DOI: 10.1534/genetics.109.107995

Closed-Form Two-Locus Sampling Distributions: Accuracy and Universality

Paul A. Jenkins* and Yun S. Song*,†,1

*Computer Science Division and †Department of Statistics, University of California, Berkeley, California 94720

Manuscript received July 30, 2009
Accepted for publication September 4, 2009

ABSTRACT

Sampling distributions play an important role in population genetics analyses, but closed-form sampling formulas are generally intractable to obtain. In the presence of recombination, there is no known closed-form sampling formula that holds for an arbitrary recombination rate. However, we recently showed that it is possible to obtain useful closed-form sampling formulas when the population-scaled recombination rate ρ is large. Specifically, in the case of the two-locus *infinite-alleles* model, we considered an asymptotic expansion of the sampling formula in inverse powers of ρ and obtained closed-form expressions for the first few terms in the expansion. In this article, we generalize this result to an arbitrary *finite-alleles* mutation model and show that, up to the first few terms in the expansion that we are able to compute analytically, the functional form of the asymptotic sampling formula is common to all mutation models. We carry out an extensive study of the accuracy of the asymptotic formula for the two-locus parent-independent mutation model and discuss in detail a concrete application in the context of the composite-likelihood method. Furthermore, using our asymptotic sampling formula, we establish a simple sufficient condition for a given two-locus sample configuration to have a *finite* maximum-likelihood estimate (MLE) of ρ. This condition is the first analytic result on the classification of the MLE of ρ and is instantaneous to check in practice, provided that one-locus probabilities are known.

FOR a given population genetics model, the probability of observing a sample of DNA sequences plays a fundamental role in various applications, including parameter estimation and ancestral inference. However, except in the one-locus case with a special model of mutation such as the infinite-alleles model or the finite-alleles parent-independent mutation (PIM) model, closed-form sampling formulas are generally unknown. In particular, when recombination is involved, obtaining an analytic formula for the sampling distribution has so far remained an intractable problem, and most research has focused on developing Monte Carlo methods, such as importance sampling (Griffiths and Marjoram 1996; Stephens and Donnelly 2000; Fearnhead and Donnelly 2001; De Iorio and Griffiths 2004a,b; Griffiths *et al.* 2008) and Markov chain Monte Carlo (Kuhner *et al.* 2000; Nielsen 2000; Wang and Rannala 2008). These Monte Carlo methods have led to useful tools for population genetics analysis, but they tend to be computationally intensive—in some cases prohibitively so—and it is difficult to provide a theoretical characterization of their accuracy. In the case of the infinite-sites model of mutation, Lyngsø *et al.* (2008) recently proposed a new approach to compute likelihoods using ideas from parsimony, but that method is not scalable.

The main mathematical framework that underlies the above-mentioned computational methods is the coalescent with recombination (Griffiths 1981; Kingman 1982a,b; Hudson 1983), which models the genealogical history of sample chromosomes. When the rate of recombination is large, the genealogies sampled by Monte Carlo methods are typically very complicated, containing many recombination events. However, in contrast to this increase in complexity in the coalescent, we in fact expect the dynamics to be easier to study for large recombination rates, since the loci under consideration would then be less dependent. It seems plausible that a stochastic process exists that is simpler than the standard coalescent with recombination, which describes the dynamics of the relevant degrees of freedom for large recombination rates. To study whether such a "dual" description of weakly interacting loci might exist, we (Jenkins and Song 2009) recently revisited the problem of obtaining a closed-form sampling formula for a model with recombination and obtained useful analytic results in the case that the population-scaled recombination rate ρ is large. More precisely, we considered the diffusion limit of the two-locus *infinite-alleles* model with population-scaled mutation rates θ_A and θ_B at the two loci. For a given sample configuration $\mathbf{n}$ (defined later in the text), we found the asymptotic expansion of the sampling probability $q(\mathbf{n} \mid \theta_A, \theta_B, \rho)$ in inverse powers of ρ,

$$q(\mathbf{n} \mid \theta_A, \theta_B, \rho) = q_0(\mathbf{n} \mid \theta_A, \theta_B) + \frac{q_1(\mathbf{n} \mid \theta_A, \theta_B)}{\rho} + \frac{q_2(\mathbf{n} \mid \theta_A, \theta_B)}{\rho^2} + O\left(\frac{1}{\rho^3}\right), \tag{1}$$

where q_0, q_1, and q_2 are independent of ρ. The zeroth-order term $q_0(\mathbf{n} \mid \theta_A, \theta_B)$ corresponds to the completely

[1] *Corresponding author:* Department of Electrical Engineering and Computer Sciences, University of California, 683 Soda Hall #1776, Berkeley, CA 94720-1776. E-mail: yss@cs.berkeley.edu

unlinked case, given by a product of marginal one-locus sampling distributions found by Ewens (1972). Our main contribution was to derive a closed-form formula for the first-order term $q_1(\mathbf{n} \mid \theta_A, \theta_B)$ and to show that the second-order term $q_2(\mathbf{n} \mid \theta_A, \theta_B)$ can be decomposed into two parts, one for which we obtained a closed-form formula and the other that satisfies a simple recursion that can be easily evaluated using dynamic programming.

The goal of this article is to generalize the above result to an arbitrary *finite-alleles* model of mutation and to study the accuracy of our asymptotic sampling formula (ASF). The main theoretical result presented here is that the functional form of our ASF is *universal* in the following sense: For all mutation models, the first two terms q_0 and q_1 in the expansion can be expressed as linear combinations of products of marginal one-locus probabilities, with coefficients depending on the sample configuration but not on mutation parameters. Hence, q_0 and q_1 for different mutation models can be obtained simply by substituting the corresponding one-locus sampling probabilities into the formulas. Whether this universality property extends to higher-order terms in the expansion is an open question. As in the infinite-alleles case, the second-order term q_2 for an arbitrary finite-alleles mutation model can be decomposed into two parts, one for which we obtain a closed-form formula and the other that satisfies a simple recursion relation. Since we are not able to obtain a closed-form solution to the recursive part, we do not know whether the entire second-order term q_2 will obey the universality property.

We carry out an extensive study of the accuracy of our ASF for the two-locus PIM model, in which case we can numerically solve for the true sampling probabilities for small sample sizes. Since any diallelic recurrent mutation model can be reduced to a PIM model (reviewed below), our results should have practical applications. We discuss a concrete application in the context of the composite-likelihood method (Hudson 2001a), which uses two-locus sampling probabilities as the basic building blocks. LDhat (McVean *et al.* 2002, 2004), a widely used software package for estimating recombination rates, is based on this composite-likelihood approach. LDhat assumes the symmetric diallelic model of mutation and relies on the importance sampling scheme developed by Fearnhead and Donnelly (2001) for the coalescent with recombination, to generate exhaustive lookup tables containing two-locus sampling probabilities for all inequivalent diallelic sample configurations. This process of generating exhaustive lookup tables is very computationally expensive. The developers of LDhat suggest using a range of integral ρ-values from 0 to 100, in which case the two-locus sampling probability for $\rho > 100$ is approximated by that at $\rho = 100$. An alternative to this truncation approach is to use our ASF for large ρ-values. In this article, we examine the effect that the truncation used by LDhat has on the two-locus sampling probability.

Another application of our work is classification of the maximum-likelihood estimate (MLE) of the recombination rate ρ. Specifically, we establish a simple sufficient condition for a given two-locus sample configuration to have a *finite* MLE of ρ. To our knowledge, this is the first analytic result on the classification of the MLE of ρ. Further, the sufficient condition is instantaneous to check, provided that one-locus probabilities are known.

We have written a C++ program, called ASF, that computes the first three terms in the expansion (1) for either the infinite-alleles model or the finite-alleles PIM model. This program is publicly available at http://www.eecs.berkeley.edu/~yss/software.html.

REVIEW OF ONE-LOCUS SAMPLING DISTRIBUTIONS

Here we briefly review two mutation models for which one-locus sampling formulas are known in closed form. As mentioned above, our two-locus sampling formula is given in terms of one-locus sampling formulas.

Notation: The sample configuration at a locus is denoted by a vector $\mathbf{n} = (n_1, \ldots, n_K)$, where n_i denotes the number of gametes with allele i at the locus. Note that K takes on slightly different meanings depending on the assumed model of mutation. In an infinite-alleles model, K refers to the number of observed alleles, whereas in a finite-alleles model it refers to the number of possible alleles in the model. We use n to denote the total sample size $\sum_{i=1}^{K} n_i$. The probability of an *unordered* sample configuration is denoted by $p(\mathbf{n})$; the dependence on parameters is not shown for ease of notation.

Let $\mathscr{A}_n$ denote an *ordered* configuration of n sequentially sampled gametes such that the corresponding unordered configuration is given by $\mathbf{n}$. By exchangeability, the probability of $\mathscr{A}_n$ does not change under an arbitrary permutation of the sampling order, so we can write this probability of an ordered sample as $q(\mathbf{n})$ without ambiguity; again, we suppress the dependence on parameters. In what follows, our theoretical results are presented in the case of ordered samples.

Throughout, we consider the diffusion limit of a neutral haploid exchangeable model of random mating with constant population size $2N$. Let u denote the probability of mutation at a locus per gamete per generation. Then, in the diffusion limit, $N \to \infty$ and $u \to 0$ with the population-scaled mutation rate $\theta = 4Nu$ held fixed. Mutation events occur according to a Poisson process with rate $\theta/2$.

Finally, given a nonnegative real number x and a positive integer n, we use $(x)_n := x(x+1)\ldots(x+n-1)$ to denote the nth ascending factorial of x.

Infinite-alleles model: Under the infinite-alleles model, any two gametes can be compared to determine whether or not they have the same allele, but it is not possible to determine how the alleles are related when they are different. Therefore, allelic label is arbitrary. Each mutation gives rise to a new allele that has never been seen before in the population. For this model, the sampling distribution of an ordered sample is given by

$$q(\mathbf{n}) = \frac{\theta^K}{(\theta)_n} \prod_{i=1}^{K} (n_i - 1)! \qquad (2)$$

(Ewens 1972). The probability $p(\mathbf{n})$ of an unordered sample configuration is related to $q(\mathbf{n})$ as

$$p(\mathbf{n}) = \frac{n!}{n_1! \ldots n_K!} \frac{1}{\alpha_1! \ldots \alpha_n!} q(\mathbf{n}),$$

where α_i denotes the number of allele types represented i times; *i.e.*, $\alpha_i = |\{k : n_k = i\}|$.

Finite-alleles PIM model: Assume that allelic changes are described by a Markov chain with transition matrix $\mathbf{P} = (P_{ij})$, with P_{ij} being the probability of transition from type i to type j. For a given sample configuration $\mathbf{n} = (n_1, \ldots, n_K)$, $q(\mathbf{n})$ satisfies the recursion

$$\begin{aligned} &n(n-1+\theta)q(\mathbf{n}) \\ &= \sum_{i=1}^{K} n_i(n_i - 1)q(\mathbf{n} - \mathbf{e}_i) + \theta \sum_{i=1}^{K}\sum_{j=1}^{K} n_j P_{ij} q(\mathbf{n} - \mathbf{e}_j + \mathbf{e}_i), \quad (3) \end{aligned}$$

with boundary conditions $q(\mathbf{e}_i) = \pi_i$, for $i = 1, \ldots, K$, where $\mathbf{e}_i$ is a K-dimensional unit vector with a 1 at the i entry and $\boldsymbol{\pi} = (\pi_i)$ is the stationary distribution corresponding to $\mathbf{P}$. See for example Lundstrom *et al.* (1992) and Griffiths and Tavaré (1994) for ways to obtain (3) [Griffiths and Tavaré's $q(\mathbf{n})$ corresponds to our $p(\mathbf{n})$]. A general solution to the above recursion is not known, but a closed-form solution can be found for the parent-independent mutation model. As its name suggests, the PIM model satisfies $P_{ij} = P_j$ for all i, and the stationary distribution of $\mathbf{P}$ is $\boldsymbol{\pi} = (P_1, \ldots, P_K)$. (In some sense the infinite-alleles model may be viewed as a PIM model. However, henceforth in this article we take "PIM" to mean a finite-alleles model whose transition matrix satisfies the above condition.) Further, the stationary sampling probability of an ordered configuration is given by

$$q(\mathbf{n}) = \frac{1}{(\theta)_n} \prod_{i=1}^{K} (\theta P_i)_{n_i} \qquad (4)$$

(Wright 1949), while the probability of an unordered sample configuration is related to $q(\mathbf{n})$ by

$$p(\mathbf{n}) = \frac{n!}{n_1! \ldots n_K!} q(\mathbf{n}).$$

The diallelic case (*i.e.*, with $K = 2$) is of particular interest. The mutation process is specified by $\theta(\mathbf{P} - \mathbf{I})$, where $\mathbf{I}$ is the identity matrix, so two models with the same $\theta(\mathbf{P} - \mathbf{I})$ are in fact equivalent. Hence, given any diallelic model with mutation parameter $\theta_{\text{Diallelic}}$ and transition matrix

$$\mathbf{P}^{\text{Diallelic}} = \begin{pmatrix} 1 - P_{12} & P_{12} \\ P_{21} & 1 - P_{21} \end{pmatrix},$$

it can be transformed into an equivalent PIM model with mutation parameter $\theta_{\text{PIM}} = \theta_{\text{Diallelic}}(P_{12} + P_{21})$ and transition matrix

$$\mathbf{P}^{\text{PIM}} = \begin{pmatrix} P_1 & 1 - P_1 \\ P_1 & 1 - P_1 \end{pmatrix},$$

where $P_1 = P_{21}/(P_{12} + P_{21})$.

