Available online at www.sciencedirect.com

SCIENCE DIRECT®

Infection, Genetics and Evolution 6 (2006) 163–170

Infection, Genetics and Evolution

www.elsevier.com/locate/meegid

Discussion

Molecular epidemiology of clonal diploids: A quick overview and a short DIY (do it yourself) notice

Thierry De Meeûs [a,*], Laurent Lehmann [b], François Balloux [c]

[a] *Génétique et Evolution des Maladies Infectieuses, Equipe Evolution des Systèmes Symbiotiques, UMR2724 CNRS-IRD, BP 64501, 911 Av. Agropolis, 34394 Montpellier Cedex 5, France*

[b] *Laboratory of Ecological and Evolutionary Dynamics, Department of Biological and Environmental Science, University of Helsinki, Niemenkatu 73, 15140 Lahti, Finland*

[c] *Department of Genetics, Downing Street, University of Cambridge, Cambridge CB2 3EH, UK.*

Received 15 July 2004; received in revised form 15 February 2005; accepted 20 February 2005
Available online 9 November 2005

Abstract

In this short review we report the basic notions needed for understanding the population genetics of clonal diploids. We focus on the consequences of clonality on the distribution of genetic diversity within individuals, between individuals and between populations. We then summarise how to detect clonality in mainly sexual populations, conversely, how to detect sexuality in mainly clonal populations and also how genetic differentiation between populations is affected by clonality in diploids. This information is then used for building recipes on how to analyse and interpret genetic polymorphism data in molecular epidemiology studies of clonal diploids.

Keywords: Clonality; Parthenogenesis; Population genetics; Linkage disequilibrium; Diploids; Heterozygosity; Population differentiation

1. Introduction

The population genetics of clonal organisms and its application in epidemiological studies has been the focus of much work and controversies as testified by the almost endless list of reviews on the topic (e.g. Suomolainen et al., 1976; Génermont, 1980; Silander, 1985; Tibayrenc et al., 1991; Carvalho, 1994; Tibayrenc, 1995; Milgroom, 1996; Judson and Normark, 1996; Milgroom and Fry, 1997; Milgroom, 1997; Anderson and Kohn, 1998; Tibayrenc, 1998; Taylor et al., 1999; Tibayrenc, 1999; Maynard-Smith et al., 2000; Tibayrenc and Ayala, 2002; Halkett et al., 2005). Facing such an impressive heap of literature one may ask what might be the point of another review dealing with the subject? Let us reassure the reader and make clear that our aim is not to cover once more the entirety of the topic, but rather to pinpoint some essential points and then rapidly move on to recent results that shed new light on the amount and apportionment of genetic variance expected in clonal and partially clonal diploids. We will also restrict us to clonal diploids, with emphasis on parasitic protozoa and pathogenic fungi. A recent more general review on the population genetics of clonal organisms can be found elsewhere (Halkett et al., 2005). We will conclude by an attempt to highlight the simple, robust analyses that optimise biological inference from genetic data in epidemiological studies on such organisms. The paper is subdivided into six short sections. In the first section we review the theoretical background, some old beliefs as well as recent advances. In the second section we will see how clonality can be detected in a mainly sexual population. The third section deals with means to uncover the presence of sex in essentially clonal populations. The fourth section dwells into the effect of clonality on population differentiation in diploids. In the fifth section we delineate a guide for drawing sound inferences from molecular co-dominant markers in clonal diploids, and in the final section (unsurprisingly) the conclusion, we briefly highlight some issues that shall be addressed by future research.

2. Theory of clonal genetics in diploids: old beliefs and new advances

2.1. The effective population size of diploid clones

The effective size of a population, usually designated by N_e, allows quantifying the rate at which a population looses its

[*] Corresponding author. Tel.: +33 467 41 63 10; fax: +33 467 41 62 99.
E-mail address: demeeus@mpl.ird.fr (T. De Meeûs).

1567-1348/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2005.02.004

genetic diversity. Indeed, the reciprocal of the effective size ($1/N_e$) gives the long-term probability that two randomly sampled genes in the population are replicates (or descend) from a single gene in the parental generation. Such repeated ‘‘coalescence events’’ of several genes in one gene imply that other genes do not contribute to the future of the population. Hence, genetic diversity is lost. The ratio of the actual census size N_c to the effective size N_e of a population is a measure of the dynamics of quantities linked to genetic diversity (e.g. heterozygosity) in the population under scrutiny compared to an ‘‘ideal’’ population. This ‘‘ideal’’ population is in fact a population that loses genetic diversity at rate ($1/N_c$) per generation so that its effective size is equal to its census size. Such a condition is met for populations of semelparous monoecious individuals mating at random and living in a constant environment with no selective pressure. As an example of a population, which loses genetic diversity faster than the ‘‘ideal’’ one, we can consider 100 dioecious individuals with uneven sex ratio. The effective population size of a herd of one bull ($N_m = 1$) and 99 cows ($N_f = 99$) yields $N_e = 4N_mN_f/N_c \approx 4$ (e.g. Hartl and Clark, 1989, p. 86), i.e. a 25-fold decrease compared to the census size ($N_c = N_f + N_m = 100$). Under such a scenario, genetic diversity is lost very rapidly. Other factors such as population subdivision might also be important; let us for example consider a population of total size $N_c = N_n = 10{,}000$, structured as an island model (Wright, 1951; Box 1) with $n = 1000$ demes of $N = 10$ individuals each, and where migration occurs at rate $m = 0.001$. This population will be characterized by a Wright’s (1965; Box 1) $F_{st} = 1/(4Nm + 1) \approx 0.96$, at equilibrium (e.g. Hartl and Clark, 1989, p. 310), translating into an effective population size of $N_e = N_c/(1 - F_{st}) \approx 26{,}000$ (Rousset, 2004, page 14), i.e. a 2.6-fold increase. Population structure thus slows down the rate of loss of genetic diversity.

