Accurate and flexible estimation of effective population size history

Zhendong Huang[1], Yao-ban Chan[1], and David Balding[1]

[1]Melbourne Integrative Genomics, School of Mathematics & Statistics, University of Melbourne, Australia

July 10, 2024

Abstract

Current methods for inferring demographic history from DNA sequences often impose a heavy computational burden, they can be marred by sequencing errors or uncertainty about recombination rates and the quality of inference is often poor in the recent past. We propose "InferNo" for flexible, nonparametric inference of effective population size. It requires modest computing resources and little prior knowledge of the recombination and mutation maps, and is robust to sequencing error and gene conversion. We illustrate the statistical and computational advantages of InferNo over previous approaches using a range of simulation scenarios. In particular, we demonstrate the ability of InferNo to exploit biobank-scale datasets for accurate inference of population size changes in the recent past. We then apply InferNo to worldwide human data, finding remarkable similarities in inferences from different populations in the same region. Unlike previous studies, we show at least two bottlenecks in the ancestral populations of most of the non-African sequences.

Keywords: effective population size, semi-analytical inference.

1 Introduction

Many methods have been developed to infer historic effective population sizes from population samples of genome sequences, which can provide clues about major events such as migrations, plagues and climatic extremes. Some approaches use the sequential Markov coalescent (SMC) model, including PSMC (Li and Durbin, 2011), MSMC (Schiffels and Durbin, 2014), diCal (Druet et al., 2014), and SMC++ (Terhorst et al., 2017). When the sample size is small, these approaches can provide good inferences for ancient population sizes, but they perform poorly over recent history and are computationally demanding for large sample sizes. Moreover, these approaches usually require good knowledge of the mutation and recombination rate maps across the genome to accurately specify the SMC, and they can be sensitive to sequencing errors.

A second class of approaches involves using the allele (or site) frequency spectrum (AFS, or SFS) of the genome sequences (Gutenkunst et al., 2009; Excoffier et al., 2013; Bhaskar et al., 2015; Kamm et al., 2020). Variation in mutation rate along the genome presents a particular difficulty for these methods, and they can be sensitive to sequencing errors. The AFS ignores the ordering of sites along the genome and so avoids any assumption about recombination, but the resulting loss of information leads to an ill-posed inverse problem (an AFS can arise from different demographic histories), especially for small sample sizes (DeWitt et al., 2021).

Other approaches are based on inference of the ancestral recombination graph (ARG) which summarises the genealogical history, or related summary statistics such as inferred identity by descent (IBD) relationships or linkage disequilibrium (LD) coefficients (Palamara et al., 2012; Boitard et al., 2016; Speidel et al., 2019; Fournier et al., 2023). However, inferring the ARG or IBD segments is difficult and imprecise, in part due to uncertainty about recombination and mutation rates. IBD and LD based methods typically focus on recent history (Mezzavilla et al., 2015; Browning and Browning, 2015; Ragsdale and Gravel, 2019; Santiago et al., 2020) and perform less well in the distant past. Computational performance is

another concern for many of these methods, despite advances in ARG inference (Rasmussen et al., 2014; Speidel et al., 2019; Kelleher et al., 2019; Mahmoudi et al., 2022; Deng et al., 2024).

To overcome these difficulties, we propose "InferNo" to infer effective population size history. InferNo uses a single ARG simulated under the standard coalescent model to estimate the distribution of mutation events in time and genome intervals, which is compared with the AFS and pairwise site differences to infer the time scaling (inversely proportional to population size) that maps the standard coalescent onto a genealogy for the observed data. Features of InferNo:

- It is computationally and statistically efficient for small and large sample sizes and for recent and ancient times.
- The inference is nonparametric, requiring no pre-specified model for population size.
- Mutation rate: variation along the genome is accommodated, only a genome-wide average needs to be pre-specified.
- Recombination rate: an average rate is pre-specified, but InferNo is robust both to variation along the genome and misspecification of the average rate.
- It is robust to sequencing errors, requiring no information about the error rate.

In simulation studies under different population-size models, we find that InferNo is overall more accurate and computationally faster than the ARG-based Relate (Speidel et al., 2019), AFS-based Mushi (DeWitt et al., 2021), and SMC-based PSMC (Li and Durbin, 2011) and CHIMP (Upadhya and Steinrücken, 2022).

2 Methods

Our goal is to estimate $N(g)$, the effective haploid population size g generations ago, based on a sample C_i, $i = 1, \ldots, n$, of homologous genome sequences each of length ℓ. For $s = 1, \ldots, \ell$,

each $C_i(s)$ is either 0 (ancestral) or 1 (derived); ancestral allele states can often be accurately inferred from related species.

2.1 Inference of mutation rate

In the derivations below we assume at most one mutation since the most recent common ancestor (MRCA) at any site, but we demonstrate robustness to this assumption by not incorporating it into the data generation models of the simulation study. Site-specific per-generation rates $\mu(s)$, $s = 1, \ldots, \ell$, can be pre-specified if desired, otherwise $\mu(s)$ is estimated from the number of polymorphic sites in the neighbourhood of s, scaled so that the average rate across the genome is $\bar{\mu}$, which is required input. Specifically, we partition the sites into L^* intervals $J_1^*, \ldots, J_{L^*}^*$ each of length ℓ/L^* and estimate $\mu(s)$ by a piece-wise constant map fitted to n_l^*, the number of polymorphic sites in J_l^*:

$$\hat{\mu}(s) = \frac{\sum_{l=1}^{L^*} n_l^* \mathbf{1}(s \in J_l^*)}{\sum_{l=1}^{L^*} n_l^*} \bar{\mu} L^*, \tag{1}$$

where $\mathbf{1}()$ is the indicator function.

2.2 Inferring the time scaling using AFS

We assume the coalescent with recombination model, with coalescence rate $1/N(g)$ for each pair of lineages, where $N(g)$ is the population size g generations ago. A key idea underlying InferNo is that the variable-size model is equivalent to a constant-size model with time scaled in proportion to $1/N(g)$ (Griffiths and Tavaré, 1994). The mutation rate is assumed constant over time in the variable-size model, but the mapping onto the constant-size model makes it proportional to $N(g)$.

To exploit this relationship we simulate under a "null model" with constant population size $\tilde{N}$, sample size n, sequence length ℓ, and recombination rate $r = 10^{-8}$ per site per generation. Here we fix $\tilde{N} = 20\,000$, but other values can be chosen based on prior knowledge

about $N(g)$. For ℓ large given the value of r, there can be enough internal replication due to recombination that a single simulation suffices and is used in all analyses reported here, but multiple simulations of the null model can be performed, if desired for greater accuracy.

The mutation rate in the null model is $\mu(s), s = 1, \ldots, \ell$, either pre-specified or estimated using (1), and is constant over time. Inference about $N(g)$ proceeds by inferring a mapping from mutations in the null model to observed allele frequency data.

To this end we partition the time domain for the null model into intervals $\tilde{I}_t = [\tilde{g}_{t-1}, \tilde{g}_t)$, $t = 1, \ldots, T$, where $\tilde{g}_0 = 0$ and $\tilde{g}_T = \infty$. Throughout this paper we set the $\tilde{g}_t$ to approximately equalise the number of null-simulation mutations during each $\tilde{I}_t$, but they can also be pre-specified, for example, based on historical periods of interest or prior knowledge about $N(g)$. We approximate $N(g)$ by a piece-wise constant function $N(g) \approx N_t$ for $g \in I_t = [g_t, g_{t-1})$, where $g_0 = 0$, $g_T = \infty$ and I_t is defined so that the cumulative per-site coalescence rate over $\tilde{I}_t$ in the null model equals that over I_t in the variable-size model, i.e.,

$$\frac{\tilde{g}_t - \tilde{g}_{t-1}}{\tilde{N}} = \frac{g_t - g_{t-1}}{N_t}, \qquad t = 1, \ldots, T-1. \tag{2}$$

The inference task is now, given our choices for $\tilde{N}$ and the $\tilde{g}_t$, to estimate the N_t and g_t from mutation events inferred from the sequence data.

The AFS is a vector Y_1 with kth element the number of sites at which the derived allele has frequency k, for $k = 1, \ldots, n-1$. From a null simulation we construct $\tilde{X}_1$, an $(n-1) \times T$ matrix with $\{k, t\}$ element equal to the number of sites with a frequency-k derived allele where the mutation arose during the time interval $\tilde{I}_t$ (oldest mutation if there is more than one). Because mutations underlying the observed data are expected to arise during I_t in proportion to N_t, whereas expected mutations in the null model during $\tilde{I}_t$ are proportional to $\tilde{N}$, we have the regression predictor

$$\tilde{X}_1 \beta = Y_1, \tag{3}$$

where $\beta_t = N_t/\tilde{N}$, $t = 1, \ldots, T$. The inference task now becomes estimation of β, which can proceed from (3) but the resulting inference is poor for small n. We therefore seek more information about β by refining the AFS into sequence pair mismatches in genome intervals.

2.3 Localising mutations to sequence pairs and genome intervals

We partition the genome $[0, \ell)$ into intervals $J_l = [h_{l-1}, h_l)$, $J_1, \ldots, J_L$, such that approximately c mutations are expected per lineage per generation in each J_l, which is obtained by minimizing $|c - \sum_{s \in J_l} \hat{\mu}(s)|$ for $l = 1, \ldots, L$. Here we fix $c = 2.6 \times 10^{-4}$, but InferNo is robust to the choice of c, which controls the number of intervals L.

Let Z denote a length-nL vector with $\{L(i-1)+l\}$th element the number of site differences in J_l, $l = 1, \ldots, L$, between C_i and C_{i+1}, for $i = 1, \ldots, n$, defining $C_{n+1} = C_1$. We only use neighbouring pairs of sequences to reduce computational effort, so the resulting inferences can depend on sequence order. However, this effect is weak because every mutation is represented by a site difference in at least two of the neighbour-pairs.

