Estimating evolutionary and demographic parameters via ARG-derived IBD

Zhendong Huang[1*], Jerome Kelleher[2], Yao-ban Chan[1], David J. Balding[1]

1 Melbourne Integrative Genomics, School of Mathematics & Statistics, University of Melbourne, Australia
2 Oxford Big Data Institute, University of Oxford, United Kingdom

* huang.z@unimelb.edu.au

Abstract

Inference of demographic and evolutionary parameters from a sample of genome sequences often proceeds by first inferring identical-by-descent (IBD) genome segments. By exploiting efficient data encoding based on the ancestral recombination graph (ARG), we obtain three major advantages over current approaches: (i) no need to impose a length threshold on IBD segments, (ii) IBD can be defined without the hard-to-verify requirement of no recombination, and (iii) computation time can be reduced with little loss of statistical efficiency using only the IBD segments from a set of sequence pairs that scales linearly with sample size. We first demonstrate powerful inferences when true IBD information is available from simulated data. For IBD inferred from real data, we propose an approximate Bayesian computation inference algorithm and use it to show that poorly-inferred short IBD segments can improve estimation precision. We show estimation precision similar to a previously-published estimator despite a 4 000-fold reduction in data used for inference. Computational cost limits model complexity in our approach, but we are able to incorporate unknown nuisance parameters and model misspecification, still finding improved parameter inference.

Author summary

Samples of genome sequences can be informative about the history of the population from which they were drawn, and about mutation and other processes that led to the observed sequences. However, obtaining reliable inferences is challenging, because of the complexity of the underlying processes and the large amounts of sequence data that are often now available. A common approach to simplifying the data is to use only genome segments that are very similar between two sequences, called identical-by-descent (IBD). The longer the IBD segment the more informative about recent shared ancestry, and current approaches restrict attention to IBD segments above a length threshold. We instead are able to use IBD segments of any length, allowing us to extract much more information from the sequence data. To reduce the computation burden we identify subsets of the available sequence pairs that lead to little information loss. Our approach exploits recent advances in inferring aspects of the ancestral recombination graph (ARG) underlying the sample of sequences. Computational cost still limits the size and complexity of problems our method can handle, but where feasible we obtain dramatic improvements in the power of inferences.

Introduction

A common data-reduction technique when analysing samples of genome sequences is to identify identical-by-descent (IBD) genome segments [1–5]. In practice IBD is often identified by searching for regions with no evidence for recombination along two sequences since their most recent common ancestor (MRCA). Further, only IBD segments (IBDs) above a given length threshold, often 2 to 4 cM, are retained. This practice wastes valuable information, but has been necessary because the inference of short IBDs is too noisy to be useful for downstream analyses.

The ancestral recombination graph (ARG) is widely used to represent the genealogical history of a sample [6–8] and recent developments in inferring aspects of the ARG [9–13] now permit us to rapidly extract IBD directly from inferred shared ancestors, without requiring zero recombination. Further, computationally fast ARG inference and extraction of IBD can be implemented within an approximate Bayesian computation (ABC) algorithm which removes the need for an information-wasteful length threshold. Instead, we reduce computational cost by using an efficient subset of IBDs that scales linearly with sample size with little information loss relative to using all IBDs.

Our approach relies on an efficient data structure encoding features of an ARG underlying a sample of genome sequences, called the succinct tree sequence (TS) [14]. The TS minimises redundant storage of subsequences that are similar due to shared ancestry. It has led to spectacular improvements in storage and simulation of large genome datasets [15], and has recently been applied to IBD-based inferences about demographic history and evolutionary parameters [16].

We first demonstrate powerful inferences of mutation and sequencing error rates, TMRCA (time since the MRCA), and past and present population sizes, given true IBD information in simulation studies. For real datasets, we propose TSABC: ABC with statistics computed from IBDs extracted from an inferred TS. We demonstrate the performance of TSABC with inferences of the mutation rate and population size in simulation studies and real data, and we compare mutation rate estimates with previously-published results and with analyses using a range of IBD length thresholds.

We find that using IBDs extracted from an inferred ARG leads to a surprisingly small loss of precision relative to use of true IBDs. Further, even a low threshold on IBD length reduces the quality of inferences, despite the fact that short IBDs are poorly inferred. TSABC is computationally demanding, which limits the size and complexity of inference problems that can be tackled. However, TSABC can achieve comparable results to previous estimators using much smaller data sets: we show similar precision to a previously-published estimator despite a 4 000 fold reduction in data available for inference.

Methods

Definition and notations

The TS encodes genome sequence data efficiently by storing subsequences that are similar due to shared ancestry as variations of an ancestral sequence. It is defined [17] as $\{\mathcal{C}, P, E, \mathcal{M}\}$, where $\mathcal{C} = \{1, \ldots, m\}$ is the set of leaf (or tip) nodes corresponding to m observed sequences each of length ℓ, and $P = \{m{+}1, \ldots, n\}$ is the set of internal (ancestral) nodes of the TS ordered backwards in time from the present. An edge in $E = \{(c_i, p_i, l_i, r_i) : i = 1, 2, \ldots, I\}$ represents inheritance of sites in the segment $[l_i, r_i]$, with $1 \leq l_i \leq r_i \leq \ell$, from internal node $p_i \in P$ to its child $c_i \in \{1, \ldots, p_i{-}1\}$, while $\mathcal{M} = \{(c_j, s_j) : j = 1, 2, \ldots\}$ stores the set of sites s_j at which there is a sequence

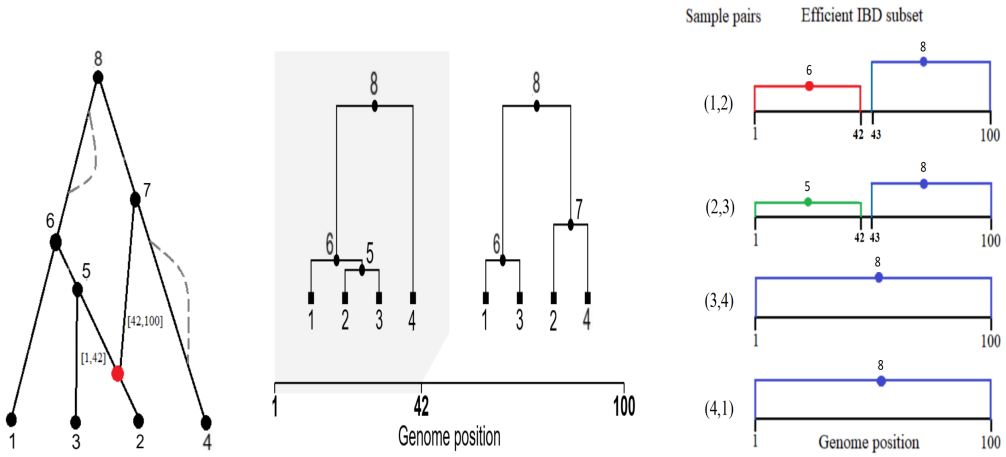

Fig 1. An ancestral recombination graph (ARG) spanning a genome sequence of length $\ell = 100$ (left), the corresponding sequence of local trees (middle) and efficient IBD subset (right). The ARG has leaf nodes $\{1, 2, 3, 4\} = \mathcal{C}$, named ancestral nodes $\{5, 6, 7, 8\} = P$, and a recombination at site 42 of an internal ancestral node (red dot). The two dashed lines in the ARG represent inheritance paths due to two ineffective recombination events, which are not represented in the TS. The efficient IBD subset includes two IBD segments for the node pair $(1, 2)$, corresponding to intervals $[1, 42]$ and $[43, 100]$ which have MRCA 6 and 8, respectively, and one IBD segment spanning the whole sequence for pairs $(3, 4)$ and $(4, 1)$.

difference between c_j and its parent, due either to a mutation or, if c_j is a leaf node, sequencing error. The TS has the "succinct" property that any tree component conserved over a genome segment is stored only once, which greatly reduces data storage requirements compared with retaining all distinct marginal trees.

