Supplemental Material

Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data

Anand Bhaskar[1,3], Y.X. Rachel Wang[2], and Yun S. Song[2,3,4,*]

[1]Simons Institute for the Theory of Computing, Berkeley, CA 94720
[2]Department of Statistics, University of California, Berkeley, CA 94720
[3]Computer Science Division, University of California, Berkeley, CA 94720
[4]Department of Integrative Biology, University of California, Berkeley, CA 94720
[*]E-mail: yss@eecs.berkeley.edu

1 Computing the expected SFS under a variable population size

In this section, we describe the details of computing the quantities $\mathbb{E}[\tau_{n,i}]$ and $\mathbb{E}[\tau_n]$, the expected branch length subtending i leaves and the expected total branch length respectively, when a coalescent tree is drawn over n individuals according to Kingman's coalescent with demographic model Φ.

Polanski and Kimmel (2003) showed that $\mathbb{E}[\tau_{n,i}]$ and $\mathbb{E}[\tau_n]$ can be computed efficiently and numerically stably using the relations,

$$\mathbb{E}[\tau_{n,i}] = \sum_{m=2}^{n} W_{n,i,m} c_m, \tag{1}$$

$$\mathbb{E}[\tau_n] = \sum_{m=2}^{n} V_{n,m} c_m, \tag{2}$$

where the coefficients $V_{n,m}$ are given by (Polanski and Kimmel 2003, Equation 12)

$$V_{n,m} = (2m-1)\frac{n!(n-1)!}{(n+m-1)!(n-m)!}(1+(-1)^m), \quad 2 \le m \le n,$$

and the coefficients $W_{n,i,m}$ are efficiently computable by the recursions (Polanski and Kimmel 2003, Equations 13–15):

$$W_{n,i,2} = \frac{6}{n+1},$$

$$W_{n,i,3} = 30\frac{(n-2i)}{(n+1)(n+2)},$$

$$W_{n,i,m+2} = -\frac{(1+m)(3+2m)(n-m)}{m(2m-1)(n+m+1)} W_{n,i,m} + \frac{(3+2m)(n-2i)}{m(n+m+1)} W_{n,i,m+1}, \quad 2 \le m \le n-2.$$

The coefficients c_m in (1) and (2) are given by the integral,

$$c_m = \int_0^\infty \exp\left[-\binom{m}{2} R(t)\right] \, dt. \tag{3}$$

Assuming $R(t) = \omega(\ln t)$ (i.e. $R(t)$ grows asymptotically faster than $\ln t$), the coefficients c_m are the expected first coalescence times for a sample of size m drawn at the present time from a population with demographic model Φ, and $R(t)$ is a time-rescaling function for the coalescent process, given by the expression,

$$R(t) = \int_0^t \frac{N_r}{N(\tau)}\, d\tau.$$

In this work, we consider the family of piecewise-exponential population size functions with M pieces. Any population size function in this family of functions can be described by $M-1$ time points, $0 < t_1 < \cdots < t_{M-1} < \infty$, where the effective population size at time $t \in [t_i, t_{i+1})$, $0 \le i \le M-1$, is given by $N(t) = N(t_i)\exp(-\beta_i(t-t_i))$. For notational consistency, we define $t_0 = 0$ and $t_M = \infty$. The times t_i are in units of N_r generations. In the piece corresponding to time interval $[t_i, t_{i+1})$, exponential population growth (decline) is encoded by $\beta_i > 0$ ($\beta_i < 0$), while $\beta_i = 0$ represents a constant population. For this family of population models, we can compute the integrals in (3) as follows. For $t \in [t_i, t_{i+1})$, $0 \le i \le M-1$,

$$R(t) - R(t_i) = \mathbf{1}[\beta_i = 0]\frac{N_r}{N(t_i)}(t-t_i) + \mathbf{1}[\beta_i \neq 0]\frac{1}{\beta_i}\frac{N_r}{N(t_i)}(\exp(\beta_i(t-t_i)) - 1)$$

$$\begin{aligned}
c_m &= \sum_{i=0}^{M-1} \int_{t_i}^{t_{i+1}} \exp\left(-\binom{m}{2} R(t)\right)\, dt \\
&= \sum_{i=0}^{M-1} \exp\left(-\binom{m}{2} R(t_i)\right) \left\{ \int_{t_i}^{t_{i+1}} \exp\left(-\binom{m}{2}(R(t) - R(t_i))\right)\, dt \right\} \\
&= \sum_{i=0}^{M-1} \exp\left(-\binom{m}{2} R(t_i)\right) \left\{ \mathbf{1}[\beta_i = 0] \int_{t_i}^{t_{i+1}} \exp\left(-\binom{m}{2}\frac{N_r}{N(t_i)}(t-t_i)\right)\, dt + \right. \\
&\qquad \left. \mathbf{1}[\beta_i \neq 0] \int_{t_i}^{t_{i+1}} \exp\left(-\binom{m}{2}\frac{1}{\beta_i}\frac{N_r}{N(t_i)}(\exp(\beta_i(t-t_i)) - 1)\right) \right\} \\
c_m &= \sum_{i=0}^{M-1} \mathbf{1}[\beta_i = 0]\frac{1}{\binom{m}{2}}\frac{N(t_i)}{N_r}\left\{\exp\left(-\binom{m}{2}R(t_i)\right) - \exp\left(-\binom{m}{2}R(t_{i+1})\right)\right\} \\
&\quad + \sum_{i=0}^{M-1} \mathbf{1}[\beta_i \neq 0]\frac{1}{\beta_i}\exp\left(-\binom{m}{2}R(t_i)\right)\exp\left(\frac{1}{\beta_i}\binom{m}{2}\frac{N_r}{N(t_i)}\right) \times \\
&\qquad \left\{\mathrm{Ei}\left(-\binom{m}{2}\frac{N_r}{N(t_i)}\frac{\exp(\beta_i(t_{i+1}-t_i))}{\beta_i}\right) - \mathrm{Ei}\left(-\binom{m}{2}\frac{N_r}{N(t_i)}\frac{1}{\beta_i}\right)\right\}, \qquad (4)
\end{aligned}$$

