J. Math. Biol. 47, 337–352 (2003)
Digital Object Identifier (DOI):
10.1007/s00285-003-0218-6

Mathematical Biology

Martin Möhle · Serik Sagitov

Coalescent patterns in diploid exchangeable population models

Received: 10 March 2003 / Revised version: 18 March 2003 /
Published online: 15 May 2003 –

Abstract. A class of two-sex population models is considered with N females and equal number N of males constituting each generation. Reproduction is assumed to undergo three stages: 1) random mating, 2) exchangeable reproduction, 3) random sex assignment. Treating individuals as pairs of genes at a certain locus we introduce the diploid ancestral process (the past genealogical tree) for n such genes sampled in the current generation. Neither mutation nor selection are assumed. A convergence criterium for the diploid ancestral process is proved as N goes to infinity while n remains unchanged. Conditions are specified when the limiting process (coalescent) is the Kingman coalescent and situations are discussed when the coalescent allows for multiple mergers of ancestral lines.

1. Introduction

The celebrated Wright-Fisher model describes an asexual (haploid) population with non-overlapping generations and fixed population size N. The reproduction law is defined by specifying the joint distribution of numbers of offspring $(\nu_1, \ldots, \nu_N)$ as the symmetric multinomial distribution $\mathrm{Mn}(N, 1/N, \ldots, 1/N)$. For this model the past genealogical tree of n sampled individuals converges as N tends to infinity to the Kingman coalescent (Kingman, 1982a,b,c). This asymptotic result says, in particular, that in a large Wright-Fisher population the sampled ancestral lines merge only pairwise when followed backward in time.

A similar convergence result holds (Möhle, 1998b) for the two-sex Wright-Fisher population model based on the following assumptions.

a) Each generation consists of $2N$ individuals, N females and N males,
b) N couples are formed by random mating, and the couple i produces d_i daughters and s_i sons, $i \in \{1, \ldots, N\}$, and

M. Möhle: Eberhard Karls University of Tübingen, Mathematics Institute, 72076 Tübingen, Germany. e-mail: martin.moehle@uni-tuebingen.de

S. Sagitov: Chalmers University of Technology and Göteborg University, School of Mathematical and Computing Sciences, S-41296 Göteborg, Sweden.
e-mail: serik@math.chalmers.se

Work supported by the Bank of Sweden Tercentenary Foundation.

Mathematics Subject Classification (2000): Primary 92F25, 60J70; Secondary 92D15, 60F17

Key words or phrases: Ancestral process – Coalescent – Diploid model – Exchangeability – Generator – Neutrality – Population genetics – Two-sex model – Weak convergence

c) the two offspring vectors $d := (d_1, \ldots, d_N)$ and $s := (s_1, \ldots, s_N)$ are independent and they have both the symmetric multinomial distribution $\mathrm{Mn}(N, 1/N, \ldots, 1/N)$.

Let $\nu_i := d_i + s_i$ denote the total number of offspring of the couple i. Then the offspring vector $\nu := (\nu_1, \ldots, \nu_N)$ has the multinomial distribution $\mathrm{Mn}(2N, 1/N, \ldots, 1/N)$. Once the d_i and s_i are given, sex assignment for the children within a generation is assumed to be performed at random under the restriction that the total number of daughters is equal to the total number of sons (and hence equal to N).

In an extension of the haploid Wright-Fisher model, introduced by Cannings (1974, 1975), the symmetric multinomial distribution $\mathrm{Mn}(N, 1/N, \ldots, 1/N)$ is replaced by an arbitrary exchangeable joint distribution for the offspring sizes $(\nu_1, \ldots, \nu_N)$ satisfying the restriction $\nu_1 + \cdots + \nu_N = N$. The analysis of the ancestral trees for this class of population models, started by Kingman (1982a,b,c), has led to more general coalescent processes allowing for multiple mergers (Sagitov, 1999) and simultaneous multiple mergers (Möhle and Sagitov, 2001) of ancestral lines.

The coalescent has been proven to be an appropriate process to analyze the ancestral history of a sample of particles, individuals, genes or DNA-sequences chosen from a large haploid population. Coalescent theory is widely used to estimate biological population parameters, for example effective population sizes or migration-, mutation- and recombination-rates (Beerlin and Felsenstein, 1999; Griffiths and Tavaré, 1994; Stephens and Donnelly, 2000). Furthermore, the coalescent provides a basis to test the hypothesis of selective neutrality against either balancing selection or the presence of advantageous alleles (Fu, 1996; Slatkin, 1994, 1996; Tajima, 1989).

Kingman, (1982a) already mentioned that it would be of great interest to seek a comparable analysis of truly diploid genealogies. In the following we analyze the ancestry of a class of diploid population models and we verify results on convergence to the coalescent. These results hence justify the usage of coalescent theory for diploid genealogies.

Combining the two features, gender and exchangeable offspring distribution, we consider a class of two-sex population models retaining three properties from the two-sex Wright-Fisher model: random mating, exchangeable joint distribution for the offspring sizes $(\nu_1, \ldots, \nu_N)$ such that

$$\nu_1 + \cdots + \nu_N = 2N, \tag{1}$$

and random sex assignment.

The key idea leading to an appropriate diploid coalescent is to consider an additional genetic level. We think of individuals as pairs of genes at a certain locus. Our focus is on the past genealogical tree (ancestral process) of n genes sampled from the current generation.

In Section 2 we recall Kingman's definition of the haploid ancestral process $(\mathcal{R}_r)_{r\in\mathbb{N}_0}$ as a Markov chain where the states are equivalence relations for n sampled genes. We introduce the diploid ancestral process $(\tilde{\mathcal{R}}_r)_{r\in\mathbb{N}_0}$ as a Markov chain where the states are extended *three-level* equivalence relations incorporating the

information that genes belong to individuals and individuals form couples. The one-step transition probabilities for the diploid ancestral process are calculated in Section 3.

Section 4 contains our main theorem on convergence to the diploid coalescent. An important corollary of this main result gives a criterion for the weak convergence to the Kingman coalescent (Section 5). The necessary and sufficient condition of this criterion uses only the second and third moments of the marginal offspring distribution:

$$\lim_{N\to\infty} \frac{\mathrm{E}((\nu_1)_3)}{N\,\mathrm{E}((\nu_1)_2)} \;=\; 0, \tag{2}$$

where $(\nu_1)_k := \nu_1(\nu_1-1)\cdots(\nu_1-k+1)$.

