Statistics in Molecular Biology
IMS Lecture Notes - Monograph Series (1999) Volume 33

THE COALESCENT WITH PARTIAL SELFING AND BALANCING SELECTION: AN APPLICATION OF STRUCTURED COALESCENT PROCESSES

BY MAGNUS NORDBORG

Lund University

As a demonstration of a generally applicable technique, a theorem based on separation of time scales in the structured coalescent is used to extend results for the coalescent process with balancing selection to allow partial selfing. The resulting model behaves like the random-mating one, but with different rates of coalescence and recombination. This result has important implications for attempts to locate selectively maintained polymorphisms. Such polymorphisms can in principle be detected through their effect on the pattern of polymorphism in the genomic region surrounding the site under selection, however this is not practically feasible unless the effected region is sufficiently large. An implication of the present results is that the region is expected to be much larger in partially selfing organisms than in outcrossing ones, suggesting that studies attempting to locate selectively maintained polymorphisms should utilize selfing organisms.

1. Introduction. If natural selection has maintained polymorphism at a certain site or locus for a long period of time, the evolutionary dynamics in closely linked regions of the chromosome will be effected. In particular, the expected pattern of neutral polymorphism will be altered in a manner that allows inference about the action of selection directly from molecular polymorphism data, without phenotypic observation [Hudson and Kaplan (1988)]. This phenomenon has been observed in a few cases [Kreitman and Akashi (1995), Hudson (1996)]: MHC (immune system) loci in human; the S (self-incompatibility) locus in plants; and *Adh* (alcohol dehydrogenase) in *Drosophila melanogaster.* Although these loci were already known to harbor selectively maintained polymorphisms, the same phenomenon could, in principle, be used to locate such polymorphisms without prior information. For this to be practical, however, the region of the chromosome in which the effects of selection are noticeable must be large enough. One of the aims of this paper is to prove earlier claims [Nordborg et al. (1996), Nordborg (1997), Charlesworth et al. (1997)] that the effected regions will be much wider in partially selfing organisms than in outcrossing

Most of the research that led to this paper was done while the author was in the Department of Ecology & Evolution at the University of Chicago, initially as a research associate, and later as a visitor. Partial financial support was provided by National Science Foundation grants DEB-9217683 to D. Charlesworth and DEB-9350363 to J. Bergelson.

AMS 1991 *subject classifications.* Primary 60G35, 92D10; secondary 60F05.

Key words and phrases. Population genetics, molecular evolution, structured coalescent, selection, selfing, time scales.

ones.

The main result is a generalization of a previous treatment of the coalescent with selection [Kaplan et al. (1988), Hudson and Kaplan (1988)] to allow partial selfing. It is based on an argument about separation of time scales in the structured coalescent [Nordborg and Donnelly (1997), Nordborg (1997)], and utilizes a convergence theorem developed for this purpose by Möhle (1998). A second aim of this paper is to demonstrate the power of this approach to greatly simplify the analysis of quite complicated coalescent models. Connections with other approaches are discussed in Section 5.1.

2. The model. The model described in this section is effectively equivalent to previously used models [Kaplan et al. (1988), Hudson and Kaplan (1988), Hey (1991)], except for allowing partial selfing. We assume a population of N (assumed to be large and constant) diploid individuals that are hermaphroditic (*i. e.*, each individual produces gametes of both types, *e. g.*, pollen and ovules) and partially selfing (individuals can fertilize themselves; a more precise definition will be given below). The population has discrete generations. In each generation, all individuals produce infinitely many gametes which unite to form infinitely many zygotes. Mutation and recombination occur during gamete formation. The zygotes are then subject to selection, after which N of them are chosen to form the next generation of adults.

2.1. *Forward dynamics at the selected locus.* Consider a locus with two alleles $\mathcal{A}_1$ and $\mathcal{A}_2$. The mutation probability (per gamete per generation) from $\mathcal{A}_i$ to $\mathcal{A}_j$ is u_{ij}. There are three possible genotypes $\mathcal{A}_1\mathcal{A}_1$, $\mathcal{A}_1\mathcal{A}_2$, and $\mathcal{A}_2\mathcal{A}_2$. Let $N_{ij}(t)$ be the (random) number of individuals in the adult population of genotype $\mathcal{A}_i\mathcal{A}_j$ in generation t: because $\sum N_{ij}(t) = N$, the $N_{ij}(t)$ are not independent. Define the genotype frequencies as $X_{ij}(t) = N_{ij}(t)/N$, and the allele frequencies as $Y_1(t) = X_{11}(t) + X_{12}(t)/2$ and $Y_2(t) = X_{22}(t) + X_{12}(t)/2$.

Let $y_i(t)$, $i \in \{1,2\}$ be the allele frequencies among the gametes produced by the adults in generation t (*i. e.*, $y_i(t)$ is the proportion of gametes carrying $\mathcal{A}_i$). Because infinitely many gametes are produced, these are functions of the $Y_i(t)$. From standard population genetics theory, we have

$$y_1(t) = (1 - u_{12})Y_1(t) + u_{21}Y_2(t), \tag{2.1}$$

$$y_2(t) = (1 - u_{21})Y_2(t) + u_{12}Y_1(t). \tag{2.2}$$

Similarly, let $x_{ij}(t)$ be the zygotic frequencies. If mating were random, these frequencies could be found by simply multiplying the gamete frequencies. For a partially selfing population, however, it is assumed that only a fraction $1 - s$ of

the available female gametes are fertilized through random mating (outcrossed), and that the remaining fraction are fertilized by male gametes from the same individual. In other words, a fraction $1 - s$ of the zygotes are produced by random union of gametes within the population, where the gamete frequencies are $y_1(t)$ and $y_2(t)$, and a fraction s are produced by random union of gametes within individuals, so that

$$\begin{aligned} x_{11}(t) &= (1-s)y_1^2(t) \\ &\quad + s\big[(1-u_{12})^2 X_{11}(t) + (1-u_{12}+u_{21})^2 X_{12}(t)/4 + u_{21}^2 X_{22}(t)\big], \end{aligned} \tag{2.3}$$

$$\begin{aligned} x_{12}(t) &= 2(1-s)y_1(t)y_2(t) + s\big[(1-u_{12}+u_{21})(1-u_{21}+u_{12})X_{12}(t)/2 \\ &\quad + 2(1-u_{12})u_{12}X_{11}(t) + 2(1-u_{21})u_{21}X_{22}(t)\big], \end{aligned} \tag{2.4}$$

$$\begin{aligned} x_{22}(t) &= (1-s)y_2^2(t) \\ &\quad + s\big[(1-u_{21})^2 X_{22}(t) + (1-u_{21}+u_{12})^2 X_{12}(t)/4 + u_{12}^2 X_{11}(t)\big]. \end{aligned} \tag{2.5}$$

The *viability* (relative chance of surviving to adulthood) of a zygote with genotype $\mathcal{A}_i\mathcal{A}_j$ is $1 - w_{ij}(t)$, where we do not exclude the possibility that $w_{ij}(t)$ is a function of the $x_{ij}(t)$ (as in frequency-dependent selection). Let $x_{ij}^\star(t)$ denote the genotype frequencies after selection. Then

$$x_{ij}^\star(t) = \frac{x_{ij}(t)(1 - w_{ij}(t))}{\bar{w}(t)}, \tag{2.6}$$

where $\bar{w}(t) = 1 - \sum w_{ij}(t)x_{ij}(t)$.

