Gene Genealogies When the Sample Size Exceeds the Effective Size of the Population

John Wakeley and Tsuyoshi Takahashi

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts

We study the properties of gene genealogies for large samples using a continuous approximation introduced by R. A. Fisher. We show that the major effect of large sample size, relative to the effective size of the population, is to increase the proportion of polymorphisms at which the mutant type is found in a single copy in the sample. We derive analytical expressions for the expected number of these singleton polymorphisms and for the total number of polymorphic, or segregating, sites that are valid even when the sample size is much greater than the effective size of the population. We use simulations to assess the accuracy of these predictions and to investigate other aspects of large-sample genealogies. Lastly, we apply our results to some data from Pacific oysters sampled from British Columbia. This illustrates that, when large samples are available, it is possible to estimate the mutation rate and the effective population size separately, in contrast to the case of small samples in which only the product of the mutation rate and the effective population size can be estimated.

Introduction

Although the history of population genetics dates back more than one hundred years, the genealogical approach that characterizes modern work emerged only during the 1970s (Ewens 1972; Karlin and McGregor 1972; Watterson 1975) in response to newly available genetic data (Harris 1966; Lewontin and Hubby 1966). It was soon formalized as the coalescent by Kingman (1982*a*, 1982*b*) and studied extensively from a more biological standpoint by Hudson (1983) and Tajima (1983). The coalescent is intuitively appealing, has a relatively simple mathematical structure, and is easily applied to data. Thus it has led to impressive advances and now frames most work in population genetics. A number of tests of the coalescent null model have been proposed, among them Tajima's (1989) *D* and the statistics of Fu and Li (1993). Because of the overwhelming historical importance of the neutral theory of molecular evolution (Kimura 1983), these tests are often mistakenly viewed as tests of selective neutrality only. However, the standard coalescent model involves a long list of assumptions, and when the model is rejected it is difficult to distinguish among several possible explanations (Simonsen, Churchill, and Aquadro 1995; Nielsen 2001).

In addition to natural selection, demographic factors like population subdivision, population growth, and population decline can cause the model to be rejected. Accepting their lack of specificity, the fact that Tajima's (1989) *D* and the statistics of Fu and Li (1993) have power to detect these deviations can be viewed as advantageous, because subdivision and changes in size are important biological properties of populations. Here we consider an assumption of the coalescent that has mostly been overlooked: the assumption that the sample size is much smaller than the effective size of the population ($n \ll N_e$). We derive expressions for the expected number of singleton polymorphisms and the expected total number of polymorphisms in a sample that can be as large or larger than the effective size of the population. Under the infinite sites model of mutation (Kimura 1969; Watterson 1975), we find that the main effect of large sample size is to increase the number of singletons in the sample relative to coalescent predictions. The increase in the relative number of singletons will give negative values of the statistics mentioned above, and thus will be indistinguishable by these tests from other factors such as population growth (Simonsen, Churchill, and Aquadro 1995). This is clearly undesirable and suggests that the genealogical approach to population genetics should be expanded to include the possibilty that the sample size is not much greater than the effective size of the population.

We use a continuous approximation for the sample size divided by the effective size ($x = n/N_e$) that was previously employed by Fisher (1930) and Watterson (1975). Fisher (1930) studied variability maintained in a large population by the introduction of a single mutant each generation. He used what is now known as the infinite sites model of mutation with free recombination between sites (Kimura 1969) and derived expected values of the numbers of mutants at low frequency (singletons, doublets, etc.), as well as the total number of polymorphisms maintained. In modern terms, Fisher's solution applies when the parameter θ is equal to 2, because θ is defined to be the mutation rate per gene copy times twice the number of gene copies in the population. Here we assume a haploid population, so $\theta = 2N_e u$, but the results can be applied to diploid organisms if $\theta = 4N_e u$. The fact that Fisher assumed exactly one mutant entered the population each generation is irrelevant in comparing predictions about expected levels of polymorphism. He simply assumed that there was no variability in the mutation process, whereas today we model muations in the population as a Poisson process with rate $\theta/2$ per generation. Another difference between Fisher's approach and the modern genealogical one concerns recombination. Under neutrality, however, the expected values derived by Fisher (1930) and Watterson (1975) and those reported by us below do not depend on the recombination rate because the marginal distribution of genealogies at every site is the same regardless of recombination. Predictions about the variances of these quantities would depend on the recombination rate.

