Available online at www.sciencedirect.com
ScienceDirect

Infection, Genetics and Evolution 7 (2007) 308–332

Infection, Genetics and Evolution

www.elsevier.com/locate/meegid

Discussion

Population genetics and molecular epidemiology or how to "débusquer la bête"

Thierry de Meeûs *, Karen D. McCoy, Franck Prugnolle, Christine Chevillon, Patrick Durand, Sylvie Hurtrez-Boussès, François Renaud

Génétique et Evolution des Maladies Infectieuses, Equipe Evolution des Systèmes Symbiotiques, UMR IRD/CNRS 2724, BP 64501, 911 Av. Agropolis, 34394 Montpellier Cedex 5, France

Received 21 March 2006; received in revised form 13 July 2006; accepted 19 July 2006
Available online 1 September 2006

Abstract

Parasites represent a great proportion of the world's living organisms and are of overwhelming significance because of their impact on hosts (evolutionarily, medically, agronomical and economically). The knowledge of the population biology of such organisms is thus of fundamental importance to population biologists. Most parasites cannot be studied by direct methods and their biology has to be assessed via indirect means, most notably using molecular markers. In this review, we present the molecular tools, the null models employed, the statistical tools available and the kinds of inferences one can make when using molecular markers to study the ecology/epidemiology of host–parasite systems (molecular ecology/molecular epidemiology). We conclude with relevant examples, most issued from our laboratory, to illustrate the pros and cons of such methods for the study of parasites, vectors, micropathogens and their hosts and briefly discuss future needs.

Keywords: Population genetics; Molecular epidemiology; Parasites; Vectors; Molecular markers

1. Introduction

Parasites represent a significant part of the described biodiversity (De Meeûs and Renaud, 2002) and despite the recent explosion of molecular studies, those conducted on such organisms are still too few (Criscione et al., 2005). This may be largely explained by the fact that infectious agents and their vectors are generally difficult to study. Because of their small size, location, biology and behaviour, direct observation of their population biology is almost impossible. Thus, the ecology, reproductive modes and/or strategies, dispersal, population sizes of parasites and vectors can mainly be assessed through what Slatkin (1985) called "indirect methods" (Nadler, 1995; De Meeûs et al., 2002a,b). In this case, the indirect methods are those that use polymorphic molecular markers and the variation of such polymorphism within and between pre-defined sub-sets of individuals (most of the time referred to sub-populations or sub-samples). The basic and reasonably realistic assumption being that the distribution of genetic variation should reflect ecologically relevant population parameters, such as those cited above. The knowledge of such parameters is not simply an academic matter or esoteric endeavour (Milgroom, 1996; Tibayrenc, 1998, 1999; Taylor et al., 1999; Criscione et al., 2005). "Population structure and mating system of pathogens are tightly linked biological phenomena with crucial consequences on the epidemiology of transmissible diseases" (Tibayrenc and Ayala, 2002). It "can be crucial for disease management" (Milgroom, 1996) as well as for "research aimed at treatment and prevention" (Taylor et al., 1999) and for the "safe evaluation and prediction of drug and antibiotic resistance" (Tibayrenc, 1999). The investigation of the population genetics of infectious diseases and their vectors has been termed genetic epidemiology or molecular epidemiology (Tibayrenc, 1998). This method has an additional bonus: it uses genetic and thus heritable information that can be informative even if the event responsible for the observed pattern is not currently occurring (i.e., it provides historic information). Intensive field observation is therefore not necessarily required to obtain useful information (e.g., Prugnolle and De Meeûs, 2002). The study of genetic variation

* 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 © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2006.07.003

in natural populations of parasites and vectors can give access to key information on their ecology and evolutionary potential, but this, of course, requires a continuously growing set of statistical and population genetics tools. The aim of this paper is to review these tools, their power and limits along with their concept bases and biological assumptions. For more general and theoretical reviews, readers may refer to the excellent papers from Criscione and Blouin (2005a) and Criscione et al. (2005). The review is subdivided into four parts. First, we briefly describe the different kinds of genetic (molecular) markers that are currently available and discuss their merit. Second, we list the different population genetics concepts and tools required. Third, we address the statistical issues associated with the analysis of these markers. Finally, we discuss different case studies, mainly chosen from work performed in our laboratory, as an illustration of how these tools can be used. A glossary where most technical words are defined will help the reader to follow the presentation.

2. What is a genetic marker?

2.1. Preliminary notions

A genetic marker is a portion of nucleic acid (e.g., DNA) or the product of a portion of nucleic acid (protein) from the organism under scrutiny. To make biological inferences, the variation of the same portion of DNA has to be studied across individuals from different sites. It is important that this portion of DNA (or its product) has the same localisation in the genome of each individual (i.e., found in the same place on the same chromosome), hence the term locus. A locus may correspond to a structural gene (coding sequence), like for isoenzymes, but this is not a requirement, and most of the time one will prefer studying a non-coding locus because it is more likely neutral (i.e., not submitted to selective forces) and thus more likely to reflect purely demographic parameters (population size, dispersal). More than one locus can be analysed, and we will see that the study of several loci of the same nature is preferable (the more the better/or at least seven). To be informative a locus must be variable (i.e., polymorphic). This means that the sequence of DNA must vary from one individual to another. These different sequences of the same locus are called alleles. The merit and differences among available molecular markers have been more thoroughly reviewed elsewhere (e.g., Roderick, 1996; Sunnucks, 2000; Caterino et al., 2000) so we will only briefly overview this topic. We have subdivided genetic markers into three different categories (cytoplasmic markers, dominant nuclear markers and codominant nuclear markers). We thus explicitly refer to diploid eukaryotic organisms.

2.2. Cytoplasmic markers

Cytoplasmic markers correspond to loci present in mitochondrial or chloroplastic genomes. These markers, and in particular mitochondrial DNA, have been extensively used in population studies (e.g., Roderick, 1996). Because it evolves rapidly and lacks recombination, mtDNA has proved very useful in phylogeographic studies (Avise et al., 1987; Avise, 2000). However, for population genetics-based studies, we would not recommend using this kind of marker, for several reasons. First, mtDNA is generally uniparentally inherited, typically maternally, but sometimes paternal lineage occurs (Xu, 2005). It is thus dependent on the population structure of one sex in dioecious parasites or vectors (many nematodes, arthropods, schistosomes), and the effective population size (Box 1) of such markers will always be difficult to grasp because it depends on several factors such as sex-specific dispersal, sex ratio and reproductive strategy (Prugnolle and De Meeûs, 2002; Prugnolle et al., 2003). Second, mtDNA might not be neutral (Gerber et al., 2001) and thus may not only be affected by demographic and geographical processes. We thus chose not to treat this class of markers in the present paper.

Box 1. The effective size of a population, usually designated by N_e, enables one to quantify the rate at which a population looses its 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 into a single 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 of $1/N_c$ 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 an uneven sex ratio. The effective population size of a herd with 1 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. A population will better maintain genetic diversity if it is subdivided. For instance, the extreme case of total subdivision (no gene flow between subpopulations) leads to an infinite effective population size because genetic diversity is frozen at the global population scale, even if it is lost locally. An excellent review on the computation of N_e in parasitic organisms can be found in Criscione and Blouin (2005a).

2.3. Dominant nuclear markers

With dominant nuclear markers, heterozygous individuals (hence in diploids) are seen as homozygous for one of the alleles present. This allele is called dominant while the other (invisible in the heterozygous state) is called recessive. Here, the phenotype (the perception we have of the genotype) does not reflect the genotype. The most well-known dominant markers are the randomly amplified polymorphic DNA (RAPD). Small primer pairs randomly amplify portions of target DNA if a sequence match is found. Thus, for diploid species, homozygous individuals with no-matching sequence are characterised by an absence of the amplified product. The heterozygote and homozygote individuals with a matching sequence display the same phenotype (presence of the amplified product). Only phenotypic frequencies can be estimated with this kind of marker. Allelic frequencies cannot be assessed in diploids and thus the local distribution of genetic information within and between individuals remains hidden. Moreover, as previously mentioned, it is always desirable studying several similar loci. However, as RAPD markers concern a random portion of DNA, there is no way to know if the different loci involved are equivalent in terms of neutrality and mutation rate. For these reasons, dominant markers in general and RAPDs in particular are far from ideal population genetics tools.

2.4. *Codominant nuclear markers*

With codominant nuclear markers, all genotypes (homozygous and heterozygous) are theoretically distinguishable. Many visualisation techniques exist for such markers. Isoenzymes, restriction fragment length polymorphisms (RFLP), amplified fragment length polymorphisms (AFLP), microsatellites, minisatellites, multilocus sequence typing (MLST) and single-stranded conformational polymorphism (SSCP) are among the best known. Here, we chose to focus on two of these: isoenzymes and microsatellites because these are generally the easiest and cheapest to implement. For other techniques, readers may refer to existing reviews (e.g., Taylor et al., 1999; Caterino et al., 2000; Sunnucks et al., 2000; Bougnoux et al., 2004). Single-nucleotide-polymorphism (SNP) markers are very useful in association studies. But as these are mainly bi-allelic loci, with heterogeneous mutation rates (there is a clear bias in favour of transitions over transversions) (Vignal et al., 2002), such markers are not ideal for population genetics studies.

Isoenzymes, also known as allozymes, are metabolic enzymes, like the glucose-phosphate-isomerase of the Kreb's cycle, contained in the cell and coded by specific genes. To use these markers, individuals, or a part of their body, are crushed in a buffer or distilled water and the extract is introduced onto or into a gel (e.g., starch, polyacrylamide, cellulose-acetate gels) that is submitted to an electric current. Proteins are generally negatively charged and will thus migrate to the anode, and more rarely to the cathode when positively charged (hence the term electrophoresis). Depending on the electric charge, different enzymes will migrate at different speed. At the end of migration, enzymes are revealed with a solution containing the substrate (or an analogue) of the enzyme and a mix of molecules that precipitates and stains the product of the enzymatic reaction. If the locus under study displays sequence variation that affects the charge of the enzyme but does not abolish its function, then such variation will be visible on the gel. For the same locus, different bands will be revealed corresponding to different alleles (hence the term allozymes). Homozygous individuals for different alleles will display banding patterns that differ in migration. Depending on the structure of the enzyme, heterozygous individuals will display two, three or five bands for monomeric, dimeric or tetrameric enzymes, respectively. More details on enzyme electrophoresis can be found in Pasteur et al. (1987) and Ben Abderrazak et al. (1993). Enzymatic loci are very easy to study and probably represent the cheapest way to study population genetics. However, as only about one-third of mutations in the DNA sequences of the considered enzyme are visible with electrophoresis (Shaw, 1970), these markers generally show low polymorphism. Moreover, the technique requires working with relatively large amount of live material and is thus not appropriate for small and uncultivable (in the cloning sense) organisms. Finally, because enzymatic function is required for visualisation, the material must be fresh or properly conserved. Samples must be kept frozen which is often difficult in many countries and for many species of medical or agronomic interest. These different reasons explain why isoenzymes are rarely used in molecular epidemiology studies, with a few (important) exceptions such as recent studies on some organisms as cockroaches (Corley et al., 2001), spearwinged flies (Niklasson et al., 2004), pathogenic fungi (Arnaviehle et al., 2000; Badoc et al., 2002; De Meeûs et al., 2002b; Nébavi et al., 2006), and kinetoplastid parasites and their vectors (Barnabé et al., 2000; Borges et al., 2000; Hide et al., 2001; Brenière et al., 2003; Njiokou et al., 2004).

Microsatellites are short tandemly repeated sequences of DNA, generally of two, three or four base pairs (more rarely five) (e.g., AC, CGT, GATA). The polymorphism of these markers consists of variation in the number of repeats of the sequence. To be visualised, a microsatellite locus requires knowledge of the sequences that flank the locus, so that primers can be designed to amplify this portion of DNA. Thus, after DNA extraction and PCR amplification, with the integration of labelling substance (e.g., fluorescein-labelled PCR), the product is revealed by electrophoretic migration in a gel and/or buffer. Variants are then differentiated because longer products (with more repeats) migrate more slowly than shorter ones. Homozygotes have one band or peak (if an automated sequencer is used), while heterozygous individuals display two bands or peaks. Microsatellite loci are generally considered to be highly polymorphic, codominant, abundant throughout the genome, and relatively easy to score (Lehmann et al., 1996). Because these markers are DNA sequences, storing samplings in alcohol prior to use is generally not a problem. In addition, with recent improvements in amplification techniques, it is possible to work out from extremely small amounts of organic material. For instance, Razakandrainibe et al. (2005) genotyped single *Plasmodium* oocysts for seven microsatellite loci. For these reasons, microsatellite loci are very interesting for use in molecular epidemiology.

3. Basic concepts in population genetics

3.1. Estimating allelic frequencies from a sample of genotyped individuals

Let us now assume that we are studying diploid individuals sampled from a natural population and genotyped at several codominant genetic markers. Let us take the example of one sample of size N with one locus and two alleles labelled 1 and 2. Let N_{11}, N_{12} and N_{22} be the number of genotypes 1/1, 1/2 and 2/2, respectively found in the sample. The allelic frequencies p_1 and p_2 of alleles 1 and 2 in the sample can therefore be computed as

$$p_1 = \frac{2N_{11} + N_{12}}{2N} = \frac{N_{11} + (1/2)N_{12}}{N} \qquad (1)$$

and

$$p_2 = \frac{2N_{22} + N_{12}}{2N} = \frac{N_{22} + (1/2)N_{12}}{N} = 1 - p_1 \qquad (2)$$

Because we are using codominant markers, these frequencies are estimates of those of the population from which the individuals were sampled.

3.2. Hardy–Weinberg assumptions and equilibrium

Most discussions on the genetics of populations start with the "Hardy–Weinberg model". Hardy (1908), a British mathematician, and Weinberg (1908), a German physician, independently produced this model the same year (1908), hence the appellation "Hardy–Weinberg". This equilibrium model is the theoretical basis for most (if not all) population genetics analyses.

The assumptions of the model are the following:

- The population studied is of infinite size (infinite number of individuals);
- There is no mutation;
- There is no migration (dispersal) (the population is isolated);
- There is no selection;
- Reproduction is sexual and zygotes are produced by the random association of gametes (panmixia).

For a locus with two alleles (1 and 2) with frequencies p_1 and $p_2 = 1 - p_1$ respectively, in such an ideal population, the expected genotypic frequencies in zygotes are simply the product of the frequencies of the different alleles. Thus, the frequency of genotype 1/1 is p_1^2, the frequency of genotype 1/2 is p_1p_2, the frequency of genotype 2/1 is p_2p_1 and the frequency of genotype 2/2 is p_2^2. This naturally leads to the classical p_1^2, $2p_1p_2$, p_2^2. It is easily seen that the sum of these frequencies equals 1 and that the frequency of each allele remains unchanged in the zygotes (e.g., using Eqs. (1) and (2)). Because the population is infinite, the random sampling of gametes and zygotes that will form the next generation does not alter these frequencies (no random genetic drift). Because there is no selection, migration or mutation, these frequencies remain constant, hence the term equilibrium. It is noteworthy that panmixia will always produce a genotypic distribution of the form p_1^2, $2p_1p_2$, p_2^2 at each generation even if the population does not fulfil all the other assumptions of Hardy–Weinberg.

3.3. Relaxing Hardy–Weinberg assumptions

3.3.1. Finite-sized populations

The best illustrations being often the most caricatured ones, let us assume a population of size $N = 2$ with two heterozygous individuals 1/2. The allelic frequencies are thus $p_1 = 1/2$ and $p_2 = 1/2$. With a random union of gametes $p_1^2 = (1/4)$ 1/1, $2p_1p_2 = (1/2)$ 1/2 and $p_2^2 = (1/4)$ 2/2 are produced in the zygotes (here, the number of zygotes is considered to be very large). If the population is to stay at a constant size, regulation must occur such that two adults are obtained from these zygotes. If the surviving zygotes are chosen at random, we have a 1/8 chance to get one allele fixed (two 1/1 or two 2/2), 1/4 chance to see p increasing to 3/4 (one 1/1 and one 1/2 or one 1/2 and one 1/1), 1/4 chance to see it decrease to 1/4 (one 1/2 and one 2/2 or one 2/2 and one 1/2) and only 3/8 chance that p remains unchanged (two 1/2 or one 1/1 and one 2/2 or one 2/2 and one 1/1). It is easily seen that allelic frequencies will change in most situations (5/8). This phenomenon is called genetic drift. It is also obvious that this process leads to a loss of diversity in the population and, providing that no other force is acting on the population (no mutation, no migration), all finite-sized populations should tend to fix a single allele at each locus. The smaller the population, the faster genetic drift acts.

3.3.2. Mutation

This occurs when a mistake is made during DNA duplication. Different kinds of mutations can affect DNA sequences and different mutation models exist as well. The mutation is called recurrent when it constantly changes one allele in the same way (i.e., leads to the same allele). This is for instance the case for many deleterious mutations like the albinism in human populations, which occurs with a rate of about 2.5×10^{-5} (e.g., Hedrick, 2003). In the K alleles model or KAM, mutation randomly transforms one existing allele into one of K possible alternatives. For instance, if the locus under consideration is one base pair long, then there are theoretically four possible allelic states (A, T, G or C). Because the number of distinct alleles is limited, two individuals can share the same allele even though they do not share a recent common ancestor. This phenomenon is called homoplasy. In the infinite allele model or IAM, each mutation transforms an existing allele into a new one that was not previously present in the population. This special case is very useful in theoretical population genetics as it allows many analytical simplifications without loosing too much realism, particularly compared to the KAM model (when K is big). In the IAM model, homoplasy is not possible and all identical alleles are identical by descent (inherited from a common ancestor). Finally, the stepwise mutation model (SMM) (Kimura and Ohta, 1978) was specially designed for microsatellite-like loci. Here, mutation corresponds to an addition or deletion of a single repeat of the elementary motif. If this occurs then homoplasy is likely common, because we characterise alleles based only on their size. However, the difference in repeat number between two alleles can be used to measure their relatedness. Some other more complex mutation models exist (e.g., combining SMM and KAM). Regardless of the mutation model, mutation will change the allele frequencies. However, most mutation rates, except for some microsatellite loci, are very low and mutation alone is a weak micro-evolutionary force, although it should be noted that when associated to drift and selection it represents the key of evolution.