The above equivalence can be checked algebraically, but the intuition behind it is perhaps more illuminating. The diagonal entries P_{ii} represent transitions from allele i to the same allele i, and they are effectively invisible in the genealogy. These "invisible" mutations provide an extra degree of freedom in the transition matrix, and when the matrix is 2 × 2 this is sufficient to force it into a parent-independent form. The appropriate choice for the rate of invisible mutations is to set $P_{11} = P_{21}$ and set $P_{22} = P_{12}$. This ensures that the entries in any column of $\mathbf{P}$ are all the same (*i.e.*, we have parent independence). Then, one normalizes each entry by $(P_{12} + P_{21})$ (also adjusting $\theta_{\text{Diallelic}}$ in a corresponding manner) so that $\mathbf{P}$ is still a stochastic matrix. This explains the expressions for θ_{PIM} and P_1 given above.

ASYMPTOTIC SAMPLING FORMULA FOR THE TWO-LOCUS MODEL

We now consider two-locus models with recombination. We denote the two loci by A and B and use θ_A and θ_B to denote the respective population-scaled mutation rates. In the case of an infinite-alleles model, we use K and L to denote the number of distinct allelic types observed at locus A and locus B, respectively, while for a finite-alleles model, K and L denote the number of possible allelic types at the two loci. The population-scaled recombination rate is denoted by $\rho = 4Nr$, where r is the probability of recombination between the two loci per gamete per generation. In the diffusion limit, $N \to \infty$ and $r \to 0$, while ρ is held fixed.

Extended sample configuration for two loci: To obtain a closed system of recursion relations satisfied by two-locus sampling probabilities, the type space must be extended to allow some gametes to be specified at only one of the two loci. The two-locus sample configuration is denoted by $\mathbf{n} = (\mathbf{a}, \mathbf{b}, \mathbf{c})$, where $\mathbf{a} = (a_1, \ldots, a_K)$ with a_i being the number of gametes with allele i at locus A and unspecified alleles at locus B, $\mathbf{b} = (b_1, \ldots, b_L)$ with b_j being the number of gametes with unspecified alleles at locus A and allele j at locus B, and $\mathbf{c} = (c_{ij})$ is a $K \times L$ matrix with c_{ij} being the multiplicity of gametes with allele i at locus A and allele j at locus B. Throughout, we use the following notation:

$$a = \sum_{i=1}^{K} a_i, \quad c_{i\cdot} = \sum_{j=1}^{L} c_{ij}, \quad c = \sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij},$$
$$b = \sum_{j=1}^{L} b_j, \quad c_{\cdot j} = \sum_{i=1}^{K} c_{ij}, \quad n = a + b + c.$$

We also use $\mathbf{c}_A = (c_{i\cdot})$ and $\mathbf{c}_B = (c_{\cdot j})$ to denote the marginal sample configurations of $\mathbf{c}$ restricted to locus A and locus B, respectively. It is important to emphasize the distinction between the vectors $\mathbf{a}$ and $\mathbf{b}$, which represent gametes with alleles specified at only one of the loci, and the vectors $\mathbf{c}_A$ and $\mathbf{c}_B$, which represent the one-locus marginal configurations of gametes with both alleles observed. Even if we observe both alleles of every gamete in a sample [in the form $(\mathbf{0}, \mathbf{0}, \mathbf{c})$], the vectors **a** and **b** are still needed when we think about the sample's ancestry. An ancestral gamete might transmit genetic material to extant gametes in the sample at only one of the two loci, whereupon we use **a** or **b** to avoid specifying the allele at the nonancestral locus.

Two-locus recursion: We use $q(\mathbf{a}, \mathbf{b}, \mathbf{c})$ to denote the sampling probability of an ordered sample with configuration $(\mathbf{a}, \mathbf{b}, \mathbf{c})$. As in the one-locus case, we suppress the dependence on parameters. Golding (1984) considered generalizing the infinite-alleles model to include recombination and constructed a recursion relation satisfied by $q(\mathbf{a}, \mathbf{b}, \mathbf{c})$ at stationarity in the diffusion limit. Ethier and Griffiths (1990) later undertook a more mathematical analysis of the model and provided several theoretical results. For a *finite-alleles* model with transition matrices $\mathbf{P}^A = (P^A_{ij})$ and $\mathbf{P}^B = (P^B_{ij})$ at the two loci, one can show that $q(\mathbf{a}, \mathbf{b}, \mathbf{c})$ satisfies the following recursion at stationarity,

$$\begin{aligned}
&[n(n-1) + \theta_A(a+c) + \theta_B(b+c) + \rho c]q(\mathbf{a},\mathbf{b},\mathbf{c})\\
&\quad = \sum_{i=1}^{K} a_i(a_i - 1 + 2c_{i\cdot})q(\mathbf{a}-\mathbf{e}_i,\mathbf{b},\mathbf{c})\\
&\qquad + \sum_{j=1}^{L} b_j(b_j - 1 + 2c_{\cdot j})q(\mathbf{a},\mathbf{b}-\mathbf{e}_j,\mathbf{c})\\
&\qquad + \sum_{i=1}^{K}\sum_{j=1}^{L}[c_{ij}(c_{ij}-1)q(\mathbf{a},\mathbf{b},\mathbf{c}-\mathbf{e}_{ij})\\
&\qquad\qquad + 2a_ib_jq(\mathbf{a}-\mathbf{e}_i,\mathbf{b}-\mathbf{e}_j,\mathbf{c}+\mathbf{e}_{ij})]\\
&\qquad + \theta_A\sum_{i=1}^{K}\left[\sum_{j=1}^{L}c_{ij}\sum_{t=1}^{K}P^A_{ti}q(\mathbf{a},\mathbf{b},\mathbf{c}-\mathbf{e}_{ij}+\mathbf{e}_{tj})\right.\\
&\qquad\qquad \left. + a_i\sum_{t=1}^{K}P^A_{ti}q(\mathbf{a}-\mathbf{e}_i+\mathbf{e}_t,\mathbf{b},\mathbf{c})\right]\\
&\qquad + \theta_B\sum_{j=1}^{L}\left[\sum_{i=1}^{K}c_{ij}\sum_{t=1}^{L}P^B_{tj}q(\mathbf{a},\mathbf{b},\mathbf{c}-\mathbf{e}_{ij}+\mathbf{e}_{it})\right.\\
&\qquad\qquad \left. + b_j\sum_{t=1}^{L}P^B_{tj}q(\mathbf{a},\mathbf{b}-\mathbf{e}_j+\mathbf{e}_t,\mathbf{c})\right]\\
&\qquad + \rho\sum_{i=1}^{K}\sum_{j=1}^{L}c_{ij}q(\mathbf{a}+\mathbf{e}_i,\mathbf{b}+\mathbf{e}_j,\mathbf{c}-\mathbf{e}_{ij}),
\end{aligned} \tag{5}$$

with boundary conditions $q(\mathbf{e}_i, \mathbf{0}, \mathbf{0}) = \pi^A_i$ for all $i \in \{1, \ldots, K\}$ and $q(\mathbf{0}, \mathbf{e}_j, \mathbf{0}) = \pi^B_j$ for all $j \in \{1, \ldots, L\}$, where $\boldsymbol{\pi}^A$ and $\boldsymbol{\pi}^B$ are stationary distributions corresponding to $\mathbf{P}^A$ and $\mathbf{P}^B$, respectively. There are several ways to derive the above recursion. One can argue directly by considering the probabilities of the most recent event going backward in time in the coalescent with recombination. Alternatively, the recursion can be obtained from the underlying diffusion process (Ethier and Griffiths 1990; Griffiths and Tavaré 1994).

Asymptotic sampling formula: In the case of the two-locus infinite-alleles model with a large ρ, we (Jenkins and Song 2009) found an asymptotic expansion of the form

$$q(\mathbf{a},\mathbf{b},\mathbf{c}) = q_0(\mathbf{a},\mathbf{b},\mathbf{c}) + \frac{q_1(\mathbf{a},\mathbf{b},\mathbf{c})}{\rho} + \frac{q_2(\mathbf{a},\mathbf{b},\mathbf{c})}{\rho^2} + O\left(\frac{1}{\rho^3}\right), \tag{6}$$

where q_0, q_1, and q_2 are independent of ρ. Below we generalize this work to an arbitrary finite-alleles mutation model. Our closed-form formulas are given in terms of the marginal one-locus sampling formulas q^A and q^B, where q^A depends only on $\theta_A, \mathbf{P}^A$, and the marginal sample configuration of **n** at locus *A*, while q^B depends only on $\theta_B, \mathbf{P}^B$, and the marginal sample configuration of **n** at locus *B*.

We now state our main results. All proofs are deferred to the appendix.

Proposition 1. *In the asymptotic expansion* (*6*) *of the two-locus sampling formula for either an infinite-alleles or an arbitrary finite-alleles model, the zeroth order term* $q_0(\mathbf{a}, \mathbf{b}, \mathbf{c})$ *is given by*

$$q_0(\mathbf{a},\mathbf{b},\mathbf{c}) = q^A(\mathbf{a}+\mathbf{c}_A)q^B(\mathbf{b}+\mathbf{c}_B). \tag{7}$$

Proposition 1 is intuitive, since the first term in the asymptotic expansion should give the exact solution for $\rho = \infty$. When the two loci are unlinked, the complete sampling probability is then simply the product of two marginal one-locus sampling probabilities, each applied to the marginal sample configuration at that locus. The proceeding terms in the asymptotic expansion are described below.

Theorem 1. *In the asymptotic expansion* (*6*) *of the two-locus sampling formula for either an infinite-alleles or an arbitrary finite-alleles model, the first-order term* $q_1(\mathbf{a}, \mathbf{b}, \mathbf{c})$ *is given by*

$$\begin{aligned}
q_1(\mathbf{a},\mathbf{b},\mathbf{c}) = \frac{1}{2}\Bigg[&c(c-1)q^A(\mathbf{a}+\mathbf{c}_A)q^B(\mathbf{b}+\mathbf{c}_B)\\
&- q^B(\mathbf{b}+\mathbf{c}_B)\sum_{i=1}^{K}c_{i\cdot}(c_{i\cdot}-1)q^A(\mathbf{a}+\mathbf{c}_A-\mathbf{e}_i)\\
&- q^A(\mathbf{a}+\mathbf{c}_A)\sum_{j=1}^{L}c_{\cdot j}(c_{\cdot j}-1)q^B(\mathbf{b}+\mathbf{c}_B-\mathbf{e}_j)\\
&+ \sum_{i=1}^{K}\sum_{j=1}^{L}c_{ij}(c_{ij}-1)q^A(\mathbf{a}+\mathbf{c}_A-\mathbf{e}_i)\\
&\times q^B(\mathbf{b}+\mathbf{c}_B-\mathbf{e}_j)\Bigg],
\end{aligned} \tag{8}$$

for arbitrary configurations **a**, **b**, **c** *of nonnegative integers.*

This theorem generalizes our previous result (Jenkins and Song 2009) to an arbitrary finite-alleles model of mutation. This result is *universal* in the following sense: The functional form of q_1 in terms of q^A and q^B is given by Equation 8, regardless of whether we are considering an infinite or a finite set of possible alleles at each locus. If the one-locus sampling probabilities are known in closed form, then (8) is instantaneous to evaluate. One simply applies the applicable one-locus solutions for q^A and q^B:

1. For an infinite-alleles model, q^A and q^B are given by the Ewens sampling Equation 2, with θ replaced with θ_A and θ_B, respectively.
2. For a finite-alleles model, q^A and q^B are the solutions to (3) with appropriate θ and P_{ij} for each locus. Closed-form expressions to these equations are not known in general, unless mutation is parent independent, in which case q^A and q^B are given by (4) with appropriate θ and P_{ij}.