where $\mathrm{Ei}(x)$ is the exponential integral special function, given by,

$$\mathrm{Ei}(x) = -\int_{-x}^{\infty} \frac{e^{-t}}{t}\, dt.$$

Equation (4) can be further simplified to get

$$c_m = \sum_{i=0}^{M-1} \mathbf{1}[\beta_i = 0]\frac{1}{\binom{m}{2}}\frac{N(t_i)}{N_r}\left\{\exp\left(-\binom{m}{2}R(t_i)\right) - \exp\left(-\binom{m}{2}R(t_{i+1})\right)\right\} +$$

$$\sum_{i=0}^{M-1} \mathbf{1}[\beta_i \neq 0] \frac{1}{\beta_i} \left\{ \exp\left(-\binom{m}{2} R(t_{i+1})\right) \exp\left(\binom{m}{2} \frac{N_r}{N(t_i)} \frac{\exp(\beta_i(t_{i+1}-t_i))}{\beta_i}\right) \times \right.$$
$$\left. \text{Ei}\left(-\binom{m}{2} \frac{N_r}{N(t_i)} \frac{\exp(\beta_i(t_{i+1}-t_i))}{\beta_i}\right) \right.$$
$$\left. - \exp\left(-\binom{m}{2} R(t_i)\right) \exp\left(\binom{m}{2} \frac{N_r}{N(t_i)} \frac{1}{\beta_i}\right) \text{Ei}\left(-\binom{m}{2} \frac{N_r}{N(t_i)} \frac{1}{\beta_i}\right) \right\}. \quad (5)$$

We evaluate terms of the form $\exp(x)\text{Ei}(-x)$ that appear in (5) for large values of x using the following asymptotic expansion,

$$\exp(x)\text{Ei}(-x) = -\frac{1}{x} \sum_{k=0}^{\infty} (-1)^k \frac{k!}{x^k}. \quad (6)$$

For the results described in the main text, we truncated the divergent expansion (6) after 10 terms for $x \geq 45$.

2 Supplementary figures

Figures S1-S5 are supplementary figures referred to in the main text.

3 Impact of the SFS binning procedure on inference accuracy

For the results reported in the main text, when computing the likelihood of a given demographic model, we used the first k entries of the observed SFS which account for a fraction $f = 90\%$ of the segregating sites in the observed data while collapsing the remaining $n - k - 1$ SFS entries into one class. To test the sensitivity of our inference procedure to this fraction f, we simulated 100 datasets under Scenario 1 with a growth duration of 100 generations and per-generation growth rate of 6.4%. We ran our inference algorithm with several different values of f (Figure S6). As expected, using a larger fraction f of entries of the SFS in the inference procedure leads to less variance in the parameter estimates. However, the reduction in this variance quickly tapers off with this fraction f. As seen in Figure S6, the violin plot for the inferred parameters look very similar for $f = 96\%$ and $f = 97\%$. At the same time, the number of leading entries of the SFS which account for a given fraction of the total number of segregating sites rapidly increases with this fraction f. In particular, while the first 41 entries of the SFS account for 96% of the segregating sites, an additional 130 entries are needed to account for 97% of the segregating sites. In general, it seems that one can trade inference accuracy for computational cost, and more research is needed to theoretically characterize the informativeness of the individual entries of the SFS for demographic inference.

4 Impact of errors in the SFS on inference accuracy

In large sample sizes sequenced at low coverage, various biases and errors in the genotype calling procedure can lead to low frequency segregating variants being missed. Similarly, sequencing errors can cause monomorphic segregating sites to be called as polymorphic. More generally, various known and unknown sources of error can lead to an SFS that differs from the true SFS of the sample under consideration, and it would be interesting to understand the effect of such errors on

demographic inference (both for our specific method, and for an arbitrary algorithm). While we do not have a rigorous way to analyze this dependence of demographic inference on uncertainty in the SFS, we empirically examine the effect of errors on our demographic estimates for one of the scenarios considered in the main text. Since the proportion of rare variants is strongly influenced by the parameters of exponential growth in the effective population size (Keinan and Clark 2012), we consider the effect of overestimating or underestimating the true number of singletons for SCENARIO 1 with a growth rate of 6.4% per gen for 100 gens. Using a mutation rate of 2.5×10^{-8} per bp per gen per haploid, we simulated 100 datasets of 100 loci of 10kb each according to SCENARIO 1 with a growth rate of 6.4% per gen for the most recent 100 gens. To simulate missed singleton variants, we removed a fraction of the singleton sites and added them to the set of monomorphic sites. Figure S7 shows the inferred demographic parameters as a function of the proportion of singleton variants that are missed. As the fraction of singletons missed increases, the inferred duration and rate of growth both decrease. To simulate the effect of incorrect singleton variant calls, we removed a fraction of the monomorphic sites and called them as singletons. Figure S8 shows the inferred demographic parameters as a function of the proportion of monomorphic variants that are misclassified as singletons, potentially due to sequencing errors. As this proportion of incorrectly called singletons increases, the inferred duration and rate of growth both increase due to the excess of singletons observed. These inferred parameters also depend on the mutation rate; a higher mutation rate will lead to less sites being monomorphic, and thus decrease the impact of such errors on the inferred demographic parameters.

5 Details of the optimization procedure

The parameter optimization for the results reported in the main text were performed using the non-linear optimization program `IPOPT` (Wächter and Biegler 2006). The gradient of the log-likelihood function with respect to the demographic parameters were calculated through automatic differentiation (AD) using the C++ library `ADOL-C` (Walther and Griewank 2012). A description of AD can be found in the main text. `IPOPT` can approximate the Hessian matrix of the Lagrangian of the optimization problem using the supplied gradient information and the L-BFGS algorithm. Hence, we do not explicitly compute the exact Hessian matrix even though this is possible in principle using AD. To deal with local optima in the likelihood landscape, we performed the optimization for each of the simulated demographic models in SCENARIOS 1 and 2 by starting the optimization from 10 randomly chosen initialization points for the demographic parameters for each dataset. We found that for most of the datasets, the inferred demographic models did not depend on the starting point, suggesting that these likelihood landscapes have a unique local optimum. For the neutral regions dataset of Gazave et al. (2014) and the exome-sequencing dataset of Nelson et al. (2012), we performed the optimization with 100 random initialization points for the demographic parameters. For these datasets as well, most values of the initialization points resulted in the same inferred demographic parameters.