Identity by descent and efficient subsets

We denote the ith IBD segment in the TS by $\text{IBD}_i = (c_{i1}, c_{i2}, l_i, r_i, p_i, M_i)$, $i = 1, \ldots, I$, ordered such that c_{i1} is non-decreasing in i. Here c_{i1} and c_{i2} are the leaf nodes of the two sequences, $[l_i, r_i]$ is the IBD genome segment, p_i is the MRCA node of c_{i1} and c_{i2} for this segment, and M_i denotes the set of sites in $[l_i, r_i]$ at which c_{i1} and c_{i2} differ. As there is no length threshold, the IBDs of any sequence pair partition the genome: every sequence site is included in exactly one of the IBD segments.

Each IBD_i has the same MRCA at each site in $[l_i, r_i]$, and a different MRCA at adjacent sites. Imposing a no-recombination requirement as part of the definition of IBD would be more restrictive, since the absence of recombination implies a common MRCA but the reverse does not hold (see Figure 1, left, for examples of recombinations that do not change the MRCA).

To reduce computational effort, we use for inference only an "efficient" subset of IBDs. After fixing an arbitrary order for the sequences, we include in the subset only the IBDs of the sequence pairs $(1, m)$ and $(c, c{+}1)$ for $c = 1, \ldots, m{-}1$ (see Figure 1, right, and Appendix S1). An efficient subset has the property that each edge of the TS is included in a descent path from the MRCA for at least one IBD segment in the subset, which ensures that information is retained in the subset about every mutation.

Imposing a length threshold on IBDs is also a form of data reduction but we show below that it can lead to high information loss, because mutations are ignored if they occur at sites not contained in a sufficiently long IBD segment.

Estimation

Let μ and ϵ be the per-site per-generation mutation rate and the per site sequencing error rate, both assumed constant over sites. For $i = 1, \ldots, I$, let g_i denote the age of p_i in generations, and let $N(g)$, $g = 0, 1, 2, \ldots$, be the population size g generations in the past. In Appendix S2 we derive method-of-moment estimators for μ and ϵ, and non-parametric estimators of $g_i, i = 1, \ldots, I$, and $N(g)$, $g = 0, 1, 2, \ldots$, based on statistics computed from IBD lengths. We investigate the performance of these estimators when true IBD information is available in simulation studies. The recombination rate r is assumed constant over sites and known for all inferences; the extension to a known recombination map is straightforward (a recombination at site s means between sites s and $s + 1$).

For observed sequence data, true IBD information is not available and we extract IBDs from an inferred TS. TSABC uses summary statistics derived from these IBDs and related to the method-of-moments estimators. For inference of μ and ϵ, we use the statistics $\bar{M} \times I$ and C_1 (Appendix S2.1) which are linear transformations of the method-of-moments estimators $\hat{\mu}$ and $\hat{\epsilon}$. Nonparametric estimation of $N(g)$ is not feasible, but we can estimate the parameters of a demographic model, which allows powerful inference provided that the model is adequate. We use as statistics the mean and standard deviation (SD) of IBD lengths $r_i - l_i$, $i = 1, \ldots, I$.

Simulation study design: true IBD available

Evolutionary parameters and sample properties		
Symbol	Definition	Value(s) in simulations
μ	mutation rate	1.3×10^{-8} per site per generation
ϵ	sequencing error rate	up to 10^{-3} per site
r	recombination rate	10^{-8} per site per generation
m	sample size	10, 20, 40, 60, 80, 100, 160 or 200 sequences
ℓ	sequence length	10^6, 10^7 or 10^8 sites
Demographic models ($N(g)$ = population size g generations ago)		
Model C	$N(g)$ constant	$N(g) = 2 \times 10^4$
Model G	$N(g) = N(0) \times e^{-\tau g}$	$N(0) = 10^6$, $\tau = 10^{-4}$ (Model Ga) $N(0) = 2 \times 10^5$, $\tau = 10^{-3}$ (Model Gb)
Model S	$N(g) = N(0)$ for $0 < g < G$ $= N(G)$ for $g \geq G$	$N(0) = G = 4 \times 10^4$ $N(G) = 10^4$
Model EA	European-American demographic history [18]	See Figure S2 for $N(g)$ values; gene conversion: tract length 100 bp rate 10^{-8}/site/generation.

Table 1. Parameter values, sample properties and demographic models for the simulation study. Unless otherwise stated, 25 simulation replicates were generated in each scenario. Model Ga is used for inferences given true IBD and Model Gb is used for inferences from inferred IBD. The value of r is assumed known for all inferences, whereas μ, ϵ and $N(g)$, $g \geq 0$, are targets of inference.

We jointly estimated μ, ϵ, g_i, $i = 1, \ldots, I$, and $N(g)$, $g \geq 0$, using our novel estimators. We used msprime [19] to generate TS under the coalescent with recombination model [20, 21], assuming demographic models C, Ga and S (Table 1). From each TS we extracted an efficient subset of IBDs (Algorithm 1). Sequencing error was simulated by adding elements to $\mathcal{M}$ at leaf nodes of the generated TS. At the largest error rate ($\epsilon = 10^{-3}$), any singleton variant is a few times more likely to arise from sequencing error rather than a mutation.

Simulation study design: inferred IBD

We used msprime to generate simulated sequences, recoded them as binary strings using 0 and 1 for the ancestral and derived alleles and added sequencing errors by assigning 1 to randomly selected sites at rate ϵ (see [22] for alternative models of sequencing error). We choose tsinfer [10] to infer the TS from the resulting sequence data; speed is critical for an ABC algorithm, and tsinfer is the fastest of the current methods, while retaining high accuracy [13, 23]. Unless otherwise stated, in each scenario we used $\eta = 2\,500$ simulations with ABC acceptance rate 0.05 (125 acceptances).

We first use simulations to confirm previous reports [24] that the quality of IBD inference is often poor, particularly for short IBDs. We compared the number of true and inferred IBDs for datasets simulated under Model C with μ ranging from 1 to 20 units of 10^{-8} per site per generation and $m = 10$, 20 and 160. We also compared the length distribution of true and inferred IBDs for $m = 160$ and $\mu = 1.3 \times 10^{-8}$.

To investigate the effect of including short IBDs, both true and inferred, we also modified TSABC to include only IBDs with length greater than a threshold of 1, 2 or 4 units of 10^4 bp. When a threshold was applied, we included all IBDs satisfying the threshold, rather than using only the efficient subset of IBDs.

We next investigated TSABC estimation of μ under Model C and Model Gb with $\ell = 10^7$. The $N(g)$ values and $\epsilon = 0$ were assumed known for the inference and we adopted a Uniform(10^{-8}, 2×10^{-8}) prior distribution for μ. For the Model C simulations with $m = 10$, we also applied TSABC after thresholding on IBD length and repeated using true IBD extracted from the msprime simulations.

To study TSABC estimation of the population size $N(g)$, we used $m = 200$ and $\ell = 10^6$ under each of Model C and Model Gb. For both data simulation models, the TSABC inference used Model G but with different prior distributions. When the simulation model was Model C, we fitted Model G with independent prior distributions Uniform(10^4, 3×10^4) for $N(0)$ and Uniform(-2×10^{-5}, 2×10^{-5}) for τ. Whenever $\tau < 0$, we impose a population size limit $N(g) \leq 2 \times N(0)$. With simulation model Model Gb, the independent prior distributions were Uniform(10^5, 3×10^5) for $N(0)$, and Uniform(0, 0.002) for τ. All parameters were treated as known except the targets of inference $N(0)$ and τ.

We performed additional simulations to allow comparison with the inferences of μ reported by [18]. Data were simulated under Model EA, which aims to capture key features of the demographic history of European-Americans (Table 1), and Model C modified to include sequencing errors. The Model C simulations of [18] used $\epsilon = 10^{-4}$ but no gene conversion, while for Model EA they set $\epsilon = 0$ and included gene conversion. We include both sequencing error and gene conversion in both Model C and Model EA simulations. We used a 400-fold smaller sample size than [18] ($m = 10$ versus $m = 4 \times 10^3$) and 10-fold smaller genome length ($\ell = 10^7$ per chromosome, versus $\ell = 10^8$).

We did not include gene conversion in the TSABC inference, thus challenging it with model misspecification. As a further challenge, we treated $N(g)$ as unknown when inferring μ, and misspecified the model for $N(g)$ in the TSABC simulations.