Table 1 shows Fisher's (1930) predictions for the expected numbers of mutants in one through five copies in the entire population. Fisher used what is now known as

Key words: coalescent theory, genealogies, effective population size, multiple mergers.

E-mail: wakeley@fas.harvard.edu.

Mol. Biol. Evol. 20(2):208–213. 2003
DOI: 10.1093/molbev/msg024

Table 1
Coalescent and $x = 1$ Predictions for the Expected Number of Mutant Factors Maintained in Low Count in the Population when $\theta = 2$

Mutant Count	Coalescent	$x = 1$	Excess
1	1.000000	1.120458	0.120458
2	0.500000	0.476888	−0.023112
3	0.333333	0.335932	0.002599
4	0.250000	0.250548	0.000548
5	0.200000	0.199881	−0.000119

NOTE:—This is adapted from the table on page 215 in Fisher (1930).

the Wright-Fisher model, in which the effective size of the population is identical to the census size. Thus, table 1 predicts the pattern of variability in a sample whose size is the same as the effective size of the population. The values in the table are scaled in terms of θ. That is, they hold for $\theta = 1$, and predictions for other values are obtained simply by multiplying these values by θ. The column marked "Coalescent" shows what is now clear are the predictions of the standard coalescent model: that θ singletons are expected, θ/2 doublets, θ/3 triplets, and so on (Tajima 1989; Fu 1995). The coalescent predictions are surprisingly close to the actual values, even when the entire population is sampled. They are off by a little more than 12% for singletons, 4.6% for doublets, and by less than 1% for all other classes of mutations. This is surprising because a fundamental property of the coalescent—that at most one common ancestor event can occur in a single generation—does not hold for large samples. We show below, however, that these differences between coalescent predictions and reality can be quite large when the sample size is greater than the effective size of the population.

It is generally accepted that in many cases the effective size of a population will be less than the its actual size (Hartl and Clark 1997; Hedrick 2000), although one exception to this is when the population is subdivided (Wright 1943). This raises the possibility that the sample size in empirical population genetics studies might exceed the effective size of the population. This is likely already the case for hypervariable region 1 of human mitochondrial DNA (mtDNA), for which there are $n = 9388$ sequences available (as of June 2002; see http://db.eva.mpg.de/hvrbase/) and N_e may only be about 5000 (Takahata 1995; Hawks *et al.* 2000). The work we present here shows that the main effect of this will be to increase the proportion of singleton polymorphisms in the sample. Beckenbach (1994) proposed that sample sizes larger than the effective population size could explain such seemingly odd patterns of genetic variation in samples of mtDNA data from Pacific oysters, *Crassostrea gigas*, from British Columbia. We reanalyze their data below and show that they are in fact consistent with small N_e. However, the mutation rate needed to reconcile the dichotomy between abundant polymorhisms and small N_e indicates that $n > N_e$ is not the only explanation for the observed pattern.

Theory

Let $x = n/N_e$ be the scaled sample size from a population of effective size N_e. To allow that x could be greater than 1, we assume a population of constant size N in which only N_e individuals ($N \geqslant N_e$) reproduce and the other $N - N_e$ die without reproducing. Generations are assumed to be discrete; each generation all adults die and are replaced by offspring. We assume that the types of these N offspring are obtained by random sampling with replacement among the N_e individuals that do reproduce. Although we assume a haploid organism, another way to think of this is that N_e individuals each produce a very large number of "gametes" and the next generation (of N individuals) is a random sample from this gamete pool. If $N = N_e$, then this model is identical to the usual Wright-Fisher model. We assume that mutations occur at rate u per gene copy per generation, and we use the scaled mutation rate $\theta = 2N_e u$, because any mutations that happen in the germ lines of the $N - N_e$ indiviuals that do not reproduce are lost.

