The Population Genetics of Clonal and Partially Clonal Diploids

François Balloux,*†,1 Laurent Lehmann‡ and Thierry de Meeûs§

*IC.A.P.B., University of Edinburgh, Edinburgh EH9 3JT, Scotland, United Kingdom; †Department of Genetics, University of Cambridge, Cambridge CB2 3EH, United Kingdom; ‡Institute of Ecology (Zoology and Animal Ecology), University of Lausanne, 1015 Lausanne, Switzerland and §Centre d’Étude du Polymorphisme des Microorganismes, Equipe ESS, UMR 9926 CNRS-IRD, BP64501, 34394 Montpellier Cedex 5, France

ABSTRACT

The consequences of variable rates of clonal reproduction on the population genetics of neutral markers are explored in diploid organisms within a subdivided population (island model). We use both analytical and stochastic simulation approaches. High rates of clonal reproduction will positively affect heterozygosity. As a consequence, nearly twice as many alleles per locus can be maintained and population differentiation estimated as $F_{ST}$ value is strongly decreased in purely clonal populations as compared to purely sexual ones. With increasing clonal reproduction, effective population size first slowly increases and then points toward extreme values when the reproductive system tends toward strict clonality. This reflects the fact that polymorphism is protected within individuals due to fixed heterozygosity. Contrarily, genotypic diversity smoothly decreases with increasing rates of clonal reproduction. Asexual populations thus maintain higher genetic diversity at each single locus but a lower number of different genotypes. Mixed clonal/sexual reproduction is nearly indistinguishable from strict sexual reproduction as long as the proportion of clonal reproduction is not strongly predominant for all quantities investigated, except for genotypic diversities (both at individual loci and over multiple loci).

The essential feature of sexual reproduction is that genetic material from different ancestors is brought together in a single individual. If sexual reproduction is dominant in eukaryotic organisms (e.g., Charlesworth 1989; West et al. 1999), many organisms of major medical or economical importance are known to reproduce mainly or strictly clonally (e.g., Milgroom 1996; Taylor et al. 1999; Tibayrenc 1999). The presence or absence of a sexual process will crucially determine the genetics at both the individual and the population level and leads to several straightforward predictions. At the individual level, clonality will produce a strong correlation between alleles within individuals at different loci, as they share a common history within a clonal lineage. Sex on the other hand will break these associations, allowing for many more potential genetic combinations. Further, in diploids, absence of sex will promote divergence between alleles within loci, as the two copies will accumulate different mutations over time. This effect has been termed the “Meselson effect” and has recently been experimentally documented in bdelloid rotifers, which are believed to have been reproducing strictly clonally over long evolutionary time (Butlin 2000; Mark Welch and Meselson 2000, 2001). Heterozygosity is thus expected to increase indefinitely under clonal propagation (Birky 1996; Judson and Normark 1996).

In another respect, theoretical considerations predict that the effective population size of clonal organisms should be lower than that of panmictic ones (e.g., Orive 1993; Milgroom 1996). However, the few theoretical population genetics studies that we are aware of provide ambiguous conclusions on that topic (Orive 1993; Berg and Lascoux 2000) and numerous field observations support this ambiguity (e.g., Butlin et al. 1998; Gabrielsen and Brochmann 1998; Cywinska and Hebert 2002). Thus, “whether organisms with clonal reproduction necessarily have lower genetic diversity is unclear” (Orive 1993, p. 337). These ambiguities illustrate what little is known on the population genetics consequences of clonal reproduction. In the absence of theoretical models providing clear expectations, estimating the rate of clonal reproduction in natural populations appears problematic (e.g., Anderson and Kohn 1998) and even the detection of purely clonal populations is often controversial (e.g., Tibayrenc 1997; Vigalys et al. 1997).

Clonality is not just an academic matter (Tibayrenc 1997). Many diploid organisms believed to reproduce mainly or strictly clonally are of major medical, veterinary, and economical importance, including pathogenic fungi such as Candida or protozoans such as Trypanosoma. A better understanding of the reproductive system of such organisms might be crucial for planning successful long-term drug administration or vaccination programs (Tibayrenc et al. 1991; Milgroom 1996; Taylor et al. 1999).

Here we present both analytical and stochastic simula-
tion results for the population genetics of clonally and partially clonally reproducing populations. We focus on a simple population subdivision model (island model) and restrict our work to neutral mutations. We derive the identities by descent, \( F \)-statistics, and mean coalescence times of alleles and genotypes for variable rates of clonal reproduction. We also investigate the allelic and genotypic diversities maintained under different rates of clonal reproduction.

