The Sampling Distribution of Disease-Associated Alleles

Montgomery Slatkin and Bruce Rannala

Department of Integrative Biology, University of California, Berkeley, California 94720-3140

Manuscript received May 22, 1997
Accepted for publication August 15, 1997

ABSTRACT

A theory is developed that provides the sampling distribution of low frequency alleles at a single locus under the assumption that each allele is the result of a unique mutation. The numbers of copies of each allele is assumed to follow a linear birth-death process with sampling. If the population is of constant size, standard results from theory of birth-death processes show that the distribution of numbers of copies of each allele is logarithmic and that the joint distribution of numbers of copies of k alleles found in a sample of size n follows the Ewens sampling distribution. If the population from which the sample was obtained was increasing in size, if there are different selective classes of alleles, or if there are differences in penetrance among alleles, the Ewens distribution no longer applies. Likelihood functions for a given set of observations are obtained under different alternative hypotheses. These results are applied to published data from the *BRCA1* locus (associated with early onset breast cancer) and the factor VIII locus (associated with hemophilia A) in humans. In both cases, the sampling distribution of alleles allows rejection of the null hypothesis, but relatively small deviations from the null model can account for the data. In particular, roughly the same population growth rate appears consistent with both data sets.

MANY genetic diseases and disorders are caused by numerous different mutations. For example, cystic fibrosis is known to be caused by at least 500 different mutations at the *CFTR* locus (CF Genetic Analysis Consortium 1994). The number of disease-associated alleles and the number of copies of each allele in a sample can provide information about the population genetic forces affecting those alleles and about the population from which the sample was taken. To use that information, a theory predicting the sampling distribution under different assumptions is needed. In this article, we find the sampling distribution of such alleles under the assumption that they are rare. We show that if alleles are equivalent and the population from which the sample is drawn is of constant size, the sampling distribution is given by the Ewens sampling formula (ESF) (EWENS 1972). If one of those assumptions is relaxed, the ESF no longer applies. We will consider three ways in which deviations from the ESF can result: population growth, differential selection, and differential penetrance. We illustrate the use of our methods by applying them to samples of the *BRCA1* and the factor VIII loci in humans.

THEORETICAL ANALYSIS

General assumptions: Throughout, we will be concerned with a single locus in a diploid species in a population of size $N(t)$ at time t. There are then $2N(t)$ copies of the locus and each has a probability μ of mutating to an allele that is associated with an identifiable disorder or condition. We assume that each mutant is different from any other in the population (the "infinite alleles" model of mutation). Once an allele appears, it remains sufficiently rare that its numbers are governed by a linear birth-death process. That is, we assume that each copy has a probability Bdt of duplicating itself in a time interval of length dt and a probability Ddt of dying in the same time interval. In the terminology of birth-death processes, B is the instantaneous birth rate and D is the instantaneous death rate. In a previous study (SLATKIN and RANNALA 1997), we have shown that the linear birth-death process provides a good approximation for the dynamics of a rare allele in a large population. In the theoretical analysis, we will not need to require that $D > B$, but that assumption is necessary for the model to be realistic because mutant alleles would not remain rare on average if $B > D$.

KENDALL (1948) developed the theory of linear birth-death processes that we use here. He showed that the distribution of numbers of descendants at time T of a single copy that arises at time t is $P_i(T - t)$ where

$$P_0(x) = \frac{D(e^{(B-D)x} - 1)}{Be^{(B-D)x} - D}$$

$$P_i(x) = (1 - P_0(x))(1 - u)u^i \qquad (1)$$

with

$$u = u(x) = \frac{B(e^{(B-D)x} - 1)}{Be^{(B-D)x} - D} \qquad (2)$$

Corresponding author: Montgomery Slatkin, Department of Integrative Biology, VLSB 3060, University of California, Berkeley, California 94720-3140. E-mail: slatkin@socrates.berkeley.edu

That is, there is a probability $P_0(T-t)$ of leaving no descendant copies at T and, given that at least one copy remains, the distribution is geometric with parameter $u(T-t)$.

We need to extend KENDALL's (1948) theory to account for the fact that only a fraction of a population is represented in a sample. To model sampling, we can assume that at T, each copy has a probability f of being sampled, which implies that if there are i copies present, the number in the sample will be approximately binomially distributed with parameters f and i (assuming sampling with replacement). Sampling can be introduced by computing the probability generating function (pgf) for (1), $G(s) = \sum_{i=0}^{\infty} s^i P_i$, and replacing s by $1 - f + fs$. That leads to the result that, with sampling, the distribution of the number of copies is given by

$$P_0(x) = \frac{fD - [B(1-f) - D]e^{(B-D)x}}{fB + [B(1-f) - D]e^{(B-D)x}}$$

$$P_i(x) = (1 - P_0(x))(1-u)u^i \qquad (3)$$

where now

$$u = u(x) = \frac{fB(e^{(B-D)x} - 1)}{fBe^{(B-D)x} + B(1-f) - D}. \qquad (4)$$

The distribution of the number of copies is still a modified geometric but with different parameters. This result was derived by NEE *et al.* (1994) using another method.