We seek expressions for the expected number of singleton polymorphisms $E[\eta_1]$ and the expected total number of polymorphic or segregating sites $E[S]$. Because of the Poisson nature of the mutation process, we have

$$E[\eta_1] = \theta\tau_1(x) \tag{1}$$

$$E[S] = \theta\tau(x) \tag{2}$$

where τ_1 and τ are the expected lengths of all the external branches in the genealogy of the sample and the expected total length of the genealogy of the sample, respectively, measured in units of $2N_e$ generations. Under the standard coalescent model (in which $x \to 0$), we have $\tau_1 = 1$ and $\tau = \sum_{i=1}^{n-1} 1/i$, and these results can be obtained in a number of different ways (Watterson 1975; Tajima 1989; Fu and Li 1993; Fu 1995). Here, we take a backwards-looking "balls in boxes" approach. That is, the genealogy of the sample is generated by throwing n balls into N_e boxes, allowing for coalescent events, and repeating this procedure each generation with the remaining ancestral lineages until the most recent common ancestor of the sample is reached. This is a standard method under the coalescent, but when n is large, multiple coalescent events can occur in the same generation.

To obtain $\tau_1(x)$ and $\tau(x)$ here we follow Fisher (1930) and Watterson (1975) and consider a continuous approximation of the scaled sample size as N_e goes to infinity for a given $x = n/N_e$. In this case, we can use the fact that the scaled number of ancestors of the sample of size x converges in probability to its asymptotic mean $1 - e^{-x}$ as N_e goes to infinity; see page 267 in Watterson (1975). Therefore, in the case of $\tau(x)$ we have the following recursion over a single generation,

$$\tau(x) = \frac{x}{2} + \tau(1 - e^{-x}), \tag{3}$$

in which the time parameter is suppressed because we assume that the population is at equilibrium. In words, equation (3) says that the expected total branch length of the genealogy of a sample of size n in this limit is equal to the lengths of branches between now and the previous generation, $n/(2N_e) = x/2$, when time is measured in units of $2N_e$ generations, plus the expected total branch length of the genealogy of the $N_e(1 - e^{-x})$ lineages remaining one generation in the past.

Fisher (1930) and Watterson (1975) found solutions for $\tau(x)$ using series approximations near $x = 0$. These solutions do not hold when x is large but are quite good for $x < 2$ (see Simulations, below). Here we use the fact that $1 - e^{-x}$ is less than 1 for all x, together with Watterson's (1975) results and equation (3) to make predictions for any value of x. In the present notation, Watterson's (1975) equation 1.4b gives

$$\tau^*(x) = \log(n) + \gamma + \frac{1}{2}g(x) \tag{4}$$

in which $g(x)$ is given by Watterson's (1975) equation 2.24, and where $\gamma = 0.57721566\ldots$ is Euler's constant. Thus, we use equation (3), but replace the second term on the right with $\tau^*(1 - e^{-x})$ to make $\tau(x)$ accurate for all x.

Watterson (1975) did not consider mutant allele frequencies, and Fisher (1930) derived the expectations only for mutants in low copy number and assuming $x = 1$ (table 1). However, a solution for $\tau_1(x)$ can be obtained, again via a recursive equation over a single generation. In this case it is necessary to weight the contributions of ancestral lineages by the probability that they have just one descendent in the sample. We obtain

$$\tau_1(x) = \frac{x}{2} + \frac{xe^{-x}}{1 - e^{-x}}\tau_1(1 - e^{-x}). \tag{5}$$

The first term on the right represents the increment to τ_1 in the first generation looking back. It is the same as the first term on the right in equation (3) because all these first-generation branches have just one descendent in the sample. Some proportion of the ancestors of the sample will have one descendent in the sample, but others will have two, three, four, etc. The number of ancestors that have a single descendent in the sample is the same as the number of boxes that contain exactly one ball when n balls are thrown into N_e boxes. Like the case of the total scaled number of ancestors $(1 - e^{-x})$ above, the scaled number of ancestors that have one descendent in the sample of size x converges in probability to its asymptotic mean xe^{-x} as N_e goes to infinity; see Feller (1968, p. 59). Thus, the term multiplying $\tau_1(1 - e^{-x})$ on the right side of equation (5) is equal to the proportion of ancestral lineages that have just one descendent in this limit.

