Wright–Fisher Revisited: The Case of Fertility Correlation

Alexandre Sibert
Laboratoire d’Anthropologie Biologique, CNRS FRE 2292, Musée de l’Homme, 17 Place du Trocadero, Muséum National d’Histoire Naturelle 75116 Paris, France
E-mail: sibert@mnhn.fr

Frédéric Austerlitz
Laboratoire de Génétique et d’Amélioration des Arbres Forestiers, INRA, Cestas, France

and

Évelyne Heyer
Laboratoire d’Anthropologie Biologique, CNRS FRE 2292, Musée de l’Homme, 17, Place du Trocadéro, Muséum National d’Histoire Naturelle 75116 Paris, France

Received November 1, 2001

We study the non-genetic inheritance of fertility from parents to offspring. For this purpose, we propose an exchangeable extension of the Wright–Fisher model. This extension allows us to introduce non-genetic fertility correlation in the forward in time process and to study its effects on the genealogies of individuals (or genes) samples. Since it is independent of the gene considered, this effect is uniform on the genome, even in diploid populations. For values of fertility correlation observed in human populations, we show that coalescence times are strongly but inhomogenously reduced and that the shape of gene genealogies is markedly unbalanced. Despite the fact that our simulations concern stationary populations, the former non-genetic effect is very similar to what has been described for populations of variable size such as populations passing through demographic bottleneck. However, additional strong tree imbalance due to non-genetic causes is reported here for the first time.

Key Words: non-genetic fertility correlation; Wright–Fisher model; coalescent; tree topology.

1. INTRODUCTION

The concept of neutrality in population genetics is sometimes ambiguous. As far as the original theory of coalescence (Kingman, 1982b) or its extensions (e.g., Slatkin and Hudson, 1991; Griffiths and Tavaré, 1994) are concerned, at least two definitions coexist. The first is derived from the historical definition that, under neutrality, the effect of demography and mutation can be separated, that is the genealogical process and the mutational process are independent (Donnelly and Tavaré, 1995). The second states that the individuals present at a given time are exchangeable as regards their propensity to reproduce, and that any other information

0040-5809/02 $35.00
© 2002 Elsevier Science (USA)
All rights reserved.
about them or their parentage is irrelevant (Kingman, 1982a). According to both definitions, fertility selection is considered non-neutral since the propensity to reproduce depends on the genotypes of individuals. Although the first statement is a subcase of the second, they are not equivalent and one can think of a type of fertility inheritance which will not be genetic, but cultural for example (Cavalli-Sforza and Feldman, 1981). Such a phenomenon would be called neutral according to the first definition, but non-neutral according to the second one.

Such non-genetic fertility inheritance is not only a speculative counter example: it has been clearly ascertained in several human populations (Huestis and Maxwell, 1932; Williams and Williams, 1974), in the sense that offspring number distributions of parents and offspring are correlated: if an individual belongs to a sibship of size \( s \), namely if its parents have \( s \) offspring, the probability distribution of its offspring number is not independent of \( s \). It was described as a socio-demographic phenomenon by Austerlitz and Heyer (1998), who investigated its effect on the effective population size \( N_e \) by means of conditioned branching processes and showed a major reduction of \( N_e \) as correlation increases. Other models have been studied that may describe fertility correlation such as direct product branching processes (Karlin and McGregor, 1963), multitype branching processes as mentioned in Harris’ (1963) founding book and later developed in Mode (1971), Donnelly and Marjoram’s (1989) extension of the Moran model, Campbell’s coalescence with correlation (Campbell, 1999) or Cabaillero’s (1994) quantitative approach. The present approach is based on an extension of the Wright–Fisher model, yielding new results to be compared with the previous ones.

In Sections 2 and 3 we introduce a formalism for the notion of propensity to reproduce in relation with the simulation of an extension of the Wright–Fisher model for haploid populations where individuals are exchangeable with regard to their offspring number distribution. This enables us to model offspring number correlation from one generation to the next as a correlation of sibship size and propensity to reproduce.

In Section 4, we study the effects of non-genetic fertility correlation on coalescence times, on coalescent tree topology and on the shape of gene genealogies as a whole. In Section 5, we discuss these results, compare them with various extensions of the coalescent process and suggest how to simulate coalescents under fertility correlation.

## 2. Extension of the Wright–Fisher Model

### 2.1. Wright–Fisher Model

The Wright–Fisher model aims at representing the forward in time evolution of a haploid population of constant size \( N \) with non-overlapping generations. For this purpose, all individuals in a given generation are assumed to produce an infinite number of gametes, \( N \) of which are randomly sampled to form the next generation. This scheme may not seem appropriate for many species, such as mammals, for which the number of gametes per reproducing individual is not large.

And yet in order to simulate the Wright–Fisher model, one must consider this reproduction phase backward in time: for two successive generations, the parent of a given individual of the offspring generation is uniformly chosen among individuals from the previous generation. This yields a very simple algorithm for the Wright–Fisher model where generations are simulated one after the other.

Hence, although the Wright–Fisher model, in its original forward in time version may not always seem appropriate in population genetics, the backward in time sampling scheme is strictly equivalent to simulating a symmetric multinomial distribution for the parents’ progeny numbers \( (v_1, v_2, \ldots, v_N) \) on \( D_N = \{(v_1, \ldots, v_N) \mid v_i \geq 0 \text{ for all } i, \text{ and } \sum_{i=1}^{N} v_i = N\} \). It is far too restrictive to consider this distribution on \( D_N \) since its only relevant property is its invariance under any relabeling of individuals, i.e., the fact that individuals are indistinguishable (or symmetric) under this progeny distribution. Therefore, one can reasonably try to find a good model for the reproductive phase in a given species with any other distribution of progeny number that has this property.

Tavaře (1984) recalls that reproductive symmetry of the individuals is reflected in the fact that the r.v. \( (v_1, \ldots, v_N) \) on \( D_N \) is assumed exchangeable, a common and often implicit assumption in population genetic models. Exchangeable models comprise Wright–Fisher, Moran, and more generally Cannings’ (1974) synthetic models. As pointed out by Möhle (1998), the exchangeability is mostly used "in an assumption that the individuals present at a given time have the same propensity to reproduce".

Hereafter we show how to construct a class of such exchangeable distributions on \( D_N \) other than symmetric multinomial. This makes it possible to simulate an exchangeable population process with a marginal
offspring number distribution with mean 1 not necessarily binomial $\mathcal{B}(N, 1/N)$. It must be stressed that the shape of this marginal distribution itself is not of great interest in the framework of exchangeable models: the quantity $E(v_1(v_1 - 1))$ suffices to describe the fixation process (Cannings, 1974). This quantity is exactly the variance in progeny number when population size is constant.

Although the original Wright–Fisher scheme precludes any correlation in the progeny number from one generation to the next, we shall show how the above-mentioned exchangeable processes constitute a natural way to introduce this correlation.

For the last two decades the focus has turned to the \textit{backward in time} process, also called genealogical or coalescence process. It was shown that after an appropriate change of time scale, the limiting process as $N$ tends to infinity has the same structure for any exchangeable neutral model (for a review, see Nordborg, 2001). Moreover, recent developments (Mohle, 1998) showed that the ancestral structure can be approximated in the same way for a large class of non-exchangeable models. However, these results in the absence of correlation do not help us to predict the behavior of the process when correlation is introduced. On the contrary, simulations of the correlated process allow to draw conclusions concerning both the fixation process and its \textit{backward in time} dual.

### 2.2. Propensity to Reproduce

Caballero (1994) recalls that “in real populations, parents may have different probabilities of contributing offspring because of differences in their fertility or in the viability of their offspring or, perhaps, because of impositions by the breeder”. That is, contributions to the next generation differ between parents not only because of random sampling, as is the case in the standard Wright–Fisher scheme, but also because of additional stochastic variations. This yields a natural extension of the usual model.

