AN ASYMPTOTIC SAMPLING FORMULA FOR THE COALESCENT WITH RECOMBINATION

By Paul A. Jenkins* and Yun S. Song*,†

University of California, Berkeley

Ewens sampling formula (ESF) is a one-parameter family of probability distributions with a number of intriguing combinatorial connections. This elegant closed-form formula first arose in biology as the stationary probability distribution of a sample configuration at one locus under the infinite-alleles model of mutation. Since its discovery in the early 1970s, the ESF has been used in various biological applications, and has sparked several interesting mathematical generalizations. In the population genetics community, extending the underlying random-mating model to include recombination has received much attention in the past, but no general closed-form sampling formula is currently known even for the simplest extension, that is, a model with two loci. In this paper, we show that it is possible to obtain useful closed-form results in the case the population-scaled recombination rate ρ is large but not necessarily infinite. Specifically, we consider an asymptotic expansion of the two-locus sampling formula in inverse powers of ρ and obtain closed-form expressions for the first few terms in the expansion. Our asymptotic sampling formula applies to arbitrary sample sizes and configurations.

1. Introduction. The probability of a sample configuration provides a useful ground for analyzing genetic data. Popular applications include obtaining maximum likelihood estimates of model parameters and performing ancestral inference [see Stephens (2001)]. In principle, model-based full-likelihood analyses, such as that based on the coalescent [Kingman (1982a,b)], should be among the most powerful methods since they make full use of the data. However, in most cases it is intractable to obtain a closed-form formula for the probability of a given dataset. A well-known exception to this hurdle is the Ewens sampling formula (ESF), which describes the stationary probability distribution of a sample configuration under the one-locus infinite-alleles model in the diffusion limit [Ewens (1972)]. Notable biolog-

*Supported in part by an NIH grant R00-GM080099

†Supported in part by an Alfred P. Sloan Research Fellowship and a Packard Fellowship for Science and Engineering

AMS 2000 subject classifications: Primary 92D15; secondary 65C50,92D10

Keywords and phrases: Ewens sampling formula, coalescent process, recombination, two-locus model, infinite alleles, Golding's recursion

ical applications of this closed-form formula include the test of selective neutrality [see Slatkin (1994); Watterson (1977)]. Hoppe (1984) provided a Pólya-like urn model interpretation of the formula, and recently Griffiths and Lessard (2005) provided a new combinatorial proof of the ESF and extended the framework to obtain new results for the case with a variable population size. We refer the reader to the latter paper for a nice summary of previous works related to the ESF. Note that the ESF also arises in several interesting contexts outside biology, including random partition structures and Bayesian statistics; see Arratia et al. (2003) for examples of intricate combinatorial connections. The ESF is a special case of the two-parameter sampling formula constructed by Pitman (1992, 1995) for exchangeable random partitions.

Golding (1984) considered generalizing the infinite-alleles model to include recombination and constructed a recursion relation satisfied by the two-locus sampling probability distribution at stationarity in the diffusion limit. Ethier and Griffiths (1990) later undertook a more mathematical analysis of the two-locus model and provided several interesting results. However, to date a general closed-form formula for the two-locus sampling distribution remains unknown. Indeed, it is widely recognized that recombination adds a formidably challenging layer of complexity to population genetics analysis. Because obtaining exact analytic results in the presence of recombination is difficult, recent research has focused on developing sophisticated and computationally-intensive Monte Carlo techniques. Examples of such techniques applied to the coalescent include Monte Carlo simulations [see Hudson (1985, 2001)], importance sampling [see De Iorio and Griffiths (2004a,b); Fearnhead and Donnelly (2001); Griffiths and Marjoram (1996); Griffiths et al. (2008); Stephens and Donnelly (2000)], and Markov Chain Monte Carlo methods [see Kuhner et al. (2000); Nielsen (2000); Wang and Rannala (2008)].

Being the simplest model with recombination, the two-locus case has been extensively studied in the past [Ethier and Griffiths (1990); Golding (1984); Griffiths (1981, 1991); Hudson (1985)], and a renewed wave of interest was recently sparked by Hudson (2001), who proposed a composite likelihood method which uses two-locus sampling probabilities as building blocks. LDhat, a widely-used software package developed by McVean and colleagues, is based on this composite likelihood approach, and it has been used to produce a fine-scale map of recombination rate variation in the human genome [McVean et al. (2004); Myers et al. (2005)]. LDhat relies on the importance sampling scheme proposed by Fearnhead and Donnelly (2001) for the coalescent with recombination, to generate exhaustive lookup tables containing

two-locus probabilities for all inequivalent sample configurations and a range of relevant parameter values. This process of generating exhaustive lookup tables is very computationally expensive. A fast and accurate method of estimating two-locus probabilities would be of practical value.

In this paper, we revisit the tantalizing open question of whether a closed-form sampling formula can be found for the coalescent with recombination. We show that, at least for the two-locus infinite-alleles model with the population-scaled recombination rate ρ large but not necessarily infinite, it is possible to obtain useful closed-form analytic results. Our work generalizes previous results [Ethier and Griffiths (1990); Golding (1984)] for $\rho = \infty$, in which case the loci become independent and the two-locus sampling distribution is given by a product of one-locus ESFs. Our main results can be summarized as follows.

Main results. Consider 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 sample configuration $\boldsymbol{n}$ (defined later in the text), we use $q(\boldsymbol{n}|\theta_A, \theta_B, \rho)$ to denote the probability of observing $\boldsymbol{n}$ given the parameters θ_A, θ_B, and ρ. For an arbitrary $\boldsymbol{n}$, our goal is to find an asymptotic expansion of $q(\boldsymbol{n}|\theta_A, \theta_B, \rho)$ in inverse powers of ρ, i.e., for large values of the recombination rate ρ, our goal is to find

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

where q_0, q_1, and q_2 are independent of ρ. As mentioned before, $q_0(\boldsymbol{n}|\theta_A, \theta_B)$ is given by a product of one-locus ESFs. In this paper, we derive a closed-form formula for the first-order term $q_1(\boldsymbol{n}|\theta_A, \theta_B)$. Further, we show that the second-order term $q_2(\boldsymbol{n}|\theta_A, \theta_B)$ can be decomposed into two parts, one for which we obtain a closed-form formula and the other that satisfies a simple strict recursion. The latter can be easily evaluated using dynamic programming. Details of these results are described in Section 3. In a similar vein, in Section 4 we obtain a simple asymptotic formula for the joint probability distribution of the number of alleles observed at the two loci.

2. Preliminaries. In this section, we review the ESF for the one-locus infinite-alleles model, as well as Golding's (1984) recursion relation for the two-locus generalization. Our notational convention generally follows that of Ethier and Griffiths (1990).

Given a positive integer k, $[k]$ denotes the k-set $\{1, \ldots, k\}$. For a nonnegative real number x and a positive integer n, $(x)_n := x(x+1)\ldots(x+n-1)$ denotes the nth ascending factorial of x. We use $\mathbf{0}$ to denote either

a vector or a matrix of all zeroes; it will be clear from context which is intended. Throughout, we consider the diffusion limit of a neutral haploid exchangeable model of random mating with constant population size $2N$. We refer to the haploid individuals in the population as gametes.

2.1. *Ewens sampling formula for the one-locus model.* In the one-locus model, a sample configuration is denoted by a vector of multiplicities $\boldsymbol{n} = (n_1, \ldots, n_K)$, where n_i denotes the number of gametes with allele i at the locus and K denotes the total number of distinct allelic types observed. We use n to denote $\sum_{i=1}^{K} n_i$, the total sample size. 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. The probability of a mutation event at the locus per gamete per generation is denoted by u. In the diffusion limit, $N \to \infty$ and $u \to 0$ with the population-scaled mutation rate $\theta = 4Nu$ held fixed. Each mutation gives rise to a new allele that has never been seen before in the population. For the one-locus model just described, Ewens (1972) obtained the following result:

PROPOSITION 2.1 (Ewens). *At stationarity in the diffusion limit of the one-locus infinite-alleles model with the scaled mutation parameter θ, the probability of an unordered sample configuration $\boldsymbol{n} = (n_1, \ldots, n_K)$ is given by*

$$p(\boldsymbol{n} \mid \theta) = \frac{n!}{n_1 \ldots n_K} \frac{1}{\alpha_1! \ldots \alpha_n!} \frac{\theta^K}{(\theta)_n}, \tag{2.1}$$

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

Let $\mathscr{A}_n$ denote an *ordered* configuration of n sequentially sampled gametes such that the corresponding unordered configuration is given by $\boldsymbol{n}$. By exchangeability, the probability of $\mathscr{A}_n$ is invariant under all permutations of the sampling order. Hence, we can write this probability of an ordered sample as $q(\boldsymbol{n})$ without ambiguity. It is given by

$$q(\boldsymbol{n} \mid \theta) = p(\boldsymbol{n} \mid \theta) \left[\frac{n!}{\prod_{i=1}^{K} n_i!} \frac{1}{\alpha_1! \ldots \alpha_n!} \right]^{-1} = \left[\prod_{i=1}^{K} (n_i - 1)! \right] \frac{\theta^K}{(\theta)_n}, \tag{2.2}$$

which follows from the fact that there are $\frac{n!}{\prod_{i=1}^{K} n_i!} \frac{1}{\alpha_1! \ldots \alpha_n!}$ orderings corresponding to $\boldsymbol{n}$ [Hoppe (1984)]. It is often more convenient to work with an

ordered sample than with an unordered sample. In this paper, we will work with the former; i.e., we will work with $q(\boldsymbol{n} \mid \theta)$ rather than $p(\boldsymbol{n} \mid \theta)$.

