Statistical inference in population genetics using microsatellites

Katalin Csilléry

Ph.D.
The University of Edinburgh
2009

To my parents, and to the memory of my grandfather, Japi

Declaration

I composed this thesis, the work is my own.

No part of this thesis has been submitted for any other degree or qualification.

Katalin Csilléry

January 27, 2009.

Acknowledgments

I worked on this PhD in three different countries, and at four different locations: a rather unusual schedule, which would have been more difficult to carry though without the help and encouragement of many. This is the place to thank them all.

I thank my supervisor at the Institute of Evolutionary Biology (IEB, Edinburgh), Josephine Pemberton, for managing my PhD work, giving me advise and feedback always rapidly. I also thank her for taking me to St Kilda, which was not only a very unique experience, but also it allowed me to see the hard work behind the data collection.

I thank Toby Johnson, who was my supervisor in Edinburgh, in email, and then in Lausanne. TJ, it was not easy to have you as a supervisor, but I admit I was not an easy student either! We had some very good and less good phases, but I owe you a lot: when I was talking to fellow students and colleagues, when I had to present at conferences and when I had my job interviews, these were the times when I came to realize again and again how much I learned from you.

Funding for this PhD was provided by the Principal's Studentship from the School of Biological Sciences (University of Edinburgh) and also benefited from a travel grant from the James Rennie Bequest.

Parts of this thesis are based on data that was generated by others. I am grateful to have had access to excellent data sets from many long term projects and I appreciate the work of many, who organized and raised funds over many years to keep these valuable projects going. I especially thank Josephine Pemberton, Bengt Hansson, Dennis Hasselquist, Staffan Bensch, Tim Clutton-Brock, Marco Festa-Bianchet, and David Coltman. I also thank the numerous field workers and genotypers.

The ideas presented in this thesis have been improved by comments and discussions with many senior colleagues. I thank Nick Barton, Kevin Davson, Arnaud Estoup, Bill Hill, François Rousset, Jon Slate, Alastair Wilson and, Penny Kukuk and Allen Moore; to the latter two for also encouraging me to start a PhD.

I spent six month as a visiting PhD student at the Centre de Biologie et de Gestion des Populations (CBGP, INRA, France). I thank my host, Arnaud Estoup, for his help and support with my project, the results of which are presented in this thesis. I also thank CBGP for hosting me and letting me use their computing cluster and Filipe Santos for

computing support.

I am grateful to have had such good colleagues at work. I thank to my fellow PhD students and post-docs for many good work discussions, particularly “the boudoir” members: Dario, Sylvia, Will, Roberta, Allan, Markus, and Helen at IEB, and Stuart and Filipe at the CBGP.

I did a short (five month) “post-doc” during my PhD at the Institut Universitaire de Médecine Sociale et Préventive (IUMSP, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland). Although work at the canton hospital was independent from my PhD, I still want to thank here my boss Murielle Bochud, especially for her understanding that I had to quit in order to finish my PhD.

I worked on my PhD in Lausanne while visiting Toby Johnson at the Département de Génétique Médicale (DGM, Université de Lausanne, Switzerland). I thank DGM, and especially Sven Bergman, for giving me desk space and very a friendly and accommodating work environment.

I would also like to thank here Zoltán Barta and Szabolcs Lengyel, my MSc supervisors at the University of Debrecen (Hungary), and Jon Graham, my instructor in statistics at the University of Montana (USA), who were the first to open my eyes to statistics, and generally, had a great influence on my thinking.

Climbing was the most serious distraction from work, nevertheless, it gave me the best moments during my PhD years. So, I hope it is appropriate to thank here some of my closest climbing partners: Barney, Phil, Majka, Uwe, Anita (Edinburgh); Julien and Jerôme, and all the others of the CAF (Montpellier), Ben (Genève), Sven et al. (Lausanne), and Anne-Claire (Chamonix). I also happened to write up my PhD in Chamonix, so I thank the mountains of Chamonix valley for being silent witnesses of my work here.

Life would have been much more difficult without the help and support of friends. I thank Emily and Emma (Edinburgh), Kriszta (Villefontaine), the family of Vera, Lél and Hanga (Edinburgh) and the family of Rita, Martim and Francisca (Montpellier): they became my second and third families, and their homes my (sometimes only) homes.

Finally, I thank my family in Hungary: my sisters, Otti and Julka, who I could always call when I needed encouragement, my Dad, who, I know, was always very proud of me and thinking of me a lot, even if he did not tell me this, and finally, but foremost I thank you, Kismama. Thank you for being alive! I still need you.

January 2009, Chamonix

Abstract

Statistical inference from molecular population genetic data is currently a very active area of research for two main reasons. First, in the past two decades an enormous amount of molecular genetic data have been produced and the amount of data is expected to grow even more in the future. Second, drawing inferences about complex population genetics problems, for example understanding the demographic and genetic factors that shaped modern populations, poses a serious statistical challenge.

Amongst the many different kinds of genetic data that have appeared in the past two decades, the highly polymorphic microsatellites have played an important role. Microsatellites revolutionized the population genetics of natural populations, and were the initial tool for linkage mapping in humans and other model organisms. Despite their important role, and extensive use, the evolutionary dynamics of microsatellites are still not fully understood, and their statistical methods are often underdeveloped and do not adequately model microsatellite evolution. In this thesis, I address some aspects of this problem by assessing the performance of existing statistical tools, and developing some new ones. My work encompasses a range of statistical methods from simple hypothesis testing to more recent, complex computational statistical tools. This thesis consists of four main topics.

First, I review the statistical methods that have been developed for microsatellites in population genetics applications. I review the different models of the microsatellite mutation process, and ask which models are the most supported by data, and how models were incorporated into statistical methods. I also present estimates of mutation parameters for several species based on published data.

Second, I evaluate the performance of estimators of genetic relatedness using real data from five vertebrate populations. I demonstrate that the overall performance of marker-based pairwise relatedness estimators mainly depends on the population relatedness composition and may only be improved by the marker data quality within the limits of the population relatedness composition.

Third, I investigate the different null hypotheses that may be used to test for independence between loci. Using simulations I show that testing for statistical independence (i.e. zero linkage disequilibrium, LD) is difficult to interpret in most cases, and instead a null hypothesis should be tested, which accounts for the "background LD" due to finite population size. I investigate the utility of a novel approximate testing procedure to circumvent this problem, and illustrate its use on a real data set from red deer.

Fourth, I explore the utility of Approximate Bayesian Computation, inference based on summary statistics, to estimate demographic parameters from admixed populations. Assuming a simple demographic model, I show that the choice of

summary statistics greatly influences the quality of the estimation, and that different parameters are better estimated with different summary statistics. Most importantly, I show how the estimation of most admixture parameters can be considerably improved via the use of linkage disequilibrium statistics from microsatellite data.

Contents

Chapter 1

Statistical inference in population genetics using microsatellites

1.1 Introduction

Microsatellites have been intensively used over the past two decades in population genetics applications. Since their discovery in the early eighties (SPRITZ, 1981), it was rapidly recognized that the codominant microsatellites were potentially more useful than previously used genetic markers, such as allozymes (ESTOUP *et al.*, 1998). Microsatellites are perfect or near perfect tandem iterations of short sequence motifs, and extremely abundant in all eukaryotic genomes, but with widely varying levels of polymorphism and allele length (ELLEGREN, 2004). Most microsatellites are non-coding DNA, either in intergenic sequence regions or in introns. Thus, they can generally be assumed to evolve neutrally, so their level of polymorphism is proportional to the underlying mutation rate. Genotyping large samples for microsatellites became fast and cost effective near the end of the eighties with the advent of PCR technology (LITT and LUTY, 1989; WEBER and MAY, 1989; TAUTZ, 1989). Since then, microsatellites have become the marker of choice in population genetic studies of natural populations and they have had a great impact on human genetics as well, for example, the first detailed linkage map of the human genome was created using microsatellites (WEISSENBACH *et al.*, 1992).

Starting from the availability of the first large pedigrees with microsatellite data (WEBER and WONG, 1993) to the more recent microsatellite sequence comparisons between related species (SAINUDIIN *et al.*, 2004), a lot has been revealed about the evolution of these ubiquitous sequences. However, their evolutionary dynamics is still not fully understood, and it seems that every new study discovers further complexities

regarding their evolution (*e.g.* CORNUET *et al.*, 2006; SEYFERT *et al.*, 2008). From an inference point of view, the most relevant question is the sensitivity of our conclusions to the assumed model of microsatellite evolution. Although considerable effort has been put into the development of biologically realistic models of microsatellite evolution, only a few of these models have been incorporated into statistical methods, and often with a delay, especially in the past decade. Even more importantly, very little sensitivity analysis has been carried out to evaluate to what extent the final biological conclusions are sensitive to model assumptions. For example, even though many studies have suggested that the simple stepwise mutation model (SMM) cannot sufficiently explain the observed data patterns (*e.g.* WHITTAKER *et al.*, 2003), most recently published methods still use the SMM because of its simplicity. The question of whether the SMM is a sufficiently good "working" model or not, unfortunately, in most cases remains unclear.

The aim of this Chapter is to provide a general overview of the statistical methods that have been developed for microsatellites from the late eighties to today. I will distinguish between two phases of method development. The first phase started around the late eighties when microsatellites started becoming the most popular marker for natural populations and in human genetics. At the time, many classical population genetic methods were re-visited and some new methods were developed, taking into account the fact that microsatellites evolve in a stepwise fashion. The second phase started around the mid-nineties with the availability of fast computers, which shifted the focus of method development from marker type *per-se* to using computationally intense methods. I will start my review with summarizing our up-to-date knowledge on the evolution of microsatellites with a focus on the mutation models. I do so because, first, it helps understanding the motivation of the early method development (first phase), and second, it points to possible further improvements to the modern, computationally intense methods (second phase). To close, I will present some original data. I will contrast estimates of the mutation rates for different published data sets which represent a wide range of species. Then, I fit a popular version of the stepwise mutation model to the data, which could motivate method improvement and also give guidance for choosing model parameters when using microsatellite data with some of the modern computational tools.

1.2 The evolution of microsatellites

1.2.1 Mutational mechanism

Two kinds of mutational events have been proposed to occur in microsatellites: length and point mutations. Length mutations, i.e. copy (or repeat) number mutations, that lead to new allelic variants are very common and have been studied in detail. The strand slippage model was first proposed by OKADA *et al.* (1966) for the mechanism of copy number mutations. The idea is that template-primer slippage within microsatellites leads to misaligned intermediates, whose number and stability increases with increasing repeat unit number. In the light of this mechanism it is also easy to interpret why longer microsatellites tend to have higher mutation rates, as has been observed in many empirical systems (*e.g.* BROHEDE *et al.*, 2002).

Much less is known about point mutations, mainly because they are not as easy to observe as length mutations due to the lack of multi-generation sequence data. However, they might play an important role in the evolutionary dynamics of microsatellites, as for example, it has been suggested for yeast (*e.g.* KRUGLYAK *et al.*, 2000). Recent cross-species comparisons have estimated the rate of point mutations (*e.g.* SAINUDIIN *et al.*, 2004), and the ratio of the two mutation rates (PUMPERNIK *et al.*, 2008). Analyzing human-chimpanzee sequence alignments PUMPERNIK *et al.* (2008) found that replication slippage mutations outnumber point mutations by one to two orders of magnitude, but also point mutations occur about twice as frequently as expected.

1.2.2 Mutation models

Mutation models can be used to describe the evolutionary mechanism of microsatellite alleles. The infinite alleles model (IAM) (KIMURA and CROW, 1964) and the K-allele model (KAM) (CROW and KIMURA, 1970) are two classical models of mutation. Under the IAM, mutations always generate an allele that is new to the population, thus there is an infinite number of possible allelic states. Under KAM there are K distinct alleles in the population and probability of mutating from any one of them to any other of them is equal, thus the mutational history is always erased. Even though these models miss many of the key features of the evolution of microsatellites they had an impact on the development of some of the early statistical methods in population genetics that are used for microsatellites. For example, WRIGHT's (1931) F-statistics , which are some of the most frequently calculated statistics for microsatellites, assume the IAM.

The stepwise mutation model (SMM), which is the most commonly used mutation model for microsatellites, was originally proposed by OHTA and KIMURA (1973) and then re-discovered for microsatellites by VALDES *et al.* (1993). The SMM assumes a symmetric forwards and backwards random walk, where a new mutation leads to a new allele, which differs from the parental allele by one repeat unit. Thus, the mutational history is conserved for one step. The SMM is an attractive model for microsatellites because of it simplicity, but it has many shortcomings. First of all, it does not have a stationary distribution of allele lengths and, thus there is nothing to stop microsatellites to grow infinitely large or shrink to zero (DIRIENZO *et al.*, 1994). Secondly, there is abundant of evidence that mutations of more then one repeat unit commonly occur (*e.g.* HUANG *et al.*, 2002), and for some species and some loci there are other significant mutational biases. Thus, it became clear that the SMM, as a one parameter mutation model, cannot explain all aspects of the evolution of microsatellites and the observed data patterns.

Two main model classes have been proposed to explain the theoretical problem of how microsatellites can be maintained in the genome without expanding to infinite size. First, the idea of long alleles breaking down to short alleles was proposed (BELL and JURKA, 1997; DIRIENZO *et al.*, 1994), which was later developed into the so-called two-phase model (TPM) (KRUGLYAK *et al.*, 1998). The TPM assumes the presence of two kinds of mutations: first, strand slippage mutations, which add or delete one repeat unit and depend on allele length, and, second, point mutations that interrupt the microsatellite repeats, thus generating two short alleles from one long one. (Note that KRUGLYAK *et al.* (1998) kept track of only one of the daughter alleles.) Thus, alleles could slowly "grow" via small step length mutations and long alleles would break down via point mutations. Under the TPM there exists an equilibrium distribution of repeat lengths via the balance between the two mutational mechanisms (KRUGLYAK *et al.*, 1998).

The second class of models is centered around the idea of applying some sort of constraint on allele length. The simplest idea is to apply an upper bound on allele length, which solves the problem of unlimited growth. GARZA *et al.* (1995) proposed a more elaborate model with a linear mutational bias towards a focal length. Under this model microsatellites below the focal length tend to expand, and those above it tend to contract. Even though the biological reality of such a focal length is unclear, it has been incorporated in many statistical models (*e.g.* FELDMAN *et al.*, 1997).

Other mutational biases that have been observed in empirical data also motivated the developments of refinements to the SMM. For example, to account for mutations involving changes of more than one repeat unit, KIMMEL and CHAKRABORTY (1996)

suggested the generalized stepwise model (GSM). The GSM has two parameters, the mutation rate and the number of repeats added or removed in one mutation event, which follows a Geometric distribution with mean p. The GSM is important because it has been used in many studies and incorporated into software packages (*e.g.* LAVAL and EXCOFFIER, 2004; CORNUET *et al.*, 2008). Length dependent mutational bias has also been incorporated in many models, using different functions of allele length to define the mutation rate. For example, in the TPM, mutations of alleles consisting of k repeat units occur at rate $(k-1)b$, where b is the per repeat unit slippage rate. WHITTAKER *et al.* (2003) proposed a class of nested models, where the most complex model include parameters controlling the rate, step size, dependence on parental allele size and direction of mutations.

1.2.3 Evidence from empirical data

A number of different approaches have been used to study the mutational patterns at microsatellite loci. The first large scale studies to gain reliable estimates of the mutation rates have been direct observations of allele transmissions in parent-offspring pairs. However, even these estimates of the mutation rates can be biased, because mutations can only be identified when the offspring genotype could not be generated by transmission from its parents' genotypes. As a result, the probability that a mutation is detected depends on allele lengths (WHITTAKER *et al.*, 2003). Another approach to gain estimates of microsatellite mutation rates is to count the mutational events in highly inbred, so-called mutation-accumulation (MA) lines, over many generations. The advantage of MA lines is that they do not rely on the assaying of many parent-offspring pairs. However, they are only feasible for model organisms. Finally, comparative genomics could also provide useful insight into the evolution of microsatellites. Most notably, comparative studies are the only ones so far in which point mutations can also be detected. Thus, they are complementary to the pedigree and MA studies.

What are the general patterns emerging from these studies so far? Which is the best supported mutation model? The two main competing mutation model classes are the TPM versus models with constraint on allele length. Mutational patterns from large human and avian pedigrees, and also from *Drosophila* MA lines, support the idea of length-dependent mutation bias (*e.g.* WHITTAKER *et al.*, 2003; SCHLÖTTERER *et al.*, 1998; ELLEGREN, 2000a; HUANG *et al.*, 2002) and evidence from yeast, *D. melanogaster*, and humans suggests that long microsatellite alleles have more contractions (ELLEGREN, 2000a; HARR and SCHLÖTTERER, 2000; XU *et al.*, 2000; HARR *et al.*, 2002). For example, WHITTAKER *et al.* (2003) showed that contractions

are more likely for microsatellites larger than 20 repeats while expansions are frequent for shorter "large" alleles. Thus, the model with constraint on allele length seems to have overwhelming support.

Species comparisons reveal that the length at which there is a significant over-representation of "long" microsatellites is dependent on repeat type and species (DIERINGER and SCHLÖTTERER, 2003), suggesting that the constraint is species dependent. A recent paper provides more insight into these findings: AMOS and CLARKE (2008) found that in mammals, the maximum repeat number is inversely correlated with body temperature, with warmer-blooded species having shorter "long" microsatellites (AMOS and CLARKE, 2008). The authors suggest a mechanism for the allele length constraint, namely that maximum length is limited by a temperature-dependent stability threshold.

One potential problem with these conclusions is that we cannot exclude the possibility that point mutations also play a role in the evolutionary dynamics of microsatellites. However, while length mutations are easy to observe via genotyping a large number of individuals, point mutations are not. Here is where sequence comparisons between species could play an important role. Using data from homologous microsatellite loci in humans and chimpanzees SAINUDIIN *et al.* (2004) compared several microsatellite mutation models. In agreement with the pedigree studies, the authors found that GARZA *et al.*'s (1995) model with a linear bias toward a focal length has the most support, and also argued that taking length dependent mutational bias into account is essential for realistic models of evolution of pure dinucleotide repeats.

Finally, I mention some further mutational biases that could be important for applications in statistical models. For example, the mutation rate has been shown to depend on specific properties of the microsatellite in question, including repeat type (mono, di, tetra etc) (CHAKRABORTY *et al.*, 1997), nucleotide composition, locus (*e.g.* DIB *et al.*, 1996; ELLEGREN, 2004; CORNUET *et al.*, 2006), species (*e.g.* AMOS and CLARKE, 2008), and sex (*e.g.* BROHEDE *et al.*, 2002). How important it is to account for these effects? Since the data that we encounter is so heterogeneous, the answer to this question often depends on the data. For example, CORNUET *et al.* (2006) fitted three mutation models (SMM, GSM, and a model where the mutation rate grows exponentially with the number of repeats) to data from a parasitic mite, and found heterogeneity across loci regarding which model had the most support. Generally, the GSM received the most support, but actually there was only one locus out of 19 for which SMM could be rejected. In another study, WHITTAKER *et al.* (2003) emphasized the importance of length dependent models, which provided a significantly

better fit to a large human data set of AC repeats. In conclusion, more studies, where alternative mutation models are fitted to real data would be desirable in order to draw a more complete picture of the microsatellite mutation process.

1.3 Statistical methods for microsatellites

1.3.1 Classical population genetics

When microsatellites were becoming commonly used in population genetics at the beginning of the nineties, many of the classical population genetics problems were re-visited. Many classic population genetics statistics, which are justified under the IAM or KAM, such as F-statistics (WEIR and COCKERHAM, 1984) or Nei's genetic distance measure (NEI and ROYCHOUDHURY, 1974), seemed to be inadequate for microsatellites. This is because microsatellites have a much higher mutation rate than, for example, the previously used allozymes, for which KAM was a reasonable approximation. Also, since microsatellites evolve in a stepwise fashion, they did not conform with the "memoryless" property of IAM and KAM, where the state of the mutant allele is independent of the state of the parental allele. At microsatellite loci, alleles that have the same allele size (i.e. identical by state or IBS) are not necessarily inherited from the same ancestor (i.e. not necessarily identical by descent or IBD). This feature of microsatellites, which is called allele size homoplasy (ESTOUP *et al.*, 2002), has to be taken into account when making inferences from microsatellite data. Also, microsatellites are multiallelic, and much of the classical population genetics theory and statistical approaches were developed for biallelic markers. Although the extension to multiallelic loci is in many cases conceptually simple, the properties of the new measures are not instantly obvious. Further, the number of alleles, and hence (as neutral markers) the mutation rates may well vary between microsatellite loci, thus weighting loci might be advantageous. Here, I discuss the methodological developments by subject area, which is a somewhat arbitrary, albeit hopefully useful, grouping since there is considerable overlap between areas.

The methods discussed in this section are mainly simple tests and estimators that dominated the population genetics literature at the time. Most developments for microsatellite markers involved modifications of existing measures to incorporate the stepwise mutation model, which meant, in practice, the use of allele size information in some form. These developments were influenced by the coalescent theory (KINGMAN, 1982a), which is a stochastic model for the genealogical tree of the ancestral relationships of a sample of DNA sequences. The most general framework

for microsatellites was described by PRITCHARD and FELDMAN (1996), who extended the coalescent framework for the SMM, though their work was preceded by other similar studies (*e.g.* GARZA *et al.*, 1995; SLATKIN, 1995; GOLDSTEIN *et al.*, 1995a). PRITCHARD and FELDMAN (1996) studied the properties of pairwise differences in repeat numbers between randomly chosen microsatellite alleles, and how it relates to coalescent times between genes, which is the fundamental idea behind many of the developments.

1.3.1.1 Tests of disequilibrium

Both testing for Hardy-Weinberg and for linkage equilibrium could be easily extended to multiallelic markers (WEIR, 1996). The main difference from the biallelic case is that we are testing for independence in a larger contingency table, whose size depends on the number of alleles. MAISTE and WEIR (2004) investigated testing for Hardy-Weinberg equilibrium in large contingency tables, and found that with multiallelic markers the power of the Monte Carlo permutation test they used increased with the number of alleles. Multiallelic markers may also contain more information about linkage disequilibrium (LD) than biallelic markers (ZHAO *et al.*, 1999), and thus may have a higher power to detect LD (SLATKIN, 1994). However, as I will argue in Chapter 3, using more polymorphic loci does not lead to a true power gain in the biological sense. This is because testing for statistical independence (zero LD) in a large contingency table does not correspond to any biologically meaningful null hypothesis. As a result, a power comparison that uses the Monte Carlo permutation null cannot be used to conclude that there is more power to detect LD with more polymorphic loci as was shown in ZHAO *et al.* (1999).

1.3.1.2 Relatedness and inbreeding

The development of various relatedness estimators was clearly motivated by the availability of highly polymorphic markers (*e.g.* RITLAND, 1996a; LYNCH and RITLAND, 1999; WANG, 2002; MILLIGAN, 2003). On the short time-scale of familial relationships, mutations can be ignored, thus the assumption of a mutation model does not pose a problem here. However, in order to evaluate the performance of estimators, generally, data sets are simulated with known relatedness, which implies an assumption about the allele frequencies. The commonly used allele frequency distributions ranged from uniform to allele frequencies estimated from real population data (WANG, 2002; VAN DE CASTEELE *et al.*, 2001; MILLIGAN, 2003; CSILLÉRY *et al.*, 2006). Some authors have argued that the Dirichlet distribution is the most realistic choice (*e.g.* WANG, 2002), since this is the distribution of allele frequencies at equilibrium under

the joint effects of drift and mutation or migration (WRIGHT, 1951). The Dirichlet approximation is used in many forensic applications as well to account for coancestry in assessing the forensic evidence from microsatellite data (*e.g.* GRAHAM *et al.*, 2000). However, the Dirichlet distribution may not be adequate for microsatellites because it does not take into account the positive correlation between frequencies of distinct alleles of similar length that arise due to the stepwise mutation process. As a result, for example, it was shown in forensic applications that match probabilities based on the Dirichlet model tend to overstate the evidence against the suspect (GRAHAM *et al.*, 2000).

Inbreeding depression has traditionally been studied using pedigree based estimates of the inbreeding coefficient (f). However, the availability of polymorphic markers, along with the lack of pedigree information in natural populations (*e.g.* MARSHALL *et al.*, 2002; PEMBERTON, 2008), motivated the use of various marker-based proxies (*e.g.* BENSCH *et al.*, 1994; COULSON *et al.*, 1998). For example, the squared difference in repeat units between two alleles at a locus, averaged over all loci, d^2, was suggested as a marker based proxy for f (COULSON *et al.*, 1998), which, under the SMM, is a linear function of time since coalescence between the two alleles (GOLDSTEIN *et al.*, 1995a).

1.3.1.3 Population structure and history

The first and still most widely used statistic to summarize genetic variability between different populations is the inbreeding coefficient, F_{ST}, which was originally developed by WRIGHT (1951). Without aiming for completeness, here I will describe various modifications of F_{ST} and related statistics that were specifically developed to take the microsatellite mutation process into account.

It has been recognized that the assumptions of IAM or KAM of the classic F_{ST} (WRIGHT, 1951) statistic does not conform to the microsatellite mutation process. SLATKIN (1995) argued that irrespective of the details of the mutation model, for microsatellites, there is clearly some memory to the mutational process, thus F_{ST} will yield biased estimates of demographic parameters. In particular, F_{ST} tends to underestimate population differentiation. SLATKIN (1995) developed the R_{ST} statistic, which is analogous to F_{ST} but assumes a stepwise mutation process. R_{ST} is defined as the fraction of the total variance of allele size that is between populations, thus it is similar to WEIR and COCKERHAM's (1984) θ, which is also a between population component of variance. The difference is that R_{ST} takes allele sizes into account, while in θ, only identity or non-identity of allelic states enter. SLATKIN (1995) found that R_{ST} generally performed better than F_{ST}, and especially when mutations are more

important than drift. F_{ST} is unbiased under the SMM type mutation models only when a very short time scale is considered, thus mutations can be ignored. ROUSSET (1996) further examined the effect of mutation models and rates on F-statistics, and derived expressions for IBS under different mutation models in an island model.

Even though R_{ST} generally outperforms F_{ST} in large simulation studies, the latter may still be preferred for most real data applications. This is because a large number of independent loci are needed for microsatellites to attain low variance estimates of R_{ST}. Therefore, if the number of individuals and loci are moderate or small the overall performance of F_{ST} is better, because although it is biased, it has lower sampling variance (BALLOUX and GOUDET, 2002; GAGGIOTTI *et al.*, 1999).

Genetic distance measures were also modified for microsatellites. Nei's classic genetic distance (NEI and ROYCHOUDHURY, 1974) is derived by assuming the IAM, and thus it does not capture information about the time since the common ancestor that is carried in the difference in repeat numbers. GOLDSTEIN *et al.* (1995a) suggested a new statistic to measure genetic distance, which is based on allele size differences between populations and is also independent of the population size. The new distance measure was found to be uniformly superior to Nei's measure and to a simple allele sharing measure (GOLDSTEIN *et al.*, 1995b).

Apart from the modifications of various forms of F statistics, the availability of microsatellites has also inspired the development of new statistics. For example, GARZA and WILLIAMSON (2001) developed the M statistic, which is the ratio of the number of alleles and the range of allele sizes averaged over all loci. GARZA and WILLIAMSON (2001) argued that M can be used to detect reductions in population size. They indeed found that the M statistic predicts the reported demographic history of various natural populations. They use the TPM in their analysis and also discuss how the mutation process affects the analysis, and find that single loci estimates are very sensitive to extreme allele frequencies (GARZA and WILLIAMSON, 2001).

1.3.1.4 Selection

Detecting the signature of selection in natural populations has long been of interest. The idea behind tests for selection is that a selective process would only affect certain regions of the genome, while demographic processes would affect the whole genome uniformly; a simple idea, which was originally proposed by CAVALLI-SFORZA (1966). Later, a formal test of neutrality was proposed by LEWONTIN and KRAKAUER (1973) based on the comparison of the variance of F statistics across populations, which was criticized because it did not take account of patterns of migration or population history. The LEWONTIN and KRAKAUER (1973) type tests were then re-discovered

(BEAUMONT, 2005), however, they seem particularly ill-suited for microsatellites because of their high mutation rates. Instead, SCHLOTTERER (2002) proposed a new multilocus test statistic, ln RV, to detect the signature of selection using microsatellite data, which is based on the ratio of observed variances in repeat number in two groups of populations. ln RV has a normal distribution under neutrality, so the test of neutrality is a test of deviation from normality. SCHLOTTERER (2002) also tested the robustness of the procedure to the mutation model, in particular, to deviations from SMM, and found that using the TPM only slightly increases the variance of ln RV.

1.3.2 Dealing with genotyping errors

Microsatellite genotyping can be error prone, which could potentially undermine the conclusions of population genetic analysis. In fact, a number of studies have suggested that even moderate error rates can seriously effect estimates of population genetic parameters that I have discussed above, such as relatedness and parentage or population history and structure (*e.g.* MARSHALL *et al.*, 1998; HOFFMAN and AMOS, 2005; TABERLET and LUIKART, 1999). Genotyping errors may arise due to a variety of sources and at different stages of the study. For example, low sample DNA quality commonly leads to "allelic dropout", where heterozygotes appear to carry one allele due to one allele failing to amplify. The so called, "misprinting" arises when false amplification products are mis-interpreted as alleles. Even with high quality DNA samples, the presence of "null alleles", i.e. allele non-amplification due to primer binding site mutation (*e.g.* PEMBERTON *et al.*, 1995), is commonly observed. Finally, at the allele scoring stage, mis-scoring of alleles is commonly resulting from false, so-called "stutter bands".

The ultimate goal is obviously to minimize the error rates to as little as possible. This is generally best achieved by blind repeat genotyping (HOFFMAN and AMOS, 2005), but when such an approach is economically not feasible, statistical tests can be carried out to reveal the presence of genotyping errors. For example, testing for Hardy-Weinberg equilibrium can help to reveal the presence of homozygote excess due to null alleles. When genotyping errors cannot be completely eliminated, but are suspected to be present, they, at the best, have to be taken account of in the statistical analysis. Specific developments of statistical methods to take genotyping errors into account mostly come from the field of human genetics, for example, in genetic mapping studies (*e.g.* THOMPSON *et al.*, 2005), and from studies of natural populations as well, for example in parentage analysis (*e.g.* MARSHALL *et al.*, 1998).

1.3.3 Computational statistics in population genetics

1.3.3.1 An overview

The aim of all population genetic analysis is to try to understand the underlying biological process that has generated the observed data. However, extracting the relevant information from patterns in genetic data is extremely challenging. Inferences in the past were often based on simple null hypothesis tests, thus summarizing the data with a single statistic and comparing it to the distribution of the statistic under a simplified null model. By doing so, we do not necessarily make use of the full data, and results are often difficult to interpret because deviations from the null could be due to a range of different biological processes. Thanks to the advances in computer technology, starting from the mid-nineties many advances have been made towards basing inferences on full data likelihood in population genetics (KUHNER *et al.*, 1995; GRIFFITHS and TAVARÉ, 1994; BEERLI and FELSENSTEIN, 1999; WILSON and BALDING, 1998; NIELSEN and WAKELEY, 2001; CHIKHI and BEAUMONT, 2001). These were the first methods that tried to tackle complex, analytically intractable, population genetics problems with the aim of understanding an underlying biological process. The use of these new computationally intense simulation approaches opened a new phase of method development in population genetics.

Parameter estimation was achieved in most cases using Markov chain Monte Carlo (MCMC) methodology, which is a convenient tool to study the properties of a probability distribution that is analytically not tractable. MCMC, and other related methods such as importance sampling (IS), are so-called Bayesian statistical methods, which have become popular not only in population genetics (Figure 1.1), but in many other fields of genetics and science as well (BEAUMONT and RANNALA, 2004; SHOEMAKER *et al.*, 1999). As opposed to classical or frequentist inference, where parameters (that we are interested in estimating) are viewed as unknown but fixed, in Bayesian statistical inference parameters are regarded as random variables, thus have a probability distribution (such as the assumed prior distribution). The focus of Bayesian inference is to compute the probability distribution of the parameter having observed the data (the posterior).

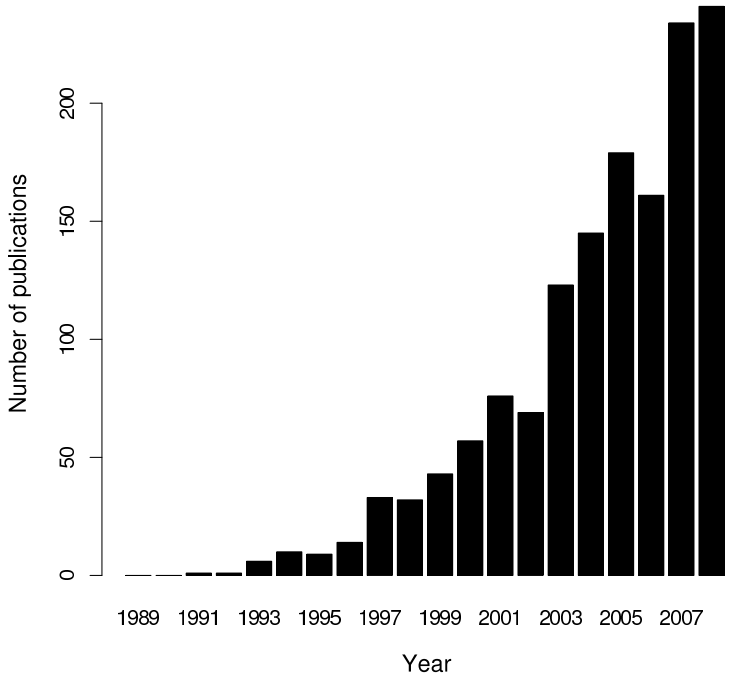

Figure 1.1: Number of publications with “Bayesian” or “likelihood” and “microsatellite” in their title or abstract from 1989 to 2008 according to the Web of Knowledge `http://apps.isiknowldge.com`.

Bayesian statistics offers many other advantages, such as the ability to easily incorporate background information via the use of the prior. However, BEAUMONT and RANNALA (2004) argue that the main reason for the “Bayesian revolution” in genetics was not the ability to incorporate background information, but the fact that complex likelihood problems that we often encounter in genetics can be tackled by MCMC methods. Indeed, most frequently, so-called objective priors are chosen, as opposed to priors that reflect our subjective beliefs about the parameter. For example, uniform priors that place equal weight on all possible values of the parameter are frequently used in the absence of biological background information. There exist only a few exceptions where non-genetic background information was incorporated in the analysis via priors (*e.g.* GAGGIOTTI *et al.*, 2004)

The other cornerstone of the new computationally intensive methods is the use of the coalescent (KINGMAN, 1982b). This is a different use of the coalescent in comparison to classic population genetics, where only certain properties of the coalescent, for example, the difference in coalescence times within and between populations, were used to justify a new statistic describing population structure (*e.g.* SLATKIN, 1995) or history (*e.g.* GARZA *et al.*, 1995). In modern computational statistics, the coalescent is used as tool for designing simulations. The coalescent is based on the idea of studying only the properties of a sample, which delivers

computational efficiency and also enables the simultaneous modelling of sampling and genetic drift. Thus, coalescent simulations can be very efficient in comparison to, for example, a Wright–Fisher forward simulation approach. When correlations between loci become important, the coalescent with recombination has to be employed, which can be computationally daunting. In such settings, often, a forward Wright-Fisher approach is coupled with the MCMC (*e.g.* FALUSH *et al.*, 2003). Fast approximations to the coalescent with recombination are promising alternatives (*e.g.* STEPHENS *et al.*, 2001).