We assume that each element z of Z is a Poisson$(2c\tau_l^i)$ random variable, conditional on τ_l^i, the average TMRCA of C_i and C_{i+1} in the interval J_l. Write f_τ and p_z for the probability density and mass functions of τ_l^i and z, then

$$p_z(m) = \int_0^\infty \mathbb{P}_{2ct}(m) f_\tau(t) \mathrm{d}t \approx \frac{1}{2c} \sum_{t=0}^{Z_{\max}} \mathbb{P}_t(m) f_\tau(t/2c), \tag{4}$$

for $m = 0, 1, 2, \ldots$, where $\mathbb{P}_a(b)$ is the probability mass function at b of the Poisson(a) distribution and $Z_{\max} = \max\{Z\}$. Omitting the low-probability event $z > Z_{\max}$, we obtain the matrix equation $p_z = A f_\tau / 2c$, where A is a matrix with $A_{i+1,j+1} = \mathbb{P}_j(i), i, j = 0, \ldots, Z_{\max}$ and the vector version of f_τ is estimated at $t/2c, t = 0, \ldots, Z_{\max}$. We approximate p_z by $\hat{p}_z$, with kth element

$$\hat{p}_z^k = \sum_{m=1}^{nL} \mathbf{1}(Z_m = k-1)/(nL),$$

and solve $\hat{f}_\tau = 2cA^{-1}\hat{p}_z$.

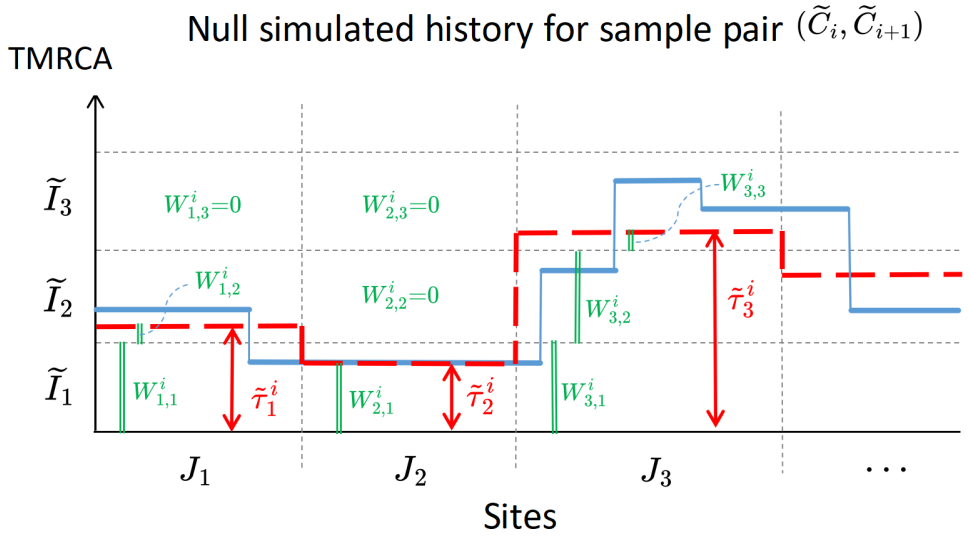

Figure 1: TMRCA in a null simulation (solid blue) and its approximation (dashed red) which restricts the TMRCA to be fixed over each interval J_l at the average blue value in J_l. The green vertical lines correspond to elements of $\tilde{W}^i$ defined in the text.

Let $\{\tilde{C}_1, \ldots, \tilde{C}_n\}$ be the sequences generated in the null simulation. As we did for the variable-size model, for each sequence pair $(\tilde{C}_i, \tilde{C}_{i+1})$ we investigate the average TMRCA denoted $\tilde{\tau}^i_l$ over J_l (see Figure 1). Since TMRCA are observed in the null simulation, we can further break down by time. Let $\tilde{W}^i$ be an $L \times T$ matrix with $\{l, t\}$ entry the contribution to $\tilde{\tau}^i_l$ from time interval $\tilde{I}_t = [\tilde{g}_{t-1}, \tilde{g}_t)$,

$$\tilde{W}^i_{l,t} = \begin{cases} \tilde{g}_t - \tilde{g}_{t-1} & \text{if } \tilde{\tau}^i_l > \tilde{g}_t \ , \\ \tilde{\tau}^i_l - \tilde{g}_{t-1} & \text{if } \tilde{g}_{t-1} < \tilde{\tau}^i_l < \tilde{g}_t \ , \\ 0 & \text{if } \tilde{\tau}^i_l < \tilde{g}_{t-1} \ , \end{cases} \tag{5}$$

so that row l of $\tilde{W}^i$ has sum $\tilde{\tau}^i_l$.

Let $\tilde{X}_2$ be the $nL \times T$ matrix constructed by vertically stacking the $\tilde{W}^i$, $i = 1, \ldots, n$, then sorting the rows to have increasing row sums. The $\{m, t\}$ element of $\tilde{X}_2$ satisfies $x_{m,t} \leq x_{m+1,t}$ for all $m = 1, \ldots, nL-1$ and $t = 1, \ldots, T$. In particular, if row m has t' nonzero entries, then row $m+1$ has $\geq t'$ nonzero entries, with $|\tilde{g}_t - \tilde{g}_{t-1}|$ in both rows for $t = 1, \ldots, t'-1$.

The matrix $\tilde{X}_2$ informs us about the average TMRCA $\tilde{\tau}^i_l$ in the null simulation, and we

have used (4) to estimate the density f_τ of the average TMRCA τ_l^i in the variable-size model. The next step is to link the two via β, the vector of population size ratios which from (2) is also the ratio of durations of the time intervals in variable-size and null models. Let Y_2 be a length-nL vector with mth element the estimated $m/(nL+1)$ quantile of f_τ. Although $\tilde{\tau}_l^i$ is independent of τ_l^i, because $\beta \geq 0$ the elements of $\tilde{X}_2\beta$ and Y_2 are both in increasing order and so we can infer the time-scaling relationship by assuming equality in the quantile approximation

$$\tilde{X}_2\beta \approx Y_2. \tag{6}$$

2.4 Estimation of population size

The first rows of $\tilde{X}_1$ and Y_1 correspond to the number of frequency-one alleles, providing overlapping information with $\tilde{X}_2$ and Y_2, so we improve robustness by removing the first rows of $\tilde{X}_1$ and Y_1 and then define $\tilde{X} = (\tilde{X}_1', \omega\tilde{X}_2')'$ and $Y = (Y_1', \omega Y_2')'$, where $'$ denotes matrix transpose. The dimensions of $\tilde{X}$ are $K \times T$, where $K = (L+1)n-2$, and (3) and (6) imply that $\tilde{X}\beta \approx Y$.

We set by default $\omega = \sqrt{400c/(nL)}$ in InferNo, which we found to work well for human data. It gives decreasing importance to $\tilde{X}_2$ and Y_2 as n increases since the de novo mutations in sample pairs recorded in Y_2 are less informative about recent population size, as found in many SMC-based approaches. Other choices for ω can be specified in InferNo, with higher ω tending to improve precision in the distant past at a cost to recent inferences.

Let S be a diagonal matrix with kth diagonal element $s_k = (y_k+1)^{-1/2}$, $k = 1, \ldots, K$, where y_k is the kth element of Y. Thus $1/s_k$ approximates the (Poisson) standard deviation of y_k, adjusted to ensure finite values. We estimate β by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \|S(\tilde{X}\beta - Y)\|_2^2 + \lambda \sum_{t=2}^{T-1} \alpha_{t-1} \left[\frac{\beta_{t+1} - \beta_t}{\tilde{g}_{t+1} - \tilde{g}_{t-1}} - \frac{\beta_t - \beta_{t-1}}{\tilde{g}_t - \tilde{g}_{t-2}} \right]^2 \right\}, \tag{7}$$

subject to $\beta_t \geq 0$ for $t = 1, \ldots, T$. Here $(\beta_{t+1} - \beta_t)/(\tilde{g}_{t+1} - \tilde{g}_{t-1})$ approximates the derivative

(trend) of the population size at g_t, thus the last term of (7) penalizes the trend-difference of the population size at neighbouring time breakpoints of $g_1, \ldots, g_{T-1}$. InferNo allows user-specified weights α_t, here we use the default values $\alpha_{t-1} = \tilde{g}_{t+1} - \tilde{g}_{t-1} + \tilde{g}_t - \tilde{g}_{t-2}$, for $t = 1, \ldots, T-2$, so that different penalties are applied to the trend-differences based on their time spans. We further adjust $\hat{\beta}_T$ by letting $(\hat{\beta}_T - \hat{\beta}_{T-1})/[2(\tilde{g}_{T-1} - \tilde{g}_{T-2})] = (\hat{\beta}_{T-1} - \hat{\beta}_{T-2})/(\tilde{g}_{T-1} - \tilde{g}_{T-3})$, so that the trend at g_{T-1} remains approximately the same as that at g_{T-2}. InferNo allows the time breakpoints $\tilde{g}_1, \ldots, \tilde{g}_{T-1}$ to be user-specified so that $\tilde{I}_T$ corresponds to the remote past, beyond the period of interest, minimising the effect of α_{T-2}. We recursively obtain $g_1, \ldots, g_{T-1}$ from $\hat{\beta}_t = (g_t - g_{t-1})/(\tilde{g}_t - \tilde{g}_{t-1})$ for $t = 1, \ldots, T-1$. Substituting g_t in (2), $N(g)$ is finally estimated by $\hat{N}_t = \tilde{N}\hat{\beta}_t$ for all $g \in I_t$.

The optimisation (7) is similar to ridge regression, but the penalty term involves a weighted second difference of β so that larger values of λ lead to smoother $\hat{\beta}$. We propose to select λ by minimising over a set of candidate values the adjusted Bayesian information criterion (BIC) (Schwarz, 1978; Ye, 1998):

$$\text{BIC} = K \log\left(\frac{1}{K}\|S(\tilde{X}\hat{\beta} - Y)\|_2^2\right) + d\log(K),$$

where d is the estimated degrees of freedom,

$$d = \text{trace}\{S\tilde{X}(\tilde{X}'S'S\tilde{X} + \lambda P'P)^{-1}\tilde{X}'S'\},$$

and P is a $(T-2) \times T$ matrix such that

$$P_{i,j} = \begin{cases} 2\sqrt{\alpha_i}/(g_{t+1} - g_{t-1}) & \text{if } j = i, \\ -2\sqrt{\alpha_i}[1/(g_{t+1} - g_{t-1}) - 1/(g_{t+2} - g_t)] & \text{if } j = i+1, \\ 2\sqrt{\alpha_i}/(g_{t+2} - g_t) & \text{if } j = i+2, \\ 0 & \text{otherwise.} \end{cases}$$

2.5 Design of Simulation studies

		g (in units of 10^3 generations)												
Model	$N(0)$	0	0.025	0.03	0.3	0.4	1	1.8	1.9	3	9	18	18.1	30
1	20 000	0			0	0	0	0	0	0	0	0	0	0
2	40 000	10	10	10	10	10	10	10	10	10	10	10	10	0
3	5 000	30	30	30	30	30	30	30	30	8	4	2	2	0
4	40 000	20	20	20	20	20	20	20	20	10	−3	−3	−3	0
5	40 000	0	0	0	1 400	0	0	−1 200	0	0	0	500	0	0
6	100 000	700	700	700	700	−270	0	0	0	27	−9	−9	−9	0
7	300 000	5000	−15000	10	10	10	10	10	10	10	10	10	10	10

Table 1: **Growth rates (per 10^5 generations) for the simulation models.** The rates in each column apply from the generation shown in the top row of that column until the generation shown in the next column, or indefinitely in the case of the final column. $N(0)$ is the present-day (haploid) population size.