In Section 6 we discuss diploid coalescent patterns allowing for multiple mergers of ancestral lines. Loosely speaking, multiple mergers of ancestral lines are only possible when large families (with the number ν_i of children of order N) occur sufficiently often. Examples of real populations whose past genealogical trees might have multiple mergers of ancestral lines are fish populations, populations with dominant males or populations with artificial insemination.

Finally, Section 7 comments on a number of related papers.

2. The diploid ancestral process

We start this section with the forward description of the population models under consideration and proceed with the backward structure of the gene inheritance process. Afterwards we introduce the ancestral process $\mathcal{R}$ as well as the diploid ancestral process $\tilde{\mathcal{R}}$ which is a convenient tool to analyze the ancestral process.

We consider a class of two-sex population models with random mating, with exchangeable joint distribution for the offspring sizes $(\nu_1,\dots,\nu_N)$ satisfying (1), and with random sex assignment, conditioned that the number of sons and the number of daughters in each generation is equal to N. To be precise, this conditional random sex assignment means that the joint distribution of the daughter and son offspring vectors (d,s) satisfies

$$P(d=k, s=l) \;=\; \frac{\binom{k_1+l_1}{k_1}\cdots\binom{k_N+l_N}{k_N}}{\binom{2N}{N}}\,P(\nu=k+l),$$

where $k=(k_1,\dots,k_N)$, $l=(l_1,\dots,l_N)\in\{0,\dots,N\}^N$. In particular,

$$P(d_1+\cdots+d_j=a, s_1+\cdots+s_j=b) \;=\; \frac{\binom{a+b}{a}\binom{2N-a-b}{N-a}}{\binom{2N}{N}}\,P(\nu_1+\cdots+\nu_j=a+b)$$

for all $j\in\{1,\dots,N\}$ and $a,b\in\{0,\dots,N\}$. Hence $d_1+\cdots+d_N=s_1+\cdots+s_N=N$ almost surely, i.e., each generation consists of exactly N daughters and N sons almost surely.

We view individuals as *pairs of genes* (since we are in the one-locus case) so that the corresponding backward dynamics of the gene inheritance process is described as a three-level combinatorial experiment: first, on the generation level,

individuals choose parent couples according to the distribution of the offspring sizes $(\nu_1, \ldots, \nu_N)$; then, on the family level, genes make a mother-or-father choice at random; finally, on the individual level, genes choose at random one of the two available gene copies.

Following Kingman (1982a,b,c) we introduce the *ancestral process* as a chain $\mathcal{R} = (\mathcal{R}_r)_{r\in\mathbb{N}_0}$ with the state space

$$\mathcal{E}_n \;=\; \text{the set of all equivalence relations on } \{1, \ldots, n\}.$$

Here n stands for the number of genes sampled from the current generation and $\mathcal{R}_r$ is the equivalence relation such that $i \sim j$ if and only if the genes i and j in the sample have a common ancestor r generations backward in time.

The process $\mathcal{R}$ is a mathematical description of the past genealogical tree tracing the ancestral lines of the sampled genes. To visualize such a tree we can think of the sampled genes and their ancestors as being marked. Obviously the graph linking the marked children to the marked parents has a tree structure.

Observe, that the process $\mathcal{R}$ is not necessarily a Markov chain in the diploid case. The Markov property can be recovered by grouping the marked genes on the individual level as follows. Consider a state

$$\xi \;=\; \{C_1, \ldots, C_\beta\}$$

of the process $\mathcal{R}$, where the C_i's are the equivalence classes representing β $(\leq n)$ marked genes. Suppose that the ordering of these genes is such that the genes $\{C_{2i-1}, C_{2i}\}$ belong to the same individuals for all $1 \leq i \leq \beta - b$ and the rest of the marked genes stand alone on the individual level. The notation

$$\hat{C}_i \;:=\; \begin{cases} \{C_{2i-1}, C_{2i}\} \text{ for } 1 \leq i \leq \beta - b, \\ \{C_{\beta-b+i}\} \quad \text{for } \beta - b + 1 \leq i \leq b, \end{cases} \tag{3}$$

will be used for representing the *marked individuals* – individuals hosting at least one marked gene. Here $\beta - b$ is the number of *2-marked individuals*, hosting two marked genes, and $2b - \beta$ is the number of *1-marked individuals*, hosting one marked gene. Going from the state ξ to the enriched version

$$\hat{\xi} \;=\; \{\hat{C}_1, \ldots, \hat{C}_b\}$$

we arrive at a process $\hat{\mathcal{R}} = (\hat{\mathcal{R}}_r)_{r\in\mathbb{N}_0}$ which can be viewed as the ancestral process on the individual level.

Due to the random mating feature of the model the process $\hat{\mathcal{R}}$ is a Markov chain. However, it turns out that it is more convenient to work with a more detailed Markov chain $\tilde{\mathcal{R}} = (\tilde{\mathcal{R}}_r)_{r\in\mathbb{N}_0}$ built on the level of couples, which we call the *diploid ancestral process*. Using the set $\hat{\xi}$ of the marked individuals $\hat{C}_i$ we introduce the set

$$\tilde{\xi} \;:=\; \{\tilde{C}_1, \ldots, \tilde{C}_B\},$$

of *marked couples* – couples with at least one marked individual involved:

$$\tilde{C}_i \;:=\; \begin{cases} \{\hat{C}_{2i-1}, \hat{C}_{2i}\} \text{ for } 1 \leq i \leq b - B, \\ \{\hat{C}_{b-B+i}\} \quad \text{for } b - B + 1 \leq i \leq B \end{cases} \tag{4}$$

(Note that the diploid ancestral process does not distinguish between males and females.)

In order to indicate that a state $\tilde{\xi}$ has a structure corresponding to B marked couples, b marked individuals and β marked genes we write $\tilde{\xi} \sim (B, b, \beta)$. In the important particular case where $\tilde{\xi} \sim (\beta, \beta, \beta)$ we write $\tilde{\xi} = \xi$ (abusing somewhat notation rules) and call such states *simple states*.