Given an individual from the current generation, we assume now that all members of the previous generation are not equally likely to be its parent, but rather that individual $i$ has probability $\lambda_i$ of being chosen, with $(\lambda_1, \ldots, \lambda_N)$ a r.v. taking value in the $(N - 1)$-dimensional simplex $\Delta_N = \{(x_1, \ldots, x_N) | x_i \geq 0 \text{ for all } i, \text{ and } \sum_{i=1}^N x_i = 1\}$. In what follows $\lambda_i$ is called \textit{propensity to reproduce} of individual $i$ and is chosen anew every generation. The sampling scheme being of the same nature as in the Wright–Fisher model, it is clear that the resulting joint distribution of the offspring numbers conditional on $(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is again multinomial on $D_N$:

\begin{equation}
(v_1, \ldots, v_N)(\lambda_1, \ldots, \lambda_N) \sim \mathcal{M}(\lambda_1, \ldots, \lambda_N).
\end{equation}

Since $E(v_i|\lambda_i) = N\lambda_i$, that is the mean progeny number of individual $i$ given its propensity, $\lambda_i$ summarizes all the possible additional causes of variation in contribution of individual $i$ to the next generation, as suggested by Caballero (1994). Moreover, as \textit{propensities to reproduce} are chosen anew every generation, they may be correlated between parents and offspring, without making any assumptions about the mode of inheritance (genetic or cultural) in what follows. It is simply materialized by a correlation between the actual progeny number of one individual and its children’s \textit{propensities to reproduce}. Hence, since it is not a priori related to individual $i$’s genotype, $\lambda_i$ should not be thought of as a selection coefficient in the sense of population genetics, but rather as the relative contribution of individual $i$ to the next generation.

In the absence of fertility correlation, the distribution of the above propensities to reproduce is independent of the actual offspring numbers in the previous generation, $(v_1, \ldots, v_N)$ (see Table I). Then the r.v. $(\lambda_1, \ldots, \lambda_N)$ is

<table>
<thead>
<tr>
<th>TABLE I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table of Notations</strong></td>
</tr>
<tr>
<td>$v_i$, progeny number of individual $i$ (current generation)</td>
</tr>
<tr>
<td>$v_j$, progeny number of individual $j$ (previous generation)</td>
</tr>
<tr>
<td>$s_i$, sibship size of current individual $i$</td>
</tr>
<tr>
<td>$\Lambda$, intermediate r.v.</td>
</tr>
<tr>
<td>$\lambda_i$, \textit{propensity to reproduce} current individual $i$</td>
</tr>
<tr>
<td>$a$, shape parameter of the Gamma distribution $G_a$</td>
</tr>
<tr>
<td>$x$, parameter in $h_0(x) = x^x$, related to fertility correlation</td>
</tr>
<tr>
<td>$\sum_{i=1}^N v_i = N$</td>
</tr>
<tr>
<td>$\sum_{i=1}^N v_j = N$</td>
</tr>
<tr>
<td>$\sum_{i=1}^N s_i = \sum_{i=1}^N (v_j)^2$</td>
</tr>
<tr>
<td>$\Lambda_i</td>
</tr>
<tr>
<td>$\lambda_i = \Lambda_i/\sum_{i=1}^N \Lambda_i$</td>
</tr>
<tr>
<td>$a = 1$: geometric distribution</td>
</tr>
<tr>
<td>$a = \infty$: degenerate at 1</td>
</tr>
<tr>
<td>$a = 0$: uncorrelated</td>
</tr>
<tr>
<td>$a &gt; 0$: correlated</td>
</tr>
</tbody>
</table>
assumed to be exchangeable on $\Delta_N$ since, on average, there is no reason why an individual should contribute more or less than any other: individuals are exchangeable as regards their propensity to reproduce.

Let $\varphi$ be the density function of $(\lambda_1, \ldots, \lambda_N)$ on $\Delta_N$. Then the unconditional joint distribution of $(v_1, \ldots, v_N)$ on $D_N$ is a compound multinomial distribution with measure $\varphi$:

$$
P(v_1, \ldots, v_N) = \int_{\Delta_N} \left( \begin{array}{c} N \\ v_1, \ldots, v_N \end{array} \right) \lambda_1^{v_1} \cdots \lambda_N^{v_N} \\
\varphi(\lambda_1, \ldots, \lambda_N) \, d\lambda_1 \cdots d\lambda_N
$$

$$
= \left( \begin{array}{c} N \\ v_1, \ldots, v_N \end{array} \right)^{m_{v_1,\ldots,v_N}}(\varphi),
$$

(2)

where $m_{v_1,\ldots,v_N}(\varphi)$ is the moment of order $(v_1, \ldots, v_N)$ of $\varphi$. Since the distribution of $(\lambda_1, \ldots, \lambda_N)$ is assumed exchangeable on $\Delta_N$, Eq. (2) implies that the unconditional distribution of $(v_1, \ldots, v_N)$ is also exchangeable on $D_N$. In this case, mean and variance of the marginal offspring number are given by $E(v_i) = E\{v_i|\lambda_i\} = N\E(\lambda_i) = 1$ and the conditional variance formula yields

$$
V(v_i) = V(E(v_i|\lambda_i)) + E(V(v_i|\lambda_i))
$$

$$
= \frac{N-1}{N}(V(N\lambda_i) + 1).
$$

(3)

The above extension does not correspond to Möhle’s (1998) generalized Wright–Fisher model because the $\lambda_i$’s are not constant in time. However, in the absence of correlation, our model is a particular case of the mixing or coupling method described by Möhle (1998), whose convergence results then apply normally. It is also important to stress that our extension constitutes only a subset of the set of exchangeable models. This is due to the fact that all exchangeable distributions on $D_N$ are not compound multinomials. Indeed, a necessary condition is that $V(v_i) \geq (N-1)/N$, the minimum being reached by the symmetric multinomial on $D_N$. For example, the distribution on $D_N$ associated with the Moran model provides a counter-example for $N \geq 3$, since $V(v_i) = 2(N-1)/N^2$ (Cannings, 1974).

Despite these remarks, the notion of propensity to reproduce allows us to introduce a correlation of fertility between parents and their offspring: it suffices, for example, that an individual’s propensity to reproduce $\lambda_j$ depends on its actual sibship size $s_j$, i.e., on its parent’s actual progeny number $v_j$ (see Table 1). This is implemented in what follows.

To finish with, one can show that under mild conditions on this relationship, that is on the type of fertility correlation, there exists a non-trivial stationary distribution of $(v_1, \ldots, v_N)$ on $D_N$, as is the case in the absence of correlation. Indeed, given the state $v = (v_1, \ldots, v_N)$ in the previous generation, there exists a probability $\pi_v$ that the state in the current generation is $v = (v_1, \ldots, v_N)$. Under the hypotheses of the Perron–Frobenius theorem for non-negative matrices, the eigenvalue of greatest modulus is simple, real and associated with an eigenvector whose components are all of the same sign. This vector corresponds precisely to the stationary distribution. It is interesting to note that the stationary probability of a configuration $v$ does not depend on the labeling of $(v_1, \ldots, v_N)$. So the stationary distribution is exchangeable. The following results concern situations where fertility correlation yields a non-degenerate stationary progeny number distribution on $D_N$.

