

TRACTABLE STOCHASTIC MODELS OF EVOLUTION FOR LOOSELY LINKED LOCI

BY PAUL A. JENKINS*

University of Warwick
AND

BY PAUL FEARNHEAD

Lancaster University
AND

BY YUN S. SONG†

University of California, Berkeley

Of fundamental importance in statistical genetics is to compute the sampling distribution, or likelihood, for a sample of genetic data from some stochastic evolutionary model. For DNA sequence data with inter-locus recombination, standard models include the Wright-Fisher diffusion with recombination and its dual genealogical process, the ancestral recombination graph. However, under neither of these models is the sampling distribution available in closed-form, and their computation is extremely difficult. In this paper we *derive* two new stochastic population genetic models, one a diffusion and the other a coalescent process, which are much simpler than the standard models, but which capture their key properties for large recombination rates. In the former case, we show that the sampling distribution is available in closed form. We further demonstrate that when we consider the sampling distribution as an asymptotic expansion in inverse powers of the recombination parameter, the sampling distributions of the two models agree with the standard ones up to the first two orders.

1. Introduction. The basis of many important problems in genetics is to find an expression for a sampling distribution or likelihood. Valuable tools in this endeavour are stochastic models of allele frequency evolution forwards in time, and their dual genealogical processes backwards in time. In

*Supported in part by EPSRC Research Grant EP/L018497/1 and an NIH Grant R01-GM094402.

†Supported in part by an NIH Grant R01-GM094402, and a Packard Fellowship for Science and Engineering.

MSC 2010 subject classifications: Primary 92D15; secondary 65C50, 92D10

Keywords and phrases: population genetics, recombination, sampling distribution, diffusion, coupling

particular, the numerous variants of the Wright-Fisher diffusion and Kingman’s coalescent, respectively, have focused attention on the scaling limit as the population size goes to infinity, leading from a (complicated) finite-population model of reproduction to a (simpler) infinite-population limit. At a single genetic locus, the problem of computing sampling distributions in these models is well studied, with even some closed-form formulas available (Wright, 1949; Ewens, 1972; Jenkins and Song, 2011; Bhaskar, Kamm and Song, 2012). However, with ongoing technological developments in high-throughput DNA sequencing, large genomic datasets are becoming available and it is necessary to consider multi-locus models. Inter-locus recombination quickly makes such models intractable; for neither the Wright-Fisher diffusion with recombination nor the coalescent with recombination—or *ancestral recombination graph* (ARG)—is it possible to obtain a closed-form expression for the sampling distribution. This has remained a notoriously difficult problem, and to make progress using these models it has usually been necessary to resort to computationally-intensive techniques such as importance sampling (Griffiths and Marjoram, 1996; Fearnhead and Donnelly, 2001; Griffiths, Jenkins and Song, 2008; Jenkins and Griffiths, 2011), Markov chain Monte Carlo (Kuhner, Yamato and Felsenstein, 2000; Nielsen, 2000; Wang and Rannala, 2008), or other numerical approximations (Boitard and Loisel, 2007; Miura, 2011). Denoting the population-scaled recombination parameter by ρ , only in the special cases of $\rho = 0$ or $\rho = \infty$ is it possible to make progress analytically, since then we are back to a single locus, or to many independent single loci, respectively.

In another direction, we have considered an analytic approach to the problem, as follows. Denote the observed sample configuration at two loci by \mathbf{n} and its sampling probability by $q(\mathbf{n}; \rho)$ (to be defined precisely below). Consider the asymptotic expansion in inverse powers of ρ :

$$(1) \quad q(\mathbf{n}; \rho) = q_0(\mathbf{n}) + \frac{q_1(\mathbf{n})}{\rho} + \frac{q_2(\mathbf{n})}{\rho^2} + \cdots,$$

where for convenience we suppress the dependence of these terms on other parameters of the model. Under an infinite-alleles type of mutation, we obtained closed-form formulas for $q_0(\mathbf{n})$ and $q_1(\mathbf{n})$ in terms of the marginal *one*-locus sampling probabilities, and a decomposition of $q_2(\mathbf{n})$ into a closed-form term plus a second part which is evaluated easily by dynamic programming (Jenkins and Song, 2010). (The result is stated more precisely in Theorem 2.1 below.) This provides the first closed-form extension of Ewens’ Sampling Formula (Ewens, 1972) to handle finite amounts of recombination. It has been extended subsequently to include more general models of

mutation (Jenkins and Song, 2009), natural selection (Jenkins and Song, 2012), higher-order terms (Jenkins and Song, 2012), and more than two loci (Bhaskar and Song, 2012), and has had practical implications for genomic inference (Chan, Jenkins and Song, 2012). One particularly appealing conclusion of these works is that both $q_0(\mathbf{n})$ and $q_1(\mathbf{n})$ are *universal*; that is, their functional form is invariant to our assumptions about mutation and selection acting marginally at each locus. The effects of these marginal processes are entirely subsumed into the relevant one-locus sampling distributions.

The simple and universal forms for $q_0(\mathbf{n})$ and $q_1(\mathbf{n})$ provide strong circumstantial evidence that there exists an underlying stochastic process which is much simpler than the standard models for finite amounts of recombination. In particular, we previously conjectured (Jenkins and Song, 2010) the existence of a process which is both much simpler than the standard models based on the Wright-Fisher diffusion or on the ARG, and is in agreement with the sampling distribution (1) up to $O(\rho^{-2})$. The goal of this paper is to describe such a process. In fact, using different arguments we describe two such processes, obtaining both a limiting diffusion and a coalescent process with these properties. In the diffusion approximation, the key idea is to suppose that the probability r of a recombination per individual per generation scales as $N^{-\beta}$ as the population size $N \rightarrow \infty$, for $0 < \beta < 1$, rather than the usual choice of $\beta = 1$. Our diffusion in this scaling is intimately related to the Gaussian diffusion approximation of Norman (1975a). A closely related scaling in the context of Ξ -coalescent processes was recently explored by Birkner, Blath and Eldon (2013) (in that paper $\beta = 1$ but with timescale N^2). The coalescent approach, meanwhile, uses a coupling argument.

The paper is organized as follows. In Section 2 we specify our notation and summarize previous research. Novel diffusion and coalescent processes are introduced in Sections 3 and 4, respectively, and we conclude in Section 5 with a brief discussion.

2. Notation and previous results. For $M \in \mathbb{N} = \{0, 1, 2, \dots\}$, let $[M] := \{1, 2, \dots, M\}$. The complement of a set J is written J^c . Denote the Kronecker delta by δ_{ij} which takes the value 1 if $i = j$ and 0 otherwise. Let \mathbf{e}_i denote a unit vector whose j th entry is δ_{ij} , and let \mathbf{e}_{ij} denote a matrix with (k, l) th entry equal to $\delta_{ik}\delta_{jl}$. Let $\binom{n}{\mathbf{n}}$ denote the multinomial coefficient for a vector or matrix \mathbf{n} whose entries are all nonnegative and sum to n ; for a matrix this is defined as $n! / \prod_{i,j} n_{ij}!$. We also extend the Euclidean norm to a matrix by the componentwise definition $|\mathbf{n}| = \sqrt{\sum_{i,j} n_{ij}^2}$. Denote the $k \times l$ zero matrix by $\mathbf{0}_{k \times l}$ and the $k \times k$ identity matrix by \mathbf{I}_k . We will

replace a subscript with a “.” to denote summation over that index. A prime symbol ‘ $'$ will denote vector or matrix transpose. For $z \in \mathbb{R}_{\geq 0}$ and $n \in \mathbb{N}$, $(z)_{n\uparrow} := z(z+1)\cdots(z+n-1)$ denotes the n th ascending factorial of z . Finally, for a random element Z we write $\mathcal{L}(Z)$ to denote its law.

Consider the usual diffusion limit of an exchangeable model of random mating with constant population size of $2N$ haplotypes. Our interest will be in a sample from this population at two loci, which we call A and B, with the probability of mutation per haplotype per generation denoted by u_A and u_B respectively. In the diffusion limit we let $N \rightarrow \infty$ and $u_A, u_B \rightarrow 0$ while the population-scaled parameters $\theta_A = 4Nu_A$ and $\theta_B = 4Nu_B$ remain fixed. In this paper we will suppose a *finite-alleles* model of mutation such that a mutation to an allele i in type space $E_A = [K]$, $K \in \mathbb{N}$, takes it to allele $k \in [K]$ with probability P_{ik}^A , with $E_B = [L]$ and P_{jl}^B , $j, l \in [L]$ defined analogously. (As we discover below, the mutation model is not important and we could pose something more complicated with little extra effort.) The probability of a recombination between the two loci per haplotype per generation is denoted by r , and we assume that $\rho_\beta = 4N^\beta r$ is fixed as $N \rightarrow \infty$, for some fixed $\beta \in (0, 1]$. Previous work has focused on the case $\beta = 1$ with time measured in units of $2N$ generations. For consistency with the usual notation, we write $\rho = \rho_1$.

A sample from this model comprises a haplotypes observed only at locus A, b haplotypes observed only at locus B, and c haplotypes observed at both loci. The sample configuration is denoted by $\mathbf{n} = (\mathbf{a}, \mathbf{b}, \mathbf{c})$ where $\mathbf{a} = (a_i)_{i \in [K]}$ and a_i is the number of haplotypes observed to exhibit allele i at locus A; $\mathbf{b} = (b_j)_{j \in [L]}$ where b_j is the number of haplotypes observed to exhibit allele j at locus B; and $\mathbf{c} = (c_{ij})_{i \in [K], j \in [L]}$ where c_{ij} is the number of haplotypes with allele i at locus A and allele j at locus B. Thus,

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

and we let $n = a + b + c$. We further write $\mathbf{c}_A = (c_{i\cdot})_{i \in [K]}$ and $\mathbf{c}_B = (c_{\cdot j})_{j \in [L]}$ to denote the marginal sample configurations of \mathbf{c} restricted to locus A and locus B respectively. Finally, we use $q(\mathbf{a}, \mathbf{b}, \mathbf{c})$ to denote the probability that when we sample n haplotypes in some order from the population at stationarity we obtain the unordered configuration $(\mathbf{a}, \mathbf{b}, \mathbf{c})$; by sampling exchangeability this is indeed a function only of the unordered configuration $(\mathbf{a}, \mathbf{b}, \mathbf{c})$. For convenience we suppress the dependence of this quantity on the model parameters. The main result motivating this work, for $\beta = 1$, is as follows.

THEOREM 2.1 (See [Jenkins and Song \(2009\)](#)). *Consider the asymptotic expansion*

$$q(\mathbf{a}, \mathbf{b}, \mathbf{c}) = q_0(\mathbf{a}, \mathbf{b}, \mathbf{c}) + \frac{q_1(\mathbf{a}, \mathbf{b}, \mathbf{c})}{\rho} + O\left(\frac{1}{\rho^2}\right), \quad \text{as } \rho \rightarrow \infty,$$

with q_0, q_1, \dots independent of ρ . Then the zeroth order term is given by

$$(2) \quad q_0(\mathbf{a}, \mathbf{b}, \mathbf{c}) = q^A(\mathbf{a} + \mathbf{c}_A)q^B(\mathbf{b} + \mathbf{c}_B),$$

and the first order term is given by

$$(3) \quad \begin{aligned} q_1(\mathbf{a}, \mathbf{b}, \mathbf{c}) = & \binom{c}{2} q^A(\mathbf{a} + \mathbf{c}_A)q^B(\mathbf{b} + \mathbf{c}_B) \\ & - q^B(\mathbf{b} + \mathbf{c}_B) \sum_{i=1}^K \binom{c_{i\cdot}}{2} q^A(\mathbf{a} + \mathbf{c}_A - \mathbf{e}_i) \\ & - q^A(\mathbf{a} + \mathbf{c}_A) \sum_{j=1}^L \binom{c_{\cdot j}}{2} q^B(\mathbf{b} + \mathbf{c}_B - \mathbf{e}_j) \\ & + \sum_{i=1}^K \sum_{j=1}^L \binom{c_{ij}}{2} q^A(\mathbf{a} + \mathbf{c}_A - \mathbf{e}_i)q^B(\mathbf{b} + \mathbf{c}_B - \mathbf{e}_j), \end{aligned}$$

where q^A, q^B are the marginal sampling distributions at locus A and locus B , respectively.

REMARK 2.1. (i) Under a neutral, finite-alleles model of mutation, if mutation is parent-independent—that is, $P_{ki}^A = P_i^A$, $i, k \in [K]$, and $P_{lj}^B = P_j^B$, $j, l \in [L]$, then $q^A(\mathbf{a})$ and $q^B(\mathbf{b})$ are known in closed-form:

$$q^A(\mathbf{a}) = \frac{1}{(\theta_A)_{\mathbf{a}\uparrow}} \prod_{i=1}^K (\theta_A P_i^A)_{a_i\uparrow}, \quad \text{and} \quad q^B(\mathbf{b}) = \frac{1}{(\theta_B)_{\mathbf{b}\uparrow}} \prod_{j=1}^L (\theta_B P_j^B)_{b_j\uparrow}.$$

These expressions follow, for example, from the moments of the Wright-Fisher diffusion with parent-independent mutation, whose stationary distribution at locus A is $\text{Dirichlet}(\theta_A P_1^A, \dots, \theta_A P_{K-1}^A)$ ([Wright, 1949](#)), and similarly at locus B .

(ii) The zeroth-order decomposition is well known (e.g. [Ethier, 1979](#)) and also intuitive, since the two loci become independent as $\rho \rightarrow \infty$.

Theorem 2.1 can be obtained by diffusion ([Jenkins and Song, 2012](#)) or by coalescent ([Jenkins and Song, 2009, 2010](#)) arguments. In this paper we address both approaches in further detail.