3. One-step transition probabilities

We illustrate the transition mechanism of the diploid ancestral process with Figures 1 and 2. In the framework of the diploid population model Figure 1 depicts an example of the one-step transition graph drawn on a mixed level: on the parent side we have $A = 4$ couples and on the offspring side we have $b = 7$ individuals. The corresponding transition graph on the gene level is a collection of A (maximal connected) family subgraphs which contain all necessary information about gene inheritance. Each subgraph (indexed by i) corresponds to an parental couple.

Each family subgraph for an parental couple (see Figure 2) is specified by six parameters b_i, β_i, β_{i1}, β_{i2}, β_{i3}, β_{i4}, where b_i is the number of marked offspring individuals of the couple, β_i is the number of marked offspring genes, while β_{ij}, $j \in \{1, 2, 3, 4\}$, denotes the number of edges coming from the parental gene j. It is assumed that the parental genes 1 and 2 belong to the first parental individual and

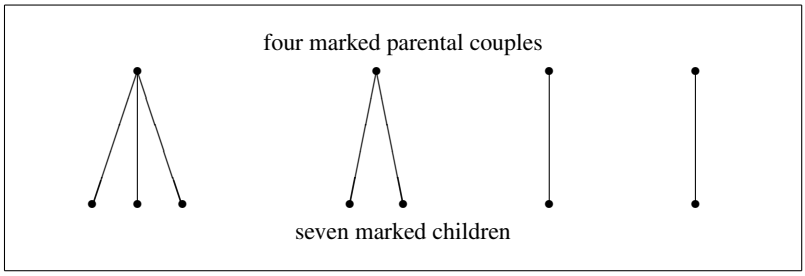

Fig. 1. Example of a transition graph in the haploid case with $A = 4, b = 7, b_1 = 3, b_2 = 2, b_3 = 1, b_4 = 1$.

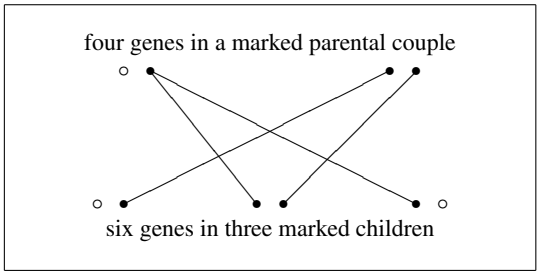

Fig. 2. Example of a family subgraph with $b_i = 3$, $\beta_i = 4$, $\beta_{i1} = 0$, $\beta_{i2} = 2$, $\beta_{i3} = 1$, $\beta_{i4} = 1$.

that the genes 3 and 4 belong to the other parental individual. Obviously

$$\beta_{i1} + \beta_{i2} + \beta_{i3} + \beta_{i4} = \beta_i, \qquad i \in \{1, \dots, A\}, \tag{5}$$

and

$$b_1 + \cdots + b_A = b, \qquad \beta_1 + \cdots + \beta_A = \beta. \tag{6}$$

Using these parameters we are able to specify the number of the marked genes after the transition as

$$\sum_{i=1}^{A} \sum_{j=1}^{4} I(\beta_{ij} \neq 0) \; = \; \alpha, \tag{7}$$

as well as the number of marked individuals after the transition as

$$\sum_{i=1}^{A} (I(\beta_{i1} + \beta_{i2} \neq 0) + I(\beta_{i3} + \beta_{i4} \neq 0)) \; = \; a. \tag{8}$$

Note that if $b = A$ holds then $\alpha = a = \beta$ due to the fact that individuals inherit their two genes from two different sources.

Finally, a family subgraph might embrace up to four (possibly multiple) mergers of gene ancestral lines, so that the total number of mergers of marked ancestral lines equals $\sum_{i=1}^{A} \sum_{j=1}^{4} I(\beta_{ij} \geq 2)$.

We write $\tilde{\xi} \prec \tilde{\eta}$ to indicate that the transition $\tilde{\xi} \to \tilde{\eta}$ in the diploid ancestral process is possible and we denote the one-step transition probability for the diploid ancestral process defined in the previous section by

$$\tilde{p}_{\tilde{\xi}\tilde{\eta}} := P(\tilde{\mathcal{R}}_{r+1} = \tilde{\eta} | \tilde{\mathcal{R}}_r = \tilde{\xi}).$$

Lemma 3.1. *If $\tilde{\xi} \prec \tilde{\eta}$, $\tilde{\xi} \sim (B, b, \beta)$, $\tilde{\eta} \sim (A, a, \alpha)$ and the transition $\tilde{\xi} \to \tilde{\eta}$ is characterized by the decompositions (5), (6), then*

$$\tilde{p}_{\tilde{\xi}\tilde{\eta}} \; = \; 2^{A+a-\beta-b} \frac{(N)_A}{(2N)_b} \mathrm{E}((\nu_1)_{b_1} \cdots (\nu_A)_{b_A}). \tag{9}$$

Proof. The proof of (9) is based on the backward structure of the gene inheritance process presented in Section 2. During the proof a parental couple with j children is viewed as a box with j cells, and genes are called balls.

On the generation level $2N$ individuals are allocated at random among $2N$ cells which are distributed in N boxes. If an individual belongs to the box i this means that the individual is a child of the couple i, i.e. it belongs to the i-th family. Assume that b of the $2N$ individuals are colored in A different ways so that b_i objects have color i, $i \in \{1, \dots, A\}$. For a favorable allocation individuals of the same color are in the same box and no box contains two individuals of different color. The probability of a favorable outcome, conditional that the box i has ν_i cells, $i \in \{1, \dots, N\}$, is

$$\sum_{\substack{k_1, \dots, k_A = 1 \\ \text{all distinct}}}^{N} \frac{(\nu_{k_1})_{b_1} \cdots (\nu_{k_A})_{b_A}}{(2N)_b}.$$