3.3.3. Migration

Natural populations are not totally isolated from each other and tend to exchange propagules between them. These propagules can be larvae, adults, spores or gametes, seeds or pollen, they can be haploid or diploid. The resulting gene flow tends to homogenise the allelic frequencies at all loci across connected populations. Migration can be strong and is a major evolutionary force. Its interaction with genetic drift and mutation can lead to neutral polymorphic equilibrium, without the need of selection.

3.3.4. Selection

Selection is, of course, a major evolutionary force. It can strongly affect allelic and genotypic frequencies. However, selection is likely to affect only those loci concerned by this force (or those tightly linked to such loci). Directional selection will increase or decrease allelic frequencies in a population and can increase (or decrease) the genetic differentiation between populations at the selected locus. If gametes fuse randomly, then the typical Hardy–Weinberg genotypic distribution (p_1^2, $2p_1p_2$, p_2^2) should be observed at each generation, with varying values of p_1 and p_2. Overdominance, increasing the fitness of heterozygous individuals, will create an excess of heterozygous adults compared to Hardy–Weinberg expectations. This is what is observed in human populations at the drepanocytose locus (a very serious genetic disease for homozygous humans) in highly endemic zones of the malaria agent *Plasmodium falciparum* (e.g., Ridley, 1996), where heterozygous humans are more resistant to malaria than mutation free homozygotes. Underdominance is the reverse phenomenon and is not expected to be frequently met in nature as it is highly unstable (the rarest allele tends to disappear). The rhesus system in humans, where a heterozygous (Rh+/Rh−) foetus (if not the first borne) in a homozygous (Rh−/Rh−) mother is endangered by the maternal immune system (e.g., Hartl and Clark, 1989), can be taken as an approximative example of underdominance, where a heterozygote deficit is expected. Selection can also be frequency-dependent. In this case, the fitness of a genotype depends on its frequency in the population, the rarest has the highest fitness. This selection is typically met in host–pathogen interactions when it acts under a gene for gene (with selective costs) or a matching allele model (e.g., Agrawal and Lively, 2002). This kind of selection will generally tend to homogenise the allelic frequencies across populations, although its interaction with migration can lead to more complex patterns (Gandon et al., 1996; Gandon, 2002; Morgan et al., 2005). Heterosis (or hybrid vigor) is a global phenomenon that should affect the whole genome. It results from a genomically widespread overdominance or because the presence of many deleterious recessive alleles produces a significant inbreeding depression and thus favours the most heterozygous individuals (e.g., Prugnolle et al., 2004).

3.3.5. Non-random union of gametes

Several phenomena, with different consequences, can affect how gametes meet. With selfing, one hermaphroditic individual can fertilise its own ovules with its own spermatozoids. This will affect the genotypic distribution at all loci and will decrease heterozygosity compared to that expected under panmixia. It can be easily demonstrated that, with a selfing rate of s (corresponding to the proportion of zygotes produced by selfing) the expected heterozygosity for two alleles of frequencies p_1 and p_2 becomes, at equilibrium (e.g., Hartl and Clark, 1989):

$$H_e = 2p_1p_2\left(1 - \frac{s}{2-s}\right) \qquad (3)$$

From Eq. (3), it is obvious that when $s = 0$ the expectation returns to the Hardy–Weinberg situation, and when $s = 1$, as in *Taenia solium* (e.g., Kunz, 2002; De Meeûs et al., 2003), no heterozygote is expected to be found in a population at equilibrium. It should be noted that hermaphroditic organisms are not necessarily in deviation from panmixia. For instance, using microsatellite markers, Hurtrez-Boussès et al. (2004) found that the monoecious liver fluke *Fasciola hepatica* was panmictic. Sib mating will have the same consequence as selfing (all loci loose heterozygosity), albeit the decrease in heterozygosity is slower (e.g., Hartl and Clark, 1989). Sib mating is met several times in nature as, for example, in the parasitoid wasp *Nasonia vitripenis* (Shuker et al., 2004) or, in the special case of some royal or imperial human families (e.g., Pharaoes, European Kings). Likewise, homogamic loci, which lead individuals (or gametes) to mix when carrying the same allele, will experience a decrease in heterozygosity. Traits such as size at maturity or pathogen resistance typically have a genetic basis and assortative pairing on such characters are well documented (e.g., Thomas et al., 1995). Symmetrically, heterogamic loci that favour mating between individuals carrying different alleles, like MHC loci (e.g., Roberts et al., 2005), will undergo a dramatic increase in heterozygosity. In this case, it is worth noting that frequency-dependent selection cannot be disentangled from heterogamy because rare individuals are compatible with most other individuals in the population. Homogamy and heterogamy only affect the genotypic composition of the concerned loci along with those closely linked to them. It is worth noting that, in a strictly panmictic population of size N, $1/N$ individuals are expected to be produced by selfing (e.g., Rousset, 1996). This means that dioecious organisms can never be truly panmictic as those genes contained in females can only encounter those contained in males. This will have insignificant consequences in populations of reasonable size, but will produce a significant excess of heterozygotes in small dioecious populations or in self-incompatible hermaphrodites (Balloux, 2004). A heterozygote excess is thus likely to occur in many parasites such as schistosomes or most monogenean flatworms (mostly self-incompatible) and has indeed been reported for *Schistosoma mansoni* (Prugnolle et al., 2002). Finally, asexual reproduction, or clonality, in conjunction with drift and mutation will result in an increase of heterozygosity at all loci (see Balloux et al., 2003; De Meeûs and Balloux, 2005).

3.4. The notion of a heterozygote deficit

As seen above, most deviations from Hardy–Weinberg assumptions will alter the genotypic composition observed in the population, especially when there are deviations from random mating. Wright (1965) designed a standardised index that measures this deviation, and is theoretically comparable across loci and populations. This index is the local fixation index, F_{IS}, which corresponds to the excess homozygosity of individuals in a sub-population (hence the subscripts I and S) resulting from the non-random union of gametes in that sub-population. If gametes do not fuse randomly, then F_{IS} corresponds to the proportion of heterozygotes that are lost or gained and equitably redistributed as homozygotes. If, for a given locus with two alleles 1 and 2 of respective frequencies p_1 and p_2, D_o, H_o and R_o are the genotypic frequencies 1/1, 1/2 and 2/2, respectively observed in the sub-population and D_e, H_e and R_e are their expected frequencies under panmixia, then we can write:

$$\begin{aligned} D_o &= p_1^2 + p_1 p_2 F_{IS} = D_e + \frac{H_e}{2} F_{IS}; \\ H_o &= 2p_1 p_2 - 2p_1 p_2 F_{IS} = 2p_1 p_2 (1 - F_{IS}) = H_e (1 - F_{IS}); \\ R_o &= p_2^2 + p_1 p_2 F_{IS} = R_e + \frac{H_e}{2} F_{IS} \end{aligned} \qquad (4)$$

which leads to:

$$F_{IS} = 1 - \frac{H_o}{H_e} = \frac{H_e - H_o}{H_e} \qquad (5)$$

When gametes fuse randomly (panmixia), $F_{IS} = 0$ ($H_e = H_o$). Negative values correspond to an excess of heterozygotes and positive values to a deficit. Note that $F_{IS} = -1$ is only possible with two alleles of equal frequencies ($p_1 = p_2$) and if the population is only composed of one kind of heterozygote (e.g., fixed for 1/2), whereas $F_{IS} = 1$ means there is only homozygous individuals in the sub-population, whatever the number of alleles and their frequency.

Using Eqs. (3) and (4), we can see that:

$$F_{IS} = \frac{s}{2 - s}$$

and thus that the selfing rate can be inferred from the estimation of F_{IS} as

$$s = \frac{2F_{IS}}{1 + F_{IS}} \qquad (6)$$

if one assume that the population has reached equilibrium and that F_{IS} only comes from self-reproduction. For instance, this was successfully used to estimate selfing rate (between 0.8 and 1) in the fresh water snail *Galba truncatula*, the intermediate host of the liver fluke *F. hepatica* (Meunier et al., 2004a).

For more than two alleles, more than one locus and more than one sub-population there are as many measures of F_{IS}. One may be more interested in mean values of F_{IS} if the purpose is to infer the reproductive strategy of the studied species. Mean "heterozygosities" should then be used to compute F_{IS}:

$$F_{IS} = \frac{H_s - \overline{H_o}}{H_s}$$

where H_s is the mean expected heterozygosity overall alleles, loci and sub-populations, or more appropriately speaking the gene diversity observed in the different sub-populations, and $\overline{H_o}$ is the mean heterozygosity observed in these sub-populations. For the sake of generality and conformity with modern notation, we will now express this fixation index (and all that follow) in terms of probabilities of identity. Let us define Q_I as the probability that the two alleles (at one locus) of an individual from one sub-population are identical and Q_S as the probability that two alleles drawn at random from two distinct individuals from the same sub-population are identical. Because $Q_I = 1 - \overline{H_o}$ and $Q_S = 1 - H_s$ we obtain:

$$F_{IS} = \frac{1 - Q_S - 1 + Q_I}{1 - Q_S} = \frac{Q_I - Q_S}{1 - Q_S} \qquad (7)$$

which corresponds to the general definition for F_{IS} (see Rousset, 2004).

4. Population structure, Wahlund effect and *F*-statistics

Living organisms generally are not homogeneously distributed across their vital domain. Most natural populations are subdivided into sub-populations of limited size. The structure of a population has much influence on the distribution of genetic information. To explore the consequences of population structure, Wright (1951) imagined a particular model he called the infinite island model. In this model, the population is composed of an infinity of sub-populations (or demes) of equal size *N*. At each non-overlapping generation, a deme is built with $(1 - m)N$ philopatric individuals coming from the same sub-population and mN migrants coming from all existing sub-populations. Because of their limited size, these sub-populations will tend to drift. Because drift is a random process, it should lead to genetic divergence between sub-populations. Migration will have the reverse effect, and will tend to homogenise allelic frequencies across sub-populations. Let us study the polymorphism at one locus with two alleles. If $\bar{p}$ and $(1 - \bar{p})$ are the mean allelic frequencies of the first and second alleles over all sub-populations, then the mean expected homozygosity across sub-populations will be $\overline{p^2 + (1 - p)^2}$, if all sub-populations are locally panmictic. The global expected homozygosity will be $\bar{p}^2 + \overline{(1 - p)}^2$. Thus, if we ignore the structure of the population, using Eq. (7) we obtain a F_{IS}-like fixation index:

$$F'_{\mathrm{IS}} = \frac{\overline{p^2} + \overline{(1-p)^2} - \bar{p}^2 - \overline{(1-p)}^2}{1 - \bar{p}^2 - \overline{(1-p)}^2}$$

which can be written in the more compact form:

$$F'_{\mathrm{IS}} = \frac{\overline{p^2} - \bar{p}^2}{\bar{p}(1-\bar{p})} \tag{8}$$

The numerator is a variance-like term and is thus always positive (or null when allelic frequencies are identical across sub-populations). Thus, if allelic frequencies are not identical then a heterozygote deficit is expected for the whole population. This is a Wahlund effect (Wahlund, 1928). This effect corresponds to the inbreeding (homozygosity) due to the subdivision of the population into separated sub-populations.

In this kind of hierarchical framework (individuals within sub-populations, sub-populations within total population, total population), three F-statistics can thus be defined (Wright, 1965). F_{IS} measures the inbreeding of individuals that is due to the local non-random union of gametes in each sub-population. F_{ST} reflects the inbreeding resulting from the subdivision of the population into sub-populations of limited size that do not freely exchange migrants; it is thus a measure of the Wahlund effect along with a measure of genetic differentiation between sub-populations. F_{IT} is the inbreeding of individuals in the total population resulting from both the previous phenomena. A general expression for these indices can also be expressed as functions of probabilities of identity of two alleles within individuals Q_{I}, between two individuals within sub-populations Q_{S} and, between sub-populations QT (e.g., Rousset, 2004):

$$F_{\mathrm{IS}} = \frac{Q_{\mathrm{I}} - Q_{\mathrm{S}}}{1 - Q_{\mathrm{S}}}; \qquad F_{\mathrm{ST}} = \frac{Q_{\mathrm{S}} - Q_{\mathrm{T}}}{1 - Q_{\mathrm{T}}}; \qquad F_{\mathrm{IT}} = \frac{Q_{\mathrm{I}} - Q_{\mathrm{T}}}{1 - Q_{\mathrm{T}}} \tag{9}$$

Note that from Eq. (8) under an infinite island model we obtain the original formulation of Wright's (1965) F_{ST}:

$$F_{\mathrm{ST}} = \frac{\sigma^2(p)}{\sigma^2_{\max}(p)} \tag{10}$$

where $\sigma^2_{\max}(p)$ is the maximum possible variance of allelic frequencies across sub-populations, i.e., when each sub-population is fixed for different alleles (e.g., for two alleles $\bar{p}$ populations and $(1-\bar{p})$ are fixed for alleles 1 and 2, respectively).

Other models of structured populations that help study the effects of different ecological constraints also exist. Such models involve geographical distances in a continuous population framework (neighbourhood models) (e.g., Rousset, 2000; Leblois et al., 2004) or in discrete patterns (stepping stone models) (e.g., Slatkin, 1985).

From Eq. (9), we can see that F_{ST} varies between $F_{\mathrm{ST}} = 0$, when genetic identity between individuals is independent from the sub-population where they are (no differentiation) and $F_{\mathrm{ST}} = 1$, when all individuals of the same sub-population are identical ($Q_{\mathrm{S}} = 1$) but differ in different sub-populations ($Q_{\mathrm{T}} < 1$), meaning a complete independence among sub-populations (e.g., expected if sub-populations are completely isolated for a long period of time). F_{IT} varies between $F_{\mathrm{IT}} = -1$, when all individuals are heterozygous for the same two alleles and $F_{\mathrm{IT}} = 1$ when all individuals are homozygous with at least two alleles. When the probability of sampling two identical genes over all the meta-population is independent of where an individual comes from, then $Q_{\mathrm{I}} = Q_{\mathrm{S}} = Q_{\mathrm{T}}$ and a global conformity to Hardy–Weinberg expectations is observed with $F_{\mathrm{IS}} = F_{\mathrm{ST}} = F_{\mathrm{IT}} = 0$. From Eq. (9) one can see that the three F-statistics are connected by the famous (at least for population geneticists) relationship: $(1 - F_{\mathrm{IT}}) = (1 - F_{\mathrm{IS}})(1 - F_{\mathrm{ST}})$.

Similar to the way that F_{IS} can be translated into ecologically relevant characteristics, F_{ST} can be linked to the structure of the meta-population in terms of number of migrants. In an infinite island model of locally panmictic sub-populations, the probability of identity between two demes is null (in fact $Q_{\mathrm{T}} \approx 1/\infty$ because of the infinite number of demes) and thus F_{ST} is equal to the probability of identity between individuals within sub-populations (Q_{S}). At any generation t, this probability of identity is $Q_{\mathrm{S}(t)}$, which is the proportion of identical genes in a sub-population. At $t + 1$ this proportion will be increased by the proportion of different genes $(1 - Q_{\mathrm{S}(t)})$ that were randomly sampled twice. In a sub-population of size N this probability is $(1/2N)^2$. As sampling must be repeated $2N$ times for building the N diploid individuals of the next generation, the general increase in identity is $1/2N$. Thus, at $t + 1$, the proportion of identical genes in any sub-population will be $Q_{\mathrm{S}(t)}+(1 - Q_{\mathrm{S}(t)})/2N$, providing none of these genes came from another sub-population, which is true with probability $(1 - m)^2$. Knowing all this, and hoping we have not yet lost all readers, we can write:

$$Q_{\mathrm{S}(t+1)} = (1-m)^2\left[Q_{\mathrm{S}(t)} + (1 - Q_{\mathrm{S}(t)})\frac{1}{2N}\right] \tag{11}$$

At equilibrium $Q_{\mathrm{S}(t+1)} = Q_{\mathrm{S}(t)} = \hat{Q}_{\mathrm{S}}=((1 - m)^2/2N)/(1-(1 - m)^2+((1 - m)^2/2N))$, which gives:

$$\hat{Q}_{\mathrm{S}} = \frac{(1-m)^2}{2Nm(2+m) + 1 - 2m + m^2}$$

Assuming small values of m and replacing Q_{S} by F_{ST} we reach the classical formula:

$$F_{\mathrm{ST}} \approx \frac{1}{4Nm + 1} \tag{12}$$

From this, the number of migrants can theoretically be estimated as $N_{\mathrm{m}} = (1 - F_{\mathrm{ST}})/4F_{\mathrm{ST}}$. If mutation must be taken into account, with an IAM and a mutation rate u, Eq. (12) becomes:

$$F_{\mathrm{ST}} \approx \frac{1}{4N(m+u) + 1} \tag{13}$$

From here it is easy to see that with high mutation rates F_{ST} will never equal 1, even when $m = 0$. It is worth noting that Eqs. (12) and (13) assume equilibrium between migration (and mutation) and drift in an infinite island model. Relaxing these assumptions can strongly limit our ability to make inferences regarding the effective number of migrants (e.g., Whitlock and McCauley, 1998). For this reason, other models, alternatives to the infinite island model, of subdivided populations also exist. When the number of sub-populations and alleles become small then equilibrium values for F-statistics are altered (e.g., Rousset, 1996). The migration model can also be different and, for instance, depends on the geographical distances that separate individuals (neighbourhood models) or sub-populations (stepping stone models). In this case, it is more appropriate to study the correlation between genetic differentiation of pairs of individuals or sub-populations and their geographical distances (Rousset, 1997, 2000).

