k}^{n} A_{k,j} \int_0^\infty r_j(t)\mu(t)dt \end{aligned} \quad [5]$$

where in the last line we exchange integration order (by Fubini's theorem) and define the inhomogeneous Poisson survival function

$$r_j(t) \equiv \int_0^\infty q_j(t')\mathbb{1}_{[0<t<t']}dt' = \exp\left[-\binom{j}{2}\int_0^t \frac{dt'}{\eta(t')}\right] \quad [6]$$

corresponding to density $q_j(t)$.

Using Eq. (5) in Eq. (2) gives

$$\begin{aligned} \xi_b &= \sum_{k=2}^{n} B_{b,k} \sum_{j=k}^{n} A_{k,j} \int_0^\infty r_j(t)\mu(t)dt \\ &= \sum_{j=2}^{n} \left(\sum_{k=2}^{j} B_{b,k}A_{k,j}\right) \int_0^\infty r_j(t)\mu(t)dt, \end{aligned} \quad [7]$$

exchanging summation order in the last line. We then have a linear expression for the expected SFS as a function of the mutation intensity history $\mu(t)$:

$$\boldsymbol{\xi} = \mathbf{C}\,\mathbf{d}(\eta, \mu), \quad [8]$$

where the $(n-1) \times (n-1)$ matrix $\mathbf{C} = \mathbf{BA}$ is constant in μ and η, and

$$d_j(\eta, \mu) \equiv \int_0^\infty r_j(t)\mu(t)dt = \int_0^\infty \exp\left[-\binom{j}{2}\int_0^t \frac{dt'}{\eta(t')}\right]\mu(t)dt, \quad [9]$$

for $j = 1, \ldots, n-1$, is a linear functional of μ and a nonlinear functional of η.

Given the boundedness assumptions that we have on η and μ, we now prove boundedness of the map from joint history functions (η, μ) to expected SFS vectors ξ. The vector $\mathbf{d}(\eta, \mu)$ may be viewed as a nonlinear operator $\mathbf{d} : L^\infty(\mathbb{R}_{\geq 0}) \times L^\infty(\mathbb{R}_{\geq 0}) \to \ell^\infty_{n-1}$ of rank $n-1$, and is bounded element-wise. In the diffusion timescale (equation 3 of main text) , d_j is the Laplace transform of the bounded function $\tilde{\eta}\tilde{\mu}$ evaluated at $\binom{j}{2}$, and is thus bounded. In particular,

$$0 \leq d_j \leq \frac{\eta_{\max}\mu_{\max}}{\binom{j}{2}}, \text{ for } j = 1, \ldots, n-1, \quad [10]$$

where $\eta_{\max}$ and $\mu_{\max}$ are the respective bounds on η and μ. Boundedness of the full operator mapping (η, μ) to the expected SFS $\boldsymbol{\xi}$ follows from the fact that $\mathbf{C}$ is a matrix with bounded norm. This completes the proof of emphTheorem .

B. Computing the elements of C. We next develop an efficient recursive procedure for computing the matrix $\mathbf{C}$. Using Eq. (3)

$$\begin{aligned} C_{b,j} &= \sum_{k=2}^{j} k p_{n,k}(b) A_{k,j} - \sum_{k=3}^{j} (k-1) p_{n,k-1}(b) A_{k,j} \\ &= W^{(1)}_{b,j} - W^{(2)}_{b,j}, \end{aligned}$$

where

$$W^{(1)}_{b,j} \equiv \sum_{k=2}^{j} k p_{n,k}(b) A_{k,j} \quad [11]$$

$$W^{(2)}_{b,j} \equiv \sum_{k=3}^{j} (k-1) p_{n,k-1}(b) A_{k,j}. \quad [12]$$

Polanski et al. (3), eqn. 11, show that the nonzero entries of $\mathbf{A}$ can be expressed as

$$A_{k,j} = \frac{n!(n-1)!}{(j+n-1)!(n-j)!} \cdot \frac{(2j-1)}{j(j-1)} \cdot \frac{(j+k-2)!}{(k-1)!(k-2)!(j-k)!} \cdot (-1)^{j-k}.$$

Given the form of $p_{n,k}(b)$ in Eq. (1), we see that Eq. (11) and Eq. (12) are definite sums over hypergeometric terms. We used Zeilberger's algorithm (4, 5), which finds polynomial recurrences for definite sums of hypergeometric terms, to procedurally generate the following second-order recursions in j:

$$\begin{aligned} W^{(1)}_{b,2} &= \frac{6}{(n+1)} \\ W^{(1)}_{b,3} &= \frac{10(5n-6b-4)}{(n+2)(n+1)} \\ W^{(1)}_{b,j+2} &= -\Big[(2j+3)\big(-(2j-1)W^{(1)}_{b,j+1}\big(2j(j+1)\big(b^2\big(j^2+j-2\big)-6b-j(j+1)-2\big) \\ &\quad - j(j+1)n\big(3b\big(j^2+j+2\big)+j^2+j-2\big)+\big(j(j+1)\big(j^2+j+6\big)+4\big)n^2+4n\big) \\ &\quad - (j-1)(j+1)^2(j-n)W^{(1)}_{b,j}(4(n+1)-j(j+2)(b-n-1))\big)\Big] \\ &\quad \Big/\Big[j^2(j+2)(2j-1)(j+n+1)\left(-bj^2+b+\left(j^2+3\right)(n+1)\right)\Big] \end{aligned}$$

and

$$\begin{aligned} W^{(2)}_{b,2} &= 0 \\ W^{(2)}_{b,3} &= \frac{20(n-2)}{(n+1)(n+2)} \\ W^{(2)}_{b,j+2} &= \frac{(2j+3)(j-n+1)}{j}\left(\frac{(j+1)}{(2j-1)(j+n)}W^{(2)}_{b,j} - \frac{(j(j+1)(2b-n+1)-2(n+1))}{(j-1)(j+2)(j-n)(j+n+1)}W^{(2)}_{b,j+1}\right). \end{aligned}$$

These formulae are used to numerically compute the entries in $\mathbf{C}$.