Note that $q_1(\mathbf{a}, \mathbf{b}, \mathbf{0}) = 0$. However, it turns out that $q_2(\mathbf{a}, \mathbf{b}, \mathbf{0})$ does not vanish in general, and we do not have an analytic expression for it. The following theorem shows that $q_2(\mathbf{a}, \mathbf{b}, \mathbf{c})$ can be decomposed into two parts, one for which we have a closed-form expression and the other $q_2(\mathbf{a} + \mathbf{c}_A, \mathbf{b} + \mathbf{c}_B, \mathbf{0})$:

Theorem 2. *In the asymptotic expansion* (6) *of the two-locus sampling formula for an arbitrary finite-alleles model, the second-order term* $q_2(\mathbf{a}, \mathbf{b}, \mathbf{c})$ *is of the form*

$$q_2(\mathbf{a}, \mathbf{b}, \mathbf{c}) = q_2(\mathbf{a} + \mathbf{c}_A, \mathbf{b} + \mathbf{c}_B, \mathbf{0}) + \sigma(\mathbf{a}, \mathbf{b}, \mathbf{c}), \qquad (9)$$

where $\sigma(\mathbf{a}, \mathbf{b}, \mathbf{c})$ *is given by the analytic formula shown in the* appendix, *and* $q_2(\mathbf{a}, \mathbf{b}, \mathbf{0})$ *satisfies the recursion relation*

$$\begin{aligned}
&[a(a+\theta_A-1)+b(b+\theta_B-1)]q_2(\mathbf{a},\mathbf{b},\mathbf{0})\\
&\quad=\sum_{i=1}^{K} a_i(a_i-1)q_2(\mathbf{a}-\mathbf{e}_i,\mathbf{b},\mathbf{0})\\
&\qquad+\sum_{j=1}^{L} b_j(b_j-1)q_2(\mathbf{a},\mathbf{b}-\mathbf{e}_j,\mathbf{0})\\
&\qquad+\theta_A\sum_{i=1}^{K} a_i\sum_{t=1}^{K} P^A_{ti}q_2(\mathbf{a}-\mathbf{e}_i+\mathbf{e}_t,\mathbf{b},\mathbf{0})\\
&\qquad+\theta_B\sum_{j=1}^{L} b_j\sum_{t=1}^{L} P^B_{tj}q_2(\mathbf{a},\mathbf{b}-\mathbf{e}_j+\mathbf{e}_t,\mathbf{0})\\
&\qquad+4\sum_{i=1}^{K}\sum_{j=1}^{L} a_ib_j[(a-1)(b-1)q^A(\mathbf{a})q^B(\mathbf{b})\\
&\qquad\qquad-(b-1)(a_i-1)q^A(\mathbf{a}-\mathbf{e}_i)q^B(\mathbf{b})\\
&\qquad\qquad-(a-1)(b_j-1)q^A(\mathbf{a})q^B(\mathbf{b}-\mathbf{e}_j)\\
&\qquad\qquad+(a_i-1)(b_j-1)q^A(\mathbf{a}-\mathbf{e}_i)q^B(\mathbf{b}-\mathbf{e}_j)],
\end{aligned} \qquad (10)$$

with boundary conditions $q_2(\mathbf{e}_i, \mathbf{0}, \mathbf{0}) = 0$ *for* $i = 1, \ldots, K$ *and* $q_2(\mathbf{0}, \mathbf{e}_j, \mathbf{0}) = 0$ *for* $j = 1, \ldots, L$.

If one-locus sampling distributions are known in closed form, evaluating the analytic part $\sigma(\mathbf{a}, \mathbf{b}, \mathbf{c})$ in $q_2(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is instantaneous for all sample configurations. Unfortunately, we do not have a closed-form expression for $q_2(\mathbf{a}, \mathbf{b}, \mathbf{0})$, so we are not able to conclude whether the above-mentioned universality result extends to the second-order term. However, for the PIM model, one can sum over the index t in (10), so $q_2(\mathbf{a}, \mathbf{b}, \mathbf{0})$ can be evaluated efficiently using dynamic programming. In practice, this procedure is orders of magnitude faster than solving (5) directly. Furthermore, as discussed in the next section, the relative contribution of the recursive term $q_2(\mathbf{a} + \mathbf{c}_A, \mathbf{b} + \mathbf{c}_B, \mathbf{0})$ to the asymptotic sampling distribution is generally very small, so we may safely discard that term without a considerable loss of accuracy.

ACCURACY OF THE ASYMPTOTIC SAMPLING FORMULA

Our asymptotic sampling formula is applicable to a range of models of genetic data. Equipped with such a formula, it is natural to consider when it can be used to approximate the probability of a given sample configuration. As discussed before, evaluating the likelihood associated with a sample configuration is at the heart of many problems in population genetics. While obtaining such likelihoods is computationally expensive for most samples of interest, the ASF can be evaluated almost instantaneously.

Test setup: To test the accuracy of the ASF, we focus on the diallelic recurrent mutation model, which has been used to model single-nucleotide polymorphism (SNP) data. In what follows, each locus represents a single site and we make the following distinction: We use the term *diallelic* to refer to a model in which there are two possible alleles at a locus. This is in contrast to the term *dimorphic*, which refers to a sample in which precisely two alleles are observed. When we refer to a sample from a two-locus model as being dimorphic, we mean that both loci are dimorphic.

We address the following simple model of sequence evolution. Assume a diallelic, symmetric mutation model for each locus, with the same mutation rate $\theta_{\text{Diallelic}}$ and transition matrix $\mathbf{P}^{\text{Diallelic}} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. As discussed earlier, it can be cast as a PIM model with mutation rate $\theta = 2\theta_{\text{Diallelic}}$ and transition matrix $\mathbf{P} = \left(\begin{smallmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{smallmatrix}\right)$. We use this PIM model at each locus in our calculation of the ASF, so $\theta_A = \theta_B = \theta$. In what follows we condition on segregation at both sites by considering only dimorphic samples and normalize where appropriate.

Realistic choices for θ depend on the biological sample of interest. Here, we consider two values of θ. First, for many neutral regions of the human genome, typical mutation rates per base are in the range $0.001 \leq \theta \leq 0.01$ (Nachman and Crowell 2000; McVean *et al.* 2002). Below, we present results for $\theta = 0.01$; results for $\theta = 0.001$ were almost identical (not shown). When mutation rates are low, the infinite-alleles assumption is also reasonable, and we obtained very similar results using the ASF under an infinite-alleles model with $\theta = 0.01$. For brevity these results are also omitted. Second,

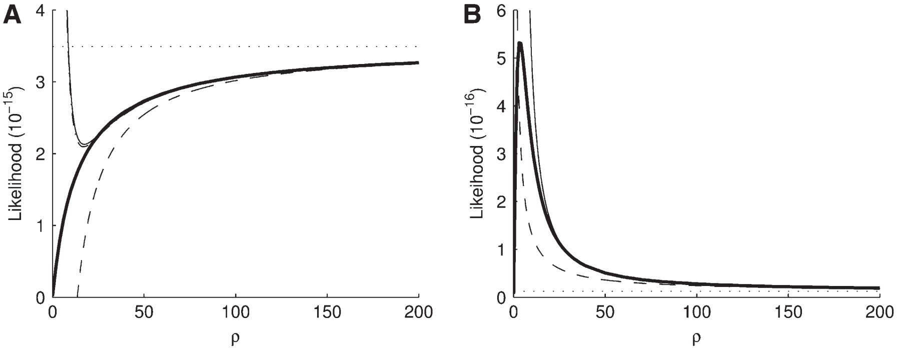

FIGURE 1.—Likelihood curves for ρ, comparing different levels of approximation. A symmetric, finite-alleles model with $\theta = 0.01$ is assumed, as described in the text. The true likelihood is shown as a thick solid line. Formulas compared are $q_0(\mathbf{0},\mathbf{0},\mathbf{c})$ (dotted line), $q_0(\mathbf{0},\mathbf{0},\mathbf{c}) + q_1(\mathbf{0},\mathbf{0},\mathbf{c})/\rho$ (dashed line), $q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ (thin solid line), and $\tilde{q}_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ (dotted-dashed line, largely concealed by the thin solid line). (a) $\mathbf{c} = \begin{pmatrix} 10 & 7 \\ 2 & 1 \end{pmatrix}$. (b) $\mathbf{c} = \begin{pmatrix} 10 & 1 \\ 2 & 7 \end{pmatrix}$.

we consider $\theta = 1.0$. This model might be assumed for the higher levels of polymorphism seen, for example, at synonymous sites in human immunodeficiency virus (McVean *et al.* 2002).

Throughout this section, we ignore missing data by considering sample configurations only of the form $\mathbf{n} = (\mathbf{0}, \mathbf{0}, \mathbf{c})$.

A first look into accuracy: We want to answer the following broad question: For a given sample size c and recombination rate ρ, how well does the ASF approximate the true likelihood? In our study, we truncate the asymptotic expansion (6) at the second order and define $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ as

$$q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c}) = q_0(\mathbf{0},\mathbf{0},\mathbf{c}) + \frac{q_1(\mathbf{0},\mathbf{0},\mathbf{c})}{\rho} + \frac{q_2(\mathbf{0},\mathbf{0},\mathbf{c})}{\rho^2}.$$

By comparing the performance of $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ against that of $q_0(\mathbf{0}, \mathbf{0}, \mathbf{c})$ and $q_0(\mathbf{0}, \mathbf{0}, \mathbf{c}) + q_1(\mathbf{0}, \mathbf{0}, \mathbf{c})/\rho$, we can gain an idea of how much improvement is provided by each additional term in $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$. As motivation, consider the likelihood curves shown in Figure 1, a and b, which compare different levels of approximation with the true likelihood for the configurations $(\mathbf{0}, \mathbf{0}, \mathbf{c})$ with $\mathbf{c} = \begin{pmatrix} 10 & 2 \\ 7 & 1 \end{pmatrix}$ and $\mathbf{c} = \begin{pmatrix} 10 & 2 \\ 1 & 7 \end{pmatrix}$, respectively, for $\theta = 0.01$. The former is illustrative of a configuration for which the MLE of ρ is infinite, and the latter is an example for which the MLE is finite. These configurations have sample sizes small enough that the true likelihoods can be solved directly from Equation 5. We can make a number of preliminary observations from Figure 1. As more terms are added to the ASF, it sustains a higher accuracy, as we would hope. In particular, $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ seems to approximate the true likelihood very accurately for $\rho \geq 20$. As illustrated in Figure 1, a similar level of accuracy is achieved by $\tilde{q}_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$, an approximation to $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ defined by

$$\tilde{q}_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c}) = q_0(\mathbf{0},\mathbf{0},\mathbf{c}) + \frac{q_1(\mathbf{0},\mathbf{0},\mathbf{c})}{\rho} + \frac{\sigma(\mathbf{0},\mathbf{0},\mathbf{c})}{\rho^2}.$$

Recall that $q_2(\mathbf{0}, \mathbf{0}, \mathbf{c})$ comprises two terms (*cf.* Theorem 2): $q_2(\mathbf{c}_A, \mathbf{c}_B, \mathbf{0})$, which must be solved by dynamic programming, and $\sigma(\mathbf{0}, \mathbf{0}, \mathbf{c})$, for which we derived an analytic expression. Calculating $q_2(\mathbf{c}_A, \mathbf{c}_B, \mathbf{0})$ imposes a small computational burden compared to solving (5) directly, but the associated running time grows with sample size. The approximation $\tilde{q}_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ simply ignores this potentially burdensome term. As we discuss later, in general this seems to lead to a negligible sacrifice in accuracy. Empirically, for most configurations we examined, $q_2(\mathbf{c}_A, \mathbf{c}_B, \mathbf{0})$ is at least an order of magnitude smaller than the analytic term $\sigma(\mathbf{0}, \mathbf{0}, \mathbf{c})$. We return to this below and carry out a more thorough study of the relative contribution of $q_2(\mathbf{c}_A, \mathbf{c}_B, \mathbf{0})$ to $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$.

Distribution of unsigned relative errors: To ensure that the above results were not unique to the particular choices of data set illustrated, we performed a more systematic study as follows. For a given $\mathbf{c}$, we measure the accuracy of $q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ by the *unsigned relative error*

$$\left| \frac{q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c}) - q(\mathbf{0},\mathbf{0},\mathbf{c})}{q(\mathbf{0},\mathbf{0},\mathbf{c})} \right| \times 100\%,$$

with corresponding definitions for the unsigned relative errors of $q_0(\mathbf{0}, \mathbf{0}, \mathbf{c})$, $q_0(\mathbf{0}, \mathbf{0}, \mathbf{c}) + q_1(\mathbf{0}, \mathbf{0}, \mathbf{c})/\rho$, and $\tilde{q}_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$. By drawing $(\mathbf{0},\mathbf{0},\mathbf{c})$ according to the true sampling probability, we may speak of the distribution of the unsigned relative error. This distribution forms the basis of our measure of accuracy, and when calculating it, we must be sure to weight each sample correctly. The correct weight for $(\mathbf{0},\mathbf{0},\mathbf{c})$ is $q(\mathbf{0},\mathbf{0},\mathbf{c})$ times a combinatorial factor capturing the number of distinct orderings of the sample; this product is denoted by $p(\mathbf{0}, \mathbf{0}, \mathbf{c})$. To evaluate $q(\mathbf{0},\mathbf{0},\mathbf{c})$, we looked at sample sizes sufficiently small so that the true likelihood could be calculated directly by solving Equation 5. We implemented a computer program for solving the recursion numerically.