5. Unbiased estimators of *F*-statistics

The F-statistics we have just seen correspond to the parametric definitions used if all individuals of all populations are sampled. If not all individuals are sampled, we can only estimate the true parameter value. This estimate must be unbiased (the average of the estimate must be equal to the true parameter value) and (as far as possible) of low variance. For instance, it is for this reason that we estimate the variance of a statistic x with:

$$s^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{14}$$

instead of with the parametric definition:

$$\sigma^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{15}$$

As illustrated in Eq. (10), Wright's F-statistics can be assimilated to variances ratios. Unbiased estimators should be able to take into account redundancy of information used while estimating these variance ratios. Weir and Cockerham's (1984) f, θ and F are unbiased estimators of F_{IS}, F_{ST} and F_{IT}, respectively (see also Balloux and Goudet, 2002). These values are obtained from a nested analysis of variance of gene frequencies. If σ_{a}^2, σ_{b}^2 and σ_{w}^2 are respectively, the among sub-populations, the among individuals within sub-populations and the among alleles within individuals components of variance of the allele frequencies, then we have:

$$f = \frac{\sigma_{\mathrm{b}}^2}{\sigma_{\mathrm{b}}^2 + \sigma_{\mathrm{w}}^2}; \qquad \theta = \frac{\sigma_{\mathrm{a}}^2}{\sigma_{\mathrm{a}}^2 + \sigma_{\mathrm{b}}^2 + \sigma_{\mathrm{w}}^2};$$
$$F = \frac{\sigma_{\mathrm{a}}^2 + \sigma_{\mathrm{b}}^2}{\sigma_{\mathrm{a}}^2 + \sigma_{\mathrm{b}}^2 + \sigma_{\mathrm{w}}^2} \tag{16}$$

Full expression of these variance components are quite cumbersome and can easily be found in the literature. It is worth noting that if F and f range from −1 to +1 as F_{IT} and F_{IS} do, θ can vary from −1 to +1. Negative values, which are impossible for the parameter F_{ST}, are met when the allelic frequencies of the samples differ less than expected by chance (through sampling variance).

6. Other measures and estimators of population differentiation

As seen from Eq. (13), the maximum value of F_{ST} is lower than 1 for high mutation rates. Mutation can also follow a strict SMM. In these cases, F_{ST} may poorly reflect the amount of gene flow. Several other measures have thus been proposed. R_{ST} was proposed by Slatkin (1995) to measure population differentiation with loci following a strict SMM. The same principle can be applied to local heterozygote deficits (Rousset, 1996) and unbiased estimators that take into account the variance in allele size have been designed (e.g., Rousset, 1996). If the mutation model does not follow a strict SMM, it is wiser to disregard such measures and use Weir and Cockerham's (1984) estimators (Balloux et al., 2000; Balloux and Goudet, 2002). Hedrick (1999, 2005) provided a tool to estimate how far a measured F_{ST} can be from the maximum possible value that would be observed with no migration when the number of alleles at the loci is big. This quantity can be estimated as a function of $Q_{\mathrm{S}} = (1 - H_{\mathrm{s}})$, which is in fact the maximum possible differentiation if the number of sub-populations is large.

Population differentiation is often measured through genetic distances between population pairs. Several measures exist. F_{ST} can of course be estimated between a pair of populations, but will suffer from some caveats (Rousset, 1997; Balloux and Goudet, 2002) and other estimators may be preferred depending on the aim of the study. For isolation by distance studies, $\theta/(1-\theta)$ will be of use (Rousset, 1997), while for others (e.g., tree construction) Cavalli-Sforza and Edwards chord distance (Cavalli-Sforza and Edwards, 1967) may be preferable (Takezaki and Nei, 1996; Kalinowski, 2002). This distance is obtained by the following formula:

$$D_{\mathrm{c}} = \frac{2}{r\pi}\sum_{j=1}^{r}\sqrt{2\left[1 - \sum_{i=1}^{mj}\sqrt{x_{ij}y_{ij}}\right]} \tag{17}$$

where r is the number of loci, j the locus name (from 1 to r), i the allele name (from 1 to mj), mj the number of alleles at locus j, x_{ij} and y_{ij} are the frequencies of allele i at locus j for sub-populations x and y, respectively.

When the distance has to be computed between individuals, the shared allelic distance (Bowcock et al., 1994) can be more appropriate (see Prugnolle et al., 2005c). If N_{sa} is equal to the number of shared alleles between two individuals over all the L loci, then the shared allelic distance is $1 - N_{\mathrm{sa}}/2L$.

7. Linkage disequilibrium

When looking at more than one locus (and, as previously mentioned, this is desirable!), a problem may arise because alleles at different loci can be correlated. Let us assume that there are two loci A and B with two alleles each and that the respective genotypic frequencies are D_A, H_A and R_A for A_1A_1, A_1A_2 and A_2A_2 at locus A and D_B, H_B and R_B for B_1B_1, B_1B_2 and B_2B_2 at locus B. If the two loci are statistically independent (i.e., unlinked), we expect the occurrence of genotypes to equal the product of the corresponding single locus genotypic frequencies. For instance, if there is linkage equilibrium the frequency of the genotype $A_1A_1\_B_1B_1$ should be equal to D_AD_B, and so on for the other combinations. If this is not the case, the two loci are in linkage disequilibrium. Of course, physical linkage between loci can be the cause of an observed linkage, but selfing and clonality are also expected to generate this pattern. Population structure, especially when sub-populations are small and have very low dispersal rates, can also generate linkage disequilibrium among loci. Selection can potentially generate linkage between some loci as well. Because we only have access to the diploid phase most of the time, composite linkage measures must be applied ($A_1A_2\_B_1B_2$ cannot be distinguished from $A_2A_1\_B_2B_1$) (Weir, 1979, 1996). Because some reproductive systems, in particular clonality, lead to global linkage, some authors have developed multilocus measures of linkage disequilibrium (e.g., Agapow and Burt, 2001). However, the behaviour of such measures in different conditions of population structure has been poorly investigated so far, even if it was suggested that such studies are worth considering (e.g., De Meeûs and Balloux, 2004). Linkage equilibrium is often important to assume for statistical tests based on multilocus averages. If the different loci are strongly linked, the multilocus information is redundant and may lead to decision errors (see below).

8. Statistical tests

8.1. Basic notions

Because we can never (or extremely rarely) work on all individuals from all populations at all generation times, sampling has to be undertaken. We therefore expect a sampling variance to be introduced to our measure. This means that the value calculated from a sample of individuals (e.g., F_{IS}, F_{ST} estimators) has little chance of equalling the true (parametric) value of the natural population it comes from. This is where the statistical analysis comes into play, by offering a criterion (the probability of the test or *P*-value) for deciding if the deviation from an expected value can be explained by the sampling variance alone. An expected value must therefore be defined. This expected value is the value expected under a null model or null hypothesis (H_0). For instance, if one wants to know if an observed F_{IS} is in agreement with the value expected under random mating, then the null hypothesis is ‘‘the observed F_{IS} is not significantly different from that expected under panmixia, i.e., 0’’. The *P*-value of the test will be the probability with which the sampling variance can explain a deviation as big or bigger than the one observed and is also called the type I error. Type I error, also noted α, corresponds to the probability of falsely rejecting H_0 when it is true. Type II error (or β) is the probability of accepting H_0 when it is false. The limit for significance is classically and arbitrarily set at 0.05, but we will see that sometimes a lower bound is needed. If the *P*-value is lower or equal to this threshold, the data are considered to deviate significantly from the null hypothesis. Depending of the nature of the alternative hypothesis H_1 (Box 2), two different families of tests can be distinguished: bilateral and unilateral tests. From Box 2, one can see that the alternative hypothesis has important consequences. If one chooses to perform a unilateral test, it has to be chosen a priori (of course), but also has to be chosen wisely. For instance, the significance of F_{IS} is tested unilaterally most of the time (alternative hypothesis, H_1; $F_{IS} > 0$) because heterozygote deficits are what is typically expected. However, for clonal organisms where a heterozygote excess is expected (e.g., Balloux et al., 2003), another H_1 is relevant. It is worth noting that the example in Box 2 also illustrates (with an extreme case) the lack of power to detect a signal with small sample sizes (here the most extreme possible *P*-value is 0.25). This illustrates what is called the type II error. Obviously, type II error is very high in the example from Box 2.

Box 2. Imagine a coin that is thrown twice. Four different events can occur: two heads, one heads and one tails, one tails and one heads and two tails. Each of these events has an equal probability 1/4 to happen. When we flip the coin, two heads are obtained. We would now like to test for a possible bias in the result (e.g., an unbalanced coin). Three different tests, with different null (H_0) and alternative (H_1) hypotheses can be performed:

- *Unilateral test 1*: H_0; the coin is balanced, H_1; the coin gives more heads than expected by chance.
- *Unilateral test 2*: H_0; the coin is balanced, H_1; the coin gives fewer heads than expected by chance.
- *Bilateral test*: H_0; the coin is balanced, H_1; the coin is not balanced.

For the first test, the probability that we are looking for is the sum of the probabilities of events with as many or more heads than the observed one (here 1/4), divided by the sum of probabilities of all possible events (here 1). The *P*-value is thus 0.25. Similarly, for the second test the *P*-value is 1/4 + 1/4 + 1/4 + 1/4 = 1. Finally, for the bilateral test the *P*-value is the sum of probabilities of all events that are as rare or rarer than the observed, that is the probability of getting two heads plus the probability of getting two tails, thus the *P*-value = 0.5. Note that the bilateral *P*-value equals twice the minimum unilateral *P*-value.

8.2. Resampling and randomisation tests

In the example described in Box 2, an exact probability could be computed because the different possible events could easily be enumerated one after the other. This will rarely be possible for the analysis of genetic data from natural populations. Nevertheless, the use of computerised calculations enables us to test our data and also obtain excellent approximations of exact probabilities. Different procedures with different properties exist.

8.2.1. Bootstrap and Jackknife

The aim of both methods is to generate a distribution of values based on resampling the data and then estimate confidence intervals. The principle of the bootstrap is to

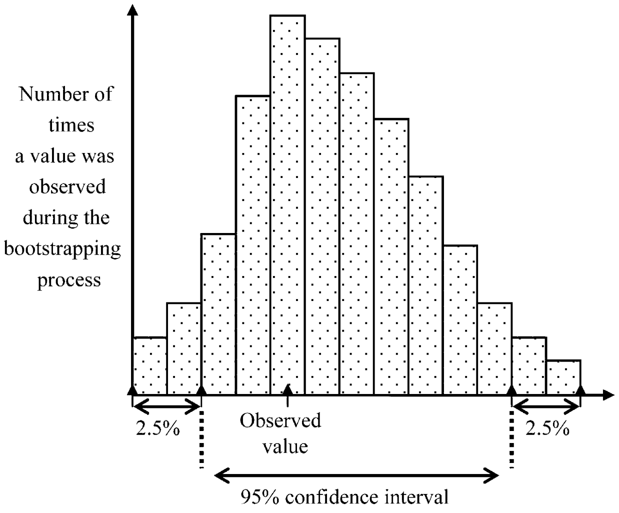

Fig. 1. Schematic representation of the bootstrap distribution of values for a given statistic and how a 95% confidence interval is obtained.

resample with replacement within the data in order to obtain a new data set. Repeating the operation a great number of times will provide a bootstrap distribution of the measured statistic (say F_{ST}) from which a confidence interval can be extracted (as shown in Fig. 1). In population genetics studies, we generally bootstrap over loci (as in Fstat, Goudet, 1995, freely downloadable at http://www.unil.ch/izea/softwares/fstat.html). There must be a certain number of loci for the confidence interval to reflect a biological reality (e.g., Raymond and Rousset, 1995a). This is why Fstat will compute such confidence intervals only when the number of loci is greater than four. The jackknife principle is to extract one item from the data (sub-sample, locus) and compute the statistic of interest with the remaining items. This provides as many statistic estimates as there are items (number of sub-samples or loci). These estimates can then be used to compute the standard error of the statistic. The standard error of a statistic x is simply:

$$s_{\bar{x}} = \sqrt{\frac{s^2(x)}{n}} \tag{18}$$

where n is the number of measures (e.g., of loci) and $s^2(x)$ is the estimated variance of x as in Eq. (14). From the standard error, the estimation of the confidence intervals is straightforward (Box 3). Fstat will not jackknife a data set when the number of samples or loci is not greater than four.

8.2.2. Randomisation tests

Two different kinds of randomisations can be described: the permutations and the Markov chain methods both are based on the Monte Carlo principle (see Metropolis, 1987 for an explanation of the name) of resampling the data. The permutation procedure is what is implemented in Fstat (Goudet, 1995). To carry out this procedure, the null hypothesis is simulated a great number of times (say 10,000) within the data set. For instance, alleles are randomly exchanged among individuals within each sub-samples to test for panmixia. F_{IS} is computed at each permutation and the proportion of values as big or bigger than the one observed (H_1 is $F_{IS} > 0$) gives the P-value of the test. In the same way, to test for population subdivision (H_0: individuals are randomly distributed among sub-samples), individuals are randomly assigned to the different sub-samples a great number of times and a statistic reflecting the variation in allelic frequencies is computed. The proportion of times a value as big or bigger than the observed one is obtained provides the P-value of the test. With the Markov chain (used in Genepop, Raymond and Rousset, 1995b, freely downloadable at http://wbiomed.curtin.edu.au/genepop), the randomisation procedure is different. The principle is to define a random walk between different possible contingency tables with identical marginal sums. The probability of occurrence of each table is computed and compared to that of the observed contingency table. The probability of the test is obtained by summing the number of times a probability as small or smaller than the observed one is obtained divided by the total number of steps. A more detailed description can be found in Raymond and Rousset (1995a). The P-values obtained are excellent approximations of exact probabilities, providing that the number of randomisations is large enough (say 10,000 for permutations and 1,000,000 for the Markov chain), which should not be a problem with the latest computers.

Box 3. Let us assume that we have a data set of eight samples with five loci. Over all loci $F_{ST} = 0.004$, and for locus 1 $F_{ST1} = 0.002$. The jackknife over loci (five values) gave a standard error of StdErrLoci(F_{ST}) = 0.003. The jackknife over populations (eight values) gave a standard error StdErrPop(F_{ST1}) = 0.001 for locus 1. Then, assuming the jackknife generates normal distributions we can use the t distribution with $\alpha = 0.05$ (Table 1 available at http://gemi.mpl.ird.fr/cepm/SiteWebESS/GB/deMeeus/SuplMat.html) to compute the 95% confidence interval for F_{ST} and F_{ST1} as:

$$\mathrm{CI}(F_{ST}) = F_{ST} \pm t_{0.05,\gamma L}\,\mathrm{StdErrLoci}(F_{ST});$$

$$\mathrm{CI}(F_{ST1}) = F_{ST1} \pm t_{0.05,\gamma P}\mathrm{StdErrLoci}(F_{ST1})$$

where $\gamma L = 5 - 1 = 4$ and $\gamma P = 8 - 1 = 7$ are the degrees of freedom for the jackknife procedures over loci and populations, respectively. From Table 1 (which gives values for intermediate degrees of freedom without the need to interpolate like in classic textbooks) we obtain $t_{0.05,\gamma L} = 2.776$ and $t_{0.05,\gamma P} = 2.365$. Thus, CI($F_{ST}$) = 0.004 ± 0.008 and CI(F_{ST1}) = 0.002 ± 0.002.

8.3. The problem of repeated tests

It is sometimes desirable to take into account multiple P-values, either because it is not possible to implement a global test or because one is interested in which of the available tests are significant. This can occur, for example, while exploring published results where raw data are not accessible. It is also the case for systematic testing between pairs of items like differentiation between pairs of samples or linkage disequilibrium between pairs of loci. There are two ways to treat this.

Combining all the tests provides one with a global testing for decision making and can be done relatively easily using Fisher's procedure (Fisher, 1970). If there are n P-values to be combined, then the quantity:

$$\chi^2_{\text{obs}} = -2\sum_{i=1}^{i=n}\ln(P_i) \qquad (19)$$

is a χ^2 variable with $2n$ degrees of freedom (to be compared to the corresponding expected χ^2 distribution). Applying this method requires the independence and the knowledge (not always given in publications) of all P_i. The binomial test, as applied in Prugnolle et al. (2002), is another solution for combining independent tests. If the null hypothesis is true, then a proportion of no more than 5% of significant tests is expected by chance at the 5% level of significance (by definition). We thus just need to adjust the observed proportion of significant tests to the expected one under the null hypothesis with an exact binomial test. In any case, it is always preferable to use a real global test instead of the Fisher procedure or the binomial test because such procedures cannot take into account the real weight of each individual test. Weights depend on sample sizes and the degree of polymorphism in each data set.

Determining what tests are significant is possible with Holm's (1979) methods called the Bonferroni procedure. In the case of multiple tests, the chance of finding a significant P-value (say $P \leq 0.05$) by chance alone is increased. The rationale behind the Bonferroni procedure is that if 100 tests are performed in a population supporting the null hypothesis (e.g., a panmictic population) then approximately five tests are expected to be significant by chance 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). It may be too conservative and be used with caution (e.g., see Nakagawa, 2004 and Section 8.6).