C. Discretization of history functions and computation of $\mathbf{d}(\eta,\mu)$. We represent histories as piecewise constant functions of time on m pieces $[t_0,t_1),[t_1,t_2),\ldots,[t_{m-1},t_m)$, where $0=t_0<t_1<\cdots<t_{m-1}<t_m=\infty$. The grid is common to $\eta(t)$ and $\mu(t)$. We take the boundaries of the pieces as fixed parameters and in practice use a logarithmically-spaced dense grid of hundreds of pieces to approximate infinite-dimensional histories. Let column vector $\mathbf{y}=[y_1,\ldots,y_m]^\intercal$ denote the constant population size $\eta(t)$ during each piece, and let $\mathbf{w}=[w_1,\ldots,w_m]^\intercal$ denote the constant mutation rate $\mu(t)$ during each piece.

With this we can follow the proof of Proposition 1 in (6), *mutatis mutandis*, with our Eq. (9) to arrive at

$$\mathbf{d} = \mathbf{M}(\mathbf{y})\mathbf{w} \quad [13]$$

where

$$\mathbf{M}(\mathbf{y}) \equiv \begin{bmatrix} 1 & & & \\ & \frac{1}{3} & & \\ & & \ddots & \\ & & & \frac{1}{\binom{n}{2}} \end{bmatrix} \begin{bmatrix} 1 & u_1 & \ldots & \prod_{i=1}^{m-1} u_i \\ 1 & u_1^3 & \ldots & \prod_{i=1}^{m-1} u_i^3 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & u_1^{\binom{n}{2}} & \ldots & \prod_{i=1}^{m-1} u_i^{\binom{n}{2}} \end{bmatrix} \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \operatorname{diag}(\mathbf{y}), \quad [14]$$

and $u_l \equiv \exp(-(t_l - t_{l-1})/y_l)$ for $l = 1, \ldots, m$. Note that the $(n-1) \times m$ matrix $\mathbf{M}(\mathbf{y})$ is a nonlinear function of the demographic history $\mathbf{y}$ because the u_l are nonlinear functions of $\mathbf{y}$. This reflects the fact that it is a discretization of the nonlinear operator $\mathbf{d}(\cdot, \mu)$. Combining Eq. (13) with Eq. (8) gives the discretized forward model

$$\boldsymbol{\xi} = \mathbf{L}(\mathbf{y})\mathbf{w}, \tag{15}$$

where $\mathbf{L}(\mathbf{y}) \equiv \mathbf{CM}(\mathbf{y})$.

D. Proof of *Proposition* . The distribution of independent Poisson random variables factorizes into an aggregate Poisson random variable and a multinomial, a well-known result often called "Poissonization" (7). Poissonization over the mutation type index j gives

$$\begin{aligned}\mathbb{P}(\mathbf{X} \mid \boldsymbol{\Xi}) &= \prod_{i=1}^{n-1} \prod_{j=1}^{K} \underbrace{\mathbb{P}(X_{i,j} \mid \Xi_{i,j})}_{\text{Poisson}} = \prod_{i=1}^{n-1} \underbrace{\mathbb{P}(x_i \mid \xi_i)}_{\text{Poisson}} \underbrace{\mathbb{P}\left([X_{i,1}, \ldots, X_{i,K}] \;\middle|\; x_i, \left[\frac{\Xi_{i,1}}{\xi_i}, \ldots, \frac{\Xi_{i,K}}{\xi_i}\right]\right)}_{\text{multinomial}} \\ &= \underbrace{\mathbb{P}(\mathbf{x} \mid \boldsymbol{\xi})}_{\text{PRF}} \quad \underbrace{\mathbb{P}(\mathbf{X} \mid \mathbf{x}, \hat{\boldsymbol{\Xi}})}_{\substack{\text{multinomial random} \\ \text{field over } i}} . \end{aligned} \tag{16}$$

This completes the proof of *Proposition* .

E. Proof of *Lemma* . Fix the mutation type i, and consider the multinomial over j

$$\mathbb{P}\left([X_{i,1}, \ldots, X_{i,K}] \;\middle|\; x_i, \left[\frac{\Xi_{i,1}}{\xi_i}, \ldots, \frac{\Xi_{i,K}}{\xi_i}\right]\right).$$

We must show that any element of the multinomial vector

$$\hat{\Xi}_{i,j} \equiv \frac{\Xi_{i,j}}{\xi_i}$$

can be formulated without reference to η. From elementary properties of the multinomial, the conditional expection value of $X_{i,j}$ given x_i is

$$\mathbb{E}[X_{i,j} \mid x_i] = x_i \; \hat{\Xi}_{i,j}.$$

Now, since mutation events are independent we perform a thinning operation on each of the x_i mutation events

$$\mathbb{E}[X_{i,j} \mid x_i] = x_i \; \mathbb{P}\,(\text{a mutation of sample frequency } i \text{ is of type } j) \tag{17}$$

$$= x_i \int_0^\infty \frac{\tilde{\mu}_j(\tau)}{\mu_0} a_i(\tau) d\tau, \tag{18}$$

where $a_i(\tau)$ is the pdf of a mutation's age τ measured in expected coalescent events (diffusion time) conditioned on its sample frequency i. So

$$\hat{\Xi}_{i,j} = \int_0^\infty \frac{\tilde{\mu}_j(\tau)}{\mu_0} a_i(\tau) d\tau.$$

This is independent of η by definition of the diffusion time scale as the intensity measure of the coalescent process. This completes the proof of *Lemma* .