We assume that the rate of mutation to alleles in the mutant class is μ per gamete per generation. Mutations to alleles that are not associated with a disorder or disease are ignored. The net flux to the mutant class at time t is $m(t) = 2\mu N(t)$. In making this assumption, we are allowing mutant alleles to mutate to another disease-associated allele. Such mutations would contribute very slightly to the death rate, D. We assume that the population of interest was founded at time $t = 0$, at which there were no mutant alleles. The goal is to find the sampling distribution of mutant alleles at some time T generations later.

Population of constant size: If we assume N does not change between 0 and T, results are easily obtained from KENDALL's (1948) analysis. He showed that if a mutation is equally likely to occur at any time during the interval, the probability that it will leave no descendants is

$$Q_0 = \frac{1}{T}\int_0^T P_0(T-t)\,dt = 1 + \frac{1}{BT}\ln(1-U), \qquad (5)$$

where $\ln(.)$ denotes the natural logarithm and $U = u(T)$. The probability of $i > 0$ descendants at T is a logarithmic distribution

$$Q_i = \frac{1}{T}\int_0^T P_i(x)\,dx = \frac{U^i}{BTi}. \qquad (6)$$

Sampling does not alter these results provided that (4) is used instead of (2) to define U.

With a constant flux of mutants at rate m in the interval $(0, T)$, the probability that there are j mutations during that interval is a Poisson distribution with mean mT. Each of those mutants has a probability $1 - Q_0$ of surviving to T, so the probability that there are k mutant lineages at T is Poisson distributed

$$\Pr(k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad (7)$$

with $\lambda = mT(1 - Q_0)$. The finding of KARLIN and MCGREGOR (1967) that k has an asymptotically normal distribution with mean and variance $\log(1 + \lambda t)$ when $B = D$ is a special case of this result. To find the joint distribution of numbers of copies of each allele, we follow the analysis of WATTERSON (1974). Given that k alleles are found at T, each of them is independently and identically distributed with probability $Q_i/(1 - Q_0)$. Thus

$$\Pr(i_1,\ldots,i_k|n,k) = \frac{Q_{i_1}\cdots Q_{i_k}}{(1-Q_0)^k} = \frac{U^{i_1}\cdots U^{i_k}}{(1-Q_0)^k(BT)^k(i_1\cdots i_k)}. \qquad (8)$$

The joint distribution conditioned on the total number of mutants in the sample, $n = \sum_{j=1}^{k} i_j$, is obtained from the probability generating function (pgf) of $Q_i/(1 - Q_0)$, which is

$$G(s) = E(s^i) = \frac{-1}{(1-Q_0)(BT)}\ln(1 - Us). \qquad (9)$$

Because n is the sum of k independent random variables, the pgf of n is the k-fold convolution of $G(s)$ is

$$[G(s)]^k = \frac{k!}{(1-Q_0)^k(BT)^k}\sum_{j=k}^{\infty}(-1)^{j-k}S_j^{(k)}\frac{(Us)^j}{j!}, \qquad (10)$$

where $S_j^{(k)}$ is Stirling's number of the first kind, and the equality follows from a standard result of combinatorial theory (ABROMOWITZ and STEGUN 1965, p. 824; WATTERSON 1974). The probability of n given k is given by the coefficient of s^j in (10):

$$\Pr(n|k) = \frac{U^n k!}{(1-Q_0)^k(BT)^k n!}(-1)^{n-k}S_n^{(k)} \qquad (11)$$

and hence

$$\Pr(i_1,\ldots,i_k|n,k) = \frac{\Pr(i_1,\ldots,i_k|n,k)}{\Pr(n|k)} = \frac{n!}{k!(-1)^{(n-k)}S_n^{(k)}(i_1\cdots i_k)}. \qquad (12)$$

The expression on the right hand side is the sampling

distribution of neutral alleles in a finite population of constant size (EWENS 1972; JOHNSON *et al.* 1997) and is called the Ewens sampling formula (ESF). In this case, the ESF applies not to all the alleles at a locus but only to those alleles associated with a particular genetic disorder or disease. BERTRANPETIT and CALAFELL (1995) assumed that the ESF applies to this situation but did not provide a derivation. Note that we have not assumed that an equilibrium has been reached. The ESF applies regardless of the value of T. KANNALA (1996) noted this property of the sampling distribution under a linear birth-death model. If T becomes large and $D > B$, then U will approach its limiting value, $fB/[D - (1 - f)B]$, but if $D \leq B$, then U will approach 1 and no equilibrium will be reached. As we have said, the case with $D > B$ is the most reasonable biologically and in practice if $T \gg (D - B)$, U will be sufficiently close to its limiting value that further increase in T will not affect the results.

