Electron. J. Probab. **0** (2014), no. 0, 1–26.
ISSN: 1083-6489 DOI: 10.1214/EJP.vVOL-PID

Tractable diffusion and coalescent processes for weakly correlated loci

Paul A. Jenkins* Paul Fearnhead† Yun S. Song‡

Abstract

Widely used models in genetics include the Wright-Fisher diffusion and its moment dual, Kingman's coalescent. Each has a multilocus extension but under neither extension is the sampling distribution available in closed-form, and their computation is extremely difficult. In this paper we *derive* two new multilocus 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. The diffusion model is based on a central limit theorem for density dependent population processes, and we show that the sampling distribution is a linear combination of moments of Gaussian distributions and hence available in closed-form. The coalescent process is based on a probabilistic coupling of the ancestral recombination graph to a simpler genealogical process which exposes the leading dynamics of the former. 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 new models agree with the standard ones up to the first two orders.

Keywords: diffusion; sampling distribution; coupling; population genetics; recombination.
AMS 2010 Subject Classification: Primary 92D15, Secondary 65C50; 92D10.

Submitted to EJP on May 29, 2014, final version accepted on May 28, 2015.
Supersedes arXiv:1405.6863v2.

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 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

*University of Warwick, UK. E-mail: p.jenkins@warwick.ac.uk
†Lancaster University, UK. E-mail: p.fearnhead@lancaster.ac.uk
‡University of California, Berkeley, USA. E-mail: yss@stat.berkeley.edu

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 et al., 2012). However, with ongoing technological developments in high-throughput DNA sequencing, large genomic datasets are becoming available and it is necessary to consider multilocus 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 et al., 2008; Jenkins and Griffiths, 2011), Markov chain Monte Carlo (Kuhner et al., 2000; Nielsen, 2000; Wang and Rannala, 2008; Rasmussen et al., 2014), 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 $\boldsymbol{n}$ and its sampling probability by $q(\boldsymbol{n}; \rho)$ (to be defined precisely below). Consider the asymptotic expansion in inverse powers of ρ:

$$q(\boldsymbol{n}; \rho) = q_0(\boldsymbol{n}) + \frac{q_1(\boldsymbol{n})}{\rho} + \frac{q_2(\boldsymbol{n})}{\rho^2} + \cdots , \tag{1.1}$$

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 $q_0(\boldsymbol{n})$ and $q_1(\boldsymbol{n})$ in terms of the marginal *one*-locus sampling probabilities, and a decomposition of $q_2(\boldsymbol{n})$ 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 et al., 2012). One particularly appealing conclusion of these works is that both $q_0(\boldsymbol{n})$ and $q_1(\boldsymbol{n})$ 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 $q_0(\boldsymbol{n})$ and $q_1(\boldsymbol{n})$ 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 (1.1) 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 $N \to \infty$, for $0 < \beta < 1$, rather than the usual choice of $\beta = 1$. Interest

in asymptotically large recombination rates is reasonable because of extensive recombination rate heterogeneity along chromosomes in e.g. humans, strong recombination rates in some species such as *Drosophila melanogaster* (Chan et al., 2012), and because of the need to understand the long-range dependencies between well-separated loci. Our diffusion in this scaling is intimately related to the central limit theorem for density dependent population processes (see Ethier and Kurtz, 1986, Theorem 11.2.3), which has been analyzed in genetics—for models of strong mutation rather than strong recombination—by Feller (1951) and Norman (1975). A closely related scaling in the context of Ξ-coalescent processes was also recently explored by Birkner et al. (2013) (in that paper $\beta = 1$ but with timescale N^2). The coalescent approach, meanwhile, uses a coupling argument. Intuitively, we would like to couple the ARG to the limiting case of two independent coalescent trees ($\rho = \infty$). To account for contributions to the sampling distribution of $O(\rho^{-1})$, we must quantify the "leading order reasons" for such a coupling to fail. When ρ is large but finite, lineages in the ARG ancestral to both loci undergo recombination backwards in time very rapidly, until the first time U that no such lineage survives. In this paper we show that, roughly speaking, in order to recover the sampling distribution up to $O(\rho^{-1})$ we need consider only the following type of exceptional event: *a coalescence occurs more recently than time U in the ARG, and the coalescence is between two lineages each of which is ancestral to both of the two loci*. This observation enables us to *define* a simple coalescent process which allows for at most one of these events but is otherwise very similar to the easy limiting process corresponding to $\rho = \infty$.

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 $M \in \mathbb{N} = \{0, 1, 2, \ldots\}$, let $[M] := \{1, 2, \ldots, M\}$. The complement of a set J is written $J^{\complement}$. Denote the Kronecker delta by δ_{ij} which takes the value 1 if $i = j$ and 0 otherwise. Let $\boldsymbol{e}_i$ denote a unit vector whose jth entry is δ_{ij}, and let $\boldsymbol{e}_{ij}$ denote a matrix with (k, l)th entry equal to $\delta_{ik}\delta_{jl}$. For a vector $\boldsymbol{v} \in \mathbb{R}^d$ we denote by $|\boldsymbol{v}|$ the usual Euclidean norm. Denote the $k \times 1$ zero vector by $\mathbf{0}_k$ and the $k \times 1$ all-one vector by $\mathbf{1}_k$. We will replace a subscript with a "$\cdot$" 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 nth ascending factorial of z. Finally, for a matrix $\boldsymbol{R}$ of processes we let $[\boldsymbol{R}]_t = ([R_i, R_j]_t)_{i,j}$ denote the matrix of corresponding covariation processes.

Consider the usual diffusion limit of an exchangeable model of random mating with constant population size of N 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 $N \to \infty$ and u_A, $u_B \to 0$ while the population-scaled parameters $\theta_A = 2Nu_A$ and $\theta_B = 2Nu_B$ 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^A_{ik}, with $E_B = [L]$ and P^B_{jl}, $j, l \in [L]$ 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 $\rho_\beta = 2N^\beta r$ is fixed as $N \to \infty$, for some fixed $\beta \in (0, 1]$. Previous work has focused on the case $\beta = 1$ with time measured in units of N 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 $\boldsymbol{n} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ where $\boldsymbol{a} = (a_i)_{i\in[K]}$ and a_i is the number of haplotypes observed to exhibit allele i at locus A; $\boldsymbol{b} = (b_j)_{j\in[L]}$ where b_j is the number of haplotypes observed to exhibit allele j at locus B; and $\boldsymbol{c} = (c_{ij})_{i\in[K],j\in[L]}$ 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 $\boldsymbol{c}_A = (c_{i\cdot})_{i\in[K]}$ 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. Finally, we use $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ to denote the probability that when we sample n haplotypes from the population at stationarity we obtain a given ordered configuration consistent with $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$; by sampling exchangeability this is indeed a function only of the unordered configuration $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. For convenience we suppress the dependence of this quantity on the model parameters and on β. The main result motivating this work is an expansion for $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ for the case of $\beta = 1$, and later we will show that this expansion holds for all $\beta \in (0, 1]$.

Theorem 2.1 (See Jenkins and Song (2009))**.** *Consider the following asymptotic expansion for $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ under the diffusion limit with $\beta = 1$:*

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

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

$$q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = q^A(\boldsymbol{a} + \boldsymbol{c}_A) q^B(\boldsymbol{b} + \boldsymbol{c}_B), \tag{2.1}$$

and the first order term is given by

$$\begin{aligned} q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = {} & \binom{c}{2} q^A(\boldsymbol{a} + \boldsymbol{c}_A) q^B(\boldsymbol{b} + \boldsymbol{c}_B) \\ & - q^B(\boldsymbol{b} + \boldsymbol{c}_B) \sum_{i=1}^{K} \binom{c_{i\cdot}}{2} q^A(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{e}_i) \\ & - q^A(\boldsymbol{a} + \boldsymbol{c}_A) \sum_{j=1}^{L} \binom{c_{\cdot j}}{2} q^B(\boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{e}_j) \\ & + \sum_{i=1}^{K}\sum_{j=1}^{L} \binom{c_{ij}}{2} q^A(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{e}_i) q^B(\boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{e}_j), \end{aligned} \tag{2.2}$$

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

Remark 2.2. Under a neutral, finite-alleles model of mutation, if mutation is *parent independent*—that is, $P^A_{ki} = P^A_i$, $i, k \in [K]$, and $P^B_{lj} = P^B_j$, $j, l \in [L]$, then $q^A(\boldsymbol{a})$ and $q^B(\boldsymbol{b})$ are known in closed-form:

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

These expressions follow, for example, from the moments of the Wright-Fisher diffusion with parent-independent mutation, whose stationary distribution at locus A is Dirichlet$(\theta_A P^A_1, \ldots, \theta_A P^A_{K-1})$ (Wright, 1949), and similarly at locus B.

Remark 2.3. The zeroth-order decomposition is well known (e.g. Ethier, 1979) and also intuitive, since the two loci become independent as $\rho \to \infty$.

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 (2.1) and (2.2). For simplicity we will obtain our diffusion as the limit of an appropriately rescaled Moran 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 Moran model

A population of N haploid, monoecious individuals evolves as a multitype birth-and-death process in continuous time. Each individual carries a haplotype comprising a pair of alleles $(i, j) \in [K] \times [L]$, one at locus A and one at locus B. Let $Z_{ij}(\tau) \in \{0, 1, \ldots, N\}$ denote the number of (i, j) haplotypes in the population at time $\tau \in \mathbb{R}_{\geq 0}$, and $\boldsymbol{Z}(\tau) = (Z_{ij}(\tau))_{i\in[K],j\in[L]}$. The population evolves as follows. At rate $N^2/2$ a reproduction event occurs, in which an individual is chosen uniformly at random from the population to die. It is replaced by a copy of another individual also chosen uniformly at random (the same individual could be chosen; whether sampling is with or without replacement does not affect the diffusion limit). Independently, each locus of each haplotype undergoes mutation: any locus A mutates at rate $\theta_A/2$ and its allele is updated according to the transition matrix $\boldsymbol{P}^A = (P^A_{ik})_{i,k\in[K]}$; similarly any locus B mutates at rate $\theta_B/2$ and its allele is updated according to $\boldsymbol{P}^B = (P^B_{jl})_{j,l\in[L]}$. Finally, each haplotype independently undergoes recombination at rate $\rho/2$: at such an event, it is replaced by a haplotype formed by sampling two alleles (one for each locus) independently from the population. Putting all this together, the rate at which a haplotype (i, j) dies and is replaced by a haplotype (k, l) when $\boldsymbol{Z}(\tau) = \boldsymbol{z}$ is given by

$$\lambda^{(N)}_{ij,kl}(\boldsymbol{z}) = \frac{z_{ij}}{N}\left[\frac{N^2}{2}\frac{z_{kl}}{N} + N\left(\frac{\theta_A}{2}P^A_{ik}\delta_{jl} + \frac{\theta_B}{2}P^B_{jl}\delta_{ik} + \frac{\rho}{2}\frac{z_{k\cdot}}{N}\frac{z_{\cdot l}}{N}\right)\right], \qquad (i, j), (k, l) \in [K]\times[L].$$

Notice that, as is standard (e.g. Baake and Herms, 2008), we decouple the mutation and recombination mechanisms from reproduction (and from each other). This simplifies the analysis without unduly affecting the diffusion limit.