6 Comparison with the inference method of Gutenkunst *et al.*

Gutenkunst et al. (2009) developed a diffusion-based method, ∂`a`∂`i`, that can infer the joint demography of multiple subpopulations with changing population sizes and complex patterns of migration between subpopulations. ∂`a`∂`i` computes the expected SFS of a given demographic model by solving the Kolmogorov forward PDE for the density of derived mutations as a function of allele frequency and time. It solves this PDE using numerical methods that discretize the allele frequency and time

dimensions using suitably chosen grid points, and by performing extrapolation of an increasing sequence of discretization grid points. Excoffier et al. (2013) performed extensive comparisons of their coalescent simulation based method, `fastsimcoal`, with ∂a∂i for several demographic scenarios. Here, we briefly compare the efficiency and inference accuracy of ∂a∂i with our method for very large samples using several simulated datasets generated under SCENARIO 1 where the population size grew exponentially at a rate of 6.4% per generation for the last 100 generations (see Figure 1A in the main text). Averaged over 10 simulated datasets with 20,000 haploid individuals each, ∂a∂i estimated a mean exponential growth rate of 7% per generation lasting for about 70 generations. We used ∂a∂i with 500 discretization grid points for these comparisons, and the average runtime on these datasets was 80 CPU minutes (compared to an average of about 1 CPU minute for our method). Using ∂a∂i with 20,000 discretization grid points took almost 3 CPU days and did not improve the inference accuracy. To investigate the above bias in parameter inference, we used ∂a∂i to compute the expected SFS of the true demographic model in SCENARIO 1 for samples of size 20, 200, 2,000, and 20,000 haploids. Figure S9 shows the first few leading entries of the expected SFS for the above demographic model, computed in three ways — using ∂a∂i, using our method, and using coalescent simulations with the `ms` program. The expected SFS computed by our method agrees with the empirical SFS from `ms` simulations very closely. On the other hand, ∂a∂i seems to err in computing the number of singletons and doubletons, with this discrepancy getting worse as the sample size increases. For these expected SFS computations, we used ∂a∂i with 200,000, 300,000, and 400,000 grid points, and applied ∂a∂i's extrapolation procedure to estimate the entries of the SFS as the number of grid points goes to infinity.

References

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M., 2013. Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**(10):e1003905.

Gazave, E., Ma, L., Chang, D., Coventry, A., Gao, F., Muzny, D., Boerwinkle, E., Gibbs, R. A., Sing, C. F., Clark, A. G., *et al.*, 2014. Neutral genomic regions refine models of recent rapid human population growth. *Proceedings of the National Academy of Sciences*, **111**(2):757–762.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**(10):e1000695.

Keinan, A. and Clark, A. G., 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, **336**(6082):740–743.

Keinan, A., Mullikin, J. C., Patterson, N., and Reich, D., 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics*, **39**(10):1251–1255.

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., Jean, P. S., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., *et al.*, 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, **337**(6090):100–104.

Polanski, A. and Kimmel, M., 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, **165**(1):427–436.

Schaffner, S., Foo, C., Gabriel, S., Reich, D., Daly, M., and Altshuler, D., 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, **15**(11):1576–1583.

Wächter, A. and Biegler, L. T., 2006. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, **106**(1):25–57.

Walther, A. and Griewank, A., 2012. Getting started with ADOL-C. In und O. Schenk, U. N., editor, *Combinatorial Scientific Computing*, pages 181–202. Chapman-Hall CRC Computational Science.

