Title

Distortion of genealogical properties when the sample is very large

Permalink
https://escholarship.org/uc/item/6gp3d1qn

Journal
Proceedings of the National Academy of Sciences of the United States of America, 111(6)

ISSN
0027-8424

Authors
Bhaskar, Anand
Clark, Andrew G
Song, Yun S

Publication Date
2014-02-11

DOI
10.1073/pnas.1322709111

Peer reviewed

Distortion of genealogical properties when the sample is very large

Anand Bhaskar[a], **Andrew G. Clark**[b,1], **and Yun S. Song**[a,c,1]

[a]Computer Science Division and [c]Department of Statistics, University of California, Berkeley, CA 94720; and [b]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853

Contributed by Andrew G. Clark, December 6, 2013 (sent for review July 20, 2013)

Study sample sizes in human genetics are growing rapidly, and in due course it will become routine to analyze samples with hundreds of thousands, if not millions, of individuals. In addition to posing computational challenges, such large sample sizes call for carefully reexamining the theoretical foundation underlying commonly used analytical tools. Here, we study the accuracy of the coalescent, a central model for studying the ancestry of a sample of individuals. The coalescent arises as a limit of a large class of random mating models, and it is an accurate approximation to the original model provided that the population size is sufficiently larger than the sample size. We develop a method for performing exact computation in the discrete-time Wright–Fisher (DTWF) model and compare several key genealogical quantities of interest with the coalescent predictions. For recently inferred demographic scenarios, we find that there are a significant number of multiple- and simultaneous-merger events under the DTWF model, which are absent in the coalescent by construction. Furthermore, for large sample sizes, there are noticeable differences in the expected number of rare variants between the coalescent and the DTWF model. To balance the trade-off between accuracy and computational efficiency, we propose a hybrid algorithm that uses the DTWF model for the recent past and the coalescent for the more distant past. Our results demonstrate that the hybrid method with only a handful of generations of the DTWF model leads to a frequency spectrum that is quite close to the prediction of the full DTWF model.

Human genetics has entered a new era in which the study sample sizes regularly exceed 10,000, a number commonly cited as the effective population size of humans (1–4). A consistent finding arising from recent large-sample studies (5–8) is that human genomes harbor a substantial excess of rare variants compared with that predicted using previously applied demographic models. For example, Nelson et al. (6) found that over 70% of single-nucleotide variants are singletons and doubletons, which corresponds to a minor allele frequency on the order of 0.01% for their study sample. There are several factors that may contribute to the discrepancy between observations in the data and theoretical predictions, including the following possible explanations:

i) Previously applied demographic models are wrong. In particular, the observed polymorphism patterns are indicative of a recent rapid growth of the effective population size, much more rapid than in previously applied demographic models. This conclusion would be consistent with historical records of census population size (9).

ii) Population substructure (10, 11) and natural selection have distorted the observed polymorphism patterns while previous demographic inference studies have failed to adequately account for these factors.

iii) Theoretical predictions for a given demographic model are inaccurate when the sample size is very large. Coalescent theory, which arises as a limit of a large class of discrete-time random-mating models, provides an accurate approximation to the original discrete-time model only if the effective population size is sufficiently larger than the sample size. Violation of this assumption may distort genealogical properties in a way that may inflate rare variants relative to the predictions of coalescent theory.

The goal of this paper is to investigate the last possibility in detail, by examining the deviation between the coalescent and a well-known discrete-time random model, namely the Wright–Fisher (WF) model.

Kingman's coalescent (12–14), henceforth simply referred to as the coalescent, is a central model in modern population genetics for studying the ancestry of a sample of individuals taken from a large randomly mating population. The coalescent is a continuous-time Markov process that can be constructed as a scaling limit of a discrete-time Wright–Fisher (DTWF) model, by taking the population size to infinity while rescaling the unit of time by the population size. The dynamics of a DTWF model can be complicated, in which multiple sets of lineages can find common ancestors in a single generation. In contrast, at most two lineages can find a common ancestor at any given time under the coalescent, and hence it is a mathematically and algorithmically more tractable model. The coalescent is an excellent approximation to the original discrete-time model if, for all times, the population size is sufficiently large relative to the number of ancestral lineages of a sample, in which case multiple and simultaneous mergers of lineages in a single generation are unlikely.