When the data were simulated under Model C, TSABC used independent prior distributions Uniform(10^{-8}, 2×10^{-8}) for μ and Uniform(0.6×10^{-4}, 1.6×10^{-4}) for ϵ. For $N(g)$, we adopted Model G with independent priors $N(0) \sim$ Uniform(14 000, 30 000) and $\tau \sim$ Uniform(-2×10^{-5}, 10^{-5}).

When the data were simulated under Model EA, TSABC used a Uniform(10^{-8}, 2×10^{-8}) prior distribution for μ. For inference of $N(g)$, we adopted Model S with independent prior distributions $N(0) \sim$ Uniform(11 000, 15 000), $G \sim$ Uniform(4500, 6500) and $N(G) \sim$ Uniform(45 000, 49 000).

Mutation and growth rates in the 1000 Genome Project

We analyse chromosomes 20 and 21 from 8 of the 26 human populations of the 1000 Genomes Project (1KGP) [25] making use of the demographic model of [26] which we refer to as the 1KGP model. See Figure S2 for plots of the 1KGP model and Appendix S3 for details of the data analysis. Separately for each chromosome, we use TSABC to infer μ assuming the prior Uniform$(10^{-8}, 2 \times 10^{-8})$ and the 1KGP model. The 16 sets of 125 accepted values were analysed in a two-way ANOVA to assess differences in μ across chromosomes and over populations.

Next, we use chromosome 20 and 21 data to estimate population size $N(g)$ assuming the 1KGP model for $g \geq 1000$ and fitting demographic Model G for $0 \leq g \leq 1\,000$, constrained such that $N(1000)$ in Model G matches the 1KGP model value. The constrained Model G has one free parameter $N(0)$, for which we adopt a Uniform$(10000, 240000)$ prior distribution. To reduce computational effort with little loss of information, in both the observed dataset and TSABC simulations we removed SNPs with a minor allele count > 40, which typically arose at $g \gg 1\,000$. We estimate $N(g)$ from each chromosome separately and average the results.

Results

Simulation study results: true IBD available

While use of the efficient subset of IBDs reduces computational cost in proportion to the reduction in sequence pairs from $m(m-1)/2$ to m, the average estimated SD of $\hat{\mu}$ in our simulation study increased only slightly, from 0.017 to 0.019 units of 10^{-8} (see also Figure S3, left panel). This gain in computation time is typically worth the small loss of statistical efficiency.

Both $\hat{\mu}$ and $\hat{\epsilon}$ are well estimated in all demographic models, with no indication of bias (Figure 2). Increasing m has only a modest effect on the SD of estimators, whereas ℓ has a larger effect (SD scales with $\sqrt{\ell}$, Figures 2, S3 (right) and S4). Sequencing errors only inflate the number of singleton variants, so $\hat{\mu}$ is little affected by increasing ϵ (Figure 2).

Although individual $\hat{g}_i$ are not precise, the empirical and theoretical densities obtained from all $\hat{g}_i$, $i = 1 \ldots, I$, are close (Figure 3) despite the TS used for input only including information about the order of the coalescent events, and not their times. The population size estimator $\hat{N}(g)$ is accurate under all models, at least for $g \leq 5 \times 10^5$ (Figure 4). Figures S3 (right) and S4 show more precise estimates of $\hat{\mu}$ and $N(g)$, with a longer sequence length.

Simulation study results: inferred IBD

The number of inferred IBDs tends to increase with both μ and m, but except for very high μ (over 10 times the average human value when $m = 160$) it remains well below the true number of IBDs (Figure 5, left). Correspondingly, the length distribution of inferred IBDs is highly skewed towards larger values relative to the true distribution (Figure 5, right), as previously reported [27, 28]. Despite this poor detection of small IBDs, and consequent tendency for inferred IBDs to be longer than the true IBDs, Table 2 shows that each increase in the length threshold reduced the precision of inference, both for inferred and true IBD, so that even poorly-inferred short IBDs do contribute useful information for inference. We also see in Table 2 (final column) further evidence that use of the efficient subset of IBDs leads to only a small loss of statistical efficiency. As expected, the use of true IBD improves TSABC compared with using inferred IBD, but the magnitude of the improvement is modest in the case of standard

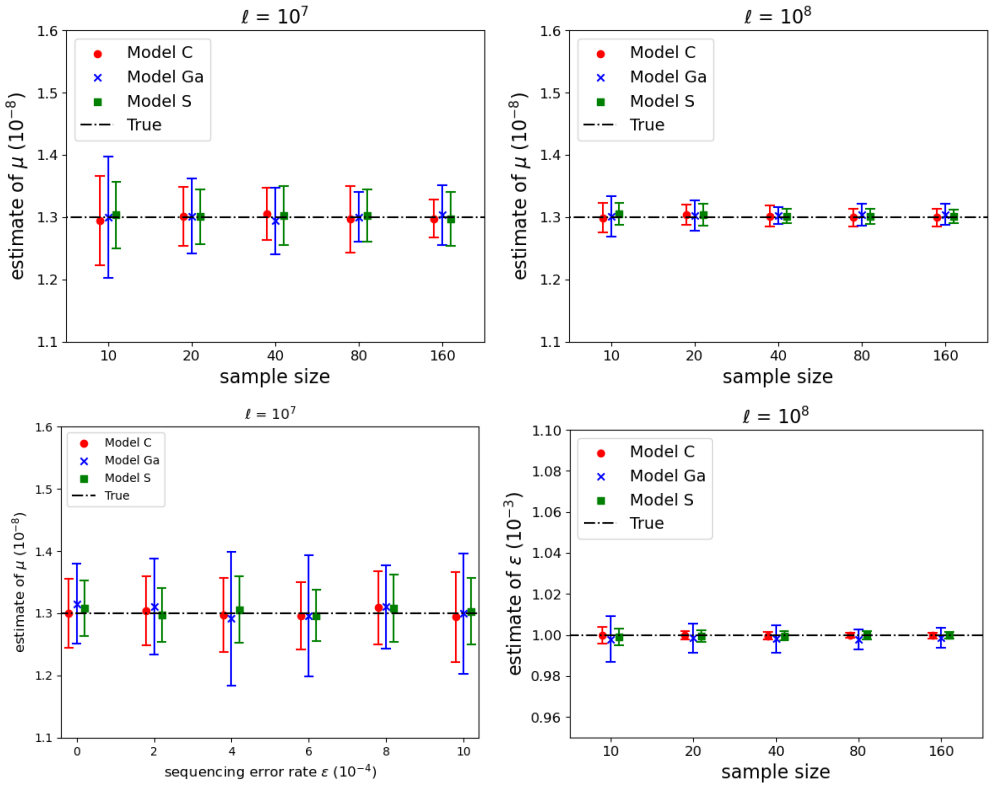

Fig 2. Inference of mutation rate μ and sequencing error rate ϵ with two sequence lengths (columns), when true IBD was available for inference. Line segments show indicative 95% CIs computed from the average estimate (indicated by a symbol, see legend box) and the empirical SD of the estimates from 25 simulated datasets in each scenario. Bottom left panel shows the impact of ϵ on $\hat{\mu}$ when $m = 10$, in the other three panels $\epsilon = 10^{-4}$.

TSABC (threshold = 0). For higher thresholds, bias can be high due to low precision of inference and the prior boundary at 10^{-8}.

Although TSABC can provide approximations to the full posterior distribution, we report here only posterior mean estimates of unknown parameters. For inference of μ, it appears that any bias of TSABC is small for both models (Figure 6). Some under-estimation is expected because the binarisation of the sequence data obscures instances of multiple mutations at the same site, but this effect is negligible.

Parametric estimation of $N(g)$ also performs well (Figure 7). When the data simulation model was Model C, the average estimate of $N(0)$ (true value 20 000) over the 25 replicates is 20 931 with standard error (SE) $1\,428/\sqrt{25} = 286$, while for the growth rate τ (true value 0) the average estimate $\pm$ SE is $2.10 \pm 1.8/\sqrt{25} = 0.36$ (in units of 10^{-6}). When the data simulation model was Model Gb, for $N(0)$ (true value 200 000) we obtained $202\,534 \pm 2\,173$ while for τ (true value 1) we obtained 1.08 ± 0.07 (in units of 10^{-3}).