We have found that the unconditional distribution of k is Poisson, but is easy to show that the distribution conditional on the sample size, n, is the same as found by EWENS (1972) for the infinite neutral alleles model. To see this, we use Bayes theorem

$$\Pr(k|n) = \frac{\Pr(n|k)\Pr(k)}{\Pr(n)}. \quad (13)$$

We have already found the two terms in the numerator, Equations 7 and 11, and we can compute the denominator using

$$\Pr(n) = \sum_{k=1}^{n} \Pr(n|k)\Pr(k) = (-1)^n e^{-\lambda} \frac{U^n}{n!} \sum_{k=1}^{n} (-x)^k S_n^{(k)}, \quad (14)$$

where $x = 2N\mu/B$. The right hand side of (14) is closely related to the generating function for Stirlings numbers of the first kind (ABROMOWITZ and STEGUN 1965, p. 824). Substituting (7), (11), and (14) and simplifying, we obtain

$$\Pr(k|n) = \frac{(-1)^{n-k} x^k S_n^{(k)}}{x(x+1)\cdots(x+n-1)}, \quad (15)$$

which is the same as found by EWENS (1972) for the model of infinite neutral alleles. In this case x plays the role of $\theta = 4N\mu$ in the infinite alleles model.

For later purposes, it will be useful to summarize the derivation of the ESF in slightly different way. Under the assumption of independence of different alleles, the joint distribution of the numbers of copies conditional on k is proportional to the product of the Q_i

$$\Pr(i_1,\ldots,i_k|n,k) = CQ_{i_1}\cdots Q_{i_k}, \quad (16)$$

regardless of the functional form of Q_i, where C is the normalization constant. The assumption of constant population size and the equivalence of the alleles lead to a logarithmic distribution of i but other assumptions presented below lead to other distributions. To condition on the total sample size n, we have to sum (16) over all configurations of the data $(i_1,\ldots,i_k)$ constrained so that they sum to n:

$$\Pr(i_1,\ldots,i_k|n,k) = \frac{Q_{i_1}\cdots Q_{i_k}}{\sum Q_{i_1}\cdots Q_{i_k}}, \quad (17)$$

where the sum in the denominator is over all suitable configurations.

In general the denominator cannot be expressed analytically and instead must be evaluated numerically. Although there are large number of suitable configurations, in fact $\binom{n}{k}$ of them, the sum can be evaluated even for relatively large values of n and k. The trick is to note that it can be expressed as the solution to a recursion equation

$$Y(n,k) = \sum_{i=1}^{n-k+1} Q_i Y(n-i,k-1), \quad (18)$$

where $Y(n,k)$ is the sum over all configurations of k alleles that sum to n copies and $Y(n,2) = \sum_{j=1}^{n-1} Q_j Q_{n-j}$. Repeated use of (18) will allow the calculation of $Y(n,k)$ in roughly kn^2 steps.

We will proceed by first discussing how the ESF is applied and then consider the effect of various modifications of the assumptions that led to the ESF.

Application of the Ewens sampling formula to *BRCA1*: The ESF is an especially useful result because it shows that conditional on n and k, the joint distribution of the numbers of alleles in the sample is independent of the other parameters of the model, particularly μ, the mutation rate to disease-associated alleles, N the population size, and B and D, the birth and death rates of the mutant allelic class. In fact, EWEN's (1972) theory shows that k, the number of alleles observed in a sample, is a sufficient statistic for estimating the composite parameter $x = 2N\mu/B$.