In this paper, we investigate whether the coalescent continues to be a good approximation to the DTWF model in the case in which the sample size increases to the point where the coalescent assumptions may be violated. We compare the two models under certain demographic scenarios previously considered in the literature, including the case of recent rapid population growth for humans (7, 15). We examine several key genealogical statistics of

Significance

Sample sizes in population genomic studies are rapidly increasing to the point where assumptions underlying analytical tools may be violated. This theoretical work examines the accuracy of a widely used probabilistic model, called the coalescent, for describing the ancestry of a sample of individuals. A method for performing exact computation of various genealogical quantities is developed here, and it is shown that the coalescent prediction of rare variants can be noticeably inaccurate when the sample is very large. A hybrid algorithm, which combines discrete- and continuous-time models, is proposed to balance the trade-off between accuracy and computational efficiency.

Author contributions: A.B., A.G.C., and Y.S.S. designed research; A.B. and Y.S.S. performed research; A.B. and Y.S.S. contributed new reagents/analytic tools; and A.B., A.G.C., and Y.S.S. wrote the paper.

The authors declare no conflict of interest.

[1] To whom correspondence may be addressed. E-mail: ac347@cornell.edu or yss@cs.berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1322709111/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1322709111

interest such as the number of multiple and simultaneous mergers in the DTWF model, the number of lineages as a function of time (NLFT), and the sample frequency spectrum. A key feature of our work is that all our results, under both the coalescent and the DTWF model, are based on exact deterministic computations rather than Monte-Carlo simulations.

To perform exact computation in the DTWF model, we exploit the Markov property of the model and devise dynamic programming algorithms to compute various genealogical quantities of interest exactly. These algorithms are computationally expensive, so we also consider a hybrid method that uses the DTWF model for the recent past and the coalescent for the more distant past. We demonstrate that this hybrid approach produces substantially more accurate predictions than does the coalescent, while being more efficient than performing computation in the full DTWF model.

Results

Demographic Models. In addition to the case of a constant population size, we consider three models of variable population size. The details of the demographic models we consider are provided below and illustrated in *SI Appendix*, Fig. S1.

Model 1. This model has a constant population size of 10,000 diploid individuals (3, 4).

Model 2. Proposed by Keinan et al. (16), this model has two population bottlenecks, the most recent of which lasted for 100 generations starting from 620 generations in the past, and a more ancient bottleneck lasting 100 generations, starting from 4,620 generations in the past. Further back in time, the population size is fixed at 10,085 diploids.

Model 3. This demographic model was inferred by Gravel et al. (15) for the CEU subpopulation from the 1000 Genomes (17) exon pilot data. In this model, a population expansion in the last 920 generations occurs at a rate of 0.38% per generation.

Model 4. This demographic model was inferred by Tennessen et al. (ref. 7, figure 2*B*) for the CEU subpopulation from exome sequencing of 2,440 individuals. The ancient demography is similar to that in model 3. However, following the most recent bottleneck, there are two epochs of exponential expansion in the most recent 920 generations—a slower expansion phase for 716 generations at 0.307% per generation, followed by a rapid expansion rate of 1.95% per generation for 204 generations.

Although other models exhibiting recent rapid population expansion have been inferred (6, 8), we focus on model 3 and model 4 in this paper because we want to consider sample sizes that are on the order of the current effective population size while also considering demographic models that incorporate realistic changes in the population size. Due to computational limitations, the largest sample size for which we can perform exact computation in the DTWF model is on the order of 10^5 haploids, which is of the same order of magnitude as the effective population size in model 3, and about 10% of the current effective population size of model 4.

Using the above four demographic models, we examine deviations in the following quantities between the coalescent and the DTWF model: (*i*) multiple and simultaneous mergers in the DTWF model, (*ii*) NLFT, and (*iii*) expected sample frequency spectrum.

Multiple and Simultaneous Mergers in the DTWF Model. For a given demographic model in the DTWF framework, it becomes more likely that multiple lineages may be lost in a single generation as the sample size n increases. The first-order approximations used in the derivation of the coalescent from the DTWF model assume that the sample size n is smaller in order than $\sqrt{N}$, with N being the population size. For example, consider a sample of size $n = 250$ with an effective population size of $n = 20{,}000$ haploids. *SI Appendix*, Fig. S2, shows the probability distribution of the number of parents of the sample in the previous generation. There is a high probability that the sample will have less than $n - 1$ parents in the previous generation, an event that is ignored in the asymptotic calculation used in the coalescent derivation from the DTWF model. *SI Appendix*, Fig. S3, shows the expected fraction of lineages (relative to $n - 1$) that are lost due to either multiple or simultaneous mergers, from the present up to time t in the past. *SI Appendix*, Table S1, shows the numerical values of the expected fraction as $t \to \infty$. The sharp jump in the plot for model 2 (*SI Appendix*, Fig. S3*b*) corresponds to the beginning (backward in time) of population bottlenecks when the population size declines substantially, thus instantaneously increasing the rate at which lineages find common ancestors and are lost. For small sample sizes relative to the population size, it is unlikely for more than one lineage to be lost in a single generation, as can be seen in the plots for $n = 20$ and $n = 200$. In contrast, for large sample sizes ($n = 2 \times 10^4$), almost all of the lineages are lost in generations when more than one lineage is lost.