In the coalescent process going backwards in time, at each event a lineage is lost either by coalescence or mutation. By consideration of the most recent event back in time, one can show that $q(\boldsymbol{n} \mid \theta)$ satisfies

$$n(n-1+\theta)q(\boldsymbol{n} \mid \theta) = \sum_{i=1}^{K} n_i(n_i-1)q(\boldsymbol{n}-\boldsymbol{e}_i \mid \theta)+\theta \sum_{i=1}^{K} \delta_{n_i,1} q(\boldsymbol{n}-\boldsymbol{e}_i \mid \theta), \tag{2.3}$$

where $\delta_{n_i,1}$ is the Kronecker delta and $\boldsymbol{e}_i$ is a unit vector with the ith entry equal to one and all other entries equal to zero. The boundary condition is $q(\boldsymbol{e}_i \mid \theta) = 1$ for all $i \in [K]$, and $q(\boldsymbol{n} \mid \theta)$ is defined to be zero if $\boldsymbol{n}$ contains any negative component. It can be easily verified that the formula of $q(\boldsymbol{n} \mid \theta)$ shown in (2.2) satisfies the recursion (2.3).

Ewens (1972) also obtained the following result regarding the number of allelic types:

PROPOSITION 2.2 (Ewens). *Let K_n denote the number of distinct allelic types observed in a sample of size n. Then,*

$$\mathbb{P}(K_n = k \mid \theta) = \frac{s(n,k)\theta^k}{(\theta)_n}, \tag{2.4}$$

where $s(n,k)$ are the unsigned Stirling numbers of the first kind. Note that $(\theta)_n = s(n,1)\theta + s(n,2)\theta^2 + \ldots + s(n,n)\theta^n$.

It follows from (2.1) and (2.4) that K_n is a sufficient statistic for θ.

2.2. *Golding's recursion for the two-locus case.* Golding (1984) first generalized the one-locus recursion (2.3) to two loci, and Ethier and Griffiths (1990) later undertook a more mathematical study of the model. We denote the two loci by A and B, and use θ_A and θ_B to denote the respective population-scaled mutation rates. We use K and L to denote the number of distinct allelic types observed at locus A and locus B, respectively. The population-scaled recombination rate is denoted by $\rho = 4Nr$, where r is the probability of a recombination event between the two loci per gamete per generation. A key observation is that to obtain a closed system of equations, the type space must be extended to allow some gametes to be specified only at one of the two loci.

DEFINITION 2.1 (Extended sample configuration for two loci). *The two-locus sample configuration is denoted by $\boldsymbol{n} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$, where $\boldsymbol{a} = (a_1, \ldots, a_K)$*

with a_i being the number of gametes with allele i at locus A and unspecified alleles at locus B, $\boldsymbol{b} = (b_1, \ldots, b_L)$ with b_j being the number of gametes with unspecified alleles at locus A and allele j at locus B, and $\boldsymbol{c} = (c_{ij})$ 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. Further, we define

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

We use $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ to denote the sampling probability of an ordered sample with configuration $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. For ease of notation, we do not show the dependence on parameters. For $0 \leq \rho < \infty$, Golding's (1984) recursion for $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ takes the following form:

$$\begin{aligned}
&[n(n-1) + \theta_A(a+c) + \theta_B(b+c) + \rho c] q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \\
&\quad \sum_{i=1}^{K} a_i(a_i - 1 + 2c_{i\cdot}) q(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{c}) + \sum_{j=1}^{L} b_j(b_j - 1 + 2c_{\cdot j}) q(\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c}) \\
&\quad + \sum_{i=1}^{K}\sum_{j=1}^{L} [c_{ij}(c_{ij} - 1) q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} - \boldsymbol{e}_{ij}) + 2a_i b_j q(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c} + \boldsymbol{e}_{ij})] \\
&\quad + \theta_A \sum_{i=1}^{K} \left[\sum_{j=1}^{L} \delta_{a_i + c_{i\cdot}, 1} \delta_{c_{ij}, 1} q(\boldsymbol{a}, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}) + \delta_{a_i, 1} \delta_{c_{i\cdot}, 0} q(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{c}) \right] \\
&\quad + \theta_B \sum_{j=1}^{L} \left[\sum_{i=1}^{K} \delta_{b_j + c_{\cdot j}, 1} \delta_{c_{ij}, 1} q(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{c} - \boldsymbol{e}_{ij}) + \delta_{b_j, 1} \delta_{c_{\cdot j}, 0} q(\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c}) \right] \\
&\quad + \rho \sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij} q(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}).
\end{aligned} \tag{2.5}$$

Relevant boundary conditions are $q(\boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}) = q(\boldsymbol{0}, \boldsymbol{e}_j, \boldsymbol{0}) = 1$ for all $i \in [K]$ and $j \in [L]$. For notational convenience, we deviate from Ethier and Griffiths (1990) and allow each summation to range over all allelic types. To be consistent, we define $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = 0$ whenever any entry in $\boldsymbol{a}$, $\boldsymbol{b}$, or $\boldsymbol{c}$ is negative.

For ease of discussion, we define the following terms:

DEFINITION 2.2 (Degree). *The degree of $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is defined to be $a + b + 2c$.*

DEFINITION 2.3 (Strictly recursive). *We say that a recursion relation is strictly recursive if it contains only a single term of the highest degree.*

Except in the special case $\rho = \infty$, a closed-form solution for $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is not known. Notice that the terms $q(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c} + \boldsymbol{e}_{ij})$ and $q(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij})$ on the right-hand side of (2.5) have the same degree as $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ on the left-hand side. Therefore, (2.5) is not strictly recursive. For each degree, we therefore need to solve a system of coupled equations, and this system grows very rapidly with n. For example, for a sample with $a = 0, b = 0$, and $c = 40$, computing $q(\mathbf{0}, \mathbf{0}, \boldsymbol{c})$ requires solving a system of more than $20,000$ coupled equations [Hudson (2001)]; this is around the limit of sample sizes that can be handled in a reasonable time. In the following section, we revisit the problem of obtaining a closed-form formula for $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ and obtain an asymptotic expansion for large ρ.

3. An asymptotic sampling formula for the two-locus case. For large ρ, our objective is to find an asymptotic expansion of the form

$$q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) + \frac{q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\rho} + \frac{q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\rho^2} + O\left(\frac{1}{\rho^3}\right), \tag{3.1}$$

where q_0, q_1, and q_2 are independent of ρ. Our closed-form formulas will be expressed using the following notation:

DEFINITION 3.1. *For a given multiplicity vector $\boldsymbol{a} = (a_1, \ldots, a_K)$ with $a = \sum_{i=1}^K a_i$, we define*

$$q^A(\boldsymbol{a}) = \left[\prod_{i=1}^K (a_i - 1)!\right] \frac{\theta_A^K}{(\theta_A)_a}. \tag{3.2}$$

Similarly, for a given multiplicity vector $\boldsymbol{b} = (b_1, \ldots, b_L)$ with $b = \sum_{i=1}^L b_i$, we define

$$q^B(\boldsymbol{b}) = \left[\prod_{j=1}^L (b_j - 1)!\right] \frac{\theta_B^L}{(\theta_B)_b}. \tag{3.3}$$

As discussed in Section 2.1, q^A (respectively, q^B) gives the probability of an ordered sample taken from locus A (respectively, B).

DEFINITION 3.2 (Marginal configuration). *We use $\boldsymbol{c}_A = (c_{i\cdot})_{i \in [K]}$ and $\boldsymbol{c}_B = (c_{\cdot j})_{j \in [L]}$ to denote the marginal sample configurations of $\boldsymbol{c}$ restricted to locus A and locus B, respectively.*

The leading order term $q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is equal to $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ when $\rho = \infty$, in which case the two loci are independent. Theorem 2.3 of Ethier and Griffiths (1990) states that $q_0(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{c}) = q^A(\boldsymbol{c}_A)q^B(\boldsymbol{c}_B)$. More generally, one can obtain the following result for the leading order contribution:

PROPOSITION 3.1. *In the asymptotic expansion (3.1) of the two-locus sampling formula, the zeroth order term $q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is given by*

$$q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = q^A(\boldsymbol{a} + \boldsymbol{c}_A)q^B(\boldsymbol{b} + \boldsymbol{c}_B). \tag{3.4}$$

Although this result is intuitively obvious, in Section 5.1 we provide a detailed new proof, since it well illustrates our general strategy. One of the main results of this paper is a closed-form formula for the next order term $q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. The case with $\boldsymbol{c} = \boldsymbol{0}$ admits a particularly simple solution:

LEMMA 3.1. *In the asymptotic expansion (3.1) of the two-locus sampling formula, the first order term satisfies*

$$q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0}) = 0$$

for arbitrary $\boldsymbol{a}$ and $\boldsymbol{b}$.

That $q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$ vanishes is not expected *a priori*. Below we shall see that $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0}) \neq 0$ in general. For an arbitrary configuration matrix $\boldsymbol{c}$ of non-negative integers, we obtain the following closed-form formula for $q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$:

THEOREM 3.1. *In the asymptotic expansion (3.1) of the two-locus sampling formula, the first order term $q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is given by*

$$\begin{aligned} q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \;=\; \frac{1}{2}\Bigg[& c(c-1)q^A(\boldsymbol{a} + \boldsymbol{c}_A)q^B(\boldsymbol{b} + \boldsymbol{c}_B) \\ & - q^B(\boldsymbol{b} + \boldsymbol{c}_B)\sum_{i=1}^{K} c_{i\cdot}(c_{i\cdot} - 1)q^A(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{e}_i) \\ & - q^A(\boldsymbol{a} + \boldsymbol{c}_A)\sum_{j=1}^{L} c_{\cdot j}(c_{\cdot j} - 1)q^B(\boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{e}_j) \\ & + \sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij}(c_{ij} - 1)q^A(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{e}_i)q^B(\boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{e}_j)\Bigg], \end{aligned} \tag{3.5}$$

for arbitrary configurations $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ of non-negative integers.