By successively taking derivatives with respect to x on both sides of equation (5), we can obtain a series approximation to the function $\tau_1(x)$ near $x = 0$. This is the method Watterson (1975) used to obtain his equation (2.24) for $g(x)$. Here we obtain

$$\begin{aligned}\tau_1^*(x) = 1 &+ \frac{1}{12}x + \frac{1}{18\times 2!}x^2 + \frac{37}{720\times 3!}x^3 \\ &+ \frac{41}{1080\times 4!}x^4 - \frac{865}{18144\times 5!}x^5 \\ &- \frac{1891}{5670\times 6!}x^6 - \frac{67543}{155520\times 7!}x^7 \\ &+ \frac{601633}{116640\times 8!}x^8 + \cdots,\end{aligned} \tag{6}$$

and this number of terms is sufficient to give $\tau_1^*(1) = 1.120439$ which is close to the value obtained by Fisher (1930) shown in table 1. Of course, we cannot expect a series approximation near $x = 0$ to be accurate for larger x, so we use equation (5), but put $\tau_1^*(1 - e^{-x})$ on the right in place of $\tau_1(1 - e^{-x})$. This gives $\tau_1(1) = 1.120458$ which matches Fisher's (1930) result to six decimal places and makes $\tau_1(x)$ accurate for any x.

Because $\tau(x)$ and $\tau_1(x)$ can be computed, we can use a simple moment method to jointly estimate θ and x. Namely, we equate the observed values of η_1 and S with their expectations (1) and (2), and solve numerically for θ and x. Because $x = n/N_e$ and n is always known, estimating θ and x is equivalent to estimating N_e and u. It is also possible, using the simulations described in the next section, to estimate the likelihood surface for the observed S and η_1 or a posterior distribution of θ and x by Monte Carlo integration over genealogies. We apply both these methods to some mtDNA data from Pacific oysters (Beckenbach 1994; Boom, Boulding, and Beckenbach 1994) under Application to Oyster Data, below.

Simulations

We performed simulations to assess the accuracy of these analytical approximations over a range of values of N_e and to investigate other properties of these large-sample genealogies. The simulations built sample genealogies under the discrete-generations model by randomly choosing the parents of all ancestral lineages each generation. If there are k lineages, this is equivalent to throwing k balls into N_e boxes. The number of balls in each occupied box determines the number of common ancestor or coalescent events, and the full genealogy of the sample was recorded. While k is not small relative to N_e, there can be many coalescent events per generation. The program is written in the C programming language and is available at http://www.oeb.harvard.edu/faculty/wakeley/.

Figure 1 compares simulation results with the predictions from equations (3) and (5) using expressions (4) and (6) on the right-hand sides as described above. The results (4) and (6) using series the approximations near $x = 0$ are also shown. In the case of $\tau(x)$, the predictions of the standard ($n \ll N_e$) coalescent are shown as well. The coalescent prediction for $\tau_1(x)$ is equal to 1 for all x. The simulations presented in figure 1 were performed with $N_e = 1000$ over a range of n from 500 to 10,000 ($x = 0.5$ to $x = 10$). As expected, equations (4) and (6) do not perform well when x is large. In addition, the predictions of the standard coalescent are good only for small x. The predictions using equations (3) and (5) together with equations (4) and (6) are accurate for all x.