We used msprime (Kelleher et al., 2016) to generate sequences of length $\ell = 10^7$ under population-size Model 1-6 (Table 1), and of length $\ell = 5\times10^7$ under Model 7, which all include exponential growth/decline from ancient population sizes with rates varying over time intervals. Model 1 is the special case of a constant population size. Model 2 (Model 3) represents monotonic increasing (decreasing) population size at varying rates. Model 4 involves a turning point, where the population size declines and then grows slowly (going backwards in time), while Model 5 approximates a piecewise constant population size with a bottom and a peak in the middle. The population of Model 6 experienced two bottlenecks, allowing us to test whether the signal from an older bottleneck can be detected despite a subsequent bottleneck. Model 7 involves the most recent bottleneck at 25 generations ago, which is used to challenge InferNo in detecting recent extreme events, such as the black death, using biobank scale data. We use 25 replicate datasets for each setting, and compare performance using RMISE (root mean integrated squared error with "integration" being a sum over g).

With some exceptions described below, the data simulations used the mutation and recombination maps shown in Figure 2. Where required for the analyses, the genome-wide average mutation rate and the recombination rate were set to the correct values, 1.3×10^{-8} and 10^{-8} per site and per generation, but the variation along the genome is not input to

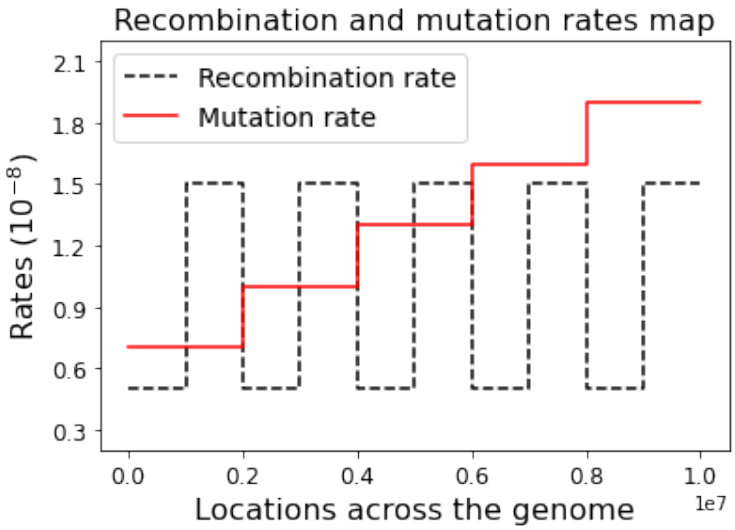

Figure 2: Recombination and mutation maps used in the simulation study. InferNo requires genome-wide average values, which for these maps are $r = 10^{-8}$ and $\mu = 1.3{\times}10^{-8}$ per site and per generation.

InferNo. To further challenge inferences with model misspecification, we include sequencing errors in the data simulations, with rate $\epsilon = 1$ per 10^5 sequence sites, and gene conversions with rate 2×10^{-8} per site per generation and tract length 300 sites which are similar to the estimates of human gene conversion reported in Williams et al. (2015).

Each InferNo analysis outputs a BIC value and we select the smoothing parameter λ to minimise the BIC over a range of values: we illustrate this approach for Models 4 and 5 with $n = 100$ and $2\log_{10}(\lambda) \in \{3,5,7,9,11,13,15,17\}$ (Figure S1). For the remainder of the simulation study we minimised BIC over these λ values. In practice, the extreme values 3 and 17 were never chosen.

We compared the performance and the computing time of InferNo with Relate, PSMC, Mushi and CHIMP using sample sizes $n = 10$ and 200 under population size Model 1-6. As a special case, we further compared InferNo with PSMC when $n = 2$ (a single diploid individual). For Relate, we set $N(0) = 30\,000$ for both ARG inference and population size estimation. For PSMC, we set 20 free parameters, otherwise default settings were used. For Mushi, we set the maximum time investigated to 60 000 generations and the maximum number of iterations to 300. We obtain piece-wise constant estimates by fixing the trend order

parameter to 0 and selecting the trend penalty parameter over $\{1, 3, 5, 10, 20, 40, 60, 100\}$ in each setting. In practice, the extreme values 1 and 100 were never chosen. For CHIMP, we set base_n = 2 and the right endpoints of the 20 time intervals as $\exp\{3 + 7.6(t-1)/19\}$ for $t = 1, \ldots, 20$. To accelerate the convergence of CHIMP, we further initialise the population sizes at each time interval to be $N(g) + \varepsilon$, where g is the starting time of the interval and ε is a normal error with mean zero and standard derivation $N(g)/7$. By doing so, we provided additional information of the demography for CHIMP, in return for faster computing time. The other inputs are set by default.

We use $T = 20$ time intervals for $N(g)$ in all six models and methods except for Relate, which automatically determines the time partition. The original PSMC only uses a pair of sequences regardless of n. We investigate the performance of PSMC by partitioning the sequences into $n/2$ pairs and then averaging the resulting $n/2$ estimates.

We then undertook a series of experiments to assess the performance and computational speed of InferNo, over sample sizes ($n = 10$, 50, 100, 500 and 1000 sequences) and under two variations to the data generating model that we now describe.

In the case $n = 100$ we checked the robustness of InferNo to misspecification of the recombination rate across the genome, by simulating datasets with (1) the recombination map given in Figure 2, which has average rate $r = 10^{-8}$ per site and per generation, (2) a constant rate $r = 10^{-8}$ (which matches the InferNo analysis model) and (3) a constant rate $r = 1.5{\times}10^{-8}$.

Again with $n = 100$, we next checked the robustness of InferNo to sequencing error. We ran InferNo on simulated datasets contaminated by sequencing error with rates $\epsilon = 0$, 1, 2 and 3 errors per 10^5 sites.

Then we applied InferNo to pseudo datasets in biobank scales to test its feasibility for inference close to the present day, using large data. Sequences of length $\ell = 5 \times 10^7$, similar to the size of Chromosome 21, are generated under Model 7 using msprime, with sample size $n = 2{\times}10^5$, recombination rate 10^{-8} and mutation rate $1.3{\times}10^{-8}$. Sequencing error is set

to 10^{-5}, and gene conversion is not included since it dramatically increases the computing time of the data-generating process. For InferNo inference, we set $\tilde{N} = 100\,000$. To focus on recent population size, we set $\tilde{g}_t{=}t/2$ for $t = 1, \ldots, 50$, and $\tilde{g}_t{=}30\exp\{1.6(t-51)\}$ for $t = 51, \ldots, 56$. Since the default value of parameter ω is very small when $n = 2{\times}10^5$, we manually set $\omega = 0$ which will dramatically accelerate InferNo.

Finally, we further validated InferNo for recent population size inference, by applying it to the same demographic models and settings as in Browning and Browning (2015), which aimed at recent population size estimation with IBD.

2.6 Real data analysis

Population	Description	Sample size	Region
LWK	Luhya in Webuye, Kenya	198	
MSL	Mende in Sierra Leone	170	Africa
ESN	Esan in Nigeria	198	(AFR)
YRI	Yoruba in Ibadan, Nigeria	216	
CHB	Han Chinese in Beijing, China	206	
JPT	Japanese in Tokyo, Japan	208	East Asia
CHS	Southern Han Chinese	210	(EAS)
CDX	Chinese Dai in Xishuangbanna, China	186	
FIN	Finnish in Finland	198	
GBR	British in England and Scotland	182	Europe
IBS	Iberian Population in Spain	214	(EUR)
TSI	Toscani in Italia	214	
BEB	Bengali from Bangladesh	172	
ITU	Indian Telugu from the UK	204	South Asia
PJL	Punjabi from Lahore, Pakistan	192	(SAS)
STU	Sri Lankan Tamil from the UK	204	

Table 2: Populations from the 1 000 Genomes Project (Fairley et al., 2020) analysed here.

We estimated the population sizes $N(g)$ using the Chromosomes 20 ($\ell = 63\,025\,522$ sites) and 21 ($\ell = 48\,129\,897$ sites) data from 16 populations of the 1000 Genomes Project (1KGP) (Fairley et al., 2020), including four populations from each of Africa, East Asia, Europe, and South Asia (Table 2). On average over the 16 populations, 296 762 and 185 178 sites are polymorphic on Chromosome 20 and 21, respectively.

The sequence data were downloaded as .vcf files from ftp.1000genomes.ebi.ac.uk. Then, we converted them to the .samples format for the Python implementation of InferNo. Specifically, we first cloned the Github repository from the website github.com/awohns/unified_genealogy_paper (Wohns et al., 2022) and installed the software, packages and modules listed in the requirements.txt file and the tools sub-folder. Separately for Chromosome 20 and 21, in the all-data sub-folder we used Makefile to build the tree sequence from the variant data, and produce a .samples file from the .vcf file.

3 Results

3.1 Simulation study results

Figure 3 shows that all methods perform relatively poorly at recent times ($g < 10^3$) and when g is very large ($g \gg 10^4$). In the former case, the small number of recent recombination events limits the number of distinct lineages that can contribute to inference, while in the latter case the number of lineages has been reduced by coalescences. Rapid changes of $N(g)$ with g also challenge every method.

Model	InferNo		Relate		PSMC		Mushi		CHIMP	
1	**2.6**	(2.6)	7 087	(10.9)	781	(4.5)	737	(**2.1**)	7.2	(7.2)
2	**5.4**	(**5.3**)	15 575	(8.8)	959 530	(10.6)	1 236	(8.8)	9153	(9149)
3	**8.7**	(8.7)	3 756	(14.4)	254 540	(**6.2**)	820	(12.9)	11.6	(11.6)
4	**5.2**	(**5.2**)	9 340	(10.7)	3 543 030	(7.8)	832	(7.5)	8.2	(7.9)
5	**10.1**	(**9.5**)	6 864	(17.7)	4 794	(11.0)	371	(10.0)	16.4	(16.4)
6	**11.3**	(**11.1**)	7 181	(21.0)	4 441 128	(12.1)	1 691	(20.4)	13.5	(13.5)

Table 3: **RMISE (in units of 10^5) for $g \in [200, 35\,000]$ ($g \in [1\,000, 35\,000]$) for five $N(g)$ inference methods**. RMISE=root mean of the sum over g of squared errors. Values are averages over the 25 curves of Figure 3.