We can make one further efficiency saving. Computing the true probability $q(\mathbf{0},\mathbf{0},\mathbf{c})$ for every configuration of size $c \geq 30$ is computationally demanding. However, when mutation rates are the same for the two loci, observe that, by symmetry, distinct configurations exist with the same sampling probability. For example, the samples $\mathbf{c} = \begin{pmatrix} x & y \\ y & x \end{pmatrix}$ and $\mathbf{c} = \begin{pmatrix} y & x \\ x & y \end{pmatrix}$ are distinct observations when x, y are distinct nonnegative integers, but their sampling probabilities are the same. We can therefore collapse our space of distinct sample configurations into a smaller one of equivalence classes, where two distinct samples are considered equivalent if they have the same

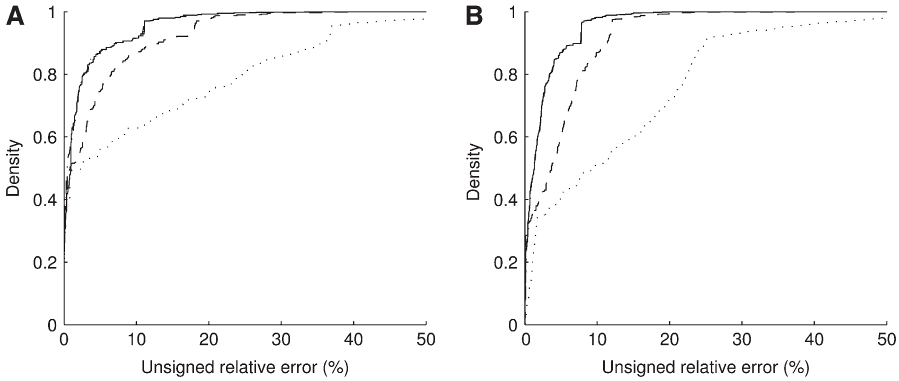

Figure 2.—Cumulative distribution of unsigned relative error of $q_0(\mathbf{0},\mathbf{0},\mathbf{c})$ (dotted line), $q_0(\mathbf{0},\mathbf{0},\mathbf{c})+q_1(\mathbf{0},\mathbf{0},\mathbf{c})/\rho$ (dashed line), $\tilde{q}_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ (dotted-dashed line, almost concealed by the solid line), and $q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ (solid line). We considered dimorphic sample configurations $(\mathbf{0},\mathbf{0},\mathbf{c})$ drawn from $p(\mathbf{0},\mathbf{0},\mathbf{c})$ in the finite-alleles model with $c = 20$ and $\rho = 50$. (a) $\theta = 0.01$. (b) $\theta = 1$.

probability. For $c = 30$, this meant that for 5456 distinct configurations only 752 probabilities required evaluation. Similar arguments can be made for the infinite-alleles model.

For the finite-alleles PIM models described above, the cumulative distribution of the unsigned relative error across all dimorphic samples of size $c = 20$ is shown in Figure 2. As is clear from the plot, the ASF offers substantial accuracy even for ρ as "low" as 50. For example, for a dimorphic sample $(\mathbf{0},\mathbf{0},\mathbf{c})$ of size 20 drawn at random from $p(\mathbf{0},\mathbf{0},\mathbf{c})$ with $\theta = 0.01$, the probability that its ASF $q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ is within 1% of the truth is 0.61, while the probability that $q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ is within 5% of the truth is 0.87. The corresponding probabilities when using $q_0(\mathbf{0},\mathbf{0},\mathbf{c})$ alone are 0.36 and 0.43, respectively. Note the spike of probability mass around zero even for the curve for $q_0(\mathbf{0},\mathbf{0},\mathbf{c})$. This suggests that assuming $\rho = \infty$ will still be accurate in about half of cases, but with substantial errors in the rest. This is not the case when employing $q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ instead.

The distributions for the two mutation parameters are rather similar, suggesting a certain degree of robustness in the ASF to θ. The case with $\theta = 1$ seems to have a lower probability of achieving a high accuracy (say within 1%) but a higher probability of achieving a moderate accuracy (within 10%).

Effect of sample size and recombination rate: To assess the sensitivity of $q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ to the sample size c and the recombination rate ρ, we repeated the above plots for varying values of these parameters; results are shown in Figures 3 and 4. As we might expect, accuracy improves with increasing ρ. For example, for $\theta = 0.01$ and $\rho = 200$, the probability of drawing a sample $(\mathbf{0},\mathbf{0},\mathbf{c})$ with $q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ within 1% of the truth reaches 1.00, compared with a probability of 0.54 when using $q_0(\mathbf{0},\mathbf{0},\mathbf{c})$ alone. On the other hand, for a fixed ρ, the accuracy of $q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ generally begins to diminish with increasing c. One possible explanation for this trend is that larger samples may have more subtle shapes to their likelihood curves, such as multiple turning points. The ASF effectively has to fit a curve with at most one turning point to approximate $q(\mathbf{0},\mathbf{0},\mathbf{c})$ with high accuracy for large ρ. This might be at the expense of accuracy for smaller ρ (see also Figure 9 later).

Signed relative error: Using the unsigned relative error as a measure of accuracy loses information on the question of whether the ASF systematically under- or overestimates the true sample probability. To investigate this, we calculated the expected signed relative error $(q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c}) - q(\mathbf{0},\mathbf{0},\mathbf{c}))/q(\mathbf{0},\mathbf{0},\mathbf{c}) \times 100\%$ across all configurations for each combination of parameters considered above. The signed relative errors of $q_0(\mathbf{0},\mathbf{0},\mathbf{c})$, $q_0(\mathbf{0},\mathbf{0},\mathbf{c})+q_1(\mathbf{0},\mathbf{0},\mathbf{c})/\rho$, and $\tilde{q}_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ are defined similarly. The results are given in Tables 1 and 2. Although most of these expectations are positive, they are generally very small, and all fall within ±1.25%. This suggests that if deviations from the true likelihood tend to err in a particular direction, then this skewness is negligible.

Monomorphic samples: One might argue that excluding nonsegregating sites from the analyses above is unwarranted, since the observation that a site is nonsegregating also contains information about the parameters

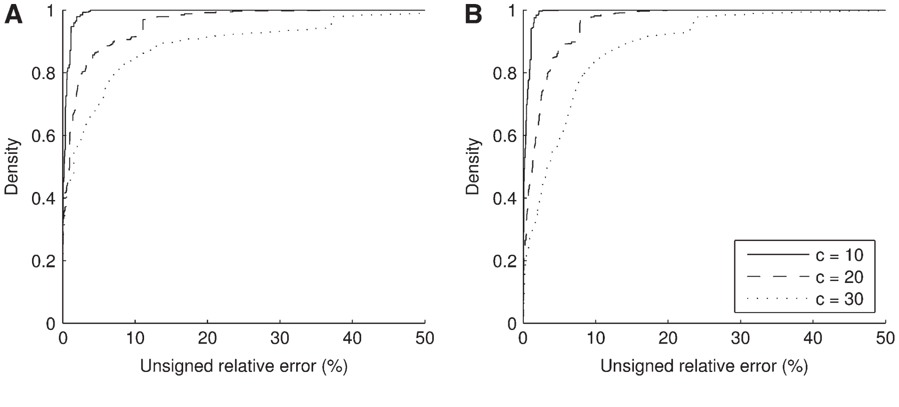

Figure 3.—The effect of varying c (fixing $\rho = 50$). Plotted is the cumulative distribution of the unsigned relative error of $q_{\mathrm{ASF}}(\mathbf{0},\mathbf{0},\mathbf{c})$ across all dimorphic sample configurations of size $c = 10$ (solid line), $c = 20$ (dashed line), and $c = 30$ (dotted line). (a) $\theta = 0.01$. (b) $\theta = 1$.

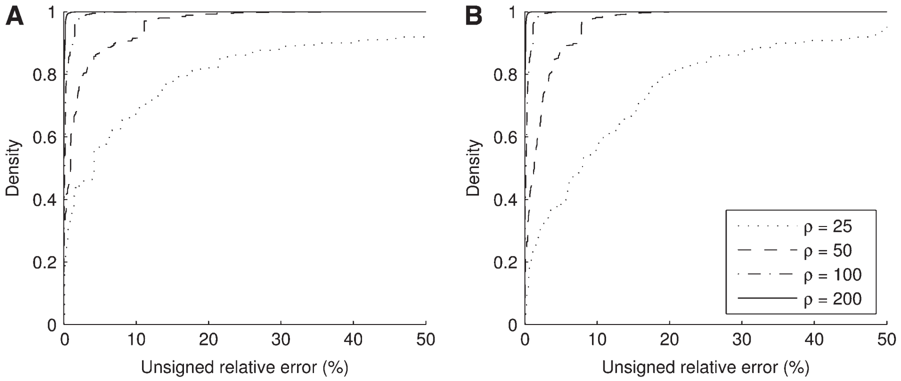

FIGURE 4.—The effect of varying ρ (fixing $c = 20$). Plotted is the cumulative distribution of the unsigned relative error of $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ across all dimorphic sample configurations of size $c = 20$ for $\rho = 200$ (solid line), $\rho = 100$ (dotted-dashed line), $\rho = 50$ (dashed line), and $\rho = 25$ (dotted line). (a) $\theta = 0.01$. (b) $\theta = 1$.

of the model. We therefore reanalyzed the distribution of the unsigned relative error across all samples of size $c = 20$ (with $\rho = 50$), rather than just dimorphic samples. Results are shown in Figure 5.

Compare Figure 5 with Figure 2 (note the different scales on the x-axes). Clearly, across all samples there is now a much greater level of accuracy, particularly when the mutation rate is small. This may be explained by the following two observations. First, the probability that one or both loci are monomorphic is relatively large: for $\theta = 0.01$ it is >0.99 and for $\theta = 1.0$ it is 0.44. Hence, these configurations will have a substantial effect on the error distribution, particularly for smaller mutation rates. Second, the ASF seems to be much more accurate precisely for those configurations with one or both loci monomorphic. This raises the question: Can we identify any further patterns for which the ASF is more accurate? When the true likelihood is unavailable, this might provide a guide as to how accurate the ASF is likely to be.

For which samples is the ASF most accurate? The figures described above suggest that for most configurations the relative error of the ASF will be very small, with a few having larger errors. It would be of interest to determine whether there is any feature of a given sample configuration that indicates how accurate the ASF will be. Figure 6 plots on a log-scale the unsigned relative error of the ASF of a sample against its sampling probability $p(\mathbf{0}, \mathbf{0}, \mathbf{c})$, across all configurations of sample size $c = 20$ (with $\theta = 0.01$ and $\theta = 1$, and $\rho = 50$). Encouragingly, there is a clear negative correlation between unsigned relative error and sampling probability (Pearson's correlation coefficient $= -0.76$ for Figure 6a and -0.61 for Figure 6b). That is, those configurations for which the ASF is most accurate are also observed most often. However, the point of this comparison is to determine which configurations have a more accurate ASF *without* having to calculate $p(\mathbf{0}, \mathbf{0}, \mathbf{c})$. We therefore identified the following categories for the samples that appear to best explain most of the variation in accuracy (the category of each sample is annotated in Figure 6):

TABLE 1

Expected signed relative error (percentage) for different approximations across all dimorphic samples with $c = 20$

	ρ			
	25		50	
	$\theta = 0.01$	$\theta = 1.0$	$\theta = 0.01$	$\theta = 1.0$
$q_0(\mathbf{0}, \mathbf{0}, \mathbf{c})$	−0.48	−0.12	−0.21	−0.06
$q_0(\mathbf{0}, \mathbf{0}, \mathbf{c}) + q_1(\mathbf{0}, \mathbf{0}, \mathbf{c})/\rho$	−0.32	−0.06	−0.12	−0.03
$\tilde{q}_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$	−0.10	0.01	−0.07	−0.01
$q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$	0.93	0.28	0.19	0.06

1. ▽, the sample is monomorphic at both loci.
2. △, the sample is monomorphic at one locus.
3. ○, for at least one allele at a locus, the sample has multiplicity one; *i.e.*, a row or column of $\mathbf{c}$ sums to one.
4. +, the sample has very low linkage disequilibrium (LD) as measured by r^2 (HUDSON 2001b), satisfying $r^2 \leq 0.02$. (This LD measure is also often denoted by Δ^2.)
5. ◇, the sample is in perfect LD; *i.e.*, it is of the form $\mathbf{c} = \begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix}$ or $\begin{pmatrix} 0 & y \\ x & 0 \end{pmatrix}$, where $x + y = c$, for $x \geq 0$ and $y \geq 0$.
6. □, the sample is in perfect LD except for one haplotype; *i.e.*, it is of the form $\mathbf{c} = \begin{pmatrix} x & 1 \\ 0 & y \end{pmatrix}$, $\begin{pmatrix} x & 0 \\ 1 & y \end{pmatrix}$, $\begin{pmatrix} 1 & y \\ x & 0 \end{pmatrix}$, or $\begin{pmatrix} 0 & y \\ x & 1 \end{pmatrix}$, where $x + y = c - 1$, for $x \geq 0$ and $y \geq 0$.
7. ●, the sample does not fall into any of the above categories.

Some configurations fell into more than one category, in which case the higher category in this list was given priority.

