An Ancestral Recombination Graph for Diploid Populations with Skewed Offspring Distribution

Matthias Birkner,* Jochen Blath,† and Bjarki Eldon†,[1]
*Institut für Mathematik, Johannes-Gutenberg-Universität Mainz, 55099 Mainz, Germany, and †Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany

ABSTRACT A large offspring-number diploid biparental multilocus population model of Moran type is our object of study. At each time step, a pair of diploid individuals drawn uniformly at random contributes offspring to the population. The number of offspring can be large relative to the total population size. Similar "heavily skewed" reproduction mechanisms have been recently considered by various authors (*cf. e.g.*, Eldon and Wakeley 2006, 2008) and reviewed by Hedgecock and Pudovkin (2011). Each diploid parental individual contributes exactly one chromosome to each diploid offspring, and hence ancestral lineages can coalesce only when in distinct individuals. A separation-of-timescales phenomenon is thus observed. A result of Möhle (1998) is extended to obtain convergence of the ancestral process to an ancestral recombination graph necessarily admitting simultaneous multiple mergers of ancestral lineages. The usual ancestral recombination graph is obtained as a special case of our model when the parents contribute only one offspring to the population each time. Due to diploidy and large offspring numbers, novel effects appear. For example, the marginal genealogy at each locus admits simultaneous multiple mergers in up to four groups, and different loci remain substantially correlated even as the recombination rate grows large. Thus, genealogies for loci far apart on the same chromosome remain correlated. Correlation in coalescence times for two loci is derived and shown to be a function of the coalescence parameters of our model. Extending the observations by Eldon and Wakeley (2008), predictions of linkage disequilibrium are shown to be functions of the reproduction parameters of our model, in addition to the recombination rate. Correlations in ratios of coalescence times between loci can be high, even when the recombination rate is high and sample size is large, in large offspring-number populations, as suggested by simulations, hinting at how to distinguish between different population models.

DIPLOIDY, in which each offspring receives two sets of chromosomes, one from each of two distinct diploid parents, is fairly common among natural populations. Mathematical models in population genetics tend to assume, however, that all individuals in a population are haploid, simplifying the mathematics. Mendel's laws describe the mechanism of inheritance as composed of two main steps, equal segregation (first law) and independent assortment (second law). The first law proclaims gametes are haploid, *i.e.*, carry only one of each pair of homologous chromosomes. Most models in population genetics are thus models of chromosomes or gene copies. Mendel's second law proclaims independent assortment of alleles at different genes, or loci, into gametes. Linkage of alleles on chromosomes, resulting in nonrandom association of alleles at different loci into gametes, is of course an important exception to the second law.

Coalescent processes (Kingman 1982a,b; Hudson 1983b; Tajima 1983) describe the ancestral relations of chromosomes (or gene copies) drawn from a natural population. The coalescent was initially derived from a Cannings (1974) haploid exchangeable population model. Related ancestral processes take into account population structure (Notohara 1990; Herbots 1997), selection (Krone and Neuhauser 1997; Neuhauser and Krone 1997; Etheridge *et al.* 2010), and recombination between linked loci (Hudson 1983a; Griffiths 1991; Griffiths and Marjoram 1997). The coalescent has proved to be an important advance in theoretical population genetics and a valuable tool for inference of evolutionary histories of populations.

Ancestral recombination graphs (ARGs) (Hudson 1983a; Griffiths 1991; Griffiths and Marjoram 1997) trace ancestral lineages of gene copies at linked loci, in which linkage is broken up by recombination. An ARG is a branching–coalescing graph, in which recombination leads to branching

doi: 10.1534/genetics.112.144329
Manuscript received July 25, 2012; accepted for publication October 30, 2012
[1]Corresponding author: Institute für Mathematik, Technische Universität Berlin, Str. des 17 Juni 136, 10623 Berlin, Germany. E-mail: eldon@math.tu-berlin.de

of ancestral chromosomes and coalescence to segments rejoining. Coalescence events in an ARG may not lead to coalescence of gene copies at individual loci. An example ARG for two linked loci is given below, labeled as ARG(1), with notation borrowed from Durrett (2002). The labels a and b refer to the two alleles (types) at loci 1 and 2, respectively. A single chromosome with two linked alleles is denoted by (ab), while chromosomes carrying ancestral alleles at only one locus are denoted (a) and (b). When coalescence occurs at either locus, the number of alleles at the corresponding locus is reduced by one. The absorbing state, either (ab) or $(a)(b)$, is reached when alleles at both loci have coalesced:

$$\begin{aligned}\text{ARG}(1) : (ab)(ab) &\xrightarrow{r} (a)(b)(ab) \xrightarrow{c} (ab)(b)\\ &\xrightarrow{r}(a)(b)(b) \xrightarrow{c} (a)(b)\\ \text{ARG}(2) : (ab)(ab) &\xrightarrow{r}(a)(b)(ab) \xrightarrow{r} (a)(b)(a)(b) \xrightarrow{c} (a)(b).\end{aligned}$$

In ARG(1), the first transition is a recombination, denoted by $\xrightarrow{r}$, followed by a coalescence ($\xrightarrow{c}$), in which the two alleles at locus 1 coalesce. Graph ARG(1) serves to illustrate two important concepts we are concerned with, namely correlation in coalescence times between alleles at different loci and the restriction to binary mergers of ancestral lineages.

Correlation in coalescence times between types at different loci follows from linkage. Alleles at different loci can become associated due to a variety of factors, including changes in population size, natural selection, and population structure. Within-generation fecundity variance polymorphism induces correlation between a neutral locus and the locus associated with the fecundity variance (Taylor 2009). Sweepstake-style reproduction (Hedgecock *et al.* 1982; Avise *et al.* 1988; Palumbi and Wilson 1990; Beckenbach 1994; Hedgecock 1994; Árnason 2004; Hedgecock and Pudovkin 2011), in which few individuals produce most of the offspring, has also been shown to induce correlation in coalescence times between loci (Eldon and Wakeley 2008). Understanding genome-wide correlations in coalescence times becomes ever more important as multilocus genetic data become ubiquitous.

The ARG exemplified by ARG(1) is characterized by admitting only binary mergers of ancestral lineages; *i.e.*, exactly two lineages coalesce in each coalescence event. The restriction to binary mergers follows from bounds on the underlying offspring distribution, in which the probability of large offspring numbers becomes negligible in a large population (Kingman 1982a,b). Sweepstake-style reproduction, in which few individuals contribute very many offspring to the population, has been suggested to explain the "shallow" gene genealogy observed for many marine organisms (Hedgecock *et al.* 1982; Avise *et al.* 1988; Palumbi and Wilson 1990; Beckenbach 1994; Hedgecock 1994; Árnason 2004; Hedgecock and Pudovkin 2011). Large offspring-number models are models of extremely high variance in individual reproductive output. Namely, individuals can have very many offspring or up to the order of the population size with nonnegligible probability (Sagitov 2003; Schweinsberg 2003; Eldon and Wakeley 2006; Sargsyan and Wakeley 2008; Birkner and Blath 2009). Such models do predict shallow gene genealogies and can be shown to give better fit to genetic data obtained from Atlantic cod (Árnason 2004) than the Kingman coalescent (Birkner and Blath 2008; Birkner *et al.* 2011; Eldon 2011; Steinrücken *et al.* 2012). Different large offspring-number models will no doubt be appropriate for different populations, and the identification of large offspring-number population models for each population is an open problem. For the sake of simplicity and mathematical tractability, the simple large offspring-number model considered by Eldon and Wakeley (2006) is adapted to our situation.

The coalescent processes derived from large offspring-number models belong to a large class of multiple-merger coalescent processes introduced by Donnelly and Kurtz (1999), Pitman (1999), and Sagitov (1999). Multiple-merger coalescent processes (Λ-coalescents), as the name implies, admit multiple mergers of ancestral lineages in each coalescence event, in which any number of active ancestral lineages can coalesce, and at most one such merger occurs each time. In simultaneous multiple-merger coalescent processes (Schweinsberg 2000a; Möhle and Sagitov 2001), any number of multiple mergers can occur each time; *i.e.*, distinct groups of active ancestral lineages can coalesce each time. The ancestral recombination graph derived from our diploid large offspring-number model admits simultaneous multiple mergers of ancestral lineages, as exemplified in ARG(2). The last transition in ARG(2) is a simultaneous multiple merger, in which the two types at each locus coalesce to separate ancestral chromosomes.

To investigate correlations in coalescence times among loci due to skewed offspring distribution, we *formally* derive an ancestral recombination graph, or a coalescent process for many linked loci, from our diploid large offspring-number model. The key to the proof of convergence to an ancestral recombination graph from our diploid model lies in resolving the separation-of-timescales phenomenon we observe. Following Mendel's laws, the two chromosomes of an offspring come from distinct diploid parents. Chromosomes can therefore coalesce only when in distinct individuals. The ancestral process consists of two phases, a dispersion phase occurring on a "fast" timescale and a coalescence and recombination phase occurring on a "slow" timescale. In the dispersion phase, chromosomes paired together in diploid individuals disperse into distinct individuals. Coalescence and recombination occur only on the slow timescale. Similar separation-of-timescales issues arise in models of populations structured into infinitely many subpopulations (demes) (Taylor and Véber 2009). When viewing the diploid individuals in our model as "demes," our scenario departs from those describing structured populations by allowing only active

ancestral lineages residing in *separate* demes to coalesce. A simple extension of a result of Möhle (1998) yields convergence in our case.

The limiting process we formally obtain is an ancestral recombination graph for many loci admitting *simultaneous* multiple mergers of ancestral chromosomes (lineages). In simultaneous multiple-merger coalescent processes, so-called Ξ-coalescents, different groups of active ancestral lineages can coalesce to different ancestors at the same time. Such coalescent processes were first studied as more abstract mathematical objects by Schweinsberg (2000a) and derived from general single-locus population models by several authors (Möhle and Sagitov 2001; Sagitov 2003; Sargsyan and Wakeley 2008; Birkner *et al.* 2009). A Ξ-coalescent with necessarily up to quadruple simultaneous multiple mergers arises at each marginal locus (*i.e.*, considering each locus separately) in our model, since four parental chromosomes are involved in each reproduction event. This structure is intrinsically owed to our diploidy assumptions.

Formulas for the correlation in coalescence times between two alleles at two loci are obtained using our ARG. As predicted by J. E. Taylor (personal communication), these correlations will not necessarily be small even for loci separated by a high recombination rate. This is a novel effect not visible in classical models. The correlation structure will of course depend on the underlying coalescent parameters introduced by the large offspring-number model we adopt. An approximation of the expected value of the statistics r^2, commonly used to quantify linkage disequilibrium, is also investigated using our ARG. In addition, we employ our ARG to investigate correlations in ratios of coalescence times between loci for samples larger than two at each locus, using simulations.

A Diploid Population Model with Multilocus Recombination and Skewed Offspring Distribution

The forward population model

Consider a population consisting of $N \in \mathbb{N} \equiv \{1, 2, \ldots\}$ diploid individuals, meaning that each individual contains two *chromosomes*. Each chromosome is structured into $L \in \mathbb{N}$ loci. We assume Moran-type dynamics: At each time step ("generation"), either a *small* or a *large* reproduction event occurs. In a small reproduction event, a single individual chosen uniformly at random from the population dies, and two other distinct individuals are chosen as *parents*. A diploid *offspring* is then formed by choosing one chromosome from each parent (see Figure 1). The parents always persist. A small reproduction event occurs with probability $1 - \varepsilon_N$, in which $\varepsilon_N \in (0, 1)$ depends on N. In a large reproduction event, a fraction $\psi \in (0, 1)$ of the population perishes, meaning that $\lfloor \psi N \rfloor$ individuals die ($\lfloor x \rfloor$ for $x \geq 0$ denotes the largest integer smaller than x). Two distinct individuals are then chosen uniformly from the remaining $N - \lfloor \psi N \rfloor$ individuals to act as parents of $\lfloor \psi N \rfloor$ offspring, and each offspring is formed independently by choosing one (potentially recombined) chromosome from each parent (see Figure 1). The population size always stays constant at N diploid individuals. Individuals that neither reproduce nor die simply persist.

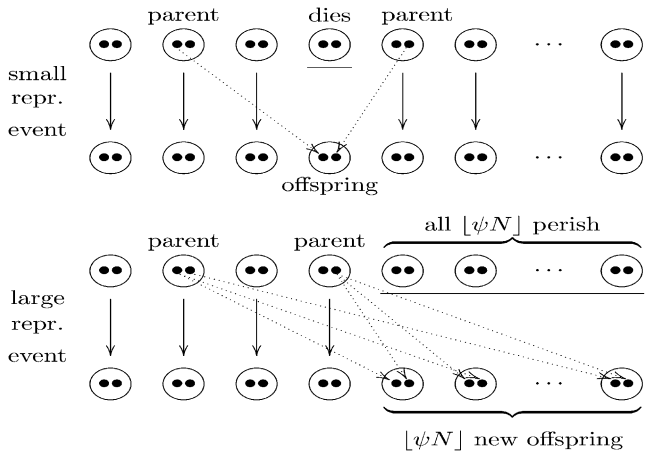

Figure 1 Illustration of "small" and "large" reproduction events without recombination. The dotted arrows indicate the copying of parental chromosomes into offspring chromosomes. The solid arrows indicate individuals that persist.

Given the two parents, genetic types of the offspring individuals are then obtained as follows. Each parent generates a large number of potential offspring chromosomes, of which a fraction $1 - r_N$ are exact copies of the original parental chromosomes and a fraction r_N are *recombinants*. Each chromosome is structured into L loci. Recombination occurs only between loci and never within. If recombination between a pair of chromosomes in a parent occurs between loci ℓ and $\ell + 1 \in \{1, \ldots, L\}$ (where we say that $X \in \{1, \ldots, L - 1\}$ is the *crossover point*), the two chromosomes exchange types at all loci from $\ell + 1$ to L. Only one crossover point is allowed in each recombination event. Let $r_N^{(\ell)}$ denote the probability of recombination between loci ℓ and $\ell + 1$ (*i.e.*, the probability that the potential crossover point X equals ℓ). An offspring chromosome is a recombinant with probability $r_N = r_N^{(1)} + \cdots + r_N^{(L-1)}$. Given that recombination happens, we thus have

$$\mathbb{P}\{X = \ell\} = \frac{r_N^{(\ell)}}{r_N^{(1)} + \cdots + r_N^{(L-1)}}, \quad 1 \leq \ell \leq L - 1.$$

Each pair of recombined chromosomes is formed independently of all other pairs. From this large pool of chromosomes, each new offspring is randomly assigned (independently of all other offspring in the case of a large reproduction event), one potentially recombined chromosome generated by each parent. In addition, the reproduction mechanism in different generations is assumed to be independent.

Ancestral relationships—notation

Now we switch from the forward population model to its ancestral process, running backward in time. Our sample

consists of $n \in \{1, \dots, 2N\}$ chromosomes, each subdivided into L loci. Hence, we need to keep track of the ancestry of nL segments (types/alleles). This implies that the different segments could end up on up to nL distinct chromosomes in nL distinct ancestral individuals. The required notation is now introduced, and our discourse will therefore necessarily become a little bit technical. However, we believe that a precise description of the objects we are working with is essential. The key to understanding our notation is that we are working with enumerated chromosomes and ordered loci on chromosomes.

At present (that is, time step $m = 0$), assume that we consider an even number n of chromosomes carried by $n/2$ individuals. The chromosomes are enumerated from 1 to n, attaching consecutive numbers to chromosomes found in the same individual. Our ancestral process keeps track of the chromosomal ancestral information, that is, which locus is ancestral to which set of sampled chromosomes. That is, in each generation $m \in \mathbb{N}_0$ (backward in time), we record all chromosomes that are *active* in the sense that they carry at least one locus that is ancestral to the same locus of at least one chromosome in generation 0. Denote the number of active chromosomes in generation $m \in \mathbb{N}_0$ by $\beta(m) \in \mathbb{N}$. The number $\beta(m)$ of active chromosomes can both increase, due to recombination, and decrease, due to coalescence, going back in time.

Now we explain our notation for the loci. For each chromosome $j \in [n] := \{1, \dots, n\}$, denote by $\mathbb{L}_\ell^{(j)}(m)$ locus $\ell \in [L]$ on chromosome j at time m. The subsets $\mathbb{L}_\ell^{(j)}(m)$ of $[n]$ contain all the numbers of chromosomes at present (time-step 0) to which locus ℓ on active chromosome number j at time step m is *ancestral*. With this convention, and for each $m \in \mathbb{N}$ and $\ell \in [L]$, the collection

$$\left\{ \mathbb{L}_\ell^{(\mathrm{j})}(m), j = 1, \dots, \beta(m) \right\}.$$

which describes the configuration of segments (*i.e.*, which ones have coalesced and which ones have not) at locus ℓ at time m, is a partition of $[n]$; *i.e.*,

$$\mathbb{L}_\ell^{(\mathrm{j})}(m) \cap \mathbb{L}_\ell^{(\hat{j})}(m) = \emptyset \quad \text{for } j \neq \hat{j};$$

and

$$\bigcup_{j=1}^{\beta(m)} \mathbb{L}_\ell^{(j)}(m) = [n].$$

Thus, with our notation we can correctly describe the configuration of segments among chromosomes at any given time. By $C^{(j)}(m)$ we denote chromosome number j at time m. At time $m = 0$,

$$C^{(j)}(0) := \left\{ \mathbb{L}_1^{(j)}(0), \dots, \mathbb{L}_L^{(j)}(0) \right\} := \{\{j\}, \dots, \{j\}\}.$$

For $m > 0$, consider the jth active chromosome at generation m, where $j \in [\beta(m)]$. The corresponding ancestral information at generation m is encoded via an ordered list of subsets of $[n]$, setting

$$\begin{aligned} C^{(j)}(m) &:= \left\{ \mathbb{L}_1^{(j)}(m), \dots, \mathbb{L}_L^{(j)}(m) \right\}, \\ \mathbb{L}_\ell^{(j)}(m) &\subset [n], \quad \ell \in [L]. \end{aligned} \tag{1}$$

Chromosomes are carried by diploid *individuals*. Keeping track of the grouping of active chromosomes into individuals is important, since by our diploid reproduction mechanism, chromosomal lineages can coalesce only when in *distinct* individuals (see Example B below). In analogy with our previous nomenclature for our ancestral process, an active individual will carry at least one (and at most two) active chromosome(s). Let $b(m)$ denote the number of active individuals at generation m, where $\beta(m)/2 \leq b(m) \leq \beta(m)$ for all m. The ordered list of active chromosomes and the number of active individuals (called a "configuration") at time $m \geq 0$ are denoted by

$$\xi^{n,N}(m) := \left\{ C^{(1)}(m), \dots, C^{(\beta(m))}(m); b(m) \right\}. \tag{2}$$

An individual number i at generation m is denoted by $\mathbb{I}_i(m)$, for $i \in [b(m)]$. An active individual is *single marked*, if carrying one active chromosome, and is *double marked*, if carrying two active chromosomes. Specifying the arrangement of chromosomes in individuals completes our description of the (prelimiting) ancestral process. However, since all active individuals are single marked in the limiting process, our description of the arrangement of chromosomes in individuals is given in section A1.1 in the *Appendix*. That is, each configuration $\xi^{n,N}(m)$ begins with the $2(\beta(m) - b(m))$ ordered consecutive chromosomes of the $\beta(m) - b(m)$ double-marked individuals, followed by the $2b(m) - \beta(m)$ chromosomes contained in single-marked individuals. With this convention, the set of single- and double-marked individuals and the grouping of chromosomes into individuals at generation m are uniquely determined by a configuration $\xi^{n,N}(m)$ of form (2). For notational convenience, the time index m is omitted if there is no ambiguity.

For a given sample size n, the set of all possible ancestral configurations $\xi^{n,N}$ is denoted by $\mathscr{A}_n$. The subset $\mathscr{A}_n^{\mathrm{sm}} \subset \mathscr{A}_n$ of all configurations $\xi^{n,N} = \{C^{(1)}, \dots, C^{(\beta)}; b\}$ with $b = \beta$, *i.e.*, configurations consisting only of single-marked individuals, will play an important role later on. Indeed, all configurations in the limiting model will be confined to the set $\mathscr{A}_n^{\mathrm{sm}}$, and the pairing of chromosomes in individuals will become irrelevant.

The mapping "complete dispersion" (cd),

$$\mathrm{cd} : \mathscr{A}_n \to \mathscr{A}_n^{\mathrm{sm}}$$

breaks up the pairing of chromosomes into diploid double-marked individuals. More precisely, we define

$$\mathrm{cd}\left(\left\{ C^{(1)}, \dots, C^{(\beta)}; b \right\} \right) := \left\{ C^{(1)}, \dots, C^{(\beta)}; \beta \right\}. \tag{3}$$

Configurations in $\mathscr{A}_n^{\mathrm{sm}}$ describe configurations in which all active individuals are single marked, *i.e.*, carry only one active chromosome.

The effects of recombination and coalescence on the ancestral configurations in the case of two typical situations are now illustrated. Example A illustrates recombination, and Example B illustrates coalescence of two chromosomes.

Example A.

Suppose the most recent previous event in the history of a given configuration $\xi^{n,N}(m)$ was a *small reproduction event* (at time $m+1$), and suppose that the resulting offspring individual is currently part of our configuration at time m, but neither of its parents is, and that the offspring individual is single marked, *i.e.*, carries one active chromosome. We obtain $\xi^{n,N}(m+1)$ as follows:

If there is no recombination during the reproduction event, then the configuration in the previous generation remains unchanged; *i.e.*, $\xi^{n,N}(m+1) = \xi^{n,N}(m)$.