On the criterion of minimising RMISE over $g \in [200, 35\,000]$, InferNo is superior to all other methods, in most cases by orders of magnitude (Table 3). This primarily reflects the poor performance of the comparison methods over recent times ($g \in [200, 1\,000)$). When

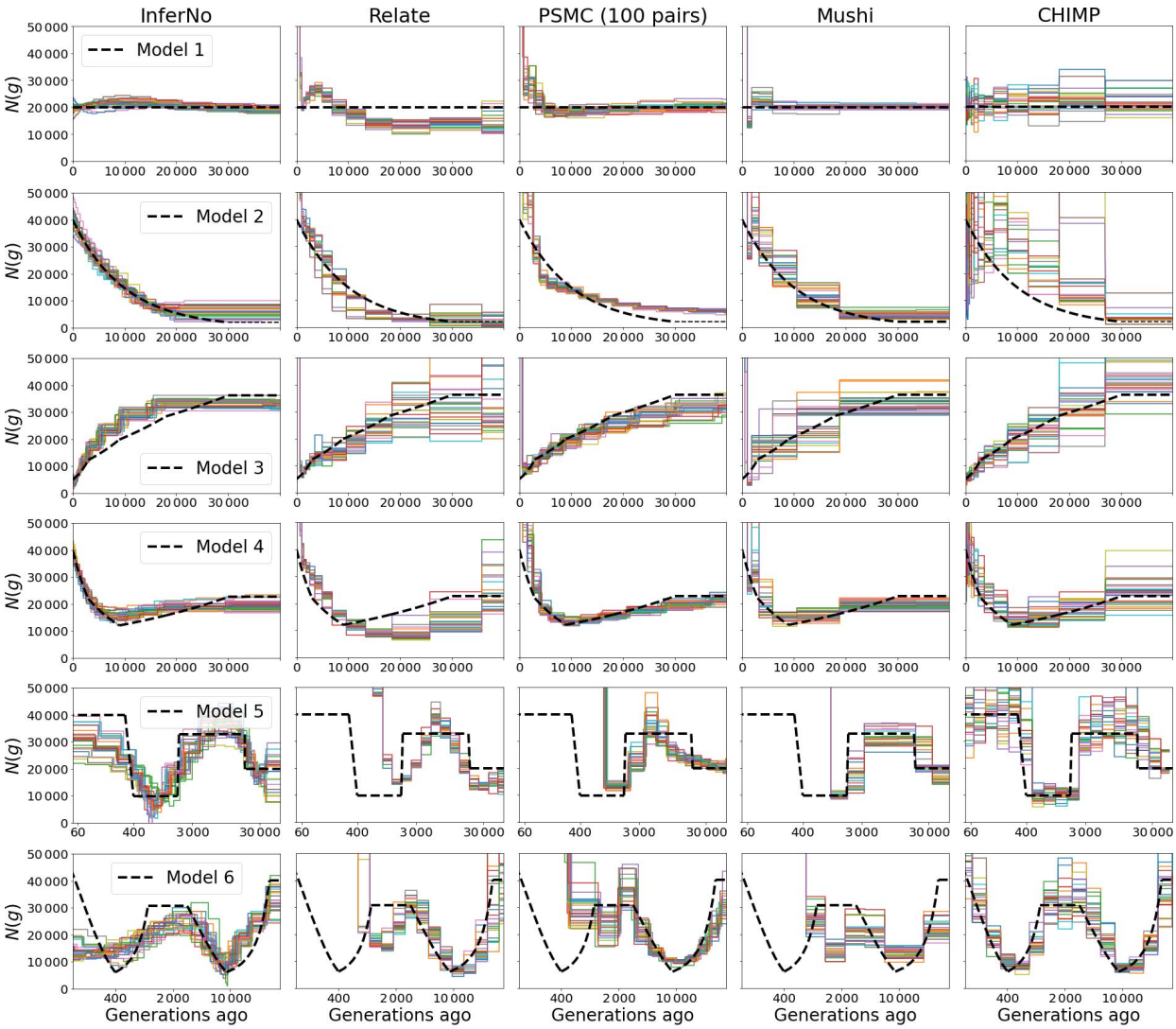

Figure 3: **Comparison of InferNo with four alternative methods for estimating population size** $N(g)$. Sample size $n = 200$ and sequence length $\ell = 10^7$ sites. Note logarithmic time scale for Models 5 and 6. See Table 3 for quantitative comparison based on RMISE, Figure S2 for corresponding plots when $n = 10$ and Table S1 for computing times.

we remove this time interval and re-compute RMISE over $g \in [1\,000, 35\,000]$, the methods become more comparable, but InferNo is best in 4 of 6 models and second in the other 2, overall better than any other method. In the special case of a single diploid individual ($n = 2$), InferNo greatly outperforms PSMC for $g < 5\,000$ (Table 4). However for $g > 5\,000$ the two methods show similar accuracy.

InferNo is also the fastest among the five methods, taking on average (over the six models

n=2		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
$g \in [200, 35\,000]$	InferNo	11	11	8	13	16	22
	PSMC	8M	10M	4M	10M	7M	6M
$g \in [1\,000, 35\,000]$	InferNo	10	11	8	13	15	21
	PSMC	31K	3.0M	24	2.9M	42	1.9K
$g \in [5\,000, 35\,000]$	InferNo	4.3	7.9	7.5	10.0	9.9	14.8
	PSMC	6.2	6.9	8.2	5.0	11.9	10.5

Table 4: **RMISE (in units of 10^5) of three time intervals for InferNo and PSMC when $n = 2$.** RMISE=root mean of the sum over g of squared errors. M = million, K = thousand. Values are averages over the 25 curves of Figure S3.

and 25 replicate datasets for each model) 8.5 seconds per analysis for $n = 200$, just faster than 9.2 seconds for Mushi and much faster than 386 seconds for CHIMP, 406 seconds for Relate and 983 seconds for PSMC (Table S1). Mushi has a computing cost similar to InferNo when $n = 200$, and is slower when $n = 10$.

	Sample Size n						Recombination			Error rate ϵ (per 10^5 sites)			
Model	10		100		1 000		map	$r = 1$	$r = 1.5$	0	1	2	3
1	0.9	(2.7)	0.3	(2.7)	0.1	(2.4)	2.6	3.1	2.5	2.5	4.9	5.2	6.2
2	2.6	(8.8)	1.0	(4.7)	0.3	(4.8)	5.4	5.4	4.2	5.6	5.7	6.0	5.7
3	0.8	(6.6)	0.6	(11.9)	0.5	(9.1)	12.4	10.7	10.4	8.4	8.3	8.5	8.0
4	1.7	(6.8)	1.0	(5.0)	0.4	(5.4)	5.7	5.3	6.9	5.3	5.8	6.3	6.6
5	4.5	(11.4)	2.7	(11.0)	1.8	(8.2)	10.8	10.9	11.2	9.5	9.5	9.9	10.9
6	11.6	(17.1)	1.9	(12.8)	1.3	(17.0)	12.9	13.0	14.7	17.0	16.7	16.9	17.0
Mean	ref	(ref)	0.34	(0.90)	0.20	(0.88)	1.03	ref	1.03	ref	1.05	1.09	1.13

Table 5: **RMISE under perturbations of the data generating model**. RMISE=root mean integrated squared error (in units of 10^5). Entries in the bottom row are expressed relative to a baseline indicated by ref. The sample size values correspond to the curves in Figure S4, with integration over $g \in [200, 1\,000]$ ($g \in [1\,000, 35\,000]$). The recombination values correspond to the curves in Figure S6 (see its caption for details of the recombination models) and the sequencing error values correspond to the curves in Figure S7; in both cases the integration is over $g \in [200, 35\,000]$.

Larger n improves inference more for recent generations ($g < 1\,000$) than the distant past (Figure S4), because many lineages coalesce in recent generations. The average RMISE for $g \in [1\,000, 35\,000]$ decreases by only 12% when n increases 100-fold from 10 to 1 000 (Table 5), but for $g \in [200, 35\,000]$ RMISE decreases by a factor of 5.

Because the computing time is approximately linear with n for each model (Figure S4), InferNo can exploit large sample sizes to obtain accurate estimates of $N(g)$ for small g.

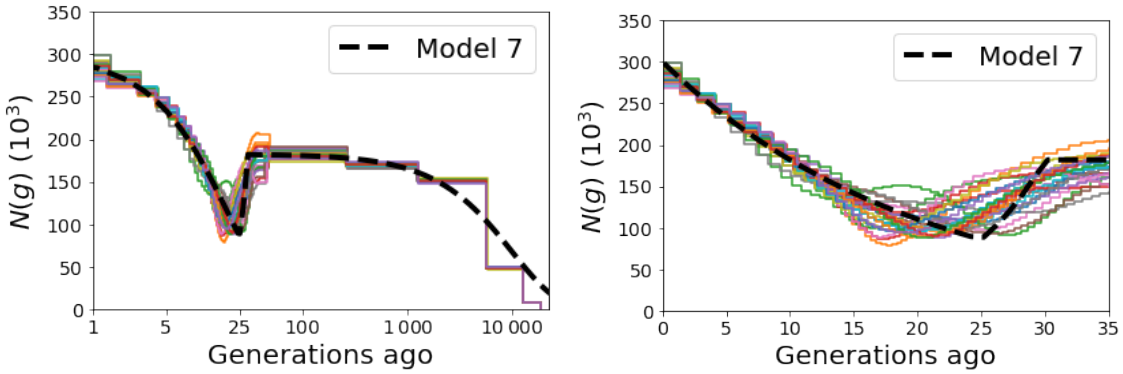

Figure 4: **InferNo estimates of population size $N(g)$ with large sample size under Model 7**. Sample size $n = 2{\times}10^5$ and sequence length $\ell = 5{\times}10^7$ sites. The left plot shows results on a logarithmic time scale for $g \in [1, 22\,000]$, while the right plot shows the same results for $g \in [0, 35]$.