While the influence on the effective size of factors such as sex ratio or population subdivision have been extensively analysed and are discussed in any population genetics textbook, the effect of clonality has been essentially neglected leading to statements such as: ‘‘the prevailing assumption has been that more genetic diversity is to be found within predominantly sexual populations and species’’ (Silander, 1985) or ‘‘strongly contrasting viewpoints have appeared in the literature’’ (Marshall and Weir, 1979). An illustration of this perplexity is culminating in the statement within the otherwise pioneering work by Orive (1993): ‘‘both of the clonal examples resulted in ratios of effective population size to census size that were generally lower than those published for non-clonal organisms. Whether this means that organisms with clonal reproduction necessarily have lower genetic diversity is unclear’’. Probably the main reason why this problem has been seen as challenging is that classical definitions of effective population size strictly focus on the loss of alleles. Under clonal reproduction, segregation and recombination, the two fundamental consequences of meiosis, are not realised, leading to an accumulation of heterozygosity at all loci and to an accumulation of identical multilocus genotypes in populations (see the next paragraph). The situation gets thus much clearer when disentangling between the amount of genetic diversity

Box 1. Basic definitions

The Island model (Wright, 1951): This model considers a population of individuals living in n demes (or sub-populations) each of finite size N. The life-cycle is usually assumed to be the following: (1) each adult individual produces independently a large number of juveniles. All adults die. (2) Each juvenile disperses randomly to another deme with probability m. With complementary probability ($1 - m$) a juvenile remains in its natal deme. (3) Regulation occurs, among all juvenile competing in a deme, only N individuals reaching adulthood.

Wright’s (1965) fixation indices: In a hierarchical population structure with two levels, such as the Island model described above, three fixation indices can be defined. F_{is} is a measure of the inbreeding of individuals resulting from the deviation from panmixia (random union of gametes) within each deme. F_{st} is a measure of the relatedness between individuals due to the structure of the population (non-random distribution of individuals among demes); F_{st} thus quantifies the differentiation between demes. Finally, F_{it} is a measure of the inbreeding of individuals resulting both from non-random union of gametes within demes and from population structure (deviation from panmixia of all individuals of the total population).

These fixation indices are generally defined (Nei, 1977; Cockerham, 1969, 1973; Rousset, 2004) as:

$$\begin{cases} F_{is} = \dfrac{F_i - F_s}{1 - F_s} \\ F_{st} = \dfrac{F_s - F_T}{1 - F_T} \\ F_{it} = \dfrac{F_i - F_T}{1 - F_T} \end{cases} \qquad \text{(a)}$$

where F_i is the probability that two homologous genes (e.g. the paternal and the maternal gene) drawn from an individual are identical, F_s is the probability that two randomly sampled genes from two different individuals within a sub-population are identical and finally, F_T is the probability that two randomly sampled genes from two individuals in two different sub-populations are identical. One can also define the fixation indices in terms of ‘‘heterozygosities’’ by using the relationship $H_i = 1 - F_i$ which are then substituted into Eq. (a). Finally, from Eq. (a) one can easily retrieve the classical equation $(1 - F_{it}) = (1 - F_{is})(1 - F_{st})$.

The value of F_{is} can vary between −1 (all individuals heterozygous for the same allele pair), 0 (random distribution of alleles within individuals) to +1 (all individuals are homozygous). By contrast, the value of F_{st} varies between 0 (random distribution of individuals between demes) to +1 (all demes fixed for one of the available alleles). Except when all individuals from all populations are sampled and genotyped, these F-statistics are biased if applied to real data. Weir and Cockerham (1984) have defined unbiased estimators of F-statistics: f for F_{is}, θ for F_{st} and F for F_{it}. The range of values taken by these estimates are the same as those of the parametric F-statistics, except for θ which can be negative when the different samples share allelic frequencies that are more similar than expected under the null hypothesis. Sampling error is indeed expected to generate some variance between samples drawn out of the same population.

Linkage disequilibrium measures: Linkage disequilibrium occurs when the different alleles at different loci are not randomly associated. Ideally, if two loci with two allele each (alleles *A* and *a* at the first locus, and alleles *B* and *b* at the second locus) are in linkage equilibrium, then the gamete *AB* should occur at frequency $P_{AB} = p_A p_B$ in the population, p_A and p_B being the allelic frequencies of *A* and *B*, respectively in the population. If not, then $P_{AB} = p_A p_B \pm D$, where *D* is the linkage disequilibrium between the two loci. Linkage may occur and be maintained because the different loci are physically linked, because selfing rate or clonal rate is significant, because the population has not reached equilibrium since the last disturbance (e.g. bottleneck) or because the sample is composed of individuals belonging to different units. Note that a statistical association between loci is always expected in populations of finite size (e.g. Hedrick, 1987). Linkage disequilibrium may be measured between pairs of loci. If *L* loci are analysed, this lead to $L(L-1)/2$ possible measures. Linkage disequilibrium can also be measured overall loci (multilocus linkage disequilibrium). R_{GGD} (Garnier-Géré and Dillmann, 1992) is a correlation coefficient of allele occurrence between a pair of loci. $\bar{r}_{\mathrm{D}}$ by Agapow and Burt (2001) is the standardised correlation coefficient of allele occurrence at a multilocus scale. $\bar{r}_{\mathrm{D}}$ is unbiased in panmictic populations but strongly variable in clonal populations. R_{GGD} behaves to some extent (but unfortunately not always) better than $\bar{r}_{\mathrm{D}}$ in purely clonal populations but is biased in panmictic populations (de Meeûs and Balloux, 2004).

Wahlund effect: Whenever a sample consists of individuals that were sampled from genetically differentiated sub-populations, a loss of heterozygosity is expected. We can illustrate this with an extreme example. Let us assume two populations that are completely isolated and display private alleles at one locus, say allele *A* for population 1 and *a* for Population 2. Population 1 is thus only composed of *AA* and population 2 of *aa.* Sampling individuals from these two populations into one sample will lead to a sample composed of *AA* and *aa* individuals only (complete lack of heterozygotes, $F_{\mathrm{is}} = 1$).