Generation $t + 1$ is formed by drawing N individuals from the surviving zygotes. Conditional on the $N_{ij}(t)$, the $N_{ij}(t + 1)$ are thus multinomially distributed with parameters N and $x_{ij}^\star(t)$.

2.2. *Genealogy at a linked neutral locus.* Our aim is to describe the gene genealogy at a locus linked to the selected site with recombination rate r. The locus is assumed to be neutral, *i. e.*, the only selection in the model is that on the locus described in Section 2.1. Recombination is only allowed *between* the loci. From a biological point of view, this requires that the length of the DNA sequences defined as "loci" be small relative to the distance between them. As is usual when tracing genealogies, time will be run backwards (the reverse direction from that in the previous section), so that generation ancestral to t is $t + 1$.

Consider a single chromosome, sampled from the adult population. With respect to the genotype of this adult, the sampled *instance* of the neutral gene is in one of three *genotypic* states. It is of course also characterized by being

linked to either an $\mathcal{A}_1$ or an $\mathcal{A}_2$ allele, which we will refer to as its *haplotypic* state. With respect to both classifications jointly, there are four possible states. Given this joint state in the present generation, what was the state of its ancestor in the previous generation? It should be clear that the transition probabilities with respect to the genotype of the ancestral individual can be calculated exactly from equations (2.3)–(2.5). To account for the haplotypic state in the previous generation we also need that if the current haplotypic state is $i \in \{1, 2\}$, then

1. if the ancestral individual was an $\mathcal{A}_i$ homozygote, the haplotypic state cannot have changed;
2. if the ancestral individual was a $\mathcal{A}_j$ homozygote ($j \neq i$), the haplotypic state must have changed (because a mutation occurred at the selected locus);
3. if the ancestral individual was a heterozygote, then the haplotypic state changed if there was either a mutation or a recombination event, but not both.

From this, the exact transition probabilities can be calculated.

Next, consider a sample of size n such instances. Each occupies one of the four states just described. In addition, they may or may not occupy the same individual as another instance. The possible genotypes for any individuals that harbor two instances is determined by the joint genotypic and haplotypic classification: for example, it is not possible for two instances to occupy the same heterozygote unless the sample contains two instances in heterozygotes, one linked to an $\mathcal{A}_1$ allele and the other linked to an $\mathcal{A}_2$ allele. The total number of states for a sample of size n is

$$\sum_{i=0}^{[n/2]} \binom{n-2i+3}{3} \binom{i+2}{2}.$$

What about the transition probabilities from such a state in the current generation t to the possible states in the previous generation $t+1$? Under the assumptions of the model, any instance that occurs singly in an individual will "pick" its ancestral state independently of all other instances. If j, $2 \leq j \leq n$ instances pick the same genotype $\mathcal{A}_i\mathcal{A}_j$ in the generation $t+1$, then, by standard arguments, the probability that two of them pick the same parental individual is

$$\frac{\binom{j}{2}}{N_{ij}(t+1)} + O\left(\frac{1}{N_{ij}^2(t+1)}\right), \tag{2.7}$$

the probability that more than two pick the same individual is $O\left(1/N_{ij}^2(t+1)\right)$, and the probability that they all pick different individuals is simply one minus

expression (2.7). Whenever two instances pick the same individual, one of two things happen. They either pick distinct haplotypes and thus their ancestors occupy the same individual in generation $t+1$, or they pick the same haplotype, in which case they *coalesce*, and the number of distinct ancestors of the sample decreases permanently by one.

Two instances that currently occupy the same individual also pick their ancestral states independently of each other and all other ancestors, if we condition on that individual having resulted from outcrossing, which by definition is equivalent to random mating. If, on the other hand, an individual in the current generation that harbors two instances was the product of selfing, the two instances will behave precisely as if they had chosen a common parental individual through random mating, and they will thus either coalesce or continue to occupy the same individual as just described.

We will return to the transition probabilities below, however, two important remarks should be made in this context. The first is that, if $s = 0$, all individuals will always have resulted from random mating, and there is therefore no need to keep track of whether instances occupy the the same individual or not. The second is that, going backwards in time, the ancestors of two instances that currently occupy the same individual will not continue doing so very long. For homozygotes this is so because if they result from selfing, then with probability one half the ancestors coalesce, and if they result from outcrossing, then the ancestors no longer occupy the same individual. For heterozygotes coalescence is considerably less likely, because it necessitates a recombination or mutation event, but it will become clear that heterozygotes are almost always the result of recent outcrossing.

2.3. *Approximate model.* Conditional on $\{N_{ij}(t)\}_{t\in\{0,1,\ldots\}}$, the genealogy of the n sampled copies can clearly be described by a discrete-time Markov chain with finite state space S_n of size

$$|S_n| = \sum_{k=1}^{n}\sum_{i=0}^{[k/2]}\binom{k-2i+3}{3}\binom{i+2}{2}. \tag{2.8}$$

How to describe the genealogy without conditioning on the allele frequencies is considerably less clear, and I refer to Section 5.1 for further discussion of this.

The issue can be avoided by assuming that selection and/or mutation are/is strong enough relative to drift (*i. e.*, relative to $1/N$) and of the correct form (*i. e.*, allowing a stable point equilibrium) for the allele frequencies to be treated as constant. In this paper, I will assume that it is selection alone that maintains constant frequencies. I will furthermore be assuming that selection is *balancing*,

by which I simply mean that the equilibrium is polymorphic. More precisely then, the model I will investigate is the discrete-time Markov chain that results from assuming that $Y_i(t) = \hat{Y}_i > 0$, $\forall i, t$. Define its transition matrix $\Pi_N = (\pi_{ij})_{i,j \in S_n}$. The state space S_n will be described in Section 3.2.

The model may be characterized as a generalized structured n-coalescent with discrete generations [Notohara (1990), Herbots (1994), Nordborg (1997)]. Selection has "disappeared", and only enters the model indirectly by determining the size of the "subpopulations" and the rates of "migration" between them. Although simple in principle, the model is difficult to work with because of the size and complexity of the state space. For example, for $n = 1, 2, 3, 4, 5$, and 6, we have $|S_n| = 4, 17, 49, 120, 260$, and 519. The remainder of this paper will be devoted to demonstrating that, through the use of a time-scales approximation, the model can be reduced to a much simpler, continuous-time structured n-coalescent with a state space of size

$$\sum_{m=1}^{n}(m+1) = \frac{1}{2}n(n+3), \tag{2.9}$$

which for the sample sizes just given equals $2, 5, 9, 14, 20$ and 27. Furthermore, the new process will be shown to be identical to the one previously obtained for random mating [Kaplan et al. (1988), Hudson and Kaplan (1988)], except that the coalescent rate is increased by a constant factor, and the rate of exchange between the haplotypic states (through recombination and mutation), is decreased by another constant factor, where both factors have useful intuitive interpretations.