8.4. Testing panmixia

Testing panmixia means testing for a deviation from random mating. This can be done using the Haldane (1954) exact test and its generalisation for more than two alleles (Guo and Thompson, 1992) as implemented in Genepop (Raymond and Rousset, 1995a,b). However, there are several problems associated with this method because it tests the exact distribution of all genotypic classes in relation to Hardy–Weinberg expectations separately for each locus in each sub-sample. First, for loci with more than two alleles, the test can be significant because one genotypic class is in excess at the expense of another class, while all other classes are in agreement with Hardy–Weinberg expectations. This result may be difficult to interpret biologically. Second, a true global test, overall loci and sub-samples, is not available to date. We thus must deal with multiple P-values, which, in addition to the first problem, can be particularly disconcerting. The alternative is to consider F_{IS}, the value of which can be estimated and tested over all loci and sub-samples and which is easy to translate into biologically sound features (e.g., reproductive strategy). Two alternatives exist. In Genepop, the statistic used for testing is related to the Robertson and Hill (1984) estimator of F_{IS}. This estimator is biased but has a lower variance (a desirable statistical property for the power of a test), than the alternative Weir and Cockerham's estimator, which is the statistic used in Fstat (see Rousset and Raymond, 1995). To the experience of one of us (TdM) the difference between the two tests is subtle. In all cases, except when an exact Haldane test is possible, panmixia is simulated by randomising allele association within each sub-sample.

It should be noted that all of these procedures are indirect tests of panmixia. It is however sometimes possible to test it more directly when there is access to information about copulating adults. In such cases, it is possible to test if adults are pangamic, i.e., mate independently of their genotype. If male and female genotypes of each pair are available, one can compute the relatedness between males and females using, for example, the software KinshipV.1.2. (module relatedness) developed by Goodnight (http://gsoft.smu.edu/GSoft.html) (see also Queller and Goodnight, 1989). Then a Mantel correlation test (see Section 8.7) can be performed between the relatedness matrix and the matrix describing actual pair (belonging (1) and not belonging (0)). To our knowledge this has only been tested once in natural populations (Prugnolle et al., 2004).

8.5. Testing differentiation

For tests of genetic differentiation, there are again two possibilities. For the first possible test one has to assume that all loci are in agreement with Hardy–Weinberg expectations because alleles, irrespective of their correlation within individuals, are independently randomized among sub-samples. The exact test designed by Raymond and Rousset (1995a) and performed in Genepop uses this principle and is the most powerful test. However, because it is likely that some correlation exists within individuals it is nevertheless safer to undertake a test that takes into account the particular genotypic structure within each sub-sample, i.e., randomising individuals among sub-samples. Here the statistic used is the log-likelihood statistic G, the best statistic in most situations (see Goudet et al., 1996). The statistic is computed on allelic counts but randomisations involve genotypes (hence the term genotypic used for this test). A description of the algebraic formula for the G statistic can be found in any statistical textbook (e.g., Sokal and Rohlf, 1981) (see also Box 4).

Box 4. Let us assume we have sampled N individuals in two locations (Samples 1 and 2 of size N_1 and N_2, respectively). These individuals where genotyped at one locus with two alleles of frequencies p_1 and q_1 in Sample 1 and p_2 and q_2 in Sample 2, respectively. This gives the quantities given in the following table:

	Observed number of alleles		
	Allele 1	Allele 2	Sum
Sample 1	$2N_1p_1$	$2N_1q_1$	$2N_1(p_1+q_1)=2N_1$
Sample 2	$2N_2p_2$	$2N_2q_2$	$2N_2(p_2+q_2)=2N_2$
Sum	$2N_1p_1+2N_2p_2$	$2N_1q_1+2N_2q_2$	$2(N_1+N_2)=2N$

If we assume all individuals were sampled in the same population (no differences in allelic frequencies between the two samples expected) then the best estimate of the true allele frequencies are the mean of allele frequencies across samples. Thus, the expected numbers of alleles become:

	Expected number of alleles		
	Allele 1	Allele 2	Sum
Sample 1	$\frac{2N_1p_1+2N_2p_2}{2N}2N_1$	$\frac{2N_1q_1+2N_2q_2}{2N}2N_1$	$2N_1$
Sample 2	$\frac{2N_1p_1+2N_2p_2}{2N}2N_2$	$\frac{2N_1q_1+2N_2q_2}{2N}2N_2$	$2N_2$
Sum	$2N_1p_1+2N_2p_2$	$2N_1q_1+2N_2q_2$	$2(N_1+N_2)=2N$

P_{MO} is the multinomial probability of observing the cell frequencies if the observed data set is correct and P_{ME} is the multinomial probability of observing the cell frequencies if the expected data set is correct.

$$P_{\mathrm{MO}}=\frac{2N!}{2N_1p_1!2N_1q_1!2N_2p_2!2N_2q_2!}\left(\frac{2N_1p_1}{2N}\right)^{2N_1p_1}\times\left(\frac{2N_1q_1}{2N}\right)^{2N_1q_1}\left(\frac{2N_2p_2}{2N}\right)^{2N_2p_2}\left(\frac{2N_2q_2}{2N}\right)^{2N_2q_2}$$

$$P_{\mathrm{ME}}=\frac{2N!}{2N_1p_1!2N_1q_1!2N_2p_2!2N_2q_2!}\times\left[\frac{2N_1(2N_1p_1+2N_2p_2)}{(2N)^2}\right]^{2N_1p_1}\times\left[\frac{2N_1(2N_1q_1+2N_2q_2)}{(2N)^2}\right]^{2N_1q_1}\times\left[\frac{(2N_1p_1+2N_2p_2)N_2}{(2N)^2}\right]^{2N_2p_2}\times\left[\frac{(2N_1q_1+2N_2q_2)N_2}{(2N)^2}\right]^{2N_2q_2}$$

The log-likelihood ratio or G is twice the natural logarithm of the likelihood ratio or:
$G=2\ln(P_{\mathrm{MO}}/P_{\mathrm{ME}})$ which writes (see p. 736 and Box 17.6 in Sokal and Rohlf, 1981):

$$\begin{aligned}G={}&2N_1p_1\ln(2N_1p_1)+2N_1q_1\ln(2N_1q_1)\\&+2N_2p_2\ln(2N_2p_2)+2N_2q_2\ln(2N_2q_2)\\&+2N\ln(2N)-2N_1\ln(2N_1)\\&-(2N_1p_1+2N_2p_2)\ln(2N_1p_1+2N_2p_2)\\&-(2N_1q_1+2N_2q_2)\ln(2N_1q_1+2N_2q_2)-2N_2\ln(N_2)\end{aligned}$$

This quantity has additive properties, which means that the G's computed for several loci can be summed in order to get a global G value. Randomizing individuals across samples and computing the resulting global G gives a possible value under the null hypothesis of free migration of individuals across samples. Doing this a great number of times offers a distribution of possible G's under the null hypothesis to which the observed one can be compared. This is the G-based test implemented in Fstat to test for population differentiation. Please note that the two sample case is far from ideal and stands here for the sake of simplicity.

Another advantage of this statistic is its additivity, enabling a global test overall loci (as in Fstat).

8.6. Testing linkage between loci

There are two ways for testing linkage: between pairs of loci or overall loci. For the test between pairs of loci, there is either the exact test defined in Genepop and computed with the Markov chain method, or the G-based test performed in Fstat. The advantage of the G-based test is that a multi-sub-sample test is available. But if all P-values in all sub-samples are desired the two procedures give equivalent results. The ‘‘between-loci-pair’’ tests are designed when the knowledge of ‘‘which pair is in linkage disequilibrium’’ is of interest. In all cases, a Bonferroni correction must be applied because of the repetitive nature of the procedure. However, the number of tests may be so great that such a correction may render the test extremely conservative. For instance, with seven loci and nine sub-samples, the potential number of tests is $(7!/(6-2)!2!)10=210$, and thus the corrected level of significance for $\alpha=0.05$ becomes $\alpha'=0.05/210=0.0002$. Depending on sub-sample sizes and the degree of polymorphism of the loci, this limit may lie beyond the reach. It may therefore be wise to define a lower limit to the degree of polymorphism required in order for the loci to be tested. For example, loci with an allele present at more than 90% will rarely give a significant test of association with any other locus and can thus be disregarded for the analysis. What we need to do is to check if loci are not ‘‘too’’ linked and thus provide information that is not too redundant. There is no general agreement on how to deal with this kind of multiple testing and it is the choice of each empiricist to decide what to do (no correction, correction by the number of loci pair, total correction, . . .). Multilocus tests (e.g., Agapow and Burt, 2001) are specifically designed to test for a global effect, like the one expected under a clonal mode of reproduction. As suggested from simulations (De Meeûs and Balloux, 2004), the most accurate measure of multilocus linkage disequilibrium is Agapow and Burt's $\overline{r_{\mathrm{D}}}$ (Agapow and Burt, 2001). It is based on the index of association I_{A} (Brown et al., 1980; Maynard-Smith and Smith, 1998; Haubold et al., 1998), but renders the index independent of the number of loci. This measure is also used as a statistic in the randomisation test implemented in Multilocus (Agapow and Burt, 2001). The test is more powerful than the between pairs procedure within each sub-sample, but cannot be performed over all sub-samples and can become significant because of the linkage between some, but not all (as expected for clones) loci. The use of multilocus has thus some restrictions. We should also remember that unless the studied population is very large and at equilibrium, statistical linkage between loci is always expected to occur, even if totally panmictic.

8.7. Correlations between distance matrices

When a more or less linear, or at least monotonous, relationship is expected between genetic differentiation and another distance that may differentiate the same sub-samples, a special procedure, called the Mantel test (Mantel, 1967), must be used to detect and test it. Indeed, a classic regression analysis should not be used because of the non-independence of the data. A correlation is measured between the two matrices, the data in one of the two matrices are then randomized and the correlation between the randomized matrix and the other is measured each time. After a certain number of randomisations, the *P*-value is calculated as the number of times a value as great or greater than that observed was found. The test may be uni- or bi-lateral (Box 2). More discussion about the principle of the Mantel test can be found in Manly (1985) (see also Box 5). When examining correlation between matrices, two kinds of tests may be considered: isolation by distance tests and correlation tests between other kinds of matrices.

Isolation by distance was deeply investigated by Rousset (1997) who showed that the statistic to use for genetic distance should be $F_{ST}/(1-F_{ST})$ rather than F_{ST}. He also showed that wherever the studied species is distributed in two dimensions, geographic distance should be transformed into its natural logarithm, whereas no transformation is required for species colonising one-dimensional (linear) habitat. It is worthy of note that such a distinction is independent from the sampling design but strictly refers to the habitat occupied by the species under study (Rousset, 1997). Thus, linear habitats look rarer than two-dimensional ones but are perfectly accurate in cases such as the inter-tidal snail *Bendicium vitatum* (see Rousset, 1997 for a re-analysis) or the seabird tick *Ixodes uriae* (see below Section 9.2) that are distributed within narrow seashore bands. The rationale of Rousset's recommendations is as follows. Isolation by distance occurs whenever the amount of gene flow exchanged between two populations, and hence the probability Q_T of genetic identity between these populations, are decreasing functions of the geographical distance. Now, according to Eq. (9), Q_T appears both in numerator and denominator of the ratio defining F_{ST} but only in the numerator of that defining $F_{ST}/(1-F_{ST})$ [note that $F_{ST}/(1-F_{ST}) = (Q_S - Q_T)/(1-Q_S)$]. As a result, the simplest linear function is expected for the increase in $F_{ST}/(1-F_{ST})$ with geographical distance. This function is directly connected to the species demography at sampling time (Rousset, 1997; Leblois et al., 2004). Let σ be the average distance between the parents' and offspring's birthplaces, i.e., the mean gene dispersal range per generation. Let D be the average density of reproducing adults onto the sampled area. Let G_d the geographical distances separating two sampled populations. Then, for diploids, $F_{ST}/(1-F_{ST}) \approx A + \ln(G_d)/(4\pi D\sigma^2)$ and $F_{ST}/(1-F_{ST}) \approx A' + G_d/(4D\sigma^2)$ in two- and one-dimensional habitats, respectively. For haploids, regression slopes become $1/(2\pi D\sigma^2)$ and $1/(2D\sigma^2)$ in two- and one-dimensional habitats, respectively (Rousset, 2004). Slopes estimates can thus be used for estimating $D\sigma^2$ with additional biological information being required to disentangle D and σ. It is noteworthy that the accuracy in $D\sigma^2$ estimation is maximal for 'intermediate' geographical distances: given μ the mutation rate of the assessed loci, the maximal accuracy is obtained for geographical distances ranging from σ to $0.56\sigma/\sqrt{2\mu}$ in two-dimensional habitats, and from σ to $0.2\sigma/\sqrt{2\mu}$ in linear habitats (Rousset, 1997, 2004; Leblois et al., 2004). This technique was used to estimate the density and dispersal of the cattle tick (*Boophilus microplus*) in New-Caledonia (Koffi et al., in press).

Box 5. Let M_1 and M_2 two distance matrices between the same pairs of items:

$$M_1 = \begin{bmatrix} m1_{11} & m1_{12} & m1_{13} & m1_{14} \\ & m1_{22} & m1_{23} & m1_{24} \\ & & m1_{33} & m1_{34} \\ & & & m1_{44} \end{bmatrix} \quad \text{and}$$

$$M_2 = \begin{bmatrix} m2_{11} & m2_{12} & m2_{13} & m2_{14} \\ & m2_{22} & m2_{23} & m2_{24} \\ & & m2_{33} & m2_{34} \\ & & & m2_{44} \end{bmatrix}$$

A measure of the correlation between the two matrices can be given by

$$Z = \sum_i \sum_j m1_{ij} m2_{ij}$$

Z can be used as a statistic in the Mantel test. In that case the items of one matrix are randomized a number of times and the corresponding *Z* measured between the randomized matrix and the other one. The observed *Z* can then be compared to the distribution of randomized *Z*. Other statistics, as the ordinary Pearson coefficient of correlation, can of course be used for the Mantel test.

It has been argued several times that the F_{ST} (or its estimator θ) is not the best suited measure for examining the genetic differentiation between pairs of samples (e.g., Takezaki and Nei, 1996; Tomiuk et al., 1998; Kalinowski, 2002). Thus, depending on the hypothesis to be evaluated, using alternative measures of population differentiation may be wiser. However, it is probable that most measures will always converge on the same result. For instance, the correlation between different genetic distance measures, like between parasite infra-populations and between their hosts (e.g., Prugnolle et al., 2005c), was successfully assessed using a Mantel test between Cavalli-Sforza and Edwards (1967) genetic distances between worms infra-populations and the shared allele distance (Bowcock et al., 1994) between individual rats (P-value = 0.0005). The software MSA (Dieringer and Schlotterer, 2003) can compute such distances. If F_{ST} estimates are both used for hosts and parasites the test is not significant anymore (P-value = 0.15) and when Cavalli-Sforza and Edwards are studied for both hosts and parasites the P-value is 0.0113 (unpublished results). This illustrates that the choice of the statistic is not completely neutral.

8.8. Assigning individuals to sub-samples

The multilocus genotype of an individual, may help to compute its probability of belonging to a given sub-population (Rannala and Mountain, 1997; Waser and Strobeck, 1998; Cornuet et al., 1999). This probability is simply the expected multinomial probability of observing a particular multilocus genotype given the allelic frequencies at each locus in the sub-population. Obviously, the quality of this probability depends on that of the estimated allelic frequencies. The allelic frequencies therefore should be computed based on a sufficient number of individuals (≥30) and a good number of polymorphic loci (≥10). This probability is usually referred to as an assignment probability or assignment index. An individual with a low probability of belonging to the population from which it was sampled is likely to be a recent immigrant. Such probabilities may therefore be used to detect recent immigrants or to identify the population of origin (providing that all possible populations were sampled) by comparing the probability that an individual belongs to different populations. For more details about assignment probabilities and their applications, you may refer to Manel et al. (2005).

8.9. Biased dispersal tests

It is sometimes desirable to compare the population genetic structure of different categories of individuals that are encountered in each sub-sample. This is typically the case of males and females for dioecious species, or for infected and uninfected hosts. There are specific procedures to test if males disperse more or less than females or if infected hosts disperse less (or more) than uninfected ones. The most general test is implemented in Fstat and relies simply on the randomisation of the status of individuals (e.g., male or female) within each sub-sample, keeping the ratio (e.g., sex ratio) in each sub-sample unchanged. It then measures the difference obtained between the two categories for a chosen statistic (e.g., F_{ST}). The P-value of the test will be the proportion of times a difference as great or greater than the observed difference was obtained. This test can be done using various statistics: the assignment index and its variance corrected for population effects (AI_c and vAI_c), F_{IS}, F_{ST}, relatedness r computed as $2F_{ST}/(1 + F_{IT})$ (Goudet et al., 2002), H_s or H_o. The most dispersive category of individuals are indeed expected to display a lower AI_c, a higher vAI_c, a higher F_{IS}, a lower F_{ST}, r and H_s and a higher H_o. Depending on the circumstances some of these statistics behave better than others, but it seems that the F_{ST} is best in most situations, followed by AI_c and vAI_c (Goudet et al., 2002). In all cases, the most philopatric category must be almost immobile and the other category fairly mobile for a signal to be detected at least for a sex-biased dispersal (Goudet et al., 2002).

Testing for sex-biased dispersal may also be achieved by comparing autosomal markers to sex-linked markers. For instance, in species where the female is the philopatric sex, mitochondrial markers should be more structured than autosomal ones. However, because effective population sizes (Box 1) and mutation rates of these markers may significantly differ, the comparison may be difficult (e.g., see Prugnolle et al., 2002, 2003). Thus, comparing statistics obtained from each sex for the same autosomal markers may be easier. In this case, it should be noted that that the signal of a sex-biased dispersal disappears at each generation and its observation thus only reflects the dispersal events that occurred just before sampling.

In some cases (e.g., not enough sub-samples) and for certain statistics (e.g., F_{ST}), randomisation will not provide reliable results or may not even be possible. In these cases, some statistics, like F_{IS}, H_s or linkage disequilibrium, can still be compared between male and female individuals or between infected and uninfected individuals using other procedures. For instance, different loci (or pair of loci) can be used as independent replicates (hence the need for independent markers) and a Wilcoxon signed ranks test for paired data can be performed (e.g., Siegel and Castellan, 1988), the pairing unit being the locus or pair of loci (e.g., see Nébavi et al., 2006).