When multiple lineages are lost in a single generation of the DTWF model, there are several ways this could happen. For example, suppose two out of m lineages are lost in one generation. This could be the result of three lineages finding the same parent in the previous generation, or two pairs of lineages each finding a common parent, with the two parents being different. In general, there are $S(m, j)$ different ways that m labeled lineages can have j distinct parents in the previous generation, where $S(m, j)$ is the Stirling number of the second kind, counting the number of ways of partitioning a set of m labeled objects into j nonempty subsets. A particular pattern of mergers of m lineages that leads to j distinct parents, where $\lceil \frac{m}{2} \rceil \le j \le m$, is illustrated in *SI Appendix*, Fig. S4. Here, $m - j$ pairs of lineages each find a common parent distinct from all other parents, and the remaining $2j - m$ lineages do not merge with any other lineages. There are j ancestral lineages left after this type of merger. We call this an "$(m - j)$-pairwise simultaneous merger." For $k \ge 2$, we use the term "k-merger" to denote an event where exactly k lineages find the same common parent in the previous generation. It is possible to have several multiple merger events in a single generation. For example, a j-pairwise simultaneous merger is equivalent to there being exactly j 2-merger events and no other merger events in a single generation.

In the coalescent, because at most two lineages find a common ancestor in any given time, the only kind of possible merger is a single 2-merger (or equivalently, a 1-pairwise simultaneous merger). However, in a DTWF model with m lineages at a given time, there are $\frac{1}{2}\binom{m}{2}\binom{m-2}{2}$ possible 2-pairwise simultaneous mergers, and $\binom{m}{3}$ possible 3-mergers, yielding the following expression for the total number of different ways for m lineages to find $m - 2$ distinct parents in the previous generation:

$$S(m, m-2) = \binom{m}{3} + \frac{1}{2}\binom{m}{2}\binom{m-2}{2}. \qquad [1]$$

Because the second term is $O(m^4)$, whereas the first term is $O(m^3)$, for large m we expect 2-pairwise simultaneous mergers to be the dominant reason for losing two lineages in a single generation.

SI Appendix, Fig. S5, illustrates the ratio of the sum of the expected number of lineages lost due to k-pairwise simultaneous mergers, for $k \ge 2$, to the results shown in *SI Appendix*, Fig. S3, the expected number of lineages lost due to multiple or simultaneous mergers, from the present up to time t in the past. As *SI Appendix*, Fig. S5, shows, a substantial fraction of the lineages that are lost in generations with multiple lost lineages (i.e., in generations with mergers forbidden in the coalescent) are due to pairwise-simultaneous mergers. Incidentally, that the curves for $n = 20$ start out near 0.93 can be attributed to the fact that the ratio of the second term in the right-hand side of Eq. **1** to $S(m, m-2)$ is $\frac{51}{55}$ for $m = 20$.

The expected fraction (relative to $n-1$) of lineages lost due to k-mergers is shown in Table 1. A substantial number of lineages are lost to 3-mergers in model 1, model 2, and model 3 for $n = 2 \times 10^4$ because the sample size is of the same order as the population size at time 0. Even in model 4, about 1.9% of lineages participate in 3-mergers. *SI Appendix*, Fig. S6, shows the fraction of 3-mergers up to time t relative to the total expected number of 3-mergers as $t \to \infty$. As expected, in model 1, model 2, and model 3, due to the large sample size relative to the population size at time 0, a substantial portion of the 3-mergers take place very early when the number of surviving lineages drops quickly. It is rather surprising that, in model 4, where there is a rapid exponential population growth, a large fraction of the 3-mergers in fact take place during this period of growth. In particular, more than 25% of the expected 3-mergers for $n = 2 \times 10^4$ occur in the most recent 32 generations when the effective population size is at least 5.5×10^5.

Based on the results described above, one would expect that the number of ancestral lineages remaining at a given time decreases more rapidly under the DTWF model than under the coalescent, and we investigate this quantity next.