3. SIMULATIONS OF AN EXTENDED WRIGHT–FISHER MODEL

3.1. Simulation of a General Exchangeable Model without Correlation

The simulation of our exchangeable extended Wright–Fisher model first requires to draw $N$ non-negative random numbers $(\lambda_1, \ldots, \lambda_N)$ adding up to 1 and then proceed to the classic Wright–Fisher multinomial sampling. The first step can be achieved by drawing $N$ i.i.d. non-negative real numbers $\Lambda_i$ and “normalizing” them as follows:

$$
(\lambda_1, \ldots, \lambda_N) = \left( \frac{\Lambda_1}{\sum_{i=1}^N \Lambda_i}, \ldots, \frac{\Lambda_N}{\sum_{i=1}^N \Lambda_i} \right).
$$

(4)

$\Lambda_i$'s are intermediate r.v. only used to generate $(\lambda_1, \ldots, \lambda_N)$. In the case where the $\Lambda_i$'s follow a two-parameter Gamma distribution with shape parameter $a_i$ and identical scale parameter $\gamma$, that is with density function

$$
\frac{\gamma^{a_i}}{\Gamma(a_i)} x_1^{a_i-1} \exp\left(-\frac{\gamma}{x_1}\right)
$$

(5)

$(\lambda_1, \ldots, \lambda_N)$ follows the well-studied Dirichlet distribution with parameters $(a_1, \ldots, a_N)$, and density function

$$
\mathcal{D}(a_1, \ldots, a_N)(x_1, \ldots, x_N) = \frac{\Gamma(a_1 + \cdots + a_N)}{\Gamma(a_1) \cdots \Gamma(a_N)} x_1^{a_1-1} \cdots x_N^{a_N-1}
$$

(6)

on $\Delta_N$. It is obvious that $\mathcal{D}(a_1, \ldots, a_N)$ is independent of $\gamma$ since any linear change of variable $(\Lambda_i \rightarrow \kappa \Lambda_i$ for all $i$) would result in the same $(\lambda_1, \ldots, \lambda_N)$. Therefore, without
Non-genetic Fertility Correlation

loss of generality, \( \gamma \) can be set to 1 and we consider the standard one-parameter Gamma distribution, denoted by \( G_a \) hereafter. We chose this particular distribution of \((\lambda_1, \ldots, \lambda_N)\) because the simulation of the Dirichlet distribution is rather simple and that other distributions of the \( \lambda_i \)'s (e.g., uniform) do not yield a closed form for the density of \((\lambda_1, \ldots, \lambda_N)\).

Now, let us assume that \((\lambda_1, \ldots, \lambda_N)\) follows a symmetric Dirichlet distribution on \( \Delta_N \), that is \( a_i = a \) for all \( i \). The offspring number distribution given \((\lambda_1, \ldots, \lambda_N)\) being multinomially distributed with parameters \((\lambda_1, \ldots, \lambda_N)\), the unconditional offspring number distribution is the so-called Dirichlet-multinomial distribution:

\[
P(v_1, \ldots, v_N) = \int_{\Delta_N} \left( \sum_{i=1}^{N} \frac{\lambda_i^v}{\Gamma(a_i)} \frac{\Gamma(Na)}{\Gamma(a + v_i)} \frac{1}{\Gamma(N(a + 1))} \right) \, d\lambda_1 \ldots d\lambda_N
\]

where \( X \) is a non-negative real valued r.v. with mean 1 so that \( \mathbb{E}(\Lambda_i|s_i) = h(s_i) \). Then the \( \lambda_i \)'s are divided by their sum so that the resulting \( \lambda_i \)'s add up to 1 as before. When \( X \) is degenerate at 1, \( \Lambda_i \) and \( \lambda_i \) depend deterministically on \( s_i \). The role of the r.v. \( X \) is to allow additional variance in \( \Lambda_i \) and make it possible to check the robustness of results for various distributions of \( \Lambda_i \) given \( s_i \), and consequently of \((\lambda_1, \ldots, \lambda_N)\) given \((s_1, \ldots, s_N)\).

For \( h \) we chose the one-parameter family of functions \( h_z(s) = s^z \) with \( z \in [0, 2] \), and for \( X \) the one-parameter Gamma distribution \( G_a \) as in (5) (see Table I). Hence,

\[
\left( \frac{a}{s^z} \right)^{a-1} \frac{1}{\Gamma(a)} e^{-a/s^z} x
\]

is the explicit probability density of \( \Lambda_i|s_i \) for \( a < \infty \). When \( a \) tends to infinity this density function converges in distribution to the degenerate distribution at \( s_i^z \delta_z(x) \). In order to ease notations in what follows, we write \( a = \infty \) to refer to this limit.

When \( z = 0 \), \( h_z \) is a constant and there is no correlation between sibship size and propensity to reproduce. For \( a < \infty \) the joint distribution of \((\lambda_1, \ldots, \lambda_N)\) is the above-mentioned Dirichlet distribution \( D(a, \ldots, a) \) on \( \Delta_N \). This corresponds to the general exchangeable model whereas the classic Wright–Fisher model is related to the limit \( a \to \infty \), for which the distribution of \((\lambda_1, \ldots, \lambda_N)\) is degenerate at \((1/N, \ldots, 1/N)\). The marginal propensity to reproduce and offspring number distributions for these different cases are given in Table II.

When \( z > 0 \) on the contrary, i.e., in the case of correlation between sibship size and propensity to reproduce, there is no closed expression for the joint distribution of \((\lambda_1, \ldots, \lambda_N)\).
TABLE II
Marginal Propensity to Reproduce and Unconditional Offspring Number Distributions without Correlation and a Symmetric Dirichlet Joint Distribution of Propensities

<table>
<thead>
<tr>
<th>Marginal distributions</th>
<th>Propensity $N\lambda$</th>
<th>Offspring number $v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(N &gt; 1, a \geq 1)$</td>
<td>$E(N\lambda) = 1$</td>
<td>$E(v) = 1$</td>
</tr>
<tr>
<td>$a &lt; \infty, N &lt; \infty$</td>
<td>$\frac{1}{N^{1+a}} \Gamma((N-1)\lambda)\Gamma(N)\Gamma((N-1)\lambda^{-1})\Gamma(N-1)\Gamma(N-1)$</td>
<td>$\frac{a^{N-2}N^{a-1}}{(a+1)N-1} \frac{1}{(a+1)N-1}$</td>
</tr>
<tr>
<td></td>
<td>$V(N\lambda) = \frac{N-1}{N^{2+a}}$</td>
<td>$V(v) = \frac{a^{N-1}}{1+aN}$</td>
</tr>
<tr>
<td>$a = \infty, N &lt; \infty$</td>
<td>$\delta_1$</td>
<td>$\text{Binomial}(1/N, N)$</td>
</tr>
<tr>
<td>(Wright–Fisher)</td>
<td>$V(N\lambda) = 0$</td>
<td>$V(v) = \frac{N-1}{a}$</td>
</tr>
<tr>
<td>$a &lt; \infty, N = \infty$</td>
<td>$\frac{a^{N-2}N^{a-1}e^{-a\lambda}}{\Gamma(N-1)}$</td>
<td>$V(v) = \frac{a^{N-1}}{1+aN}$</td>
</tr>
<tr>
<td></td>
<td>$V(N\lambda) = \frac{1}{a}$</td>
<td>$V(v) = 1 + \frac{1}{a}$</td>
</tr>
<tr>
<td>$a = \infty, N = \infty$</td>
<td>$\delta_1$</td>
<td>$\text{Poisson}(1)$</td>
</tr>
<tr>
<td></td>
<td>$V(N\lambda) = 0$</td>
<td>$V(v) = 1$</td>
</tr>
</tbody>
</table>

Results for the variance can be obtained from the conditional variance formula.

3.3. Simulations

We simulated a population of constant size $N$ forward in time with non-overlapping generations. To each individual of the parent generation we associated a propensity to reproduce $\lambda_i$ as described before.

For each set of parameters $(a, \lambda)$, we proceeded as follows. We first assigned equal propensities to reproduce to all $N$ individuals and simulated $4N$ generations of the process, a time long enough for the sibship sizes distribution to reach stationarity. We used the final sibship sizes to initiate the first period. Then as we simulated the evolution of the population forward in time during $8N$ generations, parents–offspring relationships were recorded from generation to generation. The final sibship sizes were used to initiate the next period and we repeated the procedure so as to generate 4000 such periods.