After obtaining these results (in Sections 3 and 4), I will argue (Section 5.1) that the results from the analysis will also hold to a good approximation for the original model, where the allele frequencies are random variables, if only selection is of the appropriate strength and kind.

3. Continuous time. The analysis will proceed in two steps. In this section, I will switch to a continuous time scale with time measured in units of N generations, as in the standard coalescent approximation [Kingman (1982)]. To do so, I will utilize a convergence theorem that was developed by Möhle (1998) for coalescent models with transitions on several time scales, in particular for the neutral coalescent with selfing [Nordborg and Donnelly (1997)]. The resulting process is considerably simpler, however, I will show how it can be simplified further in Section 4, where I describe the final model.

3.1. *Scaling the parameters.* The rationale behind the main results of this paper is that some transitions, namely those that involve two ancestors picking

the same individual in the previous generation, always have probability $O(1/N)$, whereas others, *e. g.*, those that simply involve a transition to a new genotypic configuration, always have probability $O(1)$. As we shall see, the states can be grouped into sets in a manner so that transitions within sets occur on time scale that is $O(N)$ faster than transitions between sets. Exactly how the state space is partitioned depends on what assumptions we place on the parameters of the model, and this, in turn, depends on the selective scenario we are interested in [Nordborg (1997)]. For the case of balancing selection, it is appropriate to assume that recombination and mutation are weak forces, and scale these parameters with N. We thus assume that the finite limits

$$R = \lim_{N\to\infty} Nr \tag{3.10}$$

and

$$U_{ij} = \lim_{N\to\infty} Nu_{ij} \tag{3.11}$$

exist. We choose not to scale the selection parameters in the same manner, *cf.* Sections 4.3 and 5.1.

With these assumptions, any transition that implies a recombination or mutation event, or two instances independently picking the same parental individual has probability $O(1/N)$, and transitions that necessitate more than one such event have probability $O(1/N^2)$ or smaller.

3.2. *Partitioning the state space.* We now turn to the state space S_n. It is convenient to partition and arrange the states as follows. First, group the states in n sets by the number of ancestors that remain of the original sample. Arrange the sets in increasing order by the number of ancestors. For example, the first four states will correspond to the four possible configurations for a single instance described in Section 2.2. Since the number of ancestors in the genealogy can only decrease, this arrangement ensures that Π_N is lower block-diagonal.

Second, consider each set with a given number of ancestors, m say, in turn. The m ancestors can be arranged into $m+1$ sets with respect to haplotypic configuration. Arrange the sets in increasing order by the number of ancestors linked to $\mathcal{A}_2$ (the order here is arbitrary).

Third, consider the set with i ancestors linked to $\mathcal{A}_1$ and $j = m-i$ ancestors linked to $\mathcal{A}_2$. This set can be divided into subsets by the number of individuals that harbor two ancestors. Write $\alpha_{i,j}$ for the set of states with all ancestors in distinct individuals, $\beta_{kl,i,j}$ for the set of states with a single $\mathcal{A}_k\mathcal{A}_l$ individual that harbors two ancestors, and $\gamma_{i,j}$ for the remaining states (that have two

or more individuals harboring two ancestors). We have $|\alpha_{i,j}| = (i+1)(j+1)$, $|\beta_{12,i,j}| = ij$,

$$|\beta_{11,i,j}| = \begin{cases} (i-1)(j+1), & \text{if } i > 1, \\ 0, & \text{otherwise,} \end{cases}$$

$$|\beta_{22,i,j}| = \begin{cases} (i+1)(j-1), & \text{if } j > 1, \\ 0, & \text{otherwise,} \end{cases}$$

and a potentially very large number of states in $\gamma_{i,j}$. Arrange these five sets in order $\alpha_{i,j}$, $\beta_{22,i,j}$, $\beta_{12,i,j}$, $\beta_{11,i,j}$, and $\gamma_{i,j}$.

We now turn to the transition probabilities within and between these sets. All transitions within $\alpha_{i,j}$ have probability $O(1)$ because they do not necessitate recombination, mutation, or two ancestors choosing the same parental individual, but simply a transition to another genotypic configuration. A transition to any state outside this set, however, has probability $O(1/N)$ or smaller. The following sets can be reached with probability $O(1/N)$ (*i. e.*, there is at least one transition to the set with that probability):

- whenever $i > 1$, transitions to $\beta_{11,i,j}$ or $\alpha_{i-1,j}$ may occur because two ancestors linked to $\mathcal{A}_1$ pick a common parental individual;
- whenever $j > 1$, transitions to $\beta_{22,i,j}$ or $\alpha_{i,j-1}$ may occur because two ancestors linked to $\mathcal{A}_2$ pick a common parental individual;
- whenever $i > 0$ and $j > 0$, transitions to $\beta_{12,i,j}$ may occur because two ancestors, one linked to $\mathcal{A}_1$, one linked to $\mathcal{A}_2$, pick a common parental individual;
- whenever $i > 0$, transitions to $\alpha_{i-1,j+1}$ may occur through recombination or mutation;
- whenever $j > 0$, transitions to $\alpha_{i+1,j-1}$ may occur through recombination or mutation.

All other sets can only be reached through multiple independent events with probability $O(1/N)$, and therefore have probability $O(1/N^2)$ or smaller.

All transitions within $\beta_{11,i,j}$ have probability $O(1)$, and so do transitions "back to" $\alpha_{i,j}$ (these occur when the individual harboring two ancestors is the product of random mating), and to $\alpha_{i-1,j}$ (which occur with probability one half when the individual harboring two ancestors is the product of selfing). All other transitions have probability $O(1/N)$ or smaller. Exactly the same is true for $\beta_{22,i,j}$, except that a coalescence of course leads to $\alpha_{i,j-1}$ instead of $\alpha_{i-1,j}$.

All transitions within $\beta_{12,i,j}$ have probability $O(1)$, and so do transitions back to $\alpha_{i,j}$. All other transitions have probability $O(1/N)$ or smaller.

We let the transitions from $\gamma_{i,j}$ remain unspecified because they will be shown to be unimportant.

3.3. *Applying Möhle's theorem.* Rewrite the transition matrix $\Pi_N = \mathbf{A} + \mathbf{B}/N + O(1/N^2)$, where $\mathbf{A} = \lim_{N\to\infty} \Pi_N$ and $\mathbf{B} = \lim_{N\to\infty} N(\Pi_N - \mathbf{A})$. From Möhle's (1998) results, it follows that if $\mathbf{P} = \lim_{m\to\infty} \mathbf{A}^m$ exists, then the finite-dimensional distributions of the process converge to those of a continuous-time Markov process with time measured in units of N generations, and infinitesimal generator $\mathbf{G} = \mathbf{PBP}$.

