Detecting and managing suspected admixture and genetic drift in domestic livestock: modern Dexter cattle – a case study

Timothy C Bray

Cardiff University

A dissertation submitted to Cardiff University in candidature for the degree of Doctor of Philosophy

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Table of Contents

Abstract

This study combines a range of contemporary genetic analysis methods to analyse the Dexter cattle breed in conjunction with the development of a novel method of admixture determination. The Dexter was chosen for its heterogeneous genetic composition due to a complex population history. Comparison against other European cattle breeds showed the Dexter to be one of the most diverse breeds and clearly distinguishable from other breed populations. The levels of migrant individuals exchanged between the Dexter and other European breeds was seen to be in the middle of the range for all breeds, as was the conservation value of the Dexter as determined through the Weitzman genetic distance approach. The Dexter was shown to stand out from other European cattle breeds due to high levels of subdivision into different regions of the herd book. The hypothesis that the ancestry of subdivisions was entirely responsible for this genetic divergence could not be proven. The quantification of admixture proportions were made for two putative ancestral representative breeds, Red Devon and Kerry. It was found that a selection of carefully chosen Traditional Dexter individuals were more closely related to the Kerry breed. Admixture contributions for remaining breed populations were inconclusive with the exception of a small sample group representing the breed in America which demonstrated a higher Red Devon contribution. Genetic drift is heavily implicated in the results shown and it is notable that high levels of variance were associated with admixture contributions.

An approximate Bayesian computation approach was designed and developed to better model the admixture scenario of interest. A method allowing for two admixture events was constructed in order to calculate parental contributions and compare them to simulated datasets according to a genetic model. Initial testing proved successful using a single admixture event. The addition of a second admixture event reduced the accuracy of the method. Testing scenarios of up to half a million simulations with nine loci were unable to successfully quantify either simulated or real admixture events here. Testing suggests that the effectiveness of the approach is thought to increase with numbers of simulated datsets used. Recommendations for the successful application of the method are made.

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed .....Tim Bray..... (candidate) Date .....21.11.08.....

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of .....PhD.....(insert MCh, MD, MPhil, PhD etc, as appropriate)

Signed .....Tim Bray..... (candidate) Date .....21.11.08...

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated.
Other sources are acknowledged by explicit references.

Signed .....Tim Bray.... (candidate) Date .....21.11.08.....

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .....Tim Bray..... (candidate) Date .....21.11.08.....

Acknowledgements

Huge thanks have been earned by many for their tolerance and perseverance that have enabled the manifestation that is this thesis document. Particular support and pressure was freely given by both my supervisors, Mike and Lounés. This work could not have been completed without the financial support of the Rare Breeds Survival Trust (RBST), Dexter Cattle Society (DCS) and Cardiff University. Appreciation again goes to the RBST and DCS for their invaluable assistance in sample collection, as well as to all of the kind people that assisted the sampling process and those whose animals were used. All things cattle would have been stuck in the mud without Andrew and his unmatched herd book collection. Other technical and computational assistance cannot go unappreciated from those stretching back into my first year in Cardiff such as Amy, Andy, Ciara, Fairus and Maxime, to more recent contributors such as Geoff, 'Condor' James, Vitor, and Bárbara. Additional (and by no means lesser) thanks go to everyone else in the labs and departments which I have had the priviledge to invade and all of whom I could not hope to mention (but I'm still going to try); Clathers, Jeff, Viola, everyone called Jo, Gabi, Faye, Rita, Dan, Muya, Ian, Jon, Carlos, Dave, Renata, Luke, Sam, Liz, Ester, George, Sian, Rhys, Pati, and Wendy. This leaves only those family and friends who unwittingly found themselves on the receiving end of PhD-fueled madness, I include PhD students around the world Sabine and Alan and perhaps mostly Siân and the family Bray for ultimate patience (and the odd sagacious rant, cheers Jon).

Chapter 1.
Introduction

1. Introducing population genetics

Population genetics has been a rapidly expanding science since the advent of molecular (especially DNA based) methods and their continuous development has allowed far greater insight into population dynamics than ever before. Domestic livestock provide convenient models with which to analyse genetic variation. One of the areas dramatically improved by molecular genetic advances is that of the study of gene-flow and admixture of populations. Recognition of the increasing need for application of admixture detection methods has led to the development of a number of different models and approaches. This review provides an assessment of the demographic processes which underly admixture events and an overview of the methods applied to investigate them.

1.1. Molecular genetics in conservation

The continued development of molecular techniques has largely fuelled the advance of population biology, promoting consideration of new problems and allowing re-analysis of old ones (Pertoldi *et al.* 2007). The consequences of these advances affect studies of all levels of interactions, from species (e.g. molecular barcoding (Hebert *et al*, 2004)), intra-population molecular diversity (e.g. using microsatellites (Avise, 2004)), and individual-based approaches (e.g. genetic fingerprinting (Buntjer *et al*, 2002)). The concomitant development of mathematical models to describe and predict changes in populations both temporally (e.g. Beaumont, 1999) and spatially (e.g. Dupanloup *et al*, 2002) has further advanced the field. Both conservation biology and ecology have benefited from renewed emphasis as a result of advances in the new discipline of 'conservation genetics', now well represented in the literature (e.g. Loeschcke *et al.* 1994; Avise *et al.* 1997; Frankham *et al.* 2002). Conservation genetics encompasses aspects of molecular ecology and population genetics as well as components of both evolutionary biology and systematics.

The distribution of genetic variation in endangered populations is the basis for conservation genetic studies and is determined by the contemporary and historical processes predominantly involving genetic drift, gene flow and migration but also including the two other evolutionary forces of selection and mutation (Slatkin, 1987). Advances in efficiency of DNA technologies have allowed the non-invasive sampling of subjects where previous sampling methods were ethically undesirable or

unfeasible (Taberlet and Luikart, 1999). Techniques specialising in retrieval of small samples for genetic analysis allow the use of hair, faeces, and body fluids. Genetic information from a set of polymorphic marker regions within the genome allow the identification of variation in distribution of alleles (Avise, 2004). According to the rate of sequence evolution in these marker regions, relationships can be inferred to the population and individual level. The focus of this chapter is the following; genetic diversity and structure, how assessment of that variation allows attribution of conservation value, and how it is possible to use methods to quantify genetic variation in the context of particular approaches (e.g. estimating genetic admixture).

1.2. Population genetic diversity

1.2.1. Microsatellites

In order to measure and describe the genetic variation present within individuals and populations, an appropriate marker system is required. Genetic diversity at marker loci is generated by mutations which cause changes in the DNA sequence of the marker region generating novel allelic states. Microsatellites are one of the most rapidly evolving genomic elements used as genetic markers in population biology, with a mutation rate of around 10^{-5}per generation (Wan et al, 2004). It is because of this high mutation rate compared with other genomic regions that microsatellites are often employed in population studies. Microsatellite are currently the marker of choice in population genetics due to their high mutation rate, ease of application, and availability although there are many examples of population studies that have employed other markers; minisatellites (Jeffreys et al. 1990), mitochondrial DNA (Michaux et al. 2005); as well as anonymous markers in the form of RFLP and AFLP data (Yan et al. 1999). Microsatellites have been seen to replace protein markers such as allozymes where differential expression and lack of neutrality reduced the effectiveness of their application (Beebee and Rowe, 2004). There is evidence that single nucleotide polymorphisms (SNPs) will soon replace microsatellites in population studies as the cost of sequencing decreases, for example systems exist which are capable of analysing a thousand SNPs across a haploid genome (Wang et al, 2005).

A microsatellite locus usually comprises a one to five base-pair motif repeated up to 40 times (Willard 1989). As predominantly autosomal markers, microsatellite loci

applied in tandem can be variable enough to allow the unequivocal identification of individuals in animal and plant populations (Paetkau et al, 1998). Microsatellites have been widely accepted as useful tools for measuring genetic diversity and divergence within and among populations (e.g. Bowcock et al, 1994; Blott et al, 1999). Microsatellite loci have also been shown to be much more useful than less variable markers, such as proteins, in determination of genetic variation and differentiation among closely related populations such as cattle breeds (Arranz et al, 1996). Hence microsatellite studies are now prolific and numbers of available loci identified in commercial species had reached the thousands by the mid nineties (Mommens et al, 1998) with applications from parentage testing (e.g. Liron et al, 2004) to population studies (e.g. Kantanen et al, 2000).

1.2.2. Within-population variability

For most population studies, selectively neutral markers are desirable. The distribution of alleles for these neutral markers is ultimately compared to those frequencies expected according to applied evolutionary models. The conformation of allele frequencies to those expected under Hardy Weinberg equilibrium (HWE) (Hardy 1908; Weinberg 1908) is commonly assumed in genetic analysis. The concept of HWE is that, in a large randomly mating population with non-overlapping generations, the genotype frequencies are the product of the allele frequencies and remain constant between generations irrespective of allelic dominance. Implicit in HWE are several properties; infinite population size, random mating, equal mutation rates between alleles, and negligible migration rates (Avise, 2004). Deviations from HWE can themselves be used to make inferences about a population and selective forces, for example clustering algorithms can use these deviations in calculating individual assignments between populations (e.g. Pritchard et al. 2000).

In order to give an unbiased representation of population processes, molecular markers must be chosen carefully, particularly in populations potentially under strong selection pressure. Selection can influence large areas of the genome, for example chromosome-specific selection can result from female mate choice acting on the Y chromosome (i.e. male traits) in Poeciliid fish (Lindholm and Breden, 2002). Close physical association of the genetic basis of a trait with another locus would be likely to increase the incidence of certain alleles alongside the beneficial alleles of the trait as selection drives it toward fixation. This is known as Linkage Disequilibrium (LD)

and whilst this is true of large randomly mating populations where there is low recombination, LD can be linked to other sources; variation in recombination rates, selection, genetic drift (founder effects and population bottlenecks), and population admixture (e.g. Slatkin, 1994 ; McKeigue, 1998 ; Pritchard and Przeworski, 2001 ; Ardlie et al., 2002).

Genetic variation within a population can be characterised using measures such as expected heterozygosity (H_e) and allele numbers (n_A). Providing relative levels of diversity in a study as well as for populations typed using the same loci, H_E and n_A can provide indications of departures from demographic stability and selective neutrality (Chikhi and Bruford, 2005). In addition to contemporary diversity, historical demographic events are also reflected in population genetic data, but in order to do this, genetic models of evolution have to be adopted. There are three prevailing models describing the mutational processes of microsatellite loci; The Infinite Allele Model (IAM) (Kimura and Crow, 1964) is based on the assumption that each new mutation creates a unique allele. The continuation of this assumption is that all equivalent alleles are identical by descent. The Stepwise Mutation Model (SMM) (Kimura and Ohta, 1978) adopts the assumption that a mutation changes the allele by a single repeat unit. Under this model alleles of the same size are more closely related but not necessarily identical by descent. Based on the SMM is the Two Phase Model (TPM) (Di Rienzo *et al*, 1994) which has the addition of a proportion of multi-step mutations in order to better simulate marker behaviour. Whilst it is accepted that the mode of mutation of microsatellites is that of a stepwise process, the allelic variation at many loci conforms better to predictions based on the IAM and not the SMM (Neff *et al*, 1999). The suggestion by Di Rienzo *et al* (1994) is that a multi-step process giving rise to novel alleles is the cause of this apparent fit to the IAM. Which model is adopted has consequences when considering demographic events such as population bottlenecks (Cornuet and Luikart, 1996).

1.2.3. Population bottlenecks

The effective size of a population affects genetic composition with particular influence on processes such as inbreeding and genetic drift. Populations can be subject to fluctuation in effective size over time and this can result in population bottlenecks, i.e. periods of variable duration where there are low numbers of breeding individuals.

During these periods there is commonly a reduction in genetic variation due to genetic drift and the increased probability of losing rare alleles (Nei et al., 1975), sometimes resulting in extreme changes in allele frequencies in comparison with the pre-bottlenecked population. This is both as a result of a sampling effect of those alleles perpetuated in the fewer remaining individuals, as well as the increased chances of fixation of alleles (Nei and Maruyama, 1975). Population bottlenecks are associated with a correlative and progressive reduction of both allele number and heterozygosity, the relationship between these is one way of identifying a bottleneck event having occurred (Cornuet and Luikart, 1996). A more rapid loss of allelic diversity relative to heterozygosity can sometimes be detected and is a consequence of the loss of rare alleles which have a less immediate effect on heterozygosity. This results in a transient deficiency of numbers of alleles found in the sample population, meaning that the observed allele numbers is less than numbers expected from the observed heterozygosity (for a population at mutation-drift equilibrium). The allele deficiency is dependent on four parameters concerning the bottleneck event; time since the start of the bottleneck, the population size ratio before and after the start of the event, the mutation rate of the locus, and the sample size of genes involved. Severe and long-term population bottlenecks can often be identified, but the power of these methods is limited and perhaps only improved through the use of temporal data in the form of ancient sample information (Beaumont, 1999). A lack of bottleneck signature from a known event may infer that the bottleneck was sufficiently ancient for the effects to be undetectable. Alternatively the introduction of introgression from another source post-bottleneck may have obscured the evidence of the bottleneck entirely.

1.2.4. Population differentiation

Nowadays, analysis of population genetic differentiation can be carried out in a wide variety of ways according to the particular question, assumptions made, and computational limitations. At perhaps the most basic level, Wright's F_{ST} (Wright, 1951) is a standard measure of population differentiation used which determines the relative fixation of alleles in subpopulations. F_{ST} has the advantages of ease of calculation and a relatively low dependency on the underlying assumptions of either stepping-stone or n-island models making it a good summary statistic. Although comparisons of F_{ST} values among studies can be limited in the information that they provide (especially due to sampling differences and where populations are of

unequal size), F_{ST} is an easy and generally effective measure of general relationships between sample groups within a study (Chikhi and Bruford, 2005). One additional problem with F_{ST} is the weighting given to rare alleles, although these can be calculated separately according to the prescribed models (Nei, 1987 ; Weir and Cockerham, 1984 ; Robertson and Hill, 1984). The same is true for other summary statistics such as genetic distances which can be calculated across a similar model range (Nei 1987; Cavalli-Sforza and Edwards, 1967).

Population genetic differentiation accumulates as a result of the temporal and geographical separation of populations (Wright, 1943). Populations that are very close to one another geographically which exchange individuals or which share a very recent common ancestor will usually have a greater proportion of shared allelic combinations than those which diverged long ago and which do not exchange genes (including reciprocally). A simple scenario of unidirectional transfer is an example where unreciprocated gene-flow can dramatically affect one population and not the other, but population interactions are commonly far more complex (Hansson *et al*, 2000). The spatial distribution of genetic diversity often varies even within the range of single populations. Through selective advantage for phenotypic characters as a result of environmental heterogeneity, variation can be partitioned within regions in a complex manner and as clines on a variety of spatial scales (Merila and Crnokrak, 2001). Such patterns may be a consequence of directional selection (Endler, 1986) and the high level of linkage of traits with a heritable component (Houle, 1992). The amount and distribution of genetic variation found in populations develops over time, for example commonly populations in the heart of a species distribution will be more diverse than those at the edge of its range (Merila *et al.* 1997). The geographic history of populations can result in complex relationships and can, for example, provide evidence tracing range origin back to glacial refugia (Valdiosera et al. 2007). But population differentiation is not always geographically correlated and anthropogenically managed populations can have even more complicated relationships.

Populations do not always require extended time periods or strong selection to become differentiated, small populations in particular can be altered rapidly through genetic drift. Genetic drift is a stochastic process by which allele frequencies vary from one generation to the next as a consequence of finite populations size (Kimura,

1971). Genetic drift has greater consequences in smaller populations and hence is an important factor in endangered species. When populations are at low numbers the effects of drift can be important, because the likelihood of rare alleles becoming fixed, and of common alleles being lost, increases. In such populations, genetic diversity tends to diminish as alleles are more likely to be lost and less likely to be generated spontaneously by mutation. A combination of drift and selective processes will act to increase the differentiation between separated populations whereas gene flow between them will act to stabilise this in a dynamic fashion (Beebee and Rowe, 2004).

Summary methods can be used to reduce genetic information from individuals in a population into single values, in order to measure inter-population differentiation. These are often quick to perform and require relatively low levels of computational resources. However because the methods reduce the dimensions of the data in such a way, information is inevitably lost. An alternative to using summary statistic based methods is to adopt an approach that co-assigns individuals into populations, or populations into groups, with an associated proportion or likelihood. More complex scenarios can also be considered such as those accounting for geographic information (e.g. Dupanloup et al. 2002), using admixture models (Pritchard et al. 2000), or classifying individuals as first generation migrants (Piry et al. 2004). The application of many of these methods can be made in the same way as F-statistics, so as to guide subsequent analytical approaches or as an indication of general patterns of gene flow across a physical or temporal barrier. The admixture model in Pritchard et al. (2000) for example, allows the consideration of many populations and attributes admixture proportions based on each inter-breeding population (at Hardy Weinberg Equilibrium) in accordance with contemporary allele frequencies.

1.3. Assignment of conservation value

The diversity hierarchy commonly represents the three levels of importance as being of ecosystems, species, and genes, but this is commonly questioned with regard to the importance of populations (Bowen, 1999). Similarly the functional units in conservation are chiefly concerned with considerations of scale (Crandall et al, 2000). Conservation scenarios involving wild organisms often use species or subspecies as the unit of interest, but this does not preclude consideration at the population or individual levels. Mayr (1963) demonstrated the value of variation in

ecologically important traits between populations. If adaptively significant gene combinations differ between populations within a species, they could be considered separately if that species is to be managed as a conservable entity with its evolutionary potential intact. Problems arise with the determination of how best to assess both genetic diversity and the myriad of factors pertinent to its conservation, and this is exacerbated when prioritisation calculations need to be made.

Studies investigating how genetic composition can allow identification of threats to population persistence have highlighted some important factors affecting populations such as low levels of variation (Ujvari *et al.* 2002), the accumulation of deleterious alleles (Bataillon and Kirkpatrick, 2000), and introgression of genes from other species or populations (Randi and Lucchini, 2002). Genetic threats to populations may include ongoing effects on persistence, such as the expression of deleterious genes which can manifest themselves in phenotype (Land and Lacy, 2000). But no less important are those factors which may act over the longer term such as low genetic variation diminishing the capacity of a population to maintain its evolutionary potential in a changing environment (Frankel and Soulé, 1981). The loss of genetic adaptation specific to a population can be more difficult to identify or quantify making management problematic (Garcia-Moreno *et al*, 1996). In particular it can be these gradual processes that appear not to warrant immediate conservation action that can result in population extinction through allee effects (Allee, 1931). Populations can become caught in a feedback loop where persistence deteriorates rapidly towards extinction (the so-called extinction vortex).

Avise (2004) suggested that programmes for protecting threatened species requires the identification of unambiguous units of management which reflect evolutionarily important lineages. Taxonomic distinctiveness can be investigated by summarising genetic data into a single 'genetic distance', where greater separation of one species in a group of species increases its value as a conservable entity. However, this kind of method has been shown to be ineffective at a finer scale where distances between populations are compromised by within population diversity effects (Cabellero and Toro, 2002; Bruford, 2004). When used in studies of populations within a species it can provide a good indication of genetic value but without the resolution to adequately resolve relationships between them (Laval *et al*, 2000). Many traditional allele frequency or genetic distance based methods are similarly limited when applied

to closely related populations due to incomplete use of the data through their reliance on summary statistics. Simianer (2005) suggests that genetic drift is the major cause of loss of allelic diversity in domestic cattle and predicted risk of extinction based on allele numbers alone. An interesting result of this approach is the suggestion of using marker loci to indicate genetic erosion as a parallel to quantitative characters, where loss of an allele represents the loss of a beneficial trait or significant variation within that trait. Populations that have recently experienced a severe reduction in size are particularly important to identify for conservation due to increased extinction risk, and can be evaluated based specifically on this loss of alleles when compared to the loss of heterozygosity (Cornuet and Luikart, 1996).

1.4. Genetic admixture

Information from admixture scenarios can be used in identifying genetic linkage and heritability as in congenital diseases (e.g. Chakraborty and Weis, 1988; Stephens et al. 1994), biogeography and historical population origins (e.g. Shriver et al. 2003), and in more contemporary population genetics and conservation contexts (e.g. Susnik et al. 2004). It is the latter of these that is of particular interest here. Admixture occurs as a result of gene-flow between two genetically differentiated populations. When admixed individuals subsequently backcross into their own or a new population the genes acquired through the admixture event are said to have been introgressed into that population (Rhymer and Simberloff, 1996). The dynamics of this interaction depend on the levels of differentiation between the populations. The difference between admixture and the general concept of gene-flow between populations depends largely on context. The long term associations of taxa as described by hybrid zone dynamics can be complex (Beebee and Rowe, 2004). Genetic exchange can spread back into the parent populations over huge areas or can be restricted to the formation of temporary intermediate hybrid populations not spreading into either parent population. Over brief periods of contact, there may just be a few mating events and little genetic exchange. These transferred genes can still be detected in the populations for many generations after the event even if this degree of exchange is limited (Hansen, 2002).

Gene flow between differentiatied populations can occur both naturally as a result of Pleistocene range shifting reconnecting populations (Arnold, 1997), and with anthropogenic influence such as between the domestic dog and Ethiopian wolf

(Gottelli et al, 1994). This degree of differentiation of individuals that can still produce offspring, and more importantly fertile offspring, is not necessarily directly related to the time since coalescence of their populations. The process of reproductive isolation can develop incidentally when there are no forces maintaining compatibility as well as being a result of drift and selective pressures in each population (Turelli et al, 2001). Such isolation will occur to a lesser extent if there is gene flow between the populations. This can lead to complex situations in which populations are able to exchange genes with those adjacent to them but not with more distant populations. In this situation intermediate populations can therefore facilitate gene flow across the whole geographical range as seen in 'ring-species' (Irwin et al, 2001).

Introgression resulting in the exchange of genes with another differentiated population can allow the creation of new allelic combinations. If the new alleles entering a population demonstrate a large selective advantage to the individuals carrying them they may perpetuate. If the new allelic combinations demonstrate maladaptive traits then the individuals carrying them may suffer disadvantage and the new alleles are likely to remain excluded or at low frequency. If two populations are very different phenotypically or genetically then it may be that the probability of a successful cross-population mating and subsequent back-crossing is very low because of this creation of maladaptive combinations. If the populations are more divergent still, genetic exchange may be excluded through incompatibilities in physiology or behaviour (Turelli et al, 2001). Introgression may be prevented in an indirect manner through the exclusion of the F1 generation due to low survival probability from a maladaptive phenotype, sterility, or lack of sexually selected traits. If this selection against hybrids is strong it can create a barrier against introgression of negatively selected and neutral alleles, but in some circumstances strongly positively selected alleles may be able to cross such a barrier (Pialek and Barton, 1997).

Gene flow is not always balanced, an effect occurring increasingly as inter-population differentiation increases. In some instances unidirectional transfer results from an aspect of behavioural or physiological biology. A consequence of this is that one population will receive genetic material without reciprocal transfer. A good example of this in the context of inter-population genetic exchange is between bison, *Bison bison*, populations and domestic cattle, *Bos taurus* (Ward et al, 2001). In this case

the differential is a direct result of asymmetry of mating preference and male hybrid sterility resulting in the exclusion of male cattle haplotypes from entering the bison population. In a Scottish hybrid zone between Sika and Red deer, the much larger red deer were thought to be unlikely to mate with the introduced Sika (Goodman et al, 1999). But despite the classification of all 246 animals into one of two distinct types; Red type and Sika type, admixture was detected. This showed a higher rate of backcrossing into the Sika population than into the red deer population per generation. This rate was seen to be very dependant on the demographic fluctuations in each population and perhaps any imbalance can be explained in this way. Although low in incidence here, admixture can vary according to population dynamics and could increase with expansion of population ranges.

1.4.1. Admixture affecting conservation

Conservation status is paramount in the management of populations and can be compromised through genetic introgression. Applying conservation status to a taxon can be dependent on a great variety of factors, among them extant numbers and genetic diversity being notable (King and Burke, 1999). Where introgression is suspected, careful consideration has to be taken as to the extent and the consequences of the introduction of this new material. Devaluation of populations of proposed conservation status can readily occur, often through admixture with a common taxon. Examples of this include; the Scottish wildcat populations that coexist with domestic cats (Beaumont et al, 2001), the rare Red wolf and the common coyote in Southeastern United States (Miller et al, 2003), and the wild wolf and domestic dog in Europe (Randi and Lucchini, 2002).

Preventing the loss of genetic diversity is among the most important contemporary issues in conservation management (Frankham 1995). Increasing the variation of a population through addition of variation from another source has the consequence of homogenisation of the two populations involved. In such a situation the loss of a unit of diversity, whether population or species, occurs when populations are no longer distinguishable from each other. This effect is the reason for much resistance in programs of species reconstruction, such as that involving the Florida panther that seek to reverse the decline of small inbred populations through crossing or upgrading (Land and Lacy, 2000). As well as being able to monitor admixed individuals in an upgrade programme of this kind it is also possible to use this ability of admixture

detection for the exclusion of admixed individuals entirely. This is an invaluable tool if reintroduction or founding populations are to be created in a habitat. One example of molecular selection with exclusion of hybrid animals is that of the Siamese crocodile, *Crocodylus siamensis*, for reintroduction into Vietnam (Fitzsimmons *et al*, 2002). Hybrids formed with the Cuban crocodile, *C. rhombifer*, could not be readily distinguished through phenotype so specific genetic markers were used instead. If marker loci produce alleles from a different species animals are excluded from the reintroduction or breeding programme. Benefits of this kind of study extend to the suggestion that marker heterozygosity may reflect wider heterozygosity within a genome (Moritz, 1999) therefore the most genetically diverse non-introgressed individuals could be chosen.

The introduction of new genetic material into a population can also be a beneficial process, particularly if the population has low prior genetic variation. This kind of introgression can increase the probability of persistence of a population, usually with greatest effect in small or declining populations suffering from inbreeding depression. This was seen in an adder introduction in Sweden (Madsen et al, 1999). In this example an isolated declining population was picked for an introduction of genetically variable males from other populations. The population was left for four generations and then surviving introduced males removed. Subsequent to the maturation of the progeny from these breeding seasons there was a significant increase in recruitment due to a sharp decrease in juvenile mortality. The corresponding increase in genetic variability of the newly recruited individuals supports the increased recruitment as a release from inbreeding depression. Other examples of introduction of genetic material between sub-species can be found in management applications, as was the case for additions of Texas puma to a Florida panther population (Land and Lacy, 2000). Detailed strategies can be formed for the genetic upgrading without loss of adaptive variation as was considered in this case (Hedrick, 1995).

1.5. Quantification of admixture

The study of admixture events and genetic introgression has become well established since initial concepts of introgressive hybridisation (Anderson, 1949) and mixture proportions in a hybrid population (Glass and Li, 1953). But early solutions to the admixture problem did not consider changes in gene frequencies between the contemporary samples and the ancestral (pre-admixture) populations (e.g. Elston,

1971). Therefore no account was made for processes with a stochastic element such as selection and genetic drift. Thompson (1973) subsequently introduced such a component in the hybrid population only, in order to model drift. But early studies had to assume that the admixed population had undergone enough time (in generations) of random mating for elimination of the allelic associations due to the admixture event. More recently statistical methods have been developed to overcome this and elucidate structure in recently admixed populations on an individual basis (Rannala and Mountain 1997; Paetkau et al.1995; Pritchard, et al. 2000). There has been a concurrent development in methods of determining admixture proportions at the population level (Bertorelle and Excoffier 1998; Chikhi et al. 2001; Wang, 2003) using several major approaches, these are detailed below.

1.5.1. Different methods of determining admixture proportions

There are a number of methods that have been developed and applied in the estimation of parental contributions after an admixture event. In general they are based on a simplified model scenario in which two parental populations, which have diverged from a common origin, subsequently meet to create a hybrid population before separating again (Figure 1.1).

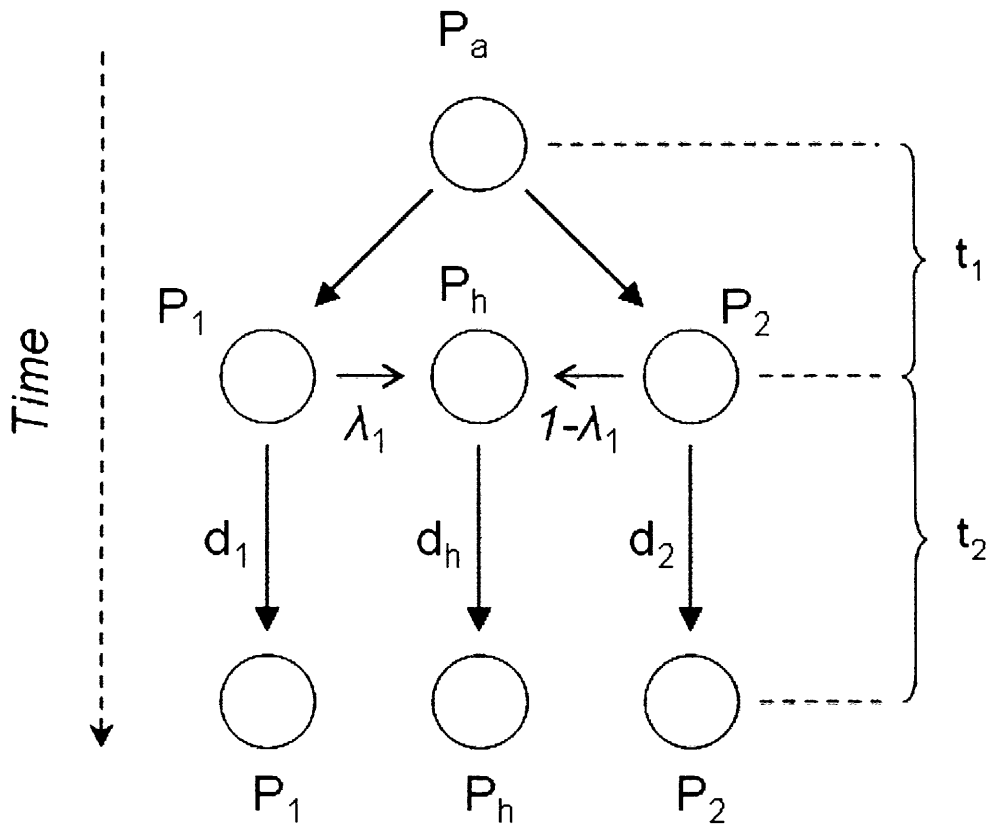


Figure 1.1. A common representation of a simplified admixture scenario. Pa represents the ancestral origin of the parental populations P1 and P2. The hybrid Ph is the consequence of genetic input from each parental population, λ1 and 1- λ1 respectively. The time since separation of the parental populations is represented by t1, and the time between the admixture event and the present is t2. The time-scaled demographic factors affecting the evolution of the populations are represented by d1, dh, and d2 respectively (adapted from Choisy et al. 2004).

The methods that can be applied to genotypic data are varied but all use allele frequency information in the calculations. What follows is a brief summary description of the premise behind each of the general methodological approaches that can be applied to investigate admixture. The first four of these methods use allele frequencies directly and are based on the application of a linear projection of the allele frequencies in the parental populations to the hybrid. This is from Bernstein's equation (1931) and shows how the frequency $p_i^{(h)}$ of allele i in the hybrid is a result of the allele frequencies $p_i^{(1)}$ and $p_i^{(2)}$ in the parental populations (Equation 1).

$$p_i^{(h)} = \lambda p_i^{(1)} + (1 - \lambda) p_i^{(2)} \qquad \textbf{(Equation 1)}$$

1.5.1.1. Gene identities

This method derives a relationship with gene identities that reflects that of allele frequencies (Chakraborty, 1975). Using arithmetic means of the probabilities of gene identity over all loci within population P_1, between populations P_1 and P_2 , and between populations P_1 and the hybrid. The estimation of gene identity coefficients allows estimation of the admixture proportions.

1.5.1.2. Private alleles Madansky's regression

Bernstein's equation can be rearranged for any allele to take the form of a linear equation as originally suggested by Roberts and Hiorns (1962, 1965).

$$p_i^{(h)} = \lambda(p_i^{(1)} - p_i^{(2)}) + p_i^{(2)} \qquad \textbf{(Equation 2)}$$

Therefore the admixture proportion λ can be seen as being the slope of the linear regression (Equation 2). This slope can then be estimated through a least square formula across I independent alleles (total allelic number minus number of loci).

Using the concept of the private allele, one whose presence is found in only one group of individuals (Neel, 1973), this method simplifies the Bernstein equation to $p_i^{(h)} = \lambda p_i^{(1)}$. The original estimator developed by Madansky (1959) is a least square regression with private allele frequencies in parental and hybrid populations as covariates but is estimated with potentially large errors. This error is reduced substantially where estimates are possible from alleles private to P_1 and P_2 respectively.

1.5.1.3 Maximum likelihood

The maximum likelihood approach to admixture estimation has been continually improved since early application by authors such as Roberts and Hiorns (1965). The likelihood of the hybrid sample genotypes are expressed as a function of the allele frequencies in the parental populations and the admixture proportion. Across independent loci the likelihood of any particular multilocus genotype is calculated through the likelihood at each locus, with the likelihood of the total hybrid population sample being a multiplication of likelihoods across individuals. The advantage of the method is its optimum use of the data (i.e. without losing data through summary statistics) but this carries with it the consequence of high computational demand. A recent application by Wang (2003) applies a pseudo maximum likelihood method to simplify calculations and reduce computational expenditure. The multi-dimensional

joint probabilities of the data are approximated using the product of the marginal probabilities as well as employing a transition matrix and using a hidden Markov chain algorithm (Wang, 2001).

1.5.1.4. Coalescence times

In order to account for both allele frequencies and molecular information these methods use a coalescent approach to estimate the mean coalescence time for a pair of genes. Ancestral lines coalesce when two ancestors in the sample share a common ancestor, in this case one allele from the hybrid and one from a parental population (Griffiths and Tavaré, 1994). The coalescence event can manifest in one of two ways; the ancestral gene line for the hybrid switches to the same parental population as the second gene of the pair and the two coalesce back to the admixture event, the alternative is that the gene switches to the other parental population and coalescence will only occur in the ancestral population (P_a). Therefore the probability of each of these routes of coalescence is either λ or $1-\lambda$, (or $1-\lambda$ and λ when considering the other parent population (Choisy et al. 2004). An example of a recent application of this method is that of Bertorelle and Excoffier (1998) that measures two estimators m_x and m_y, the latter of which considers the coalescence of the genes in the parental populations in addition to those of the hybrid. The use of this additional information of molecular differences between parent and hybrid populations improves the performance of the proportion estimate.

1.5.1.5. Monte Carlo Markov chain method

The approach to full likelihood coalescent simulation is one which is designed for application onto larger datasets without the problems associated with direct simulation. The application of a Monte Carlo Markov chain is a statistical approach which avoids independent simulations at different parameter values. The likelihood function is derived through approximations from the method of Griffiths and Tavaré (1994). The parameter space is explored through a stepwise process whereby the transition between states is dependant on the likelihood of the data at those parameter values. This is the method used in Chikhi *et al.* (2001), to estimate the combined probability of the three observed gene samples of parental and hybrid populations. The process involves simulating a coalescence tree of gene lineages that ends just before the time of the admixture event. The importance sampling

scheme computation is completed through consideration of the vectors of allelic distributions in the parent populations as draws from a specific distribution. The parental proportion is taken directly from the posterior distribution.

1.5.2. Methodological comparisons

The five methods highlighted above were compared as to their relative performance under a range of admixture circumstances by Choisy *et al.* (2004). Similar accuracy was seen between all of the methods in the analysis of recent admixture with highly differentiated parental populations, determined as being 'optimal conditions', but performance differences varied across a range of other applied circumstances. Using the Gene Identity method for an admixture event 1000-generations ago, under 20% of runs resulted in the actual parental proportion falling within the confidence interval of the method. As well as the age of admixture, parental differentiation was particularly important in method performance. The overall conclusions to be drawn from the Choisy et al. (2004) comparisons were that the Gene Identity and Private Allele Regression methods are limited in their applicability and suffer a general reduction in performance with increasing time since the admixture event. The remaining approaches were all found to be highly applicable, hence could be used in a wide variety of scenarios, and were shown to perform well in the majority of applied situations. Of the three latter approaches there are associated advantages and limitations; the likelihood method gave consistently good estimation apart from those for low differentiation between parentals and ancient admixture, the method based on coalescence times (Bertorelle and Excoffier, 1998) did have slightly higher variance of estimations, and the Markov chain method (Chikhi et al. 2001) was time consuming but accurate for low parental differentiation.

It is desirable for the method applied to be efficient in its use of the data, but this often has implications on the computational expenditure. Approaches like Bertorelle and Excoffier (1998) do not use all of the information in the data, but instead summarise into 'moments' which attempt to capture the properties of the distributions into single values. Notably, Wang (2003) summarises admixture approaches into two broad categories, moment and likelihood estimators. The former, as mentioned previously, has reduced statistical power due to only using a few moments of a distribution (allele frequency or coalescence time) but is simpler to calculate because of this. Likelihood based methods can be further separated into full likelihood and

partial likelihood methods. In a Bayesian approach, using a partial likelihood provides a convenient approximation and simplification process which allows bypassing of the prior distribution of nuisance parameters (Cox, 1975). In contrast, the full likelihood method of Chikhi et al. (2001) uses the Metropolis-Hastings algorithm to integrate over the nuisance parameters. It is the time and computational expenditure that have limited the wide testing of this full likelihood approach.

1.5.3. Approximate Bayesian computation

Following the increased use of Bayesian methods in biology there has been a drive toward reducing computational costs associated with different methodological approaches. Approximate Bayesian Computation (ABC) is a relatively recently developed approach. ABC attempts to address the computational problems caused by large numbers of nuisance parameters which currently limits Bayesian methods (Beaumont *et al.* 2002). The ABC approach is based on using a rejection-sampling algorithm that generates an approximate posterior for a given parameter set. This is possible due the ability of advances in stochastic simulation methods to generate sample data using a particular model (e.g. Hudson, 2002). Once the summary statistic is calculated for the observed dataset, simulated data is generated according to a model and a summary statistic is calculated and accepted or rejected according to a given tolerance. Despite this reliance on summary statistics, which do not use all of the data, it has been shown to be accurate when compared to full likelihood methods (Beaumont *et al.* 2002; Marjoram *et al.* 2003). As such the ABC method can be applied to more complex scenarios, provided that the model can simulate the data. An approximate Bayesian method has been applied previously by Excoffier *et al.* (2005) and has shown a matching performance to recent maximum likelihood methods with an increased accuracy in calculating ancient admixture. This method has a very similar model basis to the other methods discussed previously and is specifically designed to take mutations into account. The method is also comparable in computational expense with previous methods suggesting that ABC is a potentially advantageous approach.

1.6. Domestic animal populations

The results of anthropogenic utilisation of animals and plants can be seen in domesticated forms found worldwide, animal populations alone comprising several thousand breeds across some sixteen species (DAD-IS,2007). Selective breeding

and population isolation has produced great diversity from a small selection of species over a relatively short timeframe. In cattle (~1500 breeds) the first evidence of domestication can be found at sites such as Çatal Hüyük in Turkey around 7800 year ago (Perkins, 1969). Many breeds have resulted from high degrees of specialisation in utility and preference through both human-mediated artificial selection and environmental adaptation. This combination of high levels of differentiation between populations of the same species, as well as unparalleled levels of parentage data and ease of study have made these domestic populations invaluable subjects for study. Genetic investigations that have been applied to domestic livestock include straightforward genotype-phenotype relationships (Andersson, 2001), including quantitative traits such as milk production in cattle (Georges *et al.* 1995), as well as studies using livestock as models for human congenital diseases (Patterson et al. 1982). Population-level studies are also commonplace, focussing on within-breed genetic variation and inbreeding (e.g. Blott et al. 1998a; Machugh et al. 1998; Hanslik et al. 2000; Kantanen et al. 2000; Giovambatista et al. 2001; Cleveland et al. 2005), among breed variation (e.g. Beja-Pereira et al. 2003) and using a variety of measures to re-construct breed relationships (e.g. Casellas et al. 2004) and even origins (Caramelli, 2006). Applications of these methods are being used to make management recommendations (e.g. Freeman et al, 2004) as well as to refine genetic models of population level processes (van Hooft *et al*, 1999). The changing socio-economic environment can put breeds at risk creating the need for conservation methods to be applied to domestic populations (Taberlet et al. 2008). Rather than providing simple models for natural systems however, domestic populations also provide an insight into studies of selection and advancing methods in animal breeding (Bijma et al. 2001).

1.7. Man-made populations – a brief history of domestic cattle

Modern cattle have been shown to have originated from at least two major morphological groups of the widespread Pleistocene species, the wild aurochs *Bos primigenius,* with a distribution throughout Europe, Asia, and North Africa (Clutton-Brock, 1987). These two groups are recognised as the humped *Bos primigenius indicus*, and the humpless *Bos primigenius taurus* (Manwell and Baker, 1980). *B. indicus* or zebu cattle are thought to have expanded from the Baluchistan region, whilst *B. taurus* is from the Near East, possibly North Africa (Kumar *et al*, 2003). This

has given rise to their current distribution as they spread out from these regions (Figure 1.2). Mitochondrial sequence data has placed their divergence from a common ancestor between 200,000 and 1 million years ago (Loftus *et al*, 1994). From archaeological and genetic evidence it is thought that domestic cattle originated from each of these progenitors through separate domestication events (Kumar *et al*, 2003). Subsequent to their separate domestications, there is evidence of genetic exchange between these taurine-type and indicine-type cattle resulting in the current distributions and forms (Hanotte *et al*. 2002). This distribution gives us valuable insight into the initial dynamics of domesticated cattle in and around Africa and the Near East.

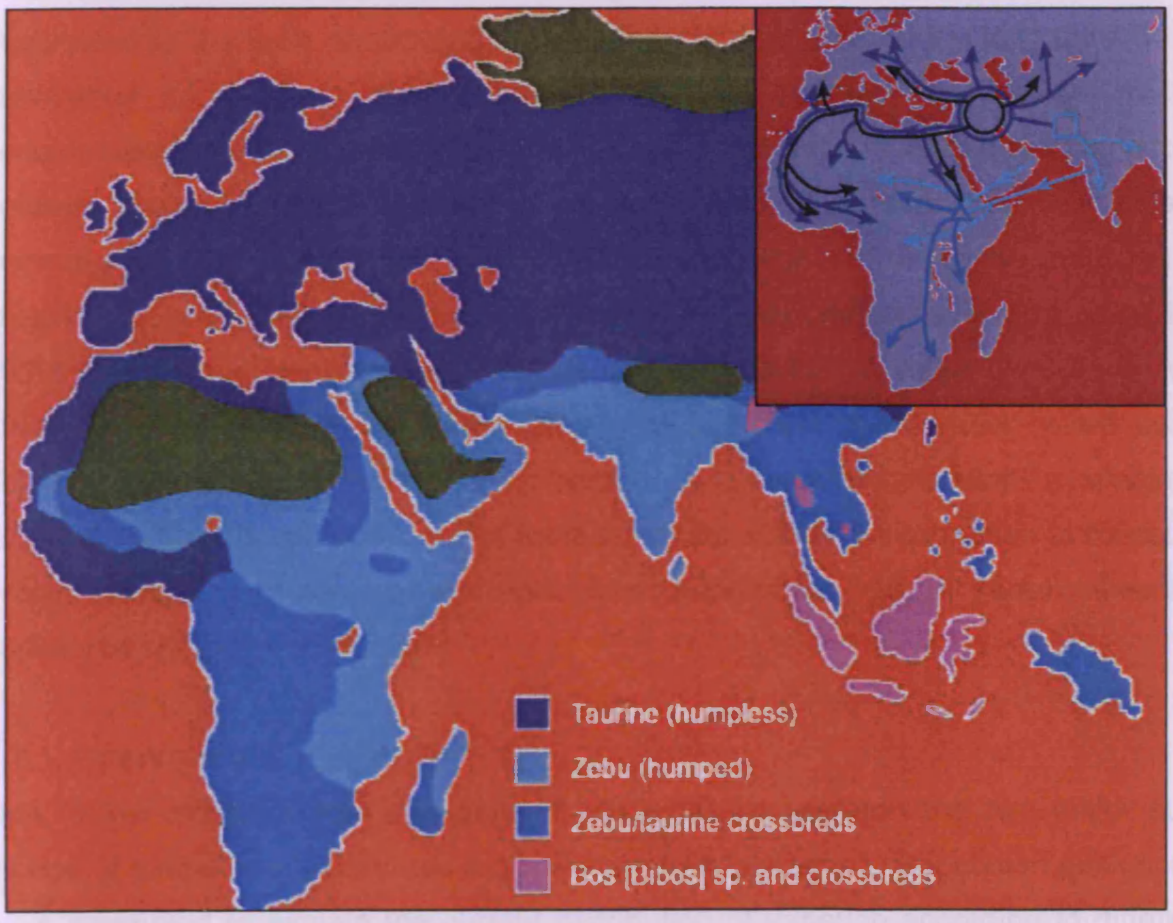


Figure 1.2. A geographical representation of the distribution of the major types of extant domestic cattle, including the proposed pattern of dispersal (inset). The patterns of dispersal for the two different taurine forms; shorthorn (black) and longhorn (dark blue) are identified separately (Adapted from: Payne, 1970; Epstein and Mason, 1984).

The domestication of cattle was associated with a rapid recent population expansion relative to other bovids (Finlay *et al*. 2007). The development of different morphology

since the use of cattle for meat, milk and as a utility animal has resulted in a diversity of extant forms. There are now a great many types of cattle worldwide estimated at around 800 distinct breeds in the mid-20th century (Moazami-Goudarzi *et al*, 1997) the current figures stand at around 1400 (Ajimone-Marsan, 2008). The cultural and environmental driving forces behind this diversification process are dynamic. Selection for characteristics such as trypanosomiasis and drought resistance are particularly important in breed specialisation in parts of Africa and India.

More recently, and particularly in more technologically developed countries, strong emphasis has been placed on increasing livestock productivity. This has been manifested in the form of strong trait selection and has promoted a shift toward the prevalence of a few specialist breeds, such as the Holstein-Friesian. Breed development has been accelerated by the ease of exchange of breeding stock between regions and the introduction of new reproductive techniques, particularly artificial insemination (AI). This has led to increased concern about the reduction of genetic diversity, particularly through the loss of traditional breeds, and since the 1980s there have been renewed efforts of conservation of cattle diversity (FAO, 1981). It is important to determine breeds of particular conservation value using objective criteria so that the loss of variation does not restrict the options available to tackle unknown future scenarios. As there is not yet sufficient information to measure breeds according to all genes of agricultural importance, overall genetic diversity conserved should be maximised.

1.7.1. Coarse-scale gene flow

Due to the morphological and genetic differentiation between the two major sub-groups of cattle, much work has focussed on their interaction. Population geneticists have attempted to explain the current distribution of variation according to different models of gene flow. In 2000 Hanotte *et al.* showed a clustered pattern of microsatellite genotypes across Africa implying definite structure to cattle distribution. They identified geographical areas which corresponded to those animals of 'indicine' morphology and implied regions where high levels of introgression from Zebu-type cattle had occurred. This potentially indicates the regional origins of cattle types and infers likely patterns of movement between regions. The use of out-crossing for trait acquisition and the subsequent selection process is commonplace in domestic organisms. The crossing between indicine and taurine cattle to confer beneficial traits

has been used extensively across the African and Indian regions. Desirable traits, such as the Zebu tolerance for heat and humidity, have been exchanged between populations in this way. A study of admixture across Asian breeds similarly confirmed taurine influence in the Indian subcontinent, again showing patterns of movement to and from Africa (Kumar *et al*, 2003). The level of taurine alleles in the Asian cattle breeds is greater in those populations geographically proximal to the Near East. The temporal aspect of this admixture has not been determined however, and the present genetic distribution may have developed over a long time period or could even be the result of high levels of recent upgrading crosses (Felius, 1995). However, faunal and artistic evidence in these regions from the third millennium BC (Zeuner, 1963) discounts recent gene flow as the sole explanation. Diagrammatic representations confirm that there was movement of many human-associated species between the two regions as well as clear depictions of the different cattle types themselves.

Despite this high degree of genetic exchange across the Near East, there is very little evidence of Indicine contribution seen in the cattle that were brought into Europe. The two routes of entry of taurine cattle were from North Africa and across from Western Asia (Figure 1.2; Caramelli, 2006). The original *B. primigenius* found in Europe has been extinct for 400 years, but there is suggestion that this represented a third cattle domestication event (Loftus *et al.* 1994; Bradley *et al.* 1996), prior to the influx into Europe of stock from Western Asia. Furthermore, there is evidence that this original *B. primigenius* was introgressed into the introduced *B. taurus* before becoming extinct (Bailey *et al.* 1996). Contemporary European breeds express a high degree of differentiation from both the taurine, and particularly indicine breeds found closer to the Near Eastern origins of domestication.

1.8 European cattle

The high number of breeds across Europe has resulted in a large proportion of domestic livestock studies focussing on the region. Europe not only has the highest breed diversity in the world, but as might be expected from this, it is also the region with the greatest number of extinct breeds (Scherf, 2000). As a result of the colonisation of many parts of the world by European human populations there has been an associated distribution of domestic livestock. Many breeds originating in Europe now have populations distributed across several continents e.g. Aberdeen Angus (Pimentel et al. 2003), Hereford (Blott et al. 1998a), and Holstein-Friesian

(Hanslik et al. 2000). The value in understanding the dynamics of breeds in Europe is clear, particularly if warnings of breed extinctions (Taberlet et al. 2007) are to be heeded.

1.8.1. Assessing genetic variation

An extensive literature has developed in characterising the genetic diversity in domestic cattle in Europe. The progression of molecular assessment of this diversity has meant a degree of standardisation of methodology. Indeed, microsatellite markers from a set recommended by the Food and Agriculture Organisation (FAO) were used in 79% of domestic livestock studies in a review by Baumung et al. (2004). This enables a greater degree of comparison between studies alongside other practises such as the combination of datasets through the overlap of genotyping (Freeman et al. 2006). There has concurrently been development of techniques allowing greater numbers of markers and individuals to be used in studies, a recent study typing over 1000 cattle across 81 bi-allelic loci (Negrini et al. 2007). One consequence of this is that of the expansion of studies, previously restricted to specific regions or breed groups, allowing studies to encompass large geographic regions and numerous breeds. Consideration of the population dynamics of many breed populations can help to explain the effects of gene flow across regions. Diversity patterns between large groups of breeds can be important in determining breed composition which may be absent on a smaller scale. The consequence of commercialisation leading to differential upgrading of populations in Eastern Europe and the Balkans is a good example (Li et al. 2007). In this case there is a separation of the native traditional populations from those that have been introgressed with commercial breeds across a wide area. It is possible to determine breed membership as being to either commercial or traditional breed-groups, with commercial breeds often displacing less productive breed populations (e.g. Kantanen et al. 2000).

Many cattle breed studies are performed according to country, with examples including; Argentina (Giovambattista 2001), Australia (Harper et al. 1998), Belgium (Peelman et al. 1998), Chirikof Island, Alaska (MacNeil et al. 2007), France (Boichard et al. 1996), Hungary (Bartosiewicz, 1997), Italy (Ciampolini, 1995; Moioli et al. 2004), Jersey Island, U.K. (Chikhi et al. 2004), Portugal (Mateus et al. 2004), Spain, (Arranz et al. 2006), Switzerland (Schmid et al. 1999), and the United States of America (Cleveland, 2005). This is the level at which application of national

management and conservation actions are typically applied. Studies at this scale regularly include measures of heterozygosity at neutral markers and a measure of breed distinction often through a genetic distance or principal component analysis. Morphological characteristics can be used to indicate breed relationships and colour type has been found to be a strong indicator of variation in underlying loci and useful for describing breeds (Klungland, 2000). However, there are lower levels of genetic differentiation found between breeds than their divergent morphology might suggest (Wiener et al. 2004). Even so, studies of cattle breed relationships have found many breeds to be significantly differentiated at the genetic level (Machugh, 1996) but this does not mean that the relationships between them are clear. Traditionally applied inter-specific methods may not be appropriate to describe relationships between breeds. This has led to difficulty in the meaningful application of genetic distances between breeds, particularly when there are high levels of gene flow, which results in poor statistical support for relationship clusters (Blott et al. 1998b). Contradictory findings across different studies can be found; for example the relationships between the Holstein, Simmental, and Swiss Brown breeds vary between authors (Machugh et al. 1998; Gryzbowski et al. 2004; Schmid et al. 1999). Breed relationships can also be vulnerable to choice of population, in this last example the German and Swiss Simmental populations differ markedly through the stepwise weighted genetic distance method (Gryzbowski et al. 2004) affecting higher order relationships.

There has been a relatively recent transition towards newer clustering algorithms in cattle studies (worldwide examples) as well as migrant assignment methods (e.g. as applied in Moioli et al. 2004). This transition follows concern that using traditional phylogenetic methods based on genetic distances to set conservation priorities are inappropriate for breed conservation (Caballero and Toro, 2002). But simple genetic distance methods are still commonly applied on inter-breed relationships (Tapio *et al.* 2006; Li *et al.* 2007; MacNeil *et al.* 2007). There can be consensus in breed relationships as seen by Edwards *et al.* (2000) where a genetic distance approach was taken to elucidate relationships between the endangered Pustertaler-Sprinzen and three other European breeds. In this case the breed relationships found through distance methods were shown to be supported by historical information about breed relationships. But even those distance methods which account for drift, such as Reynold's distance, are prone to major inflation due to genetic substructure, inbreeding and extreme drift. The result is many breeds with similarly high distance

values. Consequently any trees produced are unstable (i.e. have no statistical support) and random in the connections they make. Trees, which largely assume no reticulation and often equal rates of evolution, are simply inappropriate for domestic populations (Bruford, 2004). This has subsequent consequences for other methods that rely on genetic distance determinations such as measures of breed distinction (e.g. WEITZPRO by Derban et al. 2002). Not only can clustering algorithms allow the specification of populations but can be used to indicate assignment of individuals to populations as has been used in studies of breed admixture (Freeman et al. 2006). It is also recognised that assignment tests are becoming important in other areas such as through tracing livestock products (Ciampolini et al. 2006).

Domestic lineages have been subject to various evolutionary and demographic processes at different stages throughout their histories. Contemporary populations have been shaped through repeated incidences of founder effects and population bottlenecks, combined with small population sizes and a variety of selective forces both natural and anthropogenic (Blott *et al*, 1998a). Extant diversity and inbreeding measures can help assess how recent demographic fluctuations have affected current populations but can also allow predictions to be made. Kantanen et al. (1999) demonstrated the difference between temporally removed populations using a genetic distance method to compare contemporary and historical genetic information. In this case the allelic frequencies observed showed very little change over time. It was also possible to estimate migration rates between breeds demonstrating that the populations were not closed and displayed relatively high levels of gene flow. Population founder effects and isolation can result in the differentiation of same-breed populations across international borders. The difficulties in characterisation of particular breed populations spread between countries can lead to separation in phylogenies as suggested previously in German and Swiss Simmental populations (Gryzbowski et al. 2004). Many breeds in the UK enforce a closed herd book, Jersey cattle are a good example. The population on the island of Jersey has demonstrated that a closed herd book can maintain an overall genetic diversity that is well represented across its island range (Chikhi *et al*, 2004). The Jersey comprises approximately 4000 individuals and was investigated under concern that loss of genetic diversity and increased inbreeding may result from the absence of imported individuals onto the island. This was not the case and the success of traditional management practises in the Jersey shows the potential to maintain diversity in a

closed herd book. This is increasingly important with the advent of new technologies that have the potential to reduce the effective population size. From being simply a record keeping system, the herd book now has the potential to allow rapid reference to the incidence of individuals and bloodlines in a herd and can allow choice of sires to introduce new blood, preventing inbreeding effects particularly during selection (e.g. Woolliams and Bijma, 2000).

1.8.2. Breed conservation

In addition to European breed literature, recent studies of cattle breed conservation have been made worldwide, examples ranging from Asia (Kim *et al*, 2003), Africa (Hanotte *et al*, 2000), to the New World (Russell *et al*, 2000). Concern for domestic livestock conservation is increasing (e.g. MacHugh, 1997; Taberlet et al. 2007) alongside a growing recognition of the need to conserve domestic cattle in less developed nations (Reist-Marti *et al*, 2005). Livestock breeds worldwide are recognised as being important to biodiversity (Hall and Ruane, 1993), and around a third of these are considered under threat of extinction (Simianer et al. 2003).

Optimisation of allocation of resources for livestock conservation has been developed in order to minimise loss of genetic diversity between breeds (Simianer, 2003), but this has been without being based on molecular analysis (Lenstra et al. 2006). There has been considerable discussion as to the most appropriate approach for assigning value according to genetic diversity measures. Ruane (1999) criticised the commonly used genetic distance methods and their implications for breed individuality and conservation and suggested that a molecular approach should be one of a number of criteria applied. Simianer *et al.* (2003) continued this argument, saying that there should be conservation priority attributed according to a more utilitarian concept of value in addition to classic diversity measures. It is generally accepted as being essential to employ a wide range of criteria in order to make an informed decision about value and determination of conservation priorities. However, due to the difficulty in measurement and lack of standardisation of many of these factors, studies often concentrate on particular criteria with greater perceived importance. To date these have included; level of endangerment (Danell *et al.* 1998; Simianer, 2005), economic viability (Rege and Gibson, 2003), cultural value (Gandini and Villa, 2003), and socioeconomic functions of breeds (Tisdell, 2003).

The most commonly applied conservation assessment tool in cattle studies is the Weitzman approach (1992, 1993). Being genetic distance based, the main disadvantage of the Weitzman method is its failure to account for within-population genetic variation (Thaon D'arnoldi *et al.* 1998). Weitzman-based approaches to conservation priority assessment are still the most common tool applied to genetic data in cattle literature (e.g. Canon et al. 2001; Reist-Marti et al. 2003). Eding et al. (2001) developed an alternative method to rank breeds according to estimated minimised kinship and this has been further developed (Caballero and Toro, 2002). Both the Weitzman and minimised kinship methods have been applied to pig and cattle breeds repectively (Fabuel et al. 2004; Lenstra et al. 2006) reinforcing the vulnerability of the genetic distance based phylogenetic approach to inbreeding and genetic drift. The global coancestry approach performed particularly well against the Weitzman method when comparing diversity contributions by individual populations (Fabuel et al. 2004). These marginal value assessments were also questioned as to their utility in short or medium term application in assessment of diversity value related to extinction risk in Garcia et al. (2005). It is apparent that the ease of both application and interpretation of the Weitzman methodology encourages its use, and despite the advancement of alternative approaches it is still being applied in breed population studies (Tapio *et al.* 2006; Li *et al.* 2007; MacNeil *et al.* 2007). A potential solution to this may lie in the application of a set of independently applied summary methods that can be used easily and in parallel. The combination of heterozygosity measures (Weir and Cockerham, 1984), estimators of historical demographic events (e.g. BOTTLENECK (Cornuet and Luikart, 1996)), and an estimate of how important gene-flow or drift have been in shaping the population (e.g. GENECLASS2 (Piry et al. 2004), ESTIM (Vitalis and Couvet, 2001)) can all be used. There are also a range of commonly applied clustering algorithms that can be used to give indications of within-population segregation (STRUCTURE (Pritchard et al. 2000)) or between population groupings (PARTITION (Belkhir and Dawson, 2001), SAMOVA, BAPS4).

Successful conservation measures to ameliorate diminishing breed populations have been mounted, sometimes irrespective of their position as regards value in the wider breed community. Where breeds have been identified as threatened with imminent extinction, breeding programs have been applied as in the case of the Hungarian Grey example mentioned previously. The Hungarian Grey was subject to a sharp decline after the Second World War and subsequent recovery was facilitated by the

use of Maremman crossings to eliminate the risks of inbreeding and upgrade the traditional characteristics of the breed (Bartosiewicz, 1997). The White park breed similarly recovered from low population numbers through outcrossing with the Longhorn in the early 1900s (Hall and Clutton-Brock, 1989). But there is inherent reluctance in mounting management programs for introgression of genes from other breed populations due to the fear of loss of breed distinction. It is often only in the situation where a breed is in immediate risk of extinction that measures are taken. Whilst inter-breed gene flow can maintain genetic diversity, particularly where small populations or intensive selection are involved, it can also erode genetic distinctiveness realising the problem of 'extinction by hybridisation and introgression' coined by Rhymer and Symberloff (1996). Outbreeding for improved production traits remains a contentious issue, but increasing commercial viability may be the only way to maintain popularity and therefore retain some breed populations (Vollema and Groen, 1997).

1.8.3. Breed management

The breeds recognised across Europe today represent a process of environmental and anthropogenic selection across hundreds of years. The genesis of breed populations has until recently been a process of development of stock to efficiently fulfil the often combined needs of production and draught, in a local physical and economic environment. Contemporary breed evolution is driven through both production traits and by a selection process for stereotypical characteristics representing breed standards. The balance between these two major driving forces of largely commercial against largely cultural is greatly dependant on the breed itself. For many traditional breed populations there are a significant proportion of breeders aiming towards a particular phenotype of animal, represented through performance at individual animal competition level. As pointed out in Bartosiewicz (1997) this encompasses a culturally idiosyncratic concept based on how animals 'should look' (Kroeber and Richardson, 1940). Many characteristics attributed to a successful competition animal do not necessarily relate to production value but at the same time act as a (potentially dynamic) target for selective breeding.