If there is recombination, say at a crossover point $X \in \{1, \ldots, L-1\}$, suppose the (single) offspring chromosome is

$$C^{(j)}(m) = \left\{ \mathbb{L}_1^{(j)}(m), \ldots, \mathbb{L}_L^{(j)}(m) \right\}.$$

Necessarily, the two parental chromosomes will be part of the configuration $\xi^{n,N}(m+1)$, residing in the same double-marked individual. More precisely, the two parental chromosomes, say $C^{(\tilde{j})}(m+1)$ and $C^{(\tilde{j+1})}(m+1)$, are determined by (for $\ell \in [L]$)

$$\mathbb{L}_\ell^{(\tilde{j})}(m+1) = \begin{cases} \mathbb{L}_\ell^{(j)}(m): & 1 \le \ell \le X, \\ \emptyset: & X+1 \le \ell \le L, \end{cases}$$

and

$$\mathbb{L}_\ell^{(\tilde{j+1})}(m+1) = \begin{cases} \emptyset: & 1 \le \ell \le X, \\ \mathbb{L}_\ell^{(j)}(m): & X+1 \le \ell \le L, \end{cases}$$

in which $\emptyset$ denotes loci not carrying any ancestral segments. The offspring chromosome is of course not part of $\xi^{n,N}(m+1)$. This transition can be partially trivial (a "silent recombination" event), if the crossover point is not in an "active" area, *i.e.*, if $\mathbb{L}_\ell^{(j)} = \emptyset$ for $X+1 \le \ell \le L$ (or for all $1 \le \ell \le X$). By way of example, with $L=3$, if chromosome $C^{(j)} = \{\{j\}, \{j\}, \{j\}\}$ was a recombinant, and the crossover point occurred between loci 2 and 3, the two parental chromosomes are given by $C^{(\tilde{j})} = \{\{j\}, \{j\}, \emptyset\}$ and $C^{(\tilde{j+1})} = \{\emptyset, \emptyset, \{j\}\}$.

Example B.

Suppose the most recent previous event in the history of a given configuration $\xi^{n,N}(m)$ of chromosomes at generation m is a small reproduction event at time $m+1$, leading to a coalescence of lineages. This is the case, *e.g.*, if a single-marked offspring individual with active chromosome $C^j(m)$ is in our configuration $\xi^{n,N}(m)$, as well as its single-marked parent [say with currently active chromosome $C^j(m)$], from which it actually obtained its active chromosome. Then, to obtain the configuration $\xi^{n,N}(m+1)$, the offspring chromosome $C^{(j)}(m)$ is deleted, and the resulting ancestral chromosome $C^{(j)}(m+1)$ is given by the family of the union of the sets $\mathbb{L}_\ell^{(j)}$ and $\mathbb{L}_\ell^{(\tilde{j})}$,

$$C^{(j)}(m+1) = \left\{ \mathbb{L}_1^{(j)}(m) \cup \mathbb{L}_1^{(\tilde{j})}(m), \ldots, \mathbb{L}_L^{(j)}(m) \cup \mathbb{L}_L^{(\tilde{j})}(m) \right\}. \tag{4}$$

All other chromosomes in $\xi^{n,N}(m+1)$ are copied from $\xi^{n,N}(m)$. Again, taking $L=3$, if chromosomes $C^{(j)} = \{\{j\}, \{j\}, \{j\}\}$ and $C^{(k)} = \{\{k\}, \{k\}, \{k\}\}$ coalesce, the resulting ancestral chromosome is given by $C^{(j)} = \{\{j,k\}, \{j,k\}, \{j,k\}\}$.

Scaling and classification of transitions

To obtain a nontrivial scaling limit for $\{\xi^{n,N}(m)\}$ as $N \to \infty$, the limit theorem of Möhle and Sagitov (2001) (*cf.* also the special case considered in Eldon and Wakeley 2006) suggests one should, for some constant $c > 0$, choose probability $1 - c/N^2$ for the small reproduction events, choose c/N^2 for the large reproduction events, *i.e.*, setting

$$\varepsilon_N = \frac{c}{N^2}, \tag{5}$$

and speed up time by N^2. For the recombination rate to be nontrivial in the limit (*i.e.*, neither 0 nor infinitely large), we require that all recombination values $r_N^{(\cdot)}$ scale in units of N; *i.e.*, for each crossover point $\ell \in [L] \backslash \{L\}$,

$$r_N^{(\ell)} := \frac{r^{(\ell)}}{N}, \quad 0 < r^{(\ell)} < \infty. \tag{6}$$

Thus, even though our timescale is in units of N^2 time steps, recombination is scaled in units of N time steps. On the level of single lineages the probability of recombination is of the order $O(N^{-2})$. Indeed, after a small reproduction event, the probability of drawing an offspring is $1/N$. The probability that the offspring carries a recombined chromosome is of order $O(1/N)$.

Given the cornucopia of possible transitions from $\xi^{n,N}(m)$ to $\xi^{n,N}(m+1)$, it is important to identify those transitions that are expected to be visible in the limiting process.

All possible transitions fall into the following three regimes:

Those transitions that happen at probability of order $O(N^{-2})$ per generation, which will be visible in the limit (since time is scaled by N^2): They are called *effective transitions* and will appear at a finite positive rate in the limit.

Further, there are transitions that happen less frequently, typically with probability of order $O(N^{-3})$ or smaller per generation, which will thus become negligible as $N \to \infty$ and hence be invisible in the limit. These are called *negligible transitions.*

Finally, there are transitions that happen much more frequently [with probability of order $O(N^{-1})$ or even $O(1)$ per generation]. At first sight, one might think that their presence might lead to chaotic behavior in the limit. However, this is not the case. Instead, these transitions will happen "instantaneously" in the limit and result in a projection of the states of our process from $\mathscr{A}_n$ into the subspace $\mathscr{A}_n^{\mathrm{sm}}$, which will be the limiting state space. This is proved below. Such transitions are called *projective* or *instantaneous* transitions. The identity transition is a special case of a projective transformation.

In the *Appendix* (section A1), a full classification of all transitions into the above groups is provided.

Instantaneous and effective transitions

The most important transitions and their effect for the limiting process are now described in detail. Consider the following most recent events in the history of a set of lineages, *i.e.*, events occurring at time $m + 1$, from the perspective of the ancestral process $\xi^{n,N}(m)$ at time m:

Event 1 (silent): A small reproduction event occurs, but the offspring is not active. This is the most likely event and is of the order $O(1)$, but does not affect our ancestral configuration process $\xi^{n,N}(m)$; *i.e.*, $\xi^{n,N}(m+1) = \xi^{n,N}(m)$. This event leads to an identity transition (a trivial instantaneous transition).

Event 2 (dispersion): A small reproduction event occurs, the offspring is active in our sample but neither parent is, and recombination does not occur. This is a relatively frequent event that occurs with a probability of the order $O(N^{-1})$ per generation [since the probability that the offspring is in the sample is $b(m)/N$]. If the offspring carries only one active chromosome, we again see an identity transition; *i.e.*, $\xi^{n,N}(m + 1) = \xi^{n,N}(m)$. If the offspring carries two active chromosomes, *i.e.*, is a double-marked individual, the two active chromosomes will disperse to two separate individuals, who will then become single-marked individuals. Formally, for $\xi = \{C^{(1)}, \ldots, C^{(\beta)}; b\} \in \mathscr{A}_n$ with at least one double-marked individual ($b < \beta$), define the map $\mathrm{disp}_i(\cdot)$: $\mathscr{A}_n \to \mathscr{A}_n$ dispersing the chromosomes paired in individual i,

$$\begin{aligned}\mathrm{disp}_i(\xi) = \{&C^{(1)}, \ldots, C^{(2i-2)}, C^{(2i+1)}, C^{(2i+2)},\\ &\ldots, C^{(2(\beta-b))}, C^{(2i-1)}, C^{(2i)}, C^{(2(\beta-b)+1)},\\ &\ldots, C^{(\beta)}; b+1\}\end{aligned} \tag{7}$$

if $1 \le i \le \beta - b$ and $\mathrm{disp}_i(\xi) := \xi$ otherwise. Recall that the ith double-marked individual has chromosomes labeled $2i - 1$ and $2i$. For $\xi^{n,N}(m)$, if the ith double-marked individual is affected, we have the transition $\xi^{n,N}(m + 1) = \mathrm{disp}_i(\xi^{n,N}(m))$.

The dispersion events will happen instantaneously as $N \to \infty$ (recall we are speeding time up by N^2) and thus will, in the limit, lead to an immediate complete dispersion of all chromosomes paired in double-marked individuals. If in the course of events, a new double-marked individual emerges due to pairing of active chromosomes in the same diploid individual, a dispersion of the chromosomes will occur immediately. Event 2 will hence result in a permanent instantaneous transition, mapping our current state $\xi \in \mathscr{A}_n$ into the subspace $\mathscr{A}_n^{\mathrm{sm}}$ by means of the map cd defined in (3). Our limiting process will thus live, with probability one for each given $t > 0$, in $\mathscr{A}_n^{\mathrm{sm}}$, even if we start with a configuration from $\mathscr{A}_n \backslash \mathscr{A}_n^{\mathrm{sm}}$ at time $t = 0$.

Event 3 (recombination): A small reproduction event occurs, a single-marked offspring but neither parent is in our sample, and recombination affecting the active chromosome at a crossover point x. This event has probability of the order $O(N^{-2})$ per generation and will thus be visible with finite positive rate in the limit. It is an effective transition, which can be described formally as follows. Define the recombination operation recomb acting on chromosome j and crossover point x for a configuration $\xi \in \mathscr{A}_n^{\mathrm{sm}}$ as

$$\begin{aligned}\mathrm{recomb}_{j,x}(\xi) := \Big\{&C^{(1)}, \ldots, C^{(j-1)}, \tilde{C}^{(j,1)},\\ &\tilde{C}^{(j,2)}, C^{(j+1)}, \ldots, C^{(\beta)}; \beta+1\Big\},\end{aligned} \tag{8}$$

where

$$\tilde{C}^{(j,1)} = \left\{\tilde{\mathbb{L}}_1^{(j,1)}, \ldots, \tilde{\mathbb{L}}_L^{(j,1)}\right\}$$

with

$$\tilde{L}_\ell^{(j,1)} = \begin{cases}\mathbb{L}_\ell^{(j)} & : 1 \le \ell \le x-1,\\ \emptyset & : x \le \ell \le L,\end{cases}$$

and

$$\tilde{C}^{(j,2)} = \left\{\tilde{\mathbb{L}}_1^{(j,2)}, \ldots, \tilde{\mathbb{L}}_L^{(j,2)}\right\}$$

with

$$\tilde{\mathbb{L}}_\ell^{(j,2)} = \begin{cases}\emptyset : & 1 \le \ell \le x-1,\\ \mathbb{L}_\ell^{(j)} : & x \le \ell \le L\end{cases}$$

(if one of $\tilde{C}^{(j,1)}, \tilde{C}^{(j,2)} = \{\emptyset, \ldots, \emptyset\}$, we define $\mathrm{recomb}_{j,\alpha}(\xi) := \xi$, giving rise to a *silent* recombination event).

Event 4 (pairwise coalescence): A small reproduction event occurs, one single-marked parent and a single-marked offspring are in the sample, the active chromosome is inherited from the parent in the sample, and recombination does not occur. This event occurs with probability of order $O(N^{-2})$ and will therefore be visible in the limit

with finite positive rate and hence gives rise to an effective transition. It will lead to a binary coalescence of lineages and can formally be described as follows. The ancestral chromosome $\tilde{C}^{(j_1)}$ formed by the coalescence of chromosomes j_1 and j_2 is given by

$$\tilde{C}^{(j_1)} = \left\{ \mathbb{L}_1^{(j_1)} \cup \mathbb{L}_1^{(j_2)}, \ldots, \mathbb{L}_L^{(j_1)} \cup \mathbb{L}_L^{(j_2)} \right\} \tag{9}$$

if $1 \le j_1 < j_2 \le \beta$. Define the binary coalescence operation pairmerge acting on chromosomes j_1 and j_2 ($1 \le j_1 < j_2$) in a configuration $\xi \in \mathscr{A}_n^{\mathrm{sm}}$ as

$$\begin{aligned} \mathrm{pairmerge}_{j_1 j_2}(\xi) := \Big\{ & C^{(1)}, \ldots, \tilde{C}^{(j_1)}, \ldots, \\ & C^{(j_2-1)}, C^{(j_2+1)}, \ldots, C^{(\beta)}; \beta - 1 \Big\} \end{aligned} \tag{10}$$

if $1 \le j_1 < j_2 \le \beta$ (otherwise, we put $\mathrm{pairmerge}_{j_1 j_2}(\xi) := \xi$).

Event 5 (multiple-merger coalescence): A large reproduction event occurs, neither parent but (possibly several) single-marked offspring are in our sample, and recombination does not occur. This is again an event with probability of order $O(N^{-2})$ per generation and therefore will be visible in the limit with finite positive rate and hence gives rise to an effective transition. The offspring chromosomes are assigned their parental chromosomes independently and uniformly at random, since due to an immediate complete dispersion via Event 2 each offspring individual will carry precisely one active chromosome. Now we formally define the multiple-coalescence operation groupmerge for $\xi \in \mathscr{A}_n^{\mathrm{sm}}$ and pairwise disjoint subsets $J_1, J_2, J_3, J_4, \subset [\beta]$ in which either at least one $|J_i| \ge 3$ or at least two of the $|J_i| \ge 2$. This transition is, thus, really different from a pairmerge transition. Let J_j denote the set of offspring chromosomes derived from parental chromosome j. Then

$$\begin{aligned} & \mathrm{groupmerge}_{J_1, J_2, J_3, J_4}(\xi) \\ & := \Big\{ \tilde{C}^{(1)}, \tilde{C}^{(2)}, \tilde{C}^{(3)}, \\ & \qquad \tilde{C}^{(4)}, C^{(j)}, j \in [\beta] \setminus (J_1 \cup J_2 \cup J_3 \cup J_4); \tilde{\beta} \Big\} \end{aligned} \tag{11}$$

with $((x)^+ := \max(x, 0))$

$$\tilde{\beta} = \beta - \sum_{j=1}^{4} \big(|J_j| - 1 \big)^+$$

and the four parental chromosomes, at least one of which is involved in a merger, are given by ($1 \le i \le 4$),

$$\tilde{C}^{(i)} = \left\{ \bigcup_{j \in J_i} \mathbb{L}_1^{(j)}, \ldots, \bigcup_{j \in J_i} \mathbb{L}_L^{(j)} \right\}.$$

The chromosome(s) $C^{(j)}$ appearing in $\mathrm{groupmerge}_{J_1, J_2, J_3, J_4}(\xi)$ denote the chromosomes in ξ that are not involved in a merger.

All other events: These will either not affect our ancestral process or have a probability of order smaller than N^{-2} so that they will be absent in the limit after rescaling. A complete classification of these events is given in the *Appendix* (section A1).

The limiting dynamics and state space

The expected dynamics of the limiting continuous-time Markov chain $\{\xi(t), t \ge 0\}$, taking values in $\mathscr{A}_n$, as $N \to \infty$, is now briefly discussed:

Complete dispersion (Event 2) of the sampled chromosomes is the first event to occur (between times $t = 0$ and $t = 0^+$). By $\mathbb{I}_i$ we denote individual number i (see section A1.1 in *Appendix*). At time $t = 0$ when $\xi(0) \in \mathscr{A}_n$ we assume all n sampled chromosomes are paired in double-marked individuals (n even):

$$\xi(0) = \left\{ \mathbb{I}_i : \mathbb{I}_i = \left\{ C_0^{(2i-1)}, C_0^{(2i)} \right\}, \ 1 \le i \le \frac{n}{2} \right\}. \tag{12}$$

Immediately (at time 0^+), the chromosomes disperse into single-marked individuals,

$$\begin{aligned} \xi(0^+) &= \mathrm{cd}(\xi(0)) \\ &= \left\{ \mathbb{I}_i : \mathbb{I}_i = \left\{ C_0^{(i)}, \emptyset \right\}, \ 1 \le i \le n \right\} \\ &= \left\{ C_{0^+}^{(1)}, \ldots, C_{0^+}^{(n)}; n \right\} \in \mathscr{A}_n^{\mathrm{sm}}. \end{aligned} \tag{13}$$

Throughout the evolution of the process, whenever double-marked individuals appear (*e.g.*, from a coalescence-of-lineages event), Event 2 will immediately change our configuration to the corresponding "all dispersed" configuration; *i.e.*, for each $t > 0$,

$$\xi(t^+) = \mathrm{cd}(\xi(t)) \in \mathscr{A}_n^{\mathrm{sm}}.$$

Such "flickering" states will not affect any quantities of interest of our genealogy, so we can assume that they will be removed from the limit by choosing the càdlàg modification of $\{\xi(t), t \ge 0\}$, taking only values in $\mathscr{A}_n^{\mathrm{sm}}$ for all $t > 0$ (this modification does not affect the finite-dimensional distributions of $\{\xi(t), t \ge 0\}$).

Recombination (Event 3) appears in the limiting process at total rate $r = r^{(1)} + \cdots + r^{(L-1)}$, where a certain recombination involving a given crossover point ℓ appears with rate $r^{(\ell)}$ on any lineage. Indeed, from our scaling considerations, we have that the probability of not seeing a recombination at ℓ in a small resampling event for more than $N^2 t$ scaled time units for a given single-marked individual satisfies ($r_N^{(\ell)} = r^{(\ell)}/N$)

$$\left(1 - \left(1 - \frac{c}{N^2} \right) r_N^{(\ell)} \frac{1}{N} \right)^{N^2 t} \to e^{-r^{(\ell)} t},$$

as $N \to \infty$ [recall (6); the probability for any given individual to be the child in a small reproduction event is $1/N$]; hence the waiting time for this event to happen is exponential with rate $r^{(\ell)}$.

Coalescences appear according to the effective transitions described by Events 4 and 5. From the point of view of a given pair of active chromosomes in different individuals, a single pairwise coalescence will occur at rate $1 + c(\psi^2/4)C_{\beta;2;\beta-2}$ with $C_{\beta;2;\beta-2}$ from (15) (with $r = 1$, $s = \beta - 2$), where the 1 comes from a pairwise coalescence according to a small reproduction event and the $c(\psi^2/4)C_{\beta;2;\beta-2}$ from a large merger event (the rates can be easily derived from considerations similar to the recombination rate r above), recalling that both coalescing chromosomes have to "successfully flip a ψ-coin" to take part in the large coalescence event and then are uniformly distributed into four groups according to the choice of any of the four potential parental chromosomes.

Given that large coalescence events (involving at least three individuals or at least two simultaneous pairwise mergers) happen with overall rate $c(\psi^2/4)$ times the corresponding coalescence rate of a Ξ-coalescent, obtained from the number of individuals taking part in the merger independently with probability ψ, the participating individuals are then distributed uniformly into four groups according to the chosen parental chromosome. The corresponding rate is given in the third line of (14) [*cf*. also (15)].

The limiting ancestral process

According to the above consideration, it is now plausible to consider the following limiting Markov chain as the ancestral limiting process. This fact is proved below, with most computations provided in the *Appendix*. The mth falling factorial is given by $(a)_m := a(a-1)\dots(a-m+1)$, $(a)_0 := 1$. The operations pairmerge, recomb, and groupmerge for elements of $\mathscr{A}_n^{\mathrm{sm}}$ were defined above in the section on scaling. Now we define the generator of the continuous-time ancestral recombination graph derived from our model.

Definition 1.1 (limiting multilocus diploid ancestral recombination graph). *The continuous-time Markov chain* $\{\xi(t), t \geq 0\}$ *with values in* $\mathscr{A}_n^{\mathrm{sm}}$, *initial condition* $\xi(0) := \mathrm{cd}(\xi)$ *for* $\xi \in \mathscr{A}_n$ *and transition matrix* G, *with entries for elements* $\xi', \xi \in \mathscr{A}_n^{\mathrm{sm}}$, $\xi' \neq \xi$, *is given by* $(J := (J_1, \dots, J_4))$

$$G(\xi,\xi') = \begin{cases} 1 + c\frac{\psi^2}{4}C_{\beta;2;\beta-2} & \text{if} \quad \xi' = \mathrm{pairmerge}_{j_1,j_2}(\xi) \\ r^{(\ell)} & \text{if} \quad \xi' = \mathrm{recomb}_{j,\ell}(\xi) \\ c\frac{\psi^2}{4}C_{\beta;|J|} & \text{if} \quad \xi' = \mathrm{groupmerge}_J(\xi) \\ 0 & \text{for all other} \quad \xi' \neq \xi \end{cases} \tag{14}$$

(*where in the penultimate line we consider only cases where either at least one* $|J_i| \geq 3$ *or at least two of the* $|J_i| \geq 2$), *with*

$$C_{\beta;|J|} := C_{\beta;|J_1|,|J_2|,|J_3|,|J_4|;\beta-(|J_1|+|J_2|+|J_3|+|J_4|)}$$

and $(s = b - k_1 - \dots - k_r \geq 0$, $x \wedge y := \min(x,y))$,

$$C_{b;k_1,\dots,k_r;s} = \frac{4}{\psi^2}\sum_{l=0}^{s\wedge(4-r)}\binom{s}{l}\frac{(4)_{r+l}}{4^{k_1+\dots+k_r+l}}\cdot (1-\psi)^{s-l}(\psi)^{k_1+\dots+k_r+l} \tag{15}$$

For the diagonal elements, one has of course

$$G(\xi,\xi) = -\sum_{\xi'\neq\xi,\xi'\in\mathscr{A}_n^{\mathrm{sm}}} G(\xi,\xi'). \tag{16}$$

The rates in (15) are the transition rates of the Ξ-coalescent (a simultaneous multiple-merger coalescent) with

$$\Xi = \delta_{(\psi/4,\psi/4,\psi/4,\psi/4,0,0,\dots)},$$

when r distinct groups of ancestral lineages merge. The number of lineages in each group is given by $k_1, \dots, k_r$, given β active ancestral lineages. The number $s = \beta - (k_1 + \dots + k_r) \geq 0$ gives the number of lineages (ancestral chromosomes) unaffected by the merger (*cf*. Schweinsberg 2000a, Theorem 2). The particular form of Ξ given above follows from the fraction ψ of the population replaced by the offspring of the two parents in a large reproduction event and our assumption that each parent contributes exactly one chromosome to each offspring. We have the following convergence result:

Theorem 1.2. *Let* $\{\xi^{n,N}(m), m \geq 0\}$ *be the ancestral process of a sample of n chromosomes in a population of size N and assume the scaling relations* (5) and (6). Then, starting from $\xi^{n,N}(0) \in \mathscr{A}_n$, *we have that*

$$\{\xi^{n,N}(\lfloor N^2 t\rfloor)\} \to \{\xi(t)\}, \quad \text{as} \quad N \to \infty,$$

in the sense of the finite-dimensional distributions on the interval $(0, \infty)$. The initial value *of the limiting process is given by*

$$\xi(0) = \mathrm{cd}(\xi^{n,N}(0)) \in \mathscr{A}_n^{\mathrm{sm}}.$$

A proof can be found in the *Appendix*. If $c = 0$, the classical ancestral recombination graph for a diploid population with recombination in the spirit of Griffiths and Marjoram (1997) results.