Figure 4 illustrates the ability of InferNo to infer population size changes right up to the present day with sample size $n = 2{\times}10^5$. The average processing time for each of the 25 estimates shown is only 326 seconds. Thus, it is feasible to use InferNo on even the largest available human biobanks to infer effects of famines, plagues and migrations over recent centuries. Figure S5 further demonstrates its capability for accurate estimates of recent population sizes.

Figure S6 shows little effect on inference of misspecifying recombination rates, with average RMISE increasing 3% when the recombination map of Figure 2 or a constant rate 50% higher than that assumed by InferNo is used for the data simulation (Table 5), relative to constant-rate recombination with rate equal to the average rate assumed by InferNo.

Similarly, Figure S7 shows that InferNo is robust to sequencing errors, with RMISE (Table 5) increasing by 5% when sequencing errors are introduced at a low rate (1 error per 10^5 sites), then RMISE increases by approximately 9% and 13% as the rate is increased to 2 and then 3 errors per 10^5 sites.

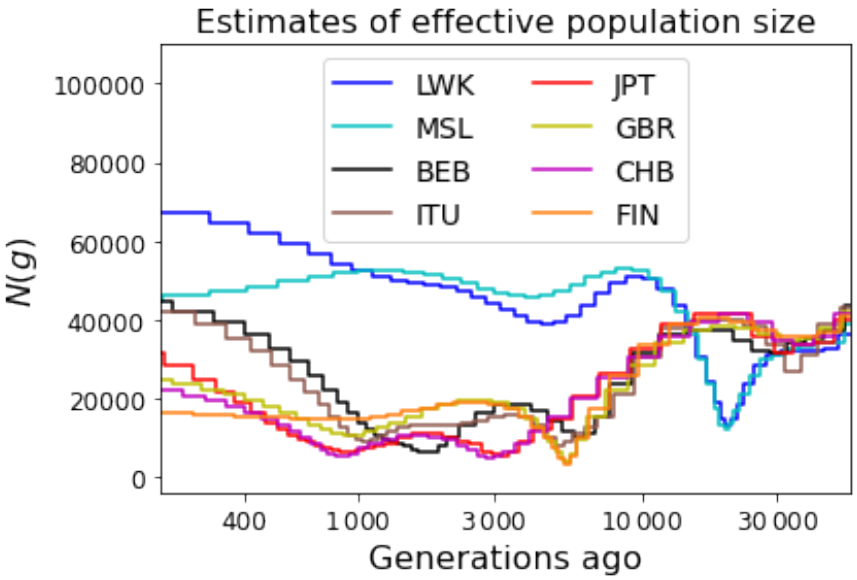

Figure 5: **InferNo estimates of population size $N(g)$ for eight 1KG populations**. See Table 2 for the population codes. The results are weighted averages of the estimates from Chromosomes 20 and 21 (see Figure S8 for results from each chromosome), with weights proportional to the numbers of sites that are polymorphic in the eight populations combined. The time axis is on a logarithmic scale.

3.2 Results from 1KG data analysis

The $N(g)$ curves inferred from Chromosomes 20 and 21 are similar (Figure S8), and we focus on the combined results (Figure 5) which show the populations falling into two groups: two African and six non-African. The African populations have $N(g) > 40\,000$ ($= 20\,000$ individuals) for $g < 12\,000$ (up to 324 000 years ago if we assume 27 years per generation (Wang et al., 2023)) and $N(g)$ only briefly dips below 20 000, at about $g = 18\,000$. In contrast, the non-African populations mostly have $N(g) < 40\,000$ for $g < 12\,000$ and they each experience bottlenecks, which may be due to migration events, plagues and/or climatic extremes. The most severe bottleneck arose in the ancestry of the two EUR populations, with $N(g) \approx 3\,000$ at $g \approx 5\,300$ ($\approx 143\,100$ years ago). The African and non-African populations remain distinct at $g = 20\,000$, well before the date usually given for the migrations of modern humans out of Africa, which is typically around $g = 3\,000$.

Figure 6 shows a striking similarity of the inferred $N(g)$ curves for $g > 1\,000$ across

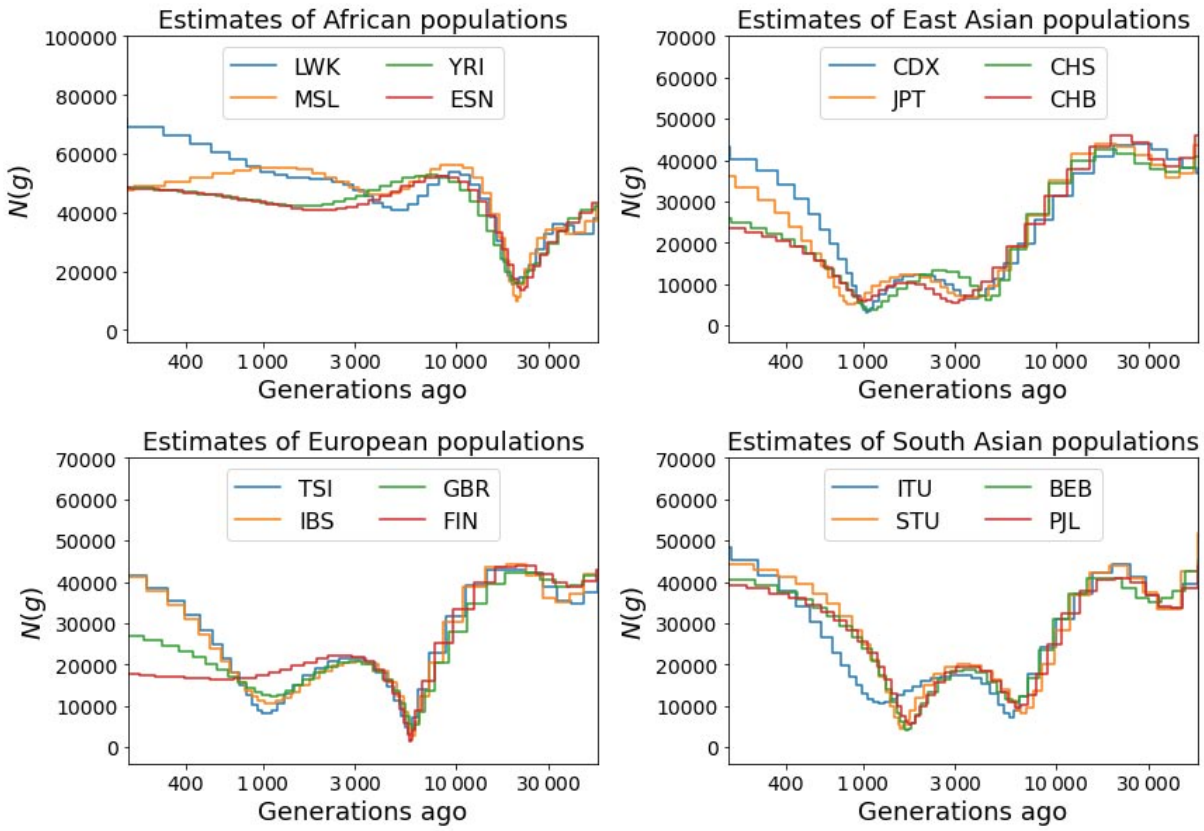

Figure 6: **InferNo estimates of population size $N(g)$ for four populations in each of four worldwide regions.** Based on data from Chromosome 20 only.

the four populations within each of four worldwide regions, despite the InferNo analysis being performed independently for each population. Each of the 12 non-African populations except FIN shows two bottlenecks, with the most severe bottleneck in EAS and SAS populations occurring when $g \approx 1\,000$ (with $N(g) \approx 3\,600$) and $g \approx 1\,400$ (with $N(g) \approx 5\,400$), respectively, prior to the last glacial maximum.

Our results show broad similarity, but also important differences, to the results of (Bergström et al., 2020, Fig 4), who analysed samples from different populations in the same regions, and Upadhya and Steinrücken (2022) who analysed three of the 1KGP populations studied here. Bergström et al. (2020) used MSMC2 (Schiffels and Wang, 2020) which is related to PSMC and was found by Upadhya and Steinrücken (2022) to be inferior to their CHIMP software. Both these studies identified only a single bottleneck in the histories of EUR and

EAS populations, dated around $g = 1\,400$ and $g = 3\,700$, respectively. There is no ground-truth available to conclude that any method is most accurate, but we point to the better performance of InferNo in simulations, including under Model 6 which includes two bottlenecks. We suggest that signals from the multiple bottlenecks shown in the InferNo results may have been misinterpreted by the other methods as single bottleneck at a compromise time. InferNo also infers a single bottleneck when its smoothing parameter λ is increased from the value that minimises BIC.

4 Discussion

InferNo is a new method for estimating effective population sizes $N(g)$, $g \geq 0$. It achieves higher statistical and computational efficiency than existing approaches by combining analytical steps with a single genealogical history simulation. The simulation model assumes constant population size, which is equivalent to the target variable population-size model under a time rescaling that we infer by comparing the observed AFS, and site differences for sequence pairs in different genome regions, with results from the null simulation.

We showed robustness of InferNo to misspecification of the recombination rate and map, the presence of sequencing errors and gene conversions, and variation in the mutation rate along the genome. The better performance of InferNo over four alternative approaches is particularly marked for recent times ($g < 1\,000$). The computational cost of InferNo is low, and increases linearly with n. We illustrated the potential of InferNo to infer recent population sizes up to the present day, using simulated biobank-scale datasets.

We used InferNo to estimate historical $N(g)$ for 16 worldwide human populations from the 1000 Genomes Project. We found a striking similarity of the inferred $N(g)$ curves over the four populations within each of four worldwide regions for $g > 1\,000$. As expected, we found the greatest differences between African and non-African populations, with the former showing larger $N(g)$ for $g < 12\,000$ and the latter showing bottlenecks with $N(g) < 10\,000$.

While InferNo avoids the challenging task of inferring recombination events, it requires a sequence length that is long relative to the recombination rate to generate replication along the genome, which is feasible for humans and many other species. InferNo allows $\mu(s)$ to vary over sites s, but like all other methods for inferring $N(g)$, it assumes that $\mu(s)$ is constant over time which is unlikely to be strictly accurate.

Here we have inferred $N(g)$ for a single population, but the approach can be extended to multiple populations. Our real data example used human data, but InferNo can be applied to any populations that are approximately panmictic. More precise estimates of $N(g)$ can benefit inference for other evolutionary parameters, such as species divergence times.

Data availability

The software InferNo as well as the data and computer code used in this paper are available at: https://github.com/ZhendongHuang/Inference_for_demographic_history.