Lemma 3.1 is used in proving Theorem 3.1. A proof of Theorem 3.1 is provided in Section 5.2, while a proof of Lemma 3.1 is given in Section 5.3. In principle, similar arguments can be used to find the $(j+1)$th order term given the jth, although a general expression does not seem to be easy to obtain. In Section 5.4, we provide a proof of the following result for $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$:

THEOREM 3.2. *In the asymptotic expansion (3.1) of the two-locus sampling formula, the second order term $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is of the form*

$$q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = q_2(\boldsymbol{a} + \boldsymbol{c}_A, \boldsymbol{b} + \boldsymbol{c}_B, \boldsymbol{0}) + \sigma(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}), \tag{3.6}$$

where $\sigma(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is given by the analytic formula shown in Appendix A, and $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$ satisfies the following strict recursion:

$$\begin{aligned}
&[a(a+\theta_A-1)+b(b+\theta_B-1)]q_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0}) = \\
&\quad \sum_{i=1}^{K} a_i(a_i-1)q_2(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b},\boldsymbol{0}) + \sum_{j=1}^{L} b_j(b_j-1)q_2(\boldsymbol{a},\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{0}) \\
&\quad +\theta_A \sum_{i=1}^{K} \delta_{a_i,1} q_2(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b},\boldsymbol{0}) + \theta_B \sum_{j=1}^{L} \delta_{b_j,1} q_2(\boldsymbol{a},\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{0}) \\
&\quad +4\left[a\theta_A - (\theta_A+a-1)\sum_{i=1}^{K}\delta_{a_i,1}\right]\left[b\theta_B - (\theta_B+b-1)\sum_{j=1}^{L}\delta_{b_j,1}\right] q^A(\boldsymbol{a}) q^B(\boldsymbol{b}),
\end{aligned} \tag{3.7}$$

with boundary conditions $q_2(\boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}) = q_2(\boldsymbol{0}, \boldsymbol{e}_j, \boldsymbol{0}) = 0$ for all $i \in [K]$ and $j \in [L]$.

In contrast to $q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$ (c.f., Lemma 3.1), it turns out that $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$ does not vanish in general. We do not have an analytic solution for $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$, but note that (3.7) is strictly recursive and that it can be easily solved numerically using dynamic programming. Deriving an analytic expression for $\sigma(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ in (3.6) is a laborious task, as the long equation in Appendix A suggests. We have written a computer program to verify numerically that our analytic result is correct.

Numerical study (not shown) suggests that the relative contribution of $q_2(\boldsymbol{a} + \boldsymbol{c}_A, \boldsymbol{b} + \boldsymbol{c}_B, \boldsymbol{0})$ to $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is in most cases extremely small. For very large sample sizes, solving the dynamic programming problem for $q_2(\boldsymbol{a} + \boldsymbol{c}_A, \boldsymbol{b} + \boldsymbol{c}_B, \boldsymbol{0})$ could become prohibitive, and so we might consider simply omitting it for an analytic estimate of $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. For a given sample configuration it would therefore be desirable to estimate the contribution of this

term without having to calculate it directly. In Appendix B we obtain a close upper bound on $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$ which can be found using an $a \times b$ dynamic programming table. This is much simpler than the dynamic programming problem for (3.7), whose table has $[\prod_{i=1}^{K} a_i][\prod_{j=1}^{L} b_j]$ entries.

4. Joint distribution of the number of alleles at the two loci in a sample. Following the same strategy as in the previous section, we can obtain the asymptotic behavior of the joint distribution of the number of alleles observed at the two loci in a sample. To make explicit the dependence of these numbers on the sample size, write the number of alleles at locus A as $K_{a,b,c}$ and the number of alleles at locus B as $L_{a,b,c}$. Ethier and Griffiths (1990) proved that the probability $p(a,b,c;k,l) := \mathbb{P}(K_{a,b,c} = k, L_{a,b,c} = l)$ satisfies the recursion

$$\begin{aligned}
&[n(n-1) + \theta_A(a+c) + \theta_B(b+c) + \rho c]p(a,b,c;k,l) = \\
&\quad a(a-1+2c)p(a-1,b,c;k,l) + b(b-1+2c)p(a,b-1,c;k,l) \\
&\quad + c(c-1)p(a,b,c-1;k,l) + 2abp(a-1,b-1,c+1;k,l) \\
&\quad + \theta_A\left[ap(a-1,b,c;k-1,l) + cp(a,b+1,c-1;k-1,l)\right] \\
&\quad + \theta_B\left[bp(a,b-1,c;k,l-1) + cp(a+1,b,c-1;k,l-1)\right] \\
&\quad + \rho cp(a+1,b+1,c-1;k,l), \qquad (4.1)
\end{aligned}$$

where $p(a,b,c;k,l) = 0$ if $a < 0$, $b < 0$, $c < 0$, $k < 0$, $l < 0$, $a = b = c = 0$, or $k = l = 0$. Equation (4.1) has a unique solution satisfying the initial conditions

$$p(1,0,0;k,l) = \delta_{k,1}\delta_{l,0}, \quad p(0,1,0;k,l) = \delta_{k,0}\delta_{l,1},$$

for $k, l = 0, 1, \ldots, n$.

As with Golding's recursion, equation (4.1) can be solved numerically, but quickly becomes computationally intractable with growing n. The only exception is the special case of $\rho = \infty$, for which the distribution is given by the product of (2.4) for each locus. In what follows, we use the following notation in writing an asymptotic series for $p(a,b,c;k,l)$:

DEFINITION 4.1. *For loci A and B, respectively, we define the analogues of (2.4) as*

$$p^A(a;k) = \frac{s(a,k)\theta_A^k}{(\theta_A)_a}, \qquad (4.2)$$

and

$$p^B(b;l) = \frac{s(b,l)\theta_B^l}{(\theta_B)_b}, \qquad (4.3)$$

where $s(a,k)$ and $s(b,l)$ are the Stirling numbers of the first kind.

We pose the expansion

$$p(a,b,c;k,l) = p_0(a,b,c;k,l) + \frac{p_1(a,b,c;k,l)}{\rho} + O\left(\frac{1}{\rho^2}\right), \tag{4.4}$$

for large ρ. Then, in Section 5.5 we prove the following result for the zeroth order term:

PROPOSITION 4.1. *For an asymptotic expansion of the form (4.4) satisfying the recursion (4.1), $p_0(a,b,c;k,l)$ is given by*

$$p_0(a,b,c;k,l) = p^A(a+c;k)p^B(b+c;l). \tag{4.5}$$

Similar to Lemma 3.1, we obtain the following vanishing result for the first order term in the case of $c = 0$:

LEMMA 4.1. *For an asymptotic expansion of the form (4.4) satisfying the recursion (4.1), we have*

$$p_1(a,b,0;k,l) = 0.$$

Using this lemma, it is then possible to obtain the following result for an arbitrary c:

PROPOSITION 4.2. *For an asymptotic expansion of the form (4.4) satisfying the recursion (4.1), $p_1(a,b,c;k,l)$ is given by*

$$p_1(a,b,c;k,l) = \frac{c(c-1)}{2}\left[p^A(a+c;k) - p^A(a+c-1;k)\right] \times \left[p^B(b+c;l) - p^B(b+c-1;l)\right]. \tag{4.6}$$

Proofs of Proposition 4.2 and Lemma 4.1 are provided in Section 5.6 and Section 5.7, respectively.

5. Proofs of main results. In what follows, we provide proofs of the results mentioned in the previous two sections.

5.1. *Proof of Proposition 3.1.* First assume $c > 0$. Substitute the expansion (3.1) into Golding's recursion (2.5), divide by ρc and let $\rho \to \infty$. We are then left with

$$q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{c_{ij}}{c} q_0(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}). \tag{5.1}$$

Now, applying (5.1) repeatedly gives

$$q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \sum_{\text{orderings}} \frac{\prod_{(i,j)\in[K]\times[L]} c_{ij}!}{c!} q_0(\boldsymbol{a} + \boldsymbol{c}_A, \boldsymbol{b} + \boldsymbol{c}_B, \boldsymbol{0}),$$

where the summation is over all distinct orderings of the c gametes with multiplicity $\boldsymbol{c} = (c_{ij})$. There are $\frac{c!}{\prod_{(i,j)} c_{ij}!}$ such orderings and since the summand is independent of the ordering, we conclude

$$q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = q_0(\boldsymbol{a} + \boldsymbol{c}_A, \boldsymbol{b} + \boldsymbol{c}_B, \boldsymbol{0}). \tag{5.2}$$

Clearly, (5.2) also holds for $c = 0$. From a coalescent perspective, this equation tells us that any gamete with specified alleles (i.e., "carrying ancestral material") at both loci must undergo recombination instantaneously backwards in time.