3. Diffusion model. In this section we extend the above results by obtaining a full description of a simple diffusion process such that its sampling distribution is known *exactly* and has a Taylor expansion about $\rho = \infty$ consistent with (2) and (3). For simplicity we will obtain our diffusion as the limit of an appropriately rescaled Wright-Fisher model, although we expect our results to hold for a more general class of discrete models of reproduction within the domain of convergence of the Wright-Fisher diffusion.

3.1. Neutral Wright-Fisher model. A population of N diploid, monoecious individuals reproduces in discrete, non-overlapping generations. Each individual carries two haplotypes, and each haplotype comprises a pair of alleles $(i, j) \in [K] \times [L]$, one at locus A and one at locus B.

Let $Z_{ij}(\tau) \in \{0, 1, \dots, 2N\}$ denote the number of (i, j) haplotypes in the population in generation $\tau \in \mathbb{N}$, and $\mathbf{Z}(\tau) = (Z_{ij}(\tau))_{i \in [K], j \in [L]}$. The $(\tau + 1)$ th generation is formed from the τ th as follows. Each individual contributes equally to an infinite pool of haploid gametes, and each gamete is formed by sampling a parental haplotype at random. Recombination occurs with probability r —specifically, a parent with diplotype $\{(i, j), (k, l)\} \in ([K] \times [L])^2$ contributes gametes with haplotypes (i, j) , (k, l) , (i, l) , (k, j) in relative frequencies $(1-r)/2$, $(1-r)/2$, $r/2$, and $r/2$, respectively. Each allele of each gamete also undergoes mutation independently with probability u_A at locus A, mutating according to the transition matrix $\mathbf{P}^A = (P_{ik}^A)_{i,k \in [K]}$; and with probability u_B at locus B, mutating according to $\mathbf{P}^B = (P_{jl}^B)_{j,l \in [L]}$. The frequency of gametes of type (i, j) in the resulting population pool is given approximately by

$$\begin{aligned} \phi_{ij} = & \sum_{k=1}^K \sum_{l=1}^L \left[(1-r) \frac{Z_{kl}(\tau)}{2N} + r \frac{Z_{k\cdot}(\tau)}{2N} \frac{Z_{\cdot l}(\tau)}{2N} \right] [u_A P_{ki}^A u_B P_{lj}^B \\ (4) \quad & + \delta_{ik}(1-u_A)u_B P_{lj}^B + \delta_{jl}u_A P_{ki}^A(1-u_B) + \delta_{ik}\delta_{jl}(1-u_A)(1-u_B)], \end{aligned}$$

In the terminology of Ewens (2004), this is the random union of gametes approximation to a random union of *zygotes* model, in which we assume that the frequency of $\{(i, j), (k, l)\}$ diploypes in generation τ is given by $Z_{ij}(\tau)Z_{kl}(\tau)/(2N)^2$. The approximation is reasonable for large N and exact in the diffusion limit (Ewens, 2004, p130, p227). We work under this assumption so that we may follow haplotype, rather than diplotype, frequencies.

Finally, the $2N$ haplotypes of generation $\tau + 1$ are formed by multinomial sampling from this pool:

$$(5) \quad \mathbf{Z}(\tau + 1) \mid \mathbf{Z}(\tau) \sim \text{Multinomial}(2N, \boldsymbol{\phi}), \quad \boldsymbol{\phi} := (\phi_{ij})_{i \in [K], j \in [L]}.$$

We will change variables by introducing the collection

$$\mathbf{M}^{(N)}(\tau) := \{\mathbf{X}^{(N)}(\tau), \mathbf{Y}^{(N)}(\tau), \mathbf{D}^{(N)}(\tau)\},$$

where

$$\begin{aligned} \mathbf{X}^{(N)}(\tau) &= (X_i^{(N)}(\tau))_{i \in [K]} = \left(\frac{Z_{i\cdot}(\tau)}{2N} : i \in [K] \right), \\ \mathbf{Y}^{(N)}(\tau) &= (Y_j^{(N)}(\tau))_{j \in [L]} = \left(\frac{Z_{\cdot j}(\tau)}{2N} : j \in [L] \right), \\ \mathbf{D}^{(N)}(\tau) &= (D_{ij}^{(N)}(\tau))_{i \in [K], j \in [L]} = \left(\frac{Z_{ij}(\tau)}{2N} - \frac{Z_{i\cdot}(\tau)}{2N} \frac{Z_{\cdot j}(\tau)}{2N} : i \in [K], j \in [L] \right). \end{aligned}$$

That is, we describe the state of the Wright-Fisher model in generation τ by the marginal allele frequencies and the coefficients of linkage disequilibrium. The process $(\mathbf{M}^{(N)}(\tau) : \tau = 0, 1, \dots)$ is Markov on a $(KL - 1)$ -dimensional shifted simplex Δ_{KL-1} defined via the conditions

$$(6) \quad \begin{aligned} 0 \leq X_i^{(N)} \leq 1; \quad 0 \leq Y_j^{(N)} \leq 1; \quad -1 \leq D_{ij}^{(N)} \leq 1; \\ X_{\cdot}^{(N)} = 1; \quad Y_{\cdot}^{(N)} = 1; \quad D_{\cdot j}^{(N)} = 0; \quad D_{i\cdot}^{(N)} = 0; \end{aligned}$$

for all $i \in [K]$ and $j \in [L]$. To find the diffusion limit we first need the conditional means and covariances of the increments

$$\Delta \mathbf{M}^{(N)}(\tau) := \mathbf{M}^{(N)}(\tau + 1) - \mathbf{M}^{(N)}(\tau).$$

For convenience we drop the dependence on τ .

PROPOSITION 3.1. *In the neutral two-locus Wright-Fisher model with mutation and recombination, the conditional means and covariances of increments of $\mathbf{M}^{(N)}$ are given by*

$$(7) \quad \mathbb{E}[\Delta X_i^{(N)} \mid \mathbf{M}^{(N)}] = \frac{\theta_A}{4N} \sum_{k=1}^K (P_{ki}^A - \delta_{ik}) X_i^{(N)},$$

$$(8) \quad \mathbb{E}[\Delta Y_j^{(N)} \mid \mathbf{M}^{(N)}] = \frac{\theta_B}{4N} \sum_{l=1}^L (P_{lj}^B - \delta_{jl}) Y_j^{(N)},$$

$$\begin{aligned} \mathbb{E}[\Delta D_{ij}^{(N)} \mid \mathbf{M}^{(N)}] &= -\frac{\rho\beta}{4N\beta} D_{ij}^{(N)} + \frac{1}{2N} \left(\frac{\theta_A}{2} \sum_{k=1}^K (P_{ki}^A - \delta_{ik}) D_{kj}^{(N)} \right. \\ &\quad \left. + \frac{\theta_B}{2} \sum_{l=1}^L (P_{lj}^B - \delta_{jl}) D_{il}^{(N)} - D_{ij}^{(N)} \right) \end{aligned}$$

$$(9) \quad + O\left(\frac{1}{N^{1+\beta}}\right),$$

$$\begin{aligned} \text{cov}[\Delta X_i^{(N)}, \Delta X_k^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{2N} X_i^{(N)} (\delta_{ik} - X_k^{(N)}) + O\left(\frac{1}{N^2}\right), \\ \text{cov}[\Delta Y_j^{(N)}, \Delta Y_l^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{2N} Y_j^{(N)} (\delta_{jl} - Y_l^{(N)}) + O\left(\frac{1}{N^2}\right), \\ \text{cov}[\Delta X_i^{(N)}, \Delta Y_j^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{2N} \left(1 - \frac{\rho_\beta}{4N^\beta}\right) D_{ij}^{(N)} + O\left(\frac{1}{N^2}\right), \\ \text{cov}[\Delta X_i^{(N)}, \Delta D_{kl}^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{2N} \left(D_{kl}^{(N)} (\delta_{ik} - X_i^{(N)}) - X_k^{(N)} D_{il}^{(N)}\right) \\ &\quad + O\left(\frac{1}{N^{1+\beta}}\right), \\ \text{cov}[\Delta Y_j^{(N)}, \Delta D_{kl}^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{2N} \left(D_{kl}^{(N)} (\delta_{jl} - Y_j^{(N)}) - Y_l^{(N)} D_{kj}^{(N)}\right) \\ &\quad + O\left(\frac{1}{N^{1+\beta}}\right), \\ \text{cov}[\Delta D_{ij}^{(N)}, \Delta D_{kl}^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{2N} \left(X_i^{(N)} Y_j^{(N)} (\delta_{ik} - X_k^{(N)}) (\delta_{jl} - Y_l^{(N)}) \right. \\ &\quad + D_{kj}^{(N)} X_i^{(N)} Y_l^{(N)} + D_{il}^{(N)} X_k^{(N)} Y_j^{(N)} \\ &\quad + D_{ij}^{(N)} (X_k^{(N)} Y_l^{(N)} - \delta_{ik} Y_l^{(N)} - \delta_{jl} X_k^{(N)}) \\ &\quad + D_{kl}^{(N)} (X_i^{(N)} Y_j^{(N)} - \delta_{ik} Y_j^{(N)} - \delta_{jl} X_i^{(N)}) \\ &\quad \left. + D_{ij}^{(N)} (\delta_{ik} \delta_{jl} - D_{kl}^{(N)})\right) + O\left(\frac{1}{N^{1+\beta}}\right). \end{aligned}$$

PROOF. To simplify notation we write $\mathbf{Z}' \mid \mathbf{Z}$ for $\mathbf{Z}(\tau+1) \mid \mathbf{Z}(\tau)$. Incremental moments can then be expressed as follows:

$$\begin{aligned} \mathbb{E}[\Delta X_i^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{2N} \mathbb{E}[Z'_i \mid \mathbf{Z}] - \frac{X_i^{(N)}}{2N}, \\ \mathbb{E}[\Delta D_{ij}^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{2N} \mathbb{E}[Z'_{ij} \mid \mathbf{Z}] - \frac{1}{(2N)^2} \mathbb{E}[Z'_i Z'_{\cdot j} \mid \mathbf{Z}] - D_{ij}^{(N)}, \\ \mathbb{E}[\Delta X_i^{(N)} \Delta X_k^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{(2N)^2} \left[\mathbb{E}[Z'_i Z'_{k\cdot} \mid \mathbf{Z}] - \mathbb{E}[Z'_i \mid \mathbf{Z}] Z_{k\cdot} \right. \\ &\quad \left. - \mathbb{E}[Z'_{k\cdot} \mid \mathbf{Z}] Z_i + Z_i Z_{k\cdot} \right], \\ \mathbb{E}[\Delta X_i^{(N)} \Delta Y_j^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{(2N)^2} \left[\mathbb{E}[Z'_i Z'_{\cdot j} \mid \mathbf{Z}] - \mathbb{E}[Z'_i \mid \mathbf{Z}] Z_{\cdot j} \right. \end{aligned}$$

$$\begin{aligned}
& - \mathbb{E}[Z'_{\cdot j} \mid \mathbf{Z}] Z_{i\cdot} + Z_{i\cdot} Z_{\cdot j}], \\
\mathbb{E}[\Delta X_i^{(N)} \Delta D_{kl}^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{(2N)^2} \left[\mathbb{E}[Z'_{i\cdot} Z'_{kl} \mid \mathbf{Z}] - \frac{1}{2N} \mathbb{E}[Z'_{i\cdot} Z'_{k\cdot} Z'_{\cdot l} \mid \mathbf{Z}] \right] \\
& - \frac{D_{kl}^{(N)}}{2N} \mathbb{E}[Z'_{i\cdot} \mid \mathbf{Z}] - \frac{Z_{i\cdot}}{2N} \mathbb{E}[\Delta D_{kl}^{(N)} \mid \mathbf{M}^{(N)}], \\
\mathbb{E}[\Delta D_{ij}^{(N)} \Delta D_{kl}^{(N)} \mid \mathbf{M}^{(N)}] &= \frac{1}{(2N)^2} \mathbb{E}[Z'_{ij} Z'_{kl} \mid \mathbf{Z}] - \frac{1}{(2N)^3} \mathbb{E}[Z'_{ij} Z'_{k\cdot} Z'_{\cdot l} \mid \mathbf{Z}] \\
& - \frac{1}{(2N)^3} \mathbb{E}[Z'_{kl} Z'_{i\cdot} Z'_{\cdot j} \mid \mathbf{Z}] \\
& + \frac{1}{(2N)^4} \mathbb{E}[Z'_{i\cdot} Z'_{\cdot j} Z'_{k\cdot} Z'_{\cdot l} \mid \mathbf{Z}] \\
& - D_{ij}^{(N)} \mathbb{E}[\Delta D_{kl}^{(N)} \mid \mathbf{M}^{(N)}] \\
& - D_{kl}^{(N)} \mathbb{E}[\Delta D_{ij}^{(N)} \mid \mathbf{M}^{(N)}] + D_{ij}^{(N)} D_{kl}^{(N)},
\end{aligned}$$

with the remaining terms involving $Y_j^{(N)}$ following similarly. Substitute the expectations into the right-hand sides of the above equations by summing over the first four moments of the multinomial distribution in (5), then use the known form (4) for ϕ_{ij} , discard higher order terms, and finally express covariances in terms of the first two moments. We omit the straightforward but lengthy algebraic details. \square