The variance in progeny number was computed for the last generation of each period and averaged over all 4000 repetitions. Fertility correlation of sibship size between parents and offspring ($\rho_s$) was computed over the last four generations taking into account only individuals having at least one offspring and averaging over all repetitions. In fact, we applied the same procedure as described in Austerlitz and Heyer (1998) to allow comparison.

Finally, a sample of $n$ individuals was uniformly drawn from the last generation and its genealogy reconstructed until the most recent common ancestor (MRCA) of the sample. For each simulation, the times between successive coalescent events were recorded and the mean and variance of these coalescence times over all repetitions were computed.

4. RESULTS

4.1. Variance and Correlation of Offspring Number Distributions

In the sequel we use $V(a, \lambda)$ to denote the average variance in offspring number $E[V(v)]$ for a given set of parameters $(a, \lambda)$.

The relationship between $V(a, \lambda)$ and the set $(a, \lambda)$ is displayed in Fig. 1. Curves correspond to different values of $a$, that is different shapes of the Gamma distribution. Whatever be the value of $a$, the offspring number variance increases with $\lambda$. Intuitively, a higher value of $\lambda$ yields a higher frequency of large sibships and since the mean number of children is 1 (constant population size), individuals with no offspring also
become more frequent. Hence the cumulative effect of fertility correlation yields an increased variance of the progeny number. Observed values of \( V(\alpha) \) for \( \alpha = 0 \) agree with theoretical variances (Table II): for \( N = 50 \), \( V(\alpha, 0) = (a + 1)(N - 1)/(1 + aN) = 0.98 \), 1.46 and 1.92 with \( a = \infty \), 2 and 1, respectively. Moreover, the ratio of \( V(\alpha, \alpha) \)'s corresponding to different values of \( a \) is almost independent of \( \alpha \) (results not shown) which allows us to write simply

\[
V(\alpha, \alpha) = (1 + g(\alpha))V(\alpha, 0),
\]

where \( g(\alpha) \) is an increasing function on \( \mathbb{R}^+ \) with \( g(0) = 0 \).

The relationship between average offspring number correlation \( E(\rho_u) \) and \( \alpha \) is less intuitive as shown in Fig. 2. For \( \alpha \leq 0.6 \), \( E(\rho_u) \) grows linearly with \( \alpha \) and is independent of \( a \) (\( E(\rho_u) \approx \alpha/4 \)). Hence the offspring number variance has no effect on \( E(\rho_u) \). Put in other words, for a fixed small value of \( \alpha \) the same offspring correlation \( E(\rho_u) \) is observed, whatever be the type of the marginal offspring number distribution (binomial or geometric). For higher values of \( \alpha \), \( E(\rho_u) \) tends to an upper bound (approximately 1/2 for \( a = \infty \), 1/3 for \( a = 2 \) and 1/4 for \( a = 1 \)). Upper bound values increase with \( a \), which is relevant since an increase in \( a \) corresponds to a decrease in the variance of the offspring number conditional on the individual sibship size (Table II), and therefore implies a higher correlation. Due to this upper asymptote \( E(\rho_u) \) is not an appropriate parameter to describe results of our simulations. In what follows, we keep \( \alpha \) as a measure of fertility correlation intensity.

### 4.2. \( T_{\text{MRCA}} \) and Branch Lengths in Coalescent Trees

#### 4.2.1. Time to the most recent common ancestor (\( T_{\text{MRCA}} \)).

For a given distribution of \((v_1, \ldots, v_N)\), the probability \( c_N \) that two offspring belong to the same sibship is (Möhle, 2000):

\[
c_N = \frac{1}{N(N - 1)} \sum_{i=1}^{N} E(v_i(v_i - 1)). \tag{12}
\]

This is exactly the unconditional probability that two lineages coalesce in one generation. It must be stressed that for exchangeable models and a wide range of non-exchangeable ones (Möhle, 1998), \( c_N \) is the appropriate time scale to describe the limit of the backward in time process as \( N \to \infty \); this limit is known as the coalescent. If there is no offspring number correlation and if \( c_N \) is independent of time, it can be shown that the expected time until the most recent common ancestor \( T_{\text{MRCA}} \) is

\[
E(T_{\text{MRCA}}) = \frac{2}{c_N} \left( \frac{1 - \frac{1}{n}}{n} \right). \tag{13}
\]

Since all \( v_i \)'s have the same marginal distribution in the exchangeable case, (12) becomes \( c_N = E(v_i(v_i - 1))/(N - 1) \). One can check that \( c_N = 2/(N(N - 1)) \) in the Moran model and \( 1/N \) in the exchangeable Wright–Fisher model for example.

Moreover, for a population of constant size (i.e., \( E(v_i) = 1 \)) \( E(v_i(v_i - 1)) = V(v_i) \). In our simulations, we computed the average value of \( V(v_i) \) together with the actual mean time until the MRCA of the sample. As is
shown in Fig. 3, the ratio \( \frac{\mathbb{E}(T_{MRCA}^{\alpha,0})}{\mathbb{E}(T_{MRCA}^{\alpha,2})} \)—the mean time until the MRCA without correlation (\( \alpha = 0 \)) divided by the mean time until the MRCA with correlation—depends linearly on \( V(a, x) \). A linear regression yields high correlation coefficients (\( \approx 0.99 \)) that suggest the following relationship:

\[
\mathbb{E}(T_{MRCA}^{\alpha}) = \frac{\mathbb{E}(T_{MRCA}^{\alpha,0})}{A(a)V(a, x) + B(a)},
\]

(14)

where \( A(a) > 0 \) is the slope and \( B(a) \) the \( Y \)-axis intercept of the least-squares line.

When there is no correlation, the variance of the number of offspring is given by \( V(a, 0) = \frac{(1 + a(N - 1))}{1 + aN} \) (Table II) and its value is 0.98 for \( a = \infty \), 1.46 for \( a = 2 \) and 1.92 for \( a = 1 \), close to observed values. Moreover, putting \( x = 0 \) in (14) yields \( B(a) = 1 - A(a)V(a, 0) \).

Since \( V(a, x) = (1 + g(x))V(a, 0) \) (see above) and \( \mathbb{E}(T_{MRCA}^{\alpha,0}) = 2N-1\frac{1}{F(a,0)}(1 - \frac{1}{n}) \) (standard coalescent theory), we get:

\[
\mathbb{E}(T_{MRCA}^{\alpha,2}) = \frac{2(N - 1)(1 - \frac{1}{n})}{V(a, 0)[1 + A(a)V(a, 0)g(x)]}.
\]

(15)

Let us denote \( V_\alpha \) the denominator in (15). It can be decomposed into \( V_\alpha = V_1 + V_2 + V_3 \) with:

\[
V_1 = V(a, 0),
\]

\[
V_2 = V(a, 0)g(x),
\]

\[
V_3 = V(a, 0)g(x)[A(a)V(a, 0) - 1].
\]

(16)

\( V_1 \) is the mean variance in offspring number in absence of correlation and \( V_1 + V_2 \) is the observed mean variance under correlation. Since \( g(x) \) ranges from 0 to 1, \( V_1 + V_2 \) is of the same order of magnitude as \( V_1 \).

If the coalescent process under correlation was a mere homogenous rescaling of the process without correlation (as is the case in exchangeable models for example), the expected time to the most recent common ancestor would be proportional to \( 1/(V_1 + V_2) \). From the results shown in Fig. 3, \( T(a)V(a, 0) \) ranges from around 10 to more than 15 (see Table III). Hence, the reduction of \( T_{MRCA} \) is not only due to the increase of variance in offspring number \( V_2 \), but also to an additional component \( V_3 \). Since the latter is one order of magnitude greater than the former, the process under correlation differs dramatically from the standard exchangeable coalescent. In the following we investigate this discrepancy.