General Diploid Moran-Type Models: "Random" ψ

One of the aims of the present work is to understand the genome-wide correlations in gene genealogies induced by sweepstake-style reproduction. So far, we have discussed this for a very simple example of a sweepstake mechanism (analog to the one considered in Eldon and Wakeley 2006). More precisely, the *fraction* $\psi \in (0, 1)$ of the population replaced by the offspring of a single pair of individuals in a large offspring-number event has hitherto been assumed to be (approximately) constant. Along the lines of the previous discussion, an ancestral recombination graph with a *randomized* offspring distribution can be derived (a comprehensive discussion of single-locus haploid Moran models in the domain of attraction of Λ-coalescents can be found in a recent article by Huillet and Möhle 2011). Even though ψ is now considered

a random variable, the population size stays constant at N diploid individuals. Allowing ψ to be random may be biologically more realistic than taking ψ to be a constant. On the other hand, the problem of identifying suitable classes of probability distributions for ψ, reflecting the specific biology of given natural populations, is still open and an area of active research.

To explain the convergence arguments when ψ is random, let the random variable Ψ_N, taking values in $[N-2]$, denote the random number of diploid offspring contributed by the single reproducing pair of parents at each time step; a new realization of Ψ_N is drawn before each reproduction event. Again, we consider the effect of such a reproduction mechanism on coalescence events in a *sample*. The probability that two given chromosomes residing in two single-marked individuals in the sample coalesce in the previous time step given the value of Ψ_N is

$$\begin{aligned}&\mathbb{P}(\{\text{pair coalescence}\}\,|\,\Psi_N=k)\\&\quad=\frac{1}{4}\delta_{\{k=1\}}\frac{4}{N(N-1)}+\frac{1}{4}\delta_{\{k>1\}}\left(\frac{4k}{N(N-1)}+\frac{k(k-1)}{N(N-1)}\right),\end{aligned}\tag{17}$$

where the first and second terms on the right-hand side describe the case where one parent and one offspring are drawn, the third term covers the case where two offspring are drawn, and the 1/4 accounts for the probability that the two chromosomes in question must descend from the same parental chromosome. Define

$$c_N:=4\mathbb{P}(\{\text{pair coalescence}\})\tag{18}$$

$$=4\sum_{k=1}^{N-2}\mathbb{P}(\{\text{pair coalescence}\}\,|\,\Psi_N=k)\mathbb{P}(\Psi_N=k)$$

$$=\mathbb{E}\left[\frac{\Psi_N(\Psi_N+3)}{N(N-1)}\right]\tag{19}$$

(the factor 4 facilitates comparison with the haploid case). The sequence of laws $\mathcal{L}(\Psi_N)$, $N\in\mathbb{N}$, is assumed to satisfy the following three conditions,

$$c_N\to 0\quad\text{as}\quad N\to\infty,\tag{20}$$

$$\begin{aligned}\frac{c_N}{\mathbb{E}[\Psi_N/N]}=\frac{1/\mathbb{E}[\Psi_N/N]}{1/c_N}&=\frac{\mathbb{E}[\Psi_N(\Psi_N+3)]}{(N-1)\mathbb{E}[\Psi_N]}\\&\to 0\quad\text{as}\quad N\to\infty,\end{aligned}\tag{21}$$

and there exists a probability measure F on $[0,1]$ such that

$$\frac{1}{c_N}\mathbb{P}(\Psi_N>Nx)\xrightarrow[N\to\infty]{}\int_x^1\frac{1}{y^2}F(dy)\tag{22}$$

for all continuity points $x\in(0,1]$ of F.

Condition (20) is necessary for any limit process of the genealogies to be a continuous-time Markov chain, condition (21) ensures that a separation-of-timescales phenomenon occurs, and (22) fixes the limit dynamics of the large merging events [it is analogous to Sagitov 1999, necessary condition (13) in the haploid case]. In the proof of convergence to a limit process we recall equivalent conditions to (22) (see *Appendix*, section A4). Condition (20) implies (see section A4 in the *Appendix*)

$$\mathbb{E}\left[\frac{\Psi_N}{N}\right]\to 0\quad\text{as}\quad N\to\infty;\tag{23}$$

i.e., the probability for a given individual to be an offspring in a given reproduction event becomes small. Hence, (23) and (21) together show that there will be two diverging timescales: The "short" timescale $1/\mathbb{E}[\Psi_N/N]$ on which chromosomes paired in double-marked individuals disperse into single-marked individuals and the "long" timescale $1/c_N$ over which we observe nontrivial ancestral coalescences.

To obtain a nontrivial genealogical limit process, we then speed up time by a factor of $4/c_N$; *i.e.*, $4/c_N$ reproduction events correspond to one coalescent time unit (see Theorem 1.3 below). This time rescaling is chosen for two chromosomes to coalesce at rate 1 in the limit. The required scaling relation for the recombination rates is now

$$r_N^{(\ell)}\sim\frac{c_N}{4\mathbb{E}[\Psi_N/N]}r^{(\ell)}\quad\text{as}\quad N\to\infty\tag{24}$$

with $r^{(\ell)}\in[0,\infty)$ fixed for $\ell=1,\dots,L-1$ [where $f(N)\sim g(N)$ means $\lim_{N\to\infty}f(N)/g(N)=1$]. An intuitive explanation for the requirement (24) is that since the probability for a given individual to be an offspring in a given reproduction event is $\mathbb{E}[\Psi_N/N]$, after speeding up time by $4/c_N$, on any lineage recombination events between loci ℓ and $\ell+1$ occur as a Poisson process with rate $r^{(\ell)}$.

A simple sufficient condition for (21) is the following: For any $\varepsilon>0$,

$$N\mathbb{P}(\Psi_N>\varepsilon N)\to 0\quad\text{as}\quad N\to\infty.\tag{25}$$

Indeed, we have, by assuming $N>\varepsilon N$,

$$\begin{aligned}\mathbb{E}\left[\Psi_N^2\right]&=\sum_{k=1}^{\lfloor\varepsilon N\rfloor}k^2\mathbb{P}(\Psi_N=k)+\sum_{k=\lfloor\varepsilon N\rfloor+1}^{N}k^2\mathbb{P}(\Psi_N=k)\\&\le\sum_{k=1}^{\lfloor\varepsilon N\rfloor}k\varepsilon N\mathbb{P}(\Psi_N=k)+\sum_{k=\lfloor\varepsilon N\rfloor+1}^{N}N^2\mathbb{P}(\Psi_N=k)\\&\le\varepsilon N\mathbb{E}[\Psi_N]+N^2\mathbb{P}(\Psi_N>\varepsilon N).\end{aligned}$$

Dividing by $N\mathbb{E}[\Psi_N]$ gives

$$\frac{\mathbb{E}\left[\Psi_N^2\right]}{N\mathbb{E}[\Psi_N]}\le\varepsilon+\frac{N\mathbb{P}(\Psi_N>\varepsilon N)}{\mathbb{E}[\Psi_N]},$$

and, since $\mathbb{E}[\Psi_N]>1$,

$$\limsup_{N\to\infty}\frac{\mathbb{E}\left[\Psi_N^2\right]}{N\mathbb{E}[\Psi_N]}<\varepsilon+\limsup_{N\to\infty}N\mathbb{P}(\Psi_N>\varepsilon N)=\varepsilon.$$

Condition (21) is now obtained since we can choose ε to be as small as we like.

The limiting genealogical process will then be a continuous-time Markov chain on $\mathscr{A}_n^{\mathrm{sm}}$ with generator matrix G whose off-diagonal elements are given by [for the values on the diagonal we again have (16)]

$$G\left(\xi,\xi'\right)=\begin{cases} C_{\beta;2} & \text{if} \quad \xi'=\text{pairmerge}_{j_1,j_2}(\xi) \\ r^{(\ell)} & \text{if} \quad \xi'=\text{recomb}_{j,\ell}(\xi) \\ C_{\beta;|J|} & \text{if} \quad \xi'=\text{groupmerge}_{J_1,J_2,J_3,J_4}(\xi) \\ 0 & \text{for all other} \quad \xi'\neq\xi, \end{cases} \tag{26}$$

where

$$C_{\beta;|J|}:=C_{\beta;|J_1|,|J_2|,|J_3|,|J_4|;\beta-(|J_1|+|J_2|+|J_3|+|J_4|)},$$

$k=(k_1,\ldots,k_r)$, $|k|=k_1+\ldots+k_r$, and

$$\begin{aligned} C_{b;k;s}&=4\sum_{l=0}^{s\wedge(4-r)}\binom{s}{l}\frac{(4)_{r+l}}{4^{|k|+l}} \\ &\quad\cdot\int_{[0,1]}x^{|k|+l}(1-x)^{s-l}\frac{1}{x^2}F(dx) \\ &=F(\{0\})\delta_{\{r=1,k_1=2\}}+4\sum_{l=0}^{s\wedge(4-r)}\binom{s}{l}\frac{(4)_{r+l}}{4^{|k|+l}} \\ &\quad\cdot\int_{(0,1]}x^{|k|+l}(1-x)^{s-l}\frac{1}{x^2}F(dx) \end{aligned} \tag{27}$$

with F from (22). As in the case of constant ψ, the third line in (26) gives the transition rates for a given merger into r (≤ 4) groups of sizes $k_1,\ldots,k_r$ when β active ancestral lineages are present, with $s=\beta-|k|\geq 0$ lineages unaffected by a given merger of the Ξ-coalescent with

$$\Xi=\int_{[0,1]}\delta_{(x/4,x/4,x/4,x/4,0,0,\ldots)}F(dx)$$

(*cf.* Schweinsberg 2000a, Theorem 2). By way of example, $C_{2;2;0}=1$. Now we can state the convergence of our ancestral recombination graph process with random ψ. The analog of Theorem 1.2 is the following:

Theorem 1.3. *Let $\{\xi^{n,N}(m),\ m\geq 0\}$ be the ancestral process of a sample of n chromosomes in a population of size N with offspring laws $\mathcal{L}(\Psi_N)$ that satisfy* (20), (21), *and* (22), *and assume the scaling relation* (24) *for the recombination rates. Then, starting from $\xi^{n,N}(0)\in\mathscr{A}_n$, we have that*

$$\left\{\xi^{n,N}\left(\left\lfloor\frac{4t}{c_N}\right\rfloor\right)\right\}\to\{\xi(t)\},\quad as\quad N\to\infty,$$

in the sense of the finite-dimensional distributions on the interval $(0,\ \infty)$. *The process $\{\xi(t)\}$ is the Markov chain with generator matrix* (26) *and initial value $\xi(0)$ given by*

$$\xi(0)=\mathrm{cd}\left(\xi^{n,N}(0)\right)\in\mathscr{A}_n^{\mathrm{sm}}.$$

The proof is given in section A4 in the *Appendix*.

While $c_N\geq 1/N^2$ by definition, in principle any decay behavior of c_N that is consistent with $\liminf_{N\to\infty}N^2c_N\geq 1$, and hence any there-from–derived scaling relation between coalescent timescale and model census population size, is possible via a suitable choice of the family $\mathcal{L}(\Psi_N)$, $N\in\mathbb{N}$.

For an extreme example, let $\Psi_N=\lfloor N^\gamma\rfloor$ for some $\gamma\in(0,1)$; then $c_N\sim N^{-2(1-\gamma)}$ and (22) is satisfied with $F=\delta_0$.

The relation with the "fixed-ψ" model is as follows: For Theorem 1.2, we used the simple mixture distribution for Ψ_N,

$$\mathbb{P}\left(\Psi_N=\lfloor\psi N\rfloor\right)\ =1-\mathbb{P}(\Psi_N=1)=\frac{c}{N^2} \tag{28}$$

for Ψ_N, in which $\psi\in(0,1)$ and $c>0$ are both constants. Our choice (28) of law for Ψ_N gives, using (17),

$$\begin{aligned} c_N&=\mathbb{E}\left[\frac{\Psi_N(\Psi_N+3)}{N(N-1)}\right] \\ &=\left(1-\frac{c}{N^2}\right)\ \frac{4}{N(N-1)}+\frac{c}{N^2}\frac{\psi N(\psi N+3)}{N(N-1)} \\ &\sim\frac{1}{N^2}\left(4+c\psi^2\right). \end{aligned}$$

Define $1_{(0,\psi)}(x)=1$ if $x\in(0,\psi)$ and $1_{(0,\psi)}(x)=0$ otherwise. Our choice (28) further gives

$$\begin{aligned} \mathbb{P}\left(\Psi_N>\lfloor Nx\rfloor\right)&=1_{(0,\psi)}(x)\mathbb{P}(\Psi_N>Nx) \\ &=1_{(0,\psi)}(x)cN^{-2}, \end{aligned}$$

and therefore

$$\begin{aligned} &\frac{1}{c_N}\mathbb{P}\left(\Psi_N>\lfloor Nx\rfloor\right)\ \to 1_{(0,\psi)}(x)\frac{c}{4+c\psi^2} \\ &=\int_{(x,1]}y^{-2}F(dy) \end{aligned}$$

with

$$F=\frac{4}{4+c\psi^2}\delta_0+\frac{c\psi^2}{4+c\psi^2}\delta_\psi.$$

Furthermore, $\mathbb{E}[\Psi_N/N]=1/N+O(1/N^2)$; thus

$$\frac{c_N}{4\mathbb{E}[\Psi_N/N]}\sim\frac{1}{N}\frac{4+c\psi^2}{4}$$

and Theorem 1.2 follows from Theorem 1.3 [after rescaling time in the limit process $\{\xi(t)\}$ by a factor of $(4+c\psi^2)/4$].

The constant $C_{b;k}:=C_{b;k_1,\ldots,k_r;s}$ (27) depends on the probability measure F. The form of F will no doubt be different for different populations. We reiterate that resolving the mechanism of sweepstake-style reproduction will require detailed knowledge of the reproductive behavior and the ecology of the organism in question, along with comparison of model predictions to multilocus genetic data. A candidate for F may be the beta distribution with parameters $\vartheta>0$ and $\gamma>0$, in which case the constant $C_{b;k}$ in (26) takes the form ($|k|:=k_1+\cdots+k_r$)

$$C_{b;k} = 4\sum_{\ell}\binom{s}{\ell}(4)_{r+\ell}\left(\frac{1}{4}\right)^{|k|+\ell} \cdot \frac{B(|k| + \ell + \vartheta - 2, s + \gamma - \ell)}{B(\vartheta, \gamma)}, \tag{29}$$

$B(\cdot,\cdot)$ being the Beta function.

Different Scaling Regimes

The mechanism of sweepstake-style reproduction may be different for different populations, and the frequency of large offspring-number events may also be different. The particular timescale of the large reproduction events (we chose $\varepsilon_N = c/N^2$) results in a separation of timescales of the limit process. Resolving the separation-of-timescales problem results in the ARG with generator (14). Different scalings of ε_N result in different limit processes. By way of example, if $N^2\varepsilon_N \to 0$, large offspring-number events are negligible in a large population, and we obtain the ARG associated with the usual Wright–Fisher reproduction, which can be read off Equation 14 by taking $c = 0$. One other scaling regime may seem reasonable, namely taking large offspring-number events to be more frequent than in assumption (5), but not too frequent. In mathematical notation, $N^2\varepsilon_N \to \infty$ and $N\varepsilon_N \to 0$. The ancestral process in this regime is again characterized by instantaneous separation of marked chromosomes into single-marked individuals, followed by coalescence and recombination occurring on the slow timescale. The probability of recombination is proportional to $N\varepsilon_N$ since the slow timescale must be in units proportional to $1/\varepsilon_N$. Hence, small reproduction events become negligible in the limit, and the generator of the limit process is given by

$$G(\xi,\xi') = \begin{cases} \dfrac{\psi^2}{4}C_{\beta;2;\beta-2} & \text{if} \quad \xi' = \text{pairmerge}_{j_1 j_2}(\xi) \\ r^{(\ell)}/r & \text{if} \quad \xi' = \text{recomb}_{j,\alpha}(\xi) \\ \dfrac{\psi^2}{4}C_{\beta;|J|} & \text{if} \quad \xi' = \text{groupmerge}_J(\xi) \\ 0 & \text{for all other} \quad \xi' \neq \xi \end{cases} \tag{30}$$

in which $C_{\cdot;\cdot;\cdot}$ is given by Equation 15. The requirement $N\varepsilon_N \to 0$ is needed to prevent an unreasonably high rate of recombination.

Haploid Analogs

A haploid version of the above model, where only one parent contributes offspring at each time step, is a specific example of a Λ-coalescent, where

$$\Lambda(dx) = \delta_0(dx) + c\psi^2\delta_\psi(dx), \quad \psi \in (0,1), \quad c \in [0,\infty)$$

(see, *e.g.*, Eldon and Wakeley 2006 and Birkner and Blath 2009). More precisely, as the population size N tends to infinity, assume probability $1 - c/N^2$ for the small reproduction events and c/N^2 for the large reproduction events (*i.e.*, choose $\varepsilon_N = c/N^2$), and speed up generation time by N^2. Again, by randomizing ψ and/or switching to different scaling regimes, it is possible to obtain any given Λ-coalescent as limiting genealogy.

Two-Sex Extensions

Recent studies of the spawning behavior of Atlantic cod indicate that cod adopts a lekking behavior, in which males compete for females, and females exercise mate choice (Nordeide and Folstad 2000). Direct microsatellite DNA analysis indicates that although multiple paternity is sometimes detected, the reproductive success is highly skewed among the males; *i.e.*, most of the successfully fertilized eggs can be attributed to a single male (Hutchings *et al.* 1999). Our model thus seems a good approximation to the actual reproduction mechanism of cod. Modifications to allow two distinct genders, and multiple paternity, are in principle straightforward.

More General Recombination Models

Our model can easily be enriched to allow also more general recombination events involving more than one crossover point at a time. Furthermore, by letting the number L of loci tend to infinity, a continuous model, where [0, 1] represents a whole chromosome (as in Griffiths and Marjoram 1997), can be accommodated into our framework.

Correlations in Coalescence Times

The marginal process

Every marginal process (marginal with respect to one fixed locus under consideration) of our ancestral recombination graph is a Ξ-coalescent (see Schweinsberg 2000a for notation and details) with

$$\Xi = \delta_0 + c\frac{\psi^2}{4}\delta_{(\psi/4,\psi/4,\psi/4,\psi/4,0,0,\ldots)}.$$

For $r = 0$, all marginals are identical (realization-wise), in particular times to the most recent common ancestor for different loci have correlation 1. However, in contrast to the classical setting, for $r \to \infty$ one expects that the loci will not completely decorrelate, but instead keep positive correlations, as pointed out to us by J. E. Taylor (personal communication). In particular, one will not obtain the product distribution. This observation is a potential starting point for designing tests for the presence of large reproduction events, by comparing correlations for loci at large distance (hence with high recombination rate) under a Kingman- and a Ξ-coalescent–based ARG.

Correlation in coalescence times at two loci

Correlations in coalescence times between two loci have been considered in the context of quantifying association between loci (McVean 2002). Eldon and Wakeley (2008) consider correlations in coalescence times for a haploid population model, admitting large offspring numbers, in which the ancestral process admits only asynchronous multiple mergers of ancestral lineages. To illustrate the effects of the reproduction parameters on the coalescence times, we also consider the probability that coalescence occurs at the same time at the two loci, as well as the expected time until coalescence.

The calculations to obtain the correlations for a sample of size two at two loci (following the approach and notation of Durrett 2002) are shown in the *Appendix*, section A5. As we are now considering the gene genealogy of unlabeled lineages, let us briefly state the sample space. Let a and b denote the types at loci a and b, respectively. The three sample states before coalescence at either locus has occurred can be denoted as $(ab)(ab)$, $(ab)(a)(b)$, and $(a)(b)(b)(b)$. By $(ab)(ab)$ we denote the state of two chromosomes, each carrying ancestral material at both loci. By $(ab)(a)(b)$ we denote the state of one (ab) chromosome in addition to two chromosomes (a) and (b) carrying ancestral types at loci 1 and 2 only, respectively. The notation $(a)(a)(b)(b)$ denotes the state of four chromosomes, each carrying ancestral types at only one locus. Let

$$h(i) := \mathbb{P}(\{T_a = T_b\} \mid i), \quad i \in \{0, 1, 2\}$$

denote the probability that coalescence at the two loci occurs at the same time, given that the process starts in state i, in which i refers to the number of double-marked chromosomes (2, 1, or 0). As we are working with the limiting model, all marked individuals are effectively single marked. Under the usual (Kingman-coalescent–based) ARG, $\lim_{r\to\infty} h(i) = 0$ as one would expect. Our model yields

$$\lim_{r\to\infty} h(i) = \frac{c\psi^4}{32 + 8c\psi^2 - c\psi^4}, \quad i \in \{0, 1, 2\}, \tag{31}$$

indicating that even unlinked loci remain correlated due to sweepstake-style reproduction. Figure 2 shows graphs of $h(i)$ as a function of ψ for different values of c and r. As expected, $h(i)$ increases with ψ, at a rate that increases with c.