3.3.1. *Finding* $\mathbf{A}$ *and* $\mathbf{P}$. From Section 3.2, it follows that $\mathbf{A}$ has the form

$$\mathbf{A} = \left(\begin{array}{ccccccc|ccccc|cc}
\cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \mathbf{A}_{i,j-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mathbf{A}_{i-1,j} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \mathbf{A}_{i+1,j-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{A}_{i,j} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \mathbf{A}'_{22,\wedge} & 0 & 0 & 0 & 0 & 0 & \mathbf{A}'_{22,\vee} & \mathbf{A}'_{22} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{A}'_{12,\vee} & 0 & \mathbf{A}'_{12} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mathbf{A}'_{11,\wedge} & 0 & 0 & 0 & \mathbf{A}'_{11,\vee} & 0 & 0 & \mathbf{A}'_{11} & 0 & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{A}'_{\gamma} & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{A}_{i-1,j+1} & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right),$$

where $\mathbf{A}_{i,j}$ contains all transitions within $\alpha_{i,j}$, and the $\mathbf{A}'_{\ldots}$ denote the transitions involving states with one or more individuals harboring two ancestors, with the dependence on i and j suppressed to save space. Thus $\mathbf{A}'_{kl}$ contains all transitions within $\beta_{kl,i,j}$, where i and j should be understood to be the ones of the of first $\mathbf{A}_{i,j}$ encountered when following the diagonal towards the upper left corner, and analogously for $\mathbf{A}'_{\gamma}$. The meaning of the off-diagonal elements is evident from their position. The dots represent blocks that may contain non-zero elements. It is clear that $\mathbf{A}$ is stochastic because it can be interpreted as Π_N conditional on no events with probability $O(1/N)$ or lower taking place.

It is easy to show through induction that all blocks of zeros in $\mathbf{A}$ remain blocks of zeros in $\mathbf{A}^m$. Furthermore, the diagonal blocks of $\mathbf{A}^m$ are simply the diagonal blocks of $\mathbf{A}$ raised to the power m. Since $\mathbf{A}_{i,j}$ is a positive stochastic matrix containing the transition probabilities within $\alpha_{i,j}$ conditional on not leaving that set, $\mathbf{P}_{i,j} = \lim_{m\to\infty} \mathbf{A}^m_{i,j}$ is a matrix with all rows identical and equal to the stationary distribution within $\alpha_{i,j}$. The diagonal blocks for states involving one or more individuals harboring two ancestors are even easier: these blocks are all non-negative matrices with modulus less than one and thus vanish as $m \to \infty$.

$\mathbf{A}^m$ also contains two types of sub-diagonal blocks, namely those that correspond to transitions from states with a single individual harboring two ancestors to states with all ancestors in different individuals, and those that correspond to transitions from states with more than one individual harboring two ancestors to states with a single such ancestor. Both types of blocks have the general form

$$\sum_{k=0}^{m-1} \mathbf{Q}_{22}^k \mathbf{Q}_{21} \mathbf{Q}_{11}^{m-1-k}, \tag{3.12}$$

where $\mathbf{Q}_{11}$ stands for the diagonal block containing transitions within the set with the lower number of individuals harboring two ancestors, $\mathbf{Q}_{22}$ stands for the diagonal block containing transitions within the set with the higher number of such individuals, and $\mathbf{Q}_{21}$ contains the transitions from higher to lower. It can be shown that

$$\lim_{m\to\infty} \sum_{k=0}^{m-1} \mathbf{Q}_{22}^k \mathbf{Q}_{21} \mathbf{Q}_{11}^{m-1-k} = (\mathbf{I} - \mathbf{Q}_{22})^{-1} \mathbf{Q}_{21} \mathbf{Q}_{11}^{\infty}. \tag{3.13}$$

Thus, when $\mathbf{Q}_{11}$ stands for a diagonal block of transitions within a set with all ancestors in distinct individuals, $\mathbf{Q}_{11}^{\infty}$ is a matrix with identical rows containing the stationary distribution, as described in the previous paragraph, whereas when $\mathbf{Q}_{11}$ stands for a diagonal block of transitions within a set with a single individual containing two ancestors, we have $\mathbf{Q}_{11}^{\infty} = \mathbf{0}$ so that the right hand side of equation (3.13) vanishes.

We thus have

$$\mathbf{P} = \left(\begin{array}{ccccccc|cccc|cc}
\cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \mathbf{P}_{i,j-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mathbf{P}_{i-1,j} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \mathbf{P}_{i+1,j-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{P}_{i,j} & 0 & 0 & 0 & 0 & 0 \\
0 & \mathbf{P}'_{22,\wedge} & 0 & 0 & 0 & 0 & 0 & \mathbf{P}'_{22,\vee} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{P}'_{12,\vee} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mathbf{P}'_{11,\wedge} & 0 & 0 & 0 & \mathbf{P}'_{11,\vee} & 0 & 0 & 0 & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{P}_{i-1,j+1} & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right),$$

where the blocks are labeled as in $\mathbf{A}$, and the off-diagonal elements are obtained through equation (3.13).

3.3.2. *Finding* $\mathbf{B}$ *and* $\mathbf{G}$. The matrix $\mathbf{B}$ contains the coefficients of the terms $O(1/N)$ from a series expansion of Π_N in N^{-1}. It is neither stochastic nor non-negative. From Section 3.2, it follows that it has the structure

$$\mathbf{B} = \left(\begin{array}{ccccccc|ccccc|cc}
\cdot & \cdot & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\cdot & \mathbf{B}_{i,j-1} & \cdot & \cdot & \cdot & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\cdot & \cdot & \cdot & \mathbf{B}_{i-1,j} & \cdot & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{B}_{i+1,j-1} & \cdot & \cdot & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{0} & \mathbf{0} \\
\hline
\mathbf{0} & \mathbf{B}_{0,-1} & \mathbf{0} & \mathbf{B}_{-1,0} & \mathbf{0} & \mathbf{B}_{+1,-1} & \mathbf{0} & \mathbf{B}_{i,j} & \mathbf{B}_{22} & \mathbf{B}_{12} & \mathbf{B}_{11} & \mathbf{0} & \mathbf{B}_{-1,+1} & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{0} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{0} \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{i-1,j+1} & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right),$$

where the diagonal blocks are denoted as before, and the meaning of the off-diagonal blocks can be inferred from their position in the matrix. Again, the dependence of these blocks on i and j has been suppressed to save space.

Turning to $\mathbf{G} = \mathbf{PBP}$, we first note that the diagonal blocks for all sets that contain one or more individuals harboring two ancestors are zero, as are the blocks containing transitions into such sets from sets with all ancestors in distinct individuals. Since $\mathbf{G}$ is the infinitesimal generator for the continuous-time version of the Markov process describing the genealogy of the sample, this means that, as one would intuitively expect, all such states are instantaneous on a time-scale measured in units of $O(N)$. These states can therefore be eliminated, leaving us with a process involving only the sets $\{\alpha_{i,j}\}_{i,j\in\{1,\dots,m\},i+j=m,m\in\{1,\dots,n\}}$. The blocks of $\mathbf{G}$ that correspond to transitions within and between these sets are as follows.