A possible counterpoint to a 'breed standard' type of selective process for breed management and development is that of an outcrossing regime in order to actively select for particular characters from other populations. Because of the density of differentiated breed populations across Europe there is a high potential for breed

interaction. Gene flow between neighbouring breed populations is a dynamic process reducing the separation of European cattle breeds, as well as maintaining genetic variation within them. Whilst much inter-breed gene flow is thought to be accidental or at least unrecorded in herd books, there is a practice of active out-breeding programs. Discrete admixture events have occurred multiply in some breed populations in order to acquire novel traits. A good example of this admixture to maintain competitive production is that of the Lincoln Red breed. The Lincoln Red has had two managed admixture events in its recent history, once for the purpose of polling in 1963, and a further continental influence in the mid seventies to facilitate a shift from dual-purpose into beef production (S.J. Hall, pers. comm.). Many modern breed populations have originated from crosses between older breeds. The Murray Grey is a good example of this, being the product of Aberdeen Angus crossed with Shorthorn. Other breeds have been subject to far more recent introgression as in the Dairy Shorthorn (Blott *et al*, 1998b) and Holstein Friesian (Hanslik *et al*, 2000). European Holstein Friesian cattle provide a good example of a breed having experienced recent introgression, in this case from imported and genetically distinct New World Holstein Friesians. After separation of around 200 years there has been gene-flow since the 1960's into Europe and in some cases this has been substantial, as is the case in some German herds. Despite this introgression from the New World most European populations are still significantly differentiated. Breed distinction also has a definite importance economically and it is not always the case that high production commercial breeds command the market. There is an economic climate of increased interest in products from rare and endangered cattle breeds. Specialist commercial value is now attached to the milk and beef of such breeds as Guernsey and Aberdeen Angus respectively as well as many others (Blott *et al*, 1999). This benefit of breed individuality can be seen as directly conflicting with the selection for those economically beneficial traits that act to homogenise them (Ciampolini *et al*, 1995).

1.8.4. Agricultural progression

The selective pressures acting on a population are important in determining the characteristics of its constituent individuals. For important commercial cattle breeds the selection pressures come from artificial manipulation towards the development of economically desirable traits. The selection for commercial traits in many traditional breeds is generally not as strong due to the constraints of breed identity and the fact

that many are kept in small herds as subsidiary income or for hobby purposes. The relaxation of anthropogenic selective pressures in this way often allow breeds to become more genetically diverse and could even allow a degree of local environmental adaptation (Giovambattista *et al*, 2001).This could be facilitated through schemes that specifically promote use of traditional breeds such as for conservation grazing. Rough pastures and difficult terrain are often not compatible with the larger and less agile commercial breeds and their production systems so in the use of traditional breeds, maintenance of pasture in a traditional manner is still possible (Hobbs, 1992).

Artificial Insemination (AI) has been proposed as causing a reduction in overall reproductive population sizes seen in a breed where the same males are used far more than would have been possible previously. Reducing effective population size can then have a detrimental effect on extant diversity in a population. The effect of AI and embryo transfer methods was tested in this respect in the Aberdeen Angus in the United Kingdom (Vasconcellos *et al*, 2003), a highly specialised beef breed that uses these technologies more than many others. In comparison with eight other breeds the Angus did not show any reduction in average heterozygosity and gene diversity. But there may be more of an effect seen in smaller populations where AI technology is used, and problems may arise where individuals with mixed breed ancestry are somehow allowed onto an AI register. The concerns over AI and its effects have also been touched upon in the Dexter through pedigree analysis (Sheppy, 1998) and in the Holstein by molecular means (Miglior and Burnside, 1995). The former study examined the result of the effective reduction of numbers of bulls to those widely available through AI. The latter example supports this and refers to the consequence of inbreeding due to intense selective pressure combined with preferential use of few highly ranked bulls through AI. This difference of consequence of AI use may demonstrate the variation in outcome from the differential application of AI technology but it is clear that overuse of few males in a national herd will act towards homogenisation of the greater population. There is also the potential to increase the frequency of specific recessive genetic disorders present in the few males used which will be revealed when homozygosity increases. Taberlet et al. (2008) commented on the particularly extreme effects of AI on effective population sizes of commercial breeds. The most extreme example

is that of the French population of Holstein cattle composed of 2.5 million animals and estimated as having an effective population size of 46 (Boichard et al. 1996).

1.9. Introducing an out-bred British cattle breed: the Dexter

The modern Dexter cattle breed is one of a number of traditional specialist British cattle breeds still in existence. The breed is described in the literature as usually being black, but also can be red and dun (Moss, 1890; Mason, 1988). The Dexter has been the focus of a number of studies alongside the traditional historical breed accounts detailing breed origin and early characteristics (e.g. Hooper, 1898; Moyles, 1959). Subjects covered include elucidation of the genetic basis of coat colour (Berryere et al. 2003) and investigation of the genetic disorder known as chondrodysplasia, not seen in other cattle populations (Harper, 1998). There have also been population genetic studies that have included Dexter cattle, with conflicting conclusions (Buys and Chiperzak, 1992; Blott et al. 1998b; Wiener et al. 2004). This is almost certainly due to a unique breed history and management resulting in a breed with high variation therefore leading to problems with such effects as ascertainment bias.

1.9.1. Breed origins

The Dexter is the smallest of the cattle breeds in the United Kingdom and records of its existence began in Ireland in the late 19th century with its inclusion into the Kerry and Dexter herd book in 1890. It spread throughout the UK and can now be found worldwide including populations in Australia, North America, and South Africa. Described as both beef and milk producing the breed was originally found in two colours; red and black. Size is particularly variable according to a heritable trait which produces leg length variants of short, medium, and long (Curran, 1990). The modern Dexter is largely true to the historical descriptions concerning the above characters as well as details of head, neck and body proportions.

Historical information is largely contradictory on the major influences in the Dexter breed since its origin from within the Kerry breed in Ireland. There are various accounts to suggest a wide range of different breeds involved but with no supporting information. A major difficulty lies in a huge demographic contraction of the Dexter breed in the UK between 1965 and 1975 where yearly numbers registered with the breed society fell dramatically. By the 1969-1970 herd book, numbers had fallen from

an average of around 25 to only 3 bull registrations, with cow registrations falling from a previous average of approximately 100 to a low of 35 in 1970. This population bottleneck prompted a relaxation of breed regulations to allow 'upgrading' outcrosses to be registered in the breed society herd book. An experimental register, established in the 1960's, brought both Aberdeen Angus and Jersey into the Dexter. By allowing introgression of other breeds in this way, there was an almost exponential rise in numbers of registrations over the following decade (Sheppy, 1998). Those breeds thought to have been used in this process include; Aberdeen Angus, Red Poll, Red Devon, Guernsey, Fresian, and Jersey (A.J. Sheppy, pers. comm.). Due to the paucity of information available for the breeds involved in this outcrossing the list may not be complete. At the point of the process of 'upgrading' there can be perceived to be a number of primary breed lineages distinguished through their founding sires as seen in Figure 1.3.

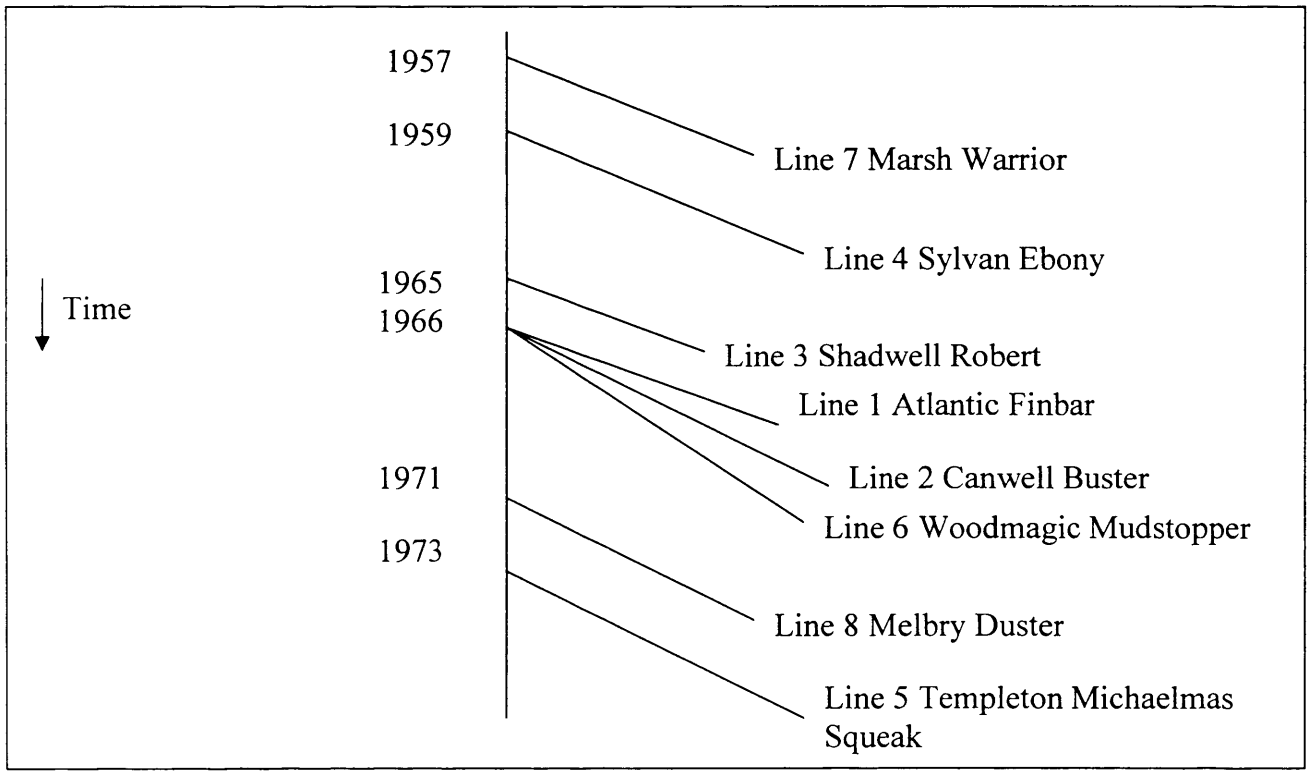


Figure1.3. The major identifiable lineages in the Dexter prior to the expansion of the breed after the 1965-1975 bottleneck, arranged with respect to birth years of the line founders.

In this basic scenario the eight principal lineages represent the extant Dexter diversity present prior to the majority of the introgression of the hypothesised breeds. These lineages can be identified less easily with each passing generation as admixture

occurs. Only one of these lineages can be identified with ease due to its historical isolation, that of Woodmagic. Due to this isolation it is thought to be entirely free from any upgraded animals to date and as such is an interesting measure against other purebred Dexters. However, this isolation has allowed the maintenance of characters abnormal to the breed and may have removed its importance as a possible breed standard in the same way as might be assumed from the other historical lineages (Sheppy, 1998). Woodmagic Dexters are smaller than those animals with the genetically determined long leg length without carrying the achondroplasia gene responsible for the short leg length condition (Rutherford, 2005).

Modern Dexter numbers are higher than ever previously recorded (~9000), resulting in their withdrawal from the list of 'rare' cattle held by the Rare Breeds Survival Trust (RBST) (Sheppy, 1998). However, the proportion of admixed individuals is unknown in the population and this has led to concern about the integrity and authenticity of the modern Dexter as a breed. Some conditions are accepted as historically Dexter such as the dwarfism condition whereby animals of both short and long legs can be identified (Brenig *et al*, 2003). In the modern Dexter there are various traits that appear to be previously unrecorded, and as such could be taken as only recently occurring. Two particularly obvious examples would be the incidence of hornless or polled individuals and a novel colour, 'dun'. In this latter example the genetic basis has now been established as a mutation in the Tyrosinase related protein 1 (TYRP1) which is involved in the coat colour pathway (Berryere *et al*, 2003), but it is unknown whether this originated an ancestor of the contemporary South Devon breed (A separate breed population to the Red Devon). Research into breed influences and origins of the Dexter are limited, but examples can be found in which close ties with the historically documented progenitor, Kerry, are questioned (Buys and Chiperzak, 1992) and closer relationships shown with other contemporary breeds such as the Aberdeen Angus (Blott, 1997). This is of particular interest due to the implication of introgression into the Dexter through outcrosses with breeds such as the polled Aberdeen Angus. Polled members of the Dexter breed are presumed to be from this source as Wilde (1858) stated that the Dexter and Kerry breeds were unrelated to any polled breeds in Ireland.

Through its inclusion in studies of breed relationships the British Dexter can be seen to be unremarkable with respect to levels of genetic diversity and genetic distances to

other breeds (Blott *et al*, 1998b; Wiener *et al*, 2004). This may be a consequence of the upgrading process with a number of separate breeds. In fact in the principal component analysis of Blott *et al*. the Dexter commanded a very central position between other breeds, but remained a separate entity perhaps due to the one way nature of Dexter crosses with other breeds. This noted, the Dexter was seen to have the lowest number of private alleles, twinned with the Highland, of three across 30 loci. Maximum numbers of private alleles in the study was ten with an average of about six (Wiener *et al*, 2004). The closest relationship to the Dexter was found to be with the Jersey, although it must be noted that the Kerry was absent from Wiener *et al*,'s study. Again the proximity of Jersey to Dexter is interesting because of implication of the Jersey in upgrading crosses.

1.9.2. Introgression in the Dexter

The most instantly intuitive method of admixture detection is that of visual determination through phenotypic characters. Traits unique to each parent population can be investigated for presence in the progeny of the cross. The drawbacks of such a method might include; lack of objectivity, potential for unusual expression of traits not found in the parent populations which may mislead the observer. Pelage has been used in the Scottish wildcat (Beaumont *et al*, 2001) and it was demonstrated that groups constructed in this way possessed informative differences in allele frequencies. However, the level of differentiation can be highly variable between species and the reliance on aspects such as pelage to differentiate two black cattle breeds, for example, has obvious limitations. Other more informative characters as given in breed descriptions may be useful if applied as a complete set and the presence of population specific traits will help determination in this way. Whilst not universally informative, this methodology can be swift to employ and give indications on the provenance of individuals of uncertain parentage that can be subsequently tested with other methods. It is not possible to know for certain whether an individual is a result of cross population mating even if it does not demonstrate abnormal morphology. Confounding this problem is that of the numbers of generations between an individual and a crossed ancestor. The proportion of nuclear genes from the other population diminishes (although this is not necessarily the case for mitochondrial or Y-chromosome DNA) and is increasingly unlikely to be represented in obvious characters.

Accurate pedigree information can be used in the identification of admixture since the proportion of genes likely to be present in an individual downstream from an introgression event can be calculated. Using the known number of generations between a parent and an individual and given that there is equal parental contribution for each parent in each progeny can be crudely numerically represented as a 50:50 split each generation (Chesser and Baker, 1996). This can be used to calculate average proportional influence of parental populations in an individual. As distance from a cross event changes per generation the proportion of introgressed material decreases by a half; 0.5, 0.25, 0.125 etc (assuming pure matings since the introgression event). This approach, however, can be time and information intensive and is subject to the validity of the pedigree itself which in itself is its greatest disadvantage. If the value of an individual would be compromised by the declaration of influence from another source outside the breed, it may be that parentage information is altered to disguise the fact.

1.10. Statement of aims

This project will apply a variety of molecular markers to genotype a sub-sample of the Dexter breed as well as several other breeds thought to have been introgressed into the breed at origin or since the population bottleneck. Using the allelic configuration found in individuals using these markers, the aim is to understand the recent admixture or introgression history of the Dexter breed. To carry out this analysis I typed individuals at a number of microsatellite loci. I also worked on the development of new Bayesian methods to determine relationships between the breeds involved. Using this type of modelling approach I aim to identify introgressed lineages within the Dexter breed.

This will be separated into several parts;

- Dexter relationships within a wide selection of European breeds.
- Dexter breed origins, the founder population contributions will be investigated as to their relationship with the identified subdivisions in the Dexter breed.
- An improved admixture detection method based on Approximate Bayesian Computation will be developed in order to explore breed dynamics more accurately.
- The novel method of admixture detection will be tested on an accurately documented case of introgression in the Lincoln Red cattle breed.
- The novel method will be applied to the Dexter breed to extract more information on the more complex admixture scenario that the breed presents.

1.11. References

Ajmone-Marsan P. (2008) Are cattle, sheep, and goats endangered species? *Molecular Ecology* **17**(1), 275-284.

Allee, W. C. (1931) Animal Aggregations. A study in General Sociology. University of Chicago Press, Chicago.

Andersson L. (2001) Genetic dissection of phenotypic diversity in farm animals. *Nature Reviews Genetics* **2**, 130-138.

Ardlie K.G., Kruglyak L., & Seielstad M. (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* **3**, 299-309.

Arnold M.L. (1997) *Natural Hybridisation and Evolution*. Oxford University Press, New York.

Arranz J.J., Bayón T., & San Primitivo F. (1996) Genetic variation at five microsatelllite loci in four breeds of cattle. *Journal of Agricultural Science* **127**, 533-528.

Avise J.C., & Hamrick J.L. (1997) *Conservation genetics case histories from nature*. Springer publishing.

Avise J.C. (2004) *Molecular markers, natural history and evolution*. Sunderland Massachusetts. Sinauer Associates.

Bailey J.F., Richards M., Macaulay V.A., Colson I.B., James T., Bradley D.G., Hedges R.E.M., & Sykes B. (1996) Ancient DNA suggests a recent expansion of European cattle from a diverse wild progenitor species. *Proclamations of the Royal Society of London B* **263,** 1467-1473.

Bartosiewicz L. (1997) The Hungarian grey cattle: a traditional European breed. *Agriculture* **21**, 49-60.

Bataillon T., & Kirkpatrick M. (2000) Inbreeding depression due to mildly deleterious mutations in finite populations: size does matter. *Genetic Research* **75**, 75-81.

Baumung R., Simianer H., & Hoffmann I. (2004) Genetic diversity studies in farm animals – a survey. *Journal of Animal Breeding Genetics* **121**, 361-373.

Beaumont M.A. (1999) Detecting population expansion and decline using microsatellites. *Genetics* **153**, 2013-2029.

Beaumont M, Barratt E.M., Gottelli D., Kitchener A.C., Daniels M.J., Pritchards J.K., & Bruford M.W. (2001) Genetic Diversity and introgression in the Scottish wildcat. *Molecular Ecology* **10**, 319-336.

Beaumont M.A., Zhang W., & Balding D. (2002) Approximate Bayesian Computation in population genetics. *Genetics* **162**, 2025-2035.

Beebee T. & Rowe G. (2004) *Molecular Ecology*. Oxford University Press.

Beja-Pereira A., Alexandrino P., Bessa I., Carretero Y., Dunner S., Ferrand N., Jordana J., Laloe D., Moazami-Goudarzi K., Sanchez A., & Canon J. (2003) Genetic characterisation of Southwestern European bovine breeds: A historical and biogeographical reassessment with a set of 16 microsatellites. *Journal of Heredity* **94**(3), 243-250.

Belkhir K., & Dawson K.J. (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research* **78**, 59-77.

Bernstein F. (1931) Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung. *Comitato Italiano Per Lo Studio Dei Problemi Della Populazione*, 227–243. Instituto Poligraphico Dello Stato, Rome.

Berryere T.G., Schmutz S.M., Schimpf R.J., Cowan C.M., & Potter J. (2003) TYRP1 is associated with dun coat colour in Dexter cattle or how now brown cow? *Animal Genetics* **34**(3), 169-175.

Bertorelle G., & Excoffier L. (1998) Inferring admixture proportion from molecular data. *Molecular Biology and Evolution* **15**, 1298–1311.

Bijma P., Van Arendonk J.A.M., & Woolliams J.A. (2001) Predicting rates of inbreeding for livestock improvement schemes. *Journal of Animal Science* **79**, 840-853.

Blott S.C., Williams J.L., & Haley C.S. (1998a) Genetic variation within the Hereford breed of cattle. *Animal Genetics* **29**(3), 203-211.

Blott S.C., Williams J.L., & Haley C.S. (1998b) Genetic relationships among European cattle breeds. *Animal Genetics* **29**, 273-282.

Blott S.C, Williams J.L, & Haley C.S. (1999) Discriminating among cattle breeds using genetic markers. *Heredity* **82**, 613-619.

Boichard D., Maignel L., Verrier E. (1996) Analyse généalogique des races bovines laitières françaises. *INRA Productions Animales* **9**, 323–335.

Bowcock A.M., Ruiz-Linares A., Tomfohrde J., Minch E., Kidd J.R., & Cavalli-Sforza L.L. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **369**, 455-457.

Bowen B.W. (1999) Preserving genes, species, or ecosystems? Healing the fractured foundations of conservation policy. *Molecular Ecology* **8**, 1-5.

Brenig B., Baumgartner B.G., Kriegsman B., Habermann F., Fries R., and Swalve H.S. (2003) Molecular cloning, mapping, and functional analysis of the bovine sulfate transporter SLC26a2 gene. *Gene* **319**, 161-166.

Bruford M.W. (2004) Conservation genetics of UK livestock: from molecules to management. In: G. Simm, B. Villanueva and S. Townsend (eds). '*Conservation of genetic resources*', University of Nottingham Press, Nottingham, UK, 151-169.

Buntjer J.B, Otsen M., Nijman I.J., Kuiper M.T.R., & Lenstra J.A. (2002) Phylogeny of bovine species based on AFLP fingerprinting. *Heredity* **88**, 46-51.

Buys C. & Chiperzak J. (1992) A comparative study of blood groups in the Kerry and Dexter cattle breeds, In: *Genetic Conservation of Domestic Livestock* (ed. By L. Alderson and L. Bodo), CAB international, Wallingford, UK.

Caballero A, and Toro M. (2002) Analysis of genetic diversity for the management of conserved subdivided populations. *Conservation genetics* **3**, 289-299.

Caramelli D. (2006) The origins of domesticated cattle. *Human Evolution* **21**, 107-122.

Casellas J., Jiminez N., Fina M., Tarres J., Sanchez A., & Piedrafita J. (2004) Genetic diversity measures of the bovine Alberes breed using microsatellites: variability among herds and types of coat colour. *Journal of Animal Breeding* **121**, 101-110.

Cavalli-Sforza L.L., & Edwards W.F. (1967) Phylogenetic analysis: models and estimation procedures. *Evolution* **21**(3), 550-570.

Chakraborty R. (1975) Estimation of race admixture – a new method, American *Journal of Physical Anthropology* **42**, 507-511.

Chakraborty R., & Weiss K.M. (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proclamations of the National Academy of Sciences, USA* **85**, 9119–9123.

Chesser R.K., & Baker J.B. (1996) Effective sizes and dynamics of uniparentally and diparentally inherited genes. *Genetics* **144**, 1225-1235.

Chikhi L., Bruford M. W., & Beaumont M.A. (2001) Estimation of admixture proportions: A likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**, 1347-1362.

Chikhi L., Goossens B., Treanor A., & Bruford M.W. (2004) Population genetic structure of and inbreeding in an insular cattle breed, the jersey, and its implications for genetic resource management. *Heredity* **92**, 396-401.

Chikhi L., & Bruford M.W. (2005) Mammalian population genetics and genomics. In "*Mammalian Genomics*" Eds. A. Ruvinsky and J. Marshall-Graves. CABI Publishing, Oxford, UK.

Choisy M., Franck P., & Cornuet J.M. (2004) Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Molecular Ecology* **13**, 955-968.

Ciampolini R., Moazami-Goudarzi K., Vaiman D., Dillmann C., Mazzanti E., Foulley J-L., Leveziel H., & Cianci D. (1995) Individual Multilocus Genotypes Using Microsatellite Polymorphisms to Permit the Analysis of the Genetic Variability Within and Between Italian Beef Cattle Breeds. *Journal of Animal Science* **73**, 3259-3268.

Cleveland M.A., Blackburn H.D., Enns R.M., & Garrick D.J. (2005) Changes in inbreeding of U.S. Herefords during the twentieth century. *Journal of Animal Science* **83**, 992-1001.

Clutton-Brock J. (1987) *A natural history of domesticated mammals.* Cambridge University Press.

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

Cox D.R. (1975) Partial Likelihood. *Biometrika* **62**(2), 269.

Crandall K.A., Bininda-Edwards O.R.P., Mace G.M., & Wayne R.K. (2000) Considering evolutionary processes in conservation biology. *Trends in Evolution and Ecology* **15**, 290-295.

Curran P. L. (1990) *Kerry and Dexter cattle*. Royal Dublin Society.

Danell B., Distl O., Gandini G., Georgoudis A., Groeneveld E., Martyniuk E., Ollivier L., van Arendonk J., & Woolliams J. (1998) The development of criteria for evaluating the degree of endangerment of livestock breeds in Europe. *Paper produced for EEAP working group on animal genetic resources*.

Derban S., Foulley J-L., & Ollivier L. (2002) WEITZPRO: a software for analysing genetic diversity.

Di Rienzo A., Peterson, A.C., Garza J.C., Valdes A.M., Slatkin M., & Freimer N.B. (1994) Mutational processes of simple sequence repeat loci in human populations. *Proclamations of the National Academy of Sciences U.S.A.* **91**, 3166–3170.

Dupanloup I., Schneider S., & Excoffier L. (2002) A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology* **11**(12), 2571-81.

Eding H., Crooijmans P.M.A., Groenne M.A.M., & Meuwissen T.H.E. (2002) Assessing the contribution of breeds to genetic diversity in conservation schemes. *Genetic Selection and Evolution* **34**, 613–633.

Edwards C.J., Dolf G., Looft C., Loftus R.T., & Bradley D.G. (2000) Relationships between the endangered Pustertaler-Sprinzen and three related European cattle breeds as analysed with 20 microsatellite loci. *Animal Genetics* **31**, 329-332.

EFABIS – European Farm Animal Biodiversity Information System http://efabis.tzv.fal.de/

Elston R.C. (1971). Estimation of admixture in racial hybrids. *Annals of Human Genetics* **35**, 9-17.

Endler J.A. (1986) Natural selection in the wild, Princeton University Press, Princeton.

Epstein H., & Mason I.L. (1984) Cattle. In: *Evolution of Domesticated Animals*. (1st ed.). (Mason I.L., ed.) London: Longman. pp. 6-27.

Excoffier L., Estoup A., & Cornuet J-M. (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**, 1727-1738.

Fabuel E., Barragan C., Silio L., Rodriguez M.C., & Toro M.A. (2004) Analysis of genetic diversity and conservation priorities in Iberian pigs based on microsatellite markers. *Heredity* **93**:104–113.

FAO (1981) Animal production and health. *Paper* **24**, 388, FAO, Rome , Italy.

Felius M. (1995) *Cattle breeds: an encyclopaedia*. Misset: Doetinchem, Netherlands.

Finlay E.K., Gaillard C., Vahidi S.M.F., Mirhoseini S.Z., Jianlin H., Qi X.B., El-Barody M.A.A., Baird J.F., Healy B.C., & Bradley D.G. (2007) Bayesian inference of population expansions in domestic bovines. *Biological Letters* **3**(4), 449-452.

Fitzsimmons N.N., Buchan J.C., Lam P.V., Polet G., Hung T.T., Thang N.Q., & Gratten J. (2002) Identification of purebred Crocodylus siamensis for reintroduction in Vietnam. *Journal of Experimental Zoology* **294**, 373-381.

Frankel O. H., & Soulé M. E. (1981) Conservation and Evolution. Cambridge University Press.

Frankham R. (1995) Conservation genetics. *Annual Review of Genetics* **29**, 305-327.

Freeman A.R., Bradley D.G., Nagda S., Gibson J.P., & Hanotte O. (2006) Combination of multiple microsatellite data sets to investigate diversity and admixture of domestic cattle. *Animal Genetics* **37**(1), 1-9.

Freeman A.R., Meghen C.M., MacHugh D.E., Loftus R.T., Achukwi M.D., Bado A., Sauveroche B., & Bradley D.G. (2004) Admixture and diversity in West African cattle populations. *Molecular Ecology* **13**, 3477-3487.

Gandini G.C., & Villa E. (2003) Analysis of the cultural value of local livestock breeds: a methodology. *Journal of Animal Breed Genetics* **120**, 1-11.

Garcia-Moreno J., Matocq M.D., Roy M.S., Geffen E., & Wayne R. (1996) Relationships and genetic purity of the endangered Mexican wolf based on analysis of microsatellite loci. *Conservation Biology* **10**(2), 376-389.

Georges M,D., Nielsen M., Mackinnon A., Mishra R., Okimoto A.T., Pasquino L.S., Sargeant A., Sorensen M.R., Steele X., Zhao J.E., Womack R. & Hoeschele I. (1995) Mapping Quantitative Trait Loci Controlling Milk Production in Dairy Cattle by Exploiting Progeny Testing. *Genetics* **139**(2), 907–920.

Giovambattista G., Ripoli M.V., Peral-Garcia P., & Bouzat J.L. (2001) Indigenous domestic breeds as reservoirs of genetic diversity: the argentinian creole cattle. *Animal Genetics* **32**(5), 240-248.

Glass B., & Li C.C. (1953). The dynamics of racial intermixture - an analysis based upon the American Negro. *American Journal of Human Genetics* **5**, 1-20.

Goodman S.J., Barton N.H., Swanson G., Abernethy K., & Pemberton J.M. (1999) Introgression through rare hybridisation: a genetic study of a hybrid zone

between red and sika deer (genus cervus) in Argyll, Scotland. *Genetics* **152**, 355-371.

Gottelli D., Sillero-Zubiri C., Applebaum G.D., Roy M.S., Girman D.J., Garcia-Moreno J., Ostrander E.A., & Wayne R.K. (1994) Molecular genetics of the most endangered canid: the Ethipian wolf Canis simensis. *Molecular Ecology* **3**(4), 301-312.

Griffiths R.C., & Tavaré S. (1994) Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131-159.

Gryzbowski G. & Prusak B. (2004) Genetic variation in nine European cattle breeds as determined on the basis of microsatellite markers. II. Gene migration and genetic distance. *Animal Science Papers and Reports* **22**(1), 37-44.

Hall S.J.G., & Clutton-Brock J. (1989) *Two Hundred Years of British Farm Livestock*. Natural History Museum, London.

Hall S. J. G, and Ruane J. (1993) Livestock breeds and their conservation: a global overview. *Conservation Biology* **7**, 815-825.

Hanotte O., Bradley D.G., Ochieng J.W., Verjee Y., & Hill E.W. (2002) African pastoralism: genetic imprints of origins and migrations. *Science* **296**, 336-339.

Hanotte O., Bradley D.G., Okomo M., Verjee Y., Ocieng J., & Rege J.E.O. (2000) Geographic distribution and frequency of a Bos Taurus and an indicine Bos indicus Y specific allele amongst sub-Saharan African cattle breeds. *Molecular Ecology* **9**, 387-396.

Hansen M.M. (2002) Estimating the long term effects of stocking domesticated trout into wild brown trout (Salmo trutta) populations: an approach using microsatellite DNA analysis of historical and contemporary samples. *Molecular Ecology* **11**(6), 1003.

Hanslik S., Harr B., Brem G., & Schlotterer C. (2000) Microsatellite anlaysis reveals substantial genetic differentiation between contemporary New World and Old World Holstein Friesian populations. *Animal Genetics* **31**(1), 31-44.

Hansson B., Bensch S., Hasselquist D., Lillandt B.G., Wennerberg L., & Von Shantz T. (2000) Increase of genetic variation over time in a recently founded population of great reed warblers (Acrocephalus arundinaceus) revealed by microsatellites and DNA fingerprinting. *Molecular Ecology* **9**, 1529-1538.

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

Harper P.A.W., Latter M.R., Nicholas F.W., Cook R.W., & Gill P.A. (1998) Chrondrodysplasia in Australian Dexter cattle. *Australian Veterinary Journal* **76**(3), 199-202.

Hebert P.D.N., Stoeckle M.Y., Zemlak T.S., & Francis C.M. (2004) Identification of birds through DNA barcodes. *Public Library of Science Biology* **2**(10), 1657-1663.

Hedrick P.W. (1995) Gene Flow and Genetic Restoration: The Florida Panther as a Case Study. *Conservation Biology* **9**(5), 996-1007.

Hobbs R.J. (1992) Disturbance, diversity, and invasion: implications for conservation. *Conservation Biology* **6**(3), 334-337.

Hooper W. (1898) Kerry and Dexter cattle; origin and history. *Journal of the Royal Agricultural Society of England* **9**, 667-677.

Houle D. (1992) Comparing evolvability and variability of quantitative traits. *Genetics* **130**, 195-204.

Hudson R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**(2), 337-338.

Irwin D.E., Irwin J.H., & Price T.D. (2001) Ring species as bridges between microevolution and speciation. *Genetica* **112-113**, 223-243.

Jeffreys A., Neumann R., & Wilson V. (1990) Repeat Unit Sequence Variation in Minisatellites: A Novel Source of DNA Polymorphism for Studying Variation and Mutation by Single Molecule Analysis. *Cell* **60**, 473-485.

Kantanen J., Olsaker I., Holm L-E., Lien S., Vilkki J., brusgaard K., Eythorsdottir., Danell B., & Adalsteinsson S. (2000) Genetic diversity and population structure of 20 northern European cattle breeds. *Journal of Heredity* **91**(6), 446-457.

Kim K.S., Yeo J.S., & Choi C.B. (2003) Genetic diversity of north-east asian cattle based on microsatellite data. *Animal genetics* **33**, 201-204.

Kimura M., & Crow J. F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725-738.

Kimura M., & Ohta T. (1971) *Theoretical aspects of populations genetics*. Princeton University Press.

Kimura M., & Ohta T. (1978) Stepwise mutation model and distribution of allele frequencies in a finite population. *Proclamations of the National Academy of Science U. S. A.* **75**, 2868-2872.

King T.L., & Burke T. (1999) Special issue on gene conservation: identification and management of genetic diversity. *Genetic Ecology* **8**, S1-S3.

Klungland H., Olsen H.G., Hassanane M.S, Mahrous K., & Iva D. (2000) Coat colour genes in diversity studies. Animal *Breeding Genetics* **117**; 217-224.

Kroeber A.L. & Richardson J. (1940) Three centuries of women's dress fashions: A quantitative analysis. *Anthropological Records* **5**(2), i-iv, 111-153.

Kumar P., Freeman A.R., Loftus R.T., Gaillard C., Fuller D.Q., & Bradley D.G. (2003) Admixture analysis of south asian cattle. *Heredity* **91**, 43-50.

Land E.D. & Lacy R.C. (2000) Introgression level achieved through Florida Panther genetic restoration. *Endangered Species Update* **17**(5), 99-103.

Laval G., Lannuccelli N., Legault C., Milan D., Groenen M.A.M., Giuffra E., Andersson L., Nissen P.H., Jorgensen C.B., Beeckmann P., Geldermann H., Foulley J., Chevalet C., & Ollivier L. (2000) Genetic Diversity of eleven European pig breeds. *Genetics of Selection and Evolution* **32,** 187-203.

Lenstra H. (2006) Marker-assisted conservation of European cattle breeds: an evaluation. *Animal Genetics* **37**, 475-481.

Li M-H., Tapio I., Vilkki J., Ivanova Z., Kiselyova T., Marzanov N., Cinkulov M., Stojanovic S., Ammosov I., Popov R., & Kantanen J. (2007) The genetic structure of cattle populations (*Bos taurus*) in northern Eurasia and the neighbouring Near Eastern regions: implications for breeding strategies and conservation. *Molecular Ecology* **16**, 3839-3853.

Lindholm A., & Breden F. (2002) Sex Chromosomes and Sexual Selection in Poeciliid Fishes. *American Naturalist* **160**, 214-222.

Liron J.P., Ripoli M.V., Garcia P.P., & Giovambattista G. (2004) Assignment of paternity in a judicial dispute between two neighbour Holstein dairy farmers. *Journal of Forensic Science* **49**, 96-98.

Loeschcke V., Tomiuk J., & Jain K. Eds. (1994) *Conservation genetics.* Birkhauser.

Loftus R.T., MacHugh D.E., Bradley D.G., & Sharp P.M. (1994) Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences of the USA* **91**, 2757-2761.

MacHugh D.E. (1996) *Molecular biogeography and genetic structure of domestic cattle* [PhD Thesis], University of Dublin.

MacHugh D.E., Loftus R.T., Cunningham P., & Bradley D.G. (1998) Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. *Animal Genetics* **29**, 333-40.

MacNeil M.D., Cronin M.A., Blackburn H.D., Richards C.M., Lockwood D.R., & Alexander L.J. (2007) Genetic relationships between feral cattle from Chirikof Island Alaska and other breeds. *Animal Genetics* **38**, 193-197.

Madansky A. (1959) The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association* **54**, 173–205.

Madsen T., Shine R., Olsson M., & Wittzell H. (1999) Restoration of an inbred adder population. *Nature* **402**(6757), 34-35.

Manwell C., & Baker C.M.A. (1980) Chemical classification of cattle. 2. Phylogenetic tree and specific status of the Zebu. *Animal Blood Groups and Biochemical Genetics* **11**, 151-162.

Marjoram P., Molitor J., Plagnol V., & Tavare S. (2003) Markov chain Monte Carlo without likelihoods. *Proclamations of the National Academy of Sciences U.S.A.* **100**(25), 15324-15328.

Mateus J.C., Penedo M.C.T., Alves V.C., Ramos M., & Rangel-Figueiredo T. (2004) Genetic diversity and differentiation in Portuguese cattle breeds using microsatellites. *Animal Genetics* **35**(2), 106–113.

Mayr E. (1963) *Animal species and evolution.* Belknap Press, Harward.

McKeigue P.M., Carpenter J.R., Parra E.J., & Shriver M. D. (2000) Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Annals of Human Genetics* **64**, 171-186.

Merila J., & Crnokrak P. (2001) Comparison of genetic differentiation at marker loci and quantitative traits. *Journal of Evolutionary Biology* **14**, 892-903.

Michaux J.R., Hardy O.J., Justy F., Fournier P., Kranz A., Cabria M., Davison A., Rosoux R., & Libois R. (2005) Conservation genetics and population history of the threatened European mink *Mustela lutreola*, with an emphasis on the west European population. *Molecular Ecology* **14**(8), 2373-2388.

Miglior F., & Burnside E. B. (1995) Inbreeding of Canadian Holstein cattle. *Journal of Dairy Science* **78**, 1163-1167.

Miller C.R., Adams J.R., & Waits L.P. (2003) Pedigree-based assignment tests for reversing coyote (Canis latrans) introgression into the wild red wolf (Canis rufus) population. *Molecular Ecology* **12**, 3287-3301.

Moazami-Goudarzi K., Laloe D., Furet J.P., & Grosclaude F. (1997) Analysis of genetic relationships between 10 cattle breeds with 17 microsatellites. *Animal Genetics* **28**, 338-45.

Moioli B., Napolitano F., & Catillo G. (2004) Genetic Diversity between Piedmontese, Maremmana, and Podolica Cattle Breeds. *Journal of Heredity* **95**(3), 250–256.

Mommens G., Van Zeveren A., & Peelman L.J. (1998) Effectiveness of bovine microsattelites in reslving paternity cases in American bison, Bison bison L. *Animal Genetics* **29**, 12-18.

Moritz C. (1999) Conservation units and translocations: strategies for conserving evolutionary processes. *Hereditas* **13**, 217-228.

Moyles M.G. (1959) 'Dexter cattle: an historical sketch'. *Journal of the Department of Agriculture* (Ireland) **56**, 109-113.

Neel J.V. (1973) "Private" genetic variants and the frequency of mutation among South American Indians. *Proclamations of the National Academy of Sciences of the U.S.A.* **70**(12), 3311-3315.

Neff B.D., Fu P., & Gross M.R. (1999) Microsatellite evolution in sunfish. *Canadian Journal of Fish and Aquatic Science* **56**, 1198-1205.

Negrini R., Nijman I.J., Milanesi E., Moazami-Goardzi K., Williams J.L., Erhardt G., Dunner S., Rodellar C., Valentini A., Bradley D.G., Olsaker I., Kantanen J., Ajmone_Masan P., & Lenstra J.A. (2007) Differentiation of European cattle by AFLP fingerprinting. *Animal Genetics* **38**(1), 60-66.

Nei M., Maruyama T., & Chakraborty R. (1975) The bottleneck effect and genetic variability of populations. *Evolution* **29**, 1–10.

Nei M. (1987) *Molecular evolutionary genetics*. New York: Columbia University Press.

Paetkau D., Shields G.F., & Strobeck C. (1998) Geneflow between insular, coastal and interior populations of brown bears in Alaska. *Molecular Ecology* **7**, 1283-1292.

Patterson D.F., Haskins M.E., & Jezyk P.F. (1982) Models of human genetic disease in domestic animals. *Advances in Human Genetics* **12**, 263-339.

Payne W.A. (1970) *Cattle Production in the Tropics*. (1st ed.) London: Longman.

Peelman L.J., Mortiaux F., Van Zeveren A., Dansercoer A., Mommens G., Coopman F., Bouquet Y., Burny A., Renaville R., & Portetelle D. (1998) Evaluation of the genetic variability of 23 bovine microsatellite markers in four Belgian cattle breeds. *Animal Genetics* **29**(3), 161-167.

Perkins D. Jr. (1969) Fauna of Çatal Hüyük: Evidence for early cattle domestication in Anatolia. *Science* **164**(3875), 177–178.

Pertoldi C., Bijlsma R., & Loeschcke V. (2007) Conservation genetics in a globally changing environment: present problems, paradoxes and future challenges. *Biodiversity Conservation* **16**, 4147–4163.

Pialek J., & Barton N.H. (1997) The spread of an advantageous allele across a barrier: the effects of random drift and selection against heterozygotes. *Genetics* **145**, 493-504.

Pimentel de Mello L., Vasconcellos K., Tambasco-Talhari D., Pereira A.P., Coutinho L.L., & Correia de Almeida Regitano L. (2003) Genetic characterization of Aberdeen Angus cattle using molecular markers. *Genetics and Molecular Biology* **26**(2), 133-137.

Piry S., Alapetite A., Cornuet J–M., Paetkau D., Baudouin L., & Estoup A. (2004) GENECLASS2: A Software for Genetic Assignment and First-Generation Migrant Detection. *Journal of Heredity* **95**(6), 536–539.

Pritchard J.K., & Przeworski M. (2001) Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69**(1), 1-14.

Pritchard J.K., Stephens M., & Donnelly P.J. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Randi E., & Lucchini V. (2002) Detecting rare introgression of domestic dog genes into wild wolf (Canis lupus) populations by Bayesian admixture analyses of microsatellite variation. *Conservation Genetics* **3**, 31-45.

Rannala B., & Mountain J.L. (1997), Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the USA* **94**, 9197-9102.

Rhymer J.M, & Simberloff D. (1996) Extinction by hybridisation and introgression. *Annual Review of Ecological Systematics* **27**, 83-109.

Rege J.E.O., & Gibson J.P. (2003) Animal genetic resources and economic development: issues in relation to economic valuation. *Ecological Economics* **45**(3), 319-330.

Reist-Marti S.B., Abdulai A., & Simianer H. (2005) Conservation programmes for cattle: design, cost and benefits. *Journal of Animal Breeding Genetics* **122**, 95-109.

Roberts D.F., & Hiorns R.W. (1965) Methods of analysis of the genetic composition of a hybrid population. *Human Biology* **37**, 38–43.

Robertson A., & Hill W. (1984) Deviations from Hardy Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics* **107**, 703-718.

Ruane J. (1999) A critical review of genetic distance studies in conservation of animalgenetic resources. *Journal of Animal Breed Genetics* **116**, 317-323.

Scherf B.D. (2000) *World Watch List for Domestic Animal Diversity*, 3rd edn. Food and Agriculture Organization of the United Nations, Rome.

Schmid M., Saitbekova N., Gaillard C., & Dolf G. (1999) Genetic diversity in Swiss cattle breeds. *Journal of Animal Breeding Genetics* **116**, 1-8.

Sheppy A.J. (1998) Bloodlines, breed structure, and the influence of artificial insemination in Dexter cattle. *Proceedings of the first world congress on Dexter cattle*. Dexter cattle society, UK.

Shriver M.D., Parra E.J., Dios S., Bonilla C., Norton H., Jovel C., Pfaff C., Jones C., Massac A., Cameron N., Baron A., Jackson T., Argyropoulos G., Jin L., Hoggart C.J., McKeigue P.M., & Kittles R.A. (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. *Human Genetics* **112**, 387–399.

Simianer H. (2005) Using expected allele number as objective function to design between and within breed conservation of farm animal biodiversity. *Journal of Animal breeding and Genetics* **122**(3): 177-187.

Simianer H., Marti S.B., Gibson J., Hanotte O., & Rege J.E.O. (2003) An approach to the optimal allocation of conservation funds to minimize loss of genetic diversity between livestock breeds. *Ecological Economics* **45**(3), 377-392.

Slatkin M. (1987) Gene Flow and the Geographic Structure of Natural Populations. *Science* **236**, 787-792.

Slatkin M. (1994) Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331-336.

Stephens J.C., Briscoe D., & O'Brien S.J. (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Journal of Human Genetics* **55**(4), 809–824.

Susnik S., Berrebi P., Dovc P., Hansen M.M., & Snoj A. (2004) Genetic introgression between wild and stocked salmonids and the prospects for using molecular markers in population rehabilitation: the case of the Adriatic grayling (Thymallus thymallus L. 1785). *Heredity* **93**, 273–282.

Taberlet P., & Luikart G. (1999) Non-invasive genetic sampling and individual identification. *Biological journal of the Linnean Society* **68**(12), 41-55.

Tapio I., Varv S., Bennewitz J., Maleviciute J., Fimland E., Grislis Z., Meuwissen T.H.E., Miceikiene I., Olsaker I., Viinalass H., Vilkki J., & Kantanen J. (2006) Prioritisation of northern European cattle breeds based on analysis of microsatellite data. *Conservation Biology* **20**(6), 1768-1779.

Thaon d'Arnoldi C., Foulley J.L., & Ollivier L. (1998) An overview of the Weitzman approach to diversity. *Genetics, Selection, and Evolution* **30**, 149-161.

Thompson (1973) The Icelandic admixture problem. *Annals of Human Genetics* **37**, 69.

Tisdell C. (2001) Socioeconomic causes of loss of animal genetic diversity: analysis and assessment. *Nota Di Lavoro*, Q200.

Turelli M., Barton N.H., & Coyne J. A. (2001) Theory and speciation, *Trends in Ecology and Evolution* **16**(7), 330-343.

Ujvari B., Madsen T., Ketnko T., Olsson M., Shine R., & Witzell H. (2002) Low genetic diversity threatens imminent extinction for the Hungarian meadow viper (*Vipera ursinii rakosiensis*). *Biological Conservation* **105**(1), 127-130.

Valdiosera C.E., Garcia N., Anderung C., Dalen L., Cregut-Bonnoure E., Kahlke R-D., Stiller M., Brandstrom M., Thomas M.G., Arsuaga J.L., Gotherstrom A., & Barnes I. (2007) Staying out in the cold: glacial refugia and mitochondrial DNA phylogeography in ancient European brown bears. *Molecular Ecology* **16**(24), 5140–5148.

van Hooft W.F., Hanotte O., Wenink P.W., Groen A.F., Sugimoto Y., Prins H.H.T., & Teale A. (1999) Applicability of bovine microsatellite markers for population genetic studies on African buffalo (Syncerus caffer). *Animal Genetics* **30**, 214-220.

Vasconcellos L.P.M.K., Tambasco-Talhari D., Pereira A.P., Coutinho L.L., & Regitano L.C.A. (2003) Genetic characterisation of Aberdeen Angus cattle using molecular markers. *Genetics and Molecular Biology* **26**(2), 133-137.

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

Vollema A.R., & Groen A.F. (1997) Genetic correlations between longevity and conformation traits in an upgrading dairy cattle population. *Journal of Dairy Science* **80**, 3006-3014.

Wan Q.H., Wu H., Fujihara T., & Fang S-G. (2004) Which genetic marker for which conservation genetics issue. *Electrophoresis* **25**(14), 2165-2171.

Wang H.J., Luo M., Tereschenko I.V. , Frikker D.M., Cui X., Li J.Y., Hu G., Chu Y., Azaro M.A., Lin Y., Shen L., Yang Q., Kambouris M.E., Gao R., Shih W., and Li H. A genotyping system capable of simultaneously analyzing >1000 single nucleotide polymorphisms in a haploid genome. *Genome Research* **15**, 276-283.

Wang J. (2001) A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetic Research* **78**, 243-257.

Wang J. (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**, 747-765.

Ward T.J., Skow L.C., Gallagher D.S., Schnabel R.D., Nall C.A., Kolenda D.E., Davis S.K., Taylor J.F., & Derr J.N. (2001) Differential introgression of uniparentally inherited markers in bison populations with hybrid ancestries. *Animal Genetics* **32**, 89-91.

Weir B.S., & Cockerham C.C. (1984) Estimating F-Statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Weitzman M.L. (1992) On diversity. *Quarterly Journal of Economics* **CVII**, 157-183.

Weitzman M.L. (1993) What to preserve? An application of diversity theory to crane conservation. *Quarterly Journal of Economics* **107**(2), 363-405.

Weinberg W. (1908) On the demonstration of heredity in man. In: Boyer SH, trans (1963) *Papers on human genetics*. Prentice Hall, Englewood Cliffs, NJ.

Wiener P., Burton D., & Williams J.L. (2004) Breed relationships and definition in British cattle: a genetic analysis. *Heredity* **93**, 597-602.

Wilde W. (1858) in *Proceedings of the royal Irish academy* vol. vii, Dublin, Royal Irish Academy.

Willard H.F. (1989) The genomics of long tandem arrays of satellite DNA in the human genome. *Genome* **31**, 737-744.

Wright S. (1951) The genetical structure of populations. *Annals of Eugenics* **15**, 323-354.

Wright S. (1943) Isolation by distance. *Genetics* **28**, 114.

Yan G., Romero-Severson J., Walton M., Chadee D.D., & Severson D.W. (1999) Population genetics of the yellow fever mosquito in Trinidad: comparisons of amplified fragment length polymorphism (AFLP) and restriction fragment length polymorphism (RFLP) markers. *Molecular Ecology* **8**(6), 951–963.

Zeuner F.E. (1963) The history of the domestication of cattle, In: Mourant A. E, Zeuner F. E, (eds) *Man and cattle*. Royal Anthropological Institute of Great Britain and Ireland, London, Volume 18, 9-20.

Chapter 2.
Materials and methods

I An introduction into the methods used

is section provides a complete description of the processes involved in the actical and analytical work that is applied in the subsequent chapters. This includes e concepts included in software descriptions as well as sampling, practical data anipulation, and additional software used in formatting. Initial methods describe the oratory processes of DNA extraction and creation of genotypic information for ch animal, following this are the analytical methods applied for population genetics well as the admixture analysis including the newly developed Approximate yesian method.

2. Sampling

mpling from the Dexter breed was conducted through a selection process cording to pedigree analysis carried out in collaboration with Andrew Sheppy obthorn Trust). In order to satisfy the requirements for the admixture analysis imals were sampled from several major demographic groups within the herd book d included a selection of animals with a closer relationship to (i.e. as far as ssible, directly descended from) putative ancestral lines within the breed. aditional' dexters were also selected where the herd-book indicated they were free documented introgression and were therefore chosen as an accurate presentation of the historical breed. Appendix 2.1 provides an anonymised record the samples analysed including their farm name, approximate location (by county) d the herdbook category to which they were assigned. The precise details erdbook numbers, owners etc) is proprietory information belonging to the Dexter ttle Society, which can be applied for on request.

mpling also targeted longstanding herds known to have had considerable uence on the formation of the contemporary Dexter. The Ypsitty herd group was med from what is thought to be the oldest Dexter herd population and Ypsitty scendents are commonly found in contemporary pedigrees. The closed oodmagic herd group was chosen due to unique selective breeding for the removal an adverse genetic condition (achondroplasia) and its demographic isolation. The oodmagic herd has subsequently been extensively used as breeding stock in the odern Dexter and represents an important component of the breed. The ntemporary Dexter population as it predominantly exists today was sampled ough a diverse selection of individuals from throughout the herdbook chosen to

represent the breed average. A further sample of animals of American descent were included as a comparison of a semi-isolated breeding population originating from common Dexter stock.

Dexter sample groups comprised between 12 and 91 individuals (Table 2.2). DNA was obtained from combination of plucked hairs and cryogenically stored semen samples. The British individuals sampled originated from farms over a wide geographical area (Figure 2.1). The American samples were chosen to represent a selection of older breeding lines and also included some animals with high levels of early Woodmagic herd ancestry.

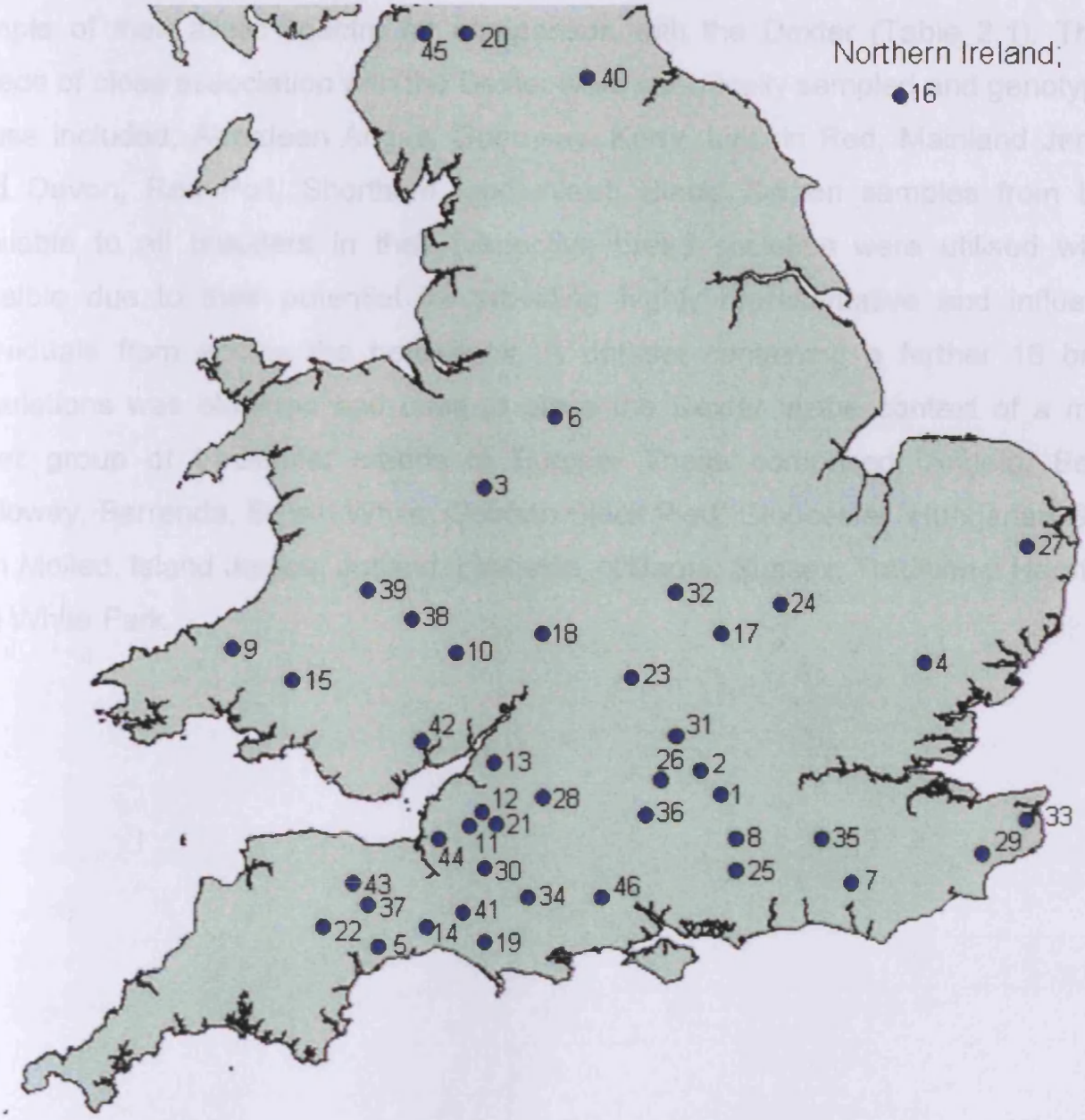


Figure 2.1. Distribution of sampled animals according to natal farm. Farm names; (1)Apple, (2)Atlantic, (3)Beeches, (4)Bolyns, (5)Bookhams, (6)Bradash, (7)Butterbox, (8)Byfield, (9)Canwell, (10)Chalicewood, (11)Clivedon, (12)Cobthorn, (13)Coppinswell, (14)Cullaford Vale, (15)Dolau, (16)Donard grange, (17)Elmwood, (18)Frith, (19)Godshill, (20)Harron, (21)Honely hall, (22)Ilsington, (23)Kidmore, (24)Knotting, (25)Lowercombe, (26)Melbry, (27)Minden, (28)Mindoro, (29)Moomin, (30) Parndon, (31)Saltaire, (32)Sarum, (33)Statenboro, (34)Sunnyside, (35)Swanthorpe, (36)Sylvestor, (37)Templeton, (38)The Gaer, (39)Vatch, (40)Vycanny, (41)Wantsley, (42)Whitegates, (43)Woodmagic, (44)Woodmanor, (45Ypsitty, & (46)Ytene.

Other cattle breed populations were sampled broadly to construct a representative sample of their allelic spectra for comparison with the Dexter (Table 2.1). Those breeds of close association with the Dexter were specifically sampled and genotyped. These included; Aberdeen Angus, Guernsey, Kerry, Lincoln Red, Mainland Jersey, Red Devon, Red Poll, Shorthorn, and Welsh Black. Semen samples from bulls available to all breeders in their respective breed societies were utilised where possible due to their potential for providing highly representative and influential individuals from across the herd book. A dataset containing a further 16 breed populations was obtained and used to place the Dexter in the context of a much wider group of traditional breeds in Europe. These comprised; Angeln, Belted Galloway, Berrenda, British White, German Black Pied, Gloucester, Hungarian Grey, Irish Moiled, Island Jersey, Jutland, Limousin, N'Dama, Sussex, Traditional Hereford, and White Park.

Table 2.1. Sample group sizes, Numbers of typed loci, and origins of DNA

Sample population	Sample group size	Numbers of typed loci	origin of DNA
American Dexter	13	22	All plucked hair
Traditional Dexter	12	22	2 plucked hair/ 10 semen
Woodmagic Dexter	19	22	15 plucked hair/ 4 semen
Ypsitty Dexter	13	22	12 plucked hair/ 1 semen
Dexter breed population	91	22	71 plucked hair/ 20 semen
Beef Devon	20	22	8 plucked hair/ 12 semen
Milking Devon	32	22	All plucked hair
Kerry	32	12/22	10 plucked hair/ 10 semen
Aberdeen Angus	20	22	6 plucked hair/ 14 semen
Beef Shorthorn	11	22	All semen
Lincoln Red	60	23	All plucked hair
Red Poll	20	22	13 plucked hair/ 7 semen
Welsh Black	17	22	All semen
UK Mainland Jersey	21	22	All plucked hair
Guernsey	12	22	6 plucked hair/ 6 semen
Shetland	31	12	All plucked hair
Belted Galloway	15	12	All plucked hair
Irish Moiled	20	12	All plucked hair
British White	11	12	All plucked hair
Traditional Hereford	19	12	All plucked hair
Sussex	50	12	All plucked hair
White Park	33	12	All plucked hair
Gloucester	14	12	All plucked hair
Island Jersey	25	12	All plucked hair
Jutland	17	12	All plucked hair
Angeln	24	12	All plucked hair
German Black Pied	19	12	All plucked hair
Hungarian Grey	16	12	All plucked hair
Limousin	23	12	All plucked hair
Berrenda	31	12	All plucked hair
N'dama	9	12	All plucked hair

Due to the dynamic and heterogeneous nature of cattle populations it is important to use those populations which are most similar to the historical type involved in the admixture events. For this reason two populations of Devon cattle were sampled. These were both the beef producing Red Devon, and the Milking Devon. This latter population is now extinct in Britain and therefore a surviving American population was sampled.

2.3. DNA extraction

Both plucked hair and semen samples were extracted using one of the following three methods;

2.3.1. Chelex100 method

The chelex 100 protocol described in Walsh et al (1991) and adapted by Goossens et al (1998) was applied; the numbers of hairs used depended on hair size and availability, this typically ranged from ten to twenty. One centimetre of hair was cut from the root end and placed in a 1.5ml eppendorf tube with 200µl of a five percent chelex (Bio-Rad, Catalogue:142-1253) suspension. Tubes were then incubated for six hours at 56 °C with agitation. Tubes were then placed in boiling water for 8 minutes to deactivate the chelex. Extracts were allowed to cool and stored at -20 °C.

2.3.2. Buffer-based extraction method

This method is a simplified cell lysis approach as detailed by Vigilant (1999). One µl of proteinase k is added to 20µl Qiagen PCR buffer and 79µl de-ionised water and incubated overnight with agitation. Extracts were allowed to cool and stored at -20 °C.

2.3.3. Kit method

Semen samples were extracted using the DNeasy Tissue extraction kit method using a modification of the inclusion of 1µl Sodium Diothreitol (DTT) in the incubation stage of each extraction. From the extractions 1.5µl was added as template in each PCR reaction. Extracts were allowed to cool and stored at -20 °C.

The Qiagen DNeasy Micro extraction kit method was also used in samples for which the previous methods proved unsuccessful, or where there was a very limited amount of material available. The extraction was performed as in the Qiagen handbook.