We will write the results of Proposition 3.1 succinctly first by arranging the variables in a linear order:

$$(X_1^{(N)}, \dots, X_K^{(N)}, Y_1^{(N)}, \dots, Y_L^{(N)}, D_{11}^{(N)}, \dots, D_{KL}^{(N)})',$$

and thinking of $\mathbf{M}^{(N)}(\tau)$ as a vector of length $\Lambda := (K + L + KL)$. Now introduce the conditional mean vector \mathbf{w} and conditional covariance matrix \mathbf{s} , defined by

$$(10) \quad \mathbb{E}[\Delta \mathbf{M}^{(N)} \mid \mathbf{M}^{(N)}(\tau)] = \frac{1}{2N^\beta} \mathbf{w}(\mathbf{M}^{(N)}(\tau)) + O\left(\frac{1}{N}\right),$$

$$(11) \quad \text{cov}[\Delta \mathbf{M}^{(N)} \mid \mathbf{M}^{(N)}(\tau)] = \frac{1}{2N} \mathbf{s}(\mathbf{M}^{(N)}(\tau)) + O\left(\frac{1}{N^{1+\beta}}\right),$$

with entries determined by Proposition 3.1. (Only leading order terms are retained in \mathbf{w} and \mathbf{s} .) Thus for example, equations (7)–(9) show that

$$(12) \quad \mathbf{w}(\mathbf{M}^{(N)}(\tau)) = \left(\underbrace{0, \dots, 0}_K, \underbrace{0, \dots, 0}_L, \underbrace{-\frac{\rho_\beta}{2} D_{11}^{(N)}(\tau), \dots, -\frac{\rho_\beta}{2} D_{KL}^{(N)}(\tau)}_{K \times L} \right)',$$

with \mathbf{s} determined in a similar fashion. Notice the different leading orders of the two quantities: the mean increments are of $O(N^{-\beta})$ while the covariances are of $O(N^{-1})$. It is this difference, which is a consequence of our assumption that the recombination probability r is $O(N^{-\beta})$ for $\beta < 1$, that leads to a novel diffusion limit. Under the usual choice of $\beta = 1$ it is well known that the Wright-Fisher model converges to a diffusion process after a linear rescaling of time. In the special case $K = L = 2$, the diffusion limit for $\mathbf{M}^{(N)}(\lfloor N\tau \rfloor)$ as $N \rightarrow \infty$ was obtained by [Ohta and Kimura \(1969a,b\)](#). Our interest is however in $0 < \beta < 1$, for which r is larger, and the loss of linkage disequilibrium (LD) is subsequently much faster. Intuitively, we should expect such loss to resemble the exponential decay predicted in an infinitely large population, but with small fluctuations about this deterministic behaviour. The diffusion process we define below quantifies these fluctuations precisely.

3.2. Diffusion limit. We first introduce the process $(\gamma^{(N)}(\tau) : \tau = 0, 1, \dots)$ defined recursively by the mapping

$$(13) \quad \gamma^{(N)}(\tau + 1) := \gamma^{(N)}(\tau) + \frac{1}{2N^\beta} \mathbf{w}(\gamma^{(N)}(\tau)), \quad \gamma^{(N)}(0) = \mathbf{M}^{(N)}(0).$$

This mapping closely approximates the deterministic, exponential decay in LD of (10) and (12). Fluctuations about this process are captured on an appropriate spatial scale by defining

$$(14) \quad \widetilde{\mathbf{M}}^{(N)}(\tau) := [\mathbf{M}^{(N)}(\tau) - \gamma^{(N)}(\tau)]N^{(1-\beta)/2}.$$

The scaling $N^{(1-\beta)/2}$ can be regarded as the one on which both recombination and genetic drift are observable ([Jenkins and Song, 2012](#)). The factor of $1/2$ appearing in the exponent is also closely related to the central limit theorem ([Norman, 1972](#), p116). We measure time in units of $2N^\beta$ generations by introducing the continuous time parameter t defined via $\tau = \lfloor 2N^\beta t \rfloor$, and the linearly interpolated continuous time process $\widehat{\mathbf{M}}^{(N)} := (\widehat{\mathbf{M}}^{(N)}(t) : t \in [0, T])$ for fixed $T \in (0, \infty)$ given by

$$\widehat{\mathbf{M}}^{(N)}(t) := \widetilde{\mathbf{M}}^{(N)}(\tau) + \Delta \widetilde{\mathbf{M}}^{(N)}(\tau) \cdot (2N^\beta t - \tau).$$

Our main result is the following.

THEOREM 3.1. *Suppose $\text{cov}(\widetilde{\mathbf{M}}^{(N)}(0)) = 0$. Then:*

1. $\mathbb{E} \left[\max_{\tau \leq 2N^\beta T} \left| \widetilde{\mathbf{M}}^{(N)}(\tau) \right|^3 \right]$ is bounded.

2. *There exists a diffusion process $(\mathbf{V}^{(N)}(t) : t \geq 0)$ with $\mathbf{V}^{(N)}(0) = \mathbf{0}$ almost surely, and with Gaussian transition probabilities such that the law $\mathcal{L}(\mathbf{V}^{(N)}(t+u) \mid \mathbf{V}^{(N)}(t) = \mathbf{v})$ is that of a normal distribution with mean $\mathbf{B}(u)\mathbf{v}$ and covariance $\Sigma(u, f(t, \mathbf{M}^{(N)}(0)))$, where for a realization $\mathbf{M}^{(N)} = \mathbf{m} = (\mathbf{x}, \mathbf{y}, \mathbf{d})'$,*

$$\mathbf{B}(u) = \begin{bmatrix} \mathbf{I}_K & \mathbf{0}_{K \times L} & \mathbf{0}_{K \times KL} \\ \mathbf{0}_{L \times K} & \mathbf{I}_L & \mathbf{0}_{L \times KL} \\ \mathbf{0}_{KL \times K} & \mathbf{0}_{KL \times L} & \mathbf{I}_{KL} \exp\left(\frac{-\rho_\beta u}{2}\right) \end{bmatrix},$$

$$f(t, \mathbf{m}) = (\mathbf{x}, \mathbf{y}, \mathbf{d}e^{-\rho_\beta t/2})',$$

$$\Sigma(u, \mathbf{m}) = \begin{bmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} & \Sigma_{\mathbf{XD}} \\ \Sigma_{\mathbf{XY}} & \Sigma_{\mathbf{YY}} & \Sigma_{\mathbf{YD}} \\ \Sigma_{\mathbf{XD}} & \Sigma_{\mathbf{YD}} & \Sigma_{\mathbf{DD}} \end{bmatrix},$$

and

$$\begin{aligned} [\Sigma_{\mathbf{XX}}]_{ik} &= x_i(\delta_{ik} - x_k)u, \\ [\Sigma_{\mathbf{YY}}]_{jl} &= y_j(\delta_{jl} - y_l)u, \\ [\Sigma_{\mathbf{XY}}]_{ij} &= \frac{2}{\rho_\beta} d_{ij}(1 - e^{-\rho_\beta u/2}), \\ [\Sigma_{\mathbf{XD}}]_{i,kl} &= \frac{2}{\rho_\beta} [d_{kl}(\delta_{ik} - x_i) - x_k d_{il}](1 - e^{-\rho_\beta u/2}), \\ [\Sigma_{\mathbf{YD}}]_{j,kl} &= \frac{2}{\rho_\beta} [d_{kl}(\delta_{jl} - y_j) - y_l d_{kj}](1 - e^{-\rho_\beta u/2}), \\ [\Sigma_{\mathbf{DD}}]_{ij,kl} &= \frac{1}{\rho_\beta} \left[x_i y_j (\delta_{ik} - x_k)(\delta_{jl} - y_l)(1 - e^{-\rho_\beta u}) \right. \\ &\quad + 2[d_{kj} x_i y_l + d_{il} x_k y_j + d_{ij}(x_k y_l - \delta_{ik} y_l - \delta_{jl} x_k) \\ &\quad + d_{kl}(x_i y_j - \delta_{ik} y_j - \delta_{jl} x_i) \\ &\quad + d_{ij} \delta_{ik} \delta_{jl}](e^{-\rho_\beta u/2} - e^{-\rho_\beta u}) \\ &\quad \left. - d_{ij} d_{kl} u e^{-\rho_\beta u} \right]. \end{aligned}$$

The finite-dimensional distributions of $(\widetilde{\mathbf{M}}^{(N)}(\tau) : \tau = 0, 1, \dots)$ converge to those of $(\mathbf{V}^{(N)}(t) : t \geq 0)$ in the following uniform sense: For any $J \geq 1$ and any bounded, continuous function $F : \Delta_{KL-1}^J \rightarrow \mathbb{R}$,

$$\max_{\tau_1 \leq \tau_2 \leq \dots \leq \tau_J \leq 2N^\beta T} \left| \mathbb{E}[F(\widetilde{\mathbf{M}}^{(N)}(\tau_1), \dots, \widetilde{\mathbf{M}}^{(N)}(\tau_J))] \right. \\ \left. - \mathbb{E}\left[F\left(\mathbf{V}^{(N)}\left(\frac{\tau_1}{2N^\beta}\right), \dots, \mathbf{V}^{(N)}\left(\frac{\tau_J}{2N^\beta}\right)\right)\right] \right| \rightarrow 0,$$

as $N \rightarrow \infty$. Furthermore,

$$\max_{\tau \leq 2N^\beta T} \left| \mathbb{E}[\widetilde{\mathbf{M}}^{(N)}(\tau)] \right| \rightarrow 0, \quad \text{and} \\ \max_{\tau \leq 2N^\beta T} \left| \text{cov}(\widetilde{\mathbf{M}}^{(N)}(\tau)) - \Sigma\left(\frac{\tau}{2N^\beta}, \widetilde{\mathbf{M}}^{(N)}(0)\right) \right| \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

3. In fact, for any bounded, continuous function $H : C[0, T] \rightarrow \mathbb{R}$ (with domain the space of continuous functions on $[0, T]$),

$$\mathbb{E}[H(\widehat{\mathbf{M}}^{(N)})] - \mathbb{E}[H(\mathbf{V}^{(N)})] \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

To prove Theorem 3.1, we need the following lemma, a proof of which may be found in Norman (1972, p261–262):

LEMMA 3.1. Suppose $Y \sim \text{Binomial}(N, \frac{x}{N})$ for $x \in \{0, \dots, N\}$. Then

$$\mathbb{E} \left[\left| \frac{Y}{N} - \frac{x}{N} \right|^3 \right] = O \left(\frac{1}{N^{3/2}} \right),$$

as $N \rightarrow \infty$, uniformly over x .

PROOF OF THEOREM 3.1. This is an application of multidimensional versions of Theorems 1–3 of Norman (1975a, p241). To apply those results we need to check the following conditions, (i)–(ix).

(i) The random vector $\mathbf{M}^{(N)}(\tau)$ takes values in a closed, convex subset of \mathbb{R}^k , for some k .

This holds since Δ_{KL-1} is a closed, convex subset of \mathbb{R}^Λ .

(ii) $\mathbf{M}^{(N)}(\tau)$ is measurable with respect to a σ -field $\mathcal{F}_\tau^{(N)}$, $\mathcal{F}_{\tau+1}^{(N)} \supseteq \mathcal{F}_\tau^{(N)}$, and the conditional incremental means and covariances are of the form

$$(15) \quad \mathbb{E}[\Delta \mathbf{M}^{(N)} \mid \mathcal{F}_\tau^{(N)}] = \epsilon^{(N)} \mathbf{w}^{(N)}(\mathbf{M}^{(N)}(\tau)) + \mathbf{e}_{1,\tau}^{(N)},$$

$$(16) \quad \text{cov}[\Delta \mathbf{M}^{(N)} \mid \mathcal{F}_\tau^{(N)}] = \tau^{(N)} \mathbf{s}^{(N)}(\mathbf{M}^{(N)}(\tau)) + \mathbf{e}_{2,\tau}^{(N)},$$

where $\epsilon^{(N)} > 0$; $\tau^{(N)} > 0$; $\epsilon^{(N)} \rightarrow 0$ and $\tau^{(N)}/\epsilon^{(N)} \rightarrow 0$ as $N \rightarrow \infty$.

These forms follow from (10) and (11): it suffices to take the natural filtration associated with $(\mathbf{M}^{(N)}(\tau) : \tau = 0, 1, \dots)$, $\epsilon^{(N)} = 1/(2N^\beta)$, and

$\tau^{(N)} = 1/(2N)$. There is some freedom over whether to absorb higher order terms into $\mathbf{w}^{(N)}$ and $\mathbf{s}^{(N)}$, or into $\mathbf{e}_{1,\tau}^{(N)}$ and $\mathbf{e}_{2,\tau}^{(N)}$; this involves a trade-off on the conditions on the four quantities we need to check below. We choose to match (10) and (11) so that $\mathbf{w}^{(N)} \equiv \mathbf{w}$ and $\mathbf{s}^{(N)} \equiv \mathbf{s}$ do not depend on N .
 (iii) $\mathbf{w}(\mathbf{m})$ and $\mathbf{s}(\mathbf{m})$ are a vector and nonnegative definite matrix, respectively, whose entries are real-valued functions with domain Δ_{KL-1} .

The forms of $\mathbf{w}(\mathbf{m})$ and $\mathbf{s}(\mathbf{m})$ follow by construction. Nonnegative definiteness is verified as follows. The error term $2N\mathbf{e}_{2,\tau}^{(N)} \rightarrow \mathbf{0}_{\Lambda \times \Lambda}$ in L^1 as $N \rightarrow \infty$ [see (viii) below], so if $\mathbf{s}(\mathbf{M}^{(N)}(\tau))$ fails to be nonnegative definite then we can find an N such that with nonzero probability $\mathbf{s}(\mathbf{M}^{(N)}(\tau)) + 2N\mathbf{e}_{2,\tau}^{(N)}$ fails too. But the right-hand side of (16) is nonnegative definite by construction, so $\mathbf{s}(\mathbf{M}^{(N)}(\tau)) + 2N\mathbf{e}_{2,\tau}^{(N)}$ is also nonnegative definite a.s., leading to a contradiction.