NLFT. Here, we compare the expected NLFT in the coalescent and in the DTWF model. In what follows, we let $A_n^C(t)$ and $A_n^D(t)$ denote the random variables for the number of lineages at generation t in the coalescent and the DTWF model, respectively, starting with a sample of size n at time 0. Under the coalescent, the expectation and SD of the NLFT, $\mathbb{E}[A_n^C(t)]$ and $\sigma(A_n^C(t))$, can be computed exactly in a numerically stable fashion for an arbitrary variable population size model as described in *SI Appendix*, *SI Text*. An algorithm to compute $\mathbb{E}[A_n^D(t)]$ and $\sigma(A_n^D(t))$ under the DTWF model is also described there.

For the four demographic models considered, *SI Appendix*, Fig. S7, shows the expectation and SD of the NLFT under the DTWF model, whereas *SI Appendix*, Figs. S8 and S9, show the relative differences in the expectation and SD, respectively, of the NLFT in the coalescent with respect to the NLFT in the DTWF model. For large sample sizes under model 1, model 2, and model 3, it can be seen that the lineages are lost at a faster rate in the DTWF model than in the coalescent. This pattern is consistent with the fact that these demographic models exhibit a substantial number of 3-mergers in the DTWF model for large sample sizes (Table 1), although the deviation in the expected NLFT is still substantially less than the expected number of 3-mergers. The deviation disappears after about 1,000 generations when enough time has passed for the number of ancestral lineages to become sufficiently small that the coalescent approximation holds.

For model 4, the expected NLFT in the coalescent provides a fairly good approximation to that in the DTWF model for all times and for all sample sizes considered. This is because the population size remains much larger than the number of ancestral lineages at all times.

Expected Sample Frequency Spectrum. Given a sample of n haploid (or $n/2$ diploid) individuals, a common summary of the sample used in various population genetic analyses is the sample frequency spectrum, $\hat{\boldsymbol{\tau}}_n = (\hat{\tau}_{n,1}, \ldots, \hat{\tau}_{n,n-1})$. Under the infinite-sites model of mutation, the kth entry $\hat{\tau}_{n,k}$ corresponds to the number of polymorphic sites in the sample that have k derived alleles and $n-k$ ancestral alleles, where $1 \le k \le n-1$. For a sample of n haploids randomly drawn from the population, we denote the expected value of $\hat{\tau}_{n,k}$ in the coalescent and the DTWF models by $\tau_{n,k}^C$ and $\tau_{n,k}^D$, respectively. In the case of a constant population size, $\tau_{n,k}^C$ under the infinite-sites model of mutation is given exactly by the following expression:

$$\tau_{n,k}^C = \frac{\theta}{k}, \qquad [2]$$

where θ denotes a population-scaled mutation rate. (Mutations arise according to a Poisson process with intensity $\theta/2$ in each lineage, independently of all other lineages.) For variable population size models, the results of Polanski and Kimmel (18) can be used to compute the expected sample frequency spectrum numerically stably under the coalescent. In *SI Appendix*, *SI Text*, we develop an algorithm to compute the expected sample frequency spectrum under the DTWF model, denoted by $\boldsymbol{\tau}_n^D = (\tau_{n,1}^D, \ldots, \tau_{n,n-1}^D)$.

Fig. 1 illustrates the relative difference between the coalescent and the DTWF model in the number of singletons ($\tau_{n,1}$) and doubletons ($\tau_{n,2}$) as a function of the sample size (n). As the figure shows, the number of singletons predicted by the DTWF model is larger than the coalescent prediction by as much as 11% when the sample size is comparable to the current population size. It is interesting to note that, even though there are a substantial number of 3-mergers and 4-mergers in model 1 and model 2, the deviations in the frequency spectrum are not nearly as large as one might have expected. This is probably because even though the coalescent forbids multiple mergers by construction, successive 2-mergers can be separated by arbitrarily small amounts of time (as opposed to being separated by at least one generation in a discrete model). This allows the coalescent to simulate the effect of multiple mergers without explicitly allowing them, leading to fairly similar frequency spectra as a DTWF model.

The deviations in the singletons and doubletons for model 1 match those computed by Fisher (19) (and tabulated in ref. 20, table 1) when the sample size equals the population size and in the limit that the population size tends to infinity. For model 4 (Fig. 1*D*), we could not consider sample sizes $>10^5$ because of

Table 1. Expected percentage of lineages (relative to $n - 1$, where n is the sample size) lost due to k-mergers in models 1–4

	Model 1		Model 2		Model 3		Model 4	
k	$n = 2 \times 10^3$	$n = 2 \times 10^4$	$n = 2 \times 10^3$	$n = 2 \times 10^4$	$n = 2 \times 10^3$	$n = 2 \times 10^4$	$n = 2 \times 10^3$	$n = 2 \times 10^4$
2	96.68%	68.70%	96.66%	68.93%	98.77%	89.99%	98.96%	98.11%
3	3.24%	23.03%	3.26%	22.93%	1.22%	9.25%	1.03%	1.87%
4	0.08%	6.44%	0.08%	6.36%	0.01%	0.72%	0.01%	0.02%

In model 1 and model 2 for $n = 2 \times 10^4$, a substantial number of lineages are involved in 3-mergers, and more than 6% of the lineages are involved in 4-mergers, because the sample size is of the same order as the current population size. Even in model 3 and model 4 for $n = 2 \times 10^4$, around 9% and 2% of the lineages participate in 3-mergers, respectively.