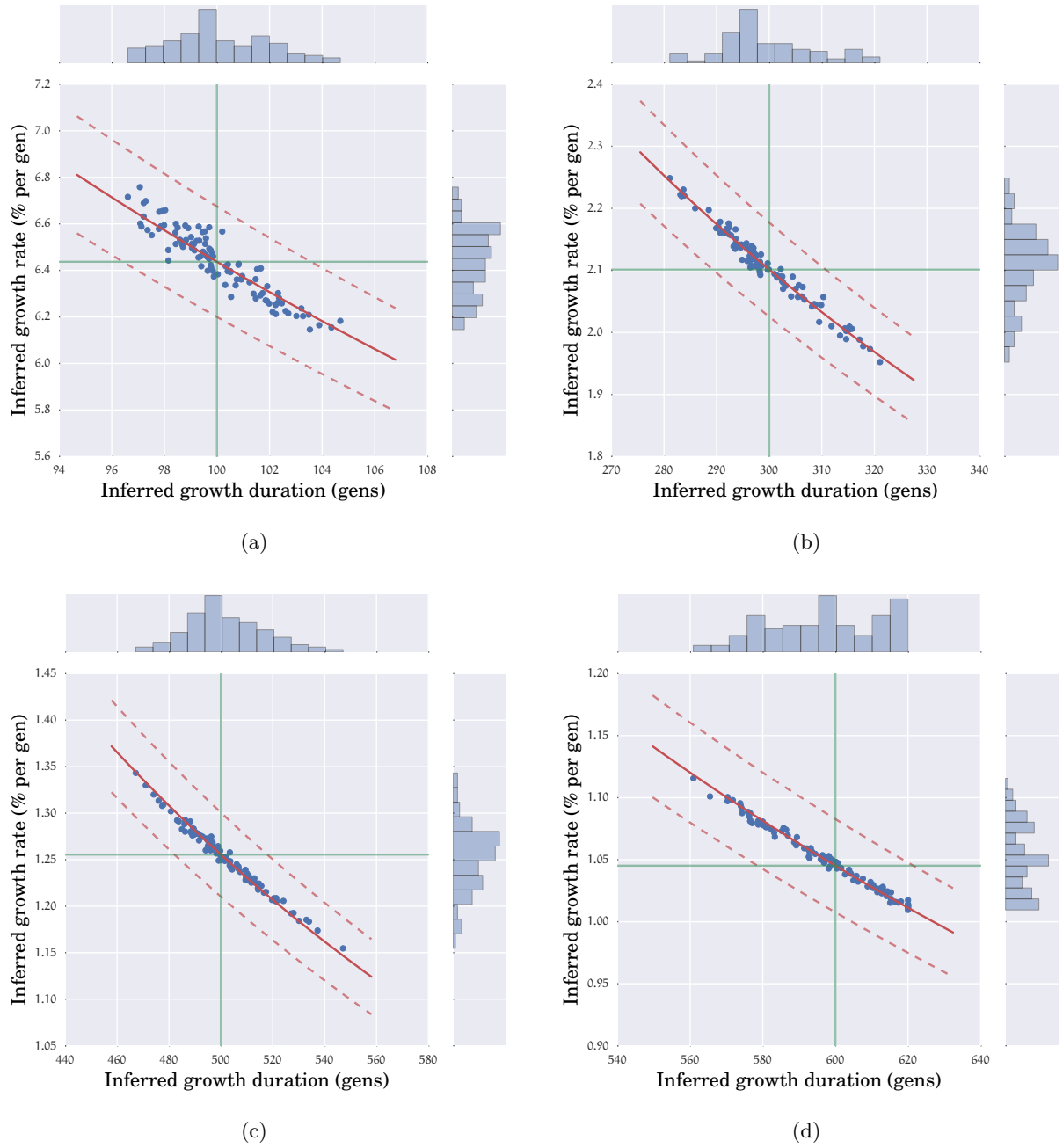
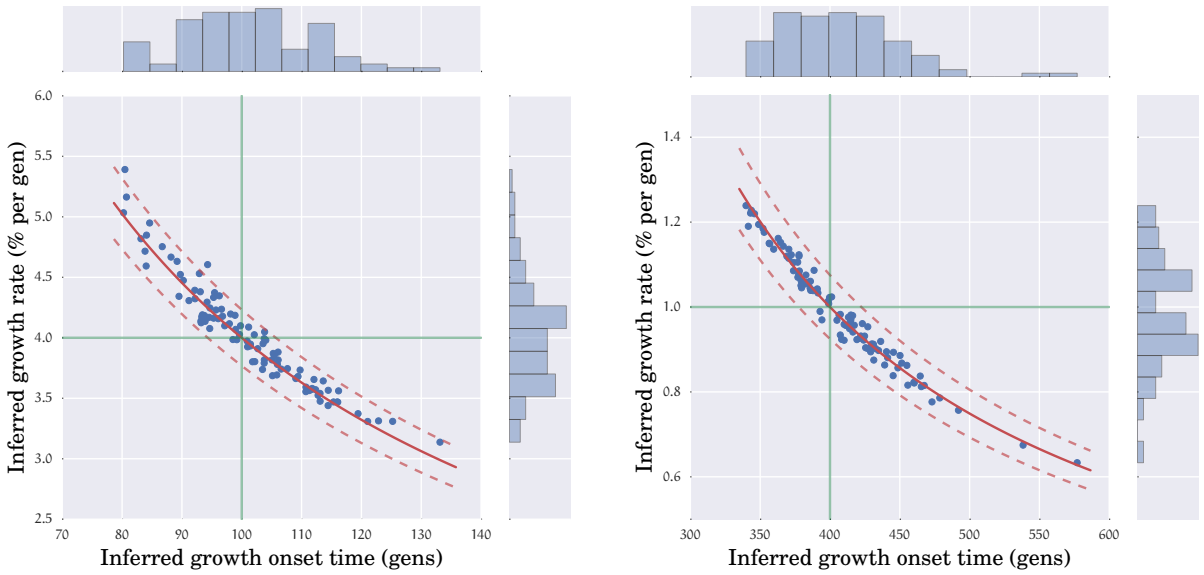


Figure S1: Joint and marginal distributions of the inferred duration and rate of exponential growth for several parameter settings in SCENARIO 1. Each point in each plot represents a simulated dataset with 100 unlinked loci of length 10 kb each over 10,000 diploid individuals. The green vertical and horizontal lines denote the true simulation parameter values for the duration and rate of exponential population growth. In SCENARIO 1, the present population size is 512 times larger than the ancestral population size. The solid red curves (dashed red curves) in each plot are the combinations of parameter values which have this same ratio (resp., higher and lower by 25%) of present to ancestral population size. (a) $t_1 = 100$ gens, $r_1 = 6.44\%$ per gen, (b) $t_1 = 300$ gens, $r_1 = 2.10\%$ per gen, (c) $t_1 = 500$ gens, $r_1 = 1.26\%$ per gen, (d) $t_1 = 600$ gens, $r_1 = 1.05\%$ per gen.



(a) (b)

Figure S2: Joint distribution of the inferred growth onset time and growth rate in Scenario 2 for (a) epoch 1 with $t_1 = 100$ gens and $r_1 = 4\%$ per gen, and for (b) epoch 2 with $t_2 = 400$ gens and $r_2 = 1\%$ per gen. These plots were generated using 100 simulated datasets with 100 unlinked loci of 10 kb each over 10,000 diploid individuals. The green lines indicate the true values for the simulation parameters. In Scenario 2, the population expands about 50-fold and 20-fold in epochs 1 and 2, respectively. The solid red curves are the combinations of parameter values which have these same population expansion factors in epochs 1 and 2, while the dashed red curves are the combinations of parameter values which are higher and lower by 25% compared to the true population expansion factors in each epoch. The inferred parameter combinations reflect the population expansion factors very well.