Taking the expectation and using exchangeability we arrive at the second factor in (9)

$$\frac{(N)_A}{(2N)_b}\mathrm{E}((\nu_1)_{b_1}\cdots(\nu_A)_{b_A}).$$

On the family level (mother or father choice) it is crucial to distinguish two types of individuals: $\beta - b$ individuals are pairs of balls glued together (2-marked individuals) and $2b - \beta$ individuals are single balls (1-marked individuals). There are A independent experiments performed on this level corresponding to the A subgraphs, indexed by $i \in \{1, \ldots, A\}$. We describe the i-th experiment. Consider a box containing $\beta_i - b_i$ single balls and $2b_i - \beta_i$ glued balls, corresponding to b_i individuals of color i. A single ball chooses one of the two individuals of the parental couple i at random, glued balls are separated and choose different parental individuals of the parental couple i at random. Every ball is labelled with a two-digit number, where each digit is either 1 or 2. At this stage we are interested in the outcome when balls with the same first digit choose the same parental individual. For the i-th experiment the probability in question is $2 \cdot 2^{-b_i}$, so that the overall contribution of the second level experiment is (cf. (6))

$$\prod_{i=1}^{A}(2 \cdot 2^{-b_i}) \;=\; 2^{A-b}.$$

Finally, on the third (gene choice) level, the balls which have chosen the same parental individual, are separated in two groups at random. The probability that all balls with the same second digit end up in the same group equals (cf. (5), (6), (8))

$$\prod_{i=1}^{A}(2^{I(\beta_{i1}+\beta_{i2}\neq 0)}2^{-\beta_{i1}-\beta_{i2}} \cdot 2^{I(\beta_{i3}+\beta_{i4}\neq 0)}2^{-\beta_{i3}-\beta_{i4}}) \;=\; 2^{a-\beta}.$$

To finish the proof of (9) it remains to multiply the three probabilities found so far. □

Example. In the two-sex Wright-Fisher model the joint distribution of the offspring sizes $(\nu_1, \ldots, \nu_N)$ is $\mathrm{Mn}(2N, \frac{1}{N}, \ldots, \frac{1}{N})$ and therefore

$$\mathrm{E}(s_1^{\nu_1}\cdots s_A^{\nu_A}) \;=\; \left(\frac{s_1 + \cdots + s_A}{N} + \frac{N-A}{N}\right)^{2N}.$$

Hence $\mathrm{E}((\nu_1)_{b_1}\cdots(\nu_A)_{b_A}) = (2N)_b/N^b$ and according to (9)

$$\tilde{p}_{\tilde{\xi}\tilde{\eta}} \;=\; 2^{A+a-\beta-b}\frac{(N)_A}{N^b}.$$

For simple states $\tilde{\xi}$ and $\tilde{\eta}$ ($\alpha = a = A$ and $\beta = b = B$) we obtain

$$\tilde{p}_{\tilde{\xi}\tilde{\eta}} \;=\; \frac{4^\alpha (N)_\alpha}{(4N)^\beta}.$$

This expression is asymptotically equal to $(4N)_\alpha/(4N)^\beta$ which corresponds to the transition probability of the ancestral process in the haploid Wright-Fisher model with population size $4N$.

4. General limit theorem

The main Theorem 4.2 is a weak convergence result for the diploid ancestral process as the population size N tends to infinity. The time scale c_N used in this result is defined as the coalescence probability – the probability that two genes, chosen randomly without replacement from different individuals of the same generation, have the same parental gene one generation backward in time. According to (9), with $a = A = 1$ and $\beta = b = 2$, we have

$$c_N \;=\; \frac{N}{4(2N)_2}\mathrm{E}((\nu_1)_2) \;\sim\; \frac{\mathrm{E}((\nu_1)_2)}{16N},\; N\to\infty.$$

Note that $\mathrm{E}((\nu_1)_2) \geq (\mathrm{E}(\nu_1))_2 = 2$ and hence $c_N > 0$. We need a condition saying that the limits

$$\lim_{N\to\infty} \frac{N^j \mathrm{E}((\nu_1)_{k_1}\cdots(\nu_j)_{k_j})}{(2N)^{k_1+\cdots+k_j} c_N} \;=\; \phi_j(k_1,\ldots,k_j),\; k_1\geq\cdots\geq k_j\geq 2 \tag{10}$$

exist for all $j\in\mathbb{N}$. This condition is an adjusted version of a similar assumption for the haploid case from Möhle and Sagitov, (2001). By repeating the argument from Möhle and Sagitov, (2001) it is shown that this condition implies the existence of the limits in (10) also for the wider set of parameters $k_1,\ldots,k_j\in\mathbb{N}$, $j\in\mathbb{N}$ satisfying $k_1+\cdots+k_j > j$. Furthermore, it is straightforward to verify that (10) implies also the existence of the limits

$$\gamma_j \;:=\; \lim_{N\to\infty} \frac{1-\frac{(N)_j}{(2N)_j}\mathrm{E}(\nu_1\cdots\nu_j)}{c_N},\quad j\geq 1, \tag{11}$$

which can be expressed in terms of the functions ϕ_i as

$$\gamma_j \;=\; \sum_{i=1}^{j-1} i\cdot\phi_i(2,1,\ldots,1). \tag{12}$$

To state the theorem we need to introduce four $(\tilde{\mathcal{E}}_n\times\tilde{\mathcal{E}}_n)$-matrices. Write $\tilde{P}_N$ for the one-step transition matrix with the elements $\tilde{p}_{\tilde{\xi}\tilde{\eta}}$. Let $\tilde{J}$ be the matrix of the indicators $1_{\{\tilde{\xi}\prec\tilde{\eta},b=A\}}$. Put $\tilde{H} := \tilde{J}^2$ and observe that the elements of the matrix $\tilde{H}$ are the indicators $1_{\{\tilde{\eta}=\eta=\xi\}}$. Let the entries of the matrix $\tilde{G}$ be defined by the formula

$$\tilde{g}_{\tilde{\xi}\tilde{\eta}} \;:=\; \begin{cases} -\gamma_b & \text{if } \tilde{\xi}\prec\tilde{\eta}, b=A,\\ 2^{A+a-\beta-b}\phi_A(b_1,\ldots,b_A) & \text{if } \tilde{\xi}\prec\tilde{\eta}, b>A,\\ 0 & \text{otherwise.}\end{cases} \tag{13}$$

Theorem 4.2. *If the limits (10) exist for all $j\in\mathbb{N}$, and $c_N\to 0$ as $N\to\infty$, then*

$$\tilde{P}_N = \tilde{J} + c_N\tilde{G} + o(c_N),\; N\to\infty, \tag{14}$$

and, moreover, the weak convergence

$$(\tilde{\mathcal{R}}_{[t/c_N]})_{t\geq 0}\to(\tilde{R}_t)_{t\geq 0},\; N\to\infty \tag{15}$$

holds in the Skorohod sense, where the limit process $(\tilde{R}_t)_{t\geq 0}$ is a continuous time Markov chain with the transition matrix $\tilde{H}e^{t\tilde{Q}}$ with the generator $\tilde{Q} := \tilde{H}\tilde{G}\tilde{H}$.