Bonferroni correction for multiple tests: In the case of multiple testing the chance of finding a significant *P*-value (say $P < 0.05$) is increased. The rationale behind this correction is that if 100 tests were handled in a population that verifies the null hypothesis (e.g. a panmictic population) then five tests are expected to be significant at the 5% level (by definition). The Bonferroni correction is a conservative but efficient way to avoid this caveat. It simply consists in multiplying the observed *P*-values by the total number of tests, or dividing the level of significance (e.g. 0.05) by the number of tests (see Holm, 1979 or Rice, 1989 for more details).

maintained in terms of alleles per locus and the genetic diversity maintained in terms of multilocus genotypes. Recent results on the neutral polymorphism in clonal diploids with a simple life cycle reveal that polymorphism is considerably enhanced at individual loci, but that at the same time multilocus genotypic diversity is reduced (Balloux et al., 2003, de Meeûs and Balloux, 2004). We will come back to the underlying causes of this phenomenon in Sections 2.2 and 2.3. Note however that more complicated life cycles, in particular where the variance in reproductive success is variable, may lead to different observations (e.g. Yonezawa et al., 2004).

2.2. Heterozygosity of clones and the Meselson effect

In the absence of segregation redistributing alleles, diploid clones accumulate heterozygosity over time through mutation events at each locus (e.g. Lokki, 1976; Pamilo, 1987; Tibayrenc and Ayala, 2002) thus leading to extremely negative values of F_{is} (see Box 1; Balloux et al., 2003). This interesting property leads over long evolutionary times to what has been termed the Meselson effect (Judson and Normark, 1996), where, in ancient strictly clonal lineages, the divergence between the two alleles at a same locus within an individual is expected to exceed divergence between different clonal lineages. The absence of segregation in strictly clonal lineages leads to the result that effective size tends towards infinity because coalescence is not possible between the two homologous genes of an individual. In other words, heterozygosity is maintained because it becomes fixed within individuals and a single allele cannot reach fixation; as long as the population does not go extinct there will be at least two alleles present. It is however important to mention that both the effective size and the parameter F_{is} converge towards their value expected under random mating whenever there is a small amount of sexual reproduction (Balloux et al., 2003). In Fig. 1, the expected values of F_{is} are presented as a function of both the size of demes (*N*) and the number of demes (*n*) for purely clonal populations. It can be

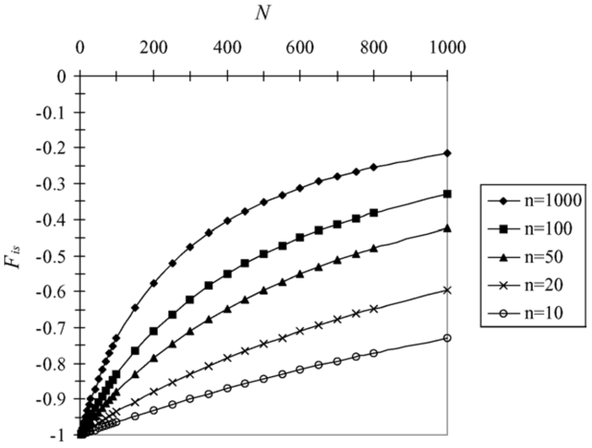

Fig. 1. Expected F_{is} values at equilibrium for an organism reproducing strictly clonally and subdivided in an island model. Migration rate was set to $m = 10^{-3}$ and mutation rate to $u = 10^{-5}$. $K = 99$ possible allelic states are assumed. Number of sub-populations (*n*) and individuals per sub-populations (*N*) were varied. These are analytical results obtained from Eq. (10) in Balloux et al. (2003).

seen that, even if the observed heterozygosity is maximal for all situations, the minimum value for F_{is} is bounded by the number of alleles that are maintained, and thus by both global and sub-population sizes.

2.3. Linkage disequilibrium in diploid clones

The absence of recombination in clonal organisms generates a statistical association of allele between different loci, as clonality is equivalent to absolute physical linkage over the entire genome. As a consequence, clonality will reduce multilocus genetic diversity (number of different multilocus genotypes), despite increasing the number of alleles found at each individual locus. Linkage disequilibrium progressively decreases with increasing rates of sexual reproduction in the population. Thus, even for reasonable amounts of sexual reproduction (e.g. 50%) the signature of clonality should be detectable. Indeed, contrarily to Hardy-Weinberg proportions that are reached in one round of panmixia, linkage equilibrium requires many generations to build up and will thus never be reached if for instance 50% of clonal individuals are recurrently produced at each generation. However, the increased heterozygosity characterising diploid clones seems to interfere considerably with linkage disequilibrium measures such as I_A (multilocus association index; Brown et al., 1980; Maynard-Smith et al., 1993) or some of Ohta's (1982) components of linkage disequilibrium in subdivided population (de Meeûs and Balloux, 2004), and the biases and/or variances associated to linkage disequilibrium estimators tend in general to lead to inaccurate estimation of clonal rates on empirical data (de Meeûs and Balloux, 2004; see also Box 1).

2.4. Population differentiation in diploid clones

For diploid clones with simple life cycle, the expected level of differentiation is lower than for sexual organisms, everything else being equal (Balloux et al., 2003). In an infinite island model (Box 1), with mutation rate u and migration rate m being small enough, the equilibrium value for F_{st} (see Box 1 for the definitions of Wright's F-statistics or fixation indices) read:

$$F_{st} \cong \frac{1}{4N(m+u)+2}$$

for purely clonal populations instead of $F_{st} \cong (1/(4N(m+u)+1))$ expected under panmixia.

We can note that when $m \rightarrow 0$, the clonal $F_{st} \rightarrow 0.5$, instead of 1 in the panmictic case because exactly two alleles are maintained in each individual (Balloux et al., 2003).