Remark 3.1. It is worth mentioning that the parameters of the Moran model—θ_A, θ_B, and ρ—are defined directly, without reference to u_A, u_B, and r as in Section 2. The two sets of definitions are reconciled provided we interpret one *generation* of the Moran model as N^{-1} units of time. Then for example the expected number of recombination events per individual per unit of time in the Moran model is $Nr = \rho/2$, as before. This interpretation of one "generation" is reasonable since the expected lifetime of an individual is $2/N$ units of time (a factor of two difference arising because the effective population size of the Moran model is $N/2$ (Ewens, 2004, p121)). The advantage of introducing u_A, u_B, and r in this way is that the new scaling described below, given by $\rho_\beta = 2N^\beta r$, provides both biological intuition and a means of comparison with the Wright-Fisher model. We can now continue to speak of the asymptotic behaviour of the parameter r; equivalently, we are going to rescale ρ by writing $\rho = 2Nr = 2N^\beta r \times N^{1-\beta} = \rho_\beta N^{1-\beta}$ and fixing ρ_β as $N \to \infty$.

We next change variables by introducing the collection

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

where

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

That is, we describe the state of the Moran model at time τ by the marginal allele frequencies and the coefficients of linkage disequilibrium (see, e.g. Ewens, 2004, p69, p227). We will write this succinctly by arranging the variables in a linear order:

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

and thinking of $\boldsymbol{M}^{(N)}(\tau)$ as a vector of length $\Lambda := K + L + KL$. The process $(\boldsymbol{M}^{(N)}(\tau) : \tau \geq 0)$ is then Markov on a state space we denote by $\Delta_{KL-1}^{(N)}$, which is a rational subset (those points consistent with $\sum_{i=1}^K \sum_{j=1}^L Z_{ij} = N$) of the $(KL-1)$-dimensional shifted simplex

$$\Delta_{KL-1} = \left\{(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{d}) \in [0,1]^K \times [0,1]^L \times [-1,1]^{KL} : \sum_{i=1}^K x_i = 1 = \sum_{j=1}^L y_j, \sum_{i=1}^K d_{ij} = 0 = \sum_{j=1}^L d_{ij}\right\}.$$

To find the diffusion limit we first need the conditional means and covariances of the increments

$$\Delta\boldsymbol{M}^{(N)}(\tau) := \boldsymbol{M}^{(N)}(\tau + \mathrm{d}\tau) - \boldsymbol{M}^{(N)}(\tau).$$

From these, and under the assumption that θ_A, θ_B, and ρ are fixed as $N \to \infty$, it is possible to show that the model converges to a (Wright-Fisher) diffusion limit (Ethier and Kurtz, 1986, Example 10.3.9, p433). Recall however that our interest is when ρ_β, rather than ρ, is fixed, so below we write these increments in terms of ρ_β using the fact that $\rho = \rho_\beta N^{1-\beta}$, as noted in Remark 3.1.

In the following, for convenience we drop the dependence on τ.

Proposition 3.2. *In the neutral two-locus Moran model with mutation and recombination, the conditional means and covariances of increments of $\boldsymbol{M}^{(N)}$ are given by*

$$\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{E}[\Delta X_i^{(N)} \mid \boldsymbol{M}^{(N)}] = \frac{\theta_A}{2}\sum_{k=1}^K (P_{ki}^A - \delta_{ik})X_k^{(N)}, \tag{3.1}$$

$$\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{E}[\Delta Y_j^{(N)} \mid \boldsymbol{M}^{(N)}] = \frac{\theta_B}{2}\sum_{l=1}^L (P_{lj}^B - \delta_{jl})Y_l^{(N)}, \tag{3.2}$$

$$\begin{aligned}
\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{E}[\Delta D_{ij}^{(N)} \mid \boldsymbol{M}^{(N)}] = &-\frac{\rho_\beta}{2N^{\beta-1}}D_{ij}^{(N)} - D_{ij}^{(N)} + \frac{\theta_A}{2}\sum_{k=1}^K (P_{ki}^A - \delta_{ik})D_{kj}^{(N)}\\
&+ \frac{\theta_B}{2}\sum_{l=1}^L (P_{lj}^B - \delta_{jl})D_{il}^{(N)} + O\left(\frac{1}{N^\beta}\right),
\end{aligned} \tag{3.3}$$

$$\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{C}ov[\Delta X_i^{(N)},\Delta X_k^{(N)}\mid \boldsymbol{M}^{(N)}] = X_i^{(N)}(\delta_{ik}-X_k^{(N)})+O\left(\frac{1}{N^\beta}\right),$$

$$\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{C}ov[\Delta Y_j^{(N)},\Delta Y_l^{(N)}\mid \boldsymbol{M}^{(N)}] = Y_j^{(N)}(\delta_{jl}-Y_l^{(N)})+O\left(\frac{1}{N^\beta}\right),$$

$$\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{C}ov[\Delta X_i^{(N)},\Delta Y_j^{(N)}\mid \boldsymbol{M}^{(N)}] = D_{ij}^{(N)}+O\left(\frac{1}{N^\beta}\right),$$

$$\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{C}ov[\Delta X_i^{(N)},\Delta D_{kl}^{(N)}\mid \boldsymbol{M}^{(N)}] = D_{kl}^{(N)}(\delta_{ik}-X_i^{(N)})-X_k^{(N)}D_{il}^{(N)}+O\left(\frac{1}{N^\beta}\right),$$

$$\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{C}ov[\Delta Y_j^{(N)},\Delta D_{kl}^{(N)}\mid \boldsymbol{M}^{(N)}] = D_{kl}^{(N)}(\delta_{jl}-Y_j^{(N)})-Y_l^{(N)}D_{kj}^{(N)}+O\left(\frac{1}{N^\beta}\right),$$

$$\begin{aligned}\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{C}ov[\Delta D_{ij}^{(N)},\Delta D_{kl}^{(N)}\mid \boldsymbol{M}^{(N)}] &= X_i^{(N)}Y_j^{(N)}(\delta_{ik}-X_k^{(N)})(\delta_{jl}-Y_l^{(N)})\\ &\quad + D_{kj}^{(N)}X_i^{(N)}Y_l^{(N)}+D_{il}^{(N)}X_k^{(N)}Y_j^{(N)}\\ &\quad + D_{ij}^{(N)}(X_k^{(N)}Y_l^{(N)}-\delta_{ik}Y_l^{(N)}-\delta_{jl}X_k^{(N)})\\ &\quad + D_{kl}^{(N)}(X_i^{(N)}Y_j^{(N)}-\delta_{ik}Y_j^{(N)}-\delta_{jl}X_i^{(N)})\\ &\quad + D_{ij}^{(N)}(\delta_{ik}\delta_{jl}-D_{kl}^{(N)})+O\left(\frac{1}{N^\beta}\right).\end{aligned}$$

Higher order moments of order $m\geq 2$ are $O(N^{-(m-2)})$.

Proof. These expressions follow directly from the first four moments of $\boldsymbol{Z}(\tau+\mathrm{d}\tau)\mid\boldsymbol{Z}(\tau)$, which are easily computed by noting that

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{Z}(\tau+\mathrm{d}\tau))\mid\boldsymbol{Z}(\tau)=\boldsymbol{z}] &= \sum_{(i,j)}\sum_{(k,l)} f(\boldsymbol{z}-\boldsymbol{e}_{ij}+\boldsymbol{e}_{kl})\lambda_{ij,kl}^{(N)}(\boldsymbol{z})\mathrm{d}\tau\\ &\quad + f(\boldsymbol{z})\left[1-\frac{N}{2}(N+\theta_A+\theta_B+\rho)\mathrm{d}\tau\right]+o(\mathrm{d}\tau).\end{aligned}$$

For example, choosing $f(\boldsymbol{z})=z_{uv}$ we find

$$\begin{aligned}&\mathbb{E}[Z_{uv}(\tau+\mathrm{d}\tau)\mid\boldsymbol{Z}(\tau)=\boldsymbol{z}] = z_{uv}\\ &+N\left[\frac{\theta_A}{2}\sum_{k=1}^{K}(P_{ku}^A-\delta_{ku})\frac{z_{kv}}{N}+\frac{\theta_B}{2}\sum_{l=1}^{L}(P_{lv}^A-\delta_{lv})\frac{z_{ul}}{N}+\frac{\rho}{2}\left(\frac{z_{u\cdot}}{N}\frac{z_{\cdot v}}{N}-\frac{z_{uv}}{N}\right)\right]\mathrm{d}\tau+o(\mathrm{d}\tau),\end{aligned}$$

and hence we recover (3.1) via

$$\begin{aligned}\mathbb{E}[\Delta X_u\mid\boldsymbol{M}^{(N)}] &= \frac{1}{N}\sum_{v=1}^{L}(\mathbb{E}[Z_{uv}(\tau+\mathrm{d}\tau)\mid\boldsymbol{Z}(\tau)]-Z_{uv})\\ &= \frac{\theta_A}{2}\sum_{k=1}^{K}(P_{ku}^A-\delta_{ku})X_k^{(N)}\mathrm{d}\tau+o(\mathrm{d}\tau).\end{aligned}$$

The remaining terms follow similarly; we omit the straightforward but lengthy algebraic details. □