Acknowledgment

ZH is funded by Australian Research Council grant DP210102168 awarded to YBC and DJB.

Supplementary Tables and Figures

	$n = 10$					$n = 200$				
Model	InferNo	Relate	PSMC	Mushi	CHIMP	InferNo	Relate	PSMC	Mushi	CHIMP
1	0.64	20.2	52.9	6.7	190	9.2	464	985	9.5	271
2	0.65	17.8	53.3	6.8	850	8.7	380	979	9.2	947
3	0.65	17.0	52.9	7.2	113	7.8	345	982	9.6	243
4	0.64	22.2	52.8	7.3	144	8.5	479	982	9.4	252
5	0.68	25.7	52.2	7.3	192	9.9	483	985	8.8	270
6	0.56	15.7	52.7	3.0	130	6.7	287	984	8.8	334
Mean	0.64	19.8	52.8	6.4	270	8.5	406	983	9.2	386

Table S1: **CPU time (seconds)**. Average computing time (over the 25 replicates in each setting) for the results in Figure 3 ($n = 200$) and Figure S2 ($n = 10$).

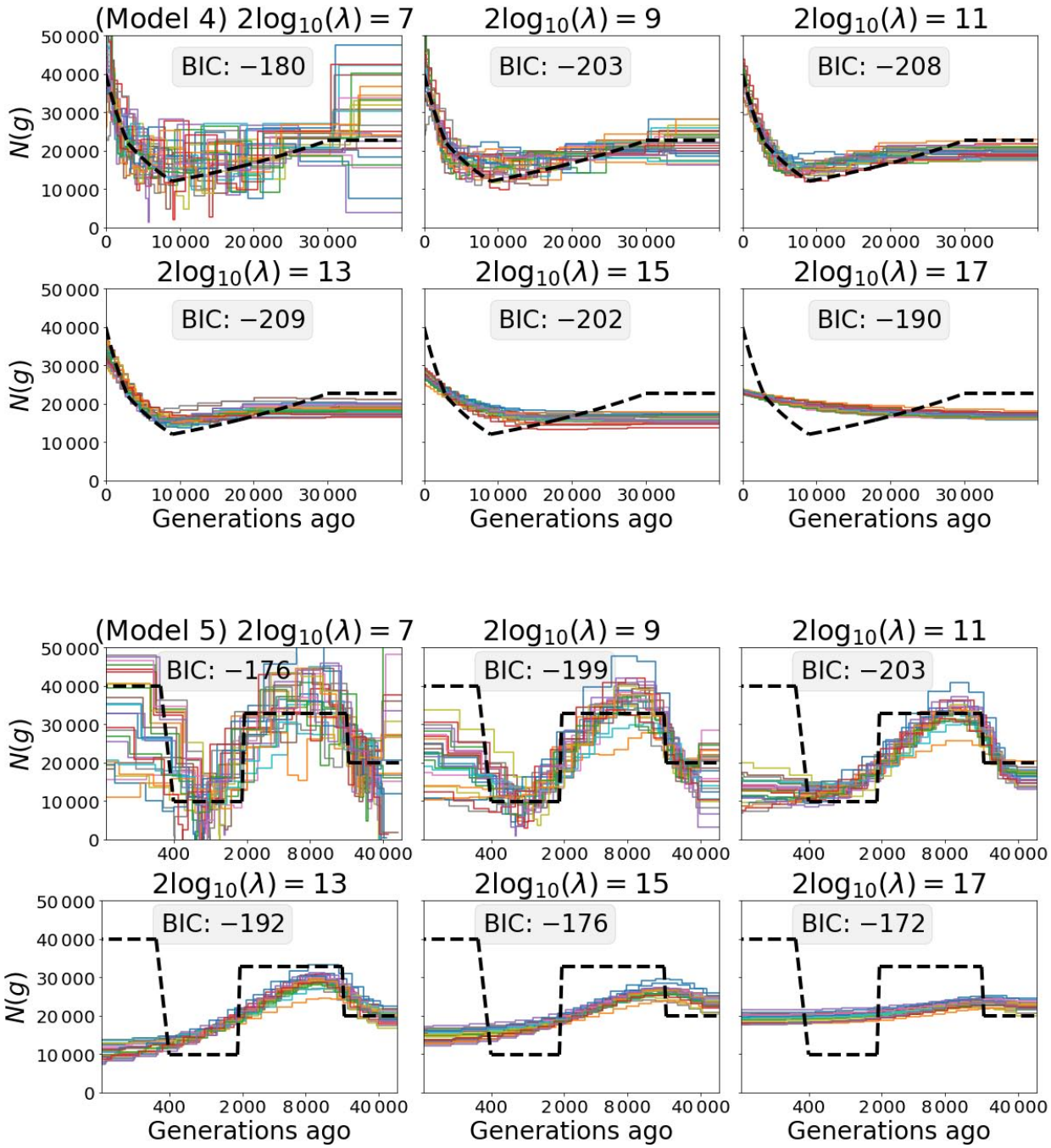

Figure S1: **Effects of smoothing parameter λ and the corresponding BIC (in units of 10^3)**. Estimates of population size $N(g)$, with $n = 100$, for Model 4 (top two rows) and Model 5 (bottom two rows, logarithmic time scale). For each model the six panels each corresponding to a value for λ. The best fit (minimum BIC) is obtained for $\lambda = 10^4$ in Model 4 and $\lambda = 10^2$ in Model 5.

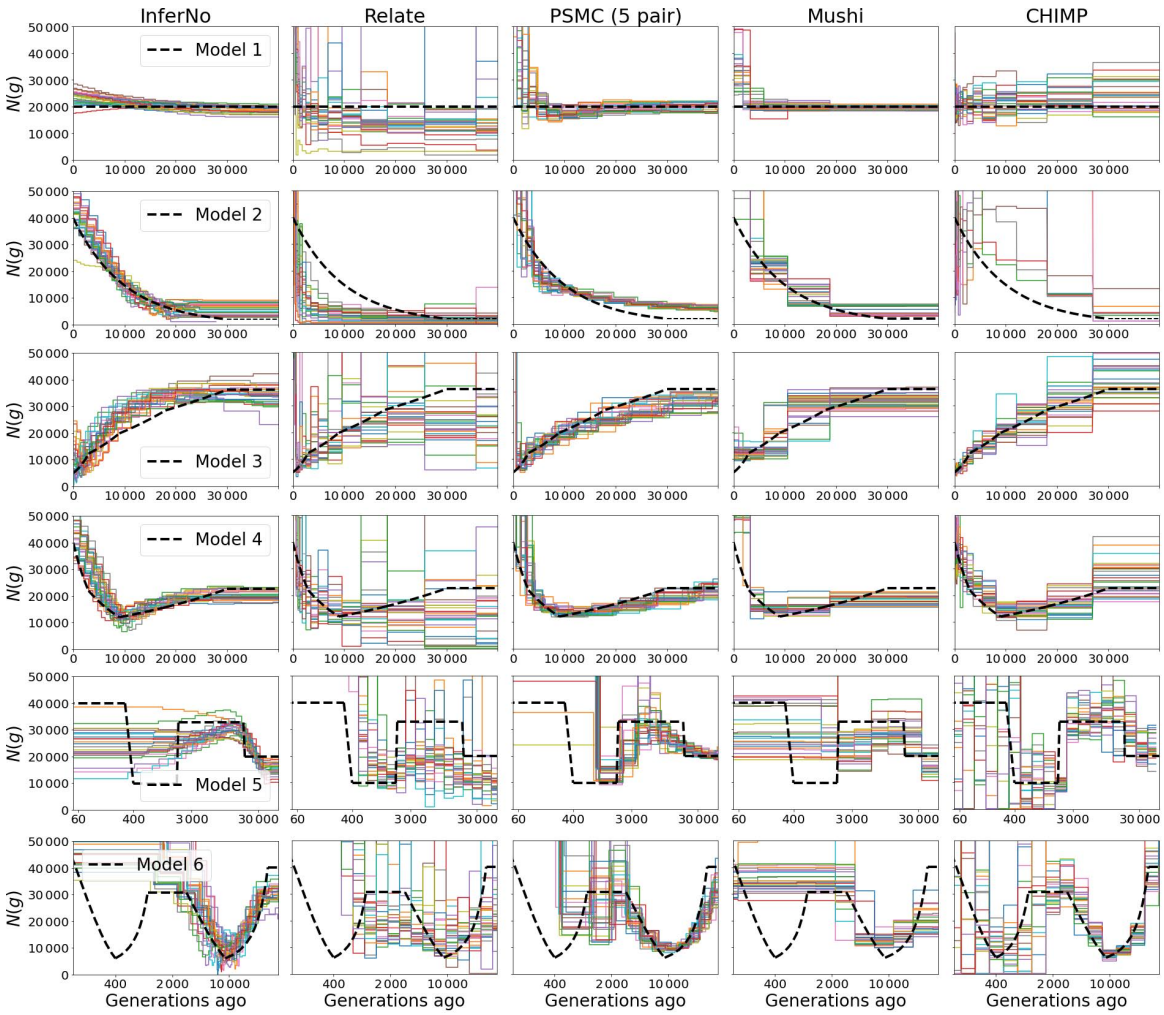

Figure S2: **Comparison with other methods** ($n = 10$). Other details are the same as for Figure 3.

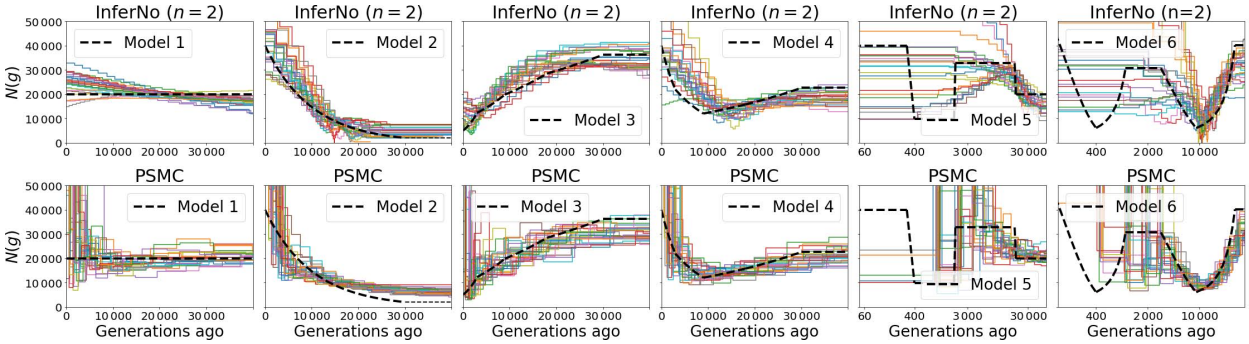

Figure S3: **Comparison of InferNo with PSMC when Sample size** $n = 2$. Sequence length $\ell = 10^7$ sites. Note logarithmic time scale for Models 5 and 6. See Table 4 for quantitative comparison based on RMISE.