(iv) The vector \mathbf{w} and the matrices $\partial\mathbf{w}$, \mathbf{s} are bounded (uniformly in N , which is automatic here):

$$\sup_{\mathbf{m} \in \Delta_{KL-1}} |\mathbf{w}(\mathbf{m})| < \infty; \quad \sup_{\mathbf{m} \in \Delta_{KL-1}} |\partial\mathbf{w}(\mathbf{m})| < \infty; \quad \sup_{\mathbf{m} \in \Delta_{KL-1}} |\mathbf{s}(\mathbf{m})| < \infty;$$

where $\partial\mathbf{w} := \left(\frac{\partial w_i}{\partial m_j} \right)_{ij}$ for $i, j \in [\Lambda]$.

These follow immediately from the fact that entries of \mathbf{w} , $\partial\mathbf{w}$, and \mathbf{s} are polynomials on a closed and bounded set.

(v) $\partial\mathbf{w}$ and \mathbf{s} are Lipschitz (again, uniformly in N):

$$\sup_{\mathbf{m}_1 \neq \mathbf{m}_2} \frac{|\partial\mathbf{w}(\mathbf{m}_1) - \partial\mathbf{w}(\mathbf{m}_2)|}{|\mathbf{m}_1 - \mathbf{m}_2|} < \infty; \quad \sup_{\mathbf{m}_1 \neq \mathbf{m}_2} \frac{|\mathbf{s}(\mathbf{m}_1) - \mathbf{s}(\mathbf{m}_2)|}{|\mathbf{m}_1 - \mathbf{m}_2|} < \infty.$$

These hold as in (iv).

(vi) $\gamma^{(N)}$ [equation (13)] maps Δ_{KL-1} into Δ_{KL-1} .

We can verify this directly by solving (13) to yield

$$\gamma^{(N)}(\tau) = \left(\mathbf{X}^{(N)}(0), \mathbf{Y}^{(N)}(0), \mathbf{D}^{(N)}(0) \left(1 - \frac{\rho_\beta}{4N^\beta} \right)^\tau \right)',$$

which is in Δ_{KL-1} provided $0 \leq \frac{\rho_\beta}{4N^\beta} \leq 1$, as must be the case since $r = \rho_\beta/(4N^\beta)$ is a probability.

(vii) $\mathbb{E}[|\mathbf{e}_{1,\tau}^{(N)}|^3]^{1/3} \leq R^{(N)} \sqrt{\epsilon^{(N)} \tau^{(N)}} (= R^{(N)} \frac{1}{2} N^{-(\beta+1)/2} \text{ here}),$ where $R^{(N)}$ is a positive sequence such that $R^{(N)} \rightarrow 0$ as $N \rightarrow \infty$.

From (7)–(9), the entries of $\mathbf{e}_{1,\tau}^{(N)}$ are of $O(N^{-1})$ and are all bounded since coefficients are always polynomials on a closed and bounded set. Thus,

$\mathbb{E}[|\mathbf{e}_{1,\tau}^{(N)}|^3]^{1/3}$ is bounded by a positive sequence $S^{(N)}$ of $O(N^{-1})$, and the requirement is satisfied if we choose $R^{(N)} = S^{(N)}N^{(1+\beta)/2} = O(N^{(\beta-1)/2})$. Then $R^{(N)}$ converges to 0 for $\beta < 1$.

(viii) $\mathbb{E}[|\mathbf{e}_{2,\tau}^{(N)}|] \leq R^{(N)}\tau^{(N)} (= R^{(N)}/(2N) \text{ here})$, where $R^{(N)}$ is a positive sequence such that $R^{(N)} \rightarrow 0$ as $N \rightarrow \infty$.

From Proposition 3.1, the entries of $\mathbf{e}_{2,\tau}^{(N)}$ are of $O(N^{-(1+\beta)})$ and bounded as before. Thus, $\mathbb{E}[|\mathbf{e}_{2,\tau}^{(N)}|]$ is bounded by a positive sequence $S^{(N)}$ of $O(N^{-(1+\beta)})$, and the requirement is satisfied if we choose $R^{(N)} = S^{(N)}N = O(N^{-\beta})$. Then $R^{(N)}$ converges to 0 for $\beta > 0$.

(ix) *The error term*

$$\begin{aligned} \mathbf{e}_{3,\tau}^{(N)} &:= \Delta \mathbf{M}^{(N)} - \mathbb{E}[\Delta \mathbf{M}^{(N)}(\tau) \mid \mathbf{M}^{(N)}(\tau)] \\ &= \mathbf{M}^{(N)}(\tau + 1) - \mathbb{E}[\mathbf{M}^{(N)}(\tau + 1) \mid \mathbf{M}^{(N)}(\tau)] \end{aligned}$$

satisfies $\mathbb{E}[|\mathbf{e}_{3,\tau}^{(N)}|^3]^{1/3} \leq C\sqrt{\tau^{(N)}} (= C/\sqrt{2N} \text{ here})$, for some constant C .

By Minkowski's inequality,

$$\mathbb{E}[|\mathbf{e}_{3,\tau}^{(N)}|^3]^{1/3} \leq \sum_{k=1}^{\Lambda} \mathbb{E}[|(\mathbf{e}_{3,\tau}^{(N)})_k|^3]^{1/3}.$$

Thus, it suffices to consider each component of $\mathbf{e}_{3,\tau}^{(N)}$ separately. The first $K + L$ components are $O(N^{-1/2})$ by Lemma 3.1. For the remaining components involving $D_{ij}^{(N)}$, we invoke Minkowski's inequality again to write

$$\begin{aligned} \mathbb{E}[|(\mathbf{e}_{3,\tau}^{(N)})_k|^3 \mid \mathbf{M}^{(N)}]^{1/3} &\leq \mathbb{E} \left[\left| \frac{Z'_{ij} - \mathbb{E}[Z'_{ij} \mid \mathbf{Z}]}{2N} \right|^3 \mid \mathbf{Z} \right]^{\frac{1}{3}} \\ &\quad + \mathbb{E} \left[\left| \frac{Z'_{i \cdot} Z'_{\cdot j} - \mathbb{E}[Z'_{i \cdot} Z'_{\cdot j} \mid \mathbf{Z}]}{(2N)^2} \right|^3 \mid \mathbf{Z} \right]^{\frac{1}{3}}, \end{aligned}$$

when the k th component in $\mathbf{e}_{3,\tau}^{(N)}$ corresponds to $D_{ij}^{(N)}$. The first term on the right-hand side is $O(N^{-1/2})$, again by Lemma 3.1, while the second term

can be decomposed as

$$\begin{aligned}
\mathbb{E} \left[\left| \frac{Z'_{i \cdot} Z'_{\cdot j} - \mathbb{E}[Z'_{i \cdot} Z'_{\cdot j} | \mathbf{Z}]}{(2N)^2} \right|^3 \mid \mathbf{Z} \right]^{\frac{1}{3}} &\leq \mathbb{E} \left[\left| \frac{Z'_{i \cdot} (Z'_{\cdot j} - \mathbb{E}[Z'_{\cdot j} | \mathbf{Z}])}{(2N)^2} \right|^3 \mid \mathbf{Z} \right]^{\frac{1}{3}} \\
&\quad + \mathbb{E} \left[\left| \frac{\mathbb{E}[Z'_{i \cdot} | \mathbf{Z}] (Z'_{\cdot j} - \mathbb{E}[Z'_{\cdot j} | \mathbf{Z}])}{(2N)^2} \right|^3 \mid \mathbf{Z} \right]^{\frac{1}{3}} \\
&\quad + \mathbb{E} \left[\left| \frac{\mathbb{E}[Z'_{i \cdot} | \mathbf{Z}] \mathbb{E}[Z'_{\cdot j} | \mathbf{Z}] - \mathbb{E}[Z'_{i \cdot} Z'_{\cdot j} | \mathbf{Z}]}{(2N)^2} \right|^3 \mid \mathbf{Z} \right]^{\frac{1}{3}}, \\
&\leq \mathbb{E} \left[\left| \frac{Z'_{\cdot j} - \mathbb{E}[Z'_{\cdot j} | \mathbf{Z}]}{2N} \right|^3 \mid \mathbf{Z} \right]^{\frac{1}{3}} + \mathbb{E} \left[\left| \frac{Z'_{i \cdot} - \mathbb{E}[Z'_{i \cdot} | \mathbf{Z}]}{2N} \right|^3 \mid \mathbf{Z} \right]^{\frac{1}{3}} \\
(17) \quad &\quad + \mathbb{E} \left[\left| \frac{\phi_{ij} - \phi_{i \cdot} \phi_{\cdot j}}{2N} \right|^3 \mid \mathbf{Z} \right]^{\frac{1}{3}}.
\end{aligned}$$

The first two terms on the right of (17) are again $O(N^{-1/2})$ by Lemma 3.1, while the third term is $O(N^{-1})$. Putting all this together we find that

$$\mathbb{E}[|(\mathbf{e}_{3,\tau}^{(N)})_k|^3 \mid \mathbf{M}^{(N)}] = O(N^{-3/2}),$$

and hence $\mathbb{E}[|(\mathbf{e}_{3,\tau}^{(N)})_{D_{ij}^{(N)}}|^3]^{1/3} = O(N^{-1/2})$, as required.

The conclusions of Theorems 1–3 in Norman (1975a) are as in the conclusions in the statement of our theorem, except that $f(t, \mathbf{m})$, $\mathbf{B}(u, \mathbf{m})$, and $\mathbf{\Sigma}(u, \mathbf{m})$ are not given explicitly; they are defined only as the unique solutions of a system of first order ordinary differential equations. In our case these equations can be solved directly, leading to the expressions given in the statement of the theorem; we omit the details. \square

REMARK 3.1. (i) The theory of Norman (1975a) has been used to study strong mutation and selection (Norman, 1972, 1975a; Kaplan, Darden and Hudson, 1988; Nagylaki, 1990; Wakeley and Sargsyan, 2009), but to the best of our knowledge this is the first time it has been used to quantify the effects of high rates of recombination.

(ii) The exponential decay of linkage disequilibrium implied by $\mathbf{B}(u)$ is a classical result; the above theorem further quantifies the fluctuations about this deterministic behaviour in a fully time-dependent manner. In particular, the definition of $\widehat{\mathbf{M}}^{(N)}(t)$ shows that fluctuations are of order $N^{(1-\beta)/2}$ on a timescale of $2N^\beta$ generations.

- (iii) The covariance matrix $\Sigma(u, \mathbf{m})$ is composed of a superposition of “harmonics”, with different terms decaying at different exponential rates. This corresponds to the decomposition of the infinitesimal generator of the Wright-Fisher diffusion studied in [Jenkins and Song \(2012\)](#).

3.3. Stationary distribution. [Norman \(1975b\)](#) provides conditions similar to those of Theorem 3.1 for convergence of a scalar diffusion to a stationary distribution. Those conditions are satisfied by the marginal limiting diffusion corresponding to each $D_{ij}^{(N)}$ but not by the diffusions corresponding to each $X_i^{(N)}$ and $Y_j^{(N)}$, which undergo Brownian motions (with nonunit volatility). Letting $t \rightarrow \infty$ in Theorem 3.1, we see that the marginal diffusion $V_{\tilde{D}_{ij}}^{(N)} := (V_{\tilde{D}_{ij}}^{(N)}(t) : t \geq 0)$ corresponding to $\tilde{D}_{ij}^{(N)}$ has transition probability

$$\mathcal{L}(V_{\tilde{D}_{ij}}^{(N)}(u) \mid V_{\tilde{D}_{ij}}^{(N)}(0) = v) \stackrel{d}{=} \text{Normal} \left(v e^{-\rho_\beta u/2}, \frac{1 - e^{-\rho_\beta u}}{\rho_\beta} \sigma_{ij,ij}^{(N)} \right),$$

where

$$\sigma_{ij,kl}^{(N)} := X_i^{(N)}(0)Y_j^{(N)}(0)[\delta_{ik} - X_k^{(N)}(0)][\delta_{jl} - Y_l^{(N)}(0)].$$

In other words, $V_{\tilde{D}_{ij}}^{(N)}$ is an Ornstein-Uhlenbeck process with damping towards linkage equilibrium at rate $\rho_\beta/2$ and constant volatility $(\sigma_{ij,ij}^{(N)})^{1/2}$. The stationary distribution of such a process is $\text{Normal}(0, \sigma_{ij,ij}^{(N)}/\rho_\beta)$. Similarly, $\mathbf{V}_{\tilde{\mathbf{D}}}^{(N)} = (\mathbf{V}_{\tilde{\mathbf{D}}}^{(N)}(t) : t \geq 0)$ is an Ornstein-Uhlenbeck process marginally along each coordinate, and with stationary distribution

$$(18) \quad \text{Normal} \left(\mathbf{0}_{KL \times 1}, \frac{1}{\rho_\beta} \boldsymbol{\sigma}^{(N)} \right), \quad \boldsymbol{\sigma}^{(N)} := (\sigma_{ij,kl}^{(N)}).$$

Thus, we have *derived* a Gaussian diffusion approximation $\mathbf{V}_{\tilde{\mathbf{D}}}^{(N)}$ for $\tilde{\mathbf{D}}^{(N)}$. We can exploit this to obtain a simple approximation of the usual two-locus *Wright-Fisher* diffusion limit, as follows. From (14), $\mathbf{D}^{(N)} = \tilde{\mathbf{D}}^{(N)} N^{(\beta-1)/2}$ at stationarity, and from (18) its diffusion approximation is normally distributed with mean $\mathbf{0}_{KL \times 1}$ and covariance $N^{\beta-1} \boldsymbol{\sigma}^{(N)}/\rho_\beta = \boldsymbol{\sigma}^{(N)}/\rho$. Notice that this description does not depend on the particular choice of β . Furthermore, under the usual “Wright-Fisher” regime with ρ fixed, it depends on N only through the initial conditions $\mathbf{X}^{(N)}(0)$ and $\mathbf{Y}^{(N)}(0)$. Letting $\mathbf{X}^{(N)}(0) \rightarrow \mathbf{X}(0)$ and $\mathbf{Y}^{(N)}(0) \rightarrow \mathbf{Y}(0)$ as $N \rightarrow \infty$ with ρ fixed, we obtain a limiting diffusion $\mathbf{V}_{\mathbf{D}} = \lim_{N \rightarrow \infty} (N^{(\beta-1)/2} \mathbf{V}_{\tilde{\mathbf{D}}}^{(N)}(t) : t \geq 0)$ with stationary distribution

$$(19) \quad \text{Normal} \left(\mathbf{0}_{KL \times 1}, \frac{1}{\rho} [X_i(0)Y_j(0)[\delta_{ik} - X_k(0)][\delta_{jl} - Y_l(0)]]_{ij,kl} \right).$$

Now further suppose that the marginal allele frequencies, \mathbf{X} and \mathbf{Y} , have reached their (independent) stationary distributions, which we refer to as π_A and π_B , respectively (and whose respective sampling distributions are q^A and q^B). Then we can complete the picture for (19) by specifying $(\mathbf{X}(0), \mathbf{Y}(0)) \sim \pi_A \otimes \pi_B$.

