



Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Original research

An instantaneous coalescent method insensitive to population structure

Zeqi Yao¹, Kehui Liu¹, Shanjun Deng, Xionglei He^{*}

State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

ARTICLE INFO

Article history:

Received 7 December 2020

Received in revised form

23 February 2021

Accepted 5 March 2021

Available online 9 April 2021

Keywords:

Coalescent

Population structure

Demographic history

Population size

Phylogenetic tree

ABSTRACT

Conventional coalescent inferences of population history make the critical assumption that the population under examination is panmictic. However, most populations are structured. This complicates the prevailing coalescent analyses and sometimes leads to inaccurate estimates. To develop a coalescent method unhampered by population structure, we perform two analyses. First, we demonstrate that the coalescent probability of two randomly sampled alleles from the immediate preceding generation (one generation back) is independent of population structure. Second, motivated by this finding, we propose a new coalescent method: *i*-coalescent analysis. The *i*-coalescent analysis computes the instantaneous coalescent rate by using a phylogenetic tree of sampled alleles. Using simulated data, we broadly demonstrate the capability of *i*-coalescent analysis to accurately reconstruct population size dynamics of highly structured populations, although we find this method often requires larger sample sizes for structured populations than for panmictic populations. Overall, our results indicate *i*-coalescent analysis to be a useful tool, especially for the inference of population histories with intractable structure such as the developmental history of cell populations in the organs of complex organisms.

Copyright © 2021, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, and Genetics Society of China. Published by Elsevier Limited and Science Press. All rights reserved.

Introduction

Coalescent theory is widely used to infer the demographic history of populations (Strimmer and Pybus, 2001; Drummond et al., 2005; Li and Durbin, 2011; Liu and Fu, 2015). Conventional coalescent inference, for example, Kingman's *n*-coalescent, uses the waiting time between coalescent events to estimate effective population size (Kingman, 1982). Such inferences often require the critical assumption that the population under examination is panmictic. Inferences of population history with violated assumption can lead to highly inaccurate estimations (Wakeley, 1999; Leblois et al., 2006; Chikhi et al., 2010; Mazet et al., 2016). For example, while Li and Durbin (2011) suggest that several population size reductions occurred in recent human history, Mazet et al. (2016) show that population structure could explain these signals. Because most populations are not panmictic, it is often necessary to incorporate population structure into coalescent models. This can greatly increase model complexity and requires the use of much larger data sets to make reliable statistical inferences (Peter et al., 2010; Grusea et al., 2019).

Furthermore, if structural information is intractable, model construction and data collection become infeasible. These complications necessitate a coalescent method that is insensitive to population structure.

In this article, we present a new coalescent method to circumvent the influence of population structure, *i*-coalescent analysis. In contrast to other methods which focus on the waiting time between coalescent events (such as Kingman's *n*-coalescent), *i*-coalescent analysis computes the instantaneous coalescent rate (iCR) of every generation. Mathematical deductions and computational simulations show that the *i*-coalescent analysis can be used to reconstruct the dynamics of population size for highly structured populations. Notably, same as skyline-plot methods (Strimmer and Pybus, 2001; Drummond et al., 2005), *i*-coalescent analysis would apply to genealogical history with no recombination. In addition, it requires a sufficiently large sampling of the population and dense mutation information.

Results

The coalescent probability at the immediate prior generation is insensitive to population structure

To show how population structure influences coalescent analysis, we consider a group of haploid individuals of population size *N*

^{*} Corresponding author.

E-mail address: hexiongli@mail.sysu.edu.cn (X. He).

¹ These authors contributed equally to this work.

(equivalent to a diploid group of size $N/2$) under a classical n -island model. The subpopulation of each island evolves as a Wright-Fisher process with no migration allowed between islands. We first consider two simple scenarios (A and B). In scenario A, $n = 1$. Therefore, the entire population is panmictic with no structure. In scenario B, $n = 2$ and the size of the two subpopulations are N_1 and N_2 , respectively, in which $N_1 + N_2 = N$. In accordance with standard coalescent theory in a discrete-generation model, for scenario A:

$$P_A(t) = \frac{1}{N} \left(1 - \frac{1}{N}\right)^{t-1}, \quad (1)$$

where $P_A(t)$ is the probability that two alleles randomly sampled from the population coalesce at generation t . Note that $t = 0$ depicts the present, and larger t indicates further back in the past. Similarly, in scenario B:

$$P_{B1}(t) = \frac{1}{N_1} \left(1 - \frac{1}{N_1}\right)^{t-1} \quad (2)$$

$$P_{B2}(t) = \frac{1}{N_2} \left(1 - \frac{1}{N_2}\right)^{t-1}, \quad (3)$$

where $P_B(t)$ is the probability that two alleles sampled randomly from the i th island coalesce at generation t , where $i = 1, 2$. Accordingly, for two alleles sampled randomly from the two islands, the coalescent probability at generation t is

$$P_B(t) = \left[\binom{N_1}{2} P_{B1}(t) + \binom{N_2}{2} P_{B2}(t) \right] / \binom{N}{2} \quad (4)$$

For large N , equation (4) can be approximated by

$$P_B(t) = \frac{1}{N^2} \left[N_1 \left(1 - \frac{1}{N_1}\right)^{t-1} + N_2 \left(1 - \frac{1}{N_2}\right)^{t-1} \right] \quad (5)$$

The difference between $P_A(t)$ and $P_B(t)$ (equations [1] and [5]) highlights the influence of population structure on coalescent analysis.

Importantly, $P_A(t)$ and $P_B(t)$ would be identical for $t = 1$, as $P_A = P_B = 1/N$. We plot the difference between $P_A(t)$ and $P_B(t)$ by assuming $N = 100,000$ with different N_1 and N_2 combinations (Fig. 1A). As expected, $P_A(t)$ and $P_B(t)$ are identical for $t = 1$ and their difference increases as a function of t . Hence, the coalescent probability in

scenario B is independent of population structure at and only at the immediate preceding generation ($t = 1$).

We obtain similar results for when $n > 2$ (a greater number of islands and, consequently, a higher complexity of the population structure). Extending equation (5) to scenarios with $n > 2$ yields

$$P_B(t) = \frac{1}{N^2} \sum_{i=1}^n N_i \left(1 - \frac{1}{N_i}\right)^{t-1}, \quad (6)$$

where N_i is the size of each subpopulation, B_i . For $t > 1$, the difference between $P_A(t)$ and $P_B(t)$ increases as n increases (Fig. 1B). Thus, the impact of population structure on coalescent analysis increases with the former's complexity. However, the key observation is that $P_A(t)$ and $P_B(t)$ are always identical for $t = 1$, irrespective of the number of subpopulations. This again demonstrates that population structure has no impact on the coalescent probability on the immediate prior generation under the constant-size model.

We extend these results to the exponential-growth model, in which

$$N(t) = N(0)e^{-rt}, \quad (7)$$

where $N(t)$ is the size of population at generation t , $N(0)$ is the population size at the present, and r is the population growth rate. Under this model, $P_A(t)$ and $P_B(t)$ can be calculated by the following expressions:

$$P_A(t) = \frac{1}{N(t)} \prod_{x=1}^{t-1} 1 - \frac{1}{N(x)} = \frac{e^{rt}}{N(0)} \prod_{x=1}^{t-1} 1 - \frac{e^{rx}}{N(0)} \quad (8)$$

$$P_B(t) = \frac{1}{N(0)^2} \sum_{i=1}^n \frac{N_i(0)^2}{N_i(t)} \prod_{x=1}^{t-1} 1 - \frac{1}{N_i(x)} \\ = \frac{1}{N(0)^2} \sum_{i=1}^n N_i(0) e^{rt} \prod_{x=1}^{t-1} 1 - \frac{e^{rx}}{N_i(0)} \quad (9)$$

Similar to the constant-size model, the difference between $P_A(t)$ and $P_B(t)$ increases with t under different initial population sizes in the two-island case ($n = 2$; Fig. 2A) and under different values of n (Fig. 2B). Also as per the previous case, the difference between $P_A(t)$ and $P_B(t)$ increases with increasing population structure complexity (greater n ; Fig. 2B). Most importantly, $P_A(t)$ and $P_B(t)$ are identical if and only if $t = 1$, assuming each subpopulation has the same growth rate. Thus, as in the constant-size model, the coalescent probability