Given that x is estimated by k, the ESF provides a test of fit of the data to the model. As in the application of the ESF to the neutral mutation theory, several tests are possible. WATTERSON (1978) proposed using the computed homozygosity, $F = \sum_{j=1}^{k} (i_j/n)^2$ as a test statistic, and SLATKIN (1994, 1996) proposed an exact test using, in effect, the ESF itself as a test statistic. To illustrate, we can apply these tests to data from the *BRCA1* locus associated with early onset breast cancer. We used data from Table 3 of SHATTUCK-ELDENS *et al.* (1995), which listed 37 alleles distinguished by sequence difference in a total sample of 63 individuals with early onset breast cancer attributable to mutations at this locus. Thus, $n = 63$ and $k = 37$ in our notation. The configuration of the data is $(2 \times 8, 5, 8 \times 2, 26 \times 1)$, which means that there eight copies of each of two alleles (185 del AG and 5382 ins C), five copies of another

(4184 del 4), eight alleles with two copies each and 26 alleles with one copy each. Using the program described by SLATKIN (1996), Watterson's homozygosity test rejects the null hypothesis (that the ESF applies) with a tail probability of $P_H = 0.969898$, and the exact test rejects the null hypothesis with a tail probability of $P_E = 0.973983$. Thus, even with a very small sample size, we can conclude that these data are not consistent with the null hypothesis that the alleles are governed by the same mutation and selection processes in a population of constant size.

MODIFICATIONS TO THE BASIC MODEL

The assumptions made in deriving Equation 12 (the ESF) can all be relaxed. Doing so will both inform us about the robustness of the theory and guide us in searching for alternative models when the ESF can be rejected. We will begin with modifications to the basic theory that do not result in a deviation from the ESF and then to those modifications that require the development of a new sampling theory.

Different rates of mutation to different alleles: The null model assumes that all mutations occur at the same rate μ, which determines $\lambda = 2\mu NT(1 - Q_0)$. That assumption is not biologically reasonable because, for example, insertions and deletions are probably less likely than substitutions, and transversion substitutions are much less likely than transition substitutions. Allowing for differences in mutation rates does not change the null model other than in requiring a change in the definition of λ. Assume that there are several classes of mutations to disease-associated alleles and that the rate to class j is μ_j. The distribution of the number of alleles in class j at T is Poisson with parameter $\lambda_j = 2N\mu_j(1 - Q_0)$. The total number of alleles at T will be the sum over the different mutational classes. The sum of Poisson distributed random variables is Poisson distributed with a parameter equal to the sum of the parameters, so the distribution of k is unchanged if we interpret μ as the sum of the mutation rates of the different mutational classes. Even with variation in mutation rate among classes, the distribution of the numbers of each allele is logarithmic with the same parameter, U, and hence the ESF still applies. Thus, no modification of the null model is needed to account for variation in mutation rates and any deviations from the ESF cannot be attributable to such variation.

Differential selection: It is possible that mutant alleles are not equivalent in their phenotypic effects, even though they are all identified as producing the same clinical symptoms. An alternative model that allows for this possibility is one in which there are different classes of alleles that differ in their birth and death rates, with values B_j and D_j for class j.

One possibility is that there is an *a priori* way to distinguish among classes of alleles. For example, in the *BRCA1* data set, there are frameshift, missense, nonsense, splice site, and inferred regulatory mutations, and each of these classes might differ in effect. If alleles within each class are homogeneous, then each class separately should fit the ESF. In the *BRCA1* data set, only the frameshift mutations have a large enough sample size to allow a test of the ESF. The configuration of the frameshift mutations alone (2×8, 5, 5×2, 14×1; $n = 45$, $k = 22$) does not lead to rejection of the null hypothesis ($P_H = 0.898631$, $P_E = 926914$), although the sample size is small enough that the power of this test is in doubt. Nevertheless, this result does provide a working hypothesis that can be applied to larger samples as they become available.

Another possibility is that different classes of mutations cannot be distinguished. If we can model the variation in the birth and death rates among alleles, we can derive the Q_i for different parameter values and hence find the likelihood of the data using Equation 17. To illustrate, we assume two classes of mutations. The mutation rate is μ and when a mutation occurs the probability that it is in class 1 is α and the probability that it is in class 2 is $1 - \alpha$. The distribution of the numbers of copies of an allele is a mixture of two logarithmic distributions

$$Q_i = C\left(\frac{\alpha U_1^i}{i} + \frac{(1 - \alpha)\, U_2^i}{i}\right), \qquad (19)$$

where U_1 and U_2 are the based on the birth and death rates in each class and C is a normalization constant. We can then use (17) to find the sampling distribution.