Now, by substituting the asymptotic expansion (3.1) with $\boldsymbol{c} = \boldsymbol{0}$ into Golding's recursion (2.5) and letting $\rho \to \infty$, we obtain

$$\begin{aligned}
&[n(n-1) + \theta_A a + \theta_B b]\, q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0}) = \\
&\quad \sum_{i=1}^{K} a_i(a_i - 1) q_0(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{0}) + \sum_{j=1}^{L} b_j(b_j - 1) q_0(\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{0}) \\
&\quad + 2 \sum_{i=1}^{K} \sum_{j=1}^{L} a_i b_j q_0(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{e}_{ij}) \\
&\quad + \theta_A \sum_{i=1}^{K} \delta_{a_i,1} q_0(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{0}) + \theta_B \sum_{j=1}^{L} \delta_{b_j,1} q_0(\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{0}).
\end{aligned} \tag{5.3}$$

Equation (5.2) implies $q_0(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{e}_{ij}) = q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$, so with a bit of rearranging we are left with

$$\begin{aligned}
&[a(a + \theta_A - 1) + b(b + \theta_B - 1)]\, q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0}) = \\
&\quad \sum_{i=1}^{K} a_i(a_i - 1) q_0(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{0}) + \sum_{j=1}^{L} b_j(b_j - 1) q_0(\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{0}) \\
&\quad + \theta_A \sum_{i=1}^{K} \delta_{a_i,1} q_0(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{0}) + \theta_B \sum_{j=1}^{L} \delta_{b_j,1} q_0(\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{0}),
\end{aligned} \tag{5.4}$$

with boundary conditions $q_0(\boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}) = q_0(\boldsymbol{0}, \boldsymbol{e}_j, \boldsymbol{0}) = 1$ for all $i \in [K]$ and $j \in [L]$. Noting that (5.4) is the sum of two independent recursions of the form (2.3), one for each locus and each with appropriate boundary condition, we conclude that $q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$ is given by

$$q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0}) = q^A(\boldsymbol{a}) q^B(\boldsymbol{b}), \tag{5.5}$$

a product of two (ordered) ESFs. It is straightforward to verify that (5.5) satisfies (5.4). Finally, using (5.2) and (5.5), we arrive at (3.4). □

5.2. *Proof of Theorem 3.1.* First assume $c > 0$. Substitute the asymptotic expansion (3.1) into Golding's recursion (2.5), eliminate terms of order ρ by applying (5.1), and let $\rho \to \infty$. After applying (5.2) to the remaining terms and invoking (5.4), with some rearrangement we obtain

$$\begin{aligned} &cq_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) - \sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij} q_1(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}) = \\ &\quad c(c-1) q_0(\boldsymbol{a} + \boldsymbol{c}_A, \boldsymbol{b} + \boldsymbol{c}_B, \boldsymbol{0}) - \sum_{i=1}^{K} c_{i\cdot}(c_{i\cdot} - 1) q_0(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{c}_B, \boldsymbol{0}) \\ &\quad - \sum_{j=1}^{L} c_{\cdot j}(c_{\cdot j} - 1) q_0(\boldsymbol{a} + \boldsymbol{c}_A, \boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{e}_j, \boldsymbol{0}) \\ &\quad + \sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij}(c_{ij} - 1) q_0(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{e}_j, \boldsymbol{0}). \end{aligned}$$

Now, by utilizing (5.5), this can be written in the form

$$q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = f(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) + \sum_{i=1}^{K}\sum_{j=1}^{L} \frac{c_{ij}}{c} q_1(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}), \tag{5.6}$$

where

$$\begin{aligned} f(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \;:=\; & (c-1) q^A(\boldsymbol{a} + \boldsymbol{c}_A) q^B(\boldsymbol{b} + \boldsymbol{c}_B) \\ & - q^B(\boldsymbol{b} + \boldsymbol{c}_B) \sum_{i=1}^{K} \frac{c_{i\cdot}(c_{i\cdot} - 1)}{c} q^A(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{e}_i) \\ & - q^A(\boldsymbol{a} + \boldsymbol{c}_A) \sum_{j=1}^{L} \frac{c_{\cdot j}(c_{\cdot j} - 1)}{c} q^B(\boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{e}_j) \\ & + \sum_{i=1}^{K}\sum_{j=1}^{L} \frac{c_{ij}(c_{ij} - 1)}{c} q^A(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{e}_i) q^B(\boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{e}_j). \end{aligned} \tag{5.7}$$

Above we assumed $c > 0$. We define $f(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = 0$ if $\boldsymbol{c} = \boldsymbol{0}$. Iterating the recursion (5.6), we may write $q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ as

$$q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = f(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) + \sum_{i=1}^{K}\sum_{j=1}^{L} \frac{c_{ij}}{c} \Bigg[f(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}) + \sum_{i'=1}^{K}\sum_{j'=1}^{L} \frac{c_{i'j'} - \delta_{ii'}\delta_{jj'}}{c-1} q_1(\boldsymbol{a} + \boldsymbol{e}_i + \boldsymbol{e}_{i'}, \boldsymbol{b} + \boldsymbol{e}_j + \boldsymbol{e}_{j'}, \boldsymbol{c} - \boldsymbol{e}_{ij} - \boldsymbol{e}_{i'j'}) \Bigg].$$

Similarly, repeatedly iterating (5.6) yields

$$\begin{aligned} q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) &= q_1(\boldsymbol{a} + \boldsymbol{c}_A, \boldsymbol{b} + \boldsymbol{c}_B, \boldsymbol{0}) \\ &\quad + f(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \\ &\quad + \sum_{i_1 j_1} \frac{c_{i_1 j_1}}{c} f(\boldsymbol{a} + \boldsymbol{e}_{i_1}, \boldsymbol{b} + \boldsymbol{e}_{j_1}, \boldsymbol{c} - \boldsymbol{e}_{i_1 j_1}) \\ &\quad + \sum_{i_1 j_1, i_2 j_2} \frac{c_{i_1 j_1}}{c} \frac{c_{i_2 j_2} - \delta_{i_1 j_1, i_2 j_2}}{c-1} \\ &\qquad\qquad \times f(\boldsymbol{a} + \boldsymbol{e}_{i_1} + \boldsymbol{e}_{i_2}, \boldsymbol{b} + \boldsymbol{e}_{j_1} + \boldsymbol{e}_{j_2}, \boldsymbol{c} - \boldsymbol{e}_{i_1 j_1} - \boldsymbol{e}_{i_2 j_2}) \\ &\quad + \ldots + \sum_{i_1 j_1, \ldots, i_c j_c} \frac{\prod_{ij} c_{ij}!}{c!} f(\boldsymbol{a} + \boldsymbol{c}_A, \boldsymbol{b} + \boldsymbol{c}_B, \boldsymbol{0}). \end{aligned} \tag{5.8}$$

The key observation is that the right-hand side of (5.8) has a nice probabilistic interpretation which allows us to obtain a closed-form formula. To be more precise, consider the first summation

$$\sum_{i_1 j_1} \frac{c_{i_1 j_1}}{c} f(\boldsymbol{a} + \boldsymbol{e}_{i_1}, \boldsymbol{b} + \boldsymbol{e}_{j_1}, \boldsymbol{c} - \boldsymbol{e}_{i_1 j_1}).$$

For a fixed sample configuration $\boldsymbol{c}$, this can be interpreted as the sum over all possible ways of throwing away a gamete at random and calculating f based on the remaining subsample, which we will denote $\boldsymbol{c}^{(c-1)}$. Equivalently, it is the expected value of f with respect to subsampling without replacement $c-1$ of the gametes in $\boldsymbol{c}$. Write this as

$$\mathbb{E}[f(\boldsymbol{A}^{(c-1)}, \boldsymbol{B}^{(c-1)}, \boldsymbol{C}^{(c-1)})],$$

where $\boldsymbol{C}^{(c-1)}$ is the random subsample obtained by sampling without replacement $c-1$ gametes from $\boldsymbol{c}$, and $\boldsymbol{A}^{(c-1)} := \boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{C}_A^{(c-1)}$, $\boldsymbol{B}^{(c-1)} := \boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{C}_B^{(c-1)}$. Note that once the subsample $\boldsymbol{c}^{(c-1)}$ is obtained, then $\boldsymbol{a}^{(c-1)}$

and $\boldsymbol{b}^{(c-1)}$ are fully specified. More generally, consider the $(c-m)$th sum in (5.8). A particular term in the summation corresponds to an ordering of $c-m$ gametes in $\boldsymbol{c}$, which, when removed leave a subsample $\boldsymbol{c}^{(m)}$. With respect to this subsample, the summand is

$$\prod_{i=1}^{K}\prod_{j=1}^{L}\frac{c_{ij}!}{c_{ij}^{(m)}!}\frac{m!}{c!}f(\boldsymbol{a}^{(m)},\boldsymbol{b}^{(m)},\boldsymbol{c}^{(m)}),$$

and for each such subsample $\boldsymbol{c}^{(m)}$ there are $\binom{c-m}{\boldsymbol{c}-\boldsymbol{c}^{(m)}}$ distinct orderings of the remaining types in $\boldsymbol{c}$, with each ordering contributing the same amount to the sum. Here, $\binom{c-m}{\boldsymbol{c}-\boldsymbol{c}^{(m)}}$ denotes the multinomial coefficient:

$$\binom{c-m}{\boldsymbol{c}-\boldsymbol{c}^{(m)}}=\frac{(c-m)!}{\prod_{i=1}^{K}\prod_{j=1}^{L}(c_{ij}-c_{ij}^{(m)})!}.$$

Gathering identical terms, the $(c-m)$th sum in (5.8) can therefore be written over all distinct subsamples of $\boldsymbol{c}$ of size m:

$$\begin{aligned}
&\sum_{\boldsymbol{c}^{(m)}}\binom{c-m}{\boldsymbol{c}-\boldsymbol{c}^{(m)}}\prod_{i=1}^{K}\prod_{j=1}^{L}\frac{c_{ij}!}{c_{ij}^{(m)}!}\frac{m!}{c!}f(\boldsymbol{a}^{(m)},\boldsymbol{b}^{(m)},\boldsymbol{c}^{(m)})\\
&\quad=\sum_{\boldsymbol{c}^{(m)}}\frac{1}{\binom{c}{m}}\prod_{i=1}^{K}\prod_{j=1}^{L}\binom{c_{ij}}{c_{ij}^{(m)}}f(\boldsymbol{a}^{(m)},\boldsymbol{b}^{(m)},\boldsymbol{c}^{(m)})\\
&\quad=\mathbb{E}[f(\boldsymbol{A}^{(m)},\boldsymbol{B}^{(m)},\boldsymbol{C}^{(m)})],
\end{aligned}$$

where, for a fixed m, $\boldsymbol{C}^{(m)}=(C_{ij}^{(m)})$ is a multivariate hypergeometric$(c,\boldsymbol{c},m)$ random variable; that is,

$$\mathbb{P}\left(\cap_{(i,j)\in[K]\times[L]}\left[C_{ij}^{(m)}=c_{ij}^{(m)}\right]\right)=\frac{1}{\binom{c}{m}}\prod_{(i,j)\in[K]\times[L]}\binom{c_{ij}}{c_{ij}^{(m)}}.$$