Under the usual ARG, the expected time $\mathbb{E}_i[T_s]$ until coalescence at either locus, starting from state i is given by $\mathbb{E}_i[T_s] = (1 + h(i))/2$. The random variable T_s can be viewed as the minimum of the time until coalescence occurs at the two loci. As $r \to \infty$, the times T_1 and T_2 until coalescence at the two loci, respectively, become independent and identically distributed exponentials (i.i.d.e.) with rate 1, whose minimum has expected value 1/2. Under our model, the mean of T_s is *not* the minimum of two i.i.d.e. with rate $1+c\psi^2/4$, another reflection of the correlation in gene genealogies induced by sweepstake-style reproduction. Indeed, our model gives

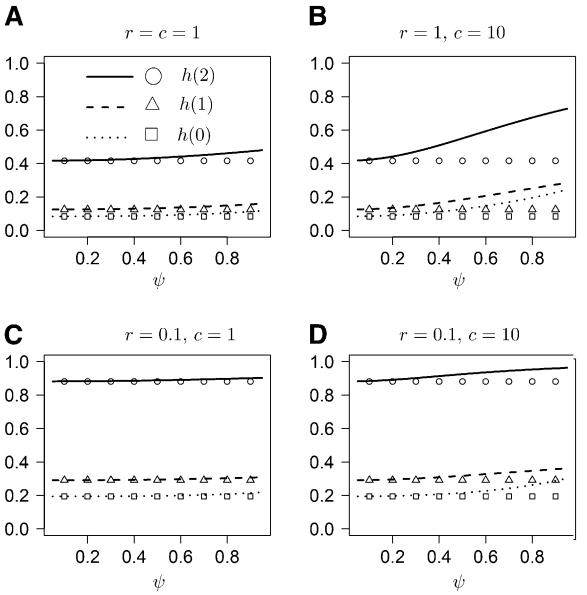

Figure 2 (A and B) The probabilities $h(2)$, $h(1)$, and $h(0)$ as functions of ψ (lines) for different values of r and c. Values of $h(\cdot)$ obtained from the usual Moran model are shown for reference (symbols).

$$\lim_{r\to\infty} \mathbb{E}_i[T_s] = \frac{1}{2}\left(\frac{1}{1 + \chi c\psi^2/4}\right), \quad i \in \{0, 1, 2\},$$

in which $\chi = 1 - \psi^2/8$.

Under our model, $\mathbb{E}_i[T_s]$ decreases with ψ, and the rate of decrease increases with c (Figure 3). The same pattern holds for the expected time $\mathbb{E}_i[T_l]$ until coalescence has occurred at both loci (Figure 4). As $r \to \infty$, $\mathbb{E}_i[T_l]$ associated with the usual ARG approaches the expected value (3/2) of the maximum of two i.i.d.e. with rate 1. Under our model,

$$\lim_{r\to\infty} \mathbb{E}_i[T_l] = \frac{3}{2}\frac{1}{1 + c\psi^2/4}\frac{1}{1 + c\psi^2/4 - c\psi^4/32} + \frac{c\psi^2(6 - \psi^2)}{(c\psi^2 + 4)(4 + c\psi^2 - c\psi^4/8)}$$

while the maximum of two i.i.d.e. with rate λ has expected value $3/(2\lambda)$.

The correlation $\mathrm{cor}_i(T_1, T_2)$ between T_1 and T_2 when starting from one of the three possible sample states $i \in \{0, 1, 2\}$ (see *Appendix*) increases with ψ and more so if c is large (Figure 5). One obtains the following limit relations between $h(i)$ and $\mathrm{cor}_i(T_1, T_2)$ for $i \in \{0, 1, 2\}$:

$$\lim_{r\to\infty} \mathrm{cor}_i(T_1, T_2) = \lim_{r\to\infty} h(i) \quad \text{(see Equation 31)};$$

$$\lim_{r\to 0} \mathrm{cor}_i(T_1, T_2) = \lim_{r\to 0} h(i) \quad \text{(see Equation A35)};$$

$$\lim_{c\to\infty} \mathrm{cor}_i(T_1, T_2) = \lim_{c\to\infty} h(i) \quad \text{(see Equation A34)}.$$

Quantifying the association between alleles at different loci can give insight into the evolutionary history of populations.

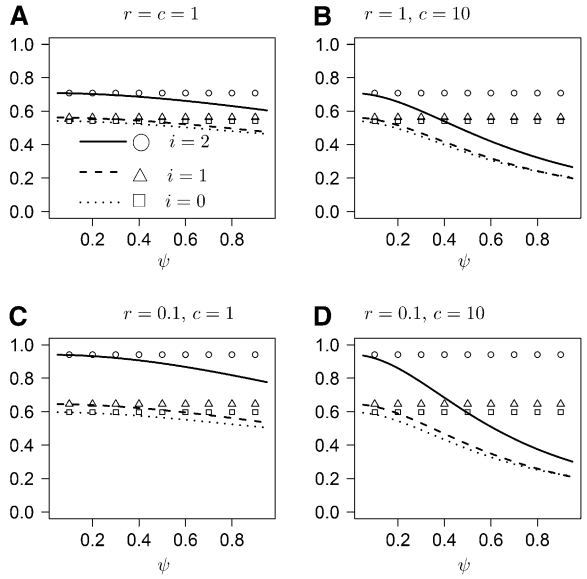

Figure 3 The expected time $\mathbb{E}_i[T_s]$ as a function of ψ for different values of c and r. Values of $\mathbb{E}_i[T_s]$ associated with the case $c = 0$ are shown for reference (symbols).

Let f_a and f_b denote the frequencies of alleles a at locus 1 and b at locus 2, and let f_{ab} denote the frequency of chromosome ab in the total population. The statistic $D_{ab} := f_{ab} - f_a f_b$ measures the deviation from independence, since if the two loci were evolving independently, $f_{ab} = f_a f_b$. A related quantity is the r^2 statistic, defined as

$$r^2 := \frac{D^2}{f_a(1-f_a)f_b(1-f_b)}$$

(Hill and Robertson 1968), assuming $f_a, f_b \notin \{0, 1\}$. In applications, one wants to compare observed values of r^2 calculated from data to the expected value $\mathbb{E}[r^2]$, obtained under an appropriate population model. Calculating the expected value of r^2 is not straightforward, since r^2 is a ratio

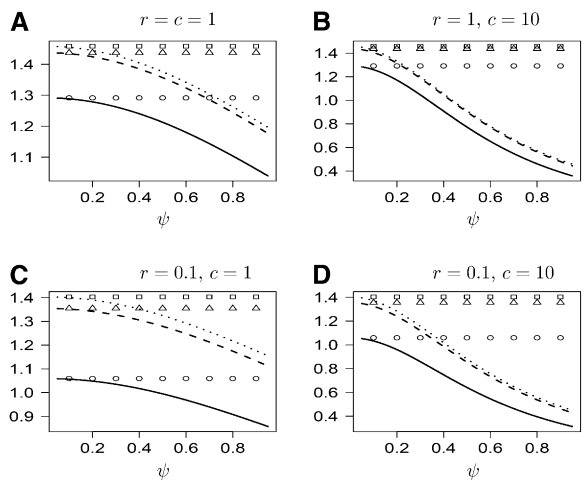

Figure 4 The expected time $\mathbb{E}_i[T_i]$ as a function of ψ for different values of c and r. For explanation of symbols, see Figure 3.

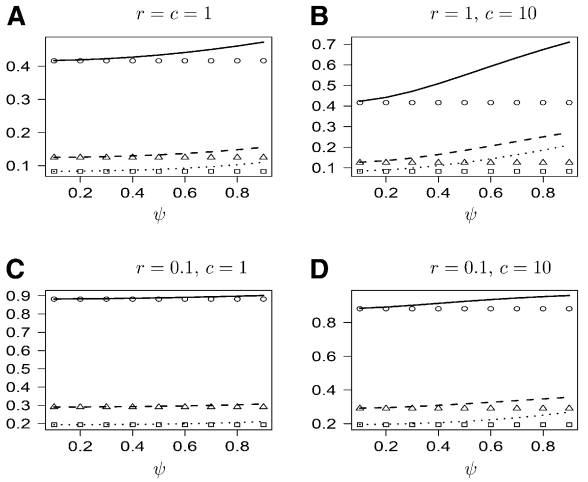

Figure 5 Correlation of the time to coalescence at two loci as a function of ψ, for different values of c and r. For explanation of symbols, see Figure 3.

of correlated random variables. The expected value of r^2 is, instead, approximated by the ratio $\mathfrak{D} = \mathbb{E}[D^2]/\mathbb{E}[f_a(1 - f_a) f_b(1 - f_b)]$ (Ohta and Kimura 1971).

A prediction $\mathfrak{D}$ of linkage disequilibrium in the population can be framed in terms of correlations in coalescence times between two loci for a sample of size two, assuming a small mutation rate (McVean 2002). The prediction rests on approximating the expected value $\mathbb{E}[r^2]$ of the squared correlation statistic r^2 (Hill and Robertson 1968) of association between alleles at two loci by the ratio of expected values (Ohta and Kimura 1971). Following, *e.g.*, Durrett (2002) one can obtain expressions for correlations in coalescence times between two loci for a sample of size two (see *Appendix*). Under our model, one obtains the limit results

$$\lim_{r\to\infty} \mathfrak{D} = 0,$$

$$\lim_{c\to\infty} \mathfrak{D} = \frac{\psi^3 - 16\psi^2 + 56\psi - 80}{\psi^3 - 10\psi^2 + 88\psi - 176}.$$

When ψ is small but c large, one obtains

$$\mathfrak{D} = \frac{5 - 7\psi/2}{11 - 11\psi/2} + O(\psi^2).$$

Under the usual ARG, $\lim_{r\to 0}\mathfrak{D} = 5/11$. Thus, even in the presence of a *high* recombination rate, if large offspring-number events are frequent enough, one may see only evidence of *low* recombination rate in data. Further, the prediction $\mathfrak{D}$ can be substantially higher than Kingman-coalescent–based predictions if c is large and the recombination rate is not too small (Figure 6).

For particular examples of probability measures F from Equation 27 associated with the generator derived from our random offspring distribution model one can compute the quantities considered above in relation to fixed ψ. One such

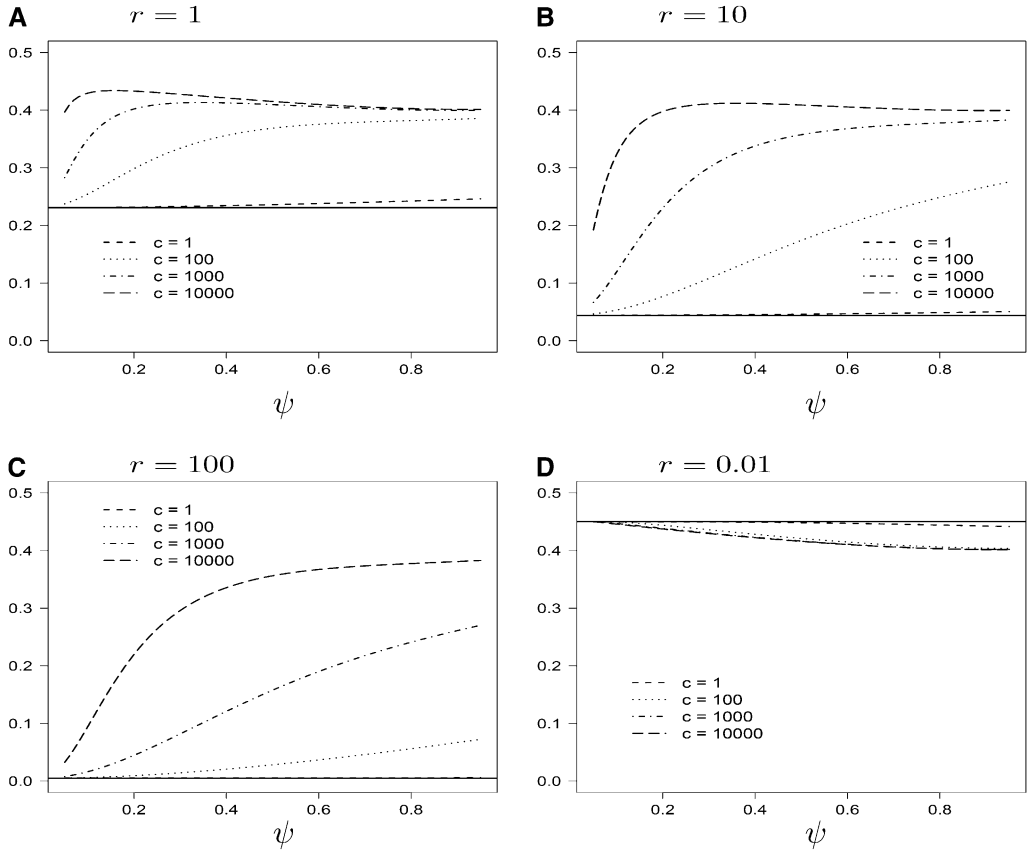

Figure 6 The estimate $\mathfrak{D}$ of the expected value $\mathbb{E}[r^2]$ as a function of ψ for different values of c and r. The solid lines represent the value of $\mathfrak{D}$ associated with the usual Wright–Fisher model.

example distribution can be the Beta(ϑ, γ) distribution (see Figure 7 for $\mathfrak{D}$). One obtains for $i \in \{0, 1, 2\}$,

$$\lim_{r\to\infty} h(i) = \frac{4\gamma(1+2\vartheta+\gamma)}{8\gamma(1+\gamma)+10\gamma\vartheta+7\vartheta(1+\vartheta)}.$$

Define $\tilde{h}(i) := \lim_{r\to\infty} h(i)$. For $i \in \{0, 1, 2\}$ one obtains

$$\lim_{r\to\infty} \mathbb{E}_i[T_s] = 4\tilde{h}(i) + \frac{4\gamma(1+2\vartheta+\gamma)}{8\gamma(1+\gamma)+10\gamma\vartheta+7\vartheta(1+\vartheta)},$$
$$\lim_{r\to\infty} \mathbb{E}_i[T_l] = \frac{3}{2} - \frac{1}{2}\tilde{h}(i) + \frac{3\gamma}{2(8\gamma(1+\gamma)+10\gamma\vartheta+7\vartheta(1+\vartheta))}. \quad (32)$$

The form of the relation shown in (32) between $h(i)$ and $\mathbb{E}_i[T_s]$ and $\mathbb{E}_i[T_l]$ resembles the one obtained for the Kingman-coalescent–based ARG, with the addition of a "correction" term due to simultaneous multiple mergers.

Variance of pairwise differences

The expected variance of pairwise differences was employed by Wakeley (1997) to estimate the recombination rate in low offspring-number (Wright–Fisher) populations, under the usual ancestral recombination graph. Let the random variable K_{ij} denote the number of differences between sequences i and j, with $K_{ii} = 0$. The average number π of pairwise differences for n sequences is

$$\pi = \frac{2}{n(n-1)} \sum_{i<j} K_{ij}.$$

The (empirical) variance S_π^2 of pairwise differences is defined as

$$S_\pi^2 = \frac{2}{n(n-1)} \sum_{i<j} (K_{ij} - \pi)^2.$$

In the *Appendix* we derive the expected variance of pairwise differences $\mathbb{E}[S_\pi^2]$ under the ancestral recombination graph described by the generator G (Equation 14) derived from our large offspring-number model. Under our model, $\mathbb{E}[S_\pi^2]$ is a function of the parameters c and ψ, in addition to being a function of r and θ (Figure 8 and Figure 9). In Figure 8, $\mathbb{E}[S_\pi^2]$, when only two loci are considered, is graphed as a function of the recombination rate and in Figure 9 as a function of sample size. Figures 8 and 9 show that $\mathbb{E}[S_\pi^2]$ is primarily influenced by the mutation rate (θ), when the values of c and ψ are fairly modest. However, $\mathbb{E}[S_\pi^2]$ can be quite low when both c and ψ are large, even when θ is also large (Figure 9). When c and ψ are both large, two sequences are more likely to coalesce before a mutation separates them.

The variance of pairwise differences alone will not suffice to yield estimates of r if both c and ψ are unknown. To jointly estimate the four parameters (c, ψ, r, θ) of our model one probably needs to employ computationally heavy likelihood

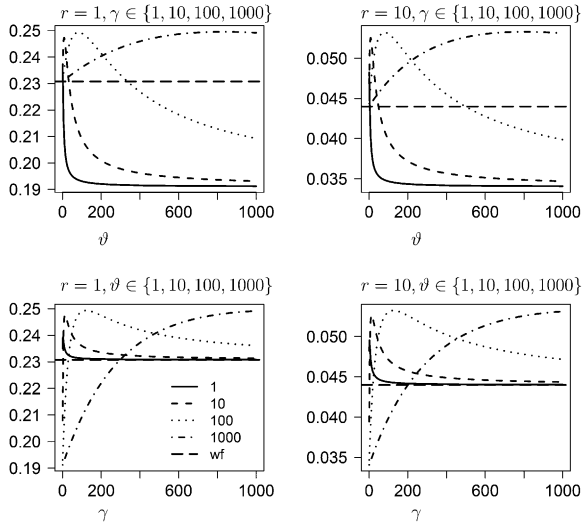

Figure 7 The prediction $\mathfrak{D}$ of linkage disequilibrium obtained from the ARG associated with the Beta(ϑ, γ) distribution. The different lines represent different values of γ (top panels) or ϑ (bottom panels). The broken horizontal line represents the prediction obtained from the usual ARG.

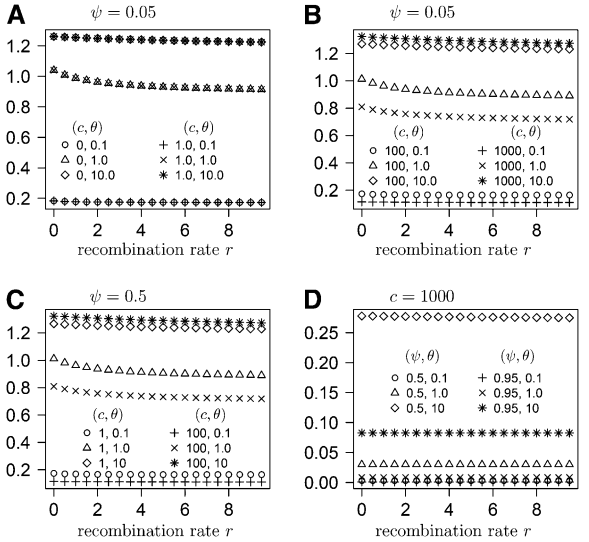

Figure 8 (A–D) The expected variance of pairwise differences for sample size 50 as a function of the recombination rate r for different values of the parameters c, ψ, and θ as shown.

and importance sampling methods in the spirit of Fearnhead and Donnelly (2001). However, given knowledge of c and ψ, one can, in principle, use the variance of pairwise differences to quickly obtain estimates of the recombination rate.

Correlations in ratios of coalescence times

The behavior of the correlations in ratios of coalescence times for sample sizes larger than two is investigated using Monte Carlo simulations.

Let L_i denote the total length of branches ancestral to i sequences at one locus, let L denote the total length of the genealogy at the same locus, and define $R_i := L_i/L$. Thus, R_1 is the total length of external branches to the total size of the genealogy. The idea behind estimating the expected value $\mathbb{E}[R_i]$ is as follows. Assuming the infinitely many sites mutation model, let S_i denote the total number of mutations in i copies and S denote the total number of segregating sites, and define $V_i := S_i/S$. The key idea behind deriving the coalescent was to separate the (neutral) mutation process from the genealogical process. The same principle also applies to predicting patterns of genetic variation using the coalescent: First, one constructs the genealogy and then superimposes mutations on the genealogy. The shape of the genealogy is thus a deciding factor in the genetic patterns one predicts. The relative lengths R_i of the different types of branches should therefore predict the relative number V_i of mutations of each class. This idea is exploited by Eldon (2011) to estimate coalescence parameters in the large offspring-number models introduced by Schweinsberg (2003) and Eldon and Wakeley (2006). Namely, the claim is

$$\lim_{n\to\infty} \mathbb{E}[R_i] = \lim_{n\to\infty} \mathbb{E}[V_i] = f(\varpi, i), \tag{33}$$

where n denotes the sample size, and ϖ denotes the coalescence (reproduction) parameters. Indeed, it follows from the results of Berestycki *et al.* (2007, 2008) that $(1 < \alpha < 2)$

$$\lim_{n\to\infty} \mathbb{E}[R_i] = \lim_{n\to\infty} \mathbb{E}[V_i] = \frac{\Gamma(i+\alpha-2)(\alpha-1)(2-\alpha)}{\Gamma(\alpha)i!}$$

when associated with the Beta($2 - \alpha$, α) coalescent derived by Schweinsberg (2003) from a population model in which the offspring law is stable with index α. A key feature of expression (33) is the absence of mutation rate in the function $f(\varpi, i)$; thus given a large number of DNA sequences

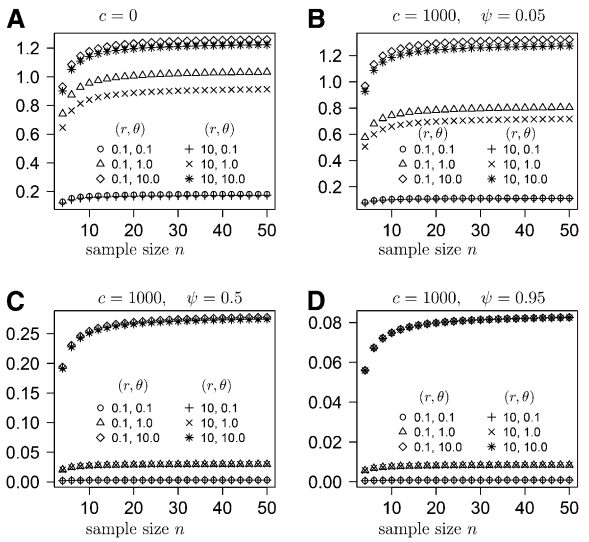

Figure 9 (A–D) The expected variance of pairwise differences as a function of sample size for different values of the parameters c, ψ, r, and θ as shown.