Conversely, if $c_N \to 0$ and if (14) holds for some $\tilde{G}$ then the limits (10) exist for all $j \in \mathbb{N}$ and all $k_1, \ldots, k_j \in \mathbb{N}$ with $k_1 + \cdots + k_j > j$ and the entries of $\tilde{G}$ are of the form (13).

Proof. According to (9), (10), (11)

$$\tilde{p}_{\tilde{\xi}\tilde{\eta}} \to 2^{A+a-\beta-b} I(\tilde{\xi} \prec \tilde{\eta}, b = A) = I(\tilde{\xi} \prec \tilde{\eta}, b = A), \ N \to \infty, \qquad (16)$$

and moreover

$$\frac{\tilde{p}_{\tilde{\xi}\tilde{\eta}}}{c_N} \to 2^{A+a-\beta-b}\phi_A(b_1, \ldots, b_A), \ b > A, \ \tilde{\xi} \prec \tilde{\eta},$$

and

$$\frac{1 - \tilde{p}_{\tilde{\xi}\tilde{\eta}}}{c_N} \to \gamma_b, \ b = A, \ \tilde{\xi} \prec \tilde{\eta},$$

as N tends to infinity. Since $\tilde{p}_{\tilde{\xi}\tilde{\eta}} = 0$ unless $\tilde{\xi} \prec \tilde{\eta}$, we conclude that the asymptotic formula (14) holds with the matrix $\tilde{G}$ specified by (13).

Using Theorem 2.2 from Möhle (1998a) we derive from (14) that the limiting Markov chain has the transition matrix $\tilde{H}e^{t\tilde{Q}}$ with the generator $\tilde{Q} = \tilde{H}\tilde{G}\tilde{H}$. Finally, the weak convergence in the Skorohod sense is obtained as in Möhle (1999a).

Conversely, assume now that (14) holds. Then the relation (9) ensures that the limits (10) exist for all $j \in \mathbb{N}$ and all $k_1, \ldots, k_j \in \mathbb{N}$ with $k_1 + \cdots + k_j > j$. It remains to apply the first part of the Theorem (which is already proven) to finish the proof of the Theorem. □

Remark. The special structure of the projection $\tilde{H}$ implies that the entries $\tilde{q}_{\tilde{\xi}\tilde{\eta}}$ of the generator $\tilde{Q} = \tilde{H}\tilde{G}\tilde{H}$ have the form

$$\tilde{q}_{\tilde{\xi}\tilde{\eta}} = \sum_{\tilde{x},\tilde{y}} 1_{\{\tilde{x}=x=\xi\}}\tilde{g}_{\tilde{x}\tilde{y}}1_{\{\tilde{\eta}=\eta=y\}} = \sum_{\tilde{y}} \tilde{g}_{\xi\tilde{y}}1_{\{\tilde{\eta}=\eta\}}1_{\{y=\eta\}} = 1_{\{\tilde{\eta}=\eta\}} \sum_{\tilde{y}:y=\eta} \tilde{g}_{\xi\tilde{y}}. \qquad (17)$$

This special form of the generator shows that the limiting process jumps instantaneously from a state $\tilde{\xi} \in \tilde{\mathcal{E}}_n$ to the simple state ξ. To explain this effect note that due to (14) we have $\tilde{P}_N \to \tilde{J}$, where $\tilde{J}$ is a stochastic matrix. The Markov chain with the one-step transition matrix $\tilde{J}$ transforms every non-simple state to its simple counterpart in two steps ($\tilde{H} = \tilde{J}^2$). The following corollary concerning the ancestral process $(\mathcal{R}_{[t/c_N]})_{t\geq 0}$ follows from Theorem 4.2.

Corollary 4.3. *Under the condition (10) the time-scaled ancestral process $(\mathcal{R}_{[t/c_N]})_{t\geq 0}$ converges weakly to the diploid coalescent process $(R_t)_{t\geq 0}$ which is a continuous time Markov chain with the generator Q with the entries*

$$q_{\xi\eta} = \sum_{\tilde{y}:y=\eta} \tilde{g}_{\xi\tilde{y}}, \qquad \xi, \eta \in \mathcal{E}_n.$$

5. Convergence to the Kingman coalescent

In this section we use Theorem 4.2 to verify that the process $(\tilde{R}_t)_{t\geq 0}$ is a diploid version of the Kingman coalescent if and only if the limits (10) are equal to zero whenever $k_1+\cdots+k_j > j+1$. In this case the entries (13) of the matrix $\tilde{G}$ have the simpler form

$$\tilde{g}_{\tilde{\xi}\tilde{\eta}} = \begin{cases} -4\cdot\binom{b}{2} & \text{if } \tilde{\xi}\prec\tilde{\eta}, b=A,\\ 2^{1+a-\beta} & \text{if } \tilde{\xi}\prec\tilde{\eta}, b=A+1,\\ 0 & \text{otherwise.}\end{cases} \tag{18}$$

The entries of the corresponding generator $\tilde{Q}$ satisfy the formula

$$\tilde{q}_{\tilde{\xi}\tilde{\eta}} = \begin{cases} -\binom{\beta}{2} & \text{if } \xi=\eta=\tilde{\eta},\\ 1 & \text{if } \tilde{\xi}\prec\tilde{\eta}, \tilde{\eta}=\eta \text{ and } \beta=\alpha+1,\\ 0 & \text{otherwise.}\end{cases} \tag{19}$$

According to Corollary 4.3 the corresponding diploid coalescent $(R_t)_{t\geq 0}$ has intensities of the Kingman coalescent:

$$q_{\xi\eta} := \begin{cases} -\binom{\beta}{2} & \text{if } \xi=\eta,\\ 1 & \text{if } \xi\subseteq\eta \text{ and } \beta=\alpha+1,\\ 0 & \text{otherwise,}\end{cases} \qquad \xi,\eta\in\mathcal{E}_n. \tag{20}$$