Furthermore, marginally we have

$$\begin{aligned}
C_{ij}^{(m)}&\sim\text{hypergeometric}(c,c_{ij},m),\\
C_{i\cdot}^{(m)}&\sim\text{hypergeometric}(c,c_{i\cdot},m),\\
C_{\cdot j}^{(m)}&\sim\text{hypergeometric}(c,c_{\cdot j},m).
\end{aligned}$$

In summary, (5.8) can be written as

$$q_1(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})=q_1(\boldsymbol{a}+\boldsymbol{c}_A,\boldsymbol{b}+\boldsymbol{c}_B,\boldsymbol{0})+\sum_{m=1}^{c}\mathbb{E}[f(\boldsymbol{A}^{(m)},\boldsymbol{B}^{(m)},\boldsymbol{C}^{(m)})]. \tag{5.9}$$

According to Lemma 3.1, the first term $q_1(\boldsymbol{a}+\boldsymbol{c}_A, \boldsymbol{b}+\boldsymbol{c}_B, \boldsymbol{0})$ vanishes, so we are left with

$$q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \sum_{m=1}^{c} \mathbb{E}[f(\boldsymbol{A}^{(m)}, \boldsymbol{B}^{(m)}, \boldsymbol{C}^{(m)})]. \tag{5.10}$$

Finally, since $\boldsymbol{A}^{(m)} + \boldsymbol{C}_A^{(m)} = \boldsymbol{a} + \boldsymbol{c}_A$ and $\boldsymbol{B}^{(m)} + \boldsymbol{C}_B^{(m)} = \boldsymbol{b} + \boldsymbol{c}_B$, (5.7) and (5.10) together imply

$$\begin{aligned}
q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = & \\
& \sum_{m=1}^{c} \Bigg[(m-1) q^A(\boldsymbol{a}+\boldsymbol{c}_A) q^B(\boldsymbol{b}+\boldsymbol{c}_B) \\
& - q^B(\boldsymbol{b}+\boldsymbol{c}_B) \frac{1}{m} \sum_{i=1}^{K} \mathbb{E}[C_{i\cdot}^{(m)}(C_{i\cdot}^{(m)} - 1)] q^A(\boldsymbol{a}+\boldsymbol{c}_A - \boldsymbol{e}_i) \\
& - q^A(\boldsymbol{a}+\boldsymbol{c}_A) \frac{1}{m} \sum_{j=1}^{L} \mathbb{E}[C_{\cdot j}^{(m)}(C_{\cdot j}^{(m)} - 1)] q^B(\boldsymbol{b}+\boldsymbol{c}_B - \boldsymbol{e}_j) \\
& + \frac{1}{m} \sum_{i=1}^{K} \sum_{j=1}^{L} \mathbb{E}[C_{ij}^{(m)}(C_{ij}^{(m)} - 1)] q^A(\boldsymbol{a}+\boldsymbol{c}_A - \boldsymbol{e}_i) q^B(\boldsymbol{b}+\boldsymbol{c}_B - \boldsymbol{e}_j) \Bigg].
\end{aligned}$$

The moments in this equation are easy to compute and one can sum them over m to obtain the desired result (3.5). □

5.3. *Proof of Lemma 3.1.* First note that for any sample $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ and any subsample of the form $(\boldsymbol{a}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{c}^{(1)})$, we have $f(\boldsymbol{a}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{c}^{(1)}) = 0$, since every term on right-hand side of (5.7) has a vanishing coefficient. So, equation (5.9) implies

$$q_1(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{e}_{ij}) = q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0}), \tag{5.11}$$

for any $(i, j) \in [K] \times [L]$. Now, substitute the asymptotic expansion (3.1) with $\boldsymbol{c} = \boldsymbol{0}$ into Golding's recursion (2.5). Note that terms of order ρ are absent since $\boldsymbol{c} = \boldsymbol{0}$. Eliminate terms with coefficients independent of ρ by applying (5.3), multiply both sides of the recursion by ρ, and let $\rho \to \infty$ to

obtain the following:

$$\begin{aligned}
&[n(n-1)+\theta_A a+\theta_B b]q_1(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0}) = \\
&\quad \sum_{i=1}^{K} a_i(a_i-1)q_1(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b},\boldsymbol{0}) + \sum_{j=1}^{L} b_j(b_j-1)q_1(\boldsymbol{a},\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{0}) \\
&\quad + 2\sum_{i=1}^{K}\sum_{j=1}^{L} a_i b_j q_1(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{e}_{ij}) \\
&\quad + \theta_A \sum_{i=1}^{K} \delta_{a_i,1} q_1(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b},\boldsymbol{0}) + \theta_B \sum_{j=1}^{L} \delta_{b_j,1} q_1(\boldsymbol{a},\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{0}),
\end{aligned}$$

with boundary conditions $q_1(\boldsymbol{e}_i,\boldsymbol{0},\boldsymbol{0}) = q_1(\boldsymbol{0},\boldsymbol{e}_j,\boldsymbol{0}) = 0$ for all $i \in [K]$ and $j \in [L]$. This equation can be made strictly recursive by applying (5.11) to $q_1(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{e}_{ij})$. It therefore follows from the boundary conditions (for example by induction) that $q_1(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0}) = 0$. □

5.4. *Proof of Theorem 3.2.* Here, we provide only an outline of a proof; details are similar to the proof of Theorem 3.1. Substitute the asymptotic expansion (3.1) into Golding's recursion (2.5), eliminate terms with coefficients proportional to ρ or independent of ρ. Then, multiply both sides of the recursion by ρ and let $\rho \to \infty$ to obtain

$$\begin{aligned}
&cq_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) - \sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij} q_2(\boldsymbol{a}+\boldsymbol{e}_i,\boldsymbol{b}+\boldsymbol{e}_j,\boldsymbol{c}-\boldsymbol{e}_{ij}) = \\
&\quad \sum_{i=1}^{K} a_i(a_i-1+2c_{i\cdot})q_1(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b},\boldsymbol{c}) + \sum_{j=1}^{L} b_j(b_j-1+2c_{\cdot j})q_1(\boldsymbol{a},\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{c}) \\
&\quad + \sum_{i=1}^{K}\sum_{j=1}^{L} [c_{ij}(c_{ij}-1)q_1(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}-\boldsymbol{e}_{ij}) + 2a_i b_j q_1(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{c}+\boldsymbol{e}_{ij})] \\
&\quad + \theta_A \sum_{i=1}^{K} \left[\sum_{j=1}^{L} \delta_{a_i+c_{i\cdot},1}\delta_{c_{ij},1} q_1(\boldsymbol{a},\boldsymbol{b}+\boldsymbol{e}_j,\boldsymbol{c}-\boldsymbol{e}_{ij}) + \delta_{a_i,1}\delta_{c_{i\cdot},0} q_1(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b},\boldsymbol{c}) \right] \\
&\quad + \theta_B \sum_{j=1}^{L} \left[\sum_{i=1}^{K} \delta_{b_j+c_{\cdot j},1}\delta_{c_{ij},1} q_1(\boldsymbol{a}+\boldsymbol{e}_i,\boldsymbol{b},\boldsymbol{c}-\boldsymbol{e}_{ij}) + \delta_{b_j,1}\delta_{c_{\cdot j},0} q_1(\boldsymbol{a},\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{c}) \right] \\
&\quad - [n(n-1)+\theta_A(a+c)+\theta_B(b+c)]q_1(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})
\end{aligned} \tag{5.12}$$

By substituting our expression (3.5) for $q_1(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$, the right-hand side can be expressed as a function $g(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$ which is completely known but rather

cumbersome to write down. As in the proof of Theorem 3.1, the same 'unwrapping' maneuver can be applied to rearrange (5.12) into the form (3.6), where

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

This time $\mathbb{E}[g(\boldsymbol{A}^{(m)},\boldsymbol{B}^{(m)},\boldsymbol{C}^{(m)})]$ is a function of fourth-order moments of the multivariate hypergeometric distribution. The formula shown in Appendix A is obtained by evaluating the expectations and summing over m.

We now show that $q_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0})$ satisfies the recursion shown in (3.7). We will use the fact that for a sample $(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{e}_{ij})$, we have

$$\begin{aligned} g(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{e}_{ij}) &= 2(a-1)(b-1)q^A(\boldsymbol{a})q^B(\boldsymbol{b}) \\ &\quad - 2(b-1)(a_i-1)q^A(\boldsymbol{a}-\boldsymbol{e}_i)q^B(\boldsymbol{b}) \\ &\quad - 2(a-1)(b_j-1)q^A(\boldsymbol{a})q^B(\boldsymbol{b}-\boldsymbol{e}_j) \\ &\quad + 2(a_i-1)(b_j-1)q^A(\boldsymbol{a}-\boldsymbol{e}_i)q^B(\boldsymbol{b}-\boldsymbol{e}_j). \end{aligned} \tag{5.13}$$

For an arbitrary $\boldsymbol{c}$, $g(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$ is much more complicated.