Table 3 shows that TSABC performs similarly to the results reported by [18] despite a 4 000-fold reduction in data used for inference, and despite the challenges we imposed on TSABC: gene conversion was incorporated in data simulation models but not the ABC inference simulations, and the latter also used a misspecified demographic model.

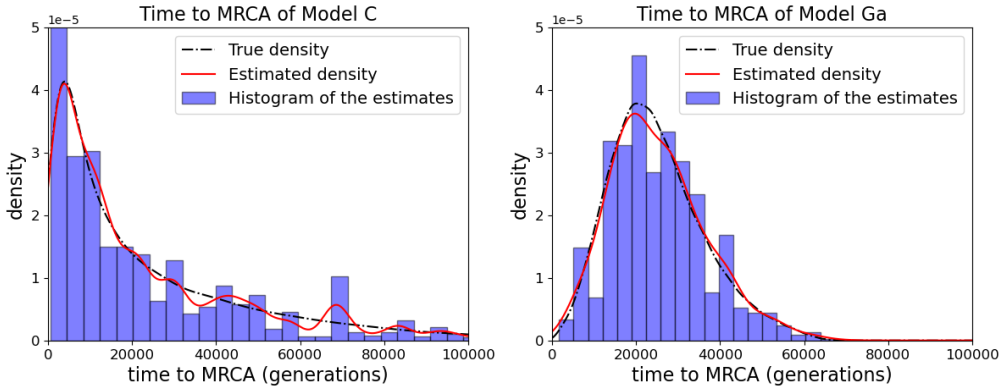

Fig 3. Histogram of the $\hat{g}_i$, $i = 1, \ldots, I$, obtained from one sample simulated under each of Model C (left) and Model Ga (right), with sample size $m = 80$, sequence length $\ell = 10^8$ and sequencing error rate $\epsilon = 10^{-3}$. Also shown is a probability density obtained by kernel smoothing of the $\hat{g}_i$ together with the true density. True IBD was available for inference but no time information.

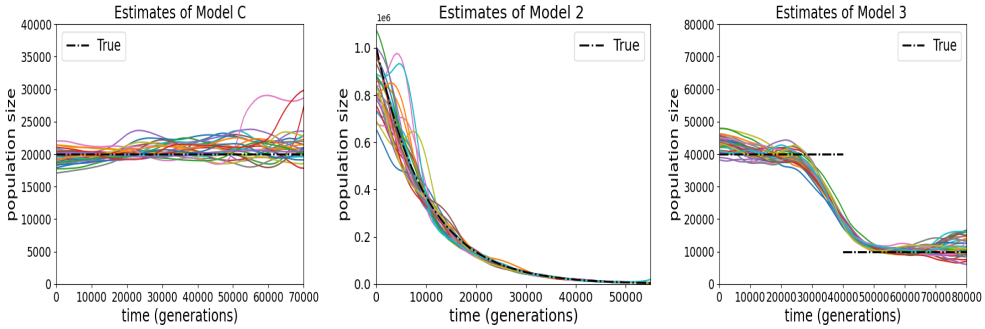

Fig 4. Estimates of the population size $N(g)$, $g \geq 0$, from each of 25 simulation replicates under Model C, Model Ga and Model S, when true IBD was available for inference. Sequence length is $\ell = 10^8$, sequencing error rate is $\epsilon = 10^{-3}$ and sample size is $m = 80$.

1000 Genomes data analysis

The global mean $\hat{\mu}$ over the two chromosomes and eight populations is 1.27×10^{-8} (Table 4), similar to previous estimates assuming μ to be constant over populations [26, 29, 30], and also those finding small between-family differences in μ [31, 32]. A two-way ANOVA revealed no significant difference between the two chromosomes, but highly significant differences across populations, which may be due to differences in heritable factors or environmental exposures.

Figure 8 shows positive growth in the past 1 000 generations for all eight populations. CHB and BEB (both in Asia) have the highest $N(0)$ while MSL and LWK (both in Africa) have the lowest $N(0)$ despite having the highest values of $N(1000)$. These findings are consistent with the results of [26] for $400 \leq g \leq 1\,000$, and the recent growth estimates obtained using Relate [11, Figure 3].

Discussion

We have shown that ARG-derived IBD combined with ABC can deliver big advantages over previous IBD-based methods for inferring evolutionary and demographic

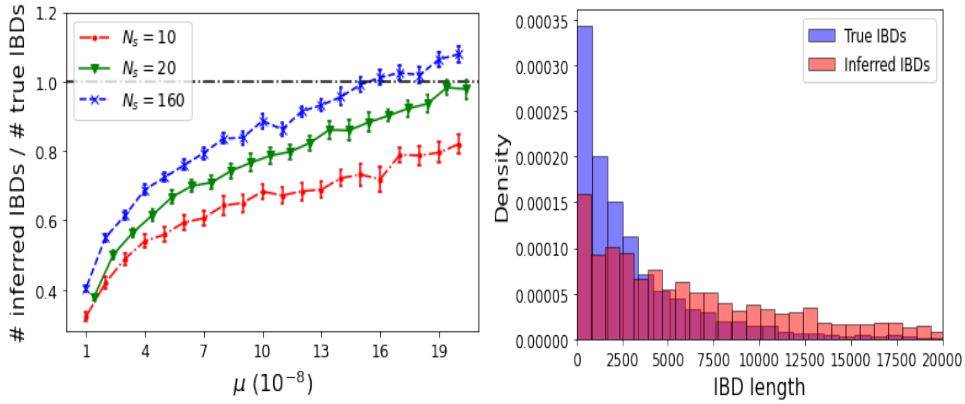

Fig 5. Comparison of true and inferred IBDs. Left: each symbol and vertical line segment shows the mean and 95% CI of the mean ratio of IBD counts over 25 Model C simulations with sample sizes $m = 10$, 20 and 160. Right: histograms of true and inferred IBD length distributions for a Model C simulated dataset with $m = 160$ and sequence length $\ell = 10^6$.

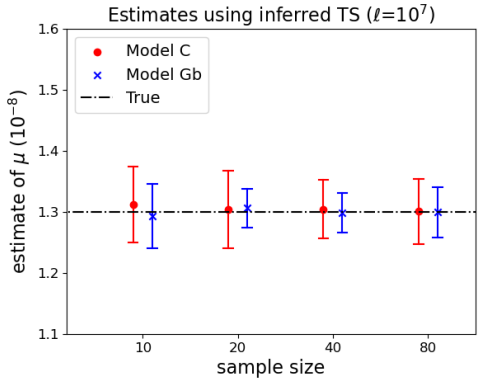

Fig 6. TSABC estimation of mutation rate μ. Symbols and line segments show mean and 95% CI over 25 simulations with no sequencing error ($\epsilon = 0$) and sequence length $\ell = 10^7$.

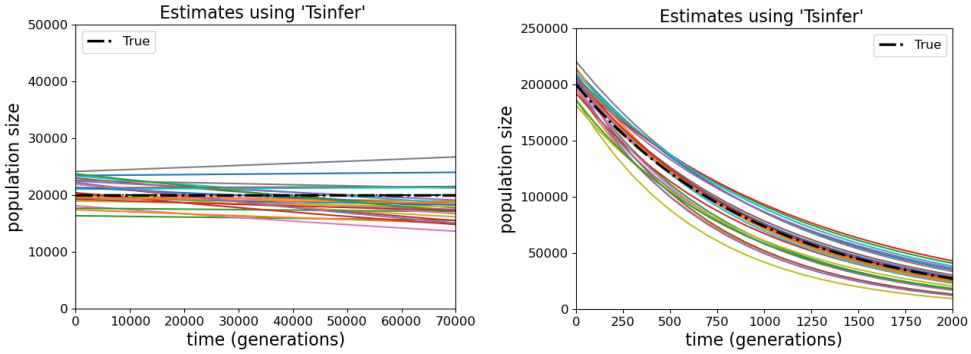

Fig 7. Fitted exponential curves for the population size $N(g)$ obtained using TSABC. Each of the 25 curves corresponds to a dataset simulated under Model C (left) and Model Gb (right) with no sequencing error ($\epsilon = 0$), sample size $m = 200$ and sequence length $\ell = 10^6$.