4.2.2. Correlation of coalescence times. For a sample of \( i \) genes, exchangeable coalescent theory states that the time \( T_i \) until the next coalescence event is geometrically distributed with mean \( \frac{1}{\mathbb{E}(T_i)} \). Moreover, for a large class of non-exchangeable processes, the associated coalescent process also corresponds to this uniform time rescaling of the classic coalescent (Kingman, 1982b) by a factor \( \frac{1}{\mathbb{E}(T_i)} \). In our simulations, this factor is equal to \( \frac{n - 1}{n} = \frac{2}{1 + \frac{1}{F(a,0)}} \) (Table II).

Denote \( T_i(a, x) \) the mean time until the next coalescence event for a sample of \( i \) genes and a set of parameters \((a, x)\). Without correlation, we checked (results not shown) that the ratio \( T_i(a, 0)/T_i(\infty, 0) \) is independent of \( i \) and that its value is close to the theoretical value \( \frac{1 + \frac{1}{n}}{1 + \frac{1}{F(a,0)}} \): rescaling is time homogenous as predicted by theory.

On the contrary, when correlation is effective (i.e., \( \alpha > 0 \)) the ratios \( T_i(a, 0)/T_i(a, x) \) (see Fig. 4 for the case

| Table III |
|-----------------|-----------------|-----------------|
| Values for \( A(a) \) (after Linear Regression) and \( A(a)V(a, 0) \) for Various \( \alpha \)'s |
| \( \alpha \) | \( \infty \) | 2 | 1 |
| \( A(a) \) | 15.8 | 8.0 | 5.1 |
| \( A(a)V(a, 0) \) | 15.4 | 11.7 | 9.9 |
The ratio $E(T_\infty(\omega)) / E(T_\infty(0))$ for a set of parameters $(\omega, 0)$ increases: long branches (i.e., for $i$ small) are more affected by correlation than smaller ones. Hence, the effect of correlation on coalescence times is not homogenous: it is all the stronger as there are fewer lineages. Similar results were obtained for different values of $\omega$.

Moreover, the classical coalescent process as described by Kingman (1982b) states that coalescence times are independent random variables. For that case, since $T_{\text{MRCA}} = \sum_{i=2}^n T_i$, the variance of the observed $T_{\text{MRCA}}$ is equal to the sum of the $T_i$’s variances:

$$V(T_{\text{MRCA}}) = \sum_{i=2}^n V(T_i).$$

To study departure from this expectation, we consider the ratio

$$Z(\omega, 2) = \frac{\sqrt{V(T_{\text{MRCA}})} - \sqrt{\sum_{i=2}^n V(T_i)}}{\sqrt{V(T_{\text{MRCA}})}}$$

for $\omega = \infty$. For a given set of parameters $(\omega, z)$, $Z(\omega, z)$ represents the proportion of the $T_{\text{MRCA}}$’s standard error due to the correlation of coalescence times. As shown in Fig. 5, $Z(\omega, z)$ is approximately null for $z \leq 1$ whatever be the value of $\omega$, but increases significantly with higher values of $z$. A geometric distribution of propensities to reproduce ($\omega = 1$) enhances this effect.

4.3. Topology of Coalescent Trees

In this section we focus on the shape of genealogical trees, without taking branches lengths into account. Here we are only interested in what is called the combinatorial structure of genealogical trees, or cladograms (Aldous, 2001). The neutral coalescent yields a specific probability distribution on these topologies. Here we try to show to what extent fertility correlation yields a departure from this distribution.

4.3.1. Multiple and simultaneous coalescences. Coalescent theory deals with populations large enough to preclude coalescence events involving more than two lineages or simultaneous coalescence events. In fact, since the probability of these events is $o(1/N)$, they can be neglected for large $N$: only two lineage coalescences are taken into account, whose probability is $O(1/N)$. However, the size of simulated populations being small due to finite computer memory, multiple or simultaneous coalescences are common during simulations. Since correlation shrinks genealogical trees this effect is enhanced.

At least the last coalescences backward in time are less likely to be simultaneous or multiple than any other. This is why we study the ultimate coalescence event, that is the very last coalescence resulting in a unique lineage. For values of $z \in [0, 2]$ we computed the frequencies $f_2$, $f_3$ and $f_4$ of runs for which, respectively, 2, 3 and 4 lineages are involved in the ultimate coalescence. In the absence of correlation, these theoretical frequencies are
of order $O(1)$, $O(1/N)$ and $O(1/N^2)$, respectively. Figure 6 gives simulation results for $a = \infty$ (+) and 1 (*). Although multiple coalescences are rare when $x = 0$, i.e., in the absence of correlation, their frequency increases markedly with $x$. The effect is stronger in the case of a geometric propensity distribution (*): in this case, the variance of the offspring number $v$ is roughly twice as much as in the case $a = \infty$. This factor 2 is obvious for the frequencies of 3- and 4-lineage coalescences when $x \leq 1$. Beyond $x = 1$ the effect remains markedly stronger, as shown by the decrease in frequency of 2-lineage coalescences.

4.3.2. Balance statistics. As mentioned in Aldous (2001), an important and often addressed question concerning the topology of phylogenetic trees concerns their degree of balance or imbalance. Several summary statistics have been suggested to quantify imbalance in a whole tree. In principle, they can be used to test departure in tree topologies from the so-called "neutral scenario" resulting from various causes. We studied the six summary statistics described by Kirkpatrick and Slatkin (1993). Our results (not shown) did not allow to clearly discriminate between coalescent trees under the standard uncorrelated model ($x = 0$) and the correlated one ($x > 0$). Hence we focused more precisely on the degree of balance for each node in the tree.

Consider a binary tree. Unless a node is a leaf, it subtends $m \geq 2$ leaves distributed into two subtrees (left

**FIG. 6.** Effect of fertility correlation on frequencies of ultimate coalescences involving 2-, 3- and 4-lineages (resp. $f_2$, $f_3$ and $f_4$ in the text). $f_2$, $f_3$ and $f_4$ are plotted (top to bottom) vs $x$ for $a = \infty$ (+) and 1 (*).

**FIG. 7.** Distribution of imbalance for nodes subtending $m = 15$ (a), 13 (b), 11 (c) and 9 (d) leaves (as described in the text). V’s and solid lines correspond to $x = 0$, Δ’s and dashed lines to $x = 0.5$, O’s and dashed and dotted lines to $x = 1$ and ¥’s and dotted lines to $x = 1.5$. ($a = \infty$).
and right) with say \( m_L \) and \( m_R \) leaves, respectively. The absolute difference \( |m_L - m_R| \) is a simple measure of node’s imbalance. Its range lies between 0 (when \( m \) is even) or 1 (when \( m \) is odd) and \( m - 2 \). Low values correspond to most balanced nodes and vice versa. One can extend this measure to \( p \)-ary nodes with \( \sum_{i=1}^{p} |\bar{m}_i - m_i| \) where \( \bar{m}_i \) is the ratio of the number of leaves subtended by the node divided by the number of branches emerging from the node (\( p \)), and \( m_i \) is the number of leaves subtended by the \( i \)th branch (or the corresponding subtree).

In order to study the degree of imbalance of simulated trees we chose to fix the number of sampled individuals to \( n = 15 \) among \( N = 100 \). For each node in each tree we recorded both the number of leaves subtended and the degree of imbalance. Figure 7 shows the distribution of imbalance for different values of \( z \). The graphs correspond to nodes subtending \( m = 15 \) (root), 13, 11 and 9 leaves, respectively. Frequencies of odd imbalance values are plotted from 1 to \( m - 2 \). Other imbalance values (even or non-integer) result from multiple coalescences and remain rare enough not to be plotted here.

First of all, as expected in absence of correlation (\( \forall \)'s and solid lines) all values of imbalance are equiprobable. For odd \( m \), this is equivalent to stating that subtrees subtend all numbers of leaves (between 1 and \( m - 1 \)) with equal probability. Since there are \((m - 1)/2\) possible values when \( m \) is odd, the frequency of each imbalance value is \( 2/(m - 1) \). That imbalance distribution is uniform for \( z = 0 \) can be thought of as a counterpart of the neutral coalescence equiprobability for each lineage (or pair of lineages) to coalesce.