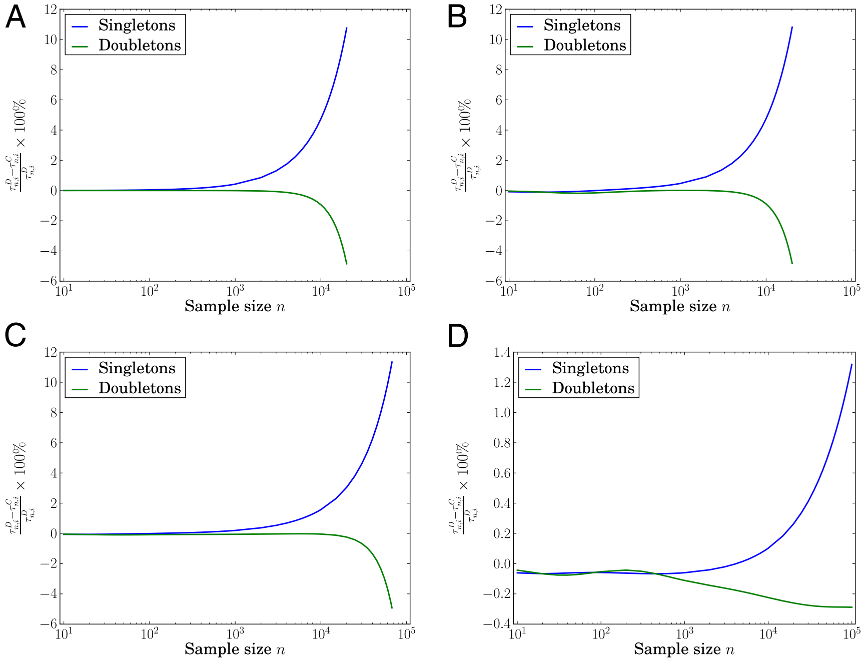

Fig. 1. The percentage relative error in the number of singletons and doubletons between the coalescent and DTWF models, as a function of the sample size *n*. When the sample size is comparable to the current population size, the number of singletons predicted by the DTWF model is larger than the coalescent prediction by as much as 11%, whereas the number of doubletons predicted by the DTWF model is smaller than the coalescent prediction by about 4.8%. In model 4, we could not consider a sample size comparable to the population size (10^6) because of computational burden, but we expect a similar extent of deviation as in models 1–3 as *n* increases. Note that the *y*-axis scale for model 4 is different from that for models 1–3. (*A*) Model 1. (*B*) Model 2. (*C*) Model 3. (*D*) Model 4.

computational burden, but the results for models 1–3 suggest that we should expect to observe ≥10% deviation when the sample size n is increased to 10^6, the current population size in model 4. The deviation in the number of doubletons is also significant when the sample size is comparable to the current population size; the DTWF prediction for doubletons is smaller than the coalescent prediction by about 4.8%.

The findings described above are especially important given that rare variants comprise a large fraction of segregating sites when the sample size is large. In *SI Appendix*, Fig. S10, we plot the cumulative distribution of the frequency spectrum in the DTWF model for models 1–4. The number of singletons in models 3 and 4 is higher than in models 1 and 2 due to exponential population growth. The rapid population expansion in model 4 results in about 51% of the segregating sites being singletons and over 80% of the segregating sites having less than five copies of the derived allele in a sample of size $n = 2 \times 10^4$. *SI Appendix*, Fig. S11, shows the expected proportion of rare variants (derived allele frequency $\leq 0.01\%$) as a function of the sample size n for models 3 and 4 under the coalescent. It can be seen that as n approaches the current population size, the proportion of rare variants increases substantially. *SI Appendix*, Fig. S12, shows the expected proportion of segregating sites that are singletons as a function of n for models 1–4 under the coalescent. For small sample sizes (say, $n < 100$), the proportion of singletons in models 3 and 4 (which incorporate rapid recent population expansion) is not much larger than that in models 1 and 2. However, the difference increases considerably as the sample size goes beyond a few hundred individuals, illustrating the need for large sample sizes to infer recent population expansion from frequency spectrum data.