Although MCMC approaches are very efficient, they can be computationally intractable with large and complex data sets and for models with many parameters (*e.g.* MARJORAM *et al.*, 2003). In recent years, with advances in molecular biology, genotyping hundreds of individuals has become not only feasible but cost effective. As a result, many more large and complex data sets started appearing, for which the computational tractability of MCMC can drop drastically. For example, it has been reported that the run time of the software Structure can significantly increase for large and complex data sets, especially under the linkage model (FALUSH *et al.*, 2003). Thus, interest has turned from full data likelihood to approximate methods. For example, Approximate Bayesian Computation (ABC) (BEAUMONT *et al.*, 2002) is a method where the data are replaced by summary statistics, so that it can remain computationally tractable for highly complex models provided that simulation of data under the model is feasible. Due to the relative ease of use and the attractive concept of ABC, it has gained popularity in recent years in various applications (*e.g.* HICKERSON *et al.*, 2006; ROSENBLUM *et al.*, 2007; ESTOUP *et al.*, 2004; INGVARSSON, 2008). Another approximate approach that has recently been explored in some population genetics problems is the "product of approximate conditionals" or PAC (LI and STEPHENS, 2003). PAC can be used to efficiently approximate the likelihood. The utility of PAC models have not been explored in a great detail for microsatellites yet, but, for example, PAC has been shown to be accurate with microsatellite data in a recent study (CORNUET and BEAUMONT, 2007).

Although the computationally intensive likelihood-based and approximate Bayesian methods have brought great progress in our understanding in many areas of population genetics, they have limitations as well. A general problem of both full likelihood and approximate methods is that for complex models with many parameters it is often not feasible to systematically study the sensitivity of models to priors, which is of key importance in Bayesian inference (GELMAN *et al.*, 2003). An MCMC-specific problem is the difficulty of qualitatively assessing the convergence of the Markov chain to the target distribution (the posterior) that we sample from, and thus, more generally,

the reliability of the parameter estimates (GELMAN, 1996). A similar problem arises for ABC. In simple problems, where a full data likelihood approach is also feasible, ABC is generally much faster and performs only slightly worse (BEAUMONT *et al.*, 2002). However, in more complex problems there are indications that the performance can be unpredictable. This is a problem that I will investigate for the case of population admixture in Chapter 4 of this thesis.

1.3.3.2 Microsatellite data in simulation based methods

From the point of view of microsatellite data, a great advantage of the new simulation-based methods is the ease with which, in theory, any kinds of mutation models can be incorporated. This is particularly true for the coalescent-based and approximate methods. Despite this possibility, mutation models more elaborate than the SMM have rarely been implemented (Table 1.1), and software that adopt a forward approach (*e.g.* FALUSH *et al.*, 2003; CORANDER *et al.*, 2004; PELLA and MASUDA, 2006; PRITCHARD *et al.*, 2000a) use a Dirichlet model for the allele frequencies (Table 1.1), with which I have already pointed out some problems in the previous section. Some authors base their justification for using the SMM on a data set. For example, WILSON *et al.* (2003) consider using GSM in software Batwing, but then argues that the SMM is good approximation based on a recent study of PRITCHARD *et al.* (1999). However, I suggest that the main reason for using SMM is because a mutation model with more parameters can significantly increase the computational time. Actually, even the SMM can be computationally demanding in an MCMC scheme, thus some packages adopted a Brownian motion approximation to the SMM, which has been shown to be accurate, but much faster (*e.g.* KUHNER, 2006; STEPHENS *et al.*, 2001).

Table 1.1: The different mutation models implemented in recently developed population genetics analysis methods that are also implemented as software packages. Abbreviations of the mutation models: SMM: Stepwise mutation model, BM: Brownian motion approximation to SMM, SMM+KAM: a mixture of SMM and KAM (K-alleles model). Although Simcoal2 is a simulation and not an estimation tool I added to the list because it can be (and has been) used in approximate estimation procedure.

Software	**What it does**	**How it models microsatellites**	**Reference**
MCMC and forwards Wright–Fisher model			
NewHybrids	identifies species hybrids	Dirichlet prior	ANDERSON and THOMPSON (2002)
Structure	infers population structure	Dirichlet prior	FALUSH *et al.* (2003)
BAPS2	infers population structure	Dirichlet prior	CORANDER *et al.* (2004)
Geneland	infers population structure	Dirichlet prior	GUILLOT *et al.* (2005a)
Structurerama	infers population structure	Dirichlet prior	PELLA and MASUDA (2006)
MCMC and coalescent			
Bottleneck	detects recent reductions in population size	SMM, TMP	CORNUET and LUIKART (1997)
Micsat	estimates population size	SMM	WILSON and BALDING (1998)
MIGRATE-N	estimates population size and geneflow	SMM, BM	BEERLI and FELSENSTEIN (1999)
PHASE	infers haplotypes	SMM, SMM+KAM	STEPHENS *et al.* (2001)
Batwing	estimates population size	SMM	WILSON *et al.* (2003)
LAMARC	infers population history and structure	SMM, BM, SMM+KAM	KUHNER (2006)
IM and IMa	estimates demographic parameters	SMM	HEY and NIELSEN (2007)
Approximate (ABC)			
DIYABC	estimates demographic parameters	SMM, GSM	CORNUET *et al.* (2008)
Simcoal2	coalescent simulation tool	SMM, GSM	LAVAL and EXCOFFIER (2004)

The importance of implementing mutation models other than the SMM is not because they are necessarily more "realistic" models of the microsatellite mutation process, but the fact that the sensitivity of the mutation model assumptions to the final conclusions could be investigated. Although many software manuals and papers emphasize the importance of sensitivity analysis, almost none have been carried out so far. For example, the software IM (HEY and NIELSEN, 2007) implements only the SMM, but comments in the manual that it needs to be investigated whether analyzing loci that do not fit the SMM, with an assumption of a SMM, leads to significant bias in the estimates of demographic parameters or not. Not surprisingly, it is mostly the packages that implement the faster approximate methods that accommodate more elaborate mutation models, for example the GSM is available in DIYABC (CORNUET *et al.*, 2008). Also, the software PHASE (STEPHENS *et al.*, 2001) and LAMARC (KUHNER, 2006), implements a modification to the SMM, which can be regarded as a mixture between SMM and KAM, since with some user specified probability the mutant allele may change to any other present alleles with equal probability. The availability of these options may enhance sensitivity analysis in the future studies.

When a mutation model with more parameters than just the mutation rate, such as in SMM, is applied the choice of the mutation parameters could pose a challenge as well. For example, how to choose the switching rate from SMM to KAM in PHASE? Or, how to choose the Geometric parameter in GSM? Here, I estimate the Geometric parameter of GSM for many systems using published data, which may give guidance for choosing parameters in future studies. I choose the GSM because it is a relatively simple, two parameter model, and some studies suggest that it is generally better supported then the SMM (CORNUET *et al.*, 2006; WHITTAKER *et al.*, 2003), and also it has been implemented in some software packages (Table 1.1).

Data came from two types of studies: pedigrees and mutation accumulation (MA) lines. For estimates of the mutation rates, μ, I either took the estimate that was reported in the paper or occasionally combined estimates from different publications or calculated the mutation rates from the available information. I assumed that in the pedigree-type studies when a parent-offspring mismatch is detected the mutation is derived from the parental allele which requires the lesser change in the number of repeat units. If two values were available for the length mutations and both were reported I chose the smaller change, because this was the information that all papers have consistently reported. Then, I estimated the parameter p, which is the parameter of a Geometric distribution of the variable $x - 1$, where x is the size of the change in repeat numbers in a single mutation event. Thus, $p = 0$ corresponds to the simple SMM. This is the same parameter that is used by the software Simcoal2, for example

(Table 1.1). Finally, I performed a goodness of fit test to study if the data conforms with the Geometric model.

Table 1.2 shows that both μ and p vary greatly depending on the study system. Although the sizes of the studies, both regarding the number of transmissions/meioses and the number of loci, differ by up to two orders of magnitude, there seems to be no trend with the size of the study and μ or p (not shown). In many cases, I detected significant deviation from the Geometric model (Table 1.2). The deviations were mainly caused by that fact that the data is overdispersed, i.e. there are more multi-step changes than predicted by the Geometric model. These findings provide further evidence for the complexity of the mutation process at microsatellite loci and emphasize the importance of sensitivity analysis when making inferences from microsatellite data.

Table 1.2: Mutation rates and estimates of the geometric parameter of the GSM for various species. Data come from two types of studies (Type) pedigrees (P) and mutation accumulation lines (MA). L is the number of loci, N is the number of transmissions for pedigree studies and the number of lines and generations, respectively, for mutation accumulation lines. p is the estimate of parameter of a Geometric distribution for $x - 1$, where x is the change in step size in a single mutation event. p-values correspond to a goodness-of-fit test from fitting a Geometric distribution with parameter p to the observed data.

Species	**Type**	**Nucleotide**	**L**	**N**	μ	p	p**-value**	**Reference**
human	P	di, tri, tetra	28	20000	$1.2 \times 10^{3-}$	0.043	<0.001	Weber and Wong 1993
human	P	di	362	499650	1.94×10^{-4}	0.623	0.192	Huang et al. 2003
human	P	di	400	118866	4.5×10^{-4}	0.271	<0.001	Whittaker et al. 2003
D. melanogaster	MA	di	10	122,–	1.03×10^{-4}	0.579	<0.001	Harr and Schlotterer 2000
Egernia stokesii	P	tetra	7	5980	1.22×10^{-2}	0.458	0.221	Gardner et al. 2000
Malurus cyaneus	P	di and tetra	2	5980	1.25×10^{-2}	0.292	<0.001	Beck et al. 2003
Hirundo rustica	P	tetra, penta	3	5973	1.91×10^{-2}	0.295	0.032	Brohede et al. 2002, 2004
Daphnia pulex	MA	di	22	268, 27	9.6×10^{-5}	0.611	0.135	Seyfert et al. 2008
C. elegans	MA	di, tri, tetra	23	80, 140	2.01×10^{-3}	0.732	<0.001	Seyfert et al. 2008

Chapter 2

Performance of marker-based relatedness estimators in natural populations of outbred vertebrates

The material presented in this Chapter closely resembles my publication in Genetics, CSILLÉRY et al. *(2006). I declare that the data analysis and the simulation study were performed by me, and also I wrote the paper. My co-authors contributed with data and ideas. I would further like to acknowledge Bill Hill, Penny Kukuk, Allen Moore, Jon Slate and Alastair Wilson for useful discussions and for comments on my manuscript.*

2.1 Introduction

Inferring relatedness among pairs of individuals plays a central role in our understanding of many areas of genetics and population biology. For example, the extent of relatedness between individuals is important in the study of social evolution (*e.g.* HAMILTON, 1964; CHEVERUD, 1985) and studies incorporating measures of relatedness have influenced our understanding of the mechanism of kin selection in natural populations (*e.g.* CHOE and CRESPI, 1997). In quantitative genetics the estimation of genetic variance components, allowing estimation of heritability and genetic correlation, requires pairs of individuals with known relatedness (LYNCH and WALSH, 1998). In conservation biology, knowledge of relatedness is essential in captive management, where the goal is to preserve the genetic variation of the wild population from which the founders were drawn (*e.g.* LACY, 1994). Relatedness estimates are also used when testing hypotheses about inbreeding avoidance (*e.g.* REUSCH *et al.*, 2001; RICHARDSON *et al.*, 2004) and isolation by distance (*e.g.*

MATOCQ and LACEY, 2004).

Relatedness has traditionally been estimated from pedigrees. Given an outbred source population and good recording, laboratory or managed populations can instantly provide pedigree information. However, many relevant ecological and evolutionary questions can only be addressed in free-living populations with the help of molecular marker data (KRUUK, 2004). When relatedness estimation can be simplified to hypothesis testing over candidate genetic relationships on the basis of some prior life history or partial pedigree information, maximum likelihood methods have been successfully applied (THOMAS, 2005). However, in the absence of prior information, inferences will need to be based solely on marker data. In such cases, with the most commonly available marker numbers (5-20 microsatellite loci), the method of moments estimators are preferred because the ideal properties of the maximum likelihood estimators are only achieved asymptotically, i.e. as the number of loci typed becomes vary large (LYNCH and RITLAND, 1999; WANG, 2002; MILLIGAN, 2003). Thus, the moments estimators developed by QUELLER and GOODNIGHT (1989); LI *et al.* (1993); RITLAND (1996a); LYNCH and RITLAND (1999); WANG (2002) have became the most commonly used; hereafter abbreviated QG, L, R, LR, and W, respectively. There is a considerable interest in the performance of these relatedness estimation methods because, in theory, they could make any species accessible for estimating pairwise relatedness (BLOUIN, 2003).

Most previous studies have evaluated the performance of the estimators for the most common first and second order genetic relationships in isolation using Monte Carlo simulations (*e.g.* LYNCH and RITLAND, 1999; WANG, 2002; VAN DE CASTEELE *et al.*, 2001; MILLIGAN, 2003). These studies, using either theoretical or empirical allele frequency distributions, demonstrated that first, the performance of the estimators depends on many factors, including the number of loci and alleles, the shape of the allele frequency distribution, and the relatedness itself (QUELLER and GOODNIGHT, 1989; RITLAND, 1996a; LYNCH and RITLAND, 1999; WANG, 2002; MILLIGAN, 2003), second, that estimators generally exhibit a high sampling variance (VAN DE CASTEELE *et al.*, 2001), and third, as a result, the best performing estimators are different depending on the population under investigation (VAN DE CASTEELE *et al.*, 2001).

Although it is useful to know the performance of the estimators for individual relationships, in practice, relatedness is usually estimated among all pairs of individuals in a sample. Thus, we may want to quantify the performance of the estimators across all pairs, i.e. across a range of genetic relationships, what we may call “average performance”. Two measures have been proposed that may be used to measure the

average performance.

Although pairwise relatedness estimators were not developed to classify pairs of individuals to simple genetic relationships, BLOUIN *et al.* (1996) suggested the estimation of the misclassification rate between relationship categories to estimate the error rate if the estimators were used in such way. BLOUIN *et al.* (1996) defined the misclassification rate between full-sibs and half-sibs as the proportion of pairs that belong to one of the relationships but would be classified as the other using the QG estimator. The mean relatedness of the two relationship categories, given some allele frequencies, was determined via simulations and then, the midpoint between the means was used as the cutoff point to classify individuals of unknown relatedness as either full- or half-sibs. Error rates were estimated in both directions and used as a performance measure, so that for example, studied as a function of the number of markers. This method has been applied in many recent studies comparing the common first and second order genetic relationships (*e.g.* RUSSELLO and AMATO, 2004; FRASER *et al.*, 2005). Some authors have advocated that the moment estimators can accurately discriminate first order relationships from microsatellite marker data (*e.g.* GERLACH *et al.*, 2001; RUSSELLO and AMATO, 2004; SEKINO *et al.*, 2004; FRASER *et al.*, 2005).

The other approach was proposed by VAN DE CASTEELE *et al.* (2001), who estimated the proportion of variance explained in the marker-based relatedness estimates by true relatedness, given a population relatedness composition. In the absence of knowledge of the true population relatedness composition they simulated various arbitrary population compositions as a mixture of unrelated, half-sib, full-sib, and parent-offspring pairs. They found that the variance explained by true relatedness was generally high, ranging from 25%-79% (median: 52%) over ten possible relatedness compositions and using the estimator that performed the best for the given set of observed allele frequencies. Of particular note, the r^2 was the smallest for the population with the highest proportion of unrelated pairs (60% unrelated). RUSSELLO and AMATO (2004) applied the same method but tried to estimate the population relatedness composition using a likelihood-based method. The fraction of a particular relationship in the population was estimated as the likelihood of drawing the observed distribution of the relatedness estimates from the sampling distributions of any of the four relationships, unrelated, half-sib, full-sib, parent-offspring. The resulting population compositions were rather similar to those of VAN DE CASTEELE *et al.* (2001), and thus the proportion of variance explained by true relatedness was also high, ranging from 35%-52% (median: 50%) over 11 possible population relatedness compositions.

A major limitation of the previous studies is that the average performance of the relatedness estimators was investigated without reliable estimates of the true population relatedness composition. This might have led to inaccurate estimates of estimator performance for two reasons. First, the proportions of different relationships may well be different from those assumed, and second, there may be a non-negligible proportion of higher order relationships in natural populations. In fact, the proportions of highly related pairs were unrealistically high in previous studies; e.g. at least 40% of the pairs had relatedness 0.25 or higher in the simulations of VAN DE CASTEELE *et al.* (2001), which is unlikely to be the case in most natural populations of outbred diploids.

In this Chapter I assess the average performance of the available method of moments marker-based relatedness estimators using five unique data sets from long term projects of outbreeding vertebrates. Since both deep pedigree and marker data are available, I have reliable knowledge of the population relatedness composition as well as the allele frequencies. In particular, I address the point that the population relatedness composition of natural populations may often differ greatly from that assumed by previous investigators (*e.g.* VAN DE CASTEELE *et al.*, 2001; RUSSELLO and AMATO, 2004). The effect of marker data quality, particularly the number of loci and level of polymorphism, has been emphasized previously (*e.g.* LYNCH and RITLAND, 1999; WANG, 2002). Here, I also examine the effect of the marker data quality, but in the specific context of its importance relative to the population relatedness composition.

2.2 Methods

2.2.1 Observed populations

2.2.1.1 Long term projects

The starting point for this study is five outbred vertebrate populations, the meerkats, great reed warbler, bighorn sheep, red deer, and Soay sheep, which have each been subjected to intensive individual-based research over several (overlapping) generations. Since the five data sets differ in some underlying features (e.g. mating system) that may effect performance of estimators, e.g. through influences on relatedness composition of the population, here I provide more details about each species.

Data on meerkats (*Suricata suricatta*) were collected from semi-arid savannah near Vanzylsrus in the Northern Cape of South Africa, where a long term study has been

established since 1993 (CLUTTON-BROCK *et al.*, 1998). The study population is spatially continuous and the migration rate is high from the neighboring areas (i.e. unstudied groups). Meerkats have a nearly monogamous mating system and live in groups of three to 20 adults and sub-adults accompanied by dependent young (CLUTTON-BROCK *et al.*, 1999). Each group is composed of a dominant pair, who produce most, but not all pups, subordinates that were born in the group and a variable number of immigrant males. The life span of meerkats is up to 5-15 years (VAN STAADEN, 1994).

A breeding population of the migratory great reed warbler (*Acrocephalus arundinaceus*) has been studied at Lake Kvismaren in Southern central Sweden since 1983 (HASSELQUIST, 1998; BENSCH, 1996). The population was founded by a few individuals in 1978. Great reed warblers are facultatively socially polygynous, with males forming new pair bonds with up to five females each season while about 20% of the males remain unpaired (HASSELQUIST, 1998). The median clutch size is five. Great reed warblers had an average life span of 2.7 years in the study population (HANSSON *et al.*, 2004).

The bighorn sheep (*Ovis canadensis*) population at Ram Mountain, Alberta, Canada has been intensively monitored since 1975 (JORGENSON *et al.*, 1998; COLTMAN *et al.*, 2002). Bighorn sheep are highly polygynous, and rams have up to 22 mating partners through their life time in the study population. Reproduction is highly seasonal, and after maturation (at age three or four), females produce a single offspring per year. Bighorn sheep may live for over 10 years (FESTA-BIANCHET, 1999), but the lifespan of males is significantly shortened by trophy hunting in the study population (COLTMAN *et al.*, 2003).

Red deer (*Cervus elaphus*) are the subject of a long-term study in the North Block of the Isle of Rum, Scotland since 1973 (CLUTTON-BROCK *et al.*, 1982). The population was founded by introductions starting in 1845, and was sourced from at least four different British mainland populations. Red deer have a polygynous mating system and males had up to 26 mating partners through their life time in the study population. Reproduction is highly seasonal. Females typically become mature when three or four years old and do not necessarily breed every year; if they do they produce a maximum of one calf per year. Average life-span is 10-12 years.

The Soay sheep (*Ovis aries*) population on the island of Hirta (St. Kilda Archipelago, Scotland) has been intensively studied since 1984 (CLUTTON-BROCK and PEMBERTON, 2004). The population was founded in 1932 by introduction from a neighboring island. The Soay sheep is a primitive domestic sheep with a promiscuous, polygynous mating system. Reproduction is seasonal, with ewes reproducing once a year. Females start breeding either in the first or second year of their life and

produce one or two offspring per year (CLUTTON-BROCK and PEMBERTON, 2004). The twinning rate is density dependent and fluctuates between 3% and 25%. The life-span is also density dependent and averages to three years.

2.2.1.2 Pedigrees

In all data sets maternity was determined by observation and paternity was assigned using microsatellite genotypes and the likelihood-based paternity inference software CERVUS (MARSHALL *et al.*, 1998), with the exception of great reed warblers (see ARLT *et al.* 2004 for details). Pedigrees were further corroborated by identifying parent-offspring mismatches in both maternal and paternal lines and removal of some dubious paternities. In order to justify the comparability of the five pedigrees in terms of their quality, we estimated the depth of the pedigrees by counting the number of ancestors that were present in the pedigree for each individual and averaged over all individuals. Table 2.1 summarizes the microsatellite marker and pedigree data and demonstrates that the pedigrees are similar in quality.

Table 2.1: Summary of the microsatellite markers and pedigrees of the five data sets from natural populations. Marker data quality parameters for the great reed warbler and Soay sheep populations in parenthesis correspond to the mapping data sets (see text for details). Pedigree depth was estimated by counting the number of ancestors that were present in the pedigree for each individual, averaged over all individuals. The population relatedness composition based on the pedigrees is, for simplicity, summarized as proportion of pairs that would be classified as unrelated, half-sib, or full-sib (parent-offspring) if only a two generation deep pedigree was available. Note that this is great simplification of the true relatedness composition (see text and Figure 2.2 for details).

	Meerkat	**Great reed warber**	**Bighorn sheep**	**Red deer**	**Soay sheep**
Marker data quality					
Number of loci	8	15 (62)	30	8	19 (101)
Median of number of alleles	12.5	10 (5)	5	9	6 (5)
Range of number of alleles	5-17	4-26 (2-78)	3-10	6-10	4-8 (3-8)
Pedigree quality					
Number of individuals	792	802	910	3544	3550
Number of individuals with both parents known	586	706	382	1003	1720
Pedigree depth	6.3	5.4	4.6	4.6	5.9
Simplified relatedness composition					
Percentages of pairs with relatedness					
- less than 0.25	90.8	96.6	98.9	99.6	99.7
- less than 0.5, but at least 0.25	7.5	2.5	0.9	0.3	0.2
- at least 0.5	1.7	0.9	0.2	0.1	0.1

2.2.1.3 Microsatellite markers

For each population a set of polymorphic unlinked markers was selected (see Table 2.1). For great reed warbler and Soay sheep, a larger set of markers was available, which had been scored for linkage and QTL mapping (HANSSON *et al.* 2005; D. Beraldi, unpublished data). Although many of the markers in these larger data sets are linked, it did not matter for the sake of this study because I only used allele frequency information from these markers for simulating marker genotypes, i.e. as hypothetical sets of unlinked markers.

2.2.2 Simulated populations

In order to simulate realistic populations I estimated the allele frequencies and the relatedness compositions from the five data sets and treated them as known (see the different steps of data treatment in Figure 2.1). I estimated the population relatedness composition as the relative frequencies of the relatedness coefficients among all possible pairs in the observed pedigrees. The number of different kinds of relationships was large, and many of them were represented by only a few pairs. In order to simplify the analysis, but still keep the complexity of the population relatedness structure, I took a relationship into consideration if it was represented by at least 50 pairs in the sample. In this way less than one percent of the total pairs were lost. Relatedness coefficients were calculated using the *kinship* package of R (ATKINSON, 2005).

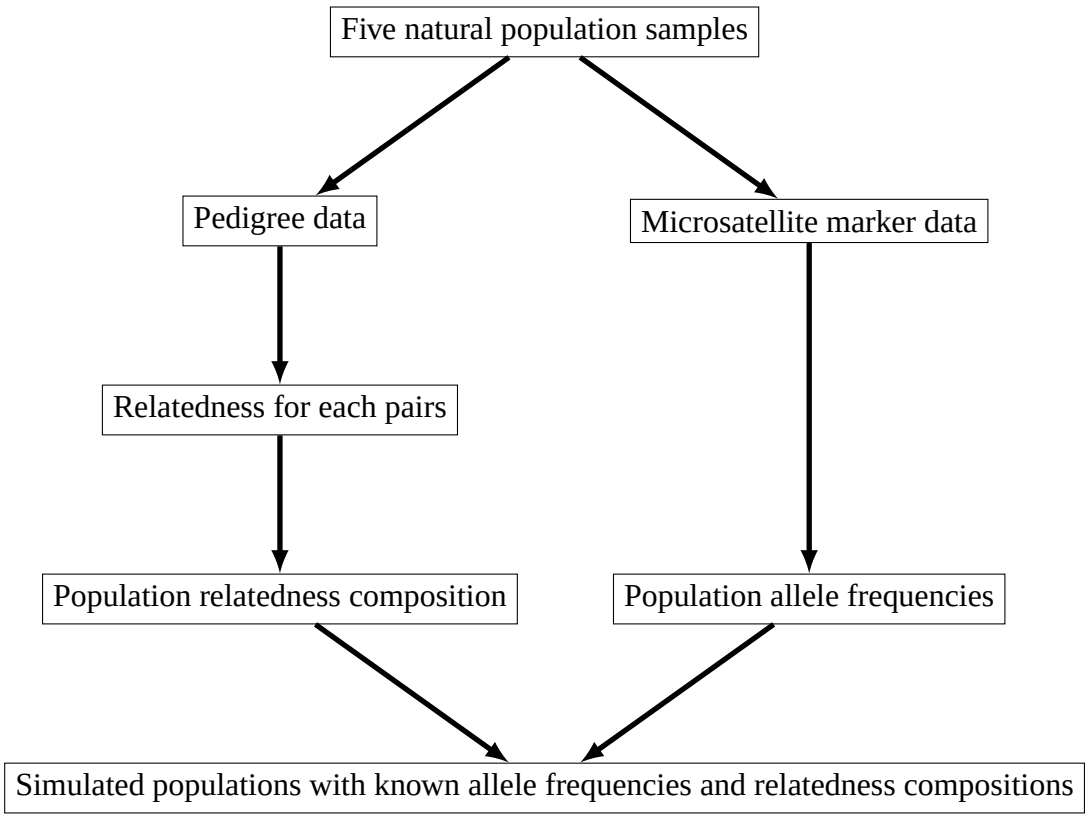

Figure 2.1: The steps of the data simulation to estimate the average performance of relatedness estimators across different population relatedness compositions. The analysis was designed to avoid using the data in a circular way: the pedigree data was used only to estimate the population relatedness composition, while the marker data were used only to estimate the allele frequencies.

Using the observed population allele frequencies for the marker loci 10,000 multilocus genotype pairs were simulated as follows. Reference genotypes were drawn randomly according to their Hardy-Weinberg and linkage equilibrium frequencies and genotypes of the pairs were drawn randomly from their conditional genotypic distribution given a particular genetic relationship. All observed genetic relationships were simulated for each of the five populations, using the corresponding population allele frequencies. Since any given non-zero relationship can be simulated with more than one pedigree configuration (which greatly increases the number of relationships to be simulated) I assumed, for simplicity, that, for a given kind of relationship, the members of the pair were related via a single common ancestor, and thus have zero probability of having two genes identical by descent. There was one exception, relatedness 0.5, where I simulated both the parent-offspring and full-sib relationship, as both of them were present in all five pedigrees.

The marker-based relatedness estimates were calculated for all five published

relatedness estimators, but results are presented only for the QG, LR, and W estimators, for simplicity. I choose these estimators because the QG is the most commonly used estimator, the LR estimator is an improved version of the R estimator, and the W can be considered as an improved version of the L estimator. The assumptions of the marker-based relatedness estimators hold in the simulated populations, such that the populations are in Hardy-Weinberg equilibrium, there are no genotyping errors, loci are unlinked and selectively neutral, and population allele frequencies are known. Marker-based relatedness estimators were used in their published form; for formulae consult (*e.g.* WANG, 2002). Estimates were calculated using functions written in R (R DEVELOPMENT CORE TEAM, 2005).

2.2.3 Measuring estimator performance

I quantified the average performance of the estimators using two different measures. First, I used the method suggested by BLOUIN *et al.* (1996) and calculated the misclassification rates among the unrelated, half-sib, full-sib, and parent-offspring relationships. Error rates were calculated as the proportion of pairs for a given relationship that were misclassified as another relationship or the proportion of pairs that belong to the latter relationship, but classified as the former, based on a cutoff point defined as the midpoint between the sampling distributions of the two relationships (subsequently called "naive estimates"). I discourage the use of BLOUIN *et al.*'s (1996) original terms, Type I and Type II error for the misclassification rates in the two directions because we are not testing a hypothesis of one relationship over another. Then we re-estimated the misclassification rates using knowledge of the population relatedness composition, i.e. knowing that each of the four simple relationships actually incorporate many other higher order relationships. Thus, the sampling distributions of the four relationships were substituted with that of a mixture of higher order relationships observed from the pedigrees. The misclassification rates were determined using the same cutoff points as before, since these are what one could determine before the study was conducted (subsequently called "real estimates"). Comparing the naive and real estimates allowed us to estimate the error of the misclassification rates caused by the assumption that the population is composed of the four simple relationships.

Second, I calculated the proportion of variance explained in the marker-based relatedness estimates by true relatedness (VAN DE CASTEELE *et al.*, 2001), for which I used the observed (pedigree-based) population relatedness composition. In order to generate a population of given relatedness composition, 10,000 pairs of different relationships were drawn according to the observed population proportions (see Table

2.1 for a simplified version of the five population relatedness compositions). The variance explained by the true relatedness was estimated as the between group sum of squares divided by the total sum of squares (r^2) and averaged over 500 realizations of the given relatedness composition.

Any variation in the proportion of variance explained between populations either arises from differences in the population relatedness composition or differences in the marker data (number of loci and/or levels of polymorphism). In order to address how these two factors play a role in driving the proportion of variance explained by true relatedness, I analyzed in greater detail the great reed warbler and Soay sheep data sets, which turned out to be the most and least favorable populations in terms of maximizing the r^2, and were also the two populations for which the largest sets of marker data were available (Table 2.1). I randomly selected five different sets of five, ten, 20, 30, 40 marker allele frequencies from the available 101 and 62 loci in the Soay sheep and great reed warbler populations, respectively and studied r^2 as a function of the number of markers. Since there was variation in the level of polymorphism between markers, I also investigated the effect of polymorphism on the average performance. I randomly selected 50 different sets of five markers, calculated the mean number of alleles as a polymorphism measure and compared the effect of polymorphism in the two populations. I choose to investigate the polymorphism effect using sets of five markers instead single locus estimates, because, in practice, we are generally interested in multilocus estimates.

2.3 Results

A simplified version of the relatedness composition of the five populations illustrates that in all populations the majority of the pairs have relatedness less than 0.25, or in other words less related than e.g. half-sibs, and would thus be classified as unrelated using a shallow pedigree (Table 2.1). The highest proportion of pairs with relatedness equal to or higher than 0.25 were in the meerkat and great reed warbler populations, reflecting the fact that these two populations have the highest number of half-sib and full-sib pairs. Further, the sampling distributions of estimated relatedness for the most common relationships illustrate that the deep pedigrees recovered a large number of relationship categories in all five populations (see Figure 2.2 for the QG or the LR estimators). The number of relationships is the highest in the meerkat and red deer populations, 347 and 471 respectively, reflecting a higher number of inbreeding events, while it is much less, 72, 103, and 76 in the great reed warbler, Soay sheep, and bighorn sheep populations, respectively. Figure 2.2 also shows that

density curves for the different relationships overlap greatly, especially for the low relationship categories, i.e. below relatedness of 0.5, and that only the density curves for the parent-offspring and full-sib relationships (at relatedness of 0.5) are, at most, somewhat distinct from the rest of the relationships. Regarding differences between the two estimators illustrated, the QG has smaller sampling variance for the high relationship categories (density curves are peaked), while when the LR estimator is used the sampling variance is smaller for the low relationship categories. The W estimator is similar to the QG estimator, thus has smaller sampling variance for the high relationship categories (data not shown).

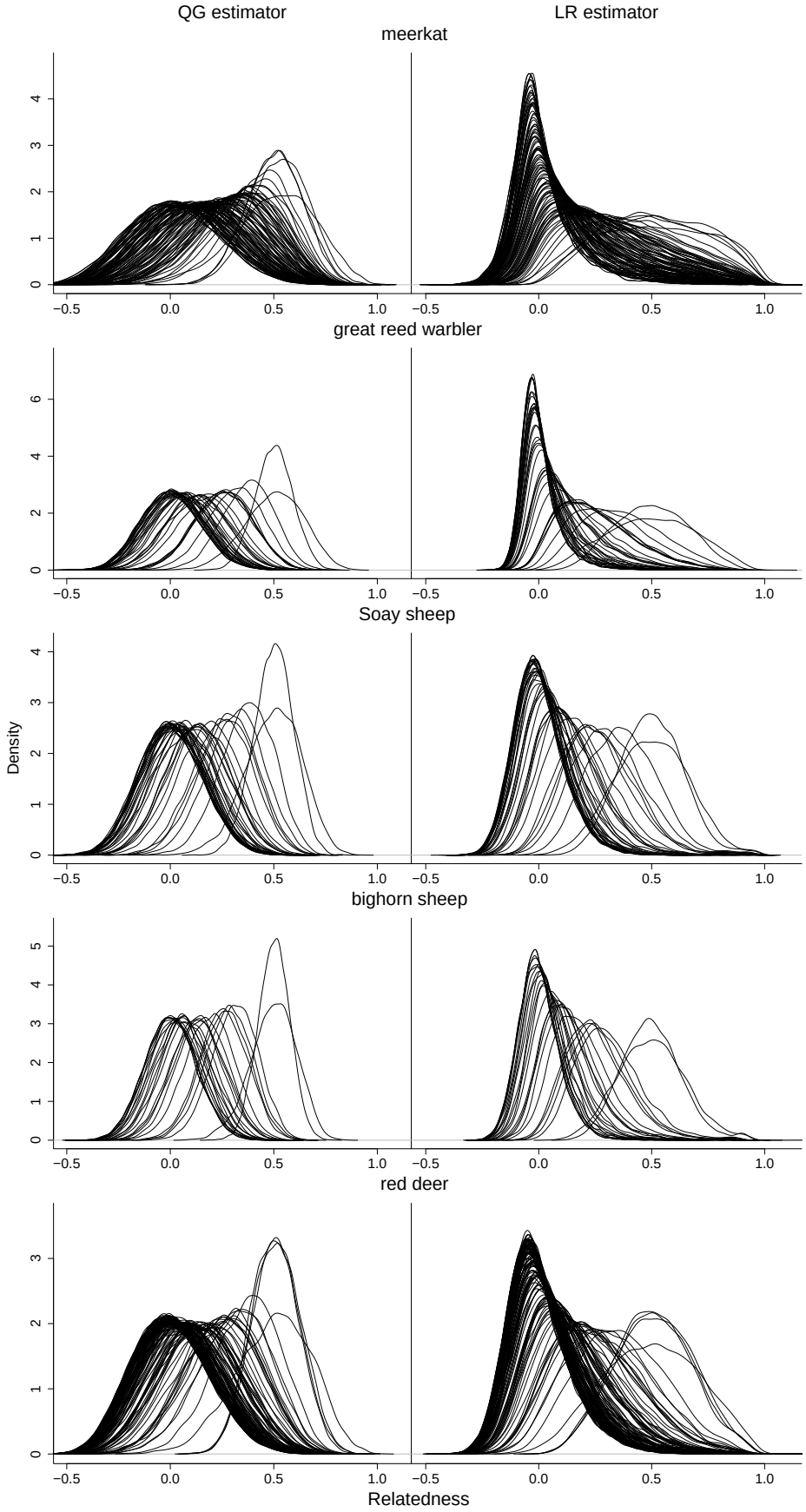

Figure 2.2: Sampling distributions of the most common pedigree-based relationships in the five natural populations when pairwise relatedness is calculated using the QG or the LR estimators. Each density curve is based on 10,000 simulated pairs of the most common genetic relationship in each of the five pedigrees. The observed population allele frequencies were used estimating relatedness.