3. Little clonality in a mainly sexually reproducing population

While strongly negative F_{is} estimates are a clear signature for clonal reproduction, this quantity is not informative in organisms where sexual reproduction is frequent. F_{is} is a highly non-linear function of the sexuality rate. F_{is} estimates are indistinguishable from panmixia except when the rate of clonal reproduction becomes dominant (Balloux et al., 2003; de Meeûs and Balloux, 2004). Further note that slightly negative F_{is} are also expected in small dioecious or self incompatible monoecious populations that reproduce sexually (Balloux, 2004). A more promising measure might thus be the standardised multilocus linkage disequilibrium $\bar{r}_D$ of Agapow and Burt (2001) since this quantity is strictly unbiased in panmictic populations, for which it is centred on 0, and progressively increases as clonality rate increases (de Meeûs and Balloux, 2004). However, it becomes very unstable in purely clonal populations, where negative values (only expected under panmixia) can be observed.

4. Little sexuality in a mainly clonal population

Rare sexual reproduction in a mainly clonal population is easier to deal with, as in that case, strong variance of F_{is} across loci (Balloux et al., 2003) and strong linkage disequilibria are expected (de Meeûs and Balloux, 2004). Linkage disequilibrium can be estimated by the correlation between pairs of loci R_{GGD} of Garnier-Géré and Dillmann (1992). Nonetheless one should interpret those results while keeping in mind that such a linkage disequilibrium measure is strongly biased in panmictic populations (significantly above 0; de Meeûs and Balloux, 2004). It is also necessary to be able to distinguish biological causes from technical artefacts. Null alleles or short allele dominance are known to strongly bias the detection of heterozygotes and thus F_{is} estimates (e.g. Brookfield, 1996; Wattier et al., 1998; de Meeûs et al., 2004). In purely clonal populations where all loci are expected to display strongly negative F_{is} (low variance), the presence of such artefacts may inflate the variance of F_{is} across loci and may in turn lead to incorrect conclusions being drawn. Note however that under such a scenario, only a subset of loci will be responsible for the increased variance. The same subset of loci will systematically generate a similar bias in the different sub-samples. Such artefacts can thus be detected when working with a sufficient number of loci and sub-samples. For small data sets there is a genuine risk that problematic loci cannot be recognised as such and that purely clonal populations will be indistinguishable from mainly clonal ones (i.e. with a low rate of sexual reproduction).

5. Population differentiation in mainly or purely clonal populations

The effect of clonal reproduction on population differentiation has been hardly addressed in the literature. In order to avoid pseudo-replication, which is believed to inflate population differentiation, most authors only consider one isolate per unit of sampling (typically the patient for medical surveys) or delete repeated genotypes from the data set (e.g. Shaw et al., 1994; Boerlin et al., 1996; Arnavielhe et al., 2000; Fundyga et al., 2002; Delmotte et al., 2002). We will see that this sampling strategy may not be ideal in most cases. In Fig. 2, we illustrate how biased the results may be when one only considers the first isolate of each deme (patient) or when one deletes repeated

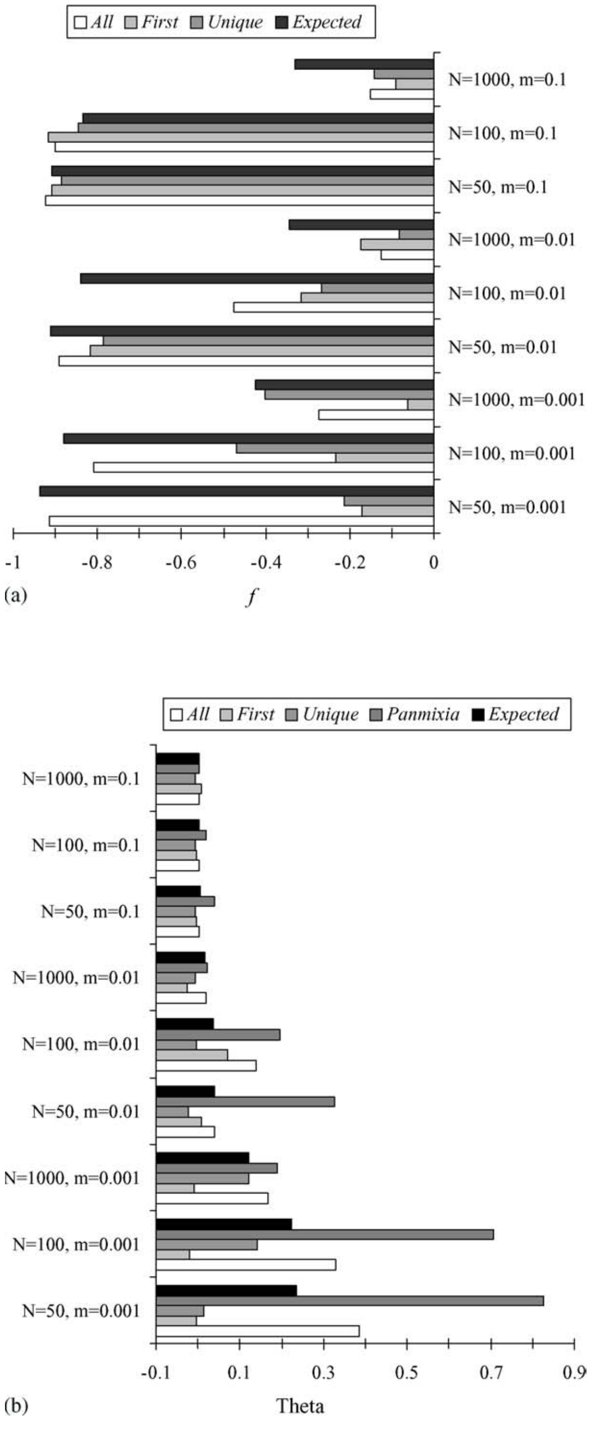

Fig. 2. Results obtained for *f*(*a*) and *θ*(*b*), F_{is} and F_{st} estimators, respectively (Weir and Cockerham, 1984), for different simulations and with different sampling strategies. For all simulations, the number of demes (Island model, see Box 1) was set to 50, the mutation rate per locus to 10^{-5} (20 loci), the number of possible allelic states was limited to 99. All simulations assume a clonal rate of 100% except for (b) where panmictic simulations were also analysed (with the same parameter set otherwise) for comparison with 100% clonal simulations. For each simulation, 20 demes, and 50 individuals per deme were sampled. All simulations were run with Easypop (version 1.8) (Balloux, 2001). The data without repeated genotypes were obtained by Clonality (version 1) (Prugnolle et al., 2004) and all data sets were analysed with Fstat version 2.9.3.2 (Goudet, 1995). *N* stands for the number of individuals per deme, and *m* for migration rate, All indicates the statistics have been computed on the entire data set and first that we only considered the first individual of each deme (20 individuals), grouped in two artificial sub-samples (of 10 individuals each), *Unique* indicates computed after removal of genotypes repeated in each demes (repetition across demes allowed), *Panmixia* stands for a panmictic population with identical parameters, *Expected* indicates the expected value at equilibrium obtained from Eq. (10) (for F_{is}) and 14 (for F_{st}) in Balloux et al. (2003).

