THE UNIVERSITY *of* EDINBURGH

Genetic diversity and structure of livestock breeds

Samantha Wilkinson

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

University of Edinburgh

2011

TABLE OF CONTENTS

DECLARATION

I declare that this thesis is my own composition and that the research described in it was carried out by me. Specific contributions of others are acknowledged.

Samantha Wilkinson

May 2012

ACKNOWLEDGEMENTS

Primarily, I would like to thank my supervisors Dr Pam Wiener, Dr Chris Haley and Dr Paul Hocking for their immense guidance, support and input over the past four years. I acknowledge the financial support of the Biological and Biotechnology Sciences Research Council (BBSRC) and the Rare Breeds Survival Trust (RBST) for the PhD studentship.

The experimental chapters would not have been possible without data available from previous research projects and the work and input from several people. The second and third chapter on the British pig breeds used genetic data that was a subset from an extensive European funded project, PigBioDiv, on pig breed biodiversity. I am grateful to Andy Law for maintenance of the PigBioDiv database. I am indebted to Dr Rex Walters and the British Pig Association for kindly providing individual inbreeding coefficients calculated from pig pedigrees for the British Saddleback individuals, which were then used in chapter three. The fourth chapter on the British chicken breeds used samples that were collected and genotyped with the help of RBST. I thank Dr Dawn Teverson for identifying suitable flocks and sending requests to breeders to supply hatching eggs; Graeme Robertson for collating, incubating and archiving the eggs and preparing embryos for DNA extraction; Karen Troup and David Morrice of ArkGenomics for DNA extraction and microsatellite genotyping and Andy Law for the creation of database to which the genotypes were uploaded. The fifth and sixth chapter would not have been possible without Dr Rob Ogden, who set up the project. I thank Dr Jerry Taylor, Dr Robert Schnabel, Hendrik-Jan Megens, Dr Alan Archibald, Dr Martin Groenen and Dr Rob Ogden for

the provision of genotype data. I thank David Morrice and Heather Finlayson of ArkGenomics for additional SNP genotyping. I am indebted to several people highly active in the field livestock breed conservation in the UK. I thank everyone at RBST for the support during the course of my PhD, in particular, Dr Dawn Teverson and Claire Berber. I'd also like to thank Vaughn Byrne for providing invaluable information on the British Saddleback breed bloodlines and Andrew Sheppy, of the Cobthorn Trust, for invaluable information on British chicken breeds. Last, but not least, I thank Lawrence Alderson for his unwavering support, advice and input over the course of my PhD.

I would also like to thank my friends and colleagues at Roslin for advice and support, including those of Room 85 NWE (I was greatly enlivened and would not have been so without the camaraderie).

LIST OF PUBLICATIONS

Wilkinson S., Haley CS., Alderson L. and Wiener P. (2011). An empirical assessment of individual-based population genetic statistical techniques: application to British pig breeds. Heredity **106**:261-269. (Based on **chapter 2**).

Wilkinson S., Wiener P., Archibald AL., Law A., Schnabel A., McKay SD., Taylor JF. and Ogden R. (2011). Evaluation of approaches for identifying population informative markers from high density SNP Chips. BMC Genetics **12**:45-59. (Based on **chapter 5**).

Wilkinson S., Wiener P., Haley CS., Teverson D. and Hocking PM. (2011). Characterisation of the genetic diversity, structure and admixture in British chicken breeds. Animal Genetics (accepted) (Based on **chapter 4**).

Wilkinson S., Archibald AL., Haley CS., Megens H-J., Crooijmans RPMA., Groenen MAM., Wiener P. and Ogden R. (2011). Development of a genetic tool for product regulation in the diverse British pig breed market. BMC Genomics (accepted). (Based on **chapter 6**).

ABSTRACT

This thesis addresses the genetic characterisation of livestock breeds, a key aspect of the long-term future breed preservation and, thus, of primary interest for animal breeders and management in the industry.

First, the genetic diversity and structure of breeds were investigated. The application of individual-based population genetic approaches at characterising genetic structure was assessed using the British pig breeds. All approaches, except for Principle Component Analysis (PCA), found that the breeds were distinct genetic populations. Bayesian genotypic clustering tools agreed that breeds had little individual genetic admixture. However, inconsistent results were observed between the Bayesian methods. Primarily, BAPS detected finer genetic differentiation than other approaches, producing biologically credible genetic populations. BAPS also detected substructure in the British Meishan, consistent with prior known population information. In contrast, STRUCTURE detected substructure in the British Saddleback breed that could not wholly be explained. Further analysis of the British Saddleback revealed that the genetic subdivision did not reflect its historical origin (union of Essex pig and Wessex Saddleback) but was associated with herds. The Rainbarrow appeared to be moderately differentiated from the other herds, and relatively lower allelic diversity and higher individual inbreeding, a possible result of certain breeding strategies.

The genetic structure and diversity of the British traditional chicken breeds was also characterised. The breeds were found to be highly distinctive populations with

moderately high levels of within-breed genetic diversity. However, majority of the breeds had an observed heterozygote deficit. Although individuals clustered to their origin for some of the breeds, genetic subdivision of individuals was observed in some breeds. For two breeds the inferred genetic subpopulations were associated with morphological varieties, but in others they were associated with flock supplier. As with the British Saddleback breed, gene flow between flocks within the chicken breeds should be enhanced to maintain current levels of genetic diversity.

Second, the thesis focused on breed identification through the assignment of individuals to breed origin. Dense genome-wide assays provide an opportunity to develop tailor-made panels for food authentication, especially for verifying traditional breed-labelled products. In European cattle breeds, the prior selection of informative markers produced higher correct individual identification than panels of randomly selected markers. Selecting breed informative markers was more powerful using delta (allele frequency difference) and Wright's F_{ST} (allele frequency variation), than PCA. However, no further gain in power of assignment was achieved by sampling in excess of 200 markers. The power of assignment and number of markers required was dependent on the levels of breed genetic distinctiveness. Use of dense genome-wide assays and marker selection was further assessed in the British pig breeds. With delta, it was found that 96 informative SNP markers were sufficient for breed differentiation, with the exception of Landrace and Welsh pair. Assignment of individuals to breed origin was high and few individuals were falsely assigned, especially for the traditional breeds. The probability that a sample of a presumed origin actually originated from that breed was high in the traditional breeds.

Validation of the 96-SNP panel using independent test samples of known origin and market samples revealed a high level of breed label conformity.

CHAPTER ONE

General Introduction

1.1 Introduction

Since domestication, livestock breeds have been purposefully bred for desirable production traits that were advantageous for human society. Combined with additional evolutionary and demographic processes, such as genetic drift, founder effects, population contraction, mutation and migration, this has resulted in an enormous array of breeds and the formation of well-defined phenotypes rarely observed in mammalian species. At last count, the Food and Agricultural Organisation recorded an extraordinary 7 616 livestock breeds worldwide (FAO 2007).

Industrial consolidation of agriculture in more recent decades has driven many breeds to extinction because most commercial livestock populations have been developed from a limited number of existing breeds (FAO 2007). As a result, livestock biodiversity is threatened by the marginalisation of the less commercially important breeds. Many have argued that further extinction should be halted because livestock breeds have a large amount of accompanying phenotypic diversity which could be a valuable genetic reservoir for the agricultural industry (Ajmone-Marsan and Consortium 2010; DEFRA 2006; Hall and Bradley 1995; Smith 1984), especially when considering genetic erosion in commercial populations due to intensive selection (Muir et al. 2008).

Although many traditional livestock breeds are low in population size and may be on the verge of extinction, their long divergent evolutionary histories make them important genetic resources. Traditional breeds that are not currently of commercial

interest may have future value because of unique traits related to disease resistance, meat quality and behavioural or physiological characteristics. However, the practical use of conserved breeds in the livestock commercial industry is still questionable. Since all the variation could be selected from within a single breed, finding and using variation (e.g. disease resistance) from across multiple breeds could be a more difficult, long and expensive process than selection from within one breed (Hill 2000; Hill and Zhang 2004). Even so, there is a social, historical and cultural argument for the preservation of livestock breed diversity as it embodies the heritage of the agricultural revolution (DEFRA 2006; Hill 2000). In addition, there is an increasing awareness of the other beneficial aspects of traditional breeds. Many traditional breeds are well adapted to harsh environments which makes them suitable for grazing conservation habitats to facilitate site management (DEFRA 2006; Small 2004). Traditional breeds also tend to possess different meat quality characteristics to commercial breeds, and their produce is now becoming more available in the market place (BPA 2002). Genetic diversity and, thus, viability of livestock breeds is a necessary aspect for long-term perpetuation. The molecular characterisation of the genetic diversity and structure of livestock will help inform management and preservation of livestock breeds.

This thesis focuses on the genetic characterisation of livestock breeds. First, the exploration and description of population genetic structure and diversity using both traditional and more modern analytical approaches is pursued. Second, assignment of individuals to breed origin is examined in light of breed composition.

Below, this chapter first briefly gives an account of the genetic markers used in livestock biodiversity studies. Then, the thesis rationale is explained through an account of pertinent literature findings on the use population genetic methods for the characterisation of livestock diversity and the genetic patterns of livestock breeds that have already been discerned. The use of genetic information from individuals for the ascertainment of population membership is also discussed, and more practical applications, such as food authentication, are also highlighted. This chapter concludes by outlining the specific research aims for the subsequent chapters in this thesis.

1.2 Genetic markers

Since the development of DNA technology various genetic markers have become available for livestock biodiversity studies. Earlier molecular characterisations of livestock breeds used polymorphic protein loci such as blood group types (cattle (Blott et al. 1998a); horses (Bowling 1994); sheep (Clarke et al. 1989); goat (Tunon et al. 1989); pigs (Van Zeveren et al. 1990a; Van Zeveren et al. 1990b)). With further advancements other markers have been developed including polymorphic fragment size DNA markers, which are genotyped portions of the genomic DNA where the size or length of the portion measures the allele. These markers have also been utilised in livestock diversity studies, such as restriction fragment length polymorphisms and randomly amplified polymorphic DNA in sheep (Kunene et al. 2009) and amplified fragment length polymorphisms in goats (AFLP) (Ajmone-

Marsan et al. 2001), cattle (Ajmone-Marsan et al. 2002) and pigs (SanCristobal et al. 2006b).

The most frequently used fragment sized marker in population genetic studies are microsatellites. These are repetitive portions of the genomic DNA comprised of one to six DNA base pairs repeated from 5 up to 40 times (Selkoe and Toonen 2006; Sunnucks 2000). Microsatellites possess a number of desirable characteristics over the above mentioned genetic markers. First, they are abundant in the genome and a high mutation rate makes them highly polymorphic (Bruford et al. 2003). Also, they can be easily genotyped even when there is DNA degradation of a sample (Selkoe and Toonen 2006). Thousands of microsatellites have been found throughout livestock genomes and species-specific panels of microsatellites have been recommended by the Food and Agriculture Organisation and used, for example, in sheep, chicken and pig diversity studies (Glowatzki-Mullis et al. 2009; Granevitze et al. 2007; SanCristobal et al. 2006a). The characterisation of European pig breeds diversity using microsatellites (SanCristobal et al. 2006a) produced a similar clustering of pig populations to that using AFLP markers, but with a higher resolution (SanCristobal et al. 2006b). The authors concluded that microsatellites are the more preferable genetic marker due to the bi-allelic and dominant nature of AFLPs.

With advances in genome-sequencing technologies Single Nucleotide Polymorphisms (SNP) markers, which are DNA sequence variants, are being discovered in abundance in genomes (Eck et al. 2009; Lindblad-Toh et al. 2005;

Wong et al. 2004). Though not as variable as microsatellite loci, there are a number of biological advantages of SNPs over microsatellites. First, the mutation model of SNPs is simple in comparison to that of microsatellites (Ellegren 2004). Second, microsatellites can suffer from homoplasy (parallel evolution where alleles are of the same size but from different lineages, such that the two alleles are identical by state but not identical by descent) (Selkoe and Toonen 2006), whereas homoplasy is virtually absent in SNPs minimising its potential effects on estimated levels of genetic diversity and divergence. There are also technological advantages for the use of SNPs such as robust methods of discovery and straightforward automation of SNP genotyping through assay design (Kim and Misra 2007; Morin et al. 2004). SNP data are also easily comparable between laboratories, circumventing the technical difficulties associated with developing large datasets from fragment sized DNA markers across laboratories. Dense genome-wide SNP assays have been developed available for many livestock species, enabling the rapid automated large-scale production of genomic data (Kijas et al. 2009; Matukumalli et al. 2009; Ramos et al. 2009; Van Tassell et al. 2008; vonHoldt et al. 2010). The new assays are highly informative resources; the SNP chips have already been used to investigate breed genetic structure in cattle, sheep and dogs (Decker et al. 2009; Gautier et al. 2010; Gibbs et al. 2009; Kijas et al. 2009; vonHoldt et al. 2010).

Both microsatellites and SNPs continue to be employed in livestock biodiversity studies. As will be described in more detail in the following sections of this chapter, the use of genetic markers in livestock biodiversity studies has revealed the

complexity surrounding animal domestication and breed development, and certain broad genetic patterns across livestock species have been discerned.

1.3 The genetic effects of breed development

1.3.1 Genetic diversity within breeds

One of the primary applications of molecular markers in the context of livestock biodiversity studies is to quantify the levels of genetic diversity within breeds. In general, studies have found that livestock breeds are genetically diverse populations (goat, (Canon et al. 2006); horse (Glowatzki-Mullis et al. 2006); sheep (Lawson Handley et al. 2007); cat (Menotti-Raymond et al. 2008); pig (SanCristobal et al. 2006a); cattle (Wiener et al. 2004)). This could be due to the development of a broad genetic base during domestication (Andersson 2001). First, recent studies have shown that domestication has been a complex and recurrent phenomenon (Bruford et al. 2003) debunking an earlier suggestion that domestication was a single event, which resulted in a strong bottleneck (Clutton-Brock 1999). Through mitochondrial sequencing and wide geographic sampling the presence of multiple maternal lineages with moderate geographical partitioning in most livestock species has emerged, indicating that multiple domestication events of divergent populations have occurred (pig (Larson et al. 2005); chicken (Liu et al. 2006); goat (Luikart et al. 2001); sheep (Meadows et al. 2007)). Second, different subspecies may have also contributed to the founding populations of domesticated animals. For example, genetic evidence suggests that the *Bos taurus* (humpless cattle, originating in the Middle East) and the *Bos indicus* (humped cattle, originating in the Indian subcontinent) were derived from two distinct subspecies of the wild progenitor (Loftus et al. 1994; MacHugh et

al. 1997). Similarly, European and Asian pig breeds appear to have been independently domesticated from two different subspecies of wild boar (Giuffra et al. 2000). Large genetic diversity in livestock breeds could be due to a broad ancestral genetic base, the consequence of multiple domestication events and the contribution of more than one divergent populations and/or (sub)species (Andersson 2001).

Following on from the original domestication events, farm animal breeds probably then adapted to their local environments (FAO 2007). The commencement of artificial selection in breed development, pioneered by the likes of Robert Bakewell in the 18th century, dramatically altered the genetic landscape of livestock breeds and the agricultural industry (DEFRA 2006). Many livestock breeds were upgraded by the introduction of favourable genetic material from another breed. The native British pig breeds had alleles introgessed from Asian pigs to impart favourable traits such as earlier maturation and increased prolificacy (Darwin 1868). Since the Asian and European breeds come from disparate origins (Giuffra et al. 2000) new Asian alleles could have influenced the genetic diversity of the British pig breeds. The broad genetic base in many livestock breeds was arguably maintained by human-mediated gene flow between distinctive populations. MacHugh et al (1997) stated that the high genetic diversity in Charolais and Friesian cattle could be attributed to genetic introgression from other breeds.

The system of selective animal breeding that started in the 18th century imparted rapid genetic change in livestock breeds. This resulted in the development of distinctive breeds for a variety of purposes, as Moll noted the morphological

diversification and specialisation of European cattle breeds in the mid-19th century (Moll 1860). Darwin also observed changes in the British pig breeds, as a " consequence of so much crossing, some well-known breeds have undergone rapid changes; thus, according to Nathusius, the Berkshire breed of 1780 is quite different from that of 1810" (Darwin 1868). The importance of the many phenotypically diverse and distinct livestock breeds that were formed during the agricultural revolution was acknowledged by the creation of breed societies in the late 19th and early 20th century. By keeping the herdbooks closed the organisations were instrumental in preserving the genetic integrity of many breeds. Although the genetic introgression from one breed into another can be viewed as eroding or contaminating the genetic integrity of the affected breed, gene flow is an important process as it introduces new genetic material. Therefore, in the absence of gene flow the genetic diversity within a population could be considerably narrowed. If a population is isolated other contributing factors, such as small population size, severity of artificial selection and time since isolation, can also adversely affect genetic diversity within populations. These demographic and genetic processes have occurred in certain domesticated populations and, as a consequence, there are exceptions to the pattern of high genetic diversity. Granevitze et al (2007) reported extremely low genetic diversity in German native fancy chicken breeds. It was presumed that this was due to positive assortative mating and small population sizes. A more extreme case of reduced genetic diversity in livestock breeds is the feral herd of Chillingham cattle (Visscher et al. 2001). A combination of 300-year isolation and a severe population reduction, which increased the chance probability of random genetic drift driving alleles to fixation, has created a herd extremely lacking in genetic diversity.

Though most livestock breeds have a high level of genetic diversity, with the exception of isolated (small) populations, there is a subtle pattern of genetic diversity across geographic clines. Following on from the original domestication events, the generally held view is that when humans migrated in the past they took a small sample of their diverse original animal stock. This would be a subset of the genetic diversity of the original stock and would then represent the only genetic diversity for newly founded populations (Bruford et al. 2003). Consequently, in the absence of genetic introgression, there should be a negative correlation such that genetic diversity of livestock breeds decreases with increasing geographic distance from the centre of domestication. Estimates of genetic diversity confirm that there is a higher degree of genetic variation present within breeds from or near the centre of domestication for certain livestock species (goats (Canon et al. 2006); cattle (Loftus et al. 1999); sheep (Peter et al. 2007)).