Now, one can adopt the approach used in the proof of Lemma 3.1 to obtain a strict recursion for $q_2(\boldsymbol{a}+\boldsymbol{c}_A,\boldsymbol{b}+\boldsymbol{c}_B,\boldsymbol{0})$. First, note that (3.6) and (5.13) imply

$$\begin{aligned} q_2(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{e}_{ij}) &= q_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0}) + \mathbb{E}[g((\boldsymbol{A}-\boldsymbol{e}_i)^{(1)},(\boldsymbol{B}-\boldsymbol{e}_j)^{(1)},\boldsymbol{e}_{ij}^{(1)})] \\ &= q_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0}) + g(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{e}_{ij}) \\ &= q_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0}) + 2(a-1)(b-1)q^A(\boldsymbol{a})q^B(\boldsymbol{b}) \\ &\quad - 2(b-1)(a_i-1)q^A(\boldsymbol{a}-\boldsymbol{e}_i)q^B(\boldsymbol{b}) \\ &\quad - 2(a-1)(b_j-1)q^A(\boldsymbol{a})q^B(\boldsymbol{b}-\boldsymbol{e}_j) \\ &\quad + 2(a_i-1)(b_j-1)q^A(\boldsymbol{a}-\boldsymbol{e}_i)q^B(\boldsymbol{b}-\boldsymbol{e}_j). \end{aligned} \tag{5.14}$$

As before, substitute the asymptotic expansion (3.1) for $\boldsymbol{c}=\boldsymbol{0}$ into Golding's recursion (2.5), eliminate terms with coefficients independent of ρ or proportional to ρ^{-1}, and let $\rho\to\infty$ to obtain

$$\begin{aligned} &[n(n-1)+\theta_A a+\theta_B b]q_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0}) = \\ &\quad \sum_{i=1}^{K} a_i(a_i-1)q_2(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b},\boldsymbol{0}) + \sum_{j=1}^{L} b_j(b_j-1)q_2(\boldsymbol{a},\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{0}) \\ &\quad + 2\sum_{i=1}^{K}\sum_{j=1}^{L} a_i b_j q_2(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{e}_{ij}) \\ &\quad + \theta_A \sum_{i=1}^{K} \delta_{a_i,1} q_2(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b},\boldsymbol{0}) + \theta_B \sum_{j=1}^{L} \delta_{b_j,1} q_2(\boldsymbol{a},\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{0}), \end{aligned}$$

with boundary conditions $q_2(\boldsymbol{e}_i, \mathbf{0}, \mathbf{0}) = q_2(\mathbf{0}, \boldsymbol{e}_j, \mathbf{0}) = 0$ for all $i \in [K]$ and $j \in [L]$. This equation can be made strictly recursive by applying (5.14) to $q_2(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{e}_{ij})$. After some simplification, this leads to the recursion (3.7). □

5.5. *Proof of Proposition 4.1.* The proof is similar to the proof of Proposition 3.1, working with the system (4.1) rather than Golding's recursion (2.5). First assume $c > 0$. Substitute the expansion (4.4) into the recursion (4.1), divide by ρc and let $\rho \to \infty$. We are left with

$$p_0(a, b, c; k, l) = p_0(a + 1, b + 1, c - 1; k, l),$$

which implies

$$p_0(a, b, c; k, l) = p_0(a + c, b + c, 0; k, l). \tag{5.15}$$

Clearly, (5.15) also holds for $c = 0$.

Now, by substituting the asymptotic expansion (4.4) with $c = 0$ into (4.1) and letting $\rho \to \infty$, we obtain

$$\begin{aligned}
&[n(n-1) + \theta_A a + \theta_B b]\, p_0(a, b, 0; k, l) = \\
&\quad a(a-1)p_0(a-1, b, 0; k, l) + b(b-1)p_0(a, b-1, c; k, l) \\
&\quad + 2ab p_0(a-1, b-1, 1; k, l) \\
&\quad + \theta_A a p_0(a-1, b, 0; k-1, l) + \theta_B b p_0(a, b-1, 0; k, l-1).
\end{aligned} \tag{5.16}$$

After invoking (5.15) on $p_0(a - 1, b - 1, 1; k, l)$ and rearranging, we are left with

$$\begin{aligned}
&[a(a + \theta_A - 1) + b(b + \theta_B - 1)]\, p_0(a, b, 0; k, l) = \\
&\quad a(a-1)p_0(a-1, b, 0; k, l) + b(b-1)p_0(a, b-1, 0; k, l) \\
&\quad + \theta_A a p_0(a-1, b, 0; k-1, l) + \theta_B b p_0(a, b-1, 0; k, l-1),
\end{aligned} \tag{5.17}$$

with boundary conditions $p_0(1, 0, 0; k, l) = \delta_{k,1}\delta_{l,0}$, and $p_0(0, 1, 0; k, l) = \delta_{k,0}\delta_{l,1}$. Equation (5.17) can be expressed as a linear sum of two independent recursions:

$$\begin{aligned}
(a - 1 + \theta_A)p_0^A(a; k) &= (a-1)p_0^A(a-1; k) + \theta_A p_0^A(a-1; k-1), \\
(b - 1 + \theta_B)p_0^B(b; l) &= (b-1)p_0^B(b-1; l) + \theta_B p_0^B(b-1; l-1),
\end{aligned}$$

with respective boundary conditions $p_0^A(1; k) = \delta_{k,1}$ and $p_0^B(1; l) = \delta_{l,1}$. These recursions are precisely those considered by Ewens (1972, eq. 21), with respective solutions (4.2) and (4.3). Hence, $p_0^A(a; k) = p^A(a; k)$ and $p_0^B(b; l) = p^B(b; l)$, and it is straightforward to verify that $p^A(a; k)p^B(b; l)$ satisfies (5.17). Substituting this solution into (5.15), we arrive at (4.5), as required. □

5.6. *Proof of Proposition 4.2.* First assume $c > 0$. Substitute the asymptotic expansion (4.4) into the recursion (4.1), eliminate terms with coefficients linear in ρ by applying (5.15), and let $\rho \to \infty$. After applying (5.15) to the remaining terms and invoking (5.17), with some rearrangement we obtain

$$\begin{aligned} &p_1(a,b,c;k,l) - p_1(a+1,b+1,c-1;k,l) = \\ &\quad (c-1)[p_0(a+c,b+c,0;k,l) - p_0(a+c-1,b+c,0;k,l) \\ &\quad - p_0(a+c,b+c-1,0;k,l) + p_0(a+c-1,b+c-1,0;k,l)]. \end{aligned} \tag{5.18}$$

Applying the recursion repeatedly, this becomes

$$\begin{aligned} p_1(a,b,c;k,l) \;=\; & p_1(a+c,b+c,0;k,l) + [p_0(a+c,b+c,0;k,l) \\ & - p_0(a+c-1,b+c,0;k,l) - p_0(a+c,b+c-1,0;k,l) \\ & + p_0(a+c-1,b+c-1,0;k,l)] \sum_{m=0}^{c-1} (c-1-m). \end{aligned} \tag{5.19}$$

According to Lemma 4.1, the first term $p_1(a+c,b+c,0;k,l)$ vanishes. Hence, since $p_0(a,b,c;k,l)$ is given by (4.5), the right-hand side of (5.19) is fully known. With some rearrangement we are left with (4.6). □

5.7. *Proof of Lemma 4.1.* First note that (5.18) implies

$$p_1(a-1,b-1,1;k,l) = p_1(a,b,0;k,l). \tag{5.20}$$

Now, substitute the asymptotic expansion (4.4) with c = 0 into (4.1), eliminate leading order terms by applying (5.16), and let $\rho \to \infty$. The result is made strictly recursive by invoking (5.20), and we obtain

$$\begin{aligned} &[a(a+\theta_A-1) + b(b+\theta_B-1)]p_1(a,b,0;k,l) = \\ &\quad a(a-1)p_1(a-1,b,0;k,l) + b(b-1)p_1(a,b-1,0;k,l) \\ &\quad + \theta_A a p_1(a-1,b,0;k-1,l) + \theta_B b p_1(a,b-1,0;k,l-1), \end{aligned}$$

with boundary conditions $p_1(1,0,0;k,l) = p_1(0,1,0;k,l) = 0$, for $k,l = 0,\ldots,n$. It therefore follows (for example by induction) that $p_1(a,b,0;k,l) = 0$. □

Acknowledgments. We thank Robert C. Griffiths, Charles H. Langley, and Joshua Paul for useful discussions.

APPENDIX A: EXPRESSION FOR $\sigma(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$

We use Q^A to denote $q^A(\boldsymbol{a}+\boldsymbol{c}_A)$, Q^A_i to denote $q^A(\boldsymbol{a}+\boldsymbol{c}_A-\boldsymbol{e}_i)$, Q^A_{ik} to denote $q^A(\boldsymbol{a}+\boldsymbol{c}_A-\boldsymbol{e}_i-\boldsymbol{e}_k)$, and so on. Then,