Figure 6a confirms that samples monomorphic at one or both loci have among the highest sampling probabilities and the highest accuracies. This is followed by another clearly identifiable cluster of samples of high accuracy, which have in common that one allele has multiplicity one. For the remaining samples, the amount of LD seems to be highly informative of the accuracy of the ASF; for example, at the upper end of the error distribution are those in perfect or almost perfect LD. We do not have an intuitive explanation for

TABLE 2

Expected signed relative error (percentage) of $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ across all dimorphic samples

	ρ							
	25		50		100		200	
c	θ = 0.01	θ = 1.0	θ = 0.01	θ = 1.0	θ = 0.01	θ = 1.0	θ = 0.01	θ = 1.0
10	0.60	0.27	0.11	0.05	0.02	0.01	0.00	0.00
20	0.93	0.28	0.19	0.06	0.03	0.01	0.01	0.00
30	1.21	0.28	0.26	0.06	0.05	0.01	0.01	0.00

this observation. Most of the remaining samples fall around a cluster of intermediate accuracy (~1%), with the samples with lowest LD at the center of this cluster. For higher mutation rates, illustrated in Figure 6b, we observed a similar, albeit less distinguished, pattern. When mutation rates are higher, the difference in probability between any two sample configurations tends to be attenuated. We also observed qualitatively similar patterns for different choices of sample size and recombination rate (not shown).

A quick approximation to the ASF: Recall that $q_2(\mathbf{a}, \mathbf{b}, \mathbf{c})$ contains a term—namely, $q_2(\mathbf{a} + \mathbf{c}_A, \mathbf{b} + \mathbf{c}_B, \mathbf{0})$—that needs to be evaluated recursively using dynamic programming. (See Theorem 2.) This imposes a small but growing (with sample size) computational burden on the calculation of $q_{\mathrm{ASF}}(\mathbf{a}, \mathbf{b}, \mathbf{c})$. Empirical tests (for example, Figure 2) suggested that simply ignoring $q_2(\mathbf{a} + \mathbf{c}_A, \mathbf{b} + \mathbf{c}_B, \mathbf{0})$ leads to almost no loss in accuracy. It is worth addressing the contribution of this term in more detail. We measure its contribution by the following quantity:

$$\frac{q_{\mathrm{ASF}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) - \tilde{q}_{\mathrm{ASF}}(\mathbf{a}, \mathbf{b}, \mathbf{c})}{q_{\mathrm{ASF}}(\mathbf{a}, \mathbf{b}, \mathbf{c})} \times 100\%$$

$$= \frac{1}{\rho^2} \frac{q_2(\mathbf{a} + \mathbf{c}_A, \mathbf{b} + \mathbf{c}_B, \mathbf{0})}{q_{\mathrm{ASF}}(\mathbf{a}, \mathbf{b}, \mathbf{c})} \times 100\%. \qquad (11)$$

As before, we weight this contribution by the sampling probability $p(\mathbf{a}, \mathbf{b}, \mathbf{c})$ of each configuration. Its distribution is shown in Figure 7, for dimorphic samples of size $c = 20$ and $\rho = 50$.

It is clear from Figure 7 that for these parameter values, using $\tilde{q}_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ in place of $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ affects the result by no more than 1%, and with a high probability the effect is <0.2%. Indeed, when $\theta = 1$, the contribution is <0.2% with an estimated probability of 1. These results are encouraging. They suggest that we may safely discard the $q_2(\mathbf{a} + \mathbf{c}_A, \mathbf{b} + \mathbf{c}_B, \mathbf{0})$ term without a considerable loss of accuracy. This also seems to hold for larger sample sizes. For all samples of size $c = 100$, one can still calculate the contribution (11), although it is now computationally impractical to weight each sample by its true probability $p(\mathbf{0}, \mathbf{0}, \mathbf{c})$. Instead we looked at the *number* of sample configurations within a given range of contributions. For example, when $\rho = 50$ and $\theta = 0.01$, the contribution (11) was within the range ± 0.5% for 93.2% of dimorphic samples of size $c = 100$. When $\theta = 1$, 95.7% of dimorphic sample configurations were within this range. Thus, even for larger sample sizes, when calculating the term $q_2(\mathbf{a} + \mathbf{c}_A, \mathbf{b} + \mathbf{c}_B, \mathbf{0})$ becomes impractical, using the approximation $\tilde{q}_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ seems to be an attractive option. For the infinite-alleles model, we (Jenkins and Song 2009) have obtained a close upper bound to estimate the contribution of this term whenever calculating it directly is computationally intractable.

DISCUSSION

Computing likelihoods under the coalescent with recombination is a challenging problem that has received much attention in the past. Here, building on our recent work (Jenkins and Song 2009), we provided

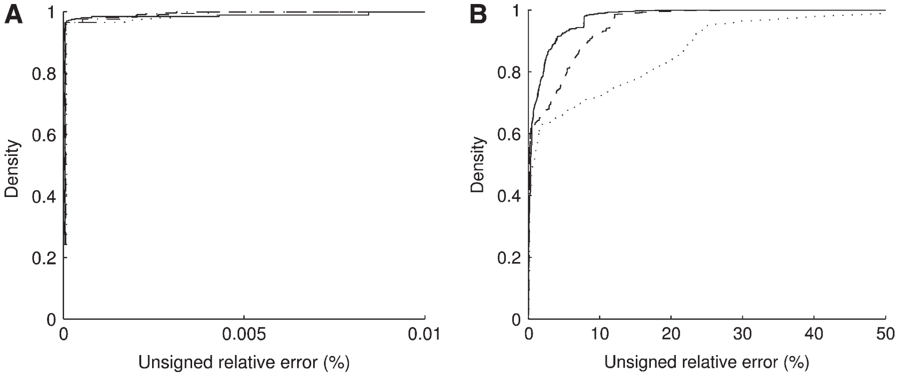

Figure 5.—A repeat of Figure 2 but across all samples rather than all dimorphic samples ($c = 20$, $\rho = 50$). Note the different scales on the x-axes. The curves show the distribution of the unsigned relative error of $q_0(\mathbf{0}, \mathbf{0}, \mathbf{c})$ (dotted line), $q_0(\mathbf{0}, \mathbf{0}, \mathbf{c}) + q_1(\mathbf{0}, \mathbf{0}, \mathbf{c})/\rho$ (dashed line), $\tilde{q}_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ (dotted-dashed line, concealed by the solid line), and $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ (solid line). (a) $\theta = 0.01$. (b) $\theta = 1$.

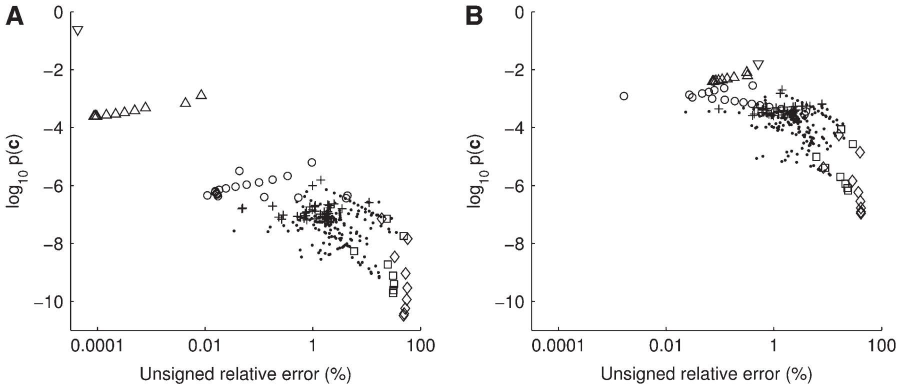

FIGURE 6.—The relationship between unsigned relative error of the ASF and a configuration's sample probability $p(\mathbf{0}, \mathbf{0}, \mathbf{c})$, across all configurations of size $c = 20$ with $\rho = 50$. Configurations are categorized as monomorphic at both loci (∇), monomorphic at one locus ($\triangle$), one or both loci have a singleton allele ($\bigcirc$), very low LD ($+$), in perfect LD ($\diamond$), in perfect LD except for one haplotype ($\square$), and the rest (●). See text for details of these definitions. (a) $\theta = 0.01$. (b) $\theta = 1.0$.

an alternative perspective on the likelihood computation and obtained analytic results for the two-locus model with an arbitrary finite-alleles model of mutation. As mentioned in the Introduction, a stochastic process may exist that is simpler than the well-established coalescent with recombination that captures the important dynamics of the relevant degrees of freedom as the rate of recombination gets large. We think that the fact that the first-order term $q_1(\mathbf{a}, \mathbf{b}, \mathbf{c})$ [see (8)] is so simple supports this speculation. For $\theta = 1$, the Ewens sampling formula for the one-locus infinite-alleles model reduces to the classic formula for the joint distribution of cycle counts in a random permutation (see ARRATIA *et al.* 2003 for a discussion of this and other combinatorial connections). It would be interesting to provide such a combinatorial interpretation of $q_1(\mathbf{a}, \mathbf{b}, \mathbf{c})$. Further, we believe that finding a closed-form expression for the first-order term for more than two loci might be possible.

As a concrete application of our asymptotic sampling formula, one can analytically compute, for large values of ρ, the expectation of linkage disequilibrium measures such as r^2. SONG and SONG (2007) showed analytically that the populationwide expectation of r^2 approaches $1/\rho$ as $\rho \to \infty$. For a finite sample size n, HILL and WEIR (1994) derived $(1/(1 + \rho))(1 - 1/n) + 1/n$ for the samplewise expectation $\mathbb{E}[r^2]$ of r^2. This result was obtained under restrictive assumptions that the populationwide marginal allele frequencies are all $\frac{1}{2}$ and that heterozygote superiority holds at each locus. For a given sample size n, it would be interesting to use our asymptotic sampling formula to derive a general asymptotic expression for the samplewise expectation $\mathbb{E}[r^2]$.

Below we discuss in detail two other concrete applications of our sampling formula. First, we discuss how we can improve the composite-likelihood method (HUDSON 2001a). Then, we provide the first analytic result on the classification of the MLE of ρ.

Composite-likelihood method: LDhat (MCVEAN *et al.* 2002, 2004) is a software package aimed at using patterns of variation in population genetic data to estimate fine-scale recombination rates. The procedure is efficient and has been applied to genomewide estimates of the human genetic map (MYERS *et al.* 2005; INTERNATIONAL HAPMAP CONSORTIUM 2007) (henceforth, we assume that recombination events are always resolved as crossovers and ignore the effects of homologous gene conversion). Part of its efficiency stems from the use of the pairwise composite likelihood first proposed by HUDSON (2001a). For a given genetic map, LDhat employs a Bayesian reversible-jump Markov chain Monte Carlo (rjMCMC) procedure to propose updates to this map and to sample from an approximate posterior distribution (the approximation resulting from the use of a pseudolikelihood defined below). To calculate the acceptance probabilities of proposed moves, one must calculate the likelihood of the data. To do so exactly (*e.g.*, by solving a multilocus version of Equation 5) or

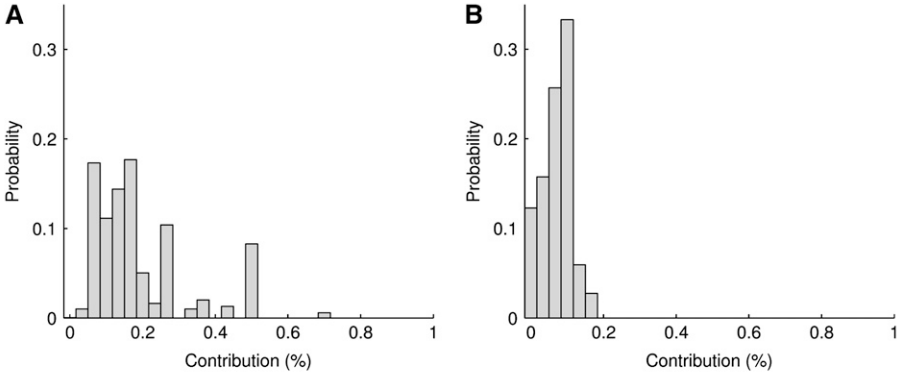

FIGURE 7.—Distribution of the contribution (11) of $q_2(\mathbf{c}_A, \mathbf{c}_B, \mathbf{0})$ to $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$, with respect to dimorphic samples of size $c = 20$ drawn at random from $p(\mathbf{0}, \mathbf{0}, \mathbf{c})$ for $\rho = 50$. (a) $\theta = 0.01$. (b) $\theta = 1$.

even by a computationally intensive full-likelihood procedure such as importance sampling (Fearnhead and Donnelly 2001) is highly impractical for the sizes of data sets of interest. To circumvent this problem, Hudson (2001a) proposed the *composite-likelihood* method and showed that it works very well as an estimator for ρ. This method uses two-locus likelihoods as building blocks and LDhat uses the following variation on the theme: Suppose that the input data set D contains m polymorphic sites, and let D_{ij} denote the data set restricted to the two sites i and j. Then, the composite likelihood is defined as

$$L_C(\rho) = \prod_{i<j} p(D_{ij}). \qquad (12)$$

In this formula, ρ is assumed to be constant across the region and the probability associated with each pair uses a rescaled value for ρ to account for the physical distance between the pair. It is straightforward to modify this definition to allow for variable recombination rates, as implemented in LDhat. The composite-likelihood approximation treats the collection $\{D_{ij} : i < j \text{ and } i, j = 1, \ldots, m\}$ as $\binom{m}{2}$ independent observations. Of course, in reality these observations are highly dependent through their shared ancestry, but often in practice little information about the MLE of ρ is lost. The key to this approach is that it is much more straightforward to calculate each term in the product (12) than it is to calculate the full likelihood, since now one has to deal only with two sites at any one time. Hudson (2001a) considered an infinite-sites model in the limit as $\theta \to 0$, conditional on the observed variation. McVean *et al.* (2002) extended this approach to finite-alleles models and finite θ and used the importance sampling scheme of Fearnhead and Donnelly (2001) to estimate each of these two-site likelihoods. This importance sampling stage is the most computationally intensive part of the LDhat estimation procedure. By considering all possible pairwise sample configurations of a given size, it is possible to precompute these likelihoods, and indeed the LDhat software package is accompanied by precomputed lookup tables for various sample sizes and mutation rates. For each sample configuration and mutation rate, the likelihood is calculated over a grid of ρ-values known as the *driving values*; the default choice is over the integers $\rho = 0, 1, \ldots, 100$. Genetic distances are then rounded to the nearest integer. For pairs of sites whose genetic distance is >100, the two-site likelihood simply truncates ρ to 100 and uses this precomputed value for the likelihood instead.