Whereas the theory of the previous section focused on singleton polymorphisms, and this is certainly the major effect, figure 2 shows that other components of the site-frequency distribution can also differ markedly from the predictions of the standard coalescent. Looking at equation (5), we can see that, when x is large, nearly all the singleton polymorphisms will be the result of mutations that occurred in the immediately previous generation. Prior to that, few lineages will have only one descendent in the sample. In fact, so many coalescent events will occur in that first generation that doublet, triplet, etc.,

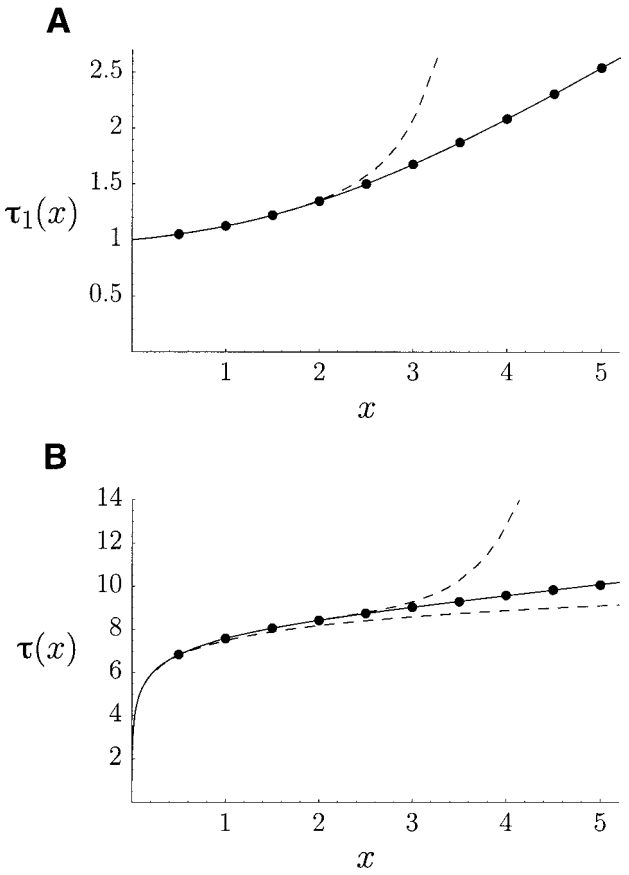

FIG. 1.—Comparison of simulations to analytical results for (*A*) the total length of external branches and (*B*) the total length of the genealogy. Dots are the average values among ten thousand simulation replicates, and solid curves plot the theoretical expectations derived in the text. The dashed curve below in (*B*) is the expectation from the coalescent, and the other dashed lines are series approximation for the expectations around $x = 0$ (see text for details).

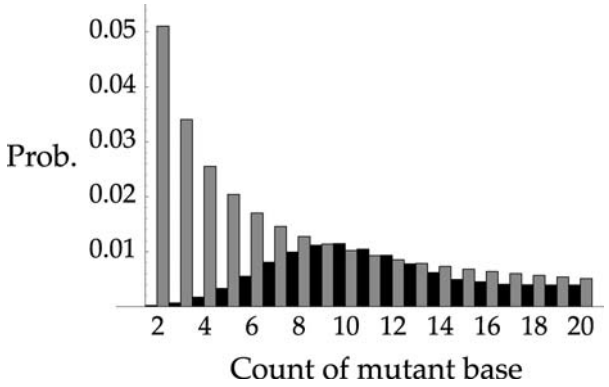

FIG. 2.—The expected proportion of segregating sites at which the mutant base is present in counts ranging from 2 to 20 in a sample of $n = 10000$. Black bars are averages of ten thousand simulation replicates with $N_e = 1000$ ($x = 10$), and grey bars are the analytical prediction of the coalescent (Fu 1995). This is just the far left edge of the distribution; mutant counts can be as large 9999. The values for singleton mutants are not shown; they are 0.397 for the simulated data and 0.102 for the coalescent prediction.

polymorphisms will be underrepresented relative to the standard coalescent. This is evident in table 1, which displays Fisher's (1930) results for $x = 1$. In general, there will be a mode in the site-frequency distribution at mutant counts close to $x(1 - e^{-x})$—approximately x when x is large—which is the expected number of balls per box when n balls are thrown into N_e boxes or, equivalently, the expected number of descendents per lineage. Figure 2 shows this effect when $x = 10$.