To prepare for our diffusion limit, we must rescale time; from (3.3) it is clear that to obtain a nontrivial limit we should let $t=N^{1-\beta}\tau$. Henceforth, to avoid a trivial rescaling of time we assume $\beta\in(0,1)$. Now introduce the conditional mean vector $\boldsymbol{w}^{(N)}$ and conditional covariance matrix $\boldsymbol{s}^{(N)}$ on this timescale, defined by

$$\begin{aligned}&\lim_{\mathrm{d}t\to 0}(\mathrm{d}t)^{-1}\mathbb{E}[\Delta\boldsymbol{M}^{(N)}(\tau)\mid\boldsymbol{M}^{(N)}(\tau)=\boldsymbol{m}]\\ &\qquad = N^{\beta-1}\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{E}[\Delta\boldsymbol{M}^{(N)}(\tau)\mid\boldsymbol{M}^{(N)}(\tau)=\boldsymbol{m}]=:\boldsymbol{w}^{(N)}(\boldsymbol{m}),\quad(3.4)\end{aligned}$$

$$\lim_{\mathrm{d}t\to 0}(\mathrm{d}t)^{-1}\mathbb{C}\mathrm{ov}[\Delta \boldsymbol{M}^{(N)}(\tau)\mid \boldsymbol{M}(\tau)=\boldsymbol{m}]$$
$$= N^{\beta-1}\lim_{\mathrm{d}\tau\to 0}(\mathrm{d}\tau)^{-1}\mathbb{C}\mathrm{ov}[\Delta \boldsymbol{M}^{(N)}(\tau)\mid \boldsymbol{M}^{(N)}(\tau)=\boldsymbol{m}] =: N^{\beta-1}\boldsymbol{s}^{(N)}(\boldsymbol{m}), \quad (3.5)$$

with entries determined by Proposition 3.2. (Notice the appearance of τ on the left-hand side of (3.4) and (3.5); we are considering the evolution of $\boldsymbol{M}^{(N)}$ over a *fixed* time interval $\mathrm{d}\tau$ as measured in two different time units, t and τ.) Thus, with $\boldsymbol{m}=(x_1,\ldots,x_K, y_1,\ldots,y_L,d_{11},\ldots,d_{KL})$, equations (3.1)–(3.3) show that

$$\boldsymbol{w}^{(N)}(\boldsymbol{m}) = \boldsymbol{w}(\boldsymbol{m}) + O(N^{\beta-1}),$$
$$\text{where}\qquad \boldsymbol{w}(\boldsymbol{m}) = \Big(\underbrace{0,\ldots 0}_{K},\underbrace{0,\ldots 0}_{L},\underbrace{-\frac{\rho_\beta}{2}d_{11},\ldots,-\frac{\rho_\beta}{2}d_{KL}}_{K\times L}\Big)', \quad (3.6)$$

with $\boldsymbol{s}^{(N)}(\boldsymbol{m})=\boldsymbol{s}(\boldsymbol{m})+O(N^{-\beta})$ determined in a similar fashion:

$$\boldsymbol{s}(\boldsymbol{m}) = \begin{bmatrix} \boldsymbol{s}_{\mathbf{XX}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{XY}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{XD}}(\boldsymbol{m}) \\ \boldsymbol{s}_{\mathbf{XY}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{YY}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{YD}}(\boldsymbol{m}) \\ \boldsymbol{s}_{\mathbf{XD}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{YD}}(\boldsymbol{m}) & \boldsymbol{s}_{\mathbf{DD}}(\boldsymbol{m}) \end{bmatrix},$$

where

$$\begin{aligned}
[\boldsymbol{s}_{\mathbf{XX}}(\boldsymbol{m})]_{ik} &= x_i(\delta_{ik}-x_k),\\
[\boldsymbol{s}_{\mathbf{YY}}(\boldsymbol{m})]_{jl} &= y_j(\delta_{jl}-y_l),\\
[\boldsymbol{s}_{\mathbf{XY}}(\boldsymbol{m})]_{ij} &= d_{ij},\\
[\boldsymbol{s}_{\mathbf{XD}}(\boldsymbol{m})]_{i,kl} &= d_{kl}(\delta_{ik}-x_i)-x_k d_{il},\\
[\boldsymbol{s}_{\mathbf{YD}}(\boldsymbol{m})]_{j,kl} &= d_{kl}(\delta_{jl}-y_j)-y_l d_{kj},\\
[\boldsymbol{s}_{\mathbf{DD}}(\boldsymbol{m})]_{ij,kl} &= x_iy_j(\delta_{ik}-x_k)(\delta_{jl}-y_l)+d_{kj}x_iy_l+d_{il}x_ky_j+d_{ij}(x_ky_l-\delta_{ik}y_l-\delta_{jl}x_k)\\
&\quad + d_{kl}(x_iy_j-\delta_{ik}y_j-\delta_{jl}x_i)+d_{ij}(\delta_{ik}\delta_{jl}-d_{kl}).
\end{aligned}$$

Notice in particular the different leading orders of the two quantities in (3.4) and (3.5): the mean increments are of $O(1)$ on this timescale while the covariances are of $O(N^{\beta-1})$. It is this difference, which is a consequence of our assumption that the recombination parameter 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 we see convergence to a diffusion process: in the special case of a Wright-Fisher model and $K=L=2$, such a diffusion limit was obtained (after a rescaling of time) by Ohta and Kimura (1969a,b). Our interest is however in $\beta\in(0,1)$, 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 Gaussian diffusion limit of fluctuations in linkage disequilibrium

We first provide a heuristic description of the diffusion limit. First observe that, for $\beta\in(0,1)$, equations (3.4) and (3.5) reduce to an ordinary differential equation as $N\to\infty$:

$$\frac{\mathrm{d}\boldsymbol{M}(t)}{\mathrm{d}t} = \boldsymbol{w}(\boldsymbol{M}(t)). \quad (3.7)$$

Thus, if $\boldsymbol{M}^{(N)}(0)\to\boldsymbol{M}(0)$ as $N\to\infty$ then we should expect $\boldsymbol{M}^{(N)}(t)$ to converge to the solution of (3.7):

$$\boldsymbol{M}^{(N)} \xrightarrow{d} \boldsymbol{M} := \Big\{(\boldsymbol{X}(0),\boldsymbol{Y}(0),\boldsymbol{D}(0)e^{-\rho_\beta t/2})' : t\geq 0\Big\}, \qquad N\to\infty, \quad (3.8)$$

the deterministic exponential decay in LD typical of an infinitely large population. See Baake and Herms (2008) for a formal statement of this law-of-large-numbers type result for the Moran model with recombination. For the corresponding central limit theorem, we seek a diffusion limit for

$$\boldsymbol{U}^{(N)}(t) := r_N[\boldsymbol{M}^{(N)}(t) - \boldsymbol{M}(t)], \tag{3.9}$$

for some rescaling $r_N \to \infty$. In our application the appropriate choice is

$$r_N := N^{(1-\beta)/2},$$

which can be regarded as the one on which both recombination and genetic drift are observable on the fastest timescale (Jenkins and Song, 2012). We will assume this scaling henceforward. To find the limit $\boldsymbol{U} = \lim_{N\to\infty} \boldsymbol{U}^{(N)}$, write

$$\boldsymbol{U}^{(N)}(t) = r_N \left[[\boldsymbol{M}^{(N)}(0) - \boldsymbol{M}(0)] + \int_0^t [\boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s)) - \boldsymbol{w}(\boldsymbol{M}(s))]\mathrm{d}s + \boldsymbol{R}^{(N)}(t) \right], \tag{3.10}$$

where

$$\boldsymbol{R}^{(N)}(t) := \boldsymbol{M}^{(N)}(t) - \boldsymbol{M}^{(N)}(0) - \int_0^t \boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s))\mathrm{d}s$$

describes the deviations of $\boldsymbol{M}^{(N)}(t)$ from its expected behaviour and is a martingale. It suffices to characterize the limits of each of the three grouped terms on the right of (3.10). For the first term we assume that it converges to a limit, $\boldsymbol{U}^{(N)}(0) \xrightarrow{d} \boldsymbol{U}(0)$ as $N \to \infty$. For the second term, from (3.6) we should expect

$$\begin{aligned}
&r_N \int_0^t [\boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s)) - \boldsymbol{w}(\boldsymbol{M}(s))]\mathrm{d}s \\
&\quad = r_N \int_0^t \left[\left(\mathbf{0}_K, \mathbf{0}_L, -\frac{\rho_\beta}{2}[\boldsymbol{D}^{(N)}(s) - \boldsymbol{D}(s)] \right)' + O(N^{\beta-1}) \right] \mathrm{d}s \\
&\quad = \int_0^t \left[-\frac{\rho_\beta}{2} (\mathbf{0}_K, \mathbf{0}_L, \mathbf{1}_{KL})' \circ \boldsymbol{U}^{(N)}(s) + O(N^{(\beta-1)/2}) \right] \mathrm{d}s \\
&\quad \xrightarrow{d} -\frac{\rho_\beta}{2} \int_0^t (\mathbf{0}_K, \mathbf{0}_L, \mathbf{1}_{KL})' \circ \boldsymbol{U}(s)\mathrm{d}s, \qquad N \to \infty,
\end{aligned} \tag{3.11}$$

where $\circ$ denotes the Hadamard (elementwise) product of two vectors. Finally, we obtain a complete description of the limit $r_N\boldsymbol{R}^{(N)} \xrightarrow{d} \boldsymbol{R}$ as $N \to \infty$ by an application of the martingale central limit theorem (Ethier and Kurtz, 1986, Theorem 7.1.4); we find

$$\boldsymbol{R}(t) = \int_0^t \boldsymbol{\sigma}(\boldsymbol{M}(s))\mathrm{d}\boldsymbol{W}(s),$$

where $\boldsymbol{\sigma}\boldsymbol{\sigma}' = \boldsymbol{s}$, and $\boldsymbol{W}$ is a $(KL-1)$-dimensional Brownian motion. In summary then, we expect $\boldsymbol{U}$ to satisfy

$$\boldsymbol{U}(t) = \boldsymbol{U}(0) - \frac{\rho_\beta}{2} \int_0^t (\mathbf{0}_K, \mathbf{0}_L, \mathbf{1}_{KL})' \circ \boldsymbol{U}(s)\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(\boldsymbol{M}(s))\mathrm{d}\boldsymbol{W}(s). \tag{3.12}$$

Our main result formalizes this argument, as follows.

Theorem 3.3. *Suppose that $\boldsymbol{U}^{(N)}(0) \xrightarrow{d} \boldsymbol{U}(0)$ as $N \to \infty$. Then for each $t > 0$, as $N \to \infty$,*

$$\sup_{s\le t} |\boldsymbol{M}^{(N)}(s) - \boldsymbol{M}(s)| \xrightarrow{d} 0;$$

$N^{(1-\beta)/2}\boldsymbol{R}^{(N)} \xrightarrow{d} \boldsymbol{R}$, *where* $\boldsymbol{R}$ *has Gaussian, independent increments with mean zero, and with*

$$\mathbb{E}[\boldsymbol{R}(t)\boldsymbol{R}(t)'] = \int_0^t \boldsymbol{s}(\boldsymbol{M}(s))\mathrm{d}s; \tag{3.13}$$

and $\boldsymbol{U}^{(N)} \xrightarrow{d} \boldsymbol{U}$, *satisfying* (3.12).