genotypes from the data set. This is so because for strongly structured populations ($N \leq 100$ and $m \leq 0.01$), F_{is} is strongly overestimated and F_{st} is underestimated with ‘‘first isolate’’ or ‘‘unique genotypes’’ sampling strategies (Fig. 2). Alternatively, the ‘‘All individuals’’ sampling strategy provides less biased F_{is} estimates but tends to slightly overestimate F_{st}. When deme size is 50 and migration rate 0.001 for instance, a situation where strongly negative values of F_{is} should testify of the strict clonality of these populations, the overestimation of F_{is} is huge for ‘‘first isolate’’ or ‘‘unique genotypes’’ sampling strategies. Such a sampling strategy will also produce very low F_{st} that are not significantly different from 0 ($P \approx 0.5$, 15000 permutations), despite the fact that the population is strongly structured (expected $F_{\text{st}} = 0.24$). This strong population subdivision is more accurately reflected by the ‘‘all individuals’’ sampling strategy ($F_{\text{st}} = 0.39$, $P \leq 0.00007$, 15000 permutations). In such populations, the existence and subsequent reproduction of repeated genotypes are part and consequence of the population process as a whole. Thus, repeated genotypes are important to consider. It can also be noticed from Fig. 2 that the observed values are often in discrepancy with expected ones, meaning that equilibrium is often not reached (see for instance the simulation with $N = 100$ and $m = 0.01$ in Fig. 2). In such cases, the ‘‘unique genotypes’’ strategy can sometimes provide more accurate results. To conclude, results obtained from the analysis including repeated genotypes are important to consider and the interpretation of data should never rely on single genotypes only.

It is noteworthy that this recommendation does not hold in the case of organisms with complex life cycles alternating between sexual and clonal reproduction, such as trematodes (Prugnolle et al., 2005), *Echinococcus* cestodes, Apicomplexa protozoans (e.g. *Plasmodium*) or most aphids. Under alternation of sexual and clonal phases, repeated genotypes are only a transient consequence of the last asexual cycle, which will be broken up again in the subsequent sexual phase (rather than a long term consequence inherent of the mating system of clonal organisms). The inclusion of those transient repeated genotypes in the analysis of organisms with complex life-cycles will obscure the patterns of the underlying population structure. For such organisms, the analysis of the data set without repeated genotypes is indispensable (e.g. see Prugnolle et al., 2005).

6. Final recipe

6.1. Sampling

Sampling should ideally be performed at the lowest possible scale. For pathogenic protozoa or fungi, the individual patient or even the different organs (when relevant) could usefully be taken as the subpopulation reference unit from which multiple isolates should be sampled. If this scale were smaller than the real reproductive unit, it is always possible later to pool isolates belonging to different sub-samples. This is not a trivial point because sampling at a scale wider than the relevant one will inflate the F_{is} estimate through Wahlund effect (Box 1) thus masking the signature of clonality, and of course precluding any accurate inference on the population structure. All loci must have been checked for null alleles. As F_{is} will be a critical criterion in diploid clones, loci with null alleles should be avoided. Null alleles are expected to introduce strong and systematic (in all samples) heterozygote deficits at the loci concerned. If some sexual reproduction exists, null alleles will produce null (homozygous) individuals.

6.2. Translating obscure quantities into biological parameters

The most useful quantities to estimate are the mean and variance across loci of F_{is}, the correlation coefficients R_{GGD} and $\bar{r}_{\mathrm{D}}$ and the F_{st}. Another useful parameter is the proportion of multilocus repeated genotypes. Taken together those statistics should constitute a minimal toolkit for drawing biological inferences from genotypic and allele frequency data in clonal or partially clonal diploids. At this point, it is useful to discuss the neutrality assumption together with hitchhiking problems. In clonal organisms, because all loci are linked together, each selective event concerns the whole genome. This however should not affect strongly the following discussion, provided a sufficient number of samples and loci are analysed, and that a global selective event did not recently affect the population sampled (see Barraclough et al., 2003, for a more extensive discussion on this topic). Let us now consider in turn a few scenarios and how to interpret those statistics.

(i) All samples and all loci display strongly negative F_{is} values with small associated variance, strongly significant linkage disequilibria at many loci (R_{GGD} and $\bar{r}_{\mathrm{D}}$), but not necessarily between the same pairs of loci in different sub-samples (as would be the case for physically linked loci) and F_{st} close to 0.5: the population under investigation is purely clonal and strongly structured (the product between subpopulation size N and migration rate m: N_{m} small as for a highly endemic or a barely contagious disease).

(ii) The same as (i) but with moderately negative F_{is} and $F_{\mathrm{st}} \ll 0.5$: the population is purely clonal but moderately structured (Nm is large, most likely to be an epidemic and/or contagious disease).

(iii) F_{is} is strongly negative but highly variable from one locus to the other, varying from nearly -1 for most loci to nearly +1 for some loci, linkage disequilibria are strong and do not involve specific loci pairs across populations and $F_{\mathrm{st}} > 0.5$: the population is strongly clonal but with a very small amount of sex at each generation (e.g. clonal rate c in [0.9999–0.99]) and strongly structured (small Nm). Note that the inter-locus variance should not be systematically due to the same loci from one sample to the other, in which case the null allele hypothesis should also be considered.