We also used simulations to examine the accuracy of the theoretical predictions when N_e is not large. It might have been expected that our results using a continuous approximation for $x = n/N_e$ would not be accurate for smaller n and n and N_e. Surprisingly, our results give accurate predictions over a very broad range of N_e. We do not display these results, but note that the worst case we examined was $n = N_e = 2$. The correct result here is $E[S] = E[\eta_1] = \theta$, whereas our results predict that $E[S] = 1.37\theta$ and $E[\eta_1] = 1.12\theta$.

Application to Oyster Data

It is typical to seek an explanation whenever data show an excess of singleton polymorphisms relative to the predictions of the coalescent, for instance whenever Tajima's (1989) D is negative. The results presented under Theory, above, show that a sample size close to or larger than the effective size of the population can explain an excess of singletons. Thus, if such a pattern is observed, for instance if Tajima's (1989) D is significantly negative, it may be appropriate to fit the model we considered here to the data. Note that if the excess of singletons is greater than about 12% (table 1, fig. 1), the model will estimate N_e to be less than the sample size n. Thus, the present model should probably not be applied if n is small.

Boom, Boulding, and Beckenbach (1994) sampled $n = 141$ Pacific oysters, *C. gigas*, from British Columbia and performed restriction enzyme digests of their mtDNA. Subsequently, Beckenbach (1994) analyzed the pattern of these restriction fragment length polymorphism (RFLP) in the context of the infinite alleles mutation model (Ewens 1972). He proposed that samples sizes larger than the effective population size could explain the overabundance of low-frequency haplotypes (*i.e.*, ones found in a single copy, or a few copies, in the sample of $n = 141$) in British Columbian *C. gigas*. Beckenbach (1994) used simulations to show that large sample size can explain such a pattern, with most single-copy haplotypes resulting from mutations in the immediately previous generation and the few middle frequency haplotypes resulting from mutations that occurred earlier in the history.

To illustrate the application of our results, we reanalyzed the data of Boom, Boulding, and Beckenbach (1994), but from the perspective of the infinite sites mutation model we have assumed. The RFLP haplotype frequency data in table 1 of Boom, Boulding, and Beckenbach (l994) and the lists of fragment sizes in their table 2 were used to estimate that the data are the result of $S = 50$ mutations and that for $\eta_1 = 31$ of these the mutant type is found in only a single copy in the sample. Equating these to their expectations (1) and (2) and solving numerically, we obtain point estimates of $\theta = 5.8$ and $x = 10.8$, and thus $N_e = n/x = 13$. We also used our simulation

program to estimate the likelihood surface for these data using Monte Carlo integration (over genealogies). A grid of paired (N_e,θ) values was examined, and for each of these we computed the log-likelihood of the data by averaging its value over 50,000 replicate genealogies. The likelihood for each simulated genealogy is easily computed by recording its values of τ and τ_1 and using the fact that, given these values, $S - \eta_1$ and η_1 are independent Poisson random variables with parameters $\theta(\tau - \tau_1)$ and $\theta\tau_1$, respectively. Figure 3 shows the result. Note that figure 3, resealed, could be interpreted as a posterior distribution of N_e and θ under a Bayesian approach.

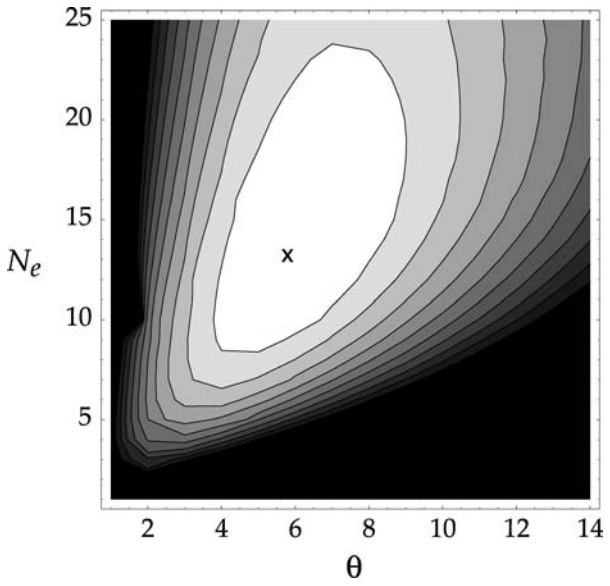

FIG. 3.—Contour plot of the likelihood surface for the data (n = 141, S = 50, η_1 = 31) of Boom, Boulding, and Beckenbach (l994). Contours are draw every three log-likelihood units from the maximum which is marked with an x.