The distribution (19) provides a simple, explicit method for the approximate simulation of haplotype frequencies under a stationary, two-locus Wright-Fisher diffusion, which we summarize in the following algorithm.

Algorithm to simulate from a Gaussian approximation to the stationary Wright-Fisher diffusion with recombination.

1. Simulate marginal allele frequencies at locus A, $\mathbf{X}(0) \sim \pi_A$.
2. Independently simulate marginal allele frequencies at locus B, $\mathbf{Y}(0) \sim \pi_B$.
3. Conditionally simulate \mathbf{D} from (19) given $\mathbf{X}(0)$ and $\mathbf{Y}(0)$.
4. Calculate two-locus haplotype frequencies via

$$X_{ij} = D_{ij} + X_i(0)Y_j(0), \quad \text{for each } i \in [K], j \in [L].$$

When mutation is parent-independent, as in Remark 2.1(i), π_A and π_B take on a particularly simple form, but we note that these distributions are not known in general.

3.4. *Sampling distribution.* The significance of the Gaussian diffusion approximation \mathbf{V}_D is further evident from the following theorem. First we need some further notation. Let

$$\mathcal{P}_m = \left\{ \mathbf{r} \in \mathbb{N}^{K \times L} : \sum_{i=1}^K \sum_{j=1}^L r_{ij} = m \right\},$$

for $m \in \mathbb{N}$, and let $\mathbf{l}^{(r)} \in ([K] \times [L])^m$ denote a sequence of m haplotypes (in some arbitrary, fixed order) with multiplicities specified by $\mathbf{r} \in \mathcal{P}_m$. Further let $\mathbf{l}^{(r)A} \in [K]^m$ denote the corresponding list of alleles obtained by looking at the first entry of each element of $\mathbf{l}^{(r)}$, and define $\mathbf{l}^{(r)B}$ similarly. For $\lambda \in \mathbb{N}$ denote by $\mathcal{Q}_{2\lambda}$ the set of partitions of $[2\lambda]$ with precisely λ blocks of size 2, and write a representative element as $\xi_{\mu\nu} = \{\{\mu_k, \nu_k\} : k = 1, \dots, \lambda\} \in \mathcal{Q}_{2\lambda}$; $\mu = (\mu_k)$ and $\nu = (\nu_k)$ are sequences of length λ . For $J \subseteq [\lambda]$, denote by μ_J, ν_J the subsequences obtained by looking only at the indices in J , and denote by $\mathbf{l}_{\mu}^{(r)}$ the subsequence of $\mathbf{l}^{(r)}$ obtained by looking only at the indices in μ . The matrix of multiplicities of $\mathbf{l}_{\mu}^{(r)}$ is denoted by

$\mathbf{r}^{(\mu)}$, so that $\mathbf{r}^{(\mu)} + \mathbf{r}^{(\nu)} = \mathbf{r}$. For example, if $\mathbf{r} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$ then a representative list of haplotypes is $\mathbf{l}^{(\mathbf{r})} = ((1, 1), (1, 2), (1, 2), (2, 2))$ with marginal allele lists $\mathbf{l}^{(\mathbf{r})A} = (1, 1, 1, 2)$ and $\mathbf{l}^{(\mathbf{r})B} = (1, 2, 2, 2)$. Here, $m = 2\lambda = 4$, and $\mathcal{Q}_4 = \{\{\{1, 2\}, \{3, 4\}\}, \{\{1, 3\}, \{2, 4\}\}, \{\{1, 4\}, \{2, 3\}\}\}$. Then for example the first element in \mathcal{Q}_4 is the partition $\xi_{\mu\nu}$ constructed from $\mu = (1, 3)$ and $\nu = (2, 4)$, and so $\mathbf{l}_\mu^{(\mathbf{r})} = ((1, 1), (1, 2))$ and $\mathbf{l}_\nu^{(\mathbf{r})} = ((1, 2), (2, 2))$.

THEOREM 3.2. *Suppose that $\mathbf{X} \sim \pi_A$, $\mathbf{Y} \sim \pi_B$ independently, and \mathbf{V}_D is stationary according to the Gaussian distribution (19). Then the sampling distribution is given by exactly by*

$$\begin{aligned}
 q_G(\mathbf{a}, \mathbf{b}, \mathbf{c}) &= \sum_{\lambda=0}^{\lfloor c/2 \rfloor} \frac{1}{\rho^\lambda} \sum_{\mathbf{r} \in \mathcal{P}_{2\lambda}} \sum_{\xi \in \mathcal{Q}_{2\lambda}} \left[\prod_{i=1}^K \prod_{j=1}^L \binom{c_{ij}}{r_{ij}} \right] \\
 &\quad \times \left[\sum_{I \subseteq [\lambda]: \mathbf{l}_{\mu_I}^{(\mathbf{r})A} = \mathbf{l}_{\nu_I}^{(\mathbf{r})A}} (-1)^{|I^c|} q^A(\mathbf{a} + \mathbf{c}_A - \mathbf{r}_A^{(\nu_I)}) \right] \\
 (20) \quad &\quad \times \left[\sum_{J \subseteq [\lambda]: \mathbf{l}_{\mu_J}^{(\mathbf{r})B} = \mathbf{l}_{\nu_J}^{(\mathbf{r})B}} (-1)^{|J^c|} q^B(\mathbf{b} + \mathbf{c}_B - \mathbf{r}_B^{(\nu_J)}) \right], \\
 &= q_0(\mathbf{a}, \mathbf{b}, \mathbf{c}) + \frac{q_1(\mathbf{a}, \mathbf{b}, \mathbf{c})}{\rho} + O\left(\frac{1}{\rho^2}\right),
 \end{aligned}$$

with q_0 and q_1 given by (2) and (3) respectively (and we impose the convention that the empty summations for $\lambda = 0$ have a single term, with $(-1)^{|\emptyset \setminus \emptyset|} = 1$).

PROOF. With respect to the diffusion in the transformed co-ordinate system, the sampling distribution is

$$\begin{aligned}
 q_G(\mathbf{a}, \mathbf{b}, \mathbf{c}) &= \mathbb{E} \left[\left(\prod_{i=1}^K X_i^{a_i} \right) \left(\prod_{j=1}^L Y_j^{b_j} \right) \left(\prod_{i=1}^K \prod_{j=1}^L [D_{ij} + X_i Y_j]^{c_{ij}} \right) \right], \\
 &= \sum_{m=0}^c \sum_{\mathbf{r} \in \mathcal{P}_m} \left[\prod_{i=1}^K \prod_{j=1}^L \binom{c_{ij}}{r_{ij}} \right] \mathbb{E} \left[\left(\prod_{i=1}^K X_i^{a_i + c_{i\cdot} - r_{i\cdot}} \right) \right. \\
 &\quad \left. \times \left(\prod_{j=1}^L Y_j^{b_j + c_{\cdot j} - r_{\cdot j}} \right) \mathbb{E} \left[\prod_{i=1}^K \prod_{j=1}^L D_{ij}^{r_{ij}} \mid \mathbf{X}, \mathbf{Y} \right] \right],
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{\lambda=0}^{\lfloor c/2 \rfloor} \sum_{\mathbf{r} \in \mathcal{P}_{2\lambda}} \sum_{\boldsymbol{\xi} \in \mathcal{Q}_{2\lambda}} \left[\prod_{i=1}^K \prod_{j=1}^L \binom{c_{ij}}{r_{ij}} \right] \mathbb{E} \left[\left(\prod_{i=1}^K X_i^{a_i + c_{i\cdot} - r_{i\cdot}} \right) \right. \\
&\quad \left. \times \left(\prod_{j=1}^L Y_j^{b_j + c_{\cdot j} - r_{\cdot j}} \right) \prod_{k=1}^{\lambda} \mathbb{E}[D_{\mathbf{l}_{\mu_k}^{(r)}} D_{\mathbf{l}_{\nu_k}^{(r)}} \mid \mathbf{X}, \mathbf{Y}] \right], \\
&= \sum_{\lambda=0}^{\lfloor c/2 \rfloor} \frac{1}{\rho^{\lambda}} \sum_{\mathbf{r} \in \mathcal{P}_{2\lambda}} \sum_{\boldsymbol{\xi} \in \mathcal{Q}_{2\lambda}} \left[\prod_{i=1}^K \prod_{j=1}^L \binom{c_{ij}}{r_{ij}} \right] \\
&\quad \times \mathbb{E} \left[\left(\prod_{i=1}^K X_i^{a_i + c_{i\cdot} - r_{i\cdot}} \right) \left(\prod_{j=1}^L Y_j^{b_j + c_{\cdot j} - r_{\cdot j}} \right) \right. \\
&\quad \left. \times \prod_{k=1}^{\lambda} X_{\mathbf{l}_{\mu_k}^{(r)A}} Y_{\mathbf{l}_{\mu_k}^{(r)B}} (\delta_{\mathbf{l}_{\mu_k}^{(r)A} \mathbf{l}_{\nu_k}^{(r)A}} - X_{\mathbf{l}_{\nu_k}^{(r)A}}) (\delta_{\mathbf{l}_{\mu_k}^{(r)B} \mathbf{l}_{\nu_k}^{(r)B}} - Y_{\mathbf{l}_{\nu_k}^{(r)B}}) \right], \\
&= \sum_{\lambda=0}^{\lfloor c/2 \rfloor} \frac{1}{\rho^{\lambda}} \sum_{\mathbf{r} \in \mathcal{P}_{2\lambda}} \sum_{\boldsymbol{\xi} \in \mathcal{Q}_{2\lambda}} \left[\prod_{i=1}^K \prod_{j=1}^L \binom{c_{ij}}{r_{ij}} \right] \\
&\quad \times \sum_{I \subseteq [\lambda]} (-1)^{|I^c|} \delta_{\mathbf{l}_{\mu_I}^{(r)A} \mathbf{l}_{\nu_I}^{(r)A}} \sum_{J \subseteq [\lambda]} (-1)^{|J^c|} \delta_{\mathbf{l}_{\mu_J}^{(r)B} \mathbf{l}_{\nu_J}^{(r)B}} \\
&\quad \times \mathbb{E} \left[\left(\prod_{i=1}^K X_i^{a_i + c_{i\cdot} - r_{i\cdot}^{(\nu_I)}} \right) \left(\prod_{j=1}^L Y_j^{b_j + c_{\cdot j} - r_{\cdot j}^{(\nu_J)}} \right) \right],
\end{aligned}$$

The second equality follows from the multinomial theorem and the tower property, the third equality follows from Isserlis' theorem ([Michałowicz et al., 2011](#)), and the fourth equality follows from [\(19\)](#):

$$\mathbb{E}[D_{ij} D_{kl} \mid \mathbf{X}, \mathbf{Y}] = \frac{1}{\rho} X_i Y_j (\delta_{ik} - X_k) (\delta_{jl} - Y_l).$$

The fifth equality follows from expanding the final product (using the convention $\delta_{\emptyset \emptyset} = 1$), while [\(20\)](#) follows from $(\mathbf{X}, \mathbf{Y}) \sim \pi_A \otimes \pi_B$. The equalities still hold for $\lambda = 0$ provided we take $\prod_{\emptyset} = 1$.