Proof of Theorem 3.3. This is an application of a central limit theorem for density dependent population processes; for textbook coverage see Ethier and Kurtz (1986, Chapter 11) and for a recent treatment see Kang et al. (2014). We apply Theorem 2.11 of Kang et al. (2014). To do so we need to validate each of the assertions that led to (3.12) above by checking the following sufficient conditions (i)–(iv). (Kang et al. (2014, Theorem 2.11) is rather more general than is required here: it permits the state space of $\boldsymbol{M}^{(N)}$ to be unbounded, and for $\boldsymbol{M}^{(N)}$ to depend on other processes that evolve on faster timescales than that of the diffusion. We omit those conditions which are not needed.)

(i) **The Moran process converges to an identifiable, deterministic limit.** *This is guaranteed by the following: the infinitesimal generator* $\mathcal{A}_N$ *of* $\boldsymbol{M}^{(N)}$ *satisfies*

$$\lim_{N\to\infty} \sup_{\boldsymbol{m}\in\Delta^{(N)}_{KL-1}} |\mathcal{A}_N f(\boldsymbol{m}) - \mathcal{A}f(\boldsymbol{m})| = 0, \qquad f \in \mathcal{D}(\mathcal{A}),$$

for a generator $\mathcal{A}$ *with domain* $\mathcal{D}(\mathcal{A})$.

(ii) **Fluctuations about the deterministic limit are well behaved.** *More precisely,* $\boldsymbol{R}^{(N)}$ *is a local martingale and the covariations processes* $[\boldsymbol{M}^{(N)}]_t \xrightarrow{d} 0$.

(iii) **Contributions of** $O(r_N^{-1})$ **to the error** $\boldsymbol{w}^{(N)} - \boldsymbol{w}$ **can be identified.** *These would contribute to the limiting drift of* $\boldsymbol{U}(t)$, *and a sufficient condition to identify them is: there exists a continuous function* $\boldsymbol{G}_0 : \Delta_{KL-1} \to \mathbb{R}^\Lambda$ *(recall* $\Lambda = K + L + KL$*) such that*

$$\lim_{N\to\infty} \sup_{\boldsymbol{m}\in\Delta^{(N)}_{KL-1}} \left| r_N \left[\boldsymbol{w}^{(N)}(\boldsymbol{m}) - \boldsymbol{w}(\boldsymbol{m})\right] - \boldsymbol{G}_0(\boldsymbol{m})\right| = 0.$$

(iv) **The martingale central limit theorem applies to** $r_N\boldsymbol{R}^{(N)}$**.** *This is guaranteed by the following:*

$$\lim_{N\to\infty} \mathbb{E}\left[\sup_{s\le t} r_N \left|\boldsymbol{M}^{(N)}(s) - \boldsymbol{M}^{(N)}(s-)\right|\right] = 0, \tag{3.14}$$

and there exists a continuous $\boldsymbol{G} : \Delta_{KL-1} \to \mathbb{R}^{\Lambda\times\Lambda}$ *such that for each* $t > 0$,

$$r_N^2[\boldsymbol{M}^{(N)}]_t - \int_0^t \boldsymbol{G}(\boldsymbol{M}^{(N)}(s))\mathrm{d}s \xrightarrow{d} 0. \tag{3.15}$$

We address each of these requirements in turn.

(i) It follows immediately from Proposition 3.2 that $\mathcal{A}_N f(\boldsymbol{m})$ converges to $\mathcal{A}f(\boldsymbol{m}) := \boldsymbol{w}.\nabla f(\boldsymbol{m})$, the generator of $\boldsymbol{M}$ [see (3.8)] with domain $\mathcal{D}(\mathcal{A}) = C^2(\Delta_{KL-1})$, the set of twice continuously differentiable functions. Convergence is uniform in $\boldsymbol{m}$ because the $O(N^{-\beta})$ terms in Proposition 3.2 have coefficients that are polynomials in $\boldsymbol{M}^{(N)}$ on a compact space.

(ii) Since the state space is bounded, for $\boldsymbol{R}^{(N)}$ to be a martingale it suffices that the jump rate is bounded across all values of $\boldsymbol{M}^{(N)}(t)$ in $\Delta^{(N)}_{KL-1}$ (Kurtz, 1971, Proposition 2.1), as is the case for the Moran process. The covariations process $[\boldsymbol{M}^{(N)}]_t \xrightarrow{d} 0$ as a consequence of (3.15), verified below.

(iii) From (3.6), $r_N[\boldsymbol{w}^{(N)}(\boldsymbol{m}) - \boldsymbol{w}(\boldsymbol{m})] = O(N^{(\beta-1)/2})$, again uniformly in $\boldsymbol{m} \in \Delta^{(N)}_{KL-1}$, so here the appropriate choice is $\boldsymbol{G}_0 \equiv \boldsymbol{0}$. Thus, the only relevant contribution to the

limit (3.11) is from the error $\boldsymbol{w}(\boldsymbol{M}^{(N)}(s)) - \boldsymbol{w}(\boldsymbol{M}(s))$ rather than from $\boldsymbol{w}^{(N)}(\boldsymbol{M}^{(N)}(s)) - \boldsymbol{w}(\boldsymbol{M}^{(N)}(s))$.

(iv) Jumps of any component of $\boldsymbol{M}^{(N)}$ are bounded in magnitude by $2/N$, so

$$\sup_{s\le t} r_N \left|\boldsymbol{M}^{(N)}(s) - \boldsymbol{M}^{(N)}(s-)\right| \le N^{(1-\beta)/2} \cdot \frac{2\Lambda^{1/2}}{N} \to 0, \qquad N \to \infty,$$

and (3.14) holds. To identify the asymptotic behaviour of $r_N^2[\boldsymbol{M}^{(N)}]_t$, let

$$\mathcal{N}_{\boldsymbol{m}}^{(N)}(t) = \mathcal{Y}_{\boldsymbol{m}}\left(\int_0^t \lambda_{\boldsymbol{m}}^{(N)}(\boldsymbol{M}^{(N)}(s))\mathrm{d}s\right)$$

denote the total number of jumps of the Moran process into state $\boldsymbol{m} \in \Delta_{KL-1}^{(N)}$ by time t, where $(\mathcal{Y}_{\boldsymbol{m}} : \boldsymbol{m} \in \Delta_{KL-1}^{(N)})$ is a collection of independent Poisson processes of unit rate and $\lambda_{\boldsymbol{m}}^{(N)}(\boldsymbol{M}^{(N)}(s))$ denotes the rate of transition of the process from current state $\boldsymbol{M}^{(N)}(s)$ to $\boldsymbol{m}$. Then

$$\begin{aligned} r_N^2[\boldsymbol{M}^{(N)}]_t &= N^{1-\beta}\int_0^t \sum_{\boldsymbol{m}\in\Delta_{KL-1}^{(N)}} [\Delta\boldsymbol{M}^{(N)}(s)][\Delta\boldsymbol{M}^{(N)}(s)]'\mathrm{d}\mathcal{N}_{\boldsymbol{m}}^{(N)}(s), \\ &\sim N^{1-\beta}\int_0^t \sum_{\boldsymbol{m}\in\Delta_{KL-1}^{(N)}} [\boldsymbol{m} - \boldsymbol{M}^{(N)}(s)][\boldsymbol{m} - \boldsymbol{M}^{(N)}(s)]'\lambda_{\boldsymbol{m}}^{(N)}(\boldsymbol{M}^{(N)}(s))\mathrm{d}s, \\ &\sim \int_0^t \boldsymbol{s}^{(N)}(\boldsymbol{M}^{(N)}(s))\mathrm{d}s, \end{aligned}$$

by (3.5). Thus we may take $\boldsymbol{G} = \boldsymbol{s}$ in (3.15) [$\boldsymbol{G}$ identifies the moments appearing in (3.13)]. □

Remark 3.4. One could obtain the same diffusion limit starting from a Wright-Fisher model rather than a Moran model, since the means and covariances of its increments are identical to leading order, up to a rescaling of time. (Specifically, $t = \lfloor 2N^\beta\tau \rfloor$ when τ counts generations of the Wright-Fisher model, differing from the Moran model by a usual factor of $2/N$.) This alternative approach is in some respects less appealing since the Wright-Fisher model, when expressed in continuous time, is non-Markovian. The additional complications raised by this approach have been addressed by Norman (1975) (see also Ethier and Nagylaki, 1980, 1988), and we have checked that the conditions of his theorems still apply when we introduce recombination to the Wright-Fisher model. The theory of Norman (1975) has been used to study strong mutation and selection (Norman, 1972, 1975; Kaplan et al., 1988; Nagylaki, 1986, 1990; Wakeley and Sargsyan, 2009), and a Gaussian diffusion approximation of a Moran model with strong selection is developed by Feder et al. (2014), but to the best of our knowledge this is the first time a central limit theorem has been obtained for strong recombination.

Remark 3.5. The exponential decay of linkage disequilibrium implied by $\boldsymbol{M}$ [equation (3.8)] 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 $\boldsymbol{U}$ [equation (3.9)] shows that fluctuations are of order $N^{(1-\beta)/2}$ on a timescale of $N^{\beta-1}$ units of the Moran process. If we designate N^{-1} units as one *generation*, as we discussed in Remark 3.1, then these fluctuations can be said to occur on a timescale of order N^β generations.