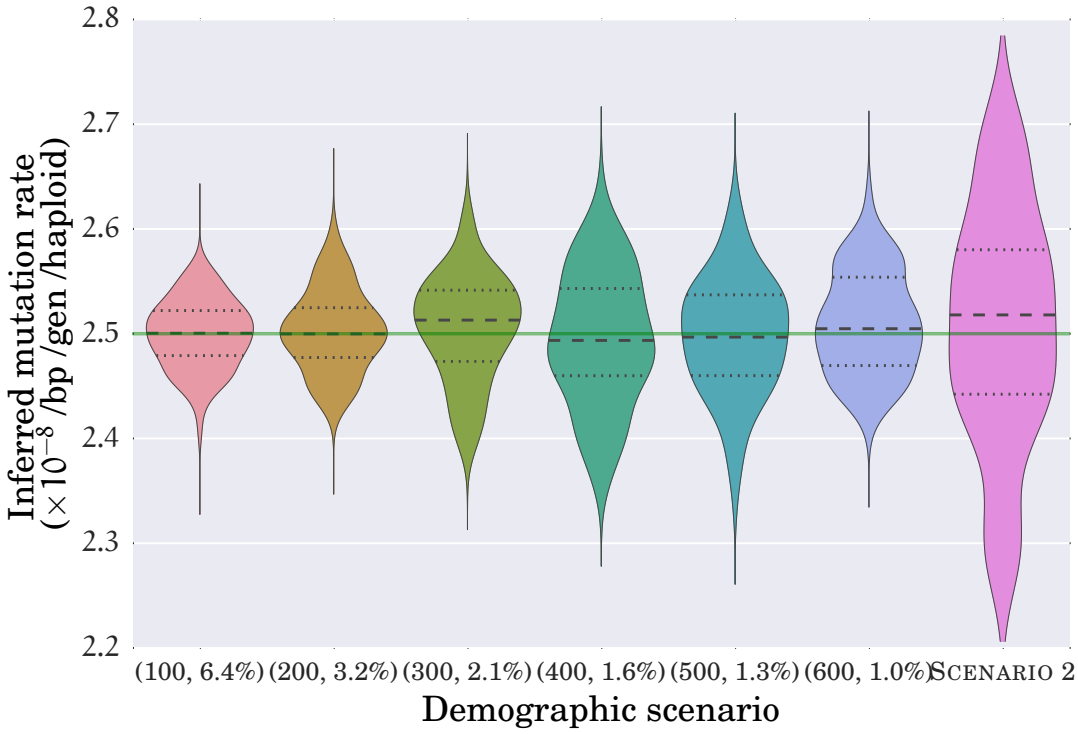


Figure S3: Violin plots of the inferred mutation rates for each of the six simulation parameter combinations of Scenario 1, and for Scenario 2. Each plot represents 100 simulated datasets with 100 unlinked loci of 10 kb each over 10,000 diploid individuals. All loci were simulated using a mutation rate of 2.5×10^{-8} per bp per gen per haploid. The uncertainty in the inferred mutation rate is significantly higher for Scenario 2 due to the higher uncertainty in the demographic parameters being simultaneously estimated (see Figures 2C and 2D in the main text).

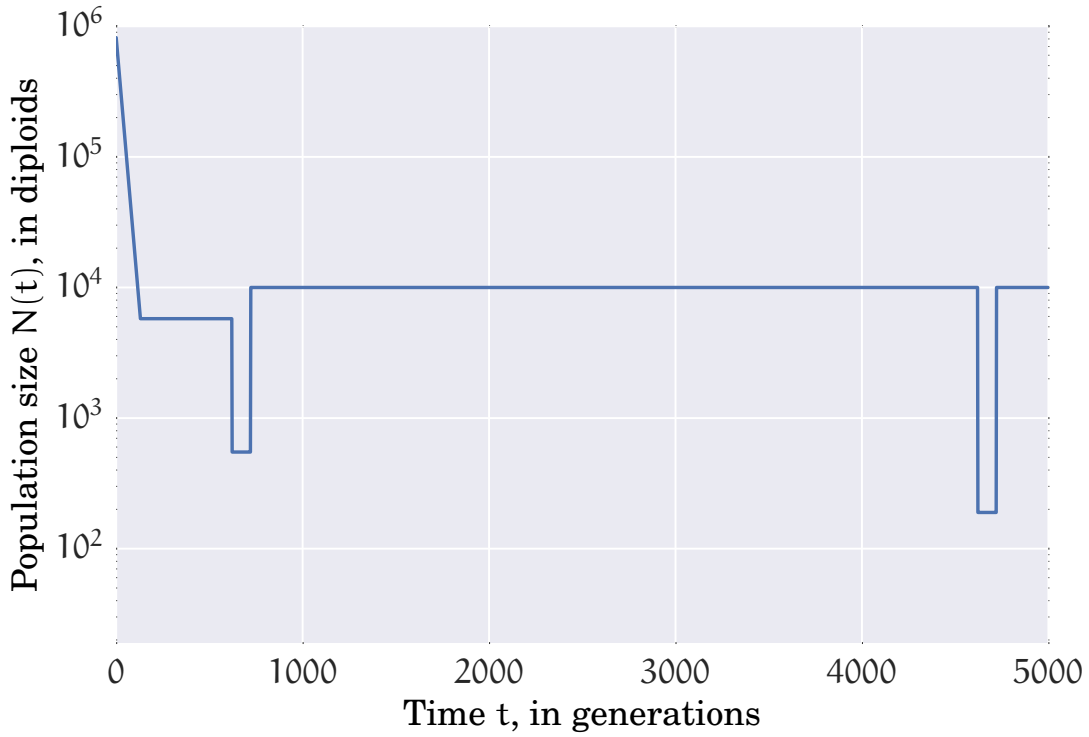


Figure S4: Demographic model inferred by our program on the neutral regions dataset of Gazave et al. (2014). We inferred a model with three parameters: the rate and onset of recent exponential growth, and the population size just before the onset of exponential growth. The ancient bottlenecks were fixed to those inferred by Keinan et al. (2007).