In general, the rows of $\mathbf{G}$ corresponding to transitions from $\alpha_{i,j}$ has five non-zero blocks. Two of these correspond to a single recombination or mutation event, namely:

$$\mathbf{G}_{\alpha_{i,j},\alpha_{i+1,j-1}} = \mathbf{P}_{i,j}\mathbf{B}_{+1,-1}\mathbf{P}_{i+1,j-1}; \tag{3.14}$$

$$\mathbf{G}_{\alpha_{i,j},\alpha_{i-1,j+1}} = \mathbf{P}_{i,j}\mathbf{B}_{-1,+1}\mathbf{P}_{i-1,j+1}, \tag{3.15}$$

two correspond to single coalescence events, namely:

$$\mathbf{G}_{\alpha_{i,j},\alpha_{i,j-1}} = \mathbf{P}_{i,j}(\mathbf{B}_{0,-1}\mathbf{P}_{i,j-1} + \mathbf{B}_{22}\mathbf{P}'_{22,\wedge}); \tag{3.16}$$

$$\mathbf{G}_{\alpha_{i,j},\alpha_{i-1,j}} = \mathbf{P}_{i,j}(\mathbf{B}_{-1,0}\mathbf{P}_{i-1,j} + \mathbf{B}_{11}\mathbf{P}'_{11,\wedge}), \tag{3.17}$$

and the final one is the diagonal block

$$\mathbf{G}_{\alpha_{i,j},\alpha_{i,j}} = \mathbf{P}_{i,j}(\mathbf{B}_{i,j}\mathbf{P}_{i,j} + \mathbf{B}_{22}\mathbf{P}'_{22,\vee} + \mathbf{B}_{12}\mathbf{P}'_{12,\vee} + \mathbf{B}_{11}\mathbf{P}'_{11,\vee}). \tag{3.18}$$

4. The collapsed process. We have demonstrated that all states with one or more individuals harboring two ancestors are instantaneous and can be eliminated. The same result is obtained for the neutral coalescent with selfing [Nordborg and Donnelly (1997), Möhle (1998)], and it greatly simplifies the process. However, it turns out that further simplifications can be made to the present model, because all transitions within the remaining sets $\{\alpha_{i,j}\}_{i,j\in\{1,\dots,m\},i+j=m}$ ($m \in \{1,\dots,n\}$) have probability $O(1)$, whereas transitions between these sets have probability $O(1/N)$ or smaller. Loosely speaking, the individual states within these sets are instantaneous, but the sets themselves are not, and we would expect the processes governing transitions within the sets to be at stationarity on the time scale on which transitions between the sets occur.

It is evident from $\mathbf{G}$ that this intuition is correct. Notice that each of the blocks given by equations (3.14)–(3.18) is multiplied from the left by $\mathbf{P}_{i,j}$, which, as we have seen, consists of identical rows, each equal to the stationary distribution for $\mathbf{A}_{i,j}$. Thus all blocks of $\mathbf{G}$ also have all rows equal, so that each state within $\alpha_{i,j}$ behaves identically, and the starting condition with respect to these states is irrelevant.

We can therefore simplify the process further by collapsing each $\alpha_{i,j}$ into a single state. This is done by summing over the rows and columns in the appropriate manner, so that in the generator for the collapsed process each block $\mathbf{G}_{\alpha_{i,j},\alpha_{k,l}}$ is replaced by a corresponding *element* $g_{\alpha_{i,j},\alpha_{k,l}}$. Let $\mathbf{p}_{i,j}$ be the stationary distribution for $\mathbf{A}_{i,j}$. The elements corresponding to the blocks (3.14)–(3.18) are

$$g_{\alpha_{i,j},\alpha_{i+1,j-1}} = \mathbf{p}_{i,j}\mathbf{B}_{+1,-1}\mathbf{P}_{i+1,j-1}\mathbf{1}^T, \tag{4.19}$$

$$g_{\alpha_{i,j},\alpha_{i-1,j+1}} = \mathbf{p}_{i,j}\mathbf{B}_{-1,+1}\mathbf{P}_{i-1,j+1}\mathbf{1}^T, \tag{4.20}$$

$$g_{\alpha_{i,j},\alpha_{i,j-1}} = \mathbf{p}_{i,j}(\mathbf{B}_{0,-1}\mathbf{P}_{i,j-1} + \mathbf{B}_{22}\mathbf{P}'_{22,\wedge})\mathbf{1}^T, \tag{4.21}$$

$$g_{\alpha_{i,j},\alpha_{i-1,j}} = \mathbf{p}_{i,j}(\mathbf{B}_{-1,0}\mathbf{P}_{i-1,j} + \mathbf{B}_{11}\mathbf{P}'_{11,\wedge})\mathbf{1}^T, \tag{4.22}$$

$$g_{\alpha_{i,j},\alpha_{i,j}} = \mathbf{p}_{i,j}(\mathbf{B}_{i,j}\mathbf{P}_{i,j} + \mathbf{B}_{22}\mathbf{P}'_{22,\vee} + \mathbf{B}_{12}\mathbf{P}'_{12,\vee} + \mathbf{B}_{11}\mathbf{P}'_{11,\vee})\mathbf{1}^T, \tag{4.23}$$

where $\mathbf{1}$ is a unit vector of appropriate length.

We thus have a new, "collapsed" continuous-time Markov process with states consisting of the number of ancestors and their haplotypic configuration. The size of the state space is that given by expression (2.9). The new process is equivalent to the one previously obtained for random mating [Kaplan et al. (1988), Hudson and Kaplan (1988)], except for the precise values of the non-zero transitions rates [and the possible initial coalescence events for chromosomes sampled from the same individual [Nordborg and Donnelly (1997)]. We now turn to these rates, noting that because the row sums of the generator matrix must be zero, and because of symmetry, it suffices to find $g_{\alpha_{i,j},\alpha_{i+1,j-1}}$ and $g_{\alpha_{i,j},\alpha_{i-1,j}}$, *i. e.*, the rate of recombination/mutation and the rate of coalescence, respectively.

4.1. *The rate of recombination/mutation.* Equation (4.19) turns out to have a simple interpretation. We have

$$g_{\alpha_{i,j},\alpha_{i+1,j-1}} = \mathbf{p}_{i,j}\mathbf{B}_{+1,-1}\mathbf{P}_{i+1,j-1}\mathbf{1}^T = \mathbf{p}_{i,j}\mathbf{B}_{+1,-1}\mathbf{1}^T.$$

Each element in the column vector $\mathbf{B}_{+1,-1}\mathbf{1}^T$ is the sum, over all states in $\alpha_{i+1,j-1}$, of the transition rates from a particular state in $\alpha_{i,j}$ to the states in $\alpha_{i+1,j-1}$. To put it another way, it is the transition rate from $\alpha_{i,j}$ to $\alpha_{i+1,j-1}$ conditional on the process currently being in a particular state in $\alpha_{i,j}$. Multiplication with the stationary distribution over $\alpha_{i,j}$ gives the total transition rate from $\alpha_{i,j}$ to $\alpha_{i+1,j-1}$. Thus, transitions between these sets due to recombination or mutation occur according to the stationary distribution within them, as predicted.