$$
\begin{aligned}
\sigma(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = \\
&\frac{c}{3}\left[\frac{(c-1)(c+1)(3c-2)}{8}+(c-1)(3a+3b+2c-1)+6ab\right]Q^AQ^B \\
&-\frac{\theta_A(c-1)}{2}\sum_{i=1}^{K}\delta_{a_i,0}\delta_{c_{i\cdot},1}Q^A_iQ^B-\frac{\theta_B(c-1)}{2}\sum_{j=1}^{L}\delta_{b_j,0}\delta_{c_{\cdot j},1}Q^AQ^B_j \\
&+\sum_{i=1}^{K}\left[\frac{\theta_A-c(c-3)+2a+4b-4}{4}c_{i\cdot}(c_{i\cdot}-1)\right. \\
&\qquad\left.-(2b+c-1)c_{i\cdot}(a_i+c_{i\cdot}-1)\right]Q^A_iQ^B \\
&+\frac{1}{2}\sum_{i=1}^{K}\left[\frac{\theta_A}{2}\delta_{c_{i\cdot},2}+\frac{5-6a_i-4c_{i\cdot}}{6}\right]c_{i\cdot}(c_{i\cdot}-1)Q^A_{ii}Q^B \\
&+\sum_{j=1}^{L}\left[\frac{\theta_B-c(c-3)+2b+4a-4}{4}c_{\cdot j}(c_{\cdot j}-1)\right. \\
&\qquad\left.-(2a+c-1)c_{\cdot j}(b_j+c_{\cdot j}-1)\right]Q^AQ^B_j \\
&+\frac{1}{2}\sum_{j=1}^{L}\left[\frac{\theta_B}{2}\delta_{c_{\cdot j},2}+\frac{5-6b_j-4c_{\cdot j}}{6}\right]c_{\cdot j}(c_{\cdot j}-1)Q^AQ^B_{jj} \\
&+\sum_{i,k=1}^{K}\frac{c_{i\cdot}(c_{i\cdot}-1)c_{k\cdot}(c_{k\cdot}-1)}{8}Q^A_{ik}Q^B+\sum_{j,l=1}^{L}\frac{c_{\cdot j}(c_{\cdot j}-1)c_{\cdot l}(c_{\cdot l}-1)}{8}Q^AQ^B_{jl} \\
&-\frac{\theta_A+\theta_B-c(c-5)+2a+2b-4}{4}\sum_{i=1}^{K}\sum_{j=1}^{L}c_{ij}(c_{ij}-1)Q^A_iQ^B_j \\
&+\sum_{i=1}^{K}\sum_{j=1}^{L}\left[\frac{c_{i\cdot}(c_{i\cdot}-1)c_{\cdot j}(c_{\cdot j}-1)}{4}+\frac{c_{ij}(c_{ij}+1-2c_{i\cdot}+2c_{i\cdot}c_{\cdot j}-2c_{\cdot j})}{2}\right. \\
&\qquad+c_{ij}b_j(c_{i\cdot}-1)+c_{ij}a_i(c_{\cdot j}-1)+2a_ib_jc_{ij} \\
&\qquad\left.+\frac{\theta_B}{2}\delta_{b_j,0}\delta_{c_{\cdot j},1}\delta_{c_{ij},1}(c_{i\cdot}-1)+\frac{\theta_A}{2}\delta_{a_i,0}\delta_{c_{i\cdot},1}\delta_{c_{ij},1}(c_{\cdot j}-1)\right]Q^A_iQ^B_j
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{2}\sum_{i=1}^{K}\left[a_i + c_{i\cdot} - 1 - \frac{\theta_A}{2}\delta_{a_i+c_{i\cdot},2}\right]\sum_{j=1}^{L} c_{ij}(c_{ij}-1)Q_{ii}^{A}Q_{j}^{B} \\
&+ \frac{1}{2}\sum_{j=1}^{L}\left[b_j + c_{\cdot j} - 1 - \frac{\theta_B}{2}\delta_{b_j+c_{\cdot j},2}\right]\sum_{i=1}^{K} c_{ij}(c_{ij}-1)Q_{i}^{A}Q_{jj}^{B} \\
&- \frac{1}{4}\sum_{i=1}^{K}\sum_{j=1}^{L}\sum_{k=1}^{K} c_{ij}(c_{ij}-1)c_{k\cdot}(c_{k\cdot}-1)Q_{ik}^{A}Q_{j}^{B} \\
&- \frac{1}{4}\sum_{i=1}^{K}\sum_{j=1}^{L}\sum_{l=1}^{L} c_{ij}(c_{ij}-1)c_{\cdot l}(c_{\cdot l}-1)Q_{i}^{A}Q_{jl}^{B} \\
&+ \frac{1}{8}\sum_{i=1}^{K}\sum_{j=1}^{L}\sum_{k=1}^{K}\sum_{l=1}^{L} c_{ij}(c_{ij}-1)c_{kl}(c_{kl}-1)Q_{ik}^{A}Q_{jl}^{B} \\
&- \frac{1}{12}\sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij}(c_{ij}-1)(2c_{ij}-1)Q_{ii}^{A}Q_{jj}^{B}.
\end{aligned}
$$

To check the correctness of the above expression, we also solved the recursion (5.12) numerically for all sample configurations of sizes $n = 10, 20,$ and 30 (with $K, L \leq 2$), and confirmed that the above analytic expression agreed in all cases. We also implemented a *Mathematica* program to solve $q(a, b, c)$ exactly in the special case $K = L = 1$. The program can return series expansions in terms of ρ^{-1} as $\rho \to \infty$ which are symbolic in θ_A and θ_B. We could then compare the first three terms against q_0, q_1, and q_2, for various sample configurations (a, b, c).

APPENDIX B: AN UPPER BOUND ON $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$

To obtain an upper bound on $q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$, we can exploit the linearity in (3.7). Let

$$
\begin{aligned}
s_1(\boldsymbol{a}, \boldsymbol{b}) &:= ab\theta_A\theta_B, \\
s_2(\boldsymbol{a}, \boldsymbol{b}) &:= -b\theta_B(\theta_A + a - 1)\sum_i \delta_{a_i,1}, \\
s_3(\boldsymbol{a}, \boldsymbol{b}) &:= -a\theta_A(\theta_B + b - 1)\sum_j \delta_{b_j,1}, \\
s_4(\boldsymbol{a}, \boldsymbol{b}) &:= (\theta_A + a - 1)(\theta_B + b - 1)\sum_i \delta_{a_i,1}\sum_j \delta_{b_j,1}.
\end{aligned} \tag{B.1}
$$

Now note that if $r_k(\boldsymbol{a},\boldsymbol{b})$ $(k=1,\dots,4)$ satisfies

$$\begin{aligned}[a(a+\theta_A-1)+b(b+\theta_B-1)]r_k(\boldsymbol{a},\boldsymbol{b}) &= \\ \sum_{i=1}^{K} a_i(a_i-1+\theta_A\delta_{a_i,1})r_k(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}) &+ \sum_{j=1}^{L} b_j(b_j-1+\theta_B\delta_{b_j,1})r_k(\boldsymbol{a},\boldsymbol{b}-\boldsymbol{e}_j) \\ &+ 4q^A(\boldsymbol{a})q^B(\boldsymbol{b})s_k(\boldsymbol{a},\boldsymbol{b}), \end{aligned} \tag{B.2}$$

then $q_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0}) = r_1(\boldsymbol{a},\boldsymbol{b})+r_2(\boldsymbol{a},\boldsymbol{b})+r_3(\boldsymbol{a},\boldsymbol{b})+r_4(\boldsymbol{a},\boldsymbol{b})$. We can then bound each $r_k(\boldsymbol{a},\boldsymbol{b})$ individually and sum them for bounds on $q_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0})$.

By iteratively unwrapping (B.2), each term that contributes to $r_k(\boldsymbol{a},\boldsymbol{b})$ can be mapped bijectively to a sequence of subsamples of $(\boldsymbol{a},\boldsymbol{b})$ with the following constraints:

1. The first term in the sequence is $(\boldsymbol{a},\boldsymbol{b})$,
2. The mth term in the sequence has total sample size $a+b+1-m$,
3. The sequence has length m such that $1 \le m \le a+b$.

For example, the sequence $((\boldsymbol{a},\boldsymbol{b}))$ is associated with the contribution

$$\frac{4q^A(\boldsymbol{a})q^B(\boldsymbol{b})s_k(\boldsymbol{a},\boldsymbol{b})}{a(a+\theta_A-1)+b(b+\theta_B-1)},$$

while the sequence $((\boldsymbol{a},\boldsymbol{b}),(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}))$ is associated with the contribution

$$\frac{a_i(a_i-1+\theta_A\delta_{a_i,1})}{[a(a+\theta_A-1)+b(b+\theta_B-1)]}\cdot\frac{4q^A(\boldsymbol{a}-\boldsymbol{e}_i)q^B(\boldsymbol{b})s_k(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b})}{(a-1)(a+\theta_A-2)+b(b+\theta_B-1)},$$

and so on. Call a sequence satisfying the constraints 1–3 a *path*. Denote a subsample of $(\boldsymbol{a},\boldsymbol{b})$ of size (m_A,m_B) by $(\boldsymbol{a}^{(m_A)},\boldsymbol{b}^{(m_B)})$. To see how to obtain a bound on $r_k(\boldsymbol{a},\boldsymbol{b})$, let us consider the contribution from a single example path,

$$\begin{aligned}&((\boldsymbol{a},\boldsymbol{b}),(\boldsymbol{a}^{(a-1)},\boldsymbol{b}),(\boldsymbol{a}^{(a-1)},\boldsymbol{b}^{(b-1)}),(\boldsymbol{a}^{(a-2)},\boldsymbol{b}^{(b-1)})) \\ &\qquad := ((\boldsymbol{a},\boldsymbol{b}),(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}),(\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j),(\boldsymbol{a}-2\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j)).\end{aligned}$$

If $a_i = 2$ and $b_j > 1$, the contribution from this path to $r_k(\boldsymbol{a},\boldsymbol{b})$ is

$$\begin{aligned}&\left[\frac{a_i(a_i-1)}{D(a,b)}\cdot\frac{b_j(b_j-1)}{D(a-1,b)}\cdot\frac{\theta_A}{D(a-1,b-1)}\right] \\ &\qquad\qquad\times\frac{4q^A(\boldsymbol{a}^{(a-2)})q^B(\boldsymbol{b}^{(b-1)})}{D(a-2,b-1)}s_k(\boldsymbol{a}^{(a-2)},\boldsymbol{b}^{(b-1)}),\end{aligned} \tag{B.3}$$

where $D(a,b) := a(a-1+\theta_A)+b(b-1+\theta_B)$. Evaluating $r_k(\boldsymbol{a},\boldsymbol{b})$ completely is problematic because of terms like

$$[D(a,b)D(a-1,b)D(a-1,b-1)D(a-2,b-1)]^{-1},$$

which differ for each path. Suppose for now we have obtained an upper bound $U_{a-2,b-1}$ on this term which depends only on the subsample size $(a-2,b-1)$ for the final term in the path, and not on the complete path. Then the contribution (B.3) can be rewritten

$$4\left[\frac{a(a-1+\theta_A)}{D(a,b)}\cdot\frac{b(b-1+\theta_B)}{D(a-1,b)}\cdot\frac{(a-1)(a-2+\theta_A)}{D(a-1,b-1)}\cdot\frac{1}{D(a-2,b-1)}\right] \times\left[\frac{a_i}{a}\frac{b_j}{b}\frac{1}{a-1}\right]q^A(\boldsymbol{a})q^B(\boldsymbol{b})s_k(\boldsymbol{a}^{(a-2)},\boldsymbol{b}^{(b-1)})$$

$$\text{(B.4)}\quad \leq 4U_{a-2,b-1}\frac{a!}{(a-2)!}\frac{b!}{(b-1)!}\frac{(\theta_A)_a}{(\theta_A)_{a-2}}\frac{(\theta_B)_b}{(\theta_B)_{b-1}} \times\left[\frac{a_i}{a}\frac{b_j}{b}\frac{1}{a-1}\right]q^A(\boldsymbol{a})q^B(\boldsymbol{b})s_k(\boldsymbol{a}^{(a-2)},\boldsymbol{b}^{(b-1)}).$$

The reason for this approach is that one can now sum (B.4):

(I) First, over all paths to this subsample, then
(II) over all paths ending at a subsample of size $(a-2,b-1)$, and finally
(III) over all subsample sizes.