MODEL ASSUMPTIONS AND GENETIC IDENTITIES

We consider a subdivided monoecious population of diploid individuals, which reproduce clonally with probability \( c \), with sexual reproduction occurring at the complementary probability \((1 - c)\). Sexual reproduction in the model follows random union of gametes, self-fertilization occurs at a rate \( s \), and a subpopulation is composed of \(N\) number of adults. In our model, individuals, rather than gametes, migrate following an island model (Wright 1951) at a rate \( m \), implying that a migrant has an equal probability to reach any of the subpopulations. We further assume stable census sizes and population structure and no selection. The life cycle involves nonoverlapping generations and juvenile migration. The precise sequence goes as follows:

1. Adult reproduction and subsequent death
2. Juvenile dispersal
3. Regulation of juveniles, the survivors reaching adulthood

Because of the symmetry of the island model, only the following probabilities of identity by descent are needed to describe the apportionment of genetic variation in a subdivided monoecious population.

\( F \): The inbreeding coefficient, defining the probability that two alleles drawn at random from a single individual are identical by descent.

\( \theta \): Coancestry of individuals drawn at random from within the same subpopulation, defined as the probability that two randomly sampled alleles from two different individuals within a subpopulation are identical by descent.

\( \alpha \): Coancestry of individuals randomly drawn from different populations. This is defined as the probability that two randomly sampled alleles from two individuals in different subpopulations are identical by descent.

The identities may be calculated in juveniles (\( F_{i}, \theta_{i}, \alpha_{i} \)), or adults (\( F_{A}, \theta_{A}, \alpha_{A} \)), or respectively before or after migration. In a first step, we express identities between adults one generation forward in time \((t + 1)\) as functions of juvenile identities \((t + \frac{1}{2})\). Adult identities are affected only by dispersal,

\[
F_{A(t+1)} = F_{\frac{1}{2}(t+1/2)}
\]

\[
\theta_{A(t+1)} = (q \theta_{\frac{1}{2}(t+1/2)} + (1 - q) \alpha_{\frac{1}{2}(t+1/2)})
\]

\[
\alpha_{A(t+1)} = (q \theta_{\frac{1}{2}(t+1/2)} + (1 - q) \alpha_{\frac{1}{2}(t+1/2)}),
\]

with \( q \) being the probabilities that two individuals taken at random within the same subpopulation after migration were born in the same deme. The exact expression for \( q \) is relatively cumbersome (see Wang 1997). However, for relatively large values of \( N, q \) reduces to the much more compact form that we use throughout the article,

\[
q \approx (1 - m)^2 + \frac{(m)^2}{n - 1},
\]

where \( m \) represents the migration rate and \( n \) the number of subpopulations. Now we can define \( q_{s} \) as the probability that two individuals sampled after migration in different subpopulations originated from the same deme:

\[
q_{s} = \frac{1 - q}{n - 1}.
\]

We then express juvenile identities as functions of adult identities in the previous generation. Here both mutation and the reproductive system will affect the genetic identities of juveniles. The mutation rate is \( u \) for all alleles and therefore the probability of two alleles that are identical by descent before mutation still being identical after mutation will be \( \gamma = (1 - u)^2 \). In the absence of any mutation event, clonal reproduction occurring at rate \( c \) will produce offspring identical to its progenitor, so that the inbreeding coefficient of a clonally produced juvenile individual will be identical to its parent’s. Selfing occurs with probability \( s \), and in that case the coancestry will be \((1 + F_{c})/2\). With a probability \( 1 - s \), nonselfing sexual reproduction occurs, the offspring will have two parents, and its inbreeding will be the parental coancestry \((\theta_{i})\). This gives us the following juvenile identities as functions of adult identities:

\[
F_{\frac{1}{2}(t+1/2)} = \gamma \left( c F_{A(t)} + (1 - \delta) \left( \frac{1 + F_{A(t)}}{2} + (1 - s) \theta_{A(t)} \right) \right)
\]

\[
\theta_{\frac{1}{2}(t+1/2)} = \gamma \left( \frac{1}{N} \left( \frac{1 + F_{A(t)}}{2} \right) + (1 - \frac{1}{N}) \theta_{A(t)} \right)
\]

\[
\alpha_{\frac{1}{2}(t+1/2)} = \gamma (\alpha_{A(t)}).
\]

Substituting Equation 4 in (1), we obtain the recurrence equations for describing the dynamics of identities among adults:

\[
F_{A(t+1)} = \gamma \left( c F_{A(t)} + (1 - \delta) \left( \frac{1 + F_{A(t)}}{2} + (1 - s) \theta_{A(t)} \right) \right)
\]

\[
\theta_{A(t+1)} = \gamma \left( q \left( \frac{1}{N} \left( \frac{1 + F_{A(t)}}{2} \right) + (1 - \frac{1}{N}) \theta_{A(t)} \right) + (1 - q) \alpha_{A(t)} \right)
\]

\[
\alpha_{A(t+1)} = \gamma \left( q \left( \frac{1}{N} \left( \frac{1 + F_{A(t)}}{2} \right) + (1 - \frac{1}{N}) \theta_{A(t)} \right) + (1 - q) \alpha_{A(t)} \right).
\]
The recurrence equations for $\theta_\lambda$ and $\alpha_\lambda$ are identical to those given by Rousett (1996, Equation 2). Only $F_s$ is affected by the variable amount of clonal reproduction and by the fact that we assume zygotic rather than gametic migration.