8.10. Comparing sub-populations

Under certain circumstances individuals are sampled in different types of sub-samples. This is the case of infra-populations of infectious agents sampled in female hosts as compared to infra-populations from male hosts. It may also correspond to organisms sampled in woods compared to others sampled in open fields within a heterogeneous landscape (e.g., bocage). In these situations, one may wonder if sub-samples differ in terms of various statistics (as above). Randomising sub-samples between categories, as undertaken in Fstat, allows one to test if the difference observed between sub-sample types is due to sampling error. Here again, when randomising the data is not an option, a Wilcoxon test for paired data is still possible (although less powerful).

8.11. Multivariate analyses

Multivariate analyses often provide convenient ways to represent the overall organisation of genetic data and sometimes include statistical inferences.

The factorial correspondence analysis (FCA), adapted for diploid organisms (She et al., 1987), positions each individual in a K dimension hyperspace (K being the total number of alleles summed over all loci) and projects each individual on the plane defined by the axes that best explain the shape of the scatter plot (same principle as for a least squares regression). It has sometimes proven useful to arrange individuals according to their genetic relatedness. For example this was done to study the occurrence of pathogens in a hybrid zone of their host (e.g., Coustau et al., 1991) and is particularly spectacular in the example extracted from Renaud (1988) and presented in Fig. 2. This technique may also help detecting cryptic population structure as in Solano et al. (2000) (see next section). The software GENETIX 4.05.4 (developed by Belkhir et al. and freely downloadable at http://www.univ-montp2.fr/~genetix/genetix/genetix.html) provides a very easy way to produce a FCA from genetic data (unfortunately, the help file is only in French!).

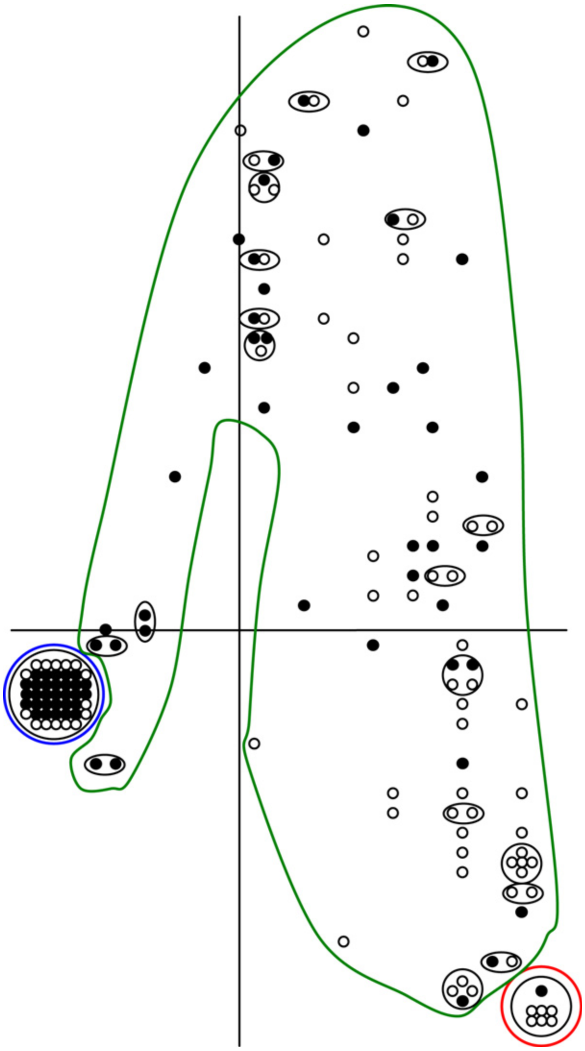

Fig. 2. FCA projection of individual *Barbus* fishes on the plane defined by the two first axes of the analysis, from genotypic data from nine enzymatic loci. Pure *Barbus barbus* genotypes are circled in red, pure *B. meridionalis* genotypes are circled in blue and hybrid between the two fish species are surrounded by a green line. Each small circle corresponds to an individual fish, with the black circles corresponding to fishes parasitized by the monogenean worm *Diplozoon gracile*. Circles surrounded by a black line are superimposed individuals (same genotypic coordinates on the two axes). The inversed U shaped scatter is typical of data that progressively change from one state to another (Guttman effect) (e.g., Wolff, 1996), like the alleles along a hybrid zone. The parasites are clearly following this trend as they increase in frequency when *B. meridionalis* allele numbers increase in individuals (Redrawn from Renaud, 1988).

Principal component analysis or PCA follows the same principle as the FCA, but use continuous ordinal data instead of disjunctive qualitative data. It aims at positioning groups of individuals (sub-samples) in a multidimensional scale. The advantage here is that the coordinates of sub-samples can be used for further statistical analysis as in Nébavi et al. (2006). The software PCA-GEN Version 1.2 (developed by Goudet and freely downloadable at http://www2.unil.ch/popgen/softwares/pcagen.html) produces this type of analysis from genetic data.

Canonical correspondence analysis (CCA) as implemented by the software CANOCO (Ter Braak, 1986, 1987; Ter Braak and Šmilauer, 2002) is a multivariate ordination method, designed to directly assess the relationship between multi-dimensional tables. It provides the advantage of integrating ordination and regression techniques and enables randomisation tests for the fit of data to environmental variables. This has been successfully used in some population genetic studies (Škalamera et al., 1999; Angers et al., 1999).

Dendrograms provide a very convenient way to present genetic data in a hierarchical arrangement. Such representations are very popular and found in any number of studies. One of the privileged fields of application of tree construction can be found in the molecular epidemiology of clonal organisms (e.g., see Taylor et al., 1999 for review).

8.12. Finding unknown population structure

There is sometimes no obvious direct evidence of population structure. In such situations, sampling strategies may not meet biological and/or ecological realities. If unknown biological/ecological factors contribute to the shape of the genetic architecture of the studied individuals it is then likely that these phenomena have left a genetic signature. Several procedures allow one to partition genetic datasets into potential sub-samples (henceforth called sub-populations). For instance, Solano et al. (2000) found huge heterozygote deficits in tsetse fly microsatellites that null alleles (see Section 8.14) could not totally explain. A FCA helped to identify the source of this observation as the result of a Wahlund effect (i.e., a hidden population structure). Other methods, based on Bayesian statistics and Markov Chain Monte Carlo simulations (Structure 2.01, Pritchard et al., 2000, http://pritch.bsd.uchicago.edu/software/structure2_1.html) or stochastic optimization (BAPS 4, Corander et al., in press, http://www.rni.helsinki.fi/~jic/bapspage.html) can also be used to infer the likelihood of hidden genetic structure in the data set as in Ravel et al. (in press).

8.13. Estimating effective population sizes and dispersal

We have already seen in Section 8.7 that demographic parameters could be estimated from genetic data from isolation by distance frameworks. In other cases, some demographic parameters can also be estimated.

Two different methods allow estimating effective population sizes (Box 1). Temporal studies allow estimating the variance effective sizes (N_e) of repetitively sampled populations (Waples, 1989). The software MACLEEPS 1.1 (Anderson et al., 2000) (downloadable at http://www.stat.washington.edu/thompson/Genepi/Mcleeps.shtml) performs maximum-likelihood estimates for different N_e using the allele frequencies shifts between generations. The computation assumes that selection, migration and mutation are negligible in changing allelic frequencies, compared to drift. A 95% confidence interval (CI) can be estimated (Anderson et al., 2000). Spatial studies allow estimating the inbreeding effective size of the populations using the software ESTIM 1.2 (Vitalis and Couvet, 2001a) (freely downloadable by anonymous FTP at ftp://isem.isem.univ-montp2.fr/pub/pc/estim). This software performs estimates of the two-locus identity disequilibria, η, within populations, together with a single locus parameter $F = Q_{1,i} - Q_2/(1 - Q_2)$, where $Q_{1,i}$ is the probability of identity of a pair of genes in sub-population i, and Q_2 the probability of identity for two genes in two different sub-populations (Vitalis and Couvet, 2001b). These two parameters, F and η, both depend on local N_e and m, the immigration rate, but not on ‘‘nuisance’’ parameters such as the mutation rate or mutation model (Vitalis and Couvet, 2001b,c). However, the selfing and recombination rates must be known. An example of how these two methods can be used and a discussion on their merits and problems can be found in Meunier et al. (2004b). Note that a synthetic method, taking into account information at both space and time scales, is now available (Wang and Whitlock, 2003) and a computer program MLNE estimating both N_e and m, can be downloaded for free at http://www.zoo.cam.ac.uk/ioz/software.html.

8.14. Some special cases

Null alleles are often met in population genetics studies, but are frequently ignored. Null alleles may be frequent in allozymes (Gaffney, 1994; Nébavi et al., 2006) and in DNA markers such as microsatellites (Paetkau and Strobeck, 1995; Pemberton et al., 1995; Brookfield, 1996). In allozymes, this would correspond to a loss of function and thus may not be neutral (i.e., deleterious) unless the organism keeps it at a heterozygous state as in clonal organisms such as *Candida albicans* (Nébavi et al., 2006). For DNA markers, it simply corresponds to a mutation in one of the primer sequences, such that the PCR amplification is not successful (Paetkau and Strobeck, 1995). The result is that many individuals that are heterozygous for a null (invisible) allele and another allele are falsely interpreted as being homozygous for the visible allele, which will artificially inflate the F_{IS} estimate. An easy way to check for null alleles is to observe the variation of F_{IS} across loci (e.g., De Meeûs et al., 2002a; Hurtrez-Boussès et al., 2004). Because null alleles are not expected to be present at the same frequency across all loci, the presence of such alleles is expected to strongly inflate the variation of heterozygosity across loci. Some relatively easy procedures are available to estimate the frequency of null alleles in a data set (e.g., Brookfield, 1996) and a software especially designed for microsatellite markers is also available (Micro-checker V 2.2.3., Van Oosterhout et al., 2004, available at http://www.microchecker.hull.ac.uk). A more subtle phenomenon, called short allele dominance or large allele drop-out, can also arise (Wattier et al., 1998; De Meeûs et al., 2004a; Ravel et al., in press). Here, for reasons that remain unclear, a competition for the polymerase exists such that shorter alleles have an amplification advantage over longer ones in heterozygous individuals. This phenomenon is nasty as it changes both F_{IS} and allele frequency estimates (De Meeûs et al., 2004a). A simple way to check for it is to regress F_{IS} measured on each allele against allele size (De Meeûs et al., 2004a). Micro-checker also identifies when this phenomenon is present.

When working from a very small amount of DNA or degraded material, it may happen that only one of the two alleles of a heterozygous individual is amplified at random, without any connection to allele size. This phenomenon is called allelic drop-out (e.g., Smith et al., 2000; Schneider et al., 2004; Criscione and Blouin, 2005b). It can be detected by inconsistencies between two amplifications of the same individual.

Clonal organisms must be treated with different tools because the consequences of this reproductive mode on the distribution of genetic information within individuals, between individuals, among populations and across loci. The theory and its application have attracted recent attention and have been the subject of recent reviews (Halkett et al., 2005; De Meeûs et al., 2006).

The classical situation with three levels of population structure (individual, sub-population and total population) is inappropriate if more hierarchical levels are involved. Instead of separating the data into independent replicates and using multiple tests, the approach proposed by HierFstat (Goudet, 2005, downloadable at http://www2.unil.ch/popgen/softwares/hierfstat.html) offers the possibility to analyse populations with any number of hierarchical levels and partitions the genetic variance into each of its within and between level components in a single analysis (e.g., see Trouvé et al., 2005, Nébavi et al., 2006).

9. Case studies

9.1. The population genetics of P. falciparum in Kenya

The population genetic structure of *P. falciparum*, the agent of malignant malaria, has been shown to be highly inbred in regions of low infectivity. In high-infectivity regions, it is often thought to be panmictic, or nearly so, although there were little supporting evidence for this claim. The matter can be settled by investigating the parasite's genetic make-up in the midgut oocysts of the mosquito vector, where diploidy occurs. Indeed, using a very original sampling design, Razakandrainibe et al. (2005) have investigated genetic polymorphism of *P. falciparum* oocysts from 145 mosquitoes in Kenya, where malignant malaria is perennial and intense. As seen in Fig. 3, there is considerable inbreeding, about 50% on average as seen from the F_{IT}. The inbreeding is due to selfing (about 25% estimated from Eq. (6)) and to the non-random distribution of oocyst genotypes among mosquito guts. All loci appeared to be in linkage disequilibrium. These results confirm that even in

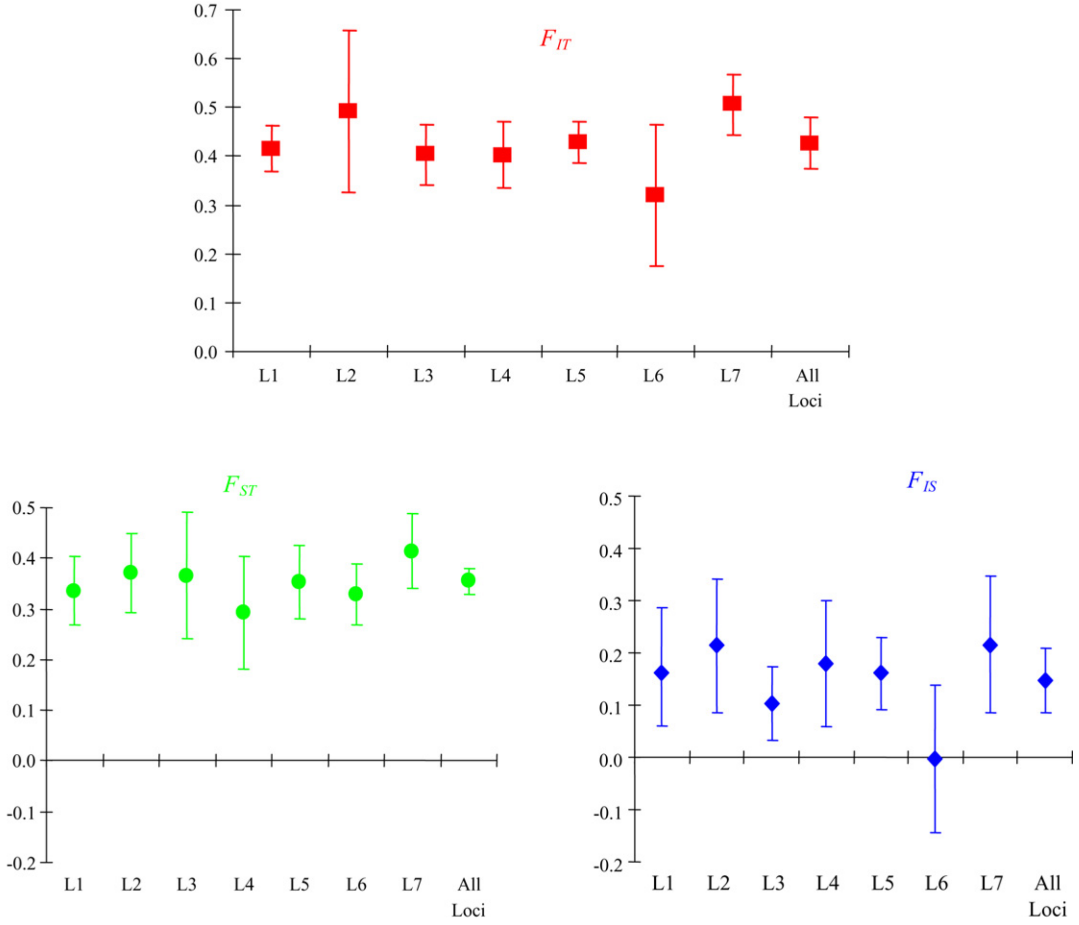

Fig. 3. Inbreeding statistics for diploid *Plasmodium falciparum* oocysts collected from *Anopheles gambiae* mosquitoes. L1–L7 refer to the seven microsatellite loci. F_{IS} measures the deficiency of heterozygous genotypes as a result of self-fertilization within individual mosquito guts; F_{ST} measures the deficiency of heterozygotes resulting from the non-random distribution of oocyst genotypes among mosquito guts; F_{IT} is the overall deficiency of heterozygotes in a population, resulting from both effects combined. From Razakandrainibe et al. (2005).

highly endemic zones, this parasite is far from panmictic and is composed of highly divergent, hence diverse, infra-populations across hosts, and are of considerable significance both evolutionarily and epidemiologically, particularly in relation to the spread of multilocus drug and vaccine resistance.