F. Tempora incognita: observability toward the coalescent horizon. The time-domain singular vectors of $\mathcal{L}(\eta)$ form an eigenbasis for solutions $\mu(t)$ that are possible, in principle, to reconstruct from the SFS. However, sampling noise about the expected SFS will corrupt information from singular vectors that are associated to smaller singular values. These corrupted components will be the directions in solution space associated with higher frequency and less smooth dynamics. Since the singular values of $\mathcal{L}(\eta)$ have a very large dynamic range (the condition number is large), the presence of noise will limit reconstruction to smoother, more slowly varying components that are least corrupted and erase information about more sudden events.

Figure S9 depicts the observability of mutation rate history via spectral analysis of $\mathcal{L}(\eta)$ for a case with $\eta(t)$ a simple bottleneck history. From Eq. (4) and Eq. (6) in Appendix A, the CDF of the TMRCA can be computed given $\eta(t)$. We see only the top few components (ranked by singular value) persist at times older than the bottleneck, and all components vanish beyond the TMRCA of the sample. Higher frequency behavior becomes more difficult to observe if it is older than the bottleneck, concretely illustrating how demographic events erase information about population history.

G. Modeling ancestral state misidentification. Computing the SFS and the k-SFS from variant data requires polarizing reference and alternative alleles to ancestral and derived alleles. Ancestral states are themselves inferred (usually by invoking parsimony criteria in a comparison to an outgroup reference genome, or a larger multi-species alignment), so ancestral state misidentification is expected at some fraction of sites. Misidentification results in allele frequency complementation: a variant at sample frequency i out of n sampled haplotypes will appear to have frequency $n - i$. Misidentification also results in mutation type reversion: e.g. a variant of triplet mutation type TCC→TTC will appear of mutation type GAA→GGA.

Under very general conditions, the expected SFS $\boldsymbol{\xi}$ is a non-increasing vector: $\xi_i \geq \xi_j \ \forall i < j$ (8). This result covers all demographic histories $\eta(t)$. Given the pointwise nonidentifiability of η and μ (equation 3 of main text) , it also covers all mutation rate histories $\mu(t)$, so all columns of the expected k-SFS are non-increasing row-wise.

Ancestral state misidentification violates this non-increasing expectation result. The SFS for the subset of misidentified sites is reflected in frequency, so the SFS $\boldsymbol{\xi}$ becomes a sum of a non-increasing vector (for the correctly identified sites) and a non-*de*creasing vector (for the misidentified sites). This contributes to the so-called "smile" in empirical SFS data. Because this smile can't be explained by η and μ with a model that produces a non-increasing SFS, the misidentification rate r is identifiable as an additional parameter. Let ξ' denote the expected SFS with misidentification, so

$$\xi'_i = (1 - r)\xi_i + r\xi_{n-i}, \quad i = 1, \ldots, n - 1.$$

In matrix form this is

$$\boldsymbol{\xi}' = (1 - r)\boldsymbol{\xi} + r\boldsymbol{J}\boldsymbol{\xi},$$

where $\boldsymbol{J}$ denotes the $(n - 1) \times (n - 1)$ exchange matrix (with 1s on the anti-diagonal and 0s elsewhere).

For the k-SFS, misidentified sites contribute to counts in the reflected frequency row, and also in a different mutation type column, corresponding to the revertant mutation type. Let $\pi(j)$ denote the revertant partner of mutation type index j (π is a permutation of the mutation type columns). Let r_j denote the misidentification rate of mutation type j. Then the expected k-SFS with misidentification is

$$\Xi'_{i,j} = (1 - r_j)\Xi_{i,j} + r_j\Xi_{n-i,\pi(j)}, \quad i = 1, \ldots, n - 1, \quad j = 1, \ldots, K.$$

In matrix form this is

$$\boldsymbol{\Xi}' = \boldsymbol{\Xi}\ (\boldsymbol{I} - \mathrm{diag}(\boldsymbol{r})) + \boldsymbol{J}\ \boldsymbol{\Xi}\ \boldsymbol{P}_\pi^\intercal\ \mathrm{diag}(\boldsymbol{r}),$$

where $\boldsymbol{P}_\pi$ is the permutation matrix corresponding to π.

We infer the total misidentification rate r jointly with η inference, then infer the rates for each mutation type $\boldsymbol{r}$ jointly with $\boldsymbol{\mu}$, constraining compositionally such that $\sum_{j=1}^{K} f_j r_j = r$ via the isometric log ratio transform, where f_j is the fraction of variants from each mutation type (the column sums of the k-SFS normalized by its total). These additional parameters allow us to obtain very good fits to empirical SFS and k-SFS data that include prominent high frequency "smiles".

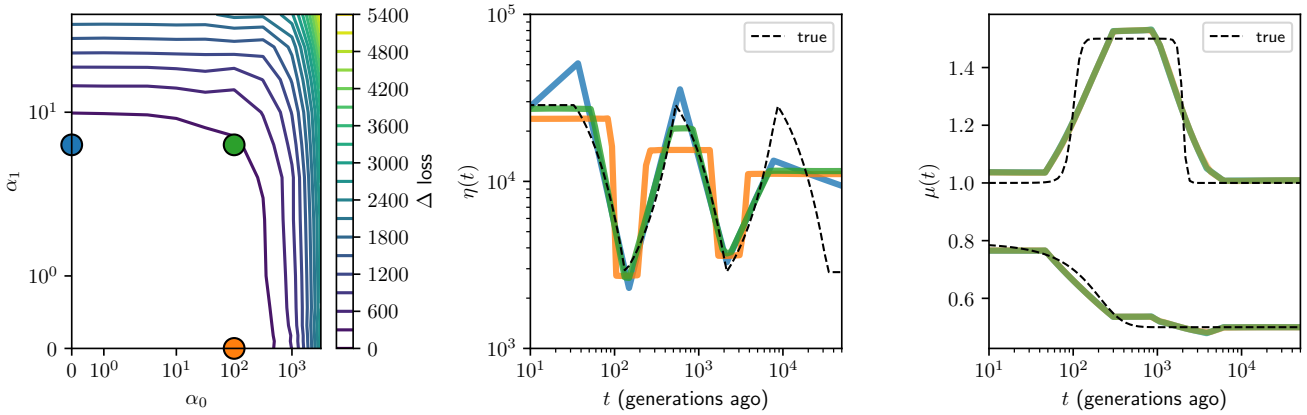

Fig. S1. The effect of demographic model selection on MuSH inference in our simulation study. Left panel shows change in loss (goodness of fit) as a function of 0-th order and 1st order trend penalties. Middle panel shows the inferred demography at each indicated penalty value (colors corresponding to points in left panel). Right panel shows the two variable components of subsequent MuSH inference for each demographic model, indicating they are very weakly effected by demographic model selection.