For analytical effectiveness, recurrence equations for identities by descent can be presented in matrix form,

$$Q_{t+1} = \gamma G Q_t + \gamma D,$$

where $Q_t$ is a column vector of the probabilities of identities at generation $t+1$. The transition matrix $G$ defines the probabilistic changes of the vector variables, and $D$ is the constant column vector. Solving (6) at equilibrium we obtain the identities from

$$Q = \gamma (I - \gamma G)^{-1} D$$

with $I$ being the identity matrix.

**INDIVIDUAL-BASED SIMULATIONS**

To obtain the variances of the quantities of interest, as well as multilocus behavior, we additionally performed stochastic individual-based simulation, as implemented in the software EASYPOP (version 1.7.4; Balloux 2001). For all simulations, we used 20 loci with a mutation rate of $10^{-5}$. Mutations had an equivalent probability to generate any of the 99 possible allelic states. This relatively high number of allelic states keeps the probability of obtaining indistinguishable alleles through different mutational events (homoplasy) low. At the start of the simulation, genetic diversity was set to the maximum possible value at the first generation and the simulation was then run for 10,000 generations, the point at which all statistics measured in EASYPOP ($F_s, F_{ST}, H_s, H_t$, and the number of alleles) had reached equilibrium. All simulations were replicated 20 times.

**$F$-STATISTICS**

Deviations from random mating are generally expressed by means of $F$-statistics (Wright 1951). They are the most commonly used tools for describing gene flow and breeding structure in both theoretical and empirical studies (reviewed in Balloux and Lugon-Moulin 2002). $F$-statistics are defined as

$$F_s = \frac{F - \theta}{1 - \theta}, \quad F_{ST} = \frac{\theta - \alpha}{1 - \alpha}, \quad F_{IT} = \frac{F - \alpha}{1 - \alpha}$$

(Cockerham 1969, 1973), where subscripts $I$, $S$, and $T$ represent individuals, subpopulations, and the total population, respectively. $F_s$ can be thought of as a measure of the identity of alleles within individuals relative to the identity between alleles randomly drawn from two different individuals from within the same subpopulation. $F_{ST}$ is the identity of alleles drawn randomly from within a subpopulation relative to alleles drawn from the entire population. $F_{IT}$ is the identity of alleles within individuals relative to randomly drawn gametes from the entire population. In more biological terms $F_s$ is interpreted in terms of deviation from random mating, caused by the breeding system of the organism under study, and $F_{ST}$ represents the heterozygote deficiency due to population subdivision. Finally $F_{IT}$ is the measure of inbreeding taking into account both deviations from random mating within subpopulations and the effects of population subdivision. The relation linking the three coefficients can be expressed as

$$(1 - F_{IT}) = (1 - F_{ST})(1 - F_s)$$


**Within-population deviations from random mating ($F_s$):** Replacing the solutions of Equation 7 in (8), we get $F_s$ after migration for subdivided populations with a mixed system of clonal and sexual reproduction (selling set to $1/N$) and zygotic migration

$$F_s = \frac{\gamma(q - c(\gamma(q - q_s) - 1) - 1)}{2N(1 - \gamma)(\gamma(q - q_s) - 1) - \gamma(q_s - c(\gamma(q - q_s) - 1) - 1)}.$$

(10)

Neglecting mutation ($\gamma = 1$), but allowing for a mixed system of clonal reproduction with arbitrary selfing rate, we obtain

$$F_s = \frac{\frac{Nn(1 - \epsilon) - 1}{Nn(1 - \epsilon)(2 - s) + 1}}.$$  

The equation shows that $F_s$ is independent of the migration rate but sensitive to the total number of individuals in the population; this occurs because we assumed zygotic rather than gametic migration. Under random mating ($s = 1/N$) we further obtain

$$F_s = \frac{(1 - \epsilon)n - 1}{(1 - \epsilon)(2N - 1)n + 1}.$$  

When reproduction is strictly sexual ($\epsilon = 0$), Equation 12 reduces to the form

$$F_s = \frac{n - 1}{(2N - 1)n + 1}.$$  

(13)

For a strictly clonal population ($\epsilon = 1$), $F_s = -1$. This reflects the fact that in the absence of sexual reproduction, all individuals are expected to be heterozygous at equilibrium $F = 0$, while $\theta = \gamma_s$

In Figure 1, we plot $F_s$ as obtained from Equation 10 against the rate of clonal reproduction. We also give values obtained from individual-based simulations. Analytical and stochastic simulation results are in excellent agreement. From Figure 1, it can be seen that for very high values of clonal reproduction, huge heterozygote excesses are obtained. However, as long as there is a small proportion of sexual reproduction, $F_s$ stays close to what is expected under panmixia; a significant excess of heterozygotes occurs only for extreme rates of asexuality. As long as there is mutation in the system, $F_s$ cannot reach $-1$ even for strict clonality. If the product
of the number of individuals in the complete population \((nN)\) times the mutation rate is high, the \(F_{IS}\) value for complete clonality can be very much offset from \(-1\). The reason for this can be seen from Equation 8. Under limited proportions of sex, there is no noticeable effect, while \(\theta\) decreases with increasing mutation rate.

