1 A novel coalescent method insensitive to population structure

2 Zeqi Yao¹, Kehui Liu¹, Shanjun Deng¹, and Xionglei He^{1*}

3 ¹State Key Laboratory of Bio-control, School of Life Sciences, Sun Yat-sen University,
4 Guangzhou 510275, China

5 *Correspondence: hexiongl@mail.sysu.edu.cn (X.H.)

6

7 Abstract

8 Conventional coalescent inferences of population history make the critical assump-
9 tion that the population under examination is panmictic. However, most populations are
10 structured. This complicates the prevailing coalescent analyses and sometimes leads to
11 inaccurate estimates. To develop a coalescent method unhampered by population struc-
12 ture, we perform two analyses. First, we demonstrate that the coalescent probability of
13 two randomly sampled alleles from the immediate preceding generation (one genera-
14 tion back) is independent of population structure. Second, motivated by this finding, we
15 propose a new coalescent method: *i*-coalescent analysis. *i*-coalescent analysis computes
16 the instantaneous coalescent rate (iCR) by using a phylogenetic tree of sampled alleles.
17 Using simulated data, we broadly demonstrate the capability of *i*-coalescent analysis to
18 accurately reconstruct population size dynamics of highly structured populations, alt-
19 hough we find this method often requires larger sample sizes for structured populations
20 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.

Introduction

Coalescent theory is widely used to infer the demographic history of populations (Drummond et al., 2005; Li and Durbin, 2011; Liu and Fu, 2015; Strimmer and Pybus, 2001). Conventional coalescent inference, e.g. 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 (Chikhi et al., 2010; Leblois et al., 2006; Mazet et al., 2016; Wakeley, 1999). 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 datasets to make reliable statistical inferences (Grusea et al., 2019; Peter et al., 2010). 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 paper, 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.

Results

We first present two key results: (1) population structure influences coalescent analysis and (2) coalescent probability is independent of population structure for the immediate preceding generation ($t = 1$). Motivated by the latter result, we derive a novel nonparametric coalescent approach, *i*-coalescent analysis, that is insensitive to population structure.

We consider a group of haploid individuals of population size N (equivalent to a diploid group of size $N/2$) under a classical *n*-island model. The sub-population 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 sub-populations are N_1 and N_2 , respectively, in which $N_1 + N_2 = N$. According to 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 $P_{Bi}(t)$ 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 (3) and (5)) highlights the influence of population structure on coalescent analysis.

Importantly, for $t = 1$, $P_A = P_B = 1/N$. However, the difference between $P_A(t)$ and $P_B(t)$ increases with increasing t for all combination of N_1 and N_2 (Fig. 1a). Therefore, the coalescent probability in scenario B is independent of population structure 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 sub-population, 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 sub-populations. 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 according to 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 sub-population has the same growth rate. Thus, as in the constant-size model, the coalescent probability at the first prior generation under an exponential-growth model

is insensitive to population structure.

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.

To validate the efficacy of *i*-coalescent analysis, we carried out simulations. We simulated two populations: (1) a panmictic population, A , of haploid individuals and (2) a structured population, B , consisting of two sub-populations, 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, respec-

tively, 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.

We also considered a scenario with a more complex structured population, consisting of five sub-populations 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.

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 instantaneous coalescent rate (iCR) of every 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 of a phylogenetic tree can be estimated (see **Methods**). However, the increasing capacity of DNA sequencing has led to an accumulation of data, and nearly one million sequenced human genomes available (Bycroft et al., 2018). Therefore, this sample size requirement does not pose an insurmountable obstacle for the application of this method in making population genetics inferences. 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 instantaneous coalescent rate (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 1000. The initial size of each sub-population was set 500 for $n = 2$, and 200 for $n = 5$ where n is the number of sub-populations. The

growth trajectory of each sub-population was designed to be independent from one another (see the dashed lines in Fig. 3b and Fig. 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 Fig. 3 and Fig. 4). All simulated populations evolved for 100 generations under a discrete-generation model. At each generation, newborn individuals was randomly assigned a parent from the previous prior generation. Each individual was then assigned a new mutation. To record the genealogical information, no recurrent mutations were allowed. After each simulation, 1%, 5%, 10% of the final individuals were sampled. Each sample was used to reconstruct phylogenies by tracing the genealogy of each individual.

References

- 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. <https://doi.org/10.1534/genetics.110.118661>
- 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. <https://doi.org/10.1093/molbev/msi103>
- 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, *Journal of Mathematical Biology*. Springer Berlin Heidelberg. <https://doi.org/10.1007/s00285-018-1272-4>
- 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. <https://doi.org/10.1371/journal.pgen.1003326>
- Kingman, J.F.C., 1982. The coalescent. *Stoch. Process. their Appl.* 13, 235–248. [https://doi.org/10.1016/0304-4149\(82\)90011-4](https://doi.org/10.1016/0304-4149(82)90011-4)
- 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. <https://doi.org/10.1111/j.1365-294X.2006.03046.x>

- 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., O'Neill, L., Stratton, M.R., Laurenti, E., Green, A.R., Kent, D.G., Campbell, P.J., 2018. Population dynamics of normal human blood inferred from somatic mutations. *Nature* 561, 473–478. <https://doi.org/10.1038/s41586-018-0497-0>
- Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496. <https://doi.org/10.1038/nature10231>
- Liu, X., Fu, Y.X., 2015. Exploring population size changes using SNP frequency spectra. *Nat. Genet.* 47, 555–559. <https://doi.org/10.1038/ng.3254>
- 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 (Edinb.)* 116, 362–371. <https://doi.org/10.1038/hdy.2015.104>
- 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. <https://doi.org/10.1111/j.1365-294X.2010.04783.x>
- Strimmer, K., Pybus, O.G., 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol. Biol. Evol.* 18, 2298–2305. <https://doi.org/10.1093/oxfordjournals.molbev.a003776>
- Wakeley, J., 1999. Nonequilibrium migration in human history. *Genetics* 153, 1863–1871.

Acknowledgments

We are grateful to the members of the He lab for comments on the manuscript.

Author Contributions: H.X., D.S., L.K. and Y.Z. conceived the study; D.S., L.K., Y.Z. and H.X. analyzed data; H.X. supervised the study and wrote the manuscript with inputs from co-authors.

Author Information: The authors declare no competing interests.

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 sub-populations. 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 sub-populations 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 sub-populations experiencing exponential-growth: $N(t) = N(0)e^{-rt}$, with $r = 0.005$ for each sub-population 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 sub-population has initial sizes ($N_i(0) = N(0)/n$ for sub-population Bi).

Fig. 3. Inference of population size dynamics for a simulated population consisting of two sub-populations.

a. A simulated panmictic population, A . The demographic history is depicted with solid black line. **b.** A simulated structured population, B , consisting of two sub-populations 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%, 10% of the final population, as shown above each subplot.

Fig. 4. Inference of population size dynamics for a simulated population consisting of five sub-populations.
a, b. The same simulations as that shown in Fig. 3, except that the simulated structured population (*B*) in **b** consists of five sub-populations, each with distinct demographic histories.