1.3.2 Population structure of breeds

Another important aspect of livestock biodiversity is the level of genetic variation amongst breeds. Variation in allele frequencies between populations can be used to measure the degree of genetic differentiation (known as F_{ST}). The quantification of F_{ST}, with a range of little ($0.00 < F_{ST} < 0.05$), to moderate ($0.05 < F_{ST} < 0.15$), to great ($0.15 < F_{ST} < 0.25$) to very great ($F_{ST} > 0.25$) (Hartl and Clark 1997), can inform on the distinctiveness of populations and whether populations contain distinctive multilocus combinations that render them genetically unique and the degree of genetic relatedness between breeds.

Livestock biodiversity studies concur that there is marked levels of genetic variation between livestock breeds, though the degree of differentiation is variable amongst livestock species. For instance, cattle, goat, horse and sheep breeds tend to exhibit moderate levels of genetic differentiation (Canon et al. 2006; Druml et al. 2007; Lawson Handley et al. 2007; MacHugh et al. 1998). Historically, domesticated animals were easily and widely transported which would allow for substantial breed intermingling (Clutton-Brock 1999) and with that a reduced breed genetic differentiation. Luikart et al (2001) found weak genetic structure of goat breeds (relative to other livestock breeds) and suggested that far greater transportation occurred in goats than in other livestock breeds. Neighbouring breeds also tend to experience enhanced gene flow, such as the closely related modern Baltic sheep breeds (Tapio et al. 2005a), resulting in relatively low breed genetic differentiation. In contrast, strict breeding practices were likely in place to isolate and preserve phenotypically distinct sets of traditional and commercially chicken and pig breeds, resulting in high levels of breed genetic differentiation (Bodzsar et al. 2009; SanCristobal et al. 2006a). An exception is the dog where extremely high levels of differentiation amongst breeds have been observed, possibly due to even stricter enforced breeding practices for favourable morphological traits, thus creating breed barriers and a lack of gene flow between breeds (Parker et al. 2004).

Another quantification of the amount of genetic variation between breeds is to estimate the genetic distance using population allele frequencies. Estimated genetic distances are then generally visualised as a phylogenetic tree which can help unravel the evolutionary history of livestock breeds. This is a commonly adopted approach to

characterise population structure of livestock breeds as it provides a useful illustration of the genetic relationships amongst breeds. Phylogenetic reconstruction has generally resulted in long breed-branch lengths indicating high levels of genetic distinction (cattle ((Li et al. 2007; Maudet et al. 2002); pigs (Megens et al. 2008; SanCristobal et al. 2006a); goats (Peter et al. 2007)). Another common observation is that where groups of breeds are identified in the evolutionary tree these tend to correspond to geographic origin (Li et al. 2007; Megens et al. 2008; Peter et al. 2007). The genetic similarities between pairs of breeds can be also uncovered and these tend to reflect common ancestry, but could also be due to past cross-breeding.

The description of the elucidated genetic patterns of variation between livestock breeds given above was derived from results using traditional methods, like F-statistics and genetic distances, which are based on population allele frequencies. Recently, a new set of population genetic approaches have been developed to infer population structure which instead use individual multilocus genotypes. These methods have been developed in a Bayesian statistical framework and aim to both partition a sample of individual genotypes into an unknown number of genetically distinct populations and to determine the proportion of an individual's genome that originates from different inferred populations (Beaumont and Rannala 2004). In brief, with specific prior probabilities, the posterior probability that an individual arises from a subpopulation can be estimated given the likelihood of the multilocus genotype. The Bayesian genotypic clustering methods are valuable and powerful tools in the elucidation of population structure and the practical applications has proven to be broad (Beaumont and Rannala 2004). These novel tools allow otherwise

neglected questions to be addressed. For instance, does a phenotypically defined breed necessarily equate to a genetic population? In other words, the genetic composition of a breed can be investigated to determine if it is broader, narrower or equivalent to that defined by its phenotypic criteria. Using Bayesian genotypic clustering approaches, studies on small sets of breeds usually sampled from a particular country have found that breeds are generally genetically differentiated populations (e.g., Swiss sheep (Glowatzki-Mullis et al. 2009); Estonian cattle (Li et al. 2011); Italian chickens (Zanetti et al. 2010)). Thus, by definition, in the absence of extensive gene flow that could create a homogenous genetic pool, most livestock breeds equate to a genetic population. Bayesian genotypic clustering tools can also be used to reconstruct hierarchical genetic structure to identify groups of related livestock breeds. By sampling many breeds across an extensive geographic area, the main genetic subdivision of breed structure into groups of closely related breeds appears to correspond to geographic divisions (goats (Canon et al. 2006); sheep (Lawson Handley et al. 2007); cats (Menotti-Raymond et al. 2008); dogs (Parker et al. 2004)). Finally, the Bayesian genotypic clustering approaches can infer genetic patterns in individuals that are the result of hybridisation or genetic introgression. This of importance in livestock biodiversity studies as it could expose cross-breeding practices, particularly if breed societies are striving to maintain breed integrity by limiting genetic introgression and hybridisation. For example, Lawson-Handley et al (2007) identified admixed individuals in Greek sheep breeds despite the attempts of breeders to maintain breed separation.

The application of Bayesian genotypic clustering approaches to livestock biodiversity has provided previously unattainable inferences on patterns present in a genetic datasets. However, sifting through the pertinent literature, it has become apparent that certain challenges have been encountered when applying these novel methods to empirical data. Peter et al (2007) found that several sheep breeds were assigned to more than one cluster and that clusters were only partially hierarchical. Another study on cattle breeds reported that multiple clustering solutions were depicted (Li et al. 2007). In some cases identifying the number of underlying populations can be difficult, as found in a study on cat breeds (Menotti-Raymond et al. 2008).

1.3.3 Identification of the origin of individuals

When populations are genetically distinct it indicates that the individuals originating from a given population are genetically different from individuals belonging to another population. If the genetic patterns of individuals from different populations are sufficiently disparate, this genetic information can be used to ascertain the population origin of individuals. Paetkau et al (1995) recognised the need to determine whether a population is the genetic source of a given individual genotype. The individual assignment test was developed to determine, given the observed allele frequencies in a set of potential source populations, the probability of an individual genotype arising in one or more populations (Paetkau et al. 1995).

In population genetic studies, the assignment of individuals to their origin is another useful analysis that is increasingly being incorporated (Davies et al. 1999; Waser and Strobeck 1998). The identification or verification of the origin of individuals can complement the traditional population genetic approaches with regards to establishing population structure. Livestock biodiversity studies have generally found that, due to the genetic distinctiveness of most livestock breeds, individuals can be assigned to their breed of origin with a high confidence using polymorphic genetic markers (e.g. cattle (Blott et al. 1999; Ciampolini et al. 2006; Negrini et al. 2009); horse (Glowatzki-Mullis et al. 2006); dog, (Koskinen 2003); pig (Ramos et al. 2011)). As with Bayesian genotypic clustering approaches, individual assignment methods simply confirm that genetic differentiation between populations (measured using F_{ST} or genetic distances) is present (Manel et al. 2005). Nonetheless, it highlights the potential for individual assignment methods to be used for more practical purposes.

With sufficient population genetic heterogeneity, genetic markers can be used to identify or verify the claimed origin of a biological sample. For example, microsatellite markers were used to determine if misconduct had occurred involving a prize winning salmon in a fishing competition. An individual assignment test indicated that there was an extremely low probability that the winning salmon could have arisen from the population found at the competition location. Based on the genetic results, it was inferred that the suspect fish was not caught at the competition location but, in fact, had originated from a local food market (Primmer et al. 2000). In the food industry knowingly substituting the biological name of a product with

another (be it species, breed, variety and/or geographic origin) for financial incentives is considered a widespread activity because certain names attract a premium value (Primrose et al. 2010; Teletchea et al. 2005; Woolfe and Primrose 2004). Products derived from traditional livestock breeds tend to be more expensive due to higher production costs, the value of rarity and different meat quality such as the highly priced Iberian pig (Garcia et al. 2006). Consequently, there is financial profit to be gained having a product derived from a commercial breed labelled under a traditional breed name. The identification or verification of the origin of breed-labelled products using genetic markers could not only address consumer confidence, but also protect the livelihoods of breeders, particularly those who keep traditional livestock breeds.

1.4 Aims and objectives

The aim of this thesis is to characterise the genetic diversity, structure, extent of genetic admixture and genetic composition of individuals in a number of livestock breeds. These analyses contribute to current work on breed characterisation in the interest of preservation of genetic diversity. In conjunction with the above, this thesis also explores the use of genetic markers for assignment of individuals to breed origin. Not only is the inference of ancestry of individuals indicative of breed integrity and distinctiveness, but an accompanying aspect is that, in the interests of more practical applications, genetic markers can be used to verify the claimed origin biological samples.

The subsequent paragraphs give a content outline and study objectives for each chapter:

Chapter 2 explores the efficiency of several individual-based population genetic statistical approaches at characterising population structure using British pig breed individual multilocus genotypes. Bayesian genotypic clustering approaches are tools that are now routinely used to describe population structure (e.g. sheep, Handley-Lawson et al., 2007). However, it is only recently that the performance and merits of these novel and popular techniques are being evaluated (e.g., Safner et al., 2011). Three Bayesian genotypic clustering approaches were therefore compared alongside individual-based phylogenetic reconstruction and Principle Component Analysis (PCA) in the characterisation of the population structure of British pig breeds.

Chapter 3 details further exploration of the genetic structure of the British Saddleback pig breed, the only British traditional pig breed that did not form a genetically distinctive and cohesive unit using certain individual-based clustering techniques (chapter 2). The British Saddleback is a relatively new British pig breed, having formed from the union of Essex pig and Wessex Saddleback breeds in 1967. The pattern of genetic substructure in the British Saddleback breeds was explored using individual-based clustering methods and independent information on individual inbreeding coefficients.

Chapter 4 is devoted to the characterisation of the current genetic state of British traditional chicken breeds. In the interests of preserving and conserving regional

diversity the genetic characterisation of livestock populations is recommended. Of the numerous livestock breeds present in Britain, there is a dearth of work on poultry species. To contribute to on-going livestock biodiversity efforts, the genetic diversity, structure and extent of admixture in the British traditional chicken breeds were characterised. Recommendations concerning the preservation of genetic diversity are proposed.

Chapter 5 explores how high density assays featuring Single Nucleotide Polymorphism (SNP) markers can be exploited to create reduced panels of informative markers for population genetic analyses. Dense genome-wide data is valuable but can be relatively costly to produce and time-consuming and computationally expensive to analyse; it is therefore often desirable to reduce the number of markers by screening and selecting according to their information content to create reduced panels for population genetic analyses. For the verification of the origin of individuals in European cattle breeds, several population genetic differentiation methods were used to determine the most appropriate selection methods to identify informative markers from the BovineSNP50 beadchip.

Chapter 6 describes the development of a multiplex-assay using markers selected from a dense SNP assay for pork authentication in the British food industry. Once DNA markers that contain high genetic information have been identified, the use of such markers extends from population demarcation to more practical applications of the verification of the origin of food products. A panel of informative markers was

developed from the PorcineSNP60 beadchip and subsequently validated using further test samples.

Chapter 7 presents an overall summary and conclusions of this thesis. An overall perspective on the findings and their relevance to the field is given. Further studies that would build on the work covered by this thesis are proposed.

CHAPTER TWO

An empirical assessment of individual-based population genetic statistical approaches: application to British pig breeds

2.1 Introduction

Traditional population genetic statistics, such as expected heterozygosity, Wright's F-statistics and genetic distances (Hartl and Clark 1997), are routinely used to describe the genetic diversity and structure of livestock breeds. However, analysis at the level of the population or breed and subsequent results may prove misleading and, increasingly, this approach has received scrutiny and criticism (Mank and Avise 2004; Pearse and Crandall 2004). First, natural populations are usually defined by geographical sampling distribution and livestock breeds by registered phenotypes. The delineation of populations might impose a subjective pre-existing structure that may not reflect the genetic reality and, thus, the *a priori* assignment of individuals to pre-defined populations might introduce unintended bias. Second, population genetic statistical estimates are summarised by averaging across individuals to produce a single number for each population. Biological processes and patterns, such as cryptic population structure and the occurrence of gene flow, would remain undetected using *a priori* defined populations and traditional population genetics (Mank and Avise 2004), leading to possible inaccurate representations of genetic diversity and structure.

These complexities and challenges can be addressed by analysing population genetic data at the level of the individual. Recent methodological advancements now conveniently allow the inference of population structure directly from individual genetic polymorphism data, instead of relying on *a priori* population information. Several individual-based genotypic clustering models have been developed in a

Bayesian statistical framework and are available in special purpose software packages (Excoffier and Heckel 2006). The methods operate by creating clusters in which the assumptions of Hardy-Weinberg and linkage equilibrium are met, and simultaneously each individual is assigned to a cluster based on a probabilistic model. Each method has slightly different underlying assumptions and different methods of searching the parameter space. The range of applications and use of Bayesian genotypic clustering approaches is broad, for example, depicting genetic relationships between populations, population structure and presence of genetic admixture (Beaumont and Rannala 2004; Mank and Avise 2004; Pearse and Crandall 2004; Rosenberg et al. 2001). These novel tools have become very popular and are now routinely used in empirical studies on both natural and livestock populations to the extent that it has been suggested that individual-based clustering techniques should be a pre-requisite part of population genetic studies (Luikart et al. 2003). However, the reliability, performance and limitations of the novel Bayesian genotypic clustering techniques are only now being tested (Ball et al. 2010; Frantz and Cellina 2009; Kalinowski 2011; Rowe and Beebee 2007; Safner et al. 2011).

In addition to Bayesian genotypic clustering approaches, two other individual-based methods are also available, principal component analysis (PCA) (Menozzi et al. 1978) and phylogenetic reconstruction (Bowcock et al. 1994). Both techniques have been used for several decades to study genetic structure and diversity, but are more often used on population-averaged data where populations are defined *a priori*. It may be preferable to adopt the individual-based approach for PCA and phylogenetic reconstruction. First, as mentioned above, within population genetic variation is

ignored as estimates are averaged across individuals. Second, the principles of phylogenetic reconstruction are not upheld when applied to admixed populations, because a key assumption is that there is no genetic exchange between populations (Toro and Caballero 2005). Like Bayesian genotypic clustering, individual-based PCA and phylogenetic reconstruction make no assumptions about the number or identity of separate populations from which individuals are drawn.

The individual-based population genetic approaches described above are potentially useful for the elucidation of livestock breed diversity and structure. However, it is still not clear how the methods differ in their power and appropriateness for particular data and questions. In addition, individual-based PCA and phylogenetic reconstruction approaches are rarely used in conjunction with Bayesian genotypic clustering methods. The objective of this study was to empirically assess various individual-based population genetic statistical methods for inference of genetic diversity and structure of livestock breeds. Using microsatellite data from British pig breeds, the genetic structure was inferred with various 'individual-based' approaches: three Bayesian genotypic clustering techniques, PCA and phylogenetic reconstruction. The applicability, efficacy and complementarity of the chosen methods were considered.

2.2 Materials and Methods

2.2.1 Data

The genotypic data of British pig breeds used in this study were a subset from an extensive European pig breed biodiversity study (PigBioDiv) (Russell et al. 2003).

The microsatellites recommended for the European pig biodiversity had high polymorphism, good genotyping performance and were well spaced across the genome (SanCristobal et al. 2006a). Forty-six microsatellites were selected for this study as they were genotyped across all of the selected populations. The proportion of missing data was 6.8%.

To determine if markers were in Hardy-Weinberg equilibrium (HWE), exact tests, which calculate the probability that an observed sample could be drawn from the population by chance, were performed. The probabilities of all possible genotypic frequencies were calculated for the observed allele frequencies. These samples were then ranked in ascending order based on their probabilities. The probability of the observed sample was obtained by summing all probabilities up to and including that of the observed sample. HWE was rejected if the total probability was less than the significance level (Weir 1996). To determine if there was a non-random association of alleles, pairs of loci were tested for genotypic linkage disequilibrium (LD) using exact tests. Since no assumptions were made about the gametic phase of heterozygotes (at a diploid locus), association between diploid genotypes was examined with a null hypothesis that genotypes at one locus were independent from genotypes at another locus. As in the case for HWE, the probability of the observed array was conditional on the cumulative probabilities of all possible arrays (Weir 1996). Deviation from HWE within loci and the presence LD between pairs of loci were tested using GENEPOP version 4.0.7 (Rousset 2008). Significance levels of the tests were adjusted for multiple comparison following standard Bonferroni

corrections (Rice 1989). No markers showed consistent evidence of LD or deviations from HWE (results not shown).

In brief, there were a total of 18 populations (Table 2.1). Twelve British pig breeds were included with two breeds represented by more than one line. The term 'population' was used to represent commercial lines sampled within a breed.

Table 2.1 The British pig breeds. [1] determined by the number of breeding females: Endangered = 100 - 200, Vulnerable = 200 - 300, At Risk = 300 - 500 and Minority = 500 - 1000; taken from Rare Breeds Survival Trust website (RBST; http://www.rbst.org.uk/).

Population	Breed	Category	Status[1]	Sample Size
1	Berkshire	Traditional	Vulnerable	50
2	British Lop	Traditional	Endangered	35
3	British Saddleback	Traditional	At Risk	41
4	Duroc (PIC)	Commercial		50
5	Gloucester Old Spots	Traditional	Minority	53
6	Hampshire (PIC)	Commercial		50
7	British Landrace (PIC1)	Commercial		50
8	British Landrace (PIC2)	Commercial		50
9	British Landrace (PIC3)	Commercial		48
10	Large Black	Traditional	Vulnerable	52
11	Large White (PIC1)	Commercial		50
12	Large White (PIC2)	Commercial		50
13	Large White (PIC3)	Commercial		50
14	Middle White	Traditional	Endangered	38
15	Pietrain	Commercial		50
16	Tamworth	Traditional	Vulnerable	42
17	Asian Meishan FR	Asian		25
18	Asian Meishan GB	Asian		36

An Asian breed (Meishan) was chosen as an outgroup, composed of two populations: the first from France and the second from Great Britain (which also consisted of two

subpopulations, sampled from the Roslin Institute and from PIC, a UK-based pig breeding company). The number of individuals sampled per population ranged from 25 to 53, giving a total of 819 individual multilocus genotypes (Table 2.1). Additional information on the sampling and genotyping of the microsatellite markers can be found at SanCristobal el al (2006).