Misclassification rates were calculated in two ways. First, a "naive estimate" was calculated between each pair of the following four relationships: unrelated (UR), half-sib (HS), full-sib (FS), and parent-offspring (PO), by assuming that the population comprises only these four relationships, i.e. there are no other, higher order relationships. Second, I calculated the misclassification rates between the same four relationships as before, but also using the information from the pedigree i.e. the fact that each of the four simple relationships are actually a mixture of higher order relationships ("real estimate", see Table 2.2). The results are illustrated using two estimators, LR and W, and two populations, the bighorn sheep and meerkats, because they represent the best and worse performances, respectively, of BLOUIN *et al.*'s (1996) misclassification method. Table 2.2 shows that, since relatedness estimates for the four relationships are lower bounds (relatedness is either that or higher), the misclassification rates to any higher relationships are actually higher (i.e. the real estimate is higher than the naive one; see upper right corners in Table 2.2) and the misclassification rates to any lower relationships are actually lower (i.e. real estimates are lower than naive; see lower left corners in Table 2.2). The difference between the naive and real estimates, however, was often very small because the proportion of pairs that are more related than the given relationship is relatively low. This is especially true for the populations of ungulates, e.g. the bighorn sheep (see Table 2.2; similar figures were obtained for Soay sheep and red deer, results not shown). The difference between the naive and real estimates is the greatest for the meerkats (see Table 2.2) and great reed warblers (data not shown). Note that these are lower bounds for the biases in the misclassification rate estimates because they are limited by the depth of the observed pedigrees. Regarding the differences between estimators, I found that in all five populations the misclassification rates are the lowest when the LR estimator is used.

Table 2.2: Misclassification rates between pairs of genetic relationships illustrated by the W and LR estimators and for the bighorn sheep and meerkat populations. Four relationships are shown on each axis, unrelated (UR), half-sib (HS), full-sib (FS), and parent-offspring (PO). Row headers indicate true relationship and column headers indicate the relationship into which pairs were misclassified. Each cell of the table compares misclassification rates calculated in two different ways, the first value is the "naive estimate" and the second value is the "real estimate" (see Methods for details). Both values are means of 500 independent draws of the population relatedness compositions from 10000 simulated pairs for each genetic relationships.

W estimator, bighorn sheep population

		Misclassified as			
		UR	HS	FS	PO
True relat.	UR	-	0.17 / 0.18	0.027 / 0.031	0.027 / 0.031
	HS	0.145 / 0.142	-	0.147 / 0.151	0.144 / 0.149
	FS	0.015 / 0.015	0.132 / 0.132	-	0.489 / 0.489
	PO	0 / 0	0.045 / 0.044	0.496 / 0.491	-

W estimator, meerkat population

		Misclassified as			
		UR	HS	FS	PO
True relat.	UR	-	0.311 / 0.366	0.143 / 0.186	0.143 / 0.187
	HS	0.261 / 0.217	-	0.285 / 0.339	0.287 / 0.341
	FS	0.103 / 0.103	0.246 / 0.246	-	0.479 / 0.479
	PO	0.017 / 0.017	0.146 / 0.146	0.513 / 0.511	-

LR estimator, bighorn sheep population

		Misclassified as			
		UR	HS	FS	PO
True relat.	UR	-	0.087 / 0.098	0.009 / 0.013	0.009 / 0.013
	HS	0.183 / 0.18	-	0.176 / 0.179	0.177 / 0.181
	FS	0.044 / 0.044	0.208 / 0.208	-	0.503 / 0.503
	PO	0.021 / 0.022	0.177 / 0.178	0.52 / 0.522	-

LR estimator, meerkat population

		Misclassified as			
		UR	HS	FS	PO
True relat.	UR	-	0.142 / 0.208	0.054 / 0.097	0.054 / 0.098
	HS	0.37 / 0.321	-	0.269 / 0.313	0.271 / 0.315
	FS	0.199 / 0.199	0.342 / 0.341	-	0.496 / 0.496
	PO	0.151 / 0.152	0.321 / 0.322	0.516 / 0.516	-

I found that the proportion of variance explained in the marker-based relatedness estimates by true relatedness (r^2) was generally low, especially in the three populations

of ungulates, but two to 14 times higher in the great reed warbler and meerkat populations (Table 2.3). There is a considerable difference between estimators; notably in all five populations the highest proportion of the variance is explained when the LR estimator is used, which reflects the fact that this estimator has the smallest sampling variance for unrelated or low relatedness pairs which are the most common, having over 90% frequency, in all five populations (see Table 2.1). The W estimator shows the poorest performance in all five populations. Table 2.3 also highlights two aspects of the populations that are potentially responsible for the between population differences: the number of loci and the variance in relatedness, a summary of the population relatedness composition.

Table 2.3: Proportion of variance explained in marker-based relatedness estimates by true relatedness in simulated populations based on the observed relatedness composition and allele frequencies of five natural population samples. The population relatedness composition is summarized as the variance in relatedness of all pairs. Values are means of 500 independent draws of the population relatedness compositions from 10000 simulated pairs for each genetic relationship.

	Meerkat	**Great reed warbler**	**Bighorn sheep**	**Red deer**	**Soay sheep**
Variance in relatedness	0.0106	0.0044	0.0015	0.0005	0.0004
Number of loci	8	15	30	8	19
Variance explained					
QG	0.176	0.172	0.086	0.015	0.015
LR	0.269	0.328	0.14	0.024	0.028
W	0.161	0.167	0.074	0.013	0.013

Some of the differences in the r^2 among populations may also be contributed to the variation in the number of loci or the quality of the marker data. For example, the great reed warbler and meerkat populations have the most polymorphic markers. I found that in the great reed warbler and Soay sheep populations, using more loci increases the proportion of variance explained, which is expected since with more loci the sampling variance of the marker-based relatedness estimates decreases (Figure 2.3). However, there is a striking difference between the two populations; the r^2 in the great reed warbler population is nine to 23 times larger than in the Soay sheep population for the studied range of marker number. Figure 2.3 also shows that the average effect of the number of markers is greater in the great reed warbler population, where adding one locus elevates the r^2 by 0.0113 on average, while the equivalent figure is only 0.0014 in the Soay sheep population. Also, in the great reed warbler population there is more variation in the values of r^2 for any given number of loci, which probably reflects the fact that there is more variation in the level of polymorphism in the great reed warbler marker set. Results on are shown only for the LR estimator (Figure 2.3), but the other estimators revealed very similar differences between the two populations (results not shown).

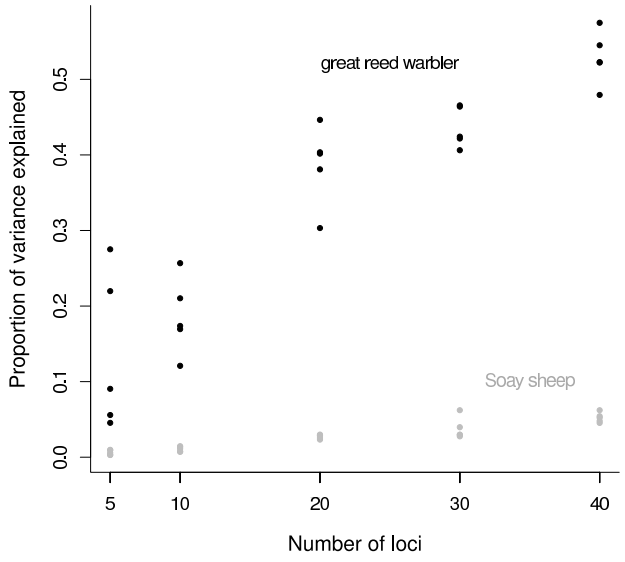

Figure 2.3: Proportion of variance explained in the marker-based relatedness estimates by true relatedness as a function of the number of loci. Populations were simulated based on the observed relatedness composition and allele frequencies of five natural population samples. For each number of loci five different loci were drawn from the available set of markers, which consisted of 62 loci for the great reed warbler and 101 for the Soay sheep. Relatedness was estimated using the LR estimator.

The increased level of polymorphism, expressed as the mean number of alleles,

in samples of five loci also had a positive effect on the proportion of variance explained (Figure 2.4). Note that results are similar using other measures of marker polymorphism, e.g. mean polymorphism information content or heterozygosity (results not shown). Since there is more variation in the great reed warbler markers, the two populations can be compared only at the lower end of the polymorphism scale. Again the Soay sheep population has a much lower r^2 at all studied polymorphism levels. Increasing the average number of alleles by one increases the r^2 in the great reed warbler population by 0.0086 on average, but by only 0.0012 in the Soay sheep population.

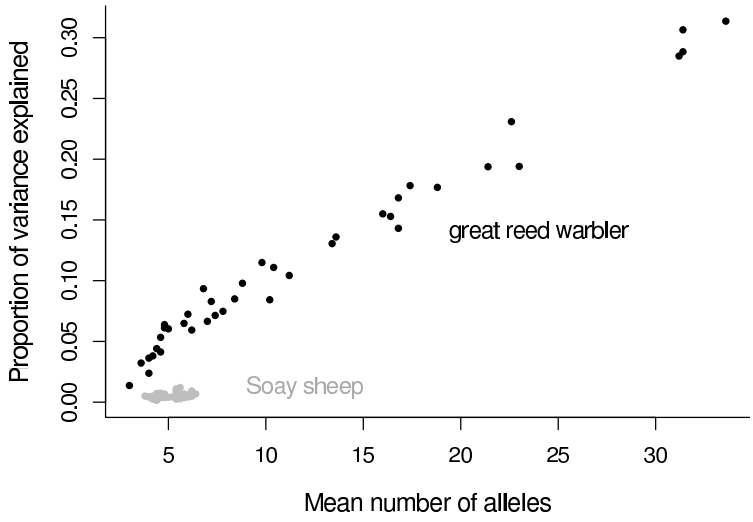

Figure 2.4: Proportion of variance explained in the marker-based relatedness estimates by true relatedness as a function of the level of polymorphism expressed as the mean number of alleles at five randomly selected loci. Populations were simulated based on the observed relatedness composition and allele frequencies of five natural population samples. Loci were drawn from the available set of markers, which consisted of 62 loci for the great reed warbler and 101 for the Soay sheep. Relatedness was estimated using the LR estimator.

2.4 Discussion

2.4.1 Relatedness composition of natural populations

My study demonstrates that in a range of natural populations of vertebrate species the population relatedness composition is different from what has been assumed in previous simulation studies and as a result the average performance of marker-based relatedness estimators, defined as the performance across all possible pairs in a sample,

is considerably lower than has been previously predicted (*e.g.* VAN DE CASTEELE *et al.*, 2001; RUSSELLO and AMATO, 2004). I estimated the population relatedness composition from four to six generation deep pedigrees established by long term studies of five species. Analysis of the population relatedness composition reveals that over 90% of the pairs have relatedness less than 0.25, and thus would be classified as unrelated using a shallow pedigree, while the proportion of pairs with relatedness of at least 0.5 (e.g. full-sib or parent-offspring pairs) was almost negligible, ranging from 0.1% to 1.7%, as opposed to the 20-50% assumed in previous simulation studies (*e.g.* VAN DE CASTEELE *et al.*, 2001). Deep pedigrees have also recovered many higher than first and second order relationships, the presence of which may have a non-negligible effect on the average performance of the relatedness estimators, in contrast to previous studies that assumed a simple population composition of unrelated, half-sib, full-sib, and parent-offspring pairs (*e.g.* BLOUIN *et al.*, 1996).

The five species represent mating systems ranging from near monogamy to highly skewed polygyny. Thus, I argue that the examples of relatedness composition are close to what one would find in many natural populations of vertebrates. Further, if we summarize the population relatedness composition as the variance in relatedness across all pairs, a comparison can be made with a study of monkey-flowers (*Mimulus guttatus*) where the variance in relatedness lies within the range of 0.0025 and 0.01 (RITLAND and RITLAND, 1996), which closely resembles my estimates (0.0004 - 0.0106). Unfortunately, other examples are scarce in the literature, perhaps because there was no interest in estimating this population parameter (RITLAND, 1996b).

2.4.2 Average performance of relatedness estimators

Here, I investigated the consequences of the observed population relatedness composition on the average performance of relatedness estimators by two methods. First, I used the misclassification rate between two relationships described by BLOUIN *et al.* (1996) and, second, I used the variance explained in the marker-based estimates by true relatedness originally suggested by VAN DE CASTEELE *et al.* (2001).

The complexity of the relatedness composition of natural populations could lead to misleading estimates of the misclassification rates between the common first and second order relationships. This is because the actual sampling distributions of the first and second order relationships are skewed to the right, towards higher relationships, and thus both their sampling variances and means are underestimated. As a result, the optimal cutoff point to classify pairs to different relationships may not be the midpoint between the empirically determined means as has been suggested by BLOUIN *et al.* (1996), but shifted towards the higher relationships to an extent that is itself dependent

on the unknown population relatedness composition. Although this effect will be negligible in some populations (e.g. the bighorn sheep), in others it will not be (e.g. the meerkats). I would argue that it is generally important to be aware that when one selects highly related pairs from a sample based on BLOUIN *et al.*'s (1996) method, it is likely that the sample will be diluted with more unrelated pairs than expected. In contrast, we can more confidently select unrelated pairs because the pre-determined error rates will be conservative. These findings are relevant to applications where the aim is to classify pairs to simple first or second order relationships or just to simply select "unrelated" or "related" pairs, e.g. when selecting founders for captive breeding. Unfortunately, endangered species that are selected for captive management are often inbred, and the bias of the pre-determined misclassification rates are expected to be magnified in such scenarios. Finally, I would also generally argue that using the pairwise relatedness estimates as a categorical measure should not only be avoided because the misclassification rate depends on the population relatedness composition, but also the estimators themselves are inherently not categorical measures. To distinguish between simple first order relationships, likelihood ratio tests are available and should be preferred (*e.g.* GOODNIGHT and QUELLER, 1999).

The effect of the radical difference between the observed population relatedness composition and what has previously been assumed is more pronounced on VAN DE CASTEELE *et al.*'s (2001) r^2 measure, the proportion of variance explained in the marker-based relatedness estimates by true relatedness. This effect is driven by the fact that the low variance in relatedness results in a generally low r^2. Since increasing the number of markers and/or choosing highly polymorphic markers decreases the sampling variance of relatedness estimates, r^2 can be improved, but I have shown that, even with a hypothetical set of 45 independent, polymorphic microsatellite loci for the Soay sheep population, in which the variance in relatedness is an order of magnitude smaller than in great reed warbler population, the r^2 is on average 10 times smaller. Thus, I suggest that the population relatedness composition sets a limit to the proportion of variance explained in the marker-based relatedness estimates and thus, the average performance may only be improved within the limits of the population relatedness composition.

Knowledge of the proportion of variance explained by true relatedness is essential when pairwise relatedness estimates are used in subsequent analysis as an explanatory variable. When the variance in relatedness is low, as it is expected to be in most natural populations, applications that require the use of relatedness estimates as an explanatory variable will not have sufficient power. This fact closely mirrors recent studies pointing out that there is not sufficient power to detect inbreeding depression in the wild using

marker heterozygosity when the variance in inbreeding is low (SLATE *et al.*, 2004; BALLOUX *et al.*, 2004). As a consequence, I argue that some studies might have falsely rejected hypotheses regarding the effect of relatedness. For example, in mate choice experiment with 46 female sticklebacks (*Gasterosteus aculeatus*) REUSCH *et al.* (2001) claimed to exclude the possibility that preferred males were less related to the females than unpreferred males, on the basis of non significant correlation between preference time and pairwise relatedness estimated using seven microsatellite markers (REUSCH *et al.*, 2001). As another example, a study found that eider ducks (*Somateria mollissima*) form non-kin brood-rearing coalitions, and thus rejected the kin selection hypothesis on the basis of comparing the relatedness estimates of 24 pairs of brood-rearing females with 24 randomly drawn pairs of females using 6-8 microsatellite markers (ÖST *et al.*, 2005).

Pairwise relatedness estimates across all pairs in a sample are again used as the independent variable when applying RITLAND's (1996b) method, which has been developed to estimate quantitative genetic parameters in natural populations. The method was published over 10 years ago, but remarkably few applications have appeared in the literature since, and apparently most of them are using one of the data sets presented in this Chapter. As an example, studies comparing heritability estimates based on RITLAND's (1996b) method and traditional pedigree-based methods found that marker-based estimates erratically either under or overestimate the pedigree-based estimates of heritability (*e.g.* THOMAS *et al.*, 2002; WILSON *et al.*, 2003; COLTMAN, 2005), for which the low variance in relatedness and the inaccurate estimate of it should at least partly be responsible. I suspect that there are many more unpublished heritability estimates, which are also out of range, and thus biologically not valid.

2.4.3 Improving the average performance

If the average performance of the relatedness estimators in natural populations is generally expected to be low, the question arises how to improve it or what alternative methods are available. Here I discuss the choice of relatedness estimator, the importance of marker data quality, the potential choice of study population and/or organism, and the combined use of marker data and pedigrees.

The most frequently addressed question is undoubtedly the choice of estimator (VAN DE CASTEELE *et al.*, 2001). Given the observed population relatedness compositions my study uniformly supports the use of the LR estimator. In all five populations the sampling variance is the lowest when using the LR (and R, data not shown) estimator for the low relationship categories, and since in all studied populations most pairs have relatedness less than 0.25, on average, across all pairs the

LR estimator minimizes the sampling variance. Thus, in terms of average performance the Lynch and Ritland estimator is recommended on the basis of these five populations and some other published studies as well (RUSSELLO and AMATO, 2004; COLTMAN, 2005). In contrast, the most recent simulation study of the moment estimators demonstrated that the single locus sampling variances for the QG, L, and W estimators asymptotically approach the minimum sampling variances (variance in identity-by-descent) with increasing number of alleles, while the R and LR estimators do not have this property (WANG, 2002), because the latter two estimators assume zero relatedness when calculating weights. WANG (2002) also pointed out that LR and R estimators are sensitive to sample sizes both in terms of variance and bias. In summary, however, when we look at the differences between estimators in terms of e.g. the r^2, they are almost negligible in relation to the differences between the populations, suggesting that the estimator choice may not be the most crucial question.

The utmost importance of good quality marker data has been emphasized by many. For example, WANG (2002) concluded that estimators asymptotically approach the minimum sampling variances with increasing number of alleles, or RITLAND (1996b) suggested that, when estimating heritability in unstructured populations, where the performance of his method is poorer, more polymorphic markers are needed. In contrast, I argue that only if the population has a high variance in relatedness acquiring more markers and more polymorphic markers deliver substantial improvements in the average performance of the relatedness estimators. In such cases, it is worthwhile to type as many markers as possible in order to achieve the best possible performance. The number of markers available, however, may well be limited by the number of chromosomes, because some of the markers will unavoidably be linked. I argue that the use of linked markers could nevertheless be useful. Relatedness estimators would lose only efficiency (i.e. have a higher variance) when applied to linked loci relative to the same number of unlinked loci, depending on the recombination rate between loci (THOMPSON and MEAGHER, 1998). This is because linked markers simply carry less information about identity by descent (THOMPSON, 1986). Regarding the level of polymorphism, over which we have even less control, this is generally specific to the species; e.g. mammals were found to have less polymorphic markers than birds in a comparative study using AC repeats (NEFF and GROSS, 2001).

Recognizing that the population relatedness composition plays the major role in the average performance of the estimators, one may choose to address questions that require the knowledge of relatedness in study organisms where the expected estimator performance is high. RITLAND (1996b) has also suggested selecting study population or taxa where the variance in relatedness is expected to be high and

described two potentially favorable situations (RITLAND, 2000). One of them is where one polygynous male is breeding within a lineage of philopatric females, a common breeding system in many mammalian social systems. Three of my populations of ungulates exhibit this mating system, but, in contrast, my results show that the variance in relatedness is rather low. RITLAND's (2000) other example is newly founded populations, with a small number of related founders. The fact that the great reed warbler population was recently founded by a few individuals might have played a role in the fact that this population has a relatively high variance in relatedness, but perhaps it is more likely to be the result of the mating system of the species. On the basis of the five populations studied, I rather recommend using information on the mating system of the species to predict the population relatedness composition, and prefer species with large full-sib families. More specifically, one may want to consider using monogamous birds with a large clutch size.

Traditional, pedigree-based methods supported by marker aided parentage inference, where required, offer a good alternative in many applications, for example, when the aim is to classify pairs to different relationships or to estimate quantitative genetic parameters in natural populations (KRUUK, 2004; THOMAS, 2005). However, when relatedness estimates are used as an explanatory variable and the variance in relatedness is low in the study population, even knowledge of the pedigree cannot directly help. In such cases pedigrees may be employed to aid the marker-based relatedness estimation. Even partial or shallow pedigrees could be used to selectively sample highly related pairs or families and, thus artificially generate a population with more favorable population relatedness composition. When only shallow pedigrees are available this is perhaps the preferred method, because marker-based estimates could potentially be more accurate than categorical measures of relatedness, assuming good marker data. This is because marker-based methods estimate the actual relatedness between two individuals, which is the realized relationship and not the mere expectation that the categorical measures estimates of relatedness, like pedigrees, provide (THOMAS, 2005).

Chapter 3

On the choice of an appropriate null hypothesis when testing for linkage disequilibrium in finite population samples

The material of this Chapter closely resembles a manuscript that has previously been submitted to Genetics with my co-author, Toby Johnson. However, it has not been published as of today. I declare that the data analysis and the simulation study were performed by me, we have equally contributed to the ideas presented, and I wrote the manuscript. The ideas presented in this Chapter have been improved by comments from Nick Barton, Bill Hill, Arnaud Estoup, François Rousset, and Kevin Dawson, mainly as comments on the above mentioned manuscript.

3.1 Introduction

Linkage disequilibrium (LD) is a non-random association between alleles at two or more different loci, in a given population. Estimating LD has been a topic of longstanding interest in evolutionary genetics, because many biological processes can play a role in shaping the LD patterns, including mutation, recombination, selection, drift and demography. For example, patterns of LD may reflect variation in the recombination rates and the presence of recombination hot spots (*e.g.* MYERS *et al.*, 2005), a recent selective sweep (*e.g.* KIM and NIELSEN, 2004), interaction between drift and directional selection (HILL and ROBERTSON, 1966), epistatic selection (*e.g.* FELDMAN *et al.*, 1980), or recent population admixture (*e.g.* PFAFF

et al., 2001; FALUSH *et al.*, 2003). The ultimate aim of studying LD patterns is almost always understanding the underlying biological processes, and not estimating the population LD per se. However, extracting the relevant information about the biological parameters from LD data poses a serious statistical challenge (*e.g.* MCVEAN, 2007), partly because many sample histories are compatible with any given sample LD pattern, and partly because the same LD patterns could indicate different biological processes. For example, REED and TISHKOFF (2005) report that selective sweeps might have been falsely detected as recombination hotspots. To date, only a few statistical approaches have been developed that are directly aimed at inferring biological parameters from LD data, for example, the recombination rates as implemented in the software LDhat (MCVEAN *et al.*, 2004; MYERS *et al.*, 2005) or demographic parameters as implemented in the software Structure (FALUSH *et al.*, 2003) or Geneland (GUILLOT *et al.*, 2005b).

Testing the null hypothesis of zero LD, i.e. testing for independence between the rows and columns of a contingency table of haplotype counts, provides a simple and attractive approach. This is because, seemingly, zero LD corresponds to the null model of many underlying biological processes. Thus, the null hypothesis of zero LD can be tested to address many different biological questions (albeit using different test statistics). In this Chapter, however, I argue that zero LD is not a biologically meaningful null hypothesis against any alternatives (e.g. linkage, admixture, epistatic selection etc.) because exactly zero LD can only arise with any appreciable probability in infinite populations. In finite populations, correlations continually arise between alleles of different loci, due to genetic drift. These correlations are broken down over time by recombination, but there will always be some low, but non-zero correlation, which I call "background LD". The fallacy of the "zero LD test" may be reflected in some recent empirical studies that found high proportions of genetically independent loci pairs in significant LD (SLATE and PEMBERTON, 2007; MCRAE *et al.*, 2002; FRANIR *et al.*, 2000).

A biologically meaningful null hypothesis is that of LD between genetically independent (i.e. freely recombining) loci in a finite panmictic population (with no mutation, selection etc.), thus a null that accounts for the "background LD". Unfortunately, testing this biologically meaningful null hypothesis in a satisfactory way is extremely difficult, while the readily available alternative, the statistical test of independence is straightforward to carry out. Briefly, this is because the so-called nuisance parameters, which are parameters not of direct interest, but which nonetheless must be taken into account, are difficult to estimate in the former case. Thus, the very practical question arises of whether the convenient statistical (infinite population)

null hypothesis is an acceptable approximation to the biologically meaningful (finite population) null hypothesis. I will illustrate the differences between testing the infinite and finite population nulls against the alternative hypothesis of genetic linkage. I choose this alternative mainly for convenience: the degree of linkage can be easily manipulated in simulations by varying a single scalar, the recombination rate. I emphasize, however, that restricting my attention to a single alternative scenario does not affect the generality of the results with respect to exposition of a more general problem.

There is reason to believe that how well the biological null is approximated by the statistical null depends on the informativeness of the data. It is well known that with sufficiently large sample sizes and sufficiently informative data, a null hypothesis that is not exactly true has a high probability of being rejected (the "paradox of the large n", SPROTT (2000)[p. 98]). Thus, we might expect that in large samples and when using highly polymorphic markers, e.g. microsatellites, the statistical null hypothesis, which is never exactly true, will have a high probability of being rejected. I will show evidence for the counter-intuitive result that one is more likely to have statistically significant but biologically irrelevant results with more informative samples.

Valid comparison of the infinite and finite population tests from less to more informative samples, and especially along the polymorphism scale from biallelic to multiallelic markers is, however, not straightforward to carry out. This is for reasons related to the change in the size and sparseness of the contingency table of haplotype counts. While in 2×2 tables (i.e. when using biallelic markers) there could be deviation from linkage equilibrium (LE) only in one direction, in large tables there could be departure from LE in many different directions, or, statistically speaking, the alternative hypotheses have many additional degrees of freedom. Thus, it is often impossible to determine the direction of the departure based on a single statistic (SABATTI and RISCH, 2002), and different statistics could well measure different aspects of the departure from independence. Thus, I will consider a number of different statistics and apply a Monte Carlo permutation method, which is suitable to empirically determine the null distribution of any given test statistic. Also, the Monte Carlo permutation does not rely on conventional χ^2 approximations, which may not be valid for the large and sparse tables of haplotype counts when using multiallelic markers (SHAM and CURTIS, 1995).

3.2 Test statistics for LD

Consider two loci, A and B, and let n and m be the numbers of alleles observed at each locus. Let p_i be the estimated frequency of the A_i allele at the first locus, where $i = 1, ..., n$, and let q_j be the estimated frequency of the B_j allele at the second locus, where $j = 1, ..., m$. Let h_{ij} be the estimated frequency of haplotype A_iB_j. Here, I consider only the situation when haplotypes have been observed or inferred. The standard coefficient of LD for any pairs of alleles, i and j is then,

$$D_{ij} = h_{ij} - p_i q_j \ .$$

General statistics that are applicable for arbitrary numbers of alleles can be defined as weighted averages of biallelic statistics, with weights specific to each pair of alleles, i and j.

The commonly used χ^2 test statistic can be written, in my notation, as,

$$\chi^2_{raw} = 2N \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{D_{ij}^2}{p_i q_j} \ ,$$

To standardize the χ^2 measure to the $[0, 1]$ range we want to divide χ^2 with its upper bound given the allele frequencies. ZHAO *et al.* (2005) found that the best measure (i.e. using the best approximation to the upper bound) was standardized as follows:

$$\chi^2 = \frac{\chi^2_{raw}}{2N(l-1)},$$

where $l = min(m, n)$. An alternative measure can be constructed by standardizing by the degrees of freedom,

$$\chi^2_{\mathrm{df}} = \frac{\chi^2}{2N(m-1)(n-1)},$$

(HEDRICK, 1987). I found that χ^2_{df} performs similarly to the χ^2, thus I do not show results separately for this statistic.

A related statistic, which is less commonly used, is the likelihood ratio test statistic, defined as,

$$LRT = 4N \sum_i \sum_j h_{ij} \ln \left(\frac{h_{ij}}{p_i q_j} \right) .$$

The most commonly used measure for multiallelic markers is the multiallelic D'

developed by HEDRICK (1987), which can be defined as,

$$D' = \sum_i \sum_j p_i q_j |D'_{ij}|,$$

which is a weighted average of individual D'_{ij} measures, where

$$D'_{ij} = D_{ij}/D_{max},$$

and

$$D_{max} = \begin{cases} \min(p_i q_j, (1-p_i)(1-q_j)), & \text{when } D_{ij} < 0 \\ \min(p_i(1-q_j), (1-p_i)q_j), & \text{when } D_{ij} > 0 \end{cases}.$$

Weighting by the observed haplotype frequencies, h_{ij}, gives an alternative version of D' (KARLIN and PIAZZA, 1981),

$$D'_{\mathrm{h}} = \sum_i \sum_j h_{ij} D'_{ij}.$$

A multiallelic r^2 can be constructed, using a generalization of the definition of multiallelic D'. The r^2 statistic for the allele pair A_i and B_j is

$${r_{ij}}^2 = \frac{(h_{ij} - p_i q_j)^2}{p_i(1-p_i)q_j(1-q_j)},$$

and weighted average of all ${r_{ij}}^2$ can be taken, using weights w_{ij}. ZHAO *et al.* (2007) suggest weighting by either the expected haplotype frequencies, $w_{ij} = p_i q_j$,

$$r^2 = \sum_i \sum_j p_i q_j {r_{ij}}^2,$$

or by the observed haplotype frequencies, $w_{ij} = h_{ij}$,

$$r^2_{\mathrm{h}} = \sum_i \sum_j h_{ij} {r_{ij}}^2.$$

Since I found that both for D' and r^2 the different weighting schemes make little difference in performance, I will present results for D' and r^2 only (i.e. when weighting with the expected haplotype frequencies).

NOTHNAGEL *et al.* (2002) proposed a multilocus measure of LD based on the concept of entropy. Motivated by this, I define a multiallelic measure of LD, denoted

here as NE,

$$NE = 1 - \frac{-\sum_{ij} h_{ij} \ln(h_{ij})}{-\sum_{ij} p_i q_j \ln(p_i q_j)}.$$

This measure compares the randomness of the observed haplotype distribution (in the numerator), to the randomness of a hypothetical haplotype distribution with the same allele frequencies but with zero LD (in the denominator). Note that missing haplotypes do not contribute to the numerator.

Based on entropy, it is possible to derive a measure (using the concept of mutual information, SOOFI (1994)) that is directly proportional to the LRT statistic, thus has identical behavior to it. In fact, I also found that NE has nearly identical behavior to the LRT test statistic, thus I will not present results separately for the two measures.

Lastly, SABATTI and RISCH (2002) and CHEN *et al.* (2006) proposed different so-called volume measures of multiallelic LD and implemented an importance sampling scheme to evaluate the measures on large contingency tables. I was unable to consider these measures because, although for an individual data set they are computationally feasible (CHEN *et al.*, 2006), in the scale of my simulations they are not.

3.3 Sampling distributions for testing null hypothesis

Testing the null hypothesis of zero LD is technically straightforward. The null hypothesis of exactly zero LD implies independence between rows and columns in a contingency table, and thus the null distribution of any test statistic can be conveniently simulated via Monte Carlo permutation. This null distribution will be conditional on the sample allele frequencies, and thus I shall refer to this null distribution as the "conditional infinite population" null distribution. Note, that in the terms of the coalescent with recombination, this null is equivalent to testing if the scaled recombination rate, $\rho = 4N_e c$ equals ∞.

In contrast, testing the biologically meaningful finite population null hypothesis is technically extremely difficult. We want to test if two loci are genetically independent in a finite population, which involves testing if $\rho = 4N_e \times 0.5$. Under this model we have nuisance parameters, which are N_e, and, for each locus, a mutation rate μ and a mutation model for the genetic marker in use. The conventional statistical way to simulate the distribution of a test statistic under this null would be to condition on sufficient statistics for the nuisance parameters. However, I am not aware of any computationally feasible method to simulate data conditioned on these nuisance parameters.

An alternative, but less sophisticated, method to deal with nuisance parameters is to

simulate the finite population null distribution using estimated values for the nuisance parameters. This is equivalent to performing a goodness of fit test: I perform the test of genetic independence by plugging in the estimated values of the parameters. The difference from a "fully conditional" approach, even when the nuisance parameters, N_e and μ, are perfectly estimated, arises because the null distribution will contain samples that do not have exactly the same allele frequencies as the observed sample. This is because samples cannot be generated to fit pre-specified allele frequencies using any population genetic model, whether forward- or backward-time simulations. Nevertheless, it is at least possible to simulate the null distribution of a test statistic under genetic independence in a finite population. Since such a null distribution will be unconditional on the data, I will refer to it as "unconditional finite population" null distribution.

Test data sets, which were loci pairs under the alternative hypotheses (genetically linked loci) were simulated using the standard coalescent with recombination (HUDSON, 1983), as implemented in the software Simcoal2 (freely available at `http://cmpg.unibe.ch/software/simcoal2`). I used a range of recombination rates, $c = \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and different effective population sizes $N_e = \{10^2, 10^3, 10^4\}$, which resulted in a wide range of values for $\rho = 4N_e c$, the scaled recombination rate. I defined the parameters in terms of N_e and c rather than ρ because a value of ρ corresponding to the biological null depends on N_e. I ran 1000 replicate simulations for each parameter combination.