9.2. The population genetics of Ixodes ticks and epidemiology of Lyme borreliosis

Tick-borne diseases make up the overwhelming majority of human vector-borne infections in the temperate zones of the Northern Hemisphere, among which Lyme borreliosis has major public health and economic effects (Gubler, 1998). To date many questions remain unanswered regarding the epidemiology of this disease and the variability of its clinical manifestations (Hubbard et al., 1998). Using microsatellites and molecular probes, new insights could be obtained about the population biology of Lyme borreliosis vectors in Western Europe, *Ixodes ricinus* on the mainlands (De Meeûs et al., 2002a, 2004b) and *I. uriae* on sea birds in coastal environments (McCoy et al., 2001, 2003, 2005a,b). For example, the use of biased dispersal tests suggested a sex-biased dispersal (female ticks are philopatric) (De Meeûs et al., 2002a) in *I. ricinus* ticks, along with a biased dispersal of infected ticks (ticks infected with *Borrelia afzelii* disperse less) (De Meeûs et al., 2004b). As dispersal in ticks is host-dependent, these results suggest a biased host specificity of ticks of different sex and infectious status. For *I. uriae*, the coastal vector of Lyme borreliosis, strong structure has been found between ticks of different sea bird species in both hemispheres (Figs. 4 and 5), suggesting that host-race formation is a recurrent process (McCoy et al., 2001, 2005b). This degree of host specificity suggested a possible effect of host species behaviour on effective gene flow in the tick. This was confirmed by isolation by distance analyses showing strong discrepancies between *I. uriae* ticks from kittiwake and puffin hosts (Fig. 6) (McCoy et al., 2003). However, this dependency on host behaviour (in particular movements between and within colonies) is not reflected by the population genetic structure of the hosts as both kittiwakes (seven microsatellite loci) and guillemots (six microsatellite loci) seem to be panmictic, or nearly, at a continental scale (McCoy et al., 2005a; Riffaut et al., 2005). Dispersal of ticks between the two hemispheres seems to be entirely blocked according to the F_{ST} measured (0.38) that was nearly equal to the maximum possible F_{ST} $(1 - H_s = 0.42)$ given the variability of the marker. These different results highlight

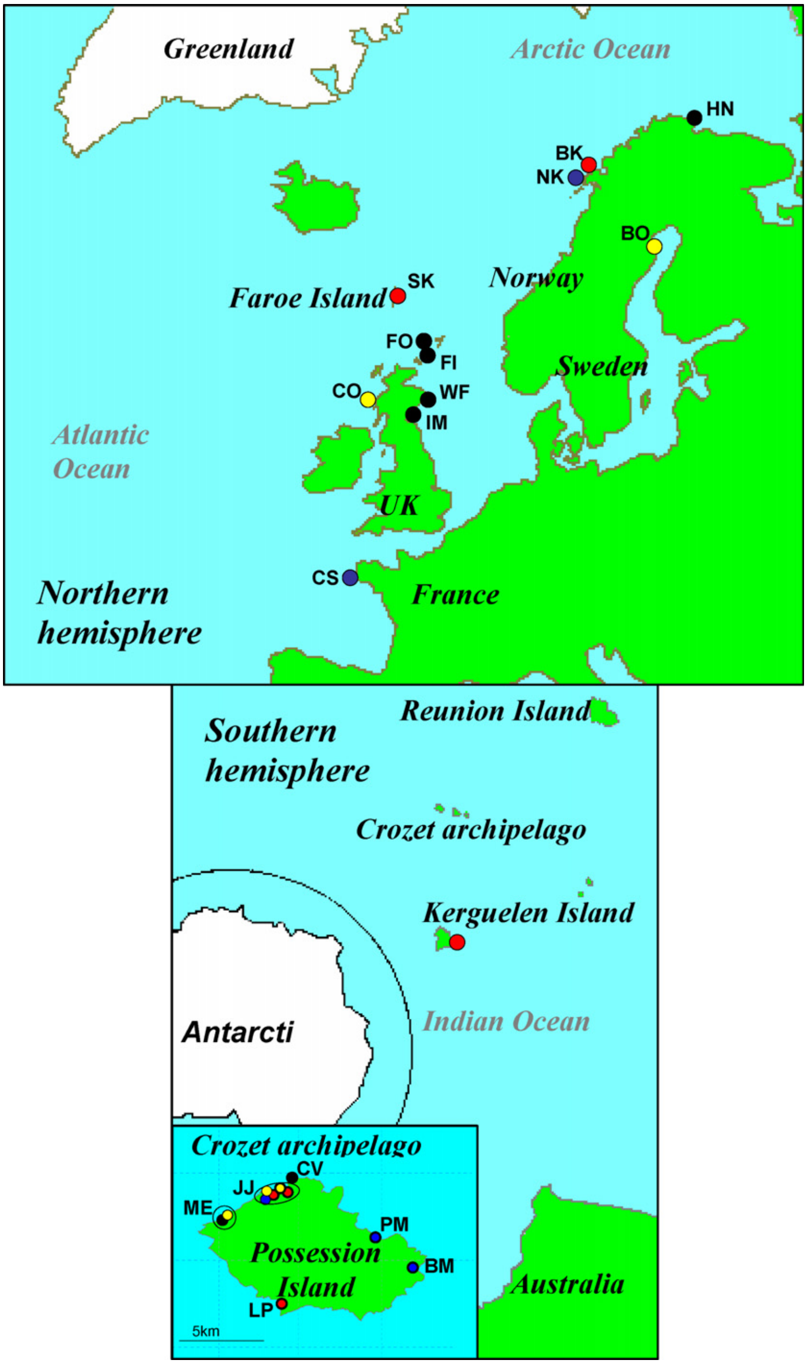

Fig. 4. Sampling sites for *Ixodes uriae* in the Northern and the Southern hemispheres. For the Northern hemisphere, black-legged kittiwake (*Rissa tridactyla*) colonies are in blue, common guillemot (*Uria aalge*) colonies are in yellow and Atlantic puffin (*Fratercula arctica*) colonies are in red. For the Southern hemisphere, king penguin (*Aptenodytes patagonicus*) colonies are in blue, macaroni and rock hopper penguins (*Eudyptes chrysocome*, *E. chrysolophus*) are in red and yellow, respectively. Sites in black correspond to mixed colonies where at least two bird species breed in sympatry. From McCoy et al. (2005b).

how complex the epidemiology of Lyme borreliosis can be and indicate the focus of future studies. The sex of the vector, the effect of the micropathogen on hosts (including the vectors) and host behaviour are all relevant. The host specificity of the coastal vector adds a new dimension to the epidemiological picture with a whole string of consequences for dispersal patterns of this pathogenic agent and the still unsolved links with its inland cycle.

9.3. The population genetics of the S. mansoni–Biomphalaria glabrata–Rattus rattus system

Schistosome flukes are dioecious trematodes (separate sexes) responsible for one of the most important human parasitic diseases (schistosomiasis, also known as bilharziasis) in tropical countries. Some 200 million people are infected, of which 20 million are thought to suffer severe consequences of

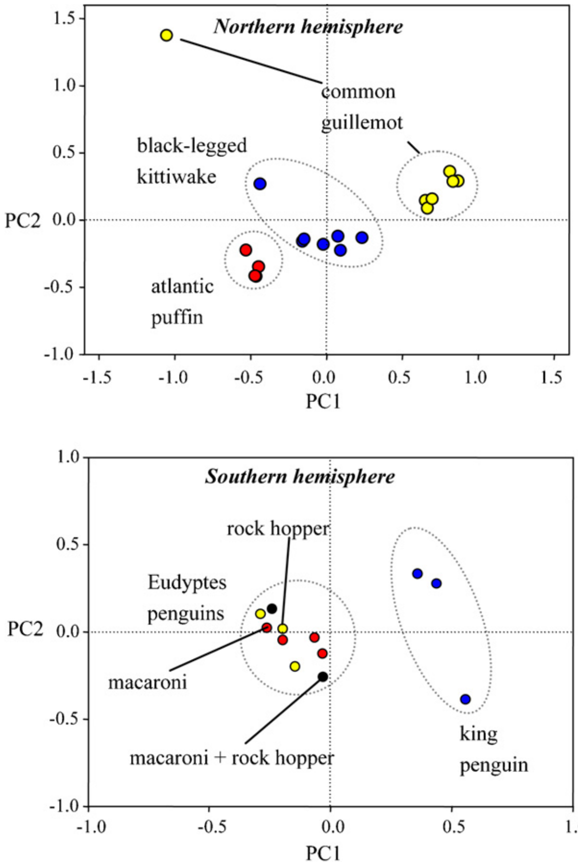

Fig. 5. PCA analysis of genotypes at eight microsatellite loci of different sub-samples of the tick *I. uriae* from different sea bird colonies. The colours populations are as in Fig. 4. For the Northern hemisphere, the first two axes explained more than 68% of the dispersion of the data whereas for the Southern hemisphere the first two axes explained more than 58% of the dispersion of the data. From McCoy et al. (2005b).

infection (Chitsulo et al., 2000), making this parasitic disease ranking second after malaria (Morel, 2000). In the insular focus of Guadeloupe (French West Indies), *S. mansoni* has permanently shifted from humans toward the black rat *R. rattus*, which is now its principal (if not only) host on this island (e.g., Prugnolle et al., 2002). This enabled us to sample adult worms that live in the mesenteric venules of the vertebrate host. In water, eggs hatch into swimming larvae (miracidia) that must infect an aquatic snail (*B. glabrata*). From the snail host, the parasite produces a massive amount of clonal cercariae (second infective stage) that must then find a rat to complete the cycle. By sampling the infra-population of worms in rats at a local scale and analysing their genotype at seven microsatellite loci, Prugnolle et al. (2002) showed a sex-biased genetic structure (females are more differentiated between infra-populations, biased dispersal test). Based on the life cycle and the strong heterozygote excesses observed ($F_{IS} = -0.1$ for female schistosomes), this could not be interpreted as the result of a sex-biased dispersal (females disperse less) alone. Further studies showed that this pattern probably emerged as an interaction between the effect of a strong variance in clonal success (Prugnolle et al., 2005a,b) along with selective processes (Prugnolle et al., 2004) and a possible role of host's sex (Caillaud et al., in press). At a regional scale, Prugnolle et al. (2005c) genotyped all protagonists with microsatellite loci and using Mantel tests demonstrated a strong pattern of isolation by distance for the snails and a structuring that was independent of geography for both rats and schistosomes. A highly significant correlation was also shown between rats and the schistosome infra-populations they harboured (Fig. 7). This confirmed that rats were the real dispersal vehicles of schistosomes.

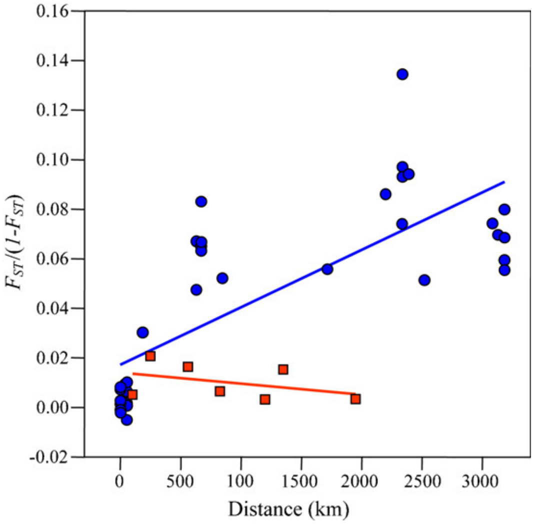

Fig. 6. Isolation by distance test between genetic differentiation, measured as $F_{ST}/(1 - F_{ST})$, between pairwise *I. uriae* sub-samples and geographic distance (km) measured along the Atlantic coast from kittiwake (in blue) and puffin (in red) colonies. The linear option of regression between $F_{ST}/(1 - F_{ST})$ and geographical distances was chosen since the habitat of *I. uriae* is limited to the narrow coastal band where seabirds reproduce. The test was significant only for kittiwake ticks (Mantel test, *P*-value = 0.004), whereas no relationship was found for puffins (*P*-value = 0.917). From McCoy et al. (2003).

9.4. The hidden population structure of tsetse flies

Glossina palpalis gambiensis is a riverine West African tsetse fly that transmits trypanosomes causing both human and animal African trypanosomiasis. A preliminary survey of this species in the site of Nyafaro in Burkina Faso revealed a strong and significant heterozygote deficit at this site. A FCA analysis

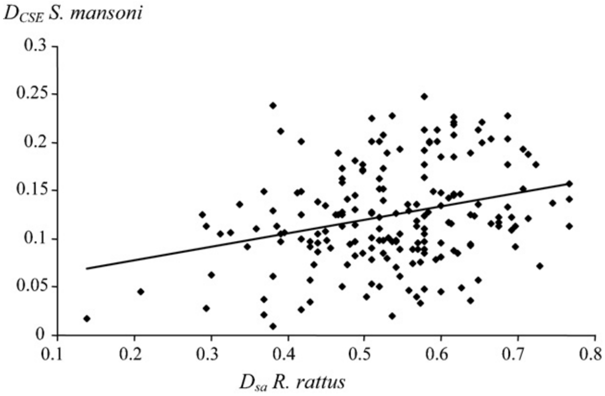

Fig. 7. Relationship between the shared allelic distance (D_{sa}) computed between individual rats and Cavalli-Sforza and Edwards genetic distance (D_{CSE}) computed between schistosome infra-populations harbored by these rats ($R = 0.29$, $p = 0.0005$) (From Prugnolle et al., 2005c).

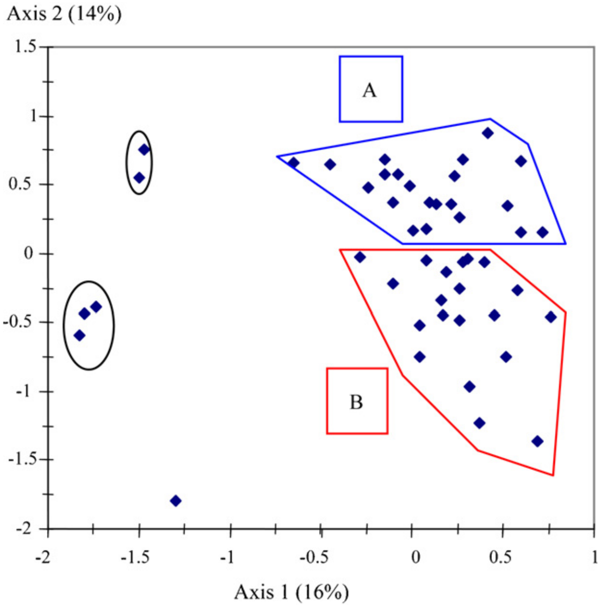

Fig. 8. Projection of tsetse fly (*G. palpalis gambiensis*) individuals, based on their genotype at three microsatellite loci, on the two first axes of FCA analysis. Two main clusters, where a F_{IS} can be computed, can be identified (A, surrounded in blue and B, surrounded in red). Within each group, F_{IS} is not significantly different from 0. However, this is not the case when all individuals are considered together ($F_{IS} = 0.2$, P-value < 0.001). The percentage of total inertia explained by each axis (which represent how well the axis fits the data), is provided (From Solano et al., 2000).

revealed that this sample was heterogeneous and composed of at least two main groups (Fig. 8) where the heterozygote deficits did not stay significant anymore (Solano et al., 2000). A re-analysis of these data with BAPS (cf. Section 8.12) clustered the data into nine putative sub-populations where the heterozygote deficits dropped to negative values (unpublished). This suggested that tsetse flies are strongly clustered into small sub-populations and that the traps used to capture these flies are sufficiently attractive to catch individuals from different sub-populations, thus creating a Wahlund effect. A recent survey in Bonon (Côte d'Ivoire) with five microsatellite loci of *Glossina palpalis palpalis* revealed that a great part of heterozygote deficits could be explained by null alleles and short allele dominance (Ravel et al., in press). However, the use of BAPS allowed to partition the data into many clusters, unrelated to sampling site (trap) in which a significant drop of F_{IS} could be observed, thus suggesting again a strong Wahlund effect in this sub-species as well (see Ravel et al., in press).

9.5. The population structure and reproductive mode of C. albicans

C. albicans is a diploid opportunistic yeast present in the gastrointestinal and genitourinary flora of most healthy humans and other mammals. It is seriously pathogenic only in immunocompromised patients (Hull et al., 2000; Berman and Sudbery, 2002). The sexual cycle was demonstrated under experimental conditions but, how diploidy is restored, how often it occurs in the wild and what kind of population structure this yeast displays, remain largely unknown. A recent study by Nébavi et al. (2006) has shed new light on all of these aspects. The authors analyzed a sample consisting of five isolates in each of 42 HIV+ patients (19 women and 23 men) genotyped for 14 enzymatic loci. This sampling design, combined with the use of PCA, HierFstat and the latest clonal genetic tools (De Meeûs and Balloux, 2005; De Meeûs et al., 2006) demonstrated an almost total (if not absolute) clonal mode of reproduction and extremely strong genetic differentiation between individual patients. Female patients apparently maintained a higher genetic diversity of *C. albicans* strains, suggesting larger infra-populations of this yeast in female environments. Nonetheless, these results were indicative of a very low level of strain exchange between patients, and/or a modest infection role of environmental strains (each individual keeps its own strains). This study also indicated the crucial importance of sampling design, particularly with these kinds of organisms, in order for correct inferences to be made (illustrated by Fig. 9). Interestingly, the Wahlund effect produced by inappropriate sampling translates into F_{IS} values expected under panmictic conditions. In particular, the classic sampling design for such organisms of one isolate per individual patient with all patients pooled into a single sample would have produced a F_{IS} very close to 0 (Fig. 9).

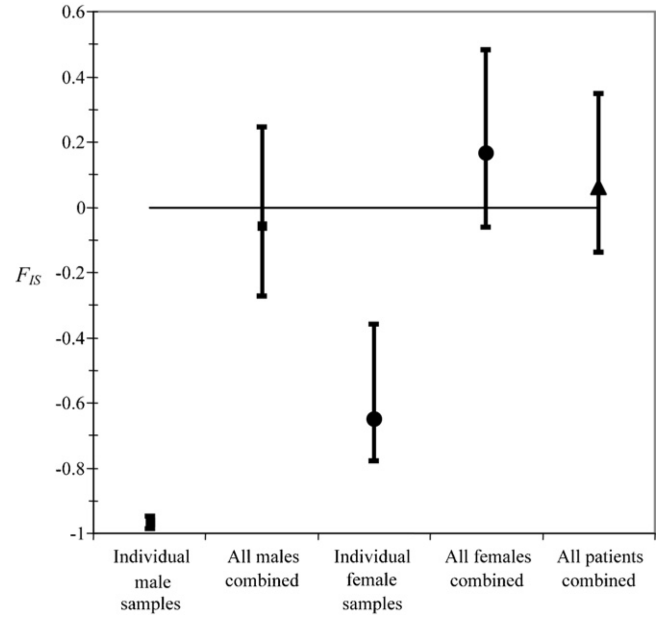

Fig. 9. Demonstration of the implications of an incorrect sampling design on estimation of the within sub-population fixation index (F_{IS}) for the *Candida albicans* data set. Samples from male patients are represented with solid squares and samples from female patients with solid circles. It is easy to see that when each male or female patient is considered as a sub-population, the picture provided by the analysis is very different from that observed when all male or all female patients are pooled, or when all patients (solid triangle) are considered together. Obviously, in this case, the real sub-populations are within patients and the other measures reflect strong Wahlund effects.