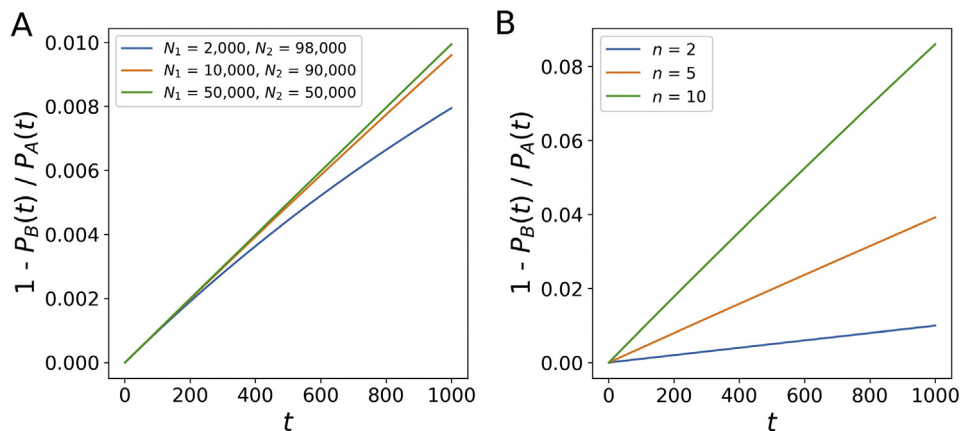


Fig. 1. The difference between $P_A(t)$ and $P_B(t)$ under a constant-size model. Figures depict a population consisting of n constant-size subpopulations. The total population size $N = 100,000$ in all cases. **A:** The curve of $1 - P_A(t)/P_B(t)$ where $n = 2$ with different combinations of N_1 and N_2 . **B:** The curve of $1 - P_A(t)/P_B(t)$ where $n = 2, 5, 10$, respectively, with equal subpopulations sizes ($N_i = N/n$ for sub-population B_i).

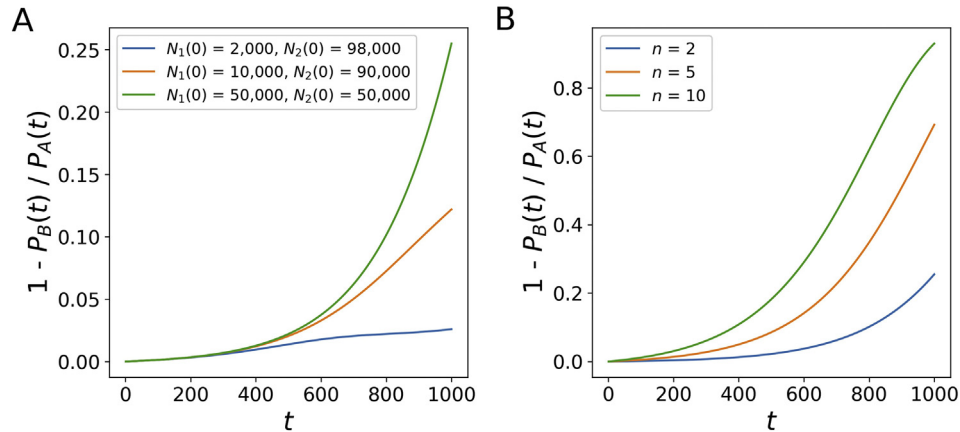


Fig. 2. The difference between $P_A(t)$ and $P_B(t)$ under an exponential-growth model. Figures depict a population consisting of n subpopulations experiencing exponential growth: $N(t) = N(0)e^{rt}$, with $r = 0.005$ for each subpopulation and $N(0) = 100,000$ for the whole population. **A:** The curve of $1 - P_A(t)/P_B(t)$ where $n = 2$ with different $N_1(0)$ and $N_2(0)$ combinations. **B:** The curve of $1 - P_A(t)/P_B(t)$ where $n = 2, 5, 10$, respectively. Each subpopulation has initial sizes ($N_i(0) = N(0)/n$ for subpopulation B_i).

at the first prior generation under an exponential-growth model is insensitive to population structure.

i-coalescent analysis

The above analysis indicates that an inference method for population history based on the coalescent probability of two randomly sampled alleles at the immediate prior generation, henceforth the instantaneous coalescent rate (iCR), may be insensitive to population structure. Therefore, we develop a novel nonparametric coalescent approach, *i*-coalescent analysis, that estimates the population size of single generation by computing the iCR. With a sufficiently large sample size, the iCR of each generation can be calculated from a phylogenetic tree using the formula

$$iCR = \frac{k}{\binom{l}{2}} = \frac{2k}{l(l-1)}, \quad (10)$$

where k is the number of coalescent events and l is the number of lineages at the examined generation on the phylogenetic tree. Accordingly, the population size N can be estimated as

$$N = \frac{1}{iCR} = \frac{l(l-1)}{2k}, \quad (11)$$

noting that N is also instantaneous, corresponding to the specific generation, t , examined. See [Methods](#) for more details.