Figure 1 shows some of the resulting likelihood curves for the *BRCA1* data set described above. In this analysis, alleles in class 1 were neutral ($D = B$) and alleles in class 2 were deleterious with selection coefficient s ($D = B + s$). We assumed $B = {}^1/_2$ per generation, which scales the birth-death process in a way compatible with the Wright-Fisher model (SLATKIN and RANNALA 1997). In this figure and others, we plotted the ratio of the likelihood under different parameter values (L) to the likelihood of the data under the null model (L_0), obtained from the ESF. The fraction f of the population represented in this sample is unknown but is probably small. Figure 1 shows that relatively weak selection against the deleterious class and a relatively small value of α maximized the likelihood of the data under this model. Figure 1 and later figures show the relatively likelihood under the modified model compared to the ESF. From these figures, maximum likelihood estimates of parameters and their support intervals can be obtained. Too much is unknown to take such estimates literally but they do indicate the kinds of processes that could lead to deviations from the null model.

Differential penetrance: Alleles may also differ in their penetrance and hence in their chance of being included in a sample. We can model this situation by

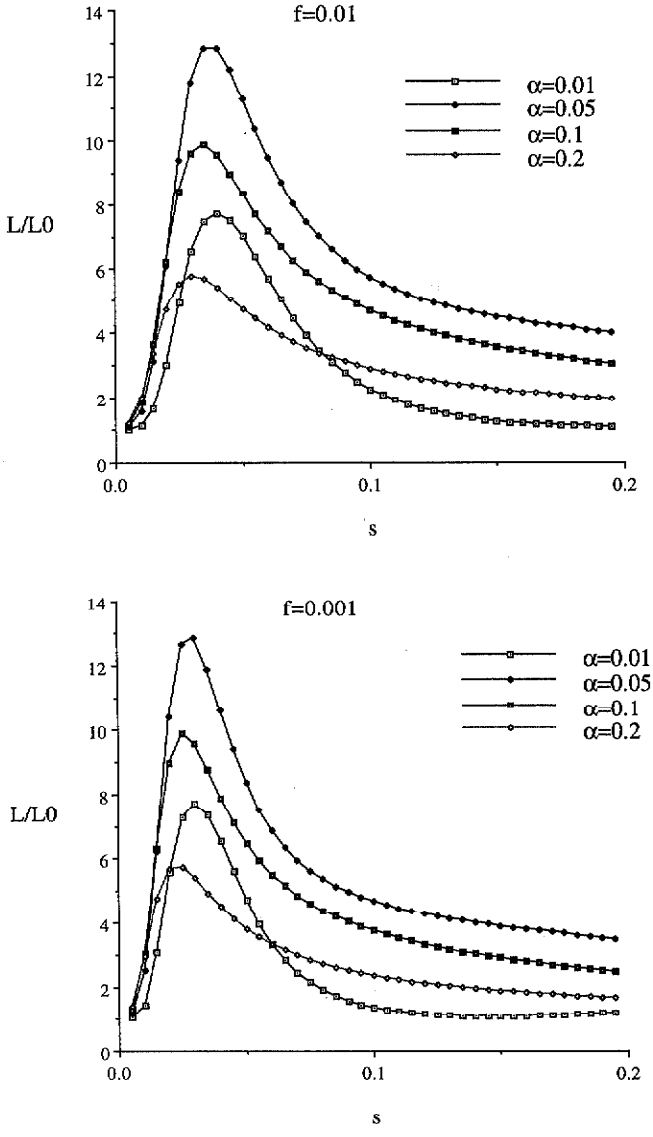

FIGURE 1.—The ratio of the likelihood (L) of the configuration of the *BRCA1* data under the heterogeneous selection model described in the text to the likelihood (L_0) under the null model. The value of L_0 is given by the Ewens sampling formula, Equation 12 in the text. The value of L was obtained for the model with two classes of alleles, one neutral ($D = B$) and the other deleterious ($D = B + s$). α is the probability that a new mutant is neutral and f is the probability that each copy of the locus will be included in the sample. In all cases $T = 100$.