3.3 Stationary distribution

Although $\boldsymbol{U}$ is described completely by (3.12), the volatility term $\boldsymbol{\sigma}(\boldsymbol{M}(t))$ is neither simple nor time-independent. On the other hand, our main interest is in stationary behaviour, and $\boldsymbol{\sigma}(\boldsymbol{M}(\infty))$ takes on a much simpler form. First note that the components of $\boldsymbol{U}(t)$ corresponding to each X_i and Y_j undergo Brownian motions (with nonunit volatility), so we restrict our attention to the stationary distribution of the component corresponding to $\boldsymbol{D}$, which we denote $\boldsymbol{U_D}$. Setting $\boldsymbol{\sigma}(\boldsymbol{M}(s)) = \boldsymbol{\sigma}(\boldsymbol{M}(\infty))$ in (3.12), we find

$$\mathrm{d}\boldsymbol{U_D}(t) = -\frac{\rho_\beta}{2}\boldsymbol{U_D}(t)\mathrm{d}t + \boldsymbol{\sigma}_\infty \mathrm{d}\boldsymbol{W}(t), \tag{3.16}$$

where $\boldsymbol{\sigma}_\infty$ is a *constant* defined by

$$\boldsymbol{\sigma}_\infty \boldsymbol{\sigma}_\infty' = \boldsymbol{s}_\infty := \boldsymbol{s_{DD}}(\boldsymbol{M}(\infty)) = [X_i(0)Y_j(0)(\delta_{ik} - X_k(0))(\delta_{jl} - Y_l(0))]_{ij,kl}.$$

The process (3.16) is much simpler to describe. Marginally, $\boldsymbol{U}_{D_{ij}}$ is an Ornstein-Uhlenbeck process with damping towards linkage equilibrium at rate $\rho_\beta/2$ and constant volatility $[\boldsymbol{\sigma}_\infty]_{ij,ij}$. $\boldsymbol{U_D}$ has stationary distribution Normal$(\boldsymbol{0}_{KL}, \boldsymbol{s}_\infty/\rho_\beta)$. This is a slightly different idea of stationarity than usual, since it depends on $\boldsymbol{X}(0)$ and $\boldsymbol{Y}(0)$, so an immediate question is: what should be the distributions for $\boldsymbol{X}(0)$ and $\boldsymbol{Y}(0)$? We address this by reconsidering the usual two-locus *Wright-Fisher* diffusion limit operating on a slower timescale. We can exploit (3.16) to obtain a simple approximation of this diffusion limit, as follows. First, we have *derived* the Gaussian diffusion approximation

$$\boldsymbol{D}(0)e^{-\rho_\beta t/2} + N^{(\beta-1)/2}\boldsymbol{U_D}(t)$$

for $\boldsymbol{D}^{(N)}(t)$. Thus the stationary distribution of this approximation is

$$\text{Normal}\left(\boldsymbol{0}_{KL}, \frac{\boldsymbol{s}_\infty}{\rho}\right). \tag{3.17}$$

(One can think of this as an approximation for the distribution of

$$\lim_{N\to\infty}\lim_{t\to\infty}\boldsymbol{D}^{(N)}(t)$$

having been obtained by exchanging the two limits.) Notice that the description (3.17) does not depend on the particular choice of β. Under the usual "Wright-Fisher" regime we treat ρ as fixed. It remains to specify the stationary distributions for the marginal allele frequencies $\boldsymbol{X}$ and $\boldsymbol{Y}$, which we suppose to have reached their usual (independent) stationary distributions in the Wright-Fisher diffusion limit, 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 (3.17) by specifying $(\boldsymbol{X}(0), \boldsymbol{Y}(0)) \sim \pi_A \otimes \pi_B$.

The distribution (3.17) therefore provides a simple, explicit method for the approximate simulation of haplotype frequencies under a stationary, two-locus Wright-Fisher diffusion, which we summarize in Algorithm 1 below. (When mutation is parent independent, as in Remark 2.2, 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 $\boldsymbol{U_D}$ is further evident from the following theorem. First we need some further notation. Let

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

Algorithm 1 Simulate from a Gaussian approximation to the stationary Wright-Fisher diffusion with recombination.

1. Simulate marginal allele frequencies at locus A, $\boldsymbol{X}(0) \sim \pi_A$.
2. Independently simulate marginal allele frequencies at locus B, $\boldsymbol{Y}(0) \sim \pi_B$.
3. Conditionally simulate $\boldsymbol{D}$ from (3.17) given $\boldsymbol{X}(0)$ and $\boldsymbol{Y}(0)$.
4. Calculate two-locus haplotype frequencies via

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

for $m \in \mathbb{N}$, and let $\boldsymbol{l}^{(\boldsymbol{r})} \in ([K] \times [L])^m$ denote a sequence of m haplotypes (in some arbitrary, fixed order) with multiplicities specified by $\boldsymbol{r} \in \mathcal{P}_m$. Further let $\boldsymbol{l}^{(\boldsymbol{r})_A} \in [K]^m$ denote the corresponding list of alleles obtained by looking at the first entry of each element of $\boldsymbol{l}^{(\boldsymbol{r})}$, and define $\boldsymbol{l}^{(\boldsymbol{r})_B}$ 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 $\boldsymbol{\xi}_{\boldsymbol{\mu\nu}} = \{\{\mu_k, \nu_k\} : k = 1, \ldots, \lambda\} \in \mathcal{Q}_{2\lambda}$; $\boldsymbol{\mu} = (\mu_k)$ and $\boldsymbol{\nu} = (\nu_k)$ are sequences of length λ. For $J \subseteq [\lambda]$, denote by $\boldsymbol{\mu}_J$, $\boldsymbol{\nu}_J$ the subsequences obtained by looking only at the indices in J, and denote by $\boldsymbol{l}^{(\boldsymbol{r})}_{\boldsymbol{\mu}}$ the subsequence of $\boldsymbol{l}^{(\boldsymbol{r})}$ obtained by looking only at the indices in $\boldsymbol{\mu}$. The matrix of multiplicities of $\boldsymbol{l}^{(\boldsymbol{r})}_{\boldsymbol{\mu}}$ is denoted by $\boldsymbol{r}^{(\boldsymbol{\mu})}$, so that $\boldsymbol{r}^{(\boldsymbol{\mu})} + \boldsymbol{r}^{(\boldsymbol{\nu})} = \boldsymbol{r}$. For example, if $\boldsymbol{r} = \left[\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right]$ then a representative list of haplotypes is $\boldsymbol{l}^{(\boldsymbol{r})} = ((1,1),(1,2),(1,2),(2,2))$ with marginal allele lists $\boldsymbol{l}^{(\boldsymbol{r})_A} = (1,1,1,2)$ and $\boldsymbol{l}^{(\boldsymbol{r})_B} = (1,2,2,2)$. Here, $m = 2\lambda = 4$, and $\mathcal{Q}_4 = \big\{\{\{1,2\},\{3,4\}\}, \{\{1,3\},\{2,4\}\}, \{\{1,4\},\{2,3\}\}\big\}$. Then for example the first element in $\mathcal{Q}_4$ is the partition $\boldsymbol{\xi}_{\boldsymbol{\mu\nu}}$ constructed from $\boldsymbol{\mu} = (1,3)$ and $\boldsymbol{\nu} = (2,4)$, and so $\boldsymbol{l}^{(\boldsymbol{r})}_{\boldsymbol{\mu}} = ((1,1),(1,2))$ and $\boldsymbol{l}^{(\boldsymbol{r})}_{\boldsymbol{\nu}} = ((1,2),(2,2))$.

Theorem 3.6. *Suppose that $\boldsymbol{X} \sim \pi_A$, $\boldsymbol{Y} \sim \pi_B$ independently, and conditional on $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\boldsymbol{D}$ is distributed according to the Gaussian distribution in* (3.17)*. Then the sampling distribution is given* exactly *by*

$$\begin{aligned} q_G(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = &\sum_{\lambda=0}^{\lfloor c/2 \rfloor} \frac{1}{\rho^\lambda} \sum_{\boldsymbol{r} \in \mathcal{P}_{2\lambda}} \sum_{\boldsymbol{\xi} \in \mathcal{Q}_{2\lambda}} \left[\prod_{i=1}^{K} \prod_{j=1}^{L} \binom{c_{ij}}{r_{ij}} \right] \\ &\times \left[\sum_{I \subseteq [\lambda] : \boldsymbol{l}^{(\boldsymbol{r})_A}_{\boldsymbol{\mu}_I} = \boldsymbol{l}^{(\boldsymbol{r})_A}_{\boldsymbol{\nu}_I}} (-1)^{|I^{\complement}|} q^A(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{r}_A^{(\boldsymbol{\nu}_I)}) \right] \\ &\times \left[\sum_{J \subseteq [\lambda] : \boldsymbol{l}^{(\boldsymbol{r})_B}_{\boldsymbol{\mu}_J} = \boldsymbol{l}^{(\boldsymbol{r})_B}_{\boldsymbol{\nu}_J}} (-1)^{|J^{\complement}|} q^B(\boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{r}_B^{(\boldsymbol{\nu}_J)}) \right], \qquad (3.18) \\ = &\, q_0(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) + \frac{q_1(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})}{\rho} + O\left(\frac{1}{\rho^2}\right), \end{aligned}$$

with q_0 and q_1 given by (2.1) *and* (2.2) *respectively (and we impose the convention that the empty summations for $\lambda = 0$ have a single term, with $(-1)^{|\varnothing \setminus \varnothing|} = 1$).*

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

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

$$
\begin{aligned}
&= \sum_{m=0}^{c} \sum_{\boldsymbol{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. \\
&\qquad\qquad \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 \boldsymbol{X},\boldsymbol{Y}\right]\right], \\
&= \sum_{\lambda=0}^{\lfloor c/2\rfloor} \sum_{\boldsymbol{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. \\
&\qquad\qquad \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_{\boldsymbol{l}_{\mu_k}^{(\boldsymbol{r})}} D_{\boldsymbol{l}_{\nu_k}^{(\boldsymbol{r})}} \mid \boldsymbol{X},\boldsymbol{Y}]\right], \\
&= \sum_{\lambda=0}^{\lfloor c/2\rfloor} \frac{1}{\rho^\lambda} \sum_{\boldsymbol{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)\left(\prod_{j=1}^{L} Y_j^{b_j+c_{\cdot j}-r_{\cdot j}}\right)\right. \\
&\qquad\qquad \left.\times \prod_{k=1}^{\lambda} X_{\boldsymbol{l}_{\mu_k}^{(\boldsymbol{r})_A}} Y_{\boldsymbol{l}_{\mu_k}^{(\boldsymbol{r})_B}} (\delta_{\boldsymbol{l}_{\mu_k}^{(\boldsymbol{r})_A}\boldsymbol{l}_{\nu_k}^{(\boldsymbol{r})_A}} - X_{\boldsymbol{l}_{\nu_k}^{(\boldsymbol{r})_A}})(\delta_{\boldsymbol{l}_{\mu_k}^{(\boldsymbol{r})_B}\boldsymbol{l}_{\nu_k}^{(\boldsymbol{r})_B}} - Y_{\boldsymbol{l}_{\nu_k}^{(\boldsymbol{r})_B}})\right], \\
&= \sum_{\lambda=0}^{\lfloor c/2\rfloor} \frac{1}{\rho^\lambda} \sum_{\boldsymbol{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^{\complement}|} \delta_{\boldsymbol{l}_{\boldsymbol{\mu}_I}^{(\boldsymbol{r})_A}\boldsymbol{l}_{\boldsymbol{\nu}_I}^{(\boldsymbol{r})_A}} \sum_{J\subseteq[\lambda]} (-1)^{|J^{\complement}|} \delta_{\boldsymbol{l}_{\boldsymbol{\mu}_J}^{(\boldsymbol{r})_B}\boldsymbol{l}_{\boldsymbol{\nu}_J}^{(\boldsymbol{r})_B}} \\
&\quad \times \mathbb{E}\left[\left(\prod_{i=1}^{K} X_i^{a_i+c_{i\cdot}-r_{i\cdot}^{(\boldsymbol{\nu}_I)}}\right)\left(\prod_{j=1}^{L} Y_j^{b_j+c_{\cdot j}-r_{\cdot j}^{(\boldsymbol{\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 (Michalowicz et al., 2011), and the fourth equality follows from (3.17):

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

The fifth equality follows from expanding the final product (using the convention $\delta_{\varnothing\varnothing} = 1$), while (3.18) follows from $(\boldsymbol{X},\boldsymbol{Y}) \sim \pi_A \otimes \pi_B$. The equalities still hold for $\lambda = 0$ provided we take $\prod_{\varnothing} = 1$.