2.4. Genotyping

Twenty two microsatellite markers used in this study were taken from the United Nations Food and Agriculture Organisation (FAO) list of cattle markers that have been used in a number of recent studies (Beja-Pereira et al. 2003; Kumar et al. 2003; Wiener et al. 2004). The microsatellites were; hel1, hel5, hel9, hel13 (Kaukinen and Varvio, 1993), ilsts005 (Brezinsky et al. 1993a), csrm60 (Moore et al. 1994), eth3, eth10 (Solinas-Toldo et al. 1993), tgla 227, tgla122, tgla126 (Georges and Massey, 1992), sps115 (Moore and Byrne, 1993) , inra032, inra037, inra063 (Vaiman et al. 1994), eth152, eth225 (Steffen et al. 1993), bm1818, bm1824 (Bishop et al. 1994), ilsts006 (Brezinsky et al. 1993b), haut27 (Thieven et al. 1997), and cssm66 (Barendse et al. 1994). The markers are from a panel of microsatellites selected by the FAO for their common application in cattle studies, polymorphism, and suitability for automated sequencing, multiplexing and readability. The 22 markers used here are distributed across 15 chromosomes with up to three on any one chromosome (Table 2.2). Known positions for those markers sharing a chromosome suggest that there is a low likelihood of linkage between them, inter-locus distances being no less than 15 kilobases (Table 2.3).

Table 2.2. Multiplex and primer details for the Microsatellite markers applied.

Multiplex	Primer	Chromosome	Primer seq. For / Rev	Allelelic range	Dye Label
1	HEL9	8	CCCATTCAGTCTTCAGAGGT/ CACATCCATGTTCTCACCAC	141-173	FAM
	ILSTS005	10	GGAAGCAATGAAATCTATAGCC/ TGTTCTGTGAGTTTGTAAGC	176-194	HEX
	CSRM60	10	AAGATGTGATCCAAGAGAGAGGCA/ AGGACCAGATCGTGAAAGGCATAG	79-115	HEX
	ETH3	19	GAACCTGCCTCTCCTGCATTGG/ ACTCTGCCTGTGGCCAAGTAGG	103-133	FAM
	ETH10	5	GTTCAGGACTGGCCCTGCTAACA/ CCTCCAGCCCACTTTCTCTTCTC	207-231	HEX
2	TGLA227	18	CGAATTCCAAATCTGTTAATTTGCT/ ACAGACAGAAACTCAATGAAAGCA	75-105	HEX
	TGLA126	20	CTAATTTAGAATGAGAGAGGCTTCT/ TTGGTCTCTATTCTCTGAATATTCC	115-131	HEX
	SPS115	15	AAAGTGACACAACAGCTTCTCCAG/ AACGAGTGTCCTAGTTTGGCTGTG	234-258	HEX
	ETH225	9	GATCACCTTGCCACTATTTCCT/ ACATGACAGCCAGCTGCTACT	131-159	FAM
	INRA063	18	ATTTGCACAAGCTAAATCTAACC/ AAACCACAGAAATGCTTGGAAG	167-189	HEX
	ETH152	5	TACTCGTAGGGCAGGCTGCCTG/ GAGACCTCAGGGTTGGTGATCAG	181-211	FAM
3	HEL1	15	CAACAGCTATTTAACAAGGA/ AGGCTACAGTCCATGGGATT	99-119	FAM
	INRA037	11	GATCCTGCTTATATTTAACCAC/ AAAATTCCATGGAGAGAGAAAC	112-148	HEX
	HEL5	21	GCAGGATCACTTGTTAGGGA/ AGACGTTAGTGTACATTAAC	145-171	NED
	INRA032	11	AAACTGTATTCTCTAATAGCTAC/ GCAAGACATATCTCCATTCCTTT	160-204	HEX
	TGLA122	21	CCCTCCTCCAGGTAAATCAGC/ AATCACATGGCAAATAAGTACATAC	136-184	FAM
	BM1824	1	GAGCAAGGTGTTTTTCCAATC/ CATTCTCCAACTGCTTCCTTG	176-197	NED
4	ILSTS006	7	TGTCTGTATTTCTGCTGTGG/ ACACGGAAGCGATCTAAACG	277-309	HEX
	HAUT27	26	TTTTATGTTCATTTTTTGACTGG/ AACTGCTGAAATCTCCATCTTA	120-158	NED
	CSSM66	14	ACACAAATCCTTTCTGCCAGCTGA/ AATTTAATGCACTGAGGAGCTTGG	171-209	FAM
	HEL13	11	TAAGGACTTGAGATAAGGAG/ CCATCTACCTCCATCTTAAC	178-200	HEX
	BM1818	23	AGCTGGGAATATAACCAAAGG/ AGTGCTTTCAAGGTCCATGC	248-278	FAM

Table 2.3. Inter-locus distance information about all homo-chromatic loci used. All values given in Kilobases (ArkDB, 2008)

Marker Name	Chromosome	Mapping Position	Chromosome range
ETH10	5	55.000	0-142.000
ETH152	5	127.000	0-142.000
CSRM60	10	77.816	0-100.800
ILSTS	10	97.400	0-100.800
INRA032	11	68.182	0-130.965
INRA037	11	-	0-130.965
HEL13	11	114.500	0-130.965
SPS115	15	-	0-94.000
HEL1	15	27.700	0-94.000
TGLA227	18	145.000	18.000-145.000
INRA063	18	93.000	18.000-145.000
HEL5	21	18.000	18.000-106.000
TGLA122	21	75.000	18.000-106.000

The markers were applied in multiplexes of up to six primer pairs, one primer from each pair synthesized with a fluorescent dye FAM, HEX, or NED on the 5' end. Where loci overlapped in size range of amplified fragments different dyes were used to label each marker (Table 2.3). Marker amplification was performed using the Qiagen multiplex kit according to manufacturers' instructions. Amplification of the loci was carried out in 6µl reactions (1 x QIAGEN PCR Multiplex Master Mix (3mM $MgCl_2$), 0.2µM each primer). Thermocycling conditions were as follows: initial denaturation at 95 °C for 15 minutes followed by 35 cycles of 60 seconds at 94°C, annealing for 90 seconds at 55°C then extension for 60 seconds at 72°C, with a final extension for 10 minutes at 60°C. All PCR products were separated using an ABI 3100 automated sequencer.

Gels were analysed using Genescan analysis 2.0™, Genotyper 1.1™ and Genemapper™ software. Scoring of alleles was made by hand using a number of reference individuals for comparison of peak morphology. Minimum fluorescence for allele peaks was set at 150 units, alleles failing to meet these criteria were not scored and the sample was repeated. Where high levels of non specific amplification was

evident, samples were also repeated or eventually discarded. Accurate scoring was maintained through upgrades of fragment analysis hardware and software using reference samples. Retyping of this reference material after an upgrade allowed the tracking of allelic size shifts and the maintenance of a consistent and comparable dataset.

2.5. Analytical approaches and software used

Specific applications were employed for creation and conversion of input files for use in analytical software as suggested by the overview of population genetic software by Excoffier and Heckel (2006). To create input files for STRUCTURE a basic format was created either by hand or using the Convert (Glaubitz, 2004) application. The Genepop (Raymond and Rousset, 1995) application was used to convert between Convert version and Genepop version formats which are accepted by most applications. The Populations (Langella, 1999) application was used to convert Genepop version formatted data into LEA version, ADMIX version, and Genetix version formats.

2.5.1. Genetic diversity measures

Genetic diversity measures included here were; Observed Heterozygosity (H_O), Expected Heterozygosity (H_E), and Non-biased Heterozygosity ($H_{n.b.}$), calculated using the GENETIX 4.03 software (Belkhir *et al.* 2002). Observed heterozygosity is the average of the heterozygosity values across loci for the sample. Expected heterozygosity is the heterozygosity that would be found in the population of origin calculated from the sample. The non-biased heterozygosity is a heterozygosity measure that is corrected for sample size.

Wright's F statistics were applied in accordance with Weir and Cockerham (1984) through the GENETIX 4.03 software (Belkhir *et al.* 2002). Departure from the null hypothesis, which was no genetic variation for F_{ST} and Hardy-Weinberg equilibrium for F_{IS} and F_{IT}, was tested over 10^4 permutations. The statistics f(= F_{IS}), F(= F_{IT}), and Theta(= F_{ST}) are generated, as seen in Equations 3-5.

F	=	1- C/(A+B+C)	**(Equation 3)**
F	=	1- C/(B+C)	**(Equation 4)**
Theta=		A/(A+B+C)	**(Equation 5)**

Equations 2-4. The basic formula for calculation of the statistics f, F, and theta in GENETIX, where A is the inter-population variation in allele frequencies, B is the within-population variation, and C is the between-individual variation

In order to investigate whether a relationship existed between the recorded number of breeding females and the diversity of the breed populations a regression was also performed in the MINITAB 3.2 software. The data for numbers of breeding females was also transformed using the logarithmic and square-root transformations in order to investigate whether any relationship found was better described than using a simple linear function.

2.5.2. Investigation of demographic processes and events

The forces of migration, mutation and selection govern genetic changes in populations over time (Wang, 2005). Genetic data give us the opportunity to investigate historical events through modelling these forces in order to determine pertinent demographic processes and infer particular events such as population bottlenecks. Measuring gene identity is a method used in estimating population substructure, as for example developed by Vitalis and Couvet (2001) and applied in the ESTIM 1.0 software. This method employs the parameter F to estimate averaged per locus gene identities which are used as a measure of within-population genetic drift. In a finite population, genetic variation is continually lost, with neutral genes this is dependant on population size so the smaller the population the greater the loss of alleles. This observation has a direct link to the time elapsed since two genes diverged from a common ancestor (the coalescence time) since this time increases with increasing effective population size. The perspective of many stochastic theories within population genetics is to assume constant population size but population fluctuations are likely to be more often the case (Sano *et al*, 2004). Similarly the assumption of isolated populations is often more unrealistic than attempting to include a degree of migration between populations. This has been tackled successfully by adopting a maximum likelihood approach (Tufto *et al*. 1996) but on the basis of known population size. The method used in the ESTIM 1.0 application

provides simultaneous estimates of both the effective population size and the migration rate using a method-of-moments approach (Vitalis and Couvet, 1996).

A method of detecting recent population size changes (bottlenecks or growth; Cornuet and Luikart 1996) implemented in the BOTTLENECK 1.2.02 program, was applied. The method relies on the patterns of genetic diversity expected for a demographically stable population (null hypothesis), using two summary statistics of the allelic frequency spectrum, namely the number of alleles (n_A) and the expected heterozygosity (H_e). Simulations were performed to obtain the distribution of H_e conditional on n_A and the sample size for each population and locus. In order to test for significant deviations from the null hypothesis, 10000 simulated H_e values were compared to those obtained from the real dataset, using the Wilcoxon Sign Rank Test, under three mutational models: infinite allele model (I.A.M.), stepwise mutation model (S.M.M.), and a two-phase model (T.P.M.), in which, 30% of mutations were allowed to occur under a multi-step manner.

2.6. Application of clustering algorithms

There are a number of methods available for clustering individuals or groups of individuals into genetically cohesive groups (populations, regions or breeds). It is often useful to compare the underlying models through a selection of methods in order to assess the effects of changing assumptions and model estimation of parameters (Beerli, 2006). Pritchard et al. (2000) suggest two broad methods by which clusters can be made of these individuals which do not include the subjective bias introduced through predetermined populations; distance-based, and model-based. The former uses a pairwise distance matrix whereby a graphical representation can be made of the distances between each pair of individuals. In this way clusters can be identified by eye. In model-based methods the assumption is that observations from each cluster are random draws from some parametric model. Inference for the parameters of each cluster is then made alongside the assignment of the cluster membership of each individual. Distance-based methods are usually easily applied but tend to be more suited to exploratory analysis than to fine statistical inference. Bayesian methods rely on the calculation of the probability of observing the data given specific values of the parameters. The Bayesian approach focuses on what information the data provides about the parameters (Lindley, 1986). Any information known about the model's parameters can be used to establish a 'prior

distribution'. The data are then combined with information from the prior distribution to give a probability distribution known as the 'posterior distribution' (Wade, 2000). Equation 1 represents Bayes' Theorem (Bayes, 1763) the first term of which represents the prior, the second term represents the probability of observing the data under the statistical model, and the third represents the probability of the data.

$$p(\Psi|D) = \frac{p(\Psi)p(D|\Psi)}{p(D)} \qquad \textbf{(Equation 6)}$$

Equation 6. A mathematical representation of Bayes' theorem describing the probability distribution of the parameter set (Ψ) given the data $p(\Psi|D)$, where D is the observation of the data (Chikhi et al, 2001).

The application of methodology to investigate underlying genetic structure in this study was used to validate the assumptions of extant pre-determined sample populations and to investigate, as far as possible, population associations. A variety of methodologies were applied: initial comparisons through F statistics, model-based clustering algorithm that operates without consideration of pre-existing populations, a method to analyse the spatial analysis of molecular variance, and three model-based assignment methods to estimate recent immigration rates.

2.6.1. Individual assignment

The populations used in this analysis are closely related, with varying presumed levels of recent and ancient genetic exchange. In order to investigate any cryptic population structure the whole dataset was examined using the Bayesian model-based clustering approach developed by Pritchard et al. (2000). This approach is implemented to detect the structure of a genetic sample without prior information affecting the origin of individuals, and so is independent of the sampling regime. The method, originally applied by Pritchard et al. (2000) and improved by Falush et al. (2003), is implemented in STRUCTURE 2.1 and uses the term K to describe homogeneous clusters of individuals that are as close to Hardy-Weinberg and Linkage equilibrium as can be identified within the parameter space of the dataset. At any given value of K, a Markov Chain Monte Carlo (MCMC) approach with a Gibbs sampler was used to obtain the posterior distribution of the parameters. This parameter distribution is dependent on the genotypes and the population value, K. This process can be performed with, or without the consideration of admixture. For

this study admixture was considered, giving the additional parameter of the proportion of each population in the ancestry of a given individual.

I therefore examined how the populations separate relative to one another and although the value of K is specified for each independent program run, an estimation of the likelihood of the resulting assignment is given and this can give an approximate indication of the most likely partitions within the dataset. Consideration of the likelihood value was originally suggested as a method to favour partition values within the data. Where the increase in likelihood of successive K values would asymptote, the optimum K had been reached. This approach was criticised due to its non-statistical basis by Evanno *et al.* (2005), who suggested the mode of the ΔK distribution as a more robust approach and incorporated the variation in the standard deviation of the likelihood estimates at K for calculating the most likely number of partitions present in the data. However this method becomes increasingly difficult to apply as the K value increases due to the increased variation of the likelihood estimates at any given value of K. This variation is at least in part due to the presence of multiple potential solutions of that particular number of partitions in the data, all with different estimated likelihoods. When this method is calculated it is important to distinguish between these multiple estimated likelihood 'peaks' in the data. If several different distributions of assignment are made at the same value of K then it may be that the variation in likelihood estimations are not informative of the particular value of K due to the presence of multiple scenarios and this will affect the calculation for estimation of the real K value. In this case the most likely scenario should be chosen and the runs resulting in any others discarded.

To take account of the above problem it is important to consider the subdivision of population groups where a dataset potentially contains large numbers of populations. We set the value of K from 1 to the number of populations in the analysis plus 2 in order to determine the uppermost hierarchical level of population structure. The runs were performed 20 times each with a different starting point for each value of K and ΔK was calculated as $\Delta K = m(|L(K+1) - 2L(K) + L(K-1)|)/s[L(K)]$, where m and s represent the average and standard deviation of the corresponding values across 20 runs (Evanno *et al.* 2005). Experimentation on the full dataset of over 500 individuals was investigated to check for convergence of the Markov Chain and it was found that a burnin of $5x10^4$ followed by $5x10^5$ steps was sufficient to give a stable estimated

likelihood. Runs of length $1x10^6$ were also used to confirm convergence of the Markov chain.

2.6.2. Spatial clustering

Partitioning of individuals accounting for geographic associations between each closed population was performed according to the method of Dupanloup *et al.* (2002). The method assumes geographic homogeneity within populations that are maximally geographically removed from one another. The algorithm sorts an initially random partition of populations into a specified K number of groups (equal to or less than the number of populations) optimally and iteratively moving geographically adjacent populations between groups until the maximum proportion of total genetic variation can be attributed to differences between groups. This application of spatial analysis of molecular variance is performed in the SAMOVA 1.0 software. In the software application a coordinate file was created using a set of regionally centred coordinates estimated for each breed. In this way the algorithm can adjust genetic relationships according to geographical ones in order to distribute partitions across the landscape.

2.6.3. Higher-order clustering

The Corander *et al.* (2004) method also utilised a Bayesian model-based clustering method and was applied through the BAPS 2 program. This method assumes populations and given a maximum value for the number of partitions in the data, uses a stochastic optimisation process in order to reach the clustering solution with the highest likelihood of K. The value of K is varied until a stable likelihood is reached over a number of repetitions. We used 20 repeats per value of K, varying K from 2 to 28.

2.6.4. Migration between clusters

Additional to the methods employed to infer population structure are those whose application can identify migrants between the populations. The investigation into migrant assignment was performed through the method described in Piry *et al.* in 2004. This method aims to go some way towards rectifying the excess exclusion of resident individuals in other methods, relying on a genetic distance criterion, an allele frequency based criterion, and a criterion based on a derivation of the Bayesian method developed by Rannala and Mountain (1997). The assignments criterion is the

genetic distance between the individual for assignment and the reference population (Cornuet et al. 1999). Frequency based assignment was based on the likelihood of an individual originating from a sample given the frequency of that individual's alleles in the population. The Bayesian method similarly calculates a likelihood, the prior distribution for allele frequency being a Dirichlet distribution for the current allele given all allelic states over all reference populations. The implementation of this method through the software GENECLASS2 differs from the precursor GENECLASS through its use of a simulation algorithm that generates population samples of the same size as the reference sample. The assignment critera are then calculated for each individual other than the one in question. Due to the inclusion of the sample size of the reference population, the sampling variance associated with the analysed dataset is better reflected.

2.7. Analysis of admixture

The admixture methodology applied here is based on the Monte Carlo Markov chain method of Chikhi *et al.* (2001) although two other methods are included for comparative purposes. A purely coalescent approach is applied in ADMIX2.0 (Dupanloup and Bertorelle, 2001) and a maximum likelihood method through LEADMIX (Wang *et al.* 2003). All three methods are used as high performance estimators of relative parental contributions (Choisy *et al.* (2004).

2.7.1. LEA

The Likelihood-based Estimation of Admixture (LEA) method is detailed in Chikhi et al. (2001). This method initially assumes two independent parental populations (P_1,P_2) of given sizes (N_1,N_2) that mix to produce a hybrid population (H) a number of generations in the past (T) (Figure 2.2). The parental populations each contribute a proportion to the hybrid population (p_1,p_2) summing to 1. At the point of hybridisation the gene frequency distributions of P_1, P_2, and H are also represented (x_1,x_2, and ($p_1x_1 + p_2x_2$)). After the admixture event, all populations evolve independently but without mutations until the present. The time since the admixture event is scaled by the effective population size for each of the populations ($t_1t_2t_h$).

In the Bayesian approach of this method the aim is to draw inferences about a parameter, or parameter set, of a model using the information contained in the data. The resultant probability density function describes the probability distribution

as a result of the given data. The prior probability density function (pdf) describes what is already known about the parameters, the posterior pdf is produced when the given data is considered. A full-likelihood approach assumes that all relevant information is contained in the posterior density function. This puts the emphasis on the complete distribution rather than using point estimates of parameters as summary statistics.

A Bayesian requirement is that a prior is provided on all model parameters. When is not possible to infer a particular prior distribution then this lack of knowledge can be represented as flat. The result is that the posterior will be proportional to the likelihood function. In this case the flat prior applies to p_1, t_1, t_2, and t_h, but the x_1 and x_2 priors are such that any possible allele frequencies are equally possible through a uniform Dirichlet distribution. This allows no dependence on the genetic distance between the parental populations, allowing any possible history of these populations to be accounted for.

Calculating the full Likelihood requires calculation of the probability of observation of allelic configurations given those of the founders just after admixture, the probability of observation of coalescent events within the timescale, and a measure of the probability of the allelic configuration of the founders given the distribution in the ancestral parental population and amount of admixture. Large numbers of allelic configurations renders the estimation of the likelihood directly is computationally expensive. An alternative is to make estimations of the allelic configurations at each coalescent event. The Griffiths and Tavaré (1994) method harnesses a Monte Carlo approach to evaluate the likelihood at specific parameter values.

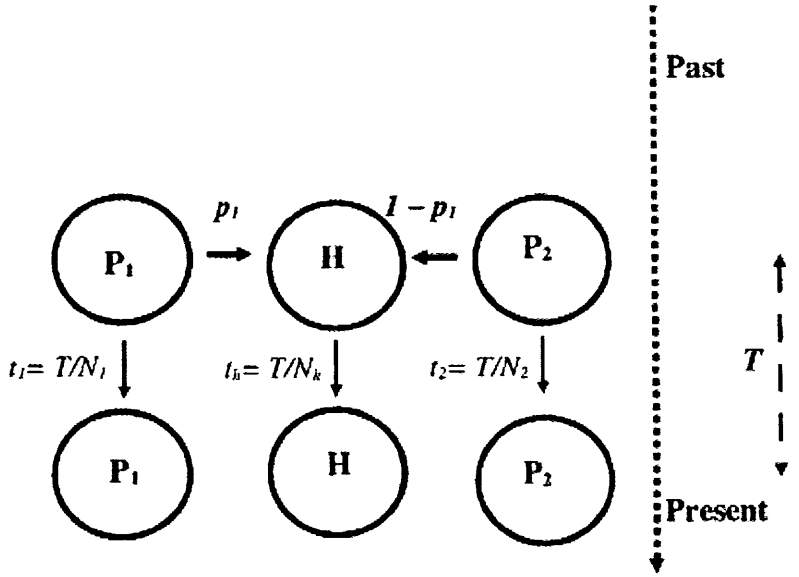


Figure 2.2. The admixture model of LEA. A single admixture event is assumed T generations ago, the populations are allowed to have different sizes N1, N2, and Nh. The contribution of the first parental population is P1.

The Markov chain has a given starting point in parameter space for each run of the LEA program. To ensure that this starting point has no effect on the analysis result there is a 'burn-in' period during which the Markov chain reaches an equilibrium position, this portion of the run must be discarded. The application of the Gelman statistic (Gelman, 1996) can be used to compare whether a number of chains have converged on the same result. Taking a variety of run lengths, this statistic can be used to compare the similarity of the result after removal of a proportion of the beginning of the Markov chain (ten percent was a usual value chosen). Typical lengths of LEA runs in this study were 5 $x10^5$ steps with checks being made using at least one repeat run of 1 $x10^6$ steps.

The output generated by LEA is in the form of six columns of data with one row for each five steps of the run. The columns represent; program iteration, likelihood, p_1, t_1, t_2, and t_h. Due to the potentially large size of output files, R language (Ihaka and Gentleman, 1996) was employed for analysis. A density plot was used to generate a distribution for each parameter. The distribution is used to calculate the estimate for the parameter and the confidence margins.

2.7.2. ADMIX2.0

The model employed in this method is developed and detailed in Bertorelle and Excoffier (1998) with an extension to include multiple parental populations made by Dupanloup and Bertorelle (2001). It introduces m_Y as an estimator of the admixture coefficient based on coalescence times between pairs of genes sampled within and between populations. The model is based on a simplified admixture scenario; an ancestral population splits into two parental populations and generates a hybrid population with proportions combining two fractions of genes (μ and (μ-1)) taken at random from each parent population (Figure 2.3). The assumption is made that the parental populations simultaneously contribute to the hybrid population. From that point the populations evolve independently for t_A generations. m_Y is a least squares estimator of μ which is derived and applied to the data. The initial term in the model considers the number of coalescence events that will occur from the present time until the admixture event, the second term is concerned with the coalescence events during the time period over which the parental populations were kept separated. In this latter calculation, coalescent events can only occur between those genes that co-migrated in the same parental population. The last terms consider the coalescences occurring in the ancestral population, these events have different probabilities depending on whether the two genes co-migrated in the same population or not.

The coalescence times between two genes are not directly accessible and means for these values must be estimated from the genetic variability. For microsatellite data the single-step stepwise model of mutation is adopted. This employs coalescence times estimated from the mutation rate and the average squared difference in allele size. The single step mutation model, whereby each mutation can increase or decrease the allele size by a single repeat, has been widely accepted and applied as an approximation of the underlying source of diversity in microsatellites (Goldstein *et al*, 1995; Slatkin, 1995). By taking into account mean coalescence times between the hybrid and additional populations this estimator can be extended to the contribution of more than two populations to the hybrid.

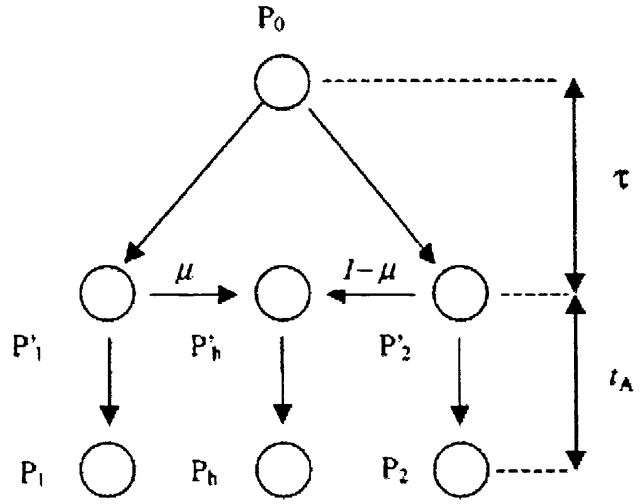


Figure 2.3. The admixture model of ADMIX2.0.

2.7.3. LEADMIX

This admixture method adopts the model applied in Bertorelle and Excoffier (1998) but with some important differences; the former method is a moment estimator only and estimates just p_1 as well as assuming an equal effective size of all populations. The method applied here through LEADMIX (Wang, 2003) has been extended for application with any number of parental populations and also takes into account the differentiation between parental populations in the admixture calculation. This latter feature is a consequence of the potential introduction of bias that could result in the likelihood methods from falsely assuming independent uniform priors for the allele frequency distributions of populations P_1 and P_2 when the admixture event occurs. The model employs eight parameters; the admixture proportion p_1, the two periods of time in generations, ξ and ψ, and the average effective sizes of the parental populations P_1 and P_2 during the period ξ (n_1 and n_2) and of the parental and hybrid populations during period ψ (N_1, N_2, and N_3) (Figure 2.4). The number of parameters is reduced to six through the rescaling of time by the effective sizes of the populations.

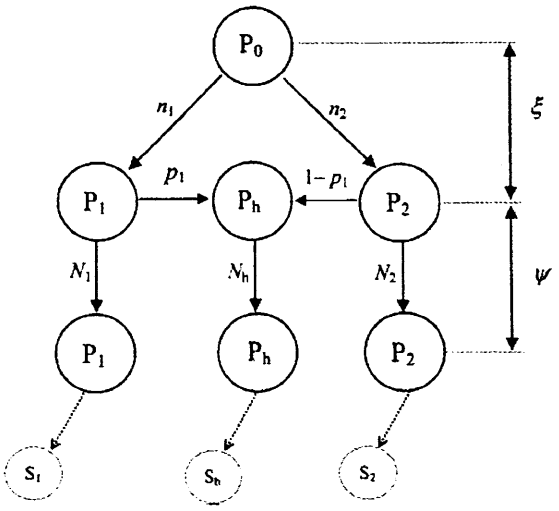


Figure 2.4. The admixture model of LEADMIX. The ancestral population P0 splits into two parental populations P1 and P2 (of sizes n1 and n2), which evolve independently before they contribute genes of proportions p1 and 1-p1 to form the hybrid population Ph. After formation of the hybrid population P1, P2, and P3 with effective sizes N1, N2, and Nh evolve independently. ξ and ψ represent the time periods between the separation of the parental populations and the admixture event, and the time since admixture event respectively

2.7.4. Approximate Bayesian computation

The MATLAB v7 (The Mathworks, 2001) platform was used for the application of the ABC method. A graphical interface initiates the computation which comprises four main stages; specification of model parameters, generation of the simulated data, calculation of summary statistics for observed and simulated data, and application of a comparison between observed and simulated data in order to predict population parameters.

2.7.4.1. Generation of parameter information

Like all admixture models the ABC approach developed here is model based and relies on this model in order to specify the parameters for the admixture event. Although the parameters required are dependent on the genetic model detailed below, the process of parameter generation is the initial step in the actual application of the approach so shall be detailed first.

In order to run potentially millions of simulations over a range of values for each parameter an automated procedure is required for practical purposes. Random

values can be generated using information about the distribution and range. A graphical user interface (GUI) can be used for the specification of these (Figure 2.5). The generated value for each parameter is then stored by row in a file, each column representing a single parameter. In this interface the distribution of the parameters in the range can be uniform, normal, gamma, or log normal. Error messages are set to be generated if the 'Run' button is pressed without correct completion of the input regions. The program executes on the selection of the 'Run' button and output files are generated.

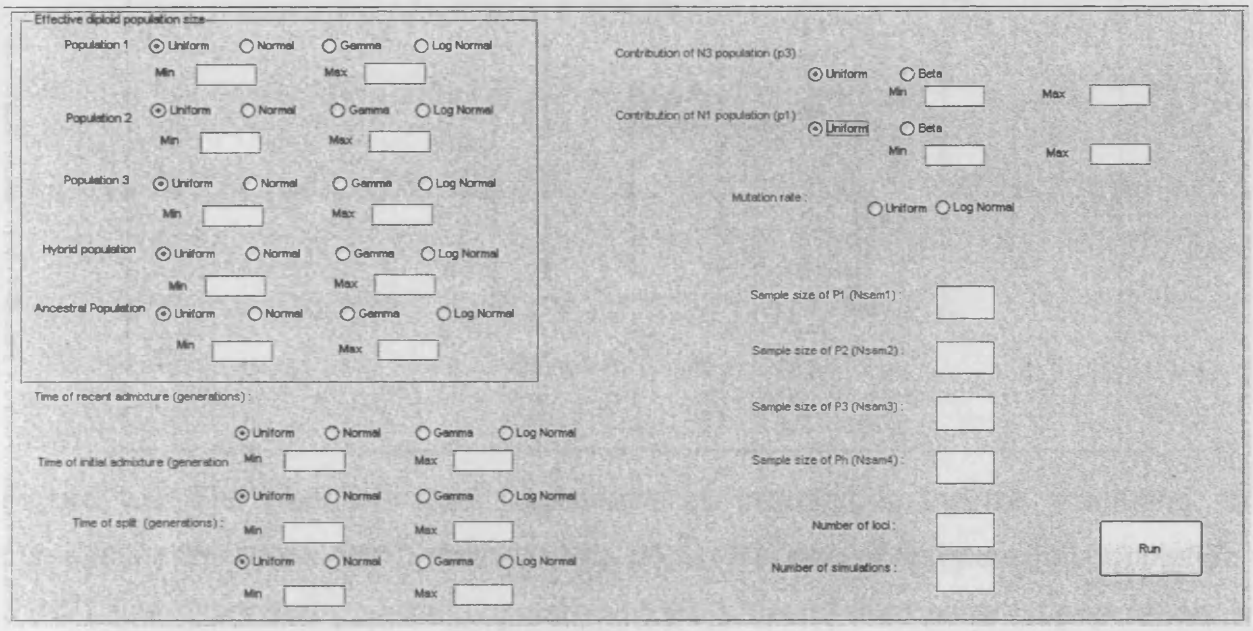


Figure 2.5. The Graphical user interface for the specification of parameter limits, number of loci, and number of simulations.

2.7.4.2. Simulation of data through the genetic model

The specification of the genetic model to use is made through the ms application (Hudson, 2002). The application requires a command line for the generation of microsatellite data under a particular genetic model. The method applied here uses the same model (i.e. parameter values) to generate data independently for as many loci as are present in the observed data. The process is repeated for each simulated dataset with different parameter values from the range. The model applied is one of an initial split event creating three populations, the first two of which admixed to form a hybrid population which later receives genes from an admixture event with the third parental population. The parameters that need to be specified in this model are; effective population sizes, mutation rate of the molecular marker, time of

coalescence, time of first admixture event, time of second admixture event, size of the ancestral population, and sizes of sample populations (Figure 2.6).

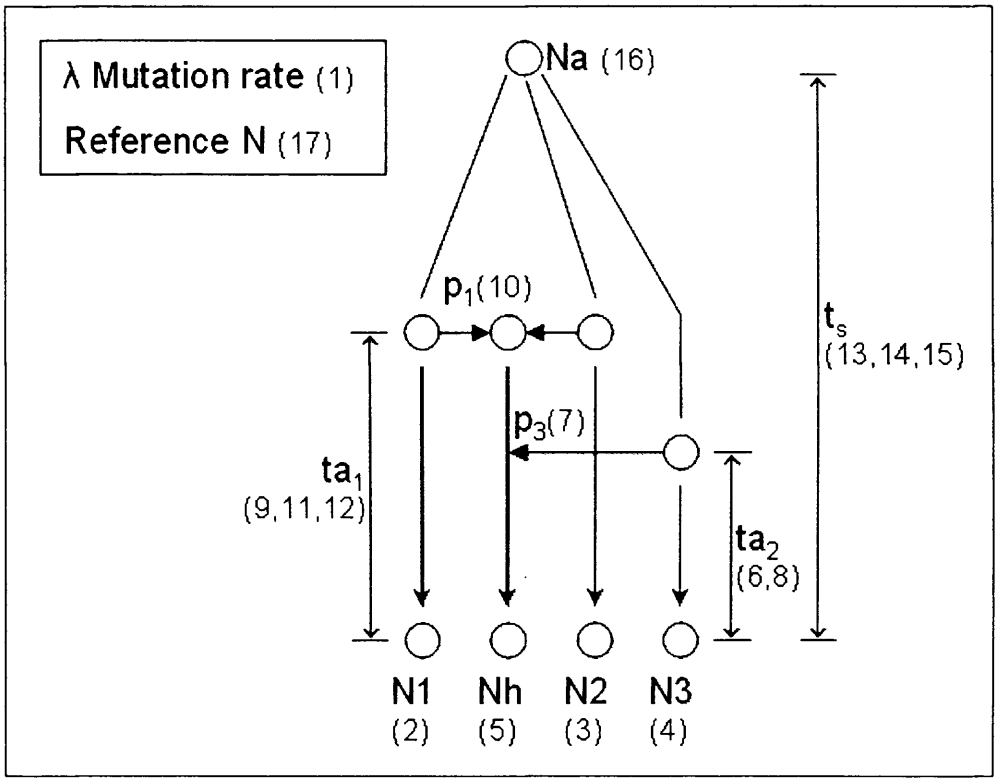


Figure 2.6. The genetic model parameters as included in the ms command. In numerical order these are; 1 mutation rate (λ), 2 (N1) size of first parental population, 3 (N2) size of second parental population, 4 (N3) size of third parental population, 5 (Nh) size of hybrid population, 6(8) (ta2) scaled time since recent admixture event (/4*Nref), 7 (1-p3) proportion contribution in recent admixture event, 9(11,12) (ta1) scaled time since older admixture event (/4*Nref), 10 (p1) proportion contribution in recent admixture event, 13(14,15) (ts) scaled time since initial split from ancestral population (/4*Nref), 16 (Na) size of ancestral population, 17 reference population size.

As detailed in the ms user information, simulations are generated according to the command entered (Figure 2.7.) in conjunction with the file of parameter information. The program is able to generate independent simulations with the use of 'tbs' which allows the values to be inserted using the same framework for each line of the parameter file. In this case although independent values are generated for each simulation, the same parameters are then used for each locus within a simulated dataset. In this way the same model is applied to all of the loci in a simulation although the resultant allelic information will be different in each locus due to independent generation.

```
command=['c:\Data_Analysis\Ms\ms.exe', num2str(Nsam),' ',
num2str(handles.Nsim*handles.nb_loci),' -t tbs -I 4 ', num2str(handles.Nsam1),' ',
num2str(handles.Nsam2),' ', num2str(handles.Nsam3), ' ', num2str(handles.Nsam4),'
-n 1 tbs -n 2 tbs -n 3 tbs -n 4 tbs -es tbs 4 tbs -ej tbs 5 3 -es tbs 4 tbs -ej tbs 6 2 -ej
tbs 4 1 -ej tbs 3 2 -ej tbs 2 1 -en tbs 1 tbs
```

Figure 2.7. The ms command for generation of simulated data under he two admixture model.

2.7.4.3. Applying summary statistics

Due to the stochastic process of generation of simulated datasets it is most convenient to summarise the data into a set of statistics. These were chosen in order to accurately represent the genetic information in the population data. For this the following statistics were applied;

- A measure of heterozygosity for each population followed by an overall heterozygosity across all populations.
- Allelic range by population and overall allelic range.
- Private alleles by population and an overall measure of private alleles among the populations.
- F_{ST} by population and overall F_{ST}.

The generation of the summary statistics is performed once for the observed data and then separately for each simulation of the data generated according to the input model parameters. This enables the independent comparison of each simulation with the observed data. Across the four populations involved in this model there are 23

statistics which need to be compared between the simulated and observed data. A measure of the difference between the summary statistics in each must be calculated to allow their concatenation into a single measure. A proportional weighting is therefore introduced to avoid a bias on any single statistic due to relative magnitude.

2.7.4.4. Prediction of parameters

The distance measure is calculated for each simulated dataset against the observed data. Lower distance measures reflect greater proximity between the observed and simulated data. In order to only include a subset of the simulated datasets corresponding to the lowest distance measures, a rejection scheme is applied accepting values according to a tolerance determined by a particular percentage of the distribution of distances. The procedure samples the first 10,000 observed-simulated distances in order to calculate a distance margin within which fall the lowest proportion of the distance values depending on the chosen tolerance (e.g. 10%). This allows comparisons to be made including the simulations whose summary statistics most closely approximate the observed data. The distribution of values for each parameter can be plotted for this subset of simulations accepted in the rejection process. From characterising the distribution of the simulated parameter values an estimate of the parameter is made.

3. References

ArkDB (2008) Website of the Roslin Institute ArkDB genome database http://www.thearkdb.org/

Barendse W., Armitage S.M., Kossarek L.M., Shalom A., Kirkpatrick B.W., Ryan A.M., Clayton D., Li L., Neibergs H.L., Zhang N., Grosse W.M., Weiss J., Creighton P., McCarthy F., Ron M., Teale A.J., Fries R., McGraw R.A., Moore S.S., Georges M., Soller M., Womack J.E., & Hetzel D.J.S. (1994).A genetic linkage map of the bovine genome. *Nature Genetics* **6**, 227-235.

Bayes T. (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* **53**, 370-418.

Beerli P. (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**(3), 341-345.

Beja-Pereira A., Alexandrino P., Bessa I., Carretero Y., Dunner S., Ferrand N., Jordana J., Laloe D., Moazami-Goudarzi K., Sanchez A., & Canon J. (2003) Genetic characterisation of Southwestern European bovine breeds: A historical and biogeographical reassessment with a set of 16 microsatellites. *Journal of Heredity* **94**(3), 243-250.

Belkhir K., Borsa P., Chikhi L., Raufaste N., & Bonhomme F. (2002) GENETIX, logiciel sous WindowsTM pour la génétique des populations. Laboratoire Génome, Populations, Interactions CNRS UMR 5000, Université de Montpellier II, Montpellier (France).

Bertorelle G., & Excoffier L. (1998) Inferring admixture proportion from molecular data. *Molecular Biology and Evolution* **15**, 1298–1311.

Bishop M.D., Kappes S.M., Keele J.W., Stone R.T., Sunden S.L.F., Hawkins G.A., Solinas –Toldo S., Fries R., Grosz M.D., Yoo J., & Beattie C.W. (1994) A genetic linkage map for cattle. *Genetics* **136**(2), 619-639.

Brezinsky L.S., Kemp J., & Teale A.J. (1993a) ILSTS005: a polymorphic bovine microsatellite. *Animal Genetics* **24**, 73.

Brezinsky L.S., Kemp J. & Teale A.J. (1993b) ILSTS006: a polymorphic bovine microsatellite. *Animal Genetics* **24**, 73.

Chikhi L., Bruford M. W., & Beaumont M.A. (2001) Estimation of admixture proportions: A likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**, 1347-1362.

Choisy M., Franck P., & Cornuet J.M. (2004) Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Molecular Ecology* **13**, 955-968.

Corander J., Waldmann P., Marttinen P., & Sillanpää M.J. (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**, 2363-2369.

Cornuet J.M. & Luikart G. (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144,** 2001-2014.

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

Dupanloup I., & Bertorelle G. (2001) Inferring admixture proportions from molecular data: extension to any number of parental populations. *Molecular Biology and Evolution* **18**(4), 672-675.

Dupanloup I., Schneider S., & Excoffier L. (2002) A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology* **11**(12), 2571-81.

Evanno G., Regnaut S., & Goudet J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L. & Heckel G. (2006) Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics* **7**, 745-758.

Falush D., Stephens M., & Pritchard J. K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**,1567-1587.

Gelman A. (1996) Inference and monitoring convergence, in *Markov Chain Monte Carlo in Practice*. eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, New York: Chapman and Hall, pp. 131-143.

Georges M. & Massey J.M. (1992) Polymorphic DNA markers in Bovidae. Patent WO 92/13102 (1992).

Glaubitz J.C. (2004) CONVERT: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Molecular Ecology Notes* **4**, 309–310.

Goldstein D.B., Linares A.R., Feldman M.W., & Cavalli-Sforza L.L. (1995) An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**, 463-471.

Goossens B., Waits L.P., & Taberlet P. (1998) Plucked hair samples as a source of DNA: reliability of dinucleotide microsatellite genotyping. *Molecular Ecology* **7**, 1237-1241.

Griffiths R.C., & Tavaré S. (1994) Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131-159.

Hudson R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**(2), 337-338.

Ihaka R., & Gentleman R. (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* **5**(3), 299-314.

Kumar P., Freeman A.R., Loftus R.T., Gaillard C., Fuller D.Q., & Bradley D.G. (2003) Admixture analysis of south asian cattle. *Heredity* **91**, 43-50.

Langella O. (1999) POPULATIONS 1.2.28 ed. Gif Sur Yvette: Laboratoire Populations, Génétique et Evolution, Centre National de la Recherche Scientifique, CNRS UPR9034.

Lindley D.V. (1986) Comment. *The American Statistician* **40**(1), 6-7.

Moore S.S., & Byrne K. (1993). Dinucleotide polymorphism at the bovine calmodulin independent adenylcyclase locus. *Animal Genetics* **24**, 150.

Moore S.S., Byrne K., Berger K.T., Barendse W., McCarthy F., Womack J.E., & Hetzel D.J.S. (1994) Characterisation of 65 bovine microsatellites. *Mammalian Genome* **5**, 84-90.

Piry S., Alapetite A., Cornuet J–M., Paetkau D., Baudouin L., & Estoup A. (2004) GENECLASS2: A Software for Genetic Assignment and First-Generation Migrant Detection. *Journal of Heredity* **95**(6), 536–539.

Pritchard J.K., Stephens M., & Donnelly P.J. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Rannala B., & Mountain J.L. (1997), Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the USA* **94**, 9197-9102.

Raymond M., & Rousset F. (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**, 248-249.

Sano A., Shimizu A., & Lizuka M. (2004) Coalescent process with fluctuating population size and its effective size. *Theoretical Population Biology* **65**(1), 39-48.

Slatkin M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457-462.

Solinas-Toldo S., Fries R., & Steffen P. (1993) Physically mapped, cosmid derived microsatellite markers as anchor loci on bovine chromosomes. *Mammalian Genome* **4**, 720-727.

Steffen P., Eggen A., Dietz A.B., Womack J.E., Stranzinger G., & Fries R. (1993) Isolation and mapping of polymorphic microsatellites in cattle. *Animal Genetics* **24**(2), 121-124.

The MathWorks Inc. (2001) *MATLAB 7*, 24 Prime Park Way, Natick MA.

Thieven U., Solinas-Toldo S., Friedl R., Masabanda J., Fries R., Barendse W., Simon D., & Harlizius B. (1997) Polymorphic CA-microsatellites for the integration of the bovine genetic and physical map. *Mammalian Genome* **8**, 52-55.

Tufto J., Engen S., & K. Hindar (1996) Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics* **144**, 1911–1921.

Vaiman D., Mercier D., Moazami-Goudarzi K., Eggen A., Ciampolini R., Lepingle A., Velmala R., Kaukinen J., Varivio S.L., Martin P. Leveziel H., & Guerin G. (1994) A set of 99 cattle microsatellites: characterization, synteny mapping, and polymorphism. *Mammalian Genome* **5**, 288-297.

Vigilant L. (1999) An evaluation of techniques for the extraction and amplification of DNA from naturally shed hairs. *Biological Chemistry* **380**, 1329-1331.

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

Walsh P.S., Metzger D.A., & Higuchi R. (1991) Chelex R100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques* **10**, 506-513.

Wang J. (2003) Maximum-likelihood estimation of admixture proportions from genetic data, Genetics 164: 747-765.

Wang J. (2005) Estimation of effective population sizes from data on genetic markers. *Philosophical Transaction of the Royal Society B* **360**, 1395-1409.

Weir B.S., & Cockerham C.C. (1984) Estimating F-Statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Wiener P., Burton D., & Williams J.L. (2004) Breed relationships and definition in British cattle: a genetic analysis. *Heredity* **93**, 597-602.

Chapter 3.

Manuscript in preparation for *Conservation Biology*

Population genetic structure, demographic history and conservation of minority British, Irish, and European cattle breeds

T.C. BRAY[a], L. CHIKHI[b], B. GOOSENS[a], G.L.H. ALDERSON, K. BYRNE, J. DE RUITER, A.J. SHEPPY, S. TOWNSEND, and M. W. BRUFORD[a],

[a]Cardiff School of Biosciences, Cardiff University, P.O. Box 915, Cardiff CF10 3TL, UK.

[b]University Paul Sabatier, Toulouse

Correspondence: M. W. Bruford, Fax: 029 20 874316; E-mail: BrufordMW@Cardiff.ac.uk

Running title:

Keywords: genetic drift, gene flow, livestock

3.1. Abstract.

Genetic diversity and breed structure were investigated in 27 distinct original traditional cattle breed populations, many of which have become minority breeds in recent times. Despite potentially high levels of gene flow between some breed populations, all maintain identifiable genetic distinctiveness. Genetic differentiation was not shown to clearly correlate with geographic origins. Previous findings partitioning variation between some British Isles and European mainland breed populations were recapitulated. The genetic signature of demographic events and migration was shown to be key to the development of contemporary breed variation. A Weitzman-based diversity measure suggested further partitioning of extant genetic diversity and the existence of two groups of British breed populations. We provide evidence for the distortion of breed importance in small populations that have undergone significant levels of genetic drift. We also suggest that there is a potentially important relationship between within breed diversity and breed distinctiveness that may prove detrimental in applying Weitzman for determining priorities for breed conservation.

3.2. Introduction

The almost ubiquitous distribution of domestic cattle worldwide reflects their historical importance to humans and has resulted in the ~1400 extant breed populations that we recognise today (Ajmone-Marsan, 2008). The spread of cattle from Near East and African origins since domestication explains much of the extant diversity at a large spatial scale (Loftus *et al.* 1994; Bradley *et al.* 1996; Hanotte *et al.* 2002; Caramelli, 2006). However, finer scale within and among-breed variation can be unrelated to geographic separation (Jordana *et al.* 2003) and commercial breed development has been noted as contributing to this phenomenon (Rendo *et al.* 2003). Even local-scale selection strategies can be shown to produce variation between herds within a breed (Beja-Pereira *et al.* 2003) contributing to breed evolution. The cultural heritage represented in traditional breed populations has been historically affected by the socio-economic environment in which the breed has been developed (Bartosiewicz, 1997). Periodic changes in emphasis on particular traits have been seen to effect homogenisation and divergence among breeds and this phenomenon can be related to temporary changes in breed popularity (Notter, 1999) often with the result of breed decline. Low levels of gene flow between populations can act to maintain diversity and reduce the inter-population differentiation that accrues through genetic drift. High

levels of gene flow can reduce inter-population differentiation and erode population distinctiveness altogether (Slatkin, 1985). This reduction in breed differentiation due to gene flow has been seen in sympatric populations that are known to exchange genetic material (Beja-Pereira *et al*, 2003). The consequences of this gene flow can therefore affect the conservation value attributed to breed populations.

Weitzman's (1992,1993) diversity concept for conservation has been influential in consideration of conservation priorities in domestic animal populations (Reist-Marti et al. 2003), although is not so well acknowledged in the wild species literature, for which the approach was originally designed. The approach combines genetic distance/diversity information (among breeds/ within a set of breeds) and estimated survival probabilities to assess the potential change in overall diversity according to a specified time horizon and breed persistence. This provides a framework for predicting which populations could be most cost effectively managed to maximise the predicted level of surviving diversity. Implicit in this approach is the marginal diversity that is attributed to populations. Those populations that, once removed from the breed group, display greater differentiation relative to the remaining group members are attributed a higher marginal diversity and greater conservation value. There has, however, been widespread criticism since the first Weitzman application to cattle by Thaon D'arnoldi *et al.* (1998), highlighting the failure of the approach to include intra-population genetic diversity with the suggestion of improved methodologies (Caballero and Toro 2002; Eding *et al.* 2002). A selection of non-genetic and combined approaches to breed conservation and prioritisation has also been suggested: Simianer *et al.* (2003) argued that there should be conservation priority attributed according to a more utilitarian concept of value in addition to classic diversity measures. This use of aspects of breed utility includes attributing value scores to qualitative concepts such as unique production traits and social, cultural, and religious roles. Producing recommendations according to a wide range of these more subjective criteria is potentially important for making an informed decision about conservation priorities but there is still merit in making more simplified breed descriptions limited to a single factor such as level of endangerment (Danell *et al.* 1998; Simianer, 2005), economic viability (Rege and Gibson, 2003), cultural value (Gandini and Villa, 2003), and socioeconomic functions of breeds (Tisdell, 2003).

European cattle breed populations have been studied widely to assess their molecular diversity and inter-breed relationships (e.g. Blott *et al.* 1998; Beja-Pereira *et al.* 2003; Wiener *et al.* 2004), including in a conservation context (Ruane, 2000; Canon *et al.* 2001; Gandini 2003; Simianer 2005). Examples exist of the recognition of threat and the successful mitigation of declining breeds (Bartosiewicz, 1997) as well as the upgrading of production traits to increase commercial viability (Vollema and Groen, 1997). What seems less clear however, is how overall breed population dynamics compare across traditional breed populations, all of which display unique histories. These traditional populations are potentially important to maintain, not least for their role as reservoirs of diversity (Giovambattista *et al.* 2001), but despite this many are classified as endangered (EFABIS, 2007). To what extent the unique diversity of these older, traditional (i.e. autochthonous breeds that have been subject to less intensive selection) populations has been maintained, particularly those that have experienced severe or extended demographic bottleneck events, is of great conservation interest. Furthermore, whether the older traditional breed populations have been robust in response to the increased potential of selection and introgression in today's competitive agricultural landscape is unknown.

In this study we apply a number of diversity measures and population clustering approaches to a selected set of traditional European cattle breed populations in order to assess inter-population differentiation and compare relationships both accounting for and ignoring geographical separation. How the extant breeds reflect their individual demographic histories with specific reference to population bottlenecks, and to what extent breed evolution can be identified through measures of genetic drift, was also investigated. Finally, we used these data to reassess the applicability of Weitzman-derived approaches towards prioritisation for conservation. Genetic distance-based approaches are still being widely applied in contemporary domestic cattle literature (Tapio *et al.* 2006; Li *et al.* 2007; MacNeil *et al.* 2007) although they may present a poor reflection of genetic relationships between breed populations (e.g. Bruford, 2004). Weitzman marginal diversity measures are genetic-distance based and commonly applied in conjunction with other approaches (eg. Glowatzki-Mullis *et al* 2008) and for comparison with newly developed methods (Caballero and Toro, 2002; Eding *et al.* 2002; Fabuel *et al.* 2004; Ollivier and Foulley, 2005). In this case the breeds used span multiple regional and country boundaries and the assessment of breed 'value' at this scale is of limited direct utility. However, we are

interested in the performance of the genetic distance approach, theoretical limitations notwithstanding, in the relative assessment of the original population traditional breeds used in this study.

The particular aims of this study are to assess the degree to which the traditional breed populations used here can be distinguished and how any distinction can be attributed to geographic or demographic determinants. In addition, we are particularly interested in the potential effects of migrant exchange between breeds and whether regional migration maxima can be identified.

3.3. Materials and methods

3.3.1. Sampling

Twenty six traditional UK and European breeds of cattle, and one breed from West Africa for comparison, were sampled using carefully chosen original populations. These were populations whose herd book records showed little or no influence of recent introgression or modernisation through intensive selection for commercial traits. Sample sizes ranged from nine to 154 individuals, and ranged from rare to non-endangered status (Table 3.1). Both plucked hair and cryogenically frozen semen samples were used. 762 individuals of British, European, and West African origin were typed at nine loci, 334 of these were also typed across an additional 3 loci. Each sample group reflects a separate breed.

Table 3.1. Sample information by breed

Breed	Sample Origin	Number of individuals	Number of loci	Breed type and specific attribute	Conservation status / trend
Shetland *	UK	31	12	dual	Rare / stable
Aberdeen Angus	UK	20	9	beef	Not endangered / stable
Belted Galloway+	UK	15	12	beef	Rare / stable
Irish Moiled	UK	20	12	dual	Rare / increasing
Beef Shorthorn *	UK	11	9	beef	Rare / stable
Lincoln Red	UK	60	9	beef	Unknown / stable
British White +	UK	11	12	dual	Unknown / stable
Kerry *	UK	32	12 / 9	dairy	Rare / stable
Traditional Hereford *	UK	19	12	beef	Rare / unknown
Red Poll *	UK	20	9	dual	Rare / stable
Welsh Black	UK	17	9	beef	Not Endangered / decreasing
Dexter	Ireland	154	9	dual	Not Endangered / stable
Sussex	UK	50	12	beef	Rare / stable
White Park *	UK	33	12	beef	Rare / increasing
Gloucester *	UK	14	12	dairy	Rare / stable
Milking Devon	UK	32	9	dairy	Not Endangered / stable
Beef Devon	UK	20	9	beef	Not Endangered / stable
UK Mainland Jersey	UK	21	9	dairy	Not Endangered / stable
Guernsey	UK	12	9	dairy	Not Endangered / stable
Island Jersey	UK	25	12	dairy	Not Endangered / stable
Jutland *	Denmark	17	12	dairy	Potentially endangered / increasing
Angeln	Germany	24	12	dairy	Not Endangered / stable
German Black Pied	Germany	19	12	dairy	Endangered / stable
Hungarian Grey	Hungary	16	12	beef	Unknown / unknown
Limousin	France	23	12	beef	Not endangered / unknown
Berrenda	Spain	31	12	beef/fighting bulls	Unknown / stable
N'dama*	West Africa	9	12	dual	Not endangered / unknown

3.3.2. DNA extraction

Plucked hair samples were extracted using either the chelex 100 protocol described in Walsh *et al* (1991) with specific details in Goossens *et al* (1998) or with a PCR buffer-based method described in Vigilant (1999). Semen samples were extracted using a modification of the Qiagen Dneasy tissue extraction method according to manufacturers' instructions. From the extractions 1.5µl was added as template in each PCR reaction.

3.3.3. Genotyping

The microsatellite markers used in this study were taken from the Food and Agriculture Organisation (FAO) list of cattle markers that have been used in a number of recent studies (Beja-Pereira *et al.* 2003; Kumar *et al.* 2003; Wiener *et al.* 2004). Data are combined, where appropriate, from studies using nine and twelve loci. The first nine microsatellite markers were; inra063 (Vaiman et al. 1994), eth225 (Steffen et al. 1993), hel5 (Kaukinen and Varvio, 1993), eth10 (Solinas-Toldo et al. 1993), bm1818 (Bishop et al. 1994), ilsts006 (Brezinsky et al. 1993), haut27 (Thieven et al. 1997), tgla227 and tgla122 (Georges and Massey, 1992). Three additional markers used in the larger dataset were inra005 (Vaiman *et al.* 1992), bm2113, and bm1314 (Bishop et al. 1994). These were amplified using the Qiagen multiplex kit according to manufacturers' instructions. One primer from each pair was synthesized with a fluorescent dye FAM, HEX, or NED on the 5' end. Amplification of the loci was carried out in 6µl reactions (1xQIAGEN PCR Multiplex Master Mix (3mM $MgCl_2$), 0.2µM each primer). Thermocycling conditions were as follows: initial denaturation at 95' for 15' followed by 35 cycles of 60s at 94°C, annealing for 90s at 55°C then extension for 60s at 72°C, with a final extension for 10' at 60°C. Markers were applied in multiplex conditions using up to seven pairs per PCR. All PCR products were analysed using an ABI 377 and 3100 semi-automated DNA analysers. Gels were analysed using Genescan analysis 2.0™, Genotyper 1.1™, and Genemapper™ software.

3.3.4. Statistical Analysis

3.3.4.1. Genetic variability and Population Structure

Genetic diversity measures included Expected Heterozygosity (H_E), and Observed Heterozygosity (H_O) and Wright's F statistics, computed following the method of Weir

and Cockerham (1984), as well as a heterozygosity measure that accounts for bias due to sample size (Nei 1978). Their departure from the null hypothesis, which was no differentiation between populations (F_{ST}) and Hardy-Weinberg equilibrium between individuals in subpopulations (F_{IS}) and individuals in the total population (F_{IT}), was tested over 10^4 permutations as implemented in the GENETIX 4.03 software (Belkhir *et al.* 2002).

The majority of populations used in this analysis are relatively closely related taurine cattle with varying levels of recent and ancient genetic exchange. In order to investigate cryptic population structure, the whole dataset was examined using the Bayesian model-based clustering approach developed by Pritchard *et al.* (2000). This method, further improved by Falush et al. (2003) is implemented in STRUCTURE 2.1. This groups individuals into K homogeneous clusters (populations) that are as close to Hardy-Weinberg and Linkage equilibrium as possible. For each K value, we performed 20 runs with different starting points, having a 10^5 burn-in period followed by 10^5 steps (we tested different run lengths ranging from 10^4 to 10^6 and found that convergence was achieved after 10^5 steps). Then, for each K value, we calculated the average and standard deviation of the 'log estimated likelihood' [$L(K)$] across the 20 runs. The values of ΔK (Evanno *et al*, 2005) statistics were obtained as $\Delta K = m(|L(K+1) - 2L(K) + L(K-1)|)/s[L(K)]$, where m and s represent the average and standard deviation of the corresponding values across 20 runs, respectively. The ΔK statistic was thus used to determine the uppermost level of population structure. In order to determine whether each cluster was itself subdivided into smaller and less differentiated units, the identified clusters were reanalysed independently. This was repeated for each subgroup until the most likely K was shown to be 1. For all these analyses, the program was run under the admixture model, considering independent allele frequencies.

The breed sample populations were largely of either British or European mainland origin. To investigate potential geographical explanations for current breed relationships, an analysis was performed using coordinates of origin for each sample population. These coordinates were obtained using approximate centralised values of either region of origin or entire country if the breed is widespread. The analysis was applied according to the method of Corander *et al.* (2003, 2004, 2007) using prior knowledge of sampling location and estimating the posterior probabilities for all

the different ways of combining populations. Stochastic optimisation is used to infer the posterior mode of the genetic structure applied in the software BAPS 4.

The investigation into migrant assignment was performed through a method described in Piry *et al.* in 2004. This method uses three assignment criteria; genetic distance, allele frequency, and a measure based on the Bayesian method developed by Rannala and Mountain (1997). Implementation of this method is through the software GENECLASS2.

3.3.4.2. *Demographic analysis*

Of particular interest to us was the presence of the signatures of demographic events in the recent histories of these populations. A method of estimating population substructure was used to identify populations in which genetic drift had played a particularly important role; populations in isolation or at small sizes for extended periods. Developed by Vitalis and Couvet (2001), this was performed through the parameter F, an averaged measure of probability of identity for each locus (equivalent to a within population F_{ST}). This has been used to measure population substructure and applied as a measure for genetic drift within a population (Sousa *et al.* In press). This method was implemented in the ESTIM 1.0 software. A method of detecting recent population size changes (bottlenecks or growth) developed by Cornuet and Luikart (1996), is implemented in the BOTTLENECK 1.2.02 program. Comparisons are made against the patterns of genetic diversity expected for a demographically stable population (null hypothesis), using two summary statistics of the allelic frequency spectrum, namely the number of alleles (n_A) and the expected heterozygosity (H_e). Simulations were performed to obtain the distribution of H_e conditional on n_A and the sample size for each population and locus. In order to test for significant deviations from the null hypothesis, 10000 simulated H_e values were compared to those obtained from the real dataset, using the Wilcoxon Sign Rank Test, under three mutational models: infinite allele model (I.A.M.), stepwise mutation model (S.M.M.), and a two-phase model (T.P.M.), in which, 30% of mutations were allowed to occur under a multi-step manner. A further investigation into the demographic dynamics of these breed populations was to see whether a relationship existed between the recorded numbers of breeding females and the diversity of the breed populations. To this end a regression was performed using the MINITAB 3.2 software.

3.3.4.3. Weitzman application

For a comparative conservation assessment, we applied a theoretical approach which was developed by Weitzman (1992) in which pairwise genetic distances between the populations are used to construct a maximum-likelihood diversity tree through the program WEITZPRO (Derban *et al.* 2002). In the Weitzman measure of diversity, populations are also ranked based on their contribution to diversity by excluding each population in turn from the original tree. This allows the relative genetic contributions to overall diversity of each breed population to be examined in the context of the total diversity among all breeds. Here we compare and contrast the findings of the Weitzman approach with other carefully selected methods in order to determine how the value assignments correspond to the genetic characteristics of the populations inferred from these methods.