9.6. The population structure and reproductive modes of G. truncatula and F. hepatica

The liver fluke *F. hepatica* is a common parasite of the liver of many different vertebrates including man (see Hurtrez-

Boussès et al., 2001 for review), causing fasciolosis, a reemerging disease. It affects ca. 17 millions of people around the world (Meunier et al., 2001) and 180 millions people are exposed at the risk of disease (e.g., Hurtrez-Boussès et al., 2001). Adult worms sexually reproduce in the vertebrate host, are monoecious and probably self-compatible. In the intermediate host, a freshwater snail (mainly *G. truncatula*), the parasite experiences an intensive asexual multiplication. Using microsatellites, Meunier et al. (2001, 2004a) evidenced a preferential selfing mode of reproduction of the mollusc host, that contrasts with the apparent panmixia of the parasite (Hurtrez-Boussès et al., 2004). It was also shown that the Bolivian Altiplano, where fasciolosis prevalences reach the highest known values in human (Mas-Coma et al., 1999), is colonized by a single *G. truncatula* genotype. This fact is most likely explained by a recent and strong bottleneck (Meunier et al., 2001). This may explain in part the success of the fluke, which has to adapt to only one kind of mollusc and suggests that control campaign against snails could easily be undertaken with a very low threat of resistance evolution (Meunier et al., 2001).

10. Conclusion

After such a long article, it is perhaps wiser to conclude with a short discussion. We hope that we have convinced the reader about the interest of using molecular and population genetics tools for examining the population biology of infectious agents, their vectors and their reservoirs. We hope also that the references and techniques described in this review will allow some scientists interested in developing such approaches to begin with a high chance of success. We would like to insist on some very important aspects that are still too often neglected in molecular epidemiological studies. The sampling design is crucial, as testified by the *Plasmodium*, *Schistosoma* and *Candida* examples. Unfortunately, these examples still represent the rarest cases. Other sampling designs would have led to completely different pictures and thus to poor (if not wrong) inferences. Using clustering techniques, as employed for the tsetse example, enables the diagnosis of a wrong sampling design, but is not a cure. Thus, much remains to be done. The growing access to whole genome sequences will provide extremely useful tools in this perspective, as it will enable researchers to compare neutral markers with markers of known utility. Such knowledge will surely bring invaluable incite into the processes involved in the epidemiology and evolutionary genetics of infectious agents. "Parasite molecular ecology is still in its infancy, but it promises to be a rewarding field for those who embrace it" (Criscione et al., 2005).

Acknowledgements

This work was partly inspired from a meeting of a workshop financed by the "Réseau Ecologie des Intéractions Durables" (REID). A part of this article is based on works financed by French Ministry of Ecology and Sustainable Development (Contract no CV2000071, MEDD, programme "Ecosystèmes Tropicaux", and program "INVABIO2", MEDD, contract no. CV02000234). All authors are financed by the Centre National Pour la Recherche Scientifique (CNRS), the Institut de Recherches pour le Développement (IRD) and belong to the team "Structure, Genetics and Adaptation in Symbiotic Systems" (SGASS).

References

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

Agrawal, A., Lively, C.M., 2002. Infection genetics: gene-for-gene versus matching alleles models and all points in between. Evol. Ecol. Res. 4, 79–90.

Anderson, E.C., Williamson, E.G., Thompson, E.A., 2000. Monte Carlo evaluation of the likelihood for Ne from temporally spaces samples. Genetics 156, 2109–2118.

Angers, B., Magnan, P., Plante, M., Bernatchez, L., 1999. Canonical correspondence analysis for estimating spatial and environmental effects on microsatellite gene diversity in brook charr (*Salvelinus fontinalis*). Mol. Ecol. 8, 1043–1053.

Arnaviehle, 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.

Avise, J.C., 2000. Phylogegraphy: The History and Formation of Species. Harvard University Press, Cambridge, MA.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., Saunders, N.C., 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Ann. Rev. Ecol. Syst. 18, 489–522.

Badoc, C., De Meeûs, T., Bertout, S., Odds, F.C., Mallié, M., Bastide, J.-M., 2002. Clonality structure in *Candida dubliniensis*. FEMS Microbiol. Lett. 209, 249–254.

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

Balloux, F., Brünner, H., Lugon-Moulin, N., Hausser, J., Goudet, J., 2000. Microsatellites can be misleading: an empirical and simulation study. Evolution 54, 1414–1422.

Balloux, F., Goudet, J., 2002. Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. Mol. Ecol. 11, 771–783.

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

Barnabé, C., Brisse, S., Tibayrenc, M., 2000. Population structure and genetic typing of *Trypanosoma cruzi*, the agent of Chagas disease: a multilocus enzyme electrophoresis approach. Parasitology 120, 513–526.

Ben Abderrazak, S., Guerrini, F., Mathieu-Daudé, F., Truc, P., Neubaeur, K., Lewicka, K., Barnabé, C., Tibayrenc, M., 1993. Isoenzyme electrophoresis for parasite characterization. In: Hyde, J.E. (Ed.), Protocols in Molecular Parasitology. Humana Press, Totowa, NJ, pp. 361–362.

Berman, J., Sudbery, P.E., 2002. *Candida albicans*: a molecular revolution built on lessons from budding yeast. Nat. Rev. Genet. 3, 918–928.

Borges, E.C., Dujardin, J.P., Schofield, C.J., Romanha, A.J., Diotaiuti, L., 2000. Genetic variability of *Triatoma brasiliensis* (Hemiptera: Reduviidae) populations. J. Med. Entomol. 37, 872–877.

Bougnoux, M.E., Aanensen, D.M., Morand, S., Théraud, M., Spratt, B.G., d'Enfert, C., 2004. Multilocus sequence typing of *Candida albicans*: strategies, data exchange and applications. Infect. Genet. Evol. 4, 243–252.

Bowcock, A.M., Ruizlinares, A., Tomfohrde, J., Minch, E., Kidd, J.R., Cavalli-Sforza, L.L., 1994. High-resolution of human evolutionary trees with polymorphic microsatellites. Nature 368, 455–457.

Brenière, S.F., Barnabé, C., Bosseno, M.F., Tibayrenc, M., 2003. Impact of number of isoenzyme loci on the robustness of intraspecific phylogenies using multilocus enzyme electrophoresis: consequences for typing of *Trypanosoma cruzi*. Parasitology 127, 273–281.

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.

Caillaud, D., Prugnolle, F., Durand, P., Théron, A., De Meeûs, T., in press. Host sex and parasite genetic diversity. Micr. Infect., in press.

Caterino, M.S., Cho, S., Sperling, F.A.H., 2000. The current state of insect molecular systematics: a thriving tower of Babel. Annu. Rev. Entomol. 45, 1–54.

Cavalli-Sforza, L.L., Edwards, A.W.F., 1967. Phylogenetic analysis: model and estimation procedures. Am. J. Hum. Genet. 19, 233–257.

Chitsulo, L., Engels, D., Montresor, A., Savioli, L., 2000. The global status of schistosomiasis and its control. Acta Trop. 77, 41–51.

Corander, J., Marttinen, P., Mäntyniemi, S., in press. Bayesian identification of stock mixtures from molecular marker data. Fish. Bull., in press.

Corley, L.S., Blankenship, J.R., Moore, A.J., 2001. Genetic variation and asexual reproduction in the facultatively parthenogenetic cockroach *Nauphoeta cinerea*: implications for the evolution of sex. J. Evol. Biol. 14, 68–74.

Cornuet, J.M., Piry, S., Luikart, G., Estoup, A., Solignac, M., 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153, 1989–2000.

Coustau, C., Renaud, F., Maillard, C., Pasteur, N., Delay, B., 1991. Differential susceptibility to a trematode parasite among genotypes of the *Mytilus edulis*/*galloprovincialis* complex. Genet. Res. Camb. 57, 207–212.

Criscione, C.D., Blouin, M.S., 2005a. Effective sizes of macroparasite populations: a conceptual model. Trends Parasitol. 21, 212–217.

Criscione, C.D., Blouin, M.S., 2005b. Eleven polymorphic microsatellite loci for the salmonid trematode *Plagioporus shawi*. Mol. Ecol. Notes 5, 562–564.

Criscione, C.D., Poulin, R., Blouin, M.S., 2005. Molecular ecology of parasites: elucidating ecological and microevolutionary processes. Mol. Ecol. 14, 2247–2257.

De Meeûs, T., Beati, L., Delaye, C., Aeschlimann, A., Renaud, F., 2002a. Sex biased genetic structure in the vector of Lyme disease, *Ixodes ricinus*. Evolution 56, 1802–1807.

De Meeûs, T., Balloux, F., 2004. Clonal reproduction and linkage disequilibrium in diploids: a simulation study. Infect. Genet. Evol. 4, 345–351.

De Meeûs, T., Balloux, F., 2005. *F*-statistics of clonal diploids structured in numerous demes. Mol. Ecol. 14, 2695–2702.

De Meeûs, T., Durand, P., Renaud, F., 2003. Species concepts: what for? Trends Parasitol. 19, 425–427.

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

De Meeûs, T., Lehmann, L., Balloux, F., 2006. Molecular epidemiology of clonal diploids: a quick overview and a short DIY (do it yourself) notice. Infect. Genet. Evol. 6, 163–170.

De Meeûs, T., Lorimier, Y., Renaud, F., 2004b. Lyme borreliosis agents and the genetics and sex of their vector, *Ixodes ricinus*. Microb. Infect. 6, 299–304.

De Meeûs, T., Renaud, F., 2002. Parasites within the new phylogeny of eukaryotes. Trends Parasitol. 18, 247–251.

De Meeûs, T., Renaud, F., Mouveroux, E., Reynes, J., Galeazzi, G., Mallié, M., Bastide, J.M., 2002b. The genetic structure of *Candida glabrata* populations in AIDS and non-AIDS patients. J. Clin. Microbiol. 40, 2199–2206.

Dieringer, D., Schlotterer, C., 2003. Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. Mol. Ecol. Notes 3, 167–169.

Fisher, R.A., 1970. Statistical Methods for Research Workers, 14th ed. Oliver and Boyd, Edinburgh.

Gaffney, P.M., 1994. Heterosis and heterozygote deficiencies in marine bivalves: more light? In: Beaumont, A.R. (Ed.), Genetic and Evolution of Aquatic Organisms. Chapman and Hall, London, pp. 146–153.

Gandon, S., 2002. Local adaptation and the geometry of host–parasite coevolution. Ecol. Lett. 5, 246–256.

Gandon, S., Capowiez, Y., Dubois, Y., Michalakis, Y., Olivieri, I., 1996. Local adaptation and gene for gene coevolution in a metapopulation model. Proc. R. Soc. Lond. B 263, 1003–1009.

Gerber, A.S., Loggins, R., Kumar, S., Dowling, T.E., 2001. Does nonneutral evolution shape observed patterns of dna variation in animal mitochondrial genomes? Ann. Rev. Genet. 35, 539–566.

Goudet, J., 1995. FSTat Version 1.2: a computer program to calculate FSTatistics. J. Hered. 86, 485–486.

Goudet, J., 2005. HierFSTat, a package for R to compute and test hierarchical F-statistics. Mol. Ecol. Notes 5, 184–186.

Goudet, J., Perrin, N., Waser, P., 2002. Tests for sex-biased dispersal using biparentally inherited genetic markers. Mol. Ecol. 11, 1103–1114.

Goudet, J., Raymond, M., De Meeüs, T., Rousset, F., 1996. Testing differentiation in diploid populations. Genetics 144, 1933–1940.

Gubler, D.J., 1998. Resurgent vector-borne diseases as a global health problem. Emerg. Infect. Dis. 4, 442–450.

Guo, S.W., Thompson, E.A., 1992. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 48, 361–372.

Haldane, J.B.S., 1954. An exact test for randomness of mating. J. Genet. 52, 631–635.

Hardy, G.H., 1908. Mendelian proportions in a mixed population. Science 28, 49–50.

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 in Population Genetics, 2nd ed. Sinauer Associates Inc., Sunderland, MA.

Haubold, B., Travisano, M., Rainey, P.B., Hudson, R.R., 1998. Detecting linkage disequilibrium in bacterial populations. Genetics 150, 1341–1348.

Hedrick, P.W., 1999. Perspective: highly variable loci and their interpretation in evolution and conservation. Evolution 53, 313–318.

Hedrick, P.W., 2003. Hopi Indians, cultural selection, and albinism. Am. J. Phys. Anthropol. 121, 151–156.

Hedrick, P.W., 2005. A standardized genetic differentiation measure. Evolution 59, 1633–1638.

Hide, M., Bañuls, A.L., Tibayrenc, M., 2001. Genetic heterogeneity and phylogenetic status of *Leishmania* (*Leishmania*) *infantum* zymodeme MON-1: epidemiological implications. Parasitology 123, 425–432.

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

Hubbard, M.J., Cann, K.J., Baker, A.S., 1998. Lyme borreliosis: a tick-born spirochaetal disease. Rev. Med. Microbiol. 9, 99–107.

Hull, C.M., Raisner, R.M., Johnson, A.D., 2000. Evidence for mating of the asexual yeast *Candida albicans* in a mammalian host. Science 289, 307–310.

Hurtrez-Boussès, S., Durand, P., Jabbour-Zahab, R., Guégan, J.F., Meunier, C., Bargues, M.D., Mas-Coma, S., Renaud, F., 2004. Isolation and characterization of microsatellite markers in the liver fluke (*Fasciola hepatica*). Mol. Ecol. Notes 4, 689–690.

Hurtrez-Boussès, S., Meunier, C., Durand, P., Renaud, F., 2001. Dynamics of host–parasite interactions: the example of population biology of the liver fluke (*Fasciola hepatica*). Microb. Infect. 3, 841–849.

Kalinowski, S.T., 2002. Evolutionary and statistical properties of three genetic distances. Mol. Ecol. 11, 1263–1273.

Kimura, M., Ohta, T., 1978. Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc. Natl. Acad. Sci. U.S.A. 75, 2868–2872.

Koffi, B.B., De Meeûs, T., Barré, N., Durand, P., Arnathau, C., Chevillon, C., in press. Founder effects, inbreeding and effective sizes in the Southern cattle tick: the effect of transmission dynamics and implications for pest management, Mol. Ecol., in press.

Kunz, W., 2002. When is a parasite species a species? Trends Parasitol. 18, 121–124.

Leblois, R., Rousset, F., Estoup, A., 2004. Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. Genetics 166, 1081–1092.

Lehmann, T., Hawley, W.A., Kamau, L., Fontenille, D., Simard, F., Collins, F.H., 1996. Genetic differentiation of *Anopheles gambiae* populations from East and West Africa: comparison of microsatellites and allozyme loci. Heredity 77, 192–208.

Manly, F.J., 1985. The Statistics of Natural Selection. Chapman and Hall, New York.

Manel, S., Giaggioti, O.E., Waples, R.S., 2005. Assignment methods: matching biological questions with appropriate techniques. Trends Ecol. Evol. 20, 136–142.

Mantel, N., 1967. The detection of disease clustering and a generalised regression approach. Cancer Res. 27, 209–220.

Mas-Coma, S., Anglés, R., Esteban, J.G., Bargues, M.D., Buchon, P., Franken, M., Strauss, W., 1999. The Northern Bolivian Altiplano: a region highly endemic for human fasciolosis. Trop. Med. Int. Health 4, 454–467.

Maynard-Smith, J., Smith, N.H., 1998. Detecting recombination from gene trees. Mol. Biol. Evol. 15, 590–599.

McCoy, K.D., Boulinier, T., Tirard, C., Michalakis, Y., 2001. Host specificity of a generalist parasite: genetic evidence of sympatric host races in the seabird tick *Ixodes uriae*. J. Evol. Biol. 14, 395–405.

McCoy, K.D., Boulinier, T., Tirard, C., Michalakis, Y., 2003. Host-dependent genetic structure of parasite populations: differential dispersal of seabird tick host races. Evolution 57, 288–296.

McCoy, K.D., Boulinier, T., Tirard, C., 2005a. Comparative host–parasite population structures: disentangling prospecting and dispersal in the black-legged kittiwake *Rissa tridactyla*. Mol. Ecol. 14, 2825–2838.

McCoy, K.D., Chapuis, E., Tirard, C., Boulinier, T., Michalakis, Y., Le Bohec, C., Le Maho, Y., Gauthier-Clerc, M., 2005b. Recurrent evolution of host-specialized races in a globally distributed parasite. Proc. Roy. Soc. Lond. B 272, 2389–2395.

Metropolis, N., 1987. The beginning of the Monte Carlo method. Los Alamos Science 15, 125–130.

Meunier, C., Hurtrez-Boussés, S., Jabbour-Zahab, R., Durand, P., Rondelaud, D., Renaud, F., 2004a. Field and experimental evidence of preferential selfing in the freshwater mollusc *Lymnaea truncatula* (Gastropoda Pulmonata). Heredity 92, 316–322.

Meunier, C., Hurtrez-Boussés, S., Durand, P., Rondelaud, D., Renaud, F., 2004b. Small effective population sizes in a widespread selfing species, *Lymnaea truncatula* (Gastropoda: Pulmonata). Mol. Ecol. 13, 2535–2543.

Meunier, C., Tirard, C., Hurtrez-Boussès, S., Durand, P., Bargues, M.D., Mas-Coma, S., Pointier, J.P., Jourdane, J., Renaud, F., 2001. Lack of molluscan host diversity and the transmission of an emerging parasitic disease in Bolivia. Mol. Ecol. 10, 1333–1340.

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

Morel, C., 2000. Foreword. Acta Trop. 77, 1.