2.2.2 Bayesian genotypic clustering techniques

Two widely used Bayesian methods (STRUCTURE and BAPS) and a newer method (STRUCTURAMA) were applied. The clustering methods perform a Bayesian analysis, using the multilocus genotypes, to probabilistically assign individuals to clusters and infer the number of genetically distinguishable populations (K). The three methods assume that all markers are in Hardy–Weinberg Equilibrium and genotypic linkage equilibrium (LE). Both BAPS and STRUCTURE allow individuals to be of mixed ancestry, proportionally assigning an individual genome to clusters (estimated individual coefficients of ancestry, 'q'), but differ in their approaches to estimating admixture.

STRUCTURE uses a Monte Carlo Markov Chain (MCMC) method and estimates the natural logarithm of the probability (Pr) of the observed genotypic array (X), given a pre-defined number of clusters (parameter K) in the data set (ln Pr(X|K)) (Pritchard et al. 2000). In a Bayesian context the estimate of ln Pr(X|K) is a direct indicator of the posterior probability of having K clusters, given the observed genotypic array. The model with correlated allele frequencies was implemented,

assuming admixture. This model assumes that frequencies from different populations are likely to be similar due to either migration or shared ancestry. The Markov Chain was run for 1,000,000 iterations, after a burn-in of 500,000 iterations, for values of K from 1 to 20, with 5 replicates for each K value. From each MCMC chain, STRUCTURE simultaneously infers the posterior probability of K and membership probabilities (q) for each individual. Individuals may be assigned probabilistically to more than one cluster, reflecting admixture. To help identify the optimal K value, an ad hoc statistic, ΔK, was calculated (Evanno and Regnaut 2005). It is based on the second order rate of change of Pr(X|K) with respect to K, where the height of the estimated values indicate the strength of the population subdivision.

The most recent version of BAPS (v 5.2) uses a 'greedy stochastic optimisation algorithm' to directly estimate the most likely K and assign individuals to clusters (Corander et al. 2008). In BAPS the value of K can be either pre-defined to investigate the clustering solutions of populations with successive K values or, unlike with STRUCTURE, the value of K can be left un-defined so that the algorithm can search for the most likely K value. For each K value BAPS searches for the optimal partitions, stores them internally, and, after all K value have been processed, it merges the stored results according to log-likelihood values. Five independent replicate runs for every level of K from 1 to 20 were conducted. Unlike STRUCTURE which estimates population assignment and admixture simultaneously, with BAPS, estimating individual admixture is a two-tiered approach. BAPS first, determines the clustering solutions of populations by assigning individuals to populations based on allele frequency distributions and then the

admixture of genotypes are quantified by establishing the ancestral sources of alleles for each individual based on the determined population clusters using a simulation procedure. In BAPS, evidence for admixture was considered significant for individuals with p values < 0.05 (Corander and Marttinen 2006).

The third Bayesian clustering approach was performed using the program STRUCTURAMA (Huelsenbeck and Andolfatto 2007). The method implemented is similar to that in STRUCTURE except that STRUCTURAMA treats K as a random variable searching for the most likely K value. In addition, at the time of implementation, the STUCTURAMA model did not take into account admixture. A prior distribution is placed on K such that the data determines the most appropriate value. The number of clusters and the assignment of genotypes to those clusters were estimated simultaneously. A Markov chain of 100,000 iterations following a burn-in of 50,000 iterations was sufficient for convergence and production of consistent results. A partition was sampled from the Markov chain every 100 iterations and the mean partition, which minimises the squared distance to the sampled partitions, was calculated in order to make assignments. This process was independently replicated 5 times.

2.2.3 Multivariate analysis

The second approach was a multivariate principal component analysis (PCA) (Menozzi et al. 1978) performed using the statistical package R (Team 2011). PCA is a statistical technique that can be used to reduce the dimension of a multivariate

dataset. The original variables are linearly transformed by PCA into a set of underlying variables ("principal components"). Each new Principal Component (PC) has an associated eigenvalue that measures the respective amount of explained variance. The PCs that explain the most genetic variation are ranked based on their eigenvalue, such that most of the original variability in the original multi-dimensional dataset may be contained in a smaller number of variables. Objects can then be distributed along axes based on their allelic compositions. The data for individual genotypes were prepared by scoring a '0' if a particular allele was not present, a '1' if it was present and '2' if two copies were present in the homozygous state (MacHugh et al. 1998; Patterson et al. 2006) and an eigenvalue decomposition of the covariance matrix of the original multilocus dataset was conducted in R. A statistical test was conducted to determine the number of PCs to retain, Horn's parallel analysis. This method is a simulation procedure that takes the original dataset and produces simulated datasets of the same dimensions as the empirical data through sampling with replacement. The simulated datasets are then subjected to PCA. Due to the lack of structure in the simulated data the eigenvalues obtained for the successive PCs should be similar, such that there is no decrease in variance with increasing PC. In contrast, due to the structure present in the empirical dataset eigenvalues should decrease with increasing PC. Once the eigenvalue of a PC from the empirical dataset is equal to or less than the eigenvalue obtained from the simulated dataset for the same PC no further PCs of the empirical data should be considered. In other words, the components to retain from the PCA on the empirical data are those that account for more variance than the components derived from

random simulation. The parallel analysis was conducted in the R package paran (Dinno 2009).

2.2.4 Phylogenetic reconstruction

The final approach was an unrooted phylogenetic analysis implemented in MICROSAT (Minch et al. 1997). Pairwise individual genetic distances (shared allele distance, DSA) were estimated using the proportion of shared alleles (PSA) (Bowcock et al. 1994). At any locus, two individuals' genotypes share either 0, 1 or 2 alleles and with a large numbers of loci, the proportion of shared alleles becomes almost continuously distributed such that it can be used as an index of genetic similarity. A population tree was then constructed from the genetic distance matrix using the neighbour-joining method (Saitou and Nei 1987), which does not require that all lineages evolve at the same rate and is statistically consistent under many models of evolution (Weir 1996). The algorithm is based on the minimum-evolution criterion where at each successive step neighbour pairs that give the smallest sum of branch lengths are chosen, ultimately producing a single topology that minimises the total branch length. An unrooted neighbour-joining cladogram was constructed from the distance matrix for all pairs of individuals using the R package APE (Paradis et al. 2004).

To determine the robustness of the tree topology a bootstrap analysis was performed. Bootstrapping is a suggested method to assess the confidence intervals of reconstructed evolutionary clades (Felsenstein 1985). For each bootstrap replicate the

original matrix of individual genotypes was randomly sampled with replacement across the loci so that the resulting data set was the same size as the original. From each bootstrap replicate a matrix of pairwise genetic distances between the individuals was calculated and a neighbour-joining tree was then reconstructed. A consensus tree was constructed whereby the robustness of each branch was evaluated by determining the percentage of times that it occurred in all the bootstrap replicates. Bootstrapping of 1000 replicates from the original dataset was performed across loci in MICROSAT and the consensus cladogram was calculated using CONSENSE (Phylip v 3.67) (Felsenstein 2008).

2.2.5 Genetic differentiation

The degree of population genetic differentiation was estimated using F_{ST}. If populations are differentiated the genotype frequencies of the total population will exhibit a deficit of heterozygotes and excess of homozygotes relative to HWE. Thus, F_{ST} measures the reduction in heterozygosity among populations (i.e. increasing variation in allele frequencies among populations) relative to the heterozygosity of the total population (Hartl and Clark 1997). An extension of Wright's F_{ST} (Wright 1943; Wright 1951) is Weir and Cockerham's, which is the ratio of variance in allele frequencies among populations to the overall variance in allele frequencies (Weir and Cockerham 1984). The degree of genetic differentiation was estimated, for the defined pig populations listed in Table 2.1 and for certain inferred clusters for each Bayesian genotypic clustering approach (at the inferred optimal K value), with Weir

and Cockerham's unbiased estimator of Wright's fixation index (F_{ST}) using FSTAT 2.9.3 (Goudet 1995).

2.3 Results

2.3.1 Bayesian genotypic clustering

2.3.1.1 Number of populations (K) and clustering solutions

Results from the STRUCTURE analysis showed steadily increasing values of log likelihoods from K = 1 to 20 subpopulations (Fig 2.1). In Figure 2.1, the rate of change in the log likelihood with successive K values, ΔK, was plotted from $K_1 - K_2$ to $K_{19} - K_{20}$. The largest ΔK value, ΔK = 6, was at $K_3 - K_2$, followed by a second and third mode at $K_8 - K_7$ and $K_{15} - K_{14}$, respectively (Fig 2.1).

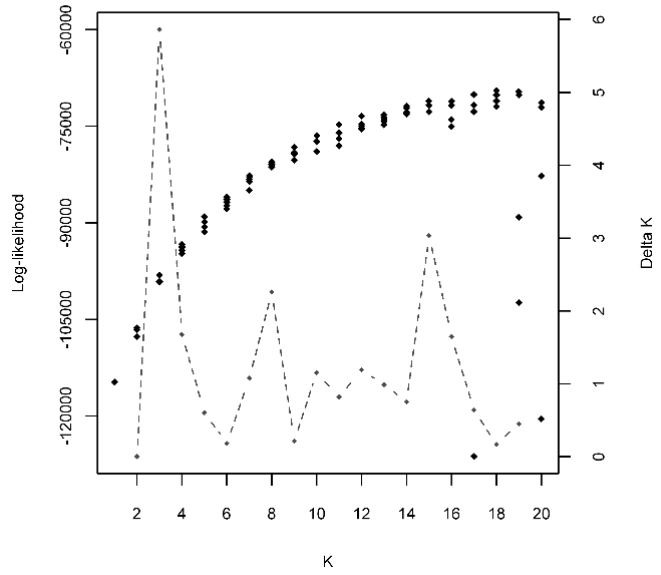


Figure 2.1 Likelihood plot of STRUCTURE results. The black points are the likelihood values and the grey points are the estimated delta values. The plot illustrates the difficulty in deciding the most likely number of subpopulations in the data set.

STRUCTURE clustering solutions at various K values are presented in Figure 2.2 and are consensus of 5 replicate runs. At K = 2 there were inconsistent clustering solutions between runs. At K = 3 the Asian lines either clustered with the British Landrace line or independently. Regarding the clustering patterns, until K = 4 the Asian lines clustered with the British populations.

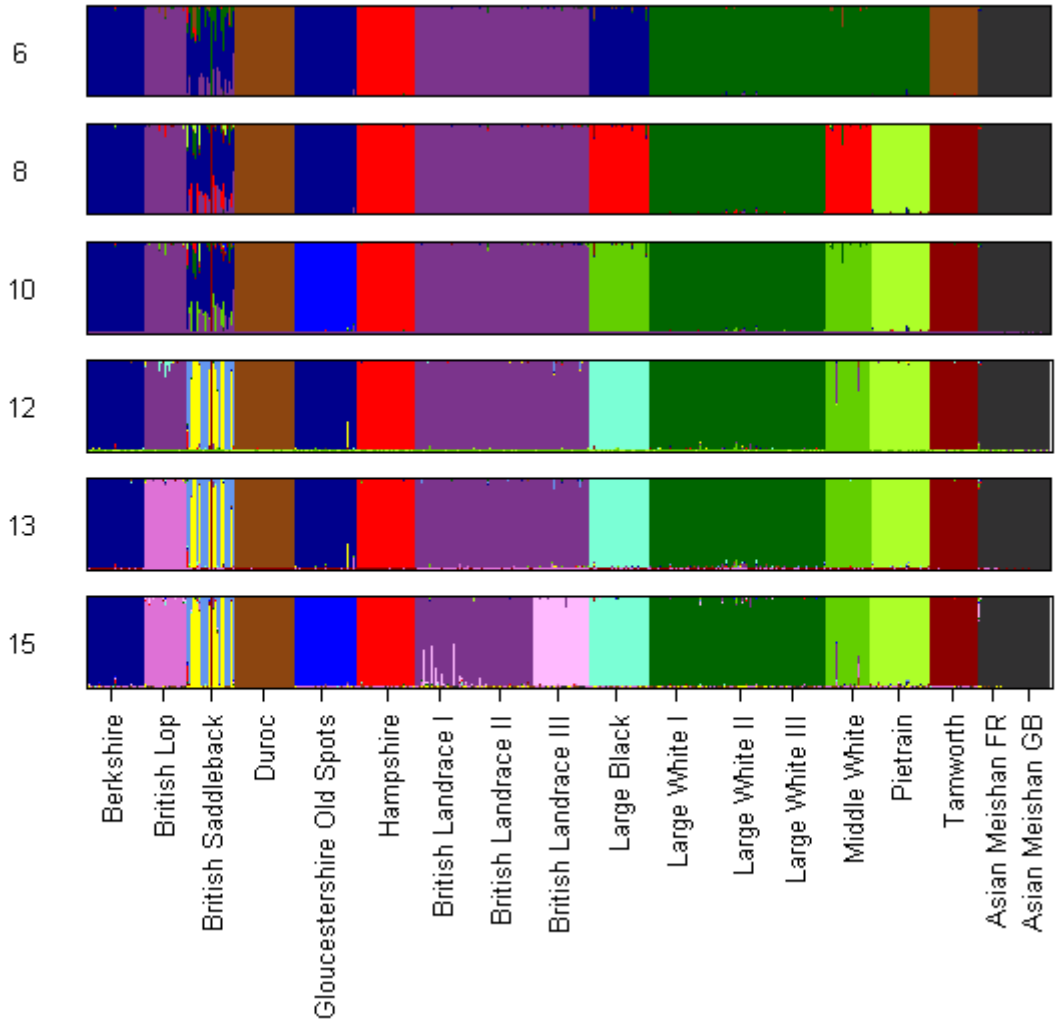


Figure 2.2 Individual assignment from Bayesian genotypic cluster analysis using STRUCTURE at various values of K. Histograms demonstrate the proportion of each individual's genome that originated from each of 18 populations. Each individual is represented by a vertical line corresponding to its membership coefficient (q). Histograms are a consensus view across 5 replicates.

The British Lop and the British Landrace lines consistently clustered together until K = 13, where the British Lop split from the British Landrace lines to form its own cluster. For the remaining breeds the results were inconsistent between the runs such that a large number of clustering solutions were depicted. At values greater than K = 16 the confidence of assignments fell dramatically, with 'ghost' or empty clusters observed. As can be seen in Figure 2.2, there was partial hierarchical splitting of clusters at each stage, but also some inconclusive splitting: for instance, Gloucestershire Old Spots splitting (K = 10) and then rejoining Berkshire (K = 12). Consequently, a sensitivity analysis was conducted where various starting parameters including ALPHA (the degree of admixture), ALPHAPROPSD (standard deviation for ALPHA that allows for better mixing in the Metropolis-Hasting chain) and LAMBDA (distribution of allelic frequencies) were varied from the default values in an attempt to produce more repeatable results. This did not decrease the variation in log likelihood estimates at high K values, nor alter the log likelihood curve or the inconsistent clustering solutions.

Unlike STRUCTURE, BAPS provides a probabilistic approximation of the number of clusters when K was left undefined and the optimal partition was identified at K = 18 (Pr (K = 18 | X) = 1.0). At K = 18, all populations formed their own independent cluster except: i) the two Large White lines formed one population and, ii) the Asian Meishan GB line split over two populations. When K was predefined the clustering solutions were identical between replicate runs at a given K value. BAPS clustering solutions at various K values are presented in Figure 2.3 and are consensus of 5 replicate runs. At K = 2, the Asian populations formed one cluster separate from the

British populations. As K increased, the commercial breeds first split away to form their own clusters: the Large White (K = 4), British Lop-Landrace lines (K = 6), Hampshire (K = 7), Duroc (K = 8) and Pietrain (K = 9).

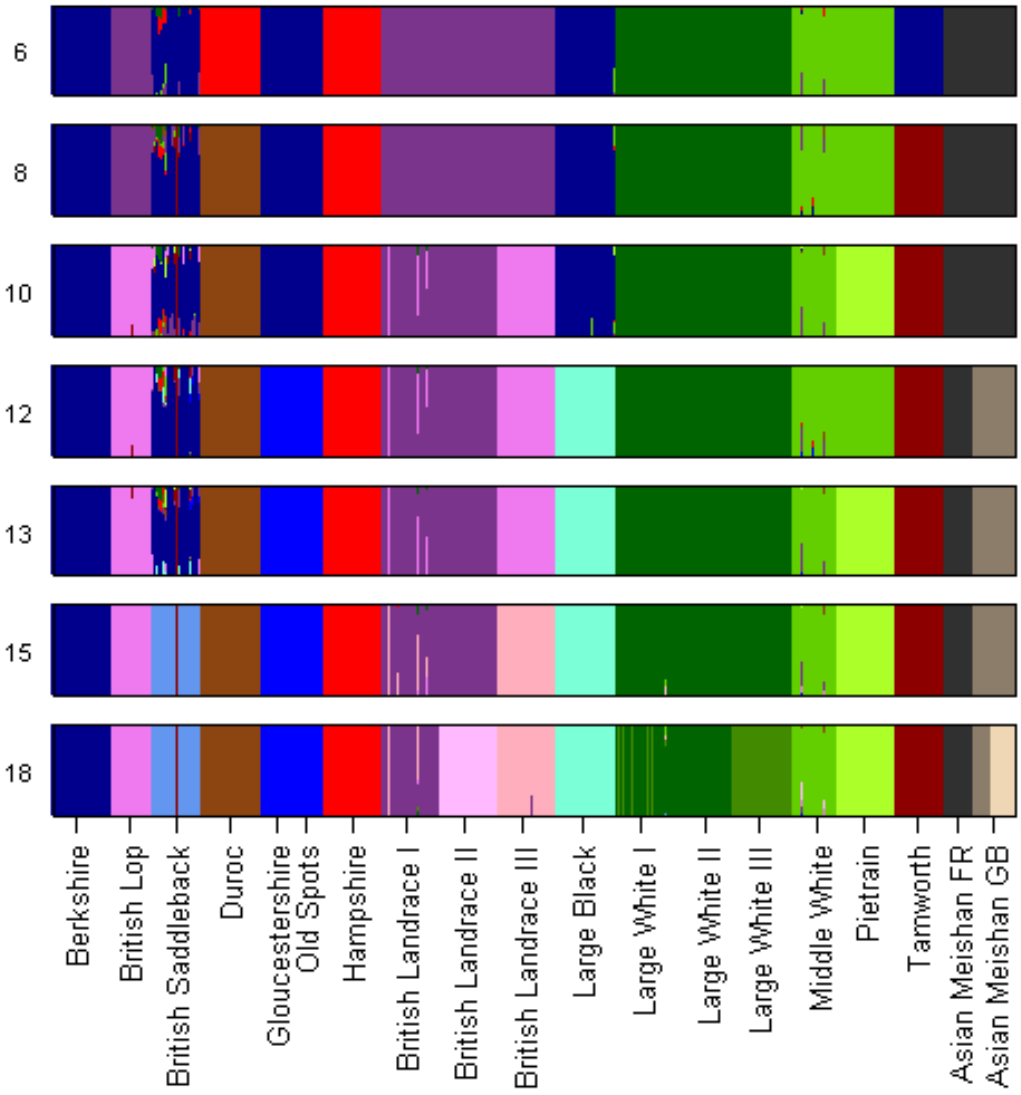


Figure 2.3 Individual assignment from Bayesian genotypic cluster analysis using BAPS at various values of K. Histograms demonstrate the proportion of each individual's genome that originated from each of 18 populations. Each individual is represented by a vertical line corresponding to its membership coefficient (q).