I simulated two types of markers, microsatellites and SNPs. Microsatellite markers were simulated using the generalized step-wise mutation model, which allows the insertion and deletion of more than one repeat unit. The number of repeat units had a Geometric distribution with a variance of 0.36 (ESTOUP *et al.*, 2001), and a range constraint of 70. I used a mutation rate of $\mu = 10^{-3}$, which is in the range of our current best estimates for microsatellites (ESTOUP *et al.*, 2002). Under different values of N_e this mutation rate resulted in a variety of polymorphism levels measured as allele numbers (Table 3.1). Polymorphism levels even higher than around 25 alleles (i.e. under the most extreme θ in Table 3.1) are not uncommon in natural populations, for example, in birds: up to 94 alleles were observed in barn swallows (BROHEDE *et al.*, 2002), up to 75 alleles in great reed warblers (*e.g.* CSILLÉRY *et al.*, 2006) or up to 71 alleles for the Principe seedeaters (MELO and HANSSON, 2006). Thus we considered $\theta = 400$ as well, which resulted in a median of 55.5 alleles (with 44 and 64 as the 5% and 95% percentiles, respectively) and I will mention some relevant results for this extreme end of the polymorphism scale. Simulations that resulted in monomorphic loci were excluded (these occurred with small N_e) and also cases when there were two

or three alleles present at a locus and the rare allele(s) appeared only in one individual. SNPs were modelled as a DNA base pair with random mutations along the genealogy of the locus with mutation rate 10^{-8} (HAAG-LIAUTARD *et al.*, 2007).

The "conditional infinite population" null distribution was determined for each pair of loci simulated under alternative hypothesis via Monte Carlo permutation. A sample of 400 haplotypes was taken from each population and 1000 permutations were used to determine the null distribution. The "unconditional finite population" null was simulated for each loci pair under the alternative hypothesis using the coalescent with recombination rate, $c = 0.5$, with known N_{e} and mutation model and parameters. In order to investigate the effect of the lack of fully conditioning on the data, I simulated another unconditional null distribution (which is analogous to the Monte Carlo permutation null) using independent coalescent models at each locus. This null distribution corresponds to $\rho = \infty$, thus I call this the "unconditional infinite population" null.

My first aim is to demonstrate the consequences of using an inadequate infinite population null distribution generated by Monte Carlo permutation. I cannot do this directly, since the analogous "conditional finite population" null distribution is unknown. I can, however, quantify the consequences of using an infinite population null with the actual false positive rates, (hereafter actual FPRs). Actual FPRs were defined as the proportion of times when the null hypothesis was rejected using the Monte Carlo permutation, when the biological null ("unconditional finite population") was in fact true. If the infinite population null is a good approximation to the finite population null, the actual FPR is expected to be the same as the nominal FPR, which I set to be 0.05.

My second aim is to show how data informativeness, i.e. level of polymorphism and sample size, influences the difference between the finite and infinite population nulls, and the power. Table 3.1 shows a summary of the different levels of polymorphism attained under different values of θ. As discussed above, I believe that these polymorphism levels are representative of microsatellite markers found in many natural populations (*e.g.* CSILLÉRY *et al.*, 2006; ELLEGREN, 2000b). As for sample size, I considered $n = 100$, 200, 500, 1000, and 2000.

My third aim is to compare the power different test statistics under different alternative hypotheses (i.e. for different degrees of genetic linkage). First, I compared the power of the test statistics in a "naive" way, by calculating the power using the p-values from the Monte Carlo permutation. In each simulation, p-values were calculated as $(1 + r)/(1 + n)$, where n is the number of permutations and r is the number of cases when the test statistic is greater than or equal to that in the original sample

Table 3.1: The median and the 5 and 95% percentiles of the number of alleles under different values of N_e and mutation rate of 10^{-3} in my simulations for microsatellite markers, using the GSM mutation model (details in the text).

N_e	**Number of alleles**		
	5%	50%	95%
	Percentiles		
100	2	3	4.5
1000	6.5	9	11.5
10000	20	25.5	32.5

(NORTH *et al.*, 2002). Then, 1000 replicate simulations were used to calculate the power of the test. For a nominal FPR of 0.05, the power of a test is the proportion of simulations where the p-value was at least as low as 0.05. However, these power comparisons may not be valid, because the actual FPRs could well be different for different test statistics and sizes of the contingency table (i.e. polymorphism levels). It is only meaningful to compare the power of different tests with fixed actual FPRs. Thus, second, I calculated power by fixing the actual FPRs to be 0.05 by using the 95% percentile of the unconditional finite population as a critical value.

3.4 Simulation results

I considered three different null distributions: the unconditional infinite and finite population nulls, and the Monte Carlo permutation null, which is conditional on the sample allele frequencies. The three null distributions were identical for SNPs, across all different values of N_e (and, thus θ), indicating that the unconditional infinite population null was a good approximation to the null distribution that would be observed in a finite population (Figure 3.1). This result was consistent across different test statistics (results not shown). When $N_e = 100$, LD was higher, on average, under the finite than under the infinite population nulls, because the difference between ρ for independent loci ($\rho = \infty$) and for loci with recombination rate 0.5 becomes more extreme as N_e becomes smaller. This is because as N_e gets smaller ρ becomes smaller as well, and thus more different from ∞. The differences in the tails of the distributions for larger N_e values are not consistent between the three null scenarios (as shown in Figure 3.1 and 3.2, i.e. finite and infinite population, and Monte Carlo nulls), and are probably due to Monte Carlo error. In contrast, when LD was measured between polymorphic microsatellite loci, the finite population null distribution was shifted towards higher LD values on average, indicating that there was an excess of LD

between freely recombining loci in a finite population, in relation to that in an infinite population (Figure 3.2). The difference between the unconditional infinite and finite population null distributions increased as the data became more informative: when loci were more polymorphic (i.e. when N_{e} was larger, Figure 3.2) and the sample size increased (Figure 3.3). Note that the pattern in Figure 3.2 stays the same for fixed N_{e} and different mutation rates as well (i.e. with more polymorphic loci the difference between the unconditional infinite and finite population null distributions increases; results not shown).

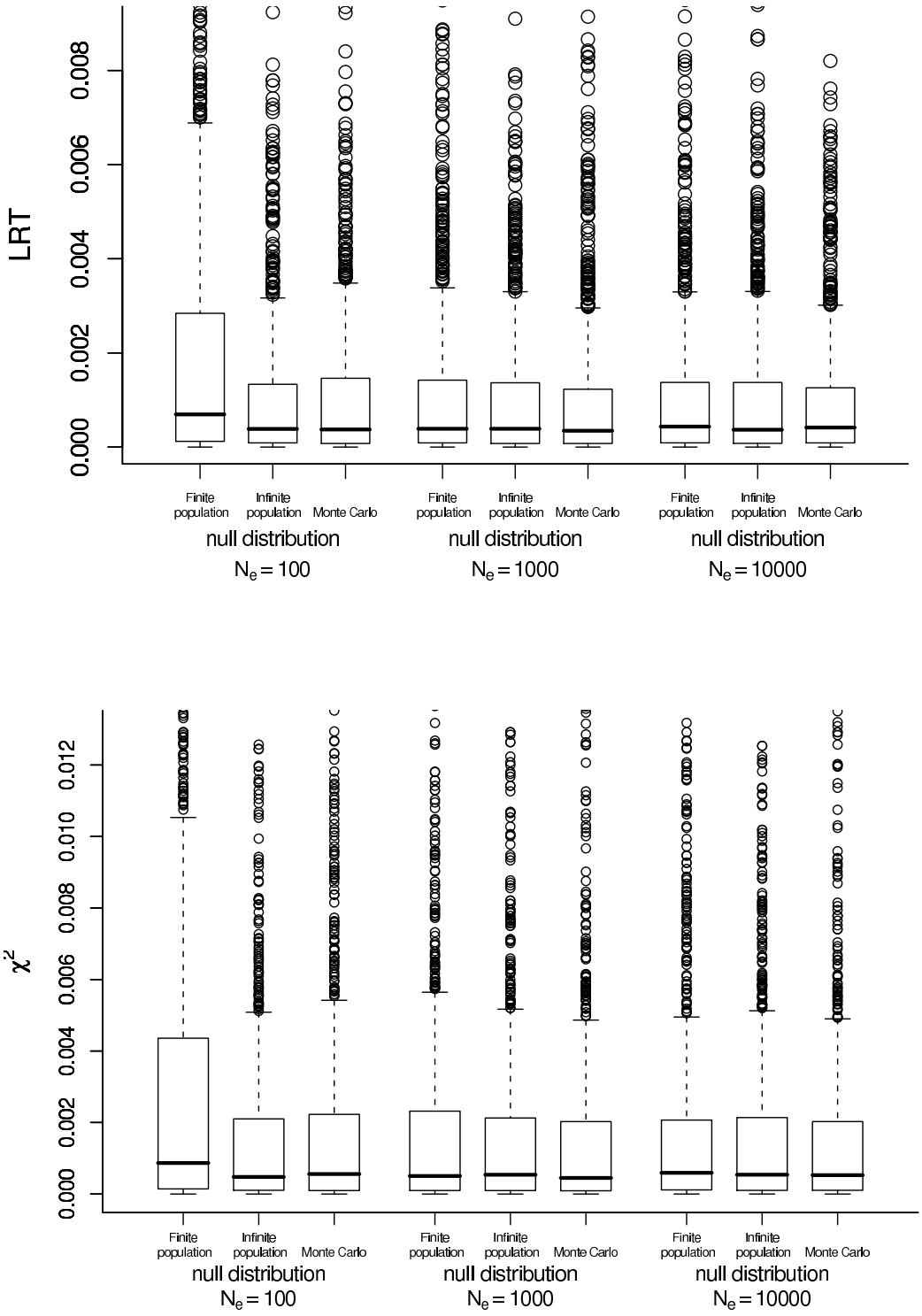

Figure 3.1: Comparison of LD under the three null distributions: the unconditional infinite and finite population and the Monte Carlo permutation null under three different values of N_{e} when using SNP loci. The Monte Carlo permutation null was generated by selecting one permuted sample from each replicate coalescent simulation, generated under a mixture of alternative scenarios. Results are shown for the likelihood ratio (LRT) and χ^2 test statistics.

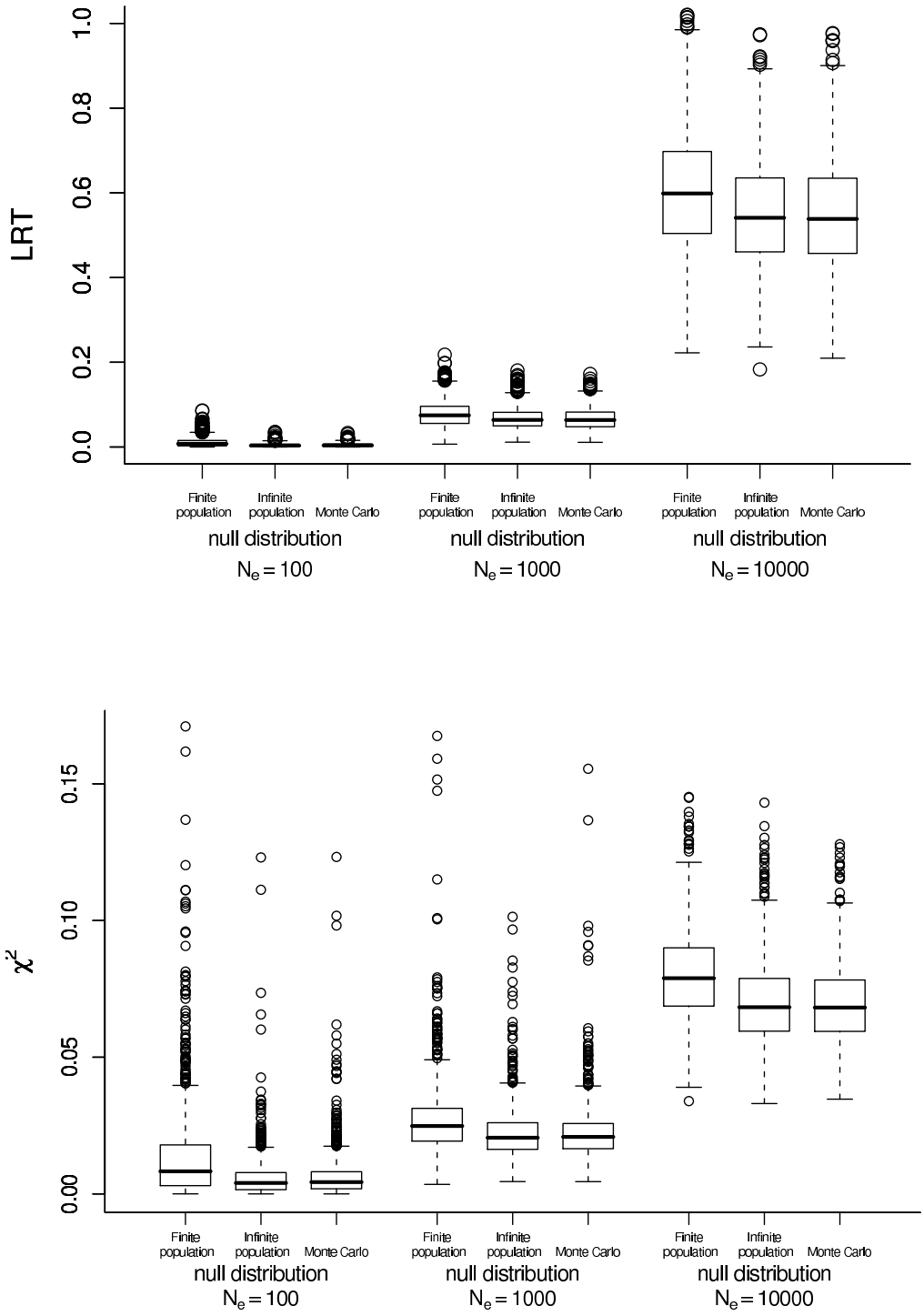

Figure 3.2: Comparison of LD under the three null distributions: the unconditional infinite and finite population and the Monte Carlo permutation null for different values of N_e when using microsatellite loci. The Monte Carlo permutation null was generated by selecting one permuted sample from each replicate coalescent simulation, generated under a mixture of alternative scenarios. Results are shown for the likelihood ratio (LRT) and χ^2 test statistics.

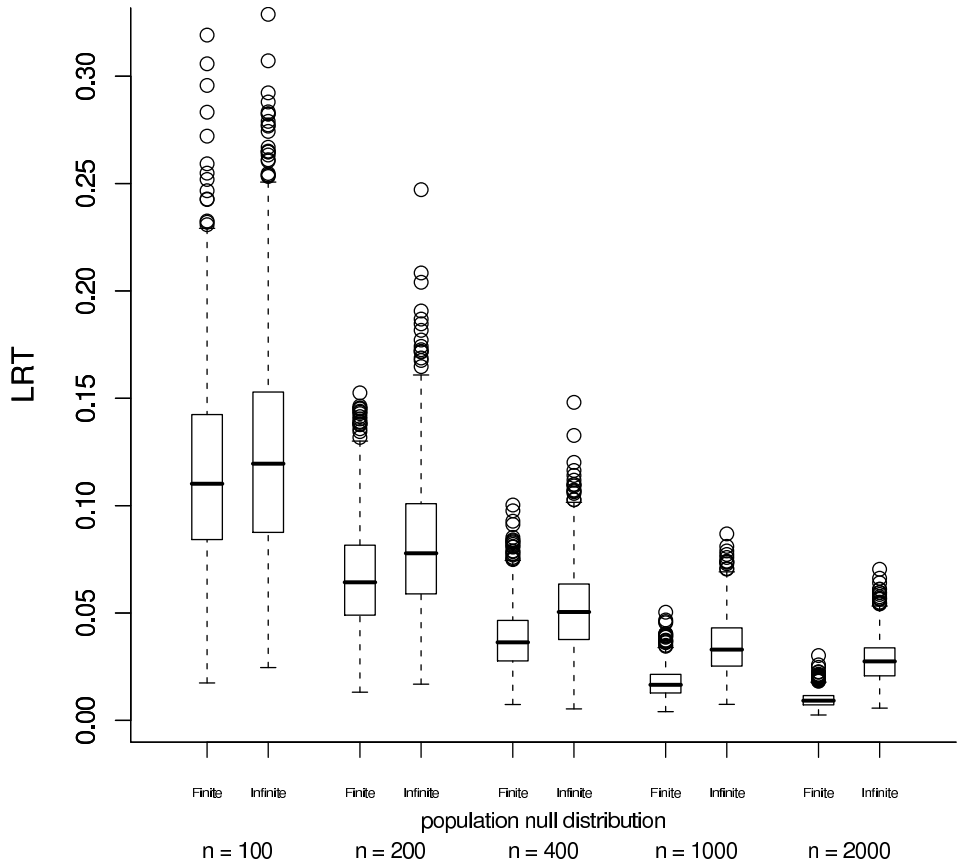

Figure 3.3: Comparison of LD under the two null distributions: the unconditional infinite and finite population for different sample sizes (n) when using microsatellite markers. Data are shown for mutation rate, $\mu = 10^{-3}$ and $N_{\mathrm{e}} = 1000$ and for the likelihood ratio test statistics (LRT).

The actual FPRs were also strongly dependent on the marker polymorphism. When using SNPs the actual FPRs were approximately equal to the nominal FPRs, i.e. 0.05 (Table 3.2) under both the finite and infinite null distributions. Thus, the effect of using the unrealistic infinite population null distribution was small. Results were consistent across all test statistics (data not shown). For polymorphic microsatellites the actual FPRs were also approximately 0.05 when the infinite population null was used (Table 3.3). Deviations are probably due to Monte Carlo error since the binomial confidence intervals always included 0.05 (Table 3.3). However, when the biological, finite population null was true the statistical (infinite population) null hypothesis was rejected much more than 5% of the time for microsatellite loci (Table 3.3). The magnitude of the actual FPRs strongly depended on the level of polymorphism: with increasing number of alleles the actual FPR also increased. For even higher levels of polymorphism, e.g. for $\theta = 400$, the actual FPRs are even higher: 0.935, 0.421,

0.744, and 0.907 for the LRT, χ^2, D', and r^2 test statistics, respectively. Different test statistics had different actual FPRs indicating that the choice of test statistic may not be negligible for microsatellites (Table 3.3). Almost always, the χ^2 had the lowest and the LRT had the highest actual FPRs (Table 3.3).

The binomial confidence interval for the actual FPRs when using SNPs often did not include 0.05 when using the infinite population null (Table 3.2). In fact, the actual FPRs were often lower than 0.05, indicating that tests would be conservative (Table 3.2), which is a well-known phenomenon for 2×2 contingency tables, and is explained by the extreme discreteness of the test statistics (*e.g.* MEHTA and HILTON, 1993). Figure 3.4 shows distribution of the p-values under the null hypothesis of statistical independence (i.e. the infinite population scenario) when using LRT test statistic for different numbers of alleles. Recall that when the null hypothesis is true the distribution of p-values is expected to be uniform on [0,1] for a continuous test statistic. Figure 3.4 illustrates that for SNPs, i.e. in 2×2 tables the distribution of the p-values has a spike at one, which gradually vanishes as the size of the contingency table increases (Figure 3.4).

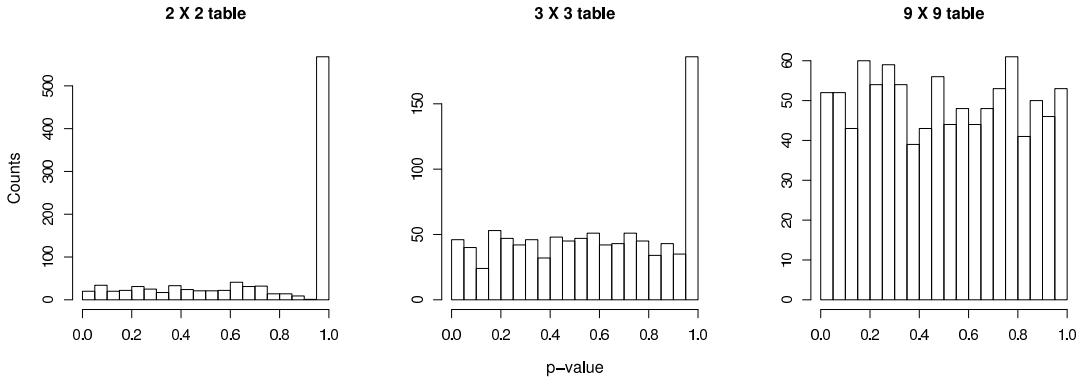

Figure 3.4: Distribution of the p-values under the unconditional infinite population null hypothesis for three different sizes of contingency tables, i.e. numbers of alleles. Table size 2 corresponds to SNP data and table sizes 3 and 9 correspond to microsatellite data, thus they are only approximate since the allele numbers very in the individual simulations. Results are shown for for mutation rate, $\mu = 10^{-8}$ for SNPs and $\mu = 10^{-3}$ for microsatellites and for $N_{\mathrm{e}} = 1000$ for both markers, and for the likelihood ration test (LTR).

The high actual FPRs led to inflation of power for highly polymorphic markers when using the Monte Carlo permutation test, i.e. the infinite population null distribution: for microsatellites, the apparent power ranged between 50 and 100% for loose linkage ($c = 10^{-1}$) and 100% or nearly 100% otherwise (Table 3.4). Thus, the null hypothesis of zero LD was almost always rejected for large numbers of alleles,

regardless of the degree of evidence against the null. Such a high power in large tables, however, is not a "useful" power since the actual FPRs were also much higher (Table 3.3). Regarding the differences between test statistics, unless all the tests had a power of one, the LRT always had the highest power and its power advantage increased with increasing c (i.e. looser linkage) and level of polymorphism.

In order to address which test statistic had the highest power, taking into account the actual FPRs, I compared the power of tests with fixed actual FPRs by using the 95% percentile of the "unconditional finite population" null distribution as the critical value. I found that the r^2 statistic had consistently the highest power, and that its power advantage was the most for $N_e = 10000$ (Figure 3.5). This result is consistent with the fact that the r^2 statistic had actual FPRs closest to the nominal FPRs (i.e. 0.05, Table 3.3).

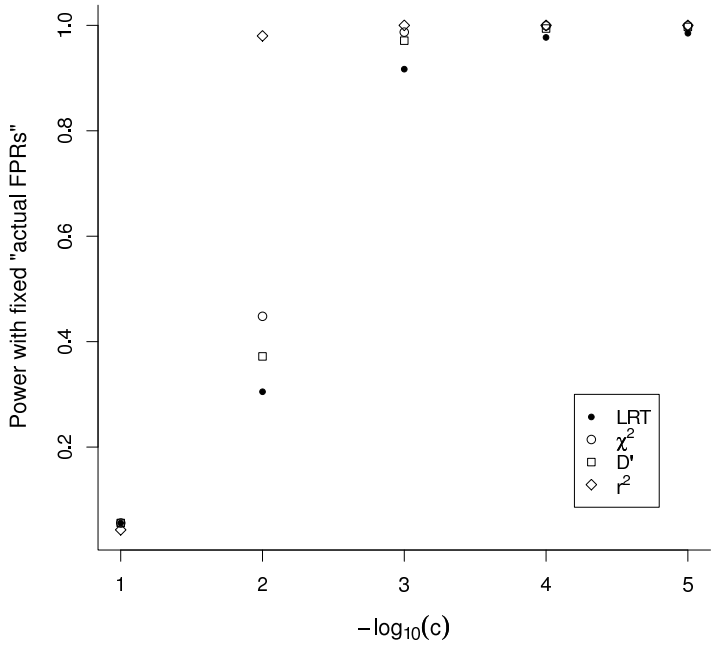

Figure 3.5: Comparison of the power of four test statistics with fixed actual FPRs. The critical value was the 95% percentile of the "unconditional finite population" null distribution. N_e was 10000 and $\mu = 10^{-3}$.

For completeness, I show the power of the Monte Carlo permutation test for SNPs (with nominal FPR of 0.05), but warn that the power is not directly comparable to that for microsatellites because the attained size of the test is different due to the severe discreteness problem discussed above. Table 3.5 shows the power for SNPs, which was roughly $20 - 30\%$ for very tightly liked loci (recombination rate under 10^{-3}). As expected, the power was higher for tighter linkage, which effect also depended on N_{e}.

The increase in power from $c = 10^{-1}$ to $c = 10^{-5}$ is more than tenfold when N_{e} was large, but only one and a half when $N_{\mathrm{e}} = 100$. There were only slight differences between tests and generally the LRT had the highest power (Table 3.5).

So far, I have shown that the effect of using the statistically convenient null instead of the biologically relevant finite population null is considerable and increases as information in the sample increases. Thus, the question arises whether we are better off using the unconditional, but finite population, null distribution, which involves estimating N_{e}, and μ for each locus and assuming a mutation model as well. In other words, I ask to what extent is testing sensitive to the specific sample allele frequencies: what is the effect of using a null distribution based on allele frequencies only roughly similar to the sample's allele frequencies. To address this question I compared the p-values from tests using the conditional (Monte Carlo) and unconditional infinite population nulls using a mixture of samples simulated under different alternative scenarios. Samples with recombination rates, $c = \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, were mixed in equal proportions (Figure 3.6).

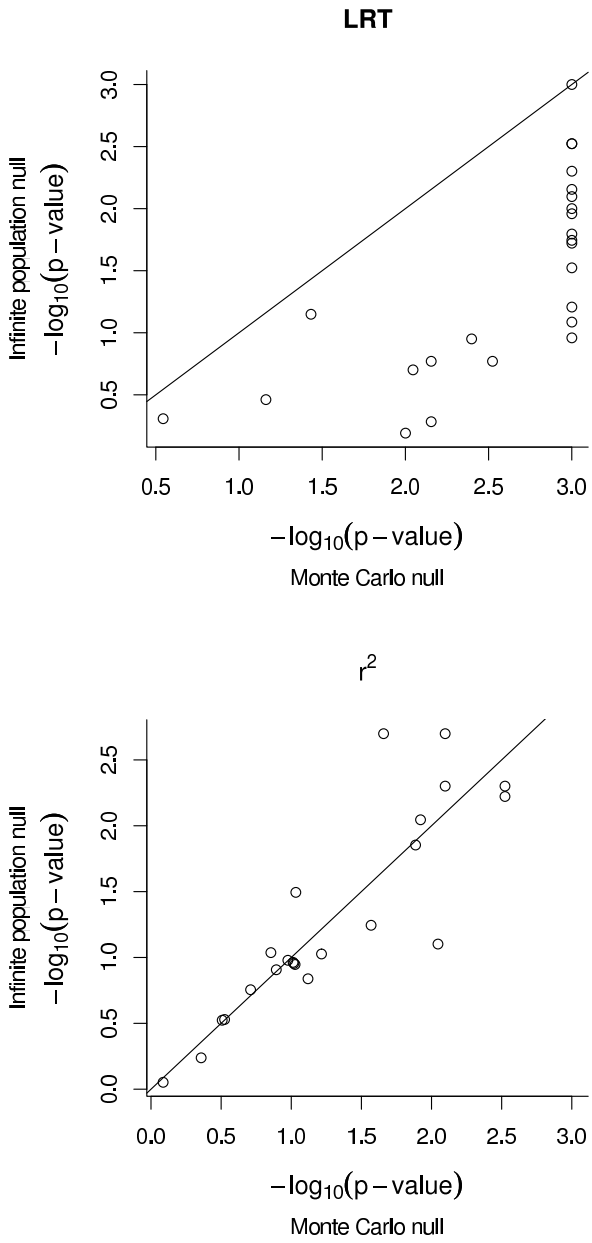

Figure 3.6: Comparison of the p-values from tests using the conditional infinite (Monte Carlo permutation) and unconditional infinite population null distributions. Lines show what we expect if conditioning did not have an effect on the p-values. Data was simulated under a mixture of recombination rates, $c = \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, which were mixed in equal proportions. Figures show data for $N_{\mathrm{e}} = 1000$ and mutation rate $\mu = 0.001$, and for two test statistics, the likelihood ratio (LRT) and r^2.

I found that the effect of conditioning on the sample allele frequencies depends on the choice of the test statistic. When the LRT is used there is no correlation between the p-values from the conditional and unconditional tests (Figure 3.6), but the

Table 3.2: Actual FPRs for SNPs under the finite and infinite population null scenarios for three different test statistics. The mutation rate is 10^{-8} in all cases. Asterisks indicate cases when the binomial confidence interval did not include 0.05.

Null scenario	N_e	**Actual FPR**		
		LRT	χ^2	D'
Infinite population	100	0.019^*	0.020^*	0.012^*
	1000	0.020^*	0.024^*	0.008^*
	10000	0.021^*	0.022^*	0.012^*
Finite population	100	0.107^*	0.106^*	0.064
	1000	0.039	0.039	0.025^*
	10000	0.040	0.040	0.021^*

Table 3.3: Actual FPRs for microsatellites under the infinite and finite population null scenarios for four different test statistics. The mutation rate is 10^{-3} in all cases. Asterisks indicate cases when the binomial confidence interval did not include 0.05.

Null scenario	N_e	θ	**Actual FPR**			
			LRT	χ^2	D'	r^2
Infinite population	100	0.4	0.046	0.038	0.032^*	0.041
	1000	4.0	0.051	0.054	0.060	0.054
	10000	40.0	0.048	0.056	0.053	0.054
Finite population	100	0.4	0.333^*	0.262^*	0.194^*	0.253^*
	1000	4.0	0.260^*	0.127^*	0.182^*	0.140^*
	10000	40.0	0.687^*	0.249^*	0.560^*	0.514^*

unconditional test was always more conservative (i.e. led to higher p-values). Thus, the outcome of the test using the LRT statistic would be unreliable if I used the unconditional null. However, when D' and χ^2 were used, there was a weak correlation (results not shown) and when r^2 was used there was a strong correlation between the p-values; thus conditional and unconditional tests gave similar results, on average (Figure 3.6). Notice, however, that the correlation between the p-values becomes weaker for small p-values (the upper right corner of Figure 3.6), which is the area where the significant results are; some of the scatter is probably due to Monte Carlo error.

3.5 Application to a real data set

Here I will illustrate on a real data set how unconditional testing may provide an alternative to the Monte Carlo permutation test. I used data on independent pairs of loci to compare the performance of tests using finite and infinite population null

Table 3.4: Power of four different tests (with nominal FPR of 0.05) for microsatellites when using the Monte Carlo permutation test, i.e. the infinite population null distribution. Different values of the recombination rate, c, indicate different degrees of evidence against the null hypothesis. The mutation rate is 10^{-3} in all cases.

c	N_e	θ	**Power**			
			LRT	χ^2	D'	r^2
10^{-5}	100	0.4	0.718	0.659	0.652	0.707
10^{-4}			0.735	0.686	0.687	0.724
10^{-3}			0.762	0.695	0.692	0.736
10^{-2}			0.697	0.614	0.606	0.644
10^{-1}			0.595	0.488	0.430	0.490
10^{-5}	1000	4	1.000	0.999	0.999	0.999
10^{-4}			1.000	0.999	0.999	1.000
10^{-3}			1.000	0.998	0.999	1.000
10^{-2}			1.000	0.974	0.993	0.989
10^{-1}			0.852	0.472	0.670	0.498
10^{-5}	10000	40	1.000	1.000	1.000	1.000
10^{-4}			1.000	1.000	1.000	1.000
10^{-3}			1.000	1.000	1.000	1.000
10^{-2}			1.000	0.931	1.000	0.999
10^{-1}			0.696	0.256	0.553	0.466

Table 3.5: Power of three different tests (with nominal FPR of 0.05) for SNPs when using the Monte Carlo permutation test, i.e. the infinite population null distribution. Different values of the recombination rate, c, indicate different degrees of evidence against the null hypothesis. The mutation rate is 10^{-8} in all cases.

c	N_e	**Power**		
		LRT	χ^2	D'
10^{-5}	100	0.285	0.273	0.256
10^{-4}		0.285	0.279	0.252
10^{-3}		0.314	0.300	0.284
10^{-2}		0.248	0.238	0.224
10^{-1}		0.198	0.188	0.150
10^{-5}	1000	0.289	0.279	0.247
10^{-4}		0.305	0.293	0.271
10^{-3}		0.266	0.257	0.231
10^{-2}		0.181	0.176	0.134
10^{-1}		0.070	0.070	0.037
10^{-5}	10000	0.316	0.305	0.287
10^{-4}		0.288	0.272	0.258
10^{-3}		0.194	0.183	0.142
10^{-2}		0.053	0.055	0.022
10^{-1}		0.031	0.033	0.016

distributions. Genetic data were used from a long-term study on red deer (*Cervus elaphus*) in the North Block of the Isle of Rum, Scotland (*e.g.* CLUTTON-BROCK *et al.*, 1982). Population census data were also available for the whole island. A subset of 200 unrelated individuals were selected for the purpose of this example, by excluding all first degree and, where they were known, second degree relatives. Data were available for 15 microsatellite loci (CP26, FCB193, FCB304, FCB5, INRA011, INRA035, JP15, JP27, JP38, MAF109, RT1, TGLA127, TGLA322, TGLA94, and VH54; see details in SLATE *et al.* (2002)). All but three pairwise comparisons between loci were considered: the pairs which ensure that locus pairs are in separate linkage groups or separated by at least 60 cM (SLATE and PEMBERTON, 2007). I used a total of 102 pairwise comparisons. Note that a related, but larger data set was used in SLATE and PEMBERTON (2007). In order to calculate LD, haplotype frequencies had to be estimated from the data. Even though the selected loci were known to be unlinked, I pretended that this information was not available, and inferred phase, i.e. the gametic frequencies, from the genotype data using the EM algorithm implemented in the R function *haplo.em* (in the *haplo.stats* package). The algorithm first expands the data for each individual into the complete set of possible pairs of haplotypes, from which the frequencies of the haplotypes can be estimated (see detailes of the algorithm in the documentation of the software SNPHAP, `http://www-gene.cimr.cam.ac.uk/clayton/software/`). I estimated the "haplotype" frequencies for each loci pair independently, i.e. no information was used from other loci.

The Monte Carlo permutation test can be readily performed, but, in order to simulate an unconditional finite population null distribution, the nuisance parameters N_{e} and μ also have to be estimated. I estimated the mutation rate using the program BATWING (WILSON *et al.*, 2003), freely available at `http://www.mas.ncl.ac.uk/~nijw/`). BATWING uses a Markov Chain Monte Carlo method based on a coalescent model to generate the posterior distribution of model parameters, including mutation rate μ and N_{e}. In a coalescent model the data do not provide information about N_{e} and μ separately, but only about their product, θ. Therefore, I estimated N_{e} using independent demographic data.

There are several formulae available to estimate N_{e} taking into account one or more of the departures from the Wright-Fisher model. Red deer have a polygynous, highly skewed mating system: male life-time reproductive success ranged from 0 to 48 offspring in the data set and the generations are overlapping. Since the effect of overlapping generations is in most cases negligible (ALLENDORF and LUIKART, 2007) I decided to ignore it, but to take into account the sex specific variance of family size. I used the following formula suggested by ALLENDORF and LUIKART (2007) to

estimate N_{e},

$$N_{\mathrm{e}} = \frac{8N - 4}{V_{km} + V_{kf} + 4}\,, \tag{3.1}$$

where V_{km} and V_{kf} are the variance in family size for males and females, which I estimated from the red deer pedigree data. V_{km} and V_{kf} were estimated to be 47.6 and 9.36, respectively. N is the population census of the whole island, which I calculated as an average over a 20 year period (1980-2000). The average N was 1470, which gives us an $\hat{N}_{\mathrm{e}}$ of 193.