Simulations show that *i*-coalescent analysis is insensitive to population structure

To validate the efficacy of *i*-coalescent analysis, simulations were carried out. We simulated two populations: (1) a panmictic population, A , of haploid individuals and (2) a structured population, B , consisting of two subpopulations, B_1 and B_2 ([Fig. 3](#); [Methods](#)). Populations A and B were parametrized to be with the same overall size dynamics for 100 generations, in which B_1 and B_2 had distinct growth patterns. After the population growth simulations, 1%, 5%, 10% of the final individuals were sampled, respectively, to reconstruct phylogenetic trees for *i*-coalescent analysis ([Methods](#)). For population A , *i*-coalescent analysis was able to accurately reconstruct population size dynamics, even with a sampling proportion of 1% ([Fig. 3A](#)). For population B , *i*-coalescent analysis yielded

accurate results with medium (5%) and large (10%) sample sizes ([Fig. 3B](#)). Therefore, *i*-coalescent analysis is capable of providing accurate population size estimates in structured populations, although structured populations require larger sample sizes than their panmictic counterparts. In addition, simulations with migration allowed also achieved the same conclusion ([Fig. S1](#)).

We also considered a scenario with a more complex structured population, consisting of five subpopulations with distinct demographic histories ([Fig. 4](#); [Methods](#)). Using *i*-coalescent analysis to infer size dynamics of the simulated populations, we obtained similar results to those shown in [Fig. 3](#). Interestingly, *i*-coalescent analysis performance in the complex scenario ([Fig. 4B](#)) was approximately as accurate as in the simpler case ([Fig. 3B](#)). Thus, *i*-coalescent analysis can reconstruct population size dynamics for highly structured populations.

In the above simulations, we directly used the genealogical information in the simulated populations to make inferences. This effectively made an assumption that there is exactly one mutation per lineage per generation. In reality, however, the number of mutation may vary among generations. To model the effects of varied mutation numbers, we carried out the same simulations as in [Figs. 3 and 4](#) except that the number of mutations acquired at each generation is set to follow a Poisson distribution with the mean equal to one. We obtained essentially the same results ([Figs. S2 and S3](#)) as in the above simulations, suggesting the robustness of the *i*-coalescent analysis.

Discussion

Population structure often hampers efforts to perform coalescent analysis. In this study, we propose a new coalescent method, *i*-coalescent analysis, to infer the dynamics of population size by computing the iCR of each single generation. Motivated by the result that the coalescent probability of two randomly sampled alleles at the immediate prior generation is insensitive to population structure, *i*-coalescent analysis provides a coalescent method insensitive to population structure. We use simulations to validate this method and show that *i*-coalescent analysis is capable of making reliable inferences of the population size dynamics of highly structurally complex populations.

A necessary prerequisite of *i*-coalescent analysis is that the sample size must be sufficiently large such that the iCR of every generation through a phylogenetic tree can be estimated (see [Methods](#)). This raises a concern on the approximations used for