The \(F_{ST}\) estimates from the stochastic simulations in Figure 1 are averaged over loci and replicates and do not reveal anything about the strong influence of the rate of clonal reproduction on the variance over loci. This huge variation among loci, in particular for low rates of sexual reproduction, is illustrated by standard errors in \(F_{ST}\) (Figure 2). The lowest variations are obtained with pure clonality and with \(<95\%\) of clonality.

**Population differentiation \((F_{ST})\):** Again by replacing the solutions of Equation 7 in (8), we obtain \(F_{ST}\) for subdivided populations with a mixed system of clonal, selfing, and sexual reproduction after migration:

\[
F_{ST} = \frac{\gamma(1 - \gamma)(q - q_d)}{N(1 - \gamma)(2 - s)(1 - (q - q_d)) + \gamma(q - q_d) + q}, \tag{14}
\]

Neglecting mutation \((\gamma = 1)\) leads to

\[
F_{ST} = \frac{(1 - c)(q - q_d)}{N(1 - c)(2 - s)(1 - (q - q_d)) + c(q - q_d) + q}. \tag{15}
\]

Finally, if only sexual reproduction is allowed \((c = 0\) and \(s = 1/N)\), we get

\[
F_{ST} = \frac{(q - q_d)}{(2N - 1)(1 - (q - q_d)) + q}. \tag{16}
\]

In Figure 3, we plot \(F_{ST}\) as obtained from Equation 14 against the proportion of clonal reproduction, as well as values obtained from the individual-based simulations.

The amount of clonal reproduction has a strong effect on population differentiation. Whereas even for very limited proportions of sex, there is no noticeable effect, when reproduction tends toward strict clonality, \(F_{ST}\) is strongly reduced. Note that in the absence of any mutation, \(F_{ST}\) would be defined but equal to 0, as all the genetic variance is within individuals and none between individuals and subpopulations. In all simulated cases the between-loci variance of \(F_{ST}\) strongly increases with the proportion of clonal reproduction (results not shown).

**EFFECTIVE POPULATION SIZE**

**Effective population size:** The effective population size (Wright 1931) is the parameter summarizing the amount of genetic drift to which a population is subjected. It is quantified as the number of idealized randomly mating individuals that experience the same amount of random fluctuations at a neutral locus as the population under scrutiny. The dynamics of idealized randomly mating individuals are described by the Wright-Fisher model, whose well-known properties lead to different definitions of the effective population size depending on whether the quantities of interest are the variance of change in allelic frequencies, inbreeding coefficients, or the rate of decline in heterozygosity (Ewens 1982; Whitlock and Barton 1997). Here we introduce a new definition of effective size called the coalescence effective size,

\[
N_c = \frac{1}{2} \tilde{t}, \tag{17}
\]

where \(\tilde{t}\) is the expected time it takes for two randomly sampled alleles in a population to coalesce to a common
ancestor. For the Wright-Fisher model \( \bar{t} = 2N \), so that the effective size reduces to the actual number of diploid individuals. This definition of effective size allows us to disentangle the allelic effective size (all classical definitions) from the genotypic effective size (see below), and we can further obtain their variance.

There is a strict relationship between identity-by-descent probabilities and coalescence times (Slatkin 1991; Rousset 1996). The probability of identity of any pair of alleles is the probability that neither allele has undergone mutation since their most recent common ancestor (Hudson 1990). Recalling Equation 7,

\[
Q = \gamma (I - \gamma G)^{-1} D. \tag{18}
\]

The matrix \( G \) is diagonalizable for \( \gamma \neq 1 \). We can represent the vector \( D \) on the basis of the right eigenvectors of the matrix \( G \) as \( D = \sum_j a_j r_j \), where \( j \) is the number of columns of \( G \), \( a_j \) the coefficient determined by the preceding system of equations, and \( r_j = (r_{i,j}, \ldots, r_{n,j})^T \) the \( j \)th right eigenvector of \( G \). Using the fact that the \( j \)th eigenvalue of the matrix \((I - \gamma G)^{-1}\) is \( 1/(1 - \gamma \lambda_j) \), and its associated right eigenvector is \( r_j \), where \( \lambda_j \) is the \( j \)th eigenvalue of \( G \), we can express Equation 18 following Rousset (2002) as

\[
Q = \sum_{j=1}^n \frac{\gamma a_j r_{j,1}}{1 - \gamma \lambda_j} = \sum_{j=1}^n \gamma \sum_{j=1}^n \lambda_j^{1-1} a_j r_{j,1}. \tag{19}
\]