Threshold (10^4 bp)	4	2	1	0	
inferred IBD					
# IBD	1 001	7 033	24 683	48 289	(10 667)
$\hat{\mu}$ (10^{-8})	1.51	1.44	1.32	1.30	(1.31)
SD (10^{-8})	0.219	0.167	0.067	0.041	(0.043)
true IBD					
# IBD	478	1 803	6 940	121 027	(26 394)
$\hat{\mu}$ (10^{-8})	1.33	1.30	1.31	1.30	(1.29)
SD (10^{-8})	0.141	0.089	0.064	0.029	(0.034)

Table 2. Comparison of TSABC inference for μ using different IBD length thresholds. Each result is an average over 25 Model C simulation replicates with $m = 10$ and $\epsilon = 0$. In the last column, values based only on IBDs in the efficient subset are given in ().

	Model C		Model EA			
	$\hat{\mu}$	SD	$\hat{\mu}$	SD	m	ℓ
[18]	1.30	0.020	1.34	0.007	4 000	10^8
TSABC	1.30	0.017	1.28	0.007	10	10^7

Table 3. Comparison of TSABC inference of μ (in units of 10^{-8}) with results reported in [18]. TSABC results are obtained from 25 simulated datasets under each model.

parameters from a sample of genome-wide sequences, including nonparametric estimation of past population sizes. Despite verifying that IBD extracted from an inferred TS is often inaccurate, we have shown that it provides powerful inferences for the mutation rate and historic population sizes. For example, we obtained similar estimation results to a previous study that used 4 000 times more data for inference. These advantages arise because we can define IBD in terms of a common MRCA, avoiding both the problem of detecting recombinations and the need for a minimum IBD length. Further, we require only IBDs from m sequence pairs, rather than all $m(m-1)/2$ pairs, which reduces computational effort with little loss of statistical efficiency.

We illustrated our TSABC approach in simple scenarios, finding that it suffers only modest loss of efficiency relative to using true IBD. Importantly, removing IBDs with length below even a low threshold reduces the precision of inferences despite the poor quality of ARG-based IBD inferences.

TSABC can be computationally demanding for complex demographic models, and

		MSL	LWK	BEB	ITU	FIN	GBR	JPT	CHB
	Sample size:	170	198	172	204	198	182	208	206
Chr 20	$\hat{\mu}$	1.27	1.24	1.23	1.22	1.32	1.36	1.32	1.20
	SE	0.004	0.004	0.004	0.004	0.004	0.011	0.004	0.004
Chr 21	$\hat{\mu}$	1.25	1.26	1.21	1.29	1.29	1.35	1.33	1.24
	SE	0.004	0.005	0.006	0.005	0.006	0.006	0.006	0.006
Combined	$\hat{\mu}$	1.26	1.25	1.22	1.25	1.31	1.36	1.33	1.22
	SE	0.003	0.003	0.003	0.004	0.004	0.006	0.004	0.004

Table 4. Estimates of the posterior mean and SE of the mutation rate per site per generation (in units of 10^{-8}) on human chromosome 20 and 21 for populations MSL (Mende in Sierra Leone), LWK (Luhya in Webuye, Kenya), BEB (Bengali from Bangladesh), ITU (Indian Telugu from the UK), FIN (Finnish in Finland), GBR (British in England and Scotland), JPT (Japanese in Tokyo, Japan), and CHB (Han Chinese in Beijing, China). The TSABC analysis assumes the 1KGP demographic model in each population.

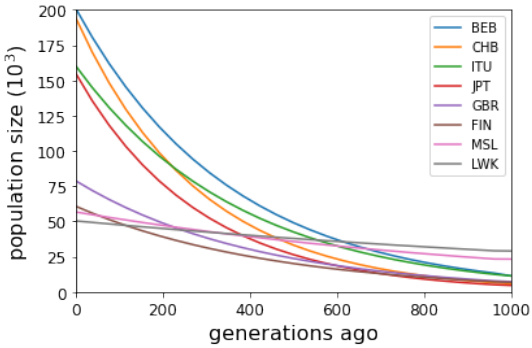

Fig 8. Estimates of recent population sizes for eight populations sampled in the 1000 Genomes Project (curves are shown in order of decreasing $N(0)$). See Table 4 caption for explanation of the population labels.

the results presented here are limited to inferring the mutation rate and two parameters of a demographic model. However, we were able to incorporate unknown nuisance parameters such as the sequence error rate and misspecification of the demographic model to challenge TSABC inference without substantial detriment to inference quality.

Our results open the way for more powerful demographic and evolutionary inferences from samples of genome sequences than have previously been available.

Data availability

Data and code used here are available at: github.com/ZhendongHuang/Estimating_evolutionary_and_demographic_parameters_Huang

Acknowledgment

ZH is funded by Australian Research Council grant DP210102168 awarded to YBC, DJB and JK. JK acknowledges support from the Robertson Foundation, US National Institutes of Health (grants HG011395 and HG012473) and UK Engineering and Physical Sciences Research Council (grant EP/X024881/1).

Supporting Information

S1 Appendix: The efficient IBD subsets and TSABC algorithms

Algorithm 1 searches the IBDs in the order of their children. For instance, the algorithm first starts to search all IBDs corresponding to the sequence pair $(1, 2)$ along the whole sequence, from left to right. Then it searches the IBDs for $(2, 3)$ and so on until the final sequence pair $(m, 1)$.

Algorithm 2 adopts a standard ABC approach, the innovation here is in the choice of summary statistics which are derived from an inferred TS.

S2 Appendix: Derivation of Estimators when TS is known

S2.1 Mutation and sequencing error rates

The right endpoint r_i of an IBD segment is usually the site of an effective recombination event for (c_1, c_2), where "effective" means that c_1 and c_2 have a different MRCA on

Algorithm 1 The efficient IBD subsets algorithm

1: Initialize $c = 0$. Let $p_1 = c$ and $p_2 = c{+}1$ be the current children and parents of the edge chains. Let $l_t = 0$ and $r_t = \ell$ be the current potential left and right end point of the IBD. Let $p = \min(p_1, p_2)$. Let $S = \emptyset$ be a set storing candidate IBDs.
2: Search edges in TS. Let c_e, p_e, l_e and r_e be the child, parent, left and right end points of an edge, respectively. If $c_e = p$ and $l_e \leq l_t < r_e$, update $r_t = \min(r_e, r_t)$. Go to Step 3.
3: If $p_e \neq \max(p_1, p_2)$, go to Step 4. Else, we find a candidate IBD segment and add it to the set S. Update $l_t = r_t$ and $r_t = \ell$. If $l_t \neq \ell$, update $p_1 = c$, $p_2 = c{+}1$ and $p = \min(p1, p2)$. Go to Step 2. Else, go to step 5
4: If $p_1 < p_2$, update $p_1 = p_e$. Else, update $p_2 = p_e$. Let $p = \min(p_1, p_2)$. Go to Step 2.
5: Combine neighboring candidate IBDs in S with the same parent. Update $c = c{+}1$, $p_1 = c$ and $p_2 = c{+}1$. If $c < m$, go to Step 2. Else, stop.

Algorithm 2 TSABC algorithm for parameter θ

1: Initialization. Set the number of simulation replicates η, and a prior distribution for θ with density f_θ. Choose summary statistics and write s_0 for their value computed from the observed data. Set a tolerance level $\varepsilon > 0$.
2: For $= 1, \ldots, \eta$, generate θ_i from f_θ. Simulate a sample of m genomes with $\theta = \theta_i$ and length ℓ, infer the TS from the simulated genomes and then compute the summary statistics s_i. Apply the linear adjustment to θ_i [33] by letting $\theta_i^* = \theta_i - (s_i{-}s_0)'\hat{\beta}$, where $(\hat{\alpha}, \hat{\beta}')' = \text{argmin}_{(\alpha,\beta)} \sum_{i=1}^{\eta} \{\theta_i - \alpha - (s_i{-}s_0)'\beta\}^2$.
3: Accept θ_i^* if $d(s_i{-}s_0) < \varepsilon$, where $d(\cdot)$ is the estimated Mahalanobis distance.
4: The accepted values of θ^* can be regarded as a sample from the posterior distribution. We use the sample mean as an estimator of the posterior mean.

either side. The exceptions are IBDs terminating at the sequence end site ℓ, which are excluded from the derivations below but, as they are rare, there is little impact if they are included in practice. For the derivation of our estimators, we also assume that at most one mutation occurs at each site since the MRCA of the sequences at that site. The effect of this assumption is minor when the mutation rates is low, which is the case for humans and many other organisms. We further assume that sequencing errors occur independently at rate ϵ at each site, and do not occur at sites with mutations.