The grouping of British Lop-Landrace into a single cluster was observed until K = 9. The Middle White and Tamworth were the only two traditional British breeds that split at lower K values to occupy independent clusters (K = 9 and K = 8, respectively). This left the remaining four traditional British breeds, Berkshire, British Saddleback, Gloucestershire Old Spots and Large Black, as a single cluster from K = 7 to 10. Once the commercial breeds inhabited independent clusters, at K = 11 the group of 4 indigenous breeds began to split. Berkshire and British Saddleback formed a single cluster from K = 11 - 14. In the BAPS analysis the British Saddleback did not split into two clusters at any point; instead all individuals formed a single genetic population from K =14 - 18. At K = 12, the two Asian Meishan lines split to occupy independent clusters and at K = 18 the Meishan GB population split into two subpopulations. At K = 19 the first 'ghost' population was observed.

The final Bayesian implementation was performed using STRUCTURAMA, which like BAPS, allows K to be a random variable and thus estimates K. The estimated number of populations was 11 across the 5 independent runs (Pr (K = 11 | X) = 0.99). STRUCTURAMA clustering solutions are given in Figure 2.4 for various fixed K values. The clustering of British Lop-Landrace lines in a single cluster was again observed and British Saddleback and Gloucestershire Old Spots were placed in one cluster. All other breeds formed independent clusters. Hierarchical splitting of clusters at lower K values was not observed. At K = 12 British Lop split from the British Landrace. STRUCTURAMA could not converge on a clustering solution for a fixed value of K = 10 and from K = 13 upwards.

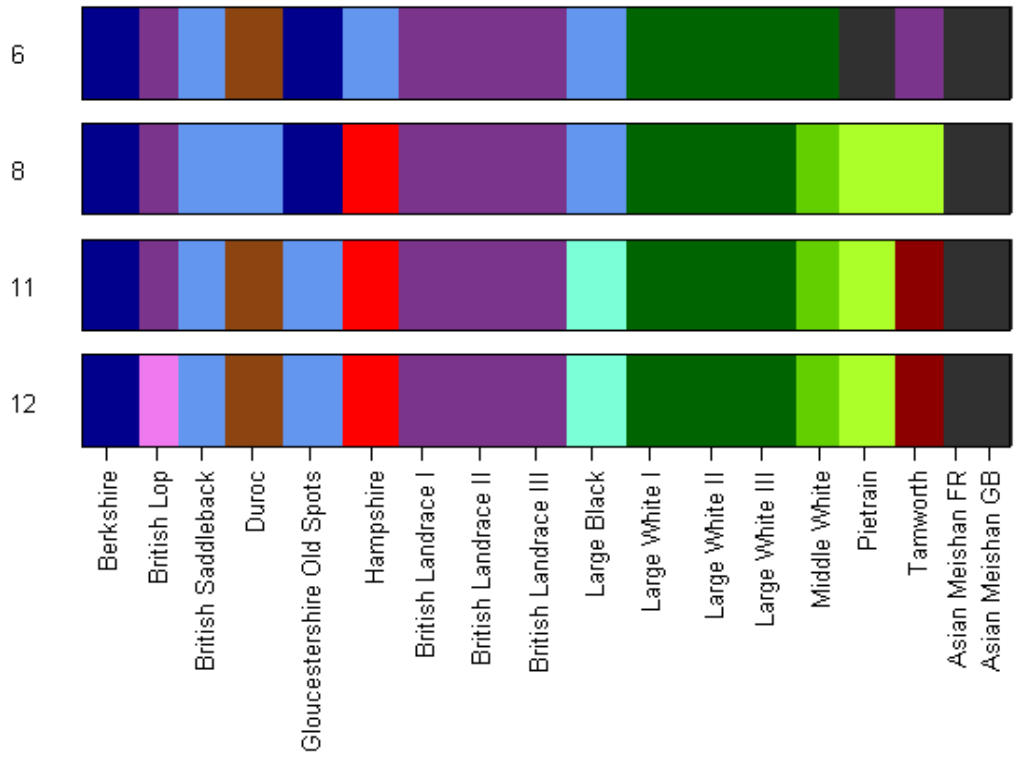


Figure 2.4 Individual assignment from Bayesian genotypic cluster analysis using STRUCTURAMA at various values of K. Histograms demonstrate the proportion of each individual's genome that originated from each of 18 populations. Each individual is represented by a vertical line corresponding to its membership coefficient (q). STRUCTURAMA could not converge on a solution for a fixed value of K = 10.

2.3.1.2 Assignment of individuals and genetic admixture

The majority of individuals clustered to the pre-labelled population origin (proportion of genome assignments, q > 0.9, Fig 2.2-2.4). At lower K values (STRUCTURE - K ≤ 11 and BAPS - K ≤ 15), British Saddleback individuals appeared to be admixed, probably a reflection of the inability of the algorithms to resolve the clustering of this breed. At higher K values STRUCTURE split the British Saddleback into two separate clusters with four individuals being admixed,

whilst BAPS retained the breed as a single genetic unit with admixed individuals present (Fig 2.2-2.3). At higher K values, (both STRUCTURE and BAPS at $K \geq 16$) the Large White individuals was split into 2 clusters, but not strictly according to population identities. Some individuals from Large White Line I clustered with individuals from Line III. Both BAPS and STRUCTURE identified the same individual labelled British Saddleback as being of Tamworth origin ($q > 0.9$) and individuals of Middle White with a proportion of DNA from other breeds ($q > 0.15$). STRUCTURE identified, from $K \geq 5$, one individual from Gloucestershire Old Spots with a substantial proportion of British Saddleback DNA ($q > 0.25$) and one individual from the French Meishan population with a proportion of Large Black DNA ($q > 0.15$). BAPS identified five individuals of British Landrace line I with proportions of DNA from British Landrace III and British Lop.

2.3.2 Principal component analysis

The first two principal components (PC) are shown plotted in Figure 2.5. The first PC accounted for 29.3% of the underlying variation and the second PC accounted for 4.3%. The first PC clearly split the British from the Asian populations. The second PC gave a coarse separation of the British breed individuals: British Landrace lines and British Lop group clustering at the top of the quadrant with Large White lines grouping at the bottom (Fig 2.5a). The third PC, which accounted for 4.1% of the variation, showed additional structuring amongst the British breeds (Fig 2.5b). The Large White lines were clearly separated from the other breeds at the top left of the plot and the Berkshire and Gloucestershire Old Spots clustered together at the bottom

left. With increasing dimensions there was further breed partitioning: PC 5 (3.3%) separated the Hampshire and Tamworth breeds (Fig 2.6a), PC 6 (2.4%) separated the Duroc breed (Fig 2.6b) and PC 7 (2.1%) separated the Pietrain, Middle White and Large Black breeds.

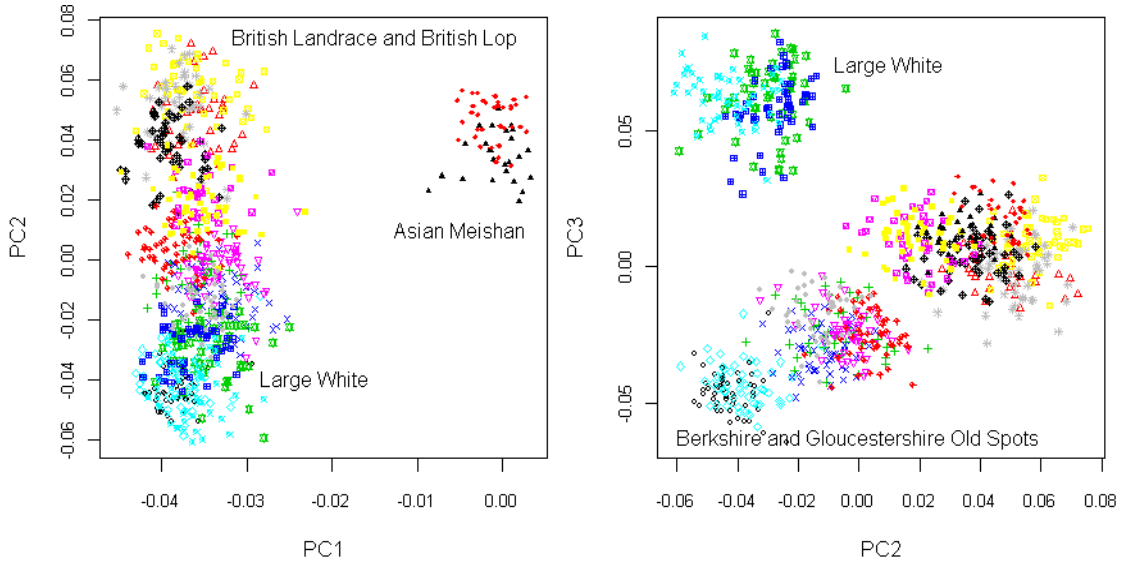


Figure 2.5 Principal component analysis projections. (a) scatterplot diagram showing the first and second PCs and allele distribution from all individuals. (b) scatterplot diagram showing the second and third PCs and allele distribution from all individuals.

No further breeds or populations within breeds were partitioned out. According to the parallel test 28 PCs should be retained. However, from PC 12 the components were noisy and non-informative as there visually appeared to be no structure present. When PCA was conducted on just the British populations, congruent results were produced. The first PC (30%) spread out individuals within the breeds and the second PC (4.6%) gave a coarse separation of the populations. The PCA projection was similar to that shown in Fig 2.5, save for the Asian populations.

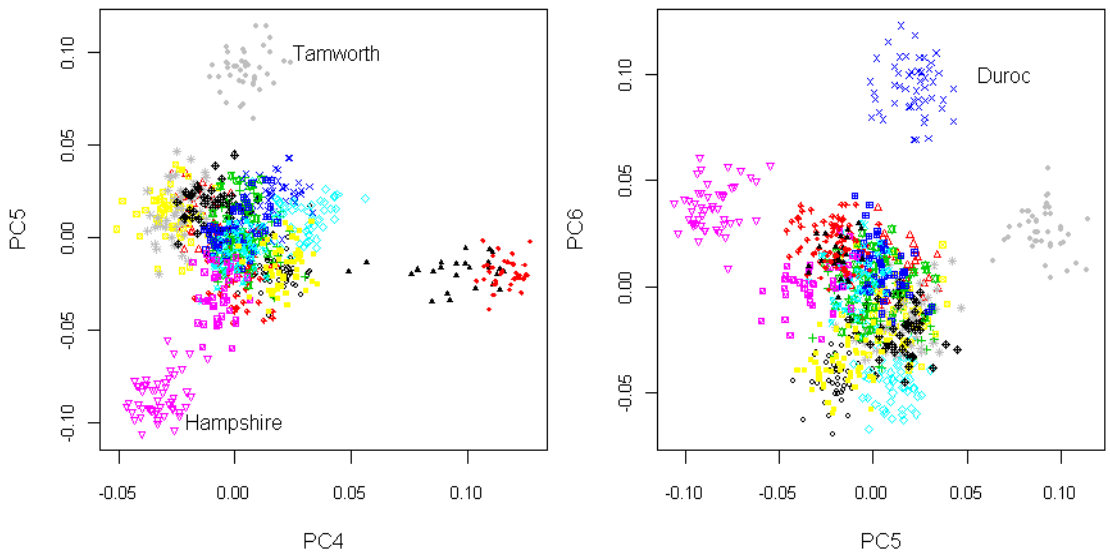


Figure 2.6 Principal component analysis projections. (a) scatterplot diagram showing the fourth and fifth PCs and allele distribution from all individuals. (b) scatterplot diagram showing the fifth and sixth PCs and allele distribution from all individuals.

2.3.3 Phylogenetic reconstruction

The phylogenetic reconstruction based on the proportion of shared alleles distance (PSA) measure is presented in Figure 2.7. All individuals clustered to their designated breed origin except for the British Saddleback in which individuals were split into two clusters. One British Saddleback individual fell within the Tamworth clade (the same individual identified in the Bayesian genotypic clustering analyses). There was high bootstrap support for individuals belonging to their breed of origin, except for the British Landrace-Lop and the British Saddleback groupings. The longest branches separated individuals within breeds implying that there is greater variation within than between breeds. There was no bootstrap support for genetic relationships between the pig breeds.

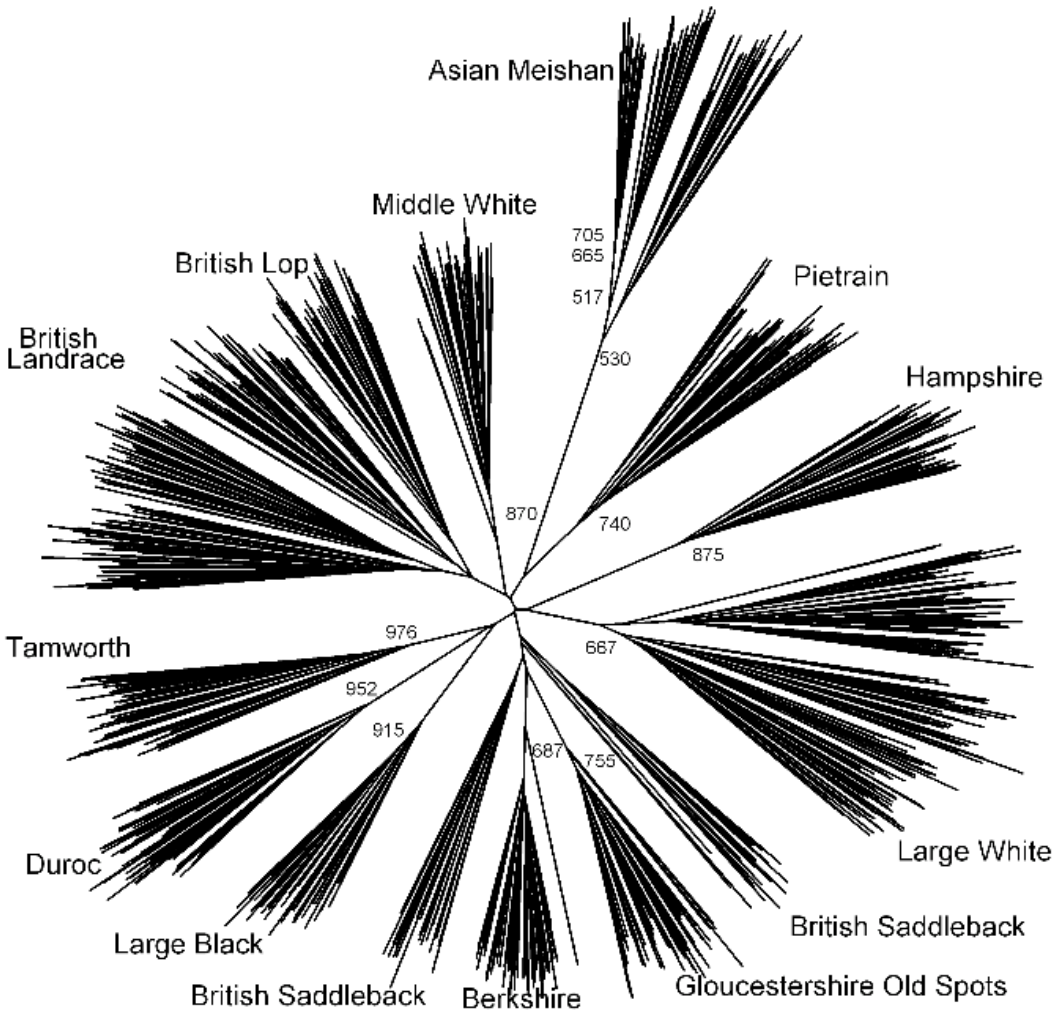


Figure 2.7 A neighbour-joining tree constructed from allele-sharing distances among all individuals. Bootstrap values greater than 500 are shown.

2.3.4 Genetic differentiation

The level of genetic differentiation (F_{ST}) for the defined populations listed in Table 2.1 and for certain clusters that were inferred beyond the defined populations, at K = 15 and K = 18 for STRUCTURE and BAPS, respectively are presented in Table 2.2.

Table 2.2 Population genetic differentiation amongst the populations. [1] an inferred clustering by STRUCTURE at K = 15, [2] an inferred clustering by BAPS at K = 18, [3] the genetic differentiation of a population calculated by averaging over all pairwise comparisons for a given population.