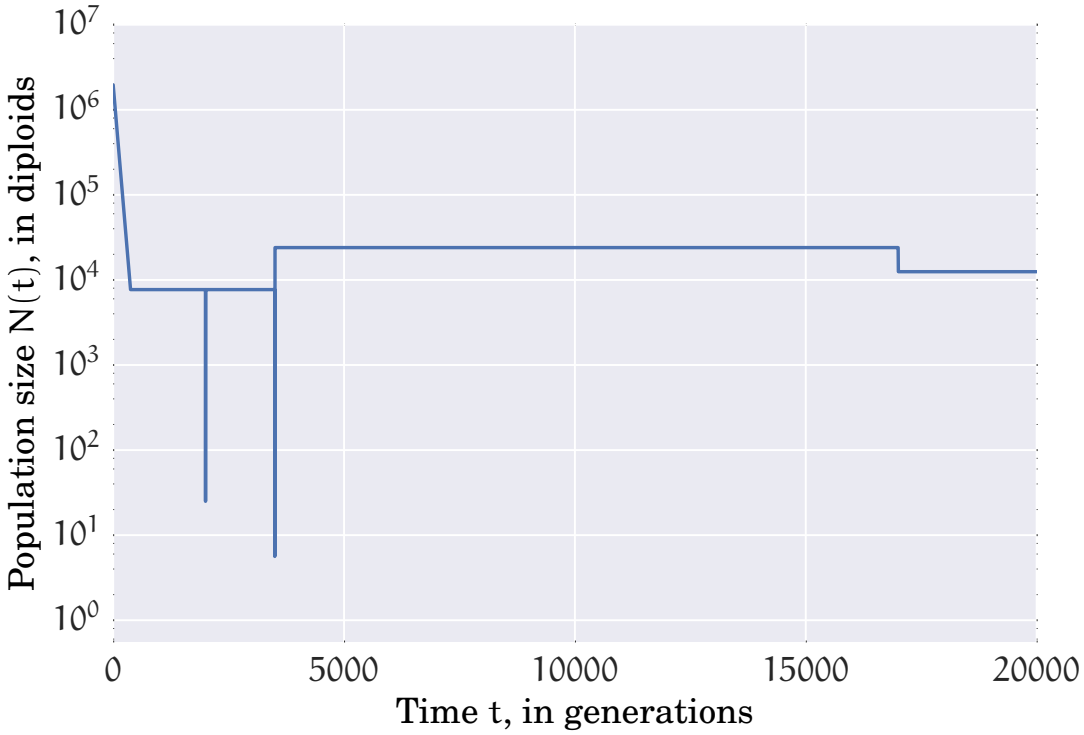
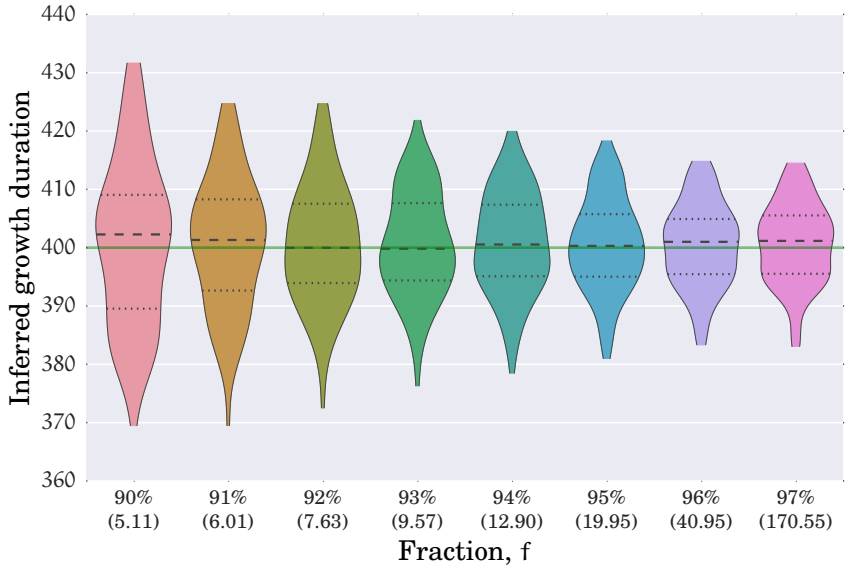
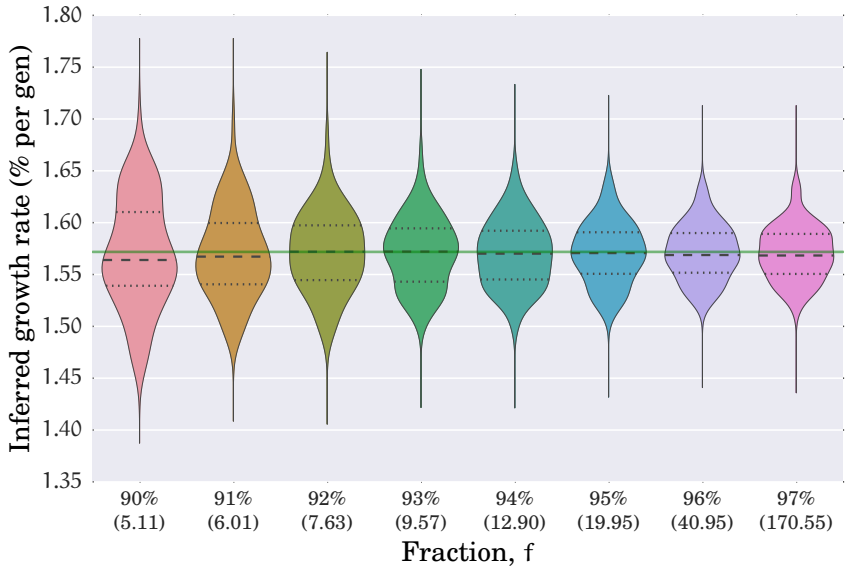


Figure S5: Demographic model of recent population expansion inferred by our program on the CEU exome-sequencing dataset of Nelson et al. (2012). The ancient demographic model before the epoch of exponential growth was fixed to that inferred by Schaffner et al. (2005). The two population bottlenecks 2,000 and 3,500 generations in the past correspond to the European-Asian population split and the out-of-Africa bottleneck, respectively. Our method infers an epoch of exponential population growth beginning 372 generations in the past resulting in the effective population size expanding from 7,700 individuals to 1.9 million individuals at a rate of 1.5% per generation.