It is interesting to mention that imbalance distribution does not depend on \( z \) and remains uniform for \( z \leq 0.5 \). For \( z > 0.5 \), all four graphs show marked departure from the equiprobable distribution: unbalanced trees are more frequent and equilibrated ones become rare as fertility correlation increases. Nevertheless, the overall magnitude of this effect decreases with the total number of subtended leaves \( m \).

Finally, Fig. 7(a) shows imbalance distribution of the root (\( m = n \)). A value of 13 corresponds to a single lineage connecting to the root. Przeworski et al. (1999) studied the frequency of this event as a statistics. In the standard coalescence theory, the probability of this event (i.e., a lineage “excluded” from all coalescences but the last one) is

\[
\frac{m}{(m - 1)} \cdot \frac{m - 2}{2} \cdot \ldots \cdot \frac{2}{2} = \frac{2}{m - 1},
\]

in accordance with the above frequency. Here the frequency of trees with an external branch connecting to the root increases dramatically with \( z \), from around 0.15 (as expected for \( z = 0 \)) to 0.35 for \( z = 1.5 \).

4.3.3. Aggregation of coalescences. In the classical coalescence theory, all lineages present at a given time have the same probability to coalesce. We study here the deviation from this expectation, namely if the lineage resulting from the last coalescence is more likely to coalesce again in the next event, and if this probability depends on its size. Hereafter we call this phenomenon aggregation of coalescences. For a given set of parameters \((a, z)\) we study the frequency of runs for which the lineage resulting from the ante-penultimate coalescence is again involved in the penultimate one. Without correlation, the probability of this event is exactly \( 2/3 \).

In order to have a clear picture of this relationship we set \( a = \infty \) and we simulated 1 million trees for each value of \( z \). We obtained a frequency of 0.667 for \( z = 0 \) which agrees with theoretical predictions. Moreover, we observed an increase of this frequency with \( z \) up to a maximum of 0.725 for \( z = 1.6 \). This implies that when fertility is correlated, coalescence events do not involve all present lineages with the same probability. In fact, lineages that have just coalesced have a higher
probability to coalesce again during the next event. Figure 8 shows to what extent this probability to coalesce again depends on the size of the lineage itself. For low correlations ($\alpha = 0.4$ and 0.8), this probability is independent of the size. On the contrary for strong correlations ($\alpha = 1.2$ and 1.6), a positive linear correlation appears: the higher the lineage size, the higher its probability to coalesce again. A linear regression yields correlation coefficients of 0.98 and 0.96, respectively.

The fact that some lineages are more likely to coalesce and that this property is “transmitted” backward in time to the resulting lineage agrees with results of the previous sections: the higher the correlation, the higher the probability that some lineages are excluded from coalescence events until the very last ones. This explains why unbalanced trees are more likely when correlation increases.

In the next section we study the global effect of coalescence times correlation and coalescences aggregation on coalescence trees.

### 4.4. Shape of Correlated Trees

In its broad sense, coalescent tree shape takes both tree topology and branch length into account. Branch lengths is partly related to the coalescence times studied above. In the past two decades several statistics have been suggested to test departure from Kingman’s neutral coalescent. Although these tests are based on the study of genetic diversity observed in specific models (e.g., structured population, population expansion or weak selection), they definitely rely on the assumption that departure from Wright–Fisher model’s assumptions (e.g., neutrality or constant size) yields particular topologies (for example, with unexpectedly long external branches in the case of population expansion or positive selection). Since coalescence and mutational processes are independent in the former case, particular topologies themselves yield specific genetic variability. In this section we focus on the direct effect of fertility correlation on the shape of gene genealogies.

Uyenoyama (1997) introduces four statistics related to the structure of gene genealogies. These statistics are based on five quantities defined as follows: the time to the most recent common ancestor of the sample $D$ (equivalent to our $T_{\text{MRCA}}$), the total length of the tree $T$, the mean time until the most recent common ancestor for pairs of individuals in the sample $P$, the total length of external branches $S$, and $B$ the mean length of base branches, that is branches connecting to the root. In the framework of the standard exchangeable coalescent
theory (Kingman, 1982a,b) expectations of these quantities are
\[ \mathbb{E}(D) = 2N(1 - 1/n), \quad \mathbb{E}(T) = 2Na_n, \quad \mathbb{E}(P) = N, \]
\[ \mathbb{E}(S) = 2N, \quad \mathbb{E}(B) = 2Nb_n, \] \hspace{1cm} (20)
where \( a_n = \sum_{i=1}^{n-1} 1/i \) and \( b_n = 1/n + \sum_{i=1}^{n-1} 1/i^2 \). It must be stressed that \( D \) and \( T \) depend on coalescence times only, whereas \( B, P \) and \( S \) depend both on coalescence times and the topology of the tree. Uyenoyama’s statistics consist of four scaled ratios with numerator in the second group and denominator in the first:
\[ R_{PT} = \frac{2Pa_n}{T}, \quad R_{ST} = \frac{Sa_n}{T}, \]
\[ R_{SD} = \frac{S(1 - 1/n)}{D}, \quad R_{BD} = \frac{B(1 - 1/n)}{Db_n}. \] \hspace{1cm} (21)

For \( a = \infty \) and different values of \( x \) we simulated 4000 populations of constant size \( N \) as described above, traced back the genealogy of a random sample of \( n \) individuals and computed the different scaled ratios. Figure 9 gives their mean values for \( (N,n) = (100,10), (100,15), (150,15) \text{ and } (150,20). \)

When \( x = 0 \) our results are in very good accordance with Uyenoyama’s: average scaled ratios are close to 1, except \( \mathbb{E}(R_{SD}) \approx 1.4 \). Moreover, \( \mathbb{E}(R_{PT}), \mathbb{E}(R_{ST}), \mathbb{E}(R_{SD}) \) and \( \mathbb{E}(R_{BD}) \) are independent of the population and sample sizes, as also pointed out by Uyenoyama.

As \( x \) increases on the contrary, the mean scaled ratios become very sensitive to sample size, and to a lesser extent to population size. Given the coalescence times in the tree, the effect of imbalance is an increase of \( S \) since external branches connect to the root with a higher probability, and a relative decrease of \( P \) since the number pairs whose MRCA coincides with the sample’s MRCA decreases when imbalance increases. These remarks suffice to explain the behavior of the scaled ratios under fertility correlation, except for \( R_{BD} \): although the higher frequency of unbalanced trees yields longer base branches, the correlation and inhomogenous reduction of coalescence times dominates and induces a decrease of \( B \) relative to \( D \).

The general trend of the above statistics is very close to that described by Schierup and Hein (2000) for expanding populations. And yet, theoretical behavior of these scaled ratios is unknown and from the results presented here it is legitimate to investigate whether the observed variations are intrinsic properties since they do not allow us to differentiate fertility correlation from population growth (where both involve the whole genome in the haploid and diploid cases), recombination or complicated departure from neutrality.

5. DISCUSSION

In the previous sections we showed how to simulate a general exchangeable haploid Wright–Fisher model and model fertility correlation by a correlation between sibship sizes and propensities to reproduce. Our main result is that the process under fertility correlation is not equivalent to a standard exchangeable model with the same stationary progeny number distribution: although stationary distributions are exchangeable, the fact that progeny numbers are not independent from one generation to the next violates exchangeable models assumptions (Möhle, 2000) and yields unexpected results. In particular, the genealogies of samples of genes (or individuals) are deeply affected by fertility correlation. The effect is not only on the expected time until the MRCA as shown by Campbell (1999), but on both the topologies of the trees and the distribution of coalescence times (or branch lengths). The typical shape of coalescent trees under fertility correlation is characterized by a shrunk height, correlated and heterogeneously reduced coalescence times, and by a strongly unbalanced topology. All these characteristics constitute major departures from the shape expected under the standard the so-called “neutral” coalescent for populations of constant size (Kingman, 1982b; Tavaré, 1984).