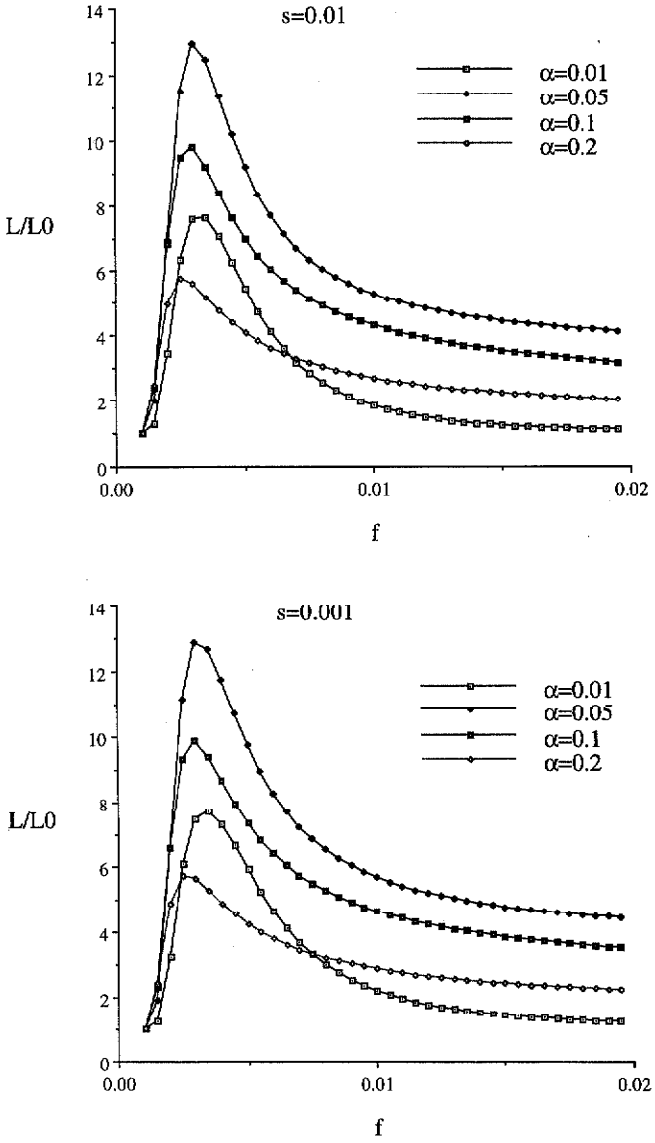

FIGURE 2.—The ratio of the likelihood (L) of the configuration of the *BRCA1* data under the heterogeneous penetrance model described in the text to the likelihood (L_0) under the null model. As in Figure 1, the value of L_0 is given by the Ewens sampling formula. The value of L was obtained for the model with two classes of alleles, one neutral ($D = B + s$, $B = 1/2$). α is the probability that a new mutant has a sampling fraction f and $1 - \alpha$ is the probability that a new mutant has a sampling fraction 0.001. In all cases $T = 100$.

allowing for heterogeneity in f, the sampling fraction. The problem is exactly the same as that of heterogeneity in selection. If different classes can be distinguished in advance, then the ESF applies to the numbers in each class separately. If, instead, the different classes cannot be distinguished, then the distribution of the number of copies of a mutant is a mixture of the distributions for each class. As in the previous discussion, we illustrate these results by assuming two classes of mutants, one with a sampling fraction f_1 and the other with a sampling fraction f_2. The discussion following Equation 19 applies, with the only difference being in the definitions of U_1 and U_2, which now reflect differences in f rather than in B and D. We then proceed in the same way to calculate the likelihood under different combinations of parameter values. Figure 2 shows some results for the *BRCA1* data. In this case, a mutant was assumed to have a probability $1 - \alpha$ of having a value $f = 0.001$ and a probability α of having a larger sampling fraction f which was allowed to vary. We assumed $B = 1/2$ and $D = B + s$. These results, like those in Figure 1, show that heterogeneity in the probability of being included in a sample can account for these data, but in this case

roughly a fivefold difference in penetrance is necessary to fit the data.

Population growth: The preceding analysis assumes the population of nondisease alleles has been of constant size. That is not true for many human populations so we need a theory that allows for past population growth. We assume a single class of alleles but allow the flux of mutants to be at a rate $2\mu N(t)$. Now the probability that a mutant has arisen in $(0, T)$ but does not leave a descendant at T is

$$Q_0 = \frac{\int_0^T N(t) P_0(T - t)\, dt}{\int_0^T N(t)\, dt}. \quad (20)$$

Equation 5 is a special case with $N(t)$ constant. The probability that the mutant has i descendant copies is

$$Q_i = \frac{\int_0^T N(t) P_i(T - t)\, dt}{\int_0^T N(t)\, dt}. \quad (21)$$

In practice, it is impossible to evaluate the integrals in the numerator of (20) and (21) analytically for functional forms of $N(t)$ that represent reasonable assumptions about population growth. In particular, both linear and exponential functions lead to integrals that appear to be intractable. Only piecewise constant forms for $N(t)$ lead to relatively simple integrals.

Because each mutation occurs independently of others, the distribution of the number of mutant alleles at T is still Poisson with mean

$$\lambda = 2\mu(1 - Q_0) \int_0^T N(t)\, dt. \quad (22)$$

Given that k mutant alleles are present at T, their joint distribution is given by (17). The likelihood can then be calculated as a function of r and the other parameters, just as in the previous sections.