The resulting rate can be found as follows. Consider the set $\alpha_{i,j}$. The states within this set can be described by (k,l), where $k \in \{0,\ldots,i\}$ ($l \in \{0,\ldots,j\}$) is the number of ancestors linked to $\mathcal{A}_1$ ($\mathcal{A}_2$) in heterozygotes. The transition probabilities between these states are those found in $\mathbf{A}_{i,j}$. As explained in Section 2.2, ancestors pick their state in the previous generation independently of one another. In particular, the probability of the transition (k_0,l_0) to (k,l) can be written $\psi_{k_0,k}\omega_{l_0,l}$. Furthermore, we see from equation (2.1) that the probability that a given ancestor, currently linked to $\mathcal{A}_1$ was linked to $\mathcal{A}_2$ in the previous generation but switched because of mutation is

$$\frac{\hat{Y}_2 u_{21}}{\hat{Y}_2 u_{21} + \hat{Y}_1(1-u_{12})} = \frac{\hat{Y}_2 U_{21}}{\hat{Y}_1 N} + O\left(\frac{1}{N^2}\right).$$

Similarly, the probability that a given ancestor switched because of recombination is

$$\frac{\hat{Y}_2 r}{\hat{Y}_2 r + \hat{Y}_1(1-r)} = \frac{\hat{Y}_2 R}{\hat{Y}_1 N} + O\left(\frac{1}{N^2}\right),$$

conditional on the ancestor picking a heterozygous parent (the probability is zero otherwise). The total probability of a transition from $(k_0, l_0) \in \alpha_{i,j}$ to *some* state in $\alpha_{i+1,j-1}$ is therefore

$$\sum_{k=0}^{i}\sum_{l=0}^{j}\frac{\hat{Y}_2}{\hat{Y}_1}\left(\frac{U_{21}}{N}j + \frac{R}{N}l\right)\psi_{k_0,k}\omega_{l_0,l} + O\left(\frac{1}{N^2}\right),$$

and the element of $\mathbf{B}_{+1,-1}\mathbf{1}^T$ that corresponds to a transition from (k_0, l_0) is

$$\begin{aligned}\sum_{k=0}^{i}\sum_{l=0}^{j}\frac{\hat{Y}_2}{\hat{Y}_1}(U_{21}j + Rl)\psi_{k_0,k}\omega_{l_0,l} &= \frac{\hat{Y}_2}{\hat{Y}_1}\sum_{l=0}^{j}(U_{21}j + Rl)\omega_{l_0,l}\\ &= \frac{\hat{Y}_2}{\hat{Y}_1}\left(U_{21}j + R\sum_{l=0}^{j} l\omega_{l_0,l}\right)\\ &= \frac{\hat{Y}_2}{\hat{Y}_1}[U_{21}j + R\mathbb{E}(l|l_0)],\end{aligned}$$

where $\mathbb{E}(l|l_0)$ is the expectation, over the transitions in $\mathbf{A}_{i,j}$, of the number of ancestors linked to $\mathcal{A}_2$ occupying heterozygotes in the previous generation, given that the number is l_0 in the present generation. By multiplying $\mathbf{B}_{+1,-1}\mathbf{1}^T$ with the stationary distribution $\mathbf{p}_{i,j}$, we are in effect calculating the unconditional expectation

$$\mathbb{E}\left(\frac{\hat{Y}_2}{\hat{Y}_1}[U_{21}j + R\mathbb{E}(l|l_0)]\right) = \frac{\hat{Y}_2}{\hat{Y}_1}[U_{21}j + R\mathbb{E}(l)],$$

where l is binomially distributed with parameters j and H_2, the probability that, in the absence of recombination and mutation, an ancestor linked to an $\mathcal{A}_2$ in the present generation occupied a heterozygote in the previous generation (this probability can be calculated exactly from the equations in Section 2.1, but does not have a simple form). Using this, the sought rate becomes

$$g_{\alpha_{i,j},\alpha_{i+1,j-1}} = \left(\frac{\hat{Y}_2}{\hat{Y}_1}U_{21} + H_2R\right)j. \tag{4.24}$$

By symmetry,

$$g_{\alpha_{i,j},\alpha_{i-1,j+1}} = \left(\frac{\hat{Y}_1}{\hat{Y}_2}U_{12} + H_1R\right)i, \tag{4.25}$$

where H_1 is defined analogously to H_2.

4.2. *The rate of coalescence.* From (4.22), and utilizing (3.13), we have

$$\begin{aligned} g_{\alpha_{i,j},\alpha_{i-1,j}} &= \mathbf{p}_{i,j}(\mathbf{B}_{-1,0}\mathbf{P}_{i-1,j} + \mathbf{B}_{11}\mathbf{P}'_{11,\wedge})\mathbf{1}^T, \\ &= \mathbf{p}_{i,j}(\mathbf{B}_{-1,0}\mathbf{P}_{i-1,j} + \mathbf{B}_{11}(\mathbf{I} - \mathbf{A}'_{11})^{-1}\mathbf{A}'_{11,\wedge}\mathbf{P}_{i,j})\mathbf{1}^T \\ &= \mathbf{p}_{i,j}\mathbf{B}_{-1,0}\mathbf{1}^T + \mathbf{p}_{i,j}\mathbf{B}_{11}(\mathbf{I} - \mathbf{A}'_{11})^{-1}\mathbf{A}'_{11,\wedge}\mathbf{1}^T. \end{aligned}$$

Consider first $\mathbf{p}_{i,j}\mathbf{B}_{-1,0}\mathbf{1}^T$. Using the arguments and notation of the previous section, and referring to equation (2.7) and the discussion following it, we note that the total probability of a single-generation transition from $(k_0, l_0) \in \alpha_{i,j}$ to *some* state in $\alpha_{i-1,j}$ can be shown to be

$$\sum_{k=0}^{i}\sum_{l=0}^{j}\left(\frac{\binom{k}{2}}{N\hat{X}_{12}} + \frac{\binom{i-k}{2}}{N\hat{X}_{11}}\frac{1}{2}\right)\psi_{k_0,k}\omega_{l_0,l} + O\left(\frac{1}{N^2}\right). \tag{4.26}$$

The corresponding element of $\mathbf{B}_{-1,0}\mathbf{1}^T$ is thus

$$\begin{aligned} \sum_{k=0}^{i}\sum_{l=0}^{j}\left(\frac{\binom{k}{2}}{\hat{X}_{12}} + \frac{\binom{i-k}{2}}{\hat{X}_{11}}\frac{1}{2}\right)\psi_{k_0,k}\omega_{l_0,l} &= \sum_{k=0}^{i}\left(\frac{\binom{k}{2}}{\hat{X}_{12}} + \frac{\binom{i-k}{2}}{2\hat{X}_{11}}\right)\psi_{k_0,k} \\ &= \mathbb{E}\left(\frac{\binom{k}{2}}{\hat{X}_{12}} + \frac{\binom{i-k}{2}}{2\hat{X}_{11}}\Big|k_0\right). \end{aligned}$$