(iv) Same as for (iii) but with moderately negative F_{is} varying from significantly negative $F_{\mathrm{is}} \gg -1$ for most loci to significantly positive $F_{\mathrm{is}} \ll +1$ for some loci and with F_{st} not strongly above 0.5: same population as in (iii) but moderately structured (large Nm).

(v) F_{is} is not significantly different from 0 with moderate associated variance across loci, linkage disequilibria are significant between some to many pairs of loci, $\bar{r}_{\mathrm{D}}$ is significantly >0: the population is moderately clonal, but the signal will strongly depend on the number of sub-samples, the number of individuals (isolates) per sub-sample and the level of polymorphism that the different loci studied display. This means that in that case the rate of clonal reproduction c may be in a range between 0.5 and 0.9. For example, in simulations with the following parameter values $n = 50$, $N = 50$, $m = 0.001$, $c = 0.8$ and sampling 20 demes and 50 individuals per demes assayed for 20 loci gives strongly significant but moderately negative $F_{\mathrm{is}} = -0.076$ ($P = 0.0001$) and 160 significant linkage disequilibrium tests after Bonferroni correction (Box 1) out of 190 between pairs of loci ($R_{\mathrm{GGD}} = 0.39$). Interestingly $\bar{r}_{\mathrm{D}} = -0.009$ in this case, thus illustrating the problems that this statistic can meet in mainly clonal populations. When considering the same data set but reducing the sample to 5 demes with 20 individuals in each, and assayed for 7 loci, F_{is} is not significantly different from 0 anymore and only one pair of loci (out of 18) is in significant linkage disequilibrium after Bonferroni corrections ($R_{\mathrm{GGD}} = 0.26$). Here $\bar{r}_{\mathrm{D}} = 0.48$ but is not significantly different from 0 ($P = 0.06$). Sampling effort can thus have considerable consequences for these kinds of intermediate scenarios.

(vi) F_{is} is close to zero for all loci, $\bar{r}_{\mathrm{D}} = 0$ and no locus pair is in significant linkage disequilibrium at the Bonferroni level, then the population behaves as a non-clonal one (within deme panmixia).

(vii) F_{is} is close to zero for all loci but there are strong and significant linkage disequilibria between many pairs of loci and the organism studied cannot (of course) be haploid. This may be the signature for an essentially clonal organism where sampling has been performed at the wrong scale, pooling individuals (isolates) that do not belong to the same reproductive unit within erroneously a priori defined sub-populations. F_{is} is overestimated to nearly 0 as a result of a Wahlund effect (Box 1). This conclusion must however rely on a sufficient number of loci and populations to exclude confounding causes such as null alleles and tight physical linkage between the loci used.

6.3. Detecting the different levels of population structure

Accurate partitioning of genetic variance at different levels requires a sampling strategy at the lowest possible scale. The detection of the different levels of population structure and their significance can then be assessed by a method based on F_{is} estimates, allowing to test where the actual levels of population

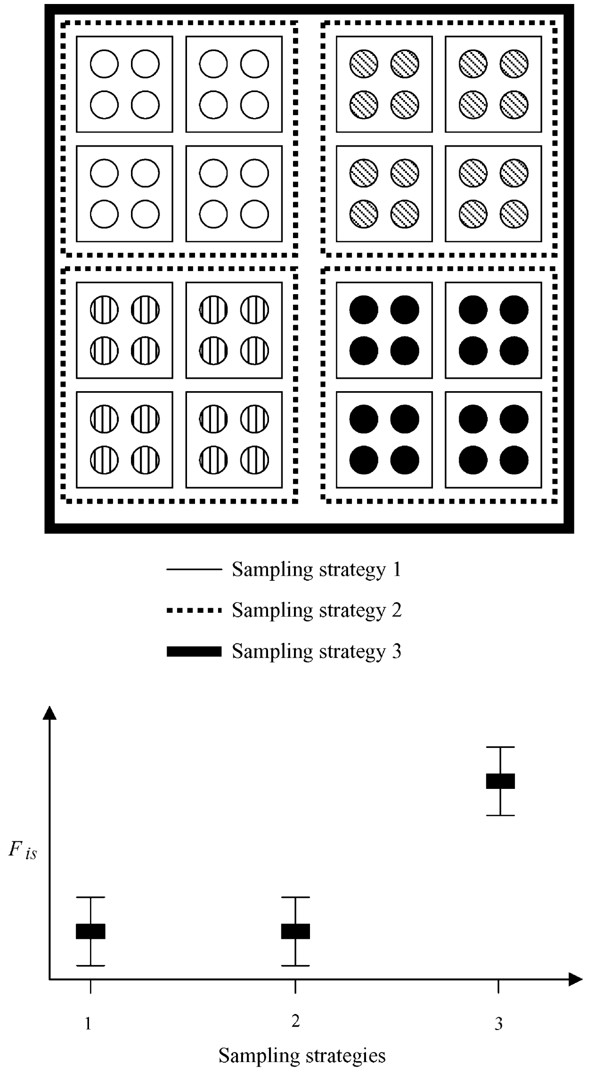

Fig. 3. Illustration of the Goudet et al. (1994) method. In this imaginary sample, 64 (circles) individuals were genotyped. There are four demes. Individuals belonging to the same deme share the same filling texture (empty, oblique stripes, vertical stripes and full black). The sub-samples where F_{is} can be computed are defined following three sampling strategies. With strategy 1 (thin squares), 16 sub-samples are defined with four individuals each. With strategy 2 (dashed squares), four sub-samples are defined with 16 individuals each. Strategy 3 pools all 64 individuals within a single sample. F_{is} computed under strategy 3 is significantly higher than under the other strategies, suggesting that the correct population structure is better defined by the sampling strategy at lower scales (i.e. Strategy 2).

structuring occur. This method was described in Goudet et al. (1994) and a schematic illustration is provided in Fig. 3. F_{is} is first estimated at the level of the smallest sub-samples (e.g. isolates from the same organ of the same patient). These sub-samples are then pooled at the next hierarchical scale (e.g. isolates from the same patient) and F_{is} is computed again. Those sub-samples are then pooled at the next level and so on (e.g. patients of the same town, towns of the same region etc.). As long as the individuals (isolates) that are pooled belong to the same reproductive unit, no change in F_{is} estimates is expected. Each time the pooling meets a significant level of population structure, F_{is} will experience an increase comparable to a Wahlund effect (Box 1). The significance of any increase can be tested by re-sampling methods (e.g. Bootstrap over loci; see Goudet, 1995).