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

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

as required. □

3.5 Accuracy of the diffusion process

A natural question to ask is: to what extent does the process of Theorem 3.6 capture the dynamics of the full process? To address this we consider the accuracy of the sampling distribution (3.18) as an approximation to the "true" distribution, $q(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$. 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 (3.18) grows rapidly with λ (as long as $\lambda \leq \lfloor \frac{c}{2} \rfloor$), so we define an approximate sampling distribution $q_G^{(\lambda)}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$ by truncating the outer sum in (3.18) at a fixed index λ. This is analogous to the asymptotic sampling formulae for the full model which are obtained by truncating equation (1.1) (Jenkins and Song, 2012). As our measure of accuracy we define the relative error,

$$\mathring{e}^{(\lambda)}_{\text{Gaussian}} = \left| \frac{Q_G^{(\lambda)}(\mathbf{0},\mathbf{0},\boldsymbol{c}) - q(\mathbf{0},\mathbf{0},\boldsymbol{c})}{q(\mathbf{0},\mathbf{0},\boldsymbol{c})} \right| \times 100\%, \tag{3.19}$$

where $Q_G^{(\lambda)}(\mathbf{0},\mathbf{0},\boldsymbol{c})$ is the staircase Padé approximant to $q_G^{(\lambda)}(\mathbf{0},\mathbf{0},\boldsymbol{c})$. (The former is used for its superior convergence properties; see Jenkins and Song, 2012, for details.) We define $\mathring{e}^{(\lambda)}_{\text{True}}$ analogously, replacing $Q_G^{(\lambda)}(\mathbf{0},\mathbf{0},\boldsymbol{c})$ in (3.19) with the Padé approximant to the partial sum of (1.1), computed up to $O(\rho^{-(\lambda+1)})$ by the method of Jenkins and Song (2012).

We computed the distribution of $\mathring{e}^{(\lambda)}_{\text{Gaussian}}$ and of $\mathring{e}^{(\lambda)}_{\text{True}}$ 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 (1.1) contributes to both errors; $\mathring{e}^{(\lambda)}_{\text{Gaussian}}$ has additional contributions reflecting its use of an approximate *model*. Of 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, $Q_G^{(\lambda)}(\mathbf{0},\mathbf{0},\boldsymbol{c})$ converges rapidly (even without Padé summation; not shown), and becomes a reasonable approximation to $q(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$. For example, for $\rho = 50$, $Q_G^{(6)}(\mathbf{0},\mathbf{0},\boldsymbol{c})$ is within 10% of $q(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$ 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, $\mathring{e}^{(\lambda)}_{\text{Gaussian}}$ 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 (2.1) and (2.2). We first provide an informal description. Let $\mathcal{C}^{(\rho)}_{a,b,c}(t)$ 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. Recombination occurs at the usual rate of $\rho c/2$, where $\rho = 2Nr$. 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 $(\mathcal{C}^{(\rho)} = (\mathcal{C}^{(\rho)}_{a,b,c}(t) : t \geq 0), \mathcal{D}^{(\rho)} = (\mathcal{D}^{(\rho)}_{a,b,c}(t)) : t \geq 0))$, where $\mathcal{D}^{(\rho)}$ is a simple process closely related to $\mathcal{C}^{(\infty)}$ and defined below. $\mathcal{C}^{(\rho)}(\omega)$ is said to be coupled to $\mathcal{D}^{(\rho)}(\omega)$ 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

Table 1: Cumulative distribution $\Phi(x) = \mathbb{P}(\mathring{e}^{(\lambda)} < x\%)$ (where $\mathring{e}^{(\lambda)}$ denotes either $\mathring{e}^{(\lambda)}_{\text{Gaussian}}$ or $\mathring{e}^{(\lambda)}_{\text{True}}$ 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

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 $\mathcal{D}^{(\rho)}$ 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 $\mathcal{D}^{(\rho)}$ 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.

Recall the two-locus ancestral process for the coalescent with recombination: Going backwards in time, each pair of lineages coalesces independently at rate 1, and each lineage ancestral at both loci recombines at rate $\rho/2$. When two lineages coalesce, they are replaced with a single lineage, and this lineage is ancestral at a given locus if either of its two progenitors were ancestral at this locus. Thus for example, with a, b, and c defined as above the total rate of coalescence involving one left-half fragment and one right-half fragment is ab, resulting in a transition of the form $(a,b,c) \mapsto (a-1,b-1,c+1)$. The remaining transitions are given in Table 2. We can now make the following concise definition.

Definition 4.1. *The ancestral process $\mathcal{C}^{(\rho)} = (\mathcal{C}^{(\rho)}_{a,b,c}(t) : t \geq 0)$ is a continuous-time Markov process on $\mathbb{N}^4$ such that $\mathcal{C}^{(\rho)}_{a,b,c}(0) = (a,b,c,c)$ a.s., and with infinitesimal generator*

$$\begin{aligned}\mathscr{L}f(a,b,c,c) = \frac{\rho c}{2} f(a+1,b+1,c-1,c-1) + \binom{c}{2} f(a,b,c-1,c-1) \\ + R_{a,b,c,c}\mathscr{G}f(a,b,c,c) - \left[\frac{\rho c}{2} + \binom{c}{2} + R_{a,b,c,c}\right] f(a,b,c,c), \qquad (4.1)\end{aligned}$$

where

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

and $f : \mathbb{N}^4 \to \mathbb{R}$ 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 representation is seemingly redundant, but it will make the coupling with the corresponding process $\mathcal{D}^{(\rho)}$ (for which we allow $c \neq d$) transparent. We will define $\mathcal{D}^{(\rho)}$ via the following recipe. First, take $\mathcal{C}^{(\rho)}$ and let $\rho \to \infty$. Ordinarily, $\mathcal{C}^{(\infty)}_{a,b,c}(0)$ moves instantaneously to the state $\mathcal{C}^{(\infty)}_{a+c,b+c,0}(0+)$ and evolves thereafter according to $\mathscr{L}f(a+c,b+c,0,0)$. However, our second step is to make a notational change: we reuse the third and fourth entries of f by separately tracking the half-fragment lineages that *originated* as full fragments: we write it as a process initiated at (a,b,c,c) and evolving according to the generator

$$\begin{aligned}\mathscr{L}^{(\infty)}f(a,b,c,d) = \binom{c}{2} f(a,b,c-1,d) + \binom{d}{2} f(a,b,c,d-1) \\ + R_{a,b,c,d}\mathscr{G}f(a,b,c,d) - \left[\binom{c}{2} + \binom{d}{2} + R_{a,b,c,d}\right] f(a,b,c,d). \qquad (4.2)\end{aligned}$$

Third, we introduce an *artificial* recombination process which induces transitions of the form $(a,b,c,c) \mapsto (a+1,b+1,c-1,c-1)$ at rate $\rho c/2$. This does not reflect any concrete evolutionary dynamic but merely acts as a mathematical device to facilitate a coupling between the two processes. (As a minor technical detail, we should like to allow the process ultimately to reach a state of the form $(a,b,0,0)$. We therefore make a minor adjustment, below, to this artificial process to allow for it to act even if one of c or d is 0.) We therefore have the following definition.

Definition 4.2. *The ancestral process $\mathcal{D}^{(\rho)} = (\mathcal{D}^{(\rho)}_{a,b,c}(t) : t \geq 0)$ is a continuous-time Markov process on $\mathbb{N}^4$ such that $\mathcal{D}^{(\rho)}_{a,b,c}(0) = (a,b,c,c)$ a.s., and with infinitesimal gener-*

Table 2: Transition rates of events in the two ancestral processes $\mathcal{C}^{(\rho)}$ and $\mathcal{D}^{(\rho)}$.

	Transition	Rate	
Type	$(a,b,c,d)\mapsto$	$\mathcal{C}^{(\rho)}$	$\mathcal{D}^{(\rho)}$
I	$(a,b,c-1,d-1)$	$c(c-1)/2^*$	0
II	$(a,b,c-1,d)$	0	$c(c-1)/2$
III	$(a,b,c,d-1)$	0	$d(d-1)/2$
IV	$(a-1,b,c,d)$	$a(a+2c-1)/2$	$a(a+2c-1)/2$
V	$(a,b-1,c,d)$	$b(b+2d-1)/2$	$b(b+2d-1)/2$
VI	$(a-1,b-1,c+1,d+1)$	ab	ab
VII	$(a+\mathbb{I}\{c>0\},b+\mathbb{I}\{d>0\},$ $c-\mathbb{I}\{c>0\},d-\mathbb{I}\{d>0\})$	$\rho c/2^*$	$\rho\max\{c,d\}/2$

*Defined only when $c=d$.

ator

$$\mathscr{H}f(a,b,c,d) := \mathscr{L}^{(\infty)}f(a,b,c,d) + \frac{\rho\max\{c,d\}}{2}[f(a+\mathbb{I}\{c>0\},b+\mathbb{I}\{d>0\},c-\mathbb{I}\{c>0\},d-\mathbb{I}\{d>0\})-f(a,b,c,d)], \tag{4.3}$$

where $f:\mathbb{N}^4\to\mathbb{R}$ is an appropriate test function.

Transitions of this process are also summarized in Table 2, and henceforth we will refer to the numberings of each type of transition given in the table. It is important to keep in mind that although ρ appears as a parameter in (4.3), the process $\mathcal{D}^{(\rho)}$ acts as if the two loci are independent. The process with rate depending on ρ is simply an artificial relabelling of lineages. A key observation is that this artificial process does not affect the distribution of the marginal coalescent trees, so $\mathcal{C}^{(\infty)}$ and $\mathcal{D}^{(\rho)}$ have the same sampling distribution.