3.4. Results

3.4.1. Genetic variability and population structure

The within-population heterozygote deficiency values (F_{IS}) are given in Appendix 3.1. Five of these populations exhibited positive F_{IS} values which diverged significantly from Hardy Weinberg Equilibrium. These populations were; Shetland, Belted Galloway, Traditional Hereford, Guernsey, and Berrenda Heterozygosity measures for the breed populations ranged widely with H_E values from 0.53 to 0.71 (Table 3.2). All F_{ST} values but one showed populations as being highly significantly differentiated from one another ($P<0.001$) (Appendix 3.2). The lowest pairwise F_{ST} was between the Guernsey and Beef Devon (F_{ST}: 0.03), but the populations were still significantly differentiated ($P<0.05$). The highest F_{ST} value was 0.37 between the White Park and Mainland Jersey breeds.

Table 3.2. European breeds sampled including; number of breeding females (N) (EFABIS, 2007; Alderson, 2007), number of genotyped individuals (n), expected, non-biased and observed heterozygosity (HE, Hn.b., HO)

Breed Population	N	n	H_E	$H_{n.b.}$	H_O
Guernsey	4,217	12	0.71	0.74	0.65
German Black Pied	1,130	19	0.70	0.72	0.72
Red Devon	3,332	20	0.69	0.71	0.71
Aberdeen Angus	20,500	20	0.69	0.70	0.67
Limousin	808,500	24	0.68	0.70	0.70
Dexter	9,000	154	0.68	0.69	0.70
Milking Devons	600	32	0.68	0.68	0.67
Angeln	220	24	0.67	0.68	0.66
Welsh Black	9,412	17	0.66	0.68	0.65
Jutland	120	17	0.66	0.68	0.65
Berrenda	870	31	0.65	0.66	0.56
Shorthorn	4,520	11	0.65	0.68	0.69
Red Poll	1,804	20	0.65	0.66	0.70
Lincoln Red	2,038	60	0.63	0.63	0.65
Island Jersey	3,400	25	0.62	0.63	0.62
Kerry (UK)	60	32	0.62	0.63	0.61
Sussex	2,467	50	0.62	0.62	0.61
Mainland Jersey	22,750	21	0.61	0.62	0.60
Hungarian Grey	1,520	15	0.59	0.61	0.57
Belted Galloway	2,608	15	0.58	0.60	0.56
British White	1,655	11	0.58	0.61	0.57
Shetland	664	31	0.57	0.58	0.54
N'dama	2,000,000	9	0.57	0.60	0.65
Traditional Hereford	616	19	0.56	0.58	0.52
Gloucester	500	14	0.55	0.57	0.52
Irish Moiled	308	20	0.53	0.54	0.54
White Park	727	33	0.53	0.54	0.52

Initial partitioning of the dataset gave the maximum likelihood value of *K* as two. This produced a group comprising 16 of the 20 British breeds and a group containing all six European breeds, the African N'Dama, and the remaining four British breed populations. Substructure within the British partition consisted of four levels of nested subdivisions resulting in 13 identifiable clusters of individuals. The European partition was divided through three nested subdivisions also resulting in 13 identifiable clusters. Breed structure is shown in Figure 3.1. Across the dataset six breeds failed to separate into distinct units (Aberdeen Angus, Beef Devon, Guernsey, Red Poll, Shorthorn, Welsh Black), and two breeds were further separated into two subpopulations each (Berrenda, White Park).

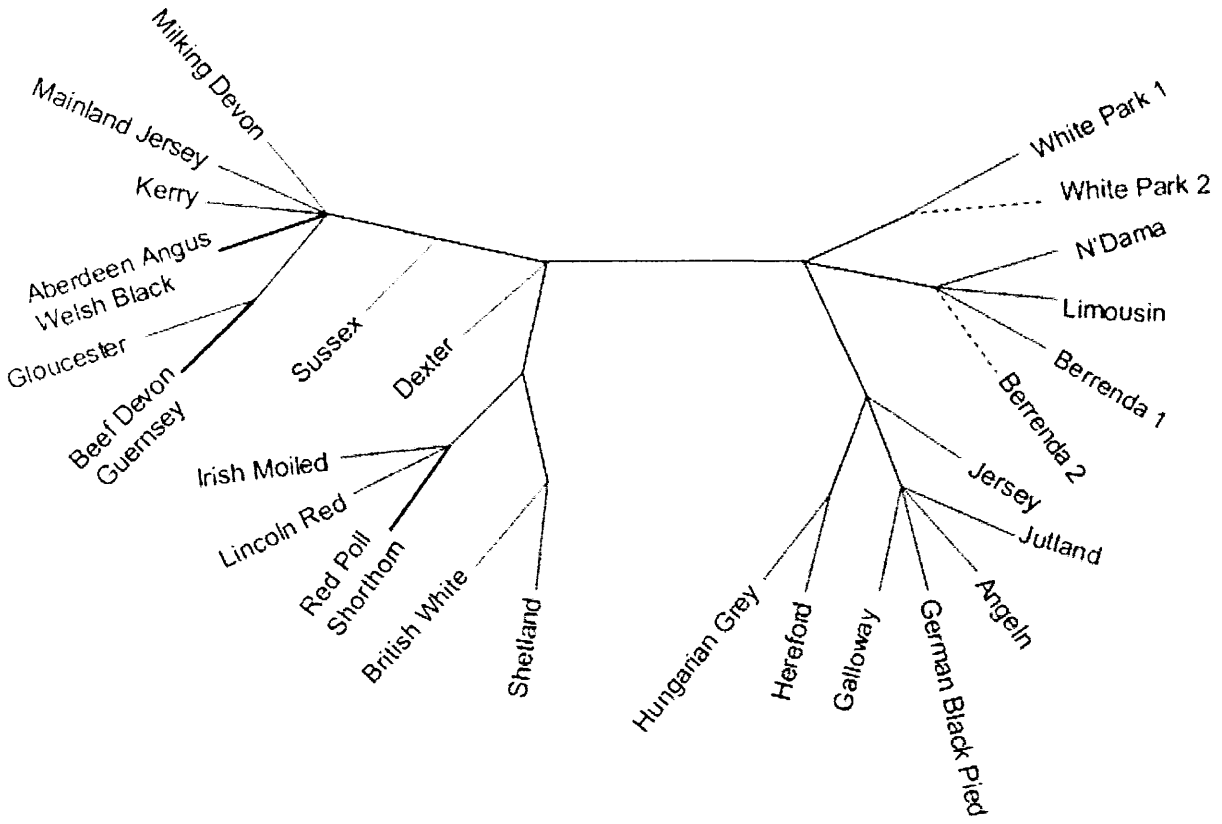


Figure 3.1. Population structure for the entire breed dataset as determined by the STRUCTURE software without using pre-defined populations. Nodes represent the points at which the data was identified as containing a split of individuals and new datasets were constructed from each subset and run through the same analysis process once more. Thickened and dotted lines represent sample groups that either failed to separate, or were the result of a sample group split respectively.

3.4.2. Geographical population clustering

Twenty-two population clusters were identified, (this result was identical when the analysis was performed excluding geographic information). Twenty of the clusters corresponded to breed sample groups. Two clusters contained multiple sample populations. One cluster comprised; Aberdeen Angus, Beef Devon, Guernsey, Shorthorn, and Welsh Black. The remaining cluster was composed of Angeln and German Black Pied. Although a breakdown of probabilities across the ten most visited clustering results is available, in this case no other results were found leaving the associated probability as one for a result of 22 clusters.

3.4.3. Migrant analysis

Under the analysis of first generation migrants 99 individuals (13%) were reclassified as being from a population other than that from which the animal truly originated. This reclassified proportion of each population varied from zero individuals in the British White, Traditional Hereford, Hungarian Grey, Island Jersey, and White Park up to half of the individuals for the Guernsey (Appendix 3.3). There was a strong geographic aspect to the genetic exchange identified here. Almost all of the migration was seen to be intra-regional within both Britain and mainland Europe. Of the 99 migrants identified, three were between British and European breeds contrasting markedly with the 10 migrants identified within the cluster of six European breeds and 85 between the greater 20 British breed group. A single migrant was also identified from the African breed N'Dama and assigned to the European Limousin.

3.4.4. Demographic history and conservation value

The F statistic calculated in ESTIM varied from low values in the Guernsey, and German Black Pied (0.07, 0.08) to high values in the White Park and Irish Moiled (0.32, 0.31) (Appendix 3.4). Under the infinite alleles model 23 of the populations showed significant heterozygote excess consistent with a population bottleneck determined using the Wilcoxon Signed Rank test using the program BOTTLENECK (Appendix 3.5). Nine of the breeds showed bottleneck signatures under TPM, and three assuming the SMM. One breed, the Kerry, demonstrated a heterozygote deficiency ($P<0.005$) under the SMM as well as a heterozygote excess under the IAM Only the Berrenda, Gloucester, and Jutland breeds showed a bottleneck signature under all three methods, another six displaying signatures in both IAM and TPM. However, a regression of estimated effective numbers of breeding females of the breed populations against H_E showed no significant relationship.

The maximum-likelihood phenogram resulting from the Weitzman approach (Figure 3.2) distinguished two initial partitions separating ten of the British breeds (Mainland Jersey, Red Poll, Milking Devon, Shorthorn, Lincoln Red, Aberdeen Angus, Guernsey, Beef Devon, Dexter, Welsh Black). The next major division partitioned the remaining British from the European breeds (and the N'Dama) with the exception of the Traditional Hereford and Island Jersey which clustered with the European breeds. Some level of agreement with the STRUCTURE method was shown in the distinction of the Belted Galloway and White Park breeds from the main British grouping as well

as the close association of the Aberdeen Angus, Beef Devon, and Guernsey breeds. The level of marginal diversity that would be lost by removal of any one sample population varied between 1.3 and 8.8 percent. The highest of these was found for the White Park breed and lowest for the Lincoln Red. A regression of this measure against the Geneclass2.0 average percentage proportional assignment within breeds was significant ($P<0.05$).

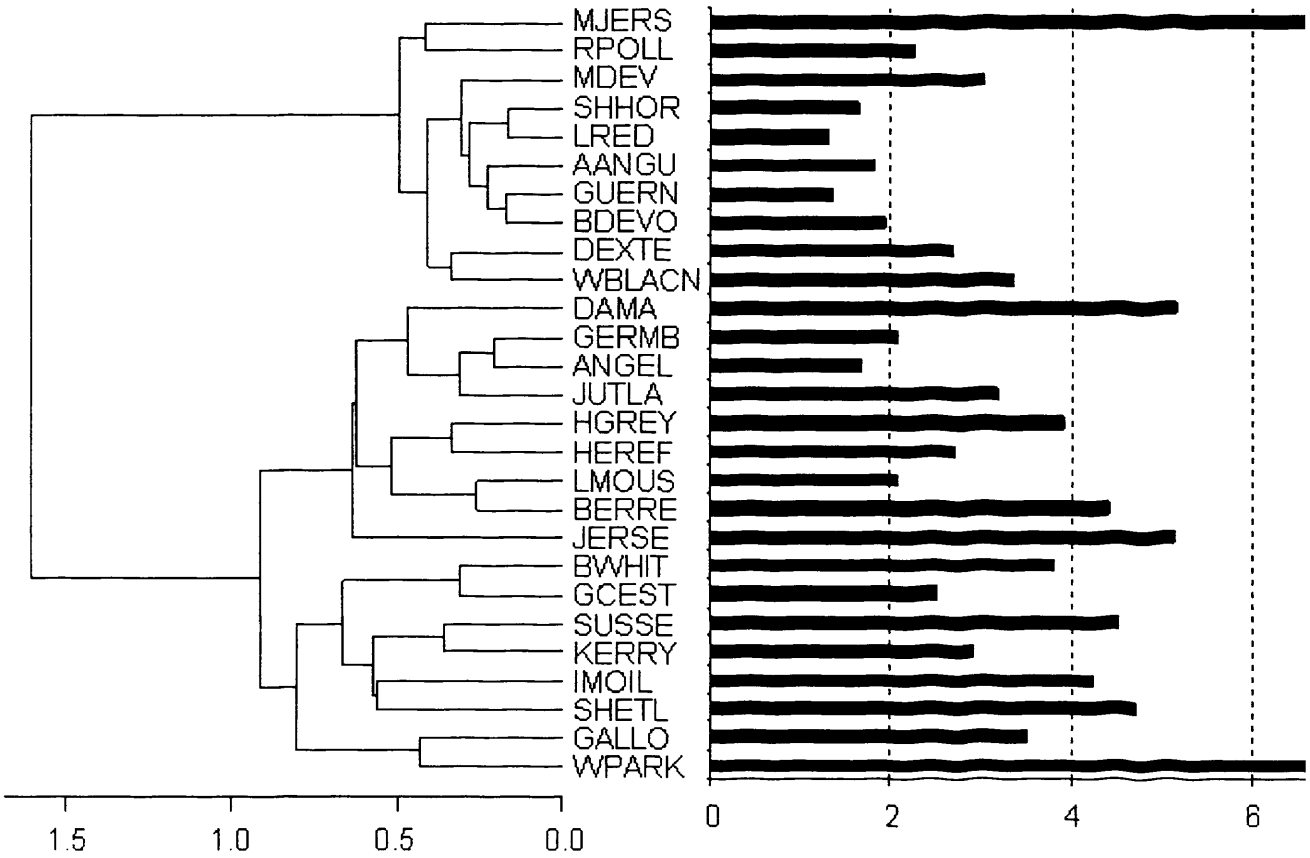


Figure 3.2. Maximum Likelihood tree and marginal diversity values (%) as determined by the WEITZPRO software (Derben et al. 2002).

3.5. Discussion

The genetic variability seen in this study is similar to the range shown for breeds over a number of recent studies based on commonly used microsatellites in European cattle populations (e.g. Kantanen *et al*, 2000; Wiener *et al*, 2004). This is true even for those populations whose effective size were estimated as below one hundred, such as the Kerry breed (H_E=0.62). The lowest breed variability here was displayed by the White Park (H_E=0.53), and this may reflect low population numbers in the 1970s (Alderson, 1997). The departure from Hardy Weinberg equilibrium seen in some breed populations may be the consequence of selective breeding, or may

simply reflect population substructure perhaps demonstrated here by the Berrenda breed, which is seen to split into separate breeding populations by coat colour.

Through the application of the clustering algorithms we were assessing how the genetic variation is distributed among the breed populations sampled. We applied the method of Pritchard *et al.* (2000) for non-spatial clustering despite the suggestion that it is unable to effectively discriminate individuals with closely related genotypes or low levels of genetic differentiation (Ibeagha-Awemu and Erhardt, 2005) as might be expected in domestic breeds. This follows its successful implementation in such studies as Negrini *et al.* (2007) and their application to AFLP data in European cattle. The method identified 21 of the 27 breed populations, finding further substructure in two of the breeds. The primary partition in the breed dataset suggested the presence of two groups of breeds; British and European. A UK-Europe split of this kind might be expected from interpretation of some previous studies (MacHugh *et al.* 1994; Blott *et al.* 1998) but is not supported by others (MacHugh *et al.* 1996; MacHugh *et al.* 1997). Re-analysis of this dataset including only the nine common loci across all breeds showed that the analysis was largely robust to the number of loci used, and that this was not influential in determining whether breeds could be individually distinguished by the clustering algorithm. The iterative application of this method allowed more fine scale elucidation of population relationships than a single application on the whole dataset, particularly given the number of samples involved. However, partitions resolved in this manner represent decreasing differentiation at each level therefore the importance of and confidence in each subsequent partition must be considered carefully. Consistent with the hypothesis that high levels of recent gene flow is likely to obscure ancestral breed relationships, we found that beyond the initial partition there appeared little reflection of co-ancestry i.e. clustering of breeds of the same colour-type or local origin. The presence of some British breeds in the European partition can be putitively explained by documented historical breed associations. There is a recognised association of the Hereford with European breeds, similarly found by Machugh *et al.* (1994), reflecting pre-founding input from northern Europe (Heath-Agnew, 1983). Likewise the Belted Galloway association in an otherwise German-only breed cluster can be traced to the northern European mainland, from its origin as a hybrid of the Galloway and the Dutch Belted (Wallace and Watson, 1923). A further breed relationship of note is that between the Island and mainland populations of the Jersey. It can be seen that the two populations are

markedly different from their separation into different clusters in both the STRUCTURE and Weitzman analyses and this is supported by a high FST of 0.25 between the populations. This is likely to be almost entirely due to the influence of migration on the mainland Jersey from the other mainland breed populations, the Island Jersey being completely isolated and closed to such influences.

The inclusion of spatial data added little further information to breed partitioning using the BAPS approach. This analysis suggested that all breeds were separate apart from two breed clusters; 1) five of the British breeds that failed to separate out individually from the STRUCTURE analysis (Aberdeen Angus, Shorthorn, Welsh Black, Beef Devon, Guernsey), and 2) the European mainland breeds of Angeln and German Black Pied. Other than the suggestion that there is a British core of similar populations and to a lesser extent a European one, the spatial approach here fails to reinforce the English Channel as an important barrier to gene flow. Migration events across the channel might involve more individuals at a time rather than the gradual exchange between neighbouring farms seen elsewhere, and due to the large width of the channel itself this might even reduce the effective geographic distance between the two landmasses. The analysis of first generation migrants however, contradicts this conclusion. The pattern seen is that of two regional migration maxima combined with very low inter-regional migration levels. Total levels of migrant assignment showed that 13% of individuals were reassigned to another population on the basis of being a first-generation migrant. Of these only 3% of all migrants were exchanged between the British and European regions. In a similar migrant analysis by Li *et al.* (2007) this type of regional association was not seen in cattle breeds across northern Eurasia and the Near Eastern Balkans. The greatest influence on migrant assignment in this case was concluded to be due to upgrading crosses for adaptive (through the Red Steppe) and commercial (through the Finnish Holstein-Friesian) purposes. Conversely, our study shows no single breed to be particularly influential in the other populations, perhaps due to the exclusion of the more commercial breeds. Our results fail to support any convergence of breeds by utility in these traditional populations by any of the methods used.

The history of a population, how it affects the distribution of extant genetic variation, and whether we can identify important demographic events is of particular value in traditional domestic populations. We would expect genetic drift to be far greater in

isolated populations that had been subjected to a period of low population size (Futuyma, 1998). Genetic drift has been shown to be the major force leading to the loss of alleles in cattle breeds (Simianer, 2005), as such we were interested in being able to identify populations in which drift may have been important. The Vitalis and Couvet (2001) gene identity measure F is a way of doing just this. The range of F measures found here are high and comparable to highly inbred captive populations of Goodeid fish (Bailey *et al.* 2007). The spectrum of values presented is particularly interesting in reference to the Island Jersey breed. The Island Jersey population has been closed to immigration for over 200 years and comprises a stable 3-4000 individuals which were shown not to be at risk from inbreeding (Chikhi *et al.* 2004). Using the Island Jersey as a comparison, those breed populations with appreciably higher F may be more at risk of the detrimental consequences of being small isolated populations. This would suggest the White Park and Irish Moiled breeds being potentially at risk of loss of diversity due to genetic drift and inbreeding and to a lesser extent also the Gloucester, Traditional Hereford, and Shetland breeds. Among British breeds there is a significant negative relationship ($P<0.05$) between the F value and the current effective size of breed populations which might suggest (naively assuming constant historical population sizes) that levels of gene flow between breed populations are comparable. However, breed populations are far from constant and this has profound consequences on estimations of demographic parameters. For this reason we looked for indications of severe demographic perturbation such as population bottlenecks or expansions in each population. We found evidence of bottlenecks in the majority of breeds under the IAM, the most highly significant of these also giving a bottleneck under the TPM and to a lesser extent the SMM. Microsatellites are better approximated as evolving under the SMM, and the TPM has been shown to be the best representative for microsatellite datasets (Di Rienzo *et al.* 1994). For these reasons we concentrate on bottlenecks that are present in those breeds showing a signature only under the SMM and TPM. It might be expected that the Irish Moiled and White Park have such high F measures due to bottleneck events, but this was not found to be the case. One potential explanation for this is that for enough time to have passed for drift to have become highly influential in population genetic structure the signature of any bottleneck that may have been present is likely to be obscured. This is likely to be the case in the White Park breed whose numbers were in the seventies in 1976 (Alderson, 1997) yet show no bottleneck signature here. Four breeds stand out from this bottleneck

analysis, the Berrenda, Gloucester, and Jutland breeds all showing the signature of a population bottleneck in TPM and SMM. and the Kerry which shows evidence of population expansion (heterozygote deficiency) under the SMM only. The Berrenda comprises two colour forms which, as identified in the clustering methods, actually appear genetically distinct and this was considered a potential reason for the significant bottleneck result. Re-analysis with the colour forms as separate sample populations showed that one of these subgroups still shows significant heterozygote excess (data not shown) inferring a genuine bottleneck event. The Gloucester and Jutland breeds show evidence of recent population bottlenecks and this is perhaps consistent with small current population sizes (~180 and ~500 respectively), but it is difficult to expand on this without further demographic information. The Kerry was the only population to show a heterozygosity deficit, implying a population expansion, and it is noted in Cornuet and Luikart (2001) that this deficit effect is more pronounced following a population size reduction. Despite the low current population numbers (~60) the UK population is known to have expanded from lower numbers in its recent history. This is similar to trends in Ireland where the population currently stands at over 450 breeding females having suffered a decline from around 1000 in the 1890's to below 300 in the mid to late 20th century (Olori and Wickham, 2004).

We were interested in a general comparative tool with which to summarise an *ad hoc* diversity value in each breed. The Weitzman-based genetic distance summary tool was first applied to cattle by Thaon d'Arnoldi *et al.* (1998) and we use it here fully aware of the criticism of distance methods and their failure to consider within-breed variation, an important part of practical breed management (Tapio *et al.* 2006). Distance-based approaches are still widely applied in cattle breed studies (Canon *et al.* 2001; Machado *et al.* 2003; Negrini *et al.* 2007) due to their providing a basic relative measure of distinctiveness, which can be informative when applied in conjunction with other methods (e.g Caballero and Toro, 2002). The Weitzman marginal diversity measure here produces a tree suggesting the existence of a subset of British breeds that separate from the rest even before the divergence of the European breeds. Subsequent to this initial divergence the method does recognise a British-European division as described previously in the other clustering methods. The high density of breed populations in Europe and particularly in the UK makes the process of characterising and explaining cryptic breed relationships problematic. Methods that rely on genetic distances are attractive due to their quick and easy

application relative to the more computationally intensive methods such as those employing Bayesian algorithms. We found that the breed clusters produced by the genetic distance based approach contrasted with the Bayesian-derived clusters in the initial separation of a subset of British breeds. Low accuracy of estimation from genetic distance approaches has been previously noted between individuals and populations (Cornuet *et al.* 1999), and has been shown to be sensitive to population size fluctuation (Estoup *et al.* 1998). This latter aspect is particularly pertinent to domestic livestock and may account for some of the differences between the outcomes of the different methods.

The calculation of a Weitzman marginal diversity contribution of each breed individually was the most problematic aspect of the application of the genetic distance approach here. Our findings contribute further speculation about the validity of this Weitzman approach within a species, adding to some previously noted criticisms, such as failure to account for effective population sizes (Caballero and Toro, 2002). Specifically, the Weitzman derived marginal diversity of each breed displayed a highly significant ($P<0.001$) inverse relationship to H_E (Appendix 3.6). This seems to reiterate the problem of the Weitzman approach promoting the maintenance of many inbred lines as suggested by Eding *et al.* (2002). In this scenario, having low genetic diversity effectively increases the genetic distance between a breed and the remaining breed cluster, somewhat counterintuitive to a conservation prioritisation scheme. To our knowledge, this inverse genetic diversity – marginal value relationship has not been reported previously. However, we calculated the same inverse relationship (for both H_E and allele number when compared against diversity loss by removal of individual breeds from a breed cluster) using data in published Weitzman analyses of pig (Laval *et al.* 2000) and goat (Glowatzki-Mullis *et al.* 2008) breeds. The low-diversity bias is illustrated particularly well in our data for the White Park breed, being the least genetically diverse of all of the breeds in this study yet having the highest marginal diversity. It can't be ruled out however, that genetic drift on this breed during a period of low population size caused the chance fixation of alleles absent or rare in the other breed populations resulting in very unique current diversity. The potential explanation of genetic drift for the attribution of marginal diversity may be argued. There is a highly significant positive relationship between the marginal diversity and Cornuet and Luikart (2001) F values (Appendix 3.7) Although this relationship is almost certainly affected by other

population parameters such as H_E and effective size. This suggests a propensity to overestimate the importance of populations in which drift has played a greater role for demographic reasons (small population size, repeated population bottlenecks etc.). Conclusions drawn about breed relationships from basic genetic distance data may be similarly affected and this sensitivity should be considered when applying any genetic distance based methodology. In cases of low population differentiation, such as in the context of domestic breeds, we would strongly recommend the avoidance of genetic distance methods for anything other than a superficial summary statistic used to guide subsequent analysis.

3.5.1. Implications for the conservation of genetic diversity

Genetic distinction was found between all of the traditional cattle breeds studied here, suggesting that they are useful as reservoirs of genetic diversity for the more commercial breeds. A logical consequence of this is that many of these breeds will contain unique genes or combinations of genes that need to be preserved. Almost all of the breeds here have high variability and it is possible to see genetic distinction between European and British breeds. The genetic legacy of European ancestry is identifiable in some British breed populations. As suggested in Giovambattista *et al.* (2001), it is the characteristic population structure in traditional breeds without intensive artificial selection that favours the maintenance of diversity, some of this adaptive. The consequences of maintenance of traditional breeds at small fluctuating population sizes do not seem to have detrimentally affected their diversity although inter-breed gene flow between some of the breeds has almost certainly contributed to this. The extreme genetic distinction between the Island and mainland Jersey populations is indicative of the consequences of maintaining separate herd populations. Equally, that populations can maintain their distinctiveness under the influence of the levels of gene flow suggested here reflects either a recent trend towards out-crossing, or simply the success of breed management in maintaining breed diversity. Recommendations from this work are that breeds demonstrating recent population bottleneck events, and in particular those at low population sizes such as the Jutland, should be regularly monitored for any signs of genetic deterioration.

Despite, and partially because of, the large numbers of extant traditional breed populations many are under threat from extinction and many of the breeds in this

study are listed as rare, potentially endangered, or endangered. Their continued existence as distinct breeds is still entirely dependent on their management. The generally high levels of variation found here suggests that these populations are likely to persist if the populations continue under careful management schemes. But the genetic status of breed populations is entirely dependent on the human interest in the breed, and it is imperative to promote this to maintain the diversity of breeds seen today.

3.6. Acknowledgements

This project was funded by DEFRA, the Zoological Society of London, the Rare Breeds Survival Trust, Dexter Cattle Society, and Cardiff University. Many thanks also go to all colleagues who provided samples.

3.7. References;

Alderson L (1997) A Breed of Distinction (breed history). Countrywide Livestock Ltd.

Alderson L (2007). A Review of Native Breeds of Cattle'. Report to Defra and RBST.

Bailey NW, Macias Garcia C, and Ritchie MG (2007) Beyond the point of no return? A comparison of genetic diversity in captive and wild populations of genetic diversity in captive and wild populations of two nearly extinct species of Goodeid fish reveals that one is inbred in the wild. Hered 98:360–367.

Bartosiewicz L (1997) The Hungarian Grey cattle: a traditional European breed. AGRI 21:49-60.

Beaumont M (1999) Detecting population expansion and decline using microsatellites. Genet 158:2013-2029.

Beja-Pereira A, Alexandrino P, Bessa I et al (2003) Genetic characterisation of Southwestern European bovine breeds: A historical and biogeographical reassessment with a set of 16 Microsatellites. J Hered 94(3):243-250.

Belkhir K, Borsa P, Chikhi L et al (2002) GENETIX 4.03, Logiciel Sous Windows^{TM} Pour la Génétique Des Populations. Laboratoire Génome Populations, Interactions. CNRS UMR 5000, Université de Montpellier II, Montpellier, France.

Bishop MD, Kappes SM, Keele JW et al (1994) A genetic linkage map for cattle. Genetics 136:619-639.

Blott SC, Williams JL, and Haley CS (1998) Genetic variation within the Hereford breed of cattle. Anim Genet 29: 202-211.

Brezinsky LS, Kemp J, and Teale AJ (1993) ILSTS006: a polymorphic bovine microsatellite. Anim Genet 24:73.

Bruford MW (2004) Conservation genetics of UK livestock: from molecules to management. In: Simm G, Villanueva B, Sinclair KD, and Townsend S (eds) Farm animal genetic resources, Brit Soc Anim Sci.

Caballero A, and Toro MA (2002) Analysis of genetic diversity for the management of conserved subdivided populations. Cons Genet 3:289-299.

Canon J, Alexandrino P, Bessa I et al (2001) Genetic diversity measures of local European beef cattle breeds for conservation purposes. Genet Sel Evol 33:311-332.

Chikhi L, Goossens B, Treanor A et al (2004) Population genetic structure of and inbreeding in an insular cattle breed, the jersey, and its implications for genetic resource management. Hered 92:396-401.

Corander J, Waldmann P, and Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. Genet 163:367-374.

Corander J, Waldmann P, Marttinen P, Sillanpää MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinform 20:2363-2369.

Corander J, Marttinen P, Mäntyniemi S (2006) Bayesian identification of stock mixtures from molecular marker data. Fish Bull 104:550–558.

Corander J, Siren J, and Arjas E (2007) Bayesian spatial modelling of genetic discontinuities in populations. Comp Stats (Prerelease).

Cornuet JM, and Luikart G (1997) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genet 144:2001-2014.

Cornuet JM, Piry S, Luikart G et al (1999) New Methods Employing Multilocus Genotypes to Select or Exclude Populations as Origins of Individuals. Genet 153:1989-2000.

Derban S, Foulley J-L, and Ollivier L (2002) WEITZPRO: a software for analysing genetic diversity. INRA, Paris.

Di Rienzo A, Peterson AC, Garza JC et al (1994) Mutational processes of simple sequence repeat loci in human populations. Proc Natl Acad Sci U.S.A. 91:3166–3170.

Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. Mol Ecol 11(12):2571-81.

Eding H, Crooijmans PMA, Groenne MAM et al (2002) Assessing the contribution of breeds to genetic diversity in conservation schemes. Genet Sel Evol 34:613–633.

EFABIS website. European Farm Animal Breed Information System. http://efabis.tzv.fal.de

Estoup A, Rousset F, Michalakis Y 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.

Evanno G, Regnaut S, and Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol 14:2611-2620.

Fabuel E, Barragan C, Silio L et al (2004) Analysis of genetic diversity and conservation priorities in Iberian pigs based on microsatellite markers. Hered 93:104–113.

alush D, Stephens M, and Pritchard JK (2003) Inference of population tructure using multilocus genotype data: linked loci and correlated allele equencies. Genet 164:1567–1587.

utuyma DJ (1998) Evolutionary Biology, 3rd ed, Sunderland: Sinauer.
eorges M, and Massey JM (1992) Polymorphic DNA markers in Bovidae. atent WO 92/13102.

iovambattista G, Ripoli MV, Peral-Garcia P et al (2001) Indigenous domestic reeds as reservoirs of genetic diversity: the argentinian creole cattle. Anim enet 32(5):240-248.

lowatzki-Mullis M-L, Muntwyler J, Baumle E et al (2008) Genetic diversity easures of Swiss goat breeds as decision-making support for conservation olicy. Small Ruminant Res 74:202-211.

oossens B, Waits LP, and Taberlet P (1998) Plucked hair samples as a ource of DNA: reliability of dinucleotide microsatellite genotyping, Molr Ecol :1237-1241.

eath-Agnew E (1983) A history of Hereford cattle and their breeders, London. uckworth.

antanen J, Olsaker I, Holm L-E, et al (2000) Genetic diversity and population ructure of 20 North European cattle breeds. Am Genet Ass 91:446-457.
aukinen J, and Varvio SL (1993) Eight polymorphic bovine microsatellites. nim Genet 24:148.

aval G, Iannuccelli N, Legault C, et al (2000) Genetic diversity of eleven uropean pig breeds. Genet Sel Evol 32:187– 203.

M-H, Tapio I, Vilkki J et al (2007) The genetic structure of cattle populations *Bos taurus*) in northern Eurasia and the neighbouring Near Eastern regions: nplications for breeding strategies and conservation, Mol Ecol 16:3839-3853.

Machado MA, Schuster I, Martinez ML et al (2003) Genetic diversity of four cattle breeds using microsatellite markers. R Bras Zootec 32(1):93-98.

Machugh DE, Loftus RT, Bradley DG et al (1994) Microsatellite DNA variation within and among European cattle breeds. Procl R Soc Lon B 256:25-31.

Machugh DE, Loftus RT, Cunningham P, et al (1996) Genetic structure of seven European cattle breeds as assessed using 20 microsatellite markers. Anim Genet 29:33-340.

Machugh DE, Shriver MD, Loftus RT et al (1997) Microsatellite DNA variation and the Evolution, Domestication and Phylogeography of Taurine and Zebu Cattle (*Bos Taurus* and *Bos indicus*). Genet 146:1071-1086.

Machugh DE, Loftus RT, Cunningham P et al (1998) Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. Anim Genet 29:333-340.

MacNeil MD, Cronin MA, Blackburn HD et al (2007) Genetic relationships between feral cattle from Chirikof Island, Alaska and other breeds. Anim Genet 38:193-197.

Negrini R, Nijman IJ, Milanesi E et al (2007) Differentiation in European cattle by AFLP fingerprinting. Anim Genet 38:60-66.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583-590.

Negrini R, Nijman IJ, Milanesi E et al (2007) Differentiation of European cattle by AFLP fingerprinting. Anim Genet 38:60-66.

Ollivier L, and Foulley J-L (2005) Aggregate diversity: new approach combining within- and between-breed genetic diversity. Livestock Production Sci 95:247–254.

Olori VE, and Wickham B (2004) Strategies for the conservation of the indigenous Kerry cattle of Ireland. AGRI 35, FAO Rome.

Piry S, Alapetite A, Cornuet J-M et al (2004) Geneclass2: A software for genetic assignment and first-generation migrant detection. J Hered 95(6):536-539.

Pritchard JK, Stephens M, and Donnelly P (2000) Inference of population structure using multilocus genotype data. Genet 155:945-959.

Rannala B, and Mountain JL (1997) Detecting immigration by using multilocus genotypes. Procl Nat Acad Sci USA 94:9197-9201.

Sheppy AJ (1998) Bloodlines, breed structure, and the influence of artificial insemination in Dexter cattle. Proceedings of the first world congress on Dexter cattle, Dexter cattle society, UK.

Simm G, Henson EL, Villanueva B et al (2002) A UK conservation success story: Longhorn cattle, a case study. Nottingham University Press.

Simianer H (2005) Using expected allele number as objective function to design between and within breed conservation of farm animal biodiversity. J Anim Breed Genet 122:177-187.

Solinas Toldo S, Fries R, Steffen P et al. (1993). Physically mapped, cosmid-derived microsatellite markers as anchor loci on bovine chromosomes. Mam Genome 4:720-727.

Sousa V, Penha F, Collares-Pereira MJ et al (In press) Do genetic data confirm the population collapse of a critically endangered fish? The case of the endemic freshwater cyprinid Chondrostoma lusitanicum.

Steffen P, Eggen A, Dietz AB et al (1993) Isolation and mapping of polymorphic microsatellites in cattle. Anim Genet 24:121-124.

Tapio I, Varv S, Bennewitz J et al (2006) Prioritisation of northern European cattle breeds based on analysis of microsatellite data. Cons Biol 20(6):1768-1779.

Thaon d'Arnoldi C, Foulley JL, Ollivier L (1998) An overview of the Weitzman approach to diversity. Genet Sel Evol 30:149-161.

Thieven U, Solinas-Toldo S, Friedl R et al (1997) Polymorphic CA-microsatellites for the integration of the bovine genetic and physical map. Mam Genome 8:52-55.

Vaiman D, Osta D, Mercier D et al. (1992) Characterisation of five new bovine microsatellite repeats. Anim Genet 23:537.

Vaiman D, Mercier D, Moazami-Goudarzi K et al (1994) A set of 99 cattle microsatellites: characterization, synteny mapping, and polymorphism. Mam Genome 5:288-297.

Vigilant L (1999) An evaluation of techniques for the extraction and amplification of DNA from naturally shed hairs. Biol Chem, 380:1329-1331.

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

Wallace R, and Watson JAS (1923) Farm Livestock of Great Britain, Oliver & Boyd, Edinburgh.

Walsh PS, Metzger DA, and Higuchi R (1991) Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. Biotechniques 10:506-513.

Weir BS, and Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evol 38:1358–1370.

Weitzman ML (1993) What to preserve? An application of diversity theory to crane conservation. Q J Econom.

Wiener P, Burton D, and Williams JL (2004) Breed relationships and definition in British cattle: a genetic analysis. Hered 93:597-602.

Wilson GA, and Rannala B (2003) Bayesian inference of recent migration rates using multilocus genotypes. Genet 163:1177–1191.

Chapter 4.

Paper submitted to *Animal Genetics*

The population genetic effects of ancestry and admixture in a subdivided cattle breed

T.C. BRAY[a], L. CHIKHI[b,c], A.J. SHEPPY[d] and M. W. BRUFORD[a],

[a] Cardiff School of Biosciences, Cardiff University, P.O. Box 915, Cardiff CF10 3TL, UK.

[b] UMR 5174 CNRS/UPS Evolution et Diversité Biologique, Université Paul Sabatier, 118 Route de Narbonne, Bât. 4R3 b2, 31062 Toulouse cédex 09, France

[c] Population and Conservation Genetics Group, Instituto Gulbenkian de Ciência, Rua da Quinta Grande, Nº6, P-2780-156 Oeiras, Portugal

[d]Cobthorn Trust, Congresbury, Bristol BS49 5JA

Correspondence:
M. W. Bruford, Fax: +44 (0)29 20 87 43 16; E-mail: BrufordMW@Cardiff.ac.uk

Running title: Ancestry and admixture in a cattle breed

Keywords: genetic diversity, admixture,

4.1. Abstract

The genetic structure of the Dexter, a minority cattle breed with complex demographic history was investigated using microsatellite markers and a range of statistical approaches designed to detect both admixture and genetic drift. Modern representatives of two putative ancestral populations, the Devon and Kerry, together with the different populations of the Dexter, which have experienced different demographic histories, were analysed. Breed units showed comparatively high levels of genetic variability (H_E=0.63-0.68), however, distinct genetic subgroups were detected within the Dexter which could be attributed to known demographic events. Much lower diversity was identified in three small, isolated Dexter populations (H_E=0.52-0.55) and higher differentiation (F_{ST} >0.13), was found. For one of these populations, where strong selection has taken place, we also found evidence of a demographic bottleneck. Three methods for quantifying breed admixture were applied and substantial method-based variation in estimates for the genetic contribution of the two proposed ancestral populations for each subdivision of the Dexter was found. Results were consistent only in the case of a group consisting of selected Traditional Dexter animals, where the ancestor of the modern Kerry breed was also determined as the greater parental contributor to the Dexter. This inconsistent estimation of admixture proportions among methods highlights the potentially confounding role of genetic drift in shaping small population structure and the consequences on accurately describing population histories from contemporary genetic data.

4.2. Introduction

In domestic animals, among-population demographic relationships can be linked to breed origins and distribution (see Caramelli, 2006 for a review) as well as to breed management practices such as upgrading and selection (Takeshima, *et al.* 2003). There are many documented accounts of such demographic events shaping well-known breed populations (Bartosiewicz, 1997). Isolation of breed populations into herd-book schemes is a relatively recent development (Taberlet *et al*, 2008), but it can rarely be assumed that gene exchange among breeds has ceased entirely. There remain relatively few completely closed populations, the island Jersey cattle currently being one (Chikhi *et al*, 2004). Conversely the management schemes of many traditional breed populations have involved an outbreeding regime to improve production value of stock, such as polling in Hereford cattle (Heath-Agnew, 1983). It remains largely untested if the genetic signature of such admixture events can be similarly quantified if undocumented, or not linked to an obvious trait, or how the relationships between a population and its ancestors are affected.

The Dexter cattle breed provides a model example of the population dynamics found in a number of minority domestic breeds. The Dexter itself is thought to have been formed in Ireland from a Celtic black cattle population, the most direct modern descendent of which is thought to be the Kerry (Wilson, 1909a). The Dexter is thought to have become gradually demographically separated from the Kerry as it became established in England, Wales and Scotland and is also reputed to have received genetic input from the old Devon breed shortly prior to the creation of the Dexter herd book in 1890 (Wilson, 1909b) (Figure 4.1). Since the creation of the Dexter herd book all three breeds have been maintained independently. As is typical of many livestock breeds, the modern Dexter is divided into different populations or herds that may be separated from each other and have had their own, sometimes independent, demographic history.

This study investigates the origin and relationships of the different subgroups within the Dexter, by (i) assessing patterns of molecular diversity and differentiation and (ii) estimating the relative contributions of the two breeds thought to be ancestral to the Dexter breed using genetic data from their modern descendents. We predicted that there would be a number of distinct partitions within the Dexter due to the

demographic isolation of some herds. Genetic distinctiveness would be expected to be greater in geographically isolated populations and for those under strong selection. We expected greater genetic drift in isolated populations due to lower effective sizes, and to find evidence of population bottlenecks and lower genetic diversity. The Dexter as a whole could, however, be expected to retain high overall levels of genetic variation since outcrossing is implicated in the expansion of the breed in the 1960s and 1970s. The Kerry breed was predicted to be the most similar to the Dexter's ancestor due to its close historical associations with the Dexter.

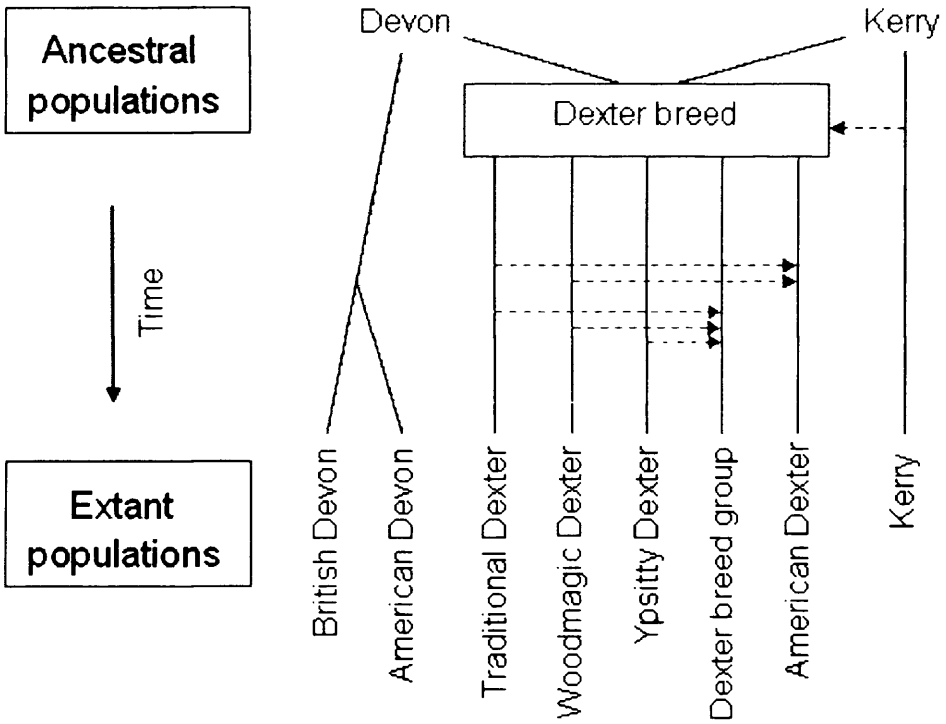


Figure 4.1. Schematic representation of a putative history for the traditional Dexter breed. Dotted horizontal lines represent gene-flow or introgression that is suspected to have taken place between populations. The order and length of these arrows is for illustration only and does not correspond to a chronological sequence.

4.3. Materials and Methods

4.3.1. Data collection

We sampled both cryogenically frozen semen and plucked hair from the Dexter, the Kerry and the Red Devon (Table 4.1). Both extant Devon types were sampled, the milking Devon which is present only in North America today, and the British beef Devon. Individuals representing the oldest form of the Dexter (not known to have

been upgraded in herdbook records) were collected to form a sample of 'Traditional' Dexters. The oldest single herd represented in the breed today was sampled as an example of a traditional yet demographically isolated population ('Ypsitty'). American Dexters were sampled to represent a population established through exports prior to 1914. An additional isolated population receiving no genes from the rest of the breed for over 25 years and subject to strong selection was also analysed ('Woodmagic'). A larger group representing the wider Dexter breed as it exists today was also sampled: this group could potentially have been introgressed by a number of other breeds during upgrading in the 1970's.

Table 4.1. Sample group sizes and description of origins of DNA from either plucked hair or cryogenically frozen semen.

Sample population	No. individuals	origin of DNA
American Dexter	13	Plucked hair
Traditional Dexter	12	2 plucked hair/ 10 semen
Woodmagic Dexter	19	15 plucked hair/ 4 semen
Ypsitty Dexter	13	12 plucked hair/ 1 semen
Dexter breed population	91	71 plucked hair/ 20 semen
Beef Devon	20	8 plucked hair/ 12 semen
Milking Devon	32	Plucked hair
Kerry	20	10 plucked hair/ 10 semen

4.3.2. DNA extraction

DNA was extracted from plucked hair samples using either chelex 100 as described in Walsh *et al.* (1991) with modifications in Goossens *et al.* (1998) or with a PCR buffer-based method described in Vigilant (1999). Semen samples were extracted using a modification of the Qiagen Dneasy tissue extraction method according to manufacturers' instructions. From the extractions, 1.5µl was added as template in each PCR reaction.

4.3.3. Genotyping

Twenty two microsatellite markers used in this study were taken from the UN Food and Agriculture Organisation (FAO) list of cattle markers. The microsatellites were; hel1, hel5, hel9, hel13 (Kaukinen & Varvio, 1993), ilsts005 (Brezinsky *et al.* 1993a),

csrm60 (Moore *et al.* 1994), eth3, eth10 (Solinas-Toldo *et al.* 1993), tgla 227, tgla122, tgla126 (Georges & Massey, 1992), sps115 (Moore & Byrne, 1993) , inra032, inra037, inra063 (Vaiman *et al.* 1994), eth152, eth225 (Steffen *et al.* 1993), bm1818, bm1824 (Bishop *et al.* 1994), ilsts006 (Brezinsky *et al.* 1993b), haut27 (Thieven *et al.* 1997), and cssm66 (Barendse *et al.* 1994). These were amplified in 4 multiplex reactions using the Qiagen multiplex kit according to manufacturers' instructions. Amplification was carried out in 6µl reactions (1xQIAGEN PCR Multiplex Master Mix (3mM $MgCl_2$), 0.2µM each primer). Thermocycling conditions were as follows: initial denaturation at 95' for 15' followed by 35 cycles of 60s at 94°C, annealing for 90s at 55°C then extension for 60s at 72°C, with a final extension for 10' at 60°C. All PCR products were electrophoresed using an ABI 3100 semi-automated DNA analyser. Gels were analysed using Genescan analysis 2.0™, Genotyper 1.1™ and Genemapper™ software.

4.3.4. Genetic variability, population size change and structure

General genetic diversity estimates were made, such as observed and expected heterozygosity under random mating (H_E) and Wright's F statistics, following Weir and Cockerham (1984). The departure of these statistics from the null hypothesis, which was no genetic differentiation for F_{ST} and Hardy-Weinberg equilibrium for F_{IS} and F_{IT}, was tested over 10^4 permutations as implemented using GENETIX 4.03 (Belkhir *et al.* 2002). Gene identity was estimated to further explore population substructure (Vitalis and Couvet, 2001) using ESTIM 1.0. The parameter F estimates average per locus gene identities as a measure of within-population genetic drift. In finite populations variation is continually lost at a rate depending on population size and coalescence times scale inversely with effective population size. ESTIM 1.0 provides simultaneous estimates of both the effective population size and the migration rate using a method-of-moments approach (Vitalis & Couvet, 1996).

A method of detecting recent population size changes (bottlenecks or growth; Cornuet & Luikart 1996) implemented in the BOTTLENECK(version 1.2.02) program, was applied. We used 10^4 simulated H_E values for comparison to those obtained from the real dataset, using the Wilcoxon Sign Rank Test, under three mutational models: infinite allele model (I.A.M.), stepwise mutation model (S.M.M.), and a two-phase model (T.P.M.), in which, 30% of mutations were allowed to occur under a multi-step manner. This analysis was performed to determine whether it was possible

to detect the bottleneck which is thought to have taken place in the recent history of the Dexter Woodmagic population. One question was whether other events such as admixture could have obscured any signal of this event.

To investigate population structure we used the Bayesian model-based clustering approach developed by Pritchard *et al.* (2000). This method, further improved by Falush *et al.* (2003) is implemented in STRUCTURE 2.1 and groups individuals into *K* homogeneous clusters (populations). We applied the Evanno *et al.* (2005) approach of selection of the *K* value corresponding to the mode of the ΔK distribution. We set *K* to vary between 1 and 8, and for each *K* value we performed 20 simulations with different starting points, having a 10^5 burn-in period followed by 10^5 steps (we tested different run lengths ranging from 10^4 to 10^6 and found that convergence was achieved after 10^5 steps). For each *K* value we calculated the average and standard deviation of the 'log estimated likelihood' [L(*K*)] across the 20 runs. In order to determine whether each cluster was itself subdivided into smaller and less differentiated units, the identified clusters were reanalysed independently. This was repeated for each subgroup until the most likely *K* was shown to be 1. The program was run under the admixture model, considering independent allele frequencies.

4.3.5. Investigating admixture

We aimed to determine the relative contributions of the two breeds that are thought to be descendent from those also ancestral to the Dexter, the Devon and Kerry. The methods described below were shown by Choisy *et al.* (2004) to perform well under different conditions. For the purpose of the admixture analyses we pooled the two Devon lineages to form a single parental population. The three admixture methods are implemented in three software packages ADMIX2.0, LEADMIX and LEA.

4.3.5.1. ADMIX2.0

The ADMIX2 (Dupanloup & Bertorelle, 2001) method uses a simple moment-based estimator m_Y, the calculation of which is based on coalescence times between pairs of genes sampled within and among populations. The method assumes a simple model where two or more parental populations diverge from an older ancestor, these parental populations then meet during an admixture event to create a third 'hybrid' population. All populations then drift from each other without exchanging genes. This

method is the only one (among those used here) that accounts for mutations, however we applied the basic method option and did not therefore incorporate inter-allelic distances into the admixture calculation.

4.3.5.2. LEADMIX

Based on the same demographic model as in ADMIX2, Wang (2003) developed a maximum likelihood method that also takes into account the genetic differentiation between parental populations in the admixture calculation. In this way the method aims to avoid falsely assuming independent allele frequency distributions of the parental populations and any resultant bias in the admixture calculation.

4.3.5.3. LEA

This method is based on a different demographic model where the two parental populations are assumed to be at demographic equilibrium and the allele frequencies prior to admixture are sampled from independent uninformative prior probability distributions (see Chikhi *et al.* 2001 for details). The difference from the previous methods is that this approach is a full-likelihood Bayesian method and hence provides posterior distributions for the parameters of the model, rather than point estimators. It also accounts for genetic drift, which is estimated through the scaled parameters $t_1=T/N_1$, $t_2=T/N_2$ $t_h=T/N_h$, where T is the time since the admixture event (in generations), and N_i is the effective size of population i (with i=1,2, h). A full-likelihood Bayesian approach assumes that all relevant information is contained in the posterior distribution. For each population either two or three independent runs were performed, using different starting values in the parameter space, in order to determine whether equilibrium had been reached. Each run had at least 500,000 steps together with a thinning interval of five. Also, a few longer runs (up to $1x10^6$ steps) were used to check for convergence.

4.4. Results and Discussion

Table 4.2 shows that genetic diversity varied widely across the populations sampled. H_E and F values were, as expected, highly negatively correlated (Pearson's r= -0.993). Diversity was lowest in the Dexter subpopulation groups that have been isolated for longer periods from the rest of the breed (Woodmagic H_E=0.52, and Ypsitty H_E=0.54, & American Dexter H_E=0.54) compared with H_E 0.63-0.68 for the remaining populations. Diversity estimates in the Dexter, Devon and Kerry breed

populations was comparable to the range of estimates seen in a number of contemporary cattle breed studies (Peelman *et al.* 1998; Loftus *et al.* 1999; Martin-Burriel *et al.* 1999; Chikhi *et al.* 2004) (Table 4.3). The Traditional Dexter and main Dexter breed populations demonstrated extremely similar levels of diversity to a previous study of Wiener *et al.* (2004) on the Dexter breed using many of the same marker loci.

Table 4.2. Observed and expected heterozygosity, mean number of alleles per locus and average one-locus identity probabilities (F) for all populations

Population	H_O	H_E	Allele/locus	F
Dexter breed population	0.71	0.68	6.41	0.05
Kerry	0.68	0.64	5.00	0.08
Traditional Dexter	0.73	0.63	4.64	0.07
Devon	0.61	0.63	5.64	0.11
American Dexter	0.56	0.55	4.18	0.20
Ypsitty Dexter	0.57	0.54	3.68	0.20
Woodmagic Dexter	0.58	0.52	3.41	0.26

Table 4.3. A comparison of previously determined breed diversities where marker loci have overlapped with this study

Overlapping loci / Total loci	Breeds (H_E)	Reference
9/23	Belgian Blue (0.65), Holstein Friesian (0.69), East Flemish (0.69), Red Pied (0.71)	Peelman *et a* 1998
12/30	Menorquina (0.56), Fighting Bull (0.59), Pyrenean (0.62), Asturian Mountain (0.67), Nordwest Brown group (0.67), Asturian Lowland (0.68)	Martin-Burrie *et al* 1999
2/12	Hereford (0.40), Jersey (0.41), Angus (0.42), Simmental (0.43), Charolais (0.46), Friesian (0.49)	MacHugh *al.* 1994
14/20	N'Dama (0.54), Hungarian Grey (0.62), Jersey (0.63), Ongole (0.64), Nellore (0.65), Charolais (0.66), Damascus (0.74), Turkish Grey (0.76), Anatolian Black (0.78), South Anatolian Red (0.78), East Anatolian Red (0.78), Egypt (0.78), Iraqi (0.78), Kurdi (0.79)	Loftus *et* 1999
8/12	Jersey (0.64)	Chikhi *et* 2004
21/30	A. Angus (0.61) Ayrshire (0.68) Dexter (0.65) Friesian (0.67) Guernsey (0.63) Hereford (0.63) Highland (0.56) Jersey (0.60)	Wiener *et* 2004

Across the three bottleneck detection methods significant signals of demographic bottlenecks were found in all groups except American Dexter (IAM), a bottleneck was again found in Woodmagic Dexter (TPM), and signals of expansion were found in American Dexter, Kerry, and Devon (SMM). The Kerry and Devon breeds simultaneously showed the signal of heterozygote excess under the IAM and deficiency under the SMM. The only consistent bottleneck signal present was detected in the Woodmagic herd which is known to have been founded by only five individuals (four females and one male) (Rutherford, 2005).

Despite the high degree of similarity between the Traditional Dexter and the Dexter breed group (theta <0.01), all other pairwise comparisons showed significant

differentiation (Table 4.4). As expected, the most divergent populations were also those that had the lowest levels of within-population variability. The high F_{ST} values between the Woodmagic Dexter and all other samples is presumed to be a result of small founder number, extended isolation from the UK herd with a small population size, and the application of strong selection (Rutherford, 2005). The level of divergence of the Woodmagic from the rest of the Dexter is at least at the level of among-breed values seen for the other samples here, and for values observed among other commonly studied European breeds (Canon *et al.* 2001). Similarly, the Ypsitty population represents the oldest Dexter herd, older extant animals represent the closed nature of the population through unique allele spectra. The American Dexter underwent a bottleneck that took place when it was founded, and showed lower F_{ST} values compared with both the Dexter breed and Traditional Dexter groups (0.08, 0.10 respectively) and the other breeds (Kerry-0.15, Devon-0.09), than the Woodmagic herd. This lower degree of differentiation from UK populations may be explained by a combination of cross-Atlantic gene flow, (at least eleven animals are documented as receiving export licenses for the United States between 1951-1988 (Dexter Cattle Society Herdbook; 1951, 1988)), chance retention of similar allelic spectra over time, or that the bottleneck was not as strong as originally believed. This latter suggestion is supported by the failure to detect the genetic signature of a bottleneck in this sample. Other breeds such as the Hereford have demonstrated significant genetic differentiation across multiple countries (Blott *et al.* 1998), the American Hereford cattle population itself experienced post-foundation inbreeding levels of up to 11.5% (Cleveland *et al.* 2005). However, uniquely within this dataset, the American sample analysed represents a low proportion of the extant total (approximately 6,000) and therefore inference may be improved by using an extended dataset. The rapid change in gene frequencies in the three Dexter populations is similarly supported by the high gene identity estimates shown for Woodmagic, Ypsitty, and American Dexter samples, contrasting with the low values seen in the Dexter breed group.

Table 4.4. Weir and Cockerham's theta (F_{ST}) values for all populations and their significances (P<0.05*,0.01**,0.005***, non-significant^{NS})

	Dexter breed	Kerry	Original population Dexter	Devon	Woodmagic Dexter	American Dexter	Ypsitty Dexter
Dexter breed		0.085	0.002	0.102	0.129	0.085	0.054
Kerry	***		0.078	0.094	0.214	0.146	0.152
Traditional Dexter	NS	***		0.12	0.158	0.103	0.067
Devon	***	***	***		0.202	0.094	0.157
Woodmagic Dexter	***	***	***	***		0.197	0.207
American Dexter	***	***	*	***	***		0.157
Ypsitty Dexter	***	***	***	***	***	*	

A summary of the hierarchical population clustering obtained using STRUCTURE can be seen in Figure 4.2, ΔK support for each level of clustering is shown in Table 4.5. Based on a simple interpretation of F_{ST} values one may have predicted that the clusters that would split first, as identified using STRUCTURE, would be the Woodmagic and Ypsitty Dexters, but this was not the case. STRUCTURE analysis identified the Devon as the most divergent set of genotypes with very little overlap with the Dexter cluster. The Kerry breed was intermediate, individuals largely belonging to the Devon cluster but with an average membership of some ~25% to the Dexter cluster. This, together with the admixture analysis suggests that the Kerry is closer to the Dexter than the Devon and hence that it probably contributed most to its genome, as would also be expected from the breed history and geography (these breeds both originate from Ireland). However, whilst the structure algorithm is a useful tool for finding hidden genetic substructure, it is important to stress that it does not account for the known demographic history of the breeds. The method of Pritchard *et al.* (2000) is a clustering method unlike the admixture methods used here that aim to account, at least to some extent for this unknown history, and they suggest that the picture is less clear cut.

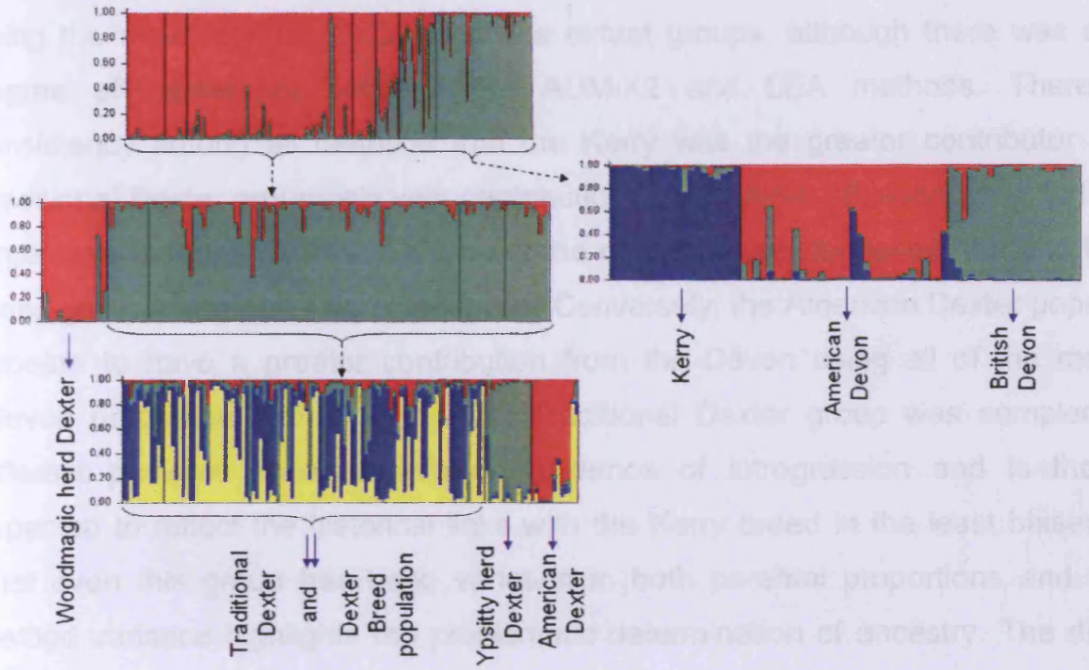


Figure 4.2. Hierarchical analysis of population clustering, determined through STRUCTURE. The data were split into K groups of populations using the Evanno et al. (2005) method. Members of each identified cluster were separated into individual datasets and the process repeated until terminal groups (indicated by solid arrows) were formed in which no further clusters were identified.

Table 4.5. Relative log likelihood and delta K values for population partitioning in STRUCTURE.