The second equality is obtained by using the property of geometric series. Then

\[
e_i(t) = \sum_j \lambda_j^{t-1} a_j r_{j,i} \tag{20}
\]

is the coalescence probability of alleles at time \( t \) at any hierarchical level \( i \) (where \( i \) stands for \( F, \theta \), and \( \alpha \)). From this, we can obtain the expected coalescence times by classical tools. However, after substituting Equation 20 into Equation 19, a closer look reveals that the vector \( Q \) defines the probability-generating functions of coalescence time at each level \( i \). These functions reduce to the calculations of expected coalescence times as

\[
\bar{t}_i = \frac{\partial Q}{\partial \gamma} \bigg|_{\gamma=1}, \tag{21}
\]

where \( \bar{t}_i \) is the expected coalescence time at level \( i \) and \( Q \) is the \( i \)th row of the equilibrium vector given by Equation 20 and their variances as

\[
\sigma^2(t_i) = \left[ \frac{\partial^2 Q}{\partial \gamma^2} + \frac{\partial Q}{\partial \gamma} \right]_{\gamma=1}. \tag{22}
\]

At this point we have all necessary tools to obtain the mean coalescence times of alleles in a subdivided population with arbitrary rates of clonal reproduction. Writing the recurrence equations (5) under the form given in Equation 18 and using Equation 21 yields the following mean coalescence times,

\[
\bar{t}_i = \frac{2(1 + (1 - \bar{c}) nN(1 - s))}{1 - \bar{c}} \tag{23}
\]

where \( \bar{t}_i \) is a low migration limit obtained by a first-degree Taylor expansion. The mean coalescence time of two randomly sampled alleles is the expectation of the \( \bar{t}_i \); in the finite-island model this yields

\[
\bar{t} = \frac{1}{nN} \bar{t}_1 + \left( 1 - \frac{1}{N} \right) \bar{t}_u + \left( 1 - \frac{1}{n} \right) \bar{t}_u. \tag{24}
\]

Substituting Equation 24 into (17), we obtain the coalescence effective population size. Note that this effective size captures the loss of allelic diversity in the population and we refer to it as \( 2N_e \), the allelic effective population size, which is equal to \( \bar{t} \) in our model. In Figure 4, we plot the effective population size as a function of clonal reproduction and selling rate (the union of gametes within individuals). Increasing the rate of clonal reproduction has no noticeable effect on most of the parameter space. However, when the reproductive system tends toward complete asexual reproduction, the effective population size suddenly tends toward infinite values. This slightly counter-intuitive result simply reflects that the genetic diversity within individuals cannot be lost in clonal organisms. Doubling of \( N_e \) compared to random mating is observed approximately when the rate of sexual reproduction is in the order of \( 10^{-4} \) with the simulation parameters used in Figure 4. Contrarily, increased rates of selling decrease effective population size. This effect is linear and the effective population size ranges
between \( N_i \) under absence of selfing to \( N_i/2 \) for strict selfing.

**Genotypic and allelic effective population size:** We have shown that increased rates of clonal reproduction will increase the allelic effective population size, and thus clonal populations are expected to maintain more alleles at neutral loci than are sexually reproducing ones. We can go a step further and address the issue of how clonal reproduction will affect the number of different genotypes maintained. The coalescence approach allows us to capture qualitatively these trends by calculating the *genotypic effective population size*. To obtain this quantity, we need, in addition to \( F \) and \( \theta \), the probabilities that three alleles randomly sampled in two different individuals are identical. These three variables are necessary to calculate the probability \( \Delta \) that two genotypes are identical. However, these higher-order coefficients are complicated and we therefore limit ourselves to a non-subdivided monoeconomic population without mutation. We follow the approach of Cockerham’s (1971, pp. 243–244) to calculate the dynamics of these four variables. Collecting the identities given in Appendix A leads to the following system of recurrence equations:

\[
F_{\alpha Neville} = cF_{\alpha} + (1 - c)\left(\frac{1}{N} + F_{\alpha} \right) + \left(1 - \frac{N}{4}\right)\theta_{\alpha}.
\]

\[
\theta_{\alpha Neville} = \left(\frac{1}{N} + E_{\alpha}\right) + \left(1 - \frac{N}{4}\right)\theta_{\theta}.
\]

\[
\rho_{\alpha Neville} = c\left(\frac{1}{N} + E_{\alpha}\right) + \left(1 - \frac{N}{4}\right)\rho_{\rho}.
\]

\[
\Delta_{\alpha Neville} = c\left(\frac{1}{N} + E_{\alpha}\right) + \left(1 - \frac{N}{4}\right)\Delta_{\Delta} + 2(1 - c)\left(\frac{1}{N} + E_{\alpha}\right) + \frac{3(N - 1)}{2N^2}\theta_{\theta} + \frac{3(N - 1)}{2N^2}\rho_{\rho} + \frac{1}{4}\Delta_{\Delta}.
\]

\[
\xi = \frac{(N - 1)(N - 2)\Delta_{\Delta}}{N^3} + (1 - c)\left(\frac{1}{4N^2} + \frac{3}{4N^2}F_{\alpha} + \frac{2(2N - 1) - \theta_{\theta} + (2N - 1)(2N - 2)\rho_{\rho}}{(2N)^2} + \frac{2(N - 1)(2N - 2)\rho_{\rho}}{(2N)^2}\Delta_{\Delta}.
\]