The mutation rates, μ_i were estimated for each locus separately using a Uniform[0,1] prior and N_{e} as a constant set to 193. Although inferences about a single locus are not reliable (WILSON and BALDING, 1998), I emphasize that the aim here is not to accurately estimate N_{e} and the μ_i, but to demonstrate how to simulate an unconditional null distribution, in which the nuisance parameters, N_{e} and the μ_i, need to be estimated as accurately as possible given the data. I ran 1,000,000 MCMC iterations for each locus, and excluded the first 300,000 iterations as burn in. The chains converged and the posterior distribution of the mutation rates were sufficiently peaked (the width of the 95% credibility intervals ranged between 0.008 and 0.023). I used the posterior median as a point estimate for μ: estimates ranged between 0.0065 and 0.0312, with a mean of 0.0163.

The unconditional null distributions were then simulated for each locus pair, using Simcoal2 with the estimated values of N_{e} and the relevant pair of μ_i values. I set the geometric parameter of the generalized stepwise mutation model to 0 since BATWING implements the stepwise and not the generalized stepwise mutation model. Simcoal2 was run 1000 times for each of the 102 locus pairs, with recombination rate 0.5 between each pair, to simulate the unconditional finite population null distributions. I also simulated the unconditional infinite population null using independent coalescent trees at each locus. Further, an indication of sensitivity of the rejection rates to the parameter estimates was gained by performing the three tests using slightly higher and slightly lower N_{e} values. Since it is not straightforward to calculate an error on the estimate N_{e}, we made an arbitrary choice of considering $\hat{N}_{\mathrm{e}} + 50 = 243$ and $\hat{N}_{\mathrm{e}} - 50 = 143$.

The null hypothesis was almost always rejected (p-value was less than 0.05) when using the Monte Carlo permutation test, and, similarly, the rejection rates are high when the "unconditional infinite population" null was tested (Table 3.6). The differences between the two cases are marked, and could be due to the lack of conditioning, to the lack of accuracy in the estimates of N_e and μ_i, and/or to the assumption of the simple stepwise mutation model. The closest agreement between the conditional (Monte Carlo) and unconditional infinite population tests were observed

Table 3.6: Proportion of times when the null hypothesis was rejected (actual FPRs) using the Monte Carlo permutation test and the two unconditional test. 102 pairwise comparisons between unlinked microsatellite loci from red deer. The unconditional tests were performed with three different values of N_e: the estimated $\hat{N}_e = 193$ for the population and an arbitrarily chosen values of $\hat{N}_e \pm 50 = [143, 243]$.

Null distribution	$\hat{N}_e$	**Actual FPRs**			
		LRT	χ^2	D'	r^2
Monte Carlo	–	1.000	0.980	1.000	0.951
Infinite	143	0.441	0.775	0.676	0.980
	193	0.431	0.794	0.725	0.980
	243	0.451	0.804	0.686	0.971
Finite	143	0.010	0.059	0.078	0.284
	193	0.069	0.098	0.147	0.412
	243	0.108	0.167	0.235	0.549

for the r^2 statistic, in accordance with the simulation results, which showed that the r^2 was the most robust to the lack to conditioning (Figure 3.6). Thus, the r^2 statistic is the best predictor of the actual FPRs for the unknown ideal test: the conditional finite population. While the other statistics, e.g. the LRT, are expected to have much higher actual FPRs if the conditional finite population test was applied (Figure 3.6). However, when the "unconditional finite population" null was used the null hypothesis was rejected less often, but still more than 5% of the times as would be expected for comparisons between independent loci in a test of nominal size 0.05 (Table 3.6).

The unconditional testing procedure was sensitive to the parameter estimates for all tests: smaller N_e resulted in actual FPRs rates that were closer to their nominal value (Table 3.6). This result could indicate that the actual N_e in the population is smaller than $\hat{N}_e = 193$ and, that more accurate results could be obtained with more accurate parameter estimates. Also, seemingly "bad" results (i.e. high actual FPRs) for r^2 may only indicate that the parameter estimates are inaccurate.

3.6 Discussion

3.6.1 The effects of "background LD" on testing

In finite populations, mutation and genetic drift constantly generate linkage disequilibrium. Thus, even in the absence of these forces, it would take an infinite amount of time for LD to decay to exactly zero. Therefore, small but non-zero amounts of LD are expected between freely recombining loci in all real populations. In this Chapter, I

have drawn a distinction between a biologically meaningful null hypothesis, which is free recombination in a finite population, and a statistically convenient null hypothesis, which is exactly zero LD, which could arise with an appreciable probability only in an infinite population. Although it is clear that these are two distinct null hypotheses, it was unclear a priori whether there would be any practical differences between them.

My simulations demonstrated that different conclusions will often be reached, depending on whether I test the finite or infinite population null hypothesis and on the informativeness of the data. While for biallelic markers, e.g. SNPs, the null hypothesis of zero LD is a good approximation for the biologically relevant, finite population null, for highly polymorphic loci, e.g. microsatellites, the nulls can be very different. This is because, in highly informative samples, i.e. in samples of highly polymorphic markers, the statistically convenient null does not take account of the non-zero "background" LD found in all finite populations. Similarly, as the sample size increases the difference between the two null hypothesis becomes greater. Two main practical conclusions arise from these results: (i) the Monte Carlo permutation test (i.e. the infinite population null) applied to highly informative samples will result in high false positive rates (actual FPRs), and thus, (ii) the power of any test statistics will be spuriously inflated.

I found that for a nominal FPR of 5% the actual FPRs were alarmingly high for microsatellites, roughly between 20% and 60%, or even around 90% for extremely polymorphic markers, in my simulated data sets using polymorphism levels and effective population sizes that are commonly found in natural populations. Analysis of a real data set with moderately polymorphic microsatellites further corroborated these results. Using the Monte Carlo permutation test, i.e. testing the infinite population null, the null hypothesis was rejected for almost all pairs of loci (FPRs of nearly 100%), even though loci were genetically independent. These actual FPRs were higher than predicted by my simulations and, were also observed in a study using a larger related data set (SLATE and PEMBERTON, 2007), which could well be due to recent admixture in the population SLATE and PEMBERTON (2007) and the sample size differences. Nevertheless, both my simulation and empirical results strongly suggest that biologically meaningful conclusions cannot be drawn by testing the statistically convenient null hypothesis. The infinite population null may be an acceptable approximation only when biallelic markers, such as SNPs, are used or, in the case of microsatellites, if there is external information that N_{e} is extremely large (i.e. when ρ tends to ∞).

The spuriously inflated power is the other main consequence of using an inadequate null hypothesis. I found that the power of all tests increased with the informativeness

of the data. I argue that this power gain is, however, spurious because the FPRs, along with the difference between the finite and infinite nulls, also increase with the informativeness. It has been suggested that multiallelic markers can contain more information about LD than biallelic markers (ZHAO *et al.*, 1999), and thus may have a higher power to detect LD (SLATKIN, 1994). Indeed, ZHAO *et al.* (1999) found that the power of the Monte Carlo permutation test was 1 when testing for LD (using the LRT statistic) under the alternative hypothesis of "weak linkage disequilibrium" in large samples (sample size of 200). However, my results suggest that this is not a true power gain. I therefore argue that a power comparison that uses the Monte Carlo permutation null cannot be used to conclude that there is more power to detect linkage with more informative (i.e. more polymorphic) data. My results might also explain the increase in power with the number of alleles in related genetics problems as well. MAISTE and WEIR (2004) investigated testing for Hardy-Weinberg equilibrium in large contingency tables, and found that with multiallelic markers the power of the Monte Carlo permutation test increased with the number of alleles.

3.6.2 The approximate testing procedure

Both the high false positive rates and the inflated power suggest that, with highly polymorphic loci the biologically relevant, finite population null should be tested. The only feasible way for testing the finite population null, to my knowledge, is a goodness of fit test, using the unconditional finite population null distribution, which requires the estimation of N_{e} and μ for each locus and assuming a specific mutation model (the nuisance parameters). The application to a real data set of a natural population illustrates this method. I recognize the potential difficulties and inaccuracies in estimating the nuisance parameters: data might have to come from a source *additional* to the genetic data in which the LD is observed. Nevertheless, even my rough and ready estimates of the nuisance parameters yielded test results much closer to the biological expectation when the finite population null was tested. This result suggests that the dangers of true mis-inference are much greater for the supposedly "exact" Monte Carlo permutation test, than for my admittedly approximate procedure.

The accuracy of the unconditional testing will depend on the level of polymorphism. I found that there is a decrease in power with increasing level of polymorphism. The power also decreases with θ even when I used an unconditional infinite population null (with fixed actual FPRs, results not shown). Thus, with more polymorphic data the unconditional test provides an increasingly poorer approximation to the conditional, which could be explained by the fact that, with more alleles, the variance in allele numbers across the simulated populations is also inevitably higher. Thus the simulated

allele and haplotype frequencies used in the unconditional procedure will approximate the true sample allele and haplotype frequencies less well.

3.6.3 On the choice of the test statistic

In order to test the biologically meaningful null hypothesis, we are forced to use an unconditional distribution of the test statistic. My simulations confirmed that in large contingency tables different test statistics extract different information about LD. This is in contrast to 2×2 tables, where all statistics perform relatively uniformly. Thus, the choice of test statistic is critical, and the test statistic which is the most robust to the lack of conditioning should be preferred. Thus, a statistic with a null distribution that is as weakly dependent as possible on the values of the nuisance parameters, and can also *separate* the information about the null hypothesis from information about nuisance parameters (SPROTT, 2000). Good test statistics will also have a high power. I found that the distribution of all statistics depends on N_{e} and θ to some extent (Figure 3.2), and that tests based on the r^2 or χ^2 statistics are the most robust. Regarding the statistical null hypothesis of zero LD the LRT consistently had the highest power, but its FPRs are also the highest. When I compared the power of different tests, considering the technical difficulties with an arbitrary, albeit with valid and biologically meaningful approach (i.e. using the finite population null with fixed actual FPRs), I found that the LRT often had the lowest power, and the r^2 tests the highest power (Figure 3.5).

3.6.4 Related studies and future directions

The application to the red deer data set illustrated how mis-inference can be avoided via testing the biologically meaningful null hypothesis of genetic independence. In other words, I illustrated how one can test whether the recombination rate is different from 0.5. I also highlighted that nuisance parameters, the effective population size, N_e, and the locus specific mutation rates, μ, are required for the estimation. The proposed approximate testing procedure is closely related to previous methods that are aimed at estimating the fine-scale recombination rates from population data (MCVEAN *et al.*, 2004; MYERS *et al.*, 2005), in the sense that they make use of LD information to make inference about the population recombination rate.

I performed a very limited sensitivity analysis of my approximate testing procedure to the estimated value of the nuisance parameter, N_e (based on three values of N_e). However, one could imagine a sensitivity analysis on a much finer scale. Such a sensitivity analysis might then be used to make inferences about the N_e itself. This

is to say, that the logic of the proposed unconditional test could be turned the other way around. For example, if loci are known to be unlinked, e.g. because they are on different chromosomes, one could use LD data to make inferences about N_e (or the mutation rate, μ). In fact, the idea of using LD data to estimate N_e was originally proposed by HILL (1981). A related method has recently been applied to genomic data by TENESA *et al.* (2007), who presented the first genome-wide estimates of the human effective population size based on LD data. Their analysis was based on the approximate relationship between the LD measure, r^2, and N_e. A similar method could be imagined for microsatellite data as well to estimate N_e, which could, for example, be implemented in a rejection sampling scheme (*e.g.* PRITCHARD *et al.*, 1999). The advantage of such a scheme would be that, multiple LD measures could be simultaneously used, which could all potentially carry different information about population demography.

Chapter 4

Using Approximate Bayesian Computation (ABC) to estimate demographic parameters from admixed population samples

Some of the ideas presented in this Chapter have been improved by discussions with Arnaud Estoup, Nick Barton, and Mark Beaumont. Arnaud Estoup also provided useful comments for the improvement of this Chapter.

4.1 Introduction

Hybridization or population admixture has long been of central interest in both theoretical and applied population genetics. Admixture of two or more previously isolated populations can increase genetic variation and also create novel genetic variation, and thus, for example, studying hybrid populations provide an important approach to the understanding of speciation (BARTON, 2001). Admixture is also a common feature of many species invasions, which often follow climate fluctuations, such as, during the Pleistocene cycles of glaciation after expansions from glacial refuges (HANSSON *et al.*, 2008). The increased genetic variance resulting from mating between previously isolated populations could also enhance adaptation (*e.g.* LAVERGNE and MOLOFSKY, 2007). Inferring the history of admixture events can be crucial in applied sciences, e.g. in the management of invasive and/or endangered species. In human genetics, there is a growing interest in applying mapping by admixture linkage disequilibrium to identify genes underlying complex traits and

diseases (*e.g.* TIAN *et al.*, 2006). Admixture mapping that takes advantage of the fact that recently admixed populations have larger regions of LD between loci than non-admixed populations.

In the context of population admixture, many estimators have been developed to estimate the admixture proportion based on the comparison of the allele frequencies of putative parental and hybrid populations (*e.g.* CHOISY *et al.*, 2004; CHIKHI and BEAUMONT, 2001), or more recently based on a coalescent model (WANG, 2006). The accuracy of any methods to infer population admixture will depend on the allele frequency differences between the parental populations. Most of these simple statistics are based on a simple evolutionary model. More recent methods use computationally intense statistical tool, such as Markov Chain Monte Carlo (MCMC) or Importance Sampling (IS) and focus on inferring the underlying processes that gave rise to the observed sample patterns (MARJORAM and TAVARÉ, 2006). Such methods are thus able to estimate multiple demographic parameters simultaneously. For inferring population structure, for example, the most well-known method is implemented in the software Structure (FALUSH *et al.*, 2003).

In the modern computationally intense methods, however, assumption of biologically realistic models often inhibits the use of full data likelihood methods (MARJORAM and TAVARÉ, 2006). For example, in order to apply a mutation model more sophisticated than the stepwise mutation model (SMM) for microsatellite data the likelihood calculation becomes very complex. Thus, approximate methods, based on summary statistics have gained popularity in recent years. Approximate Bayesian Computation (ABC) uses inferences based on summary statistics, and it holds great promise since it can potentially handle models of any complexity, with many nuisance parameters, provided only that simulation of data under the model is feasible and that suitable summary statistics can be found (BEAUMONT *et al.*, 2002). ABC has recently been applied to population demography problems (*e.g.* EXCOFFIER *et al.*, 2005; INGVARSSON, 2008; ROSENBLUM *et al.*, 2007) or to distinguish between different potential demographic scenarios (*e.g.* ESTOUP *et al.*, 2004; FAGUNDES *et al.*, 2007).

EXCOFFIER *et al.* (2005) evaluated the performance of ABC in a simple admixture scenario and found that in comparison to a recently developed maximum-likelihood method (WANG, 2003), the ABC approach leads to similarly accurate estimates of admixture proportions in the case of recent admixture events, and outperforms the ML method when the admixture is old. The ABC approach is clearly more flexible as well: parameters, such as the divergence and admixture times, and the effective population sizes, can be simultaneously estimated.

Currently the most frequently used method to analyze data from an admixed population is the software Structure, originally developed by PRITCHARD *et al.* (2000a). In Structure2 FALUSH *et al.* (2003) extended the method by allowing for linkage between loci, motivated by the fact that linked loci are potentially better than unlinked ones for inferring population admixture, since the linkage disequilibrium (LD) generated by admixture is preserved for longer than between freely recombining loci. The linkage model of FALUSH *et al.* (2003) takes account of the correlations between linked loci that arise in admixed populations, which allows the detection of admixture events further back into the past, and also inference of the population of origin of chromosomal regions.

In the ABC scheme of EXCOFFIER *et al.* (2005), unlinked loci are used, but the authors noted that the use of linked markers does not introduce any particular bias in the estimation of admixture. Thus, EXCOFFIER *et al.*'s (2005) ABC scheme does not benefit from the additional information that linked markers carry because linkage was not explicitly taken into account. However, similarly to Structure, when the linkage map is known ABC could also potentially benefit from using linked markers, since genetic data can be simulated given the map distances between markers. However, two questions arise: first, for how many generations would LD be maintained between linked loci in the admixed population, so the scheme could benefit from using linked markers, second, how tightly linked the loci should be to benefit the most. These questions could be explicitly addressed in and ABC framework. One drawback of using linked markers is the potentially large computational cost because the simulation of linked loci using coalescence with recombination is much slower.

The aim of this Chapter is to assess the quality of the estimation of the admixture proportion and the time of admixture in a simple demographic scenario of an instantaneous admixture event. Thus, two diverged parental populations admix and form a new population without any subsequent migration from the parental populations. Specifically, first, I will assess to what extent LD statistics can improve the estimation of the admixture proportion and the time of admixture when using unlinked and when using linked markers, and for how many generations admixture LD can be maintained. Second, I will compare different LD statistics to address which measures of LD are the most informative for population admixture when using multiallelic markers such as microsatellites. Third, I will contrast the quality of the estimation in the presence and absence of samples from the parental populations. My motivation to do that is that in most real situations samples are not available from the parental populations.

The most popular method that ABC has to compete with to analyze samples from

admixed populations is implemented in software Structure. ABC could be a potentially more flexible alternative to Structure if it accurately estimated the admixture proportion and time of admixture in the absence of samples from the parental populations. There are a number of reasons why this might be true, however, only a through comparison between Structure and ABC could confirm this, which is beyond the scope of this Chapter. For example, Structure uses LD data to infer about admixture. However, for old admixture events when admixture LD is not expected to be present any more, other aspects of the data might still be informative. ABC could readily accommodate statistics that capture many different aspects of the data. For example, the shape of the allele frequency distribution using GARZA and WILLIAMSON's (2001) M statistic could be informative about population history. Finally, ABC could also easily accommodate a realistic mutation model for microsatellites.

4.2 Methods

4.2.1 The demographic scenario

I considered the following simple demographic scenario (Figure 4.1), which has been used in some previous studies (*e.g.* EXCOFFIER *et al.*, 2005; CHOISY *et al.*, 2004). The model describes an instantaneous admixture event of two populations, P_1 and P_2, both diverged from from an ancient population, P_0, at T_{div} generations ago. The proportion of genes with P_1 population origin in the admixed population, P_A, is the admixture proportion, λ. The three populations are sampled T_{adm} generations after the admixture event. The N_e in all three populations and the mutation rates are nuisance parameters. I considered three scenarios for sample availability: when samples are available (i) from the admixed and the two parental populations, (ii) form the admixed and one of the source populations, and (iii) only from the admixed population.

Generally, the raw demographic parameters, such as T_{adm}, T_{div}, N_e or the mutation rate, μ, are difficult to estimate, but the scaled parameters, such as $\theta = 4N_e\mu$, or the times of historical events scaled by N_e or μ, can be much more accurately estimated (*e.g.* EXCOFFIER *et al.*, 2005). This is because it is generally difficult to separate the information in the data about N_e from that about μ, and also to distinguish the timing of the historical events from θ. My main parameters of interest were λ and T_{adm}. For estimating admixture proportions, I estimated not only λ itself, but $\lambda_{min} = min(\lambda, 1 - \lambda)$, which is identifiable even when a sample is available only from the admixed population. For estimating the time since admixture, I estimated the raw T_{adm} parameter itself, and also τ_{adm}, which is the time scaled by the mutation rate $\tau_{adm} = T_{adm}\hat{\mu}$.

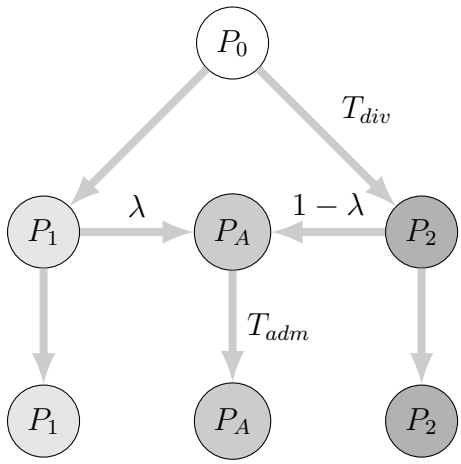

Figure 4.1: A simple demographic scenario of an instantaneous admixture event, where two previously diverged (with T_{div} generations) parental populations, P_1 and P_2, admix at time in generations, T_{adm}, and form a new population, P_A. After the admixture event there is no subsequent migration from the parental populations to P_A. The genetic contribution to P_A from P_1 is λ, the admixture proportion.

4.2.2 Parameter estimation

I used Approximate Bayesian Computation (ABC) to estimate parameters of the demographic model above as described in BEAUMONT *et al.* (2002). ABC allows approximation of the posterior distribution for the parameters, from which point estimates and credible intervals can be constructed. The parameter estimation via ABC consists of three steps. First, data are simulated using parameter values drawn from their prior distributions. Second, summary statistics are calculated from the simulated data sets and the proportion of simulations where the summary statistics are the closest to the observed data summary statistics are retained. In the third step, the posterior distributions of the parameters are estimated using a locally weighted linear regression, which weights the parameter values of the retained simulations by distance of the corresponding summary statistics from the observed summary statistics.

The ABC estimation provides an improvement the previously used rejection sampling (RS) approach (RIPLEY, 1982; PRITCHARD *et al.*, 1999) in step three. The advantage of the local linear regression applied in the ABC scheme, relative to RS, depends on the sensitivity to the choice of the tolerance. The tolerance is the proportion of simulations that are accepted, and in turn, used for parameter estimation. The choice of the tolerance generally involves a bias-variance trade-off (BEAUMONT *et al.*, 2002). This is because although increasing the proportion of accepted values reduces the variance in the local regression thanks to having a larger sample size, it also increases

bias arising from the uncorrected departures from additivity and linearity (BEAUMONT *et al.*, 2002). Thus, I studied a range of different tolerance values and also compared ABC and RS. Parameters in ABC were estimated using equation 9 in BEAUMONT *et al.* (2002).

I used the coalescent to simulate data sets under the demographic model above, using the simulation software Simcoal2 (freely available at `http://cmpg.unibe.ch/software/simcoal2`). $500,000$ data sets were simulated with a sample size of 200 diploid individuals from the admixed population and 25 from the two parental populations for different linkage scenarios. I considered both linked and unlinked loci in order to investigate to what extent can the use of linked markers (with known recombination rates) improve the estimation through the use of LD statistics. First, I considered 20 independent loci using the coalescent with recombination, with recombination rate 0.5. Thus, I used correlated genealogies between the loci to correctly simulate the non-zero LD between independent loci in finite populations. Second, I considered 20 loci in two linkage groups, 10 loci in each.

In order to simulate linked loci with pre-specified recombination rates with Simcoal2, I had to place all loci on a linear linkage map. I used the desired recombination rates between loci within a linkage group, and a recombination rate 0.5 between the last and the first loci of two consecutive linkage groups. This arrangement, will introduce a slight bias, in the sense that not all pairwise comparisons between independent loci (i.e. loci that are in different linkage groups) will have exactly a recombination rate 0.5 between them, but some will have slightly more. That is, they will behave as if they were independent loci in a population with a slightly smaller effective population size. The further the two loci are on the linear map the larger the bias gets. I estimated the extent of this bias and it is notable. However, ignoring the correlation between independent loci (and using coalescent without recombination) would introduce an even larger bias (see Chapter 3). I took account of the effect of this bias in the ABC estimation, by assuming a fixed map order, and all ABC and test samples were simulated according to that. A similar approach was used by EXCOFFIER *et al.* (2005), who also used Simcoal for ABC estimation.

4.2.3 Prior distributions

The prior distributions for all parameters were selected to represent a wide range of potential values in real populations and also to encompass the range of biologically relevant test values that I was interested in. I used a LogNormal prior for $\mathrm{S}N_e$ with mean of 6.5 and variance of 2 (Table 4.1). This prior is similar to that of PRITCHARD *et al.* (1999), with a lower mean in order to put more weight on smaller values of

N_e. I truncated this prior to an allowable range of 10 to 40000 individuals, so that no computational time was wasted on biologically unrealistic values. In each simulation I drew a value of N_e from this prior independently for the admixed and the two parental populations. Thus in some simulations at the admixture event a population expansion, or a population contraction, or constant population size was modelled.

Table 4.1: The prior distributions of the demographic and mutation model parameters in the ABC scheme. The demographic scenario is outlined on Figure 4.1. Amongst the raw demographic parameters, note that the prior for effective population size is truncated from 10 to 40000. Mutation parameters correspond to a two parameter model, the generalized stepwise mutation model, where in each mutational step the size of the change in repeat units follows a Geometric distribution. Composite parameters are scaled parameters. See further details in text.

Parameters		**Distribution**	**Quantiles**		
			5%	50%	95%
Demographic parameters					
Effective population size (P_1, P_2, P_A)	N_e	LogNormal$(6.5, 2)$	411	4202	18982
Time of divergence	T_{div}	Uniform$(10^3, 10^5)$	5895	50440	95051
Time of admixture	T_{adm}	Uniform$(1, 10^3)$	51	499	950
Admixture proportion	λ	Uniform$(0, 1)$	0.05	0.5	0.95
$min(\lambda, 1-\lambda)$	λ_{min}	Uniform$(0, 0.5)$	0.025	0.25	0.475
Mutation parameters					
Mean mutation rate	$\bar{\mu}$	Uniform$(10^{-5}, 10^{-3})$	-	-	-
Mutation rate at locus i	μ_i	Gamma$(3, 2/\bar{\mu})$	-	-	-
$\hat{\mu}$	$\sum \mu_i/n$		8.805×10^{-5}	7.442×10^{-4}	1.481×10^{-3}
Mean Geometric parameter	$\bar{p}$	Uniform$(0.3, 0.7)$	-	-	-
Geometric parameter at locus i	p_i	Beta$(0.5 + 199\bar{p}, a(1-\bar{p})/\bar{p})$	-	-	-
$\hat{p}$	$\sum p_i/n$		0.457	0.5	0.544
Composite parameters					
Scaled mutation rate	θ_{adm}	$4N_e\hat{\mu}$	0.531	10.092	64.084
Scaled T_{adm}	τ_{adm}	$T_{adm} \sum \mu_i/n$	0.016	0.282	1.062

For the admixture proportion I used a Uniform$(0, 1)$ prior, thus allowing for any possible values of λ. This prior thus translates into a Uniform$(0, 0.5)$ for λ_{min} (Table 4.1). For times of the divergence event, I used a Uniform prior with range of 1000, 100000, thus I assumed that the two parental populations were always well diverged, though the different N_e values of any particular simulations would always later the degree of divergence (i.e. smaller populations will be more diverged, Table 4.1). For the time since the admixture event I used a Uniform prior with a range of 0, 1000. Thus, the prior of τ_{adm}, which is defined as the product of T_{adm} and the average mutation rate across loci, $\hat{\mu}$, was the product of a Uniform and a mean of Gamma distributions (Table 4.1).

I used a version of the generalized stepwise mutation model (GSM), which is implemented in Simcoal2. GSM is a two parameter model, where changes of multiple repeat units are allowed. A mutation occurs at rate μ, and if x is the size of the change in repeat units at a single mutation event, $x - 1$ follows a Geometric distribution with parameter p. For both μ and p, I used hierarchical priors: a global prior for the locus averages, and a specific prior for the locus specific values of μ and p. I chose to do so, to ensure that loci were more similar in terms of their mutation parameters within than between simulations. I chose the prior distributions following ESTOUP *et al.* (2001) and the parameter values according to my literature based estimates presented in Chapter 1. I used a Uniform$(10^{-5}, 10^{-3})$ prior for the mean mutation rate, $\bar{\mu}$, while the locus-specific mutation rates (μ_i) followed a Gamma$(3, 2/\bar{\mu})$ distribution. For the mean Geometric parameter across loci, $\bar{p}$, I used a Uniform$(0.3, 0.7)$ prior, while the locus specific Geometric parameter parameter followed a Beta$(0.5 + 199\bar{p}, a(1 - \bar{p})/\bar{p})$ distribution (Table 4.1).

4.2.4 Summary statistics

In all estimations I used a set of well-established population genetics statistics that have also been used in other ABC studies to estimate demographic parameters (*e.g.* EXCOFFIER *et al.*, 2005; NEUENSCHWANDER *et al.*, 2008). This set of statistics included the number of alleles, the expected heterozygosity, and GARZA and WILLIAMSON's (2001) M statistic, which is the ratio of the number of alleles and the range of allele sizes (all of them averaged across the 20 loci). Further, I included pairwise F_{ST} and Goldstein's genetics distance (GOLDSTEIN *et al.*, 1995a) between all pairwise combinations of the admixed and the two parental populations (Figure 4.1). These statistics were used in all estimations and subsequently I will refer to this set as classic statistics. Two further sets of statistics were added to this set, which I will detail next.

LD statistics are likely to be informative when the admixture is recent and recombination has not yet broken up admixture LD. I used the following LD statistics: r^2, D', χ^2, and the likelihood ratio test statistic (LRT) that I also used in Chapter 3. All of these measures were calculated between all pairs of loci and then averaged over all locus pairs. When both linked and unlinked loci were simulated, an average was also taken within only linked and unlinked locus pairs, thus yielding two additional statistics.

Summary statistics that capture information about the shape of the allele frequency distribution could also be informative about the time of admixture. This is because, when two sufficiently diverged populations admix, the allele frequency distribution of the admixed population might have multiple modes corresponding to the modes in the parental populations. For example, if two sufficiently diverged parental populations with symmetric allele frequency distributions admix in roughly equal proportions, the allele frequency distribution of the admixed population has two modes. In this ideal scenario "mode-counting" statistics could be very informative about the admixture event, and thus they might be informative in scenarios close to this ideal case. I defined statistics that count the number of modes in the allele frequency distributions, where modes were estimated using a Gaussian kernel density estimator with three different bandwidths: SILVERMAN's (1986) (page 43) "rule of thumb" (0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power), and half and double this value.

All three sets of summary statistics were considered in three variations, depending on sample availability. When samples were assumed to be available from the admixed and the two parental populations, all statistics are available. However, when samples were available only from the admixed, and missing for one or both of the parental populations, the statistics were not available that correspond to the missing samples, nor the between population statistics, such as F_{ST}.

4.2.5 Test data sets

I evaluated the performance of the ABC approach on a series of samples with known parameter values. For each set of fixed test parameter values, I simulated 500 data sets. These were then used to calculate different accuracy measures, and also for inspection of the posterior distributions of parameters.

Since there is an enormous range of parameter combinations that could be explored, I fixed some of the parameters that were not of direct interest. I assumed that the two parental populations were always sufficiently diverged, so I fixed the time of divergence at 10000 generations. I also fixed the mutation parameters over all loci:

μ_i's to 5×10^{-4} per generation and the geometric parameter to 0.6.

My main parameters of interest were the time of admixture, T_{adm} and the admixture proportion, λ. For T_{adm} I considered admixture events at 2,5,10,100, and 500 generations before present. For the admixture proportion I considered a skewed admixture, with $\lambda = 0.1$, and symmetric admixture, with $\lambda = 0.5$. I also considered two values for N_e, 1000 and 10000, for the admixed and the source populations. Note that I also considered smaller N_es, but, these result in monomorphic data sets, so that there were not enough cases to analyze.

The simulation study was carried out using scripts written in R (R DEVELOPMENT CORE TEAM, 2005), and the parameter estimation using an R script kindly provided by M. A. Beaumont. As suggested by BEAUMONT *et al.* (2002), the retained simulated parameters were log transformed before the regression adjustment, and then parameter estimates were based on the back-transformed values. I followed BEAUMONT *et al.* (2002), who suggested using the regression fitted values as point estimates (Equation 7.), and I also considered two other commonly used statistics as point estimates, the posterior median and mode. The performance of the ABC estimation was characterized by the following accuracy measures: absolute and relative bias, where the latter is the bias divided by the true value of the parameter, the relative root mean square error (RelRMSE, square root of the mean squared error divided by the true value), the median absolute deviation (MAD), the 95% coverage (proportion of times when the true value is within the 2.5% and 97.5% quantiles), and the Factor 2 (proportion of times when the estimated value is in an interval bounded by values equal to 50 and 200% that of the true value). The RelMSE, and the absolute and relative biases were calculated using all three point estimates (i.e. regression fitted values, posterior median and mode). Generally, the three different point estimates produced very similar accuracy measures, and importantly, the relative performance under different parameter combinations was not affected by the choice of the point estimate. Generally, the bias and RelMSE were the smallest when using the posterior mode as a point estimate, thus, most results will be reported when using the posterior mode.

4.3 Results

4.3.1 Assessing the quality of estimation

I assessed how the choice of the proportion of simulations accepted for estimation (the tolerance) affects the conclusions about the accuracy of the parameter estimation. I also studied how the different accuracy measures (such as RelMSE, bias, etc.) change

along a grid of tolerance values ranging from 0.001 to 0.1, for different parts of the parameter space, and also with different sets of summary statistics. My aim was to determine an optimal level of tolerance under each set of statistics, where some suitable measure of the error is minimized. However, I generally found that the true value of a parameter of interest (and the other parameters) and the particular set of summary statistics influence the tolerance level under which the estimation is the most accurate. Thus, a single optimal level of tolerance for a particular estimation problem is difficult to determine. Next, I will illustrate these findings with examples.

First, I found that the accuracy measures change as a function of the tolerance level, and also depend on the true parameter value. Figure 4.2 and 4.3 contrasts four different accuracy measures as a function of the tolerance. Notice that while Factor 2 of one means the highest accuracy, the other measures, i.e. Relative bias, RelMSE and MAD, indicate the highest accuracy as they approach zero. For example, when estimating λ (at $\lambda = 0.1$ and $T_{adm} = 10$ generations or more recent) both RelMSE and MAD increased with the tolerance. However, for old admixture events, the estimation became more accurate when increasing the proportion of accepted simulations (Figure 4.2). When estimating T_{adm} for recent admixture events the most accurate estimates of T_{adm} were gained when using a tolerance as low as possible (here 0.001). In contrast, for old admixture events the optimal tolerance was 10 times larger (Figure 4.3).