	Breed	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	a[1]	b[1]	a[2]	Ave[3]
1	Berkshire																						0.33
2	British Lop	0.34																					0.25
3	British Saddleback	0.19	0.17																				0.21
4	Duroc	0.33	0.27	0.21																			0.31
5	Glouc. Old Spots	0.26	0.30	0.19	0.31																		0.31
6	Hampshire	0.33	0.27	0.23	0.32	0.34																	0.31
7	British Landrace I	0.34	0.12	0.19	0.30	0.33	0.28																0.24
8	British Landrace II	0.31	0.14	0.16	0.29	0.29	0.26	0.09															0.24
9	British Landrace III	0.30	0.12	0.16	0.28	0.27	0.26	0.12	0.13														0.23
10	Large Black	0.31	0.23	0.17	0.29	0.28	0.28	0.15	0.24	0.24													0.27
11	Large White I	0.30	0.23	0.18	0.29	0.28	0.28	0.22	0.22	0.20	0.26												0.24
12	Large White II	0.30	0.23	0.19	0.28	0.28	0.28	0.21	0.22	0.20	0.26	0.05											0.25
13	Large White III	0.34	0.28	0.23	0.33	0.33	0.32	0.29	0.29	0.26	0.30	0.10	0.14										0.29
14	Middle White	0.34	0.24	0.19	0.31	0.33	0.29	0.22	0.22	0.21	0.24	0.22	0.23	0.29									0.27
15	Pietrain	0.31	0.23	0.18	0.30	0.30	0.31	0.23	0.23	0.20	0.26	0.22	0.23	0.27	0.22								0.27
16	Tamworth	0.34	0.29	0.21	0.31	0.32	0.36	0.30	0.30	0.27	0.30	0.28	0.27	0.33	0.34	0.31							0.32
17	Asian Meishan FR	0.48	0.35	0.33	0.41	0.45	0.43	0.37	0.37	0.36	0.40	0.36	0.37	0.42	0.38	0.35	0.44						0.38
18	Asian Meishan GB	0.47	0.36	0.35	0.42	0.45	0.44	0.37	0.38	0.37	0.41	0.37	0.38	0.43	0.40	0.36	0.44	0.11					0.38
a[1]	British Saddleback I	0.25	0.20	0.01	0.26	0.24	0.27	0.21	0.18	0.18	0.22	0.22	0.23	0.28	0.24	0.22	0.24	0.36	0.38				
b[1]	British Saddleback II	0.21	0.19	0.01	0.21	0.21	0.23	0.20	0.18	0.18	0.17	0.17	0.19	0.23	0.20	0.18	0.23	0.33	0.35	0.08			
a[2]	Asian Meishan GB I	0.50	0.38	0.37	0.45	0.48	0.47	0.39	0.40	0.39	0.44	0.40	0.40	0.46	0.43	0.38	0.48	0.15	0.06	0.41	0.38		
b[2]	Asian Meishan GB II	0.50	0.38	0.37	0.45	0.48	0.46	0.40	0.40	0.40	0.44	0.39	0.40	0.45	0.42	0.39	0.48	0.15	0.01	0.41	0.38	0.18	

Pairwise F_{ST} for the defined populations ranged from 0.05 (Large White I vs Large White II) to 0.47 (Asian Meishan vs Berkshire), with an average F_{ST} of 0.28. Genetic differentiation of the defined population ranged from 0.21 in British Lop to 0.38 in both the GB and FR Meishan populations. STRUCTURE divided the British Saddleback into two clusters at $K \geq 12$ (Fig 2.2) and the level of genetic differentiation between the two inferred genetic clusters was $F_{ST} = 0.08$. BAPS delineated the GB Meishan population into two clusters at K = 18 (Fig 2.3) and the level of genetic differentiation between the two inferred genetic clusters was $F_{ST} = 0.18$.

2.4 Discussion

2.4.1 Population structure

Bayesian genotypic clustering methods offer the prospect of characterising population structure by inferring the number of underlying populations, K, present in an empirical data set. However, obtaining a definitive value of K for the British pig breeds proved a challenge as the three Bayesian methods, STRUCTURE, BAPS and STRUCTURAMA, yielded slightly different answers.

STRUCTURE does not provide a statistical indication of the most likely K. Instead, K is identified at a point of inflection on the log-likelihood curve that leads to a plateau or by the maximum value (Pritchard and Wen 2004). However, when there is a continual increase in the log likelihood, as observed in Figure 2.1, choosing K may be problematic and is often a subjective task (Frantz et al. 2006). The Evanno et al. (2005) ΔK method did not clarify the best value of K for the British pig breeds as the

magnitude of ΔK were not strong or conclusive compared with, for example, ΔK = 130 reported by Frantz et al. (2006). In addition, variance in the log likelihoods for a given K increased at high values of K (K > 16, Fig 2.1) as has been reported with other data sets (Evanno and Regnaut 2005; Rosenberg et al. 2001). Through visual observation of the log-likelihood curve, it is probable that the value of K lies between 10 and 15 (Fig 2.1). This is also supported by the fact that 'ghost' populations started to appear from K = 16 in the STRUCTURE analysis. Other studies on both domestic and wild species have similarly experienced difficulty in identifying K using STRUCTURE (cat breeds (Menotti-Raymond et al. 2008) and red deer populations (Frantz et al. 2006)). In the STRUCTURE analysis the very gradual increase in log-likelihood values up to an asymptote may be indicative of the presence of genetic continuity across breeds. Such a pattern could be a possible consequence of limited breed barriers due to a short history and cross-breeding (Menotti-Raymond et al. 2008). With wild populations, similar STRUCTURE results have been attributed to an isolation-by-distance relationship (Frantz et al. 2006). In that situation, populations of individuals are spatially distributed and the chance occurrence of gene flow amongst populations declines with increasing geographic distance between populations creating a pattern of decreasing genetic similarity of populations with increasing geographic distance between the populations. It may prove challenging to define genetic populations that exhibit an isolation-by-distance or other subtle structuring as the STRUCTURE model cannot easily accommodate data that display these types of patterns (Pritchard and Wen 2004).

STRUCTURAMA implements a simpler version of the STRUCTURE model but it also allows K to be a random variable. Since the manual selection of the number genetic populations may be considered a drawback this is a useful extension. STRUCTURAMA gave a value of 11 for the number of underlying populations. At this value of K, STRUCTURAMA produced a biologically credible clustering solution where all breeds were independent units except for British Lop-British Landrace and British Saddleback-Gloucestershire Old Spots (Fig 2.4).

BAPS estimated a higher value of K (18) than STRUCTURE and STRUCTURAMA, but also produced biologically credible clustering solutions. The clustering result at this value of K did not entirely correlate with the 18 identified populations (Table 2.1) in that the Asian Meishan GB line was split into two clusters, the Large White lines I and II clustered together and a few individuals of Large White I clustered with Large White III (Fig 2.3). Rowe and Beebee (2007) similarly observed that BAPS inferred a greater number of genetically distinct groups in natterjack toad populations, than did STRUCTURE. It has been suggested that BAPS infers a finer level of genetic structure, though not necessarily true structure, when patterns of isolation-by-distance are present in a dataset (Frantz and Cellina 2009; Safner et al. 2011). The BAPS algorithm may detect weak random fluctuations in allele frequencies, which could be considered as evidence of genetic differentiation amongst sub-populations (Corander et al. 2008).

It is difficult to compare the clustering solutions of STRUCTURE and BAPS at higher values of K as the former was not consistent between runs and also produced

one or more ‘ghost’ populations. Comparison of returned clustering solutions at lower K values between the Bayesian genotypic clustering approaches revealed some differences. The methods concurred that certain breeds became distinct genetic units at low K values (e.g. K = 8, Duroc, White, Tamworth, Meishan; Fig 2.2-2.3). Yet, the clustering solutions of other breeds (namely Berkshire, British Saddleback and Gloucestershire Old Spots) were largely unresolved due to inconsistent results between the methods. At a specific value of K, different pairings of breeds were observed (e.g. K = 12). Additionally, sometimes the methods returned the same clustering observations but at different levels of K. For instance, BAPS observed subdivision between the lines of British Landrace from K = 10 (Fig 2.3) whilst STRUCTURE produced this result at a higher value of K = 15 (Fig 2.2).

An additional incongruence between the Bayesian genotypic clustering methods was in the detection of substructure within breeds. The British Saddleback did not form a single cluster according to the STRUCTURE (Fig 2.2) and phylogenetic analysis (although the bootstrap support for this was low, ~30%; Fig 2.7), while BAPS did not provide any evidence of substructuring within the breed (Fig 2.3). In contrast, BAPS separated the French Meishan from the British Meishan, and furthermore, detected substructure with the British Meishan (K = 18, Fig 2.3). Moreover, there was high bootstrap support (> 50%) for the division of British Meishan into two genetic clusters (Fig 2.7). The British Meishan was known to be composed of individuals from two separate subpopulations, derived from a single importation but subsequently bred separately, thus these findings reflect true differentiation likely due to founder effects and subsequent drift and selection after importation. The

British Saddleback, on the other hand, was considered a panmictic population like the other British pig breeds.

STRUCTURE did not reveal definitive substructure beyond the level of breed in those composed of separate lines. At high values of K there was evidence for further substructure in the British Landrace and Large White (Fig 2.2). However, 'ghost' populations, inconsistent clustering solutions and a large variation in the log-likelihood were evident at these high K values (Fig 2.2). This indicates that the MCMC chain had not converged, which could suggest that genetic differentiation was weak (Waples and Gaggiotti 2006). Yet, the levels of genetic differentiation in this analysis should have been sufficient (F_{ST} > 0.05, suggested by (Latch et al. 2006)) for the detection of the substructure in this dataset (British Landrace lines F_{ST} = 0.11 (mean F_{ST} between the 3 lines); the two identified genetic clusters of British Saddleback, F_{ST} = 0.08; the two identified genetic subpopulations of British Meishan, F_{ST} = 0.18, Table 2.2).

Overall, BAPS detected genetic structure at a finer scale than the other Bayesian clustering methods. Although other studies have found that BAPS tends to 'overestimate' K (Frantz and Cellina 2009; Safner et al. 2011), for the British pig breeds the higher estimated value of K produced genetic groups that were all biologically credible (Fig 2.3). Incongruent results between different Bayesian clustering methods have also been encountered in previous studies (Ball et al. 2010; Frantz and Cellina 2009; Latch et al. 2006; Rowe and Beebee 2007) and it is not apparent what causes these inconsistencies. They may arise from differences in the

underlying models, the statistical estimators or the algorithms used (Guillot et al. 2005) or there may not be sufficient genetic information to conclusively differentiate groups within breeds (e.g. British Saddleback) and, consequently, the methods may be operating at their limits. Thus, it remains uncertain whether STRUCTURE and phylogenetic reconstruction have uncovered real genetic structure within the British Saddleback.

2.4.2 Assignment of individuals to origin and genetic diversity

The majority of the individuals were successfully assigned to their pre-designated breed origin using both the Bayesian genotypic clustering methods and phylogenetic reconstruction. This is reflected by the estimation of high membership proportions (Fig 2.2-2.4, $q > 0.9$, clustering methods) and high bootstrap values following resampling of loci (phylogenetic reconstruction). The lack of admixture indicates that the majority of the British pig breeds are distinct genetic units and that there is little hidden substructure within the breeds, which is substantiated by the estimated high levels of genetic differentiation (Table 2.2). In a study on dog breeds, Koskinen (2003) also reported that Bayesian genotypic clustering and phylogenetic reconstruction performed similarly in terms of assigning individuals to breed origin.

Phylogenetic reconstruction also indicated that a large amount of genetic variation lies within the British pig breeds. A cladogram was produced, where the longest branches separated individuals within breeds (Fig 2.7). This was substantiated by multivariate analysis of the microsatellite allele distribution. The PCA projection

showed that individuals were not tightly clustered, and were instead spread out and populations overlapping (Fig 2.5-2.6). An analysis of molecular variance (AMOVA) (Excoffier and Heckel 2006), a measurement that partitions variance among groups, also revealed that most of the variation was found within the pig breeds (~72%, $P < 0.0001$).

2.4.3 Defining the genetic boundaries of breeds

Kalinowski (2011) stated that the colour-coded plots produced by Bayesian genotypic clustering approaches (e.g. Fig 2.2-2.4) provide limited information on population structure because the relationships amongst populations cannot be described. It is true that the degree of genetic differences or similarities between inferred populations or clusters cannot be numerically quantified using these resultant plots. However, it can be argued that the Bayesian genotypic clustering approaches do illustrate certain genetic relationships: populations that split to form independent clusters at lower K values can be interpreted as being relatively genetically unique and populations that cluster together independent of others at many K values could be indicative of genetic similarity.

With Bayesian genotypic clustering, as well as PCA, data can be examined at a number of dimensions, where populations may separate to form their own independent genetic unit with each increase in principal component or value of K. This likely indicates distinctive multilocus genetic combinations for these particular populations (Rosenberg et al. 2001). In both PCA and BAPS, the Large White was

the first British breed to form a distinct cluster, with the other commercial breeds following with increasing dimensions (Fig 2.3, 2.5-2.6).

Some credible genetic groupings of pairs of breeds were also observed from the individual-based analyses. The first was the clustering of the British Lop breed with the British Landrace lines, a breed of European origin. Megens et al (2008) showed, using phylogenetic reconstruction, a genetic affinity of British Lop with other European pig breeds. This suggests that British Lop may either be a breed of European origin or has experienced substantial genetic introgression from British Landrace (Hall and Clutton-Brock 1988). The second was the pairing of Berkshire and Gloucestershire Old Spots breeds. These two observations were consistent across all the statistical approaches (Fig 2.2-2.7). In a third case, some of the methods revealed that the British Saddleback breed shared a genetic affinity with the Berkshire-Gloucestershire Old Spots grouping. STRUCTURAMA clustered the breed with Gloucestershire Old Spots and STRUCTURE placed British Saddleback with the Berkshire-Gloucestershire Old Spots cluster in the majority of the replicates for low K values. In addition, the three breeds shared an internal branch on the phylogenetic topology (Fig 2.7). The observed genetic similarities between these pig breeds are supported by historical information. Firstly, the three breeds are indigenous to Great Britain and share a common geographic origin: the counties of the south of England (BPA 2002; Porter 1993). Secondly, the Berkshire was once a popular and prevalent pig used to improve other breeds (BPA 2002; Porter 1993). Historic genetic introgression of the Berkshire could have augmented or maintained the genetic affinities between these indigenous breeds.

Beyond pairs of breeds, the genetic structure amongst the British pig breeds could not be discerned. Although the breeds were highly differentiated (Table 2.2), the cladogram had a low bootstrap support for relationships between the breeds, a star-shaped topology and short internal branches that separated individuals from those of other breeds (Fig 2.7). The lack of a robust and coherent evolutionary tree has also been observed in a larger group of European pig breeds (Megens et al. 2008; SanCristobal et al. 2006a) and the result could be due to cross-breeding during the history of development of these breeds (Eding and Bennewitz 2007). The inconsistent clustering solutions between replicate runs found using STRUCTURE could reflect the same phenomenon. Peter et al (2007) found that STRUCTURE depicted a few possible clustering solutions for sheep breeds and that clusters were only partially hierarchical. Li et al (2007) similarly reported variation between runs, suggesting a lack of high-level substructure in European cattle breeds. Historical cross-breeding for improvement may have created a genetic structure such that the individuals of the British pig breeds are equally genetically similar, leading to the existence of multiple possible clustering solutions.

2.5 Conclusion

The performance of three Bayesian genotypic clustering methods, PCA and phylogenetic reconstruction were compared in the inference of population genetic structure in a livestock breed. Except for PCA, which was only able to separate breeds into related groups and could not identify the individual breeds, the methods were similarly effective in delineating breeds and assigning individuals to breed of

origin. However, there were incongruent results between the different Bayesian genotypic clustering techniques with respect to the determination of K, clustering solutions and the detection of substructure within breeds. Of the Bayesian genotypic clustering methods, BAPS detected finer genetic differentiation within the breeds with known substructure.

CHAPTER THREE

The genetic structure of the British Saddleback pig breed

3.1 Introduction

In chapter 2, the application of several individual-based genetic clustering techniques to characterise population genetic structure was explored using microsatellite data from British pig breeds. As reported, there were inconsistent clustering results between the different methods; one of the most striking was the inferred clustering solutions of the British Saddleback breed. Of the Bayesian genotypic clustering approaches, BAPS retained the British Saddleback as a single genetic unit (Fig 2.3), whilst STRUCTURE split the breed into two genetic clusters (Fig 2.2). The latter clustering solution was supported by phylogenetic reconstruction, which also could not resolve the British Saddleback into a single clade (Fig 2.7).

Inconsistent clustering solutions between different Bayesian genotypic clustering approaches have been observed in other studies (Frantz and Cellina 2009; Rowe and Beebee 2007). Consequently, it was uncertain whether STRUCTURE and phylogenetic reconstruction were uncovering real genetic substructure within the British Saddleback. The individual-based clustering approaches may have been operating at their extremes. Nonetheless, the clustering results were intriguing especially considering the known history of the breed.

The British Saddleback is a relatively young British pig breed, the result of the amalgamation of two breeds, the Essex pig and Wessex Saddleback. The two breeds, in particular the Wessex Saddleback, were once very popular. However, after World War II the importation of European breeds with favourable production traits led to a

considerable population decline in many traditional British pig breeds, including the Essex pig and Wessex Saddleback. Consequently, based on the common possession of a white belt ('saddleback'), the two breeds were combined to establish the British Saddleback in 1967 (Alderson 2007; BPA 2002; Porter 1993).

Although the white saddleback was a common feature to both breeds, there were notable differences in the historical development and morphological characteristics of the Essex pig and Wessex Saddleback (Alderson 2007; BPA 2002; Porter 1993). The Essex pig originated from Essex and was improved with Asian pig breeds, whilst the Wessex Saddleback was found in the South and South West of England and its breed society resisted the introduction of Asian alleles. In body size, the Wessex Saddleback tended to be larger than the Essex pig. As stated earlier, both the Essex pig and Wessex Saddleback possessed a white belt that encircled the body at the shoulders and extended down the fore-legs. However, in the Wessex Saddleback the width of the belt tended to be smaller, sometimes incomplete and with black spots occasionally dotted inside it. For both breeds, the rest of the body was black, with the exception of white appearing on the hind legs and tip of the tail in the Essex pig.

Despite the similarities between the breeds it has been steadfastly maintained that the Essex and Wessex Saddleback were two genetically distinct breeds due to their disparate geographic origins, different morphological characteristics and subsequent breed development. Therefore, there have been reservations over the amalgamation based solely on the feature of a shared white belt (Alderson 2007; BPA 2002; Porter

1993). Given the known history of the British Saddleback breed and the inconclusive genetic clustering results reported in chapter 2, further analysis of the British Saddleback individual genotype data was warranted. The aim of this chapter was to further elucidate and characterise the genetic structuring within the British Saddleback pig breed.

3.2 Materials and Methods

3.2.1 Data

The genotypic data of the British Saddleback breed was a subset from an extensive European pig breed biodiversity study (SanCristobal et al. 2006a). Genotype data was available from 46 microsatellite markers for 41 British Saddleback individuals.