Discussion

The genealogies of large samples, where n is on the order of or even greater than the effective size of the population, differ from those of smaller samples because multiple coalescent events occur in single generations. Most of these occur in the first few generations looking back. Multiple coalescent events are, in fact, the sole cause of the differences between the patterns we have described and the predictions of the coalescent. The two main effects of large sample size, when only single-site patterns are considered, are that singleton polymorphisms are relatively more abundant in large samples and that there is a mode in the site-frequency distribution for mutant counts around n/N_e. These effects become quite pronounced when $n > N_e$, and are surprisingly mild when $n \leqslant N_e$. Mutations which have occurred in the immediately previous generation are the source of the excess singletons, and the expected number of these is $\theta x/2$, or nu. In the standard coalescent, this number is negligible in comparison to the expected number of singletons (θ) and the expected number of segregating sites ($\theta \sum_{i=1}^{n-1} 1/i$), but for large samples these recent mutations can account for the bulk of polymorphisms in the sample.

The mode in the site-frequency distribution is similar to the pattern recently described for samples from a single local population in a metapopulation subject to local extinction and recolonization (Wakeley and Aliacar 2001). In both cases, this is the result of multiple coalescent events in a single generation. The mutant count at this mode is equal to the expected number of descendents per ancestral lineage when ancestors are chosen by randomly throwing n balls into N_e boxes (in the metapopulation case, the propagule size k replaces N_e). This highlights a potential problem with the coalescent approach to studying population bottlenecks, in which it is assumed that the bottleneck merely rescales coalescent times. More generally, this could be a problem whenever populations change in size over time. When the sample size or the number of ancestral lineages at the time of the bottleneck is not smaller than the effective size of the bottleneck population it will be important to allow for simulaneous coalescent events. A rigorous but abstract theory of coalescents with such multiple mergers is being developed (Pitman 1999; Sagitov 1999; Schweisnberg 2000), as well as general theory of such processes both forward and backward in time (Donnelly and Kurtz 1999), but so far without attention to making predictions about measures of genetic variation.

The application of our model and results to the Pacific oyster mtDNA data of Boom, Boulding, and Beckenbach (l994) shows that an excess of singleton polymorphisms can lead to estimates of the effective size of the population that are smaller than the sample size. An interesting aspect of the present work is that, given appropriate data (*i.e.*, where $n > N_e$), it will be possible to estimate N_e and u separately, in contrast to the case of small samples, in which only the composite parameter θ can be estimated. However, in this case, the parameter estimates themselves indicate that $n > N_e$ is not the (only) explanation for the observed pattern. Namely, we estimate u to be equal to $\theta/(2N_e) = 5.8/26 = 0.2$ per generation. Although it is difficult to say how many sites in the mtDNA were effectively surveyed in the restriction digests of Boom, Boulding, and Beckenbach (1994), this value of u is unrealistically large. Some other phenomenon, such as recent population growth or natural selection, must be the source of (at least some of) the excess singletons in this sample.

We did not present an analysis of the obvious data for this: the 9388 sequences of hypervariable region 1 of human mitochondrial DNA mentioned in the Introduction. A preliminary analysis of these data revealed that, in contrast to the oyster data, they showed a deficiency of singletons rather than an excess. Still, it seems likely that $n > N_e$ for these human mtDNA data. Assuming this is so, one possible explanation for the absence of the predicted pattern is that hypervariable region 1 of human mitochondrial DNA does not conform to the infinite sites model (Wakeley 1993). If the nu mutations expected in the immediately previous generation were to occur mostly at some small number of hypermutable sites, then those sites would have mutant counts greater than 1.