Note that when \( c = 0, F = \theta, \) and the recurrence equations reduce to Cockerham’s (1971) model. Substituting these equations into a transition matrix \( G \) and a column vector of constants \( D \) following Equation 18 allows us to obtain the mean coalescence times:

\[
\bar{\xi} = \frac{2(c + N(1 - c))}{1 - c}
\]

\[
\bar{\theta} = \frac{c + 2N(1 - c)}{1 - c}
\]

\[
\bar{\rho} = \frac{N(8N + 2c(N - 1) + c(9 - 10N) - 3)}{(1 - c)(3N - c(N - 1) - 1)}
\]

\[
\bar{\Delta} = \begin{cases} c^2a_0 + c^2a_1 + c^2a_2 + a_3 & \text{for } c 
eq 1 \\ c^2b_0 + c^2b_1 + c^2b_2 + b_3 & \text{for } c = 1. \end{cases}
\]

\( \bar{\xi} \) is the mean coalescence time for genotypes and we refer to it as the genotypic effective size. This quantity is undefined for \( c = 1 \) as the matrix \( G \) is not diagonalizable in this case. However, we know that when \( c = 1 \), the genotypic identity is independent of \( F, \theta, \) and \( \rho \) and thus follows a dynamic similar to haploid genetics, with mean coalescence time \( N \). When \( c \neq 1 \), the mean genotypic coalescence time \( \bar{\xi} \) takes the form of a relatively complex polynomial (the coefficients are given in Appendix B). Note that in the absence of clonal reproduction, there is a very compact approximation for the genotypic coalescence time \( \bar{\xi} = 3N \) (see Appendix B). Mean coalescence time for alleles in a nonsubdivided population can be obtained as \( \bar{T} = (1/N)\bar{\xi} + (1 - 1/N)\bar{\theta} \) and thus reads

\[
\bar{T} = \frac{c(1 + N) + 2N^2(1 - c)}{N(1 - c)}.
\]

Note that we could have obtained the allelic effective size directly from Equation 24 assuming no migration, one subpopulation, and a selfing rate of \( 1/N \). In Figure 5, we give mean coalescence times for both alleles and genotypes. It can be seen that contrary to what is observed at the allelic level, genotypic effective size decreases with increasing clonality. The decrease is relatively smooth over the complete parameter range of \( c \) and reaches \( N \) for strict clonality. The intuitive reason behind this is that when there is no segregation at all, the two alleles within a diploid individual behave as a single haploid locus. The rate of clonal reproduction has thus an antagonistic effect on the variability of alleles and genotypes.
Figure 5.—Allelic effective population size (solid line) and genotypic effective population size (dashed line) in a non-subdivided population. The values given are for a population of 20 individuals.

GENETIC DIVERSITIES

We can now take a closer quantitative look at how genetic diversity is distributed between alleles and genotypes with the stochastic simulations. Allelic diversity can be expressed as the effective number of alleles, $n_e$, corresponding to the number of equally frequent alleles needed to observe a given genetic diversity, which is $1/(\sum p_i^2)$, where $p_i$ is the frequency of the $i$th allele. Similarly, we can express the effective number of genotypes as $G_e = 1/(\sum g_i^2)$, where $g_i$ is the frequency of the $i$th genotype. In Figure 6 we plot both the effective number of alleles and genotypes within a subpopulation. The number of alleles maintained is strongly positively affected when the reproductive system tends to be completely asexual. This effect is generated by fixed heterozygosity (i.e., under strict clonal reproduction in diploids the two alleles at each locus are behaving as two haploid loci). In contrast to allelic diversity, clonal reproduction decreases the effective number of genotypes steadily (Figure 6). To summarize, populations of clonal or subclonal organisms can maintain more allelic diversity at each single locus but fewer distinct multilocus genotypes.

DISCUSSION

We used both an analytical approach and stochastic individual-based simulations to describe the dynamics of genetic variance in subdivided populations, characterized by various levels of clonal reproduction. Higher rates of asexual reproduction will increase heterozygosity and decrease population differentiation. Diversity at single loci will be higher in clonal organisms than in sexuals, whereas the opposite is true for genotypic diversity. At the exception of genotypic diversity (both at single loci and over multiple loci), which decreases at a constant rate with increasing rates of asexual reproduction, all other quantities investigated are significantly affected only when sexual reproduction becomes rare.