A Hybrid Method for Computing the Frequency Spectrum. As detailed in *SI Appendix*, *SI Text*, computation in the DTWF model is substantially more involved than in the coalescent. In particular, although the run time of the frequency spectrum computation in the coalescent depends only on the number of piecewise-exponential epochs and not the duration of each epoch, the run time of our dynamic programming algorithm for the DTWF model actually depends on the number of generations over which the algorithm is run. Because noticeable deviation between the DTWF model and the coalescent arises when the number of ancestral lineages is not negligible compared with the population size, a reasonable trade-off between accuracy and run time would be to use the DTWF model for the recent past and the coalescent for the more distant past (when the number of ancestral lineages has decreased sufficiently).

To explore this idea, we implemented a hybrid method for computing the frequency spectrum that, for a specified switching generation t_s, uses the full DTWF model for generations $0 < t \leq t_s$, followed by the coalescent for generations $t > t_s$. In particular, for $t_s = 0$, this algorithm computes the frequency spectrum under the coalescent, whereas for $t_s = \infty$, it computes the frequency spectrum under the full DTWF model. As Fig. 2 illustrates for model 3 in the case in which the sample size n is equal to the current effective population size N_0, the difference in the frequency spectrum between the full DTWF model and the hybrid algorithm decreases rapidly as t_s increases. With $t_s = 5$ generations, the largest deviation in the number of singletons is less than 1%, which is a substantial reduction from 11% for $t_s = 0$ (Fig. 1*C*). *SI Appendix*, Fig. S13, shows these data in a different way, in which the deviations in the frequency spectrum for model 3 between the full DTWF model and the hybrid algorithm are shown as a function of sample size for several values of t_s.

Discussion

Several analyses of genomic sequence variation in large samples of humans (6–9) have found a substantial excess of rare variation compared with those predicted using previously applied

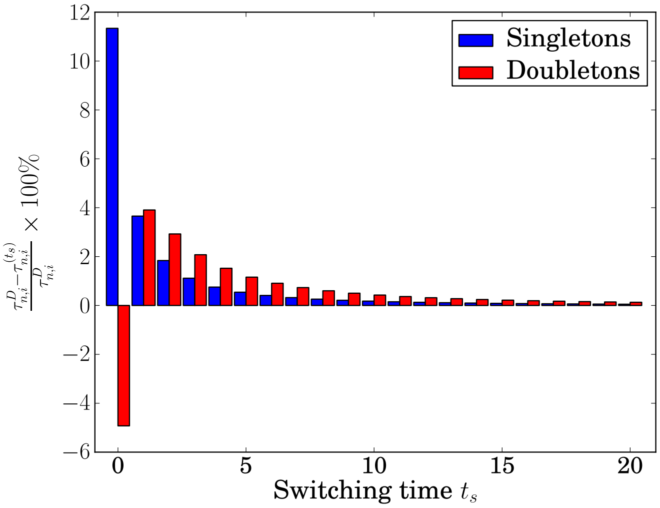

Fig. 2. The percentage relative error, with respect to the full DTWF model, in the number of singletons and doubletons in a hybrid algorithm with switching time t_s. The hybrid method uses the DTWF model for generations $\leq t_s$ and the coalescent model in generations $>t_s$. The results are for model 3 in the case in which the sample size n is equal to the current effective population size N_0 = 67,627. The case of t_s = 0 corresponds to using the coalescent model only. This plot shows that the difference in the frequency spectrum between the full DTWF model and the hybrid algorithm decreases very rapidly as the switching time t_s increases.

demographic models. The inference in these studies is that these results are consistent with a rapid growth of the effective population size in the recent past (much more rapid than in previously applied demographic models), a conclusion consistent with historical records of census population size (9). These studies also used sample sizes that would appear to be large enough to violate assumptions of the coalescent, potentially distorting genealogical properties in a way that may inflate rare variation relative to the predictions of coalescent theory. In this paper, we have investigated this issue by developing a method for performing exact computation in the DTWF model of random mating. We have studied the deviation between the coalescent and the WF model for several key genealogical quantities that are used for population genomic inference.

For several recently inferred demographic scenarios for humans, our results show that there are a significant number of multiple- and simultaneous-merger events under the WF model that are ignored by construction of the coalescent. Furthermore, there are noticeable differences in the expected number of rare variants between the coalescent and the DTWF model, especially in the regime where the sample size is on the order of the current effective population size. Even if the demographic models considered here might underestimate the true current effective population size of humans, sample sizes in population genetic studies are rapidly increasing and might grow to be large enough to cause the differences between the DTWF and the coalescent to become amplified.