Another very useful method is provided by a factorial correspondence analysis (FCA) approach as implemented in Genetix (freely downloadable at http://www.univ-montp2.fr/~genetix/genetix/genetix.htm). This method is particularly useful for determining cryptic groups when no particular clue is available for sub-population delineation. See for example Solano et al. (2000) and Ajzenberg et al. (2002).

7. Conclusion and future needs

A sound sampling strategy is paramount for drawing subsequent accurate inferences. The advantage of working on clonal diploids is the possibility to infer within population structure (F_{is}), a very useful parameter for biological inference in this context. The strong limitation of the parameter F_{is} for its wide application in various organisms is its extremely non-linear relation with the rate of clonal reproduction. Whenever there are rare events of sexual recombination, F_{is} tends towards its expectation under panmixia. This parameter is of course useless in haploids. What would thus be really needed are estimators of linkage disequilibrium displaying limited dependency of underlying population structure and with bias and variance low enough to be translatable into rates of clonal reproduction. This is even more desirable for haploids for which Linkage disequilibrium based methods represent the only means to assess clonal reproduction (see Halkett et al., 2005, for a review). Another field that has been hardly touched upon to date and should be tackled in the future is the population genetics of organisms with complex life cycles alternating asexual and sexual reproduction such as trematodes, *Echinococcus* cestodes, sporozoa (e.g. *Plasmodium*), cladocerans and aphids.

Acknowledgements

We would like to thank Franck Prugnolle, Christine Chevillon and Michel Tibayrenc for very useful discussions. Jean-Christophe Simon and one anonymous referee made extremely useful comments that considerably helped improving the manuscript. T. de Meeûs is supported by the CNRS.

References

Ajzenberg, D., Bañuls, A.L., Tibayrenc, M., Dardé, M.L., 2002. Microsatellite analysis of Toxoplasma gondii shows considerable polymorphism structured into two main clonal groups. Int. J. Parasitol. 32, 27–38.

Agapow, P.M., Burt, A., 2001. Indices of multilocus linkage disequilibrium. Mol. Ecol. Notes 1, 101–102.

Anderson, J.B., Kohn, L.M., 1998. Genotyping, gene genealogies and genomics bring fungal population genetics above ground. Trends Ecol. E 13, 444–449.

Arnavielhe, S., de Meeûs, T., Blancart, A., Mallié, M., Renaud, F., Bastide, J.M., 2000. Multicentric study of *Candida albicans* isolates from non-neutropenic patients: population structure and mode of reproduction. Mycoses 43, 109–117.

Balloux, F., 2001. Easypop (Version 1.7): a computer program for population genetics simulations. J. Hered. 92, 301–302.

Balloux, F., 2004. Heterozygote excess in small populations and the heterozygote-excess effective population size. Evolution 58, 1891–1900.

Balloux, F., Lehmann, L., de Meeûs, T., 2003. The population genetics of clonal or partially clonal diploids. Genetics 164, 1635–1644.

Barraclough, T.G., Birky, C.W., Burt, A., 2003. Diversification in sexual and asexual organisms. Evolution 57, 2166–2172.

Boerlin, P., Boerlin-Petzold, F., Goudet, J., Durussel, C., Pagani, J.L., Chave, J.P., Bille, J., 1996. Typing *Candida albicans* oral isolates from human immunodeficiency virus-infected patients by multilocus enzyme electrophoresis and DNA fingerprinting. J. Clin. Microbiol. 34, 1235–1248.

Brookfield, J.F.Y., 1996. A simple method for estimating null allele frequency from heterozygote deficiency. Mol. Ecol. 5, 453–455.

Brown, A.H.D., Feldman, M.W., Nevo, E., 1980. Multilocus structure of natural populations of *Hordeum spontaneum.* Genetics 96, 523–536.

Carvalho, G.C., 1994. Genetics of aquatic clonal organisms. In: Beaumont, A. (Ed.), Genetics and evolution of aquatic organisms. Chapman & Hall, London, pp. 291–319.

Cockerham, C.C., 1969. Variance of gene frequencies. Evolution 23, 72–84.

Cockerham, C.C., 1973. Analysis of gene frequencies. Genetics 74, 679–700.

Delmotte, F., Leterme, N., Gauthier, J.P., Rispe, C., Simon, J.C., 2002. Genetic architecture of sexual and asexual populations of the aphid *Rhopalosiphum padi* based on allozyme and microsatellite markers. Mol. Ecol. 11, 711–723.

de Meeûs, T., Balloux, F., 2004. Clonal reproduction and linkage disequilibrium in diploids: a simulation study. Inf. Genet. E 4, 345–351.

de Meeûs, T., Humair, P.F., Delaye, C., Grunau, C., Renaud, F., 2004. Non-Mendelian transmission of alleles at microsatellite loci: an example in Ixodes ricinus, the vector of Lyme disease. Int. J. Parasitol. 34, 943–950.

Fundyga, R.E., Lott, T.J., Arnold, J., 2002. Population structure of *Candida albicans*, a member of the human flora, as determined by microsatellite loci. Infect. Genet. E 2, 57–68.

Garnier-Géré, P., Dillmann, C., 1992. A computer program for testing pairwise linkage disequilibria in subdivided populations. J. Hered. 83, 239.

Goudet, J., 1995. FSTAT (version 1.2): a computer program to calculate *F*-statistics. J. Hered. 86, 485–486.