We illustrate the use of this model by assuming exponential growth at rate r from an initial size N_0: $N(t) = N_0 e^{rt}$. The integrals in the numerators of (20) and (21) were evaluated numerically. Figure 3 shows the relative likelihood under this model for $B = 1/2$, $D = B + s$, with $s = 0.01$. These results show that a relatively modest growth rate could account for the observations.

Hemophilia A: As a second application of our theory, we consider the sampling distribution of alleles at the factor VIII locus, which is associated with hemophilia A in humans. The factor VIII locus is in the most distal band of the long arm of the *X* chromosome (*Xq28*). It was cloned and sequenced by GITSCHIER *et al.* (1984). The Haemostasis Research Group at the MRC Clinical Sciences Centre in London, England, maintains a web site at the URL http://146.179.66.63/usr/WWW/WebPages/main.dir/main.htm that provides an up-to-

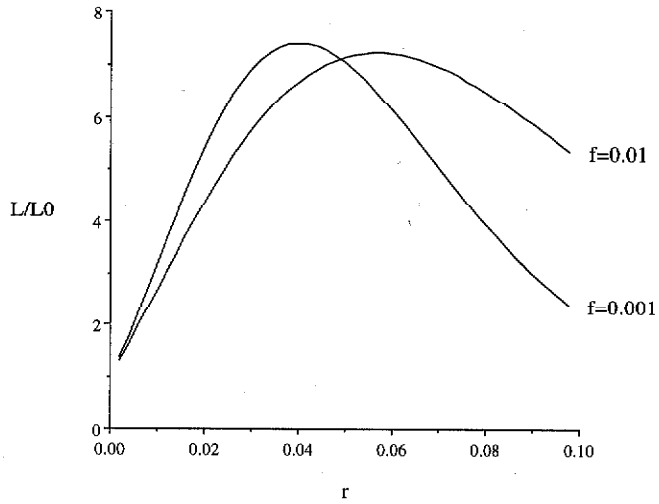

FIGURE 3.—The ratio of the likelihood (L) of the configuration of the *BRCA1* data under the population growth model described in the text to the likelihood (L_0) under the null model. As in Figure 1, the value of L_0 is given by the Ewens sampling formula. The value of L was obtained for the model in which the population was assumed to grow exponentially at rate r, $f = 0.001$, $s = 0.01$, and $T = 100$.

date list of mutations at this locus. We summarized the data obtained from Table 1 of that web site on April 15, 1997. There were $k = 234$ alleles in a sample of $n = 552$ patients. The configuration of the sample is (23, 18, 2 × 17, 16, 2 × 15, 12, 3 × 11, 2 × 10, 2 × 9, 2 × 8, 3 × 7, 6, 5 × 5, 7 × 4, 10 × 3, 31 × 2, 160 × 1). Using a Monte Carlo simulation program that performs both the exact and homozygosity tests (SLATKIN 1996), the null hypothesis that the ESF applies to these data is strongly rejected ($P_E = 1.0$, $P_H = 0.999772$ based on 10^6 replicates).

As in the case of *BRCA1*, we can consider various modifications of the null model. It is particularly interesting to allow for the possibility of population growth. Figure 4 shows the likelihood ratio as a function of the population growth rate, r, for different selection intensities against mutant alleles. It is notable that for a wide range of selection intensities, the maximum likelihood estimate of r is in the range 0.04 to 0.05 per generation, which is the same range indicated by Figure 3 for *BRCA1*. As we have shown, several modifications in the assumptions of the null model can account for the sampling distribution at *BRCA1*. The same is true for the factor VIII locus. Heterogeneity in selection or penetrance produces the same kind pattern as does population growth. When considering sampling distributions for different loci, however, population growth would be expected to have the same effect on each locus. In contrast, it is unlikely that the extent of heterogeneity in selection or penetrance would be comparable across loci. Although detailed information about the populations from which patients in these two sample were drawn is unavailable, they both appear to be from populations of predominantly western European

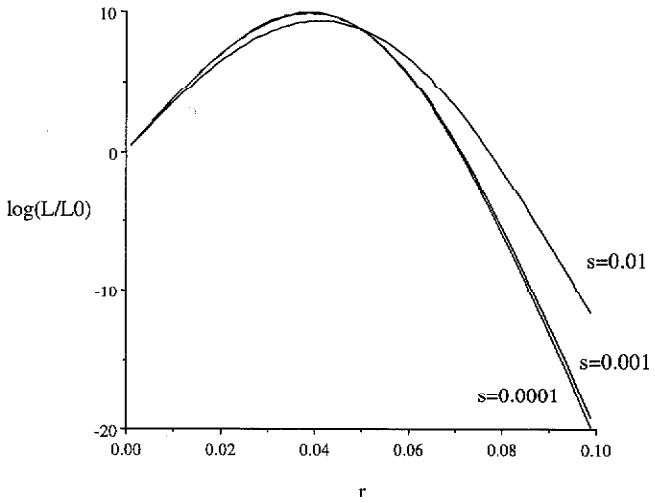

FIGURE 4.—The ratio of the likelihood of the configuration of the factor VIII locus (associated with hemophilia A) to the likelihood under the null model, the ESF. The value of L was obtained for the model in which the population was assumed to grow exponentially at rate r, $f = 0.001$, $s = 0.01$, and $T = 100$.