	Negative log likelihood	ΔK
Initial split	-12500	550
Devon-Kerry split	-3900	170
Woodmagic split	-7800	100
Second Dexter split	-6800	10

4.4.1. Parental proportions and admixture

The representation of the population relationships and potential genetic exchange (Figure 4.1) indicates how the admixture hypotheses were established to estimate the relative contributions of the ancestral populations to the Dexter breed populations (Table 4.6). The admixture methods did not converge on one ancestral population

being the major contributor to all of the extant groups, although there was a high degree of agreement between the ADMIX2 and LEA methods. There was consistency among all methods that the Kerry was the greater contributor in the Traditional Dexter group only with contributions from 0.63-0.86, even considering the large error margins (0.10-0.32). One of the methods also suggested that the Dexter breed group has greater Kerry influence. Conversely, the American Dexter population appears to have a greater contribution from the Devon using all of the methods (Devon proportion = 0.75-0.78). The Traditional Dexter group was sampled from different pedigree lineages with no evidence of introgression and is therefore expected to reflect the historical links with the Kerry breed in the least biased way. That even this group has large variation in both parental proportions and within-method variance highlights the problematic determination of ancestry. The differing amount of genetic drift that is likely to have taken place in the history of the different groups makes these results more difficult to interpret. Discordance between the contributions calculated using each program is likely to be at least in part due to their different assumptions. Whilst LEADMIX and LEA account for drift, ADMIX2.0 does not, and also assumes that mutation plays a role in population divergence. For domestic breeds in general and for the Dexter in particular, due to the short timescale in generations and the relatively small effective population sizes involved, drift is more likely to have generated the observed patterns of differentiation than mutation.

Table 4.6. Relative parental contributions from Kerry (variance in parentheses) for each Dexter population as determined using the programs, ADMIX2.0, LEADMIX, and LEA

	Hybrid population				
Method	Traditional	Breed	Ypsitty	Woodmagic	American
ADMIX2.0	0.67(0.10)	0.52(0.10)	0.50(0.10)	0.42(0.13)	0.22(0.07)
LEADMIX	0.86(0.32)	0.98(0.27)	0.58(0.40)	0.58(0.38)	0.25(0.22)
LEA	0.63(0.13)	0.46(0.11)	0.39(0.17)	0.41(0.15)	0.22(0.13)

The source of the observed variation in parental contribution in this dataset is likely to be complex, due to the wide diversity of results for different methods and populations. If the parental populations were isolated for a few generations (since divergence from the ancestral population) then most admixture methods would be expected to

estimate parental contributions close to 0.5 with a large associated variance (Bertorelle & Excoffier, 1998). Similarly if the hybrid and parental populations have drifted significantly since the admixture event, the variance in the estimation will be large and the posterior distribution for the Bayesian method would tend to be flat. It can therefore be statistically difficult to distinguish the two effects mentioned above. That there are nine (of sixteen) cases where the estimated contributions fall within 10% of 0.5 (equal contribution) and that large F_{ST} values are observed between the samples suggests that genetic drift is a major source of uncertainty. Importantly, similarity between parental populations at the point of admixture is also expected to reduce the methods' utility to estimate admixture proportions, and the F_{ST} value between the modern day Devon and Kerry is not large, although this may have been smaller or larger in the past, depending on the extent of genetic drift.

The variation in contributions for the different hybrid populations could be explained by three non-exclusive factors. First, if, as seems likely, there was a large amount of genetic variation across the Dexter herd prior to the demographic bottleneck in the 1960s, sampling (founder) effects could explain differential contribution to lineages which became demographically isolated. Second, introgression into the separate populations may have occurred during or shortly after foundation of some of the groups, though none is documented for any except the Dexter Breed group. Finally extreme genetic drift may by chance have obscured the true parental contributions of the different groups sampled. While this effect is likely to have contributed for groups such as the American Dexter sample, Woodmagic, Ypsitty and the Kerry, the LEA approach (Chikhi *et al.* 2001) is designed to account for drift yet shows similar variation in results to the other non mutation-based estimators. It should be stressed that if genetic drift is a major feature of the data, then most of the information related to the admixture (or any other) event will become lost as time increases from the point of admixture. Thus, in cases of extreme drift, it may simply have to be acknowledged that limitations to the methodology exist because information lost cannot be retrieved, at least with genetic data alone.

For management purposes it should be recognised that there are measurable differences between allele frequencies in subpopulations of the breed, and that these may be partially due to the contributions of ancestral populations. If the

Traditional Dexter sample is accepted as being representative of the original breed, the implication is that other subpopulations have potentially diverged from this specific type. However, to avoid widespread use of individuals from these subpopulations may risk the loss of potentially unique alleles (when considering present-day populations as random samples from the diversity present in the Dexter at its formation). This concern is a potentially valid one provided introgression has not contributed to this divergence of subpopulations. Genetic data such as those presented here therefore cannot by themselves answer all the questions relevant to the Dexter breed conservation.

What is clear from this study is that the admixture approaches used here did not allow us to draw uniform conclusions regarding the relative parental contributions. Developments such as the ability to account for multiple admixture events and to account for multiple parental populations would be useful in this context but whether this would increase the precision in admixture estimates has yet to be explored. Inherent small population stochasticity and the implications of genetic drift, particularly during and after population bottlenecks, still presents a major challenge for accurate admixture determination.

4.5. Acknowledgements

We thank the Rare Breeds Survival Trust, Dexter Cattle Society, Instituto Gulbenkian de Ciência, Université Paul Sabatier and Cardiff University for funding and infrastructural support for this research, which forms part of TCB's PhD. Part of this work was carried out during visits by TCB to Lisbon and Toulouse. Thanks go to A. Coutinho and B. Crouau-Roy. LEA calculations were performed using the High Performance Computing resource at the Instituto Gulbenkian de Ciência (IGC) with the help of P. Fernandes and using the Condor cluster at Cardiff University with the help of S. Adams (School of Biosciences) and J. Osborne (ARCCA). Some of this manuscript was written while L.C. was visiting IGC: the CNRS and B. Crouau-Roy are thanked for making this possible.

4.6. References

Barendse W., Armitage S.M., Kossarek L.M., Shalom A., Kirkpatrick B.W., Ryan A.M., Clayton D., Li L., Neibergs H.L., Zhang N., Grosse W.M., Weiss J., Creighton P., McCarthy F., Ron M., Teale A.J., Fries R., McGraw R.A., Moore S.S., Georges M., Soller M., Womack J.E., & Hetzel D.J.S. (1994) A genetic linkage map of the bovine genome. *Nature Genetics* **6**, 227-235.

Bartosiewicz L. (1997) The Hungarian grey cattle: a traditional European breed. *Agriculture* **21**, 49-60.

Beja-Pereira A., Alexandrino P., Bessa I., Carretero Y., Dunner S., Ferrand N., Jordana J., Laloe D., Moazami-Goudarzi K., Sanchez A., & Canon J. (2003) Genetic characterisation of Southwestern European bovine breeds: A historical and biogeographical reassessment with a set of 16 microsatellites. *Journal of Heredity* **94**(3), 243-250.

Belkhir K., Borsa P., Chikhi L., Raufaste N., & Bonhomme F. (2001) GENETIX 4.03, logiciel sous Windows™ pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier, France.

Bertorelle G. & Excoffier L. (1998) Inferring admixture proportions from molecular data. *Molecular Biology and Evolution* **15**, 1298-1311.

Bishop M.D., Kappes S.M, Keele J.W., Stone R.T., Sunden S.L.F., Hawkins G.A., Solinas –Toldo S., Fries R., Grosz M.D., Yoo J., & Beattie C.W. (1994) A genetic linkage map for cattle. *Genetics* **136**(2), 619-639.

Blott S.C., Williams J.L., & Haley C.S. (1998) Genetic variation within the Hereford breed of cattle. *Animal Genetics* **29**(3), 203-211.

Brezinsky L.S., Kemp J. And Teale A.J. (1993a) ILSTS005: a polymorphic bovine microsatellite. *Animal Genetics* **24**, 73.

Brezinsky L.S., Kemp J., & Teale A.J. (1993b) ILSTS006: a polymorphic bovine microsatellite. *Animal Genetics* **24**, 73.

Canon J., Alexandrino P., Bessa I., Carleos C., Carretero Y., Dunner S., Ferran N., Garcia D., Jordana J., Laloe D., Pereira A., Sanchez A., & Moazami-Goudarzi K. (2001) Genetic diversity measures of local European beef cattle breeds for conservation purposes. *Genetics, Selection, Evolution* **33**, 311-332.

Caramelli D. (2006) The origins of domesticated cattle, Human Evolution 21:107-122.

Chikhi L, Bruford M. W, and Beaumont M. A. (2001) Estimation of admixture proportions: A likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**, 1347-1362.

Chikhi L., Goossens B., Treanor A., & Bruford M.W. (2004) Population genetic structure of and inbreeding in an insular cattle breed, the jersey, and its implications for genetic resource management. *Heredity* **92**, 396-401.

Choisy M., Franck P., & Cornuet J. M. (2004) Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Molecular Ecology* **13**, 955-968.

Cleveland M.A., Blackburn H.D., Enns R.M., & Garrick D.J. (2005) Changes in inbreeding of U.S. Herefords during the twentieth century. *Journal of Animal Science* **83**, 992-1001.

Dexter Cattle Society Herd Book Volume LI (1951) Dexter Cattle Society

Dexter Cattle Society Herd Book Volume LXXXVIII (1988) Dexter Cattle Society

Dupanloup I. & Bertorelle G. (2001) Inferring admixture proportions from molecular data: extension to any number of parental populations. *Molecular Biology and Evolution* **18**(4), 672-675.

Evanno G., Regnaut S., & Goudet J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Falush D., Stephens M., & Pritchard J.K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587.

Georges M. & Massey, J.M. (1992) Polymorphic DNA markers in Bovidae. Patent WO 92/13102 (1992).

Goossens B., Waits LP., & Taberlet P. (1998) Plucked hair samples as a source of DNA: reliability of dinucleotide microsatellite genotyping. *Molecular Ecology* **7**, 1237-1241.

Heath-Agnew, E. (1983) *A history of Hereford cattle and their breeders.* Duckworth, London.

Kaukinen J. & Varvio S.L. (1993) Eight polymorphic bovine microsatellites. *Animal Genetics* **24,** 148.

Loftus R., Ertrugrul O., Harba A., El-Barodys M., MacHugh D., Park S., & Bradley D. (1999) A microsatellite survey of cattle from a centre of origin: the near east. *Molecular Ecology* **8**, 2015-2022.

MacHugh D.E., Shriver M. D., Loftus R.T., Cunningham, P., & Bradley D.G. (1997) Microsatellite DNA variation and the evolution, domestication and phylogeography of Taurine and Zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* **146**, 1071-1086.

Martin-Burriel I., Garcia-Muro E., & Zaragoza P. (1999) Genetic diversity analysis of six Spanish native cattle breeds using microsatellites. *Animal Genetics* **30**, 177-182.

Moore S.S. & Byrne K. (1993) Dinucleotide polymorphism at the bovine calmodulin independent adenylcyclase locus. *Animal Genetics* **24,** 150.

Moore S.S., & Byrne K., Berger K.T., Barendse W., McCarthy F., Womack J.E., & Hetzel D.J.S. (1994) Characterisation of 65 bovine microsatellites. *Mammalian Genome* **5**, 84-90.

Peelman L.J., Mortiaux F., Van Zeveren A., Dansercoer A., Mommens G., Coopman F., Bouquet Y., Burny A., Renaville R., & Portetelle D. (1998) Evaluation of the genetic variability of 23 bovine microsatellite markers in four Belgian cattle breeds. Animal *Genetics* **29**(3), 161-167.

Pritchard J.K., Stephens M., & Donnelly P.J. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Rutherford B. (2005) *My Love Affair With The Dexter* (publisher Triple D Books, Wagga Wagga, NSW, Australia, ISBN 0 9756829 0 3

Solinas-Toldo S., Fries R., & Steffen P. (1993) Physically mapped, cosmid derived microsatellite markers as anchor loci on bovine chromosomes. *Mammalian Genome* **4**, 720-727.

Steffen P., Eggen A., Dietz A.B., Womack J.E., Stranzinger G., & Fries R. (1993) Isolation and mapping of polymorphic microsatellites in cattle. *Animal Genetics* **24**(2), 121-124.

Taberlet P., Valentini A., Rezaei H. R., Naderi S., Pompanon F., Negrini R., & Ajmone-Marsan P. (2008) Are cattle, sheep, and goats endangered species? *Molecular Ecology* **17**(1), 275-284.

Takeshima S., Saitu N., Morita M., Inoko H., & Aida Y. (2003) The diversity of bovine MHC class II DRB3 genes in Japanese Black, Japanese Shorthorn, Jersey and Holstein cattle in Japan. *Gene* **316**, 111-116.

Thieven U., Solinas-Toldo S., Friedl R., Masabanda J., Fries R., Barendse W., Simon D., & Harlizius B. (1997) Polymorphic CA-microsatellites for the integration of the bovine genetic and physical map. *Mammalian Genome* **8**, 52-55.

Vaiman D., Osta D., Mercier D., Grohs C., & leveziel H. (1992) Characterization of five new bovine microsatellite repeats. *Animal Genetics* **23**, 537.

Vigilant L. (1999) An evaluation of techniques for the extraction and amplification of DNA from naturally shed hairs. *Biological Chemistry* **380**, 1329-1331.

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

Walsh P.S., Metzger D.A., & Higuchi R. (1991) Chelex R100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques* **10**, 506-513.

Wang J. (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**, 747-765.

Weir B.S., & Cockerham C.C. (1984) Estimating F-Statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Wiener P., Burton D., & Williams J.L. (2004) Breed relationships and definition in British cattle: a genetic analysis. *Heredity* **93**, 597-602.

Wilson J. (1909a) The origin of the Dexter breed of cattle. *Scientific Proceedings of the Royal Dublin Society* **12**(1) 1-17.

Wilson J. (1909b) *The evolution of British cattle and the fashioning of breeds* (London) p. 75.

Chapter 5.

Development of a novel approximate Bayesian computation method for admixture quantification

5.1. Introduction

This chapter details the development of the new Approximate Bayesian Computation admixture estimation method. Our aim was to develop a user-friendly graphical interface that could be easily used and later improved to estimate admixture proportions in a model that is likely to represent the demographic history of a number of breeds. Unlike previous methods, this application takes into account two separate admixture events under a wide range of possible parameter values. The addition of an extra admixture event allows the more accurate modeling of scenarios that were previously forced to consider admixture events independently. The process of testing and refining the working program structure is detailed. Following the testing-by-simulation approach is a section through which the working program is applied to data from two cattle breed examples. The development of the working script is summarised, especially alterations made after initial programming difficulties were resolved. Omitted from this account are those errors that have been largely due to inconsistencies in directory use, incorrect combination of functions, and use of erroneous script grammar. Included in this description are the formative steps through which the computational demand, efficiency, and program output formats were checked and finalised.

Although the development process has involved producing many functional versions of the program, only the final version is described in detail in the subsequent testing and application sections. Appendix 5.1 contains the entire program script including the main program file and all associated functions for generating summary statistics. Several important methodological improvements were made in the finalisation of the working application and these should be mentioned.

5.1.1. Reliance on the GUI

For simplicity of program use the application runs from a graphical interface that translates the parameters chosen by the user into a command used by ms, the simulation program developed by Hudson (2002). As detailed in Figure 2.5, the GUI can be time-consuming (since there are many parameters to specify, due to the complexity of the model). This is why the script for the final version allows the user to run the program without the graphical interface. To do this a 'graphical' field can be specified as on ('=1') or off ('=0'). The removal of the graphical interface not only

allows rapid testing using pre-set parameters but also provides the basis for future parallelisation of the script. This can be simply modified to enable the program file to be submitted with a datafile and text command file detailing what parameters are required onto a parallel processing system such as Condor (Litzkow, 1988). This is directly comparable to the STRUCTURE application (Pritchard et al. 2000) in which separate run jobs can be submitted onto available processors simultaneously, reducing calculation time.

5.1.2. Automatic generation of observed data

Initial versions of the application saved all file information as it was generated. The correct function of the script could be verified by checking each file to make sure that input parameters were reflected in the simulated data being generated. The large quantities of data generated in longer runs prohibited this in later versions. Therefore the use of the saved simulated allele frequencies (generated using known parameters) as observed data was no longer possible. A replacement code was included for generation of an observed data file at set parameter values. This necessitates the use of the 'make_observed' field in the final script. Where an observed file is not recognised the field will be '=0' and an observed file will be generated according to an automatic set of parameters. This will allow both determination of run times as well as a greater understanding of the program functions through investigation of the observed data and associated files: observed.txt, obs_sumstat.txt, and rel_obs_sumstat.txt.

5.1.3. Separate function files

The initial script, comprising a single file, included all of the associated functions together. The management and efficiency of this arrangement was improved by separation into the specific functions which can be called as required. Changes to functions can be tested separately and will not create errors in the main script which can often prove complicated to uncover. Initial functions were also improved through changes in the language employed (i.e. removal of unnecessary loops), increasing efficiency and reducing calculation time.

5.2. Use and testing of program application

Initiation of the program application is through the MATLAB v7 (Mathworks, 2001) platform prior to creation of a stand-alone application. Requirements to run the program are that the ms (Hudson, 2002) application for simulation of genetic data is installed into the C:\ drive of the machine to be used.

The program script can be run with or without the GUI. To ensure that all changes function with the more complicated call functions from the graphical window the testing was performed through the GUI. Before executing the script through the 'run' button, a data file must exist against which the simulated data can be compared. This file carries the identifier 'observed.txt', and takes the form of ASCII text format. As mentioned previously the observed file may be generated automatically but may also be generated manually from 'real' data. To manually generate an observed file it is possible to use any of a number of genetic analysis software to calculate allele frequencies such as GENETIX (Belkhir et al. 2002). The observed file adopts a specific format with as many lines as there are loci, it is imperative that the simulated data use the same number of loci or an error will be generated. Each row represents a single locus, for each locus the initial column contains a value 'n' representing the number of alleles at that locus. The subsequent n columns correspond to the allele counts for each allele in population 1. Following this are the allele counts for populations 2, 3, and the hybrid (H). At the end of each row are n numbers describing the relative allele sizes (measured in microsatellite repeats). A typical observed data file with 5 loci might look like Figure 5.1 (boxes have been included for the first locus only to explicitly show how the data are coded).

4 0 3 18 24 | 0 8 36 1 | 5 0 4 36 | 3 2 20 20 -4 -3 -2 -1
6 0 0 0 4 15 26 1 21 15 1 7 0 0 23 22 0 0 0 2 15 9 3 9 7 -7 -6 -5 -3 -2 -1
5 0 0 5 12 28 3 5 1 21 15 0 2 17 19 7 0 3 7 24 11 -5 -4 -3 -2 -1
4 4 0 18 23 0 8 31 6 0 6 2 37 2 2 14 27 -3 -2 -1 0
7 0 12 17 2 0 0 14 0 0 3 33 9 0 0 0 0 1 1 12 31 0 1 6 11 6 5 11 5 -8 -7 -5 -4 -3 -2 0

Figure 5.1. input file format for the program

The observed data file is transformed into the summary statistics which are saved separately for comparison against the summary statistics of the simulated datasets. The twenty five summary statistics comprise four categories; heterozygosity, allelic range, private alleles, and pairwise F_{ST}. The exact composition of each group of summary statistics is detailed in the individual program functions (Appendix 5.1).

Initial script testing was performed by fixing some parameters at set values. Due to the large number of parameters of the model, it seemed reasonable to first test the inference method by fixing some parameter values so as to reduce the uncertainty on the estimation of some parameters. Since we were particularly interested in estimating admixture and determining whether the admixture events could be located in time, we fixed all parameters of the models (Table 5.1) but the event times, and admixture proportions. This narrowing of the numbers of parameters upon which variation in simulated data depends means that only five parameters are of interest in the analytical process. Additionally, for the initial program testing a simplified scenario was applied using only a single admixture event, further reducing the initial analysis to three parameters of interest (Figure 5.2). The reason for this is that by doing that the model became similar to the admixture model of Chikhi et al (2001) and identical to that of Excoffier et al. (2005). This way we had the possibility to determine whether the inference approach chosen was giving reasonably good results compared to these two methods.

Table 5.1. Program parameters specified for the generation of simulated data in the ms application.

Parameter number	Value of observed	Simulated parameter range	Parameter description (ms value conversion)
1	0.0001	0.0001	theta (4*Ref_N*mutation rate)
2	5000	5000	relative N1
3	5000	5000	relative N2
4	5000	5000	relative N3
5	5000	5000	relative N4
6	1	0-10	time of recent admixture, or tadm2 (/4*Ref_N)
7	0	0	proportion from recent admixture event (1-p3)
8	1	0-10	time of recent admixture, or tadm2 (/4*Ref_N)
9	10	0-50	time of initial admixture event or tadm 1(/4*Ref_N)
10	0.7		contribution of p1
11	10	0-50	time of initial admixture event or tadm 1(/4*Ref_N)
12	10	0-50	time of initial admixture event or tadm 1(/4*Ref_N)
13	10000	9000-12000	time of split(/4*Ref_N)
14	10000	9000-12000	time of split(/4*Ref_N)
15	10000	9000-12000	time of split(/4*Ref_N)
16	1	1	relative size of ancestral population(/Ref_N)
17	5000	5000	Ref N

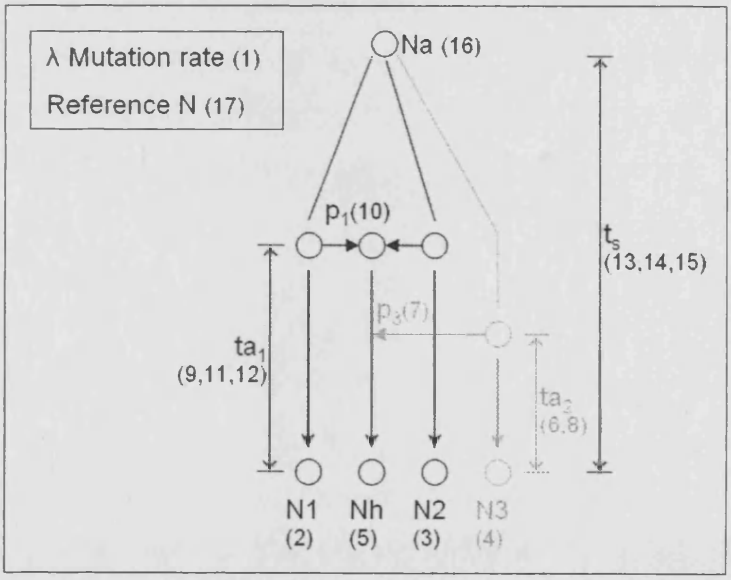


Figure 5.2. Simplified admixture scenario with the unused portion of the complete model shaded.

5.2.1. Data Collection

Once simulated data were generated through ms, summary statistics were calculated as a series of independent functions and are summarised to be saved in a results folder 'simulation_database'. Summary statistics between the mean of the observed data across all loci were compared with the same statistics across a mean of the same number of simulated loci (generated under the same specified model). A standardisation process was adopted to ensure that no single summary statistic was given a disproportionate weighting in the comparison between observed and simulated data. The method used for calculation of this standardising factor was to divide by the mean of the first 10,000 simulated summary statistics. Included in the calculation was the setting of a tolerance limit, above which distance values were discarded. The following program script gives an example of the process in an annotated Matlab language format (Figure 5.3).

```
sim_sumstats_mean_file = fopen('./simulations_database/sim_sum_means_1.txt','r');
   % read the first nbsim_tol lines
   sim_data_mean = fscanf(sim_sumstats_mean_file, '%g', [ nb_sumstats nbsim_tol]);
   % Close the simulated datafile
   fclose(sim_sumstats_mean_file);
   sim_data_mean = sim_data_mean';
   % Get the mean and standard deviation of each sum_stat over the tol_sim
   % This will be used to standardize the observed and simulated sum stat
   mean_sim_mean = mean(sim_data_mean,1);
   std_sim_mean = std(sim_data_mean,0,1);
   % Standardize the simulated data
   % The second step avoids division by zero when the std is zero
   sim_data_mean = (sim_data_mean-(ones(nbsim_tol,1)*mean_sim_mean));
   evaluate = std_sim_mean ~= 0;
   sim_data_mean(:,evaluate) = sim_data_mean(:,evaluate) ./
(ones(nbsim_tol,1)*std_sim_mean(evaluate));
   % Read the observed data for the rep observation
   obs_data_mean = fscanf(fid_mean,'%g',nb_sumstats);
   % Standardize the observed data mean
   obs_data_mean = (sumstat_obs_mean' - mean_sim_mean);
   obs_data_mean(evaluate) = obs_data_mean(evaluate) ./ std_sim_mean(evaluate);
```

Figure 5.3. An example of Matlab format program code.

Once standardised, each simulated summary statistic could be subtracted from its equivalent observed statistic. The positive mean of these values provided a measure of the similarity of simulated and observed data. This 'distance' measure allows selection of those simulated data which are most similar (shortest distance measure) to the observed data, and therefore allows the determination of corresponding parameter values.

5.2.2. Data analysis

Using the R language software (Ihaka and Gentleman, 1996), the prior parameter distributions were checked to ensure that they conformed to that specified (uniform distributions were used but distributions may also take the form of; normal, gamma, or log normal). The parameters corresponding to the shortest distances could then be plotted in a density plot. This stage was combined with a regression step (Beaumont et al., 2002; Hamilton *et al*, 2005), to increase the accuracy of parameter prediction. The specific R commands used for each of these processes can be seen in Appendix

5.2. The posterior distribution of the parameters corresponding to the shortest distances between observed and simulated data can then be plotted as a histogram. Mean, modal value, and the associated variance can be calculated for each parameter.

5.3. Results – program testing through simulation

The initial observed data file was created using a single simulation where all parameters were fixed at the desired values. For this testing scenario parameter settings were as shown in Table 5.1. The number of simulated datasets used was varied to investigate how this affected the quality of the inference of the parameters whose values were known. The probability density plots were also drawn employing different tolerances to illustrate how the prediction was affected by the proportion of the distances used. In previous test runs the large standard deviations proved problematic in the admixture determination. To initially solve this, and to test the effectiveness of the approach using a simpler scenario, the most recent admixture event was removed. This was done by setting the contribution of population 3 to zero. The following results section is divided into two sections, the first detailing the results using a single admixture event, and a second where the second admixture event was included. In all scenarios the sample sizes of the simulated datasets were set at 50 and the number of loci used was nine. The sample size was chosen to be sufficiently high to reflect the model applied whilst limiting the calculation time for generation of simulated data. The numbers of loci was again limited by calculation time and was also made to correspond with the available loci in the Lincoln Red dataset.

Where values are given in the text they refer to the means and variances of the parameter values corresponding to the accepted distances. These were calculated in R subsequent to the initial analysis and plots using a separate script (Appendix 5.3).

5.3.1. Single admixture event scenario

The results of the runs using only one admixture event between populations 1 and 2 to create the hybrid population are shown in Table 5.2. The estimates for P1 contribution where the observed value is set to 0.7 give a good accuracy for 100,000 simulations. The estimation of the 0.7 P1 proportion varied between 0.65 (+/- 0.06) and 0.0.70 (+/- 0.07). Estimates of the time of admixture varied between 3.8 (+/- 2.1) and 8.2 (+/- 4.6) generations. The estimation of time of split varied between 9110 (+/-

880) and 10310 (+/- 700) generations. The density plots for the posterior distributions and variance of P1 in the longer run can be seen in Figure 5.4 using dataset 1 as an example.

Table 5.2. An analysis of five observed files generated under the same single admixture model against 100,000 simulated datasets fixed for population size.

Parameters		Set values	Datasets 1	2	3	4	5
P1	Mean		0.70	0.69	0.65	0.70	0.69
	Adjusted mode	0.7	0.71	0.71	0.65	0.71	0.70
	Variance		0.06	0.07	0.06	0.07	0.06
Time of admixture event (generations)	Mean	10	3.83	8.24	7.36	4.90	7.92
	Variance		2.06	4.62	3.92	2.55	4.22
Time of split (generations)	Mean	10,000	9105	10120	9948	10310	10090
	Variance		880	754	778	702	739

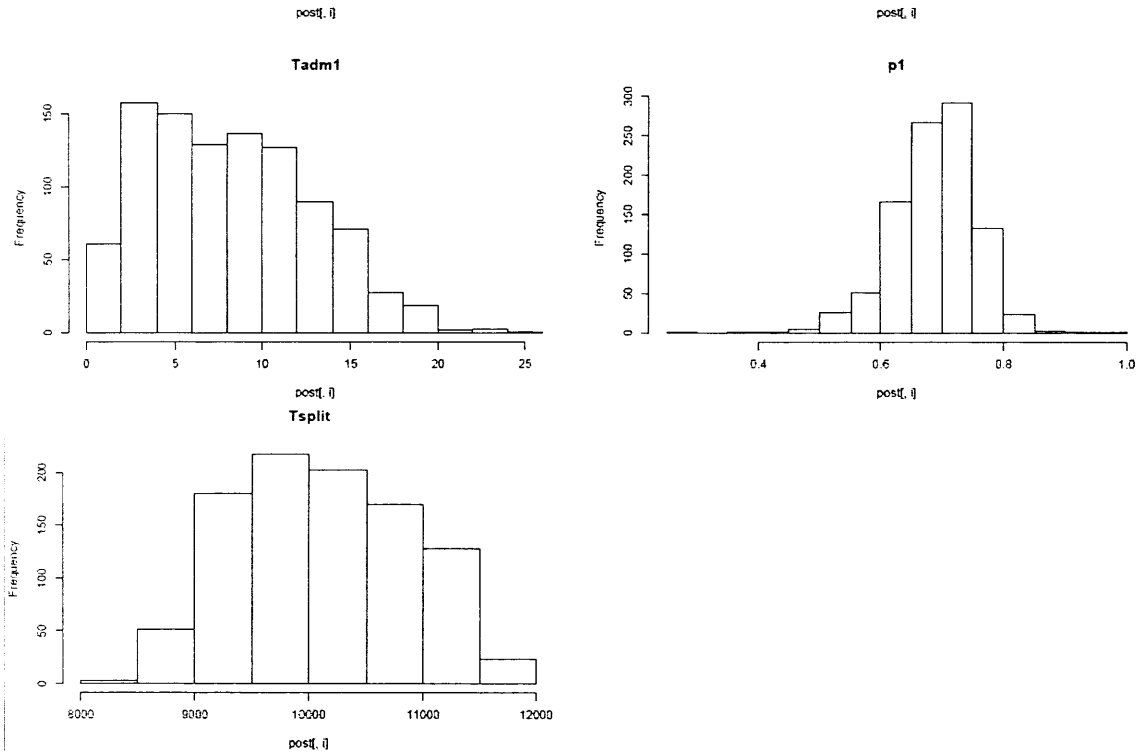


Figure 5.4. Posterior plots of the predicted parameters compared against 100,000 simulated datasets in a single admixture scenario. Tadm1 represents the time of admixture event, P1 is the admixture proportion of population 1, Tsplit is the time of split.

5.3.2. Full admixture scenario-100,000 simulations

The full model was then applied including the second admixture event but retaining fixed population sizes and mutation rates. The observed values for P1 and 1-P3 were both set at 0.7. For 100,000 simulations the P1 estimation varied between 0.453 (+/- 0.15) and 0.860 (+/- 0.07) and 1-P3 between 0.56 (+/- 0.10) and 0.65 (+/- 0.10) (Table 5.3). These estimates were more variable than for the single admixture scenario as a direct result of the added parameter and therefore complexity of the model. The mean regression histograms for p1 and 1-p3 can be seen in Appendix 5.4. The time of the most recent admixture event was overestimated and highly variable, with estimates between 1.4 (+-1.1) to 11.5 (+-7.8). The older admixture event was underestimated with estimates varying between 1.4 (+-1.0) to 10.4 (+-6.0). The time of split was underestimated in one of the scenarios (9700). In all other runs the mean values were within 60 generations either side of the true value, but with high variances of between 750 and 790 generations.

Table 5.3. An analysis of observed data generated under a two admixture event model against 100,000 simulated datasets fixed for population size

			Datasets				
Parameters		Set values	1	2	3	4	5
P1	Mean	0.7	0.64	0.66	0.45	0.86	0.62
	Adjusted mode		0.67	0.72	0.45	0.90	0.66
	Variance		0.14	0.15	0.15	0.07	0.16
1-P3	Mean	0.7	0.64	0.56	0.65	0.600	0.65
	Adjusted mode		0.65	0.59	0.65	0.590	0.67
	Variance		0.08	0.10	0.08	0.10	0.10
Time of recent admixture event (generations)	Mean	1	2.90	3.56	4.10	1.40	11.5
	Variance		1.75	2.19	2.45	1.12	7.76
Time of older admixture event (generations)	Mean	10	3.79	5.35	5.02	1.40	10.40
	Variance		2.05	2.87	2.59	1.02	6.00
Time of split (generations)	Mean	10,000	10060	10000	9950	9940	9720
	Variance		773	778	764	749	792

5.3.3. Full admixture scenario - 500,000 simulations

In this scenario the restrictions of fixing the population sizes and mutation rate were removed. The full model was then applied including the second admixture event. The observed values for P1 and 1-P3 were both set at 0.7. For 500,000 simulations the P1 estimation varied between 0.12 (+/- 0.16) and 0.70 (+/- 0.25) and 1-P3 between 0.11 (+/- 0.28) and 0.99 (+/- 0.23) (Table 5.4). The mean regression histograms for p1 and 1-p3 can be seen in Appendix 5.5, which show that the performance is much poorer, and predictions from the mean and modal values bore very little if any relationship to the true parameter that is being predicted. Figure 5.5 is an example of a test run of 500,000 simulations using 20 loci. Compared to the previous results for 9 loci, this data demonstrates a substantial improvement in the posterior distribution of the 1-p3 parameter. This data suggests that the small increase in numbers of loci used from 9 to 20 has an effect on the 1-p3 parameter estimate. Times of the admixture events were similar in accuracy to the 100,000 simulation run, with the recent admixture event slightly overestimated (0.02 – 4.8 generations) and the older admixture event underestimated (1.8 – 6.2 generations). The posterior plots in Figure 5.5 illustrate this well, showing very little difference between the two parameters. The time of split estimates were considerably worse than the previous runs and estimates varied between 2,000 and 13,000 generations with variances of up to 2400 generations.

Table 5.4. An analysis of observed data generated under a two admixture event model against 500,000 simulated datasets allowing for variation in all parameters.

Parameters		Set values	Datasets 1	2	3	4	5
P1	Mean	0.7	0.70	0.25	0.24	0.68	0.12
	Adjusted mode		0.93	0.07	0.078	1.00	0.10
	Variance		0.25	0.24	0.23	0.27	0.16
1-P3	Mean	0.7	0.48	0.24	0.99	0.71	0.11
	Adjusted mode		0.1	0.05	0.78	0.93	0.37
	Variance		0.30	0.24	0.23	0.26	0.28
Time of recent admixture event (generations)	Mean	1	0.02	4.80	1.5	0.07	2.21
	Variance		00.1	3.10	6.87	1.46	11.1
Time of older admixture event (generations)	Mean	10	1.76	6.2	3.98	2.98	3.92
	Variance		0.95	29.9	2.05	1.82	2.61
Time of split (generations)	Mean	10,000	1990	12500	13400	1210	616
	Variance		2440	2130	1260	1650	1230

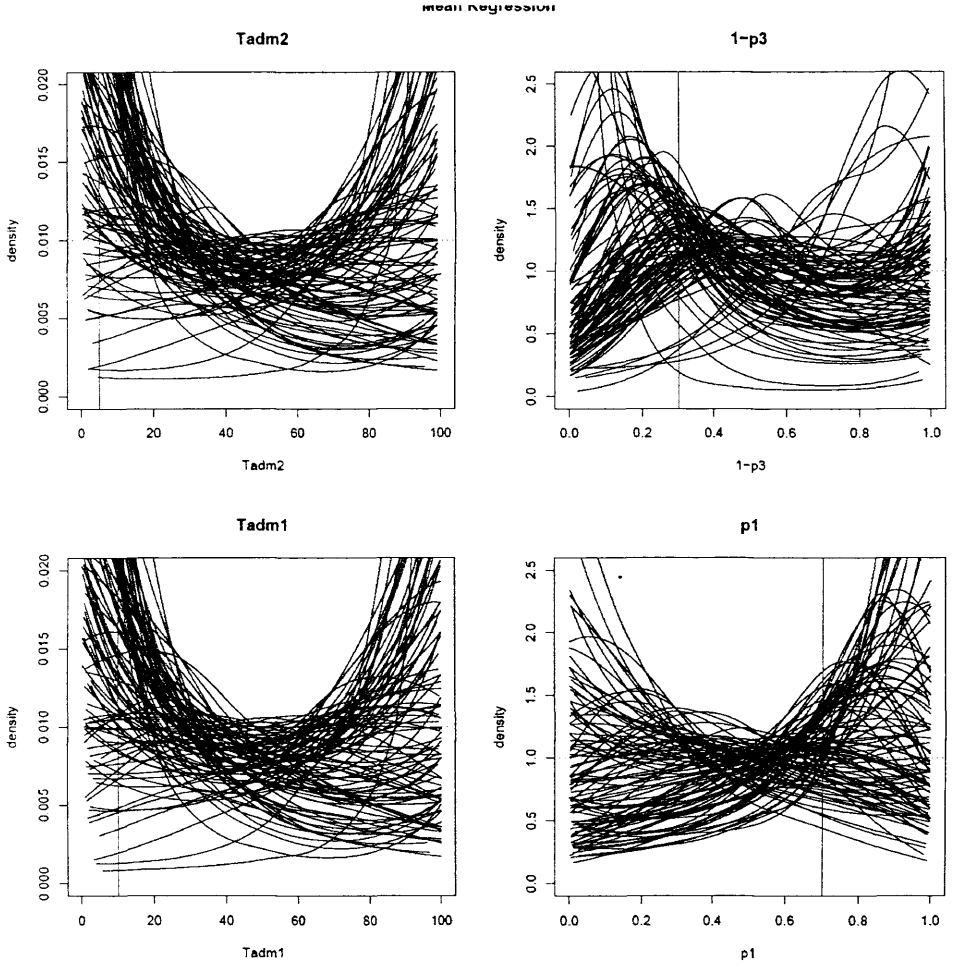


Figure 5.5. A plot of the mean regression of the shortest 1000 accepted distances using 500,000 simulations, and 20 loci. P1 = 0.7, 1-P3 = 0.3. (plots courtesy of Vitor Sousa, Gulbenkian Institute).

5.4. Detailing applied population scenarios for analysis

Using microsatellite data from actual cattle populations allows determination of the utility of the method for real data. Whether the results from the test scenarios can be upheld in real admixture situations is a true test of the methodology. For simulated testing scenarios the data were generated with a relatively ancient split since which populations have differentiated, this is often not the case with real data and decreasing differentiation time between populations may reduce the ability of this (and probably of any) approach to quantify admixture events (Bertorelle and Excoffier, 1998). In addition to the Dexter breed, a secondary population of interest in this study, is the Lincoln Red which was hoped would provide an effective test scenario for the method.

5.4.1. Lincoln Red Scenario

The Lincoln Red breed is one where the history is thought to be relatively simple, two recent admixture events having been influential in shaping the modern breed. The recent and quantifiable nature of the admixture contributions provides a natural test situation for the method developed here. Information on herd composition within the breed allowed more fine-scale investigation and comparison of relative admixture proportions than would be possible in other breed populations.

The earliest known reference to Lincolnshire cattle is thought to be that of Markham in 1695. Two types of Shorthorn cattle were initially recognised in 1822 (Coates herd book), the first Lincoln Red herd book being produced in 1896 (Lincoln Red herd book). Lincoln Red cattle were initially dual purpose but were increasingly bred for beef. The breed became polled through use of red Aberdeen Angus animals from 1963 and the subsequent breed development program reinforced the beef qualities through the Breed Development Programme of introduction of continental breeds such as (but not exclusively) Limousin from around 1973 (S.J. Hall, pers. comm.).

The Lincoln Red data available included four herds which differed in their genetic composition relative to the described ancestral breeds (Figure 5.6). For the purposes of the admixture method test analysis one herd will be used to ascertain whether the program provides accurate admixture determination. The Wardle and Harrington herds are both thought to have had only minor influence from the Limousin breed, contributions to the other herd populations are less easy to quantify at this time but the McTurk herd is though to have had the greatest level of upgrading by Limousin. From the structure plot in Figure 5.7 the Harrington herd can be seen to have a greater proportion of its members in common with other herd populations, when compared to the Wardle herd, implying a higher rate of gene flow. In order to use a population with a known (very low) Limousin proportion, the Wardle herd was therefore chosen. In addition to this the McTurk herd was used to provide a comparison of the method within the breed.

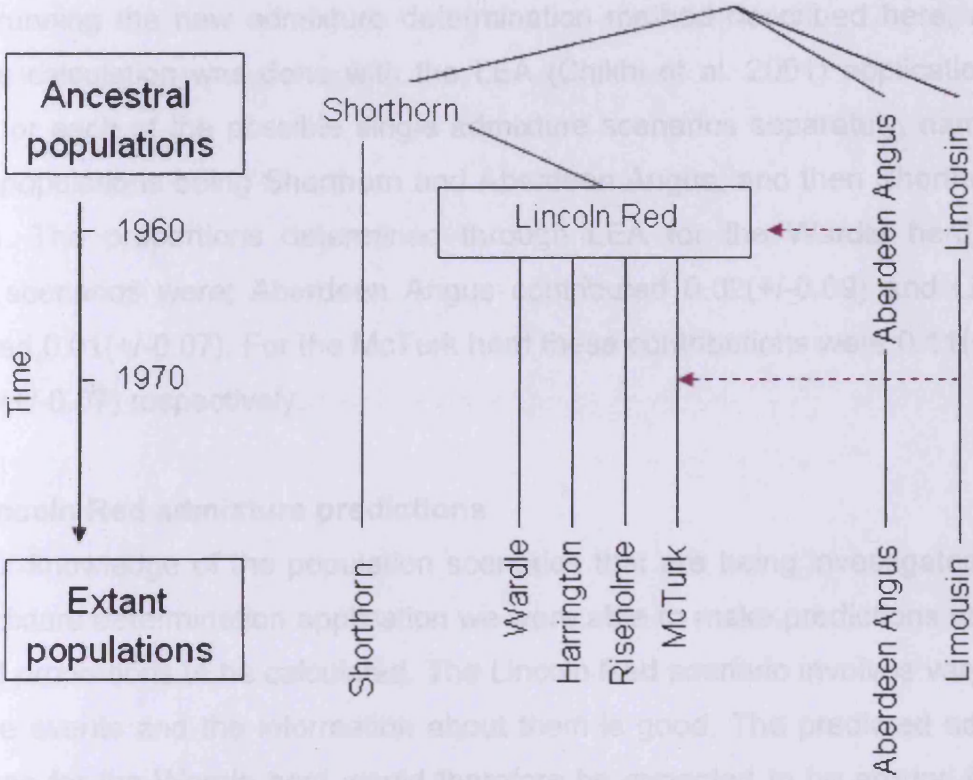


Figure 5.6. Diagrammatical representation of the historical interactions of the Lincoln Red breed with its ancestral and contributory breed populations (S.J. Hall pers. comm.).

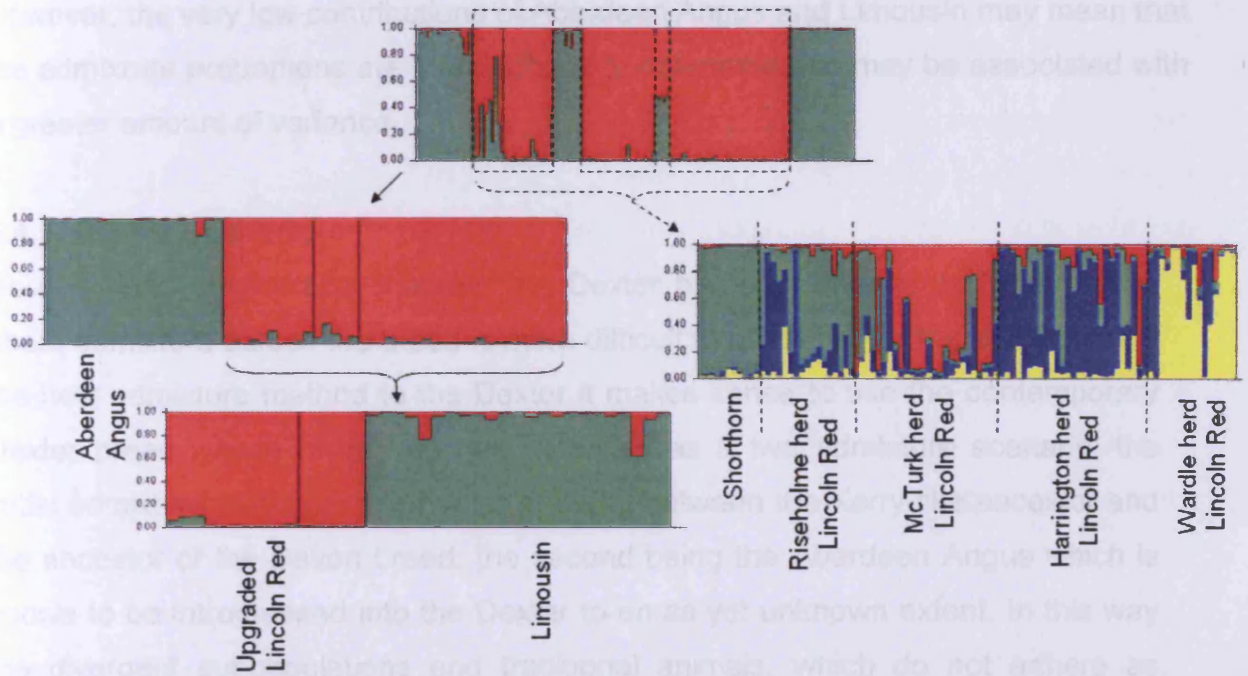


Figure 5.7. STRUCTURE representation of the genetic composition of the Lincoln Red samples taken in relation to Aberdeen Angus, Shorthorn, and Limousin breeds.

Prior to running the new admixture determination method described here, another admixture calculation was done with the LEA (Chikhi et al. 2001) application. This was run for each of the possible single admixture scenarios separately, namely the parental populations being Shorthorn and Aberdeen Angus, and then Shorthorn and Limousin. The proportions determined through LEA for the Wardle herd in the different scenarios were; Aberdeen Angus contributed 0.02(+/-0.09) and Limousin contributed 0.01(+/-0.07). For the McTurk herd these contributions were 0.11(+/-0.09) and 0.08(+/-0.07) respectively.

5.4.2. Lincoln Red admixture predictions

Given our knowledge of the population scenarios that are being investigated by the new admixture determination application we were able to make predictions about the expected proportions to be calculated. The Lincoln Red scenario involves very recent admixture events and the information about them is good. The predicted admixture proportions for the Wardle herd would therefore be expected to be around 0.05-0.1 for the contribution of the earlier admixture event by the Aberdeen Angus at around 5-10 generations ago, with the Limousin contribution being 0-0.05 up to 5 generations ago (S.J. Hall, pers. com). The McTurk herd would be expected to be marginally higher perhaps up to 0.15 for Aberdeen Angus and 0.1 for Limousin. However, the very low contributions of Aberdeen Angus and Limousin may mean that the admixture proportions are more difficult to determine and may be associated with a greater amount of variance.

5.4.3. Dexter Scenario

As described in previous chapters the Dexter breed is diverse and information about admixture across the breed is more difficult to quantify. For the application of the new admixture method to the Dexter it makes sense to use the contemporary Dexter breed whose history can be modelled as a two admixture scenario; the initial admixture event (as in Chapter 4) being between the Kerry-like ancestor and the ancestor of the Devon breed, the second being the Aberdeen Angus which is known to be introgressed into the Dexter to an as yet unknown extent. In this way the divergent subpopulations and traditional animals, which do not adhere as closely to the modelled scenario, are not included in the analysis. Using LEA to suggest potential admixture proportions for these separate scenarios (Kerry and

Devon, and Kerry and Aberdeen Angus) gives the contribution of the Devon breed as 0.54(+/-0.11) and the contribution of the Aberdeen Angus as 0.11(+/-0.12).

5.4.4. Dexter scenario predictions

Pedigree-based admixture information for the Dexter breed is more difficult to calculate than that available for the Lincoln Red due to the greater time periods involved and the inconsistency of some breeding accounts. This makes it difficult to form accurate predictions for the admixture contributions. Only very rough predictions can therefore be made which are 0.2-0.5 for the Devon breed and 0-0.1 for the Aberdeen Angus breed.

5.5. Combined comparative results from application to real data scenarios

The application of the full admixture scenario to the Lincoln Red and Dexter datasets was made using 9 loci per run. For the case of the Dexter more loci were available, two runs were performed on the same breed population. Table 5.5 details the results from a 500,000 simulation run on the two Lincoln red populations and the Dexter breed population. The parameter estimation for the Lincoln Red herds for P1 differs dramatically from 0.89 (+-0.18) in the Wardle herd to 0.08(+-0.13) in the McTurk herd. There is less variation between the 1-p3 prediction which shows the admixture proportion of the Limousin breed to be slightly greater in the McTurk herd population 0.35(+-0.29) compared to 0.25(+-0.25). In the Dexter breed population there was marked difference between the two sets of loci used. An average combining the results of both sets of loci attributes 0.25(+-0.29) proportion contribution of the Devon in the older admixture event whilst more recently a 0.34 (+-0.26) contribution from the Aberdeen Angus.

The estimated times for the admixture events in the scenarios were highly variable suggesting that the recent admixture event was between the present and just over 6 generations ago in the Lincoln Red populations and between 3 and 5 generations ago in the Dexter. The older admixture event was far more uncertain, estimates between 2 and 20 generations for the Lincoln Red and 6 to 8 generations in the Dexter. The associated variances of these estimates were also very high, up to 31 generations in the Lincoln Red and 19 generations in the Dexter. Estimates of time of split were between 3400 and 8400 generations in the Lincoln Red and 2500 and 9600 generations in the Dexter. As in the 500,000 simulation test data used

previously, the variance in the time of split was also high at 3700-3900 generations for Lincoln Red and 2900-3500 generations for the Dexter.

Table 5.5. An analysis of two cattle datasets against 500,000 simulated datasets.

		Datasets					
		Lincoln Red			Dexter breed		
Parameters		Lincoln Red predictions	Wardle herd	McTurk Herd	Dexter predictions	Loci 1-9	Loci 10-18
P1	Mean		0.89	0.08		0.48	0.97
	Adjusted mode	**0.05-0.1/0.15**	1	0	**0.5-0.8**	0.16	1.0
	Variance		0.18	0.13		0.29	0.07
1-P3	Mean		0.75	0.65		0.94	0.33
	Adjusted mode	**0-0.05/0.1**	1	1	**0-0.1**	1	0.04
	Variance		0.25	0.29		0.12	0.26
Time of recent admixture event (generations)	Mean	**0-5**	6.24	0.01	**0-10**	3.43	4.87
	Variance		3.74	0.02		1.96	2.95
Time of older admixture event (generations)	Mean	**5-10**	20.31	1.87	**10-20**	8.23	6.02
	Variance		30.85	1.07		19.41	18.42
Time of split (generations)	Mean	**50-500**	8390	3470	**50-500**	9590	2460
	Variance		3910	3720		3540	2920

5.6. Discussion

The results for the calculation of admixture contributions for the single admixture scenario were generally accurate. This was true despite the relatively small numbers of simulations being used for comparison. The best performance was seen in the P1 parameter with only slight underestimation of the true value across the five observed datasets. This underestimation was also seen in the time of admixture, and despite relatively high variance, was seen to a lesser extent in the time of split. Although this was a testing scenario and therefore only small numbers of simulations were applied, this appears to be a successful method of determining admixture proportions for a single event. However, the aim of this project was to develop methodology to interpret and describe cases of multiple admixture events and as such the single admixture application was designed for the verification of the initial approach. For this reason the further exploration of the single admixture scenario was curtailed and the

testing expanded to the full two admixture event scenario. The estimation of the parameters for the two admixture events model showed less consistent results over the five observed datasets tested here. Even though it is difficult to make strong predictions about the parameters which will be best estimated, it seemed reasonable to think that the estimation of the 1-p3 proportion corresponding to the most recent admixture event would provide the best results. Indeed, due to the reduced time period over which allele frequencies can change from their parental frequency distribution this parameter should be easier to estimate. It is difficult to draw strong conclusions from only 5 independent runs but the current results suggest that this parameter could be underestimated, at least when the mean is used as a point estimator. If that were the case this could suggest a bias toward 0.5 which has been described in the literature as a failure of methods to determine admixture proportions accurately particularly in cases where parental populations are undifferentiated (Bertorelle and Excoffier, 1998). Indeed, if there is little information in the data, the posterior mean will tend to be equal to the prior mean which is 0.5. The 1-p3 parameter variance was still high, averaging around 0.1. Whereas the estimation of p3 was fairly consistent over the five datasets there were bigger differences across the runs for the p1 parameter with estimates ranging from 0.45 to 0.86 (true value 0.7) with an average variance of 0.13.

The reduction of efficiency of parameter estimation by adding the second admixture event is expected simply because it increases the number of parameters. It also suggests that the number of simulations used here is not large enough for the number of loci applied and the scenario used. In an approximate Bayesian approach applied by Hamilton et al. (2005) they suggest that to predict three parameters one million simulations is sufficient, greatly reducing variance of estimated parameters in comparison to fewer simulations. The full model applied here, when the population sizes and mutation rate are not fixed, attempts to estimate eleven parameters. I chose to relax the previously fixed parameters to analyse the performance of the approach in a full scenario where all parameters are allowed to vary. This highlighted that as parameters such as theta vary, so the memory demand increases as more data are required to describe the variation present. What was evident from the results of the 500,000 simulation runs was that, as expected, the consistency declined from the previous run at a lower number of simulations. It is intuitive that the introduction of variation at more parameters means that far fewer simulations closely match the

state of the data (calculated through the summary statistics) despite the five-fold increase in simulated data used. The inferential results for these new runs are particularly poor and highlight the need to significantly increase the number of simulations in a 'full parameter' scenario. The mean regression histograms showed that the previous scenario was achieving consistent parameter values (Appendix 5.4), in the current scenario the posterior is not very different from the prior (Appendix 5.5). As can be seen from the plots of an additional run with 20 loci (Figure 5.5) there are suggestions that the estimates of the 1-p3 parameter were beginning to improve. Although still insufficient for consistent and accurate quantification, this shows that increasing the numbers of loci, and particularly numbers of simulations, will make an important difference to the parameter estimation as seen in other Bayesian methods (Excoffier et al. 2005).

5.6.1. Application to cattle data

Whilst the cattle data used here provided a test case for the application of the new methodology to real admixture scenarios, it is important to show that accurate estimation is possible even if longer runs are required for reliable data. The results of the admixture determination were inherently interesting, but must be accepted as being test data and therefore not yet a reliable and definitive description. It is tempting at this stage to declare that the results show a potentially accurate description of the history of the Lincoln Red Wardle Herd, for example. The admixture proportions are suggestive that, allowing for some degree of bias toward 0.5 for the admixture proportions, there is a 0.11(+/-0.18) contribution of Aberdeen Angus and 0.25 (+/-0.25) contribution by Limousin. However, the results for the Wardle herd are placed into perspective by that of the McTurk herd, expected only to have slightly differing contributions but instead suggested a completely different, almost opposite, proportion of the ancestral Shorthorn relative to Aberdeen Angus contribution. It may be logical to restrict interpretation of only the more recent admixture event having seen the results for the 20 locus dataset in Figure 5.5, and to ignore the older admixture contribution. A higher relative Limousin proportion is attributable in the McTurk herd as might be expected but caution must clearly be exercised in this interpretation.

The Dexter application of this admixture method is as diverse as might be expected for the two sets of loci. If the interpretation is restricted to the more recent admixture

event and the two results are made into an average the result is an Aberdeen Angus contribution similar to that of the Limousin contribution in the McTurk herd, 0.64 (+-~0.19). As the McTurk and Dexter breed predicted proportions are similar, the similarity of their actual results may reflect the accuracy of the methodology at this number of simulations. However, this interpretation should be reserved until further information can support it. One current interpretation is that whilst there does appear to be a degree of estimation of the true admixture contributions (parameters) there is still a high associated inaccuracy.

5.6.2. Appraisal of the methodology

The change in analysis between using simulated observed data and the actual collected data is potentially important. The ms program generated data according to the input command and relies on a model on which to base calculations. This means that the observed data are generated using exactly the same process as the simulated data against which it is being compared. This means that there are no differences in the evolutionary processes between observed and simulated data. As the model under which the ms application generates its samples is an approximation of evolutionary processes this will differ to some degree from the processes under which the observed data collected from our cattle populations was generated.

To achieve accurate admixture quantification the bounds of the parameters, within which the simulated data will vary, should be small. More accurate knowledge of the scenario will therefore allow narrower parameter margins and therefore higher accuracy of admixture estimates. A consequence of this is that to have large prior distribution margins will allow for a better testing scenario but will mean fewer simulated datasets closely reflecting the observed data. In order to maintain the same accuracy of a narrower parameter distribution more simulations are therefore necessary. In all of the simulations used in this study there was a uniform distribution of prior parameters and it is possible that a more intuitive use of parameter distribution may have been made. A gamma distribution has been shown for mutation rates in some genomic regions, for example (Schneider and Excoffier, 1999).

It must be reinforced that the testing process is a dynamic one and whilst this represents the culmination of a large number of previous program versions more testing can always be performed. The reality of the process is that analysis will be tailored to suit the users. It is strongly suggested by the results presented here that the method is increasingly effective at higher numbers of simulations. Test simulations were performed on a machine with a 1.7 gigahertz processor and 512 megabytes of RAM. Due to the relatively high specifications of new machines, it would be expected that the increase in accuracy of the method could be great indeed and it would be recommended that numbers of simulations were increased to the order of around 10 million (numbers prohibitively high to use on the test machine). Furthermore, as stated previously, accuracy of admixture determination can be improved with increased numbers of loci, but again at the cost of calculation time. The importance of the complete testing of an application of this type is paramount to the removal of any bugs and mistakes in the script. This may require the comparison against hundreds of observed datasets and not just the five employed here. From the simulations completed is it expected that the levels of accuracy in parameter prediction can be increased to a point where it will be able to be effectively employed in real data scenarios. Whilst the use of the method in a real data scenario may be considered premature in this case, it provided reassurance that the application is functional (if to a lower degree of effectiveness at the levels of simulations shown) and should be able to be implemented across a range of admixture scenarios. It is worth mentioniong here that in principle it should be possible to simulate data varying number of loci and using admixture scenarios that are close to the scenarios that are likely for the species of interest. This way, one could determine the minimum number of loci necessary to have precise estimation for the species the user is working on. For instance, for the cattle breeds analysed here we could study the properties of the ABC scheme, only for values of the parameters that are “realistic”. This way, we would know whether it is worth genotyping more loci to get reasonably good estimates.

The development of this program is the present position of work begun towards the end of this PhD thesis. As described, the application is currently slow and at present has not been tested fully or for as many simulations as would be desired. It is my intention to complete this work to construct an accurate and fully functional

multiple-admixture determination method and to describe it to a publication standard. This process is ongoing and more simulations are currently being run.

5.7. Acknowledgements

Much of the R script and advice for this program development was produced in collaboration with Vitor Sousa and Lounès Chikhi as well as much of the ideas and implementation of the final version. The original program version was based on a similar framework developed by Marie Trussart in conjunction with Lounès Chikhi. Additional advice and script improvement was contributed by Barabara Pereira. The program script was altered from the initial template by the author with suggestions and troubleshooting by Vitor Sousa. Subsequent script improvement and changes to the analytical stage as well as output files were made by Vitor Sousa including additional code improvements from Barbara Pereira. Longer testing simulations were made by Vitor Sousa.

5.8. References

Beaumont M.A., Zhang W., & Balding D. (2002) Approximate Bayesian Computation in population genetics. *Genetics* **162**, 2025-2035.

Belkhir K, Borsa P, Chikhi L., Raufaste N., & Bonhomme F. (2002) GENETIX 4.03, Logiciel Sous Windows^{TM} Pour la Génétique Des Populations. Laboratoire Génome Populations, Interactions. CNRS UMR 5000, Université de Montpellier II, Montpellier, France.

Bertorelle G., & Excoffier L. (1998) Inferring admixture proportion from molecular data. *Molecular Biology and Evolution* **15**, 1298–1311.

Bertorelle G., & Excoffier L. (1998) Inferring admixture proportion from molecular data. *Molecular Biology and Evolution* **15**, 1298–1311.

Chikhi L., Bruford M. W., & Beaumont M.A. (2001) Estimation of admixture proportions: A likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**: 1347-1362.

Excoffier L., Estoup A., & Cornuet J-M. (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* **169**, 1727-1738.

Hamilton G., Currat M., Ray N., Heckel G., Beaumont M., & Excoffier L. (2005) Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* **170**, 409-417.

Hudson R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**(2), 337-338.

Ihaka R, & Gentleman R. (1996) R: A Language for Data Analysis and Graphics Journal of Computational and Graphical Statistics, Vol. 5, No. 3 pp. 299-314.

Lincoln Red Herd Book (1896) Lincoln Red Cattle Society.

Litzkow M.J., Livny M., & Mutka M.W. (1988) Condor – a hunter of idle workstations. *Distributed Computing Systems, 8th International Conference* **13-17**, 104-111.

Markham G. (1695): *A way to get wealth*. London.

Pritchard J.K., Stephens M., & Donnelly P.J. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Schneider S., & Excoffier L. (1999) Estimation of past demographicparamters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitcchondrial DNA. *Genetics* **152**, 1079-1089.