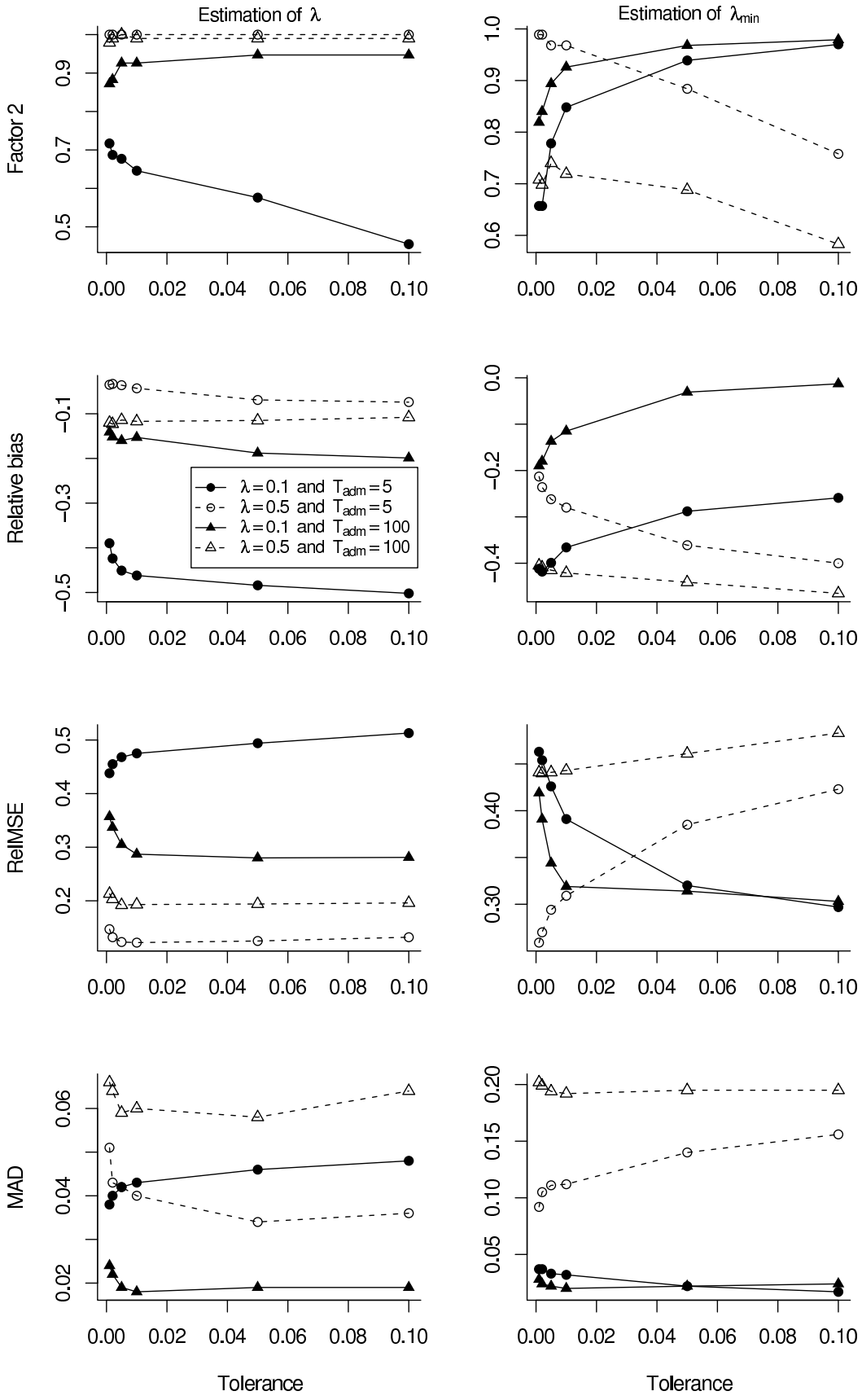

Figure 4.2: Accuracy of the estimation of the admixture proportion, λ and $\lambda_{min} = min(\lambda, 1-\lambda)$ as a function of the tolerance for two different true values of λ, 0.1 and 0.5, and two different true values of the time of admixture, T_{adm}, 5 and 100. Tolerance is the proportion of simulations accepted for parameter estimation. Accuracy is expressed in terms of four different measures, Factor 2, relative bias, relative root mean square error (RelRMSE) and median absolute deviation (MAD).

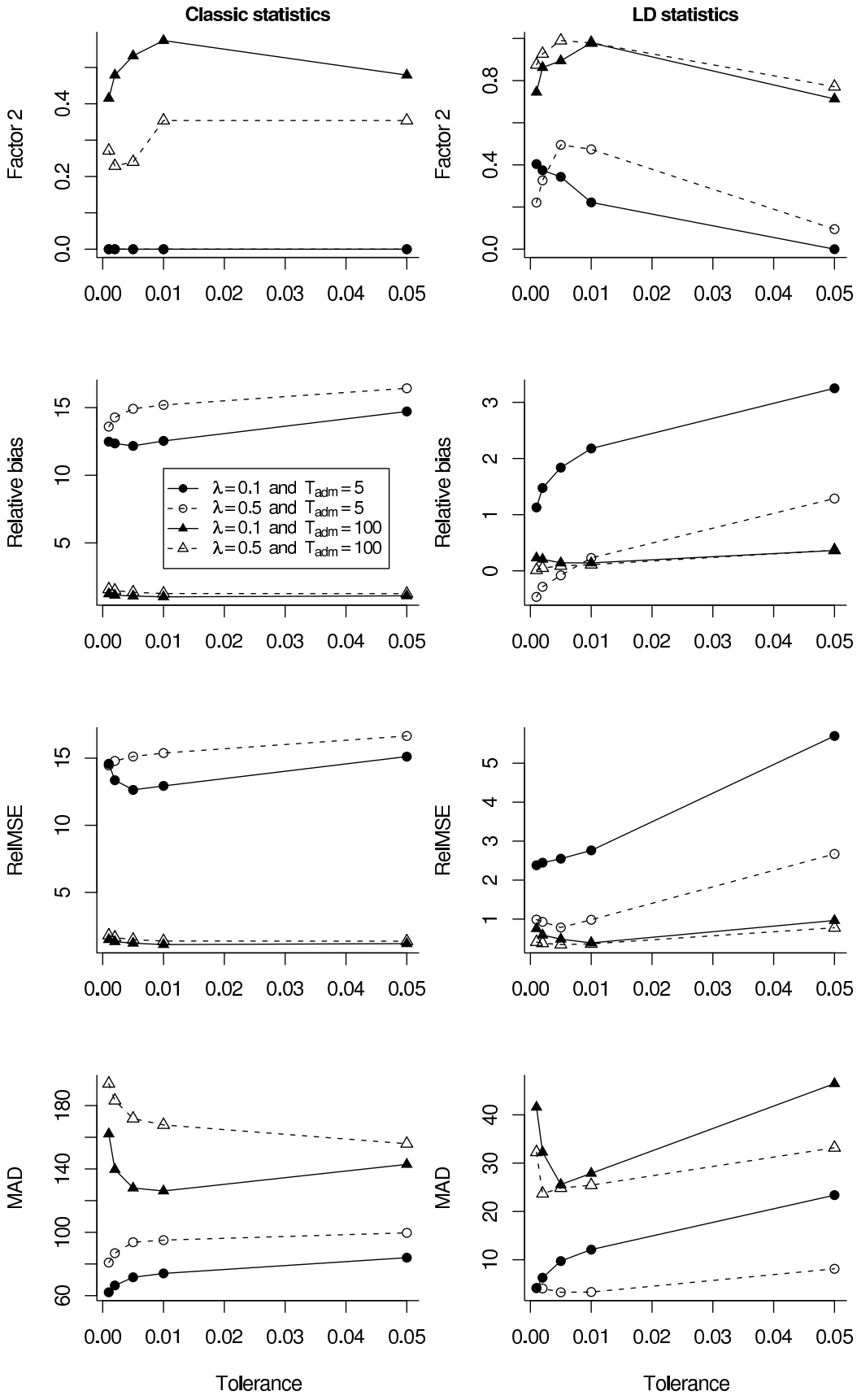

Figure 4.3: Accuracy of the estimation of the time of admixture (T_{adm}) as a function of tolerance for two different true values of λ, 0.1 and 0.5, and two different true values of the time of admixture, T_{adm}, 5 and 100, when using the classic or the LD statistics. Tolerance is the proportion of simulations accepted for parameter estimation. Accuracy is expressed in terms of four different measures, Factor 2, relative bias, relative root mean square error (RelRMSE) and median absolute deviation (MAD).

Second, I found that the optimal tolerance depends on the set of summary statistics used. For example, when estimating the time of admixture for recent and skewed admixture events with the classic statistics, the error was minimized at tolerance 0.005, but for LD statistics at 0.001 (Figure 4.3). An even more extreme example arose when the admixture was old and symmetric, here the optimal tolerance was 10 times more with LD statistics (0.01) than with the classic statistics (0.001) (Figure 4.3).

Third, I found that different accuracy measures could also suggest different optimal tolerances. For example, when estimating λ using the classic statistics (at $\lambda = 0.1$ and T_{adm}=100) Factor 2 was maximized at tolerance 0.05, while MAD was minimized at tolerance 0.01. The utility of a particular accuracy measure depends on the parameter itself and the most informative accuracy measure might be different for λ (which is a proportion) versus, for example, for the timing of a historical event that (theoretically) does not have an upper bound. As an example, for T_{adm} it is important to choose an accuracy measure, which is relative to the actual parameter value (e.g. Factor 2 or RelMSE; also see Figure 4.3), while for λ, which is a proportion, the absolute bias or MAD are the most appropriate (Figure 4.2).

How the accuracy changes with the tolerance values could be a result of multiple factors. For example, the fact that the relative bias and RelMSE for old admixture events was nearly zero for the whole range of tolerance values (Figure 4.3) does not imply highly accurate estimation, but reflects that there is not much information in the data about the parameter, so the posterior mode stays near the prior mean under the whole range of tolerance values studied. This example also demonstrates that the estimation will also depend on the priors, and especially on where the test values are in the prior parameter space. For example, when estimating λ, the true value of 0.5 is at the prior mode (and mean). However, when estimating λ_{min}, 0.5 is at the edge of the prior distribution. As a consequence, when estimating λ_{min} the absolute bias was two to three times more than for λ (at $\lambda = 0.5$).

Since no single optimal tolerance under all parameter combinations exists, I chose a compromise tolerance for each parameter that minimizes the error under most parameter combinations and did not lead to qualitatively different conclusions about the relative performance of the estimation under different parameter combinations.

4.3.2 Estimating the admixture proportion

The estimates of the two measures of the admixture proportion, λ and λ_{min}, are not directly comparable because the range of values of the two parameters are different, as well as the distances of the test values from their prior means. So, I compared λ with λ_{min} when samples were available for both the parental and the admixed populations,

and the estimation of λ_{min} with and without parental population samples.

The estimation of the admixture proportion using the classic statistics depended on the true parameter and the availability of samples from the parental populations. When samples were available from both parental populations, λ was generally well estimated. However, the performance depended strongly on the true λ and the time of the admixture event. The best estimates were gained for recent and symmetric admixture events, while estimates of skewed recent admixture events were very poor (Table 4.2). For admixture events of 100 generations ago, λ was reasonably well estimated, but performance varied slightly depending on the measure (Table 4.2). Notice that the absolute bias values of the estimates of λ were always negative, which is due to the log-transformation that was applied to all parameters. I note that, for a parameter which is a proportion, a different transformation, such as logit, would perhaps be more appropriate. Finally, comparing the estimation of λ_{min} with λ, in the presence of parental population samples, the quality of estimation of λ_{min} was much worse than that of λ when the admixture was symmetric. This is due to the fact that $\lambda_{min} = 0.5$ is the edge of the prior distribution. However, λ_{min} was just as well, or even slightly better, estimated for skewed admixture events (i.e. when $\lambda = 0.1$, see Table 4.2).

Table 4.2: Estimation of the admixture proportion λ, when $N_e = 1000$ in both parental and the admixed populations, using 20 freely recombining loci in a finite population with mutation rate $\mu = 5 \times 10^{-5}$. Estimates are based on a sample of 200 diploids from the admixed populations and 25 from each of the parental populations. ABC estimation was carried out using the classic summary statistics (see details in text) with tolerance 0.01. The accuracy of the estimation is shown under different combinations of λ and T_{adm}, based on 500 data sets and expressed in terms of the relative root mean squared error (RelMSE), the absolute bias, median absolute deviation (MAD), the 95% coverage and Factor 2 (see more details in text).

Samples	**Parameter**	λ	T_{adm}	**RelMSE**	**Abs. bias**	**MAD**	95% **Cov.**	**Factor 2**
With parental samples	λ	0.1	5	0.475	-0.046	0.043	0.667	0.646
		0.5		0.122	-0.021	0.040	1.000	1.000
		0.1	100	0.287	-0.015	0.018	0.979	0.926
		0.5		0.193	-0.058	0.060	1.000	0.990
With parental samples	λ_{min}	0.1	5	0.391	-0.037	0.032	0.980	0.848
		0.5		0.309	-0.140	0.112	1.000	0.968
		0.1	100	0.319	-0.012	0.020	1.000	0.926
		0.5		0.443	-0.211	0.192	0.958	0.719
Only admixed sample	λ_{min}	0.1	5	0.281	-0.021	0.015	1.000	0.970
		0.5		0.459	-0.220	0.205	0.968	0.747
		0.1	100	0.443	0.001	0.024	1.000	0.872
		0.5		0.490	-0.236	0.225	0.896	0.552

As expected, the estimation of λ_{min} was generally poorer when only the admixed sample was used in comparison to the case when samples were available from the parental populations. For example, the RelMSE was about 10 to 50% higher when no parental population samples were included (Table 4.2). In contrast, surprisingly, for skewed and recent admixture events, λ_{min} is slightly better estimated from only the admixed population than from all three samples (Table 4.2, Figure 4.4). I note that this rather unusual finding is not an artefact of using the posterior mode as the point estimate. The regression fitted values and the posterior median also showed the same trend. For example, using the regression fitted values as point estimates, the RelMSE was 0.457 for λ estimated from both parental and the admixed sample, 0.383 for λ_{min} when source population samples were available, and 0.239 when only the admixed sample was available.

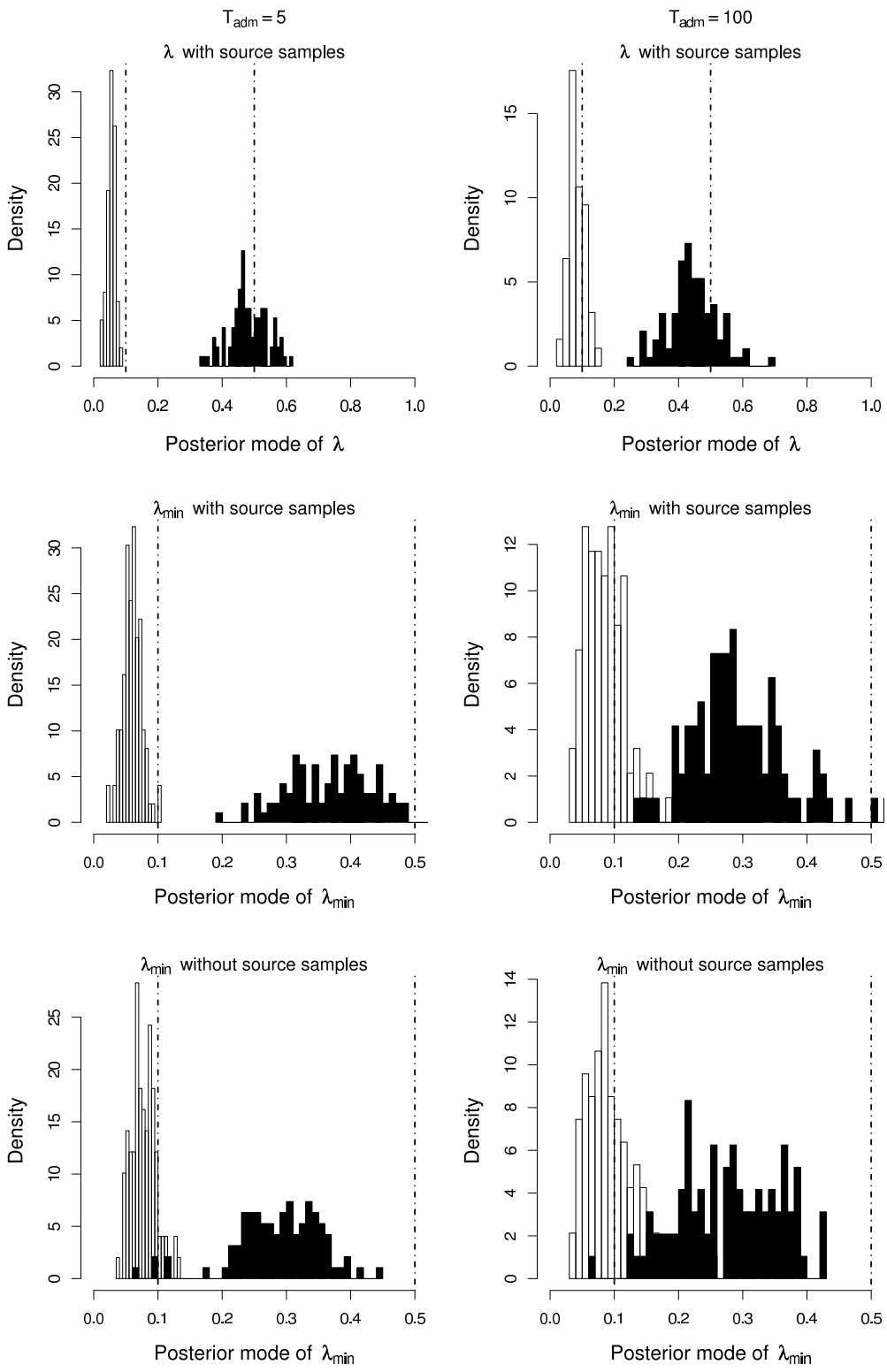

Figure 4.4: Histograms of posterior modes of λ and λ_{min} with tolerance 0.005 in the ABC scheme. White bars represent posterior modes when the true λ is 0.1, and black bars represent cases when the true λ is 0.5. Dashed vertical lines represent the true values of λ and λ_{min}. Note that prior ranges for λ and λ_{min} are $0, 1$ and $0, 0.5$, respectively.

I found that the use of LD information could improve the estimation of λ, but only when the admixture was skewed (see Figure 4.5). At the most, LD information decreased the RelMSE by 15%. For older admixture events LD statistics did not improve the estimation. In fact, for admixture 500 generations ago adding the LD statistics provided slightly poorer estimates than only the classic statistics Figure 4.5.

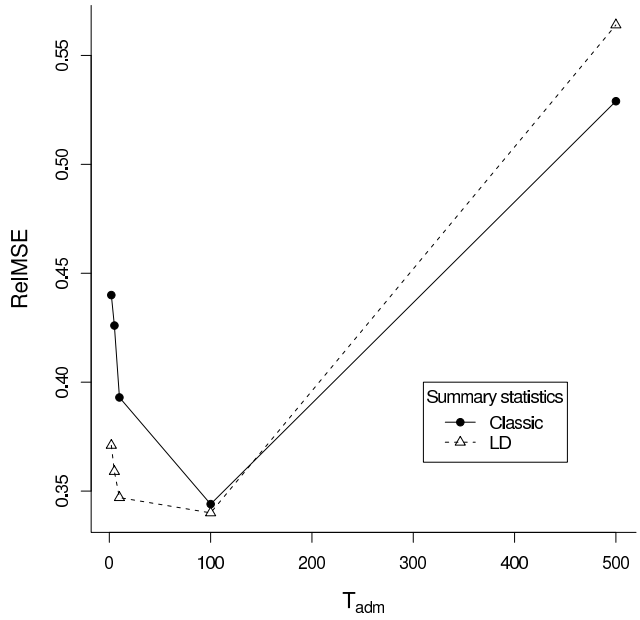

Figure 4.5: Relative mean squared error (RelMSE) of λ_{min} as a function of the time of admixture (T_{adm}) for the classic and LD statistics for $\lambda = 0.1$ and tolerance of 0.05.

4.3.3 Estimating the time of admixture

My primary aim was to assess the estimation of T_{adm} independently from N_e and μ. I found that with the classic statistics T_{adm} was greatly overestimated, especially for recent admixture events (Table 4.3). For example, when the admixture event was 5 generations ago, the posterior modes of T_{adm} were, on average, as high as 60 generations (Figure 4.6). I note that there was a similar (or sometimes even larger) bias when using the simple rejection sampling (RS) with the classic statistics. Thus the observed bias was not due to the regression adjustment (results not shown). For older admixture events, the bias lessened. However, this does not necessarily indicate more accurate estimation. The smaller bias could also indicate that little information about T_{adm} was extracted via the classic statistics, thus the bias was less because the true values were closer to the prior mean, which is 500 (Figure 4.6, Table 4.3). In fact, inspection of the posterior densities shows that the posteriors are consistently much less peaked as the admixture becomes older (Figure 4.7). When no parental population samples were available, the estimation of T_{adm} had an even higher positive

bias for recent admixture events, but the bias was much less for old admixture events in comparison to the case when parental population samples were also available (Figure 4.8).

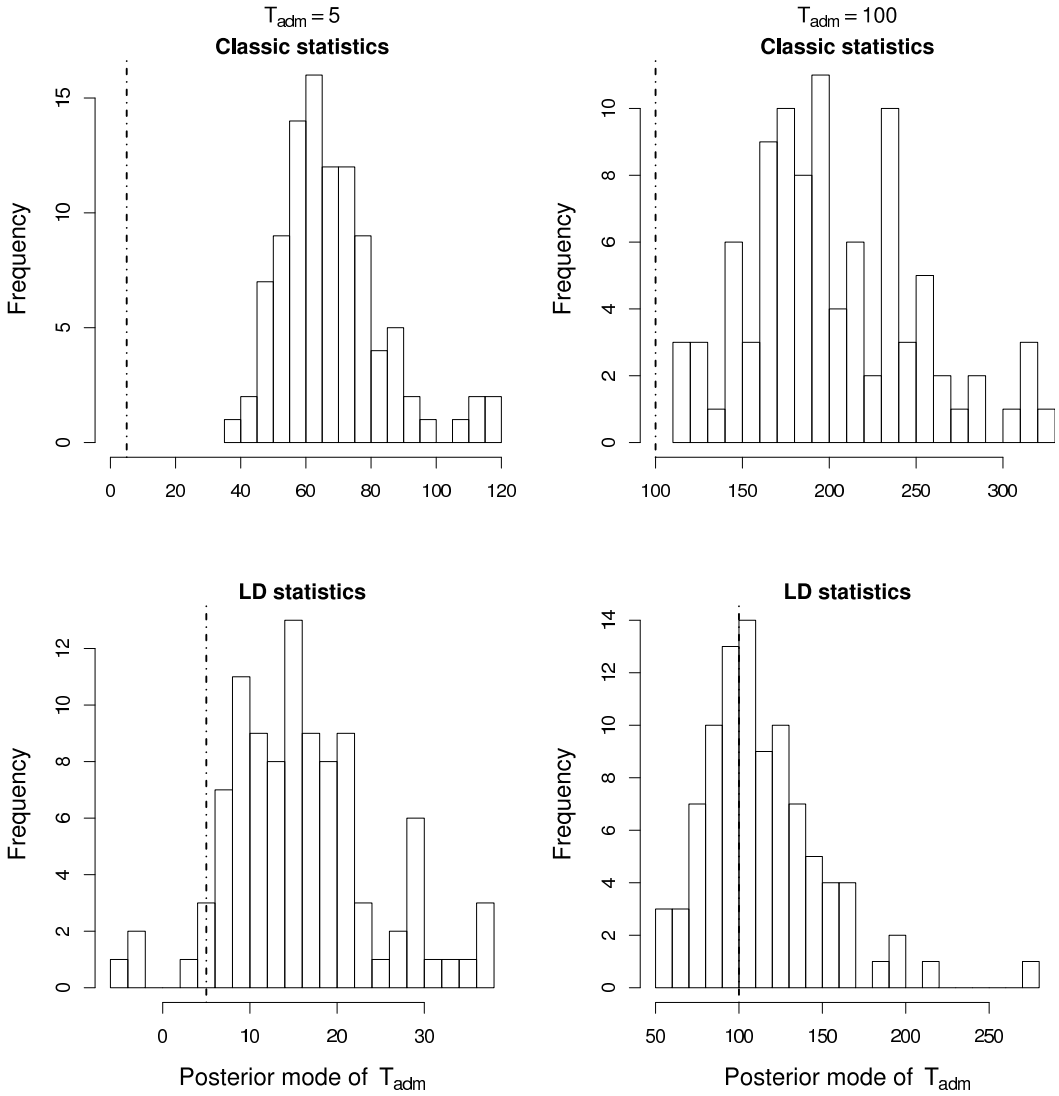

Figure 4.6: Histograms of posterior modes of T_{adm} when using the classic or LD statistics for two values of T_{adm}, 5 and 100, and with a tolerance of 0.01. Dashed vertical lines represent the true values of T_{adm}.

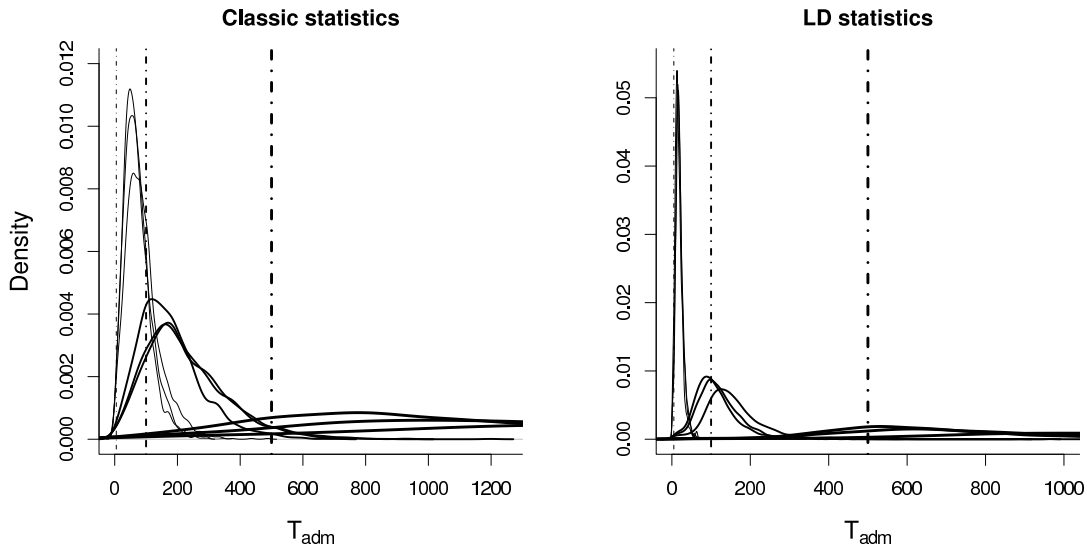

Figure 4.7: Posterior distributions of T_{adm} when using the classic or LD statistics. On each plot, three examples of the posterior distributions are shown for three true values of T_{adm} (i.e. six curves in total). The thin curves correspond to the most recent admixture event, $T_{adm} = 5$, and the thicker curves are for older admixture events ($T_{adm} = 100$ and 500 generations, respectively). The true values of T_{adm} are shown by vertical dashed lines.

The estimation of the time of admixture was greatly improved by the use of LD statistics, especially for the recent admixture events (Figure 4.6, Table 4.3). The fact that the estimation of T_{adm} was greatly improved for recent admixture events is not surprising: admixture LD due to the allele frequency differences between the two parental populations is present following an admixture event. However, it was unexpected that even for the 100 generations old admixture events, the bias of the estimation with LD statistics was reduced in comparison to classic statistics (Figure 4.8). In contrast, contrary to my expectations, the mode-counting statistics did not improve the estimation of old admixture events, but provided just as bad, or even poorer estimates, compared with the classic statistics (results not shown).

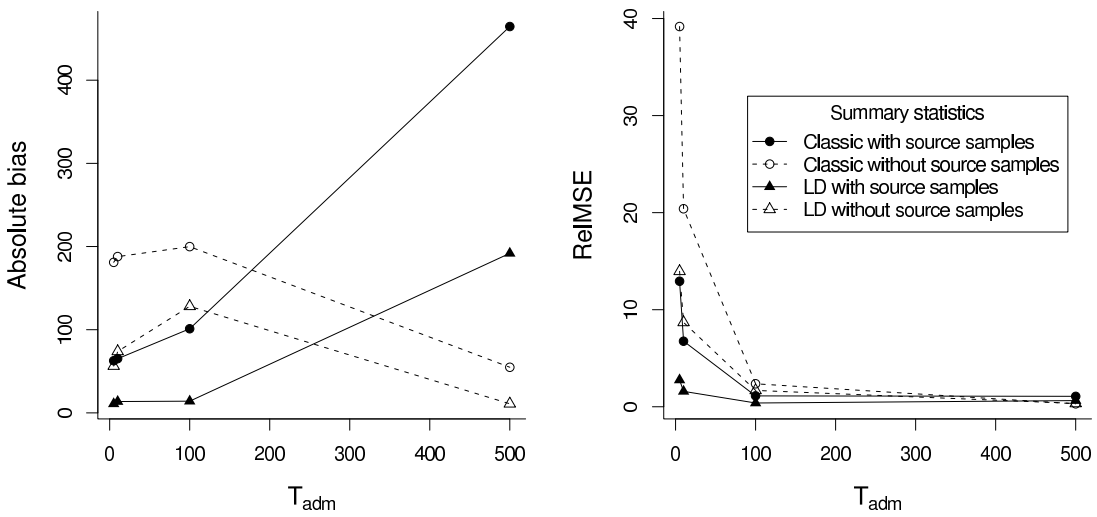

Figure 4.8: Comparing the relative mean squared error (RelMSE) for estimates of the time of admixture (T_{adm}) as a function of the parameter itself, under four different sets of summary statistics: classic and LD statistics, in the presence and in the lack of source samples. Note that $T_{adm} = 2$ is not plotted because of its high RelMSE with the classic statistics when no source population samples were available.

The estimation of the scaled time of admixture, τ_{adm}, was generally more accurate than that of T_{adm}. For recent admixture events, τ_{adm} was also better estimated with LD statistics: its RelMSE was two to five times less than with classic statistics, depending on the admixture proportion (Table 4.3). However, as opposed to the estimation of T_{adm}, there was no advantage of using LD statistics for old admixture events. In fact, the RelMSE was $20 - 25\%$ less with classic than with LD statistics (Table 4.3).

Table 4.3: Estimation of the time of admixture T_{adm} and the scaled time of admixture, τ_{adm}, when $N_e = 1000$ and samples from both parental and the admixed populations are available, using 20 freely recombining loci in a finite population with mutation rate $\mu = 5 \times 10^{-5}$. Estimates are based on a sample of 200 diploids from the admixed populations and 25 from each of the parental populations. ABC estimation was carried out using a tolerance of 0.01. Accuracy of the estimation is shown for two sets of summary statistics, classic and LD (see text for details), based on 500 data sets and expressed in terms of the relative root mean squared error (RelMSE), the absolute bias, median absolute deviation (MAD), the 95% coverage and Factor 2 (see more details in text).

Parameter	**Statistics**	T_{adm}	λ	**RelMSE**	**Relative bias**	**MAD**	95% **Cov.**	**Factor 2**
T_{adm}	Classic	5	0.1	12.932	12.536	74.033	0.000	0.000
			0.5	15.378	15.194	95.058	0.000	0.000
		100	0.1	1.119	1.011	126.096	1.000	0.574
			0.5	1.384	1.258	167.783	0.990	0.354
	LD	5	0.1	2.762	2.178	12.103	0.747	0.222
			0.5	0.982	0.229	3.310	0.874	0.474
		100	0.1	0.388	0.141	27.889	1.000	0.979
			0.5	0.365	0.110	25.443	1.000	0.979
τ_{adm}	Classic	5	0.1	3.715	3.591	0.019	0.808	0.000
			0.5	3.876	3.834	0.020	0.853	0.000
		100	0.1	0.364	−0.309	0.029	0.968	0.872
			0.5	0.331	−0.228	0.024	0.979	0.927
	LD	5	0.1	1.855	1.632	0.008	0.990	0.253
			0.5	0.924	0.666	0.003	0.979	0.684
		100	0.1	0.455	−0.413	0.041	0.830	0.617
			0.5	0.440	−0.400	0.043	0.833	0.719

Encouraged by the fact that the estimation of T_{adm} greatly improved via the use of LD statistics, I investigated the use of LD information between linked loci as well. The advantage of LD information is expected to be even stronger between linked loci, as LD is maintained for longer. However, the comparison of the parameter estimation between freely recombining and linked markers is not obvious because the genetic models are different in the two cases. In fact, the estimation has to be performed using a different set of simulations, thus, the number of simulations required could well also be different. We might expect that more simulations are required for the estimation using linked loci because this model has more parameters. I have not explored the number of simulations required for the two genetic models, but I did determine an optimal tolerance level in both cases for the same fixed number of simulations ($500,000$).

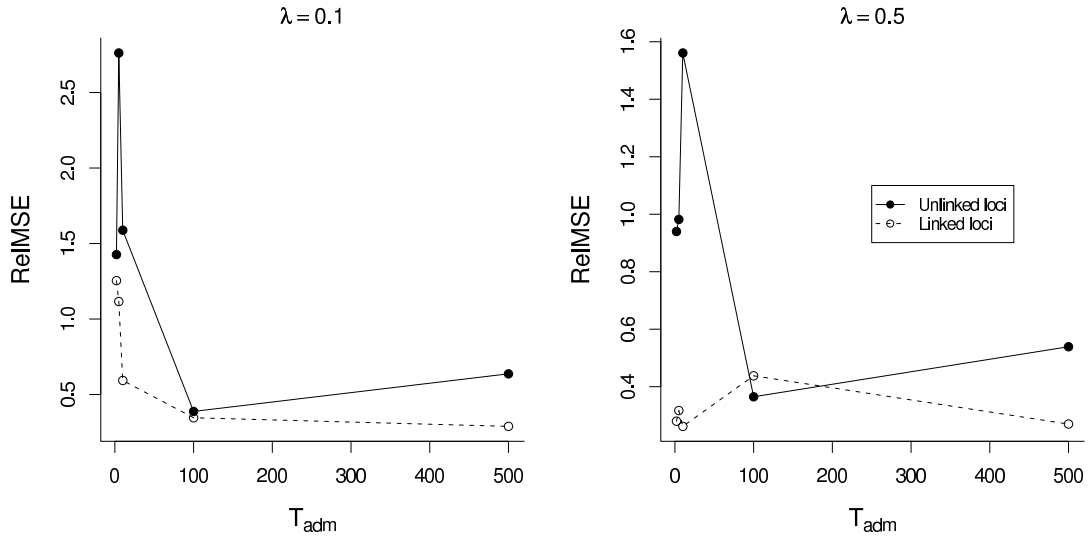

Figure 4.9: Comparing the relative mean squared error (RelMSE) of the estimates of the time of admixture (T_{adm}) as a function of the parameter itself. Two cases are compared: 20 freely recombining loci and 20 loci in two linkage groups, each with 10 loci and recombination rate, $c = 0.01$ between them. Other details are the same as for Table 4.3.

Figure 4.9 shows that, even for an admixture event as old as 500 generations, LD statistics improved the estimation of T_{adm} when linked loci were used. This is a rather surprising result, because, even between linked loci, admixture LD is generally thought to break down much more quickly, albeit based on the theoretical expectation between two independent biallelic loci (*e.g.* LYNCH and WALSH, 1998, p. 96). However, my results indicate that, between linked multiallelic loci (recombination rate, $c = 0.01$ between loci and 5-10 alleles per loci on average), admixture LD can be maintained for hundreds of generations. Ideally, we would like to tease apart whether, in an admixed population,+ LD is present because of finite population size or because of differences

in the allele frequencies in the parental populations. However, this is a challenging task because there is no way to simulate a population with pre-specified allele frequencies and a random level of LD appropriate for a finite population. In other words, we cannot simulate a "control" non-admixed population.