By the arguments of the previous section, multiplication from the left by the stationary distribution $\mathbf{p}_{i,j}$ is equivalent to calculating the unconditional expectation

$$\mathbb{E}\left(\frac{\binom{k}{2}}{\hat{X}_{12}} + \frac{\binom{i-k}{2}}{2\hat{X}_{11}}\right) = \frac{1}{\hat{X}_{12}}\mathbb{E}\left[\binom{k}{2}\right] + \frac{1}{2\hat{X}_{11}}\mathbb{E}\left[\binom{i-k}{2}\right],$$

where k is binomially distributed with parameters i and H_1. The sought rate is thus

$$\mathbf{p}_{i,j}\mathbf{B}_{-1,0}\mathbf{1}^T = \binom{i}{2}\left(\frac{1}{\hat{X}_{12}}H_1^2 + \frac{1}{2\hat{X}_{11}}(1-H_1)^2\right). \tag{4.27}$$

It remains to deal with $\mathbf{p}_{i,j}\mathbf{B}_{11}(\mathbf{I} - \mathbf{A}'_{11})^{-1}\mathbf{A}'_{11,\wedge}\mathbf{1}^T$. Consider the set $\beta_{11,i,j}$. This set contains all states with i ancestors linked to $\mathcal{A}_1$ and j ancestors linked to $\mathcal{A}_2$ such that that two of the ancestors linked to $\mathcal{A}_1$ jointly occupy a homozygote. The remaining ancestors belong to a set of the form $\alpha_{i-2,j}$. For each transition on this set, and independently of that transition, one of three things may happen to the $\mathcal{A}_1\mathcal{A}_1$ homozygote harboring two ancestors:

1. it may have resulted from outcrossing;
2. it may have resulted from the selfing of a heterozygote;
3. it may have resulted from the selfing of a homozygote.

Ignoring probabilities of $O(1/N)$ and smaller, the following then applies for these alternatives, respectively:

1. we have a transition to a state in $\alpha_{i,j}$;
2. we have a transition to $\alpha_{i-1,j}$ because the ancestors coalesce;
3. with probability one half, we have a transition to $\alpha_{i-1,j}$ because the ancestors coalesce, and with probability one half the process remains in $\beta_{11,i,j}$.

Now consider $A'_{11,\wedge}\mathbf{1}^T$. Each element of this vector is the total probability that a transition from $k_0 \in \beta_{11,i,j}$ is a transition to $\alpha_{i-1,j}$. As we have just seen, this is simply the probability that the $\mathcal{A}_1\mathcal{A}_1$ homozygote was the product of a selfed heterozygote plus one half times the probability that it was the product of a selfed homozygote. Denote the former quantity q_{12} and the latter q_{11}. We thus have

$$\mathbf{A}'_{11,\wedge}\mathbf{1}^T = \mathbf{1}^T(q_{12} + \frac{1}{2}q_{11}).$$

Next consider the matrix A'_{11}. By the above, we must have

$$\mathbf{A}'_{11} = \frac{1}{2}q_{11}A_{i-2,j},$$

and using this it is easy to show that

$$(\mathbf{I} - \mathbf{A}'_{11})^{-1}\mathbf{1}^T = \mathbf{1}^T\frac{1}{1 - \frac{1}{2}q_{11}},$$

so that

$$(\mathbf{I} - \mathbf{A}'_{11})^{-1}\mathbf{A}'_{11,\wedge}\mathbf{1}^T = \mathbf{1}^T\frac{2q_{12} + q_{11}}{2 - q_{11}}.$$

It follows from equation (2.3) that

$$q_{11} = s\frac{\hat{X}_{11}}{\hat{x}_{11}},$$

$$q_{12} = s\frac{\hat{X}_{12}}{4\hat{x}_{11}}.$$

Using this, and repeating the arguments leading to equation (4.27), we finally obtain

$$\mathbf{p}_{i,j}\mathbf{B}_{11}(\mathbf{I}-\mathbf{A}'_{11})^{-1}\mathbf{A}'_{11,\wedge}\mathbf{1}^T = \binom{i}{2}\frac{(1-H_1)^2}{2\hat{X}_{11}}\frac{s\hat{Y}_1}{2\hat{x}_{11}-s\hat{X}_{11}}.$$

Combining this with equation (4.27) leads to

$$g_{\alpha_{i,j},\alpha_{i-1,j}} = \binom{i}{2}\left(\frac{H_1^2}{\hat{X}_{12}} + \frac{(1-H_1)^2}{2\hat{X}_{11}} + \frac{(1-H_1)^2}{2\hat{X}_{11}}\frac{s\hat{Y}_1}{2\hat{x}_{11}-s\hat{X}_{11}}\right). \tag{4.28}$$

By symmetry,

$$g_{\alpha_{i,j},\alpha_{i,j-1}} = \binom{j}{2}\left(\frac{H_2^2}{\hat{X}_{12}} + \frac{(1-H_2)^2}{2\hat{X}_{22}} + \frac{(1-H_2)^2}{2\hat{X}_{22}}\frac{s\hat{Y}_2}{2\hat{x}_{22}-s\hat{X}_{22}}\right). \tag{4.29}$$

4.3. *Weak selection.* The purpose of this section is to show that the rates calculated in Sections 4.1 and 4.2 have very simple and intuitive forms if we assume that selection is weak enough for terms of the order of the selection coefficients to be ignored. Note that this is a statement about the absolute magnitude of the selection coefficients: they are still assumed to be large relative to $1/N$.

Under this approximation, then, we have from equation (2.6) that $\hat{x}_{ij} = \hat{X}_{ij}$. Furthermore, it is a classical result that

$$\hat{X}_{11} = \hat{Y}_1^2 + \hat{Y}_1\hat{Y}_2F,$$
$$\hat{X}_{12} = 2\hat{Y}_1\hat{Y}_2(1-F),$$
$$\hat{X}_{22} = \hat{Y}_2^2 + \hat{Y}_1\hat{Y}_2F,$$

where $F = s/(2-s)$. Using this, it can be shown that

$$H_1 = \hat{Y}_2(1-F),$$
$$H_2 = \hat{Y}_1(1-F),$$

so that the recombination/mutation rates (4.24)–(4.25) become

$$g_{\alpha_{i,j},\alpha_{i+1,j-1}} = \left(\frac{\hat{Y}_2}{\hat{Y}_1}U_{21} + \hat{Y}_2(1-F)R\right)j, \tag{4.30}$$

$$g_{\alpha_{i,j},\alpha_{i-1,j+1}} = \left(\frac{\hat{Y}_1}{\hat{Y}_2}U_{12} + \hat{Y}_1(1-F)R\right)i, \tag{4.31}$$

and the coalescent rates (4.28)–(4.29) become

$$g_{\alpha_{i,j},\alpha_{i-1,j}} = \binom{i}{2}\frac{1}{2\hat{Y}_1}(1+F), \tag{4.32}$$

$$g_{\alpha_{i,j},\alpha_{i,j-1}} = \binom{j}{2}\frac{1}{2\hat{Y}_2}(1+F). \tag{4.33}$$

Thus, under this additional approximation, the coalescent with balancing selection and partial selfing looks almost identical to the coalescent with balancing selection and random mating [Kaplan et al. (1988), Hudsonand Kaplan (1988)], the only difference being that the rate of coalescence within each allelic class is sped up by a factor $1 + F$, and the rate of exchange between allelic classes due to recombination is decreased by a factor $1 - F$. The former factor can be interpreted as the decrease in variance effective population size, and the latter as the decrease in heterozygosity.