The MathWorks Inc. (2001) *MATLAB 7*, 24 Prime Park Way, Natick MA, 1997.

Chapter 6.
General discussion

6.1. Contemporary population studies and the Dexter breed

Despite well recorded historical and pedigree information for modern traditional cattle breeds there is a great deal of cryptic variation that exists in these domestic populations. The aim of this thesis was to reveal in detail a particular case of introgression in the Dexter minority cattle breed. To accomplish this, the Dexter was first considered in the context of a wide set of cattle breed populations in Europe. It was previously assumed that influences in the Dexter breed were from several breed backgrounds, and this is reflected in its genetic diversity being amongst the highest in the wider European breed dataset. There have been a number of recent comparable studies of cattle breeds in Europe (Blott et al. 1998a; Loftus et al 1999; MacHugh et al 1994; Martin-Burriel et al. 1999; Peelman et al. 1998; Wiener et al., 2004 etc.) but unlike these the breeds included here were carefully chosen to be a set of traditional breeds. As such it was hoped to better reflect those with which the Dexter may have undergone genetic exchange. However, the Dexter did not display close associations with any of the breed populations in this breed dataset. Its position in the wider European context demonstrates that the Dexter has maintained a genetic uniqueness throughout the period of the herd book remaining open to appendixed animals from an out-breeding register. However, there was some information from the migrant analysis that indicated geneflow between breeds. Dexter individuals were reassigned to several breed populations; Milking Devon, Kerry, Guernsey, Mainland Jersey, Shorthorn, and Sussex. But these few migrants constituted a small proportion of the population (less then ten percent). Indeed, the proportion of correctly assigned individuals in the Dexter was actually higher than over half of the other breeds in the study. The suggestion of this analysis was that a more detailed view of the Dexter breed was required, in a narrower breed scenario.

The Dexter was then analysed specifically to examine its relationship with two historically associated breeds; the Devon and Kerry. More microsatellite loci were added to improve the genetic information available and it was possible to resolve that the Dexter was subdivided in a way not seen in the previous analysis. As the known associations of the Dexter with the Kerry and Devon were not represented in the wider breed analysis, it is thought that this may be a consequence of this highly variable composition. The variability of the Dexter itself meant that there were a number of distinct breed subgroups one of which, the Woodmagic herd, was

differentiated from other Dexter groups to a degree equal to that expected at a between-breed level. This subdivision of breed populations is usually only seen where the breed is split between separate populations in different countries (e.g. Achmann et al. 2004). There appears to be a paucity of research below the breed level, although one good example is that of the Hereford breed of cattle (Blott et al. 1998b). But generally even research specifically aimed at the analysis of subdivided populations is not seen to examine genetic structure or diversity below the breed level (Caballero & Toro, 2002). An explanation of this subdivision in the Dexter is potentially high levels of genetic drift. Similar gene identity measures to that seen in the Dexter Woodmagic herd population were also found in a fish population isolated in a single location with limited population size and also implicated as having undergone high levels of inbreeding (Bailey et al. 2007). Although the Woodmagic population has only been isolated for ~25 years the line-breeding process and small population size may have amplified the effects of genetic drift.

The analysis of the Dexter relationship with the Kerry and Devon breeds was modelled as an admixture event using several methods to estimate the parental contributions. The analysis was restricted to populations known to be good representatives of those involved in the admixture of the Dexter breed, and was performed using specific admixture applications (LEA (Chikhi et al. 2001), ADMIX2 (Dupanloup and Bertorelle, 2001), and LEADMIX (Wang 2003)) rather than the clustering algorithms commonly applied to make geographical admixture inferences (e.g. in Freeman et al. 2006). The application of specifically designed admixture approaches was intended to provide more accurate proportion estimates through specifically accounting for the effects of different evolutionary processes, such as genetic drift in LEA (Chikhi et al. 2001). The expected closer association of the Dexter with the Kerry (relative to the Devon) was suggested by the allele frequency based clustering approach of Pritchard et al. (2000), but the specific admixture determination methodology was not so clear cut. The admixture analysis methods demonstrated a clear difficulty in determining the proportional contribution of the putative ancestors. Parental proportions were calculated with large associated uncertainty and the high levels of genetic drift that are likely to be acting in these small populations are thought to be the cause of such inaccurate admixture determinations.

The high levels of genetic variation uncovered in the Dexter breed has had both potentially positive and negative effects in terms of its genetic analysis. Firstly, the Dexter appears to have maintained a high level of genetic uniqueness when compared to many other breed populations. This has been manifested in relatively high differentiation between it and other breeds, but particularly within the breed itself. Secondly, a greater potential for objective fine-scale admixture analysis is revealed through the subdivision of the breed into discrete genetically-similar groups. This has allowed investigation into relationships between distinct subdivisions within the breed. However, a third consequence is that demographic perturbation and isolation is likely to have obscured admixture proportions, particularly older contributions.

In light of the information provided by this study, management of the Dexter needs to be addressed. It is important that a few males or subpopulations of the herd book are not over-represented and this is particularly pertinent if the individuals in question originate from a genetically divergent subpopulation. The Woodmagic population in this study might be an example of this over-representation as potentially desirable characteristics are sought from a small number of males. Whilst it would be a loss of the unique allelic spectra of the subpopulation if it were not perpetuated, the clear genetic distinction of the Woodmagic individuals is such that they have the potential to disproportionately alter the composition of the wider breed. One recommendation might be to limit the use of any individuals whose semen is commercially available and perhaps to enforce a further limit where multiple males are available from a restricted herd population.

6.2. Admixture modelling

In order to attempt to provide further insight into the population dynamics of the Dexter breed, a new method was required that could overcome the problems found to be confounding admixture determination. It was theorised that accounting for multiple admixture events in a new approach may be a valid way to do this. In developing a novel method in which to investigate multiple admixture events which occur at different times we begin to open out the possibilities of more flexible applications. The difficulties in developing methodology begin with the description of, and adherence to, a particular model. The benefits to model development of recent methods and applications can be seen in programs such as ms or 'make sample'

(Hudson, 2002). The efficient use of data in population genetic algorithms can be approached in different ways, the method here is that of approximating a posterior distribution through corresponding properties of the sample as described by Beaumont et al. (2002). In this way a set of summary statistics is calculated and used as a means of comparing datasets in accordance to a rejection sampling method as in Tavaré et al. (1997). An approach is dependent on its constituent summary statistics to accurately reflect the underlying genetic processes involved in the formation of each genetic sample. It is always possible that the choice of a different or extended set of summary statistics could improve the performance of this method. Unfortunately there is an associated calculation cost to additional functions and data that must be stored in memory or saved to file. The function of the software developed here is of limited use at the numbers of simulations and loci applied here (~500,000 simulations, nine loci). It is thought that there is no technical or methodological reason that the results seen in the runs for which many of the parameters are fixed, might not be replicated in a full scenario where all parameters are varied. The process of population genetic software development is a continual one of updating and testing. Although now a complete script, the developed method for admixture detection needs testing across more scenarios, and notably in longer runs.

6.3. Domestic animal conservation genetics; present and future

As seen across much of population biology, typical breed characterisation and diversity studies have shown a trend of increasing levels of marker application from single figures (Blott et al. 1998a) to tens of marker loci applied (Zhang et al. 2007). Matching this is the increased scale of study which can encompass breed assemblages of hundreds of individuals across entire regions e.g. West Africa in Thévenon et al. (2007). A result of the increasing information available in each study has been a shift in emphasis towards analysis of demographic influences on breed structure such as population bottlenecks (Fatima et al. 2008). It is also increasingly notable that comparative applications can now be made between genetic variation studies on the same breeds with the availability of marker systems as suggested through organisations such as the Food and Agriculture Organisation (FAO). The comparison between studies can be important for the interpretation of genetic data, perhaps more so in minority domestic breeds where ascertainment bias could be particularly influential and potentially cause large discrepancies between studies.

The advent of assignment algorithms and quantification of migration has changed the way that breeds are analysed, illuminating both exchanges of migrants and phylogenetic associations between regions (Loftus et al. 1999). The analysis of detailed scenarios where migration or population subdivision is important will become more common in the literature as methods evolve to make better use of the information available. It is now possible to identify clines of upgrading crosses for increased breed productivity within a region, for example (Li et al. 2007). Assignment criteria can also be employed in situations where animal products need to be verified for reasons of food safety and authenticity (Negrini et al. 2007). The recent development of approximate Bayesian methods (Beaumont et al. 2002) is just one example of the innovations in theory that are driving the advancement of population genetic analysis. The identification of analytical limitations continues to create new areas for study, particularly so where high computational demand was the previously limiting factor.

There is also an apparent urgency for conservation action promoted by conservation assessments in domestic animals (Taberlet, 2008). The climate of increasing environmental awareness and sustainability has begun to raise the profile of minority breed populations amidst the threat of extinction to make way for their intensively produced counterparts (Tisdell, 2003). Recognition of continued threats to populations maintains the discussion of prioritisation of funds for conservation (Reist-Marti et al 2005; Tapio et al 2006; Reist-Marti et al 2006). It seems likely that increasingly accurate methods of assessing domestic breed survival probability will continue to play a part in the literature including new approaches to assessment of the genetic value of individual breed populations. It is also likely that the results of restorative breed management programs will begin to appear as a consequence of the continued recognition of the detrimental legacy of management of small populations and artificial selection programs.

6.4. Conclusions and future work

Not only are the population genetics of domestic breeds complex, they have been found here to be even more so than previously predicted. The possibilities for future genetic investigation of the Dexter are consequently significant. It has been established here that extensive population structure and subdivision exists across the

Dexter breed. The work done here will provide an excellent basis for subsequent analysis at the population genetic level. Opportunities exist for investigating phenotypic characters and genes affecting production and animal welfare such as the Major Histocompatibility Complex. The disruptive effects of anthropogenic selection acting according to each breeding unit makes the breed more dynamic than most, as does the diversity of forms of the Dexter, making it a unique breed. Genetic traits under breeders' selection include colour forms, leg length, and even presence or absence of horns. More detailed genetic work may shed further light on the complexities of the admixture history of the breed and perhaps on a relationship to some of these morphological characters. The difficulty of quantification of the introgression from even the major contributors in the breed's history means that minor influences from other breeds are still unknown. The rapid progression of technology and methodology suggests that admixture studies of this kind may become increasingly effective in the future. The development of the methodology detailed here will go some way towards achieving a solution to the problem of admixture in complex demographic scenarios. There exist many potential extensions and improvements that can be made to the current application as the need and opportunity arises. The consideration of two admixture events could potentially be increased to three as the application evolves and the computational demands can be absorbed by more powerful processors or serialisation of run jobs. Binary file formats to speed up the calculations is just one possibility that could be implemented along with other associated applications to generate input files from popular file formats.

6.5. References

Achmann R., Curik I., Dovc P., kavar T., Bodo I., Habe F., Marti E., Solkner J., & brem (2004) Micrisatellite diversity, population subdivision and gene flow in the Lipizzian horse. *Animal Genetics* **35**(4), 285-292.

Bailey NW, Macias Garcia C, and Ritchie MG (2007) Beyond the point of no return? A comparison of genetic diversity in captive and wild populations of genetic diversity in captive and wild populations of two nearly extinct species of Goodeid fish reveals that one is inbred in the wild. *Heredity* **98**, 360–367.

Beaumont M.A., Zhang W., & Balding D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035.

Blott S.C., Williams J.L., & Haley C.S. (1998a) Genetic relationships among European cattle breeds. *Animal Genetics* **29**, 273-282.

Blott S.C., Williams J.L., & Haley C.S. (1998b) Genetic variation within the Hereford breed of cattle. *Animal Genetics* **29**, 202-211.

Caballero A., & Toro M.A. (2002) Analysis of genetic diversity for the management of conserved subdivided populations. *Conservation Genetics* **3**, 289-299.

Dupanloup I., & Bertorelle G. (2001) Inferring admixture proportions from molecular data: extension to any number of parental populations. *Molecular Biology and Evolution* **18**(4), 672-675.

Fatima S. Bhong C.D., Rank D.N., & Joshi C.G. (2008) Genetic variability and bottleneck studies in Zalawadi, Gohilwadi and Surti goat breeds of Gujarat (India) using microsatellites. *Small Ruminant Research* **77**(1), 58-64.

Freeman A.R., Bradley D.G., Nagda S., Gibson J.P., & Hanotte O. (2006) Combination of multiple microsatellite data sets to investigate diversity and admixture of domestic cattle. *Animal Genetics* **37**(1), 1-9.

Hudson R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**(2), 337-338.

Li M-H, Tapio I, Vilkki J et al (2007) The genetic structure of cattle populations (*Bos taurus*) in northern Eurasia and the neighbouring Near Eastern regions: implications for breeding strategies and conservation, Mol Ecol 16:3839-3853.

Loftus R., Ertrugrul O., Harba A., El-Barodys M., MacHugh D., Park S., & Bradley D. (1999) A microsatellite survey of cattle from a centre of origin: the near east. *Molecular Ecology* **8**, 2015-2022.

MacHugh D.E., Shriver M. D., Loftus R.T., Cunningham, P., & Bradley D.G. (1997) Microsatellite DNA variation and the evolution, domestication and phylogeography of Taurine and Zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* **146**, 1071-1086.

Martin-Burriel I., Garcia-Muro E., & Zaragoza P. (1999) Genetic diversity analysis of six Spanish native cattle breeds using microsatellites. *Animal Genetics* **30**, 177-182.

Negrini R., Milanesi E., Colli L., Pellecchia M., Nicoloso L., Crepaldi P. Lenstra J.A., & Ajmone-Marsan P. (2007) Breed assignment of Italian cattle using biallelic AFLP(R) markers. *Animal Genetics* **38**(2), 147-153.

Peelman L.J., Mortiaux F., Van Zeveren A., Dansercoer A., Mommens G., Coopman F., Bouquet Y., Burny A., Renaville R., & Portetelle D. (1998) Evaluation of the genetic variability of 23 bovine microsatellite markers in four Belgian cattle breeds. Animal *Genetics* **29**(3), 161-167.

Pritchard JK, Stephens M, and Donnelly P (2000) Inference of population structure using multilocus genotype data. Genet 155:945-959.

Reist-Marti S.B., Abdulai A., & Simianer H. (2005) Conservation programmes for cattle: design, cost and benefits. *Journal of Animal Breeding Genetics* **122**, 95-109.

Reist-Marti S.B., Abdulai A., & Simianer H. (2006) Optimum allocation of conservation funds and choice of conservation programs for a set of African cattle breeds. *Genetics Selection and Evolution* **38**, 99-126.

Taberlet P., Valentini A., Rezaei H. R., Naderi S., Pompanon F., Negrini R., & Ajmone-Marsan P. (2008) Are cattle, sheep, and goats endangered species? *Molecular Ecology* **17**(1), 275-284.

Tapio I., Varv S., Bennewitz J., Maleviciute J., Fimland E., Grislis Z., Meuwissen T.H.E., Miceikiene I., Olsaker I., Viinalass H., Vilkki J., & Kantanen J. (2006) Prioritisation of northern European cattle breeds based on analysis of microsatellite data. *Conservation Biology* **20**(6), 1768-1779.

Tavaré S., Balding D.J., Griffiths R.C., & Donnelly P. (1997) Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505-518.

Thévenon S. Dayo G.K., Sylla S., Sidibe I., Berthier D., Legros H., Boichard D., Eggen A., amd Gautier M. (2007) The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies. *Animal Genetics* **38**, 277-286.

Tisdell C. (2003) Socioeconomic causes of animal genetic diversity:analysis and assessment. *Ecological Economics* **45**(3), 365-376.

Wang J. (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**, 747-765.

Wiener P., Burton D., & Williams J.L. (2004) Breed relationships and definition in British cattle: a genetic analysis. *Heredity* **93**, 597-602.

Zhang G.X., Wang Z.G., Chen W.S., Wu C.X., Han X., Chang H., Zan L.S., Li R.L., Wang J.H., Song W.T., Xu G.F., Yang H.J., & Luo Y.F. (2007) Genetic diversity a population structure of indigenous yellow cattle breeds of China using 30 microsatellite markers. *Animal Genetics* **38**(6), 550-559.

Appendices

endix 2.1. Dexter identifiers including farm and county of origin, arranged by ple group.

Animal code	Farm code	County	Analysis group
s124,	43	Devon	Woodmagic
11,	43	Devon	Woodmagic
143,	43	Devon	Woodmagic
149,	43	Devon	Woodmagic
52,	43	Devon	Woodmagic
53,	43	Devon	Woodmagic
54,	43	Devon	Woodmagic
55,	43	Devon	Woodmagic
56,	43	Devon	Woodmagic
57,	43	Devon	Woodmagic
58,	43	Devon	Woodmagic
59,	43	Devon	Woodmagic
60,	43	Devon	Woodmagic
61,	43	Devon	Woodmagic
62,	43	Devon	Woodmagic
63,	43	Devon	Woodmagic
64,	43	Devon	Woodmagic
65,	43	Devon	Woodmagic
s6,	43	Devon	Woodmagic
s13,	43	Devon	Woodmagic
s42,	43	Devon	Woodmagic
7,	12	Somerset	Traditional
8,	12	Somerset	Traditional
s1,	42	Monmouthshire	Traditional
s2,	45	Cumberland	Traditional
s4,	20	Cumberland	Traditional
s15,	32	Warwickshire	Traditional
s36,	37	Devon	Traditional
s45,	40	Yorkshire	Traditional
s47,	32	Warwickshire	Traditional

s51,	9	Pembrokeshire	Traditional
119,	27	Norfolk	Traditional
149,	3	Shropshire	Traditional
s66,	26	Berkshire	Breed
s68,	9	Pembrokeshire	Breed
66,	31	Oxfordshire	Breed
68,	31	Oxfordshire	Breed
69,	31	Oxfordshire	Breed
70,	31	Oxfordshire	Breed
165,	23	Warwickshire	Breed
176,	8	Hampshire	Breed
s46,	2	Buckinghamshire	Breed
6,	12	Somerset	Breed
21,	12	Somerset	Breed
48,	34	Dorset	Breed
49,	35	Surrey	Breed
50,	36	Wiltshire	Breed
51,	14	Devon	Breed
141,	28	Somerset	Breed
142,	15	Camarthenshire	Breed
143,	15	Camarthenshire	Breed
151,	37	Devon	Breed
159,	8	Hampshire	Breed
160,	8	Hampshire	Breed
163,	10	Gloucestershire	Breed
s5,	16	N.Ireland	Breed
s26,	5	Devon	Breed
s29,	41	Dorset	Breed
s40,	22	Devon	Breed
s43,	37	Devon	Breed
s54,	41	Dorset	Breed
2,	4	Essex	Breed
28,	1	Berkshire	Breed
31,	24	Bedfordshire	Breed
39,	24	Bedfordshire	Breed

40,	12	Somerset	Breed
45,	44	Somerset	Breed
46,	44	Somerset	Breed
s7,	7	Sussex	Breed
s8,	31	Oxfordshire	Breed
s11,	31	Oxfordshire	Breed
s34,	37	Devon	Breed
s39,	13	Gloucestershire	Breed
s41,	44	Somerset	Breed
167,	42	Monmouthshire	Breed
171,	31	Oxfordshire	Breed
172,	31	Oxfordshire	Breed
173,	31	Oxfordshire	Breed
174,	31	Oxfordshire	Breed
175,	31	Oxfordshire	Breed
176,	8	Hampshire	Breed
177,	37	Devon	Breed
178,	25	Hampshire	Breed
179,	21	Somerset	Breed
180,	17	Northamptonshire	Breed
181,	17	Northamptonshire	Breed
182,	38	Powys	Breed
183,	38	Powys	Breed
184,	46	Wiltshire	Breed
185,	46	Wiltshire	Breed
186,	46	Wiltshire	Breed
187,	24	Bedfordshire	Breed
188,	17	Northamptonshire	Breed
189,	11	Somerset	Breed
190,	11	Somerset	Breed
191,	24	Bedfordshire	Breed
192,	24	Bedfordshire	Breed
193,	33	Kent	Breed
194,	29	Kent	Breed
195,	29	Kent	Breed

196,	17	Northamptonshire	Breed
197,	6	Cheshire	Breed
198,	8	Hampshire	Breed
199,	31	Oxfordshire	Breed
200,	19	Dorset	Breed
201,	31	Oxfordshire	Breed
203,	29	Kent	Breed
208,	31	Oxfordshire	Breed
209,	31	Oxfordshire	Breed
210,	31	Oxfordshire	Breed
211,	31	Oxfordshire	Breed
212,	31	Oxfordshire	Breed
214,	45	Cumberland	Breed
244,	24	Bedfordshire	Breed
245,	39	Powys	Breed
246,	40	Yorkshire	Breed
250,	18	Worcestershire	Breed
254,	45	Cumberland	Breed
393,	21	Somerset	Breed
406,	21	Somerset	Breed
407,	12	Somerset	Breed
408,	12	Somerset	Breed
s1,	42	Monmouthshire	Breed
s100,	30	Somerset	Breed
s101,	24	Bedfordshire	Breed
s123,	33	Kent	Breed
s125,	33	Kent	Breed
109,	45	Cumberland	Ypsitty
217,	45	Cumberland	Ypsitty
218,	45	Cumberland	Ypsitty
219,	45	Cumberland	Ypsitty
231,	45	Cumberland	Ypsitty
232,	45	Cumberland	Ypsitty
233,	45	Cumberland	Ypsitty
234,	45	Cumberland	Ypsitty

235,	45	Cumberland	Ypsitty
240,	45	Cumberland	Ypsitty
241,	45	Cumberland	Ypsitty
242,	45	Cumberland	Ypsitty
243,	45	Cumberland	Ypsitty
405,	-	-	American
409,	-	-	American
220,	-	-	American
221,	-	-	American
222,	-	-	American
223,	-	-	American
224,	-	-	American
225,	-	-	American
226,	-	-	American
227,	-	-	American
228,	-	-	American
229,	-	-	American
230,	-	-	American

Appendix 3.1. F_{IS} values of European cattle per breed and locus

Marker	Shetland	Aberdeen Angus	Belted Galloway	Irish Moiled	Shorthorn	Lincoln Red	British White	Kerry	Traditional Hereford	Red Poll	Welsh Black	Dexter	Sussex	White Park
INRA063	0.18	-0.09	0.68	0.33	-0.16	0.03	-0.28	0.32	-0.13	-0.41	0.10	0.03	-0.06	0.15
INRA005	0.21	------	-0.08	-0.25	------	------	-0.23	0.02	0.05	------	------	------	-0.03	-0.35
ETH225	-0.07	-0.06	0.13	-0.16	-0.18	0.00	0.46	0.16	0.01	0.00	-0.04	0.01	-0.06	0.07
HEL5	0.24	0.03	0.30	-0.04	0.05	-0.10	-0.30	-0.04	0.54	-0.16	0.03	-0.05	-0.02	0.37
ETH10	-0.01	0.07	0.17	-0.23	0.14	-0.03	-0.21	-0.02	0.48	0.06	0.30	0.00	0.49	-0.14
BM2113	0.06	------	0.44	-0.07	------	------	0.43	-0.02	-0.01	------	------	------	0.03	0.21
BM1818	0.01	0.14	0.05	-0.10	-0.16	-0.14	0.21	-0.25	0.49	0.14	-0.01	0.11	-0.09	0.27
ILSTS006	-0.06	0.15	0.03	0.04	0.13	0.02	-0.07	-0.11	-0.07	-0.11	0.08	-0.03	0.00	-0.12
HAUT27	0.15	0.16	-0.13	0.21	0.23	0.09	-0.36	0.08	-0.01	0.04	0.12	0.01	-0.01	0.20
TGLA227	-0.05	-0.05	-0.13	0.36	-0.18	-0.14	0.63	0.02	-0.22	-0.18	-0.04	0.00	0.38	0.00
TGLA122	0.08	0.04	-0.11	0.22	0.08	0.02	0.34	0.21	0.09	0.04	-0.07	0.10	-0.13	-0.06
BM1314	-0.14	------	0.05	-0.09	------	------	0.17	0.27	-0.03	------	------	------	-0.02	0.19
Average	0.05	0.04	0.12	0.02	-0.01	-0.03	0.07	0.05	0.10	-0.06	0.05	0.02	0.04	0.07

Appendix 3.1. (continued)

Marker	Gloucester	Milking Devons	Red Devon	Mainland Jersey	Guernsey	Island Jersey	Jutland	Angeln	German Black Pied	Hungarian Grey	Limousin	Berrenda	N'dama
INRA063	0.57	-0.19	0.13	0.04	0.35	0.08	-0.25	-0.04	0.08	-0.06	-0.35	0.06	-0.46
INRA005	-0.01	------	------	------	------	0.25	-0.29	-0.08	0.18	0.09	-0.12	0.18	-0.07
ETH225	0.18	-0.06	0.20	0.02	0.49	-0.07	-0.17	0.02	-0.17	-0.32	-0.11	0.05	0.05
HEL5	-0.15	0.03	0.06	0.07	-0.02	0.15	0.15	0.07	-0.13	0.39	-0.09	-0.02	0.19
ETH10	-0.19	0.04	-0.11	-0.03	-0.27	-0.06	-0.16	-0.01	-0.07	-0.05	0.12	0.00	-0.40
BM2113	-0.19	------	------	------	------	-0.21	0.19	0.09	-0.14	0.11	0.19	0.23	-0.11
BM1818	-0.13	0.11	-0.02	0.13	0.09	0.13	0.03	0.18	-0.10	0.18	-0.03	0.26	-0.23
ILSTS006	0.19	0.04	-0.06	0.02	0.07	-0.03	0.25	-0.07	0.31	0.13	0.01	0.11	-0.06
HAUT27	-0.04	-0.02	-0.04	-0.03	0.08	0.15	-0.14	-0.04	-0.02	-0.13	-0.03	0.41	0.17
TGLA227	1.00	-0.15	-0.01	0.27	0.20	-0.13	0.23	0.21	0.00	0.03	-0.06	0.22	0.07
TGLA122	------	0.01	-0.09	-0.11	0.06	-0.06	0.21	0.12	0.10	0.19	0.37	0.16	-0.07
BM1314	-0.06	------	------	------	------	-0.06	0.23	0.11	0.06	0.18	-0.03	0.14	-0.10
Average	0.11	-0.02	0.01	0.04	0.12	0.01	0.02	0.05	0.01	0.06	-0.01	0.15	-0.08

Appendix 3.2. Weir and Cockerham (1984) Pairwise F_{ST} values (upper triangle) with their respective significance levels (lower triangle).

FST	1	2	3	4	5	6	7	8	9	10	11	12	13	14
(1)SHETL		0.19	0.30	0.24	0.15	0.20	0.24	0.16	0.27	0.17	0.21	0.19	0.21	0.31
(2)AANGU	***		0.20	0.15	0.07	0.11	0.19	0.10	0.23	0.11	0.06	0.09	0.16	0.27
(3)GALLO	***	***		0.30	0.22	0.21	0.23	0.23	0.21	0.20	0.21	0.21	0.23	0.20
(4)IMOIL	***	***	***		0.10	0.13	0.23	0.18	0.34	0.15	0.15	0.15	0.25	0.35
(5)SHHOR	***	***	***	***		0.06	0.14	0.11	0.23	0.09	0.09	0.11	0.21	0.29
(6)LRED	***	***	***	***	***		0.17	0.14	0.26	0.09	0.14	0.12	0.20	0.26
(7)BWHIT	***	***	***	***	***	***		0.18	0.18	0.22	0.21	0.22	0.21	0.29
(8)KERRY	***	***	***	***	***	***	***		0.22	0.17	0.10	0.11	0.14	0.27
(9)HEREF	***	***	***	***	***	***	***	***		0.26	0.25	0.24	0.22	0.23
(10)RPOLL	***	***	***	***	***	***	***	***	***		0.14	0.12	0.16	0.29
(11)WBLAC	***	***	***	***	***	***	***	***	***	***		0.11	0.21	0.25
(12)DEXTE	***	***	***	***	***	***	***	***	***	***	***		0.15	0.25
(13)SUSSE	***	***	***	***	***	***	***	***	***	***	***	***		0.28
(14)WPARK	***	***	***	***	***	***	***	***	***	***	***	***	***	
(15)GCEST	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(16)MDEVO	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(17)BDEVO	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(18)MJERS	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(19)GUERN	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(20)JERSE	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(21)JUTLA	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(22)ANGEL	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(23)GERMB	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(24)HGREY	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(25)LMOUS	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(26)BERRE	***	***	***	***	***	***	***	***	***	***	***	***	***	***
(27)NDAMA	***	***	***	***	***	***	***	***	***	***	***	***	***	***

Appendix 3.2. (continued)

FST	15	16	17	18	19	20	21	22	23	24	25	26	27
(1)SHETL	0.22	0.16	0.14	0.21	0.15	0.26	0.21	0.19	0.15	0.26	0.23	0.26	0.29
(2)AANGU	0.15	0.13	0.06	0.14	0.06	0.20	0.15	0.16	0.13	0.18	0.13	0.18	0.19
(3)GALLO	0.23	0.21	0.20	0.27	0.17	0.21	0.15	0.13	0.13	0.19	0.12	0.18	0.24
(4)IMOIL	0.22	0.19	0.16	0.21	0.16	0.29	0.22	0.24	0.20	0.30	0.26	0.27	0.29
(5)SHHOR	0.10	0.15	0.07	0.13	0.06	0.22	0.17	0.19	0.16	0.19	0.19	0.24	0.26
(6)LRED	0.12	0.17	0.12	0.19	0.09	0.25	0.19	0.21	0.19	0.24	0.22	0.27	0.28
(7)BWHIT	0.13	0.23	0.17	0.26	0.16	0.26	0.16	0.16	0.17	0.19	0.19	0.25	0.29
(8)KERRY	0.13	0.10	0.07	0.18	0.10	0.23	0.18	0.15	0.12	0.22	0.18	0.21	0.25
(9)HEREF	0.25	0.25	0.23	0.26	0.22	0.24	0.14	0.11	0.13	0.15	0.16	0.25	0.25
(10)RPOLL	0.18	0.14	0.12	0.15	0.05	0.24	0.19	0.20	0.16	0.24	0.19	0.23	0.26
(11)WBLAC	0.16	0.12	0.09	0.16	0.09	0.22	0.16	0.17	0.14	0.22	0.15	0.20	0.21
(12)DEXTE	0.13	0.14	0.10	0.12	0.07	0.22	0.19	0.18	0.17	0.23	0.17	0.20	0.21
(13)SUSSE	0.18	0.17	0.11	0.23	0.09	0.27	0.20	0.17	0.17	0.26	0.19	0.24	0.27
(14)WPARK	0.31	0.31	0.31	0.37	0.29	0.28	0.21	0.21	0.18	0.24	0.19	0.26	0.31
(15)GCEST		0.16	0.09	0.19	0.09	0.26	0.19	0.17	0.17	0.23	0.20	0.23	0.30
(16)MDEVO	***		0.09	0.16	0.09	0.20	0.19	0.18	0.11	0.24	0.15	0.19	0.19
(17)BDEVO	***	***		0.14	0.03	0.21	0.16	0.14	0.13	0.21	0.14	0.19	0.23
(18)MJERS	***	***	***		0.11	0.25	0.23	0.20	0.19	0.23	0.20	0.23	0.26
(19)GUERN	***	***	*	***		0.21	0.14	0.15	0.12	0.18	0.12	0.16	0.22
(20)JERSE	***	***	***	***	***		0.19	0.16	0.15	0.20	0.15	0.21	0.21
(21)JUTLA	***	***	***	***	***	***		0.08	0.09	0.17	0.10	0.18	0.16
(22)ANGEL	***	***	***	***	***	***	***		0.06	0.14	0.08	0.13	0.17
(23)GERMB	***	***	***	***	***	***	***	***		0.14	0.08	0.14	0.14
(24)HGREY	***	***	***	***	***	***	***	***	***		0.14	0.18	0.27
(25)LMOUS	***	***	***	***	***	***	***	***	***	***		0.08	0.13
(26)BERRE	***	***	***	***	***	***	***	***	***	***	***		0.20
(27)NDAMA	***	***	***	***	***	***	***	***	***	***	***	***	

Appendix 3.3. Assignment of individuals to populations and percentage of correct classification (P<0.0001) of the Bayesian approach in the program Geneclass 2.0 (Cornuet *et al.* 1999).

	total	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Correctly assigned
	31							1	1																				0.94
J	20					1			1			2								2									0.70
	15																										1		0.93
	20										1									1									0.90
?	11		1										1																0.82
	60		1			3												1		1									0.90
	11										1																		0.91
	32		1				1						1					3											0.81
	19																												1.00
L	20					2												1											0.85
.C	17		3			1			1																				0.71
E	154		3			2			1	1	1	2			1		3	2	1	5									0.86
E	50								1			1	3					1											0.88
?K	33																												1.00
T	14								1																				0.93

Appendix 3.3. (continued)

	total	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Correct assigne
)EVO	32								1				1					2		2									0.81
EVO	20		1			2					1			1			2			1									0.60
ERS	21												2							1									0.86
IERN	12		1								2		2					1											0.50
RSE	25																												1.00
TLA	17																							1					0.94
GEL	24			1																						2			0.88
:RMB	19			1																			3						0.79
;REY	15																												1.00
OUS	24																						1	1					0.92
RRE	31																									2			0.94
)AMA	9																									1			0.89
igned		0	11	2	0	11	1	1	7	1	6	5	10	1	1	0	5	11	1	13	0	0	4	2	0	5	1	0	

Appendix 3.4. F values for each breed groups as calculated in the ESTIM 1.0 software.

Breed	F
White Park	0.32
Irish Moiled	0.31
Traditional Hereford	0.26
Shetland	0.26
Gloucester	0.26
N'dama	0.24
Belted Galloway	0.23
Hungarian Grey	0.23
British White	0.22
Mainland Jersey	0.22
Sussex	0.21
Lincoln Red	0.20
Island Jersey	0.20
Kerry	0.20
Red Poll	0.17
Berrenda	0.16
Shorthorn	0.15
Welsh Black	0.15
Dexter	0.15
Jutland	0.14
Milking Devons	0.13
Angeln	0.13
Aberdeen Angus	0.12
Red Devon	0.11
Limousin	0.11
German Black Pied	0.08
Guernsey	0.07

Appendix 3.5. The presence of population bottlenecks and expansions in each population under the three models applied; Infinite Alleles, Two Phase, and Stepwise Mutation.

Breed	Bottleneck signal/significance (+=Heterozygote excess, -=Heterozygote deficiency | P<0.05,0.01,0.005 = *,**,*** respectively)		
	Infinite Allele Model	Two Phase Model	Stepwise Mutation Model
Shetland	***bottleneck		
Aberdeen Angus	*** bottleneck		
Belted Galloway	*** bottleneck	** bottleneck	
Irish Moiled	* bottleneck		
Shorthorn	*** bottleneck		
Lincoln Red	*** bottleneck		
British White	*** bottleneck		
Kerry	** bottleneck		*expansion
Traditional Hereford	*** bottleneck		
Red Poll	*** bottleneck	*** bottleneck	
Welsh Black			
Dexter			
Sussex	* bottleneck		
White Park	*** bottleneck		
Gloucester	*** bottleneck	*** bottleneck	* bottleneck
Milking Devons	*** bottleneck		
Red Devon	*** bottleneck	*** bottleneck	
Mainland Jersey	*** bottleneck		
Guernsey	* bottleneck		
Island Jersey	*** bottleneck	* bottleneck	
Jutland	*** bottleneck	*** bottleneck	* bottleneck
Angeln	*** bottleneck		
German Black Pied	*** bottleneck	**	
Hungarian Grey			
Limousin	*** bottleneck	* bottleneck	
Berrenda	*** bottleneck	** bottleneck	* bottleneck
N'dama			

Appendix 3.6. The regression of Marginal Diversity against Expected Heterozygosity for all 27 breed populations

	df	*SS*	*MS*	*F*	*Significance F*
ression	1	27.02491	27.02491	20.94095	0.000112
idual	25	32.26323	1.290529		
al	26	59.28814			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
rcept	15.12826	2.568199	5.890613	3.8E-06	9.83896	20.41757	9.83896	20.41757
ariable	-18.6858	4.083313	-4.57613	0.000112	-27.0955	-10.276	-27.0955	-10.276

Appendix 3.7. The regression of Marginal Diversity against ESTIM F values for all 27 breed populations

	df	*SS*	*MS*	*F*	*Significance F*
egression	1	30.97928	30.97928	27.35829	2.06E-05
esidual	25	28.30886	1.132354		
otal	26	59.28814			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
tercept	0.347949	0.621744	0.559633	0.58071	-0.93256	1.628455	-0.93256	1.62845
Variable	16.54727	3.163603	5.230515	2.06E-05	10.03171	23.06283	10.03171	23.0628