In order to, partly, overcome the difficulty of not being able to simulate a "control" non-admixed population I calculated a proxy for the amount of LD immediately after the admixture. My proxy for this initial LD (LD_{init}) was LD in the pooled sample of the two parental population samples, which is my best guess for the LD in the admixed population at generation zero. For this exercise, I calculated LD in a new set of test data sets with a wider range of values for T_{adm} and for fixed values for the other parameters. I used $N_e = 1000$ for all three populations and $\lambda = 0.3$. I draw large samples from the parental populations (200 diploid individuals) so that I was able to create an admixed population sample of the same size as the true admixed sample, using sampling without replacement from the parental populations. This was necessary, because LD is dependent on the sample size (Chapter 3, McRae *et al.*, 2002).

I defined LD_{diff} as the difference between LD_{init} and LD in the admixed population after T_{adm} generations. I found that, for some LD measures, the difference in LD_{diff} between linked and unlinked loci disappears between 200 and 500 generations (e.g. the LRT statistic, Figure 4.10), while for other LD measures the difference is maintained and significant even at $T_{adm} = 1000$ (e.g. χ^2 and D' statistics, Figure 4.10). The r^2 statistic is somewhat different from the other statistics (Figure 4.10). It also exhibited a difference in LD_{diff} between linked and unlinked loci, but LD_{diff} does not show the expected increase with T_{adm}.

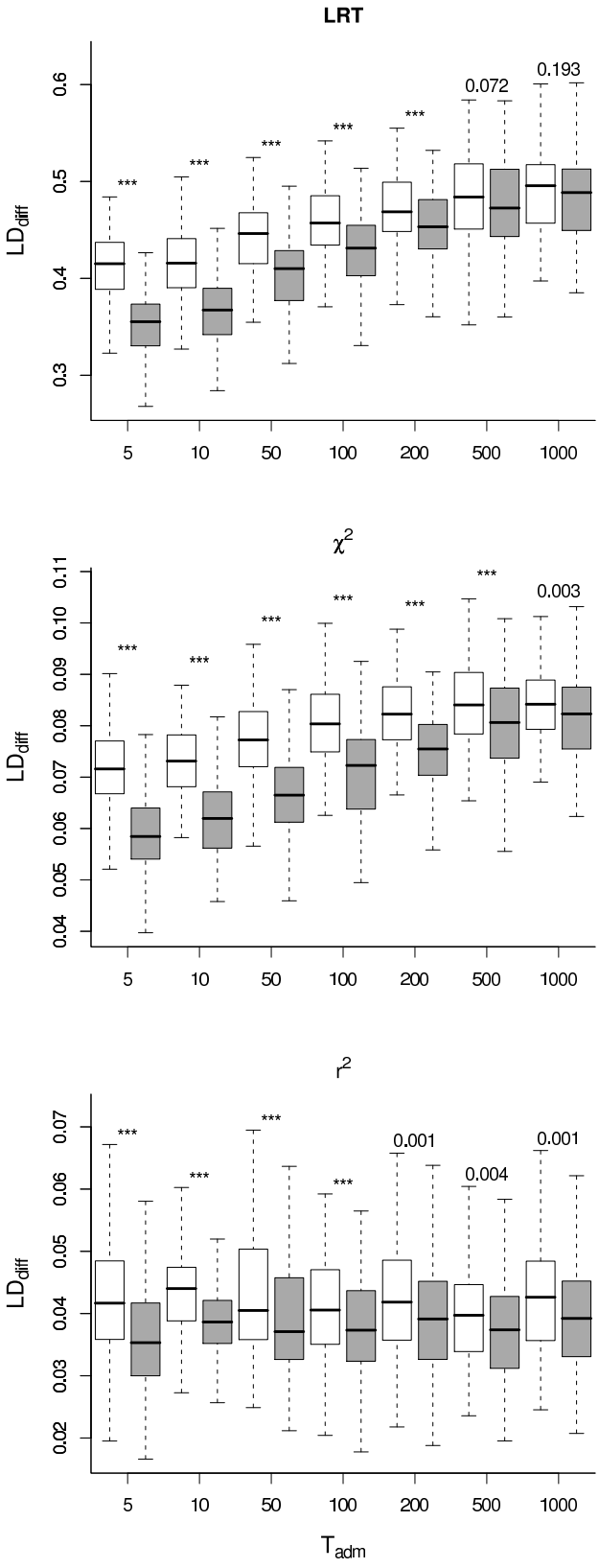

Figure 4.10: LD_{diff} as a function of T_{adm} for two measures of LD, LRT and r^2. LD_{diff} is the difference between a proxy for the LD at the admixture event (LD_{init}) and the LD in the admixed population after T_{adm} generations. White boxes show LD_{diff} between unlinked loci and grey boxes between linked loci (recombination rate 0.01). Above each pair of boxes the p-value of a Wilcoxon test is shown, "***" indicates, p-value less then 0.0001.

4.4 Discussion

4.4.1 General performance of the ABC scheme

ABC is increasingly used to infer demographic parameters and/or distinguish between alternative population history scenarios (*e.g.* EXCOFFIER *et al.*, 2005; INGVARSSON, 2008; ROSENBLUM *et al.*, 2007; ESTOUP *et al.*, 2004; FAGUNDES *et al.*, 2007). The main attractions of ABC are its flexibility and conceptual simplicity. In theory, inferences can be made under models of any complexity, as long as data can be simulated under the model and suitable summary statistics can be found (BEAUMONT *et al.*, 2002). The basic idea of the rejection sampling approach (MARJORAM *et al.*, 2003; PRITCHARD *et al.*, 1999), which underlies ABC, is conceptually simple and accessible for biologists. Along with numerous data simulation software packages, such as Simcoal2 (LAVAL and EXCOFFIER, 2004) or ms (HUDSON, 2002), the first software packages have started appearing, which combine scripts to automate the whole estimation process, such as msBayes (HICKERSON *et al.*, 2007) or DIYABC (CORNUET *et al.*, 2008), even further facilitating the use of ABC. However, the advatages of ABC are its disadvantages as well. Since ABC does not require the calculation of the likelihood, our inferences relay on summary statistics, whose effects are hard to predict, so we have no way to evaluate how far our posterior dstribution is from the true posterior (MARJORAM and TAVARÉ, 2006).

The accuracy of estimation with ABC probably depends most importantly on how much information is captured about the parameters by the summary statistics. Ideally, we want some low-dimensional sufficient or "nearly sufficient" statistics with which the simulation method is fast, but which retain all or most of the information in the data about the parameter of interest. Such statistics are, however, not available for most population genetics problems, so we are left to choose statistics based on our intuition. In most problems, researchers have selected summary statistics that are widely used in population genetics, such as heterozygosity, F_{ST}, or Tajima's D (BEAUMONT *et al.*, 2002; EXCOFFIER *et al.*, 2005; ESTOUP *et al.*, 2004; ROSENBLUM *et al.*, 2007). More recently, JOYCE and MARJORAM (2008) developed an algorithm to choose statistics based on the effect of their inclusion in an empirically calculated posterior. Nevertheless, the effects of using a particular set of statistics remains difficult to predict, because we cannot explore the space of all possible combinations of the available statistics. Perhaps due to the inherent subjectivity of the selection of statistics most empirical and simulation studies used only one set of summary statistics, and did not attempt to address the question of the choice of statistics.

Here I attempted to compare the performance of the estimation with ABC under different sets of statistics. I not only found that different sets of "sensibly" choosen summary statistics may lead to estimates of dramatically different accuracy, but also that the choice of summary statistics cannot be studied independently from other aspects of the ABC scheme. Using more statistics increases the dimensionality, which in turn requires a greater number of simulations or, for a fixed number of simulations, a higher proportion of accepted simulations (i.e. tolerance level) (BEAUMONT *et al.*, 2002). This will, however, introduce a bias towards the prior mean. The key benefit of ABC over RS is using approximations that are insensitive to the tolerance, and this can permit us to increase the number of summary statistics used, and also widen the tolerance level. BEAUMONT *et al.* (2002) showed that, in a simple model of a growing population the additional information from a larger number of statistics outweighs the increased bias towards the prior mean. In agreement with BEAUMONT *et al.* (2002) I also found that adding "good" statistics (e.g. adding LD statistics) improved the accuracy of the estimation, although this required a larger tolerance level. However, when I added statistics that, supposedly, did not extract any information from the data about the parameter (for example the mode-counting statistics) the accuracy of the estimation deteriorated.

I argue that ABC is relatively insensitive to the choice of the tolerance level only when the statistics are informative about the parameter of interest. In a demographic scenario that is more complex than that of BEAUMONT *et al.* (2002), I found that the choice of the tolerance can strongly affect the performance of the estimation. Further, often no single optimal level of tolerance exists that maximizes the accuracy for all parameters of the model in question. As a consequence, different levels of tolerance may lead to different conclusions about the relative performance of a parameter in different parts of the parameter space. There are multiple explanations for these observations, but, perhaps most importantly, by increasing the number of accepted values, the distances to the target statistics do not change smoothly. This is probably more and more true as the model gets more and more complex.

There are now a handful of papers that estimate parameters using ABC, but most of them report the accuracy of the parameter estimates for only one tolerance value, or the authors report that, based on a pilot study, estimates were only weakly dependent on the tolerance (*e.g.* INGVARSSON, 2008; PASCUAL *et al.*, 2007), in agreement with BEAUMONT *et al.* (2002). Up to now, there are only a handful of studies reporting estimation performance for different tolerance levels (*e.g.* HICKERSON *et al.*, 2006; ROSENBLUM *et al.*, 2007; HAMILTON *et al.*, 2005). For example, HICKERSON *et al.* (2006) studied two tolerance levels and two sets of summary statistics, including 8 and

32 statistics, and found that parameter estimates were slightly more biased with the larger set of statistics and slightly affected by the choice of tolerance, depending on the true parameter values (Table 3 in HICKERSON *et al.*, 2006). ROSENBLUM *et al.* (2007) reported the accuracy of parameter estimates for two tolerance levels, though only by using a randomly selected set of test data sets from the set of simulations which were used for the ABC estimation. Nevertheless, the authors found a strong dependence on the tolerance. For example, when estimating a relative population size (Figure 4 in ROSENBLUM *et al.*, 2007), estimates were biased in different directions depending on the true parameter value under different tolerance levels. Both the results of HICKERSON *et al.* (2006), and of ROSENBLUM *et al.* (2007), suggest that dependence on the tolerance is not unique to the demographic model I considered in this study.

My study also provided an example of a further potential difficulty regarding cross comparisons between different sets of summary statistics. When making comparisons between different sets of statistics, the model under which we simulate data and the priors should be held fixed. In a strict sense, I violated this assumption by making a comparison between the LD statistics calculated between freely recombining and linked loci. Although the demographic model was the same in both cases, the genetic model was not. Thus, the priors for some parameters were different, for example for $\rho = 4N_e c$.

4.4.2 Improvement via LD statistics

My study demonstrated that LD statistics greatly improve the accuracy of the estimation of the time of admixture and, under some parameter combinations, the estimation of the admixture proportion. The improvement is relative to a "sensibly" choosen set of well-known population genetic statistics that are thought to capture information about population demography. Contrary to expectations, I found that the improvement via LD statistics can be detected between freely recombining loci for up to 100 generations after the admixture events and between tighly linked loci after admixture events as many as 500 generations. Previous studies have also tried to accommodate measures of LD in an ABC scheme, but found no noticeable improvement, partly due to not studying the effect of LD statistics directly (BEAUMONT *et al.*, 2002; EXCOFFIER *et al.*, 2005). Also, LD statistics are often not considered, because unlinked loci are used. However, significant LD may be present between unlinked multiallelic loci in a finite population, as I illustrated in Chapter 3. An ABC scheme could benefit from the presence of "background LD" between unliked loci when loci are simulated with the coalescent with recombination

with recombination rate of 0.5. EXCOFFIER *et al.* (2005) also considered the used of linked loci in an admixture context, however, in their study linkage was modeled as some sort of noise because genetic map information was not used. As a result, they found that estimates of the scaled time of admixture were more biased with linked than with unlinked loci. Here I explicitly accounted for linkage, via using genetic map information, similarly to software Structure, where map information can also be supplied and it improves performance (FALUSH *et al.*, 2003).

Theory predicts that $D_{AB} = p_{AB} - p_a p_b$, the difference between the frequency of the haplotype AB and the product of the frequencies of alleles A and B, decreases with time as a function of the recombination rate: $D_{AB}(t+1) = (1-c)D_{AB}t$, where t is the time in generations and c is the recombination rate (JENNINGS, 1917). Under this simple relationship, 10, 100, and 500 generations after the admixture event 8.6, 63, and 99% of the LD is expected to be lost between loci with a recombination rate of 0.01, which does not explain why I found LD to be informative for admixture events 100's of generations old. These figures, however, only probabilistic expectations for each haplotype AB, of which there are many between multiallelic loci. Thus, the chances that LD is present after 100's of generations are much higher between multiallelic loci.

The fact that the estimation of the admixture proportion was also improved via the LD statistics was an unexpected finding. The fact that λ was estimated more accurately with LD statistics only when the admixture was skewed suggests the following explanation. The allele frequencies between the two source populations are more likely to be different for a skewed than for a symmetric admixture event due to sampling, which, in turn, results in higher admixture LD in the admixed population. However, this idea was not tested directly.

Making inferences about admixed populations without samples from the source populations is the most common scenario when analyzing real data. Nevertheless, it can be difficult and inaccurate (*e.g.* FALUSH *et al.*, 2003; PRITCHARD *et al.*, 2000b). I also found that λ and T_{adm} are more accurately estimated when samples were available from the source populations. However, I found that for recent admixture events the estimation of T_{adm} using LD statistics in the absence of source population samples was more accurate than using classic statistics with source population information (Figure 4.8). The admixture proportion, λ_{min}, was also well estimated from only the admixed population sample, in fact, with a comparable accuracy to λ from the three population samples. These findings suggest that there might be an advantage for using ABC over the software Structure when only admixed population samples are available. However, direct comparison of the two methods is needed to confirm this.

4.4.3 Future directions

Although the performance of ABC under a given model can be tested with almost no extra computational effort (BEAUMONT *et al.*, 2002), thorough classical sense performance tests are difficult to carry out for more than a limited set of parameter combinations. Here I chose to assess the choice of summary statistics, which is an increasingly common problem JOYCE and MARJORAM (2008); MARJORAM and TAVARÉ (2006). Since it is difficult to predict which statistics will provide the most information about a parameter, JOYCE and MARJORAM (2008) recently suggested an algorith to aide the choice of summary statistics. Here I found that LD information between multiallelic loci greatly improved the performance of the estimation. However, there are many different LD measures and they might well measure different aspects of the departure from linkage equilibrium (see Chapter 3 and SABATTI and RISCH, 2002). This was also well illustrated by the fact that LD measured with different statistics decayed differently with T_{adm} (Figure 4.10). Thus, a potential future project could investigate more explicitly the use of different LD statistics to distingush between different sources of LD, potnetially via an algorithm similar to JOYCE and MARJORAM's (2008).

Sensitivity analysis to the priors is an important step in all Bayesian inference, but it is often computationally demanding to carry out. Here I have not investigated the choice of the prior distributions, though, I reported many cases where the prior distribution had a strong effect on the estimation. I found that it is particularly challenging to choose a prior for the admixture proportion, because the estimates are strongly affected if the true values are either too close to the prior mean, or to the boundaries. I suggest that, in a real estimation problem, priors other than the uniform should be considered.

Another important aspect of all rejection sampling methods, or more generally all methods that use simulate data, is to perform a goodness-of-fit check, i.e. to see if the data set simulated under a simplified model are similar to the real data (MARJORAM and TAVARÉ, 2006). Here I did not analyse real data. I note, however, that with microsatellite loci it is potentially difficult to simulate "real data-like" data because of the complex mutation process (Chapter 1). For example, one could speculate that, using a different value of the Geometric parameter (for example, zero, thus the SMM) mode-counting statistics could capture more information because the allele frequency distribution would be less dispersed.

Finally, I emphasize the importance of comparisons of ABC to full-data likelihood methods, even through they available only in a limited number of cases. For example, BEAUMONT *et al.* (2002) compared ABC with MCMC in an exponential population

growth model for the human Y chromosome data (PRITCHARD *et al.*, 1999), and showed that MCMC consistently outperforms ABC. In contrast, EXCOFFIER *et al.* (2005) compared ABC with WANG's (2003) maximum likelihood method and found that estimates of admixture proportions are similarly accurate in the case of recent admixture events, and that ABC outperforms WANG's (2003) method when the admixture is old. However, WANG's (2003) method provides a somewhat limited inference by focusing on the admixture proportion only. A valuable future study would be to compare the utility of ABC for estimating admixture parameters with the widely used software Structure (FALUSH *et al.*, 2003).

Chapter 5

The future of microsatellites

5.1 Introduction

In this thesis I provided a general overview of statistical inference with microsatellites in population genetics. Specifically, I covered three topics in detail, as examples for making inferences from microsatellite marker data. My aim was to provide a balanced view of the use of microsatellites in population genetics. On the one hand, I showed that the use of microsatellites revolutionized the population genetics of natural populations by providing a highly polymorphic, presumably neutral marker that is abundant in most eukaryotic genomes organisms, and, importantly, accessible for a wide range of species. On the other hand, there are many potential pitfalls when making inferences from microsatellite data, some of which I illustrated in this thesis. Difficulties generally arise because of the complex evolution and high mutation rate of microsatellites, which are both difficult and/or inefficient to incorporate in statistical models.

At the time of writing this thesis many new molecular genetics technologies are being developed, which might soon drastically change the population genetics of natural populations (*e.g.* HUDSON, 2008). Current developments suggest that in the future not only other types of markers, such as single nucleotide polymorphisms (SNPs) will become more accessible and economic, but future genetic data might also include whole genome and whole population sequence data from natural populations. The aim of this Chapter is dual. First, I will briefly review the new technological advances and the new data types that they could produce. Second, I recall the most important findings of this thesis and I identify some related research questions that have been addressed or would be worth investigating in the future in the light of the current developments in molecular genetics.

5.2 Genetic data of the future

Starting from around the mid-nineties, more and more studies started using single nucleotide polymorphisms (SNPs), instead of microsatellites. Although SNPs are biallelic, thus a single SNP is less informative then a microsatellite, SNPs offered two main practical advantages that played the major role in driving this methodological shift. First, SNPs are much more abundant in the genome then microsatellites. For example, in the human genome, on average, there is a SNP every 100 to 300bp (Human Genome Project, `http://www.orni.gov/sci/techresources/Human_Genome/home.shtml`), which offers a whole-genome coverage on a much finer scale than microsatellites. Second, SNP genotyping can be highly automated since SNP assays only require three genotypes to be distinguished, as opposed to the often tedious and sometimes unreliable genotype scoring of microsatellites, where for instance, PCR artefacts (false, so-called "stutter bands") complicate automated genotype scoring (*e.g.* HOFFMAN and AMOS, 2005).

SNPs have indeed revolutionized human genetics: in the past decade, association mapping with the ambitious aim of mapping complex disease genes using high-throughput SNP genotyping (SNP chips) rapidly took over the place of linkage mapping using microsatellites (CARLSON *et al.*, 2004). The newborn interest and large investment in complex disease genetics was clearly motivated by the fact that whole-genome scans for genetic polymorphism could be economically carried out. SNPs for model organisms soon followed, and for species with medical or economic importance (*e.g.* VIGNAL *et al.*, 2002). However, the marked change in marker preference has not yet reached non-model organisms; the main limitation being that SNP discovery requires *a priori* knowledge of the presence of an allelic variation. Thus, either whole or, at least, partial genome sequence is required from multiple individuals, which is not readily available for most species, apart from model organisms and their close relatives.

Due to the difficulties of SNP discovery, until recently, only a few pioneer studies explored the utility of SNPs in natural populations (*e.g.* BENSCH *et al.*, 2002; SEDDON *et al.*, 2005). However, the scene is about to change: the availability of new, low cost, so-called "next-generation" sequencing technologies, might dramatically change what is possible in natural populations (*e.g.* HUDSON, 2008). The next-generation sequencing technologies have the potential to re-sequence genotypes of complex eukaryote genomes. For example, the 454 implementation of pyrosequencing (Branford, CT, USA, `http://www.454.com`) generates small sequence reads, as opposed to whole-genome sequences and has already been successfully used,

for example, for SNP discovery and microarray design in a wild population of the Glanville fritillary butterfly (*e.g.* VERA *et al.*, 2008). A potentially much cheaper, whole genome re-sequencing technology is developed by Solexa (Hayward, CA, USA, `http://www.illumina.com`). The re-sequencing technology of Illumina has been used, for example, to discover SNPs to study population structure and local adaptive trends in a natural population of white spruce (NAMROUD *et al.*, 2008).

SNP discovery, however, is just one possible use of the next-generation sequencing technologies in natural populations. Almost all marker types, and SNPs particularly, suffer from so-called ascertainment bias (AB). AB is a systematic bias due to arbitrary decisions during sampling, e.g. to using a limited panel of individuals to discover the polymorphic sites, which biases the SNP discovery process towards the identification of loci with common alleles. Ascertainment bias has been shown to notably bias estimates of various population genetic parameters, such as linkage disequilibrium or the scaled mutation rate (θ) in humans (*e.g.* KUHNER *et al.*, 2000; WAKELEY *et al.*, 2001), in model organisms (*e.g.* BRANDSTROM and ELLEGREN, 2008), and in natural populations (ROSENBLUM *et al.*, 2007). Although, AB can often be overcome using appropriate bias correction, sequence determination for multiple individuals in a population, using re-sequencing approaches, could provide fine grained information from natural populations, which would not suffer from any sampling biases. Large scale, re-sequencing studies have already started appearing in human genetics (*e.g.* KRYUKOV *et al.*, 2009).

Much of these new developments are not yet in the practice of most population genetics studies of natural populations. However, more and more researchers are facing the choice between microsatellites and SNPs, which is expected to be a more and more relevant question in the future. This is indicated by the numerous quantitative comparisons that have been carried out between the two marker types; many in a mapping context in human genetics, but for many classic population genetic parameters as well. Such comparisons are important because for many questions it is not instantly obvious if microsatellites or SNPs and how many of them are the optimal choice. Finally, although whole-genome and whole-population sequence data is not yet used for the population genetics of most species, it is worth speculating what these new data would provide in the context of the questions that I addressed in this thesis.

5.3 The future of the findings of this thesis

5.3.1 Relatedness: can we increase accuracy?

The concept of genetic relatedness is important in many theoretical and applied fields, such as agriculture, genetic mapping, conservation, and studies of kin selection. Estimating relatedness, accurately, will always be of great interest. Generally, inferences about genetic relatedness fall into two categories: first, estimating an unknown degree of relatedness and, second, distinguishing between a set of alternative genetic relationships. In both types of inferences there is a tradition of using microsatellites, which might now be challenged by the availability of SNP chips.

Estimating an unknown degree of relatedness in natural population samples using microsatellite data was the focus of Chapter 2. I investigated the average performance of relatedness estimators across all pairs of individuals. Although this is a special aspect of the performance of relatedness estimation, my findings are in accordance with earlier studies that all indicate that accurately estimate an unknown degree of relatedness from population samples is challenging (VAN DE CASTEELE *et al.*, 2001; MILLIGAN, 2003; CSILLÉRY *et al.*, 2006). This is because the sampling variance of the most common genetic relationships is high, so that in a population it is impossible to reliable distinguish between them. In particular, I showed that the proportion of variance explained in the pairwise relatedness estimates by the true population relatedness composition (r^2) is generally very low. I also showed that it is the population relatedness composition that sets a limit on average performance, and marker data quality can improve the performance only within the this limit. What remains unclear is the upper limit in performance, which is set by the availability of the marker data.

The question whether modern SNP chip data could improve relatedness estimation arises, not only when estimating the average performance of relatedness estimation in a population, but in all types of relatedness inferences. The question whether many SNPs would be better than a few microsatellites has been quantitatively addressed in many areas of relatedness estimation. Next, I will detail some examples of such comparisons between microsatellites and SNPs, and then close with two caveats that mainly affect inferences from SNP data: linkage and genomic heterogeneity.

The study of GLAUBITZ *et al.* (2003) is the most closely related to my work. It investigated the performance of the same relatedness estimation methods as I did, but in a comparative study between microsatellites and SNPs. Their result show that a panel of 100 independent moderately polymorphic SNPs will provide the discrimination

power equivalent to 16–20 independent microsatellite loci (GLAUBITZ *et al.*, 2003), which given today's SNP technology, is not a full scale comparison. Two further studies, using data from natural populations investigated the relative performance of the two marker types, but it is difficult to draw quantitative conclusions from either because of the limited set of markers they used. The study of SEDDON *et al.* (2005) concluded neither the 22 SNP nor the 20 microsatellite loci studied were sufficient to discriminate first order relationships, while the study of RENGMARK *et al.* (2006) showed that both sets of markers (16 microsatellites and 26 SNPs) were highly accurate in parentage analysis, and microsatellites performed slightly better in population assignment test.

In human as forensic and parentage analysis, more quantitative comparisons have been performed between microsatellites and SNPs. The motivation for such studies is partly economic, SNPs are cheaper, but they also provide much finer scale and more reliable data. For example, the human HapMap project (`http://www.hapmap.org`) provides over a million SNPs per individual. With such data, is there still a place for the "old fashioned" microsatellites? In the case of genetic relationship inference, there might be: a great advantage of multiallelic markers is that it is possible to distinguish between all possible classes of genotype pairs between two individuals. As a result, in forensic applications, for example, it has been actually shown that as many as eight SNPs would be required to provide the discriminatory power of one microsatellite locus (AYRES, 2005). Thus, forensic applications might continue using the well-established microsatellites (WEIR *et al.*, 2006).

There are two further caveats to mention regarding relationship inference with SNP data: linkage between SNP loci and their genomic heterogeneity. Although today a great number of SNPs is available, this is partly illusory because many of the SNPs will be linked, and thus provide less information about identity-by-descent (IBD) than microsatellites. The use of SNP haplotypes as multiallelic markers might be an efficient design in this setting, which to my knowledge has not been explored. The other concern is heterogeneity in the genome in IBD. There is inherent heterogeneity in IBD along the genome due to recombination and sampling inherent to the evolutionary process. However, SNPs could be anywhere in the genome, and are not necessarily neutral, and there could be variation in IBD due to selection as well. Particularly, SNPs maps show heterogeneity of in actual relatedness along the human genome, which might be due to selection in previous generations (WEIR *et al.*, 2005). Thus, WEIR *et al.* (2005) warn that care has to be taken that relatedness is not estimated from a small fraction of the genome.

5.3.2 Linkage disequilibrium: is there equilibrium at all?

The causes and effects of non random association between alleles at different loci (linkage disequilibrium or LD), has been of interest in population genetics for a long time. Many factors affect LD, such as mutation, drift, recombination, population history, selection, breeding system etc., and in turn, many factors are affected by LD, such as response to selection (HILL and ROBERTSON, 1966). Since LD is such a sensitive measure of many population genetic forces, linkage equilibrium, random association between alleles at different loci, will almost never be reached. In recent years, LD has attracted a lot of interest, particularly in human genetics: LD is used to human understand population history and evolution, to map disease and quantitative genetic traits and to understand gene interactions. Most of these applications use SNP data, while less attention has been paid to studying LD between other kinds of genetic variants, such as microsatellites, indels or inversions, which could also carry valuable information (SLATKIN, 2008). Here, I will briefly recall my findings of Chapter 3, which illustrated the different problems that could arise when testing for LD between microsatellite or SNP loci, and of Chapter 4, which illustrated the utility of LD between microsatellite loci when inferring the admixture history of a population. I will close with describing an idea that exploits LD between segments of the genome with dramatically different mutation rates to infer the evolutionary history. This could become feasible in natural populations with the arrival of sequence data.

In Chapter 3 I addressed a somewhat old-fashioned hypothesis question, which is well-known and can be stated simply: statistical independence between alleles of different loci (i.e. LD) is not the same as genetic independence (i.e. free recombination). However, it has not been recognized before that the difference between the two null hypothesis (statistical and genetic independence) becomes greater with data informativeness, and thus they are almost identical for SNPs, but dramatically different for polymorphic microsatellites. The practical relevance of this finding is not instantly obvious, since we are rarely interested in testing the zero LD hypothesis *per se*. Nevertheless, I illustrated using examples in the literature and analysis of a real data set how this problem could lead to mis-inference, i.e. detecting linkage between genetically independent loci.

The difference between the statistical and biological null is due to “background LD”, which is LD due to finite population size, i.e. due to genetic drift. Similarly, FALUSH *et al.* (2003) used the term, “background LD” to distinguish drift LD from “mixture LD” and “admixture LD”, which are both due to variation in ancestry among the sampled individuals. FALUSH *et al.* (2003) found that there could be substantial “background-LD”, especially between tightly linked markers. Indeed, LD breaks down

as a function of the recombination rate, and between unlinked loci it decreases with a factor of half in each generation (JENNINGS, 1917). I could take advantage of this fact in Chapter 4 to improve estimates of the time of admixture. In contrast, FALUSH *et al.* (2003) report that software Structure overestimates the time of admixture between tightly linked markers. The advantage of ABC over Structure in this context, is that it automatically takes account of "background LD".

Various measures of LD exist, and generally there is not an agreement about which is the best or the most useful measure (*e.g.* SLATKIN, 2008). The choice of LD statistics is particularly interesting for microsatellite loci, because, when there are many alleles, there could be departure from linkage equilibrium (LE) in many different directions, or, statistically speaking, the alternative hypotheses has many additional degrees of freedom. Thus, it is often impossible to determine the direction of the departure based on a single statistic (SABATTI and RISCH, 2002), and different test statistics could well measure different aspects of departure from LE. Thus, it is worthwhile to contrast the results of Chapters 3 and 4, where I compared the utility of LD measures for detecting linkage and for detecting admixture (coupled with linkage). Generally, my results confirm that different LD statistics measure different aspects of LD: I found that the likelihood ratio test statistic (LRT) was the most sensitive to genetic drift, r^2 was the most informative about the recombination rate, and the D' and χ^2 were the most informative about the admixture between linked loci (i.e. the signature of admixture could be detected for the greatest number of generations).

Throughout Chapters 3 and 4 I assumed that phase (or haplotype frequencies) was known, which is an assumption of all LD statistics. However, often only genotypes can be detected unless one has pedigree information, which is rare in most natural populations and in most human SNP surveys. Although it is common to infer haplotypes from genotype data and then treat it as observed data, this practice could lead to quite wrongly inferred frequencies for rare haplotypes (EXCOFFIER and SLATKIN, 1995). Inferring phase with microsatellite data is generally more difficult then with SNP data, first, because of their higher level of heterozygosity, there are more ambiguous genotypes, which also means that it is more computationally demanding to infer phase. The increasing amount of genomic data also means that more and more loci are genotyped, so that correlations exist between loci, and it is also expected that phasing will be an increasingly important focus of future research.

Finally, I describe an example for using LD between markers of different evolutionary histories to make inferences about population history. The classic study of TISHKOFF *et al.* (1996) used LD between a pair of linked markers, one microsatellite and a partial deletion of an *Alu* insertion, which demonstrated the utility of autosomal

haplotypes and gave evidence for the African origin of humans. Later, MOUNTAIN *et al.* (2002) suggested the general use of autosomal haplotypes composed of a microsatellite and one or more SNPs as the two markers with dramatically different mutation rates provide complementary information. PAYSEUR and CUTTER (2006) used coalescent simulations to estimate to what extent polymorphism patterns at a linked SNP and microsatellite are correlated, and found that, despite the two marker loci sharing their genealogical history, the polymorphism patterns correlate only weakly. This shows that different mutational processes of microsatellites and SNPs generate different data patterns (PAYSEUR and CUTTER, 2006). Theory has also been developed as natural extension of PRITCHARD and FELDMAN's (1996) work for microsatellites to describe the expected correlation between the number of segregating sites and the squared difference in allele size, which can be used in many population genetics applications PAYSEUR and CUTTER (2006).

A handful of empirical studies already confirm that the combined use of microsatellites and SNPs can improve estimates of various population genetic parameters. For example, RAMAKRISHNAN and MOUNTAIN (2004) shows that estimates of genetic diversity are more accurate from such SNPSTRs than solely using microsatellites, and HEY *et al.* (2004) could gain evidence for gene flow between cichlid fish species only when using "HapSTRs', a microsatellite locus and sequence polymorphisms in its immediate flanking sequence. Importantly, the practical limitation of SNPSTRs, which is that linked pairs of SNPs and a microsatellite have to be available, could be overcome with the wild-spread availability of sequencing technology in natural populations.

5.3.3 Computational statistics: can we handle large data sets?

We are witnesses of a rapid development of molecular technology as a result of which more and more data is expected to arrive. The cost of genotyping is continuously decreasing, many genome projects are on their way (e.g. 1000 Genomes Project, KAISER, 2008), and the next-generation sequencing technologies have already started producing data. As a result, importantly, the size and complexity of the data sets are increasing as well. On the one hand, this is good news because more data is always needed to make accurate inferences about parameters of interest. On the other hand, the amount of information that, for example, whole-population sequencing may provide, will be enormous. Thus, we might expect that the analysis of large data sets will be main challenge of the future, as has been already suggested in some recent papers (*e.g.* SLATKIN, 2008; JOYCE and MARJORAM, 2008). Examples where the analysis of the large amounts of data are unresolved are already abundant, here I will mention two well-known examples: multiple testing in association mapping and finding appropriate

data summaries in an ABC scheme.

Multiple testing is one of the most important challenges in association studies at the moment. This problem becomes more severe with more data. When each SNP is tested independently for association with a disease or other trait, an enormous number of tests has to be carried out. With more and more loci genotyped, partly due to re-sequencing, it is difficult to accurately determine the significance of any reported association (*e.g.* HOGGART *et al.*, 2008). This is because the so-called genome-wise significant level is affected by correlations between the single SNP tests, such as LD or SNP distances, and also by the sample size and the choice of test statistics (*e.g.* HOGGART *et al.*, 2008). Further, when one wants to move beyond single SNP analysis and consider SNP interactions, the multiple testing problem becomes completely intractable (*e.g.* CARLSON *et al.*, 2004).

Another example for difficulties arising in relation to the size and complexity of data sets is in approximate methods. The development of such methods, like ABC, were already motivated by the fact that full data likelihood methods are not feasible for large data sets (or with realistic models) (MARJORAM and TAVARÉ, 2006). However, even the approximate methods could be difficult to apply to some large data sets, due to both computational limitations and also the high dimensionality of the data is difficult to summarize. Recently, JOYCE and MARJORAM (2008) proposed an algorithm to choose summary statistics in an ABC scheme. Although the approach seems to choose sensible statistics in most settings, the order in which statistics enter the algorithm matter, and it is not computationally feasible to try all possible orders of statistics JOYCE and MARJORAM (2008).

In conclusion, I suspect that efficiently searching the large data sets for the most informative parts of the large data sets will be important in future method developments. For example, HOGGART *et al.* (2008) proposed a stochastic search method to identify the most informative SNPs in association studies. Or, it has been suggested that the comparison of the pedigree and population based estimates of the recombination rate could be used to detect selection (O'REILLY *et al.*, 2008). I also think that the combined use of different parts of the genome with different evolutionary histories could be a very fruitful approach.

Bibliography

ALLENDORF, W. F. and G. LUIKART, 2007 *Conservation and the Genetics of Populations*. Blackwell Publishing, Malden.

AMOS, W. and A. CLARKE, 2008 Body temperature predicts maximum microsatellite length in mammals. Biology Letters **4**: 399–401.