The following information was also available for each individual: sex, date of birth, farmer (geographic location), sire name and dam name. Further information was contained in each sire and dam name. The names consisted of two parts: the first was the herd prefix and the second was the bloodline, both of which indicate the ancestry of the individual. The British Saddleback has a total of 45 bloodlines, comprising of 31 female bloodlines and 14 male bloodlines. Of these lines, 11 of the female bloodlines and 3 of the male bloodlines have ancestry traced back to the old Essex pig herd books (EPS 2011). Of these Essex pig lines, most are said to contain some Wessex Saddleback ancestry, except for 3 female lines (Alvis, Duchess and Grand Duchess) and 1 male line (Dictator) that are alleged to still be ‘pure’ Essex pig. Individuals allegedly free of Wessex genetic contamination and of ‘pure’ Essex pig origin were sampled: 1 individual from the Glascote herd of sire bloodline Dictator,

1 individual of dam bloodline Alvis and 3 more individuals with sire bloodline Dictator. The Glascote herd is said to be founded by Essex pigs before the amalgamation and has remained 'pure' since its foundation.

3.2.2 Within and among population diversity

Genetic variation within a population was first measured by estimating the frequency of heterozygotes (the proportion of heterozygous individuals in a population). The observed heterozygosity (H_O) was estimated directly from the frequency of heterozygotes observed in the sample population averaged across all loci. Assuming Hardy Weinberg equilibrium (HWE), the expected heterozygosity (H_E) was estimated from the allele frequencies as $1 - \sum p^2$ averaged over all loci. In addition, the total number of alleles was obtained by counting the number of alleles occurring in a population. All calculations were performed using the program FSTAT 2.9.3 (Goudet 1995).

The amount of genetic diversity within each individual was determined by measuring the proportion of microsatellite loci that were heterozygous. Individual multilocus heterozygosity (MLH) was calculated as the number of loci at which an individual was heterozygous, divided by the total number of loci at which an individual was scored (Coltman et al. 1999; Slate et al. 2000). The calculations were performed in R (Team 2011).

To determine if populations were in HWE, exact tests were computed using GENEPOP 4.0.7 (Rousset 2008) and significance levels of the tests were adjusted for multiple comparison following standard Bonferroni corrections (Rice 1989) (for further details see section 2.2.1 in chapter 2).

Genetic differentiation of populations was measured by estimating Weir and Cockerham's F_{ST} (Weir and Cockerham 1984) (for further details see section 2.2.5 in chapter 2).

Individual inbreeding coefficients (F), defined as the probability that two homologous alleles in an individual are identical by descent from a recent common ancestor (Wright 1921; Wright 1922), were provided by Rex Walters, in association with the British Pig Association, for 37 of the British Saddleback individuals. F values were estimated by tracing the pedigree of individuals back to the common ancestors of the parents and computing the probabilities of alleles segregating at each generation. F values were estimated from the British Saddleback pedigree using the software Geneped (developed by Rare Breeds Survival Trust and Grassroots). The time period used to estimate F values was from 01/01/1977 (the start of the electronic database) to 1998/1999 (the year of birth of the sampled individuals).

3.2.3 Clustering of individuals populations

Two Bayesian genotypic clustering approaches, STRUCURE 2.1 (Pritchard et al., 2000) and BAPS 5.2 (Corander et al. 2008), were used to probabilistically assign

individual multilocus genotypes to clusters and infer the number of genetically distinguishable populations (K) (for further details see section 2.2.2 in chapter 2). For both STRUCTURE and BAPS, five replicate runs were implemented from $1 \leq K \leq 5$. The phylogenetic relationships amongst the individuals were reconstructed by estimating the proportion of shared alleles (for further details see section 2.2.4 in chapter 2).

3.3 Results

Bayesian genotypic clustering analysis using both BAPS and STRUCTURE on only the British Saddleback breed produced concordant clustering solutions at $K = 2$, where the same individuals grouped together in the two separate subpopulations. The separation of the British Saddleback individual multilocus genotypes into two genetic subpopulations was also in concordance with that found using STRUCTURE in chapter 2 at $K \geq 12$ (Fig 2.2). Individuals that clustered to either of the two genetic subpopulations did so with a high proportion ($q > 0.8$). There were four individuals whose genomes were split between the two clusters and considered as admixed ($0.2 < q < 0.8$). Phylogenetic reconstruction of the British Saddleback individuals similarly divided the breed into the two genetic clusters with all four admixed individuals positioned in one of the subpopulations. However, it was noted that certain loci genotyped in the British Saddleback breed had a large proportion of missing data. To determine if the presence of missing data influenced the inferred clustering solutions, loci with greater than 10% missing data (8) were removed leaving a total of 38 microsatellites. Re-analysis of the 38 microsatellite loci using

BAPS and STRUCTURE produced concordant clustering solutions to that of the full dataset (all 46 loci) at K = 2, except that only one individual was admixed (Fig 3.1). From K = 3 to K = 4 both BAPS and STRUCTURE split the larger of the clusters identified at K = 2, the green cluster in Figure 3.1, into further clusters. At K = 5 the individuals belonging to the green cluster identified at K = 2 were further split to produce four clusters by BAPS. At K = 5 variation in the log-likelihood between runs was observed for STRUCTURE. Three out of five runs split the individuals belonging to the green cluster identified at K = 2 to produce four clusters. The remaining two runs split individuals belonging to the pink cluster identified at K = 2 into two further clusters.

Figure 3.1 Histogram of the individual assignment from Bayesian genotypic cluster analysis using STRUCTURE at K = 2 (38 microsatellite loci). Each individual is represented by a vertical line corresponding to its membership coefficient (q).

To determine the number of underlying genetic subpopulations in the British Saddleback breed the statistic delta (ΔK), which represents the rate of change with successive K values, was estimated for the STRUCTURE analysis. In Figure 3.2, the rate of change in the log likelihood with successive K values, ΔK, was plotted from $K_1 - K_2$ to $K_4 - K_5$. As can be seen, the maximum value of ΔK showed that the greatest gain with increasing K was from K_1 - K_2 indicating that the most likely clustering solution for the British Saddleback dataset was two subpopulations. The

log-likelihood plot for BAPS results showed a notable decrease in the rate of gain in log likelihood with successive K value starting at $K_2 - K_3$ also suggesting that the optimal structuring of the British Saddleback is two subpopulations (Fig 3.2).

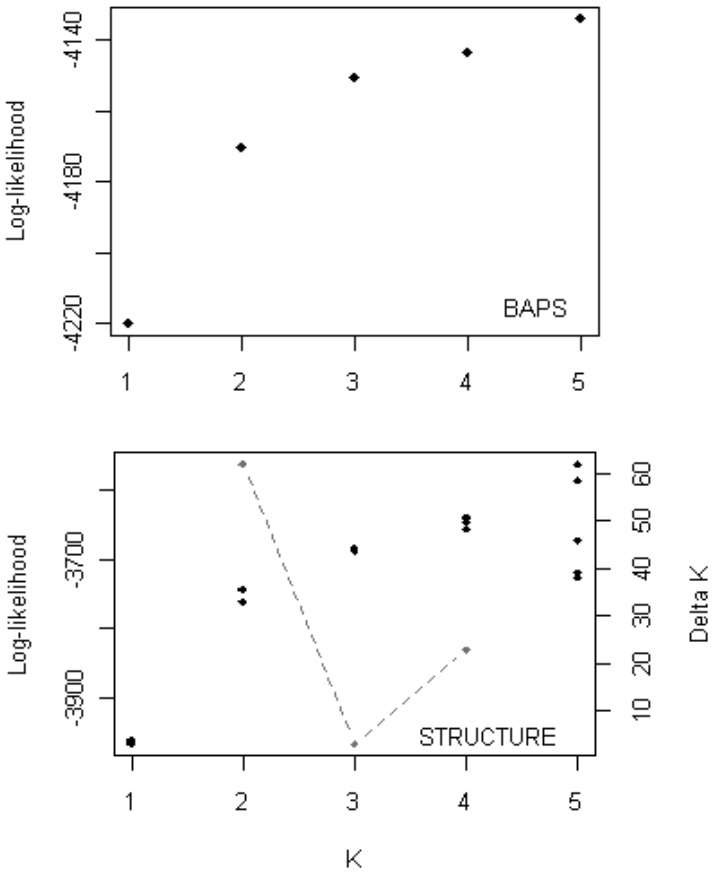


Figure 3.2 Plot of log-likelihood with increasing K value for the BAPS and STRUCTURE analyses (38 microsatellite loci). The black dots on the STRUCTURE plot represent the log-likelihood values and the grey dots connected with a dashed line represent the delta estimate.

The phylogenetic reconstruction of the British Saddleback breed with the individual genotype data of 38 microsatellite loci was visualised as a neighbour-joining tree in Figure 3.3. The reconstruction replicated the inferred genetic subdivision of the

British Saddleback using Bayesian genotypic clustering approaches, where the same individuals grouped together in the two clades. The one inferred admixed individual (Fig 3.1) fell in between the two reconstructed clades (Fig 3.3). There was a low bootstrap support (~35%) to separate the two inferred genetic subpopulations. When the one inferred admixed individual was removed from the analysis, a slightly stronger bootstrap support was obtained for the two major clades (49%).

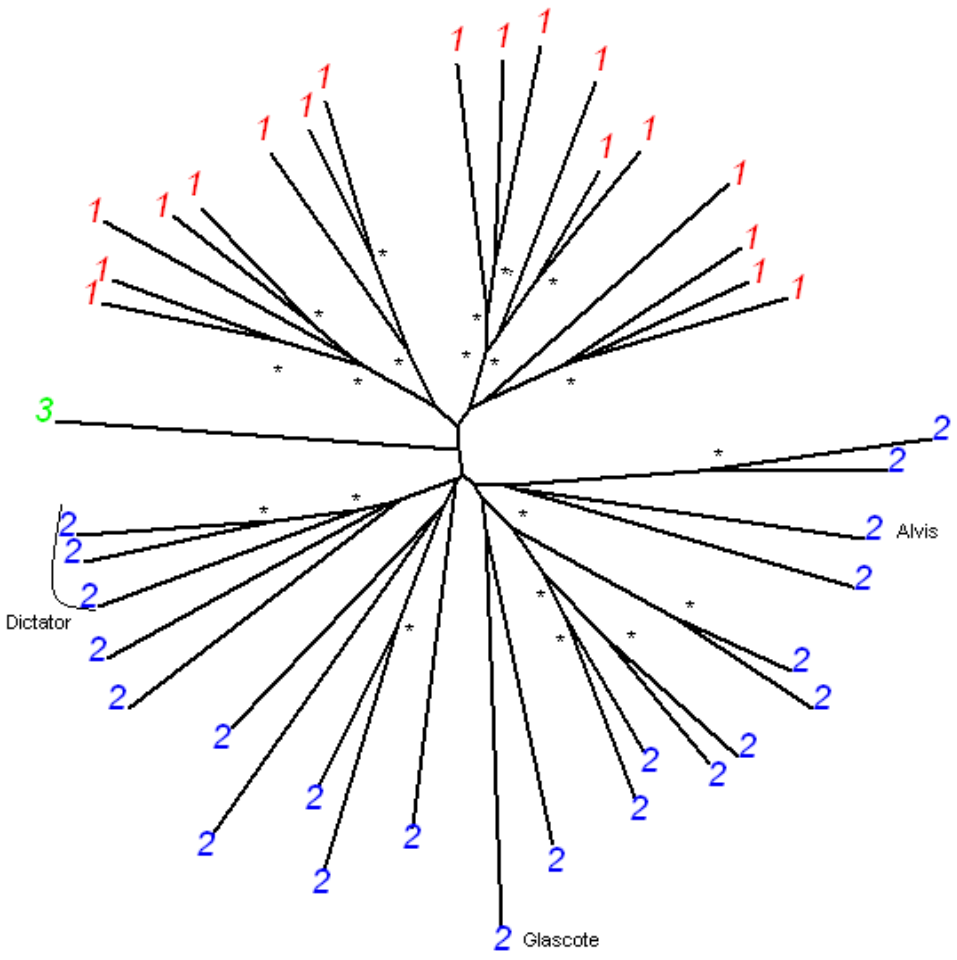


Figure 3.3 Phylogenetic reconstruction of the relationships between the British Saddleback individuals allele-sharing distances (38 microsatellite loci). Individuals are labelled according to group affiliate from the STRUCTURE results at K = 2. Bootstrap support greater than 50% are shown with an asterisk. Positions of the individuals of 'pure' Essex pig origin are indicated.

Henceforth, the two subpopulations of the British Saddleback breed, as defined by the individual-based clustering analysis using the 38 microsatellite loci dataset, were simply labelled as Subpopulation 1 and Subpopulation 2 and comprised of 18 and 22 assigned individuals, respectively. The estimated level of genetic differentiation (F_{ST}) between the two genetic subpopulations was 0.084.

Within-population estimates of genetic diversity for the whole British Saddleback breed, Subpopulation 1 and Subpopulation 2 are presented in Table 3.1. Although similar levels of observed and expected heterozygosity were found in the British Saddleback and both subpopulations, the number of alleles found in Subpopulation 1 was noticeably lower than in Subpopulation 2. Two loci deviated from HWE proportions in the British Saddleback breed.

Table 3.1 Population genetic descriptors for the British Saddleback breed, Subpopulation 1 and Subpopulation 2. H_E, H_O, N and HWE are the expected heterozygosity, observed heterozygosity, average number of alleles per locus and the number of loci that deviated from HWE after Bonferroni correction, respectively.

	H_E	H_O	N	HWE
British Saddleback	0.63	0.59	5.79	2
Subpopulation 1	0.57	0.58	3.84	0
Subpopulation 2	0.62	0.60	5.40	1

The estimated coefficient of individual inbreeding (F) averaged for the British Saddleback breed was 5.15% (s.d. 2.99). The frequency distributions of F for Subpopulation 1 and 2 are presented in Figure 3.4. Individuals from the Subpopulation 1 had significantly higher levels of F (mean F = 8.01%, s.d. 1.33) than

individuals from Subpopulation 2 (mean F = 2.60%, s.d. 1.22, Mann-Whitney U test, $p < 0.05$).

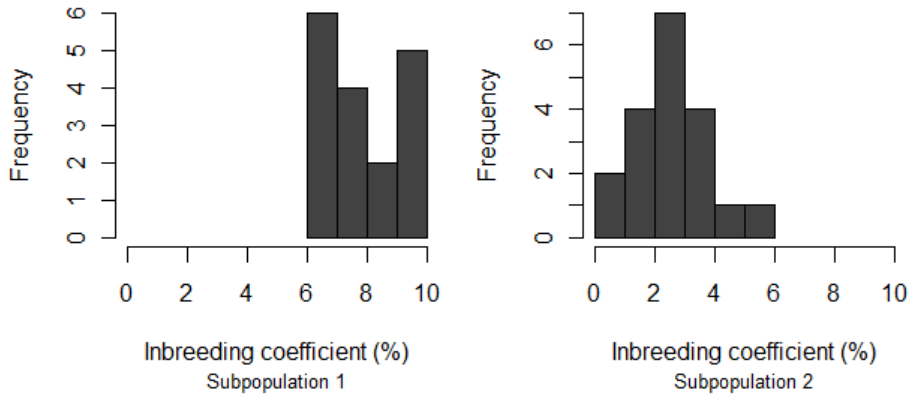


Figure 3.4 Frequency distribution of the individual inbreeding coefficients for Subpopulation 1 and 2.

Individual multilocus heterozygosity (MLH) ranged from 0.39 to 0.71 in Subpopulation 1 and were not significantly lower than that observed in Subpopulation 2 (range 0.45 to 0.74, Mann-Whitney U test, $p = 0.35$). Non-significant correlations were observed between MLH and F for both Subpopulations (Spearman's rank correlation = -0.008, $p = 0.97$ and Spearman's rank correlation = 0.24, $p = 0.3199$, for Subpopulation 1 and 2, respectively).

Information on the ancestry origin, herd prefix and bloodline, of the individuals was considered in light of the inferred genetic subdivision of the British Saddleback individuals (Table 3.2). Although Subpopulation 1 consisted predominantly of bloodlines of Wessex Saddleback ancestry, Subpopulation 2 contained a mixture of both Essex pig and Wessex bloodline ancestry. To recall, for five individuals the dam

Table 3.2 British Saddleback individuals from the two subpopulations, listed with the herd prefix and the ancestral bloodlines of the sire and dam, Essex pig (E) or Wessex Saddleback (W).

Subpopulation	Individual	Sire Herd	Sire Line	Dam Herd	Dam Line
1	s01	Rainbarrow	W	Lydling	E
	s02	Rainbarrow	W	Lydling	E
	s04	Rainbarrow	W	Lydling	E
	s05	Rainbarrow	W	Lydling	E
	s06	Rainbarrow	W	Lydling	E
	s17	Rainbarrow	W	Whithorn	E
	s24	Rainbarrow	W	Rainbarrow	W E
	s25	Rainbarrow	W	Rainbarrow	W E
	s26	Rainbarrow	W	Rainbarrow	W
	s27	Rainbarrow	W	Rainbarrow	W
	s28	Rainbarrow	W	Rainbarrow	W
	s29	Rainbarrow	W	Rainbarrow	W E
	s30	Rainbarrow	W	Rainbarrow	W
	s31	Rainbarrow	W	Rainbarrow	W
	s32	Rainbarrow	W	Rainbarrow	W
	s33	Rainbarrow	W	Rainbarrow	W
	s34	Rainbarrow	W	Rainbarrow	W
	s41	Rainbarrow	W	Rainbarrow	W
2	s09	Maddaford	W	Maddaford	W
	s10	Maddaford	W	Erriwig	W E
	s11	Maddaford	W	Maddaford	W
	s11A	Maddaford	W	Pencoed	E
	s12	Maddaford	W	Maddaford	W
	s13	Maddaford	W	Pencoed	E
	s14	Maddaford	W	Maddaford	W
	s15	Maddaford	W	Maddaford	W
	s18	Huntinghall	E	Penlas	E
	s21	Glascote	E	Glascote	E
	s22	Poplar	E	Oakleaf	W
	s23	Poplar	E	Oakleaf	W
	s35	Endsleigh	E	Colonyvalence	E
	s36	Endsleigh	E	Colonyvalence	E
	s37	Sedgefen	W	Endsleigh	W
	s38	Sedgefen	W	Endsleigh	W
	s39	Endsleigh	E	Colonyvalence	W
	s40	Colony Weeford	E	Sedgefen	W
	s42	No information			
	s43	Corella Park	W	Corella Park	W
	s44	Corella Park	W	Corella Park	W
	s45	Dalehead	W	Maddaford	W
Admixed ind.	s03	Winneyhill	E	Lydling	E

or sire were of alleged ‘pure’ Essex pig bloodline ancestry, and of four of these individuals, the other ancestral half was also an Essex pig bloodline. Although all five individuals were assigned to Subpopulation 2, they did not cluster together in the phylogenetic reconstruction (Fig 3.3) and were instead scattered amongst individuals of full and mixed Wessex Saddleback bloodline ancestry. It was not unexpected that 3 of the Dictator individuals would cluster together as they shared the same sire line, however, the two adjacent individuals to the these Dictators individuals had Wessex bloodlines from both dam and sire, although this part of the structure was not well supported (Fig 3.3).