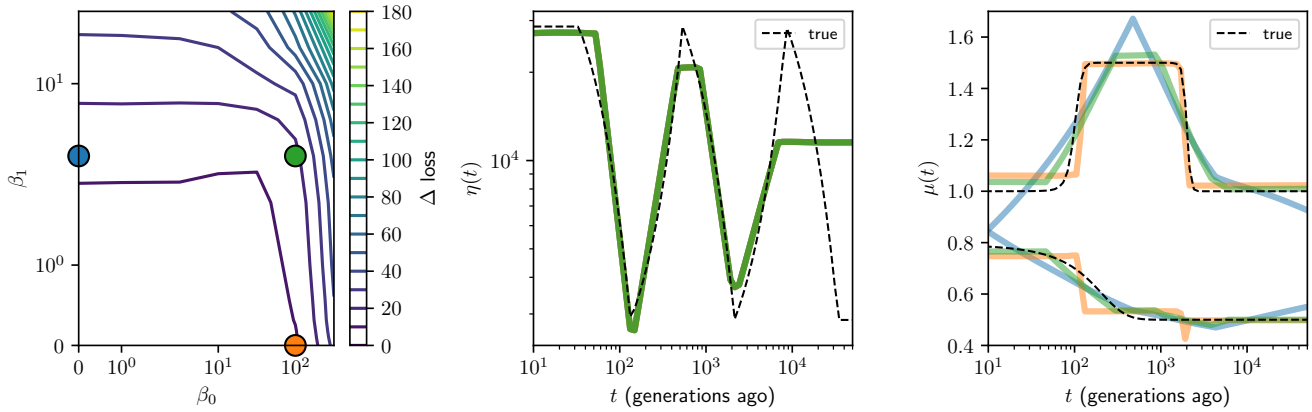

Fig. S2. The effect of MuSH regularization on MuSH inference in our simulation study. Left panel shows change in loss (goodness of fit) as a function of 0-th order and 1st order trend penalties. Middle panel shows the inferred demography (which is independent of the MuSH hyperparameters). Right panel shows the two variable components of the subsequent MuSH inference for each hyperparameter choice (colors corresponding to points in left panel).

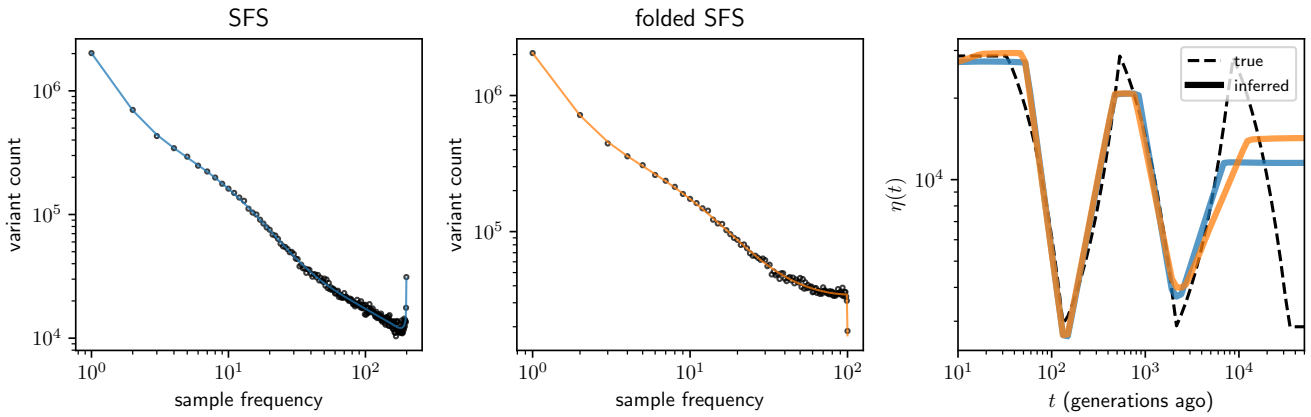

Fig. S3. Comparison of demographic inference using the folded Vs unfolded SFS in our simulation study. Using the same parameters as in Figure 1 , we fit the SFS (left), the folded SFS (middle), and show the resulting demographic histories are similar (right).