Table 1 Estimates $\bar{R}_i$ of the expected values $E[R_i]$ of the ratios $R_i := L_i/L$ for $1 \le i \le 4$ at one marginal locus, along with estimates $\hat{R}_i$ of the standard deviations of R_i

ψ	c	n	$\bar{R}_1$	$\bar{R}_2$	$\bar{R}_3$	$\bar{R}_4$	$\hat{R}_1$	$\hat{R}_2$	$\hat{R}_3$	$\hat{R}_4$
—	0	6	0.466	0.219	0.138	0.100	0.183	0.167	0.198	0.124
		10	0.378	0.180	0.117	0.085	0.156	0.132	0.120	0.110
		20	0.300	0.146	0.096	0.070	0.119	0.097	0.088	0.081
		50	0.235	0.116	0.077	0.057	0.080	0.063	0.058	0.055
0.005	1	6	0.466	0.219	0.138	0.100	0.183	0.167	0.198	0.124
		10	0.377	0.181	0.117	0.085	0.156	0.133	0.120	0.111
		20	0.299	0.146	0.095	0.071	0.118	0.097	0.088	0.082
		50	0.234	0.116	0.076	0.057	0.080	0.064	0.057	0.054
	1000	6	0.467	0.219	0.137	0.100	0.182	0.167	0.198	0.124
		10	0.377	0.181	0.117	0.085	0.156	0.133	0.120	0.110
		20	0.299	0.146	0.095	0.071	0.119	0.097	0.088	0.082
		50	0.235	0.116	0.077	0.057	0.080	0.064	0.058	0.054
0.5	1	6	0.468	0.217	0.138	0.099	0.184	0.166	0.199	0.124
		10	0.381	0.179	0.115	0.085	0.157	0.132	0.120	0.110
		20	0.304	0.145	0.095	0.070	0.120	0.097	0.088	0.081
		50	0.242	0.117	0.077	0.056	0.081	0.064	0.058	0.054
	1000	6	0.541	0.173	0.116	0.089	0.184	0.152	0.177	0.116
		10	0.566	0.117	0.078	0.058	0.159	0.101	0.090	0.082
		20	0.743	0.101	0.035	0.022	0.084	0.053	0.033	0.027
		50	0.576	0.195	0.089	0.046	0.058	0.051	0.037	0.026

Estimates are obtained from 10^5 simulated gene genealogies.

(possibly in the thousands), one hopes to be able to obtain estimates of the coalescence parameters ϖ without having to jointly estimate the mutation rate. In our model, there are four parameters to estimate, namely mutation and recombination rates, along with the coalescence parameters c and ψ. Even though full-likelihood methods exist (Birkner and Blath 2008; Birkner *et al.* 2011), applying them to large data sets consisting of thousands of sequences may represent a challenge.

Estimates of $\mathbb{E}[R_i]$ as functions of the sample size n and the coalescence parameters c and ψ are shown in Table 1. In nearly all cases the estimates of $\bar{R}_i$ decreased as sample size increased; the exception was $\bar{R}_1$ when $(c, \psi) = (1000, 0.5)$ (Table 1). When both c and ψ are large enough, we observe a nonmonotonic behavior in $\bar{R}_1$ as sample size increases (results not shown). The nonmonotonic behavior may be related to the property of the marginal haploid process (the point-mass part obtained as $c \to \infty$) of a single locus of not coming down from infinity (Schweinsberg 2000b); *i.e.*, when one starts with an infinite number of lineages (sample size), the number of lineages stays infinite. For such processes that do not come down from infinity, the ratio R_1 should go to one; *i.e.*, the gene genealogy should become completely star-shaped (see, *e.g.*, Eldon 2011). As both c and ψ increase, one expects the deviation from Kingman-coalescent–based predictions to increase. By way of example, for sample size 50 the vector $(\mathbb{E}[R_1], \ldots, \mathbb{E}[R_4])$ is estimated to be $\sim$(0.24, 0.12, 0.08, 0.06) when associated with the Kingman coalescent ($c = 0$), while being $\sim$(0.58, 0.20, 0.09, 0.05) when $(c, \psi) = (1000, 0.5)$. In all cases the estimate $\hat{R}_i$ of the standard deviation of R_i decreases as sample size increases, indicating convergence.

The rationale behind comparing the statistics in Tables 2 and 3 is as follows. As sequencing technologies advance, and the genomic sequences of more organisms become available, a case in point being the recently published genomic sequence of Atlantic cod (Star *et al.* 2011), genomic scans of thousands of individuals will become more common. Given DNA sequence data for many loci, one could calculate correlations for counts and ratios of counts of mutations and compare them to predictions based on different ancestral recombination graphs. Similarly for the single-locus statistics (Table 1), the idea is that the correlations of the coalescence time statistics (L_i and R_i) should reflect correlations of mutation counts (S_i). In particular, under the usual ARG one expects (see Tables 2 and 3)

$$\lim_{r\to\infty} \operatorname{cor}\left(L_i^{(1)}, L_j^{(2)}\right) = \lim_{r\to\infty} \operatorname{cor}\left(R_i^{(1)}, R_j^{(2)}\right) = 0,$$

where the superscript refers to locus numbers 1 and 2, respectively, while under an ARG admitting simultaneous multiple mergers one expects

$$\lim_{r\to\infty} \operatorname{cor}\left(L_i^{(1)}, L_j^{(2)}\right) = f(i, j, \varpi)$$
$$\lim_{r\to\infty} \operatorname{cor}\left(R_i^{(1)}, R_j^{(2)}\right) = g(i, j, \varpi),$$

where f and g are functions of the particular statistics indicated by i and j as well as the vector ϖ of coalescence (reproduction) parameters.

In general, the results reported in Tables 2 and 3 indicate that high values of *both* ψ and c are required for high correlations when recombination rate is high, when associated

Table 2 Estimates of the correlation cor($X^{(1)}$, $Y^{(2)}$) between $X^{(1)}$ and $Y^{(2)}$, where $X^{(1)}$ represents a statistic for locus 1 and $Y^{(2)}$ that for locus 2, as follows: the time T until the most recent common ancestor at a locus; L, the total length of the gene genealogy at a locus; and $R_i := L_i/L$, in which L_i denotes the total length of branches ancestral to i sequences

c	ψ	r	cor($T^{(1)}$, $T^{(2)}$)	cor($L^{(1)}$, $L^{(2)}$)	cor($L^{(1)}, L_1^{(2)}$)	cor($L^{(1)}, L_2^{(2)}$)	cor($L_3^{(1)}, L_3^{(2)}$)	cor($L_4^{(1)}, L_4^{(2)}$)
0	—	1	0.311	0.418	0.586	0.501	0.434	0.378
		10	0.016	0.058	0.169	0.089	0.047	0.036
1	0.005	1	0.306	0.415	0.588	0.508	0.431	0.380
		10	0.015	0.055	0.171	0.090	0.049	0.034
1000	0.005	1	0.308	0.419	0.585	0.509	0.438	0.376
		10	0.013	0.051	0.168	0.093	0.052	0.030
1	0.5	1	0.328	0.447	0.601	0.516	0.449	0.389
1		10	0.024	0.085	0.193	0.107	0.064	0.036
1000		1	0.982	0.995	0.976	0.950	0.918	0.879
		10	0.924	0.947	0.763	0.623	0.503	0.396

c	ψ	r	cor($L_1^{(1)}, L_2^{(2)}$)	cor($L_1^{(1)}, L_3^{(2)}$)	cor($L_1^{(1)}, L_4^{(2)}$)	cor($L_2^{(1)}, L_3^{(2)}$)	cor($L_2^{(1)}, L_4^{(2)}$)	cor($L_3^{(1)}, L_4^{(2)}$)
0	—	1	−0.031	−0.031	−0.021	−0.005	−0.018	0.009
		10	0.005	−0.006	−0.001	0.012	0.005	0.013
1	0.005	1	−0.035	−0.025	−0.021	−0.001	−0.019	0.009
		10	0.000	−0.002	0.008	0.009	0.005	0.014
1000	0.005	1	−0.036	−0.029	−0.021	−0.006	−0.018	0.010
		10	−0.002	−0.003	0.003	0.014	0.004	0.005
1	0.5	1	−0.022	−0.014	−0.007	0.004	−0.004	0.023
		10	0.009	0.006	0.010	0.022	0.014	0.025
1000		1	0.326	0.314	0.305	0.238	0.218	0.176
		10	0.311	0.284	0.266	0.289	0.239	0.262

Estimates are based on 10^5 simulated ancestral recombination graphs each for a sample of size 50.

with our model. In particular, the correlations between $R_i^{(1)}$ and $R_i^{(2)}$ (*i.e.*, between corresponding R_i's at different loci) can be quite high, even when recombination is high, when both c and ψ are large enough, another indicator of the genome-wide correlations induced by sweepstake-like reproduction.

A different question concerns the limit behavior as sample size n increases. Fix the recombination rate and consider the limits

$$\lim_{n\to\infty} \operatorname{cor}\left(R_i^{(1)}, R_j^{(2)}\right), \quad \lim_{n\to\infty} \operatorname{cor}\left(V_i^{(1)}, V_j^{(2)}\right) \tag{34}$$

Table 3 Estimates of the correlation cor($X^{(1)}$, $Y^{(2)}$) between $X^{(1)}$ and $Y^{(2)}$, where $X^{(1)}$ represents a statistic for locus 1 and $Y^{(2)}$ that for locus 2, as follows: the time T until most recent common ancestor at a locus; L, the total length of the gene genealogy at a locus; and $R_i := L_i/L$, in which L_i denotes the total length of branches ancestral to i sequences

c	Ψ	r	cor($R_1^{(1)}, R_2^{(2)}$)	cor($R_2^{(1)}, R_2^{(2)}$)	cor($R_3^{(1)}, R_3^{(2)}$)	cor($R_4^{(1)}, R_4^{(2)}$)
0	—	1	0.570	0.548	0.486	0.431
		10	0.116	0.089	0.052	0.042
1	0.005	1	0.566	0.552	0.487	0.435
		10	0.115	0.091	0.054	0.035
1000	0.005	1	0.570	0.551	0.491	0.434
		10	0.115	0.095	0.059	0.031
1	0.5	1	0.583	0.557	0.504	0.447
		10	0.135	0.102	0.063	0.038
1000	0.5	1	0.955	0.927	0.900	0.866
		10	0.679	0.469	0.384	0.304

c	ψ	r	cor($R_1^{(1)}, R_2^{(2)}$)	cor($R_1^{(1)}, R_3^{(2)}$)	cor($R_1^{(1)}, R_4^{(2)}$)	cor($R_2^{(1)}, R_3^{(2)}$)	cor($R_2^{(1)}, R_4^{(2)}$)	cor($R_3^{(1)}, R_4^{(2)}$)
0	—	1	−0.023	−0.040	−0.042	−0.026	−0.042	−0.014
		10	−0.022	−0.023	−0.020	0.003	−0.005	0.005
1	0.005	1	−0.024	−0.038	−0.042	−0.023	−0.046	−0.014
		10	−0.027	−0.018	−0.015	0.001	−0.007	0.011
1000	0.005	1	−0.028	−0.038	−0.038	−0.031	−0.043	−0.012
		10	−0.030	−0.024	−0.016	0.003	−0.008	−0.001
1	0.5	1	−0.023	−0.035	−0.035	−0.028	−0.034	−0.007
1	0.5	10	−0.029	−0.023	−0.015	0.004	0.000	0.016
1000		1	−0.622	−0.348	−0.112	−0.100	−0.038	−0.016
		10	−0.330	−0.255	−0.135	0.009	0.004	0.096

Estimates are based on 10^5 simulated ancestral recombination graphs each for a sample of size 50.

Under the usual ARG, one expects the limits in (34) to be only functions of the recombination rate (and i and j). If the ARG also admits simultaneous multiple mergers, one expects the limits in (34) also to be functions of ϖ. Considering unlinked loci, one would be interested in the limits

$$\lim_{r\to\infty}\lim_{n\to\infty}\operatorname{cor}\left(R_i^{(1)},R_j^{(2)}\right),\quad \lim_{r\to\infty}\lim_{n\to\infty}\operatorname{cor}\left(V_i^{(1)},V_j^{(2)}\right). \tag{35}$$

Resolving the limits (35) for different ARGs promises not only to yield insights into genome-wide correlations, but also to provide tools for inference, *e.g.*, to distinguish between different population models.

The C program written to perform the simulations was checked by comparing correlation in coalescence times for sample size two at two loci to analytical results. The program is available upon request.

Comparison with Eldon and Wakeley (2008)

Eldon and Wakeley (2008) consider correlations in coalescence times and the prediction $\mathfrak{D}$ of linkage disequilibrium, under a modified Wright–Fisher sweepstake-style reproduction model, and observe correlations in coalescence times between loci despite high recombination rate. Our work differs from theirs in important ways. To begin with, we treat diploidy in detail, in which each offspring receives its two chromosomes from two distinct diploid parents. This leads to a separation of timescales of the ancestral process. We formally derive an ancestral recombination graph that admits simultaneous multiple mergers of ancestral lineages, which naturally arise in diploid models. Eldon and Wakeley observed correlations in coalescence times when considering only sample size two at each locus in a model that contains diploid individuals only implicitly; it is not *a priori* obvious that the correlations would still hold for large sample sizes. We confirm this using our formally obtained ARG that allows us also to investigate correlations in coalescence times and in ratios of coalescence times, for sample sizes larger than two at each locus. In addition, one can apply our ARG to inference problems. Indeed, we show how the variance of pairwise differences can, in principle, be used to obtain estimates of the recombination rate. Finally, we obtain a large class of ARGs by randomizing the offspring distribution; thus one is not restricted to the simple case of fixed ψ.

Furthermore, since the estimate $\mathfrak{D}$ of the expected value of r^2 can be expressed in terms of correlations in coalescence times, Eldon and Wakeley consider $\mathfrak{D}$ under their modified Wright–Fisher model. However, $\mathfrak{D}$ is based on approximating an expected value of a ratio of correlated random variables by the ratio of expected values of the corresponding random variables and is also derived for a sample of size two at two loci. Thus, $\mathfrak{D}$ may not be the ideal quantity to quantify association between loci for large sample sizes. A more natural way may be to investigate correlations in coalescence times for samples larger than two the way we do.

Discussion

Understanding the genome-wide effects of sweepstake-like reproduction on gene genealogies was our main aim. To this end, we derived ancestral recombination graphs for many loci arising from population models admitting large offspring numbers. High variance in individual reproductive success, or sweepstake-style reproduction, has been suggested to explain the low genetic diversity observed in many marine populations (Hedgecock *et al.* 1982; Avise *et al.* 1988; Palumbi and Wilson 1990; Beckenbach 1994; Hedgecock 1994; Árnason 2004). Hedgecock and Pudovkin (2011) review the sweepstake-style reproduction hypothesis and conclude that it provides the correct framework in which to investigate many natural marine populations.

Multiple-merger (Donnelly and Kurtz 1999; Pitman 1999; Sagitov 1999) and simultaneous (Schweinsberg 2000a; Möhle and Sagitov 2001) multiple-merger coalescent models arise from population models incorporating sweepstakes reproduction by admitting large offspring numbers (Sagitov 2003; Eldon and Wakeley 2006; Sargsyan and Wakeley 2008). While multiple-merger coalescent processes describing the ancestral relations of alleles at a single locus have received the most attention from mathematicians, ancestral processes for multiple linked loci have hitherto remained unexplored. We derive an ancestral recombination graph for many loci from a diploid biparental population model, in which one pair of diploid individuals (parents) contributes offspring to the population at each time step. Thus, each offspring necessarily receives its chromosomes from distinct individuals, as diploid individuals tend to do. Incorporating diploidy into our model the way we do leads to a separation-of-timescales problem. Our limiting object is essentially a "haploid" process, in which chromosomes either coalesce or recombine. By extending a result of Möhle (1998), we show that diploidy, a fundamental characteristic of many natural populations, can thus be treated as a "black box," since the limiting object does not depend on the location of chromosomes in individuals.

By adopting a Moran-type model, in which only a single pair of individuals gives rise to offspring at each reproduction event, we chose mathematical tractability over more biologically realistic scenarios, in which, for example, many individuals contribute offspring at each time step. It should be straightforward to extend our model in many ways, for example by allowing a random number of parents or introducing population structure. Indeed, we do extend our model in one way, by taking a random offspring distribution. These extensions still leave open the question of distinguishing among different large offspring-number models. Our work on ancestral recombination graphs incorporating information from many loci is a step in this direction.

Sweepstake-style reproduction induces correlation in coalescence times even between loci separated by a high rate of

recombination. The correlation follows from the multiple-merger property of our ancestral recombination graph, since many chromosomes coalesce at the same time in a multiple-merger event. The correlation remains a function of the coalescence parameters (c and ψ) of our population model. An immediate question is the effects on predictions of linkage disequilibrium (LD). The approximation $\mathfrak{D}$ by McVean (2002) predicts low LD when the recombination rate is high. However, when the rate of large reproduction events is high ($c \to \infty$), $\mathfrak{D}$ remains a function of the coalescence parameters. The dependence of $\mathfrak{D}$ on coalescence parameters has implications for the use of LD in inference for populations exhibiting sweepstake-style reproduction. Using simulations, Davies *et al.* (2007) found little effect of multiple mergers on the prediction r^2 of linkage disequilibrium, when comparing the exact Wright–Fisher model with recombination to the usual (continuous-time) ARG. However, by directly incorporating large offspring-number events the way we do, we can show that large offspring-number events do induce correlation in coalescence times and hence influence predictions of linkage disequilibrium.

The genome-wide correlation in coalescence times (Tables 2 and 3) induced by sweepstake-style reproduction offers hints about how to distinguish between large offspring number and ordinary Wright–Fisher reproduction. We are unaware of any published multilocus methods derived to distinguish among different population models. Full-likelihood methods may be preferable to the simple moment-based methods we consider. However, likelihood-based inference tends to be computationally intensive and more so for large samples. For large samples, one should be able to quickly obtain a good idea of the underlying processes by comparing correlations in ratios of mutation counts with predictions based on different population models.

In conclusion, ancestral recombination graphs admitting simultaneous multiple mergers of ancestral lineages are derived from a diploid population model of sweepstake-style reproduction, suggested to be common in many diverse marine populations. Our calculations show that sweepstake-style reproduction results in genome-wide correlation of gene genealogies, even for large sample sizes. Estimates of linkage disequilibrium and of recombination rates are confounded by the coalescence parameters of our population model. The genome-wide correlation in gene genealogies induced by sweepstake-style reproduction implies that examining correlations between loci should provide a means of distinguishing between ordinary Wright–Fisher and sweepstake-style reproduction.

Acknowledgments

We acknowledge the comments of two anonymous referees, which helped to improve the presentation; one referee also spotted an error in our original proof of Theorem 1.3. J.B. and B.E. thank Institut für Mathematik, Johannes-Gutenberg-Universität Mainz, for hospitality. M.B. thanks Mathematisch Instituut, Universiteit Leiden, for hospitality. B.E. was supported in part by Engineering and Physical Sciences Research Council grant EP/G052026/1 and by a Junior Research Fellowship at Lady Margaret Hall, Oxford University. J.B. and B.E. were supported in part by Deutsche Forschungsgemeinschaft (DFG) grant BL 1105/3-1. M.B. was in part supported by DFG grant BI 1058/2-1 and through European Research Council advanced grant 267356 VARIS (Variational Approach to Random Interacting Systems).

Literature Cited

Árnason, E., 2004 Mitochondrial cytochrome *b* variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. Genetics 166: 1871–1885.

Avise, J. C., R. M. Ball, and J. Arnold, 1988 Current versus historical population sizes in vertebrate species with high gene flow: a comparison based on mitochondrial DNA lineages and inbreeding theory for neutral mutations. Mol. Biol. Evol. 5: 331–344.

Beckenbach, A. T., 1994 Mitochondrial haplotype frequencies in oysters: neutral alternatives to selection models, pp. 188–198 in *Non-Neutral Evolution*, edited by B. Golding. Chapman & Hall, New York.

Berestycki, J., N. Berestycki, and J. Schweinsberg, 2007 Beta-coalescents and continuous stable random trees. Ann. Probab. 35: 1835–1887.

Berestycki, J., N. Berestycki, and J. Schweinsberg, 2008 Small-time behavior of beta coalescents. Ann. Inst. H. Poincaré. Probab. Stat. 44:214–238.

Birkner, M., and J. Blath, 2008 Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. J. Math. Biol. 57: 435–465.

Birkner, M., and J. Blath, 2009 Measure-valued diffusions, general coalescents and population genetic inference, pp. 329–363 in *Trends in Stochastic Analysis*, edited by J. Blath, P. Mörters, and M. Scheutzow. Cambridge University Press, Cambridge/London/New York.

Birkner, M., J. Blath, M. Möhle, M. Steinrücken, and J. Tams, 2009 A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. ALEA Lat. Am. J. Probab. Math. Stat. 6: 25–61.

Birkner, M., J. Blath, and M. Steinrücken, 2011 Importance sampling for Lambda-coalescents in the infinitely many sites model. Theor. Popul. Biol. 79: 155–173.

Cannings, C., 1974 The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. Adv. Appl. Probab. 6: 260–290.

Davies, J. L., F. Simančík, R. Lyngsø, T. Mailund, and J. Hein, 2007 On recombination-induced multiple and simultaneous coalescent events. Genetics 177: 2151–2160.

Donnelly, P., and T. G. Kurtz, 1999 Particle representations for measure-valued population models. Ann. Probab. 27: 166–205.

Durrett, R., 2002 *Probability Models for DNA Sequence Evolution*. Springer-Verlag, New York.

Eldon, B., 2011 Estimation of parameters in large offspring number models and ratios of coalescence times. Theor. Popul. Biol. 80: 16–28.

Eldon, B., and J. Wakeley, 2006 Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172: 2621–2633.

Eldon, B., and J. Wakeley, 2008 Linkage disequilibrium under skewed offspring distribution among individuals in a population. Genetics 178: 1517–1532.

Etheridge, A. M., R. C. Griffiths, and J. E. Taylor, 2010 A coalescent dual process in a Moran model with genic selection, and the Lambda coalescent limit. Theor. Popul. Biol. 78: 77–92.

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

Griffiths, R. C., 1991 The two-locus ancestral graph, pp. 100–117 in *Selected Proceedings of the Symposium on Applied Probability*, edited by I. V. Basawa and R. L. Taylor. Institute of Mathematical Statistics, Hayward, CA.

Griffiths, R. C., and P. Marjoram, 1997 An ancestral recombination graph, pp. 257–270 in *Progress in Population Genetics and Human Evolution* (IMA Volumes in Mathematics and Its Applications 87), edited by P. Donnelly, and S. Tavaré. Springer-Verlag, New York.

Hartl, D. L., and A. G. Clark, 1989 *Principles of Population Genetics*, Ed. 2. Sinauer Associates, Sunderland, MA.

Hedgecock, D., 1994 Does variance in reproductive success limit effective population sizes of marine organisms? pp. 1222–1344 in *Genetics and Evolution of Aquatic Organisms*, edited by A. Beaumont. Chapman & Hall, London.

Hedgecock, D., and A. I. Pudovkin, 2011 Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. Bull. Mar. Sci. 87: 971–1002.

Hedgecock, D., M. Tracey, and K. Nelson, 1982 Genetics, pp. 297–403 in *The Biology of Crustacea*, Vol. 2, edited by L. G. Abele. Academic Press, New York.