We verify the formula (19) using (17) with $\tilde{\xi}\sim(B,b,\beta)$ and $\tilde{\eta}\sim(A,a,\alpha)$. If $\tilde{\xi}\prec\tilde{\eta}$, $\beta=A+1$, then the formulas (17) and (18) give $\beta=\alpha+1$ and

$$\tilde{q}_{\tilde{\xi}\tilde{\eta}} = \sum_{\tilde{y}:y=\eta}\tilde{g}_{\xi\tilde{y}}1_{\{\tilde{\eta}=\eta\}} = \tilde{g}_{\xi\tilde{\eta}}1_{\{\tilde{\eta}=\eta\}} = 2^{1+a-\beta}1_{\{\tilde{\eta}=\eta\}} = 1_{\{\tilde{\eta}=\eta\}}$$

since $\xi\sim(\beta,\beta,\beta)$, $\eta\sim(\alpha,\alpha,\alpha)$ and only the term with $\tilde{y}=\eta$ contributes to the sum. Thus, the second line of (19) is proved. The first line of (19) is obtained as follows. If $\tilde{\xi}\prec\tilde{\eta}$ and $\xi=\eta=\tilde{\eta}$, then $\alpha=\beta$ and according to (18)

$$\begin{aligned}\tilde{q}_{\tilde{\xi}\tilde{\eta}} = \sum_{\tilde{x}\sim(B',b',\beta)}\tilde{g}_{\xi\tilde{x}} &= \tilde{g}_{\xi\xi}+\sum_{\tilde{x}\sim(\beta-1,\beta-1,\beta)}\tilde{g}_{\xi\tilde{x}}+\sum_{\tilde{x}\sim(\beta-1,\beta,\beta)}\tilde{g}_{\xi\tilde{x}}\\ &= -4\tbinom{\beta}{2}+\tbinom{\beta}{2}\cdot 2^{1+\beta-1-\beta}+\tbinom{\beta}{2}\cdot 2^{1+\beta-\beta} = -\tbinom{\beta}{2}.\end{aligned}$$

Theorem 5.4. *Condition (2) is sufficient for the weak convergence (15) to hold with the limiting generator given by (19).*

Conversely, if $c_N\to 0$ and if (15) is satisfied with the limiting generator given by (19), then (2) holds.

Proof. Put

$$\Phi_A(b_1,\dots,b_A) := \Phi_A^{(N)}(b_1,\dots,b_A) := \frac{(N)_A}{(2N)_b}\mathrm{E}((\nu_1)_{b_1}\cdots(\nu_A)_{b_A}) \tag{21}$$

so that $c_N = \Phi_1^{(N)}(2)/4$. In the same manner as in Möhle and Sagitov, (2001) it is shown that the functions (21) are monotone in the sense that

$$\Phi_j(k_1, \ldots, k_j) \ \leq \ \Phi_l(m_1, \ldots, m_l) \tag{22}$$

whenever $j \geq l$ and $k_1 \geq m_1, \ldots, k_l \geq m_l$. This monotonicity together with the following Lemma 5.5 imply that under condition (2) all the limits (10) with $k_1 + \cdots + k_j > j + 1$ are equal to zero and according to Theorem 4.2 the weak convergence (15) holds with the limiting generator given by (19).

The converse assertion follows from the fact that the limiting generator given by (19) corresponds to the case $\phi_1(3) = 0$, which is equivalent to (2). □

Lemma 5.5. *If (2) holds, then*

$$\lim_{N\to\infty} \frac{N}{c_N} P(\nu_1 > N\varepsilon) \ = \ 0 \quad \forall\, \varepsilon > 0, \tag{23}$$

$$\lim_{N\to\infty} c_N \ = \ 0, \tag{24}$$

$$\phi_2(2,2) \ := \ \lim_{N\to\infty} \frac{\Phi_2^{(N)}(2,2)}{c_N} \ = \ 0. \tag{25}$$

Proof. Given (2) we have

$$\mathrm{E}((\nu_1)_3) = o(N^2 c_N), \ N \to \infty. \tag{26}$$

Fix $\varepsilon > 0$ and for $i \in \{1, \ldots, N\}$ define $A_i := \{\nu_i \leq N\varepsilon\}$ and $B_i := \{\nu_i > N\varepsilon\}$. From

$$N\, P(\nu_1 > N\varepsilon) \leq \frac{N}{(N\varepsilon)_3} \mathrm{E}((\nu_1)_3)$$

it is clear that (23) follows from (26). To verify (24) observe that

$$\begin{aligned}
4c_N = \Phi_1^{(N)}(2) \ &= \ \frac{N}{(2N)_2} \mathrm{E}((\nu_1)_2) \\
&= \frac{1}{(2N)_2} \sum_{i=1}^{N} \mathrm{E}((\nu_i)_2 1_{A_i}) + \frac{1}{(2N)_2} \sum_{i=1}^{N} \mathrm{E}((\nu_i)_2 1_{B_i}) \\
&\leq \frac{N\varepsilon}{(2N)_2} \sum_{i=1}^{N} \mathrm{E}(\nu_i 1_{A_i}) + \frac{N}{(2N)_2} \mathrm{E}((\nu_1)_2 1_{B_1}) \\
&\leq \frac{N\varepsilon}{(2N)_2} \mathrm{E}(\nu_1 + \cdots + \nu_N) + N\, \mathrm{E}(1_{B_1}) \\
&= \frac{2N^2\varepsilon}{(2N)_2} + N\, P(\nu_1 > N\varepsilon) \ \leq \ \varepsilon + \frac{N}{c_N} P(\nu_1 > N\varepsilon).
\end{aligned}$$

This inequality together with (23) leads to (24). Finally, to prove (25) use

$$\begin{aligned}\sum_{i\neq j}\mathrm{E}((\nu_i)_2(\nu_j)_2 1_{A_i}) &\le 2N^2\varepsilon\sum_j \mathrm{E}((\nu_j)_2)\\ &= 2N^3\varepsilon\mathrm{E}((\nu_1)_2)\\ &= 8N^2(2N)_2\varepsilon c_N \;\le\; (2N)^4 2\varepsilon c_N\end{aligned}$$

and

$$\begin{aligned}\sum_{i\neq j}\mathrm{E}((\nu_i)_2(\nu_j)_2 1_{B_i}) &\le (2N)^3\sum_{i,j}\mathrm{E}(\nu_j 1_{B_i})\\ &= (2N)^4\sum_i \mathrm{E}(1_{B_i}) \;=\; (2N)^4 N P(\nu_1 > N\varepsilon)\end{aligned}$$

to see that

$$(N)_2\mathrm{E}((\nu_1)_2(\nu_2)_2) \;=\; \sum_{i\neq j}\mathrm{E}((\nu_i)_2(\nu_j)_2) \;\le\; (2N)^4\left(2\varepsilon c_N + N\,P(\nu_1 > N\varepsilon)\right).$$