Extracting the two leading order terms $\lambda = 0$ and $\lambda = 1$, the expression simplifies to

$$q_G(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \mathbb{E} \left[\left(\prod_{i=1}^K X_i^{a_i + c_{i\cdot}} \right) \left(\prod_{j=1}^L Y_j^{b_j + c_{\cdot j}} \right) \right]$$

$$\begin{aligned}
& + \frac{1}{\rho} \sum_{k,u=1}^K \sum_{l,v=1}^L \frac{c_{kl}(c_{uv} - \delta_{ku}\delta_{lv})}{2} \mathbb{E} \left[\left(\prod_{i=1}^K X_i^{a_i+c_i-\delta_{iu}} \right) \right. \\
& \quad \left. \times \left(\prod_{j=1}^L Y_j^{b_j+c_j-\delta_{jv}} \right) (\delta_{ku} - X_u)(\delta_{lv} - Y_v) \right], \\
& + O\left(\frac{1}{\rho^2}\right), \\
& = q_0(\mathbf{a}, \mathbf{b}, \mathbf{c}) + \frac{q_1(\mathbf{a}, \mathbf{b}, \mathbf{c})}{\rho} + O\left(\frac{1}{\rho^2}\right),
\end{aligned}$$

as required. \square

3.5. Accuracy of the diffusion process. A natural question to ask is: to what extent does the process of Theorem 3.2 capture the dynamics of the full process? To address this we consider the accuracy of the sampling distribution (20) as an approximation to the “true” distribution, $q(\mathbf{a}, \mathbf{b}, \mathbf{c})$. For moderate sample sizes it is possible to compute the latter as the solution to a system of recursive equations (Golding, 1984; Ethier and Griffiths, 1990; Jenkins and Song, 2009). The number of summands in (20) grows rapidly with λ (as long as $\lambda \leq \lfloor \frac{c}{2} \rfloor$), so we define an approximate sampling distribution $q_G^{(\lambda)}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ by truncating the outer sum in (20) at a fixed index λ . This is analogous to the asymptotic sampling formulae for the full model which are obtained by truncating equation (1) (Jenkins and Song, 2012). As our measure of accuracy we define the relative error,

$$(21) \quad \hat{e}_{\text{Gaussian}}^{(\lambda)} = \left| \frac{Q_G^{(\lambda)}(\mathbf{0}, \mathbf{0}, \mathbf{c}) - q(\mathbf{0}, \mathbf{0}, \mathbf{c})}{q(\mathbf{0}, \mathbf{0}, \mathbf{c})} \right| \times 100\%,$$

where $Q_G^{(\lambda)}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ is the staircase Padé approximant to $q_G^{(\lambda)}(\mathbf{0}, \mathbf{0}, \mathbf{c})$. (The former is used for its superior convergence properties; see Jenkins and Song, 2012, for details.) We define $\hat{e}_{\text{True}}^{(\lambda)}$ analogously, replacing $Q_G^{(\lambda)}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ in (21) with the Padé approximant to the partial sum of (1), computed up to $O(\rho^{-(\lambda+1)})$ by the method of Jenkins and Song (2012).

We computed the distribution of $\hat{e}_{\text{Gaussian}}^{(\lambda)}$ and of $\hat{e}_{\text{True}}^{(\lambda)}$ across all sample configurations of size $c = 20$ for which both alleles are observed at each locus; results are shown in Table 1. For a collection of this size it was straightforward to compute up to $\lambda = 6$ for every possible sample configuration. Using a partial sum to approximate (1) contributes to both errors; $\hat{e}_{\text{Gaussian}}^{(\lambda)}$ has additional contributions reflecting its use of an approximate *model*. Of

TABLE 1
Cumulative distribution $\Phi(x) = \mathbb{P}(\hat{e}^{(\lambda)} < x\%)$ (where $\hat{e}^{(\lambda)}$ denotes either $\hat{e}_{\text{Gaussian}}^{(\lambda)}$ or $\hat{e}_{\text{True}}^{(\lambda)}$ as defined in the main text), for all samples of size 20 dimorphic at both loci.

		$\rho = 25$			$\rho = 50$		
λ	Type of sum	$\Phi(1)$	$\Phi(10)$	$\Phi(100)$	$\Phi(1)$	$\Phi(10)$	$\Phi(100)$
0	True	0.39	0.58	1.00	0.49	0.63	1.00
	Gaussian	0.39	0.58	1.00	0.49	0.63	1.00
1	True	0.51	0.75	0.96	0.59	0.84	0.99
	Gaussian	0.51	0.75	0.96	0.59	0.84	0.99
2	True	0.59	0.91	0.97	0.77	0.98	1.00
	Gaussian	0.50	0.73	0.97	0.50	0.86	1.00
4	True	0.83	0.99	1.00	0.95	1.00	1.00
	Gaussian	0.51	0.72	1.00	0.50	0.80	1.00
6	True	0.89	0.99	1.00	0.99	1.00	1.00
	Gaussian	0.49	0.71	0.99	0.50	0.79	1.00

		$\rho = 100$			$\rho = 200$		
λ	Type of sum	$\Phi(1)$	$\Phi(10)$	$\Phi(100)$	$\Phi(1)$	$\Phi(10)$	$\Phi(100)$
0	True	0.50	0.72	1.00	0.54	0.95	1.00
	Gaussian	0.50	0.72	1.00	0.54	0.95	1.00
1	True	0.74	0.95	1.00	0.90	0.99	1.00
	Gaussian	0.74	0.95	1.00	0.90	0.99	1.00
2	True	0.95	1.00	1.00	1.00	1.00	1.00
	Gaussian	0.64	0.99	1.00	0.85	1.00	1.00
4	True	1.00	1.00	1.00	1.00	1.00	1.00
	Gaussian	0.64	0.99	1.00	0.83	1.00	1.00
6	True	1.00	1.00	1.00	1.00	1.00	1.00
	Gaussian	0.64	0.99	1.00	0.83	1.00	1.00

course, the two errors agree up to $\lambda = 1$. However, Table 1 shows that they are comparable more broadly, particularly for large recombination rates. As λ increases, $Q_G^{(\lambda)}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ converges rapidly (even without Padé summation; not shown), and becomes a reasonable approximation to $q(\mathbf{a}, \mathbf{b}, \mathbf{c})$. For example, for $\rho = 50$, $Q_G^{(6)}(\mathbf{0}, \mathbf{0}, \mathbf{c})$ is within 10% of $q(\mathbf{a}, \mathbf{b}, \mathbf{c})$ with probability 0.79, though it is within 1% only with probability 0.50. When we consider the highest levels of accuracy, as in $\Phi(1)$ in Table 1, $\hat{e}_{\text{Gaussian}}^{(\lambda)}$ actually increases with λ when $\lambda > 1$. This suggests that the Gaussian model typically cannot approximate the true model to the same level of precision as a first order asymptotic approximation of the true model, though its behaviour as a coarser approximation (as reflected in the columns for $\Phi(100)$, for example) is comparable.

4. Coalescent process.

4.1. *A coupling argument.* In this section we derive a coalescent process which is much simpler than the ARG but whose sampling distribution agrees with (2) and (3). Let $\mathcal{C}_{a,b,c}^{(\rho)}(t)$ denote the standard, neutral, two-locus coalescent process a time t back from a sample taken at time $t = 0$, with a , b , and c counting the three types of sample as defined in Section 2. Lineages ancestral to the three types are sometimes referred to as representing *left half-fragments*, *right half-fragments*, and *full fragments*, respectively. Our strategy is to define a coupling on a joint probability space for the pair of processes $(\mathcal{C}^{(\rho)} = (\mathcal{C}_{a,b,c}^{(\rho)}(t) : t \geq 0), \mathcal{D}^{(\infty)} = (\mathcal{D}_{a,b,c}^{(\infty)}(t) : t \geq 0))$, where $\mathcal{D}^{(\infty)}$ is a simple process closely related to $\mathcal{C}^{(\infty)}$ and defined below. $\mathcal{C}^{(\rho)}(\omega)$ is said to be coupled to $\mathcal{D}^{(\infty)}(\omega)$ if the two realizations have the same marginal coalescent tree at locus A and the same marginal coalescent tree at locus B. Since it is the marginal trees which govern the mutation process at each locus, coupled processes therefore have the same sampling distribution. (There should be no ambiguity arising from the fact that our coupling is not on pairs of realizations but on pairs of equivalence classes, where an equivalence class of $\mathcal{C}^{(\rho)}$ or of $\mathcal{D}^{(\infty)}$ is a set of realizations with the same marginal tree at locus A and the same marginal tree at locus B.)

A complete description of a coalescent process is one taking values in partitions of $[n]$, as introduced by Kingman (1982), with natural extensions to incorporate recombination. We opt instead to represent $\mathcal{C}^{(\rho)}$ only by its *ancestral* process; that is, as a birth-death process on the number of each type of lineage. Such a process is studied in depth by Ethier and Griffiths (1990) and Griffiths (1991). In what follows it is understood implicitly that for any given realization of the ancestral process one could reconstruct a complete coalescent process—an ARG—given some additional independent randomness. Provided the ancestral processes of $\mathcal{C}^{(\rho)}$ and $\mathcal{D}^{(\infty)}$ remain coupled, then it is also always possible to couple their respective *coalescent* processes. For example, a decrease by one in the ancestral process corresponds to a coalescence event in the coalescent process, which can be realized by merging two uniformly chosen blocks in the partition of $[n]$. A coupling of two *ancestral* processes lets us couple the corresponding *coalescent* processes if we always pick the same pair of blocks to merge in the two processes. With this kept in mind, it is sufficient for the argument developed below to consider the simpler ancestral process representation. We define such a process on \mathbb{N}^4 ,

with $\mathcal{C}_{a,b,c}^{(\rho)}(0) = (a, b, c, c)$ and infinitesimal generator

$$(22) \quad \begin{aligned} \mathcal{L}f(a, b, c, c) &= \frac{\rho c}{2} f(a+1, b+1, c-1, c-1) + \binom{c}{2} f(a, b, c-1, c-1) \\ &+ R_{a,b,c,c} \mathcal{G}f(a, b, c, c) - \left[\frac{\rho c}{2} + \binom{c}{2} + R_{a,b,c,c} \right] f(a, b, c, c), \end{aligned}$$

where

$$\begin{aligned} R_{a,b,c,d} &= ab + ac + bd + \binom{a}{2} + \binom{b}{2}, \\ \mathcal{G}f(a, b, c, d) &= \frac{1}{2R_{a,b,c,d}} [2abf(a-1, b-1, c+1, d+1) \\ &+ a(a+2c-1)f(a-1, b, c, c) \\ &+ b(b+2d-1)f(a, b-1, c, d)], \end{aligned}$$

and $f : \mathbb{N}^4 \rightarrow \mathbb{R}$ is an appropriate test function. Regard the third and fourth entries in f as the number of left- and right- halves of full fragments; these entries are always equal. This seemingly redundant representation will make the coupling with the corresponding process $\mathcal{D}^{(\infty)}$ transparent, as follows. Ordinarily, $\mathcal{C}_{a,b,c}^{(\infty)}(0)$ moves instantaneously to the state $\mathcal{C}_{a+c, b+c, 0}^{(\infty)}(0+)$ and evolves thereafter according to $\mathcal{L}f(a+c, b+c, 0, 0)$. However, we will write this instead as a process initiated at (a, b, c, c) and evolving according to

$$(23) \quad \begin{aligned} \mathcal{L}^{(\infty)}f(a, b, c, d) &= \binom{c}{2} f(a, b, c-1, d) + \binom{d}{2} f(a, b, c, d-1) \\ &+ R_{a,b,c,d} \mathcal{G}f(a, b, c, d) - \left[\binom{c}{2} + \binom{d}{2} + R_{a,b,c,d} \right] f(a, b, c, d), \end{aligned}$$

which describes exactly the same process except that we commence by separately tracking half-fragment lineages that *originated* as full fragments. Finally, we introduce an *artificial* recombination process into $\mathcal{C}^{(\infty)}$ by defining the process $\mathcal{D}^{(\infty)}$ with $\mathcal{D}_{a,b,c}^{(\infty)}(0) = (a, b, c, c)$ and generator

$$(24) \quad \begin{aligned} \mathcal{H}^{(\infty)}f(a, b, c, d) &:= \mathcal{L}^{(\infty)}f(a, b, c, d) \\ &+ \frac{\rho \max\{c, d\}}{2} [f(a+1, b+1, \max\{c-1, 0\}, \max\{d-1, 0\}) - f(a, b, c, d)]. \end{aligned}$$

This artificial process does not affect the distribution of the marginal coalescent trees, so $\mathcal{C}^{(\infty)}$ and $\mathcal{D}^{(\infty)}$ have the same sampling distribution.

To summarize, we have defined two birth-death processes on \mathbb{N}^4 , $(\mathcal{C}_{a,b,c}^{(\rho)}(t) : t \geq 0)$ and $(\mathcal{D}_{a,b,c}^{(\infty)}(t) : t \geq 0)$, which describe two-locus ancestral processes going backwards in time and with respective generators \mathcal{L} and $\mathcal{H}^{(\infty)}$. \mathcal{L} is the generator of a standard process with recombination parameter ρ . $\mathcal{H}^{(\infty)}$ is the generator of a standard process with recombination parameter ∞ and with the additional properties that left half-fragments are recorded in two categories (of multiplicity a and c), right half-fragments are recorded in two categories (of multiplicity b and d), and there is an artificial movement of pairs from the latter to the former as if they were still full fragments. This somewhat contrived definition has an important advantage: it is a simple matter to attempt to couple the two processes by matching each kind of event in the two generators whenever possible. A recombination event in $\mathcal{C}_{a,b,c}^{(\rho)}(t)$ can be matched by an artificial recombination event in $\mathcal{D}_{a,b,c}^{(\infty)}(t)$, a coalescence $(a, b, c, c) \mapsto (a-1, b, c, c)$ in $\mathcal{C}_{a,b,c}^{(\rho)}(t)$ can be matched in $\mathcal{D}_{a,b,c}^{(\infty)}(t)$, and so on.