Herbots, H. M., 1997 The structured coalescent, pp. 231–255 in *Progress of Population Genetics and Human Evolution*, edited by P. Donnelly and S. Tavaré. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38: 226–231.

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

Hudson, R. R., 1983b Testing the constant-rate neutral allele model with protein sequence data. Evolution 37: 203–217.

Huillet, T., and M. Möhle, 2011 On the extended Moran model and its relation to coalescents with multiple collisions. Theor. Popul. Biol. (in press).

Hutchings, J. A., T. D. Bishop, and C. R. McGregor-Shaw, 1999 Spawning behaviour of Atlantic cod, *Gadus morhua*: evidence of mate competition and mate choice in a broadcast spawning. Can. J. Fish. Aquat. Sci. 56: 97–104.

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

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

Krone, S. M., and C. Neuhauser, 1997 Ancestral processes with selection. Theor. Popul. Biol. 51: 210–237.

McVean, G. A., 2002 A genealogical interpretation of linkage disequilibrium. Genetics 162: 987–991.

Möhle, M., 1998 A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. Adv. Appl. Probab. 30: 493–512.

Möhle, M., and S. Sagitov, 2001 A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. 29: 1547–1562.

Möhle, M., and S. Sagitov, 2003 Coalescent patterns in diploid exchangeable population models. J. Math. Biol. 47: 337–352.

Neuhauser, C., and S. M. Krone, 1997 The genealogy of samples in models with selection. Genetics 145: 519–534.

Nordeide, J. T., and I. Folstad, 2000 Is cod lekking or a promiscuous group spawner? Fish Fish. 1: 90–93.

Notohara, M., 1990 The coalescent and the genealogical process in geographically structured population. J. Math. Biol. 29: 59–75.

Ohta, T., and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics 68: 571–580.

Palumbi, S. R., and A. C. Wilson, 1990 Mitochondrial DNA diversity in the sea-urchins *Strongylocentrotus purpuratus* and *Strongylocentrotus droebachiensis*. Evolution 44: 403–415.

Pitman, J., 1999 Coalescents with multiple collisions. Ann. Probab. 27: 1870–1902.

Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Probab. 36: 1116–1125.

Sagitov, S., 2003 Convergence to the coalescent with simultaneous mergers. J. Appl. Probab. 40: 839–854.

Sargsyan, O., and J. Wakeley, 2008 A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theor. Popul. Biol. 74: 104–114.

Schweinsberg, J., 2000a Coalescents with simultaneous multiple collisions. Electron. J. Probab. 5: 1–50.

Schweinsberg, J., 2000b A necessary and sufficient condition for the λ-coalescent to come down from infinity. Electron. Comm. Probab. 5: 1–11.

Schweinsberg, J., 2003 Coalescent processes obtained from supercritical Galton-Watson processes. Stoch. Proc. Appl. 106: 107–139.

Star, B., A. J. Nederbragt, and S. Jentoft, U. Grimholt, M. Malstrøm *et al.*, 2011 The genomic sequence of Atlantic cod reveals a unique immune system. Nature 477: 207–210.

Steinrücken, M., M. Birkner, and J. Blath, 2013 Analysis of DNA sequence variation within marine species using Beta-coalescents. Theor. Popul. Biol. 83: 20–29.

Tajima, F., 1983 Evolutionary relationships of DNA sequences in finite populations. Genetics 105: 437–460.

Taylor, J., and A. Véber, 2009 Coalescent processes in subdivided populations subject to recurrent mass extinctions. Electron. J. Probab. 14: 242–288.

Taylor, J. E., 2009 The genealogical consequences of fecundity variance polymorphism. Genetics 182: 813–837.

Wakeley, J., 1997 Using the variance of pairwise differences to estimate the recombination rate. Genet. Res. 69: 45–48.

Communicating editor: Y. S. Song

Appendix

A1: Overview of Transitions and Their Probabilities in the Finite Population Model

A1.1: Basic setup and notation

We now classify all transitions and their probabilities of our population model relevant for the ancestral process under the scaling $\varepsilon N^2 = c/N^2$, in which N denotes the population size. Fix a sample size n for this section. Usually we suppress the dependence on the sample size in the notation below. Recall the state space $\mathscr{A}_n$ of our ancestral process (respectively $\mathscr{A}_n^{\mathrm{sm}}$ for the "effective" limiting model).

Let Π_N be the transition matrix of the Markov chain $\{\xi^{n,\,N}(m)\}_{m=0,\,1,\ldots}$ on $\mathscr{A}_n$ describing the ancestral states of an n sample in a population of size N. Our aim is to decompose Π_N into

$$\Pi_N = A_N + \frac{1}{N^2} B_N + R_N, \tag{A1}$$

where the matrix A_N contains all transitions whose probability is $O(1)$ or $O(N^{-1})$ per generation, so that they will happen "instantaneously" in the limit and are either identity transitions or projections from $\mathscr{A}_n$ to $\mathscr{A}_n^{\mathrm{sm}}$ by means of dispersing chromosomes paired in double-marked individuals. The matrix B_N contains all transition probabilities that are positive and finite after multiplication with N^2 and $N \to \infty$, that is, our "effective transitions." The remainder matrix R_N carries only transition probabilities that are of order $O(N^{-3})$ or smaller that will thus vanish after scaling.

Once we have established this decomposition, we can apply Lemma 1.7 below in a suitable way to identify the limit given in Definition 1.1 and establish the convergence result, *i.e.*, Theorem 1.2.

In Table A1, Table A2, and Table A3 we schematically deal with all possible transitions that can happen to a current sample over one time step.

Analogous to the notation and convention of Möhle and Sagitov (2003), we assume that in every configuration $\xi^{n,N}(m)$ from (2), for the order of chromosomes in individuals $\mathbb{I}_i$ for $i \in [b(m)]$ we have

$$\begin{aligned} \mathbb{I}_i(m) &= \left\{C^{(2i-1)}(m), C^{(2i)}(m)\right\} \quad \text{if} \quad 1 \le i \le \beta(m) - b(m); \\ \mathbb{I}_i(m) &= \left\{C^{(\beta(m)-b(m)+i)}(m), \emptyset\right\} \quad \text{if} \ \ \beta(m) - b(m) + 1 \le i \le b(m). \end{aligned} \tag{A2}$$

For ease of presentation, we denote by $\mathbb{I}'$ a single-marked individual carrying one active chromosome, by $\mathbb{I}''$ a double-marked individual carrying two active chromosomes, by $\widetilde{\mathbb{I}}'$ a single-marked individual (parent) whose marked chromosome is not passed on in the sample during a given reproduction event, and by $\hat{\mathbb{I}}''$ a double-marked individual (parent) where one marked chromosome is passed on and the other is not during a given reproduction event.

The symbols (A), (B), and (R) in the tables denote whether the corresponding transitions belong to A_N (A), to B_N (B), or to the "remainder term" (R) in (A1) according to the decomposition mentioned above. After that, we compute all the important probabilities explicitly. The order of the probability of each transition is also noted in Tables A1–A3.

A1.2: Transition type 1: Small or large reproduction event, no offspring in the sample

If a reproduction event takes place, say at generation m, which does not affect our sample, this will not affect the state of our ancestral process at $m + 1$, and we have $\xi^{n,N}(m) = \xi^{n,N}(m + 1)$. Hence, we see an identity transformation. We now compute the probability that our sample is not affected. Given the current state $\xi \in \mathscr{A}_n$ with b individuals and β chromosomes (hence $\beta - b$ double-marked and $2b - \beta$ single-marked individuals), the probability that no child is in the sample is

Table A1 Transitions of type 2

	Parent with marked chromosome(s) (Ø means no parent in sample)	
Offspring	Ø	$\mathbb{I}'$
$\mathbb{I}''$	$\{\mathbb{I}', \mathbb{I}'\}(A)$ $O(N^{-1})^a$	$\{\mathbb{I}', \mathbb{I}'\}$, $\{\mathbb{I}'', \mathbb{I}'\}$ $O(N^{-2})$, (B)
$\mathbb{I}'$	$\{\mathbb{I}'\}(A)$ $O(N^{-1})^b$ $\widetilde{\mathbb{I}}'$	$\{\mathbb{I}'\}$, $\{\mathbb{I}''\}$, $\{\mathbb{I}', \mathbb{I}'\}$, (B) $O(N^{-2})^c$ $\hat{\mathbb{I}}''$
$\mathbb{I}''$	$O(N^{-2})$, (B)	$O(N^{-2})$, $(B)^d$
$\mathbb{I}'$	$\{\mathbb{I}''\}$, $\{\mathbb{I}', \mathbb{I}'\}$, (B) $O(N^{-2})$	$\{\mathbb{I}', \mathbb{I}''\}$, $\{\mathbb{I}'\}$, (B) $O(N^{-2})$

[a] Offspring double-marked, no parent in sample.
[b] Offspring single marked, no parent in sample.
[c] Offspring single marked, one single-marked parent in sample.
[d] Offspring double-marked, one double-marked parent in sample.

Table A2 Transitions of type 4, neither parent in sample

Offspring	Parent: Ø
$\mathbb{I}''$	$\{\mathbb{I}'', \mathbb{I}'\}$, $O(N^{-2})$, (B)
	$\{\mathbb{I}'', \mathbb{I}''\}$, $O(N^{-3})$, (R)
$\mathbb{I}'$	$\mathbb{I}''$, $O(N^{-2})$, (B)

$$(1-\varepsilon)\frac{N-b}{N}+\varepsilon_N\frac{\binom{N-b}{\lfloor \psi N \rfloor}}{\binom{N}{\lfloor \psi N \rfloor}}=1-O\left(N^{-1}\right).$$

A1.3: Transition type 2: Small reproduction event, offspring in the sample, at most one parent in the sample, no recombination

Here, we need to distinguish only whether the offspring is single or double marked and whether there is a parent in the sample. For example, it is immediate to see that the probability of a transition from a double-marked ($\mathbb{I}''$) offspring to two single-marked ($\{\mathbb{I}',\mathbb{I}'\}$) individuals is of order $O(N^{-1})$ when no parent is in the sample and no recombination happens. Table A1 lists all corresponding events. By way of example, the state-labeled $\{\mathbb{I}', \mathbb{I}'\}$ denotes that two single-marked individuals, each carrying one active chromosome, are reached from the sample configuration. One such configuration is if the sample contains one offspring, but neither parent (Ø), and the offspring is carrying two active chromosomes ($\mathbb{I}''$).

A1.4: Transition type 3: Small reproduction event, offspring in the sample, both parents in the sample

If both parents and offspring are in the sample in a small event, this immediately gives a transition probability of order $O(N^{-3})$ or smaller (depending on the presence of recombination) and hence will be irrelevant and be part of R_N. We omit a detailed table listing the different single- and double-marked individuals.

A1.5: Transition type 4: Small reproduction event, offspring and at most one parent in the sample, recombination occurs

Table A2 lists transitions due to recombination and when neither parent is in the sample. The probability of the presence of both an offspring and at least one parent in a sample, when recombination occurs, is of order $O(N^{-3})$ and so will vanish in the limit.

A1.6: Transition type 5: Large reproduction event, offspring in the sample, no parent in the sample, no recombination

Table A3 lists all possible transitions when a large reproduction event occurs, no parent is in the sample, and recombination does not occur. The probabilities of the events listed in Table 1 in the main text are of order $O(N^{-2})$ and so will appear as effective transitions in the limit.

A1.7: Transition type 6: Large reproduction event, offspring in the sample, recombination occurs, and/or at least one parent is in the sample

The probability that a large reproduction event takes place and at least one child and at least one parent are in the sample is $O(N^{-3})$. In addition, the probability that a large reproduction event takes place, at least one child is in the sample, and also a recombination event happens in the sample is $O(N^{-3})$. Hence all such events are negligible.

A2: The Convergence Result

A2.1: The limit of the projection matrix A_N

Some care is needed to make sure $\mathbf{A}_N$ converges in the right sense to the desired projection matrix. The only relevant transitions of order $O(1)$ or $O(N^{-1})$ are transitions of types 1 and 2. The only one that is not an identity

Table A3 Transitions of type 5

Offspring	Parent: Ø
$k_1\mathbb{I}', k_2\mathbb{I}''$	$\{\mathbb{I}'', \mathbb{I}''\}$, $O(N^{-2})$, (B)
	$\{\mathbb{I}'', \mathbb{I}'\}$, $O(N^{-2})$, (B)
	$\{\mathbb{I}', \mathbb{I}'\}$, $O(N^{-2})$, (B)
	$\mathbb{I}''$, $O(N^{-2})$, (B)
	$\mathbb{I}'$, $O(N^{-2})$, (B)

transition is the first dispersion event of Table A1. For $\xi \in \mathscr{A}_n$ with $b < \beta$ (*i.e.*, at least one marked individual is double marked), that is

$$\xi \mapsto \mathrm{disp}_i(\xi).$$

This event will become part of A_N and has probability

$$A_N(\xi, \mathrm{disp}_i(\xi)) = (1-\varepsilon_N)\frac{1}{N}\frac{\binom{N-b-1}{2}}{\binom{N}{2}}(1-r_N)^2, \quad 1 \le i \le \beta - b \tag{A3}$$

(this is the probability of the event *a* listed in Table A1; event *b* in Table A1 leads to an identity transition). Otherwise, we have

$$A_N(\xi, \xi) = 1 - (1-\varepsilon_N)\frac{\beta - b}{N}\frac{\binom{N-b-1}{2}}{\binom{N}{2}}(1-r_N)^2.$$

Of course, A_N has to leave elements of the subspace $\mathscr{A}_n^{\mathrm{sm}}$ invariant; hence we set, for ξ with $b = \beta$,

$$A_N\left(\xi, \xi'\right) := 1_{\{\xi = \xi'\}}.$$

Proposition 1.4. *With the above settings, A_N is a stochastic matrix for each N and*

$$\lim_{C\to\infty}\lim_{N\to\infty}\sup_{r \ge CN}\left\|A_N^r - P\right\| = 0 \tag{A4}$$

for all $C > 0$ large enough, where P is the canonical projection from $\mathscr{A}_n$ to $\mathscr{A}_n^{\mathrm{sm}}$; i.e.,

$$P\left(\xi, \xi'\right) = 1_{\{\xi' = \mathrm{cd}(\xi)\}}.$$

Proof of Proposition 1.4. The Markov chain with transition matrix A_N can change state only by dispersing the chromosomes paired in a double-marked individual. We see from (A3) that

$$A_N(\xi, \mathrm{disp}_i(\xi)) \ge \frac{K(n,r,c)}{N}$$

for some suitable constant $K(n, r, c)$, uniformly in b and $i \le \beta - b$ and N (for all N large enough). Hence, starting from ξ with $\beta - b$ double-marked individuals, the number of A_N steps required until complete dispersion has occurred is dominated by the sum of $\beta - b$ independent geometric random variables $\gamma_1^{(N)} + \ldots + \gamma_{\beta-b}^{(N)}$, with success probability $K(n, r, c)/N$. By Markov's inequality,

$$\sup_{N\in\mathbb{N}}\mathbb{P}\left\{\gamma_1^{(N)} + \ldots + \gamma_{\beta-b}^{(N)} \ge CN\right\} \le \frac{1}{CN}\mathbb{E}\left[\gamma_1^{(N)} + \ldots + \gamma_{\beta-b}^{(N)}\right] = \frac{N(\beta-b)}{C\cdot N\cdot K(n,r,c)} \to 0 \text{ as } C\to\infty.$$

The proof can now be completed with a coupling argument, noting that two Markov chains run according to A_N resp. P, started in $\xi \in \mathscr{A}_n$, both get stuck in $\mathrm{cd}(\xi)$, and this happens after at most CN steps with high probability (for C large).

A2.2: Proof of the convergence result

With the definition of A_N from the previous section, put

$$B_N^* := N^2(\Pi_N - A_N) \tag{A5}$$

and let P be the canonical projection from $\mathscr{A}_n$ to $\mathscr{A}_n^{\mathrm{sm}}$ defined in Proposition 1.4. The following lemma identifies G as the limit containing all the effective transitions of B_N^* when projecting on the subspace $\mathscr{A}_n^{\mathrm{sm}}$:

Lemma 1.5. *We have*

$$\hat{B}_N := PB_N^* P \to G \quad \text{as} \quad N \to \infty \tag{A6}$$

with G from (14).

Remark 1.6. *We do believe that in fact the sequence of* (*formally larger*) *matrices B_N^* on $\mathscr{A}_n$ converges as well, but the statement about $\hat{B}_N$ is sufficient for our purposes below* [*see* (A13) *in Lemma* 1.7] *and simpler to prove since it allows us to restrict to the "completely dispersed" configurations in $\mathscr{A}_n^{\mathrm{sm}}$.*

Proof of Lemma 1.5. We inspect the types of events listed in Tables A1–A3 that are marked with (*B*). Events that are marked with (*R*) have probability of order at most $O(N^{-3})$; hence their total contribution to any entry of $\hat{B}_N$ is at most $O(N^{-1})$ (since we are following a finite sample, there are only finitely many possible one-step events altogether). It suffices to consider $\hat{B}_N(\xi, \mathrm{cd}(\eta))$ for $\xi = \{C^{(1)}, \dots, C^{(\beta)}; \beta\} \in \mathscr{A}_n^{\mathrm{sm}}, \eta \in \mathscr{A}_n$ (because P projects to $\mathscr{A}_n^{\mathrm{sm}}$).

Regarding $\xi' = \mathrm{pairmerge}_{j_1 j_2}(\xi)$, this transition can happen in a small reproduction event (these events are listed at *c* in Table A1; note that events listed at *d* in Table A1 lead to a trivial transition once *P* is applied) or in a large reproduction event as in Table A3 if the grouping is suitable. Up to four parental chromosomes are involved in any reproduction event. Hence, a large reproduction event can lead to a given pair merger in the sample if up to five individuals in the sample are children. Thus

$$\begin{aligned}\hat{B}_N(\xi, \xi') &= N^2(1-\varepsilon_N)(1-r_N)2 \times \frac{1}{N}\frac{1 \cdot (N-b)}{\binom{N-1}{2}}\frac{1}{2}\frac{1}{2} \\ &\quad + N^2 \varepsilon_N \sum_{c=2}^{5} (1-r_N)^c \binom{\beta-2}{c-2} \frac{\binom{N-\beta}{\lfloor N\psi \rfloor - c}}{\binom{N}{\lfloor N\psi \rfloor}} (4)_{c-1} \left(\frac{1}{4}\right)^c + O(N^{-1}).\end{aligned} \tag{A7}$$

For the first term on the right note that either j_1 or j_2 can be the child, and the two factors of $\frac{1}{2}$ come from the requirement that the chromosome in the child we are following is the one from the parent in the sample and is also the one we are following in the parent. For the second term on the right note that once we decide on *c* children in the sample [$((\beta-2)/(c-2))$ choices because j_1 and j_2 are already chosen], there are $(4)_{c-1}$ ways to assign them to the four parental chromosomes. For comparison with (15) and the first line in (14) observe

$$\frac{\binom{N-\beta}{\lfloor N\psi \rfloor - c}}{\binom{N}{\lfloor N\psi \rfloor}} = \frac{(N-\beta)!\lfloor N\psi \rfloor!(N-\lfloor N\psi \rfloor)!}{(\lfloor N\psi \rfloor - c)!(N-\beta-\lfloor N\psi \rfloor + c)!N!} \sim \frac{(N\psi)^c (N(1-\psi))^{\beta-c}}{N^\beta} = \psi^c (1-\psi)^{\beta-c}.$$

Regarding $\xi' = \mathrm{recomb}_{j,\ell}(\xi)$ [assuming that α is such that $C^{(j)}$ can be nontrivially cut into two by a recombination event between loci $\ell - 1$ and ℓ], this transition can happen in a small reproduction event as listed at *b* in Table A1 or in another event that has probability $O(N^{-3})$. Hence

$$\hat{B}_N\left(\xi, \xi'\right) = N^2(1-\varepsilon_N) \times \frac{1}{N}\frac{\binom{N-b}{2}}{\binom{N-1}{2}}\frac{r^{(\ell)}}{N} + O\left(N^{-1}\right) = r^{(\ell)} + O\left(N^{-1}\right). \tag{A8}$$

Regarding $\xi' = \mathrm{groupmerge}_{J_1, J_2, J_3, J_4}(\xi)$, this can occur only through a large reproduction event as listed in section A1.6. Write $k_i := |J_i|$; we assume $k_1 \geq \ldots \geq k_a \geq 2$ for some $a \in [4]$, $k_{a+1} = \ldots = k_4 = 0$ (if $a = 1$, $k_1 \geq 3$), and $s := \beta - (k_1 + \cdots + k_a)$ is the number of singletons (nonparticipating chromosomes) in the merger. Note that by the structure of the diploid model, with *a* groups merging there can be up to $k_1 + \cdots + k_a + (4-a)^+$ children in the sample [put differently, up to $(4-a)^+$ "nonmerging children"]. Then

$$\hat{B}_N(\xi,\xi') = N^2\varepsilon_N \sum_{c'=0}^{(4-a)^+} \binom{\beta - k_1 - \ldots - k_a}{c'} (1-r_N)^{k_1+\ldots+k_a+c'}$$

$$\times \frac{\binom{N-\beta}{\lfloor N\psi\rfloor - (k_1 + \ldots + k_a + c')}}{\binom{N}{\lfloor N\psi\rfloor}} (4)_{a+c'} \left(\frac{1}{4}\right)^{k_1+\ldots+k_a+c'} + O(N^{-1}).$$

It remains to check that the diagonal terms behave correctly, *i.e.*, that as $N \to \infty$,

$$\hat{B}_N(\xi,\xi) \to G(\xi,\xi) = -\sum_{\xi'\neq\xi,\xi'\in\mathscr{X}_n^{\mathrm{sm}}} G(\xi,\xi'). \tag{A9}$$

Because Π_N and A_N are both stochastic matrices (as is P), we have

$$\hat{B}_N(\xi,\xi) = -\sum_{\xi'\neq\xi,\xi'\in\mathscr{X}_n^{\mathrm{sm}}} \hat{B}_N(\xi,\xi') \tag{A10}$$

for each N. By inspection and the discussion above, all terms in Π_N with decay rate $1/N$ are accounted for in A_N, and all nondiagonal terms in $\Pi_N - A_N$ with decay rate $1/N^2$ appear after multiplication with N^2 in $\hat{B}_N$ with their correct limits, namely the corresponding terms in G, while terms with a faster decay rate disappear in the limit. Hence (A10) implies (A9).