Appendix 5.1. Complete Approximate Bayesian Computation method script, inclusive of functions.

```
function varargout = abc_twoadmixture(varargin)
% See also: GUIDE, GUIDATA, GUIHANDLES
% Copyright 2002-2003 The MathWorks, Inc.
graphical = 1 % 1 to run graphical interface, zero otherwise
nb_obs = 1;
sim_by_file = 100000 %how many simultions in each file?

% Begin initialization code - DO NOT EDIT
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @abc_twoadmixture_OpeningFcn, ...
                   'gui_OutputFcn',  @abc_twoadmixture_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end
% End initialization code - DO NOT EDIT
% --- Executes just before abc_twoadmixture is made visible.
function abc_twoadmixture_OpeningFcn(hObject, eventdata, handles, varargin)
handles.output = hObject;

% Update handles structure
guidata(hObject, handles);

% UIWAIT makes abc_twoadmixture wait for user response (see UIRESUME)
% uiwait(handles.figure1);

% --- Outputs from this function are returned to the command line.
function varargout = abc_twoadmixture_OutputFcn(hObject, eventdata, handles)
% varargout  cell array for returning output args (see VARARGOUT);
% hObject    handle to figure
```

```
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Get default command line output from handles structure
varargout{1} = handles.output;
%=================== PARAMETERS ========================================
% --- Executes on button press in uni_N1.
    function uni_N1_Callback(hObject, eventdata, handles)
    set(handles.norm_N1,'Value',0)
    set(handles.gamma_N1,'Value',0)
    set(handles.log_N1,'Value',0)
    set(handles.param1N1,'String','Min')
    set(handles.param2N1,'String','Max')
    set(handles.val1_N1,'Visible','on')
    set(handles.val2_N1,'Visible','on')

    % --- Executes on button press in gamma_N1.
    function gamma_N1_Callback(hObject, eventdata, handles)
    set(handles.uni_N1,'Value',0)
    set(handles.norm_N1,'Value',0)
    set(handles.log_N1,'Value',0)
    set(handles.param1N1,'String','Shape')
    set(handles.param2N1,'String','Scale')
    set(handles.val1_N1,'Visible','on')
    set(handles.val2_N1,'Visible','on')
%
    % --- Executes on button press in norm_N1.
    function norm_N1_Callback(hObject, eventdata, handles)
    set(handles.uni_N1,'Value',0)
    set(handles.log_N1,'Value',0)
    set(handles.gamma_N1,'Value',0)
    set(handles.param1N1,'String','Mean')
    set(handles.param2N1,'String','Var')
    set(handles.val1_N1,'Visible','on')
    set(handles.val1_N1,'Visible','on')
    set(handles.val2_N1,'Visible','on')

    % --- Executes on button press in log_N1.
    function log_N1_Callback(hObject, eventdata, handles)
    set(handles.uni_N1,'Value',0)
    set(handles.norm_N1,'Value',0)
    set(handles.gamma_N1,'Value',0)
    set(handles.param1N1,'String','Mean')
    set(handles.param2N1,'String','Stad Dev')
    set(handles.val1_N1,'Visible','on')
    set(handles.val2_N1,'Visible','on')

%============= N1 chosen distribution parameters ==========================
    function val1_N1_Callback(hObject, eventdata, handles)
    N1_1= str2double(get(hObject, 'String'));
    handles.N1_1 = N1_1;
    guidata(hObject,handles)
    function val1_N1_CreateFcn(hObject, eventdata, handles)
    if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
       set(hObject,'BackgroundColor','white');
    end

    function val2_N1_Callback(hObject, eventdata, handles)
    N1_2= str2double(get(hObject, 'String'));
    handles.N1_2 = N1_2;
```

```
    guidata(hObject,handles)
    function val2_N1_CreateFcn(hObject, eventdata, handles)
    if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
       set(hObject,'BackgroundColor','white');
    end

%============================== N2 ======================================
  function uni_N2_Callback(hObject, eventdata, handles)
  set(handles.log_N2,'Value',0)
  set(handles.norm_N2,'Value',0)
  set(handles.gamma_N2,'Value',0)
  set(handles.param1N2,'String','Min')
  set(handles.param2N2,'String','Max')
  set(handles.val1_N2,'Visible','on')
  set(handles.val2_N2,'Visible','on')

% --- Executes on button press in gamma_N2.
  function gamma_N2_Callback(hObject, eventdata, handles)
  set(handles.log_N2,'Value',0)
  set(handles.norm_N2,'Value',0)
  set(handles.uni_N2,'Value',0)
  set(handles.param1N2,'String','Shape')
  set(handles.param2N2,'String','Scale')
  set(handles.val1_N2,'Visible','on')
  set(handles.val2_N2,'Visible','on')

% --- Executes on button press in norm_N2.
  function norm_N2_Callback(hObject, eventdata, handles)
  set(handles.log_N2,'Value',0)
  set(handles.uni_N2,'Value',0)
  set(handles.gamma_N2,'Value',0)
  set(handles.param1N2,'String','Mean')
  set(handles.param2N2,'String','Var')
  set(handles.val1_N2,'Visible','on')
  set(handles.val2_N2,'Visible','on')

% --- Executes on button press in log_N2.
  function log_N2_Callback(hObject, eventdata, handles)
  set(handles.uni_N2,'Value',0)
  set(handles.norm_N2,'Value',0)
  set(handles.gamma_N2,'Value',0)
  set(handles.param1N2,'String','Mean')
  set(handles.param2N2,'String','Std Dev')
  set(handles.val1_N2,'Visible','on')
  set(handles.val2_N2,'Visible','on')

  %=================== N2 parameters ====================================
  function val1_N2_Callback(hObject, eventdata, handles)
  N2_1= str2double(get(hObject, 'String'));
  handles.N2_1 = N2_1;
  guidata(hObject,handles)
  function val1_N2_CreateFcn(hObject, eventdata, handles)
  if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
     set(hObject,'BackgroundColor','white');
  end

  function val2_N2_Callback(hObject, eventdata, handles)
  N2_2= str2double(get(hObject, 'String'));
  handles.N2_2 = N2_2;
  guidata(hObject,handles)
```

```
function val2_N2_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

%====================== N3 ============================================

function uni_N3_Callback(hObject, eventdata, handles)
set(handles.log_N3,'Value',0)
set(handles.norm_N3,'Value',0)
set(handles.gamma_N3,'Value',0)
set(handles.param1N3,'String','Min')
set(handles.param2N3,'String','Max')
set(handles.val1_N3,'Visible','on')
set(handles.val2_N3,'Visible','on')

function norm_N3_Callback(hObject, eventdata, handles)
set(handles.uni_N3,'Value',0)
set(handles.log_N3,'Value',0)
set(handles.gamma_N3,'Value',0)
set(handles.param1N3,'String','Mean')
set(handles.param2N3,'String','Var')
set(handles.val1_N3,'Visible','on')
set(handles.val2_N3,'Visible','on')

% --- Executes on button press in norm_N3.
function gamma_N3_Callback(hObject, eventdata, handles)
set(handles.uni_N3,'Value',0)
set(handles.norm_N3,'Value',0)
set(handles.log_N3,'Value',0)
set(handles.param1N3,'String','Shape')
set(handles.param2N3,'String','Scale')
set(handles.val1_N3,'Visible','on')
set(handles.val2_N3,'Visible','on')

% --- Executes on button press in log_N3.
function log_N3_Callback(hObject, eventdata, handles)
set(handles.uni_N3,'Value',0)
set(handles.norm_N3,'Value',0)
set(handles.gamma_N3,'Value',0)
set(handles.param1N3,'String','Mean')
set(handles.param2N3,'String','Std Dev')
set(handles.val1_N3,'Visible','on')
set(handles.val2_N3,'Visible','on')

%====================== chosen dist parameters ============================
function val1_N3_Callback(hObject, eventdata, handles)
N3_1= str2double(get(hObject, 'String'));
handles.N3_1 = N3_1;
guidata(hObject,handles)
function val1_N3_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function val2_N3_Callback(hObject, eventdata, handles)
N3_2= str2double(get(hObject, 'String'));
handles.N3_2 = N3_2;
guidata(hObject,handles)
function val2_N3_CreateFcn(hObject, eventdata, handles)
```

```
    if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
        set(hObject,'BackgroundColor','white');
    end

%============================ Nh ======================================
% --- Executes on button press in uni_Nh.
    function uni_Nh_Callback(hObject, eventdata, handles)
    set(handles.log_Nh,'Value',0)
    set(handles.norm_Nh,'Value',0)
    set(handles.gamma_Nh,'Value',0)
    set(handles.param1Nh,'String','Min')
    set(handles.param2Nh,'String','Max')
    set(handles.val1_Nh,'Visible','on')
    set(handles.val2_Nh,'Visible','on')

% --- Executes on button press in norm_Nh.
    function norm_Nh_Callback(hObject, eventdata, handles)
    set(handles.log_Nh,'Value',0)
    set(handles.uni_Nh,'Value',0)
    set(handles.gamma_Nh,'Value',0)
    set(handles.param1Nh,'String','Mean')
    set(handles.param2Nh,'String','Var')
    set(handles.val1_Nh,'Visible','on')
    set(handles.val2_Nh,'Visible','on')

% --- Executes on button press in gamma_Nh.
    function gamma_Nh_Callback(hObject, eventdata, handles)
    set(handles.log_Nh,'Value',0)
    set(handles.norm_Nh,'Value',0)
    set(handles.uni_Nh,'Value',0)
    set(handles.param1Nh,'String','Shape')
    set(handles.param2Nh,'String','Scale')
    set(handles.val1_Nh,'Visible','on')
    set(handles.val2_Nh,'Visible','on')

% --- Executes on button press in log_Nh.
    function log_Nh_Callback(hObject, eventdata, handles)
    set(handles.uni_Nh,'Value',0)
    set(handles.norm_Nh,'Value',0)
    set(handles.gamma_Nh,'Value',0)
    set(handles.param1Nh,'String','Mean')
    set(handles.param2Nh,'String','Std Dev')
    set(handles.val1_Nh,'Visible','on')
    set(handles.val2_Nh,'Visible','on')

    function val1_Nh_Callback(hObject, eventdata, handles)
    Nh_1= str2double(get(hObject, 'String'));
    handles.Nh_1 = Nh_1;
    guidata(hObject,handles)
    function val1_Nh_CreateFcn(hObject, eventdata, handles)
    if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
        set(hObject,'BackgroundColor','white');
    end

    function val2_Nh_Callback(hObject, eventdata, handles)
    Nh_2= str2double(get(hObject, 'String'));
    handles.Nh_2 = Nh_2;
    guidata(hObject,handles)
    function val2_Nh_CreateFcn(hObject, eventdata, handles)
    if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
```

```
    set(hObject,'BackgroundColor','white');
  end

%====================== Mutation rate ==================================
  function uni_mut_rate_Callback(hObject, eventdata, handles)
  set(handles.log_mut_rate,'Value',0)
  set(handles.param1mut,'String','Min')
  set(handles.param2mut,'String','Max')
  set(handles.val1_mut_rate,'Visible','on')
  set(handles.val2_mut_rate,'Visible','on')

% --- Executes on button press in log_mut_rate.
  function log_mut_rate_Callback(hObject, eventdata, handles)
  set(handles.uni_mut_rate,'Value',0)
  set(handles.param1mut,'String','Mean')
  set(handles.param2mut,'String','Std Dev')
  set(handles.val1_mut_rate,'Visible','on')
  set(handles.val2_mut_rate,'Visible','on')

function val1_mut_rate_Callback(hObject, eventdata, handles)
mut_rate_1= str2double(get(hObject, 'String'));
handles.mut_rate_1 = mut_rate_1;
guidata(hObject,handles)
function val1_mut_rate_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
  set(hObject,'BackgroundColor','white');
end

function val2_mut_rate_Callback(hObject, eventdata, handles)
mut_rate_2= str2double(get(hObject, 'String'));
handles.mut_rate_2 = mut_rate_2;
guidata(hObject,handles)
function val2_mut_rate_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
  set(hObject,'BackgroundColor','white');
end

%======================================================
% --- Executes on button press in uni_Tadm2.
  function uni_Tadm2_Callback(hObject, eventdata, handles)
  set(handles.norm_Tadm2,'Value',0)
  set(handles.gamma_Tadm2,'Value',0)
  set(handles.log_Tadm2,'Value',0)
  set(handles.param1Tadm2,'String','Min')
  set(handles.param2Tadm2,'String','Max')
  set(handles.val1_Tadm2,'Visible','on')
  set(handles.val2_Tadm2,'Visible','on')

% --- Executes on button press in norm_Tadm2.
  function norm_Tadm2_Callback(hObject, eventdata, handles)
  set(handles.uni_Tadm2,'Value',0)
  set(handles.gamma_Tadm2,'Value',0)
  set(handles.log_Tadm2,'Value',0)
  set(handles.param1Tadm2,'String','Mean')
  set(handles.param2Tadm2,'String','Var')
  set(handles.val1_Tadm2,'Visible','on')
  set(handles.val2_Tadm2,'Visible','on')

% --- Executes on button press in gamma_Tadm2.
  function gamma_Tadm2_Callback(hObject, eventdata, handles)
```

```
   set(handles.norm_Tadm2,'Value',0)
   set(handles.uni_Tadm2,'Value',0)
   set(handles.log_Tadm2,'Value',0)
   set(handles.param1Tadm2,'String','Shape')
   set(handles.param2Tadm2,'String','Scale')
   set(handles.val1_Tadm2,'Visible','on')
   set(handles.val2_Tadm2,'Visible','on')

% --- Executes on button press in log_Tadm2.
   function log_Tadm2_Callback(hObject, eventdata, handles)
   set(handles.norm_Tadm2,'Value',0)
   set(handles.gamma_Tadm2,'Value',0)
   set(handles.uni_Tadm2,'Value',0)
   set(handles.param1Tadm2,'String','Mean')
   set(handles.param2Tadm2,'String','Std Dev')
   set(handles.val1_Tadm2,'Visible','on')
   set(handles.val2_Tadm2,'Visible','on')

function val1_Tadm2_Callback(hObject, eventdata, handles)
Tadm2_1= str2double(get(hObject, 'String'));
handles.Tadm2_1 = Tadm2_1;
guidata(hObject,handles)
function val1_Tadm2_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end

function val2_Tadm2_Callback(hObject, eventdata, handles)
Tadm2_2= str2double(get(hObject, 'String'));
handles.Tadm2_2 = Tadm2_2;
guidata(hObject,handles)
function val2_Tadm2_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end

%=============== Earlier admixture time ==================================
% --- Executes on button press in gamma_Tadm1.
   function gamma_Tadm1_Callback(hObject, eventdata, handles)
   set(handles.norm_Tadm1,'Value',0)
   set(handles.log_Tadm1,'Value',0)
   set(handles.uni_Tadm1,'Value',0)
   set(handles.param1Tadm1,'String','Shape')
   set(handles.param2Tadm1,'String','Scale')
   set(handles.val1_Tadm1,'Visible','on')
   set(handles.val2_Tadm1,'Visible','on')

% --- Executes on button press in log_Tadm1.
   function log_Tadm1_Callback(hObject, eventdata, handles)
   set(handles.norm_Tadm1,'Value',0)
   set(handles.gamma_Tadm1,'Value',0)
   set(handles.uni_Tadm1,'Value',0)
   set(handles.param1Tadm1,'String','Mean')
   set(handles.param2Tadm1,'String','Std Dev')
   set(handles.val1_Tadm1,'Visible','on')
   set(handles.val2_Tadm1,'Visible','on')

% --- Executes on button press in norm_Nh.
   function norm_Tadm1_Callback(hObject, eventdata, handles)
   set(handles.log_Tadm1,'Value',0)
```

```
    set(handles.param2Tsplit,'String','Scale')
    set(handles.val1_Tsplit,'Visible','on')
    set(handles.val2_Tsplit,'Visible','on')

% --- Executes on button press in log_Tsplit.
    function log_Tsplit_Callback(hObject, eventdata, handles)
    set(handles.norm_Tsplit,'Value',0)
    set(handles.gamma_Tsplit,'Value',0)
    set(handles.uni_Tsplit,'Value',0)
    set(handles.param1Tsplit,'String','Mean')
    set(handles.param2Tsplit,'String','Std Dev')
    set(handles.val1_Tsplit,'Visible','on')
    set(handles.val2_Tsplit,'Visible','on')

%
    function val1_Tsplit_Callback(hObject, eventdata, handles)
    Tsplit_1= str2double(get(hObject, 'String'));
    handles.Tsplit_1 = Tsplit_1;
    guidata(hObject,handles)

    function val1_Tsplit_CreateFcn(hObject, eventdata, handles)
    if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
       set(hObject,'BackgroundColor','white');
    end

    function val2_Tsplit_Callback(hObject, eventdata, handles)
    Tsplit_2= str2double(get(hObject, 'String'));
    handles.Tsplit_2 = Tsplit_2;
    guidata(hObject,handles)
    function val2_Tsplit_CreateFcn(hObject, eventdata, handles)
    if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
       set(hObject,'BackgroundColor','white');
    end

%====================== p3 ================================================
    function uni_p3_Callback(hObject, eventdata, handles)
    set(handles.beta_p3,'Value',0)
    set(handles.param1p3,'String','Min')
    set(handles.param2p3,'String','Max')
    set(handles.val1_p3,'Visible','on')
    set(handles.val2_p3,'Visible','on')

% --- Executes on button press in beta_p3.
    function beta_p3_Callback(hObject, eventdata, handles)
    set(handles.uni_p3,'Value',0)
    set(handles.param1p3,'String','a')
    set(handles.param2p3,'String','b')
    set(handles.val1_p3,'Visible','on')
    set(handles.val2_p3,'Visible','on')

    function val1_p3_Callback(hObject, eventdata, handles)
    p3_1= str2double(get(hObject, 'String'));
    handles.p3_1 = p3_1;
    guidata(hObject,handles)
    function val1_p3_CreateFcn(hObject, eventdata, handles)
    if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
       set(hObject,'BackgroundColor','white');
    end

    function val2_p3_Callback(hObject, eventdata, handles)
```

```
p3_2= str2double(get(hObject, 'String'));
handles.p3_2 = p3_2;
guidata(hObject,handles)
function val2_p3_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end

%============================== p1 ========================================
function uni_p1_Callback(hObject, eventdata, handles)
set(handles.beta_p1,'Value',0)
set(handles.param1p1,'String','Min')
set(handles.param2p1,'String','Max')
set(handles.val1_p1,'Visible','on')
set(handles.val2_p1,'Visible','on')

% --- Executes on button press in beta_p1.
function beta_p1_Callback(hObject, eventdata, handles)
set(handles.uni_p1,'Value',0)
set(handles.param1p1,'String','a')
set(handles.param2p1,'String','b')
set(handles.val1_p1,'Visible','on')
set(handles.val2_p1,'Visible','on')

function val1_p1_Callback(hObject, eventdata, handles)
p1_1= str2double(get(hObject, 'String'));
handles.p1_1 = p1_1;
guidata(hObject,handles)
function val1_p1_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end

function val2_p1_Callback(hObject, eventdata, handles)
p1_2= str2double(get(hObject, 'String'));
handles.p1_2 = p1_2;
guidata(hObject,handles)
function val2_p1_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end

%============================== Nanc ======================================
function uni_Nanc_Callback(hObject, eventdata, handles)
set(handles.norm_Nanc,'Value',0)
set(handles.gamma_Nanc,'Value',0)
set(handles.log_Nanc,'Value',0)
set(handles.param1Nanc,'String','Min')
set(handles.param2Nanc,'String','Max')
set(handles.val1_Nanc,'Visible','on')
set(handles.val2_Nanc,'Visible','on')

% --- Executes on button press in norm_Nanc.
function norm_Nanc_Callback(hObject, eventdata, handles)
set(handles.uni_Nanc,'Value',0)
set(handles.gamma_Nanc,'Value',0)
set(handles.log_Nanc,'Value',0)
set(handles.param1Nanc,'String','Mean')
set(handles.param2Nanc,'String','Var')
set(handles.val1_Nanc,'Visible','on')
```

```
   set(handles.val2_Nanc,'Visible','on')

% --- Executes on button press in gamma_Nanc.
   function gamma_Nanc_Callback(hObject, eventdata, handles)
   set(handles.norm_Nanc,'Value',0)
   set(handles.uni_Nanc,'Value',0)
   set(handles.log_Nanc,'Value',0)
   set(handles.param1Nanc,'String','Shape')
   set(handles.param2Nanc,'String','Scale')
   set(handles.val1_Nanc,'Visible','on')
   set(handles.val2_Nanc,'Visible','on')

% --- Executes on button press in log_Nanc.
   function log_Nanc_Callback(hObject, eventdata, handles)
   set(handles.norm_Nanc,'Value',0)
   set(handles.gamma_Nanc,'Value',0)
   set(handles.uni_Nanc,'Value',0)
   set(handles.param1Nanc,'String','Mean')
   set(handles.param2Nanc,'String','Std Dev')
   set(handles.val1_Nanc,'Visible','on')
   set(handles.val2_Nanc,'Visible','on')

   function val1_Nanc_Callback(hObject, eventdata, handles)
   Nanc_1= str2double(get(hObject, 'String'));
   handles.Nanc_1 = Nanc_1;
   guidata(hObject,handles)
   function val1_Nanc_CreateFcn(hObject, eventdata, handles)
   if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
      set(hObject,'BackgroundColor','white');
   end

   function val2_Nanc_Callback(hObject, eventdata, handles)
   Nanc_2= str2double(get(hObject, 'String'));
   handles.Nanc_2 = Nanc_2;
   guidata(hObject,handles)
   function val2_Nanc_CreateFcn(hObject, eventdata, handles)
   if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
      set(hObject,'BackgroundColor','white');
   end


%================== samples sizes ========================================
function Nsam1_Callback(hObject, eventdata, handles)
Nsam1 = str2double(get(hObject, 'String'));
if isnan(Nsam1)
   set(hObject, 'String', 0);
   errordlg('Nsam1 must be a number','Error');
end
if Nsam1<=0
   set(hObject, 'String', 0);
   errordlg('Nsam1 must be a positive number','Error');
end
handles.Nsam1 = Nsam1;
guidata(hObject,handles)

% --- Executes during object creation, after setting all properties.
function Nsam1_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end
```

```
function Nsam2_Callback(hObject, eventdata, handles)
Nsam2 = str2double(get(hObject, 'String'));
if isnan(Nsam2)
   set(hObject, 'String', 0);
   errordlg('Nsam2 must be a number','Error');
end
if Nsam2<=0
   set(hObject, 'String', 0);
   errordlg('Nsam2 must be a positive number','Error');
end
handles.Nsam2 = Nsam2;
guidata(hObject,handles)

% --- Executes during object creation, after setting all properties.
function Nsam2_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end

function Nsam3_Callback(hObject, eventdata, handles)
Nsam3 = str2double(get(hObject, 'String'));
if isnan(Nsam3)
   set(hObject, 'String', 0);
   errordlg('Nsam3 must be a number','Error');
end
if Nsam3<=0
   set(hObject, 'String', 0);
   errordlg('Nsam3 must be a positive number','Error');
end
handles.Nsam3 = Nsam3;
guidata(hObject,handles)

% --- Executes during object creation, after setting all properties.
function Nsam3_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end

function Nsam4_Callback(hObject, eventdata, handles)
Nsam4 = str2double(get(hObject, 'String'));
if isnan(Nsam4)
   set(hObject, 'String', 0);
   errordlg('Nsam4 must be a number','Error');
end
if Nsam4<=0
   set(hObject, 'String', 0);
   errordlg('Nsam4 must be a positive number','Error');
end
handles.Nsam4 = Nsam4;
guidata(hObject,handles)

% --- Executes during object creation, after setting all properties.
function Nsam4_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end

function Nsim_Callback(hObject, eventdata, handles)
Nsim = str2double(get(hObject, 'String'));
```

```
if isnan(Nsim)
   set(hObject, 'String', 0);
   errordlg('Nsim must be a number','Error');
end
if Nsim<=0
   set(hObject, 'String', 0);
   errordlg('Nsim must be a positive number','Error');
end
handles.Nsim = Nsim;
guidata(hObject,handles)

% --- Executes during object creation, after setting all properties.
function Nsim_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end

function nb_loci_Callback(hObject, eventdata, handles)
nb_loci = str2double(get(hObject, 'String'));
if isnan(nb_loci)
   set(hObject, 'String', 0);
   errordlg('nb_loci must be a number','Error');
end
if nb_loci<=0
   set(hObject, 'String', 0);
   errordlg('nb_loci must be a positive number','Error');
end
handles.nb_loci = nb_loci;
guidata(hObject,handles)
% --- Executes during object creation, after setting all properties.
function nb_loci_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end

%================ RUN ====================================================
   function Calculate_Callback(hObject, eventdata, handles)
%    if exist('abc_twoadmixture','dir') ~= 7
%    eval(['mkdir abc_twoadmixture']);

%=========================================================================
%=
% IN CASE YOU DON'T WANT TO USE THE GRAPHICAL INTERFACE define the parameters here
graphical = 1
make_obs = 0
if graphical == 0

   handles.Nsim = 1000000
   handles.nb_loci = 2
   handles.Nsam1 = 20
   handles.Nsam2 = 20
   handles.Nsam3 = 20
   handles.Nsam4 = 20

end

% SIMULATE OBSERVED DATA
if graphical == 0 & make_obs==1

        nb_obs = 500
```

```
        % effective size
        handles.N1=10000*ones(1,nb_obs)';
        handles.N2=10000*ones(1,nb_obs)';
        handles.N3=10000*ones(1,nb_obs)';
        handles.Nh=10000*ones(1,nb_obs)';
        handles.Nanc=10000*ones(1,nb_obs)';

        % tadm2 prior (unif 1-10)
        handles.Tadm2=10*ones(1,nb_obs)';

        % tadm1 prior (unif 1-100)
        handles.Tadm1=10*ones(1,nb_obs)';

        % tsplit (unif 1000-10000)
        handles.Tsplit=50000*ones(1,nb_obs)';

        % mut rate (unif 0.00001 - 0.001)
        handles.mut_rate=0.0001*ones(1,nb_obs)';

        % p1 and p3 (unif 0-1)
        handles.p3=0*ones(1,nb_obs)';
        handles.p1=0.7*ones(1,nb_obs)';


end

nb_obs = 1;
sim_by_file = 100000
nbfiles = ceil(handles.Nsim/sim_by_file); % how many files
nbsimfile = sim_by_file*ones(1,nbfiles-1); % how many sim per file
nbsimfile(nbfiles) = handles.Nsim - sum(nbsimfile); % vector with number of simulations of each file
tic
% GO FROM FILE 1 TO NBFILES
for file = 1:nbfiles

    % SAMPLE PARAMETERS FROM THE PRIORS                %

    if graphical == 1

        %N1 :
        % If the normal distribution for the prior is selected
        if get(handles.norm_N1,'Value')==1
        %if handles.norm_N1==1
            handles.N1=zeros(nbsimfile(file),1);
            % Negative values are not allowed, and hence a new value will
            % be sampled until it is positive
            for j=1:nbsimfile(file)
                while handles.N1(j)<=0
                    handles.N1(j)=ceil(normrnd(handles.N1_1,handles.N1_2,1,1));
                end
            end
        % selecting the lognormal prior
        elseif get(handles.log_N1,'Value')==1
            handles.N1=ceil(lognrnd(handles.N1_1,handles.N1_2,nbsimfile(file),1));
        % selecting the gamma prior
        elseif get(handles.gamma_N1,'Value')==1
            handles.N1=ceil(gamrnd(handles.N1_1,handles.N1_2,nbsimfile(file),1));
        % selecting the uniform prior
        elseif get(handles.uni_N1,'Value')==1
```

```
   handles.N1=ceil(handles.N1_1+(handles.N1_2-handles.N1_1)*rand(1,nbsimfile(file)))';
else
   errordlg('N1 distribution has not been recognised ','Error');
end

%N2 :
% If the normal distribution for the prior is selected
if get(handles.norm_N2,'Value')==1
   handles.N2=zeros(nbsimfile(file),1);
   % Negative values are not allowed, and hence a new value will
   % be sampled until it is positive
   for j=1:nbsimfile(file)
      while handles.N2(j)<=0
         handles.N2(j)=ceil(normrnd(handles.N2_1,handles.N2_2,1,1));
      end
   end
% selecting the lognormal prior
elseif get(handles.log_N2,'Value')==1
   handles.N2=ceil(lognrnd(handles.N2_1,handles.N2_2,nbsimfile(file),1));
 % selecting the gamma prior
elseif get(handles.gamma_N2,'Value')==1
   handles.N2=ceil(gamrnd(handles.N2_1,handles.N2_2,nbsimfile(file),1));
 % selecting the uniform prior
elseif get(handles.uni_N2,'Value')==1
   handles.N2=ceil(handles.N2_1+(handles.N2_2-handles.N2_1)*rand(1,nbsimfile(file)))';
else
   errordlg('N2 distribution has not been recognised ','Error');
end

%N3 (same as in N1 and N2):
if get(handles.norm_N3,'Value')==1
   handles.N3=zeros(nbsimfile(file),1);
   for j=1:nbsimfile(file)
      while handles.N3(j)<=0
         handles.N3(j)=ceil(normrnd(handles.N3_1,handles.N3_2,1,1));
      end
   end
elseif get(handles.log_N3,'Value')==1
   handles.N3=ceil(lognrnd(handles.N3_1,handles.N3_2,nbsimfile(file),1));
elseif get(handles.gamma_N3,'Value')==1
   handles.N3=ceil(gamrnd(handles.N3_1,handles.N3_2,nbsimfile(file),1));
elseif get(handles.uni_N3,'Value')==1
   handles.N3=ceil(handles.N3_1+(handles.N3_2-handles.N3_1)*rand(1,nbsimfile(file)))';
else
   errordlg('N3 distribution has not been recognised ','Error');
end

%Nh (same as in N1 and N2):
if get(handles.norm_Nh,'Value')==1
   handles.Nh=zeros(nbsimfile(file),1);
   for j=1:nbsimfile(file)
      while handles.Nh(j)<=0
         handles.Nh(j)=ceil(normrnd(handles.Nh_1,handles.Nh_2,1,1));
      end
   end
elseif get(handles.log_Nh,'Value')==1
   handles.Nh=ceil(lognrnd(handles.Nh_1,handles.Nh_2,nbsimfile(file),1));
elseif get(handles.gamma_Nh,'Value')==1
   handles.Nh=ceil(gamrnd(handles.Nh_1,handles.Nh_2,nbsimfile(file),1));
elseif get(handles.uni_Nh,'Value')==1
```

```
    handles.Nh=ceil(handles.Nh_1+(handles.Nh_2-handles.Nh_1)*rand(1,nbsimfile(file)))';
  else
    errordlg('Nh distribution has not been recognised ','Error');
  end

  %Tadm2
  % (this is the time of the most recent admixture event):
  % if a normal prior is assumed
  if get(handles.norm_Tadm2,'Value')==1
    handles.Tadm2=zeros(nbsimfile(file),1);
    % Negative values are not allowed, and hence a new value will
    % be sampled until it is positive
    for j=1:nbsimfile(file)
      while handles.Tadm2(j)<=0
        handles.Tadm2(j)=ceil(normrnd(handles.Tadm2_1,handles.Tadm2_2,1,1));
      end
    end
  % if the prior is lognormal
  elseif get(handles.log_Tadm2,'Value')==1
    handles.Tadm2=ceil(lognrnd(handles.Tadm2_1,handles.Tadm2_2,nbsimfile(file),1));
  % if the prior is gamma
  elseif get(handles.gamma_Tadm2,'Value')==1
    handles.Tadm2=ceil(gamrnd(handles.Tadm2_1,handles.Tadm2_2,nbsimfile(file),1));
  % if the prior is uniform
  elseif get(handles.uni_Tadm2,'Value')==1
    handles.Tadm2=ceil(handles.Tadm2_1+(handles.Tadm2_2-
handles.Tadm2_1)*rand(1,nbsimfile(file)))';
  else
    errordlg('Tadm2 distribution has not been recognised ','Error');
  end

  %Tadm1
  % (this is the time of the oldest and first admixture event):
  % note that the values of Tadm1 must be higher than Tadm2
  %(Tadm1 > Tadm2)
  if get(handles.norm_Tadm1,'Value')==1
    handles.Tadm1=zeros(nbsimfile(file),1);
    % Values Tadm1 < Tadm2 are not allowed and new values are
    % sampled until we get Tadm1 > Tadm2
    for j=1:nbsimfile(file)
      while handles.Tadm1(j)<=handles.Tadm2(j)
        handles.Tadm1(j)=ceil(normrnd(handles.Tadm1_1,handles.Tadm1_2,1,1));
      end
    end
  % if lognormal prior
  elseif get(handles.log_Tadm1,'Value')==1
    handles.Tadm1=ceil(lognrnd(handles.Tadm1_1,handles.Tadm1_2,nbsimfile(file),1));
    for j=1:nbsimfile(file)
      while handles.Tadm1(j)<=handles.Tadm2(j)
        handles.Tadm1(j)=ceil(lognrnd(handles.Tadm1_1,handles.Tadm1_2,nbsimfile(file),1));
      end
    end
  % if gamma prior
  elseif get(handles.gamma_Tadm1,'Value')==1
    handles.Tadm1=ceil(gamrnd(handles.Tadm1_1,handles.Tadm1_2,nbsimfile(file),1));
    for j=1:nbsimfile(file)
      while handles.Tadm1(j)<=handles.Tadm2(j)
        handles.Tadm1(j)=ceil(gamrnd(handles.Tadm1_1,handles.Tadm1_2,nbsimfile(file),1));
      end
    end
```

```
    % if uniform prior
    elseif get(handles.uni_Tadm1,'Value')==1
      % this is a uniform prior between Tadm2 and max_Tadm1 (if mac
      % Tadm2 > min Tadm1)
      if handles.Tadm1_1 < handles.Tadm2_2
        handles.Tadm1=ceil(handles.Tadm2+(handles.Tadm1_2-
handles.Tadm2).*rand(1,nbsimfile(file))');
      else
        handles.Tadm1=ceil(handles.Tadm1_1+(handles.Tadm1_2-
handles.Tadm1_1)*rand(1,nbsimfile(file))');
      end
    else
      errordlg('Tadm1 distribution has not been recognised ','Error');
    end

    %Tsplit :
    % (this is the time of split of the ancestral population that
    % gave rise to the parental populations):
    % note that the values of Tsplit must be higher than Tadm1
    % (Tsplit > Tadm1 > Tadm2)
    handles.Tsplit=handles.Tadm1;
    for j=1:nbsimfile(file)
      % A new value will be sampled until Tsplit is higher than Tadm1
      while handles.Tsplit(j)<= handles.Tadm1(j)
        if get(handles.norm_Tsplit,'Value')==1
          handles.Tsplit(j)=ceil(normrnd(handles.Tsplit_1,handles.Tsplit_2,1,1));
        elseif get(handles.log_Tsplit,'Value')==1
          handles.Tsplit(j)=ceil(lognrnd(handles.Tsplit_1,handles.Tsplit_2,1,1));
        elseif get(handles.gamma_Tsplit,'Value')==1
          handles.Tsplit(j)=ceil(gamrnd(handles.Tsplit_1,handles.Tsplit_2,1,1));
        elseif get(handles.uni_Tsplit,'Value')==1
          handles.Tsplit(j)=ceil(handles.Tsplit_1+(handles.Tsplit_2-handles.Tsplit_1)*rand(1,1))';
        else
          errordlg('Tsplit distribution has not been recognised ','Error');
        end
      end
    end

    %Nanc :
    % effective size of the ancestral population
    if get(handles.norm_Nanc,'Value')==1
      handles.Nanc=zeros(nbsimfile(file),1);
      for j=1:nbsimfile(file)
        while handles.Nanc(j)<=0
          handles.Nanc(j)=ceil(normrnd(handles.Nanc_1,handles.Nanc_2,1,1));
        end
      end
    elseif get(handles.log_Nanc,'Value')==1
      handles.Nanc=ceil(lognrnd(handles.Nanc_1,handles.Nanc_2,nbsimfile(file),1));
    elseif get(handles.gamma_Nanc,'Value')==1
      handles.Nanc=ceil(gamrnd(handles.Nanc_1,handles.Nanc_2,nbsimfile(file),1));
    elseif get(handles.uni_Nanc,'Value')==1
      handles.Nanc=ceil(handles.Nanc_1+(handles.Nanc_2-handles.Nanc_1)*rand(1,nbsimfile(file)))';
    else
      errordlg('Nanc distribution has not been recognised ','Error');
    end

    %mut_rate :
    % mutation rate
    if get(handles.uni_mut_rate,'Value')==1
```

```
        handles.mut_rate=(handles.mut_rate_1+(handles.mut_rate_2-
handles.mut_rate_1)*rand(1,nbsimfile(file)))';
    elseif get(handles.log_mut_rate,'Value')==1
        handles.mut_rate=ceil(lognrnd(handles.mut_rate_1,handles.mut_rate_2,nbsimfile(file),1));
    else
        errordlg('Mutation rate distribution has not been recognised ','Error');
    end

    %p3:
    % proportion of the hybrid population that came from parental
    % population 3, in the second admixture event (Tadm2)
    if get(handles.beta_p3,'Value')==1
        handles.p3=ceil(betarnd(handles.p3_1,handles.p3_2,nbsimfile(file),1));
    elseif get(handles.uni_p3,'Value')==1
        handles.p3=(handles.p3_1+(handles.p3_2-handles.p3_1)*rand(1,nbsimfile(file)))';
    else
        errordlg('p3 distribution has not been recognised ','Error');
    end

    %p1:
    % proportion of the hybrid population that came from parental
    % population 1 in the first admixture event (Tadm1)
    if get(handles.beta_p1,'Value')==1
        handles.p1=ceil(betarnd(handles.p1_1,handles.p1_2,nbsimfile(file),1));
    elseif get(handles.uni_p1,'Value')==1
        handles.p1=(handles.p1_1+(handles.p1_2-handles.p1_1)*rand(1,nbsimfile(file)))';
    else
        errordlg('p1 distribution has not been recognised ','Error');
    end

else % if graphical is zero (when we don't want to use graphical interface)

    % effective size prior (uniform 1000-15000)
    handles.N1=ceil(1000+(15000-1000)*rand(1,nbsimfile(file)))';
    handles.N2=ceil(1000+(15000-1000)*rand(1,nbsimfile(file)))';
    handles.N3=ceil(1000+(15000-1000)*rand(1,nbsimfile(file)))';
    handles.Nh=ceil(1000+(15000-1000)*rand(1,nbsimfile(file)))';
    handles.Nanc=ceil(1000+(15000-1000)*rand(1,nbsimfile(file)))';

    % tadm2 prior (unif 1-100)
    handles.Tadm2=ceil(1+(100-1)*rand(1,nbsimfile(file)))';

    % tadm1 prior (unif 1-100)
    handles.Tadm1=ceil(handles.Tadm2+(100-handles.Tadm2).*rand(1,nbsimfile(file))');

    % tsplit (unif 1000-10000)
    handles.Tsplit=ceil(1000+(15000-1000)*rand(1,nbsimfile(file)))';

    % mut rate (unif 0.00001 - 0.001)
    handles.mut_rate=(0.00001+(0.001-0.00001)*rand(1,nbsimfile(file)))';

    % p1 and p3 (unif 0-1)
    handles.p3=(0+(1-0)*rand(1,nbsimfile(file)))';
    handles.p1=(0+(1-0)*rand(1,nbsimfile(file)))';

end; % END OF IF NOT GRAPHICAL, OTHERWISE THE PARAM WERE SAMPLED BEFORE

% STANDARDIZE THE PARAMETERS ACCORDING TO ms

param = [];
```

```
nb_pop=4;
Nsam=handles.Nsam1+handles.Nsam2+handles.Nsam3+handles.Nsam4;
ref_N=max([handles.N1 handles.N2 handles.N3 handles.Nh],[],2);%added third parental population
theta=4*ref_N.*handles.mut_rate;%theta is the mutation parameter
param(1,:)=theta;
relative_N=[handles.N1 handles.N2 handles.N3 handles.Nh]./[ref_N ref_N ref_N ref_N];
param(2:5,:)=relative_N';
time_adm2=handles.Tadm2./(4*ref_N);%time of admixture 2 (recent)
param(6,:)=time_adm2;
param(7,:)=1-handles.p3;
param(8,:)=time_adm2;
time_adm1=handles.Tadm1./(4*ref_N);%time of admixture 1 (ancient)
param(9,:)=time_adm1;
param(10,:)=handles.p1;
param(11,:)=time_adm1;
param(12,:)=time_adm1;
time_split=handles.Tsplit./(4*ref_N);%time of split (coalescence of populations)
param(13,:)=time_split;
param(14,:)=time_split;
param(15,:)=time_split; %added for the third parental population at the coalescence point
rel_size_anc_pop=handles.Nanc./ref_N;%relative size of the ancestral population
param(16,:)=rel_size_anc_pop;
param(17,:)=ref_N; %INCLUDING THE ref_N as the 17th column

% SAVE the standardized parameters sampled from the priors in a text file
% This will be read by ms with the tbs option

% Create a folder where all simulations will be saved

mkdir('simulations_database');

% Creation of a text file which contains the parameters for each simulation
% before multiplying by nb_loci
fid2=fopen(['./simulations_database/simparameters_', num2str(file), '.txt'],'w');  % open the file to save
fprintf(fid2,'%6.6f %0.6f %0.6f %0.6f %0.6f %6.10f %0.6f %6.10f %6.6f %0.6f %6.6f %6.6f %6.6f %6.6f
%6.6f %6.6f %8.0f \n',param);
fclose(fid2); % close connection to the file

% In order to simulate the data we need to repeat each line of the
% parameter file by the number of loci.
% NOTE: THIS ASSUMES THAT ALL LOCI HAVE THE SAME MUTATION RATE

% create file to save the summary means across loci
means_file = fopen(['./simulations_database/sim_sum_means_', num2str(file),'.txt'], 'w');
% create file to save the summary statistics variance across loci
var_file = fopen(['./simulations_database/sim_sum_var_', num2str(file), '.txt'], 'w');

% FOR LOOP FROM SIMULATION 1 TO NBSIM OF CURRENT FILE
for nbsim=1:nbsimfile(file);
   %nbsim

   if mod(nbsim,10000) == 0
      nbsim
   end

   %RUN MS COMMAND FOR EACH SIMULATION

   command=['"C:\ms\ms.exe" ', num2str(Nsam),' ', num2str(handles.nb_loci),' -t ',
num2str(param(1,nbsim)), ' -I 4 ', num2str(handles.Nsam1),' ', num2str(handles.Nsam2),' ',
num2str(handles.Nsam3), ' ', num2str(handles.Nsam4),' -n 1 ', num2str(param(2,nbsim)), ' -n 2 ',
```

```
num2str(param(3,nbsim)), ' -n 3 ', num2str(param(4,nbsim)), ' -n 4 ', num2str(param(5,nbsim)), ' -es ',
num2str(param(6,nbsim)), ' 4 ', num2str(param(7,nbsim)), ' -ej ', num2str(param(8,nbsim)), ' 5 3 -es ',
num2str(param(9,nbsim)), ' 4 ', num2str(param(10,nbsim)), ' -ej ', num2str(param(11,nbsim)), ' 6 2 -ej ',
num2str(param(12,nbsim)), ' 4 1 -ej ', num2str(param(13,nbsim)), ' 3 2 -ej ', num2str(param(14,nbsim)), ' 2 1 -
en ', num2str(param(15,nbsim)), ' 1 ', num2str(param(16,nbsim)), ' | c:\ms\microsat'];
        [status,ms_result] = eval(['dos(command)']); % this creates a string "ms_result"

        % transform ms_result string into a matrix
        % each line has the results for each locus
        ms_result = str2num(ms_result);

        % initialize the data and sumstat matrix
        data = {}; % this is a matrix of matrices (cell), that is the reason for the {}
        sumstat = [];

        if isempty(ms_result)

           nbsim = nbsim - 1
           command
           ms_result
           data
           sumstat

        else

           %Loop for each locus
           for locus=1:handles.nb_loci

              % CALCULATION OF SIMULATED ALLELE FREQUENCIES FOR EACH
              % LOCUS
              % get the allele frequency for each simulation
              data{locus} = [getallfreq(ms_result(locus,1:Nsam), [handles.Nsam1  handles.Nsam2
handles.Nsam3 handles.Nsam4])]';

              % CALCULATING SUMMARY STATS FOR EACH LOCUS
              % get the number of alleles for locus sim
              sim_nb_all = length(data{locus}(:,1));

              %HETEROZYGOSITY
              heterozygosity = het(data{locus}(:,1:nb_pop),  sim_nb_all);

              %ALLELIC RANGE
              all_range = a_range(data{locus}, sim_nb_all, nb_pop);

              % NUMBER OF ALLELES EACH POP AND PRIVATE ALLELES
              private_alleles = [priv_all(data{locus}(:,1:nb_pop), sim_nb_all, nb_pop)];

              %PAIRWISE FST
              pairwise_fst = pair_fst(data{locus}(:,1:nb_pop),  heterozygosity, nb_pop, sim_nb_all);

              % Save the summary statistics in a matrix (each colum is a
              % locus)
              sumstat(:,locus) = [heterozygosity,all_range,private_alleles(1:end),pairwise_fst];

           end

           % COMPUTE THE MEAN ACROSS LOCI FOR THE SIMULATED SUMSTAT
           % get the number of summary statistics
           if nbsim == 1
```

```
            nb_sumstats = sum([length(heterozygosity), length(all_range), length(private_alleles(1:end)),
length(pairwise_fst)]);
        end

        % get the mean and variance of the sumstat matrix
        sim_sum_mean = mean(sumstat,2);
        sim_sum_var = var(sumstat,0,2);

        % save the mean in the mean file
        fprintf(means_file,'%g ',sim_sum_mean);
        fprintf(means_file,'\n');

        % save the variance in the var file
        fprintf(var_file,'%g ',sim_sum_var);
        fprintf(var_file,'\n');

      end % END OF IF MS_RESULT IS EMPTY

    end % END OF LOOP TO SIMULATE DATA from sim 1 to nbsim(file)
    fclose(means_file); % close the file with the mean summary stats
    fclose(var_file); % close the file with the mean summary stats

end; % END OF LOOP FROM FILE 1 TO NBFILES

% CALCULATING THE SUMMARY STATS FOR THE OBSERVED DATA

observed_file = fopen('observed.txt','r')

nb_obs = 1;

% Open file to save the mean and variance across loci for each observed repetition
% each line corresponds to a observed repetition
fid_mean = fopen('obs_mean_sumstats.txt','w');
fid_var = fopen('obs_var_sumstats.txt','w');

% Go from repetition 1 to repetition nb_obs
for rep = 1: nb_obs
    rep

    % open file to save the observed sumstats and print the header
    mkdir(['run_', num2str(rep)])
    fid=fopen(['run_', num2str(rep), '/obs_sum_stats.txt'],'w'); %Creation of a text file
    fprintf(fid,'ov_nb_all,He_pop1, He_pop2, He_pop3,
He_pop4,Ov_He,all_range_pop1,all_range_pop2,all_range_pop3,all_range_pop4,ov_all_range,nb_all_pop1,
nb_all_pop2,nb_all_pop3,nb_all_pop4,priv_all_pop1,priv_all_pop2,priv_all_pop3,priv_all_pop4,fst_pop1_2,fs
t_pop1_3,fst_pop1_4,fst_pop2_3,fst_pop2_4,fst_pop3_4,ov_fst\n');% Print the results in the text file

    % Read the data for each locus
    for locus=1:handles.nb_loci

      % read first element which contains the number of alleles
      obs_nb_all = fscanf(observed_file, '%i', 1);
      % read the data for each locus and creates a matrix from the allele frequencies
      obs_loc = fscanf(observed_file, '%i', [obs_nb_all nb_pop+1]);

      %HETEROZYGOSITY
      heterozygosity = het(obs_loc(:,1:nb_pop), obs_nb_all);

      %ALLELIC RANGE
      all_range = a_range(obs_loc, obs_nb_all, nb_pop);
```

```
    % NUMBER OF ALLELES EACH POP AND PRIVATE ALLELES
    private_alleles = [priv_all(obs_loc(:,1:nb_pop), obs_nb_all, nb_pop)]';

    %PAIRWISE FST
    pairwise_fst = pair_fst(obs_loc(:,1:nb_pop),  heterozygosity, nb_pop, obs_nb_all);

    % Save the summary statistics
    fprintf(fid,'%g ', heterozygosity,all_range,private_alleles(1:end),pairwise_fst);
    fprintf(fid,'\n');
  end
  fclose(fid); %close file to save the obs_sum_stats

  % Get the mean across loci for the observed data
  % read the file with the observed sumstat for each locus
  fid_sumstat = fopen(['run_', num2str(rep), '/obs_sum_stats.txt'],'r');
  % skip first line containing header text
  temp = fgets(fid_sumstat);
  %create a matrix where each line is a locus and each column a statistic
  sumstat_obs = fscanf(fid_sumstat, '%g', [nb_sumstats handles.nb_loci]);
  %sumstat_obs = sumstat_obs.';

  % get the mean across loci for each sumstat
  %sumstat_obs_mean = mean(sumstat_obs,1);
  sumstat_obs_mean = mean(sumstat_obs,2);

  % get the variance across loci for each sumstat
  sumstat_obs_var = var(sumstat_obs,0,2);

  % save mean in a new file
  fprintf(fid_mean,'%g ', sumstat_obs_mean);
  fprintf(fid_mean,'\n');

  % save variance in a file
  fprintf(fid_var,'%g ', sumstat_obs_var);
  fprintf(fid_var,'\n', sumstat_obs_var);

  % close the observed file from where sumstats were read
  fclose(fid_sumstat);
end;

%close input file connection to read
fclose(observed_file);

% close the file
fclose(fid_mean);

% close the variance file
fclose(fid_var);

% REJECTION STEP                                        %
% CALCULATING THE DISTANCE BETWEEN THE SIMULATED AND OBSERVED DATA  %
% FIRST STEP: COMPUTE THE TOLERANCE as the tol_level quantile of 10000 distances
% SECOND STEP: STANDARDIZE THE SUMSTATS

tol_level = 0.01

%nbsim=handles.Nsim;
% if nbsim<=10000; %allows you to use fewer simulations to test that it is working
%    nbsim_tol=nbsimfile(1);
```

```
% elseif nbsim>10000;
%    nbsim_tol=10000;
% end

% 1ST STANDARDIZE SUMSTAT reading the sim_sum_means_1.txt

% Open the files to save the mean and variance
mean_std_file = fopen('./simulations_database/mean_mean_std.txt', 'w');
var_std_file = fopen('./simulations_database/var_mean_std.txt', 'w');

% MEAN
% Open simulated data from file 1
sim_sumstats_mean_file = fopen('./simulations_database/sim_sum_means_1.txt','r');
% Read the simulated data from file 1
sim_data_mean = fscanf(sim_sumstats_mean_file, '%g', [ nb_sumstats nbsimfile(1)]);
% Close the simulated datafile
fclose(sim_sumstats_mean_file);

% Invert the matrix to have nbsumstat * nbsim matrix
% i.e. each row is a sumstat and each column a simulation
sim_data_mean = sim_data_mean';

% Get the mean and standard deviation of each sum_stat over number of
% simulations. This will be used to standardize the observed and simulated sum stat
mean_sim_mean = mean(sim_data_mean,1);
std_sim_mean = std(sim_data_mean,0,1);

% Standardize the simulated data
% The second step avoids division by zero when the std is zero
sim_data_mean = (sim_data_mean-(ones(nbsimfile(1),1)*mean_sim_mean));
evaluate = std_sim_mean ~= 0;
sim_data_mean(:,evaluate) = sim_data_mean(:,evaluate) ./ (ones(nbsimfile(1),1)*std_sim_mean(evaluate));

% Write the relative mean and std sumstat into the file
fprintf(mean_std_file, '%5.5f ', mean_sim_mean, std_sim_mean);

% VARIANCE
% Read the simulated data from file 1
sim_sumstats_var_file = fopen('./simulations_database/sim_sum_var_1.txt','r');
% read the simulated data
sim_data_var = fscanf(sim_sumstats_var_file, '%g', [nb_sumstats nbsimfile(1) ]);
% Close the simulated datafile
fclose(sim_sumstats_var_file);
% invert the matrix
sim_data_var = sim_data_var';

% Get the mean and standard deviation of each sum_stat over the tol_sim
% This will be used to standardize the observed and simulated sum stat
mean_sim_var = mean(sim_data_var,1);
std_sim_var = std(sim_data_var,0,1);

% Standardize the simulated data var
% The second step avoids division by zero when the std is zero
sim_data_var = (sim_data_var-(ones(nbsimfile(1),1)*mean_sim_var));
evaluate = std_sim_var ~= 0;
sim_data_var(:,evaluate) = sim_data_var(:,evaluate) ./ (ones(nbsimfile(1),1)*std_sim_var(evaluate));

% Write the relative mean and std sumstat of the variance across loci into the file
fprintf(var_std_file, '%5.5f ', mean_sim_var, std_sim_var);
```

```
% Close the files to save mean and std of mean and variance sumstats
fclose(mean_std_file);
fclose(var_std_file);

% 2nd STANDARDIZE OBSERVED SUMSTAT

% Open the file to read the observed mean sumstat
fid_mean = fopen('obs_mean_sumstats.txt','r');
fid_var = fopen('obs_var_sumstats.txt','r');

% Open the file to save the relative obs sumstat and save the mean and
% standard deviation of the simulated sumstats
obs_relative_file_mean = fopen('obs_rel_sumstat_mean.txt', 'w');
obs_relative_file_var = fopen('obs_rel_sumstat_var.txt', 'w');

% go from datafile 1 to the nb_observations
for rep = 1: nb_obs

   % MEAN
   % Read the observed data for the rep observation
   obs_data_mean = fscanf(fid_mean,'%g',nb_sumstats);

   % Standardize the observed data mean
   obs_data_mean = (obs_data_mean' - mean_sim_mean);
   obs_data_mean(evaluate) = obs_data_mean(evaluate) ./ std_sim_mean(evaluate);

   % Write the relative observed sumstat into the file
   fprintf(obs_relative_file_mean, '%5.5f ', obs_data_mean);
   fprintf(obs_relative_file_mean, '\n');

   % VAR
   % Read the observed data for the rep observation
   obs_data_var = fscanf(fid_var,'%g',nb_sumstats);

   % Standardize the observed data var
   obs_data_var = (obs_data_var' - mean_sim_var);
   obs_data_var(evaluate) = obs_data_var(evaluate) ./ std_sim_var(evaluate);

   % Write the relative observed sumstat into the file
   fprintf(obs_relative_file_var, '%5.5f ', obs_data_var);
   fprintf(obs_relative_file_var, '\n');

end; % end for from repetition 1 to nbrep

% Close the files
fclose(obs_relative_file_mean);
fclose(obs_relative_file_var);
fclose(fid_mean);
fclose(fid_var);

% 3rd REJECTION STEP
% compute the distance and ACCEPT closest simulations

% Open observed files
obs_relative_file_mean = fopen('obs_rel_sumstat_mean.txt', 'r');
obs_relative_file_var = fopen('obs_rel_sumstat_var.txt', 'r');

% go from file 1 to nbfiles
for file = 1:nbfiles
```

```
% we don't need to reopen file 1 because this was done when
% standardizing the allele frequencies
if file > 1

    % MEAN
    % Open simulated datafile
    sim_sumstats_mean_file = fopen(['./simulations_database/sim_sum_means_', num2str(file),'.txt'],'r');
    % Read the simulated datafile
    sim_data_mean = fscanf(sim_sumstats_mean_file, '%g', [ nb_sumstats nbsimfile(file)]);
    % Close the simulated datafile
    fclose(sim_sumstats_mean_file);
    % transpose the matrix
    sim_data_mean = sim_data_mean';

    % Standardize the simulated data mean
    % The second step avoids division by zero when the std is zero
    sim_data_mean = (sim_data_mean-(ones(nbsimfile(file),1)*mean_sim_mean));
    evaluate = std_sim_mean ~= 0;
    sim_data_mean(:,evaluate) = sim_data_mean(:,evaluate) ./
ones(nbsimfile(file),1)*std_sim_mean(evaluate));

    % VAR
    % Open simulated data
    sim_sumstats_var_file = fopen(['./simulations_database/sim_sum_var_', num2str(file),'.txt'],'r');
    % read simulated data
    sim_data_var = fscanf(sim_sumstats_var_file, '%g', [ nb_sumstats nbsimfile(file)]);
    % Close the simulated datafile
    fclose(sim_sumstats_var_file);
    % transpose the matrix
    sim_data_var = sim_data_var';

    % Standardize the simulated data var
    % The second step avoids division by zero when the std is zero
    sim_data_var = (sim_data_var-(ones(nbsimfile(file),1)*mean_sim_var));
    evaluate = std_sim_var ~= 0;
    sim_data_var(:,evaluate) = sim_data_var(:,evaluate) ./ (ones(nbsimfile(file),1)*std_sim_var(evaluate));

end % end of IF file > 1

% go from datafile 1 to the nb_observations
for rep = 1: nb_obs

    rep

    % get the observed sumstats
    obs_data_mean = fscanf(obs_relative_file_mean, '%g', nb_sumstats);
    obs_data_var = fscanf(obs_relative_file_var, '%g', nb_sumstats);

    % Initialize dist and index matrices
    dist_1_mean = [];
    dist_1_var = [];
    dist_1 = [];

    final_dist_1_mean=[];
    final_dist_1_var=[];
    final_dist_1=[];

    index_1_mean=[];
    index_1_var=[];
    index_1=[];
```

```
    final_sim_data_mean=[];
    final_sim_data_var=[];
    final_sumstat=[];


    % MEAN
    % COMPUTE THE DISTANCE
    % Compute the distance metric 1 between observed and sum_stat vectors
    dist_1_mean = sqrt(sum( (ones(nbsimfile(file),1)*obs_data_mean' - sim_data_mean) .^ 2, 2));

    % Get the accepted values of mean as a given proportion of
    % closest simulations (tol_level)
    index_1_mean = find( dist_1_mean < quantile(dist_1_mean, tol_level));
    final_dist_1_mean = dist_1_mean(index_1_mean);
    final_sim_data_mean = sim_data_mean(index_1_mean,:);

    % VAR
    % compute the distance
    dist_1_var = sqrt(sum( (ones(nbsimfile(file),1)*obs_data_var' - sim_data_var) .^ 2, 2));

    % Get the accepted values of variance as a given proportion of
    % closest simulations (tol_level)
    index_1_var = find( dist_1_var < quantile(dist_1_var, tol_level));
    final_dist_1_var = dist_1_var(index_1_var);
    final_sim_data_var = sim_data_var(index_1_var,:);

    % MEAN + VAR
    % Get accepted distances putting together the mean and variance together
    dist_1 = sum([dist_1_mean dist_1_var], 2);

    % Get the accepted values of sum mean and variance
    index_1 = find( dist_1 < quantile(dist_1, tol_level));
    final_dist_1 = dist_1(index_1);
    final_sumstat = [sim_data_mean(index_1,:) sim_data_var(index_1,:)];

    % Save the dist and the index into a file
    text = repmat(' %g', [1 nb_sumstats]);
    file_name = ['run_', num2str(rep) ,'/distance_sum_state_mean_', int2str(file), '.txt'];
    save_dist_1_mean = fopen(file_name, 'w');
    fprintf(save_dist_1_mean, ['%i %g ', text, '\n'], [index_1_mean final_dist_1_mean
final_sim_data_mean]');
    fclose(save_dist_1_mean);

    % Save the dist and the index into a file
    file_name = ['run_', num2str(rep) ,'/distance_sum_state_var_', int2str(file), '.txt'];
    save_dist_1_var = fopen(file_name, 'w');
    fprintf(save_dist_1_var, ['%i %g ', text, '\n'], [index_1_var final_dist_1_var final_sim_data_var]');
    fclose(save_dist_1_var);

    % Save the dist and the index into a file
    file_name = ['run_', num2str(rep) ,'/distance_sum_state_', int2str(file), '.txt'];
    save_dist_1 = fopen(file_name, 'w');
    fprintf(save_dist_1, ['%i %g ', text, text, '\n'], [index_1 final_dist_1 final_sumstat]');
    fclose(save_dist_1);
  end;
end;

% Close observed files
fclose(obs_relative_file_mean);
```

```
fclose(obs_relative_file_var);

msgbox('ABC finished with no errors!')

toc


function a_range = a_range(data,numb_all,numb_pop);

        ov_all_range=abs(data(1,numb_pop+1)-data(numb_all,numb_pop+1));

        %searching for the first and last allele of each population
       for j=1:numb_pop
         done_fpop(j)=0;
         done_lpop(j)=0;
          for i=1:numb_all

             %first allele
             if (data(i,j)~=0) & (done_fpop(j)==0)
                fir_all_pop(j)=data(i,numb_pop+1);
                done_fpop(j)=1;%first allele found
             end

             %last allele
             if (data(numb_all+1-i,j)~=0) & (done_lpop(j)==0)
                last_all_pop(j)=data(numb_all+1-i,numb_pop+1);
                done_lpop(j)=1;%last allele found
             end
          end

         a_range_pop(j)=abs(fir_all_pop(j)-last_all_pop(j));

       end

       a_range = [a_range_pop ov_all_range];

       return;
       %fprintf(fid2,'%g ',ov_all_range);


function [n_gaps]= n_gaps(data)

data_zeros=find(data==0)
gap=0

for i=1:length(data_zeros)-1
   gap=gap+(data_zeros(i)+1~=data_zeros(i+1))
end

gap=gap+1

return

% GETALLFREQ
% function that returns the allelic frequency for a given number of
% populations, given the microsatellite alllele lengths and the sample size
% of each population
% ARGS
%     mstat (vector) : vector of size nsam1+nsam2+nsam3+..+nsamn with
```

```
%                   allelic length in repetitions
%    nsam (vector) : vector with the sample size of each population.
%                   the lenght of nsam is the number of populations
function result = getallfreq(msat,nsam);

   % create empty matrix
   result = [];

   % get the minimum and maximum allele repetitions
   mx=max(msat);
   mn=min(msat);

   %creates a vector with all alleles that we can find in one simulation
   all=(mn:mx);

   %creates a matrix were each line are the allellic frequencies in
   %each pop
   if length(all)==1 % if there is only one allele
      result=nsam'; % data matrix is the same as the inverted nsam vector
   else

      % get the cumulative sum of nsam
      index = [0 cumsum(nsam)];

      % go pop by pop (line by line)
      % to get the allele frequencies of each population
      for i = 1:length(nsam)
         result(i,:) = hist(msat((index(i)+1):index(i+1)),all);
      end
   end

   % add the allele length at the last row
   result = [result; all];

return;

%This scri
sam_size=100; %number of samples in each population (the same for all pops)
Nsim=1000;
nb_pop=10;
N=1000;
mig_rate=0.01;
mut_rate=0.001;
npop_Isampled=10;
directory=['MS_command_IslModel','_nb_Pop-',num2str(nb_pop),'_N-',num2str(N),'_Mig_rate-
',num2str(mig_rate),'_Number_sampled_pops-',num2str(npop_Isampled)];
nb_sample=sam_size*npop_Isampled;

fid4=fopen('ms_result_hap.txt','r');
fid3=fopen('hap_stats.txt','w');

for i=1:6
   fgetl(fid4);
end

for j=1:Nsim

%    clear all;
%    clear data;
```

```
   clear line;
   d=fgetl(fid4);
   dl=length(d);
   line(1,(1:dl))=d;

   for i=2:nb_sample %gets the alleles for all populations
   line(i,(1:dl))=fgetl(fid4);
   end

%    %Compare between each allele for each population
%    %line=num2str(line)
   result=[1:nb_sample]; %generate a sequence of Nsam

   for a=1:nb_sample-1
      for b=a+1:nb_sample
         if strcmp(num2str(line(a,(1:dl))),num2str(line(b,(1:dl))))==1
      result(b)=result(a);
           end
      end
   end

   %Matrix with number allelic frequencies for each pop
   aux=[1,(1+sam_size):sam_size:nb_sample,nb_sample+1];
   data=[];
   mx=max(result);
   mn=min(result);
   all=(mn:mx);
   for h=1:length(aux)-1
        if h==1
             data=[hist(result(aux(h):(aux(h+1)-1)),all)];
           else
             data(h,:)=[hist(result(aux(h):(aux(h+1)-1)),all)];
        end
   end
 data=[data(:,sum(data,1)~=0)]';

 %all(sum(data,1)~=0)
 nb_all=length(data(:,1))
 eval(['cd ..']);
 eval(['cd ..']);

 ht = het(data, nb_all);
 fst = pair_fst(data, ht, npop_lsampled, nb_all);


 %fid2=statistic(data,1,1,0,0,0,0,directory,fid2,j);
 eval(['cd Simulations\',directory]);
%  eval(['cd simulations']);
%  eval(['cd ' directory]);

 fprintf(fid3,'%0.3g ', ht);
 fprintf(fid3,'%0.3g ', fst);
 fprintf(fid3,'\n');


    for k=1:4
      fgetl(fid4);
   end

end
```

```
fclose(fid4)
fclose(fid3)

load hap_stats.txt
load ms_sumstats.txt
[rh,ch]=size(hap_stats)
[rm,cm]=size(ms_sumstats)

Fst_hap_data=mean(hap_stats(:,ch))
Fst_mcs_data=mean(ms_sumstats(:,cm))
Fst_kallelesM=1/(1+(4*N*mig_rate*nb_pop^2)/(nb_pop-1)^2) %Slatkin 1995

Fst_infSiteM=1/(1+4*N*mut_rate+4*N*mig_rate)

y=linspace(1,300,0.001);
hist(hap_stats(:,ch))
%axis([0 0.05 0 300])
hold on
plot(Fst_hap_data,y,'r-')
%text(Fst_data,1,'Fst data')
hold on
plot(Fst_infSiteM,y,'g-')
text(0.15,250,' red: hap data average; green: InfSiteModel Fst; blue: microssat data average; black:
kAllelModel Fst')
hold on
plot(Fst_kallelesM,y,'k-')
%text(Fst_kallelesM,1,'Fst kallelesM')
hold on
plot(Fst_mcs_data,y,'c-')
xlabel('Fst')
ylabel('simulations')
text(0.03,250,' red: hap; g: InSM; b: ms; b: kAlM')
title(['nb Pop-',num2str(nb_pop),'N:',num2str(N),' MigRate:',num2str(mig_rate),'
NumberSampledPops:',num2str(npop_lsampled),' MutRate:',num2str(mut_rate),' red: hap; g: InSM; b: ms; b:
kAlM'])

 % Compute Heterozygosity from a nb_pop by nb_all matrix
% This function returns the Heterozygosity of each population and the overall
% heterozygosity
% ARGS
%   data(matrix) : nb_all by nb_pop matrix with the absolute allele frequencies
%   num_all(integer) : number of alleles
function het = het(data, numb_all);

 frequencies=data';
 sum_freq=1./sum(frequencies,2);
 rel_freq=frequencies.*(sum_freq*ones(1,numb_all));
 Heterozygosity = 1-sum(rel_freq.^2,2);

 ov_frequencies=sum(sum(frequencies,2),1);
 ov_rel_freq=sum(frequencies,1)./ov_frequencies;
 ov_rel_freq=ov_rel_freq.^2;
 Ov_Heterozygosity = 1-sum(ov_rel_freq,2);

 het = [Heterozygosity' Ov_Heterozygosity];

return;

%Compute number of gaps from the allelic frequencies vector
%ARGS:
```

```
%n_pop: number of populations
%data(matrix): nb_allleles, nb_pops+1

function n_gaps= n_gaps(n_pop,data)

gap=zeros(1,n_pop)
all=data(:,n_pop+1);

for i=1:n_pop
   %nb_gaps(i)==0
   aux=data(:,i)
   aux=[aux(aux~=0) all(aux~=0)]

   for a=1:length(aux(:,2))-1
      if aux(a+1,2)~=aux(a,2)+1
   gap(i)=gap(i)+1
      end
   end
end

n_gaps=gap;

return

%This function returns fst pairwise values and overall fst value
% ARGS:
%  data(matrix) : nb_all by nb_pop matrix with the absolute allele frequencies
%  hetrz(matrix) : 1 by nb_pop+1 matrix with the heterozygosity of each
%  population and the overall heterozygosity in the nb_pop+1
%  nb_all(integer) : number of alleles
%  nb_pop(integer) : number of populations


function pair_fst_pop = pair_fst(data, hetrz, nb_pop, nb_all);

   freq=[];
      for i=1:nb_pop
      for j=(i+1):nb_pop
        he_local_(i,j)=(hetrz(i)+hetrz(j))/2;

            % Total heterozigoty (heterozigoty of every two pops as one)
            freq=[sum(data(:,[i j]),2)];
            relative_freq=([1./sum(freq,1)].*ones(nb_all,1));
            relative_freq=(relative_freq.*freq).^2;
            he_total(i,j)=1-sum(relative_freq,1); %returns a matrix were each element is the heterozigosity for
each two pops

            %CALCULATION OF Fst PAIRWISE
            if he_total(i,j)==0
               fst_pop(i,j)=0;
            else
               fst_pop(i,j)=(he_total(i,j)-he_local_(i,j))/he_total(i,j); %matrix (nb_pop-1)*nb_pop
            end;
         end;
      end;

      fst_result=[];
      for i=1:nb_pop-1
         fst_result=[fst_result fst_pop(i,((i+1):nb_pop))];
      end
```

```
    %Overall Fst
    ov_fst_loc=sum(hetrz(1:nb_pop))/nb_pop; %Sum of heterozygosity of each population divided by the
number of populations

    if hetrz(nb_pop+1)==0
      ov_fst=0;
    else
      ov_fst=(hetrz(nb_pop+1)-ov_fst_loc)/hetrz(nb_pop+1);
    end


   pair_fst_pop = [fst_result ov_fst]; %vector were each element is fst_pairwise values (by this order: pop 1
vs all other pops, pop 2 vs all other, etc...) followed by overall fst value

return;

% Get the number of private alleles from a nb_pop by nb_all matrix
% This function returns the number of alleles of each population and
% the number of private alleles at each population
% ARGS
%   data(matrix) : nb_all(rows) by nb_pop(column) matrix with the absolute allele frequencies
%   num_all(integer) : number of alleles
%   num_pop(integer) : number of populations
function [private] = priv_all(data, num_all, num_pop);
   zero_data=(data==0);
   nb_all_pop = num_all-sum(zero_data,1);
   sum_zero_data=sum(zero_data,2);
   private_allele=zeros(num_all,num_pop);
   %for each allele
   for i=1:num_all
      %if there is a line with a private allele
       if sum_zero_data(i)==(num_pop-1)
          private_allele(i,:)=(zero_data(i,:)==0);
       end
   end
   private=[nb_all_pop sum(private_allele,1)];
return;

% Compute and saves in the sumstats file the statistics
% asked (heterozigosity; private alleles; allelic range; FST pairwise)and
% return a file connection (this must be closed by the user)

function[fid2]=statistic(data,he,fst,priv_allele,all_range,gaps,alleles,directory,fid2,sim)


      eval(['cd Simulations\',directory]);

         %Number of alleles:
         [nb_all,nb_pop]=size(data);
         %Number of populations (last column contains the allele size):
         nb_pop=nb_pop-1;

      %Save the summary statictics in a text file
      if (exist('ms_sumstats.txt','file')~=2) || sim==1

         %Creation of a text file which contains the allelic frequencies
         fid2=fopen('ms_sumstats.txt','w');

         %First line of the text file to explain which summary
```

```
   %statictics have been chosen

   %heterogosity
   if he==1
     fprintf(fid2,'Expected heterozygosity for each pop,Overall He,');
   end
   %Private alleles
   if priv_allele==1
      fprintf(fid2,'Number of private alleles for each pop,');
   end
   %Allelic range
   if all_range==1
      fprintf(fid2,'Allelic range for each pop,Overall allelic range,');
   end
   %Fst pairwise
   if fst==1
      fprintf(fid2,'Fst pairwise and Overall Fst,');
   end

   if gaps==1
   fprintf(fid2,'Number of gaps,');
   end

   if alleles==1
   fprintf(fid2,'Number of alleles');
   end

fprintf(fid2,'\n\n')

 end

eval(['cd ..']);
eval(['cd ..']);

%================= Heterogosity ===================================
if  he==1 || fst==1
   heterozygosity = het(data(:,1:nb_pop), nb_all);
end
   % Print the results in the text file
if he==1
   eval(['cd Simulations\', directory]);
   fprintf(fid2,'%0.3g ',heterozygosity);
   eval(['cd ..']);
   eval(['cd ..']);
 end

%============== Private alleles ====================================
if priv_allele==1
   %private_alleles = priv_all(data(:,1:nb_pop), nb_all,nb_pop);
   private_alleles = [priv_all(data(:,1:nb_pop), nb_all,nb_pop)]';

    % Print the results in the text file
    eval(['cd Simulations\', directory]);
   %fprintf(fid2,'%g ',private_alleles((nb_pop+1):(2*nb_pop)));
   fprintf(fid2,'%g ',private_alleles(1:end);
   eval(['cd ..']);
   eval(['cd ..']);
 end

%============= Allelic range ======================================
```

```
 if all_range==1
   allelic_range = a_range(data, nb_all, nb_pop);

  % Print the results in the text file
  eval(['cd Simulations\', directory]);
  fprintf(fid2,'%g ',allelic_range);
  eval(['cd ..']);
  eval(['cd ..']);
 end

%================= Fst pairwise and Overall fst ==================

if  fst==1
   fst = pair_fst(data(:,1:nb_pop),heterozygosity, nb_pop, nb_all);

   % Print the results in the text file
   eval(['cd Simulations\', directory]);
   fprintf(fid2,'%0.3g ',fst);
   eval(['cd ..']);
   eval(['cd ..']);
end

%============== Number of gaps ====================================
if gaps==1
   gaps = n_gaps(nb_pop,data)

   % Print the results in the text file
   eval(['cd Simulations\', directory]);
   fprintf(fid2,'%0.3g ',gaps);
   eval(['cd ..']);
   eval(['cd ..']);
end

%================== number of alleles ============================
if alleles==1

  % Print the results in the text file
   eval(['cd Simulations\', directory]);
   fprintf(fid2,'%0.3g ',nb_all);
   eval(['cd ..']);
   eval(['cd ..']);
end

fprintf(fid2,'\n');
eval(['cd Simulations\', directory]);
```

Appendix 5.2. R script for analysis of Admixture program

```
# DEFINITION OF FUNCTIONS USED: tranform, back_transform

# TRANSFORM: Transformation of parameters (as in  Hamilton et al 2005)
transform.ham <- function(posterior, prior_min, prior_max) {

        if(min(posterior) <= prior_min){
                x.tmp <- ifelse(posterior <= prior_min,max(posterior),posterior)
                x.tmp.min <- min(x.tmp)
                posterior <- ifelse(posterior <= prior_min, x.tmp.min,posterior)

        }
        if(max(posterior) >= prior_max){
                x.tmp <- ifelse(posterior >= prior_max,min(posterior),posterior)
                x.tmp.max <- max(x.tmp)
                posterior <- ifelse(posterior >= prior_max, x.tmp.max,posterior)
        }

        y <- -log( (tan( ((posterior - prior_min)/(prior_max - prior_min)) * (pi/2) ) )^(-1) )
        y
}

# BACK TRANSFORM Tranform back into the natural parameter scale
back.transform.ham <- function(posteriorTransformed, prior_min, prior_max) {
        x <- (2/pi)*(prior_max - prior_min) * atan(exp(posteriorTransformed))
        x
}

#########################################################
# SETTINGS                                              #
# CHANGE THESE VALUES!!!!!!!!!!!!!!!!!!!!               #
#                                                       #
# define the number of sumstats                         #
nb_sumstat <- 25                                        #
# define number of parameters
nb_param <- 11                                          #
# number of "observed" datasets analysed
nb_rep <- 10                                            #
# number of files with simulated data
nb_files <- 2
# number of simulations
nbsim <- 10000
#                                                       #
#                                                       #
#########################################################

# list to save the parameters of each observed dataset analysed

acc_param_mean <- list()
acc_sumstat_mean <- list()
acc_dist_mean <-  list()

acc_param_var <- list()
acc_sumstat_var <- list()
acc_dist_var <-  list()

acc_param <- list()
acc_sumstat <- list()
acc_dist <-  list()
```

```
# Go from file 1 to nb_file and get the accepted parameters and sumstat
for(file in 1:nb_files) {

	# Read the scaled parameters
	param <- matrix(scan(paste("./simulations_database/simparameters_", file,".txt",
sep="")),ncol=17,byrow=T)

	# Get the parameters into the normal scale

	# Transform the scaled parameters into the real values
	# create a new matrix to save the parameters of interest (mut_rate, N1, N2, N3, N4, Tadm2, (1-p3),
Tadm1, p1, Tsplit, Nanc)
	# this matrix has 11 columns and the same number of rows as distance (the number of rows of
distance is given by length(distance[,1])
	new_param <- matrix(,length(param[,1]),11)
	new_param[,1] <- param[,1]/(4*param[,17])  # mut_rate = theta/(4*RefN)
	new_param[,2] <- param[,2]*param[,17]      # N1=relative N1*RefN
	new_param[,3] <- param[,3]*param[,17]      # N2=relative N2*RefN
	new_param[,4] <- param[,4]*param[,17]      # N3=relative N3*RefN
	new_param[,5] <- param[,5]*param[,17]      # N4=relative N4*RefN
	new_param[,6] <- param[,6]*(4*param[,17])  # Tadm2(gen) = Tadm1*(4*RefN)
	new_param[,7] <- param[,7]                 # 1-P3
	new_param[,8] <- param[,8]*(4*param[,17])  # Tadm1(gen) = Tadm2*(4*RefN)
	new_param[,9] <- param[,10]                # P1
	new_param[,10] <- param[,13]*(4*param[,17]) # Tsplit(gen) = Tsplit*(4*RefN)
	new_param[,11] <- param[,16]*param[,17]            # Nanc=relative Nanc*RefN

	# Clear memory (remove param)
	rm(param)

	# Make the histograms to be sure that the priors were ok
	par(mfrow=c(3,4))
	for(i in 1:11) {
		hist(new_param[,i], nclass=20, freq=F)
	}


	# Go from repetition 1 to nb_obs and get the accepted parameters from file 1
	for(rep in 1:nb_rep) {

		# Read the distances file for the mean
		# this file has:
		#	1st column: index of accepted param
		#	2nd column: corresponding distance
		#	3rd and remaining: corresponding standardized sumstats
		dist_sum_mean <- matrix(scan(paste("./run_", rep, "/distance_sum_state_mean_", file, ".txt",
sep="")), byrow=T, ncol=2+nb_sumstat)

		# Get the index, dist variables and standardized sumstat
		index_at_param <- dist_sum_mean[,1]
		dist <- dist_sum_mean[,2]
		scaled_sumstat <- dist_sum_mean[,3:(nb_sumstat+2)]
		# clear memory
		rm(dist_sum_mean)

		# Get only the closest param (already sorted according to the distance)
		# the [[]] are here because param_reg is a list of matrices
		if(file==1) {
```

```
                # save sorted param
                acc_param_mean[[rep]] <- new_param[index_at_param,]

                # Sort the simulated sumstat according to the closest distance
                acc_sumstat_mean[[rep]] <- scaled_sumstat

                # Save the sorted distance
                acc_dist_mean[[rep]] <- dist

        }
        if(file > 1) {

                # save sorted param
                acc_param_mean[[rep]] <- rbind(acc_param_mean[[rep]],
new_param[index_at_param,])

                # Sort the simulated sumstat according to the closest distance
                acc_sumstat_mean[[rep]] <- rbind(acc_sumstat_mean[[rep]], scaled_sumstat)

                # Save the sorted distance
                acc_dist_mean[[rep]] <- c(acc_dist_mean[[rep]],dist)
        }

        # Read the distances file for the var
        # this file has:
        #       1st column: index of accepted param
        #       2nd column: corresponding distance
        #       3rd and remaining: corresponding standardized sumstats
        dist_sum_var <- matrix(scan(paste("./run_", rep, "/distance_sum_state_var_", file, ".txt",
sep="")), byrow=T, ncol=2+nb_sumstat)

        # Get the index, dist variables and standardized sumstat
        index_at_param <- dist_sum_var[,1]
        dist <- dist_sum_var[,2]
        scaled_sumstat <- dist_sum_var[,3:(nb_sumstat+2)]
        # clear memory
        rm(dist_sum_var)

        # Get only the closest param (already sorted according to the distance)
        # the [[]] are here because param_reg is a list of matrices
        if(file==1) {

                # save sorted param
                acc_param_var[[rep]] <- new_param[index_at_param,]

                # Sort the simulated sumstat according to the closest distance
                acc_sumstat_var[[rep]] <- scaled_sumstat

                # Save the sorted distance
                acc_dist_var[[rep]] <- dist

        }
        if(file > 1) {

                # save sorted param
                acc_param_var[[rep]] <- rbind(acc_param_var[[rep]], new_param[index_at_param,])

                # Sort the simulated sumstat according to the closest distance
                acc_sumstat_var[[rep]] <- rbind(acc_sumstat_var[[rep]], scaled_sumstat)
```

```
                # Save the sorted distance
                acc_dist_var[[rep]] <- c(acc_dist_var[[rep]],dist)
        }

        # Read the distances file for the mean+variance
        # this file has:
        #       1st column: index of accepted param
        #       2nd column: corresponding distance
        #       3rd and remaining: corresponding standardized sumstats
        dist_sum <- matrix(scan(paste("./run_", rep, "/distance_sum_state_", file, ".txt", sep="")),
byrow=T, ncol=2+(nb_sumstat*2))

        # Get the index, dist variables and standardized sumstat
        index_at_param <- dist_sum[,1]
        dist <- dist_sum[,2]
        scaled_sumstat <- dist_sum[,3:((nb_sumstat*2)+2)]
        # clear memory
        rm(dist_sum)


        # Get only the closest param (already sorted according to the distance)
        # the [[]] are here because param_reg is a list of matrices
        if(file==1) {

                # save sorted param
                acc_param[[rep]] <- new_param[index_at_param,]

                # Sort the simulated sumstat according to the closest distance
                acc_sumstat[[rep]] <- scaled_sumstat

                # Save the sorted distance
                acc_dist[[rep]] <- dist

        }
        if(file > 1) {

                # save sorted param
                acc_param[[rep]] <- rbind(acc_param[[rep]], new_param[index_at_param,])

                # Sort the simulated sumstat according to the closest distance
                acc_sumstat[[rep]] <- rbind(acc_sumstat[[rep]], scaled_sumstat)

                # Save the sorted distance
                acc_dist[[rep]] <- c(acc_dist[[rep]],dist)
        }


    } # end of for loop between repetition 1 and nb_obs
} # end of for loop between file 1 and nb_file

# AT THIS POINT
# we have the:
#       1 - accepted parameters of each repetition in a list - acc_param[[]]
#       2 - accepted standardizes sumstat of each repetition in a list - acc_sumstat[[]]
#       3 - accepted distances of each repetition in a list - acc_dist[[]]
# All these lists have the accepted values for each file, after each other
# We need to sort the parameters, and sumstat according to closest distance
# This way, it will be straighforward to analyse different tolerance levels
```

```
# SORT the acc_param and acc_sumstat according to acc_distance
for( rep in 1:nb_rep) {

        # MEAN
        # sort the distance of observed dataset rep
        aux <- order(acc_dist_mean[[rep]])
        acc_dist_mean[[rep]] <- acc_dist_mean[[rep]][aux]

        # re-order acc_param and acc_sumstat accordingly
        acc_param_mean[[rep]] <- acc_param_mean[[rep]][aux,]
        acc_sumstat_mean[[rep]] <- acc_sumstat_mean[[rep]][aux,]

        # VAR
        # sort the distance of observed dataset rep
        aux <- order(acc_dist_var[[rep]])
        acc_dist_var[[rep]] <- acc_dist_var[[rep]][aux]

        # re-order acc_param and acc_sumstat accordingly
        acc_param_var[[rep]] <- acc_param_var[[rep]][aux,]
        acc_sumstat_var[[rep]] <- acc_sumstat_var[[rep]][aux,]

        # MEAN + VAR
        # sort the distance of observed dataset rep
        aux <- order(acc_dist[[rep]])
        acc_dist[[rep]] <- acc_dist[[rep]][aux]

        # re-order acc_param and acc_sumstat accordingly
        acc_param[[rep]] <- acc_param[[rep]][aux,]
        acc_sumstat[[rep]] <- acc_sumstat[[rep]][aux,]

}


# Read the observed sum stat mean
obs_sum_stat_mean <- matrix(scan("obs_rel_sumstat_mean.txt"), byrow=T, ncol=nb_sumstat)


# Read the observed sum stat variance
obs_sum_stat_var <- matrix(scan("obs_rel_sumstat_var.txt"), byrow=T, ncol=nb_sumstat)


# Put the mean and variance together
obs_sum_stat <- cbind(obs_sum_stat_mean, obs_sum_stat_var)

legenda_hista <- c("mut_rate", "N1", "N2", "N3", "N4", "Tadm2", "1-p3", "Tadm1", "p1", "Tsplit", "Nanc")

# Go from observed 1 to nb_obs datasets
for( rep in 1:nb_rep) {

        print(paste("rep ", rep))

        # Save the Sorted acc_param, acc_dist and acc_sumstat for each repetition
        # this files have the posteriors for the parameters obtained with the rejection step
        # with the tolerance defined in the matlab code (usually 10%)
        # MEAN
        write.table(acc_param_mean[[rep]], paste("./run_", rep, "/acc_parameters_mean.txt", sep=""),
col.names=F, row.names=F)
        write.table(acc_dist_mean[[rep]], paste("./run_", rep, "/acc_dist_mean.txt", sep=""), col.names=F,
row.names=F)
```

```
        write.table(acc_sumstat_mean[[rep]], paste("./run_", rep, "/acc_sumstat_mean.txt", sep=""),
col.names=F, row.names=F)

        # VAR
        write.table(acc_param_var[[rep]], paste("./run_", rep, "/acc_parameters_var.txt", sep=""),
col.names=F, row.names=F)
        write.table(acc_dist_var[[rep]], paste("./run_", rep, "/acc_dist_var.txt", sep=""), col.names=F,
row.names=F)
        write.table(acc_sumstat_var[[rep]], paste("./run_", rep, "/acc_sumstat_var.txt", sep=""), col.names=F
row.names=F)

        # MEAN + VAR
        write.table(acc_param[[rep]], paste("./run_", rep, "/acc_parameters.txt", sep=""), col.names=F,
row.names=F)
        write.table(acc_dist[[rep]], paste("./run_", rep, "/acc_dist.txt", sep=""), col.names=F, row.names=F)
        write.table(acc_sumstat[[rep]], paste("./run_", rep, "/acc_sumstat.txt", sep=""), col.names=F,
row.names=F)

        # define the tolerance array with the tolerance levels (closest accepted points)
        #tol.array <-
c(min(length(acc_sumstat[[rep]][,1]),length(acc_sumstat_var[[rep]][,1]),length(acc_sumstat[[rep]][,1])), 5000,
1000, 500)
        tol.array <-
c(min(length(acc_sumstat_mean[[rep]][,1]),length(acc_sumstat_var[[rep]][,1]),length(acc_sumstat[[rep]][,1])),
250, 100)

        for(tol in 1:length(tol.array) ) {

                print(paste("tolerance ", tol.array[tol]))

                # Perform the Rejection Step analysis for different tolerance values
                # MEAN
                par(mfrow=c(4,3))
                for(i in 1:nb_param) {
                        hist(acc_param_mean[[rep]][1:tol.array[tol],i], main=legenda_hista[i])
                }
                dev.print(device=pdf, width=16, height=14, paste("./run_", rep, "/acc_param_tol_median_",
tol.array[tol],".pdf", sep=""))

                # VAR
                par(mfrow=c(4,3))
                for(i in 1:nb_param) {
                        hist(acc_param_var[[rep]][1:tol.array[tol],i], main=legenda_hista[i])
                }
                dev.print(device=pdf, width=16, height=14, paste("./run_", rep, "/acc_param_tol_var_",
tol.array[tol],".pdf", sep=""))

                # MEAN + VAR
                par(mfrow=c(4,3))
                for(i in 1:nb_param) {
                        hist(acc_param[[rep]][1:tol.array[tol],i], main=legenda_hista[i])
                }
                dev.print(device=pdf, width=16, height=14, paste("./run_", rep, "/acc_param_tol_",
tol.array[tol],".pdf", sep=""))


                # Perform the Regression step for a given tolerance
                # define the tol.array (this means that first, we do the regression accepting all values)
                # and then with the closest 1000, etc
```

```
                # MEAN
                # Transform the parameters to make the regression
                # (this is done to avoid that after the regression
                # the parameter values are outside the prior limits)
                # NOTE: NEED TO CHANGE PRIOR LIMITS!!!!

                tranf_param <- matrix(,length(acc_param_mean[[rep]][1:tol.array[tol],1]),nb_param)

                tranf_param[,1] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],1], 1e-5, 1e-3)   #
mut_rate = theta/(4*RefN)
                tranf_param[,2] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],2], 1000,15000) #
N1=relative N1*RefN
                tranf_param[,3] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],3], 1000,15000) #
N2=relative N3*RefN
                tranf_param[,4] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],4], 1000,15000) #
N3=relative N3*RefN
                tranf_param[,5] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],5], 1000,15000) #
N4=relative N4*RefN
                tranf_param[,6] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],6], 1,100)         #
Tadm2(gen)
                tranf_param[,7] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],7], 0,1)  # 1-P3
                tranf_param[,8] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],8], 0,100)         #
Tadm1(gen) = Tadm2*(4*RefN)
                tranf_param[,9] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],9], 0,1)  # P1
                tranf_param[,10] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],10], 1000,15000)
        # Tsplit(gen) = Tsplit*(4*RefN)
                tranf_param[,11] <- transform.ham(acc_param_mean[[rep]][1:tol.array[tol],11], 1000,15000)
        # Nanc=relative Nanc*RefN

                # Epanechnikov Kernel Weights
                regwt <- 1-
acc_dist_mean[[rep]][1:tol.array[tol]]^2/max(acc_dist_mean[[rep]][1:tol.array[tol]]^2)

                # Perform the Regression with tolerance equal to tol
                fit0 <- lm(tranf_param ~ acc_sumstat_mean[[rep]][1:tol.array[tol],], weights=regwt)

                # Compute predicted values
                x0p <- c(1, obs_sum_stat_mean[rep,])
                predmean <- numeric(nb_param)
                for(i in 1:nb_param) predmean[i] <- sum(x0p*coef(fit0)[,i], na.rm=T)

                # Add the residuals to the predicted mean
                post <- matrix(,tol.array[tol],nb_param)
                for(i in 1:nb_param) post[,i] <- residuals(fit0)[,i] + predmean[i]

                # Back transform the regressed parameters to the original scale
                post[,1] <-  back.transform.ham(post[,1], 1e-5, 1e-3) # mute rate
                post[,2] <-  back.transform.ham(post[,2], 1000, 15000) # N1
                post[,3] <-  back.transform.ham(post[,3], 1000, 15000) # N2
                post[,4] <-  back.transform.ham(post[,4], 1000, 15000) # N3
                post[,5] <-  back.transform.ham(post[,5], 1000, 15000) # N4

                post[,6] <-  back.transform.ham(post[,6], 1, 100) # Tadm2
                post[,7] <-  back.transform.ham(post[,7], 1, 1)  # 1-p3

                post[,8] <-  back.transform.ham(post[,8], 0, 100) # Tadm1
                post[,9] <-  back.transform.ham(post[,9], 0, 1)  # p1

                post[,10] <-  back.transform.ham(post[,10], 1000, 15000) # Tsplit
                post[,11] <-  back.transform.ham(post[,11], 1000, 15000)  # Nanc
```

```
		# plot the posteriors
		par(mfrow=c(4,3))
		for(i in 1:nb_param) {
			hist(post[,i], main=legenda_hista[i])
		}
		dev.print(device=pdf, width=16, height=14, paste("./run_", rep, "/mean_posterior_reg_",
tol.array[tol],".pdf", sep=""))


		# Save the posterior distributions
		if(tol > 10000){ t = 0.1
		} else {t <- tol/nbsim}
		write.table(post, file=paste("./run_", rep, "/reg_posterior_mean_tol_", t, "_rep_", rep, ".txt",
sep=""), row.names=F, col.names=F)

		# VAR
		# Transform the parameters to make the regression
		# (this is done to avoid that after the regression
		# the parameter values are outside the prior limits)
		# NOTE: NEED TO CHANGE PRIOR LIMITS!!!!

		tranf_param <- matrix(,length(acc_param_var[[rep]][1:tol.array[tol],1]),nb_param)

		tranf_param[,1] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],1], 1e-5, 1e-3)	#
mut_rate = theta/(4*RefN)
		tranf_param[,2] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],2], 1000,15000)	#
N1=relative N1*RefN
		tranf_param[,3] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],3], 1000,15000)	#
N2=relative N3*RefN
		tranf_param[,4] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],4], 1000,15000)	#
N3=relative N3*RefN
		tranf_param[,5] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],5], 1000,15000)	#
N4=relative N4*RefN
		tranf_param[,6] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],6], 1,100)	#
Tadm2(gen)
		tranf_param[,7] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],7], 0,1)	# 1-P3
		tranf_param[,8] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],8], 0,100)	#
Tadm1(gen) = Tadm2*(4*RefN)
		tranf_param[,9] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],9], 0,1)	# P1
		tranf_param[,10] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],10], 1000,15000) #
Tsplit(gen) = Tsplit*(4*RefN)
		tranf_param[,11] <- transform.ham(acc_param_var[[rep]][1:tol.array[tol],11], 1000,15000) #
Nanc=relative Nanc*RefN

		# Epanechnikov Kernel Weights
		regwt <- 1-acc_dist_var[[rep]][1:tol.array[tol]]^2/max(acc_dist_var[[rep]][1:tol.array[tol]]^2)

		# Perform the Regression with tolerance equal to tol
		fit0 <- lm(tranf_param ~ acc_sumstat_var[[rep]][1:tol.array[tol],], weights=regwt)

		# Compute predicted values
		x0p <- c(1, obs_sum_stat_var[rep,])
		predvar <- numeric(nb_param)
		for(i in 1:nb_param) predvar[i] <- sum(x0p*coef(fit0)[,i], na.rm=T)

		# Add the residuals to the predicted var
		post <- matrix(,tol.array[tol],nb_param)
		for(i in 1:nb_param) post[,i] <- residuals(fit0)[,i] + predvar[i]
```

```
		# Back transform the regressed parameters to the original scale
		post[,1] <-  back.transform.ham(post[,1], 1e-5, 1e-3) # mute rate
		post[,2] <-  back.transform.ham(post[,2], 1000, 15000) # N1
		post[,3] <-  back.transform.ham(post[,3], 1000, 15000) # N2
		post[,4] <-  back.transform.ham(post[,4], 1000, 15000) # N3
		post[,5] <-  back.transform.ham(post[,5], 1000, 15000) # N4

		post[,6] <-  back.transform.ham(post[,6], 1, 100) # Tadm2
		post[,7] <-  back.transform.ham(post[,7], 1, 1)   # 1-p3

		post[,8] <-  back.transform.ham(post[,8], 0, 100) # Tadm1
		post[,9] <-  back.transform.ham(post[,9], 0, 1)   # p1

		post[,10] <-  back.transform.ham(post[,10], 1000, 15000) # Tsplit
		post[,11] <-  back.transform.ham(post[,11], 1000, 15000)  # Nanc



		# plot the posteriors
		par(mfrow=c(4,3))
		for(i in 1:nb_param) {
			hist(post[,i], main=legenda_hista[i])
		}
		dev.print(device=pdf, width=16, height=14, paste("./run_", rep, "/var_posterior_reg_",
tol.array[tol],".pdf", sep=""))


		# Save the posterior distributions
		if(tol > 10000){ t = 0.1}
		else {t <- tol/nbsim}
		write.table(post, file=paste("./run_", rep, "/reg_posterior_var_tol_", t, "_rep_", rep, ".txt",
sep=""), row.names=F, col.names=F)

		# MEAN + VAR
		# Transform the parameters to make the regression
		# (this is done to avoid that after the regression
		# the parameter values are outside the prior limits)
		# NOTE: NEED TO CHANGE PRIOR LIMITS!!!!

		tranf_param <- matrix(,length(acc_param[[rep]][1:tol.array[tol],1]),nb_param)

		tranf_param[,1] <- transform.ham(acc_param[[rep]][1:tol.array[tol],1], 1e-5, 1e-3)  # mut_rate
= theta/(4*RefN)
		tranf_param[,2] <- transform.ham(acc_param[[rep]][1:tol.array[tol],2], 1000,15000)          #
N1=relative N1*RefN
		tranf_param[,3] <- transform.ham(acc_param[[rep]][1:tol.array[tol],3], 1000,15000)          #
N2=relative N3*RefN
		tranf_param[,4] <- transform.ham(acc_param[[rep]][1:tol.array[tol],4], 1000,15000)          #
N3=relative N3*RefN
		tranf_param[,5] <- transform.ham(acc_param[[rep]][1:tol.array[tol],5], 1000,15000)          #
N4=relative N4*RefN
		tranf_param[,6] <- transform.ham(acc_param[[rep]][1:tol.array[tol],6], 1,100)        #
Tadm2(gen)
		tranf_param[,7] <- transform.ham(acc_param[[rep]][1:tol.array[tol],7], 0,1)# 1-P3
		tranf_param[,8] <- transform.ham(acc_param[[rep]][1:tol.array[tol],8], 0,100)        #
Tadm1(gen) = Tadm2*(4*RefN)
		tranf_param[,9] <- transform.ham(acc_param[[rep]][1:tol.array[tol],9], 0,1)# P1
		tranf_param[,10] <- transform.ham(acc_param[[rep]][1:tol.array[tol],10], 1000,15000)       #
Tsplit(gen) = Tsplit*(4*RefN)
```