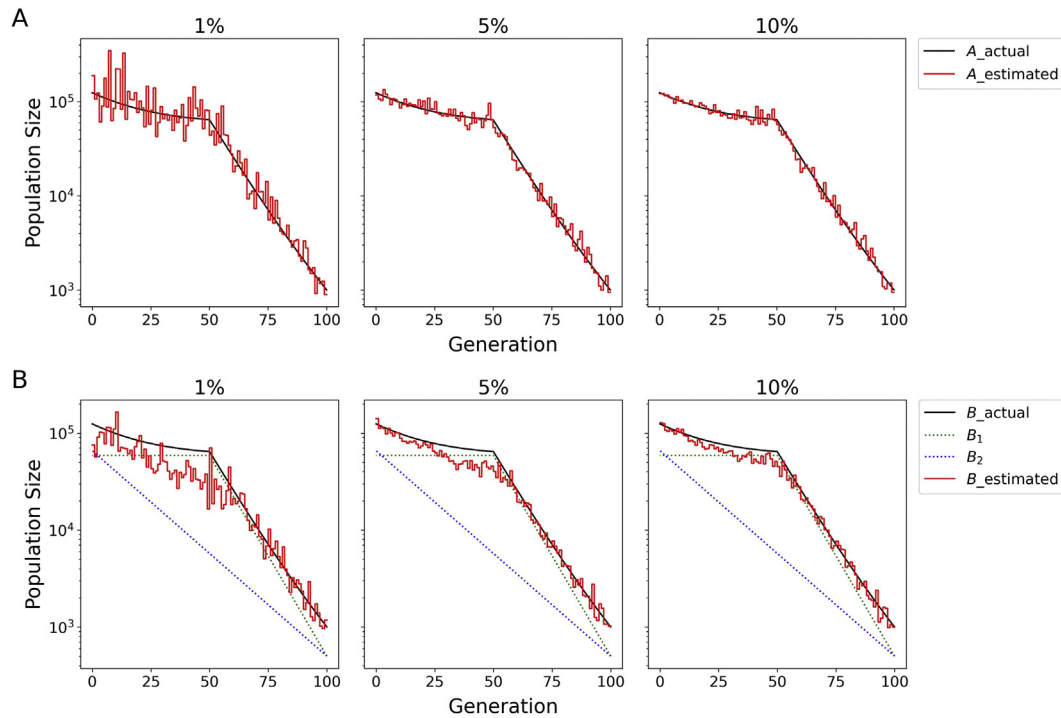


Fig. 3. Inference of population size dynamics for a simulated population consisting of two subpopulations. **A:** A simulated panmictic population, *A*. The demographic history is depicted with a solid black line. **B:** A simulated structured population, *B*, consisting of two subpopulations *B*₁ and *B*₂. The simulated demographic history of each sub-population is distinct from each other, as depicted by the dashed lines. The demographic history of the whole population *B* is the same as that of *A*. The solid red curve in each subplot represents the estimated population size dynamics, which were obtained using *i*-coalescent analysis. Estimated population size dynamics were inferred using samples of 1%, 5%, and 10% of the final population, as shown above each subplot.

calculating coalescent probability because multiple and simultaneous coalescent events would occur in a single generation for large samples (Bhaskar et al., 2014). To assess the potential bias, we compared the prediction using the discrete-time Wright-Fisher

(DTWF) model proposed by Bhaskar et al. (2014), which performs exact but computationally expensive calculation, with the prediction by *i*-coalescent analysis regarding the number of coalescent events at the immediate prior generation (Methods). We found virtually no