Suppose that a recombination occurs at site s of sequence c_1 creating two subsequences going backward in time, in the intervals $[1, s]$ and $[s{+}1, \ell]$. This recombination is effective for (c_1, c_2) if and only if one of these subsequences coalesces (reaches a common ancestor) with c_2 before coalescing with the other subsequence. By symmetry, all three possible coalescence events are equally likely, and so the recombination has probability 2/3 of being effective. Mutation and effective recombinations occur independently at each site of c_1 and c_2. Given that one of these events occurs at a site before the two sequences reach their MRCA, with probability $(2r/3)/(\mu{+}2r/3)$ it is an effective recombination. Thus the expected number of mutations that occur in the segment before it is terminated by an effective recombination is one less than the mean of a geometric distribution with parameter $(2r/3)/(\mu{+}2r/3)$, which is $3\mu/2r$. Site differences in IBDs can also arise from sequencing errors, which occur with rate 2ϵ per site.

Let $\bar{L} = \sum_{i=1}^{I}(r_i{-}l_i)/I$ and $\bar{M} = \sum_{i=1}^{I} |M_i|/I$ denote the averages of the IBD segment length and the number of sites that differ in an IBD segment, respectively. Our first estimating equation is

$$\bar{M} = \frac{3\hat{\mu}}{2r} + 2\bar{L}\hat{\epsilon}. \tag{1}$$

which can be read intuitively as site differences = mutations + sequencing errors. Our second estimating equation has a similar interpretation, but is based on site differences on each sequence relative to both its neighbours in the efficient subset, rather than between pairs of sequences:

$$C_1 = C_2 \frac{\hat{\mu}}{r} + m\ell\hat{\epsilon}, \tag{2}$$

where we define

$$C_1 = \frac{3}{2} \sum_{\substack{i,j=1 \\ i<j}}^{I} \sum_{s=1}^{\ell} \Big\{ \mathbb{I}(s \in M_i \cap M_j) \times \mathbb{I}(c_{i2} = c_{j1}) \times \mathbb{I}(p_i \neq p_j) \Big\} \tag{3}$$

$$C_2 = \frac{9}{4} \sum_{\substack{i,j=1 \\ i<j}}^{I} \sum_{\substack{i',j'=1 \\ i'<j'}}^{I} \sum_{s=1}^{\ell} \Big\{ \mathbb{I}(c_{i2} = \ c_{j1} = c_{i'2} = c_{j'1}) \times \mathbb{I}(s = r_i = r_j = l_{i'} = l_{j'}) \times$$

$$\mathbb{I}(p_i \neq p_j \cup p_{i'} \neq p_{j'}) \Big\}. \tag{4}$$

These quantities estimate, respectively, the total number of sequencing errors and mutations (C_1) and the number of recombinations (C_2) on the branch immediately above sequence c, before the first coalescence between any of $\{c-1, c, c+1\}$. See below for further explanation. Among all of the target recombinations, only 4/9 of them can be unambiguously determined from the data, so we scale this count by 9/4 in (4). The factor $\hat{\mu}/r$ in (2) converts the estimated number of recombinations to an estimated number of mutations. The final term in (2) is the expected total number of sequencing errors among the m sequences, each of length ℓ.

Combining (1) and (2), we obtain

$$\hat{\mu} = \frac{2m\ell\bar{M} - 4\bar{L}C_1}{3m\ell - 4\bar{L}C_2} r, \qquad \hat{\epsilon} = \frac{3C_1 - 2C_2\bar{M}}{3m\ell - 4\bar{L}C_2}. \tag{5}$$

S2.2 Time since MRCA of IBDs

Given g_i, the coalescence time of c_{i1} and c_{i2}, the probability of a recombination event being effective (and thus being the right end point of the IBD) is no longer 2/3, but a function of g_i. For example, if a recombination event occurs more recently than the coalescence, when g_i is small there is little opportunity for the two subsequences created by the recombination to find a common ancestor before g_i, which implies a high probability for this recombination event to be effective. For this reason, it is difficult to estimate g_i based on the distribution of the recombination events or IBD lengths.

Note that $|M_i|$ follows a Poisson distribution with parameter $2(\mu g_i + \epsilon)(r_i - l_i)$. Given $\hat{\mu}, \hat{\epsilon}$, we find the first moment estimator of g_i by solving, for $p \in P$,

$$\sum_{i=1}^{I} |M_i| \mathbb{I}(p_i = p) = (2\hat{\mu} g_p + 2\hat{\epsilon}) \sum_{i=1}^{I} (r_i - l_i) \mathbb{I}(p_i = p)$$

to obtain

$$\tilde{g}_p = \frac{1}{\hat{\mu}} \left(\frac{\sum_{i=1}^{I} |M_i| \mathbb{I}(p_i = p)}{2 \sum_{i=1}^{I} (r_i - l_i) \mathbb{I}(p_i = p)} - \hat{\epsilon} \right).$$

Let $g_0 = 0$, and noting that $g_{p-1} < g_p$ for $p = m+2, \ldots, n$, we estimate g_p by the following quadratic optimization with linear constraints,

$$\check{g}_p = \underset{g_p}{\operatorname{argmin}} \|g_p - \tilde{g}_p\|_2^2.$$

$$\text{subject to } g_{p'} - g_p \geq 0, \quad \text{for } p' > p \text{ and } p', p \in P.$$

By the nature of constrained optimization, many parent nodes will be inferred to share the same age, which is unrealistic. Numerical studies show that it will be helpful to smooth these estimates when they are used later in estimating the population size. For this reason, the final estimate $\hat{g}_p$ of g_p is acquired by further smoothing $\check{g}_p$ by a Savitzky-Golay smoothing filter [34].

S2.3 Present and past population sizes

We estimate population sizes by first estimating the density $f_{\tilde{G}}$ of $\tilde{G}$, the TMRCA at a specific site s. We estimate $f_{\tilde{G}}$ by relating it to the density f_G of G, the TMRCA of an IBD segment. We first estimate f_G empirically from the estimated TMRCAs of each IBD segment, as calculated in Section S2.2. Then the density of $\tilde{G}$ is derived by conditioning on L, the length of the IBD segment. Intuitively, the larger L is, the more sites are covered by the IBD. Hence

$$f_{\tilde{G}}(g) = \frac{\sum_{l\geq 1} l f_{G|L}(g|L)\Pr(L=l)}{\mathrm{E}(L)} = \frac{\sum_{l\geq 1} l \Pr(L=l|G=g) f_G(g)}{\mathrm{E}(L)} = \frac{\mathrm{E}(L|G=g)}{\mathrm{E}(L)} f_G(g).$$

The mean IBD length $\mathrm{E}(L)$ can be estimated as $\bar{L}$, and the conditional mean $\mathrm{E}(L|G)$ can be found by a local linear kernel regression estimator given each pair $(\hat{g}_{p_i}, r_i - l_i)$ of IBD_i, where $\hat{g}_{p_i}$ is the estimated TMRCA in Section S2.2. Thus, the estimate $\hat{f}_{\tilde{G}}$ of $f_{\tilde{G}}$ can be found by substituting the corresponding estimates of f_G, $\mathrm{E}(L|G)$ and $\mathrm{E}(L)$. We then smooth the estimate $\hat{f}_{\tilde{G}}$ by a Savitzky–Golay filter.