origin. Therefore, the similarity of the results for *BRCA1* and the factor VIII gene are at least consistent with the idea that population growth, which is thought to have occurred in the recent history of western European populations, could account for deviations from the null hypothesis at both loci. There may then be no need to invoke heterogeneity of effects of different disease-associated alleles.

DISCUSSION AND CONCLUSIONS

We have developed a theory that predicts the sampling distribution of disease-associated alleles under a null model and under several plausible alternatives. The key assumption is that the number of copies of each allele are independent. That assumption is valid for low-frequency alleles of the kind associated with most genetic diseases. Our goal is to develop a statistical framework within which different hypotheses about selection, penetrance, population growth, and other factors can be tested. In a population of constant size, alleles that are equivalent in their effects should follow the Ewens sampling formula. We have shown that deviations from that distribution can arise from several causes. For both *BRCA1* and the factor VIII locus, a realistic rate of population growth can account for deviations from the Ewens sampling formula. Part of the goal of this analysis is to encourage the collation and analysis of similar data sets. The analysis of the sampling distribution of disease-associated alleles can provide new information about factors governing the frequencies of those alleles.

We thank W. J. EWENS and S. TAVARÉ for helpful discussions and comments. This research was supported in part by a grant from the National Institutes of Health (GM-40282) to M.S. and by a Natural Sciences and Engineering Research Council of Canada postdoctoral fellowship to B.R.

LITERATURE CITED

ABRAMOWITZ, M., and I. A. STEGUN, 1956 *Handbook of Mathematical Functions.* Dover, New York.

BERTRANPETIT, J., and F. CALAFELL, 1995 Genetic and geographical variability in cystic fibrosis: evolutionary considerations, pp. 97–114 in *Variation in the Human Genome,* CIBA Foundation Symposium 197, Wiley & Co., Chichester.

Cystic Fibrosis Genetic Analysis Consortium, 1994 Population variation of common cystic fibrosis mutations. Hum. Mutat. **4:** 167–177.

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

GITSCHIER, J., W. I. WOOD, T. M. GORALKA, K. L. WION, E. Y. CHEN *et al.,* 1984 Characterization of the human factor VIII gene. Nature **312:** 326–330.

JOHNSON, N. L., S. KOTZ and N. BALAKRISHNAN, 1997 *Discrete Multivariate Distributions.* Wiley, New York.

KARLIN, S., and J. L. MCGREGOR, 1967 The number of mutant forms maintained in a population, pp. 415–438 in *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. IV, *Biology and Problems of Health,* edited by L. LECAM and J. NEYMAN, University of California Press, Berkeley.

KENDALL, D. G., 1948 On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. Biometrika **35:** 6–15.

NEE, S., R. M. MAY and P. H. HARVEY, 1994 The reconstructed evolutionary process. Phil. Trans. Roy. Soc. Lond. B **344:** 305–311.

RANNALA, B., 1996 The sampling theory of neutral alleles in an island population of fluctuating size. Theor. Popul. Biol. **50:** 91–104.

SHATTUCK-ELDENS, D., M. MCCLURE, J. SIMARD, F. LABRIE, F. S. NAROD *et al.,* 1995 A collaborative survey of 80 mutations in the *BRCA1* breast and ovarian cancer susceptibility gene. J. Am. Med. Assoc. **273:** 545–541.

SLATKIN, M., 1994 An exact test for neutrality based on the Ewens sampling distribution. Genet. Res. **64:** 71–74.

SLATKIN, M., 1996 A correction to the exact test based on the Ewens sampling distribution. Genet. Res. **68:** 259–260.

SLATKIN, M., and B. RANNALA, 1997 Estimating the age of alleles by use of intraallelic variability. Am. J. Hum. Genet. **60:** 447–458.

WATTERSON, G. A., 1974 Models for the logarithmic species abundance distributions. Theor. Popul. Biol. **6:** 639–651.

WATTERSON, G. A., 1978 The homozygosity test of neutrality. Genetics **88:** 405–417.

Communicating editor: R. R. HUDSON