Here, we investigate the effect that this truncation has on the two-site likelihood. Since the ASF applies to the diallelic, symmetric mutation model employed by LDhat, an obvious alternative to truncation is to use the ASF. As ρ increases, we expect the truncation procedure to introduce more error, while the ASF becomes more accurate. Furthermore, the ASF is a continuous function of ρ and so no further error is introduced by rounding to the nearest driving value of ρ. Assuming a recombination rate of 1 cM/Mb (Kong *et al.* 2002) and a human diploid effective population size of 10,000 (Myers *et al.* 2005), a cutoff of $\rho = 100$ corresponds to a marker separation of ~250 kb. In the presence of recombination hotspots this distance could be much lower. The study of Myers *et al.* (2005) breaks up genomewide data into windows of 2000 SNPs, using a data set with an average marker spacing of 1.9 kb (Hinds *et al.* 2005). A rough approximation to the mean window size is thus ~3.8 Mb; pairs of sites separated by a distance $>\rho = 100$ are therefore encountered frequently by the rjMCMC procedure.

TABLE 3

The mean unsigned relative error (percentage) of $L(100)$ and of $L_{\mathrm{ASF}}(\rho)$ compared to the true likelihood $L(\rho)$, across all dimorphic samples of size c

	$c = 20$: ρ				$c = 30$: ρ			
θ	100	150	200	250	100	150	200	250
0.002								
$L(100)$	0	11.3	19.0	24.5	0	22.9	42.8	59.7
$L_{\mathrm{ASF}}(\rho)$	1.6	0.7	0.4	0.2	5.3	2.8	1.7	1.1
0.02								
$L(100)$	0	11.3	18.9	24.5	0	22.9	42.6	59.5
$L_{\mathrm{ASF}}(\rho)$	1.6	0.7	0.4	0.2	5.3	2.8	1.7	1.1
1.0								
$L(100)$	0	9.9	16.4	21.1	0	20.3	37.3	51.4
$L_{\mathrm{ASF}}(\rho)$	1.1	0.4	0.2	0.1	4.2	2.2	1.3	0.8

Using its estimate of the likelihood at $\rho = 100$, LDhat encounters two sources of error: one from the truncation of the range of ρ and one from the stochastic error of the importance sampling procedure. To restrict our attention to the first of these sources, we find a lower bound on the error of LDhat by looking at the quantity

$$\left|\frac{L(\rho) - L(100)}{L(\rho)}\right| \times 100\%,$$

where $L(\rho)$ is the *true* likelihood obtained previously by solving Equation 5. The true likelihood at $\rho = 100$ will of course be at least as accurate an approximation as that reported by LDhat. Table 3 reports the mean unsigned relative error one would obtain using $L(100)$ in place of $L(\rho)$, for various choices of ρ and various choices of mutation rates used by the lookup tables in LDhat. For comparison we also show the unsigned relative error of $L_{\mathrm{ASF}}(\rho) := q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$. In Figure 8, we also show the complete distribution of these errors for one set of parameters in the table, namely, $c = 30$ and $\theta = 0.002$. (Note that this choice of mutation rate would be reported as $\theta = 0.001$ in LDhat, since it uses the mutation transition matrix $\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ at each locus instead of $\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$ used here.) Since ρ is not fixed throughout the rjMCMC procedure, it is not clear in this setting how to interpret the idea of drawing a sample configuration

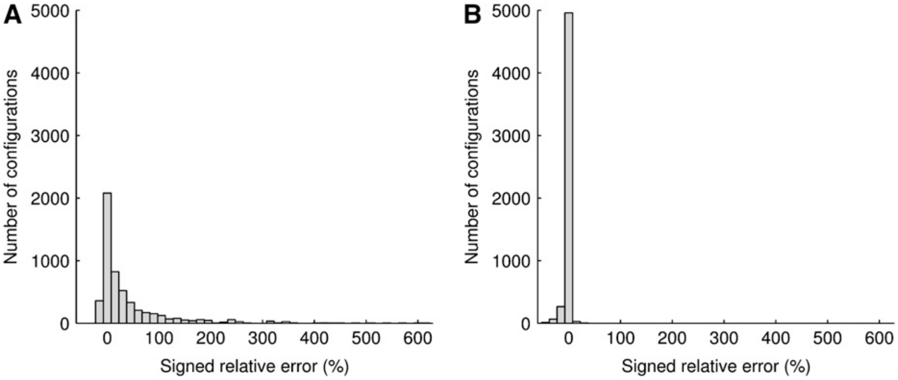

Figure 8.—Distribution of signed relative errors (%) of (a) $L(100)$ and (b) $L_{\mathrm{ASF}}(200)$ when estimating $L(200)$, across all dimorphic configurations of size $c = 30$. The mutation rate is $\theta = 0.002$.

at random. Here, we therefore take a simple arithmetic mean unsigned relative error across all dimorphic samples, rather than weighting each configuration by its sampling probability $p(\mathbf{0}, \mathbf{0}, \mathbf{c})$ as before.

As is clear from Table 2, using $L(100)$ as a proxy for $L(\rho)$ when $\rho > 100$ rapidly becomes very inaccurate with increasing ρ. Even for these moderate sample sizes, there may be a considerable amount of information in the sample so that the likelihoods at $\rho = 100$ and $\rho = 200$, say, differ substantially. For the sample sizes considered here, using the ASF instead of LDhat becomes preferable very soon above $\rho = 100$ and certainly before $\rho = 150$. For larger samples the critical point at which the ASF becomes preferable may be somewhat higher. Also, note that the error in both methods seems to be highly robust to the choice of mutation rate, at least for the set of realistic choices shown. For both methods, there is a slight decrease in error rate with increasing mutation rate.

Figure 8 confirms the wider distribution of errors when using $L(100)$ to estimate $L(200)$ rather than using $L_{\mathrm{ASF}}(200)$. Since we plot here the signed relative error rather than the unsigned relative error, we are also able to observe that the distribution of errors is not symmetric about zero; there is much more probability mass on nonnegligible positive relative errors than on negative relative errors. This suggests that LDhat could systematically overestimate the likelihood of pairs of markers for which a large recombination rate is proposed, biasing recombination rate estimates upward; proposed alterations to the genetic map that increase the recombination rate in a region will have erroneously inflated acceptance rates. When using $L_{\mathrm{ASF}}(\rho)$, the asymmetry in errors is much smaller.

Classification of the MLE of ρ: The ASF can be used to make inferences about whether or not the MLE of ρ is finite. Intuitively, when $q_1(\mathbf{a}, \mathbf{b}, \mathbf{c}) > 0$, the asymptotic approximation $q_0(\mathbf{a}, \mathbf{b}, \mathbf{c}) + q_1(\mathbf{a}, \mathbf{b}, \mathbf{c})/\rho$ approaches $q_0(\mathbf{a}, \mathbf{b}, \mathbf{c})$ from above as $\rho \to \infty$. Moreover, by virtue of it being an asymptotic approximation, for large enough ρ it must be sufficiently accurate to infer that $q(\mathbf{a}, \mathbf{b}, \mathbf{c})$ must also approach $q_0(\mathbf{a}, \mathbf{b}, \mathbf{c})$ from above. We therefore have the following theorem, for which a formal proof may be given via a short analysis argument (in the appendix):

Theorem 3. *For a given sample configuration* $(\mathbf{a}, \mathbf{b}, \mathbf{c})$, if the first-order term $q_1(\mathbf{a}, \mathbf{b}, \mathbf{c})$ [*see* (8)] *in the asymptotic expansion satisfies*

$$q_1(\mathbf{a}, \mathbf{b}, \mathbf{c}) > 0,$$

then the MLE of ρ *is finite.*

If one-locus marginal sampling probabilities are known, as in the infinite-alleles case or the PIM case, then $q_1(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is instantaneous to compute. To our knowledge,

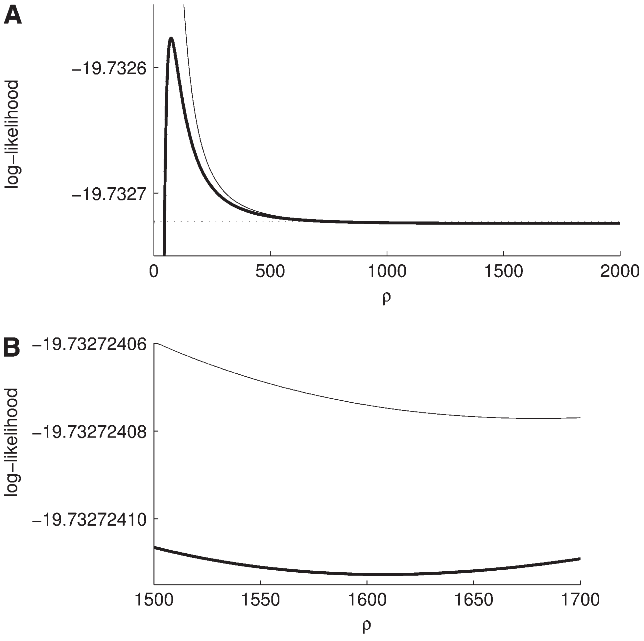

Figure 9.—(a) Example with a local minimum and a local maximum in the likelihood curve for ρ. This example is for $\mathbf{c} = \begin{pmatrix} 6 & 3 \\ 1 & 0 \end{pmatrix}$ under a finite-alleles model with symmetric mutation rates and $\theta = 0.01$. The curve has a local minimum at $\rho = 1608$ (shown in detail in b) and a local maximum at $\rho = 74$. The true likelihood is shown as a thick solid line. For comparison, also shown are $q_0(\mathbf{0}, \mathbf{0}, \mathbf{c}) = L(\infty)$ (dotted line) and $q_{\mathrm{ASF}}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ (thin solid line).

this is the first analytic result on the classification of the MLE of ρ. Note, however, that this is a sufficient condition, but not a necessary condition. That is, this condition does not identify *all* samples with a finite MLE. Indeed, it is tempting to argue for the converse to the proposition above—namely, that $q_1(\mathbf{a}, \mathbf{b}, \mathbf{c}) < 0$ implies an infinite MLE—but this is false, as the following simple counterexample demonstrates.

Consider the sample configuration $(\mathbf{0}, \mathbf{0}, \mathbf{c})$ determined by $\mathbf{c} = \begin{pmatrix} 6 & 1 \\ 3 & 0 \end{pmatrix}$, and suppose $\theta = 0.01$. It is straightforward to verify that $q_1(\mathbf{0}, \mathbf{0}, \mathbf{c}) < 0$ but that ρ_0 exists with $L(\rho_0) > L(\infty)$. The likelihood curve for this configuration can be found by solving Equation 5 directly and is illustrated in Figure 9. Although the curve approaches the asymptote $L(\infty)$ $(= q_0(\mathbf{0}, \mathbf{0}, \mathbf{c}))$ from below as $\rho \to \infty$ [as one expects when $q_1(\mathbf{a}, \mathbf{b}, \mathbf{c}) < 0$], it actually exhibits a local minimum at $\rho = 1608$, as well as a local maximum farther down at $\rho = 74$. In this example the likelihood curve is extremely flat, but minima with higher curvature may be found for larger sample sizes. To our knowledge, this is the first confirmation of a local minimum in a likelihood curve for ρ.

We thank Fulton Wang for writing a computer program that numerically solves two-locus recursions and Rasmus Nielsen and Monty Slatkin for helpful comments. This research is supported in part by a National Institutes of Health grant R00-GM080099 (to Y.S.S.), an Alfred P. Sloan Research Fellowship (to Y.S.S.), and a Packard Fellowship for Science and Engineering (to Y.S.S.).

LITERATURE CITED

Arratia, A., A. D. Barbour and S. Tavaré, 2003 *Logarithmic Combinatorial Structures: A Probabilistic Approach.* European Mathematical Society Publishing House, Zurich, Switzerland.

De Iorio, M., and R. C. Griffiths, 2004a Importance sampling on coalescent histories I. Adv. Appl. Probab. **36:** 417–433.