Our results thus suggest that strict clonality may easily be detected in diploid populations due to heterozygote excess. Furthermore, very low levels of sex (cryptic sex) may also be revealed by on average low $F_{IS}$ values with very important variance among loci, though DNA alterations may also lead to a similar pattern in a strictly clonal population. For instance, Candida albicans is known to undergo mitotic recombinations including chromosomal translocation (Lott et al. 1999). Much effort has been put into testing for evidence of strict clonal reproduction with traditional population genetics (e.g., Tibayrenc et al. 1991) or through testing for the Meselson effect (high divergence at the two alleles of a single locus within individuals; reviewed in Butlin 2000). Extreme genetic divergence at single loci within individuals has been documented in bdelloid rotifers, which are believed to be ancient asexuals (Mark Welch and Meselson 2000, 2001). The Meselson effect could, however, not be detected in other potentially old asexual lineages (Schön et al. 1998; Normark 1999). Whether this is due to rare sex or perhaps to extremely frequent gene conversion events (the copy of the DNA sequence of one chromosome on the other) is an unresolved issue to date.

Empirical data on genetic variation and its apportionment by means of $F$-statistics in clonal lineages, as compared to sexually reproducing populations of the same species, are rare. Furthermore, studies using dominant genetic markers (e.g., rapidly amplified polymorphic DNAs) do not properly allow for the disentanglement between genetic variation within loci and within geno-
types. Indeed, as can be seen from Figure 6, the absolute genetic diversity (the sum of allelic and genotypic variability) does not provide any clear prediction on the rate of clonal reproduction. Another potential problem stems from the difficulty in ruling out the presence of rare sexual reproduction. However, a recent study by Delmotte et al. (2002) comparing eight sexual with five asexual populations of the aphid Rhopalosiphum padi could provide a test for our model. Their empirical results are overall in good agreement with our analytical expectations. As we expect, Delmotte et al. (2002) report increased excess in expected heterozygotes ($F_{st}$) for asexuals and lower differentiation ($F_{st}$) between asexual populations than between sexual populations. They also report lower genotypic variation and lower allelic variation in asexuals than in sexuals. This relatively good agreement between our model and their data suggests that these asexual populations have not experienced sexual reproduction in recent times. The discrepancy in allelic diversity could be due to different factors (e.g., sampling, extinction-recolonization dynamics). However, even if we assumed all else being equal between the sexual and asexual aphid populations, selection could still reduce genetic diversities more effectively in the asexual populations. Mutations under strong directional selection make linked loci behave as if they were evolving under smaller effective population sizes (Robertson 1961). Due to the complete absence of recombination in strictly clonal organisms, any strongly deleterious dominant mutation will drive the lineage, where it appeared, to extinction. New beneficial mutations will also reduce the effective population sizes of clones, as lineages with a new beneficial mutation will displace other lineages.

Indeed our model does not include natural selection, so that our results apply strictly to neutral genetic variability or more generally to relatively weakly selected polymorphisms subject to genetic drift. Genetic drift is the main force driving allele frequencies as long as the selection differential $s$ between alleles is not much above the inverse of effective population size ($1/N_e$). For higher selection differentials, the effect of genetic drift becomes negligible. However, our predictions should hold even for relatively important selection differentials in clonal and nearly clonal organisms, as the efficacy of selection acting simultaneously at linked sites is considerably reduced (Hill and Robertson 1966).

We assumed identical fitness (in both mean and variance) for clonally and sexually produced offspring. The rate of clonal reproduction is not a heritable trait in our model, as it is a fixed property of the population (clonally produced individuals do not have a higher chance to reproduce clonally themselves). Therefore, different fecundities for sexually or clonally produced offspring would result only in increasing the variance in reproductive success and thus would decrease the effective population size. Our results are thus qualitatively robust to reasonable differences in relative fitness between clonally and sexually produced offspring.

Finally, our model could lead to the development of new approaches to infer the rate of clonal reproduction. Our results show that all estimators based on identities by descent (including linkage disequilibrium approaches) are expected to be rather insensitive to the rate of clonal reproduction as long as it does not become strongly predominant. It is therefore doubtful that such estimators will allow precise inferences on the actual rate of clonal reproduction unless it is very close or equal to 1. As genotypic diversity decreases smoothly with the rate of clonal reproduction, one promising alternative approach would be to build estimators of clonal reproduction as functions of the relative genotypic and allelic identities.

We thank Nathalie Charbonnel, Sylvain Gandon, Jérôme Goudet, Andy Overall, Franck Prugnolle, François Renaud, Max Reuter, Denis Roze, Michel Tibayrenc, and two anonymous referees for very inspiring conversations and comments; François Rouset for having given access to unpublished material; and Sylvain de l’Hérald for his strong support. F.B. was supported by the Biotechnology and Biological Sciences Research Council and by grant 823A-066716 from the Swiss National Science Foundation.