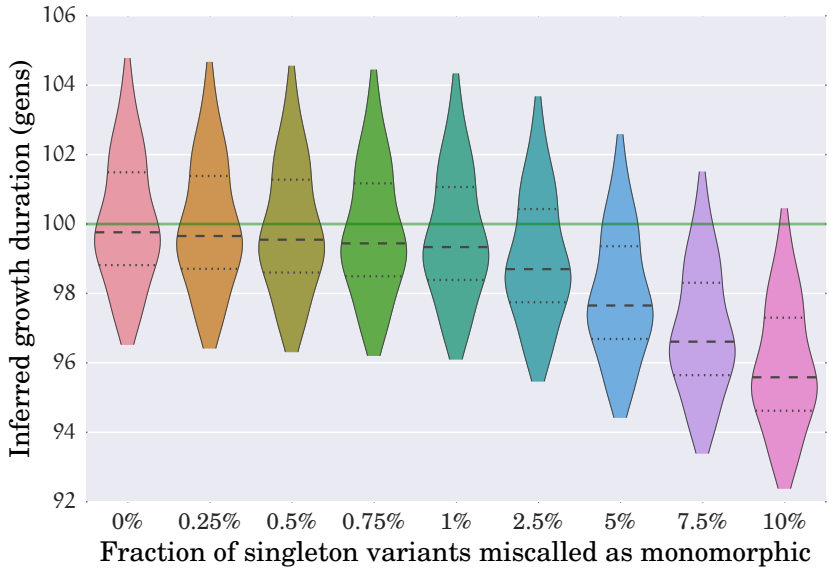


(a)

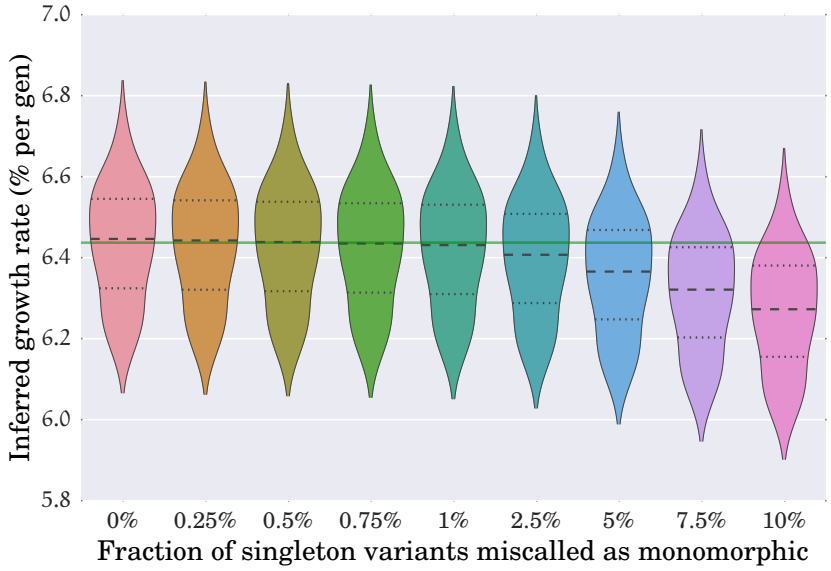


(b)

Figure S6: Violin plots of the (a) duration and (b) rate of exponential growth in the population size as a function of the fraction f of entries of SFS that are used in the inference procedure, for 100 simulated datasets of 100 loci of 10kb length under Scenario 1 with a growth duration of $t_1 = 400$ gens and a growth rate of $r_1 = 1.6\%$ per gen. The green lines indicate the true values for the simulation parameters. When computing the log-likelihood function, we used the first k entries of the SFS which account for an f fraction of the number of segregating sites, while collapsing the remaining $n - k - 1$ segregating sites into one class. The x-axis labels correspond to the fraction f used, with the numbers in parentheses giving the average value of k (over the simulation replicates) such that the leading k entries of the SFS account for an f fraction of the segregating sites.

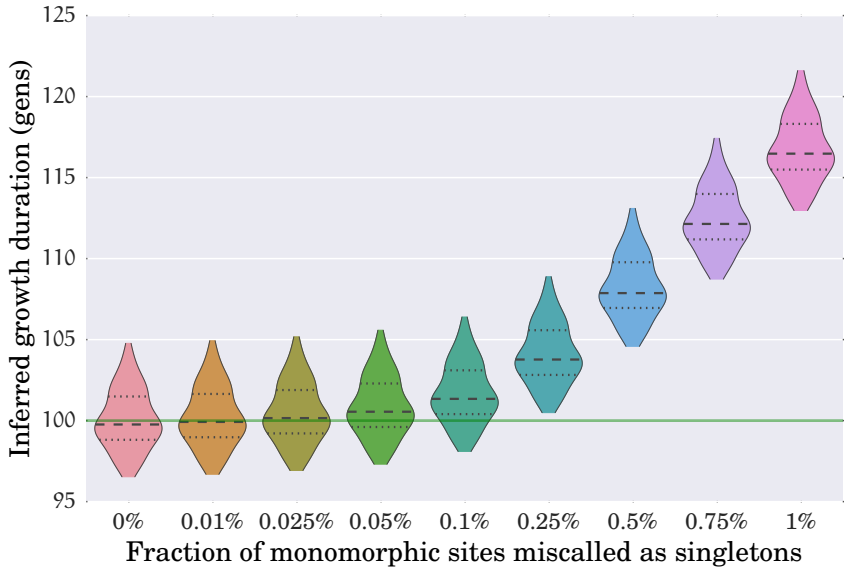


(a)

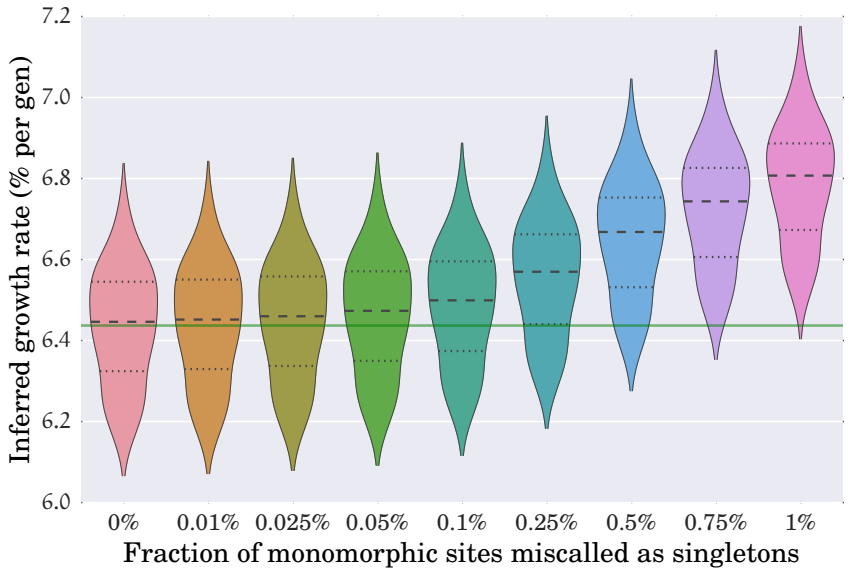


(b)