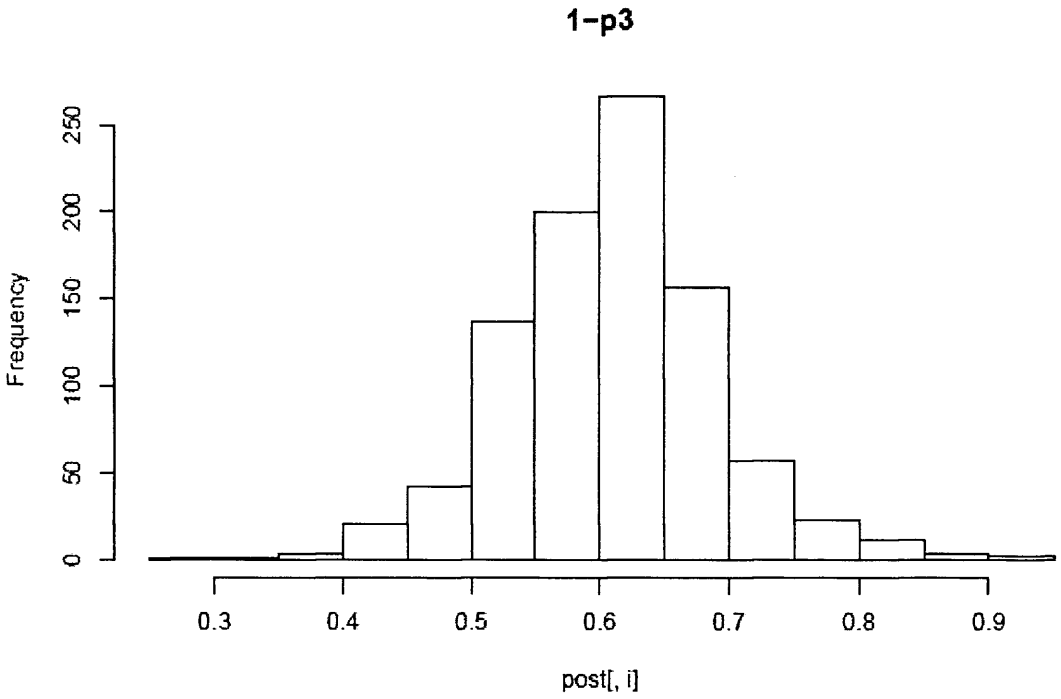
```
                tranf_param[,11] <- transform.ham(acc_param[[rep]][1:tol.array[tol],11], 1000,15000)      #
Nanc=relative Nanc*RefN

                # Epanechnikov Kernel Weights
                regwt <- 1-acc_dist[[rep]][1:tol.array[tol]]^2/max(acc_dist[[rep]][1:tol.array[tol]]^2)

                # Perform the Regression with tolerance equal to tol
                fit0 <- lm(tranf_param ~ acc_sumstat[[rep]][1:tol.array[tol],], weights=regwt)

                # Compute predicted values
                x0p <- c(1, obs_sum_stat[rep,])
                pred <- numeric(nb_param)
                for(i in 1:nb_param) pred[i] <- sum(x0p*coef(fit0)[,i], na.rm=T)

                # Add the residuals to the predicted
                post <- matrix(,tol.array[tol],nb_param)
                for(i in 1:nb_param) post[,i] <- residuals(fit0)[,i] + pred[i]

                # Back transform the regressed parameters to the original scale
                post[,1] <-  back.transform.ham(post[,1], 1e-5, 1e-3) # mute rate
                post[,2] <-  back.transform.ham(post[,2], 1000, 15000) # N1
                post[,3] <-  back.transform.ham(post[,3], 1000, 15000) # N2
                post[,4] <-  back.transform.ham(post[,4], 1000, 15000) # N3
                post[,5] <-  back.transform.ham(post[,5], 1000, 15000) # N4

                post[,6] <-  back.transform.ham(post[,6], 1, 100) # Tadm2
                post[,7] <-  back.transform.ham(post[,7], 1, 1)   # 1-p3

                post[,8] <-  back.transform.ham(post[,8], 0, 100) # Tadm1
                post[,9] <-  back.transform.ham(post[,9], 0, 1)   # p1

                post[,10] <-  back.transform.ham(post[,10], 1000, 15000) # Tsplit
                post[,11] <-  back.transform.ham(post[,11], 1000, 15000)  # Nanc

                # plot the posteriors
                par(mfrow=c(4,3))
                for(i in 1:nb_param) {
                        hist(post[,i], main=legenda_hista[i])
                }
                dev.print(device=pdf, width=16, height=14, paste("./run_", rep, "/posterior_reg_",
tol.array[tol],".pdf", sep=""))


                # Save the posterior distributions
                if(tol > 10000){ t = 0.1}
                else {t <- tol/nbsim}
                write.table(post, file=paste("./run_", rep, "/reg_posterior_tol_", t, "_rep_", rep, ".txt", sep=""),
row.names=F, col.names=F)

        }

}
```

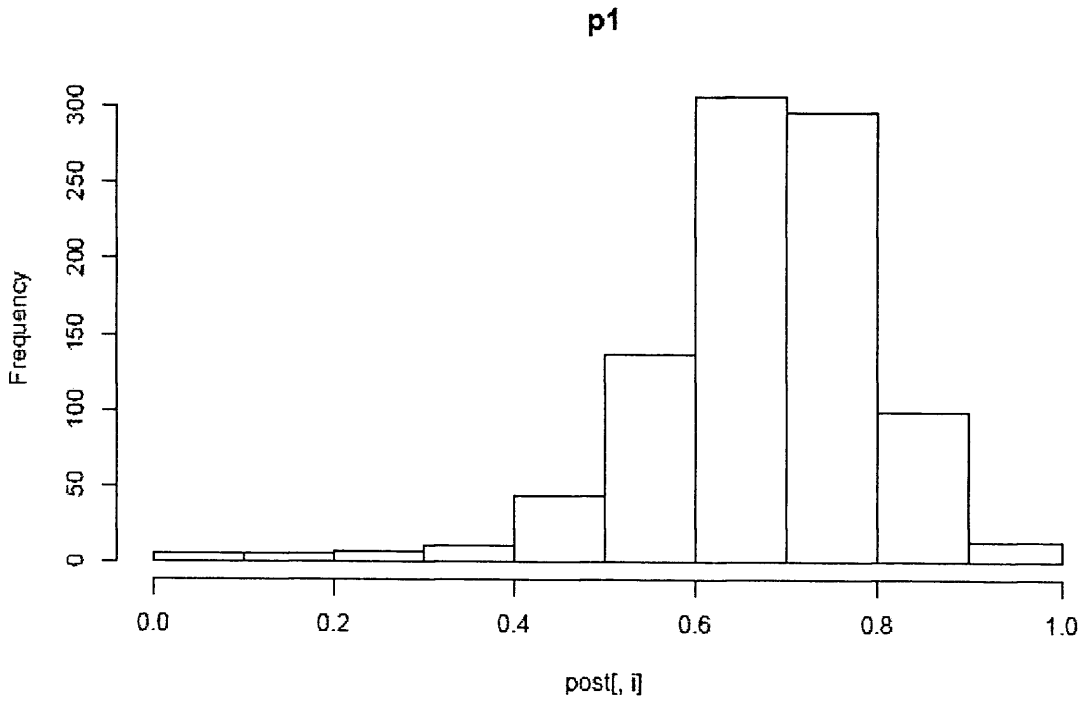
Appendix 5.3. R script for calculation of means, variance, and adjusted modal values of the posterior distribution.

```
#For analysis of 1adm and 2adm scenario
#calculates modal values (and sd) for P1, 1-P3, tadm2, tadm1, and tsplit

library(locfit)  # to use package
param9.fit1 <- locfit(~post[,9], alpha=0.7, xlim=c(min(0),max(post[,9]))) # to create density
find.mode <- function(sequence, param9.fit1) {
mode <- sequence[predict(param9.fit1,sequence)==max(predict(param9.fit1,sequence))]
mode
}
seqtest <- seq(0,1,0.001) #sets the intervals of .001
find.mode(seqtest,param9.fit1) #finds the peak at intervals of .001
sd (post[,9])
mean (post[,9])

param6.fit1 <- locfit(~post[,6], alpha=0.7, xlim=c(min(0),max(post[,6]))) # to create density
find.mode <- function(sequence, param6.fit1) {
mode <- sequence[predict(param6.fit1,sequence)==max(predict(param6.fit1,sequence))]
mode
}
seqtest <- seq(0,1,0.001) #sets the intervals of .001
find.mode(seqtest,param6.fit1) #finds the peak at intervals of .001
sd (post[,6])
mean (post[,6])

param10.fit1 <- locfit(~post[,10], alpha=0.7, xlim=c(min(0),max(post[,10]))) # to create density
find.mode <- function(sequence, param10.fit1) {
mode <- sequence[predict(param10.fit1,sequence)==max(predict(param10.fit1,sequence))]
mode
}
seqtest <- seq(0,1,0.001) #sets the intervals of .001
find.mode(seqtest,param10.fit1) #finds the peak at intervals of .001
sd (post[,10])
mean (post[,10])

param7.fit1 <- locfit(~post[,7], alpha=0.7, xlim=c(min(0),max(post[,7]))) # to create density
find.mode <- function(sequence, param7.fit1) {
mode <- sequence[predict(param7.fit1,sequence)==max(predict(param7.fit1,sequence))]
mode
}
seqtest <- seq(0,1,0.001) #sets the intervals of .001
find.mode(seqtest,param7.fit1) #finds the peak at intervals of .001
sd (post[,7])
mean (post[,7])

param8.fit1 <- locfit(~post[,8], alpha=0.7, xlim=c(min(0),max(post[,8]))) # to create density
find.mode <- function(sequence, param8.fit1) {
mode <- sequence[predict(param8.fit1,sequence)==max(predict(param8.fit1,sequence))]
mode
}
seqtest <- seq(0,1,0.001) #sets the intervals of .001
find.mode(seqtest,param8.fit1) #finds the peak at intervals of .001
sd (post[,8])
mean (post[,8])
```

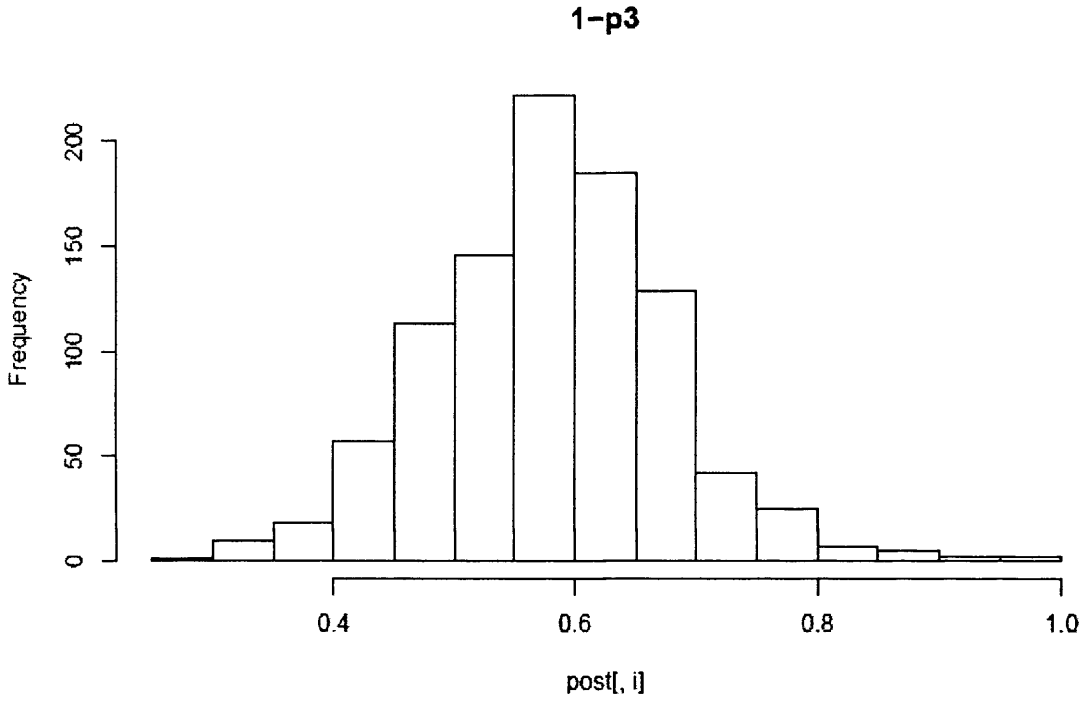
Appendix 5.4. Mean regression histograms for p1 and 1-p3 for two admixture events using 500,000 simulations.

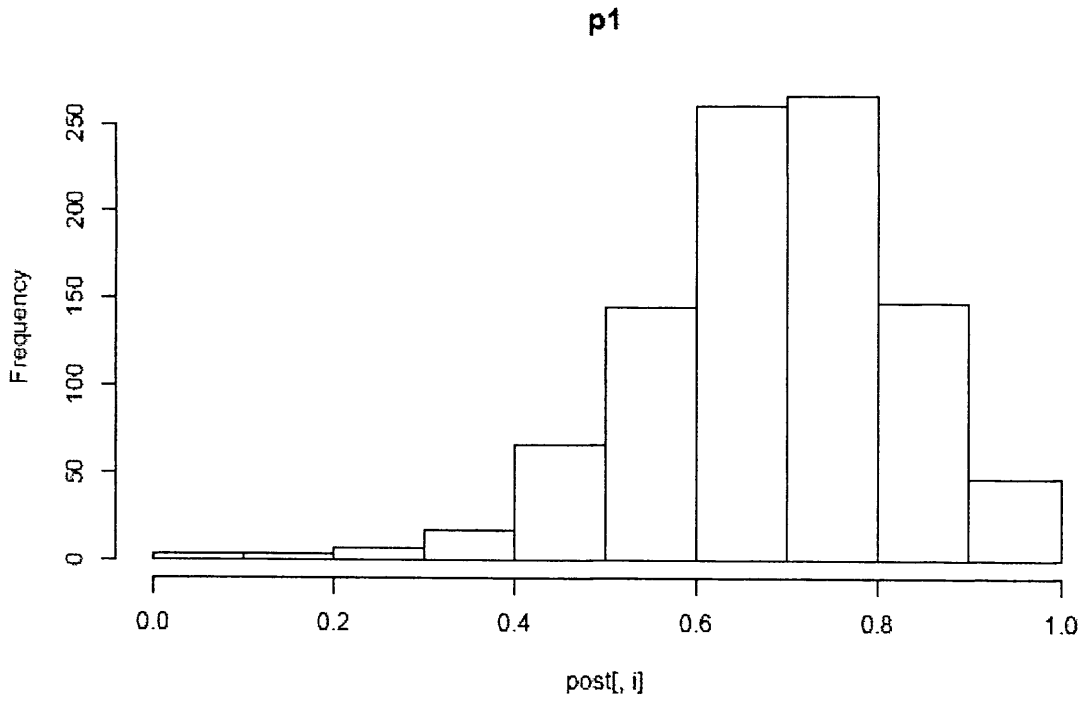
Dataset 1.



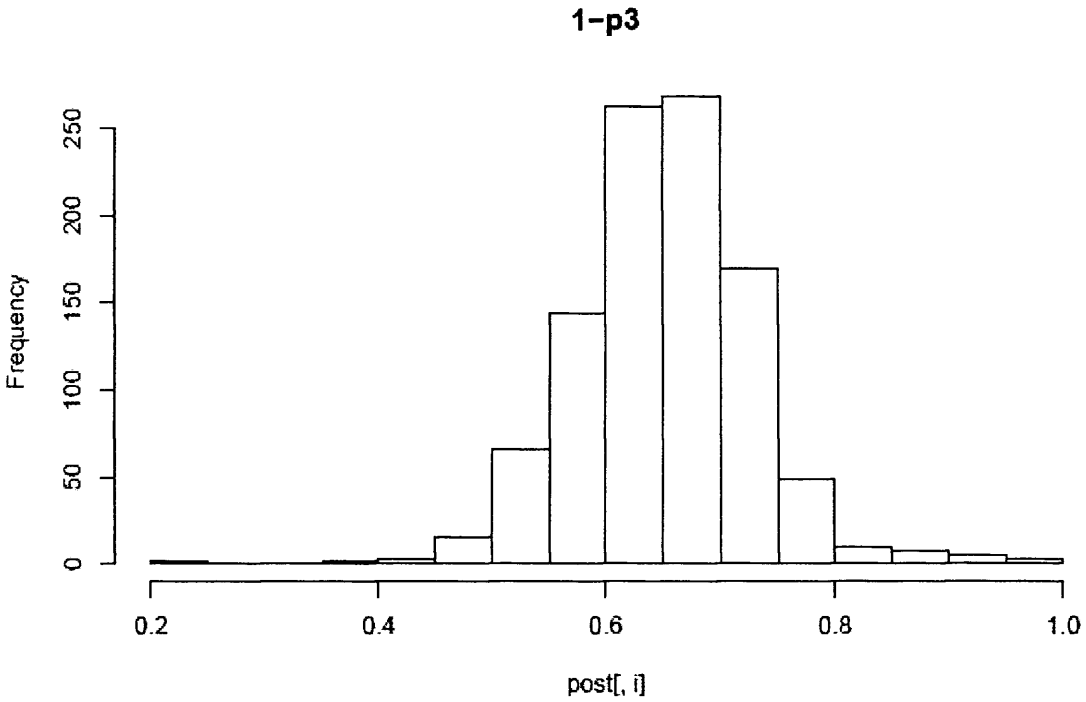


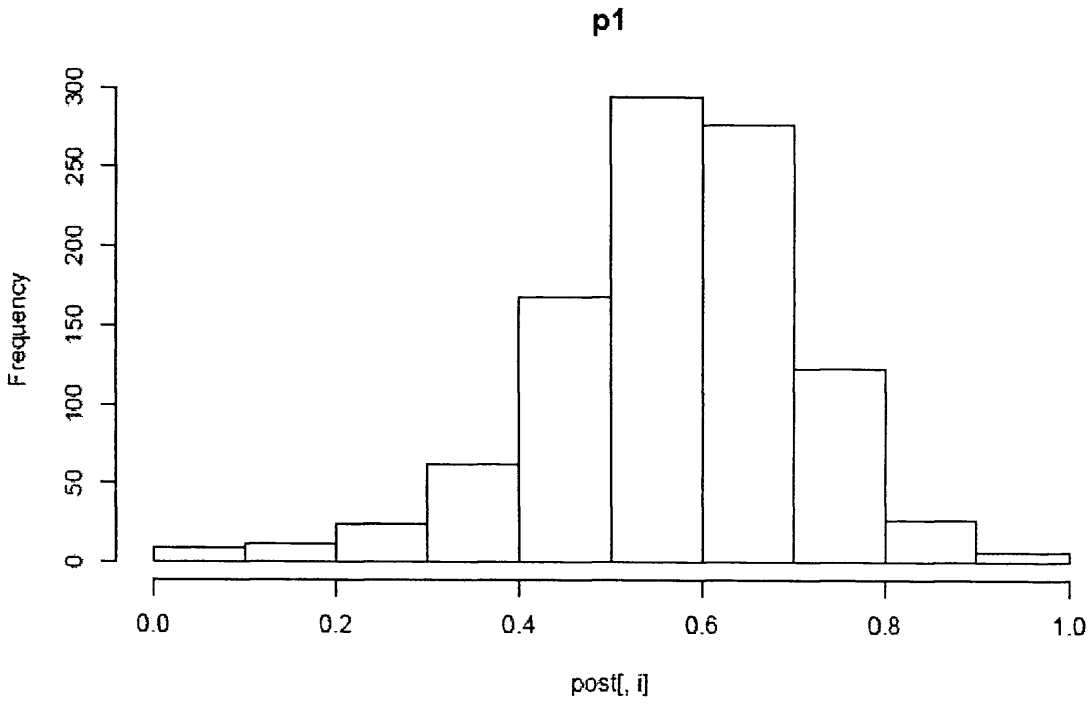
Dataset 2.


1−p3
Frequency
0
50
100
150
200
0.4
0.6
0.8
1.0
post[, i]

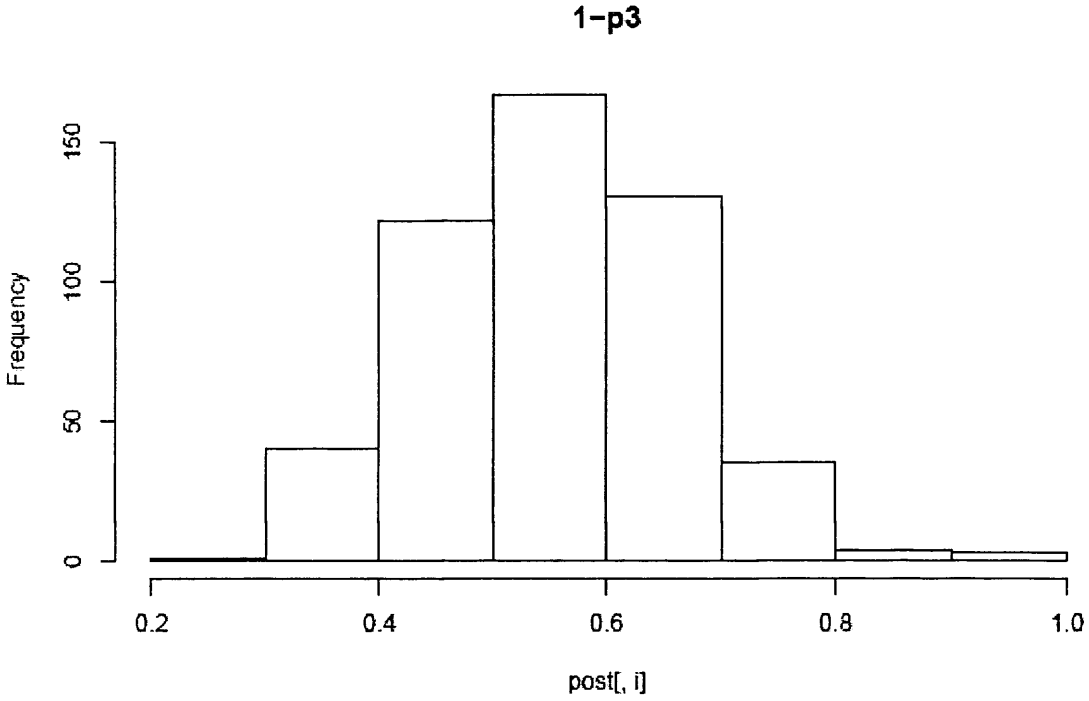

p1
Frequency
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0
post[, i]

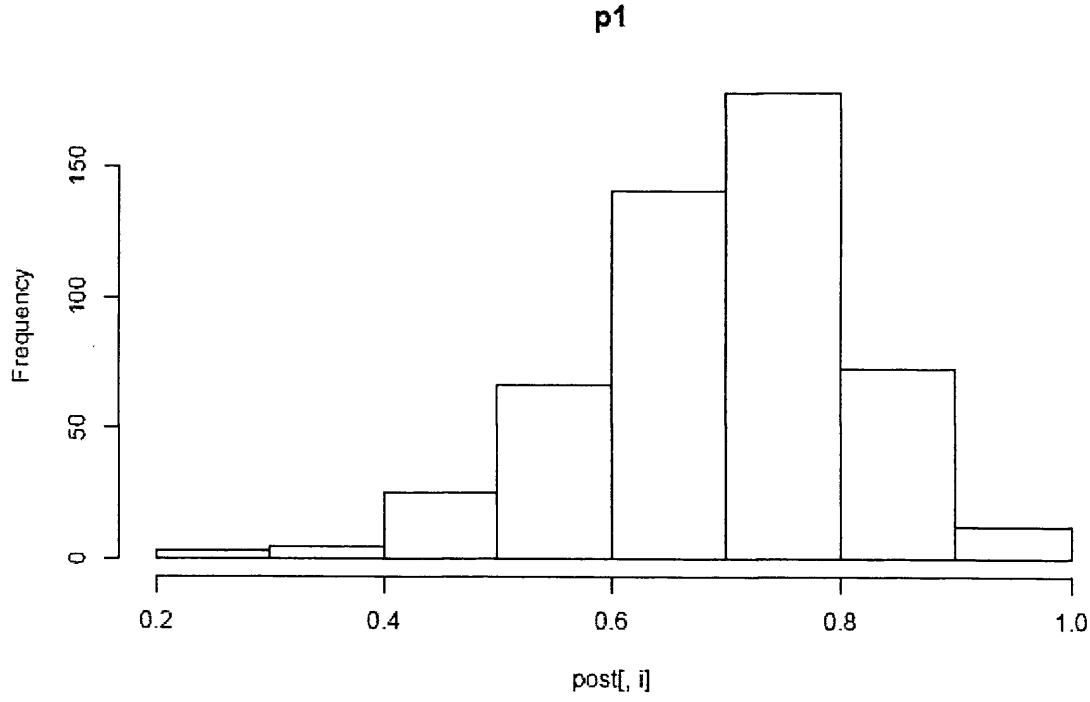
Dataset 3.


1-p3
Frequency
0
50
100
150
200
250
0.2
0.4
0.6
0.8
1.0
post[, i]

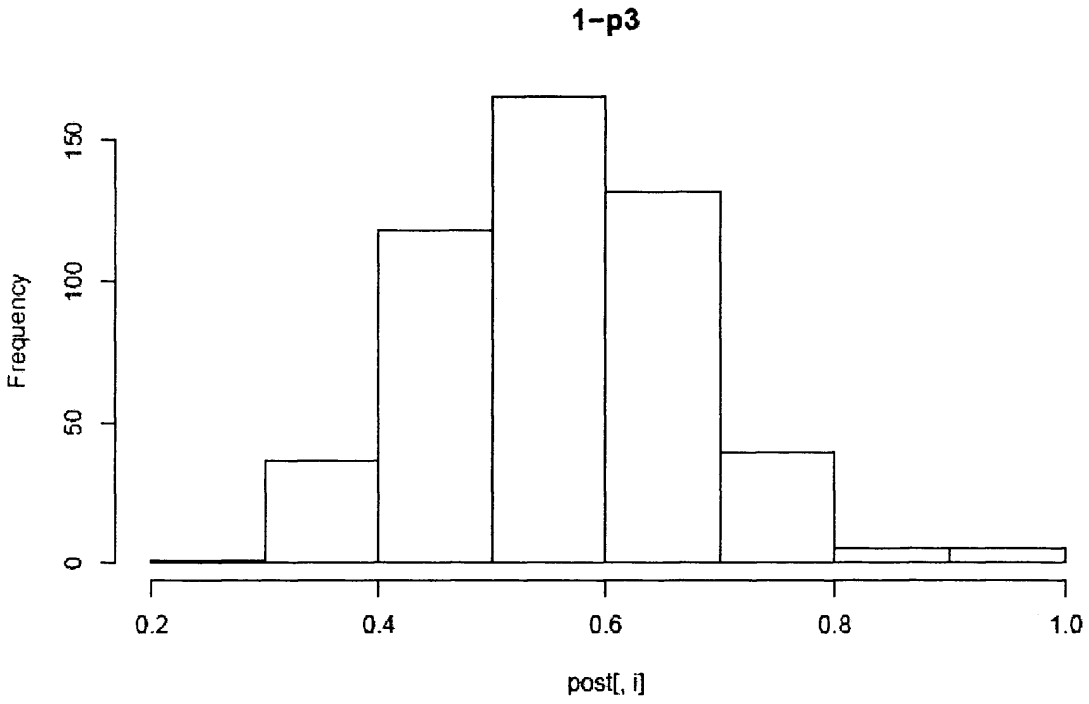

p1
Frequency
0
50
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
post[, i]

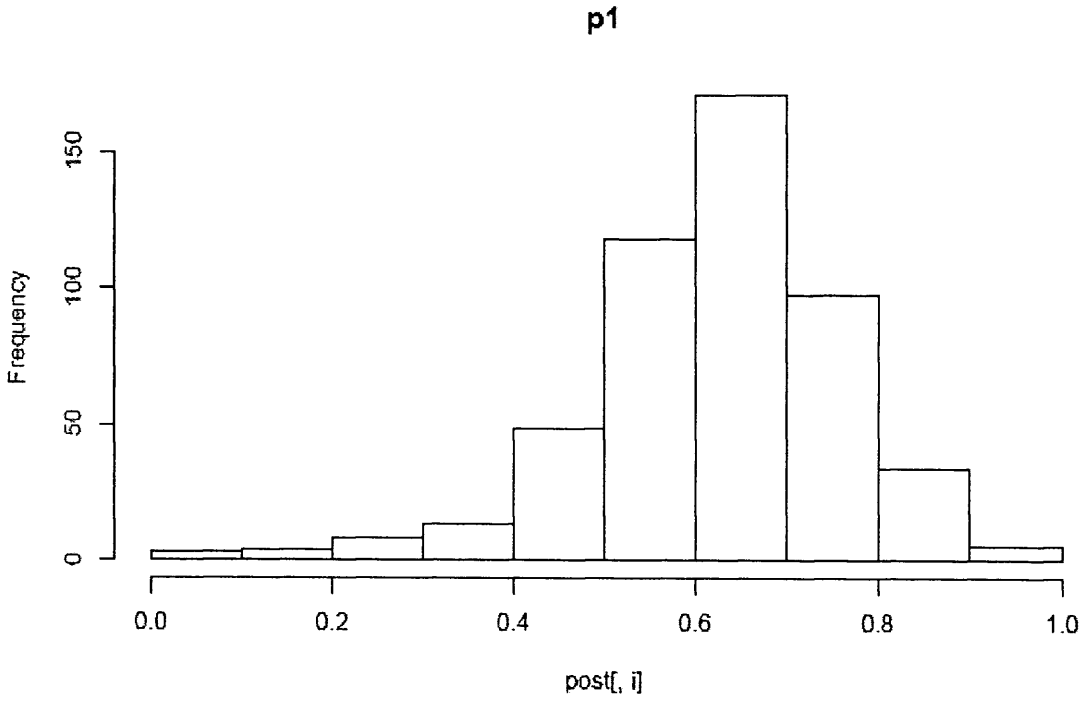
Dataset 4.


1−p3
Frequency
0
50
100
150
0.2
0.4
0.6
0.8
1.0
post[, i]


p1
Frequency
0
50
100
150
0.2
0.4
0.6
0.8
1.0
post[, i]

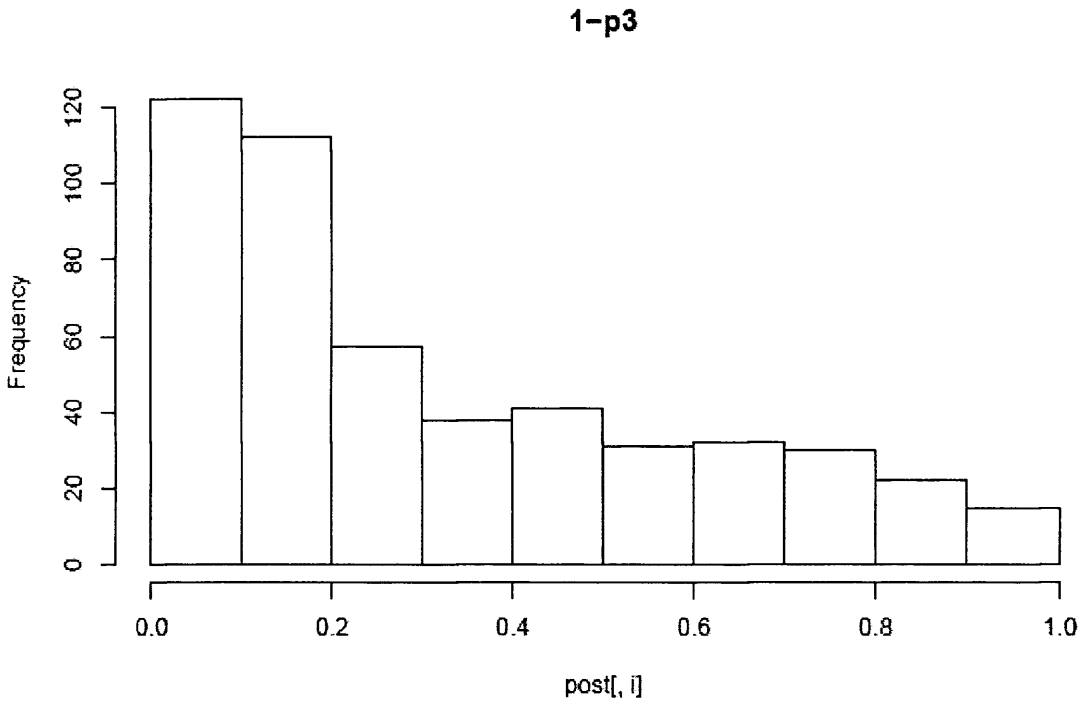
Dataset 5.

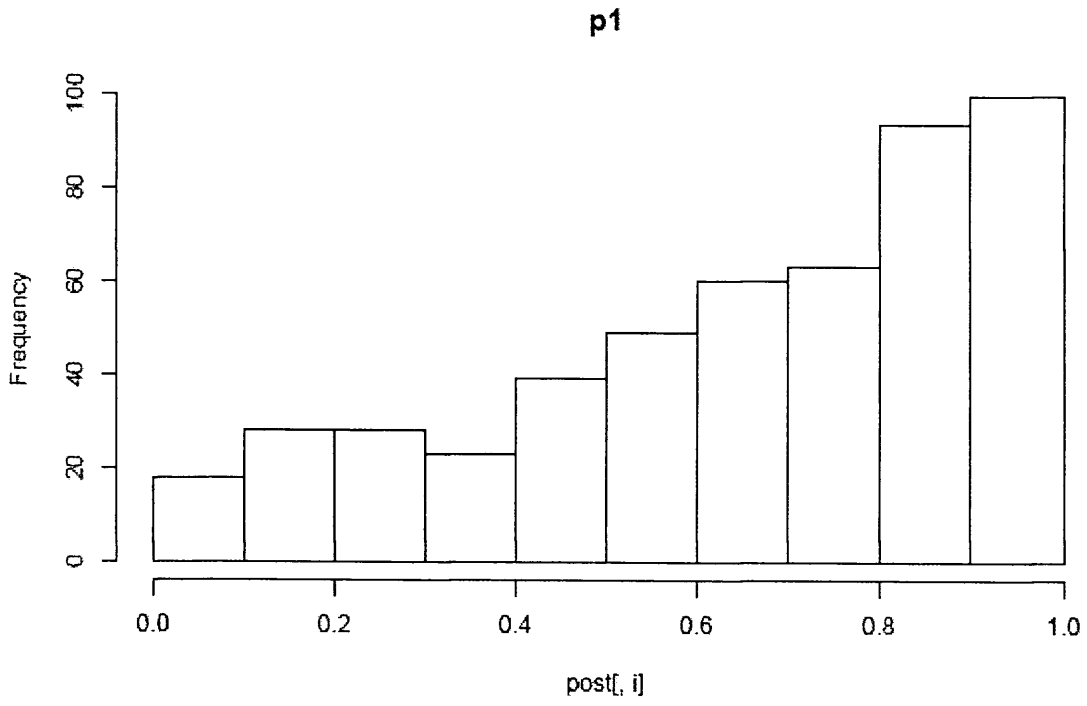

1-p3
Frequency
0
50
100
150
0.2
0.4
0.6
0.8
1.0
post[, i]


p1
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
post[, i]

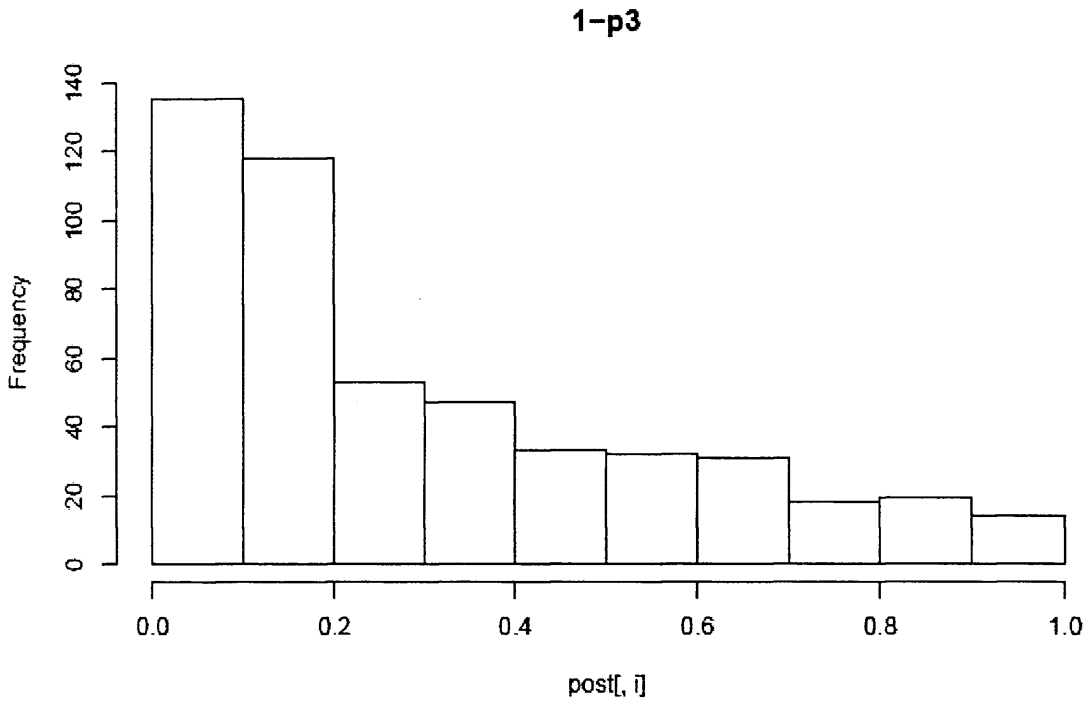
Appendix 5.5 Mean regression histograms for p1 and 1-p3 for two admixture events using 500,000 simulations.

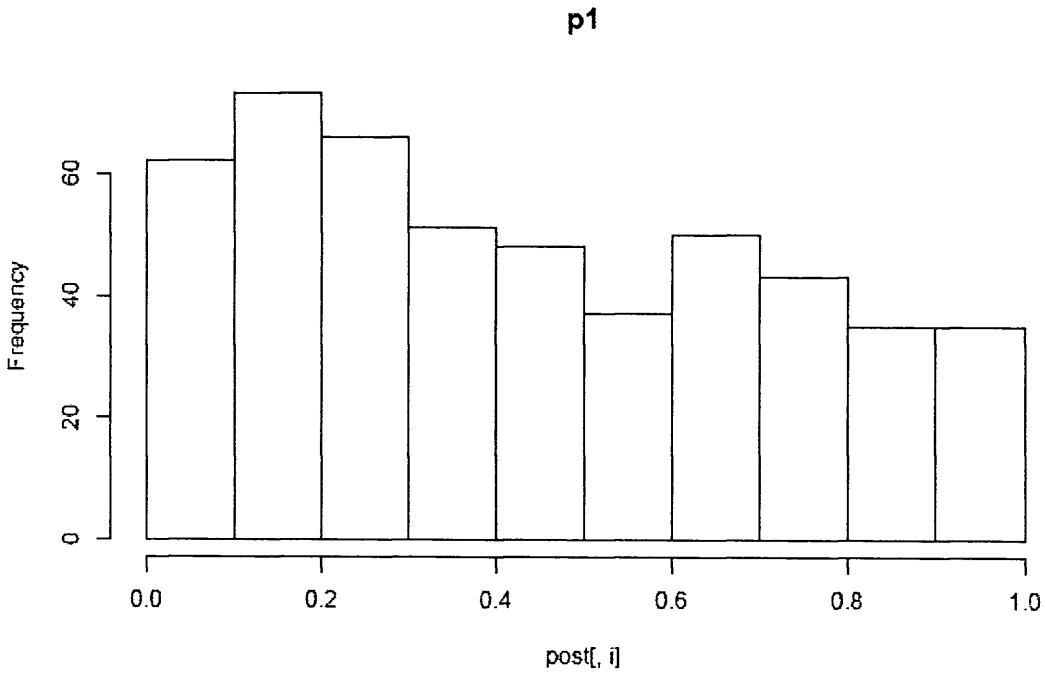
Dataset 1.



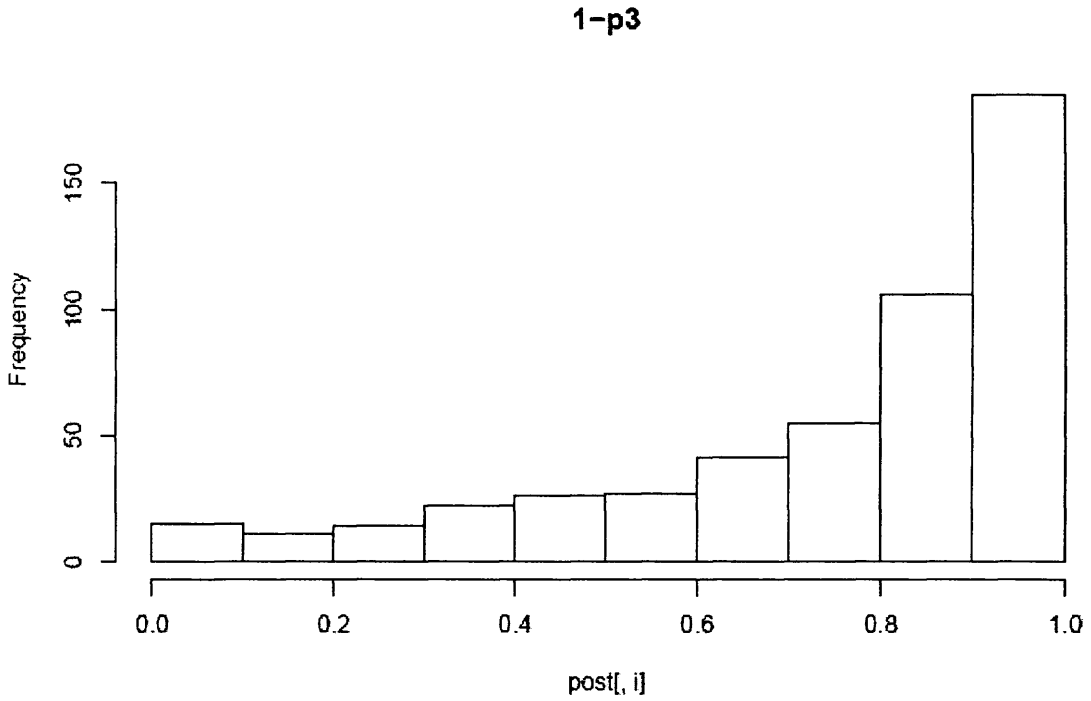


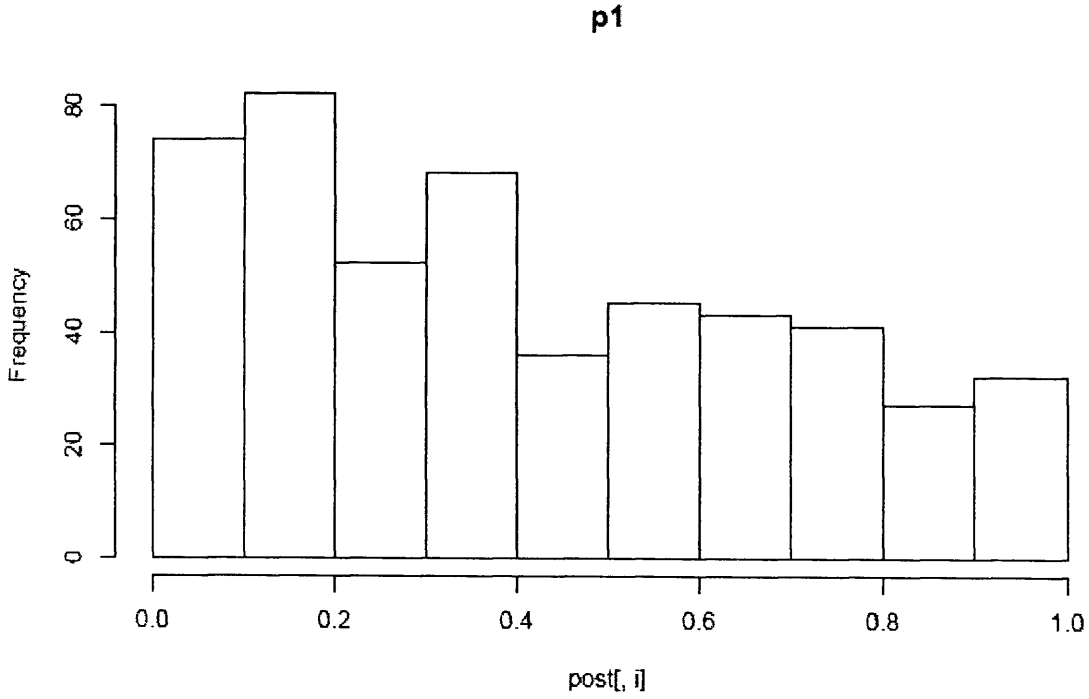
Dataset 2.


1−p3
Frequency
0
20
40
60
80
100
120
140
0.0
0.2
0.4
0.6
0.8
1.0
post[, i]

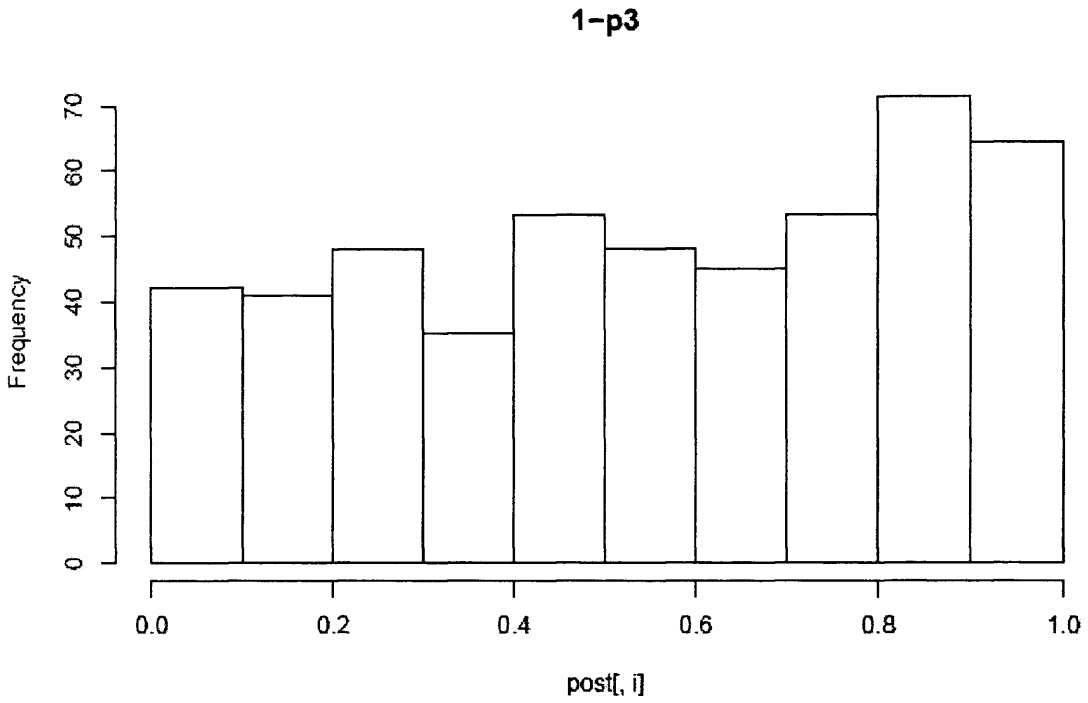

p1
Frequency
0
20
40
60
0.0
0.2
0.4
0.6
0.8
1.0
post[, i]

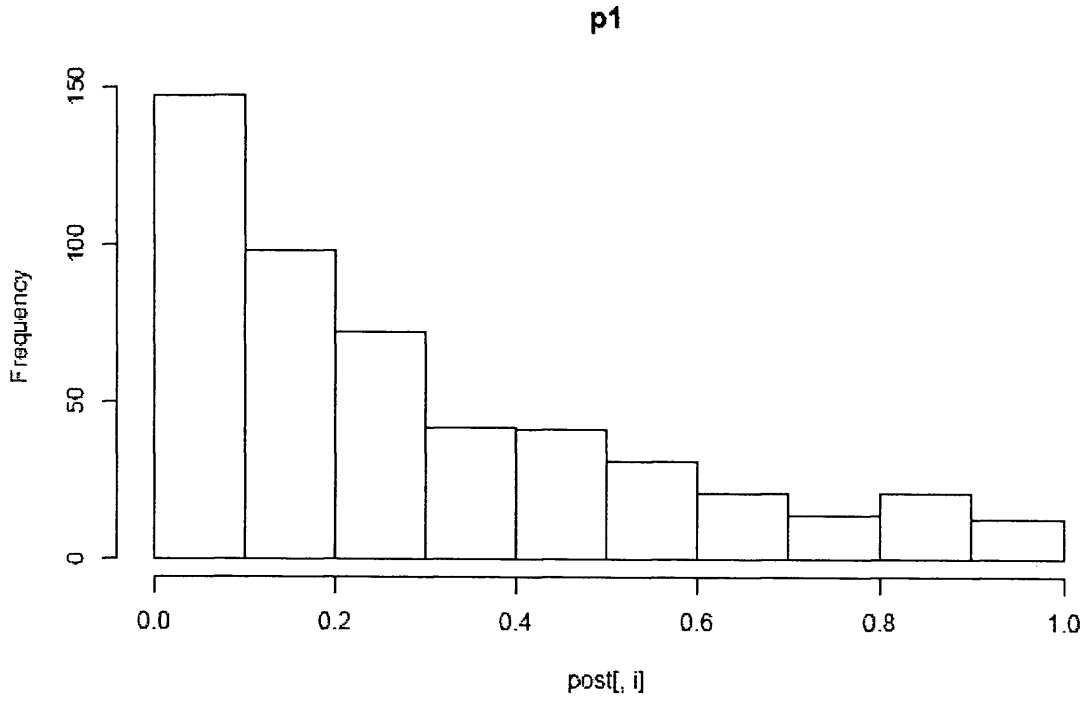
Dataset 3.


1-p3
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
post[, i]


p1
Frequency
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
post[, i]

Dataset 4.


1-p3
Frequency
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
post[, i]


p1
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
post[, i]

Dataset 5.