The genetic subdivision of the British Saddleback breed into Subpopulation 1 and Subpopulation 2 was reflected by the herd prefix (Table 3.2). All the individuals in Subpopulation 1 had one or both parents with the herd prefix Rainbarrow. All the individuals in Subpopulation 2 had a sire or dam name of other herd prefixes.

In the Bayesian genotypic clustering analysis it was Subpopulation 2 (the larger green cluster) that was divided further in to separate groups at both K = 3 and 4. These separate clusters corresponded with herd membership.

3.4 Discussion

In chapter 2 the British Saddleback was the only traditional British pig breed that had not formed a single distinct genetic unit using certain individual-based clustering tools. Further analysis in this chapter on only the British Saddleback individual

genotypes using Bayesian genotypic clustering tools resulted in a maximum delta estimate (ΔK) at K = 2, indicating that the most likely clustering solution for the breed was two subpopulations (Fig 3.2). In addition, there was a clear division of British Saddleback individuals into two clades in the phylogenetic tree, although the bootstrap support was low (Fig 3.3). By pooling the two inferred subpopulations it was likely that a Wahlund effect was created thereby resulting in the slight heterozygote deficit observed in the overall British Saddleback breed (Table 3.1)

The initial assumption was that the two genetic subpopulations in the British Saddleback breed was a reflection of the amalgamation of the Essex pig and Wessex Saddleback back in 1967 (BPA 2002). Bloodline information might have been used to make breeding decisions. In herds of other British traditional pig breeds it has been found that some individuals derived from the same bloodline have been found to more similar to one another than to individuals of different bloodlines within the same herd (Hall 1989). However, further scrutiny of the Essex pig and Wessex Saddleback historic bloodlines indicated that that this was not the case as both subpopulations had a high proportion of Wessex Saddleback ancestry (Table 3.2). From a cursory examination of a four generation pedigree of an individual available on the British Pig Association (BPA) website it was clear that many of the ancestral bloodlines of Essex pig and Wessex Saddleback have been inter-crossed repeatedly. There have been more than 30 generations of British Saddlebacks since the amalgamation and it appears that the bloodlines of the Essex pig and Wessex Saddleback have not survived as distinct groups within the current breed population. Indeed, the assertion that certain Essex pig bloodlines are free from Wessex

Saddleback blood (EPS 2011), and hence should be genetically distinct, was not supported.

On closer examination, the pattern of genetic subdivision was instead associated with herds whereby all individuals with Rainbarrow herd ancestry clustered in Subpopulation 1 while the individuals of Subpopulation 2 belonged to other herds (Table 3.2). The Rainbarrow herd is an old herd of what was originally Wessex Saddlebacks, first registered in 1953 (BPA 2008). Like all other Wessex Saddleback herds, it has since been absorbed into the established British Saddleback breed. Yet, the individuals of the Rainbarrow herd appear to be genetically different from individuals of other British Saddleback herds.

The genetic structuring observed in the British Saddleback breed implies that genetic barriers may have been imposed between the Rainbarrow herd and the other herds, inhibiting gene flow to and from the herd and thereby affecting genetic homogenisation of the breed. Hall (1989) reported that, in comparison to other British traditional pig breeds, in the British Saddleback relatively less males and females were transferred from their herd of origin to other herds for further breeding. In addition, if a subpopulation is isolated random genetic drift also influences levels of genetic differentiation, particularly if the subpopulation is small in size (Hartl and Clark 1997). Other non-mutually exclusive factors could also have contributed to the genetic heterogeneity observed in the breed, including inbreeding. There was a stark difference between the distributions of individual inbreeding coefficients estimated from the pedigree for the Rainbarrow herd and for the other herds (Fig 3.4). Very

close inbreeding appears to have been avoided in the other herds and only a small percentage of individuals have become inbred due to a more common ancestor in recent generations. The level of inbreeding in the Rainbarrow herd was higher than that found in other British traditional pig breeds (Hall 1989). While the level of inbreeding in the Rainbarrow herd was lower than that found in other livestock breeds (e.g., Lippizan horse 10.30% (Curik et al. 2003)), the high level indicates that mating between related individuals has occurred in the herd. This does not necessarily imply that consanguineous mating was intentionally practiced. Traditional pig breeds generally have small population sizes and the sizes of their herds would be even further reduced. These breeds have been shown to possess higher levels of inbreeding in individuals whose parents both originated from the same herd compared to individuals whose parents originated from different herds (Hall 1989). In other words, if a small subpopulation is closed off from any new genetic material it may be inevitable that individuals will share ancestry due to its size. Hence, the pool of closely related individuals would be genetically different from the rest of the individuals of the total population, resulting in population genetic heterogeneity and subdivision (Hartl and Clark 1997).

The mating of individuals more closely related than by chance is also expected to impact genetic diversity by reducing genome-wide heterozygosity and increasing homozygosity (relative to HWE) over successive generations (Hartl and Clark 1997). Hence, in the absence of pedigrees, as is often the case with wild populations, it has been suggested that individual multilocus heterozygosity (MLH) could be used as a proxy for inbreeding (Slate et al. 2000). However, although the range of MLH was

slightly higher in the other herds, it was not significantly different from that observed in the Rainbarrow herd. Also, there was a weak and non-significant correlation between MLH and individual inbreeding coefficients, which was not consistent with the assumed effects of inbreeding on genetic diversity. Similar results have been observed in other studies and the poor prediction of inbreeding from levels of heterozygosity of individuals has been further discussed (Pemberton 2004; Pemberton 2008; Slate et al. 2004). Moreover, the average heterozygosity of the Rainbarrow herd was similar to that of the other herds and European pig breeds (Table 3.1) (SanCristobal et al. 2006a). Even if a high proportion of individuals have become inbred in the Rainbarrow herd (Fig 3.4), it does not seem to have impacted the heterozygosity levels within the herd.

Although heterozygosity was high, there was a reduced allelic diversity in the Rainbarrow herd relative to other herds (Table 3.1). The contrasting levels of allelic diversity and heterozygosity in the Rainbarrow herd were consistent with simulation studies on the effects of a bottleneck on within-population diversity. During a bottleneck rare alleles, which have little effect on heterozygosity, are lost faster than heterozygosity and, as a consequence, there is a greater reduction in allelic diversity than heterozygosity (Allendorf 1986). The duration of a bottleneck and the size of a population also impact the severity of allelic reduction. The effects of a bottleneck on population diversity was demonstrated in a founding population of captive vultures where there was no significant reduction in heterozygosity over successive generations, but a loss in allelic diversity with fewer alleles per microsatellite locus (Gautschi et al. 2003). However, since the British Saddleback has not been sampled

at various times points there was no direct evidence of a temporal reduction in allelic diversity due to bottleneck. Nonetheless, the low allelic diversity in the Rainbarrow herd may impact the long-term viability of the subpopulation because the limit of selection response is determined by the number of alleles (Hill and Rasbash 1986).

3.5 Conclusion

Unlike the other traditional British pig breeds there appears to be subtle substructure within the British Saddleback population. The inferred genetic substructure in the breed did not reflect the historic creation of the breed from the Essex pig and Wessex Saddleback. Instead, one of the herds of the British Saddleback appeared to be genetically differentiated from the other herds. The herd in question also had relatively low allelic diversity and a high level of inbreeding, suggesting population isolation and the absence of genetic introgression. Certain breeding decisions that have been made for the Rainbarrow herd over the last few decades may have had an influence on the inbreeding rate and the allelic diversity of the subpopulation.

CHAPTER FOUR

Genetic characterisation of British traditional chicken breeds

4.1 Introduction

A recent review on farm animal genetic resources (FAnGR) in Britain detailed prior efforts undertaken to characterise the molecular biodiversity of breeds with the goal to conserve these resources (DEFRA 2009). Researchers have assessed British FAnGR in a variety of ways: participation in European-wide initiatives (e.g. cattle, goats, sheep, pigs) (Canon et al. 2006; Laloë et al. 2010; Lawson Handley et al. 2007; SanCristobal et al. 2006a), genetic analysis of a number of local breed populations (e.g. cattle) (Wiener et al. 2004) and characterisation of high priority breeds (e.g. Dexter and Jersey cattle breeds) (Bray et al. 2009; Chikhi et al. 2004). In comparison to other farm animal species, however, there has been a noticeable lack of attention directed towards characterising Britain's poultry biodiversity, with only one British chicken breed genotyped in a European-wide AVIANDIV chicken breed biodiversity study (DEFRA 2009; Hillel et al. 2003). This paucity of effort is in stark contrast to numerous studies that have characterised the genetic diversity and structure of other European chicken breeds, from extensive analyses on widely sampled sets of commercial and non-commercial chicken breeds (Granevitze et al. 2007; Granevitze et al. 2009; Hillel et al. 2003) to small-scale genetic biodiversity studies on local chicken populations (e.g. Finnish, French, Hungarian and Italian chicken breed studies) (Vanhala et al. 1998; Berthouly et al. 2008; Bodzsar et al. 2009; Zanetti et al. 2011).

The lack of any previous efforts to characterise British poultry genetic resources is surprising considering the country's history of poultry breeding. Britain has a large

number of chicken breeds due to a long tradition of breed development and exhibition of fancy breeds (Hams 2004; Roberts 1997). This reservoir of variability is a combination of old indigenous breeds, foreign introduced breeds and more recently developed breeds (Table 4.1). The immense diversity of UK chicken breeds can be observed through the many sizes, shapes and colours of morphological characteristics including: body, comb, feathers, skin, tails and feet (Roberts 1997). The large pool of British chicken breeds is potentially a major source of diversity and, considering the absence of allelic diversity among commercial chickens due to bottlenecking early in the development of the lines (Muir et al. 2008), it is imperative that an assessment of British poultry genetic resources is undertaken.

The objective of this study was to determine the levels of genetic diversity within breeds, the genetic structure of breeds and the levels of admixture in the British chicken breeds using the Food and Agriculture Organisation (FAO) recommended microsatellite marker panel. The characterisation of the state of the genetic resources of British chicken breeds will, first, inform management initiatives and help set priorities for conservation, and, second, contribute to the European-wide perspective on FAnGR.

4.2 Materials and methods

4.2.1 Data, DNA extraction and microsatellite genotyping

Owners of flocks of known provenance were approached and asked to provide up to 12 fertile hatching eggs. Eggs were mailed to the Roslin Institute and incubated for 7 days at which time the embryos were harvested. The embryo was divided and the

Table 4.1 Summary details of 24 chicken breeds. [1] sample size; [2] number of flocks contributing to sample size; [3] number of morphological types present in sample size; [4] information from (Hams 2004; Roberts 1994; Roberts 1997; Vorwald Dohner 2001) ; [5] population size status measured by the number of breeding females as follows: Critical = 100, Endangered = 200, Vulnerable = 300, At Risk = 500 (DEFRA 2010). (The adopted quantification was previously used by RBST to categorise breed status.)

Breed	N[1]	Flocks[2]	Types[3]	Origin[4]	Introduction[4]	Suspected development origins[4]	Status[5]
Appenzeller	30	9	3	Switzerland	1970s		Endangered
Araucana	29	7	3	Chile	1930s		
Brahma	25	6	3	China/USA	1800s		At Risk
Buff Orpington	26	8	1	Britain		Buff Cochin, Dorking, Hamburg, Langshan; 1800s	At Risk
Cochin	29	6	2	China	1800s		At Risk
Croad Langshan	28	8	1	China	1800s		
Derbyshire Redcap	30	9	1	Britain		Old indigenous	
Dorking	26	7	3	Britain		Old indigenous	
Hamburg	30	10	3	Britain		Old indigenous	
Indian Game	26	8	3	Britain		Asil, Old English Game, Malay; 1800s	Endangered
Ixworth	27	6	1	Britain		Indian Game, White Sussex, Wyandotte; 1930s	Critical
Leghorn (coloured)	29	7	4	Italy/USA	1800s		
Lincolnshire Buff	30	6	1	Britain		Orpington Buff, Dorking, Cochin; 1930s	Critical
Maran	30	6	2	France	Early 1900s		
Marsh Daisy	28	8	1	Britain		Old English Game, Malay, Hamburg; 1930s	Endangered
Norfolk Grey	19	4	1	Britain		Birchen Game, Silver Duckwing, Leghorn; 1930s	Critical
Old English Pheasant Fowl	28	9	1	Britain		Old indigenous	Endangered
Rhode Island Red	30	7	1	USA	1900s	Shanghai, Malay, Java, Brown Leghorn	
Scots Dumpy	28	6	1	Britain		Old indigenous	
Scots Grey	29	6	1	Britain		Old indigenous	At Risk
Silkie	29	6	3	Asia	1600s		
Spanish	28	7	1	Spain	1700s		Critical
Sussex Light	30	6	1	Britain		southern England	
Sussex	30	6	3	Britain		southern England	

head and body were stored separately at -80^{o}C. Data collection took place over a period of 3 years (2007-2009). The DNA was extracted from the thawed heads using the 'Maxwell 16' system. The number of flocks per breed was limited to a minimum of four and the number of samples per flock to a maximum of six for genetic analysis. Using these sampling criteria a total of 24 breeds were included in the study, with 19 to 30 individuals from 4 to 10 flocks (Table 4.1).

Individual samples were genotyped at 30 microsatellite loci (Table 4.2).

Table 4.2 The microsatellite loci. [1] total number of alleles; [2] average observed heterozygote frequency; [3] average expected heterozygote frequency.

	Locus	Chromosome	N[1]	H_O[2]	H_E[3]	Missing (%)
1	ADL0268	1	6	0.50	0.74	0.6
2	MCW0020	1	4	0.47	0.73	5.6
3	MCW0111	1	6	0.46	0.70	0.3
4	MCW0248	1	5	0.21	0.33	0.9
5	LEI0234	2	19	0.58	0.91	0.7
6	MCW0034	2	11	0.55	0.85	2.5
7	MCW0206	2	8	0.38	0.64	1.2
8	LEI0166	3	7	0.33	0.57	1.8
9	MCW0016	3	7	0.39	0.69	0.4
10	MCW0037	3	4	0.35	0.60	0.1
11	MCW0103	3	2	0.37	0.42	1.2
12	MCW0222	3	4	0.32	0.44	2.2
13	LEI0094	4	18	0.50	0.80	0.1
14	MCW0078	5	5	0.42	0.67	1.0
15	MCW0081	5	10	0.33	0.73	1.3
16	MCW0098	4	2	0.22	0.45	1.3
17	MCW0284	4	2	0.34	0.50	2.8
18	MCW0295	4	9	0.47	0.77	0.1
19	LEI0192	6	24	0.54	0.84	1.9
20	MCW0014	6	7	0.33	0.56	2.2
21	MCW0183	7	10	0.43	0.71	2.2
22	ADL0278	8	9	0.46	0.74	1.8
23	ADL0112	10	4	0.39	0.58	2.7
24	MCW0067	10	5	0.34	0.65	1.0
25	MCW0104	13	14	0.41	0.62	1.9
26	MCW0216	13	7	0.37	0.64	1.5
27	MCW0123	14	9	0.40	0.73	6.2
28	MCW0330	17	7	0.36	0.67	1.3
29	MCW0165	23	3	0.29	0.57	1.6
30	MCW0069	26	11	0.41	0.72	1.3

Further information on the primer sequences, annealing temperatures and Genbank accession numbers of the microsatellite loci can be found at http://aviandiv.tzv.fal.de/primer_table.html. Genotyping was conducted using the Qiagen PCR Mix Kit (p/n 206145) and the resulting products were visualised on an Applied Biosystems 3730xl genetic analyzer (Applied Biosystems/Hitachi, Applera, USA). Genemapper Software v3.5 (Applied Biosystems, Applera, USA) was used to estimate fragment sizes by comparing them to an internal size standard (LIZ500, Applied Biosystems).

4.2.2 Marker polymorphism, within and among population diversity

The genetic diversity within breeds was measured using standard population genetic diversity estimators (for further details see section 3.2.2 in chapter 3). In brief, the total number of alleles, observed heterozygosity (H_O) and expected heterozygosity (H_E) were estimated using FSTAT 2.9.3 (Goudet 1995).

Tests for deviations from Hardy-Weinberg Equilibrium (HWE) for each breed-locus combination and across all loci for each population ('global test') were performed using GENEPOP 4.0.7 (Rousset 2008) (for further details see section 2.2.1 in chapter 2). Weir & Cockerham's (1984) inbreeding coefficient (F_{IS}) was estimated for each population using FSTAT 2.9.3 (Goudet 1995). F_{IS} measures the difference between H_O and H_E relative to the magnitude of H_E. It quantifies the reduction in heterozygosity in a population relative to a random-mating population of the same allele frequencies. Positive values of F_{IS} indicate an excess of homozygotes and a deficit of heterozygotes relative to HWE.

The extent of genotypic linkage disequilibrium (LD) between pairs of loci in each breed was tested by performing probability tests using GENEPOP 4.0.7 (Rousset 2008) (for further details see section 2.2.1 in chapter 2). Significance levels of the tests were adjusted for multiple comparison following standard Bonferroni corrections (Rice 1989).