5.1. Fertility Correlation Mimicks Population Expansion

To what extent fertility correlation and population expansion are similar in their effects is a legitimate question since both phenomena affect the whole genome and yield correlated and inhomogeneously reduced coalescence times: the reduction is higher for ultimate coalescences than for earlier ones yielding relatively longer external branches. This is well documented when effective population size decreases backward in time, i.e., in the case of population expansion (Slatkin and Hudson, 1991). In the case of fertility correlation, a dramatic decrease of effective population size has already been reported by Austerlitz and Heyer (1998) and is confirmed by our results concerning the \( T_{MRCA} \) (Fig. 3). The consequence of expansion on Uyenoyama’s scaled ratios was more precisely investigated by Schierup and Hein (2000). It is interesting to notice that our results are very similar to theirs, making it practically impossible to distinguish fertility correlation in stationary population and population expansion from the study of scaled ratios only.
From this we can draw the conclusion that scaled ratios are not sensitive to tree imbalance, the main discrepancy between gene genealogies in our case and under population expansion. Hence these statistics are inappropriate to detect even strong fertility correlation alone, but they could prove very useful if associated with some statistics sensitive to tree imbalance to discriminate between population growth and fertility correlation.

5.2. Fertility Correlation and Population Structure

Population models with different subpopulations (also called classes or types) constitute the cornerstone of coalescence models applied to spatial structure (Nordborg, 1997; Wilkinson-Herbots, 1998) or selection (Hudson and Kaplan, 1995). For example, Nordborg (1997) uses separation of time scales in the coalescent process to derive the average coalescence time for two lineages, in various models of migration and selection. The underlying assumption is that migration rates between classes are either slow or fast compared to coalescence times within classes. Different approximations apply in these two cases.

Unfortunately, unless fertilities are of the same order of magnitude—and therefore fertility correlation is very low—the interpretation of our correlated Wright–Fisher model in terms of such a structured coalescent is not possible for at least two reasons. First of all because even for a relatively small number of classes, the class sizes are not of the same order: individuals in classes of high fertility are rare. The second reason is that migration rates between classes are not all of order \( N^{-1} \); migration rates to classes of lower fertility are of order 1 typically. As a result, it is not possible to discriminate between slow and fast migration rates and to decide what approximation should apply. Moreover, the above observations imply that in the case of major fertility discrepancies among classes, the sizes of low fertility classes are almost constant but high fertility classes undergo fluctuations in size as large as the mean size itself.

The general trend of our results is that changes are relatively small for \( z \leq 1 \) but increase rapidly as \( z \geq 1 \) (see \( Z(a, z) \) for example). This is most probably due to the choice of \( h_z \); almost constant for small values of \( z \) and rapidly increasing for \( z > 1 \). We conclude that the effect of fertility correlation is mostly to classes of high fertility, and that fluctuations of their sizes are essential. Hence our model cannot be reasonably approximated by constant size classes and for all these reasons populations under fertility correlation cannot be thought of as standard structured populations.

5.3. Fertility Correlation and Fertility Selection

Nordborg (1997) predicts the level of reduction in coalescence time under background selection for one deleterious locus and extends this result to several loci acting multiplicatively. Even if classes were allowed to fluctuate in size, comparison with our model would not be straightforward. In order to predict the effects of fertility correlation in diploid populations, we must stress that our model is purely demographic, in the sense that its effects on gene genealogies are independent of the genome. As a consequence, above observations apply to the genealogy of any region in the genome, and do not depend on linkage to a locus or a region under selection.

We can interpret our results concerning the time to the most recent common ancestor (\( t_{MRCA} \)) under fertility correlation in terms of reduction of population effective size, as is done by Nordborg (1997) for background selection. By analogy with the standard coalescent, we can write \( t_{MRCA} = 2N_e(1 - \frac{1}{2}) \) and quantify the reduction in effective size:

\[
\frac{N_e}{N} = \frac{1}{(V(a, 0)(1 + A(a)V(a, 0)g(z)))},
\]

with our notations. This quantity is related to the reduction in genetic diversity due to fertility correlation. Our simulation showed that \( N_e/N \) can be less than 0.1 yielding a reduction in genetic diversity of more than 90%. These figures agree with Hudson and Kaplan’s (1995) results concerning background selection for relatively high mutation rate at the deleterious zone, and neutral loci at a small recombination distance from it. Nevertheless, their conclusion concerning the invariance of neutral variants frequency spectrum under background selection is in contradiction with our results. The fact that gene trees under fertility correlation have longer external branches allows to predict an excess of rare mutations in our model.

Another approach to coalescence with selection is now known as the ancestral selection graph (Krone and Neuhauser, 1997). The underlying hypothesis of this model is that mutation rates and selection coefficients are constant in time and of order \( O(1/N) \). Our simulations give rise to propensities to reproduce that are \( O(1/N) \) on average only: these propensities vary stochastically and are far less correlated from one generation to the next than selection coefficients in the ancestral selection graph. Hence comparison of both models is difficult. Przeworski et al. (1999) simulated coalescent trees under weak selection, that is, ancestral selection graphs once pruned of their virtual branches.
They studied the effect of weak selection on the $T_{MRCA}$ and the frequency of trees with an external branch connecting to the root. Their results show that the former variable depends slightly on the selection coefficient ($\sigma \in [0, 6]$) and that the latter statistic is almost independent of $\sigma$. This is not contradictory with our results since for very small values of $\sigma$ the decrease of $T_{MRCA}$ is slight and the frequency of trees with one branch connecting to the root is almost constant (results not shown). Hence very small values of $\sigma$ may be consistent with results obtained for coalescence under weak selection. In other words, this means that loose fertility correlation is comparable to fertility selection with low selection coefficients. Nevertheless, in real cases, values of $\sigma$ may be high: for example in the Saguenay Lac-Saint Jean population (E. Heyer, unpublished) for which estimates of $\sigma$ are of order 1. In this range of the parameter, the mean $T_{MRCA}$ is reduced to a fifth of its theoretical value without correlation, and the frequency of trees with one external branch connecting to the root is almost doubled.

Santiago and Caballero (1998) focus on directional selection. Their quantitative results concerning the magnitude of the reduction in effective size for neutral loci closely linked to the locus under selection are comparable with our results for the entire genome. The authors also note that the reduction is not uniform in time for a given mutation: old mutations are more affected than younger ones. Although the above results do not allow us to compare our model with their predictions, we expect further simulations to clarify this point. Finally, Santiago and Caballero’s (1998) formalism suggests the decomposition of the spectrum of gene frequencies into an accumulation of contributions by previous generations. Their main conclusion is that the resulting spectrum cannot be explained by a single value of $N_e$ under a neutral model, indirectly corroborating present results.

### 5.4. Comparison with Other Models of Fertility Correlation

Intuitively, in the presence of correlation, alleles carried by highly reproductive lineages are more likely to reach fixation and the time to fixation is likely to be shorter than in the absence of correlation. Austerlitz and Heyer (1998) showed that in the former situation the $T_{MRCA}$-based estimator of the effective population size decreases dramatically, namely by a factor of 17 in the case they studied (Human population of Saguenay-Lac Saint Jean). The offspring number correlation they measured from genealogical data ($\rho_o$) lies in the range 0.2–0.3. In our simulations, such values are never obtained when $a = 1$, which implies that this value of $a$ is not appropriate to fit Austerlitz and Heyer’s data. This is not at all paradoxical since offspring number distribution in the Saguenay-Lac Saint Jean population is not geometric and has a variance smaller than 2 when haploid lineages are considered (E. Heyer, unpublished). Furthermore, $\rho_o = 0.3$ corresponds to $V_o = 1.2$ for $a = \infty$ and $V_o \geq 2.5$ for $a = 2$. The respective reductions of $T_{MRCA}$ are 3 and more than 10 in our model. The discrepancy with results obtained for the Quebec population can be explained by the fact that contrary to our constant size haploid model, Austerlitz and Heyer simulated a conditioned branching process for an expanding diploid population.