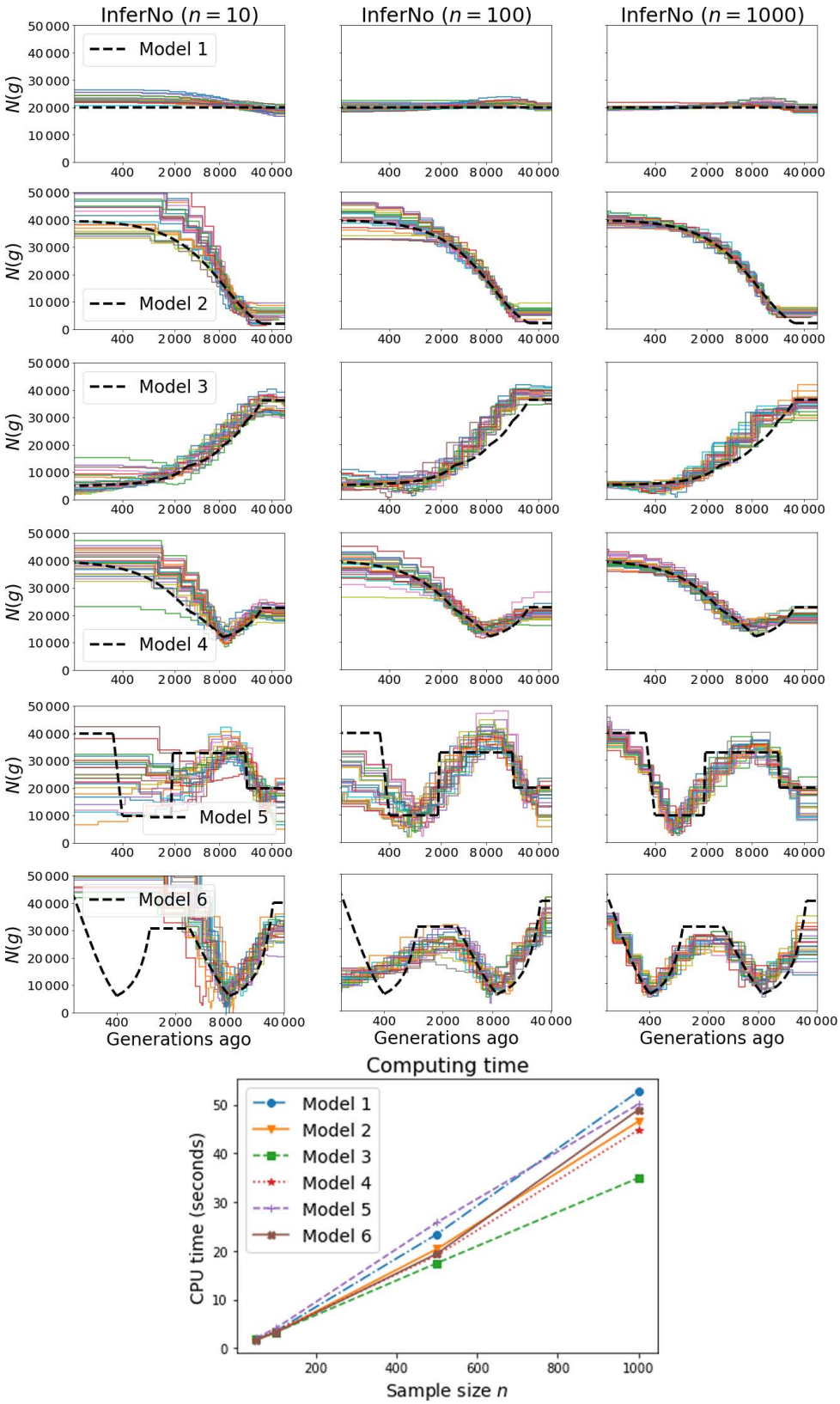

Figure S4: **Performance and computing time with different sample sizes**. InferNo estimates of population size $N(g)$ and run time with $n = 10$, 50, 100, 500 and 1 000, for Models 1-6. Each time axis is on a logarithmic scale.

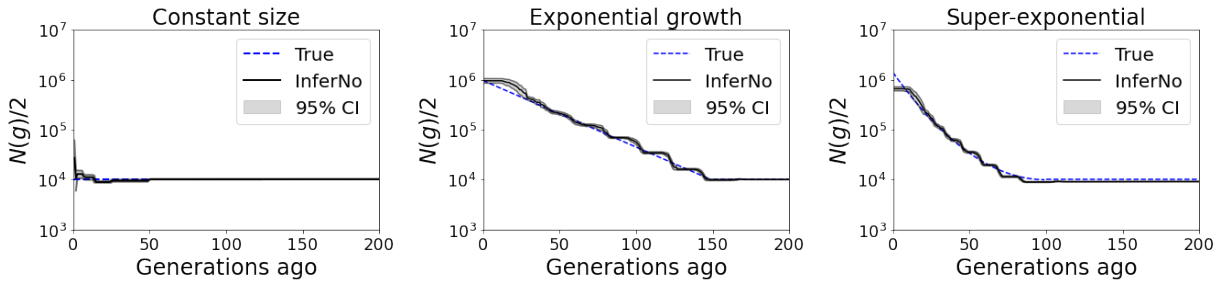

Figure S5: **Performance of InferNo under the demographic models in Browning and Browning (2015).** The left, middle and right panels correspond to the constant size, exponential growth, and super-exponential models defined in Browning and Browning (2015). Each curve is an average of estimates from 30 independent chromosomes of length ℓ=10^8. To match the settings with Browning and Browning (2015) as closely as possible, we set sample size n=2 000. Mutation and recombination rates are set as constant μ=r=10^{-8} and are treated as known. Sequencing errors and gene conversions are not included. For InferNo inference, we set $\tilde{g}_t$=$0.2\times2^{i-1}$ for $t = 1, \ldots, 9$ and $\tilde{g}_t$=$125\times5^{i-10}$ for $t = 10, \ldots, 14$, and α_t=I_t^4 to focus on the recent population size inference. To reduce computing time, we fix $\omega = 0$ and $\lambda = 0.02$, 0.01 and 0.02, respectively, for the three models.

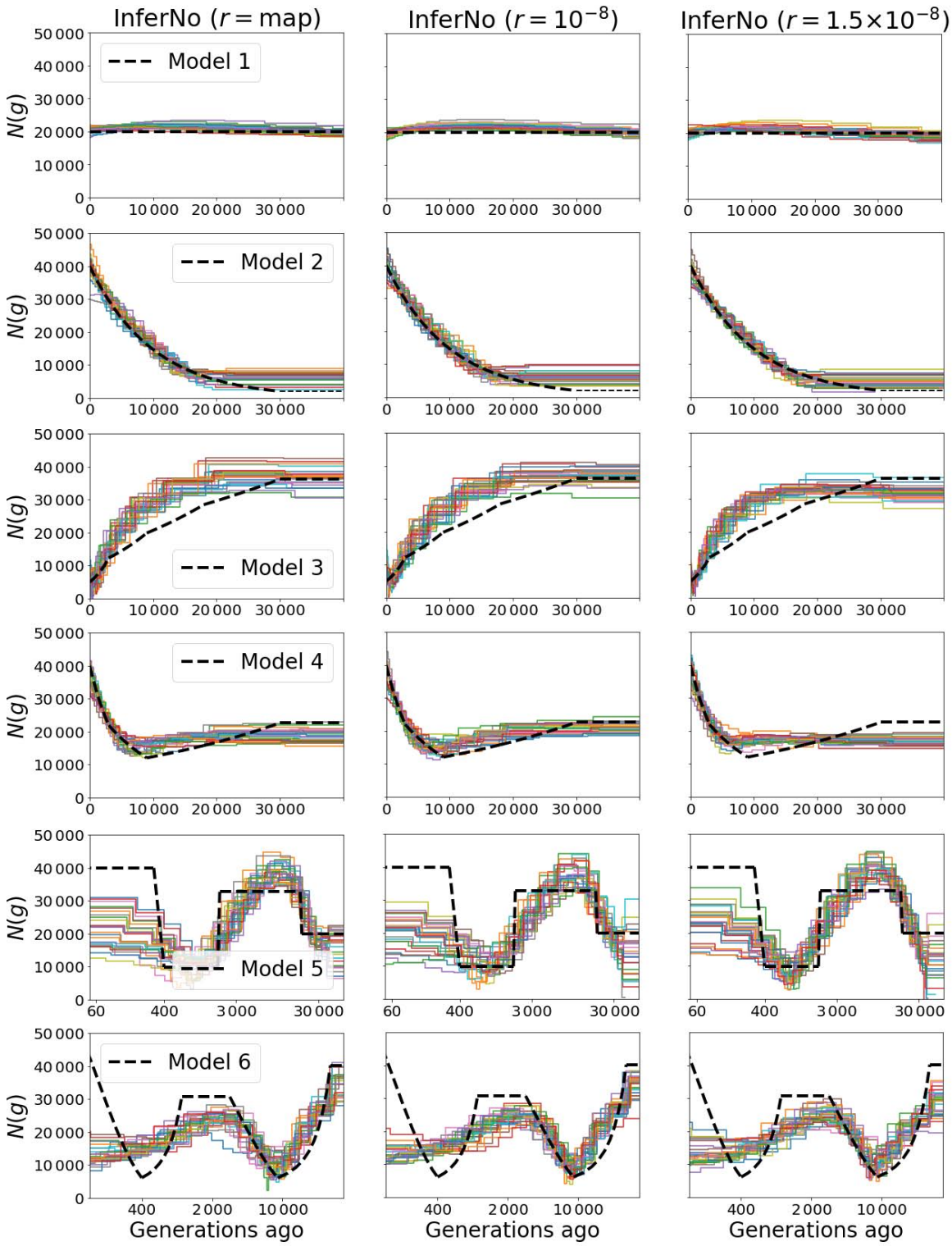

Figure S6: **Performance of InferNo with misspecified recombination rates**. Estimates of population size $N(g)$ with $n = 100$ when the observed data is generated with the recombination map in Figure 2 (column 1), and with constant rates $r = 1\times10^{-8}$ (column 2) and $r = 1.5\times10^{-8}$ (column 3). The InferNo analysis assumes $r = 1\times10^{-8}$ and so the first and last columns each correspond to a misspecified recombination model. Rows correspond to the simulation models in Table 1.