The aforementioned description is a probabilistic coupling, which may or may not succeed since not all events can be paired off in this way. Comparing (22) and (24), we see that a coupling will fail if there is a transition $(a, b, c, c) \mapsto (a, b, c-1, c-1)$ in $\mathcal{C}^{(\rho)}$ or if either of the transitions $(a, b, c, c) \mapsto (a, b, c-1, c)$ or $(a, b, c, c) \mapsto (a, b, c, c-1)$ occurs in $\mathcal{D}^{(\infty)}$. Define the failure times

$$\begin{aligned} T_{a,b,c}^{(1)} &:= \inf\{t \geq 0 : \mathcal{C}_{a,b,c}^{(\rho)}(t) = \mathcal{C}_{a,b,c}^{(\rho)}(t-) - (0, 0, 1, 1)\}, \\ T_{a,b,c}^{(2)} &:= \inf\{t \geq 0 : \mathcal{D}_{a,b,c}^{(\infty)}(t) = \mathcal{D}_{a,b,c}^{(\infty)}(t-) - (0, 0, 1, 0)\}, \\ T_{a,b,c}^{(3)} &:= \inf\{t \geq 0 : \mathcal{D}_{a,b,c}^{(\infty)}(t) = \mathcal{D}_{a,b,c}^{(\infty)}(t-) - (0, 0, 0, 1)\}, \end{aligned}$$

and

$$\begin{aligned} T_{a,b,c}^{\text{MRCA}} &:= \inf \left\{ t \geq 0 : \mathcal{C}_{a,b,c}^{(\rho)}(s) = \mathcal{D}_{a,b,c}^{(\infty)}(s) \quad \forall s \leq t, \right. \\ &\quad \left. \mathcal{C}_{a,b,c}^{(\rho)}(t) \in \{(1, 1, 0, 0), (0, 0, 1, 1)\} \right\}, \end{aligned}$$

the first time that both loci find a most recent common ancestor in the coupled processes (with the convention $\inf \emptyset = \infty$). If $T_{a,b,c}^{\text{MRCA}} < \min\{T_{a,b,c}^{(1)}, T_{a,b,c}^{(2)}, T_{a,b,c}^{(3)}\}$, we say that the coupling has been *successful*.

LEMMA 4.1. *If $c \in \{0, 1\}$, the coupling between $\mathcal{C}^{(\rho)}$ and $\mathcal{D}^{(\infty)}$ fails with probability $O(\rho^{-2})$, as $\rho \rightarrow \infty$.*

PROOF. The three events causing the coupling to fail occur at rates proportional to $\binom{c}{2}$ and thus require $c \geq 2$. For the pair $(\mathcal{C}_{a,b,1}^{(\rho)}, \mathcal{D}_{a,b,1}^{(\infty)})$, we therefore first need to see a transition of the form $(a', b', 1, 1) \mapsto (a' - 1, b' - 1, 2, 2)$ for some a', b' , followed by one of the transitions causing the coupling to fail. Reading off the rates from the generators, each of these transitions occurs with probability $O(\rho^{-1})$. The case $c = 0$ is similar, first needing a transition of the form $(a', b', 0, 0) \mapsto (a' - 1, b' - 1, 1, 1)$ whose probability is of $O(1)$. \square

LEMMA 4.2. *The coupling between $\mathcal{C}^{(\rho)}$ and $\mathcal{D}^{(\infty)}$ fails with the following probabilities:*

$$(25) \quad \mathbb{P}(I^{(k)}) = \frac{1}{\rho} \binom{c}{2} + O\left(\frac{1}{\rho^2}\right) \quad \text{as } \rho \rightarrow \infty, \quad k = 1, 2, 3,$$

where $I^{(k)} := \{T_{a,b,c}^{(k)} < T_{a,b,c}^{MRCA}\}$. Moreover, $\mathbb{P}(I^{(k_1)} \cap I^{(k_2)}) = O(\rho^{-2})$ for $k_1 \neq k_2$.

PROOF. For $k = 1$, by Lemma 4.1 it is enough to show that

$$\mathbb{P}(T_{a,b,c}^{(1)} < U_{a,b,c}^{(1)}) = \frac{1}{\rho} \binom{c}{2} + O\left(\frac{1}{\rho^2}\right),$$

where

$$U_{a,b,c}^{(1)} := \inf \left\{ t \geq 0 : \mathcal{C}_{a,b,c}^{(\rho)}(t) \in \{(a', b', 0, 0) : a', b' \in \mathbb{N}\} \right\}$$

is the first time $\mathcal{C}^{(\rho)}$ reaches $c = 0$. We proceed by induction on c ; Lemma 4.1 provides the base cases $c \in \{0, 1\}$. First note that for any $c \geq 1$,

$$(26) \quad \mathbb{P}(T_{a,b,c}^{(1)} < U_{a,b,c}^{(1)}) = O\left(\frac{1}{\rho}\right),$$

since this event requires at least one transition that is not a recombination. Reading off the relevant probabilities from (22), we have for $c \geq 2$:

$$\begin{aligned} \mathbb{P}(T_{a,b,c}^{(1)} < U_{a,b,c}^{(1)}) &= \frac{\frac{\rho c}{2}}{\frac{\rho c}{2} + \binom{c}{2} + R_{a,b,c,c}} \cdot \mathbb{P}(T_{a+1,b+1,c-1}^{(1)} < U_{a+1,b+1,c-1}^{(1)}) \\ &\quad + \frac{ab}{\frac{\rho c}{2} + \binom{c}{2} + R_{a,b,c,c}} \cdot \mathbb{P}(T_{a-1,b-1,c+1}^{(1)} < T_{a-1,b-1,c+1}^{(0)}) \\ &\quad + \frac{\binom{c}{2}}{\frac{\rho c}{2} + \binom{c}{2} + R_{a,b,c,c}} \cdot 1 + O\left(\frac{1}{\rho^2}\right), \\ &= \frac{1}{\rho} \binom{c}{2} + O\left(\frac{1}{\rho^2}\right), \end{aligned}$$

by the inductive hypothesis and using (26). By considering

$$U_{a,b,c}^{(k)} := \inf \left\{ t \geq 0 : \mathcal{D}_{a,b,c}^{(\infty)}(t) \in \{(a', b', 0, 0) : a', b' \in \mathbb{N}\} \right\}, \quad k = 2, 3,$$

the cases $k = 2, 3$ are similar. $\mathbb{P}(I^{(k_1)} \cap I^{(k_2)}) = O(\rho^{-2})$ also follows from the fact that this event requires at least two transitions which are not recombinations during the time that $c > 0$. \square

Should the coupling fail, we can say much about the sequence of events prior to $U_{a,b,c}^{(k)}$. Intuitively, the probability that *more than* one transition other than recombinations occurs is $O(\rho^{-2})$. To make this precise we denote by $\mathcal{S}_{a,b,c}^{(k)}(t)$ the jump chain up to time t of $\mathcal{C}^{(\rho)}$ if $k = 1$ and of $\mathcal{D}^{(\infty)}$ if $k = 2, 3$.

LEMMA 4.3. *Let $\mathcal{S}_{a,b,c}$ denote the set of jump chains comprising sequences which start at (a, b, c, c) , end at the first entry of the form $(a', b', 0, 0)$, $a', b' \in \mathbb{N}$, and with all transitions corresponding to recombination events, except for possibly one transition. Then*

$$\mathbb{P}(\mathcal{S}_{a,b,c}^{(k)}(U_{a,b,c}^{(k)}) \in \mathcal{S}_{a,b,c} \mid I^{(k)}) = 1 - O\left(\frac{1}{\rho}\right) \quad \text{as } \rho \rightarrow \infty, \quad k = 1, 2, 3.$$

PROOF. The non-recombination event causing $I^{(k)}$ occurs at time $T_{a,b,c}^{(k)}$. Inspection of the generators (22) and (24) shows that any further transition other than a recombination occurs with probability $O(\rho^{-1})$ during the time that $c > 0$. \square

Recall that our purpose is to obtain the sampling distribution for $\mathcal{C}^{(\rho)}$. For successful couplings, this is easy to obtain since it is the same as that of $\mathcal{D}^{(\infty)}$ and hence $\mathcal{C}^{(\infty)}$; thus $\mathcal{C}^{(\rho)} \mid I^{(1)\complement}$ has the same sampling distribution as $\mathcal{D}^{(\infty)} \mid (I^{(2)} \cup I^{(3)})^{\complement}$. Even if the coupling fails, Lemmata 4.1 and 4.3, demonstrate that the behaviour of $\mathcal{C}^{(\rho)}$ is still predictable enough to recover its sampling distribution up to $O(\rho^{-2})$. Roughly [up to $O(\rho^{-2})$], Lemma 4.3 says: if there is an event that causes the coupling to fail then this is the *only* non-recombination event in the failing process before $U_{a,b,c}^{(k)}$; by Lemma 4.1, if it has not failed by $U_{a,b,c}^{(k)}$ then the coupling will not fail after $U_{a,b,c}^{(k)}$.

The following theorem is proven in Jenkins and Song (2009); however, the following proof gives a coherent, *process-level* explanation for the result.

THEOREM 4.1. *Expressing the sampling distribution for $(\mathcal{C}_{a,b,c}^{(\rho)}(t) : t \geq 0)$ as in (1), the first two terms are given by (2) and (3).*

PROOF. Denote by $q_{\mathcal{C}(\rho)|I^{(1)}}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ the sampling distribution of the process $\mathcal{C}(\rho) \mid I^{(1)}$. By Lemmata 4.1 and 4.3, this sampling distribution is obtained up to $O(\rho^{-1})$ by picking a pair of full fragments at random to coalesce, with the remaining $c - 1$ fragments all undergoing recombination, and subsequently running the process as $\mathcal{D}_{a+c-1, b+c-1, 0}^{(\infty)} \stackrel{a.s.}{=} \mathcal{C}_{a+c-1, b+c-1, 0}^{(\infty)}$. Hence,

$$\begin{aligned} q_{\mathcal{C}(\rho)|I^{(1)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) &= \sum_{i=1}^K \sum_{j=1}^L \frac{\binom{c_{ij}}{2}}{\binom{c}{2}} q_{\mathcal{C}^{(\infty)}}(\mathbf{a}, \mathbf{b}, \mathbf{c} - \mathbf{e}_{ij}) + O\left(\frac{1}{\rho}\right), \\ (27) \quad &= \sum_{i=1}^K \sum_{j=1}^L \frac{\binom{c_{ij}}{2}}{\binom{c}{2}} q^A(\mathbf{a} + \mathbf{c}_A - \mathbf{e}_i) q^B(\mathbf{b} + \mathbf{c}_B - \mathbf{e}_j) + O\left(\frac{1}{\rho}\right). \end{aligned}$$

(We can also ignore the possibility of mutation prior to $U_{a,b,c}^{(1)}$ since, by the same argument as in Lemma 4.3, a mutation occurs during this phase with probability $O(\rho^{-1})$.) Similarly,

$$\begin{aligned} q_{\mathcal{D}^{(\infty)}|I^{(2)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) &= \sum_{i=1}^K \frac{\binom{c_{i\cdot}}{2}}{\binom{c}{2}} q_{\mathcal{C}^{(\infty)}}(\mathbf{a} + \mathbf{c}_A - \mathbf{e}_i, \mathbf{b} + \mathbf{c}_B, \mathbf{0}) + O\left(\frac{1}{\rho}\right), \\ (28) \quad &= \sum_{i=1}^K \frac{\binom{c_{i\cdot}}{2}}{\binom{c}{2}} q^A(\mathbf{a} + \mathbf{c}_A - \mathbf{e}_i) q^B(\mathbf{b} + \mathbf{c}_B) + O\left(\frac{1}{\rho}\right), \end{aligned}$$

$$\begin{aligned} q_{\mathcal{D}^{(\infty)}|I^{(3)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) &= \sum_{j=1}^L \frac{\binom{c_{\cdot j}}{2}}{\binom{c}{2}} q_{\mathcal{C}^{(\infty)}}(\mathbf{a} + \mathbf{c}_A, \mathbf{b} + \mathbf{c}_B - \mathbf{e}_j, \mathbf{0}) + O\left(\frac{1}{\rho}\right), \\ (29) \quad &= \sum_{j=1}^L \frac{\binom{c_{\cdot j}}{2}}{\binom{c}{2}} q^A(\mathbf{a} + \mathbf{c}_A) q^B(\mathbf{b} + \mathbf{c}_B - \mathbf{e}_j) + O\left(\frac{1}{\rho}\right), \end{aligned}$$

and so, together with Lemma 4.2 and the observation that

$$\begin{aligned} \mathbb{P}([I^{(2)} \cup I^{(3)}]^c) q_{\mathcal{D}^{(\infty)}|(I^{(2)} \cup I^{(3)})^c}(\mathbf{a}, \mathbf{b}, \mathbf{c}) &= q_{\mathcal{D}^{(\infty)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) \\ &\quad - \mathbb{P}(I^{(2)}) q_{\mathcal{D}^{(\infty)}|I^{(2)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) - \mathbb{P}(I^{(3)}) q_{\mathcal{D}^{(\infty)}|I^{(3)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) + O(\rho^{-2}), \end{aligned}$$

we obtain

$$\begin{aligned} (30) \quad q_{\mathcal{D}^{(\infty)}|(I^{(2)} \cup I^{(3)})^c}(\mathbf{a}, \mathbf{b}, \mathbf{c}) &= \left[1 + \frac{2}{\rho} \binom{c}{2}\right] \left[q_{\mathcal{D}^{(\infty)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) \right. \\ &\quad \left. - \frac{1}{\rho} \binom{c}{2} q_{\mathcal{D}^{(\infty)}|I^{(2)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) - \frac{1}{\rho} \binom{c}{2} q_{\mathcal{D}^{(\infty)}|I^{(3)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) \right] + O\left(\frac{1}{\rho^2}\right). \end{aligned}$$

The key decomposition is then

$$\begin{aligned}
 q(\mathbf{a}, \mathbf{b}, \mathbf{c}) &= \mathbb{P}(I^{(1)})q_{\mathcal{C}^{(\rho)}|I^{(1)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) + \mathbb{P}(I^{(1)\mathbb{L}})q_{\mathcal{C}^{(\rho)}|I^{(1)\mathbb{L}}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) \\
 (31) \quad &= \mathbb{P}(I^{(1)})q_{\mathcal{C}^{(\rho)}|I^{(1)}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) + \mathbb{P}(I^{(1)\mathbb{L}})q_{\mathcal{D}^{(\infty)}|(I^{(2)} \cup I^{(3)})^{\mathbb{L}}}(\mathbf{a}, \mathbf{b}, \mathbf{c}) \\
 &= q_0(\mathbf{a}, \mathbf{b}, \mathbf{c}) + \frac{1}{\rho}q_1(\mathbf{a}, \mathbf{b}, \mathbf{c}) + O\left(\frac{1}{\rho^2}\right),
 \end{aligned}$$

using (25), (27), (28), (29), and (30), with q_0, q_1 given by (2) and (3), respectively. \square

4.2. *A new “loose-linkage” coalescent process.* Equation (31) tells us that, up to $O(\rho^{-2})$, we can obtain the correct sampling distribution using the mixture

$$\alpha[\mathcal{C}^{(\rho)} | I^{(1)}] + (1 - \alpha)[\mathcal{D}^{(\infty)} | (I^{(2)} \cup I^{(3)})^{\mathbb{L}}], \quad \alpha = \frac{1}{\rho} \binom{c}{2},$$

provided $\alpha < 1$. The coupling used to prove Theorem 4.1 demonstrates that we can *define* a simple stochastic process, $\mathcal{E}^{(\rho)}$, as follows, whose sampling distribution agrees with (2) and (3) up to $O(\rho^{-2})$.