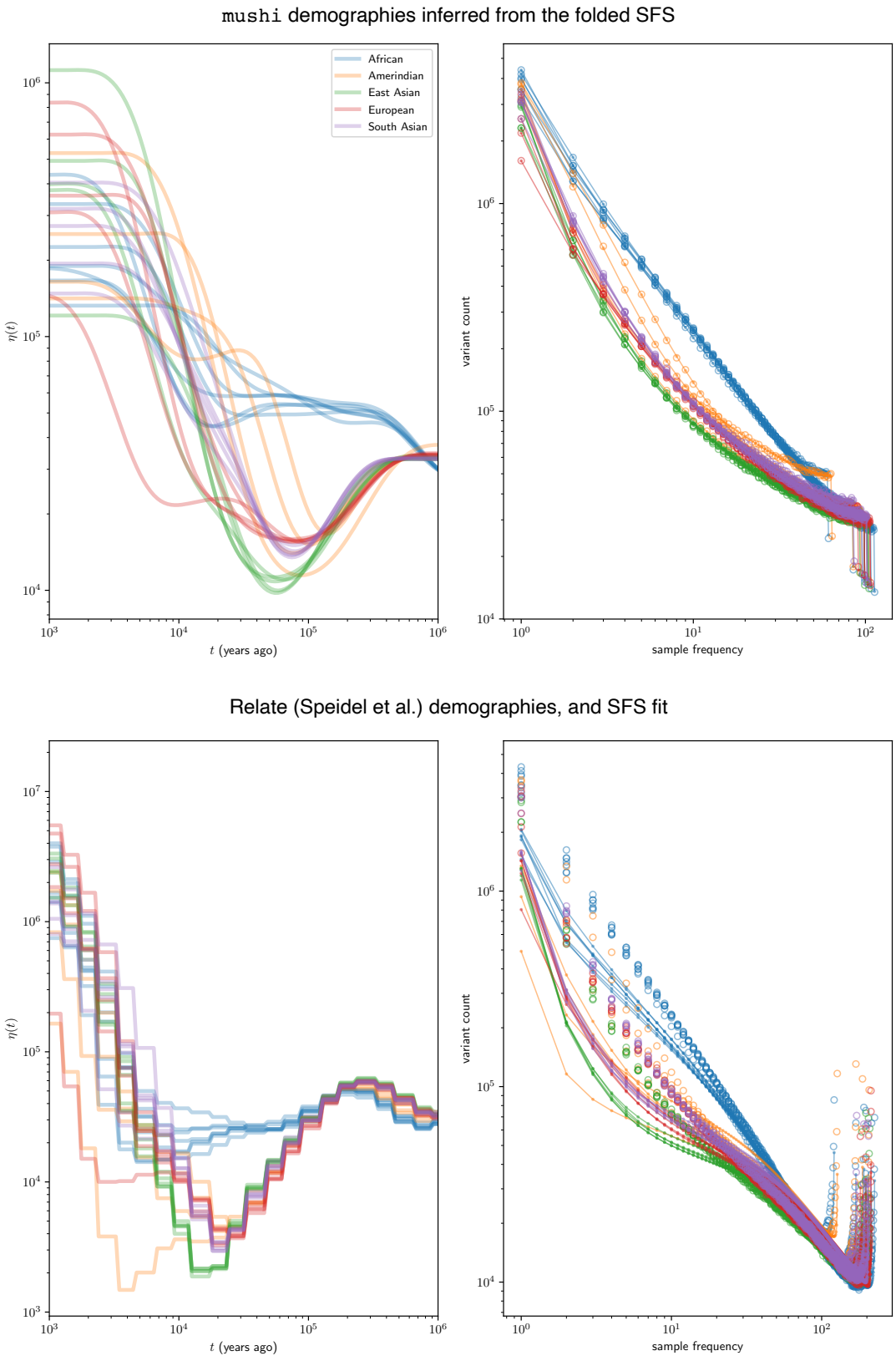

Fig. S4. Other effective population size histories for 1000 Genomes Project populations. Top panels show demographic histories estimated with `mushi` from high coverage data using the folded SFS, and the fit to the folded SFS, with no ancestral state polarization. Bottom panels show demographic histories inferred with Relate, as reported by Speidel et al. (9) (Fig S6 there), with the SFS fit displayed.