ANDERSON, E. C. and E. A. THOMPSON, 2002 A model-based method for identifying species hybrids using multilocus genetic data. Genetics **160**: 1217–1229. Software NewHybrids.

ARLT, D., B. HANSSON, S. BENSCH, T. VON SCHANTZ and D. HASSELQUIST, 2004 Breeding synchrony does not affect extra-pair paternity in great reed warblers. Behaviour **141**: 863–880.

ATKINSON, B., 2005 *kinship: mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees*. R package version 1.1.0-7.

AYRES, K., 2005 The expected performance of single nucleotide polymorphism loci in paternity testing. Forensic Sci. Int. **154**: 167–172.

BALLOUX, F., W. AMOS and T. COULSON, 2004 Does heterozygosity estimate inbreeding in real populations. Molecular Ecology **13**: 3021–3031.

BALLOUX, F. and J. GOUDET, 2002 Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. Mol. Ecol. **11**: 771–783.

BARTON, N., 2001 The role of hybridization in evolution. Mol. Ecol. **10**: 551–568.

BEAUMONT, M. A., 2005 Adaptation and speciation: what can F_{st} tell us? Trends Ecol. Evol. **20**: 435–440.

BEAUMONT, M. A. and B. RANNALA, 2004 The Bayesian revolution in genetics. Nat. Rev. Genet. **5**: 251–261.

BEAUMONT, M. A., W. ZHANG and D. J. BALDING, 2002 Approximate Bayesian computation in population genetics. Genetics **162**: 2025–2035.

BEERLI, P. and J. FELSENSTEIN, 1999 Maximum–likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152**: 763–773.

BELL, G. I. and J. JURKA, 1997 The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single step mutation process. J. Mol. Evol. **44**: 414–421.

BENSCH, S., 1996 Female mating status and reproductive success in the great reed warbler: is there a potential cost of polygyny that requires compensation? J. Anim. Ecol. **65**: 283–296.

BENSCH, S., D. HASSELQUIST and T. VON SCHANTZ, 1994 Genetic similarity between parents predicts hatching failure: nonincestuous inbreeding in the great reed warbler. Evolution **48**: 317–326.

BENSCH, S., S. ÅKESSON and D. E. IRWIN, 2002 The use of AFLP to find an informative SNP: genetic differences across a migratory divide in willow warblers. Mol. Ecol. **11**: 2359–2366.

BLOUIN, M., 2003 DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol. Evol. **18**: 503–511.

BLOUIN, M., M. PARSONS, V. LACAILLE and S. LOTZ, 1996 Use of microsatellite loci to classify individuals by relatedness. Mol. Ecol. **5**: 393–401.

BRANDSTROM, M. and H. ELLEGREN, 2008 Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. Genome Res. **18**: 881–887.

BROHEDE, J., C. R. PRIMMER, A. MØLLER and H. ELLEGREN, 2002 Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. Nucleic Acids Res. **30**: 1997–2003.

CARLSON, C. S., M. A. EBERLE, L. KRUGLYAK and D. A. NICKERSON, 2004 Mapping complex disease loci in whole-genome association studies. Nature **429**: 446–452.

CAVALLI-SFORZA, L. L., 1966 Population structure and human evolution. Proc. R. Soc. Lond. B **164**: 362–379.

CHAKRABORTY, R., M. KIMMEL, D. N. STIVERS, L. DAVISON and R. DEKA, 1997 Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA **94**: 1041–1046.

CHEN, Y., C.-H. LIN and C. SABATTI, 2006 Volume measures for linkage disequilibrium. BMC Genet. **7**: 1471–2156.

CHEVERUD, J., 1985 A quantitative genetic model of altruistic selection. Behav. Ecol. Soc. **16**: 239–243.

CHIKHI, L. and M. B. M. A. BEAUMONT, 2001 Estimation of admixture proportions: A likelihood-based approach using Markov chain Monte Carlo. Genetics **158**: 1347–1362.

CHOE, J. and B. CRESPI, 1997 *The Evolution of Social Behavior in Insects and Arachnids*. Cambridge University Press, Cambridge.

CHOISY, M., P. FRANCK and J.-M. CORNUET, 2004 Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. Mol. Ecol. **13**: 955–968.

CLUTTON-BROCK, T., D. GAYNOR, R. KANSKY, A. MACCOLL, G. MCILRATH *et al.*, 1998 Costs of cooperative behaviour in suricates (*Suricata suricatta*). Proc. R. Soc. Lond. B **265**: 185–190.

CLUTTON-BROCK, T., D. GAYNOR, G. MCILRATH, A. MACCOLL, R. KANSKY *et al.*, 1999 Predation, group size and mortality in a cooperative mongoose, *Suricata suricatta*. J. Anim. Ecol. **68**: 672–683.

CLUTTON-BROCK, T. H., F. E. GUINNESS and S. D. ALBON, 1982 *Red Deer: Behavior and Ecology of Two Sexes*. The University of Chicago Press, Chicago.

CLUTTON-BROCK, T. H. and J. M. PEMBERTON, editors, 2004 *Soay sheep: Dynamics and Selection in an Island Population*. Cambridge University Press, Cambridge.

COLTMAN, D., 2005 Testing marker-based estimates of heritability in the wild. Mol. Ecol. **14**: 2593–2599.

COLTMAN, D., M. FESTA-BIANCHET, J. JORGENSON and C. STROBECK, 2002 Age-dependent sexual selection in bighorn rams. Proc. R. Soc. Lond. B **269**: 165–172.

COLTMAN, D., P. O'DONOGHUE, J. JORGENSON, J. HOGG, C. STROBECK *et al.*, 2003 Undesirable evolutionary consequences of trophy hunting. Nature **426**: 655–658.

CORANDER, J., P. WALDMANN, P. MARTTINEN and M. J. SILLANPÄÄ, 2004 BAPS 2: enhanced possibilities for the analysis of the genetic population structure. Bioinformatics **20**: 2363–2469.

CORNUET, J.-M. and M. A. BEAUMONT, 2007 A note on the accuracy of PAC-likelihood inference with microsatellite data. J. Theor. Biol. **71**: 12–19.

CORNUET, J. M., M. A. BEAUMONT, A. ESTOUP and M. SOLIGNAC, 2006 Inference on microsatellite mutation processes in the invasive mite, *Varroa destructor*, using reversible jump markov chain monte carlo. Theor. Popul. Biol. **69**: 129–144.

CORNUET, J.-M. and G. LUIKART, 1997 Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics **144**: 2001–2014.

CORNUET, J.-M., F. SANTOS, M. A. BEAUMONT, C. P. ROBERT, J.-M. MARIN *et al.*, 2008 Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics **24**: 2713–2719.

COULSON, T. N., J. M. PEMBERTON, S. D. ALBON, M. BEAUMONT, T. C. MARSHALL *et al.*, 1998 Microsatellites reveal heterosis in red deer. Proc. R. Soc. Lond. B **265**: 265–489.

CROW, J. F. and M. KIMURA, 1970 *An introduction to population genetics theory*. Harper and Row, New York.

CSILLÉRY, K., T. JOHNSON, D. BERALDI, T. CLUTTON-BROCK, D. COLTMAN *et al.*, 2006 Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. Genetics **173**: 2091–2101.

DIB, C., S. FAURÉ, C. FIZAMES, D. SAMSON, N. DROUOT *et al.*, 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380**: 152–154.

DIERINGER, D. and C. SCHLÖTTERER, 2003 Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. **13**: 2242–2251.

DIRIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN *et al.*, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91**: 3166–3170.

ELLEGREN, H., 2000a Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. **24**: 400–402.

ELLEGREN, H., 2000b Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. **16**: 551–558.

ELLEGREN, H., 2004 Microsatellites: Simple sequences with complex evolution. Nat. Rev. Genet. **5**: 435–445.

ESTOUP, A., M. BEAUMONT, F. SENNEDOT, C. MORITZ and J. M. CORNUET, 2004 Genetic analysis of complex demographic scenarios: Spatially expanding populations of the cane toad, *Bufo marinus*. Evolution **58**: 2021–2036.

ESTOUP, A., P. JARNE and J.-M. CORNUET, 2002 Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol. Ecol. **11**: 1591–1604.

ESTOUP, A., F. ROUSSET, Y. MICHALAKIS, J.-M. CORNUET, M. ADRIAMANGA *et al.*, 1998 Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). Mol. Ecol. **7**: 339–353.

ESTOUP, A., I. J. WILSON, C. SULLIVAN, J.-M. CORNUET and C. MORITZ, 2001 Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. Genetics **159**: 1671–1687.

EXCOFFIER, L., A. ESTOUP and J.-M. CORNUET, 2005 Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics **169**: 1727–1738.

EXCOFFIER, L. and M. SLATKIN, 1995 Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Molecular Biology and Evolution **12**: 921–927.

FAGUNDES, N. J. R., N. RAY, M. BEAUMONT, S. NEUENSCHWANDER, F. M. SALZANO *et al.*, 2007 Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. USA **104**: 17614–17619.

FALUSH, D., M. STEPHENS and J. PRITCHARD, 2003 Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics **164**: 1567–1587.

FELDMAN, M. W., A. BERGMAN, D. D. POLLOCK and D. B. GOLDSTEIN, 1997 Microsatellite genetic distances with range constraints: Analytic description and problems of estimation. Genetics **145**: 207–216.

FELDMAN, M. W., F. B. CHRISTIANSEN and L. D. BROOKS, 1980 Evolution of recombination in a constant environment. Proc. Natl. Acad. Sci. USA **77**: 4838–4841.

FESTA-BIANCHET, M., 1999 Bighorn sheep. In D. Wilson and S. Ruff, editors, *The Smithsonian Book of North American Mammals*. The Smithsonian Institution Press, Washington, D.C., 348–350.

FRANIR, F. W. C., J.-J. ARRANZ, P. BERZI, N. CAMBISANO, B. GRISART *et al.*, 2000 Extensive genome-wide linkage disequilibrium in cattle. Genome Res. **10**: 220–227.

FRASER, D., P. DUCHESNE and L. BERNATCHEZ, 2005 Migratory charr schools exhibit population and kin associations beyond juvenile stages. Mol. Ecol. **14**: 3133–3146.

GAGGIOTTI, O. E., S. P. BROOKS, W. AMOS and J. HARWOOD, 2004 Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. Mol. Ecol. **13**: 811–825.

GAGGIOTTI, O. E., O. LANGE, K. RASSMANN and C. GLIDDON, 1999 A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data options. Mol. Ecol. **8**: 1513–1520.

GARZA, J. C., M. SLATKIN and N. B. FREIMER, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Molecular Biology and Evolution **12**: 594–603.

GARZA, J. C. and E. G. WILLIAMSON, 2001 Detection of reduction in population size using data from microsatellite loci. Mol. Ecol. **10**: 305–318.

GELMAN, A., 1996 Inference and monitoring convergence. In W. R. Gilks, S. Richardson and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 131–143.

GELMAN, A., J. B. CARLIN, H. S. STERN and D. B. RUBIN, 2003 *Bayesian Data Analysis*. Chapman and Hall/CRC Press.

GERLACH, G., U. SCHARDT, R. ECKMANN and A. MEYER, 2001 Kin structured subpopulations of eurasian perch (*Perca fluviatilis*). Heredity **86**: 213–221.

GLAUBITZ, J. C., O. E. RHODES and J. A. DEWOODY, 2003 Prospects for inferring pairwise relationships with single nucleotide polymorphisms. Mol. Ecol. **12**: 1039–1047.

GOLDSTEIN, D. B., A. R. LINARES, L. L. CAVALLI-SFORZA and M. W. FELDMAN, 1995a An evaluation of genetic distances for use with microsatellite loci. Genetics **139**: 463–471.

GOLDSTEIN, D. B., A. R. LINARES, L. L. CAVALLI-SFORZA and M. W. FELDMAN, 1995b Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92**: 6723–6727.

GOODNIGHT, K. F. and D. C. QUELLER, 1999 Computer software for performing likelihood tests of pedigree relationship using genetic markers. Mol. Ecol. **8**: 1231–1234.

GRAHAM, J., J. CURRAN and B. S. WEIR, 2000 Conditional genotypic probabilities for microsatellite loci. Genetics **155**: 1973–1980.

GRIFFITHS, R. C. and S. TAVARÉ, 1994 Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B **344**: 403–410.

GUILLOT, G., A. ESTOUP, F. MORTIER and J. F. COSSON, 2005a A spatial statistical model for landscape genetics. Genetics **170**: 1261–1280.

GUILLOT, G., F. MORTIER and A. ESTOUP, 2005b Geneland : A program for landscape genetics. Mol. Ecol. Notes **5**: 712–715.

HAAG-LIAUTARD, C., M. DORRIS, X. MASIDE, S. MACASKILL, D. HALLIGAN *et al.*, 2007 Direct estimation of per nucleotide and genomic deleterious mutation rates in drosophila. Nature **445**: 82–85.

HAMILTON, G., M. CURRAT, N. RAY, G. HECKEL, M. BEAUMONT *et al.*, 2005 Bayesian estimation of recent migration rates after a spatial expansion. Genetics **170**: 409–417.

HAMILTON, W., 1964 The genetical evolution of social behaviour. I. J. Theor. Biol. **7**: 1–16.

HANSSON, B., S. BENSCH and D. HASSELQUIST, 2004 Lifetime fitness of short- and long-distance dispersing great reed warblers. Evolution **58**: 2546–2557.

HANSSON, B., D. HASSELQUIST, M. TARKA, P. ZEHTINDJIEV and S. BENSCH, 2008 Postglacial colonisation patterns and the role of isolation and expansion in driving diversification in a passerine bird. PLoS ONE **3**: 2794.

HANSSON, B., M. ÅKESSON, J. SLATE and J. M. PEMBERTON, 2005 Linkage mapping reveals sex-dimorphic map distances in a passerine bird. Proc. R. Soc. Lond. B **272**: 2289 – 2298.

HARR, B. and C. SCHLÖTTERER, 2000 Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics **155**: 1213–1220.

HARR, B., J. TODOROVA and C. SCHLÖTTERER, 2002 Mismatch repair driven mutational bias in *D. melanogaster*. Mol. Cell **10**: 199–205.

HASSELQUIST, D., 1998 Polygyny in the great reed warbler: a long-term study of factors contributing to male fitness. Ecology **79**: 2376–2390.

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

HEY, J. and R. NIELSEN, 2007 Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc. Natl. Acad. Sci. USA **104**: 2785–2790.

HEY, J., Y.-J. WON, A. SIVASUNDAR, R. NIELSEN and J. A. MARKERT, 2004 Using nuclear haplotypes with microsatellites to study gene flow between recently separated cichlid species. Mol. Ecol. **13**: 909–919.

HICKERSON, M. J., E. STAHL and N. TAKEBAYASHI, 2007 msBayes: Pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. BMC Bioinformatics **8**: 268.

HICKERSON, M. J., E. A. STAHL and H. A. LESSIOS, 2006 Test for simultaneous divergence using approximate Bayesian computation. Evolution **60**: 2435–2453.

HILL, W., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. **38**: 209–216.

HILL, W. G. and A. ROBERTSON, 1966 The effect of linkage on limits to artificial selection. Genet. Res. **8**: 269–294.

HOFFMAN, J. I. and W. AMOS, 2005 Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Mol. Ecol. **14**: 599–612.

HOGGART, C. J., T. G. CLARK, M. D. IORIO, J. C. WHITTAKER and D. J. BALDING, 2008 Genome-wide significance for dense SNP and resequencing data. Genetic Epidemiology **32**: 179–185.

HUANG, Q.-Y., F.-H. XU, H. SHEN, H.-Y. DENG, Y.-J. LIU *et al.*, 2002 Mutation patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. **70**: 625–634.

HUDSON, M., 2008 Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular Ecology Resources **8**: 3–17.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23**: 183–201.

HUDSON, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.

INGVARSSON, P. K., 2008 Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. Genetics **180**: 329–340.

JENNINGS, H. S., 1917 The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked and independent, with special relation to the effects of linkage. Genetics **2**: 97–154.

JORGENSON, J., M. FESTA-BIANCHET and W. WISHART, 1998 Effects of population density on horn development in bighorn rams. J. Wildlife Man. **62**: 1011–1020.

JOYCE, P. and P. MARJORAM, 2008 Approximately sufficient statistics and Bayesian computation. Statistical Applications in Genetics and Molecular Biology **7**.

KAISER, J., 2008 DNA sequencing: A plan to capture human diversity in 1000 genomes. Science **319**: 395.

KARLIN, S. and A. PIAZZA, 1981 Statistical methods for assessing linkage disequilibrium at the HLA-A, B, C loci. Ann. Hum. Genet. **45**: 79–94.

KIM, Y. and R. NIELSEN, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167**: 1513–1524.

KIMMEL, M. and R. CHAKRABORTY, 1996 Measures of variation at DNS repeat loci under a general stepwise mutation model. Theor. Popul. Biol. **50**: 345–367.

KIMURA, M. and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49**: 725–38.

KINGMAN, J. F. C., 1982a The coalescent. Stoch. Process. Applic. **13**: 235–248.

KINGMAN, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A**: 27–43.

KRUGLYAK, S., R. DURRETT, M. D. SCHUG and C. F. AQUADRO, 2000 Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. Molecular Biology and Evolution **17**: 1210–1219.

KRUGLYAK, S., R. T. DURRETT, M. D. SCHUG and C. F. AQUADRO, 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA **95**: 10774–10778.

KRUUK, L., 2004 Estimating genetic parameters in natural populations using the 'animal model'. Proc. R. Soc. Lond. B **359**: 873–890.

KRYUKOV, G. V., A. SHPUNT, J. A. STAMATOYANNOPOULOS and S. R. SUNYAEV, 2009 Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA **106**: 3871–3876.

KUHNER, M. K., 2006 LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics **22**: 768–770.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. Genetics **140**: 1421–1430.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156**: 1393–1401.

LACY, R., 1994 Managing genetic diversity in captive populations of animals. In M. Bowles and W. C.J., editors, *Restoration of Endangered Species: Conceptial Issues, Planning and Implementation*. Cambridge University Press, Cambridge, 63–89.

LAVAL, G. and L. EXCOFFIER, 2004 Simcoal 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics **20**: 2485–2487.

LAVERGNE, S. and J. MOLOFSKY, 2007 Increased genetic variation and evolutionary potential drive the success of an invasive grass. Proc. Natl. Acad. Sci. USA **104**: 3883–3888.

LEWONTIN, R. C. and J. KRAKAUER, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74**: 175–195.

LI, C., D. WEEKS and A. CHAKRAVARTI, 1993 Similarity of DNA fingerprints due to chance and relatedness. Hum. Hered. **43**: 45–52.

LI, N. and M. STEPHENS, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165**: 2213–2233.

LITT, M. and J. A. LUTY, 1989 A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. **44**: 397–401.

LYNCH, M. and K. RITLAND, 1999 Estimation of pairwise relatedness with molecular markers. Genetics **152**: 1753–1766.

LYNCH, M. and J. WALSH, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Press, Sunderland MA.

MAISTE, P. and B. WEIR, 2004 Optimal testing strategies for large, sparse multinomial models. Comput. Stat. Data An. **46**: 605–620.

MARJORAM, P., J. MOLITOR, V. PLAGNOL and S. TAVARÉ, 2003 Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **100**: 15324–15328.

MARJORAM, P. and S. TAVARÉ, 2006 Modern computational appraches for analysing molecular genetic variation data. Nat. Rev. Genet. **7**: 759–770.

MARSHALL, T., J. SLATE, L. KRUUK and J. PEMBERTON, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. **7**: 639–655.

MARSHALL, T. C., D. W. COLTMAN, J. M. PEMBERTON, J. SLATE, J. A. SPALTON *et al.*, 2002 Estimating the prevalence of inbreeding from incomplete pedigrees. Proc. R. Soc. Lond. B **269**: 1533–1539.

MATOCQ, M. D. and E. A. LACEY, 2004 Philopatry, kin clusters, and genetic relatedness in a population of woodrats (*Neotoma macrotis*). Behav. Ecol. **15**: 647–653.

MCRAE, A. F., J. C. MCEWAN, K. D. DODDS, T. WILSON, A. M. CRAWFORD *et al.*, 2002 Linkage disequilibrium in domestic sheep. Genetics **160**: 1113–1122.

MCVEAN, G., 2007 The structure of linkage disequilibrium around a selective sweep. Genetics **175**: 1395–1406.

MCVEAN, G. A. T., S. MYERS, S. HUNT, P. DELOUKAS, D. R. BENTLEY *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304**: 581–584.

MEHTA, C. R. and J. F. HILTON, 1993 Exact power of conditional and unconditional tests: Going beyond the 2×2 contingency table. Am. Stat. **47**: 91–98.

MELO, M. and B. HANSSON, 2006 Identification of 15 polymorphic microsatellite loci in the Príncipe seedeater (*Serinus rufobrunneus*) and assessment of their utility in nine other *Serinus* species (Fringillidae, Aves). Mol. Ecol. Notes **6**: 1266–1268.

MILLIGAN, B., 2003 Maximum-likelihood estimation of relatedness. Genetics **163**: 1153–1167.

MOUNTAIN, J., A. KNIGHT, M. JOBIN, C. GINOUX, A. MILLER *et al.*, 2002 SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. Genome Res. **12**: 1766–1772.

MYERS, S., L. BOTTOLO, C. F. G. MCVEAN and P. DONNELLY, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310**: 321–324.

NAMROUD, M.-C., J. BEAULIEU, N. JUGE, J. LAROCHE and J. BOUSQUET, 2008 Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. Mol. Ecol. **17**: 3599–3613.

NEFF, B. D. and M. R. GROSS, 2001 Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. Evolution **55**: 1717–1733.

NEI, M. and A. K. ROYCHOUDHURY, 1974 Sampling variances of heterozygosity and genetic distance. Genetics **76**: 379–390.

NEUENSCHWANDER, S., C. R. LARGIADER, N. RAY, M. CURRAT, P. VONLANTHEN *et al.*, 2008 Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): Inference under a Bayesian spatially explicit framework. Mol. Ecol. **17**: 757–772.

NIELSEN, R. and J. WAKELEY, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158**: 885–896.

NORTH, B. V., D. CURTIS and P. C. SHAM, 2002 A note on the calculation of empirical *P* values from Monte Carlo procedures. Am. J. Hum. Genet. **71**: 439–441.

NOTHNAGEL, M., R. FÜRST and K. ROHDE, 2002 Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum. Hered. **54**: 186–198.

OHTA, T. and M. KIMURA, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genome Res. **22**: 201–204.

OKADA, G. S. Y., J. EMRICH, J. NEWTON, A. TSUGITA, E. TERZAGHI *et al.*, 1966 Frameshift mutations and the genetic code. Cold Spring Harbor Symp. Quant. Biol. **31**: 77–84.

O'REILLY, P. F., E. BIRNEY and D. J. BALDING, 2008 Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Res. **18**: 1304–1313.

ÖST, M., E. VITIKAINEN, P. WALDECK, L. SUNDSTRÖM, K. LINDSTRÖM *et al.*, 2005 Eider females form non-kin brood-rearing coalitions. Mol. Ecol. **14**: 3903–3908.

PASCUAL, M., M. P. CHAPUIS, F. MESTRES, J. BALANYA, R. B. HUEY *et al.*, 2007 Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. Mol. Ecol. **16**: 3069–3083.

PAYSEUR, B. A. and A. D. CUTTER, 2006 Integrating patterns of polymorphism at SNPs and STRs. Trends Genet. **22**: 424–429.

PELLA, J. and M. MASUDA, 2006 The gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. Can. J. Fish. Aquat. Sci. **63**: 576–596.

PEMBERTON, J. M., 2008 Wild pedigrees: the way forward. Proc. R. Soc. Lond. B **275**: 613–621.

PEMBERTON, J. M., J. SLATE, D. R. BANCROFT and J. A. BARRETT, 1995 Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. Mol. Ecol. **4**: 249–252.

PFAFF, C. L., E. J. PARRA, C. BONILLA, K. HIESTER, P. M. MCKEIGUE *et al.*, 2001 Population structure in admixed populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. Am. J. Hum. Genet. **68**: 198–207.

PRITCHARD, J., M. SEIELSTAD, A. PEREZ-LEZAUN and M. FELDMAN, 1999 Population growth of human y chromosomes: a study of y chromosome microsatellites. Molecular Biology and Evolution **16**: 1791–1798.

PRITCHARD, J. K. and M. W. FELDMAN, 1996 Statistics for microsatellite variation based on coalescence. Theor. Popul. Biol. **50**: 325–344.

PRITCHARD, J. K., M. STEPHENS and P. J. DONNELLY, 2000a Inference of population structure using multilocus genotype data. Genetics **155**: 945–959.

PRITCHARD, J. K., M. STEPHENS, N. A. ROSENBERG and P. DONNELLY, 2000b Association mapping in structured populations. Am. J. Hum. Genet. **67**: 170–181.

PUMPERNIK, D., B. OBLAK and B. BORSTNIK, 2008 Replication slippage versus point mutation rates in short tandem repeats of the human genome. Molecular Genetics and Genomics **279**: 53–61.

QUELLER, D. and K. GOODNIGHT, 1989 Estimating relatedness using molecular markers. Evolution **43**: 258–275.

R DEVELOPMENT CORE TEAM, 2005 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAMAKRISHNAN, U. and J. MOUNTAIN, 2004 Precision and accuracy of divergence time estimates from STR and SNPSTR variation. Molecular Biology and Evolution **21**: 1960–1971.

REED, F. A. and S. A. TISHKOFF, 2005 Positive selection can create false hotspots of recombination. Genetics **172**: 2011–2014.

RENGMARK, A. H., A. SLETTAN, O. SKAALA, O. LIE and F. LINGAAS, 2006 Genetic variability in wild and farmed Atlantic salmon (*Salmo salar*) strains estimated by SNP and microsatellites. Aquaculture **253**: 229–237.

REUSCH, T., M. HABERLI, P. AESCHLIMANN and M. MILINSKI, 2001 Female sticklebacks count alleles in a strategy of sexual selection explaining MHC polymorphism. Nature **414**: 300–302.

RICHARDSON, D. S., J. KOMDEUR and T. BURKE, 2004 Inbreeding in the Seychelles warbler: environment-dependent maternal effects. Evolution **58**: 2037–2048.

RIPLEY, B. D., 1982 *Stochastic simulation*. Wiley, New York.

RITLAND, K., 1996a Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. **67**: 175–185.

RITLAND, K., 1996b A marker-based method for inferences about quantitative inheritance in natural populations. Evolution **50**: 1062–1073.

RITLAND, K., 2000 Marker-inferred relatedness as a tool for detecting heritability in nature. Mol. Ecol. **9**: 1195–1204.

RITLAND, K. and C. RITLAND, 1996 Inferences about quantitative inheritance based upon natural population structure in the common monkeyflower, *Mimulus guttatus*. Evolution **50**: 1074–1082.

ROSENBLUM, E. B., M. HICKERSON and C. MORITZ, 2007 A multilocus perspective on colonization accompanied by selection and gene flow. Evolution **61**: 2971–2985.

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

RUSSELLO, M. and G. AMATO, 2004 *Ex situ* population management in the absence of pedigree information. Mol. Ecol. **13**: 2829–2840.

SABATTI, C. and N. RISCH, 2002 Homozygosity and linkage disequilibrium. Genetics **160**: 1707–1719.

SAINUDIIN, R., R. T. DURRETT, C. F. AQUADRO and R. NIELSEN, 2004 Microsatellite mutation models: Insights from a comparison of humans and chimpanzees. Genetics **168**: 383–395.

SCHLOTTERER, C., 2002 A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics **160**: 753–763.

SCHLÖTTERER, C., R. RITTER, B. HARR and G. BREM, 1998 High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. Molecular Biology and Evolution **15**: 1269–1274.

SEDDON, J. M., H. G. PARKER, E. A. OSTRANDER and H. ELLEGREN, 2005 SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. Mol. Ecol. **14**: 503–511.

SEKINO, M., T. SUGAYA, M. HARA and N. TANIGUCHI, 2004 Relatedness inferred from microsatellite genotypes as a tool for broodstock management of japanese flounder *Paralichthys olivaceus*. Aquaculture **233**: 163–172.

SEYFERT, A. L., M. E. A. CRISTESCU, L. FRISSE, S. SCHAACK, W. K. THOMAS *et al.*, 2008 The rate and spectrum of microsatellite mutation in *Caenorhabditis elegans* and *Daphnia pulex*. Genetics **178**: 2113–2121.

SHAM, P. C. and D. CURTIS, 1995 Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. Ann. Hum. Genet. **59**: 97–105.

SHOEMAKER, J. S., I. S. PAINTER and B. S. WEIR, 1999 Bayesian statistics in genetics: A guide for the uninitiated. Trends Genet. **15**: 354–358.

SILVERMAN, B. W., 1986 *Density Estimation*. Chapman and Hall, London.

SLATE, J., K. G. D. P. DAVID, B. A. VEENVLIET, B. C. GLASS, T. E. BROAD *et al.*, 2004 Understanding the relationship between the inbreeding coefficient and multilocus heterozygosity: theoretical expectations and empirical data. Heredity **93**: 255–265.

SLATE, J. and J. M. PEMBERTON, 2007 Admixture and patterns of linkage disequilibrium in a free-living vertebrate population. J. Evolution. Biol. **20**: 1415–1427.

SLATE, J., T. C. V. STIJN, R. M. ANDERSON, K. M. MCEWAN, N. J. MAQBOOL *et al.*, 2002 A deer (subfamily Cervinae) genetic linkage map and the evolution of ruminant genomes. Genetics **160**: 1587–1597.

SLATKIN, M., 1994 Linkage disequilibrium in growing and stable populations. Genetics **137**: 331–336.

SLATKIN, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**: 457–462.

SLATKIN, M., 2008 Linkage disequilibrium - understanding the evolutionary pastand mapping the medical future. Nat. Rev. Genet. **9**: 477–485.

SOOFI, E. S., 1994 Capturing the intangible concept of information. J. Am. Stat. Assoc. **89**: 1243–1254.

SPRITZ, R. A., 1981 Duplicationdeletion polymorphism 5≠ to the human β-globin gene. Nucleic Acids Res. **9**: 5037–5047.

SPROTT, D. A., 2000 *Statistical Inference in Science*. Springer, New York.

STEPHENS, M., N. J. SMITH and P. DONNELLY, 2001 A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68**: 978–989.

TABERLET, P. and G. LUIKART, 1999 Non-invasive genetic sampling and individual identification. Biology Journal of the Linnean Society **68**: 41–55.

TAUTZ, D., 1989 Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res. **17**: 6463–6471.

TENESA, A., P. NAVARRO, B. J. HAYES, D. L. DUFFY, G. M. CLARKE *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. **17**: 520–526.

THOMAS, S., 2005 The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. Proc. R. Soc. Lond. B **360**: 1457–1467.

THOMAS, S. C., D. W. COLTMAN and J. M. PEMBERTON, 2002 The use of marker-based relationship information to estimate the heritability of body weight in a natural population: a cautionary tale. Journal of Evolutionary Biology **15**: 92–99.

THOMPSON, C. L., D. B. D, Q. LU, G. MATHEW, Y. SONG *et al.*, 2005 Effect of genotyping error in model-free linkage analysis using microsatellite or single-nucleotide polymorphism marker maps. BMC Genet. **6(Suppl. 1)**: S153.

THOMPSON, E., 1986 *Pedigree analysis in human genetics*. The Johns Hopkins University Press, Baltimore.

THOMPSON, E. A. and T. R. MEAGHER, 1998 Genetic linkage in the estimation of pairwise relationship. Theoretical and Applied Genetics **97**: 857–864.

TIAN, C., D. A. HINDS, R. SHIGETA, R. KITTLES, D. G. BALLINGER *et al.*, 2006 A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am. J. Hum. Genet. **79**: 640–649.

TISHKOFF, S. A., E. DIETZSCH, W. SPEED, A. J. PAKSTIS, J. R. KIDD *et al.*, 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271**: 1380–1387.

VALDES, A. M., M. SLATKIN and N. B. FREIMER, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133**: 737–749.

VAN DE CASTEELE, T., P. GALBUSERA and E. MATTHYSEN, 2001 A comparison of microsatellite-based pairwise relatedness estimators. Mol. Ecol. **10**: 1539–1549.

VAN STAADEN, M., 1994 *Suricata suricatta*. In *Mammalian Species*, volume 483. Allen Press, 1–8.

VERA, C. J., C. W. WHEAT, H. W. FESCEMYER, M. J. FRILANDER, D. L. CRAWFORD *et al.*, 2008 Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol. Ecol. **17**: 1636–1647.

VIGNAL, A., D. MILAN, M. SANCRISTOBAL and A. EGGEN, 2002 A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. **34**: 275–305.

WAKELEY, J., R. NIELSEN, S. N. LIU-CORDERO and K. ARDLIE, 2001 The discovery of single-nucleotide polymorphisms and inferences about human demographic history. Am. J. Hum. Genet. **69**: 1332–1347.

WANG, J., 2002 An estimator for pairwise relatedness using molecular markers. Genetics **160**: 1203–1215.

WANG, J., 2003 Maximum-likelihood estimation of admixture proportions from genetic data. Genetics **164**: 747–765.

WANG, J., 2006 A coalescent-based estimator of admixture from DNA sequences. Genetics **173**: 1679–1692.

WEBER, J. L. and P. E. MAY, 1989 Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. **44**: 388–396.

WEBER, J. L. and C. WONG, 1993 Mutation of human short tandem repeats. Human Molecular Genetics **2**: 1123–1128.

WEIR, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, Massschusetts.

WEIR, B. S., A. D. ANDERSON and A. B. HEPLER, 2006 Genetic relatedness analysis: modern data and new challenges. Nat. Rev. Genet. **7**: 771–780.

WEIR, B. S., L. CARDON, A. D. ANDERSON, D. M. NIELSEN and W. G. HILL, 2005 Heterogeneity of measures of population structure along the human genome. Genome Res. **15**: 1468–1476.

WEIR, B. S. and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38**: 1358–1370.

WEISSENBACH, J., G. GYAPAY, C. DIB, A. VIGNAL, J. MORISSETTE *et al.*, 1992 A second-generation linkage map of the human genome. Nature **359**: 794–801.

WHITTAKER, J. C., R. M. HARBORD, N. BOXALL, I. MACKAY, G. DAWSON *et al.*, 2003 Likelihood-based estimation of microsatellite mutation rates. Genetics **164**: 781–787.

WILSON, I. J. and D. J. BALDING, 1998 Genealogical inference from microsatellite data. Genetics **150**: 499–510.

WILSON, I. J., M. E. WEALE and D. J. BALDING, 2003 Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J. Roy. Stat. Soc. C-App. **166**: 155–188.

WRIGHT, S., 1931 Evolution in mendelian populations. Genetics **16**: 97–159.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugen. **15**: 323–354.

XU, X., M. PENG and Z. FANG, 2000 The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. **24**: 396–399.

ZHAO, H., D. NETTLETON and J. C. M. DEKKERS, 2007 Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms and QTLs. Genet. Res. **88**: 1–7.

ZHAO, H., D. NETTLETON, M. SOLLER and J. C. M. DEKKERS, 2005 Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genet. Res. **86**: 77–87.

ZHAO, H., A. PAKSTIS, J. KIDD and K. KIDD, 1999 Assessing linkage disequilibrium in a complex genetic system. I. overall deviation from random association. Ann. Hum. Genet. **63**: 167–179.