There are $\binom{3}{2}=3$ paths to this subsample. Hence, summing the bound (B.4) over (I), a bound on the total contribution of paths to this subsample is

$$\text{(B.5)}\quad 4U_{a-2,b-1}\frac{a!}{(a-2)!}\frac{b!}{(b-1)!}\frac{(\theta_A)_a}{(\theta_A)_{a-2}}\frac{(\theta_B)_b}{(\theta_B)_{b-1}} \times\binom{3}{2}\mathbb{P}(\boldsymbol{A}^{(a-2)}=\boldsymbol{a}^{(a-2)})\mathbb{P}(\boldsymbol{B}^{(b-1)}=\boldsymbol{b}^{(b-1)})q^A(\boldsymbol{a})q^B(\boldsymbol{b})s_k(\boldsymbol{a}^{(a-2)},\boldsymbol{b}^{(b-1)}),$$

where the random variable $\boldsymbol{A}^{(m_A)}$ is distributed according to a multivariate hypergeometric$(a,\boldsymbol{a},m_A)$ distribution, and the random variable $\boldsymbol{B}^{(m_B)}$ is distributed according to a multivariate hypergeometric$(b,\boldsymbol{b},m_B)$ distribution.

Next, sum this bound over (II). Thus, a bound on the contribution from

all paths ending at a subsample of size $(a-2, b-1)$ is

$$4U_{a-2,b-1}\frac{a!}{(a-2)!}\frac{b!}{(b-1)!}\frac{(\theta_A)_a}{(\theta_A)_{a-2}}\frac{(\theta_B)_b}{(\theta_B)_{b-1}} \times \binom{3}{2} q^A(\boldsymbol{a})q^B(\boldsymbol{b})\mathbb{E}[s_k(\boldsymbol{A}^{(a-2)}, \boldsymbol{B}^{(b-1)})].$$

More generally, one can follow the same argument to find a bound on the total contribution from paths ending at a subsample of size (m_A, m_B):

$$4U_{m_A,m_B}\binom{a+b-m_A-m_B}{a-m_A}\frac{a!}{m_A!}\frac{b!}{m_B!}\frac{(\theta_A)_a}{(\theta_A)_{m_A}}\frac{(\theta_B)_b}{(\theta_B)_{m_B}} \times q^A(\boldsymbol{a})q^B(\boldsymbol{b})\mathbb{E}[s_k(\boldsymbol{A}^{(m_A)}, \boldsymbol{B}^{(m_B)})]. \tag{B.6}$$

The binomial coefficient $\binom{a+b-m_A-m_B}{a-m_A}$ accounts for the number of ways of interspersing the sequences $(\boldsymbol{a}, \boldsymbol{a}^{(a-1)}, \ldots, \boldsymbol{a}^{(m_A)})$ and $(\boldsymbol{b}, \boldsymbol{b}^{(b-1)}, \ldots, \boldsymbol{b}^{(m_B)})$. Referring to (B.1), the expectations in (B.6) can be evaluated:

$$\begin{aligned}
\mathbb{E}[s_1(\boldsymbol{A}^{(m_A)}, \boldsymbol{B}^{(m_B)})] &= m_A m_B \theta_A \theta_B, \\
\mathbb{E}[s_2(\boldsymbol{A}^{(m_A)}, \boldsymbol{B}^{(m_B)})] &= -m_B\theta_B(\theta_A + m_A - 1)\sum_i \frac{a_i\binom{a-a_i}{m_A-1}}{\binom{a}{m_A}}, \\
\mathbb{E}[s_3(\boldsymbol{A}^{(m_A)}, \boldsymbol{B}^{(m_B)})] &= -m_A\theta_A(\theta_B + m_B - 1)\sum_j \frac{b_j\binom{b-b_j}{m_B-1}}{\binom{b}{m_B}}, \\
\mathbb{E}[s_4(\boldsymbol{A}^{(m_A)}, \boldsymbol{B}^{(m_B)})] &= (\theta_A + m_A - 1)(\theta_B + m_B - 1) \\
&\quad \times \sum_i \frac{a_i\binom{a-a_i}{m_A-1}}{\binom{a}{m_A}} \sum_j \frac{b_j\binom{b-b_j}{m_B-1}}{\binom{b}{m_B}}.
\end{aligned} \tag{B.7}$$

Summing (B.6) over all subsample sizes and over $k = 1, \ldots, 4$, we have therefore proven the following:

Proposition B.1.

$$q_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0}) \le 4q^A(\boldsymbol{a})q^B(\boldsymbol{b}) \sum_{m_A=2}^{a}\sum_{m_B=2}^{b}(a+b-m_A-m_B)!\binom{a}{m_A}\binom{b}{m_B} \times \frac{(\theta_A)_a}{(\theta_A)_{m_A}}\frac{(\theta_B)_b}{(\theta_B)_{m_B}}U_{m_A,m_B}\sum_{k=1}^{4}\mathbb{E}[s_k(\boldsymbol{A}^{(m_A)}, \boldsymbol{B}^{(m_B)})],$$

where $\mathbb{E}[s_k(\boldsymbol{A}^{(m_A)}, \boldsymbol{B}^{(m_B)})]$ *is given by (B.7) for* $k = 1, \ldots, 4$.

In a similar fashion,

$$q_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0}) \geq 4q^A(\boldsymbol{a})q^B(\boldsymbol{b}) \sum_{m_A=2}^{a}\sum_{m_B=2}^{b}(a+b-m_A-m_B)!\binom{a}{m_A}\binom{b}{m_B!} \times \frac{(\theta_A)_a}{(\theta_A)_{m_A}}\frac{(\theta_B)_b}{(\theta_B)_{m_B}}L_{m_A,m_B}\sum_{k=1}^{4}\mathbb{E}[s_k(\boldsymbol{A}^{m_A},\boldsymbol{B}^{m_B})],$$

where L_{m_A,m_B} is a lower bound on all coefficients of the form

$$[D(a,b)\dots D(m_A,m_B)]^{-1},$$

corresponding to paths through the recursion from $(\boldsymbol{a},\boldsymbol{b})$ to any subsample $(\boldsymbol{a}^{(m_A)},\boldsymbol{b}^{(m_B)})$ of size (m_A,m_B).

It remains to choose U_{m_A,m_B}. The closest possible bound is given by the actual path which maximizes this term, and this can be found by constructing an $a \times b$ dynamic programming table as follows. Define

$$U_{m_A,m_B} = \begin{cases} \dfrac{1}{D(a,b)} & \text{if } (m_A,m_B)=(a,b), \\ \dfrac{U_{m_A+1,b}}{D(m_A,b)} & \text{if } m_B=b, \\ \dfrac{U_{a,m_B+1}}{D(a,m_B)} & \text{if } m_A=a, \\ \dfrac{\max\{U_{m_A+1,m_B},U_{m_A,m_B+1}\}}{D(m_A,m_B)} & \\ \qquad\qquad \text{if } 2\leq m_A\leq a-1, \text{ and } 2\leq m_B\leq b-1, \end{cases}$$

and similarly for L_{m_A,m_B} (replacing max with min). Note that constructing this single table finds all the relevant bounds. In fact, U_{m_A,m_B} provides upper bounds for $r_1(\boldsymbol{a},\boldsymbol{b})$ and $r_4(\boldsymbol{a},\boldsymbol{b})$; and L_{m_A,m_B} provides upper bounds for $r_2(\boldsymbol{a},\boldsymbol{b})$ and $r_3(\boldsymbol{a},\boldsymbol{b})$, since these terms are less than or equal to zero.

The bounds presented above require $O(ab)$ operations, which, while growing with a and b, require much less time than a complete evaluation of $q_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{0})$. Empirical testing of this upper bound suggest that it is generally very close. We also investigated some constant time upper bounds as well as defining lower bounds, but these proved to be less useful.

REFERENCES

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

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

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

Ethier, S. N. and Griffiths, R. C. (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 Donnelly, P. (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. The two-locus ancestral graph. In Basawa, I. V. and Taylor, R. L., editors, *Selected proceedings of the Sheffield symposium on applied probability: 18. IMS Lecture Notes—Monograph series*, volume 18, pages 100–117, 1991.

Griffiths, R. C. and Lessard, S. (2005). Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theor. Popul. Biol.*, **68**, 167–77.

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

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

Hoppe, F. (1984). Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.*, **20**, 91–94.

Hudson, R. R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics*, **109**, 611–631.

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

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

Kingman, J. F. C. (1982b). The coalescent. *Stoch. Process. Appl.*, **13**, 235–248.

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

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

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

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

Pitman, J. The two-parameter generalization of Ewens' random partition structure. Technical Report 345, Dept. Statistics, U.C. Berkeley, 1992.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Th. Rel. Fields*, **102**, 145–158.

Slatkin, M. (1994). An exact test for neutrality based on the Ewens sampling distribution. *Genet. Res.*, **64**, 71–4.

Stephens, M. Inference under the coalescent. In Balding, D., Bishop, M., and Cannings,

C., editors, *Handbook of Statistical Genetics*, chapter 8. Wiley, Chichester, UK, 2001.

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

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

Watterson, G. A. (1977). Heterosis or neutrality? *Genetics*, **85**, 789–814.

P. A. Jenkins
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720
USA
E-mail: pauljenk@eecs.berkeley.edu

Y. S. Song
Department of Statistics and
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720
USA
E-mail: yss@stat.berkeley.edu