Algorithm to simulate $\mathcal{E}^{(\rho)}$, the *loose-linkage coalescent*.

1. With probability α , choose a pair uniformly at random from the c full fragments to coalesce, and then choose uniformly from the chains in $\mathcal{S}_{a,b,c}$ compatible with $I^{(1)}$. Such chains are some permutation of a sequence corresponding to this sole coalescence and $c - 1$ recombinations. Inter-event *times* up to $U_{a,b,c}^{(1)}$ can be sampled according to the rates specified in (22). Go to step 3.
2. Otherwise (w.p. $1 - \alpha$), sample from $\mathcal{D}^{(\infty)} | (I^{(2)} \cup I^{(3)})^{\mathbb{L}}$ up to time $U_{a,b,c}^{(2)} (= U_{a,b,c}^{(3)})$, which can be achieved by running $\mathcal{D}^{(\infty)}$ as usual according to (24) but banning transitions of the form $(a, b, c, d) \mapsto (a, b, c - 1, d)$ and $(a, b, c, d) \mapsto (a, b, c, d - 1)$. (The rates of these transitions still contribute to the overall rate governing inter-event times, however.) Go to step 3.
3. Beyond time $U_{a,b,c}^{(k)}$ ($k = 1$ in the first case above and $k = 2$ in the second), construct the remainder of the process independently using $(\mathcal{C}^{(\infty)}(t - U_{a,b,c}^{(k)}) : t \geq U_{a,b,c}^{(k)})$ (with the appropriate starting configuration) back to the first time both loci have found a most recent common ancestor.

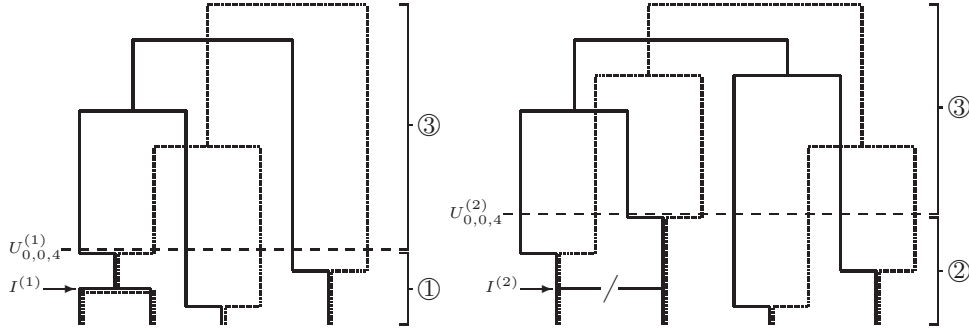


FIG 1. Sampling from the loose-linkage coalescent, $\mathcal{E}^{(\rho)}$, from an initial configuration $(0, 0, 4)$. Steps of the algorithm in the main text are denoted by circled numbers. Left: Commence from step 1 (probability α). Step 1 samples from an approximation to $\mathcal{C}^{(\rho)} \mid I^{(1)}$ which is correct to $O(\rho^{-2})$, back as far as time $U_{0,0,4}^{(1)}$. The jump chain sampled here is $\mathcal{S}_{0,0,4}^{(1)}(U_{0,0,4}^{(1)}) = ((0, 0, 4, 4), (1, 1, 3, 3), (1, 1, 2, 2), (2, 2, 1, 1), (3, 3, 0, 0))$. Thereafter (step 3) the sample is constructed from $\mathcal{C}_{3,3,0}^{(\infty)}(t - U_{0,0,4}^{(1)})$. Right: Commence from step 2 (probability $1 - \alpha$). Step 2 samples from $\mathcal{D}_{0,0,4}^{(\infty)}(t) \mid (I^{(2)} \cup I^{(3)})^c$; a transition which would cause $I^{(2)}$ is banned. Thereafter (step 3) the sample is constructed from $\mathcal{C}_{4,4,0}^{(\infty)}(t - U_{0,0,4}^{(2)})$.

An example is shown in Figure 1. Simulation and inference under $\mathcal{E}^{(\rho)}$ should be straightforward, since its dynamics are little more complicated than those of a coalescent process with $\rho = \infty$. Unlike our diffusion process of Section 3, it does not seem easy to write down its sampling distribution to all orders in closed-form, since that of $\mathcal{D}^{(\infty)} \mid (I^{(2)} \cup I^{(3)})^c$ is not so obvious.

5. Discussion. We have described two novel stochastic models of evolution for loosely linked loci, using both diffusion- and coalescent-based arguments. As a consequence we have obtained deep insight into the simple form of the asymptotic sampling formula given by (2) and (3). Our diffusion model is based on an approach of Norman (1975a), which may be viewed as a separation of the timescales N^β and N , for $0 < \beta < 1$. This contrasts with most research in this area, which focuses on the timescales $N^0 = 1$ and N . Indeed, both diffusion (Ethier and Nagylaki, 1980, 1988) and coalescent [Möhle (1998), Wakeley (2008, Ch. 6)] limits of this latter regime have been studied in detail. It is also the setting of the “loose linkage” limit of Ethier and Nagylaki (1989). Our usage of “loose linkage” therefore refers to a scaling intermediate between the usual Wright-Fisher diffusion and that of Ethier and Nagylaki (1989). That the pioneering approach of Norman (1975a) to investigate recombination does not seem to have been considered until now supports the observation that his work is “somewhat neglected” (Wakeley, 2005). It would also be of interest to find a coalescent-based ana-

logue of [Norman \(1975a\)](#) along the lines of [Möhle \(1998\)](#), or even a duality relationship in the manner of [Etheridge and Griffiths \(2009\)](#).

For simplicity we have focused on a two-locus, finite-alleles, neutral model. Most of this article does not hinge heavily on these assumptions, and it should be relatively straightforward to extend our results to incorporate things like natural selection and more sophisticated models of mutation.

Acknowledgments. We gratefully acknowledge the support of the Isaac Newton Institute. Part of this work stemmed from discussions P.F. and Y.S.S. had during the 2010 program on “Statistical Challenges Arising from Genome Resequencing.”

References.

- BHASKAR, A., KAMM, J. A. and SONG, Y. S. (2012). Approximate sampling formulae for general finite-alleles models of mutation. *Advances in Applied Probability* **44** 408–428.
- BHASKAR, A. and SONG, Y. S. (2012). Closed-form asymptotic sampling distributions under the coalescent with recombination for an arbitrary number of loci. *Advances in Applied Probability* **44** 391–407.
- BIRKNER, M., BLATH, J. and ELDON, B. (2013). An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics* **193** 255–290.
- BOITARD, S. and LOISEL, P. (2007). Probability distribution of haplotype frequencies under the two-locus Wright-Fisher model by diffusion approximation. *Theoretical Population Biology* **71** 380–391.
- CHAN, A. H., JENKINS, P. A. and SONG, Y. S. (2012). Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics* **8** e1003090.
- ETHERIDGE, A. M. and GRIFFITHS, R. C. (2009). A coalescent dual process in a Moran model with genic selection. *Theoretical Population Biology* **75** 320–330.
- ETHIER, S. N. (1979). A limit theorem for two-locus diffusion models in population genetics. *Journal of Applied Probability* **16** 402–408.
- ETHIER, S. N. and GRIFFITHS, R. C. (1990). On the two-locus sampling distribution. *Journal of Mathematical Biology* **29** 131–159.
- ETHIER, S. N. and NAGYLAKI, T. (1980). Diffusion approximations of Markov chains with two time scales and applications to population genetics. *Advances in Applied Probability* **12** 14–49.
- ETHIER, S. N. and NAGYLAKI, T. (1988). Diffusion approximations of Markov chains with two time scales and applications to population genetics, II. *Advances in Applied Probability* **20** 525–545.
- ETHIER, S. N. and NAGYLAKI, T. (1989). Diffusion approximations of the two-locus Wright-Fisher model. *Journal of Mathematical Biology* **27** 17–28.
- EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3** 87–112.
- EWENS, W. J. (2004). *Mathematical Population Genetics*, 2nd ed. Springer-Verlag, New York.
- 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. (1991). The two-locus ancestral graph. In *Selected proceedings of the Sheffield symposium on applied probability: 18. IMS Lecture Notes—Monograph series* (I. V. BASAWA and R. L. TAYLOR, eds.) **18** 100–117.
- GRIFFITHS, R. C., JENKINS, P. A. and SONG, Y. S. (2008). Importance sampling and the two-locus model with subdivided population structure. *Advances in Applied Probability* **40** 473–500.
- GRIFFITHS, R. C. and MARJORAM, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3** 479–502.
- JENKINS, P. A. and GRIFFITHS, R. C. (2011). Inference from samples of DNA sequences using a two-locus model. *Journal of Computational Biology* **18** 109–127.
- JENKINS, P. A. and SONG, Y. S. (2009). Closed-form two-locus sampling distributions: accuracy and universality. *Genetics* **183** 1087–1103.
- JENKINS, P. A. and SONG, Y. S. (2010). An asymptotic sampling formula for the coalescent with recombination. *Annals of Applied Probability* **20** 1005–1028. .
- JENKINS, P. A. and SONG, Y. S. (2011). The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theoretical Population Biology* **80** 158–173.
- JENKINS, P. A. and SONG, Y. S. (2012). Padé approximants and exact two-locus sampling distributions. *Annals of Applied Probability* **22** 576–607.
- KAPLAN, N., DARDEN, T. and HUDSON, R. R. (1988). The coalescent process in models with selection. *Genetics* **120** 819–829.
- KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications* **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.
- MICHALOWICZ, J. V., NICHOLS, J. M., BUCHOLTZ, F. and OLSON, C. C. (2011). A general Isserlis theorem for mixed-Gaussian random variables. *Statistics and Probability Letters* **81** 1233–1240.
- MIURA, C. (2011). On an approximate formula for the distribution of 2-locus 2-allele model with mutual mutations. *Genes and Genetic Systems* **86** 207–214.
- MÖHLE, M. (1998). A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Advances in Applied Probability* **30** 493–512.
- NAGYLAKI, T. (1990). Models and approximations for random genetic drift. *Theoretical Population Biology* **37** 192–212.
- NIELSEN, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154** 931–942.
- NORMAN, M. F. (1972). *Markov processes and learning models. Mathematics in science and engineering* **84**. Academic Press, New York.
- NORMAN, M. F. (1975a). Approximation of stochastic processes by Gaussian diffusions, and applications to Wright-Fisher genetic models. *SIAM Journal on Applied Mathematics* **29** 225–242.
- NORMAN, M. F. (1975b). Limit theorems for stationary distributions. *Advances in Applied Probability* **7** 561–575.
- OHTA, T. and KIMURA, M. (1969a). Linkage disequilibrium due to random genetic drift. *Genetical Research* **13** 47–55.
- OHTA, T. and KIMURA, M. (1969b). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutations. *Genetics* **63** 229–238.
- WAKELEY, J. (2005). The limits of theoretical population genetics. *Genetics* **169** 1–7.
- WAKELEY, J. (2008). *Coalescent theory: an introduction*. Roberts & Company Publishers, Greenwood Village, Colorado.

- WAKELEY, J. and SARGSYAN, O. (2009). The conditional ancestral selection graph with strong balancing selection. *Theoretical Population Biology* **75** 355–364.
- WANG, Y. and RANNALA, B. (2008). Bayesian inference of fine-scale recombination rates using population genomic data. *Philosophical Transactions of the Royal Society B* **363** 3921–3930.
- WRIGHT, S. (1949). Adaptation and selection. In *Genetics, Paleontology and Evolution* (G. L. Jepson, E. Mayr and G. G. Simpson, eds.) 365–389. Princeton University Press, Princeton.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
COVENTRY CV4 7AL
UK

E-MAIL: p.jenkins@warwick.ac.uk

DEPARTMENT OF MATHEMATICS AND STATISTICS
LANCASTER UNIVERSITY
LANCASTER LA1 4YF
UK

E-MAIL: p.fearnhead@lancaster.ac.uk

DEPARTMENT OF STATISTICS AND
COMPUTER SCIENCE DIVISION
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CA 94720
USA
E-MAIL: yss@stat.berkeley.edu