De Iorio, M., and R. C. Griffiths, 2004b Importance sampling on coalescent histories II. Adv. Appl. Probab. **36:** 434–454.

Ethier, S. N., and R. C. Griffiths, 1990 On the two-locus sampling distribution. J. Math. Biol. **29:** 131–159.

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

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

Golding, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108:** 257–274.

Griffiths, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. **19:** 169–186.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3:** 479–502.

Griffiths, R. C., and S. Tavaré, 1994 Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

Griffiths, R. C., P. A. Jenkins and Y. S. Song, 2008 Importance sampling and the two-locus model with subdivided population structure. Adv. Appl. Probab. **40:** 473–500.

Hill, W. G., and B. S. Weir, 1994 Maximum-likelihood estimation of gene location by linkage disequilibrium. Am. J. Hum. Genet. **54:** 705–714.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halprein, E. Eskin *et al.*, 2005 Whole-genome patterns of common DNA variation in three human populations. Science **307:** 1072–1079.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 2001a Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

Hudson, R. R., 2001b Linkage disequilibrium and recombination, Chap. 11 in *Handbook of Statistical Genetics*, edited by D. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

International HapMap Consortium, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature **449:** 851–861.

Jenkins, P. A., and Y. S. Song, 2009 An asymptotic sampling formula for the coalescent with recombination. Ann. Appl. Probab. (in press). Available as Technical Report 775. Department of Statistics, University of California, Berkeley, CA (http://www.stat.berkeley.edu/tech-reports).

Kingman, J. F. C., 1982a On the genealogy of large populations. J. Appl. Probab. **19:** 27–43.

Kingman, J. F. C., 1982b The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241–247.

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

Lundstrom, R., S. Tavaré and R. H. Ward, 1992 Modeling the evolution of the human mitochondrial genome. Math. Biosci. **112:** 319–335.

Lyngsø, R. B., Y. S. Song and J. Hein, 2008 Accurate computation of likelihoods in the coalescent with recombination via parsimony, pp. 463–477 in *Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB)* (Lecture Notes in Computer Science, Vol. 4955), edited by M. Vingron and L. Wong, Springer, Berlin/Heidelberg, Germany.

McVean, G., P. Awadalla and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231–1241.

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.

Myers, S., L. Bottolo, C. Freeman, G. McVean and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321–324.

Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics **156:** 297–304.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

Song, Y. S., and J. S. Song, 2007 Analytic computation of the expectation of the linkage disequilibrium coefficient r^2. Theor. Popul. Biol. **71:** 49–60.

Stephens, M., and P. Donnelly, 2000 Inference in molecular population genetics. J. R. Stat. Soc. B **62:** 605–655.

Wang, Y., and B. Rannala, 2008 Bayesian inference of fine-scale recombination rates using population genomic data. Philos. Trans. R. Soc. B **363:** 3921–3930.

Wright, S., 1949 Adaptation and selection, pp. 365–389 in *Genetics, Paleontology and Evolution*, edited by G. L. Jepson, E. Mayr and G. G. Simpson. Princeton University Press, Princeton, NJ.

Communicating editor: J. Wakeley

APPENDIX

The proofs given here are similar in structure to our previous proofs (Jenkins and Song 2009) for the infinite-alleles model of mutation. It turns out that much of the argument is independent of the particular model of mutation assumed. Therefore, we give only an outline here and refer the reader to that article for further details.

Proof of Proposition 1. The infinite-alleles case was proved in Jenkins and Song (2009), so we focus on the finite-alleles case here. First, assume $c > 0$. One can obtain a recursion satisfied by $q_0(\mathbf{a},\mathbf{b},\mathbf{c})$ by substituting our asymptotic expansion (6) into (5), dividing by ρc, and letting $\rho \to \infty$. We obtain

$$q_0(\mathbf{a},\mathbf{b},\mathbf{c}) = \sum_{i=1}^{K}\sum_{j=1}^{L}\frac{c_{ij}}{c}q_0(\mathbf{a}+\mathbf{e}_i,\mathbf{b}+\mathbf{e}_j,\mathbf{c}-\mathbf{e}_{ij}). \tag{A1}$$

By repeated application of (A1) this becomes

$$q_0(\mathbf{a},\mathbf{b},\mathbf{c}) = q_0(\mathbf{a}+\mathbf{c}_A,\mathbf{b}+\mathbf{c}_B,\mathbf{0}), \tag{A2}$$

which of course also holds when $c = 0$. Now, by substituting (6) into (5) with $\mathbf{c} = \mathbf{0}$ and then letting $\rho \to \infty$, we obtain [after some manipulation, invoking (A2) to eliminate terms for which $c > 0$]

$$\begin{aligned}
&[a(a+\theta_A-1)+b(b+\theta_B-1)]q_0(\mathbf{a},\mathbf{b},\mathbf{0})\\
&\quad=\sum_{i=1}^{K}a_i(a_i-1)q_0(\mathbf{a}-\mathbf{e}_i,\mathbf{b},\mathbf{0})+\sum_{j=1}^{L}b_j(b_j-1)q_0(\mathbf{a},\mathbf{b}-\mathbf{e}_j,\mathbf{0})\\
&\qquad+\theta_A\sum_{i=1}^{K}a_i\sum_{k=1}^{K}P^A_{ki}q_0(\mathbf{a}-\mathbf{e}_i+\mathbf{e}_k,\mathbf{b},\mathbf{0})+\theta_B\sum_{j=1}^{L}b_j\sum_{l=1}^{L}P^B_{lj}q_0(\mathbf{a},\mathbf{b}-\mathbf{e}_j,\mathbf{0}).
\end{aligned} \tag{A3}$$

Boundary conditions are $q_0(\mathbf{e}_i,\mathbf{0},\mathbf{0}) = 1$ for $i = 1,\ldots,K$ and $q_0(\mathbf{0},\mathbf{e}_j,\mathbf{0}) = 1$ for $j = 1,\ldots,L$. It is straightforward to verify that the solution to (A3) is

$$q_0(\mathbf{a},\mathbf{b},\mathbf{0}) = q^A(\mathbf{a})q^B(\mathbf{b}), \tag{A4}$$

where $q^A(\mathbf{a})$ and $q^B(\mathbf{b})$ are marginal one-locus sampling probabilities for loci A and B, respectively. Together, (A2) and (A4) give the desired result. ■

Proof of Theorem 1. Again, the infinite-alleles case was proved in Jenkins and Song (2009), so we focus on the finite-alleles case here. Substitute the asymptotic expansion (6) into the recursion (5), eliminate terms of order ρ by applying (A1), and let $\rho \to \infty$. After invoking (A3) and (7), with some rearrangement we obtain

$$q_1(\mathbf{a},\mathbf{b},\mathbf{c}) = f(\mathbf{a},\mathbf{b},\mathbf{c}) + \sum_{i=1}^{K}\sum_{j=1}^{L}\frac{c_{ij}}{c}q_1(\mathbf{a}+\mathbf{e}_i,\mathbf{b}+\mathbf{e}_j,\mathbf{c}-\mathbf{e}_{ij}), \tag{A5}$$

where

$$\begin{aligned}
f(\mathbf{a},\mathbf{b},\mathbf{c}) :=\ &(c-1)q^A(\mathbf{a}+\mathbf{c}_A)q^B(\mathbf{b}+\mathbf{c}_B)\\
&-q^B(\mathbf{b}+\mathbf{c}_B)\sum_{i=1}^{K}\frac{c_{i\cdot}(c_{i\cdot}-1)}{c}q^A(\mathbf{a}+\mathbf{c}_A-\mathbf{e}_i)\\
&-q^A(\mathbf{a}+\mathbf{c}_A)\sum_{j=1}^{L}\frac{c_{\cdot j}(c_{\cdot j}-1)}{c}q^B(\mathbf{b}+\mathbf{c}_B-\mathbf{e}_j)\\
&+\sum_{i=1}^{K}\sum_{j=1}^{L}\frac{c_{ij}(c_{ij}-1)}{c}q^A(\mathbf{a}+\mathbf{c}_A-\mathbf{e}_i)q^B(\mathbf{b}+\mathbf{c}_B-\mathbf{e}_j).
\end{aligned} \tag{A6}$$

Above we assumed $c > 0$; we define $f(\mathbf{a},\mathbf{b},\mathbf{c}) = 0$ if $\mathbf{c} = \mathbf{0}$.

Although at first sight somewhat daunting, Equation A5 can be solved via a simple probabilistic interpretation of the coefficients of the second term on the right-hand side. In particular, this term can be interpreted as the expected value of $q_1(\mathbf{a},\mathbf{b},\mathbf{c})$ after removing a gamete from $\mathbf{c}$ uniformly at random and redistributing it to $\mathbf{a}$ and $\mathbf{b}$. After this removal, the term is given recursively by discarding further gametes. More generally, we (Jenkins and Song 2009) showed that

$$q_1(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_{m=1}^{c} \mathbb{E}[f(\mathbf{A}^{(m)}, \mathbf{B}^{(m)}, \mathbf{C}^{(m)})], \tag{A7}$$

where $\mathbf{C}^{(m)}$ is the random subsample obtained by selecting m gametes from the sample $\mathbf{c}$ uniformly without replacement; that is, it is distributed according to a multivariate hypergeometric distribution with parameters $(c, \mathbf{c}, m)$. Further, we define $\mathbf{A}^{(m)} = \mathbf{a} + \mathbf{c}_A - \mathbf{C}^{(m)}$ and $\mathbf{B}^{(m)} = \mathbf{b} + \mathbf{c}_B - \mathbf{C}^{(m)}$. The expectations in (A7) are easy to compute and to sum over m, resulting in (8). ■

Proof of Theorem 2. In a similar manner to the proof of Theorem 1, one can show that

$$q_2(\mathbf{a}, \mathbf{b}, \mathbf{c}) = q_2(\mathbf{a} + \mathbf{c}_A, \mathbf{b} + \mathbf{c}_B, \mathbf{0}) + \sum_{m=1}^{c} \mathbb{E}[g(\mathbf{A}^{(m)}, \mathbf{B}^{(m)}, \mathbf{C}^{(m)})], \tag{A8}$$

where $g(\mathbf{a}, \mathbf{b}, \mathbf{c})$ has a known, but rather cumbersome, closed-form expression. Since these expectations can be evaluated, the sum $\sum_{m=1}^{c} \mathbb{E}[g(\mathbf{A}^{(m)}, \mathbf{B}^{(m)}, \mathbf{C}^{(m)})]$ composes the closed-form part of $q_2(\mathbf{a}, \mathbf{b}, \mathbf{c})$; *i.e.*,

$$\sigma(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_{m=1}^{c} \mathbb{E}[g(\mathbf{A}^{(m)}, \mathbf{B}^{(m)}, \mathbf{C}^{(m)})].$$

Unlike the form of $f(\mathbf{a}, \mathbf{b}, \mathbf{c})$, that of $g(\mathbf{a}, \mathbf{b}, \mathbf{c})$ depends on the model of mutation. We repeated the steps in Jenkins and Song (2009) using Equation 5—rather than Golding's recursion (Golding 1984), which applies to the infinite-alleles model—to obtain the following expression for $\sigma(\mathbf{a}, \mathbf{b}, \mathbf{c})$. We omit the simple but lengthy algebraic details.