A number of demographic inference methods are based on fitting the expected frequency spectra under the coalescent (5–7) or the diffusion process (15, 21–23) to observed data. For instance, the exponential growth parameters in models 3 (15) and 4 (7) were inferred using a likelihood method based on the diffusion process approximation to the DTWF model, by fitting the predicted frequency spectrum to polymorphism patterns observed in a sample size of 876 individuals and 2,440 individuals, respectively. Because the diffusion process approximation to the DTWF model is equivalent to the coalescent approximation, the differences in the frequency spectrum (Fig. 1) between the coalescent and the DTWF model indicate that we might infer different demographies if the analysis were done using the DTWF model. In particular, for a sample of size n analyzed under the DTWF model, any inferred demography will have a current effective population size of at least n. However, the coalescent imposes no such restriction on the inferred current effective population size. In fact, under the coalescent, it is even possible to estimate a current effective population size N_0 that is smaller than the sample size n. This is because one can only infer a scaling function of time in the coalescent, which is the ratio of the variable effective population size to a fixed reference population size. The inferred scaling function can then be transformed into an effective population size function by using the empirically estimated per-generation mutation rate (15), or by setting the reference population size to a specific value (5, 6) [e.g., using an ancestral population size inferred by previous studies (24)].

To balance the trade-off between accuracy and computational efficiency, we have proposed a hybrid algorithm that uses the DTWF model for the recent past and the coalescent for the more distant past. This hybrid algorithm provides a way to obtain more accurate predictions of the frequency spectrum than in the coalescent, while being computationally more efficient than considering the full DTWF model. We leave the exploration of this method for demographic inference as future research.

Wakeley and Takahashi (20) have provided asymptotically accurate expressions (as the effective population size $N \to \infty$) for the number of singletons and the number of segregating sites under a variant of the DTWF model that allows for a larger number of offspring than the effective population size, assuming that the effective population size stays constant over time. Fu (25) has also examined the accuracy of the standard coalescent model and proposed an alternative continuous-time "exact" coalescent model applicable in the regime when $N(N-1)\ldots(N-n+1) \times N^{-n} \gg 0$, where N denotes the effective population size and n the sample size. That work was restricted to the case of a constant population size, whereas in this paper we have considered several demographic scenarios inferred from recent large-scale population genomic studies. Moreover, for some of the demographic scenarios and sample sizes considered here, the assumption in Fu's work (25) that $N(N-1)\ldots(N-n+1) \times N^{-n} \gg 0$ is violated. Wakeley et al. (26) have shown that it is difficult to reject the coalescent even for data generated using fixed pedigrees with random genetic assortment. Our work is complementary to that study and compares the coalescent to the DTWF random-mating model.

In this paper, we have focused on the DTWF model for simplicity. However, it is known that under some weak conditions on the limiting probabilities of a 2-merger and a 3-merger, a large family of exchangeable random-mating models converge to the same coalescent limit as the unit of time is rescaled appropriately and the population size gets large (27, 28). The rate of convergence to the coalescent differs between different random-mating models (29), and hence the accuracy of coalescent predictions for large sample sizes depends on the random-mating model being considered. The deviation from the coalescent could be amplified for other random-mating models. It would be interesting to consider the accuracy of the coalescent for other random and realistic models of relevance to human genetics; e.g., models in which generations overlap and the distribution of offspring number more closely reflects the observed pattern for human populations [for example, the Swedish family data of Low and Clarke (30) or the Saguenay-Lac-Saint-Jean population considered by Moreau et al. (31)]. Despite having access to large samples, recent studies (6–8) have inferred much smaller current effective population sizes (on the order of millions) than the current census size (on the order of billions) of the population from which the samples were drawn. It is possible that demographic inference methods that explicitly model realistic human mating patterns might be able to infer census population size histories more accurately than does the coalescent, which assumes random mating and can only infer effective population sizes that do not have a direct census interpretation.

Furthermore, it would be interesting to compare discrete-time random models and the coalescent with respect to haplotype sharing (identity-by-descent and identity-by-state), linkage disequilibrium, and natural selection when the sample size is very large. For example, Davies et al. (32) used simulations to demonstrate that, for a constant population size model, recombination and gene conversion can increase the number of ancestral lineages of a sample of chromosomes to the extent that multiple and simultaneous mergers in the DTWF model can lead to substantial differences from the coalescent model in the rates of coalescence and in the number of sequences carrying ancestral material. It would be interesting to perform such comparisons for more realistic demographic models for humans.