Donnelly and Marjoram (1989) described a non-exchangeable extension of the Moran model by allowing the individual that has reproduced in the last time step to be more likely to reproduce again than any other individual. To some extent, their extension can be compared with ours: individuals with numerous offspring are more likely to reproduce and therefore become more frequent than in the corresponding exchangeable cases. However, this property is not inherited by offspring of highly reproductive individuals in Donnelly and Marjoram (1989) and, as a result, the correlation concerns only one individual in the population whereas in what is presented here, all individuals are simultaneously prone to fertility correlation. This may explain why Donnelly and Marjoram demonstrate analytically that, in the large population limit for both the neutral version and their extension of the Moran model, “when two lines of descent coalesce, it is equally likely to involve any of the existing lines.” This does not apply to our extension of the Wright–Fisher model.

Although we model fertility correlation for two successive generations, it is clear that offspring number correlation between ancestors and their descendants $G$ generations later tends to vanish exponentially with $G$ (Campbell, 1999). Hence aggregation of coalescences should be all the more effective as expected coalescence times are short, that is lineages more numerous or fertility correlation stronger. Since our results prove the existence of coalescence aggregation in the least favorable case, that is for the smallest possible number of lineages ($n = 3$), it appears that fertility correlation affects the tree as a whole, even on long branches.

Caballero (1994) developed a quantitative approach of the effect of fertility correlation on effective population size $N_e$. With our notations, $V_1 + V_2$ represents the observed variance in offspring number. As soon as correlation exists, this variance does not explain the
overall reduction in effective size (measured as reduction of the $T_{\text{MRCA}}$). That is, the underlying process is not equivalent to an exchangeable model with the same progeny number distribution. Indeed, a third term $V_3$ dominates, that may be interpreted as the long-term cumulative effect of correlation in previous generations, as in Caballero (1994).

5.5. Toward Simulating Coalescence under Fertility Correlation

If the effect of fertility correlation on coalescent trees was only a reduction of coalescence times, be it homogenous or not, the simulation of the genealogical process under fertility correlation would consist in the standard coalescent with a simple change in time scale, as is now common in the case of variable population size for example. But profound alterations of the shape described here are contradictory with one particular assumption making the simulation of Kingman's coalescent so powerful, namely the fact that all lineages are equally likely to coalesce, independently of their size (i.e., the number of individuals they subtend) or whether they were involved in the previous coalescence event.

To some extent, the aggregation of coalescences described here can be viewed as the *backward in time* counterpart of fertility correlation. In fact, coalescence events have a higher probability to occur in individuals with numerous offspring. Keeping this in mind, that sibship size be correlated implies reciprocally that successive ancestors of the lineage that has just coalesced are more likely to belong themselves to large families and therefore that the lineage has a higher probability to coalesce again. Also because coalescence is faster in large families, this event will occur more rapidly.

Our results concern small populations and it is not clear how they can be extended to larger ones. The model studied here cannot be simulated for populations as large as $10^5$ or so. Further investigations on this topic can either focus on an analytical description of the limiting genealogical process as $N$ tends to infinity, or indirectly, study the genetic diversity under fertility correlation. We conclude with broad ideas concerning these perspectives.

The coalescent process described by Kingman (1982b) is a particular case of a more general theory developed earlier by physicists and aimed at describing the evolution of systems of aggregating particles. In these models of coalescence or coagulation, the rate at which clusters form is not necessarily independent of their size or mass (for a review, see Aldous, 1999). On the contrary, the coalescent process familiar to population geneticists assumes that the rate of aggregation of lineages is size independent, i.e., that all lineages present at a given time have the same probability to coalesce, and coalescence events are considered independent of each other, which is obviously not the case under fertility correlation.

In particular, the fact that, under strong correlation, the probability of a lineage to coalesce again depends linearly on its size suggests that a more general coalescence process could describe the dual *backward in time* process of the correlated Wright–Fisher model studied here, provided that it can be made explicit. For example, the relevance of an additive coalescence model should be examined and if confirmed, would enable population geneticists to conceive simple algorithms for generating coalescent trees under fertility correlation.

Finally, fertility correlation can be demographically estimated only in the rare natural or artificial populations for which extensive genealogical data are available. It is therefore legitimate to simulate and study the effect of fertility correlation on genetic diversity and find criteria and tests to detect this phenomenon from genetic diversity records. This will be the subject of a forthcoming paper. Since fertility correlation is independent of the genotype and acts on the genome as a whole during short periods of time, major discrepancies with fertility selection or demographic bottleneck are again expected. For example, we can reasonably think that fertility correlation induces a departure from the Ewens sampling formula, contrary to Donnelly and Marjoram's (1989) conclusions for their correlated Moran model or Hudson and Kaplan's (1995) for background selection.

Appendix A

A.1. Compound Multinomial

In the general case (i.e., $(\lambda_1, \ldots, \lambda_N)$ not necessarily exchangeable) Mosiman (1962) gives results concerning the distribution of $(v_1, \ldots, v_N)$. Among them

$$\text{Cov}(v_i, v_j) = -N\mathbb{E}(\lambda_i)\mathbb{E}(\lambda_j) + N(N-1)\text{Cov}(\lambda_i, \lambda_j)$$

and an interesting asymptotic result stating that, as $N$ tends to infinity, all the moments of $(v_1, \ldots, v_N)/N$ tend to those of $(\lambda_1, \ldots, \lambda_N)$, in other words that the distribution of $(v_1, \ldots, v_N)/N$ approaches that of $(\lambda_1, \ldots, \lambda_N)$. 

\[ V(v_i) = N\mathbb{E}(\lambda_i)(1 - \mathbb{E}(\lambda_i)) + N(N-1)V(\lambda_i). \]
A.2. Exchangeable Distributions on $\Lambda_N$

The simulation of an exchangeable extension of the Wright–Fisher model requires to generate $N$ random numbers adding up to 1. This can be achieved by drawing $N$ random numbers i.i.d. and dividing them by their sum

$$\left(\hat{\lambda}_1, \ldots, \hat{\lambda}_N\right) = \left(\frac{\Lambda_1}{\sum_{i=1}^{N} \Lambda_i}, \ldots, \frac{\Lambda_1}{\sum_{i=1}^{N} \Lambda_i}\right).$$  \hspace{1cm} (24)

The r.v. $(\hat{\lambda}_1, \ldots, \hat{\lambda}_N)$ is obviously exchangeable. This symmetry implies that its moments are independent of the subscripts and especially $E(\hat{\lambda}_i)$, $V(\hat{\lambda}_i)$ and $\text{Cov}(\hat{\lambda}_i, \hat{\lambda}_j)$ are independent of $i$ and $j$. Hence

$$E(\hat{\lambda}_i) = \frac{1}{N},$$

$$\text{Cov}(\hat{\lambda}_i, \hat{\lambda}_j) = \frac{-1}{N-1} V(\hat{\lambda}_i) \text{ for } i \neq j,$$  \hspace{1cm} (25)

since $\sum_{i=1}^{N} \text{Cov}(\hat{\lambda}_i, \hat{\lambda}_j) = \text{Cov}(1, \hat{\lambda}_j) = 0$. The convergence in distribution of $N \hat{\lambda}_j$ to $\Lambda_j/E(\Lambda_i)$ results from the weak law of large numbers.

These results do not hold a priori if the $\Lambda_j$’s are not identically distributed.

ACKNOWLEDGMENTS

We thank Y. Vigouroux and R. C. Griffiths for useful comments on an earlier draft, and three anonymous reviewers for their valuable reports.

REFERENCES


Mosimann, J. E. 1962. On the compound multinomial distribution, the multivariate $\beta$-distribution, and correlations among proportions, Biometrika 49, 65–82.