As molecular technologies develop even further to

allow easy measurement of genetic variation, it will become even more important to model large-sample genealogies and to develop efficient methods of analysis. Although the simplicity of the standard coalescent will be lost, the work presented here shows that a continuous approximation for $x = n/N_e$, first used by Fisher (1930) then later by Watterson (1975), can give useful analytical results.

Acknowledgments

We thank Andy Beckenbach for alerting us to his very relevant work and for aid in interpreting the oyster RFLP data. We also thank Matt Hare and Simon Tavaré for helpful discussions. Two anonymous reviewers gave helpful comments on the manuscript. This work was supported by grants DEB-9815367 and DEB-0133760 from the National Science Foundation to J.W.

Literature Cited

Beckenbach, A. T. 1994. Mitochondrial haplotype frequencies in oysters: neutral alternatives to selection models. Pp. 188–198 *in* B. Golding, ed. Non-neutral evolution. Chapman & Hall, New York.

Boom, J. D. G., E. G. Boulding, and A. T. Beckenbach. 1994. Mitochondrial DNA variation in introduced populations of Pacific oyster, *Crassostrea gigas*, in British Columbia. Can. J. Fish. Aquat. Sci. **51**:1608–1614.

Donnelly, P., and T. G. Kurtz. 1999. Particle representations for measure-valued population models. Ann. Prob. **27**:166–205.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. Theor. Pop. Biol. **3**:87–112.

Feller, W. 1968. An introduction to probability theory and its applications, Vol. 1. 3rd edition. John Wiley & Sons, New York.

Fisher, R. A. 1930. The distribution of gene ratios for rare mutations. Proc. R. Soc. Edinb. **50**:205–220.

Fu, X.-Y. 1995. Statistical properties of segregating sites. Theor. Pop. Biol. **48**:172–197.

Fu, X.-Y., and W.-H. Li, 1993. Statistical tests of neutrality of mutations. Genetics **133**:693–709.

Harris, H. 1966. Enzyme polymorphism in man. Proc. R. Soc. Lond. Ser. B **164**:298–310.

Hartl, D. L., and A. G. Clark. 1997. Principles of population genetics. 3rd edition. Sinauer Associates, Sunderland, Mass.

Hawks, J., K. Hunley, S.-H. Lee, and M. Wolpoff. 2000. Population bottlenecks and Pleistocene human evolution. Mol. Biol. Evol. **17**:2–22.

Hedrick, P. W. 2000. Genetics of populations. Jones and Barlett, Sudbury, Mass.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution **37**:203–217.

Karlin, S., and J. McGregor. 1972. Addendum to paper of W. Ewens. Theor. Pop. Biol. **3**:113–116.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to the steady flux of mutations. Genetics **61**:893–903.

———. 1983. The neutral theory of molecular evolution, Cambridge University Press, Cambridge.

Kingman, J. F. C. 1982a. The coalescent. Stochastic Process. Appl. **13**:235–248.

———. 1982b. On the genealogy of large populations. J. Appl. Prob. **19A**:27–43.

Lewontin, R. C., and J. L. Hubby. 1966. A molecular aproach to the study of genic diversity in natural populations II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics **54**: 595–609.

Nielsen, R. 2001. Statistical tests of neutrality in the age of genomics. Heredity **86**:641–647.

Pitman, J. 1999. Coalescents with multiple collisions. Ann. Prob. **27**:1870–1902.

Sagitov, S. 1999. The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Prob. **36**:1116–1125.

Schweisnberg, J. 2000. Coalescents with simultaneous multiple collisions. Electron. J. Prob. **5**:1–50.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro, 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141**:413–429.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

———. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

Takahata, N. 1995. A genetic perspective on the origin and history of humans. Annu. Rev. Ecol. Syst. **26**:343–372.

Wakeley, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37**:613–623.

Wakeley, J., and N. Aliacar. 2001. Gene genealogies in a metapopulation. Genetics **159**:893–905. Corrigendum (Fig. 2): **160**:1263–1264.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7**:256–276.

Wright, S. 1943. Isolation by distance. Genetics **28**:114–138.

Brian Golding, Associate Editor

Accepted October 9, 2002