5. Discussion. I have demonstrated that essentially all the extra complexity caused by allowing partial selfing in a coalescent model with balancing selection can be removed through a time-scales approximation. The results are interesting from a theoretical as well as from a biological point of view.

5.1. *Theoretical issues.* The results of this paper demonstrate the utility of combining the structured coalescent with time-scales approximations to model complex situations. Further examples are given, albeit with much less detail, in Nordborg (1997).

A few issues related to selection in the coalescent should be commented on. As described in Section 2.3, I have assumed that selection is strong enough for the allele frequencies to be treated as deterministic, so that the coalescent with selection becomes a structured coalescent. Although I have not supplied a formal proof of convergence, this approach seems justified by the fact that the calculations can be carried out with constant selection coefficients even as we let $N \to \infty$. In their original analysis of this problem, Kaplan *et al.* (1988) derived the coalescent conditional on the allele frequencies in all generations,

and assumed that these obeyed a limiting diffusion. However, most results were then obtained assuming that the allele frequencies were "tightly controlled", which is similar to the assumption used in this paper.

Recently, theory has been developed for sample genealogies with "true" selection [Neuhauser and Krone (1997), Krone and Neuhauser (1997)], *i. e.*, without conditioning on the allele frequencies in all generations. It would be very interesting to investigate the limiting behavior of such models as selection becomes stronger. It seems clear that they will converge to structured coalescent models, however, knowing more about the conditions under which they converge could be quite important, because the exact models are considerably more difficult to analyze than the type of model analyzed here. Furthermore, although the exact models yield convenient computational algorithms for simulating samples with selection, the computational time depends exponentially on the strength of selection, whereas simulations using the structured-coalescent approach of course are independent of the strength of selection.

It is thus the view of the present author that the results in this paper, and similar results, should be seen as a "strong-selection limit" for the coalescent with selection. In this context, it is illuminating to consider the simplifications in Section 4.3. Compare equations (4.28) and (4.32). The main reason for the difference in complexity between these two expressions is that when selection is sufficiently weak, the stationary probability that an ancestor occupies a certain state is proportional to the "population size" of that state, and thus inversely proportional to the coalescent rate. As shown in Nordborg (1997), this condition is a generalized version of Nagylaki's (1980) "conservative migration" criterion (by which migration is conservative if it does not effect subpopulation sizes). Nagylaki showed that, in the strong-migration limit, a subdivided population behaves as an unstructured one if and only if migration is conservative, whereas if migration is not conservative, the population will behave as an unstructured population with a lowered variance effective population size. In the present case, we have "strong migration" between the genotypic classes within haplotypic classes, and this is clearly not conservative in general. Under some forms of balancing selection, for instance, we expect heterozygotes to be fitter than homozygotes. If we think of the genotypes as demes, then, forward in time, heterozygotes are net sources of "migrant" gametes, whereas backwards in time, "migrating" ancestors will spend a disproportional amount of time in heterozygotes. This leads to a decreased variance effective population size, and, unfortunately, to rather messy expressions.

Taking the analogy wih population structure and demography further, perhaps some forms of selection will result in a non-linear change of time-scale, just like some forms of variation in population size does? Clearly much work

remains to be done in this area of population genetics theory.

5.2. *Biological issues.* The main biological implication of the work presented here is simply stated: the dynamics of linkage disequilibrium and other forms of allelic associations in partially selfing organisms are governed (to a reasonable approximation) by $Nr(1 - F)$ rather than Nr. This is perhaps not surprising, but may be under-appreciated. Perhaps the most exciting consequence of this result is that the traces of some form of balancing selection may be detectable at a much greater distance from the actual site of selection than in a comparable outcrossing species. This suggests that studies aiming to detect selection in this way should consider using partially selfing organisms: indeed, it may even be possible to scan the genome directly for regions that show traces of balancing selection [Nordborg et al. (1996), Nordborg (1997)].

Acknowledgments. I thank T. Nagylaki and H. Nordborg for the many discussions and explanations that were essential for this paper to come into being, M. Möhle for comments on the manuscript, the American Mathematical Society for funding this conference and subsidizing my attending it, and F. Seillier-Moiseiwitsch for organizing the conference as well as for her patience.

REFERENCES

CHARLESWORTH, B., NORDBORG, M. and CHARLESWORTH, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res., Camb.* **70** 155–174.

HERBOTS, H. M. (1994). "Stochastic Models in Population Genetics: Genealogy and Genetic Differentiation in Structured Populations". PhD thesis, University of London.

HEY, J. (1991). A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theor. Pop. Biol.* **39** 30–48.

HUDSON, R. R. (1996). Molecular population genetics of adaptation. In *Adaptation*, M. R. Rose and G. V. Lauder, eds, pp. 291–309, Academic Press, San Diego.

HUDSON, R. R. and KAPLAN, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120** 831–840.

KAPLAN, N. L., DARDEN, T. and HUDSON, R. R. (1988). The coalescent process in models with selection. *Genetics* **120** 819–829.

KINGMAN, J. F. C. (1982). The coalescent. *Stochast. Proc. Appl.* **13** 235–248.

KREITMAN, M. and AKASHI, H. (1995). Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.* **26** 403–422.

KRONE, S. M. and NEUHAUSER, C. (1997). Ancestral processes with selection. *Theor. Pop. Biol.* **51** 210–237.

MÖHLE, M. (1998). A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Adv. Appl. Prob.*, to appear (available via http://www.mathematik.uni-mainz.de/Stochastik/Arbeitsgruppe/moehle.html).

NAGYLAKI, T. (1980). The strong-migration limit in geographically structured populations. *J. Math. Biol.* **9** 101–114.

NEUHAUSER, C. and KRONE, S. M. (1997). The genealogy of samples in models with selection. *Genetics* **145** 519–534.

NORDBORG, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146** 1501–1514.

NORDBORG, M., CHARLESWORTH, B. and CHARLESWORTH, D. (1996). Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. *Proc. R. Soc. Lond.* B **263** 1033–1039.

NORDBORG, M. and DONNELLY, P. (1997). The coalescent process with selfing. *Genetics* **146** 1185–1195.

NOTOHARA, M. (1990). The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.* **29** 59–75.

DEPARTMENT OF GENETICS
LUND UNIVERSITY
SÖLVEGATAN 29
223 62 LUND, SWEDEN
MAGNUS.NORDBORG@GEN.LU.SE