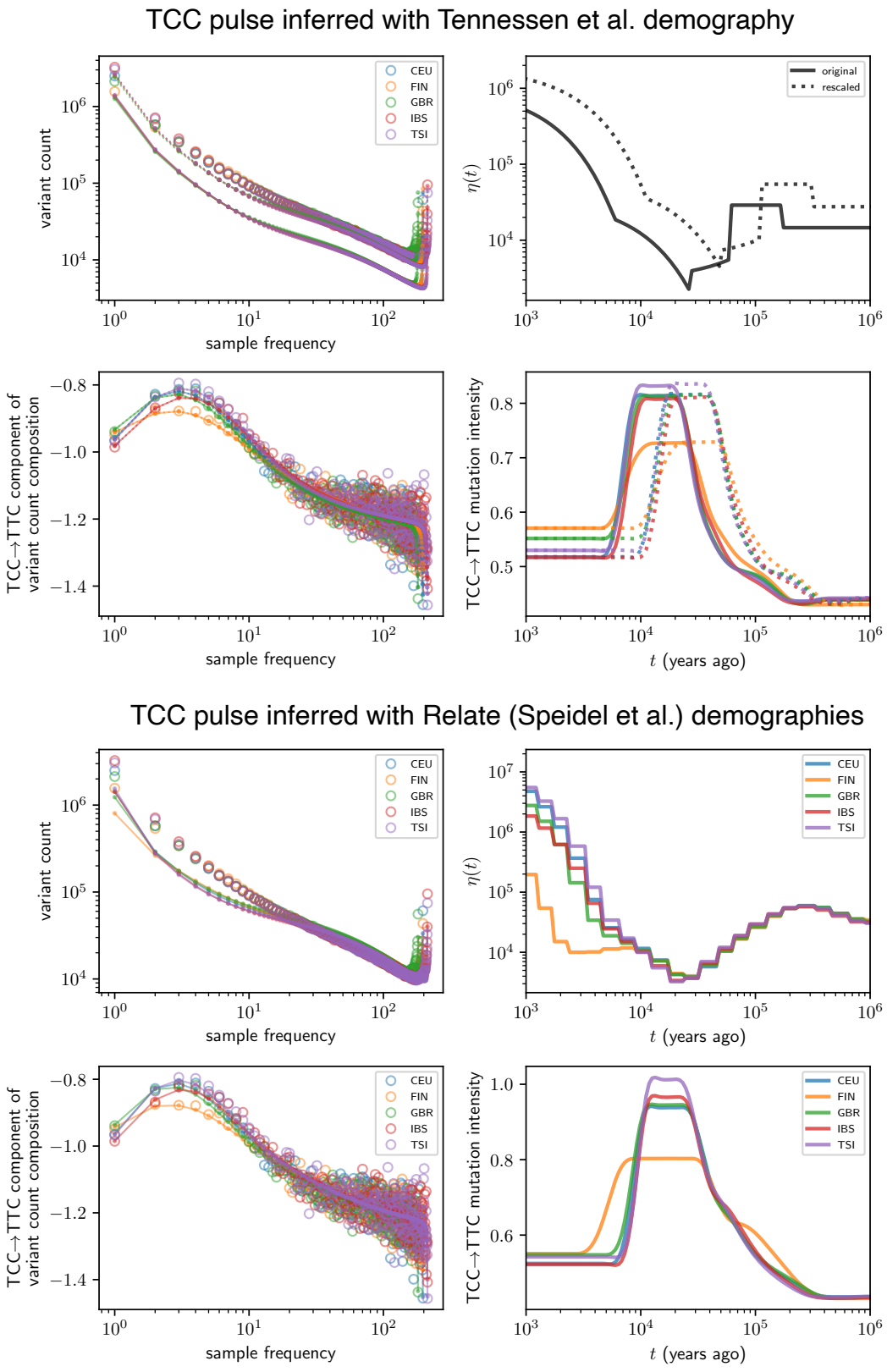

Fig. S5. Timing of TCC→TTC pulse in European populations conditioned on different demographies. Top panels show the SFS fit for each European population under the demographic history of Tennessen et al. (10), which was used by Harris and Pritchard to time the TCC→TTC pulse (11) (similar to Figure 2g), and the subsequent `mushi` fit to the TCC→TTC component of the k-SFS (similar to Figure 3). Dotted lines use a rescaled demography that accounts for the mutation rate difference between the Tennessen demographic inference and more recent de novo rate, resulting in TCC→TTC pulse that is shifted older, and better SFS fit. Bottom panels are similar, but use the European demographic histories inferred by Speidel et al. (9) using the Relate method.

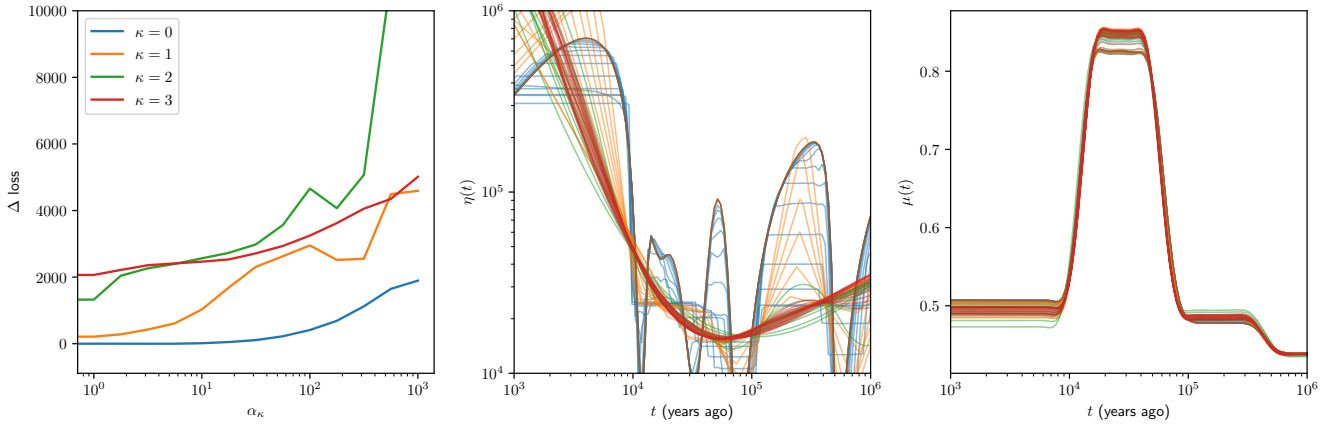

Fig. S6. The effect of demographic model selection on TCC→TTC pulse inference for CEU. Left panel shows change in loss (goodness of fit) as trend penalties of various order increase. Middle panel shows the inferred demography at each penalty value (colors corresponding to trend orders in left panel). Right panel shows the TCC→TTC component of subsequent MuSH inference for each demographic model.

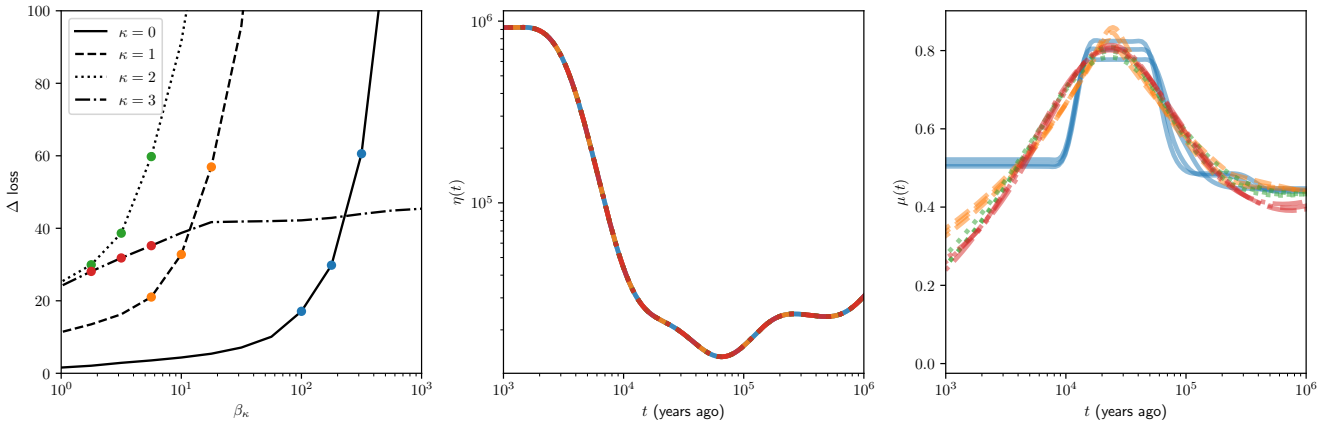

Fig. S7. The effect of MuSH regularization on TCC→TTC pulse inference for CEU. Left panel shows change in loss (goodness of fit) as the κ-th order trend penalty increases, for $\kappa = 0, 1, 2, 3$. Middle panel shows the inferred demography (which is independent of the MuSH hyperparameters). Right panel shows the TCC→TTC component of subsequent MuSH inference, with color corresponding to points in left panel.