We will soon enter an era in which it will become routine to analyze samples with hundreds of thousands if not millions of individuals. For these large sample sizes, the standard coalescent will no longer serve as an adequate model for evolution. The DTWF model is mathematically cumbersome to work with, which was one of the original motivations for adopting the coalescent for modern population genetics analyses. However, for these very large sample sizes, we will need to develop new mathematically and computationally tractable stochastic processes that better approximate realistic models of human population evolution, and under which we can efficiently compute genealogical quantities like we have been able to under the coalescent.

Materials and Methods

The computation of the various genealogical quantities in the DTWF model, such as the number of simultaneous and multiple mergers, the NLFT, and the expected frequency spectrum, rely on the Markov property of the DTWF model. By considering the types and counts of the mergers occurring in the previous generation, one can write down one-step recurrence equations relating these genealogical quantities over time and solve these recurrences by dynamic programming. The details of these recurrence equations for the various genealogical quantities considered in this manuscript are provided in *SI Appendix*, *SI Text*, and software programs implementing them can be downloaded from www.eecs.berkeley.edu/~yss/software.

ACKNOWLEDGMENTS. We are very grateful to Phil Green, John Novembre, and John Wakeley for helpful comments on an earlier version this paper. We also thank John Kamm for useful discussions. This research is supported in part by National Institutes of Health Grants R01-HG003229 and R01-GM094402, and a Packard Fellowship for Science and Engineering.

1. Takahata N (1993) Allelic genealogy and human evolution. *Mol Biol Evol* 10(1):2–22.
2. Erlich HA, Bergström TF, Stoneking M, Gyllensten U (1996) HLA sequence polymorphism and the origin of humans. *Science* 274(5292):1552–1554.
3. Harding RM, et al. (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet* 60(4):772–789.
4. Harpending HC, et al. (1998) Genetic traces of ancient demography. *Proc Natl Acad Sci USA* 95(4):1961–1967.
5. Coventry A, et al. (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* 1:131.
6. Nelson MR, et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337(6090):100–104.
7. Tennessen JA, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64–69.
8. Fu W, et al. (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493(7431):216–220, and correction (2013) 495(7440): 270.
9. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336(6082):740–743.
10. Ptak SE, Przeworski M (2002) Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet* 18(11):559–563.
11. Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P (2009) The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* 182(1):205–216.
12. Kingman JFC (1982) The coalescent. *Stochastic Process Appl* 13(3):235–248.
13. Kingman JFC (1982) Exchangeability and the evolution of large populations. *Exchangeability in Probability and Statistics*, eds Koch G, Spizzichino F (North-Holland Publishing Company, Amsterdam), pp 97–112.
14. Kingman JFC (1982) On the genealogy of large populations. *J Appl Probab* 19:27–43.
15. Gravel S, et al. (2011) Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA* 108(29):11983–11988.
16. Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* 39(10):1251–1255.
17. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.
18. Polanski A, Kimmel M (2003) New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* 165(1):427–436.
19. Fisher RA (1930) The distribution of gene ratios for rare mutations. *Proc R Soc Edinb* 50:205–220.
20. Wakeley J, Takahashi T (2003) Gene genealogies when the sample size exceeds the effective size of the population. *Mol Biol Evol* 20(2):208–213.
21. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5(10):e1000695.
22. Lukić S, Hey J, Chen K (2011) Non-equilibrium allele frequency spectra via spectral methods. *Theor Popul Biol* 79(4):203–219.
23. Lukić S, Hey J (2012) Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics* 192(2):619–639.
24. Schaffner SF, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15(11):1576–1583.
25. Fu Y-X (2006) Exact coalescent for the Wright-Fisher model. *Theor Popul Biol* 69(4): 385–394.
26. Wakeley J, King L, Low BS, Ramachandran S (2012) Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics* 190(4):1433–1445.
27. Möhle M, Sagitov S (2001) A classification of coalescent processes for haploid exchangeable population models. *Ann Probab* 29(4):1547–1562.
28. Möhle M, Sagitov S (2003) Coalescent patterns in diploid exchangeable population models. *J Math Biol* 47(4):337–352.
29. Bhaskar A, Song YS (2009) Multi-locus match probability in a finite population: A fundamental difference between the Moran and Wright-Fisher models. *Bioinformatics* 25(12):i187–i195.
30. Low B, Clarke A (1991) Family patterns in nineteenth-century Sweden: Impact of occupational status and landownership. *J Fam Hist* 16(2):117–138.
31. Moreau C, et al. (2011) Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science* 334(6059):1148–1150.
32. Davies JL, Simancík F, Lyngsø R, Mailund T, Hein J (2007) On recombination-induced multiple and simultaneous coalescent events. *Genetics* 177(4):2151–2160.