A3: Markov Chains with Two Timescales—A Variation on a Lemma of Möhle

Conceptually, our convergence result rests on a separation-of-timescales phenomenon. It can be established with the help of a variant of a well-know result; see Lemma 1 from Möhle (1998).

Let E be a finite set. We equip matrices $A = (A(x,y))_{x,y\in E}$ on E with the matrix norm $\|A\| := \max_{x\in E}\sum_{y\in E}|A(x,y)|$. Note that then $\|AB\| \le \|A\|\|B\|$ and $\|A\| = 1$ if A is a stochastic matrix.

Lemma 1.7. *Assume that for $N \in \mathbb{N}$, A_N is a stochastic matrix on E such that*

$$\lim_{C\to\infty}\lim_{N\to\infty}\sup_{r\ge CN}\left\|A_N^r - P\right\| = 0 \tag{A11}$$

for some matrix P. Then we have for any $0 < c, K, t < \infty$

$$\lim_{N\to\infty}\sup_{\|B\|\le K}\left\|\left(A_N + cN^{-2}B\right)^{[tN^2]} - \left(P + cN^{-2}B\right)^{[tN^2]}\right\| = 0. \tag{A12}$$

Furthermore, if $(B_N)_{N\in}\mathbb{N}$ is a sequence of matrices on E such that

$$G := \lim_{N\to\infty} PB_NP \text{ exists}, \tag{A13}$$

then

$$\lim_{N\to\infty}\left(A_N + cN^{-2}B_N\right)^{[tN^2]} = Pe^{ctG} \qquad \text{for all } t > 0. \tag{A14}$$

Remark 1.8. *Instead of timescales N and N^2 one can allow more generally any a_N, $b_N \to \infty$ with $b_N/a_N \to \infty$, with only notational modifications in the proof.*

Proof of Lemma 1.7. We begin with (A12). Without loss of generality assume $K = 1$; otherwise replace B by B/K and c by cK. Fix $c, t > 0$ and a matrix B with $\|B\| \le 1$, and abbreviate $m := \lfloor tN^2\rfloor$. Let $\varepsilon > 0$ and choose $C_0 < \infty$ and $N_0 \in \mathbb{N}$ such that

$$\left\|A_N^r - P\right\| \le \varepsilon \quad \text{for} \quad N \ge N_0, r \ge C_0N \tag{A15}$$

[as guaranteed by (A11)]. Note that

$$\begin{aligned}&\left\|\left(A_N+cN^{-2}B\right)^m-\left(P+cN^{-2}B\right)^m\right\|\\&\le\left\|A_N^m-P\right\|+\sum_{k=1}^{m}\left(\frac{c}{N^2}\right)^k\sum_{\substack{m_1,\dots,m_{k+1}\in\mathbb{N}_0\\ m_1+\dots+m_{k+1}=m-k}}\left\|A_N^{m_1}\prod_{j=2}^{k+1}\left(BA_N^{m_j}\right)-P^{m_1}\prod_{j=2}^{k+1}\left(BP^{m_j}\right)\right\|.\end{aligned}$$

Mimicking the proof in Möhle (1998), we split the second summand into (the ellipses refer to the term inside the large-norm brackets on the right of the last line of the previous formula)

$$S_1:=\sum_{k=1}^{m}\left(\frac{c}{N^2}\right)^k\sum_{\substack{m_1,\dots,m_{k+1}\ge C_0N\\ m_1+\dots+m_{k+1}=m-k}}\dots\quad\text{and}\quad S_2:=\sum_{k=1}^{m}\left(\frac{c}{N^2}\right)^k\sum_{\substack{m_1,\dots,m_{k+1}\in\mathbb{N}_0\\ m_1+\dots+m_{k+1}=m-k\\ \exists j:m_j<C_0N}}\dots$$

As in Möhle (1998, p. 509) we have $S_1\le 2e^t(t+1)\varepsilon$ for all N large enough, and our estimate for S_2 is a small variation of the corresponding estimate in Möhle (1998). Each of the matrix norms appearing in the big sum in S_2 is at most 2, and hence (with $x\wedge y:=\min(x,y)$)

$$\begin{aligned}S_2&\le 2\sum_{k=1}^{m}\left(\frac{c}{N^2}\right)^k\#\left\{(m_1,\dots,m_{k+1})\in\mathbb{N}_0^k:\begin{array}{c}m_1+\ \dots\ +m_{k+1}=m-k,\\ \exists j:m_j<C_0N\end{array}\right\}\\&\le 2\sum_{k=1}^{m}\left(\frac{c}{N^2}\right)^k(k+1)\sum_{m_1=0}^{C_0N\wedge(m-k)}\binom{m-m_1-1}{k-1}\\&\le 2\sum_{k=1}^{m}\left(\frac{c}{N^2}\right)^k(k+1)C_0N\binom{m-1}{k-1}\ =2C_0N\frac{c}{N^2}\sum_{k=0}^{m-1}\left(\frac{c}{N^2}\right)^k(k+2)\binom{m-1}{k}\\&\le C'\frac{1}{N}.\end{aligned}$$

[We use in the last estimate that for $|x|<1$, $n\in\mathbb{N}$, $\sum_{n=0}^{\infty}\binom{n}{k}x^k=(1+x)^n$, and $\sum_{n=0}^{\infty}k\binom{n}{k}x^k=nx(1+x)^{n-1}$.]

The derivation of (A14) from (A12) is literally the same as in Möhle (1998, pp. 509–511) (read $c_N=c/N^2$ there).

A4: The Convergence Result with General Random Ψ_N

In this section we briefly indicate how the proof of Theorem 1.2 can be modified to yield Theorem 1.3. In each reproduction event, a random number Ψ_N of individuals die and are replaced by the same number of offspring, and we recall assumptions (20), (22) and (24). By short timescale we refer to the scaling a_N given by

$$a_N=\frac{N}{\mathbb{E}[\Psi_N]}$$

and by long timescale the scaling b_N given by

$$b_N=\frac{1}{c_N}=\frac{N(N-1)}{\mathbb{E}[\Psi_N(\Psi_N+3)]}.$$

Assumption (20) yields $b_N\to\infty$ as $N\to\infty$, and $b_N/a_N\to\infty$ by assumption (21). To check (23), *i.e.*, that indeed $a_N\to\infty$, observe that Ψ_N/N is a positive random variable, bounded by 1. Condition (20) is equivalent to $\mathbb{E}[(\Psi_N/N)^2]\to\mathbb{N}0$, which implies $\Psi_N/N\to 0$ in probability and $\mathbb{E}[\Psi_N/N]\to 0$ and hence (23).

For use below, we recall implications of (22) provided that (20) holds (*cf.* Sagitov 1999):

$$\text{For all } j\ge 3:\ \frac{1}{c_N}\mathbb{E}\left[\left(\frac{\Psi_N}{N}\right)^j\right]\to\int_{[0,1]}x^{j-2}F(dx).\tag{A16}$$

Indeed, integration by parts yields

$$\begin{aligned}\frac{1}{c_N}\mathbb{E}\left[\left(\frac{\Psi_N}{N}\right)^j\right] &= \frac{1}{c_N}\int_{(0,1]} jx^{j-1}\mathbb{P}\left(\frac{\Psi_N}{N}>x\right)dx \underset{N\to\infty}{\to} \int_{(0,1]} jx^{j-1}\int_{(x,1]} y^{-2}F(dy)dx \\ &= \int_{(0,1]}\left(\int_{(0,1]} 1_{\{x\le y\}} jx^{j-1}dx\right)y^{-2}F(dy) \\ &= \int_{(0,1]} y^{j-2}F(dy).\end{aligned} \tag{A17}$$

Furthermore for the case $j = 2$ one obtains

$$\limsup_{N\to\infty}\frac{1}{c_N}\mathbb{E}\left[\left(\frac{\Psi_N}{N}\right)^2\right] = \limsup_{N\to\infty}\frac{E[\Psi_N^2]}{E[\Psi_N(\Psi_N+3)]} \le 1 < \infty. \tag{A18}$$

Let $\tilde{\Psi}_N$ have the following reweighted distribution (relative to Ψ_N):

$$\mathbb{P}\left(\tilde{\Psi}_N = k\right) = \frac{k(k+3)}{E[\Psi_N(\Psi_N+3)]}\mathbb{P}(\Psi_N = k), \quad k = 1,\dots,N-2 \tag{A19}$$

and then

$$\frac{\tilde{\Psi}_N}{N} \xrightarrow{d} F \quad \text{as} \quad N\to\infty. \tag{A20}$$

Indeed, for any $\ell \in \mathbb{N}$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\tilde{\Psi}_N}{N}\right)^\ell\right] &= \frac{N(N-1)}{\mathbb{E}[\Psi_N(\Psi_N+3)]}\mathbb{E}\left[\left(\frac{\Psi_N}{N}\right)^{\ell+1}\frac{\Psi_N+3}{N-1}\right] \\ &= \frac{1}{c_N}\mathbb{E}\left[\left(\frac{\Psi_N}{N}\right)^{\ell+2}\right]\frac{N}{N-1} + \frac{3}{(N-1)}\frac{1}{c_N}\mathbb{E}\left[\left(\frac{\Psi_N}{N}\right)^{\ell+1}\right] \underset{N\to\infty}{\to} \int_{(0,1]} y^\ell F(dy)\end{aligned} \tag{A21}$$

by (A17) and (A18), so (A20) follows because the moments characterize a probability law on [0, 1]. One can check (along the lines of Sagitov 1999) that under assumption (20), both (A17) and (A18) are in fact equivalent to (22).

The proof of Theorem 1.3 is now a relatively straightforward adaptation of the proof of Theorem 1.2 discussed in sections A1 and A2 above. Scaling by N is throughout replaced by scaling with $a_N = N/\mathbb{E}[\Psi_N]$ and scaling by N^2 becomes scaling with $b_N = N(N-1)/\mathbb{E}[\Psi_N(\Psi_N+3)]$:

i. When currently following $b \ge 1$ individuals, the probability that none of them is an offspring in the previous reproduction event (and hence the sample configuration remains unchanged) is

$$\mathbb{E}\left[\frac{\binom{N-b}{\Psi_N}}{\binom{N}{\Psi_N}}\right] = \mathbb{E}\left[\prod_{j=0}^{\Psi_N-1}\frac{N-b-j}{N-j}\right] = \mathbb{E}\left[\prod_{j=0}^{\Psi_N-1}\left(1-\frac{b}{N-j}\right)\right] = 1 - O\left(b\frac{\mathbb{E}[\Psi_N]}{N}\right) = 1 - O(a_N^{-1}).$$

This is analogous to transitions discussed in section A1.2 and happens "all the time" (leading to the projecting transitions part in the limit).

ii. When currently following $b \ge 1$ individuals, say the kth of which is double marked, the probability that the ith individual is the only offspring in the sample, and that the sample also does not contain a parent, is [we write $(x)k = x(x-1)\dots(x-k+1)$ for the ith falling factorial]

$$\mathbb{E}\left[\frac{\Psi_N(N-\Psi_N-2)_{b-1}}{(N)_b}\right] \sim \mathbb{E}\left[\frac{\Psi_N}{N}\left(1-\frac{\Psi_N}{N}\right)^{b-1}\right] = a_N^{-1}(1+o(1)).$$

The projection matrix A_N now becomes

$$A_N(\xi, \mathrm{disp}_i(\xi)) = \mathbb{E}\left[\frac{\Psi_N(N-\Psi_N-2)_{b-1}}{(N)_b}\right](1-r)^2, \quad 1 \le i \le \beta - b \tag{A22}$$

and $A_N(\xi,\xi) = 1 - (\beta - b)\mathbb{E}[\Psi_N(N-\Psi_N-2)_{b-1}/(N)_b](1-r)^2$; the analogue of Proposition 1.4 is then

$$\lim_{C\to\infty} \lim_{N\to\infty} \sup_{r \ge Ca_N} |A_N^r - P| = 0. \tag{A23}$$

iii. From now on we can work on the "projected" space $\mathscr{A}_n^{\mathbf{sm}}$. The distinction between small and large reproduction events is irrelevant in the general case. Hence, it is more suitable to distinguish whether a parent and an offspring are in the sample or whether several offspring (but no parent) are in the sample. In analogy with (A5) and (A6), we split Π_N into fast and slow parts and define

$$B_N^* := b_N(\Pi_N - A_N), \quad \hat{B}_N := PB_N^*P. \tag{A24}$$

It then remains to check that

$$\hat{B}_N \to G \quad \text{with } G \text{ defined in } (26), \tag{A25}$$

whence Theorem 1.3 follows from Lemma 1.7 together with Remark 1.8.
We now verify (A25):
iv. Recombination events give the correct limit; see the discussion below (24).
v. "Large" is the probability that exactly $k \ge 2$ individuals among b (excluding the parents) is, using (A19),

$$\mathbb{E}\left[\frac{(\Psi_N)_k(N-\Psi_N-2)_{b-k}}{(N)_b}\right] = \mathbb{E}[\Psi_N(\Psi_N+3)]\mathbb{E}\left[\frac{\left(\tilde{\Psi}_N\right)_k\left(N-2-\tilde{\Psi}_N\right)_{b-k}}{\tilde{\Psi}_N\left(\tilde{\Psi}_N+3\right)(N)_b}\right]; \tag{A26}$$

thus $1/c_N$ times this probability is

$$\begin{aligned} N(N-1)\mathbb{E}\left[\frac{\left(\tilde{\Psi}_N\right)_k\left(N-2-\tilde{\Psi}_N\right)_{b-k}}{\tilde{\Psi}_N\left(\tilde{\Psi}_N+3\right)(N)_b}\right] &= \frac{1}{(N-2)_{b-2}}\mathbb{E}\left[\left(\tilde{\Psi}_N-2\right)_{k-2}\left(N-2-\tilde{\Psi}_N\right)_{b-k}\right] + O\left(\frac{1}{N}\right) \\ &\xrightarrow[N\to\infty]{} \int_{(0,1]} y^{k-2}(1-y)^{b-k}F(dy) \end{aligned} \tag{A27}$$

by (A20). Furthermore, the probability that at least *two* offspring *and* at least one parent are in the sample is at most

$$b\binom{b-1}{2}\mathbb{E}\left[\frac{2(\Psi_N)_2}{(N)_3}\right] = O(c_N/N); \tag{A28}$$

hence such events become negligible in the limit.
vi. "Small" is a merger of a single pair, which can result either from one offspring and one parent in the sample or from two offspring but no parent in the sample: Here, the weight of $F(\{0\})$ plays a role.

The probability that exactly two given single-marked individuals in a sample of size b are offspring (and none are parents) is

$$\mathbb{E}\left[\frac{(\Psi_N)_2(N-2-\Psi_N)_{b-2}}{(N)_b}\right]; \tag{A29}$$

and the probability that among a pair of two given single-marked individuals, one is a parent, the other is an offspring, and no other element of the sample is affected by the reproduction event is

$$\mathbb{E}\left[\frac{2(2)_1(\Psi_N)_1(N-\Psi_N-2)_{b-2}}{(N)_b}\right]; \tag{A30}$$

thus, $1/c_N$ times the probability that exactly one given pair (of single-marked individuals) is involved in a reproduction event is

$$\begin{aligned}
&\frac{1}{c_N}\mathbb{E}\left[\frac{\Psi_N(\Psi_N+3)(N-\Psi_N-2)_{b-2}}{(N)_b}\right]\\
&=\mathbb{E}\left[\frac{\left(N-2-\tilde{\Psi}_N\right)_{b-2}}{(N-2)_{b-2}}\right]\\
&\underset{N\to\infty}{\to}\int_{[0,1]}(1-y)^{b-2}F(dy)=F(\{0\})+\int_{(0,1]}(1-y)^{b-2}F(dy)
\end{aligned}\tag{A31}$$

by (A20).

vii. (Combinatorial connections between participation in reproduction events and merging of ancestral chromosomes) The rest of the argument to replace (15) by (27) is purely combinatorial; it is concerned only with possible groupings of the k single-marked offspring into up to four groups depending on which of the four parental chromosomes they descend from.

In both cases considered in (6) the probability that the chromosomes actually coalesce is $\frac{1}{4}$ because they must descend from the same chromosome in the same parent or from the particular chromosome in the particular parent we are following, respectively.

A5: Correlation in Coalescence Times

In this section we outline the calculations to obtain the correlation in coalescence times T_1 and T_2 of types at two loci (1 and 2). As our sample consists of two unlabeled chromosomes typed at two loci, we sometimes find it convenient to denote an unlabeled chromosome carrying ancestral segments at both loci with the symbol $\vdash\!\dashv$ and chromosomes carrying ancestral segments at only one locus with the symbols $\vdash$ and $\dashv$. Loci at which types have coalesced are denoted by $\bullet-$ or $\bullet\dashv$. The states S of the unlabeled process for a sample of size two at two loci are also numbered as

S	In symbols
2	$(\vdash\!\dashv)(\vdash\!\dashv)$
1	$(\vdash\!\dashv)(\vdash)(\dashv)$
0	$(\vdash)(\vdash)(\dashv)(\dashv)$
−1	$(\dashv)(\dashv)$
−2	$(\vdash))(\vdash)$

in which states {0, 1, 2} denote the three possible sample states, before coalescence at either loci has occurred. States {−1, −2} will be needed when deriving the variance of pairwise differences.

Let $h(i) := \mathbb{P}(\{T_1 = T_2\} \mid i)$ denote the probability of the event $T_1 = T_2$, when B is in state i. Excluding large offspring numbers, one readily obtains ($h(i) = 0$ for $i \neq \{0, 1, 2, \}$)

$$\begin{aligned}
h(2)&=\frac{r+9}{2r^2+13r+9}\\
h(1)&=\frac{3}{2r^2+13r+9}\\
h(0)&=\frac{2}{2r^2+13r+9}.
\end{aligned}\tag{A32}$$

For each $i \in \{0, 1, 2\}$, the expression for $h(i)$ is the same as the one for the correlation between T_1 and T_2 when in state i, excluding large offspring numbers. The expected value $w(i) = \mathbb{E}[T_s]$ of the time T_s until a coalescence event at either locus starting from state $i \in \{0, 1, 2\}$ is, again excluding large offspring numbers,

$$w(2) = \frac{r+9}{2(2r^2+13r+9)} + \frac{1}{2} = \frac{1}{2}(1+h(2))$$
$$w(1) = \frac{3}{2(2r^2+13r+9)} + \frac{1}{2} = \frac{1}{2}(1+h(1)),$$
$$w(0) = \frac{1}{2r^2+13r+9} + \frac{1}{2} = \frac{1}{2}(1+h(0)),$$

obtained by solving the recursions.

$$w(2) = \frac{1+2rw(1)}{1+2r}$$
$$w(2) = \frac{1+2rw(1)}{1+2r}$$
$$w(1) = \frac{1+w(2)+rw(0)}{r+3}$$
$$w(0) = \frac{1+4w(1)}{6}.$$

Let $v(i) := \mathbb{E}_i[T_s^2]$ denote the expected value of T_s^2 when starting from state $i \in \{0, 1, 2\}$. One can follow Durrett (2002) to obtain the recursions

$$v(i) = \frac{2}{q_i^2} + \frac{2}{q_i}\sum_{k\neq i}\frac{q_{ik}}{q_i}w(k) + \sum_{k\neq i}\frac{q_{ik}}{q_i}v(k) \tag{A33}$$

in which $q_i = \sum_{k\neq i} q_{ik}$ is the sum of the transition rates out of state i. To obtain (A33) let J denote the exponential waiting time until the first transition and X_J be the state of the process immediately after the first transition. The random variables J and X_J are independent. One can write

$$\begin{aligned}\mathbb{E}\left[T_s^2 \,|\, J, X_J\right] &= \mathbb{E}\left[(T_s - J + J)(T_s - J + J) \,|\, J, X_J\right] \\ &= \mathbb{E}\left[(T_s - J)^2 + 2J(T_s - J) + J^2 \,|\, J, X_J\right] \\ &= \mathbb{E}\left[(T_s - J)^2 \,|\, J, X_J\right] + 2J\mathbb{E}[T_s - J \,|\, X_J] + \mathbb{E}\left[J^2\right].\end{aligned}$$

Taking expectations gives (A33).