Goudet, J., de Meeûs, T., Day, A.J., Gliddon, C.J., 1994. The different levels of population structuring of the dogwhelk, *Nucella lapillus,* along the south Devon coast. In: Beaumont, A.R. (Ed.), Genetics and Evolution of aquatic organisms. Chapman & Hall, London, pp. 81–95.

Génermont, J., 1980. Les animaux à reproduction uniparentale. In: Bocquet, C., Génermont, J., Lamotte, M. (Eds.), Les problèmes de l'Espèce dans le Règne Animal Tome III, Mem. Soc. Zool. Fr. 40, pp. 287–320.

Halkett, F., Simon, J.F., Balloux, F., 2005. Tackling the population genetics of clonal and partially clonal organisms. Trends Ecol. Evol. 20, 194–201.

Hartl, D.L., Clark, A.G., 1989. Principles of Population Genetics, Second ed. Sinauer associates, Sunderland, Massachussets.

Hedrick, P.W., 1987. Gametic disequilibrium measures: proceed with caution. Genetics 117, 331–341.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Judson, O.P., Normark, B.B., 1996. Ancient asexual scandals. Trends Ecol. E 11, 41–46.

Lokki, J., 1976. Genetic polymorphism and evolution in parthenogenetic animals. Hereditas 83, 57–64.

Marshall, D.R., Weir, B.S., 1979. Maintenance of genetic variation in apomictic plant populations. Heredity 42, 159–172.

Maynard-Smith, J., Smith, N.H., O'Rourke, M., Spratt, B.G., 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. USA 90, 4384–4388.

Maynard-Smith, J., Feil, E.J., Smith, N.H., 2000. Population structure and evolutionary dynamics of pathogenic bacteria. BioEssays 22, 1115–1122.

Milgroom, M.G., 1996. Recombination and the multilocus structure of fungal populations. Annu. Rev. Phytopathol. 34, 457–477.

Milgroom, M.G., 1997. Genetic variation and the application of genetic markers for studying plant pathogen populations. J. Plant. Pathol. 78, 1–13.

Milgroom, M.G., Fry, W.E., 1997. Contributions of population genetics to plant epidemiology and management. Adv. Bot. Res. 24, 1–30.

Nei, M., 1977. *F*-statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. Lond. 41, 225–233.

Ohta, T., 1982. Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc. Natl. Acad. Sci. USA 79, 1940–1944.

Orive, M.E., 1993. Effective population size in organisms with complex life-histories. Theor. Pop. Biol. 44, 316–340.

Pamilo, P., 1987. Heterozygosity in apomictic organisms. Hereditas 107, 95–101.

Prugnolle, F., Choisy, M., Théron, A., Durand, P., de Meeûs, T., 2004. Sex-specific correlation between heterozygosity and clone size in the trematode *Schistosoma mansoni*. Mol. Ecol. 13, 2859–2864.

Prugnolle, F., Liu, H., de Meeûs, T., Balloux, F., 2005. Population genetics of complex life cycle parasites: the case of monoecious trematodes. Int. J. Parasitol. 35, 255–263.

Rice, W.R., 1989. Analyzing tables of statistical Tests. Evolution 43, 223–225.

Rousset, F., 2004. Genetic Structure and Selection in Subdivided Populations. Princeton University Press, Princeton.

Shaw, P., Ryland, J.S., Beardmore, J.A., 1994. Population genetic parameters within a sea anemone family (Sagartiidae) encompassing clonal, semiclonal and aclonal modes of reproduction. In: Beaumont, A. (Ed.), Genetics and evolution of aquatic organisms. Chapman & Hall, London, pp. 351–358.

Silander Jr., J.A., 1985. Microevolution in clonal plants. In: Jackson, J.B.C., Buss, L.W., Cook, R.E. (Eds.), Population Biology of Clonal Organisms. Yale University Press, Yale, pp. 107–152.

Suomolainen, E., Saura, A., Lokki, J., 1976. Evolution of parthenogenetic insects. Evol. Biol. 9, 209–257.

Solano, P., de La Rocque, S., de Meeûs, T., Cuny, G., Duvallet, G., Cuisance, D., 2000. Microsatellite DNA markers reveal genetic differentiation among populations of *Glossina palpalis gambiensis* collected in the agropastoral zone of Sideradougou, Burkina Faso. Insect Mol. Biol. 9, 433–439.

Taylor, J.W., Geiser, D.M., Burt, A., Koupopanou, V., 1999. The evolutionary biology and population genetics underlying fungal strain typing. Clin. Microbiol. Rev. 12, 126–146.

Tibayrenc, M., 1995. Population genetics of parasitic protozoa and other microorganisms. Adv. Parasitol. 36, 47–115.

Tibayrenc, M., 1998. Genetic epidemiology of parasitic protozoa and other infectious agents: the need for an integrated approach. Int. J. Parasitol. 28, 85–104.

Tibayrenc, M., 1999. Toward an Integrated genetic epidemiology of parasitic protozoa and other pathogens. Ann. Rev. Genet. 33, 449–477.

Tibayrenc, M., Ayala, F.J., 2002. The clonal theory of parasitic protozoa: 12 years on. Trends Parasitol. 18, 405–410.

Tibayrenc, M., Kjellberg, F., Arnaud, J., Oury, B., Brénière, F., Dardé, M.L., Ayala, F.J., 1991. Are eukaryotic microorganisms clonal or sexuals? A population genetics vantage. Proc Natl. Acad. Sci. USA 88, 5129–5133.

Wattier, R., Engel, C.R., Saumitou-Laprade, P., Valero, M., 1998. Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in *Gracilaria gracilis* (Rhodophyta). Mol. Ecol. 7, 1569–1573.

Weir, B.S., Cockerham, C.C., 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38, 1358–1370.

Wright, S., 1951. The genetical structure of populations. Ann. Eugenics 15, 323–354.

Wright, S., 1965. The interpretation of population structure by *F*-statistics with special regard to system of mating. Evolution 19, 395–420.

Yonezawa, K., Ishii, T., Nagamine, T., 2004. The effective size of mixed sexually and asexually reproducing populations. Genetics 166, 1529–1539.