To estimate population sizes, note that the distribution of $\tilde{G}$ is solely determined by the coalescent rate $1/N(g)$, i.e.,

$$\Pr(\tilde{G} > g) = \exp\left\{-\int_0^g \frac{1}{N(t)}\mathrm{d}t\right\}.$$

Taking the log-derivative with respect to g on both sides, we have

$$N(g) = \frac{1 - F_{\tilde{G}}(g)}{f_{\tilde{G}}(g)}.$$

We thus calculate the estimate

$$\hat{N}(g) = \frac{\int_g^\infty \hat{f}_{\tilde{G}}(t)\mathrm{d}t}{\hat{f}_{\tilde{G}}(g)},$$

followed by another Savitzky–Golay smoothing filter.

S2.4 Interpretation of C_1 and C_2

The quantity C_1 estimates the total number of sequencing errors and mutations on the branch immediately above (i.e. backwards in time from) a sequence c, before the first coalescence between any of c, $c-1$ and $c+1$.

In the efficient IBD subset, for each sequence c we record the IBDs of the pairs $(c-1, c)$ and $(c, c+1)$. If IBDs in $(c-1, c)$ and $(c, c+1)$ covering a site s have different parent nodes, then c must coalesce with either $c-1$ or $c+1$ at site s more recently than the coalescence of $c-1$ with $c+1$. In this case, any site differences contained in both IBDs can be attributed to an event unambiguously located on the branch immediately above sequence c before the first coalescence.

We also wish to include site changes above c in the case where $c-1$ coalesces with $c+1$ first, in which site differences cannot be located in this way. By symmetry, this

case occurs 1/3 of the time, and so we scale the previous count by a factor of 3/2. See Figure S1 (top left) for illustration. The quantity C_1 in (3) thus sums the total number of sequencing errors and mutations on the branch immediately above each sequence c, before the first coalescence.

Likewise, the quantity C_2 estimates the total number of recombinations on the branch immediately above each sequence c, before the first coalescence between any of c, $c-1$ and $c+1$. Recall that p_i, l_i and r_i are the MRCA and left and right endpoints of IBD_i. Similarly to C_1, we only count recombinations that produce four adjacent IBDs IBD_i, $\text{IBD}_{i'}$, IBD_j and $\text{IBD}_{j'}$, with the first two corresponding to sequence pair $(c-1, c)$ and the latter two corresponding to $(c, c+1)$, such that $s = r_i = l_{i'} = r_j = r_{j'}$ (i.e., the breakpoint between i and i' is the same as the breakpoint between j and j'), and either $p_i \neq p_j$ or $p_{i'} \neq p_{j'}$, as shown in Figure S1 (top right). We then scale to account for the remaining cases.

If we have two IBD breakpoints at s, we must have a recombination on the c lineage since we assume that only one recombination can occur at s. If the MRCAs of these IBDs also satisfy $p_i \neq p_j$ or $p_{i'} \neq p_{j'}$, the recombination must occur before any coalescence, since:

- if $\{c-1, c+1\}$ coalesce before the recombination, we must observe $p_i = p_j$ and $p_{i'} = p_{j'}$;
- if $\{c-1, c\}$ coalesce before the recombination, there would not be an IBD breakpoint at s for the $(c-1, c)$ pair;
- likewise for when $\{c, c+1\}$ coalesce before the recombination.

Thus we do indeed count a subset of the desired recombinations.

When there is a recombination in c before any coalescences, there are 4 lineages immediately after the recombination (backwards in time), as shown in Figure S1 (bottom). There are three cases:

- If lineages 1 and 2 coalesce first, the recombination is not effective and there are no IBD breakpoints at s (probability 1/6).
- If lineages 3 and 4 coalesce first, we will have $p_i = p_{i'}$ and $p_j = p_{j'}$ and so not count the recombination (probability 1/6).
- Otherwise, we may count the recombination (probability 2/3).

As shown in Figure S1 (bottom), suppose (without loss of generality) that for the third case, lineages 1 and 3 coalesce first. If the coalesced lineage then coalesces with lineage 2, then there will not be an IBD breakpoint at s for $(c, c+1)$; otherwise the required pattern will be produced. Thus we only count $2/3 \times 2/3 = 4/9$ of the cases, and so scale by a factor of 9/4 to estimate the desired number of recombinations.

S3 Appendix: Further details for 1KGP data analysis

The chromosome lengths are $\ell_{20} = 63\,025\,522$ and $\ell_{21} = 48\,129\,897$ sites, of which 1 552 394 and 927 753 sites are polymorphic in the full dataset. The sequence data were downloaded as .vcf files from ftp.1000genomes.ebi.ac.uk. Then, we converted them to the .samples format required for input to tsinfer and adopted human reference assembly GRCh37 recombination map following the data pre-processing steps in [35]. Specifically, we first cloned the Github repository from github.com/awohns/unified_genealogy_paper and installed all of the necessary software, packages and modules listed in the "requirements.txt" file and the "tools" sub-folder. Then we redirected to the "all-data" sub-folder and conducted the "Makefile" document to build the tree sequence for 1000

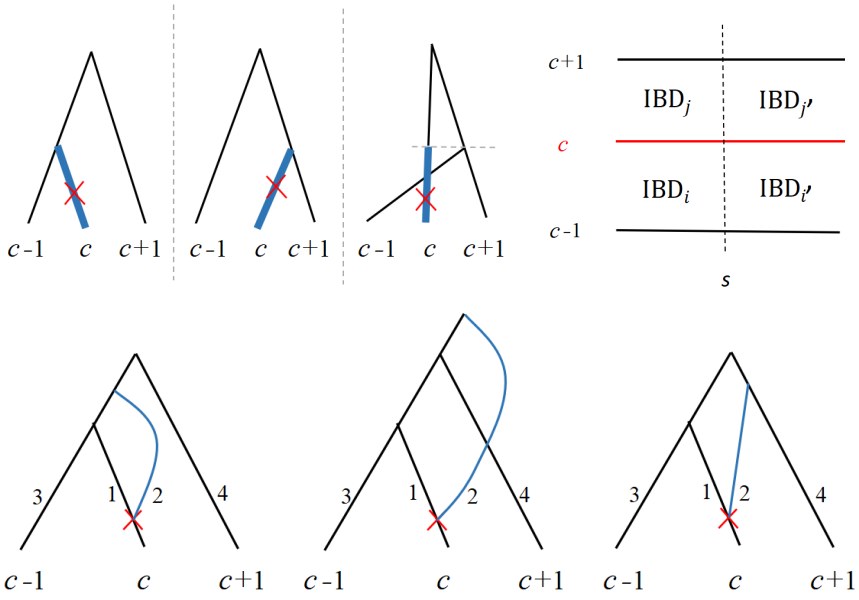

Fig S1. Top left: the three possible coalescent patterns of sequences c, $c-1$ and $c+1$ at a site s. While mutation events on the thicker edges should be included in the quantity C_1, only those in the first two patterns are counted. Top right: a sketch of four IBDs corresponding to sequence pairs $(c-1, c)$ and $(c, c+1)$. Bottom: a recombination event occurred on sequence c, which breaks the sequence before any coalescence between $c-1$, c and $c+1$, immediately resulting in a total of four segments $(1, 2, 3, 4)$. The figure shows three of the possible coalescent patterns, corresponding to the cases where segments 1 and 3 coalesce first.

Genomes chromosome 20, during which the program downloaded the chromosome 20 variant data and produced a .samples file (tsinfer input format) converted from a .vcf file. Then IBDs were extracted from the inferred TS, and TSABC was employed, as described in Section . A similar process is repeated for chromosome 21.

S4 Supplementary Figures

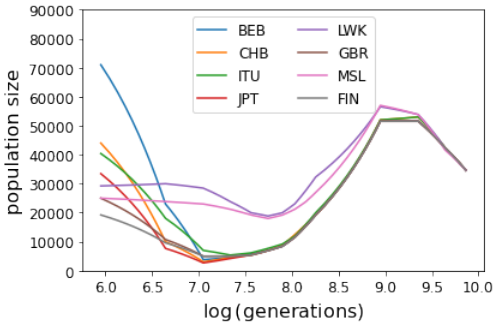

Fig S2. The 1KGP $N(g)$ models [26]. Natural logarithms of g are shown on the x-axis, with the models starting at $g = \exp(6) \approx 400$ generations in the past. The values of $N(1\,000)$ which form the right endpoints of Figure 8 correspond to $x = \log(1\,000) \approx 6.9$.