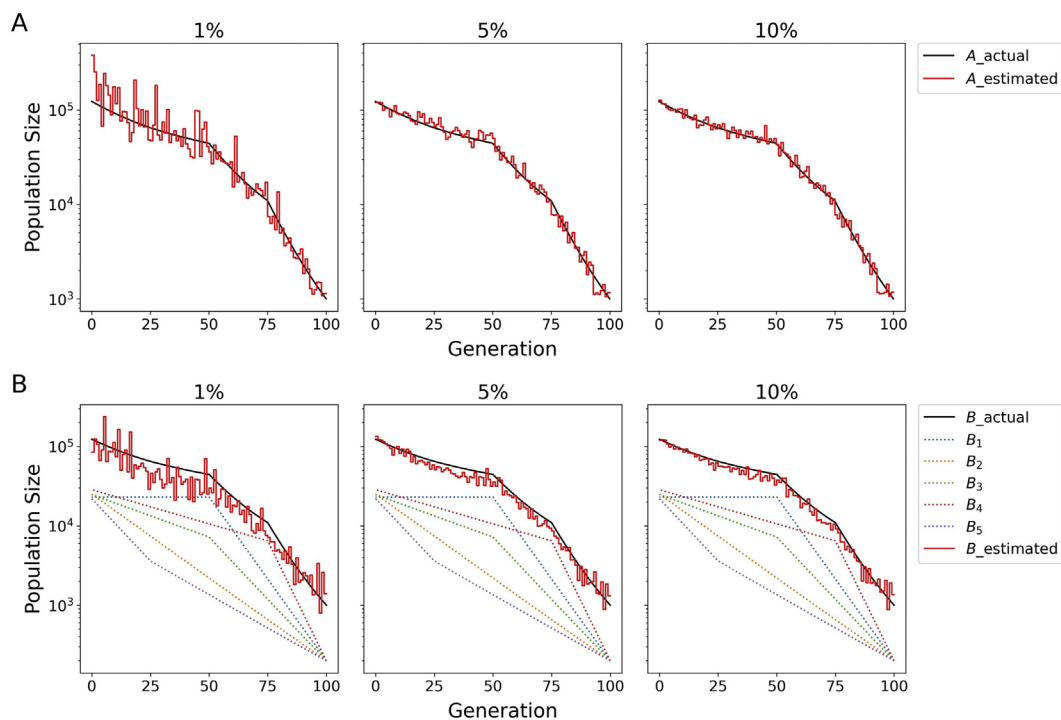


Fig. 4. Inference of population size dynamics for a simulated population consisting of five subpopulations. **A and B:** The same simulations as that shown in Fig. 3, except that the simulated structured population in **B** consists of five subpopulations, each with distinct demographic histories.

difference between the two predictions when the sampling proportion is less than 10%, although the difference becomes non-negligible after the sampling proportion reaches 20% (Fig. S4). To further address this issue, we sampled 20%, 50%, 80%, and 100% individuals in the simulated panmictic population examined in Fig. 3A. We found *i*-coalescent analysis can still accurately reconstruct the population history when the sampling proportion is 20%; there are only slight overestimations at very recent prior generations when the sampling proportion is larger than 50% (Fig. S5).

We note that, with the increasing capacity of DNA sequencing, the requirement of large sample size does not pose an insurmountable obstacle for the application of our method in making population genetic inferences. For example, there are already nearly one million sequenced human genomes available (Bycroft et al., 2018). Moreover, another potential application of *i*-coalescent analysis is the study of cell population history in the organs of multicellular organisms (Hu et al., 2013; Lee-Six et al., 2018). It is well known that the development of an organism is a process consisting of cell population growth and differentiation. Because the cell populations are highly structured, conventional coalescent methods are incapable of making reliable inferences regarding developmental history. We believe that the *i*-coalescent analysis will be a useful tool for reconstructing organ development when cell phylogeny can be mapped by the use of developing cell barcoding techniques.

Methods

Inference of population size dynamics using *i*-coalescent analysis

For each generation t , the instantaneous population size is estimated as the inverse of iCR, using the formula

$$N(t) = \frac{1}{\text{iCR}(t)} = \frac{l(t)(l(t) - 1)}{2k(t)} \quad (12)$$

where $N(t)$ is the instantaneous population size at generation t , $\text{iCR}(t)$ is the instantaneous coalescent rate at generation t , $l(t)$ is the number of examined lineages at generation t , and $k(t)$ is the number of coalescent events at generation t . The complete estimated population size dynamics are obtained from estimating $N(t)$ for each generation, from the tips ($t = 0$) to the root of the phylogenetic tree.

We note that if the sample size is insufficiently large, there may be no coalescent event at a given generation t ($k(t) = 0$). In this case, $\text{iCR}(t)$ is equal to 0 and $N(t)$ cannot be estimated by equation (12). In these cases, we search for the smallest value of g such that $k(g) > 0$ where $g > t$. We then calculate the coalescent rate for time interval $[t, g]$ as the mean of iCR of every generation in the interval. Therefore, the population size in time interval $[t, g]$ is estimated by the formula

$$N[t, g] = \frac{l(t)(l(t) - 1)(g - t + 1)}{2k(g)} \quad (13)$$

in cases for where $k(t) = 0$.

Simulations under structured and unstructured scenarios

To test the performance of *i*-coalescent analysis, we carried out a set of simulations under structured and unstructured scenarios. We compared an unstructured population to a classical n -island model for $n = 2$ and $n = 5$. In each scenario, the initial size of a population of

haploid individuals was set 1,000. The initial size of each subpopulation was set 500 for $n = 2$, and 200 for $n = 5$, where n is the number of subpopulations. The growth trajectory of each subpopulation was designed to be independent from one another (see the dashed lines in Figs. 3B and 4B). In unstructured scenarios, a panmictic population was simulated with the same demographic history as the whole population from structured scenarios (see the solid black lines in Figs. 3 and 4). All simulated populations evolved for 100 generations under a discrete-generation model. At each generation, newborn individuals were randomly assigned a parent from the previous generation. Each individual was then assigned a new mutation, with no recurrent mutations allowed. In each simulation, 1%, 5%, and 10% of the final individuals were sampled, respectively, and their genealogy in the simulated population is directly used to represent their phylogeny.