This entails the inequality

$$\frac{\Phi_2(2,2)}{c_N} \;\le\; \frac{(2N)^4}{(2N)_4}\left(2\varepsilon + \frac{N}{c_N}P(\nu_1 > N\varepsilon)\right)$$

which together with (23) shows that (25) holds. □

Corollary 5.6. *If the marginal distribution of the offspring sizes satisfies the condition (2) then the time-scaled ancestral process* $(\mathcal{R}_{[t/c_N]})_{t\ge 0}$ *converges weakly to the Kingman coalescent.*

6. The coalescent with multiple mergers

In the previous section we have seen that if multiple mergers involving three or more ancestral lines do not occur, then the diploid coalescent coincides with the Kingman coalescent. One has to leave the framework of the Kingman coalescent to see the difference between the diploid coalescent and the haploid coalescent. In this section we discuss general diploid coalescent patterns corresponding to the coalescent generator (13) stated by Theorem 4.2. As in the haploid case (Möhle and Sagitov, (2001)), in general, the coalescent tree in the diploid case allows for multiple mergers of ancestral lines.

We return to condition (10) of Theorem 4.2 which regulates the joint distribution of offspring sizes within a generation. According to Möhle and Sagitov, (2001) this condition has the following equivalent formulation. There exists a symmetric measure F_j defined on the simplex

$$\Delta_j \;:=\; \{(y_1,\dots,y_j)\in[0,1]^j \mid y_1+\cdots+y_j\le 1\}$$

such that

$$\lim_{N\to\infty}\frac{\mathrm{E}((\nu_1)_2\cdots(\nu_j)_2)}{(4N)^j c_N} = F_j(\Delta_j) \tag{27}$$

and

$$\lim_{N\to\infty}\frac{N^j}{c_N}P(\nu_1 > 2Nx_1,\ldots,\nu_j > 2Nx_j) = \int_{x_1}^1\cdots\int_{x_j}^1\frac{F_j(dy_1,\ldots,dy_j)}{y_1^2\cdots y_j^2}, \tag{28}$$

holding for all points $(x_1,\ldots,x_j)$ of continuity for the measure F_j. In particular, if $j=1$ then $\Delta_1=[0,1]$ and $F_1([0,1])=4$.

Observe that condition (28) deals with *large families* with offspring size of order N. Whenever such a family is encountered in the history of the population, then there is a positive probability that it hosts three or more marked individuals. Formula (13) can be rewritten in terms of integrals over measures F_j in a similar way as in Möhle and Sagitov, (2001).

If $F_2(\Delta_2)=0$ then (13) has the form

$$\tilde{g}_{\tilde{\xi}\tilde{\eta}} := \begin{cases} -4\displaystyle\int_{[0,1]}\frac{1-(1-x)^{b-1}(1-x+bx)}{x^2}\,F(dx) & \text{if } \tilde{\xi}\prec\tilde{\eta}, b=A,\\ 4\cdot 2^{A+a-\beta-b}\displaystyle\int_{[0,1]}x^{b_1-2}(1-x)^{b-b_1}\,F(dx) & \text{if } \tilde{\xi}\prec\tilde{\eta}, b_1\geq 2,\\ & b_2=\cdots=b_A=1,\\ 0 & \text{otherwise,}\end{cases} \tag{29}$$

with $F=F_1$ from (27). Now we can assume that the generator is given by (29) with F being an arbitrary probability measure on the unit interval $[0,1]$. In this case it makes sense to call the corresponding diploid coalescent (cf. Corollary 4.3) a *diploid F-coalescent*. The generator of the Kingman coalescent in Section 5 is the special example of the diploid F-coalescent, where the measure F is the point measure $F=\delta_0$ concentrated at zero.

It is interesting to see that the diploid F-coalescent in contrast to the haploid F-coalescent with intensities

$$q_{\xi\eta} = \begin{cases}\displaystyle\int_0^1\frac{1-(1-x)^{\beta-1}(1-x+\beta x)}{x^2}\,F(dx) & \text{if } \xi=\eta,\\ \displaystyle\int_0^1 x^{\beta_1-2}(1-x)^{\beta-\beta_1}\,F(dx) & \text{if } \xi\prec\eta,\ \beta_1\geq 2,\\ & \beta_2=\cdots=\beta_\alpha=1,\\ 0 & \text{otherwise}\end{cases} \tag{30}$$

allows for simultaneous mergers of ancestral lines. One can encounter up to four mergers within a large family.

7. Discussion

The model of Kämmerle (1991) describes a population of individuals without gender. Individuals form couples and only couples produce offspring. Hence his model is a two-parent model but not a two-sex model. He is mostly interested in the forward process. He counts the number $X_n^{(i)}$ of couples in generation n (forward in time) which are descendants of a fixed number i of couples chosen randomly from

the present generation 0. Under certain conditions the process $(X_n^{(i)})_n$ behaves asymptotically like a Galton Watson branching process as the population size N tends to infinity. This convergence property leads to asymptotical results for the extinction probability $q_i := P(X_n^{(i)} = 0 \text{ finally})$. In order to derive further results he looks also backward in time, i.e., he counts the number of ancestral couples and connects this backward process to the forward process $(X_n^{(i)})_n$ using equations which are based on the well developed principle of duality (Liggett (1985), Möhle (1999b)). Kämmerle studies numbers of ancestral couples, but he does not trace back ancestral lineages and he does not introduce coalescent structures.

Möhle (1994) generalizes the results of Kämmerle to models which distinguish between females and males. The coalescent structures for these two-sex models are introduced in Möhle (1998b). The same paper presents a convergence theorem for the two-sex Wright-Fisher model which reveals the presence of instantaneous states in the limiting two-sex ancestral process. The related question on relative compactness for the convergence-to-coalescent problem was answered in Möhle (1999a) by the modulus of continuity method.

Chang (1999) studies the two-parent Wright-Fisher model and derives asymptotical results for the number τ_N of generations back to the most recent common ancestor (MRCA) for *a*ll N current individuals. It turns out that $\mathrm{E}(\tau_N)$ is of order $\log N$ if in the ancestral graph each child is connected to both of her/his two parents. Recall that for haploid models $\mathrm{E}(\tau_N)$ is of order N. In a finite population the two-sex model produces more ancestor in comparison to a haploid model.