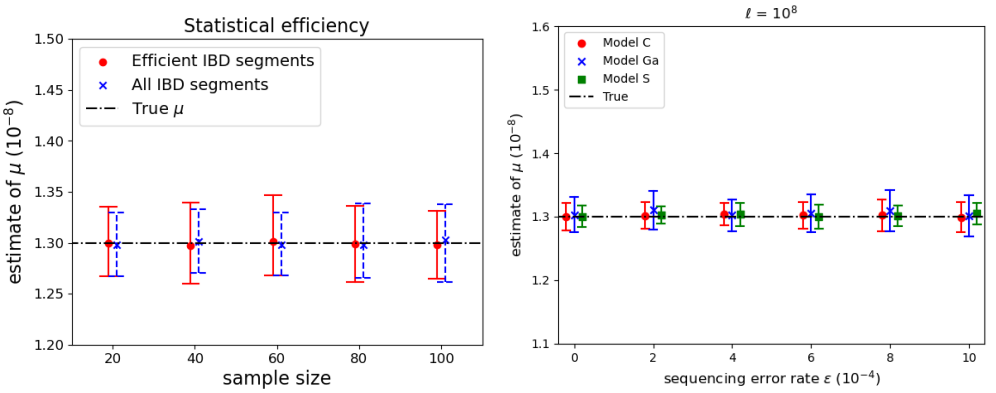

Fig S3. Left: Estimated 95% CIs for the estimation of μ when an efficient subset of IBD segments were extracted from the TS and when all IBDs were used. At each sample size, 25 replicate datasets were simulated under Model C, with sequence length $\ell = 10^7$. Right: Impact of sequencing error rate ϵ on $\hat{\mu}$ under Model C, Model Ga and Model S, from 25 replicates with sample size $m = 10$ and sequence length $\ell = 10^8$.

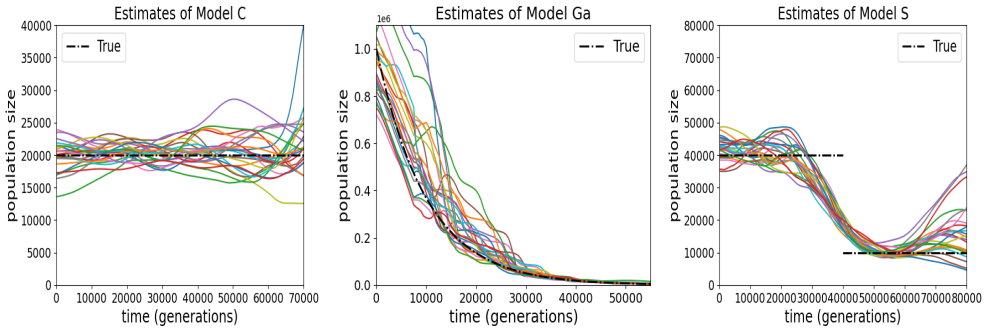

Fig S4. Estimates of the population size $N(g)$ under Model C, Model Ga and Model S, from 25 simulations at each setting. Sequence length $\ell = 10^7$, sequencing error rate $\epsilon = 10^{-3}$, sample size $m = 80$.

References

1. Browning SR, Browning BL. Identity by descent between distant relatives: detection and applications. Annual Review of Genetics. 2012;46:617–633.

2. Palamara PF, Pe'er I. Inference of historical migration rates via haplotype sharing. Bioinformatics. 2013;29(13):i180–i188.

3. Sticca EL, Belbin GM, Gignoux CR. Current developments in detection of identity-by-descent methods and applications. Frontiers in Genetics. 2021; p. 1725.

4. Tang K, Naseri A, Wei Y, Zhang S, Zhi D. Open-source benchmarking of IBD segment detection methods for biobank-scale cohorts. GigaScience. 2022;11:giac111.

5. Chen H, Naseri A, Zhi D. FiMAP: A fast identity-by-descent mapping test for biobank-scale cohorts. PLoS Genetics. 2023;19(12):e1011057.

6. Griffiths RC, Marjoram P. An ancestral recombination graph. In: Donnelly P, Tavare S, editors. IMA volume on Mathematical Population Genetics. New York: Springer–Verlag; 1997. p. 257–270.

7. Lewanski A, Grundler M, Bradburd G. The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. PLoS Genetics. 2024;20(1):e1011110.

8. Brandt DY, Huber CD, Chiang CW, Ortega-Del Vecchyo D. The promise of inferring the past using the ancestral recombination graph. Genome Biology and Evolution. 2024;16(2):evae005.

9. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-wide inference of ancestral recombination graphs. PLoS Genetics. 2014;10(5):e1004342.

10. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. Nature Genetics. 2019;51(9):1330–1338.

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

12. Mahmoudi A, Koskela J, Kelleher J, Chan Y, Balding D. Bayesian inference of ancestral recombination graphs. PLoS Computational Biology. 2022;18(3):e1009960.

13. Zhang BC, Biddanda A, Gunnarsson ÁF, Cooper F, Palamara PF. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. Nature Genetics. 2023; p. 1–9.

14. Wong Y, Ignatieva A, Koskela J, Gorjanc G, Wohns AW, Kelleher J. A general and efficient representation of ancestral recombination graphs. BioRxiv. 2023; p. 2023–11.

15. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Computational Biology. 2016;12(5):e1004842.

16. Silcocks M, Farlow A, Hermes A, Tsambos G, Patel H, Huebner S, et al. Indigenous Australian genomes show deep structure and rich novel variation. Nature. 2023;624(7992):593–601.

17. Kelleher J, Lohse K. Coalescent simulation with msprime. Statistical Population Genomics. 2020;986:191–230.

18. Tian X, Browning BL, Browning SR. Estimating the genome-wide mutation rate with three-way identity by descent. The American Journal of Human Genetics. 2019;105(5):883–893.

19. Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, et al. Efficient ancestry and mutation simulation with msprime 1.0. Genetics. 2022;220(3):iyab229.

20. Griffiths R. Neutral two-locus multiple allele models with recombination. Theoretical Population Biology. 1981;19(2):169–186.

21. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology. 1983;23(2):183–201.

22. Albers PK, McVean G. Dating genomic variants and shared ancestry in population-scale sequencing data. PLoS Biology. 2020;18(1):e3000586.

23. YC Brandt D, Wei X, Deng Y, Vaughn AH, Nielsen R. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. Genetics. 2022;221(1):iyac044.

24. Chiang CW, Ralph P, Novembre J. Conflation of short identity-by-descent segments bias their inferred length distribution. G3: Genes, Genomes, Genetics. 2016;6(5):1287–1296.

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

26. 1000 Genomes Project Consortium and others. A global reference for human genetic variation. Nature. 2015;526(7571):68.

27. Deng Y, Song YS, Nielsen R. The distribution of waiting distances in ancestral recombination graphs. Theoretical Population Biology. 2021;141:34–43.

28. Ignatieva A, Favero M, Koskela J, Sant J, Myers SR. The distribution of branch duration and detection of inversions in ancestral recombination graphs. BioRxiv. 2023; p. 2023–07.

29. Tian X, Cai R, Browning SR. Estimating the genome-wide mutation rate from thousands of unrelated individuals. The American Journal of Human Genetics. 2022;109(12):2178–2184.

30. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010;328(5978):636–639.

31. Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Cassals F, et al. Variation in genome-wide mutation rates within and between human families. Nature Genetics. 2011;43(7):712–714.

32. Harris K. Evidence for recent, population-specific evolution of the human mutation rate. Proceedings of the National Academy of Sciences. 2015;112(11):3439–3444.

33. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162(4):2025–2035.

34. Savitzky A, Golay MJ. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry. 1964;36(8):1627–1639.

35. Wohns AW, Wong Y, Jeffery B, Akbari A, Mallick S, Pinhasi R, et al. A unified genealogy of modern and ancient genomes. Science. 2022;375(6583):eabi8264.