LITERATURE CITED


Population Genetics of Clones

APPENDIX A: GENOTYPIC PROBABILITIES OF IDENTITIES BY DESCENT

We follow the same rationale as Cockerham (1971, pp. 242–243), but add the dynamics of \( \theta \). When one offspring is produced clonally, his two alleles are not independent. When we sample alleles and look back to their common parent, the two genes of a clone always stem from the same individual. Two clones are randomly sampled with probability \( \tilde{c} \); the four genes stem either from the same parent or from two different ones. The identity between genotypes and three alleles reads

\[
\Delta_{i+1} = \frac{1}{N} + \left( 1 - \frac{1}{N} \right) \Delta_i
\]

\[
\rho_{i+1} = \frac{1}{N} F_i + \left( 1 - \frac{1}{N} \right) \rho_i
\]  

(A1)

(two from one offspring, the third from the other).

When one offspring is produced clonally and one through random mating, the identity among three alleles will differ if two or only one allele stem from the clonal offspring, the former case occurs with probability \( c(1 - c) \) and the identity is as in Equation A1. The second case also occurs with probability \( c(1 - c) \), and the three genes then all stem from the same parent with probability \( P^3 \) and share identity \( \rho^3 \). The three genes might as well stem from two different parents with probability \( P^{c2} \) and then have identity \( \rho^{c2} \); finally, they can stem from three different parents with probability \( P^{111} \) and their identity is \( \rho^{111} \). The three types of identities can be expressed as

\[
\rho^{3} = \frac{1}{4} + \frac{3}{4} F_i, \quad \rho^{c2} = \frac{1}{2} \theta_i + \frac{1}{2} \rho_i, \quad \rho^{111} = \rho_i
\]

(A2.1)

For the identity between genotypes, we have with probability \( 2c(1 - c) \):

\[
\Delta_i^{k+1} = \frac{1}{2} + \frac{1}{2} F_i, \quad \Delta_i^{c2} = \frac{1}{2} \theta_i + \frac{1}{4} \rho_i + \frac{1}{4} \Delta_i, \quad \Delta_i^{111} = \Delta_i
\]

(A2.2)

When we sample two offspring issued from random mating with probability \( (1 - c)^2 \), the identities for \( \rho \) are given by Equation A2.1. The genotypic identity is obtained by summing \( \Delta_i^{k+1}, \Delta_i^{c2}, \Delta_i^{111}, \) and \( \Delta_i^{111} \), each identity weighted by its corresponding probability \( P^k \), \( P^{c2} \), \( P^{111} \), and \( P^{111} \). The five identities are written

\[
\Delta_i^{k+1} = \frac{1}{N} F_i, \quad \Delta_i^{c2} = \frac{1}{2} \theta_i + \frac{1}{4} \rho_i + \frac{3}{12} \Delta_i, \quad \Delta_i^{111} = \Delta_i
\]

(A3)

The different probabilities of gamete origins are given in Weir and Cockerham (1969) and read as

\[
\begin{align*}
p^1 &= \frac{1}{N^2}, & p^{c2} &= \frac{3(N - 1)}{N^2}, & p^{111} &= \frac{(N - 1)(N - 2)}{N^2}, \\
p^1 &= \frac{1}{N^2}, & p^{c2} &= \frac{3(N - 1)}{N^3}, & p^{111} &= \frac{4(N - 1)}{N^3}, \\
p^{c2} &= \frac{6(N - 1)(N - 2)}{N^3}, & p^{111} &= \frac{(N - 3)(N - 2)(N - 1)}{N^3}
\end{align*}
\]

(A4)
APPENDIX B: COEFFICIENT OF EQUATION 23

\[ a_0 = 2(N - 1)^2(6N^2 - 1) \]
\[ a_1 = -2(N - 1)(N(5N(2N + 3) - 19) + 2) \]
\[ a_2 = -2(1 + N(3N - 4)(N(14N - 17) + 6)) \]
\[ a_3 = 2N(N(50N^2 - 66N + 31) - 5) \quad \text{(B1.1)} \]
\[
\begin{align*}
  b_0 & = (N - 1)^2(2N + 3) \\
  b_1 & = -(19N^2 + 28N - 9) \\
  b_2 & = -(30N^3 - 65N^2 + 44N - 9)
\end{align*}
\]

If \( c = 0 \), then the mean coalescence time for two genotypes in the Wright-Fisher setting reduces to
\[
\tilde{\Delta} = \frac{2N(N(50N^2 - 66N + 31) - 5)}{(3N - 1)(N(12N - 11) + 3)}. \quad \text{(B2)}
\]
Performing a Taylor expansion of first degree under large population size and substituting some close integers yields the approximation for Equation B2:
\[
\tilde{\Delta} \approx 3N. \quad \text{(B3)}
\]