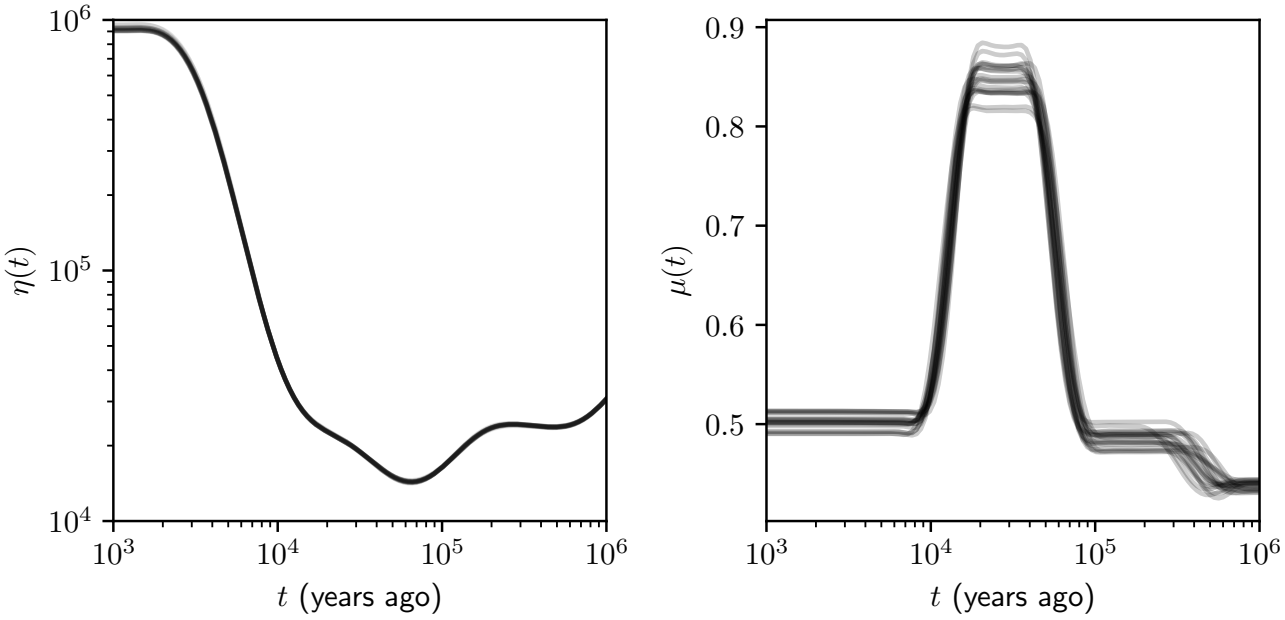

Fig. S8. Bootstrap for CEU demography and TCC→TTC pulse inference. This indicates the stability of inference under replicate data. Note that, because the penalized likelihood inference is strongly biased, this cannot be interpreted as providing confidence bounds of the histories.