Morgan, A.D., Gandon, S., Buckling, A., 2005. The effect of migration on local adaptation in a coevolving host–parasite system. Nature 437, 253–256.

Nadler, S.A., 1995. Microevolution and the genetic structure of parasite populations. J. Parasitol. 81, 395–403.

Nakagawa, S., 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. Behav. Ecol. 15, 1044–1045.

Nébavi, F., Ayala, F.J., Renaud, F., Bertout, S., Eholié, S., Moussa, K., Mallié, M., De Meeûs, T., 2006. Clonal population structure and genetic diversity of *Candida albicans* in AIDS patients from Abidjan (Côte d'Ivoire). Proc. Natl. Acad. Sci. U.S.A. 103, 3663–3668.

Niklasson, M., Tomiuk, J., Parker Jr., E.D., 2004. Maintenance of clonal diversity in *Dipsa bifurcata* (Fallén 1810) (Diptera: Lonchopteridae). I. Fluctuating seasonal selection moulds long-term coexistence. Heredity 93, 62–71.

Njiokou, F., Nkinin, S.W., Grébaut, P., Penchenier, L., Barnabé, C., Tibayrenc, M., Herder, S., 2004. An isoenzyme survey of *Trypanosoma brucei* s.l. from the Central African subregion: population structure, taxonomic and epidemiological considerations. Parasitology 128, 645–653.

Paetkau, D., Strobeck, C., 1995. The molecular basis and evolutionary history of a microsatellite null allele in bears. Mol. Ecol. 4, 519–520.

Pasteur, N., Pasteur, G., Bonhomme, F., Catalan, J., Britton-Davidian, J., 1987. Manuel Technique de Génétique par Electrophorèse des Protéïnes. Lavoisier, Paris.

Pemberton, J.M., Slate, J., Bancroft, D.R., Barret, J.A., 1995. Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. Mol. Ecol. 4, 249–252.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure from multilocus genotype data. Genetics 155, 945–959.

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., De Meeûs, T., 2002. Inferring sex-biased dispersal from population genetic tools: a review. Heredity 88, 161–165.

Prugnolle, F., De Meeûs, T., Durand, P., Sire, C., Théron, A., 2002. Sex specific genetic structure in *Schistosoma mansoni*: evolutionary and epidemiological implications. Mol. Ecol. 11, 1231–1238.

Prugnolle, F., Durand, P., Théron, A., Chevillon, C., De Meeûs, T., 2003. Sex-specific genetic structure: new trends for dioecious parasites. Trends Parasitol. 19, 171–174.

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

Prugnolle, F., Roze, D., Théron, A., De Meeûs, T., 2005b. F-statisics under alternation of sexual and asexual reproduction: a model and data from schistosomes. Mol. Ecol. 14, 1355–1365.

Prugnolle, F., Théron, A., Pointier, J.P., Jabbour-Zahad, R., Jarne, P., Durand, P., De Meeûs, T., 2005c. Dispersal in a parasitic worm and its two hosts and its consequence for local adaptation. Evolution 59, 296–303.

Queller, D.C., Goodnight, K.F., 1989. Estimating relatedness using genetic markers. Evolution 43, 258–275.

Rannala, B., Mountain, J.L., 1997. Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. U.S.A. 94, 9197–9221.

Ravel, S., de Meeus, T., Dujardin, J.P., Zézé, D.G., Gooding, R.H., Dusfour, I., Sané, B., Cuny, G., Solano, P., in press. The tsetse fly *Glossina palpalis palpalis* is composed of several genetically differentiated small populations in the sleeping sickness focus of Bonon, Côte d'Ivoire. Infect. Genet. Evol., in press.

Raymond, M., Rousset, F., 1995a. An exact test for population differentiation. Evolution 49, 1280–1283.

Raymond, M., Rousset, F., 1995b. GENEPOP (Version 1.2): population genetics software for exact tests and ecumenicism. J. Hered. 86, 248–249.

Razakandrainibe, F.G., Durand, P., Koella, J.C., De Meeüs, T., Rousset, F., Ayala, F.J., Renaud, F., 2005. "Clonal" population structure of the malaria agent *Plasmodium falciparum* in high-infection regions. Proc. Natl. Acad. Sci. U.S.A. 102, 17388–17393.

Renaud, F., 1988. Biologie et Evolution des Populations d'Helminthes Parasites: Le Modèle Helminthes-Téléostéens. Thèse d'Etat, Université Montpellier II, Montpellier, France.

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

Ridley, M., 1996. Evolution, 2nd ed. Blackwell Science Inc., Cambridge, Massachusetts.

Riffaut, L., McCoy, K.D., Tirard, C., Friesen, V.L., Boulinier, T., 2005. Population genetics of the common guillemot *Uria aalge* in the North Atlantic: geographic impact of oil spills. Mar. Ecol. Prog. Ser. 291, 263–273.

Roberts, S.C., Little, A.C., Gosling, L.M., Perrett, D.I., Carter, V., Jones, B.C., Penton-Voak, I., Petrie, M., 2005. MHC-heterozygosity and human facial attractiveness. Evol. Hum. Behav. 26, 213–226.

Robertson, A., Hill, W.G., 1984. Deviations from Hardy–Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics 107, 713–718.

Roderick, G.K., 1996. Geographic structure of insect populations: gene flow, phylogeography, and their uses. Annu. Rev. Entomol. 41, 325–352.

Rousset, F., 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics 142, 1357–1362.

Rousset, F., 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145, 1219–1228.

Rousset, F., 2000. Genetic differentiation between individuals. J. Evol. Biol. 13, 58–62.

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

Rousset, F., Raymond, M., 1995. Testing heterozygote excess and deficiency. Genetics 140, 1413–1419.

Shaw, C.R., 1970. How many genes evolve? Biochem. Genet. 4, 275–283.

She, J.X., Autem, M., Kotoulas, G., Pasteur, N., Bonhomme, F., 1987. Multivariate analysis of genetic exchanges between *Solea aegyptiaca* and *Solea senegalensis* (Teleosts, Soleidae). Biol. J. Linn. Soc. 32, 357–371.

Shuker, D.M., Reece, S.E., Whitehorn, P.R., West, S.A., 2004. Sib-mating does not lead to facultative sex ratio adjustment in the parasitoid wasp, *Nasonia vitripenis*. Evol. Ecol. Res. 473–480.

Siegel, S., Castellan Jr., N.J., 1988. Nonparametric Statistics for the Behavioral Sciences, 2nd ed. McGraw-Hill, New York.

Škalamera, J.P., Renaud, F., Raymond, M., De Meeûs, T., 1999. No evidence for genetic differentiation of the mussel *Mytilus galloprovincialis* between lagoons and the seaside. Mar. Ecol. Prog. Ser. 178, 251–258.

Slatkin, M., 1985. Gene flow in natural populations. Ann. Rev. Ecol. Syst. 16, 393–430.

Slatkin, M., 1995. A measure of population subdivision based on microsatellite allele frequency. Genetics 139, 457–462.

Smith, K.L., Alberts, S.C., Bayes, M.K., Bruford, M.W., Altmann, J., Ober, C., 2000. Cross-species amplification, non-invasive genotyping, and non-mendelian inheritance of human STRPs in savannah baboons. Am. J. Primatol. 51, 219–227.

Schneider, P.M., Bendera, K., Mayr, W.R., Parson, W., Hoste, B., Decorte, R., Cordonnier, J., Vanek, D., Morlingh, N., Karjalaineni, M., Carlotti, C.M-P., Sabatier, M., Hohoff, C., Schmitter, H., Pflug, W., Wenzel, R., Patzelt, D., Lessig, R., Dobrowolski, P., O'Donnell, G., Garafano, L., Dobosz, M., de Knijff, P., Mevag, B., Pawlowski, R., Gusmão, L., Vide, M.C., Alonso, A.A., Fernández, O.G., Nicolás, P.S., Kihlgreen, A., Bär, W., Meier, V., Teyssier, A., Coquoz, R., Brandt, C., Germann, U., Gill, P., Hallett, J., Greenhalgh, M., 2004. STR analysis of artificially degraded DNA-results of a collaborative European exercise. Forensic Sci. Int. 139, 123–134.

Sokal, R.R., Rohlf, F.J., 1981. Biometry, 2nd ed. Freeman and Co., New York.

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.

Sunnucks, P., 2000. Efficient genetic markers for population biology. Trends Ecol. Evol. 15, 199–203.

Sunnucks, P., Wilson, A.C.C., Beheregaray, L.B., Zenger, K., French, J., Taylor, A.C., 2000. SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. Mol. Ecol. 9, 1699–1710.

Takezaki, N., Nei, M., 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144, 389–399.

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

Ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167–1179.

Ter Braak, C.J.F., 1987. CANOCO—A Fortran Program for Canonical Community Ordination. Microcomputer Power, Ithaca, NY, USA.

Ter Braak, C.J.F., Šmilauer, P., 2002. CANOCO Reference Manual and CanoDraw for Widows User's Guide: Software for Canonical Community Ordination (Version 4.5) Microcomputer Power, Ithaca, NY.

Thomas, F., Renaud, F., Derothe, J.M., Lambert, A., De Meeûs, T., Cezilly, F., 1995. Assortative pairing in *Gammarus insensibilis* (Amphipoda) infested by a trematode parasite. Oecologia 104, 259–264.

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.

Tomiuk, J., Guldbrandtsen, B., Loeschcke, V., 1998. Population differentiation through mutation and drift—a comparison of genetic identity measures. Genetica 102/103, 545–558.

Trouvé, S., Degen, L., Goudet, J., 2005. Ecological components and evolution of selfing in the freshwater snail *Galba truncatula*. J. Evol. Biol. 18, 358–370.

Van Oosterhout, C., Hutchinson, W.F., Wills, D.P.M., Shipley, P., 2004. Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. Mol. Ecol. Notes 4, 535–538.

Vignal, A., Milan, D., SanCristobal, M., Eggen, A., 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. 34, 275–305.

Vitalis, R., Couvet, D., 2001a. ESTIM 1.0: a computer program to infer population parameters from one- and two-locus gene identity probabilities. Mol. Ecol. Notes 1, 354–356.

Vitalis, R., Couvet, D., 2001b. Estimation of effective population size and migration rate from one- and two-locus identity measures. Genetics 157, 911–925.

Vitalis, R., Couvet, D., 2001c. Two-locus identity probabilities and identity disequilibrium in a partially selfing population. Genet. Res. 77, 67–81.

Wahlund, S., 1928. Zusammensetzung von Populationen und Korrelationserschinungen von Standpunkt der Vererbungslehre aus betrachtet. Hereditas 11, 65–108.

Wang, J., Whitlock, M.C., 2003. Estimating effective population size and migration rates from genetic samples over space and time. Genetics 163, 429–446.

Waples, R., 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics 121, 379–391.

Waser, P., Strobeck, C., 1998. Genetic signatures of interpopulation dispersal. Trends Ecol. Evol. 13, 43–44.

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.

Weinberg, W. 1908. Über den Nachweis der Verebung beim Menschen. Jahresh. Verein f. Vaterl. Naturk in Wüttemberg 64, 368–382.

Weir, B.S., 1979. Inferences about linkage disequilibrium. Biometrics 35, 235–254.

Weir, B.S., 1996. Genetic Data Analysis. Sinauer Associates Inc., Sunderland, MA.

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

Whitlock, M.C., McCauley, D.E., 1998. Indirect measures of gene flow and migration: $FST \neq 1/(4N_{\mathrm{m}}+1)$. Heredity 82, 117–125.

Wolff, K.E., 1996. Comparison of graphical data analysis methods. In: Faulbaum, F., Bandilla, W. (Eds.), SoftStat '95 Advances in Statistical Software 5. Lucius & Lucius, Stuttgart, pp. 139–151.

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.

Xu, J., 2005. The inheritance of organelle genes and genomes: patterns and mechanisms. Genome 48, 951–958.

Glossary

Allele: Hereditary state at which a locus can be. In diploids each individual has two allele at each locus, that may be identical (homozygous) or different (heterozygous).

Assortative mating: A process that makes sexual partners to mate if phenotypically similar.

Autosome: Designs an ordinary chromosome expectedly present in pairs in each normal zygote or diploid individual (antonymous to heterosome).

Bottleneck: A demographic process where a population experiences a strong drop in individual number.

Clonality: Reproduction with no sex. The descent is identical to the parental individual.

Codominant: Describes a genetic marker where all heterozygotes are recognisable from all homozygotes.

Dioecious: Synonymous to gonochoric, it means that the species are subdivided into males and females (antonymous to monoecious).

Diploid: Characterizes an organism or a cell with a double set of nuclear genetic material (chromosomes), at the exception of sex chromosomes when there are any.

Directional selection: A selective process that tend to increase or decrease (one direction) the frequency of an allele in the population.

Dominant: Describes a genetic marker where an allele covers the expression of other alleles in heterozygous individuals. Applies also to alleles (antonymous to recessive).

Drift (random genetic): Describes the process by which allelic frequencies change from one generation to the other as a result of the random sampling of individuals (zygotes or gametes that form zygotes or adults) that survive to form the next generation in a population of finite size.

Gene: A portion of coding DNA, i.e., that is transcribed into a mRNA.

Genotype: The complete set of alleles displayed by an individual at a specific locus or a specific set of loci (when specified).

Haploid: Characterizes an organism or a cell with a simple set of nuclear genetic material (chromosomes). Gametes are typically haploid.

Heterogamy: A process where sexual partners or gametes are more likely attracted by genetically different individuals (antonymous to homogamy).

Heterosis: A genome wide selective phenomenon where some kind of advantage characterises the most heterozygous individuals.

Heterosome: Synonymous to sex chromosome. In dioecious species, a kind of chromosome the composition of which differs between male and females (e.g., XY chromosomes in mammals, ZW chromosomes in birds) (antonymous to autosome).

Heterozygous: Refers to a locus in a diploid individual for which the two alleles are different (antonymous to homozygous).

Homogamy: A process where sexual partners or gametes are more likely attracted by genetically similar individuals (antonymous to heterogamy, see also assortative mating).

Homoplasy: A phenomenon describing the identity between to alleles that do not share a common ancestry, which are then said identical by state.

Homozygous: Refers to a locus in a diploid individual for which the two alleles are identical (antonymous to heterozygous).

Infinite Allele Model (IAM): A mutation model where each mutation produces a new allele that did not exist in the population and will not be recovered if lost. Does not allow homoplasy.

Inbreeding: Characterising the proportion of identical by descent alleles within individuals as a result of closed systems of mating (selfing, sib-mating) or the limited population sizes. Note that when only due to population size, the inbreeding coefficient (probability of identity by descent in individuals) is identical to the relatedness between individuals (probability of identity by descent of genes between individuals).

Infinite island model: An island model where the number of sub-populations (n) is infinite.

Island model: A theoretical subdivided population, with non-overlapping generations, where individuals are arranged into n sub-populations (islands) of identical size N composed, at each generation, of mN migrant individuals that may come from any of the n sub-populations and $(1 - m)N$ resident individuals.

K allele model (KAM): A mutation model where each mutation changes the affected allele into any of the K possible ones, including itself, with an equal probability. The lower K the more frequent homoplasy is.

Linkage disequilibrium: A characteristic expressing the non-random association between different loci (generally by pair). Many different factors (population structure, closed system of mating, selection, etc.) can generate and maintain statistical associations between loci.

Locus: Describes a portion of DNA at a specific position in the genome. It does not necessarily correspond to a gene.

Mutation: Occurs when a mistake is made during DNA duplication.

Monoecious: Synonymous to hermaphrodite (antonymous to dioecious).

Neighbourhood model: A theoretical population where migration of each individual is limited by distance. Thus, the relatedness between individuals is a decreasing function of the distance separating them even if no subdivision exists.

Neutral: Applies to a locus the polymorphism of which is not under any kind of selective pressure (antonymous to selected).

Overdominance: A selective process where the survival and/or fertility of individuals is enhanced if heterozygous at a given locus.

Pangamy: Describes a sexually reproducing population where all individuals randomly pair for copulation.

Panmixia: Describes a sexual kind of reproduction where zygotes are produced by a random association of any pair of gametes from the population.

Phenotype: The expression of a character that may be hereditary (e.g., the colour of the eyes). A phenotype can be translated into a genotype (e.g., for isoenzymatic loci).

Polymorphism: Condition describing genetic variation (more than one allele) at a locus in a data set.

Population: A set of individuals sharing the same demographic parameters (population regulation) and more likely sharing a common ancestry as compared to members of other such populations, except for migrants.

Recessive: Describes an allele that is hidden when at a heterozygous state (antonymous to dominant).

Selection: The process by which the expected survival and/or fertility of an individual depends on its genotype in a more or less direct way.

Selected: Applies to a locus the polymorphism of which is submitted to selective pressures (antonymous to neutral).

Selfing: A sexual reproductive mode where a functional hermaphrodite individual self fertilize its proper eggs with its proper spermatozoids.

Sex ratio: The ratio of the number of males to the number of females in a population. Is equal to one when balanced.

Stepwise mutation model (SMM): A mutation process where each mutation process increases or decreases the size of the affected allele by one unit (step) with an equal probability. With such a mutation process, homoplasy is frequent and a similarity in size can be translated into a probable recent co-ancestry between the alleles compared.

Stepping-stone model: A theoretical subdivided population where migrants are only exchanged between neighbouring sub-populations.

Transition: A point mutation that changes a purine into the other purine (A ↔ G) or a pyrimidine into the other pyrimidine (C ↔ T) (antonymous to transversion).

Transversion: A point mutation that changes a purine into a pyrimidine or a pyrimidine into a purine (A ↔ T, A ↔ C, G ↔> C, G ↔ T) (antonymous to transition).

Underdominance: A selective process where the survival and/or fertility of individuals is decreased if heterozygous at a given locus.

Zygote: The result of the fusion of two gametes. The term egg is sometimes used instead.