Figure S7: Violin plots of the (a) duration and (b) rate of exponential growth as a function of the fraction of singletons that are missed, when the data is simulated according to Scenario 1 with a growth rate of 6.4% per gen lasting for the most recent 100 generations. As the fraction of singletons missed increases, the rate and duration of inferred growth are smaller.



(a)



(b)

Figure S8: Violin plots of the (a) duration and (b) rate of exponential growth as a function of the fraction of monomorphic sites that are incorrectly classified as singletons, when 100 loci of 10kb each are simulated according to Scenario 1 with a growth rate of 6.4% per gen lasting for the most recent 100 generations, and a mutation rate of 2.5×10^{-8} per bp per gen. As the fraction of monomorphic sites incorrectly classified as singletons increases, the rate and duration of inferred growth are larger due to the perceived excess of singletons.

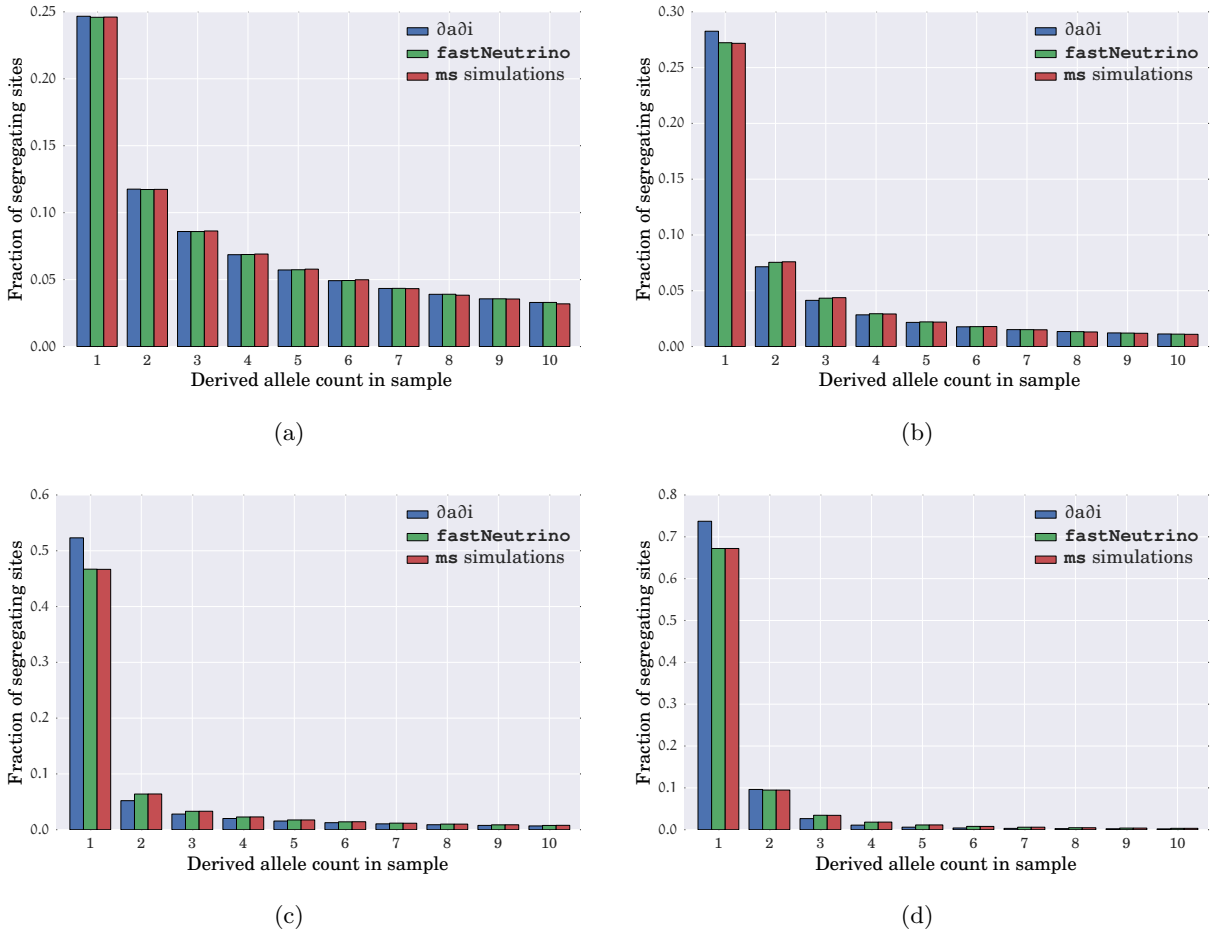


Figure S9: The first few entries of the expected SFS for a sample of (a) 20, (b) 200, (c) 2,000, and (d) 20,000 randomly sampled haploid individuals under the demographic model in SCENARIO 1, where the population has recently expanded for 100 generations at 6.4% per generation. The bars denote the entries of the SFS computed using the program ∂a∂i (Gutenkunst et al. 2009) (blue) and by our method `fastNeutrino` (green), and the empirical SFS from datasets simulated using the coalescent simulator `ms` (red). We used `ms` to simulate 10^6 unlinked loci of 100 bp each to empirically estimate the expected SFS. The entries of the expected SFS computed by our method agree with those from the `ms`simulations very closely. On the other hand, the proportion of singletons computed by ∂a∂i deviates significantly from these values for large sample sizes.