The variance $\mathbb{V}_i[T_s]$ of T_s when starting in state i is given by

$$\mathbb{V}_2[T_s] = \frac{r^3 + 31r^2/2 + 153r/2 + 81}{(2r+1)(r+6)(2r^2+13r+9)} + \frac{1}{2} - \frac{1}{4}(1+h(2))^2$$
$$\mathbb{V}_1[T_s] = \frac{r+9}{(r+6)(2r^2+13r+9)} + \frac{1}{2} - \frac{1}{4}(1+h(1))^2$$
$$\mathbb{V}_0[T_s] = \frac{r+8}{(r+6)(2r^2+13r+9)} + \frac{1}{2} - \frac{1}{4}(1+h(0))^2.$$

Hence, $\lim_{r\to\infty} \mathbb{V}_i[T_s] = 1/4$ for $i \in \{2, 1, 0\}$, and

$$\lim_{r\to 0} \mathbb{V}_2[T_s] = 1$$
$$\lim_{r\to 0} \mathbb{V}_1[T_s] = \frac{2}{9}$$
$$\lim_{r\to 0} \mathbb{V}_0[T_s] = \frac{89}{324}.$$

Denote by T_l the time until coalescence has occurred at both loci. The marginal coalescence times are exponential with rate 1, when excluding large offspring numbers. Solving the recursions

$$\mathbb{E}_2[T_l] = \frac{1 + 2r\mathbb{E}_1[T_l]}{1 + 2r}$$
$$\mathbb{E}_1[T_l] = \frac{1 + \mathbb{E}_2[T_l] + r\mathbb{E}_0[T_l] + 2}{r + 3}$$
$$\mathbb{E}_0[T_l] = \frac{1 + 4\mathbb{E}_1[T_l] + 2}{6}$$

yields

$$\mathbb{E}\left[T_l^{(2)}\right] = \frac{3}{2} - \frac{r + 9}{2(2r^2 + 13r + 9)} = \frac{1}{2}(3 - h(2))$$
$$\mathbb{E}\left[T_l^{(1)}\right] = \frac{3}{2} - \frac{3}{2(2r^2 + 13r + 9)} = \frac{1}{2}(3 - h(1))$$
$$\mathbb{E}\left[T_l^{(0)}\right] = \frac{3}{2} - \frac{1}{2r^2 + 13r + 9} = \frac{1}{2}(3 - h(0)).$$

Applying the recursions (A33) yields the variances $\mathbb{V}_i[T_l]$,

$$\mathbb{V}_2[T_l] = \frac{2r^3 + 111r^2/4 + 171r/2 - 81/4}{(2r^2+13r+9)^2} + \frac{5}{4}$$
$$\mathbb{V}_1[T_l] = \frac{4r^2 + 17r - 45/4}{(2r^2+13r+9)^2} + \frac{5}{4}$$
$$\mathbb{V}_0[T_l] = \frac{2r^2 + 7r - 10}{(2r^2+13r+9)^2} + \frac{5}{4}$$

with $\lim_{r\to\infty} \mathbb{V}_i[T_l] = 5/4$ for $i \in \{0, 1, 2\}$, and

$$\lim_{r\to 0} \mathbb{V}_2[T_l] = 1,$$
$$\lim_{r\to 0} \mathbb{V}_1[T_l] = \frac{10}{9},$$
$$\lim_{r\to 0} \mathbb{V}_0[T_l] = \frac{365}{324}.$$

Now we admit large offspring numbers, taking $\varepsilon_N = c/N^2$ and $r_N = r/N$. Ignoring the labeling of the chromosomes, the limit process has three effective sample states, depending on the number of double-marked chromosomes ($\vdash\dashv$). Denote the three sample states by

$$\begin{matrix}(\vdash\dashv)\\(\vdash\dashv)\end{matrix},$$

$$(\vdash\dashv)\begin{matrix}(\vdash)\\(\dashv)\end{matrix},$$

and

$$\begin{matrix}(\vdash) & (\dashv)\\(\vdash) & (\dashv)\end{matrix},$$

in which $\vdash$ and $\dashv$ denote single-marked chromosomes. The states of the limit process are composed of single-marked individuals only and are therefore the same as those of the haploid Wright–Fisher process. By $\bullet-$ denote a chromosome carrying a common ancestor at one locus, and $(\bullet-\bullet)$ denotes the absorbing states. The transition rates are summarized in the following table:

	$\binom{\vdash\dashv}{\vdash\dashv}$	$\binom{\vdash}{\dashv}\binom{\vdash}{\dashv}$	$\binom{\vdash}{\vdash}\binom{\dashv}{\dashv}$	$\binom{\bullet-}{\bullet-}\binom{\dashv}{\dashv}$	$(\bullet-)(\dashv)$	$(\bullet\!-\!\bullet)$
$\binom{\vdash\dashv}{\vdash\dashv}$		$2r$				$1+c\frac{\psi^2}{4}$
$\binom{\vdash}{\dashv}\binom{\vdash}{\dashv}$	$1+c\frac{\psi^2}{4}\left(1-\frac{\psi}{4}\right)$		r		$2+c\frac{\psi^2}{2}\left(1-\frac{\psi}{4}\right)$	$c\frac{\psi^3}{16}$
$\binom{\vdash}{\vdash}\binom{\dashv}{\dashv}$	$c\frac{3\psi^4}{32}$	$4+c\left(\psi^2-\frac{\psi^3}{2}-\frac{\psi^4}{8}\right)$	$2+c\left(\frac{\psi^2}{2}-\frac{\psi^3}{4}-\frac{\psi^4}{16}\right)$	$c\frac{\psi^3}{4}\left(1-\frac{\psi}{4}\right)$	$c\frac{\psi^4}{16}$	
$\binom{\bullet-}{\bullet-}\binom{\dashv}{\dashv}$				$2+c\frac{\psi^2}{2}\left(1-\frac{\psi}{4}\right)$	$1+c\frac{\psi^2}{4}$	
$(\bullet\!-\!\bullet)(\dashv)$			r		$1+c\frac{\psi^2}{4}$.	

By way of example, the rate of the transition from 1 to 2 by coalescence of the chromosomes $\vdash$ and $\dashv$ is $1 + cC_{3;2;1}$, the transition rate from 0 to 1 is $4(1 + cC_{4;2;2})$, and the transition rate from 0 to the absorbing state [$(\bullet\!-\!\bullet$ or $(\bullet-)(-\bullet)$] is $c(C_{4;4;0} + C_{4;2;2;0})$.

As before, let $h(i)$ denote the probability the two loci coalesce at the same time. One obtains limit results

$$\lim_{r\to\infty} h(i) = \frac{c\psi^4}{32 + 8c\psi^2 - c\psi^4},\ i \in \{0,1,2\}$$
$$\lim_{c\to\infty} h(2) = 1$$
$$\lim_{c\to\infty} h(1) = \frac{2}{6-\psi}$$
$$\lim_{c\to\infty} h(0) = \frac{56\psi^2/3 - 272\psi + 544}{(\psi-6)(3\psi^2+16\psi-48)} - \frac{5}{3}.$$

The first equation in (A34) tells us that the loci remain correlated due to multiple mergers even when they are far apart on a chromosome. When the recombination rate r is quite small, one obtains

$$\lim_{r\to 0} h(2) = 1$$
$$\lim_{r\to 0} h(1) = \frac{2(c\psi^2+4)}{-c\psi^3+6c\psi^2+24}$$
$$\lim_{r\to 0} h(0) = \frac{1}{3}\left(\frac{8c\psi^2+32}{-c\psi^3+6c\psi^2+24} + \frac{-80c\psi^3+208c\psi^2+832}{-3c\psi^4-16c\psi^3+48c\psi^2+192} - 5\right).$$

Let $\mathbb{E}_i[T_s]$, as before, denote the time until coalescence at either locus, starting from state i. Admitting large offspring numbers, one obtains

$$\lim_{r\to\infty} \mathbb{E}_i[T_s] = \frac{16}{32 + 8c\psi^2 - c\psi^4}, \quad i \in \{0,1,2\},$$
$$\lim_{c\to\infty} \mathbb{E}_i[T_s] = 0, \quad i \in \{0,1,2\},$$
$$\lim_{r\to 0} \mathbb{E}_2[T_s] = \frac{4}{c\psi^2+4}$$
$$\lim_{r\to 0} \mathbb{E}_1[T_s] = \frac{c(16\psi^2-2\psi^3)+64}{-c^2\psi^5+6c^2\psi^4-4c\psi^3+48c\psi^2+96}$$
$$\lim_{r\to 0} \mathbb{E}_0[T_s] = \frac{16}{3(c(6\psi^2-\psi^3)+24)} - \frac{4(\psi-8)}{(3\psi+16)(c\psi^2+4)} - \frac{32(39\psi-32)}{3(c(3\psi^4+16\psi^3-48\psi^2)-192)(3\psi+16)}.$$

Let $\mathbb{E}_i[T_l]$, as before, denote the expected value of the time T_l until coalescence has occurred at both loci, when starting from state i. Admitting large offspring numbers, one obtains the limits

$$\lim_{r\to\infty}\mathbb{E}_i[T_l]=\frac{c\left(48\psi^2-8\psi^4\right)+192}{\left(c\psi^2+4\right)\left(-c\psi^4+8c\psi^2+32\right)},\quad i\in\{0,1,2\},$$
$$\lim_{c\to\infty}\mathbb{E}_i[T_l]=0,\quad i\in\{0,1,2\},$$
$$\lim_{r\to 0}\mathbb{E}_2[T_l]=\frac{4}{c\psi^2+4}$$
$$\lim_{r\to 0}\mathbb{E}_1[T_l]=\frac{c\left(32\psi^2-6\psi^3\right)+128}{-c^2\psi^5+6c^2\psi^4-4c\psi^3+48c\psi^2+96}$$
$$\lim_{r\to 0}\mathbb{E}_0[T_l]=\frac{\left(28\psi^7-56\psi^6-800\psi^5+1600\psi^4\right)c^2+\left(-608\psi^4-3200\psi^3+12,800\psi^2\right)c+25,600}{a}$$

in which

$$a=3c^3\psi^9-2c^3\psi^8-144c^3\psi^7+288c^3\psi^6+12c^2\psi^7-80c^2\psi^6-1152c^2\psi^5+3456c^2\psi^4-288c\psi^4-2304c\psi^3+13,824c\psi^2+18,432.$$

Considering the variance $\mathbb{V}_i[T_s]$ of the time T_s when starting from state $i\in\{0,1,2\}$, and admitting large offspring numbers, one obtains

$$\lim_{r\to\infty}\mathbb{V}_i[T_s]=\frac{256}{\left(c\left(8\psi^2-\psi^4\right)+32\right)^2},\quad i\in\{0,1,2\},$$
$$\lim_{c\to\infty}\mathbb{V}_2[T_s]=0,\quad i\in\{0,1,2\},$$
$$\lim_{r\to 0}\mathbb{V}_2[T_s]=\frac{16}{\left(c\psi^2+4\right)^2}$$
$$\lim_{r\to 0}\mathbb{V}_1[T_s]=\frac{\left(12\psi^6-128\psi^5+384\psi^4\right)c^2+\left(3072\psi^2-512\psi^3\right)c+6144}{\left(c\psi^2+4\right)^2\left(-c\psi^3+6c\psi^2+24\right)^2}.$$

Correlations in coalescence times have been employed to quantify LD (McVean 2002), in which LD is quantified as the square of the correlation coefficient of types at two loci (Hill and Robertson 1968). A description of how one can quantify linkage disequilibrium as the square of the correlation coefficient of types at two loci can be found in Hartl and Clark (1989). Assuming a very small mutation rate, McVean (2002) related $\mathfrak{D}$ to covariances in coalescence times. Writing $\mathrm{Cov}_i(T_1, T_2)$ as the covariance of T_1 and T_2 when starting from state $i\in\{0,1,2\}$, McVean (2002) obtained

$$\mathfrak{D}=\frac{\mathrm{Cov}_2[T_1,T_2]-2\mathrm{Cov}_1[T_1,T_2]+\mathrm{Cov}_0[T_1,T_2]}{(\mathbb{E}[T_1])^2+\mathrm{Cov}_0[T_1,T_2]}$$
$$=1+\frac{\mathbb{E}_2[T_1T_2]-2E_1[T_1T_2]}{\mathbb{E}_0[T_1T_2]}$$

in which T_1 and T_2 denote the times until coalescence at the two loci, respectively, and the covariances are conditional on the sample configurations, as indicated. Following, *e.g.*, Durrett (2002) one can obtain the covariances under any population model. Under our population model, $\mathfrak{D}=\mathfrak{D}_1/\mathfrak{D}_2$, in which

$$\mathfrak{D}_1=640c\psi^2-224c\psi^3+32c\psi^4+80c^2\psi^4-56c^2\psi^5+16c^2\psi^6-c^2\psi^7+r\left(16c\psi^4-32c\psi^3+64c\psi^2+256\right)+1280,$$
$$\mathfrak{D}_2=1408c\psi^2-352c\psi^3+8c\psi^4+512r^2+176c^2\psi^4-88c^2\psi^5+10c^2\psi^6-c^2\psi^7+r\left(8c\psi^4-288c\psi^3+832c\psi^2+3328\right)+2816.$$

One obtains the limit results

$$\lim_{r\to\infty}\mathfrak{D}=0,$$

$$\lim_{c\to\infty}\mathfrak{D}=\frac{\psi^3-16\psi^2+56\psi-80}{\psi^3-10\psi^2+88\psi-176}.$$

A6: Correlations in Coalescence Times for Random ψ

In this section we consider the simple example of the probability measure F, evoked in relation to a random offspring distribution, taking the beta distribution with parameters ϑ and γ. The following transition rates for a sample of size two at two loci are obtained:

	$\binom{\vdash\dashv}{\vdash\dashv}$	$\binom{\vdash}{\dashv}\binom{\vdash}{\dashv}$	$\binom{\vdash}{\vdash}\binom{\dashv}{\dashv}$	$\binom{\bullet\!-}{\dashv}\binom{\dashv}{\dashv}$	$(\bullet\!-)(\dashv)$	$(\bullet\!-\!\bullet)$
$\binom{\vdash\dashv}{\vdash\dashv}$		$2r$				1
$\binom{\vdash}{\dashv}\binom{\vdash}{\dashv}$	$\frac{\gamma+3\vartheta/4}{\vartheta+\gamma}$		r		$2\frac{\gamma+3\vartheta/4}{\vartheta+\gamma}$	$\frac{\vartheta}{4(\vartheta+\gamma)}$
$\binom{\vdash}{\vdash}\binom{\dashv}{\dashv}$	$\frac{3}{8}\frac{(1+\vartheta)\vartheta}{(1+\vartheta+\gamma)(\vartheta+\gamma)}$	$\frac{4(1+\gamma)\gamma+3\vartheta\gamma+(3/2)(1+\vartheta)\vartheta}{(1+\vartheta+\gamma)(\vartheta+\gamma)}$		$\frac{2(1+\gamma)\gamma+(3/2)\vartheta\gamma+(3/4)(1+\vartheta)\vartheta}{(1+\vartheta+\gamma)(\vartheta+\gamma)}$	$\frac{\vartheta\gamma+(3/4)(1+\vartheta)\vartheta}{(1+\vartheta+\gamma)(\vartheta+\gamma)}$	$\frac{(\vartheta+1)\vartheta}{4(\vartheta+\gamma+1)(\vartheta+\gamma)}$
$\binom{\bullet\!-}{\dashv}\binom{\dashv}{\dashv}$					$2\frac{\gamma+3\vartheta/4}{\vartheta+\gamma}$	1
$(\bullet\!-\!\bullet)(\dashv)$				r		$1.$

As before, the transition rates given above can be employed to derive correlations in coalescence times. Here we consider only the probability $h(i)$. One obtains $\lim_{\vartheta\to 0}h(i)=\lim_{\gamma\to\infty}h(i)$ and the limit results are those obtained from the usual ARG (A32).

A7: Variance of Pairwise Differences

The variance of pairwise differences between DNA sequences has been employed to estimate recombination rates in low offspring number populations (Wakeley, 1997). Let the random variable K_{ij} denote the number of differences between sequences i and j, with $K_{ii}=0$. The average number π of pairwise differences for n sequences is

$$\pi=\frac{2}{n(n-1)}\sum_{i<j}K_{ij}.$$

Under the infinitely many sites mutation model, $\mathbb{E}[\pi]=\theta\mathbb{E}[T]$, in which T is the time until coalescence of two sequences. Under our model, $\mathbb{E}[T]=1/(1+c\psi^2/4)$. Define the variance S_π^2 of pairwise differences as

$$S_\pi^2=\frac{2}{n(n-1)}\sum_{i<j}\left(K_{ij}-\pi\right)^2.$$

To obtain an estimate of the recombination rate, one needs to compute the expected value $\mathbb{E}[S_\pi^2]$,

$$\mathbb{E}\left[S_\pi^2\right]=\frac{2}{n(n-1)}\sum_{i<j}\mathbb{E}\left[\left(K_{ij}-\pi\right)^2\right]=\mathbb{E}\left[\left(K_{12}-\pi\right)^2\right].$$

Thus, it suffices to consider $\mathbb{E}[(K_{12}-\pi)^2]$. Expanding, one obtains

$$\mathbb{E}\left[\left(K_{12}-\pi\right)^2\right]=\mathbb{E}\left[\left(\frac{2}{n(n-1)}\sum_{i<j}\left(K_{12}-K_{ij}\right)\right)^2\right]=\frac{4}{n^2(n-1)^2}\sum_{i<j}\sum_{\hat{i}<\hat{j}}\mathbb{E}\left[\left(K_{12}-K_{ij}\right)\left(K_{12}-K_{\hat{i}\hat{j}}\right)\right].$$

Define the event $A_{ij}^{(\ell)}$ by

$$A_{ij}^{(\ell)}:=\{\text{sequences } i \text{ and } j \text{ differ at locus } \ell\}.$$

Assuming each sequence consists of L loci, and $1_{A_{ij}^{(\cdot)}}$ are indicator functions,

$$K_{12} - K_{ij} = \sum_{\ell=1}^{L} \left(1_{A_{12}^{(\ell)}} - 1_{A_{ij}^{(\ell)}} \right)$$

yielding, in case $i = i^1 = 1$, and $j = \hat{j} = 3$,

$$\begin{aligned} \mathbb{E}\left[(K_{12} - K_{13})^2 \right] &= \sum_{\ell=1}^{L} \sum_{\hat{\ell}=1}^{L} \mathbb{E}\left[\left(1_{A_{12}^{(\ell)}} - 1_{A_{13}^{(\ell)}} \right) \left(1_{A_{12}^{(\hat{\ell})}} - 1_{A_{13}^{(\hat{\ell})}} \right) \right] \\ &= 2 \sum_{\ell=1}^{L} \sum_{\hat{\ell}=1}^{L} \mathbb{P}\left(A_{12}^{(\ell)} \cap A_{12}^{(\hat{\ell})} \right) - \mathbb{P}\left(A_{12}^{(\ell)} \cap A_{13}^{(\hat{\ell})} \right). \end{aligned}$$

In general,

$$\begin{aligned} \mathbb{E}\left[(K_{12} - K_{ij})(K_{12} - K_{\hat{i}\hat{j}}) \right] &= \sum_{\ell=1}^{L} \sum_{\hat{\ell}=1}^{L} \mathbb{E}\left[\left(1_{A_{12}^{(\ell)}} - 1_{A_{ij}^{(\ell)}} \right) \left(1_{A_{12}^{(\hat{\ell})}} - 1_{A_{\hat{i}\hat{j}}^{(\hat{\ell})}} \right) \right] \\ &= \sum_{\ell=1}^{L} \sum_{\hat{\ell}=1}^{L} \mathbb{P}\left(A_{12}^{(\ell)} \cap A_{12}^{(\hat{\ell})} \right) - \mathbb{P}\left(A_{12}^{(\ell)} \cap A_{\hat{i}\hat{j}}^{(\hat{\ell})} \right) - \mathbb{P}\left(A_{12}^{(\hat{\ell})} \cap A_{ij}^{(\ell)} \right) + \mathbb{P}\left(A_{ij}^{(\ell)} \cap A_{\hat{i}\hat{j}}^{(\hat{\ell})} \right). \end{aligned} \tag{A36}$$

Now consider the probability $\mathbb{P}(A_{12}^{(\ell)} \cap A_{12}^{(\hat{\ell})})$ of the event that sequences 1 and 2 differ at both loci ℓ and $\hat{\ell}$. Admitting mutation introduces two new states, namely the states

$$\begin{matrix} (\vdash) \\ (\vdash) \end{matrix}$$

and

$$\begin{matrix} (\dashv) \\ (\dashv) \end{matrix}.$$

Define

$$g(\mathfrak{S}) := \mathbb{P}(\text{both loci separated by mutation, starting fromstate } \mathfrak{S})$$

Thus, $\mathbb{P}(A_{12}^{(\ell)} \cap A_{12}^{(\hat{\ell})}) = g(2)$, $\mathbb{P}(A_{12}^{(\ell)} \cap A_{13}^{(\hat{\ell})}) = g(1)$, and $\mathbb{P}(A_{12}^{(\ell)} \cap A_{34}^{(\hat{\ell})}) = g(0)$, for $\ell \neq \hat{\ell}$. Now,

$$g(2) = \frac{\theta_1 g(-1) + \theta_2 g(-2) + 2r g(1)}{\theta_1 + \theta_2 + 1 + c\dfrac{\psi^2}{4} + 2r}$$

$$g(-1) = \frac{\theta_2}{\theta_2 + 1 + c\dfrac{\psi^2}{4}}$$

$$g(-2) = \frac{\theta_1}{\theta_1 + 1 + c\dfrac{\psi^2}{4}}$$

$$g(1) = \frac{\theta_1 g(-1) + \theta_2 g(-2) + r g(0) + \left(1 + c\psi^2/4\right) g(2)}{\theta_1 + \theta_2 + r + 3 + 3c\left(\psi^2/4\right)(1 - \psi/4) + c\left(\psi^3/16\right)}$$

$$g(0) = \frac{\theta_1 g(-1) + \theta_2 g(-2) + c\left(3\psi^4/32\right) g(2) + \left(c\left(\psi^2 - \psi^3/2 - \psi^4/8\right) + 4 \right) g(1)}{c\left(3\psi^4/32\right) + c\left(\psi^2 - \psi^3/2 - \psi^4/8\right) + c\left(\psi^2/2 - \psi^3/4 - \psi^4/16\right) + 6 + c\left(\psi^3/4\right)(1 - \psi/4) + c\left(\psi^4/16\right) + \theta_1 + \theta_2}.$$

In view of expression (A36), one obtains

$$
\begin{aligned}
\mathbb{P}\left(A_{12}^{(\ell)} \cap A_{12}^{(\ell)}\right) &= \mathbb{P}\left(A_{12}^{(\ell)}\right) = \frac{\theta_\ell}{\theta_\ell + 1 + c\psi^2/4}, \\
\mathbb{P}\left(A_{12}^{(\ell)} \cap A_{13}^{(\ell)}\right) &= \frac{\theta_\ell}{3\theta_\ell/2 + \lambda_3} + \frac{\lambda_2}{3\theta_\ell/2 + \lambda_3}\frac{\theta_\ell}{\theta_\ell + \lambda_2}, \\
\mathbb{P}\left(A_{12}^{(\ell)} \cap A_{34}^{(\ell)}\right) &= \frac{2\theta_\ell}{2\theta_\ell + \lambda_4}\frac{\theta_\ell}{2\theta_\ell + \lambda_4} + \frac{\lambda_{4;2}}{2\theta_\ell + \lambda_4}\left(\frac{\theta_\ell/2}{3\theta_\ell/2 + \lambda_3} + \left(\frac{\theta_\ell/2}{3\theta_\ell/2 + \lambda_3}\right)^2\right).
\end{aligned}
\tag{A37}
$$

The event $A_{12}^{(\ell)} \cap A_{34}^{(\ell)}$ (Equation A37) occurs if the first two events in the history of the four sequences are mutations on appropriate ancestral lineages or if lineages labeled 2 and 3 coalesce, followed by appropriately placed mutations.