Let Q^A denote $q^A(\mathbf{a} + \mathbf{c}_A)$, Q_i^A denote $q^A(\mathbf{a} + \mathbf{c}_A - \mathbf{e}_i)$, Q_{ik}^A denote $q^A(\mathbf{a} + \mathbf{c}_A - \mathbf{e}_i - \mathbf{e}_k)$, and so on. Then,

$$\begin{aligned}
\sigma(\mathbf{a}, \mathbf{b}, \mathbf{c}) = {} & \frac{c}{3}\left[\frac{(c-1)(c+1)(3c-2)}{8} + (c-1)(3a+3b+2c-1) + 6ab\right] Q^A Q^B \\
& + \sum_{i=1}^{K}\left[\frac{\theta_A - c(c-3) + 2a + 4b - 4}{4} c_{i\cdot}(c_{i\cdot} - 1) - (2b + c - 1)c_{i\cdot}(a_i + c_{i\cdot} - 1)\right] Q_i^A Q^B \\
& + \sum_{j=1}^{L}\left[\frac{\theta_B - c(c-3) + 2b + 4a - 4}{4} c_{\cdot j}(c_{\cdot j} - 1) - (2a + c - 1)c_{\cdot j}(b_j + c_{\cdot j} - 1)\right] Q^A Q_j^B \\
& + \frac{1}{12}\sum_{i=1}^{K}(5 - 6a_i - 4c_{i\cdot})c_{i\cdot}(c_{i\cdot} - 1) Q_{ii}^A Q^B + \frac{1}{12}\sum_{j=1}^{L}(5 - 6b_j - 4c_{\cdot j})c_{\cdot j}(c_{\cdot j} - 1) Q^A Q_{jj}^B \\
& + \sum_{i=1}^{K}\sum_{k=1}^{K}\frac{c_{i\cdot}(c_{i\cdot} - 1)c_{k\cdot}(c_{k\cdot} - 1)}{8} Q_{ik}^A Q^B + \sum_{j=1}^{L}\sum_{l=1}^{L}\frac{c_{\cdot j}(c_{\cdot j} - 1)c_{\cdot l}(c_{\cdot l} - 1)}{8} Q^A Q_{jl}^B \\
& - \frac{\theta_A + \theta_B - c(c-5) + 2a + 2b - 4}{4}\sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij}(c_{ij} - 1) Q_i^A Q_j^B \\
& + \sum_{i=1}^{K}\sum_{j=1}^{L}\left[\frac{c_{i\cdot}(c_{i\cdot} - 1)c_{\cdot j}(c_{\cdot j} - 1)}{4} + \frac{c_{ij}(c_{ij} + 1 - 2c_{i\cdot} + 2c_{i\cdot}c_{\cdot j} - 2c_{\cdot j})}{2}\right. \\
& \qquad \left. + c_{ij}b_j(c_{i\cdot} - 1) + c_{ij}a_i(c_{\cdot j} - 1) + 2a_i b_j c_{ij}\right] Q_i^A Q_j^B \\
& + \frac{1}{2}\sum_{i=1}^{K}(a_i + c_{i\cdot} - 1)\sum_{j=1}^{L} c_{ij}(c_{ij} - 1) Q_{ii}^A Q_j^B + \frac{1}{2}\sum_{j=1}^{L}(b_j + c_{\cdot j} - 1)\sum_{i=1}^{K} c_{ij}(c_{ij} - 1) Q_i^A Q_{jj}^B \\
& - \frac{1}{4}\sum_{i=1}^{K}\sum_{j=1}^{L}\sum_{k=1}^{K} c_{ij}(c_{ij} - 1)c_{k\cdot}(c_{k\cdot} - 1) Q_{ik}^A Q_j^B - \frac{1}{4}\sum_{i=1}^{K}\sum_{j=1}^{L}\sum_{l=1}^{L} c_{ij}(c_{ij} - 1)c_{\cdot l}(c_{\cdot l} - 1) Q_i^A Q_{jl}^B \\
& + \frac{1}{8}\sum_{i=1}^{K}\sum_{j=1}^{L}\sum_{k=1}^{K}\sum_{l=1}^{L} c_{ij}(c_{ij} - 1)c_{kl}(c_{kl} - 1) Q_{ik}^A Q_{jl}^B \\
& - \frac{1}{12}\sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij}(c_{ij} - 1)(2c_{ij} - 1) Q_{ii}^A Q_{jj}^B \\
& + \lambda(\mathbf{a}, \mathbf{b}, \mathbf{c}).
\end{aligned}$$

The final term, $\lambda(\mathbf{a}, \mathbf{b}, \mathbf{c})$, is a function whose form depends on the model of mutation. Using $Q_{i,+t}^A$ to denote $q^A(\mathbf{a} + \mathbf{c}_A - \mathbf{e}_i + \mathbf{e}_t)$, and so on, we obtain

$$\begin{aligned}
\lambda(\mathbf{a},\mathbf{b},\mathbf{c}) = &-\frac{\theta_A}{2}\sum_{i=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{K}P_{ti}^{A}(\delta_{tk}-\delta_{ik})c_{i\cdot}(c_{k\cdot}-\delta_{ik})Q_{ik,+t}^{A}Q^{B}\\
&-\frac{\theta_B}{2}\sum_{j=1}^{L}\sum_{l=1}^{L}\sum_{t=1}^{L}P_{tj}^{B}(\delta_{tl}-\delta_{jl})c_{\cdot j}(c_{\cdot l}-\delta_{jl})Q^{A}Q_{jl,+t}^{B}\\
&-\frac{\theta_A}{4}\sum_{i=1}^{K}\sum_{t=1}^{K}P_{ti}^{A}c_{i\cdot}(c_{i\cdot}-1)Q_{ii,+t}^{A}Q^{B}-\frac{\theta_B}{4}\sum_{j=1}^{L}\sum_{t=1}^{L}P_{tj}^{B}c_{\cdot j}(c_{\cdot j}-1)Q^{A}Q_{jj,+t}^{B}\\
&+\frac{\theta_A}{2}\sum_{i=1}^{K}\sum_{j=1}^{L}\sum_{k=1}^{K}\sum_{t=1}^{K}P_{ti}^{A}(\delta_{tk}-\delta_{ik})c_{ij}(c_{kj}-\delta_{ik})Q_{ik,+t}^{A}Q_{j}^{B}\\
&+\frac{\theta_B}{2}\sum_{i=1}^{K}\sum_{j=1}^{L}\sum_{l=1}^{L}\sum_{t=1}^{L}P_{tj}^{B}(\delta_{tl}-\delta_{jl})c_{ij}(c_{il}-\delta_{jl})Q_{i}^{A}Q_{jl,+t}^{B}\\
&+\frac{\theta_A}{4}\sum_{i=1}^{K}\sum_{j=1}^{L}\sum_{t=1}^{K}P_{ti}^{A}c_{ij}(c_{ij}-1)Q_{ii,+t}^{A}Q_{j}^{B}+\frac{\theta_B}{4}\sum_{i=1}^{K}\sum_{j=1}^{L}\sum_{t=1}^{L}P_{tj}^{B}c_{ij}(c_{ij}-1)Q_{i}^{A}Q_{jj,+t}^{B},
\end{aligned}$$

where δ_{ij} denotes the Kronecker delta, and in a PIM model this simplifies to

$$\begin{aligned}
\lambda(\mathbf{a},\mathbf{b},\mathbf{c}) = &-\frac{\theta_A(c-1)}{2}\sum_{i=1}^{K}c_{i\cdot}P_{i}^{A}Q_{i}^{A}Q^{B}-\frac{\theta_B(c-1)}{2}\sum_{j=1}^{L}c_{\cdot j}P_{j}^{B}Q^{A}Q_{j}^{B}\\
&+\frac{\theta_A}{4}\sum_{i=1}^{K}P_{i}^{A}c_{i\cdot}(c_{i\cdot}-1)Q_{ii}^{A}Q^{B}+\frac{\theta_B}{4}\sum_{j=1}^{L}P_{j}^{B}c_{\cdot j}(c_{\cdot j}-1)Q^{A}Q_{jj}^{B}\\
&+\sum_{i=1}^{K}\sum_{j=1}^{L}\left[\frac{\theta_B}{2}P_{j}^{B}c_{ij}(c_{i\cdot}-1)+\frac{\theta_A}{2}P_{i}^{A}c_{ij}(c_{\cdot j}-1)\right]Q_{i}^{A}Q_{j}^{B}\\
&-\frac{\theta_A}{4}\sum_{i=1}^{K}P_{i}^{A}\sum_{j=1}^{L}c_{ij}(c_{ij}-1)Q_{ii}^{A}Q_{j}^{B}-\frac{\theta_B}{4}\sum_{j=1}^{L}P_{j}^{B}\sum_{i=1}^{K}c_{ij}(c_{ij}-1)Q_{i}^{A}Q_{jj}^{B}.
\end{aligned}$$

For completeness, we also give the corresponding expression for the infinite-alleles model,

$$\begin{aligned}
\lambda(\mathbf{a},\mathbf{b},\mathbf{c}) = &-\frac{\theta_A(c-1)}{2}\sum_{i=1}^{K}\delta_{a_i,0}\delta_{c_{i\cdot},1}Q_{i}^{A}Q^{B}-\frac{\theta_B(c-1)}{2}\sum_{j=1}^{L}\delta_{b_j,0}\delta_{c_{\cdot j},1}Q^{A}Q_{j}^{B}\\
&+\frac{\theta_A}{2}\sum_{i=1}^{K}\delta_{c_{i\cdot},2}Q_{ii}^{A}Q^{B}+\frac{\theta_B}{2}\sum_{j=1}^{L}\delta_{c_{\cdot j},2}Q^{A}Q_{jj}^{B}\\
&+\sum_{i=1}^{K}\sum_{j=1}^{L}\left[\frac{\theta_B}{2}\delta_{b_j,0}\delta_{c_{\cdot j},1}\delta_{c_{ij},1}(c_{i\cdot}-1)+\frac{\theta_A}{2}\delta_{a_i,0}\delta_{c_{i\cdot},1}\delta_{c_{ij},1}(c_{\cdot j}-1)\right]Q_{i}^{A}Q_{j}^{B}\\
&-\frac{\theta_A}{2}\sum_{i=1}^{K}\sum_{j=1}^{L}\delta_{c_{i\cdot},2}\delta_{c_{ij},2}Q_{ii}^{A}Q_{j}^{B}-\frac{\theta_B}{2}\sum_{i=1}^{K}\sum_{j=1}^{L}\delta_{c_{\cdot j},2}\delta_{c_{ij},2}Q_{i}^{A}Q_{jj}^{B}
\end{aligned}$$

(Jenkins and Song 2009).

We now outline the steps to obtain the recursion shown in (10). One can use Equation A8 to show that

$$\begin{aligned}
q_2(\mathbf{a}-\mathbf{e}_i,\mathbf{b}-\mathbf{e}_j,\mathbf{e}_{ij}) = &\, q_2(\mathbf{a},\mathbf{b},\mathbf{0})+2(a-1)(b-1)q^{A}(\mathbf{a})q^{B}(\mathbf{b})\\
&-2(b-1)(a_i-1)q^{A}(\mathbf{a}-\mathbf{e}_i)q^{B}(\mathbf{b})\\
&-2(a-1)(b_j-1)q^{A}(\mathbf{a})q^{B}(\mathbf{b}-\mathbf{e}_j)\\
&+2(a_i-1)(b_j-1)q^{A}(\mathbf{a}-\mathbf{e}_i)q^{B}(\mathbf{b}-\mathbf{e}_j). \qquad \text{(A9)}
\end{aligned}$$

After substituting the asymptotic expansion (6) with $\mathbf{c}=\mathbf{0}$ into (5) and comparing coefficients of terms proportional to $1/\rho^2$, one obtains the recursion

$$
\begin{aligned}
&[n(n-1)+\theta_A a+\theta_B b]q_2(\mathbf{a},\mathbf{b},\mathbf{0})\\
&\quad=\sum_{i=1}^{K} a_i(a_i-1)q_2(\mathbf{a}-\mathbf{e}_i,\mathbf{b},\mathbf{0})+\sum_{j=1}^{L} b_j(b_j-1)q_2(\mathbf{a},\mathbf{b}-\mathbf{e}_j,\mathbf{0})\\
&\qquad+2\sum_{i=1}^{K}\sum_{j=1}^{L} a_i b_j q_2(\mathbf{a}-\mathbf{e}_i,\mathbf{b}-\mathbf{e}_j,\mathbf{e}_{ij})\\
&\qquad+\theta_A\sum_{i=1}^{K} a_i\sum_{t=1}^{K} P_{ti}^{A} q_2(\mathbf{a}-\mathbf{e}_i+\mathbf{e}_t,\mathbf{b},\mathbf{0})+\theta_B\sum_{j=1}^{L} b_j\sum_{t=1}^{L} P_{tj}^{B} q_2(\mathbf{a},\mathbf{b}-\mathbf{e}_j+\mathbf{e}_t,\mathbf{0}),
\end{aligned}
$$

with boundary conditions $q_2(\mathbf{e}_i,\mathbf{0},\mathbf{0})=0$ for $i=1,\ldots,K$ and $q_2(\mathbf{0},\mathbf{e}_j,\mathbf{0})=0$ for $j=1,\ldots,L$. By appealing to (A9), this can be simplified to the form given in Equation 10. ■

Proof of Theorem 3. Since $q_0(\mathbf{a},\mathbf{b},\mathbf{c})+q_1(\mathbf{a},\mathbf{b},\mathbf{c})/\rho$ is an asymptotic approximation to $q(\mathbf{a},\mathbf{b},\mathbf{c}\,|\,\rho)$, by definition

$$\frac{q(\mathbf{a},\mathbf{b},\mathbf{c}\,|\,\rho)-[q_0(\mathbf{a},\mathbf{b},\mathbf{c})+q_1(\mathbf{a},\mathbf{b},\mathbf{c})/\rho]}{q_1(\mathbf{a},\mathbf{b},\mathbf{c})/\rho}\rightarrow 0 \quad \text{as}\ \rho\rightarrow\infty.$$

Hence, using $q_1(\mathbf{a},\mathbf{b},\mathbf{c})>0$, we conclude that ρ' exists such that, for all $\rho'\leq\rho<\infty$,

$$\left|q(\mathbf{a},\mathbf{b},\mathbf{c}\,|\,\rho)-\left[q_0(\mathbf{a},\mathbf{b},\mathbf{c})+\frac{q_1(\mathbf{a},\mathbf{b},\mathbf{c})}{\rho}\right]\right|<\frac{q_1(\mathbf{a},\mathbf{b},\mathbf{c})}{\rho},$$

which implies $q(\mathbf{a},\mathbf{b},\mathbf{c}\,|\,\rho)>q_0(\mathbf{a},\mathbf{b},\mathbf{c})$ for $\rho'\leq\rho<\infty$. Since $L(\infty)=q_0(\mathbf{a},\mathbf{b},\mathbf{c})$, this implies that the MLE for ρ must be finite. ■