To summarize, we have defined two Markov processes on $\mathbb{N}^4$, $\mathcal{C}^{(\rho)}$ and $\mathcal{D}^{(\rho)}$, which describe two-locus ancestral processes going backwards in time and with respective generators $\mathscr{L}$ and $\mathscr{H}$. $\mathscr{L}$ is the generator of a standard process with recombination parameter ρ. $\mathscr{H}$ 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 $\mathcal{C}^{(\rho)}_{a,b,c}(t)$ can be matched by an artificial recombination event in $\mathcal{D}^{(\rho)}_{a,b,c}(t)$, a coalescence of type IV in $\mathcal{C}^{(\rho)}_{a,b,c}(t)$ can be matched by a coalescence of type IV in $\mathcal{D}^{(\rho)}_{a,b,c}(t)$, 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 (4.1) and (4.3), we see that a coupling will fail if there is a type I transition in $\mathcal{C}^{(\rho)}$ or if there is a type II or type III transition in $\mathcal{D}^{(\rho)}$. Define the failure times

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

and

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

the first time that both loci find a most recent common ancestor in the coupled processes (with the convention $\inf \varnothing = \infty$). If $T^{\text{MRCA}}_{a,b,c} < \min\{T^{(1)}_{a,b,c}, T^{(2)}_{a,b,c}, T^{(3)}_{a,b,c}\}$, we say that the coupling has been *successful*. We are now in a position to verify the observation made in Section 1: that we need consider whether or not a coupling has been successful only as far back as the first time that no lineages ancestral to both loci survive. For if we reach this point then, even further back in time, jointly ancestral lineages may arise again temporarily (with $c \geq 1$), but the coupling can fail only in the unlikely [i.e. $O(\rho^{-2})$] event that $c \geq 2$. We formalize this argument in the following lemma.

Lemma 4.3. *If $c \in \{0,1\}$, the coupling between $\mathcal{C}^{(\rho)}$ and $\mathcal{D}^{(\rho)}$ fails with probability $O(\rho^{-2})$, as $\rho \to \infty$.*

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 $(\mathcal{C}^{(\rho)}_{a,b,1}, \mathcal{D}^{(\rho)}_{a,b,1})$, 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 $O(1)$. □

Lemma 4.4. *The coupling between $\mathcal{C}^{(\rho)}$ and $\mathcal{D}^{(\rho)}$ fails with the following probabilities:*

$$\mathbb{P}(I^{(k)}) = \frac{1}{\rho}\binom{c}{2} + O\left(\frac{1}{\rho^2}\right) \quad \text{as } \rho \to \infty, \quad k = 1,2,3, \tag{4.4}$$

where $I^{(k)} := \{T^{(k)}_{a,b,c} < T^{\text{MRCA}}_{a,b,c}\}$. 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 4.3 it is enough to show that

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

where

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

is the first time $\mathcal{C}^{(\rho)}$ reaches $c=0$. We proceed by induction on c; Lemma 4.3 provides the base cases $c \in \{0,1\}$. First note that for any $c \geq 1$,

$$\mathbb{P}(T^{(1)}_{a,b,c} < U^{(1)}_{a,b,c}) = O\left(\frac{1}{\rho}\right), \tag{4.5}$$

since this event requires at least one transition that is not a recombination. Reading off the relevant probabilities from (4.1), we have for $c \geq 2$:

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

by the inductive hypothesis for the first term on the right and using (4.5) for the second term. By considering

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

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 $c > 0$. □

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

Lemma 4.5. *Let $\mathscr{S}_{a,b,c}$ 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*

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

Proof. The non-recombination event causing $I^{(k)}$ occurs at time $T^{(k)}_{a,b,c}$. Inspection of the generators (4.1) and (4.3) shows that any further transition other than a recombination occurs with probability $O(\rho^{-1})$ during the time that $c > 0$. □

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 $\mathcal{D}^{(\rho)}$ and hence $\mathcal{C}^{(\infty)}$; thus $\mathcal{C}^{(\rho)} \mid I^{(1)\complement}$ has the same sampling distribution as $\mathcal{D}^{(\rho)} \mid (I^{(2)} \cup I^{(3)})^{\complement}$. Even if the coupling fails, Lemmata 4.3 and 4.5, 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 4.5 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 $U^{(k)}_{a,b,c}$; by Lemma 4.3, if it has not failed by $U^{(k)}_{a,b,c}$ then the coupling will not fail after $U^{(k)}_{a,b,c}$.

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

Theorem 4.6. *Expressing the sampling distribution for $(\mathcal{C}^{(\rho)}_{a,b,c}(t) : t \geq 0)$ as in* (1.1)*, the first two terms are given by* (2.1) *and* (2.2)*.*

Proof. Denote by $q_{\mathcal{C}^{(\rho)} \mid I^{(1)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$ the sampling distribution of the process $\mathcal{C}^{(\rho)} \mid I^{(1)}$. By Lemmata 4.3 and 4.5, 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 $\mathcal{D}^{(\rho)}_{a+c-1,b+c-1,0} (\overset{a.s.}{=} \mathcal{C}^{(\infty)}_{a+c-1,b+c-1,0})$. Hence,

$$\begin{aligned} q_{\mathcal{C}^{(\rho)} \mid I^{(1)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) &= \sum_{i=1}^{K}\sum_{j=1}^{L} \frac{\binom{c_{ij}}{2}}{\binom{c}{2}} q_{\mathcal{C}^{(\infty)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}-\boldsymbol{e}_{ij}) + O\left(\frac{1}{\rho}\right), \\ &= \sum_{i=1}^{K}\sum_{j=1}^{L} \frac{\binom{c_{ij}}{2}}{\binom{c}{2}} q^A(\boldsymbol{a}+\boldsymbol{c}_A-\boldsymbol{e}_i) q^B(\boldsymbol{b}+\boldsymbol{c}_B-\boldsymbol{e}_j) + O\left(\frac{1}{\rho}\right). \end{aligned} \tag{4.6}$$

(We can also ignore the possibility of mutation prior to $U^{(1)}_{a,b,c}$ since, by the same argument as in Lemma 4.5, a mutation occurs during this phase with probability $O(\rho^{-1})$.)

Similarly,

$$q_{\mathcal{D}^{(\rho)}|I^{(2)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = \sum_{i=1}^{K} \frac{\binom{c_{i\cdot}}{2}}{\binom{c}{2}} q_{\mathcal{C}^{(\infty)}}(\boldsymbol{a}+\boldsymbol{c}_A-\boldsymbol{e}_i, \boldsymbol{b}+\boldsymbol{c}_B, \boldsymbol{0}) + O\left(\frac{1}{\rho}\right),$$
$$= \sum_{i=1}^{K} \frac{\binom{c_{i\cdot}}{2}}{\binom{c}{2}} q^A(\boldsymbol{a}+\boldsymbol{c}_A-\boldsymbol{e}_i) q^B(\boldsymbol{b}+\boldsymbol{c}_B) + O\left(\frac{1}{\rho}\right), \tag{4.7}$$
$$q_{\mathcal{D}^{(\rho)}|I^{(3)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = \sum_{j=1}^{L} \frac{\binom{c_{\cdot j}}{2}}{\binom{c}{2}} q_{\mathcal{C}^{(\infty)}}(\boldsymbol{a}+\boldsymbol{c}_A, \boldsymbol{b}+\boldsymbol{c}_B-\boldsymbol{e}_j, \boldsymbol{0}) + O\left(\frac{1}{\rho}\right),$$
$$= \sum_{j=1}^{L} \frac{\binom{c_{\cdot j}}{2}}{\binom{c}{2}} q^A(\boldsymbol{a}+\boldsymbol{c}_A) q^B(\boldsymbol{b}+\boldsymbol{c}_B-\boldsymbol{e}_j) + O\left(\frac{1}{\rho}\right), \tag{4.8}$$

and so, together with Lemma 4.4 and the observation that

$$\mathbb{P}([I^{(2)} \cup I^{(3)}]^{\complement}) q_{\mathcal{D}^{(\rho)}|(I^{(2)}\cup I^{(3)})^{\complement}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = q_{\mathcal{D}^{(\rho)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$$
$$- \mathbb{P}(I^{(2)}) q_{\mathcal{D}^{(\rho)}|I^{(2)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) - \mathbb{P}(I^{(3)}) q_{\mathcal{D}^{(\rho)}|I^{(3)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) + O(\rho^{-2}),$$

we obtain

$$q_{\mathcal{D}^{(\rho)}|(I^{(2)}\cup I^{(3)})^{\complement}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = \left[1+\frac{2}{\rho}\binom{c}{2}\right]\left[q_{\mathcal{D}^{(\rho)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})\right.$$
$$\left. - \frac{1}{\rho}\binom{c}{2} q_{\mathcal{D}^{(\rho)}|I^{(2)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) - \frac{1}{\rho}\binom{c}{2} q_{\mathcal{D}^{(\rho)}|I^{(3)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})\right] + O\left(\frac{1}{\rho^2}\right). \tag{4.9}$$

The key decomposition is then

$$q(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = \mathbb{P}(I^{(1)}) q_{\mathcal{C}^{(\rho)}|I^{(1)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) + \mathbb{P}(I^{(1)\complement}) q_{\mathcal{C}^{(\rho)}|I^{(1)\complement}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$$
$$= \mathbb{P}(I^{(1)}) q_{\mathcal{C}^{(\rho)}|I^{(1)}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) + \mathbb{P}(I^{(1)\complement}) q_{\mathcal{D}^{(\rho)}|(I^{(2)}\cup I^{(3)})^{\complement}}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) \tag{4.10}$$
$$= q_0(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) + \frac{1}{\rho} q_1(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) + O\left(\frac{1}{\rho^2}\right),$$

using (4.4), (4.6), (4.7), (4.8), and (4.9), with q_0, q_1 given by (2.1) and (2.2), respectively. □

Remark 4.7. It may be possible to use similar arguments to obtain a genealogical interpretation of the second-order term, q_2 in (1.1); for example, genealogies with *two* events that cause the coupling to fail would surely contribute. However, as is clear from the expression for q_2 given in Jenkins and Song (2009, 2010), this is not a simple endeavour and it is seems difficult to interpret some of the components of q_2.

4.2 A new "loose-linkage" coalescent process

Equation (4.10) tells us that, up to $O(\rho^{-2})$, we can obtain the correct sampling distribution using the mixture

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

provided $\alpha < 1$. The coupling used to prove Theorem 4.6 demonstrates that we can *define* a simple stochastic process for weakly correlated loci, $\mathcal{E}^{(\rho)}$, as in Algorithm 2, whose sampling distribution agrees with (2.1) and (2.2) up to $O(\rho^{-2})$.