In the simulations for Figs. S2 and S3, the number of mutations acquired at each generation was set to follow a Poisson distribution with the mean equal to one. In this scenario, the generation information cannot be directly obtained from the simulated genealogy. Hence, before applying *i*-coalescent analysis, we re-estimated the divergence time (generation) of the internal nodes in the phylogeny by using the LF method of the software r8s (version 1.7) (Langley and Fitch, 1974; Sanderson, 2003).

Comparison of the DTWF model with *i*-coalescent analysis

For the *i*-coalescent analysis, the expected number of coalescent events at the first prior generation (i.e., $t = 1$) is easy to compute by transferring equation (11) to

$$E[k] = \frac{n(n-1)}{2N} \quad (14)$$

where $E[k]$ is the expected number of coalescent events at the first prior generation, n is the sample size, and N is the population size at the first prior generation (N was set as 10,000 in Fig. S4).

For the DTWF model mentioned by Bhaskar et al. (2014), we first computed the probability that a sample of n individuals m distinct ancestors at the prior generation, which denoted as $p_{n,m}$, by using the following recursion for $1 \leq m \leq n$,

$$p_{n,m} = \frac{N-m+1}{N} p_{n-1,m-1} + \frac{m}{N} p_{n-1,m} \quad (15)$$

with the convention that $p_{0,0} = 1$, and $p_{n,m} = 0$ for all other values of m and n . Then the expected number of coalescent events at the prior generation can be computed by the formula

$$E[k] = \sum_{m=1}^n (n-m) p_{n,m} \quad (16)$$

CRedit authorship contribution statement

Zeqi Yao: Conceptualization, Methodology, Formal analysis, Writing - Original draft. **Kehui Liu:** Conceptualization, Methodology, Formal analysis. **Shanjuan Deng:** Conceptualization, Methodology, Formal analysis. **Xionglei He:** Conceptualization, Supervision, Writing - Reviewing & Editing, Funding acquisition.

Conflict of interest

There is no conflict of interest.

Acknowledgments

The authors are grateful to Drs W. Zhai, H. Chen, and Z. Hu for discussion and comments. This work was supported by National Natural Science Foundation of China (31630042 and 31970570) and Guangdong Special Support Program.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jgg.2021.03.005>.

References

- Bhaskar, A., Clark, A.G., Song, Y.S., 2014. Distortion of genealogical properties when the sample is very large. *Proc. Natl. Acad. Sci. U. S. A.* 111, 2385–2390.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al., 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209.
- Chikhi, L., Sousa, V.C., Luisi, P., Goossens, B., Beaumont, M.A., 2010. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* 186, 983–995.
- Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192.
- Grusea, S., Rodríguez, W., Pinchon, D., Chikhi, L., Boitard, S., Mazet, O., 2019. Coalescence times for three genes provide sufficient information to distinguish population structure from population size changes. *J. Math. Biol.* 78, 189–224.
- Hu, Z., Fu, Y.X., Greenberg, A.J., Wu, C.I., Zhai, W., 2013. Age-dependent transition from cell-level to population-level control in murine intestinal homeostasis revealed by coalescence analysis. *PLoS Genet.* 9, e1003326.
- Kingman, J.F.C., 1982. The coalescent. *Stoch. Process. their Appl.* 13, 235–248.
- Langley, C.H., Fitch, W.M., 1974. An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.* 3, 161–177.
- Leblois, R., Estoup, A., Streiff, R., 2006. Genetics of recent habitat contraction and reduction in population size: does isolation by distance matter? *Mol. Ecol.* 15, 3601–3615.
- Lee-Six, H., Øbro, N.F., Shepherd, M.S., Grossmann, S., Dawson, K., Belmonte, M., Osborne, R.J., Huntly, B.J.P., Martincorena, I., Anderson, E., et al., 2018. Population dynamics of normal human blood inferred from somatic mutations. *Nature* 561, 473–478.
- Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496.
- Liu, X., Fu, Y.X., 2015. Exploring population size changes using SNP frequency spectra. *Nat. Genet.* 47, 555–559.
- Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., Chikhi, L., 2016. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity* 116, 362–371.
- Peter, B.M., Wegmann, D., Excoffier, L., 2010. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Mol. Ecol.* 19, 4648–4660.
- Sanderson, M.J., 2003. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 301–302.
- Strimmer, K., Pybus, O.G., 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol. Biol. Evol.* 18, 2298–2305.
- Wakeley, J., 1999. Nonequilibrium migration in human history. *Genetics* 153, 1863–1871.