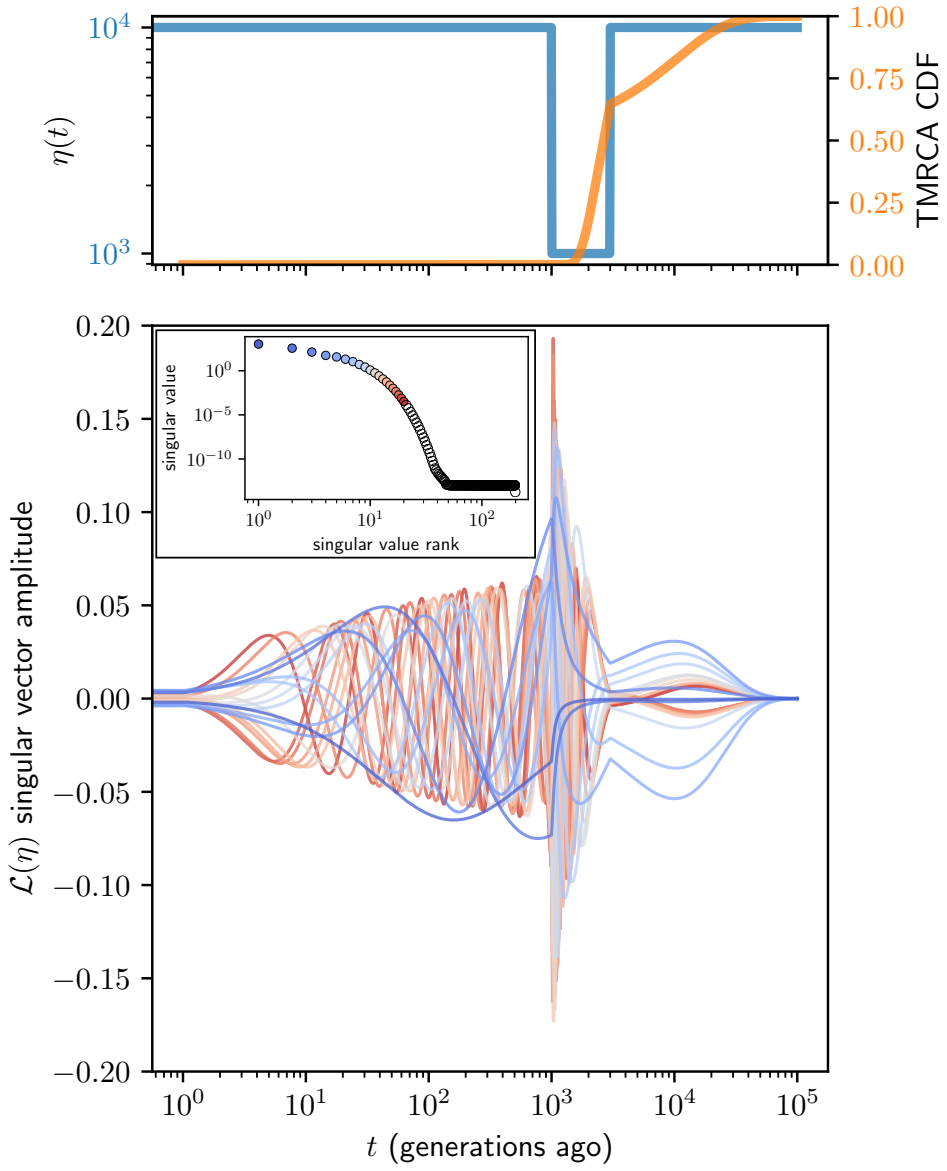

Fig. S9. Observability of mutation rate history via spectral analysis of $\mathcal{L}(\eta)$ for the case of a bottleneck history. The top panel plots demographic history with a bottleneck from about 3000 to 1000 generations ago (blue, left ordinate), and TMRCA CDF (orange, right ordinate). The bottom panel plots the top 20 time domain singular vectors, with the inset showing the corresponding ranked singular values. Time was discretized with a logarithmic grid of 1000 points, and $n = 200$ sampled haplotypes were assumed.

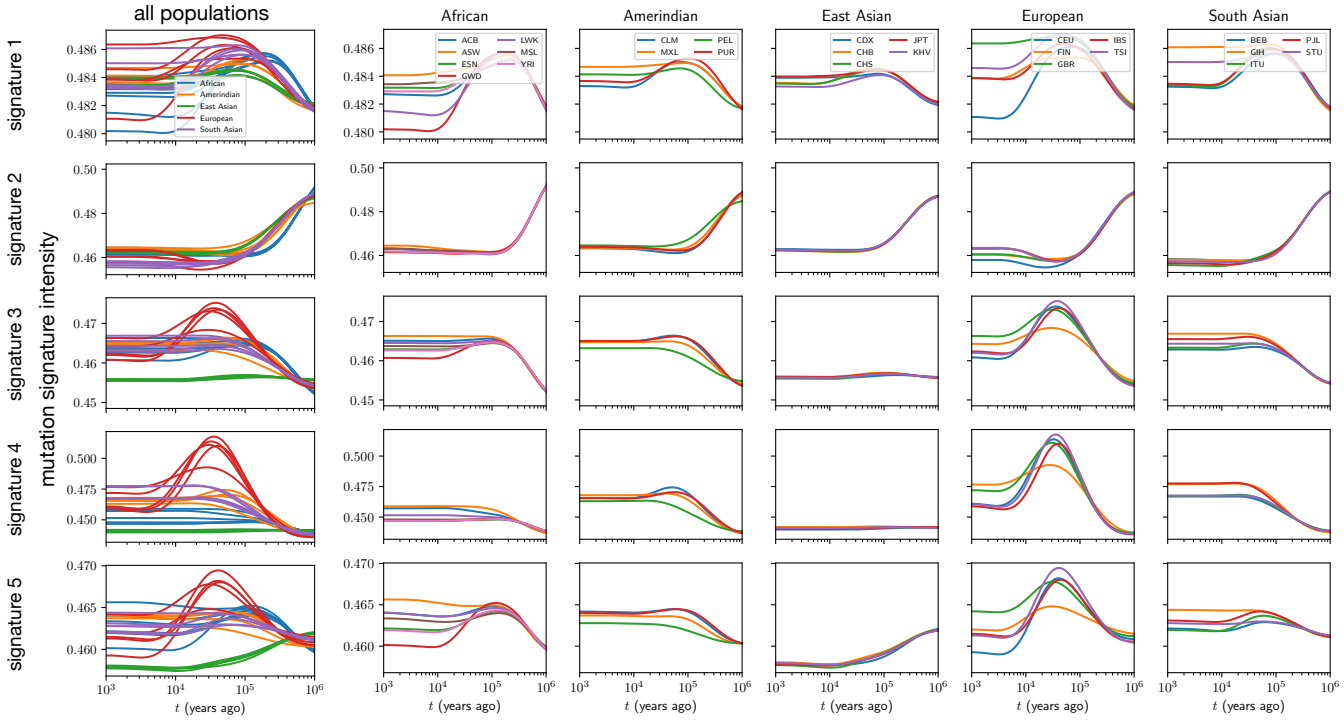

Fig. S10. **Mutation signature history for each 1000 Genome Projection super-population.** Same results as in main text, but plotted here separately for each super-population.

References

1. RC Griffiths, S Tavaré, The age of a mutation in a general coalescent tree. *Stoch. Model.* (1998).
2. A Polanski, A Bobrowski, M Kimmel, A note on distributions of times to coalescence, under time-dependent population size. *Theor. Popul. Biol.* **63**, 33–40 (2003).
3. A Polanski, M Kimmel, New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165**, 427–436 (2003).
4. M Petkovšek, HS Wilf, D Zeilberger, A= b, ak peters ltd. *Wellesley, MA* **30** (1996).
5. P Paule, M Schorn, A mathematica version of zeilberger's algorithm for proving binomial coefficient identities. *J. symbolic computation* **20**, 673–698 (1995).
6. Z Rosen, A Bhaskar, S Roch, YS Song, Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics*, genetics.300733.2018 (2018).
7. A DasGupta, *Probability for statistics and machine learning : fundamentals and advanced topics*, Springer texts in statistics. (Springer, New York), (2011).
8. O Sargsyan, J Wakeley, A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor. Popul. Biol.* **74**, 104–114 (2008).
9. L Speidel, M Forest, S Shi, SR Myers, A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
10. JA Tennessen, et al., Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
11. K Harris, JK Pritchard, Rapid evolution of the human mutation spectrum. *Elife* **6** (2017).