Pitman (1999) and Sagitov (1999) independently introduce the (haploid) F-coalescent (cf. (30)) using different approaches. Pitman defines the F-coalescent of an infinite population in the spirit of Kingman (1982a,b,c). A trajectory of such a process is a tree followed from the top to the root with the number of branches D_t decreasing as time t grows. The process $(D_t)_{t\geq 0}$ starts at $D_0 = \infty$ and has one absorbing state 1. Among other things Pitman addresses the question of finiteness of the transit time $T = \inf\{t \geq 0;\, D_t = 1\}$.

Sagitov (1999) derives the F-coalescent as a limiting process for the ancestral process of a haploid exchangeable population model (note that the coalescent time is counted as the scaled number of generations of the original population model). He shows, in particular, that $\mathrm{E}(T) < \infty$ if $F(dx) = dx^{1-\alpha}$, $0 < \alpha < 1$. The extreme case $\alpha = 1$ corresponds to the Kingman coalescent with $\mathrm{E}(T) = 2$ (Kingman, 1982a). The other extreme case $\alpha = 0$, where F is the uniform measure on $[0, 1]$, was first introduced, motivated by a problem from physics, by Bolthausen and Sznitman (1998). It is known that in the Bolthausen-Sznitman coalescent the transit time T is almost surely infinite.

Schweinsberg (2000a) solves the finiteness problem for the transit time T of the haploid F-coalescent completely. Applying the theory developed by Pitman (1999), Schweinsberg proves that T is almost surely finite if and only if $\mathrm{E}(T) < \infty$ which in turn holds if and only if $\sum_{\beta=2}^{\infty} r_\beta^{-1} < \infty$, where $r_\beta := \int_0^1 (\beta x - 1 + (1 - x)^\beta) x^{-2} F(dx)$.

The class of coalescent processes allowing for simultaneous, multiple collisions of ancestral lines is introduced as a limiting process of the backward process in

haploid exchangeable population models by Möhle and Sagitov, (2001). Schweinsberg (2000b) in particular addresses the finiteness problem of the transit time for these class of processes. Recently there is much research interest in the coalescent with simultaneous multiple mergers (Bertoin and Le-Gall, 2002; Möhle, 2001; Sagitov, 2002; Schweinsberg, 2003). The present paper adjusts the convergence results in Möhle and Sagitov, (2001) for diploid population models.

References

[1] Bertoin, J., Le-Gall, J.-F.: Stochastic flows associated to coalescent processes. Preprint 02-36, École normale supérieure, Département de mathématiques et applications (DMA), Paris, France, 2002

[2] Beerli, P., Felsenstein, J.: Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152**, 763–773 (1999)

[3] Bolthausen, E., Sznitman, A.-S.: On Ruelle's probability cascades and an abstract cavity method. Commun. Math. Phys. **197**(2), 247–276 (1998)

[4] Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. Adv. Appl. Prob. **6**, 260–290 (1974)

[5] Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach. II. Further haploid models. Adv. Appl. Prob. **7**, 264 –282 (1975)

[6] Chang, Joseph T.: Recent common ancestors of all present-day individuals. Adv. Appl. Prob. **31**, 1002–1026 (1999)

[7] Fu, Y.X.: New statistical tests of neutrality for DNA samples from a population. Genetics **143**, 557–570 (1996)

[8] Griffiths, R.C., Tavaré, S.: Ancestral inference in population genetics. Statistical Science **9**, 307–319 (1994)

[9] Kämmerle, K.: The extinction probability of descendants in bisexual models of fixed population size. J. Appl. Prob. **28**, 489–502 (1991)

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

[11] Kingman, J.F.C.: Exchangeability and the evolution of large populations. In: Koch, G., Spizzichino, F.: Exchangeability in Probability and Statistics. North-Holland Publishing Company, 1982b, pp. 97–112

[12] Kingman, J.F.C.: The coalescent. Stoch. Process. Appl. **13**, 235–248 (1982c)

[13] Liggett, T.M.: Interacting Particle Systems. Berlin: Springer-Verlag, 1985

[14] Möhle, M.: A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. Adv. Appl. Prob. **30**, 493–512 (1998a)

[15] Möhle, M.: Coalescent results for two-sex population models. Adv. Appl. Prob. **30**, 513–520 (1998b)

[16] Möhle, M.: Weak convergence to the coalescent in neutral population models. J. Appl. Prob. **36**, 446–460 (1999a)

[17] Möhle, M.: The concept of duality and applications to Markov processes arising in neutral population genetics models. Bernoulli **5**, 761–777 (1999b)

[18] Möhle, M.: Forward and backward diffusion approximations for haploid exchangeable population models. Stoch. Proc. Appl. **95**, 133–149 (2001)

[19] Möhle, M., Sagitov, S.: A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. **29**, 1547–1562 (2001)

[20] Pitman, J.: Coalescents with multiple collisions. Ann. Probab. **27**, 1870–1902 (1999)

[21] Sagitov, S.: The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Prob. **36**, 1116–1125 (1999)

[22] Sagitov, S.: Convergence to the coalescent with simultaneous multiple mergers. Preprint 2002:82, Chalmers University of Technology and Göteborg University, Sweden, 2002

[23] Schweinsberg, J.: A necessary and sufficient condition for the Λ-coalescent to come down from infinity. Electron. Comm. Probab. **5**, 2000a
[24] Schweinsberg, J.: Coalescents with simultaneous multiple collisions. Electron. J. Probab. **5**, 2000b
[25] Schweinsberg, J.: Coalescent processes obtained from supercritical Galton-Watson processes. To appear in Stoch. Proc. Appl. 2003
[26] Slatkin, M.: An exact test for neutrality based on the Ewens sampling distribution. Genet. Res. Camb. **64**, 71–74 (1994)
[27] Slatkin, M.: A correction to the exact test based on the Ewens sampling distribution. Genet. Res. Camb. **68**, 259–260 (1996)
[28] Stephens, M., Donnelly, P.: Inference in molecular population genetics. J. Roy. Statist. Soc. B **62**, 605–655 (2000)
[29] Tajima, F.: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**, 585–595 (1989)