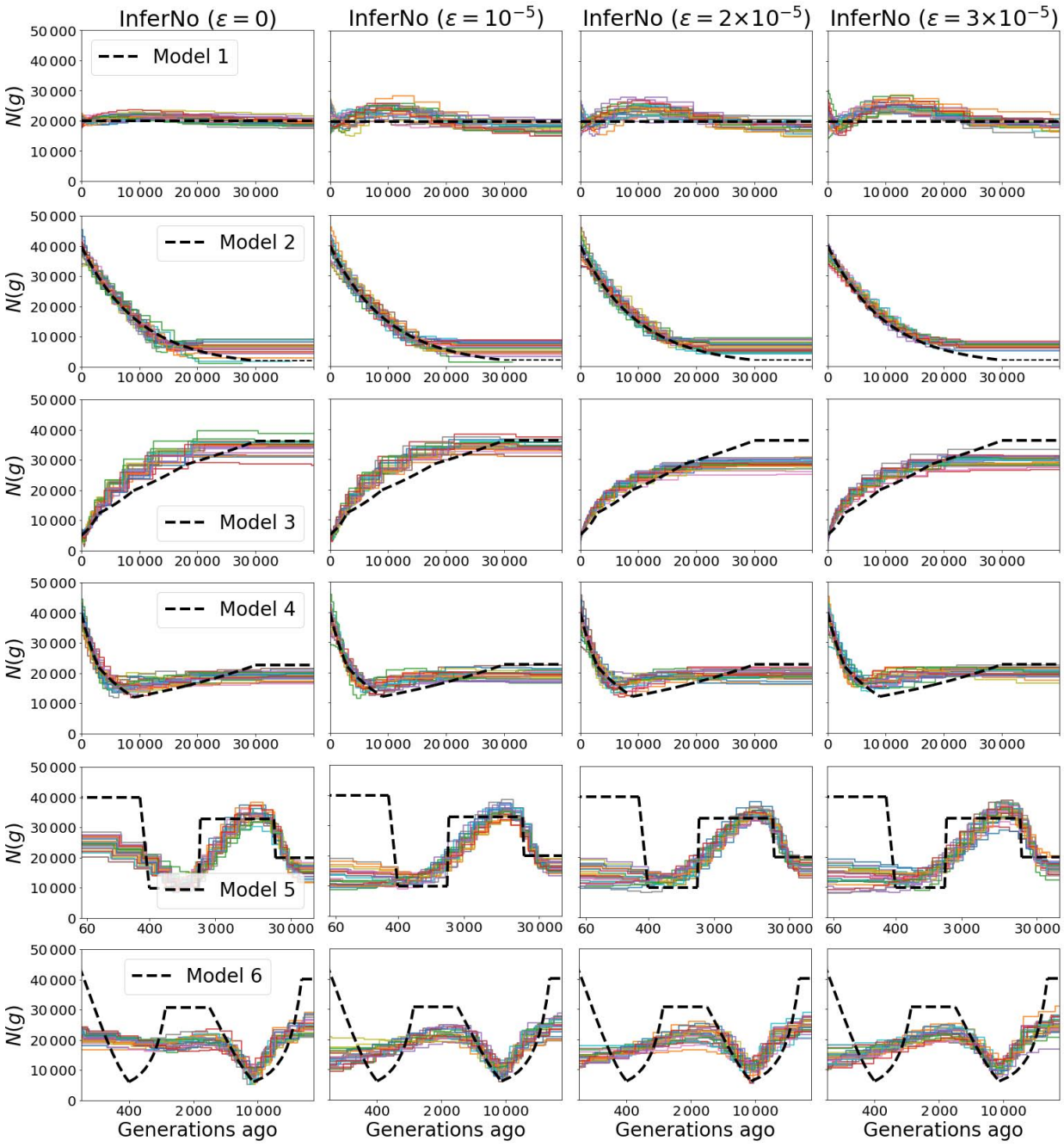

Figure S7: **Performance of InferNo with different levels of sequencing error**. Estimates of population size $N(g)$ with $n = 100$. Columns correspond to sequencing error rates $\epsilon = 0$, 1, 2, and 3 (per 10^5 generations). Rows correspond to simulation models in Table 1.

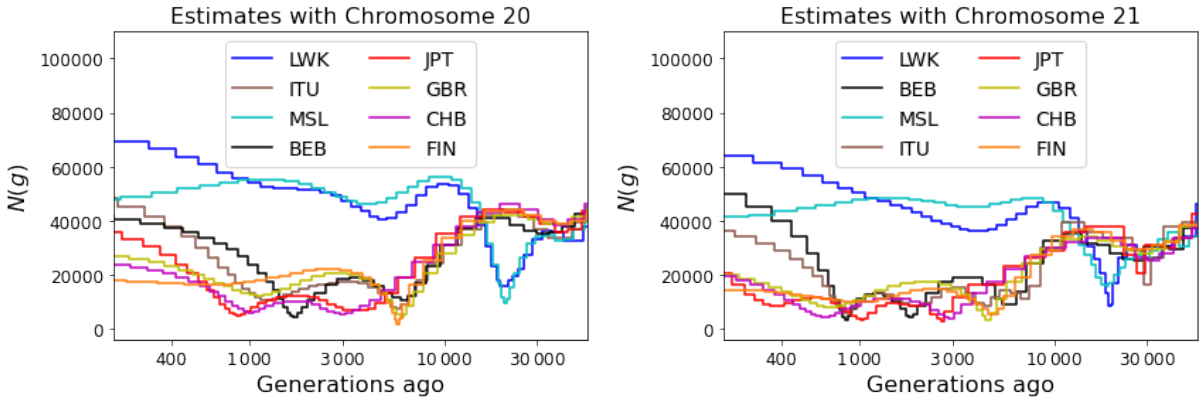

Figure S8: **InferNo estimates of population size $N(g)$ for eight 1KG populations**. The data are from Chromosome 20 (left) and Chromosome 21 (right). The time axis is on a logarithmic scale.

References

A. Bergström, S. McCarthy, R. Hui, M. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, H. Blanché, J. Deleuze, H. Cann, S. Mallick, D. Reich, M. Sandhu, P. Skoglund, A. Scally, Y. Xue, R. Durbin, and C. Tyler-Smith. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367:eaay5012, 2020.

A. Bhaskar, Y. R. Wang, and Y. S. Song. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome research*, 25(2):268–279, 2015.

S. Boitard, W. Rodríguez, F. Jay, S. Mona, and F. Austerlitz. Inferring population size history from large samples of genome-wide molecular data – an approximate Bayesian computation approach. *PLOS Genetics*, 12(3):e1005877, 2016.

S. R. Browning and B. L. Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *American Journal of Human Genetics*, 97(3):404–418, 2015.

Y. Deng, R. Nielsen, and Y. S. Song. Robust and accurate Bayesian inference of genome-wide genealogies for large samples. *bioRxiv*, pages 2024–03, 2024.

W. S. DeWitt, K. D. Harris, A. P. Ragsdale, and K. Harris. Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences*, 118(21):e2013798118, 2021.

T. Druet, I. Macleod, and B. Hayes. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity*, 112(1):39–47, 2014.

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

S. Fairley, E. Lowy-Gallego, E. Perry, and P. Flicek. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, 48(D1):D941–D947, 2020.

R. Fournier, Z. Tsangalidou, D. Reich, and P. F. Palamara. Haplotype-based inference of recent effective population size in modern and ancient DNA samples. *Nature Communications*, 14(1):7945, 2023.

R. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London* **A**, 344:403–10, 1994.

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

J. Kamm, J. Terhorst, R. Durbin, and Y. S. Song. Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, 115(531):1472–1487, 2020.

J. Kelleher, A. M. Etheridge, and G. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, 12(5):e1004842, 2016.

J. Kelleher, Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338, 2019.

H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

A. Mahmoudi, J. Koskela, J. Kelleher, Y.-b. Chan, and D. Balding. Bayesian inference of ancestral recombination graphs. *PLOS Computational Biology*, 18(3):e1009960, 2022.

M. Mezzavilla et al. Neon: An R package to estimate human effective population size and divergence time from patterns of linkage disequilibrium between SNPs. *Journal of Computer Science and Systems Biology*, 8(1), 2015.

P. F. Palamara, T. Lencz, A. Darvasi, and I. Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics*, 91 (5):809–822, 2012.

A. P. Ragsdale and S. Gravel. Models of archaic admixture and recent history from two-locus statistics. *PLOS Genetics*, 15(6):e1008204, 2019.

M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-wide inference of ancestral recombination graphs. *PLOS Genetics*, 10(5):e1004342, 2014.

E. Santiago, I. Novo, A. F. Pardiñas, M. Saura, J. Wang, and A. Caballero. Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. *Molecular Biology and Evolution*, 37(12):3642–3653, 2020.

S. Schiffels and R. Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925, 2014.

S. Schiffels and K. Wang. MSMC and MSMC2: The multiple sequentially Markovian coalescent. *Methods in Molecular Biology*, 2090:147–166, 2020.

G. Schwarz. The Bayesian information criterion. *Annals of Statistics*, 6:461–464, 1978.

L. Speidel, M. Forest, S. Shi, and S. R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, 2019.

J. Terhorst, J. A. Kamm, and Y. S. Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, 2017.

G. Upadhya and M. Steinrücken. Robust inference of population size histories from genomic sequencing data. *PLOS Computational Biology*, 18(9):e1010419, 2022.

R. J. Wang, S. I. Al-Saffar, J. Rogers, and M. W. Hahn. Human generation times across the past 250,000 years. *Science Advances*, 9:eabm7047, 2023.

A. L. Williams, G. Genovese, T. Dyer, N. Altemose, K. Truax, G. Jun, N. Patterson, S. R. Myers, J. E. Curran, R. Duggirala, et al. Non-crossover gene conversions show strong gc bias and unexpected clustering in humans. *Elife*, 4:e04637, 2015.

A. W. Wohns, Y. Wong, B. Jeffery, A. Akbari, S. Mallick, R. Pinhasi, N. Patterson, D. Reich, J. Kelleher, and G. McVean. A unified genealogy of modern and ancient genomes. *Science*, 375(6583):eabi8264, 2022.

J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998.