Algorithm 2 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 $\mathscr{S}_{a,b,c}$ 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 $U^{(1)}_{a,b,c}$ can be sampled according to the rates specified in (4.1). Go to step 3.

2. Otherwise (w.p. $1-\alpha$), sample from $\mathcal{D}^{(\rho)} \mid (I^{(2)} \cup I^{(3)})^{\complement}$ up to time $U^{(2)}_{a,b,c}$ $(= U^{(3)}_{a,b,c})$, which can be achieved by running $\mathcal{D}^{(\rho)}$ as usual according to (4.3) 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 $U^{(k)}_{a,b,c}$ ($k=1$ in the first case above and $k=2$ in the second), construct the remainder of the process independently using $(\mathcal{C}^{(\infty)}(t-U^{(k)}_{a,b,c}) : t \geq U^{(k)}_{a,b,c})$ (with the appropriate starting configuration) back to the first time both loci have found a most recent common ancestor.

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 $\mathcal{D}^{(\rho)} \mid (I^{(2)} \cup I^{(3)})^{\complement}$ is not so obvious.

5 Discussion

We have described two novel stochastic models of evolution for loosely linked, or weakly correlated, 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 (2.1) and (2.2). Our diffusion model is based on a central limit theorem for density dependent population processes, which may be viewed as a separation of the timescales N^{β} and N (in generations), for $0 < \beta < 1$, and pioneered in population genetics by Norman (1975). This contrasts with most research in this area, which focuses on separating the timescales $N^0 = 1$ and N. Indeed, both diffusion (Ethier and Nagylaki, 1980, 1988) and coalescent (Möhle, 1998; Wakeley, 2008) 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 (1975) 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 analogue of these results 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.

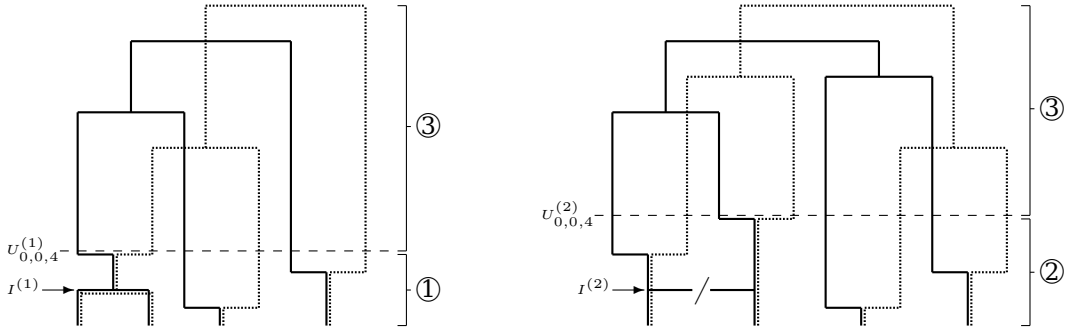

Figure 1: Sampling from the loose-linkage coalescent, $\mathcal{E}^{(\rho)}$, from an initial configuration $(0,0,4)$. Steps of the algorithm in the main text are denoted by circled numbers. *Left*: Commence from step 1 (probability α). Step 1 samples from an approximation to $\mathcal{C}^{(\rho)} \mid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

References

E. Baake and I. Herms. Single-crossover dynamics: finite versus infinite populations. *Bulletin of Mathematical Biology*, 70:603–624, 2008. MR-2389954

A. Bhaskar and Y. S. Song. Closed-form asymptotic sampling distributions under the coalescent with recombination for an arbitrary number of loci. *Advances in Applied Probability*, 44:391–407, 2012. MR-2977401

A. Bhaskar, J. A. Kamm, and Y. S. Song. Approximate sampling formulae for general finite-alleles models of mutation. *Advances in Applied Probability*, 44:408–428, 2012. MR-2977402

M. Birkner, J. Blath, and B. Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193:255–290, 2013.

S. Boitard and P. Loisel. Probability distribution of haplotype frequencies under the two-locus Wright-Fisher model by diffusion approximation. *Theoretical Population Biology*, 71:380–391, 2007.

A. H. Chan, P. A. Jenkins, and Y. S. Song. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics*, 8(12):e1003090, 2012.

A. M. Etheridge and R. C. Griffiths. A coalescent dual process in a Moran model with genic selection. *Theoretical Population Biology*, 75:320–330, 2009.

S. N. Ethier. A limit theorem for two-locus diffusion models in population genetics. *Journal of Applied Probability*, 16(2):402–408, 1979. MR-0531773

S. N. Ethier and R. C. Griffiths. On the two-locus sampling distribution. *Journal of Mathematical Biology*, 29:131–159, 1990. MR-1116000

S. N. Ethier and T. Nagylaki. Diffusion approximations of Markov chains with two time scales and applications to population genetics. *Advances in Applied Probability*, 12: 14–49, 1980. MR-0552945

S. N. Ethier and T. Nagylaki. Diffusion approximations of Markov chains with two time scales and applications to population genetics, II. *Advances in Applied Probability*, 20:525–545, 1988. MR-0955503

S. N. Ethier and T. Nagylaki. Diffusion approximations of the two-locus Wright-Fisher model. *Journal of Mathematical Biology*, 27:17–28, 1989. MR-0984223

S.N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*. Wiley, New York, 1986. MR-0838085

W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112, 1972. MR-0325177

W. J. Ewens. *Mathematical Population Genetics*. Springer-Verlag, New York, 2nd edition, 2004. MR-0554616

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

A. F. Feder, S. Kryazhimskiy, and J. B. Plotkin. Identifying signatures of selection in genetic time series. *Genetics*, 196:509–522, 2014.

W. Feller. Diffusion processes in genetics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 227–246, Berkeley, Calif., 1951. University of California Press. MR-0046022

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

R. C. Griffiths. The two-locus ancestral graph. In I. V. Basawa and R. L. Taylor, editors, *Selected proceedings of the Sheffield symposium on applied probability: 18. IMS Lecture Notes—Monograph series*, volume 18, pages 100–117, 1991. MR-1193063

R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3(4):479–502, 1996.

R. C. Griffiths, P. A. Jenkins, and Y. S. Song. Importance sampling and the two-locus model with subdivided population structure. *Advances in Applied Probability*, 40(2): 473–500, 2008. MR-2433706

P. A. Jenkins and R. C. Griffiths. Inference from samples of DNA sequences using a two-locus model. *Journal of Computational Biology*, 18(1):109–127, 2011.

P. A. Jenkins and Y. S. Song. Closed-form two-locus sampling distributions: accuracy and universality. *Genetics*, 183:1087–1103, 2009.

P. A. Jenkins and Y. S. Song. An asymptotic sampling formula for the coalescent with recombination. *Annals of Applied Probability*, 20(3):1005–1028, 2010. ISSN 1050-5164. doi: 10.1214/09-AAP646. MR-2680556

P. A. Jenkins and Y. S. Song. The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theoretical Population Biology*, 80(2): 158–173, 2011.

P. A. Jenkins and Y. S. Song. Padé approximants and exact two-locus sampling distributions. *Annals of Applied Probability*, 22(2):576–607, 2012. MR-2953564

H.-W. Kang, T. G. Kurtz, and L. Popovic. Central limit theorems and diffusion approximations for multiscale Markov chain models. *Annals of Applied Probability*, 24(2): 721–759, 2014. MR-3178496

N. Kaplan, T. Darden, and R. R. Hudson. The coalescent process in models with selection. *Genetics*, 120:819–829, 1988.

J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3): 235–248, 1982. MR-0671034

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

T. G. Kurtz. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8(2):344–356, 1971. MR-0287609

J. V. Michalowicz, J. M. Nichols, F. Bucholtz, and C. C. Olson. A general Isserlis theorem for mixed-Gaussian random variables. *Statistics and Probability Letters*, 81:1233–1240, 2011. MR-2803768

C. Miura. On an approximate formula for the distribution of 2-locus 2-allele model with mutual mutations. *Genes and Genetic Systems*, 86:207–214, 2011.

M. Möhle. A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Advances in Applied Probability*, 30(2):493–512, 1998. MR-1642850

T. Nagylaki. The Gaussian approximation for random genetic drift. In S. Karlin and E. Nevo, editors, *Evolutionary processes and theory*, pages 629–642. Academic Press, New York, 1986. MR-0954020

T. Nagylaki. Models and approximations for random genetic drift. *Theoretical Population Biology*, 37:192–212, 1990. MR-1042082

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

M. F. Norman. *Markov processes and learning models*, volume 84 of *Mathematics in science and engineering*. Academic Press, New York, 1972. MR-0423546

M. F. Norman. Approximation of stochastic processes by Gaussian diffusions, and applications to Wright-Fisher genetic models. *SIAM Journal on Applied Mathematics*, 29 (2):225–242, 1975. MR-0381749

T. Ohta and M. Kimura. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutations. *Genetics*, 63:229–238, 1969a.

T. Ohta and M. Kimura. Linkage disequilibrium due to random genetic drift. *Genetical Research*, 13(1):47–55, 1969b.

M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-wide inference of ancestral recombination graphs. *PLOS Genetics*, 10(5):e1004342, 2014.

J. Wakeley. The limits of theoretical population genetics. *Genetics*, 169:1–7, 2005.

J. Wakeley. *Coalescent theory: an introduction*. Roberts & Company Publishers, Greenwood Village, Colorado, 2008.

J. Wakeley and O. Sargsyan. The conditional ancestral selection graph with strong balancing selection. *Theoretical Population Biology*, 75:355–364, 2009.

Y. Wang and B. Rannala. Bayesian inference of fine-scale recombination rates using population genomic data. *Philosophical Transactions of the Royal Society B*, 363(1512): 3921–3930, 2008.

S. Wright. Adaptation and selection. In G. L. Jepson, E. Mayr, and G. G. Simpson, editors, *Genetics, Paleontology and Evolution*, pages 365–389. Princeton University Press, Princeton, 1949.

Acknowledgments. This work was supported in part by EPSRC Grants EP/L018497/1 (P.A.J.), EP/K014463 (P.F), NIH Grant R01-GM094402 (P.A.J., Y.S.S), and a Packard Fellowship for Science and Engineering (Y.S.S.). 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." We also thank the generous support of the Simons Institute for the Theory of Computing. This work was developed while P.A.J. and Y.S.S. were participating in the 2014 program on "Evolutionary Biology and the Theory of Computing."