Genetic differentiation of populations was measured by estimating Weir and Cockerham's F_{ST} (Weir and Cockerham 1984) using GENEPOP 4.0.7 (Rousset 2008) (for further details see section 2.2.5 in chapter 2). Levels of population differentiation were also estimated using Reynold's genetic distance. It is based on the coancestry coefficient, which is the probability that a random pair of genes at the same locus within a randomly chosen population were identical by descent (Reynolds et al. 1983). The genetic distance assumes that differences between populations arise due to genetic drift and that there is no mutation, but by incorporating the effective population size in the calculation it does not assume that population sizes have remained constant and equal in all populations. Thus, it takes into account the varying influence of random genetic drift when it comes to different population sizes. Reynold's genetic distance measure is based on Wright's F_{ST}, thus it assumes that populations have diverged due to drift alone. It is also suggested that it is appropriate for analysis of populations sampled from a single species because it also reflects the amount of gene flow between populations (Slatkin & Maddison 1990). Reynold's genetic distance was calculated between pairs of breeds from the allele frequencies using GENDIST (Phylip v 3.67) (Felsenstein 1989). An unrooted neighbour-joining tree was constructed from the genetic distance matrix using the R

package APE (Paradis et al. 2004 (for further details see section 2.2.4 in chapter 2). A total of 1000 bootstrap replicates were created in SEQBOOT, for each replicate Reynold's genetic distance was calculated between pairs of breeds in GENDIST and a consensus cladogram was calculated using CONSENSE (Phylip v 3.67) (Felsenstein 2008) (for further details see section 2.2.4 in chapter 2).

4.2.3 Clustering of individuals to populations

Population structure and admixture was investigated using the Bayesian genotypic clustering method implemented in BAPS 5.2 (Corander et al. 2008). This Bayesian genotypic clustering method was chosen over others because the results obtained in chapter 2 indicated that in comparison to the other methods BAPS was more effective at delineating the breeds and detected finer genetic differentiation within the breeds with known structuring. BAPS 5.2 uses a 'greedy' stochastic optimization algorithm to probabilistically assign individual multilocus genotypes to clusters and infer the number of genetically distinguishable populations (K) (for further details see section 2.2.2 in chapter 2). BAPS 5.2 was implemented for $2 \leq K \leq 35$ with 10 runs at each K value. The phylogenetic relationships amongst the individuals were reconstructed by estimating the proportion of shared alleles (for further details see section 2.2.4 in chapter 2).

4.2.4 Breed genetic contributions

A breed's contribution to genetic diversity was quantified following Ollivier and Foulley (Ollivier and Foulley 2005), a commonly implemented method in livestock

breed conservation analyses (e.g. pig breeds) (Laval et al. 2000). The contribution to between-breed diversity (CB), using the Weitzmann method (Weitzmann, 1992), was estimated from the Reynold's pairwise genetic distance matrix produced in GENDIST (Phylip v 3.67) (Felsenstein 2008). CB was estimated as follows: CB = 1 – V(S/*k*) / V(S) where V(S) is the total genetic distance of the whole set of breeds considered and V(S/*k*) is the total genetic distance of the set excluding breed *k*. The relative decrease in genetic distance with the removal of a breed (V_k) is the contribution of breed *k* to between-breed diversity and can be seen as a measure of the genetic originality of a breed. The contribution to within-breed diversity (CW) was estimated using average expected heterozygosity (H_E) values of breeds as CW = 1 – H_E(S/*k*) / H_E(S), where H_E(S) is the average expected heterozygosity of all the breeds and H_E(S/*k*) is the average expected heterozygosity excluding breed *k*. The relative increase or decrease in average expected heterozygosity with the removal of a breed ($H_{E(k)}$) is the contribution of breed *k* to total within-breed diversity. Aggregate diversity (D) was estimated as D = F_{ST}CB + (1- F_{ST})CW (Ollivier and Foulley 2005).

4.3 Results

4.3.1 Individual multilocus genotype data and quality cleaning

One or more loci failed to amplify in 202 out of the 705 samples subjected to PCR amplification (Fig 4.1). Every breed had at least one individual that failed to amplify at one or more loci. Total amplification failure across all 30 microsatellite loci occurred in all 4 individuals of White Dorking. Individuals with > 20% missing data (*n* = 20) were discarded (Fig 4.1). Individuals that did not conform to the breed

description, but were nonetheless included in the genotyping sample were discarded (n = 7): Buff Orpington had 5 individuals of the 'Splash' and 'Speckled' variety and Indian Game had 2 individuals of the Old English Game breed. Preliminary individual-based clustering analysis revealed that a number of individuals did not cluster to breed origin. Individuals were traced back to their supplier and if the supplier also provided samples of the breed to which these individuals clustered with, these individuals were removed from the dataset (n = 4).

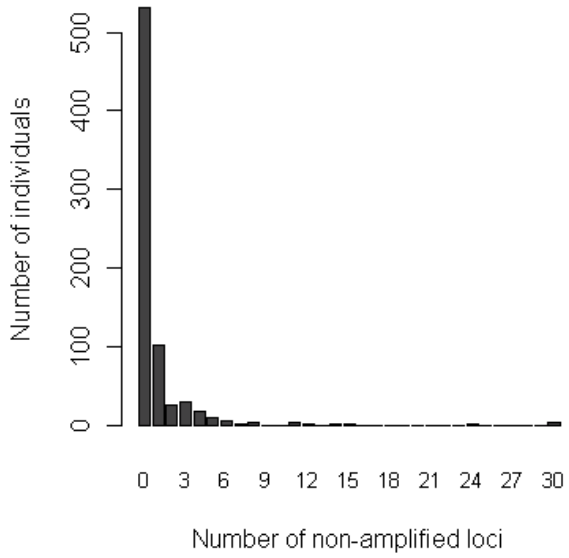


Figure 4.1 Barplot of the number of failed locus amplification per individual.

After data cleaning the total sample size was 674 individual genotypes across the 24 breeds with a range of 19 to 30 sampled individuals per breed (Table 4.1). After data cleaning the proportion of missing data per locus ranged from 0.1% (MCW0037, LEI0094 and MCW0295) to 6.2% (MCW0123) (Table 4.2).

4.3.2 Genetic diversity within and among breeds

A total of 239 alleles were found across the 30 microsatellite loci with a mean number of 7.97 alleles per locus (Table 4.2). The number of alleles per locus ranged from 2 (MCW0103, MCW0284, MCW0098) to 24 (LEI092).

The average number of alleles per locus in different breeds ranged from 2.00 in Spanish to 4.40 in Maran, with an average across all the breeds of 3.59 alleles per locus (Table 4.3). Not all markers were found to be polymorphic in each of the 24 breeds; the number of monomorphic loci per breed ranged from 0 to 10 (Table 4.3).

A total of 33 breed-specific or private alleles were detected in 15 breeds (Table 4.3). Twenty-two of the private alleles possessed frequencies < 0.1%. The remaining 11 breed-specific private alleles were found in Brahma (1), Buff Orpington (1), Cochin (1), Croad Langshan (1), Indian Game (1), Leghorn (1), Lincolnshire Buff (1), Marsh Daisy (2), Silkie (1) and Spanish (1). The average expected heterozygosity (H_E) over all loci ranged from 0.20 in Spanish to 0.62 in Araucana, while the average observed heterozygosity (H_O) varied from 0.15 in Spanish to 0.49 in Cochin (Table 4.3). Average estimates of H_E and H_O over all loci and breeds were 0.49 and 0.39, respectively.

There was a deficiency of heterozygotes compared with Hardy Weinberg Equilibrium (HWE) expectations for each breed. After Bonferroni correction, 48 out

of the 720 locus-breed comparisons revealed significant ($P < 0.00007$) departures from HWE.

Table 4.3 Population genetic diversity estimates for 24 chicken breeds. [1] total number of alleles; [2] average number of alleles per locus; [3] average observed heterozygote frequency; [4] average expected heterozygote frequency; [5] number of monomorphic loci; [6] number of breed-specific private alleles; [7] number of loci deviating from Hardy-Weinberg Equilibrium after Bonferroni correction; [8] number of significant tests of linkage disequilibrium (after Bonferroni correction) out of 435 possible pairs of loci, number in brackets indicate tests that occurred between loci pairs found on the same chromosome.

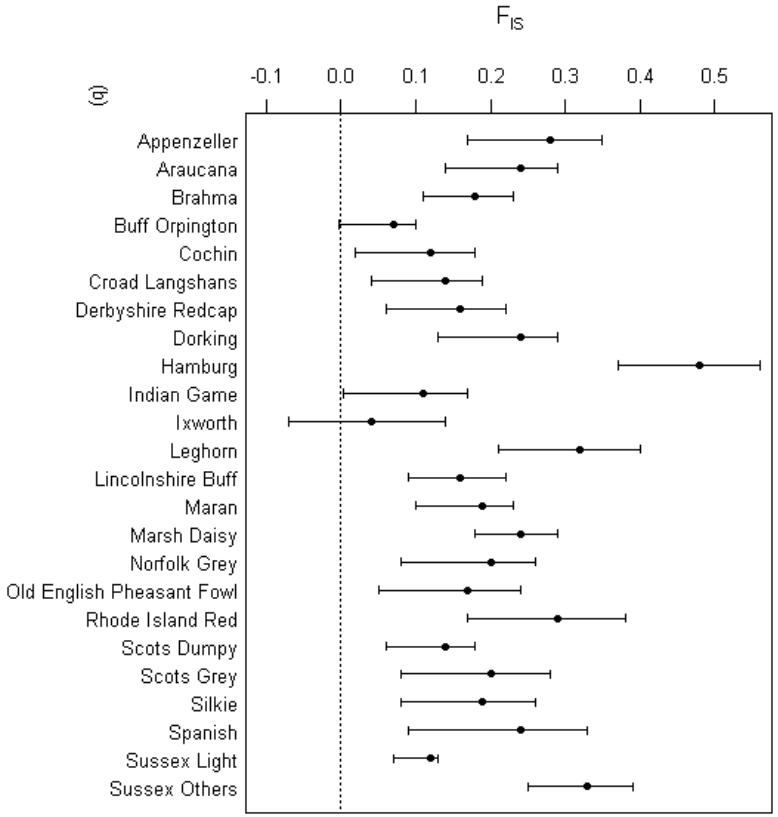
	Breed	N[1]	*n*[2]	H_O[3]	H_E[4]	M[5]	PA[6]	HWE[7]	LD[8]
1	Appenzeller	94	3.13	0.31	0.43	2	0	3	0
2	Araucana	129	4.30	0.47	0.62	0	3	1	7
3	Brahma	124	4.13	0.43	0.53	0	4	1	4
4	Buff Orpington	118	3.93	0.51	0.55	0	1	0	1
5	Cochin	124	4.13	0.49	0.56	0	3	0	3
6	Croad Langshans	107	3.57	0.42	0.49	1	3	1	1 (1)
7	Derbyshire Redcap	94	3.13	0.35	0.42	1	0	3	1 (1)
8	Dorking	107	3.57	0.36	0.47	2	0	1	4
9	Hamburg	98	2.97	0.22	0.42	4	0	8	5
10	Indian Game	100	3.33	0.44	0.49	0	1	1	0
11	Ixworth	90	3.00	0.41	0.46	0	0	0	1
12	Leghorn	119	3.97	0.37	0.54	0	2	6	43 (2)
13	Lincolnshire Buff	105	3.50	0.41	0.48	1	2	0	1
14	Maran	132	4.40	0.47	0.58	0	3	4	3
15	Marsh Daisy	96	3.20	0.37	0.49	1	3	3	2 (1)
16	Norfolk Grey	89	2.97	0.40	0.50	1	2	1	1 (1)
17	Old English Pheasant Fowl	105	3.50	0.40	0.48	1	0	0	1
18	Rhode Island Red	121	4.03	0.41	0.57	0	0	5	13
19	Scots Dumpy	126	4.20	0.44	0.51	0	1	1	2
20	Scots Grey	95	3.17	0.32	0.40	2	0	0	2
21	Silkie	115	3.83	0.45	0.56	0	1	0	14 (1)
22	Spanish	60	2.00	0.15	0.20	10	1	0	0
23	Sussex Light	121	4.03	0.48	0.54	0	0	1	5
24	Sussex	117	3.90	0.38	0.57	0	3	8	34 (4)
	Average		3.59	0.39	0.49				

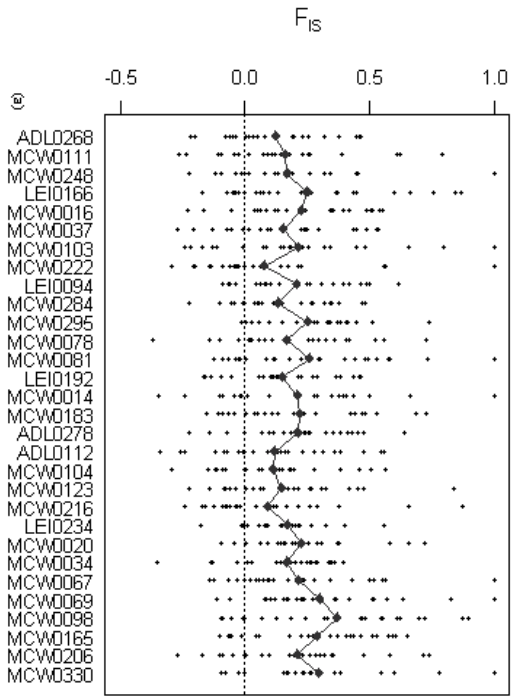
Across loci, the number of significant deviations from HWE proportions ranged from none in 7 loci (MCW0248, MCW0037, MCW0222, MCW0284, MCW0067, MCW0123 and MCW0098) to a maximum of 4 deviations in 4 loci (MCW0183, ADL0278, MCW0165 and MCW0330). Across breeds, the number of significant deviations from HWE proportions ranged from none in 8 breeds (Buff Orpington, Cochin, Ixworth, Lincolnshire Buff, Old English Pheasant Fowl, Scots Grey, Silkie and Spanish) to a maximum of 8 deviations in 2 breeds (Hamburg and Sussex) (Table 4.3). The global test for deviations from HWE indicated there was a large deficiency of heterozygotes compared with HWE expectations for 22 of the British traditional chicken breeds, with only Buff Orpington and Ixworth meeting HWE proportions ($p > 0.05$).

After Bonferroni correction, significant genotypic linkage disequlibrium (LD) was found in 148 out of 10, 440 pairs of loci. The number of locus-pairs with significant LD ranged from none in 3 breeds (Appenzeller, Indian Game and Spanish) to 43 occurrences in Leghorn (Table 4.3). Of the significant cases, 137 involved pairs of loci situated on different chromosomes. The 11 significant cases of pairs of loci on the same chromosome occurred in Croad Langshan (1), Derbyshire Redcap (1), Leghorn (2), Marsh Daisy (1) and Norfolk Grey (1), Silkie (1) and Sussex (4) (Table 4.3). The presence of related individuals in population samples could influence LD. Queller and Goodnights (1986) pairwise relatedness between individuals within breeds was calculated. The average pairwise relatedness between individuals for every breed was less than 0, except for Lincolnshire Buff (0.01), Old English

Pheasant Fowl (0.007), Scots Grey (0.02) and Spanish (0.07). Low levels of genetic diversity, not high levels of LD, were observed in these three breeds.

The presence of inbreeding (F_{IS}) was suggested by the difference between H_E and H_O and an average F_{IS} across all loci and breeds was estimated at 0.20. The loci exhibited positive average F_{IS} estimates ranging from 0.078 (locus MCW0222) to 0.37 (locus MCW0098) and there was large variation in F_{IS} within loci (Fig 4.2a). The two loci with limited number of negative within-breed F_{IS} values ($-0.01 < F_{IS} < 1$) had a low amount of missing data (MCW0069 and MCW0295 1.3% and 0.1%, respectively), suggesting that this was not an artefact of missing data. Bootstrapping across loci confirmed significant positive F_{IS} estimates in all but two breeds (F_{IS} range in Buff Orpington and Ixworth was -0.003 to 0.09 and -0.07 to 0.14, respectively, Fig 4.2b), ranging from 0.12 in Light Sussex to 0.48 in Hamburg (Fig 4.2b). Further F_{IS} analysis was conducted on five breeds (Hamburg, Leghorn, Maran, Rhode Island Red and Sussex) that exhibited extensive within-breed genetic substructure (see results on individual clustering analysis later in this section, Fig 4.5, 4.6), deviations from HWE and LD (Table 4.3). For each breed, the largest genetic subgroup defined by Bayesian genotypic clustering analysis was extracted (K = 30, Fig 4.5). Re-estimation of H_E and H_O showed a decrease in positive F_{IS} and deviations from HWE (no loci deviated from HWE) within the largest subgroup of Sussex (Speckled), Maran and Rhode Island Red (Fig 4.2c).


(a)
F_{IS}
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
Appenzeller
Araucana
Brahma
Buff Orpington
Cochin
Croad Langshans
Derbyshire Redcap
Dorking
Hamburg
Indian Game
Ixworth
Leghorn
Lincolnshire Buff
Maran
Marsh Daisy
Norfolk Grey
Old English Pheasant Fowl
Rhode Island Red
Scots Dumpy
Scots Grey
Silkie
Spanish
Sussex Light
Sussex Others


(b)
F_{IS}
-0.5
0.0
0.5
1.0
ADL0268
MCW0111
MCW0248
LEI0166
MCW0016
MCW0037
MCW0103
MCW0222
LEI0094
MCW0284
MCW0295
MCW0078
MCW0081
LEI0192
MCW0014
MCW0183
ADL0278
ADL0112
MCW0104
MCW0123
MCW0216
LEI0234
MCW0020
MCW0034
MCW0067
MCW0069
MCW0098
MCW0165
MCW0206
MCW0330

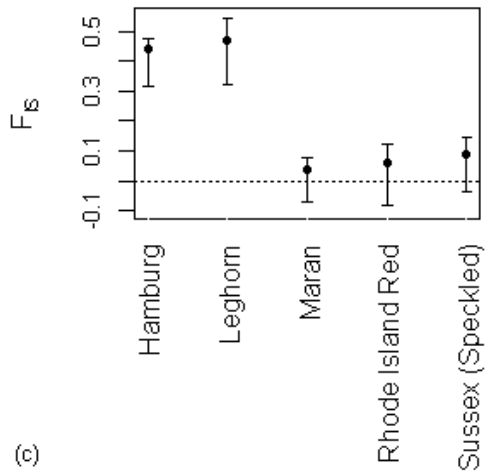


Figure 4.2 Inbreeding coefficient (F_{IS}) estimated for (a) each breed x locus combination and microsatellite locus, (b) breed average and 95% confidence intervals and (c) mean and 95% confidence intervals for breeds that exhibited genetic subdivision.

The genetic differentiation (F_{ST}) between pairs of breeds ranged from 0.10 (Rhode Island Red vs Sussex and Brahma vs Cochin) to 0.52 (Ixworth vs Spanish), with average F_{ST} across breeds ranging from 0.18 for Araucana to 0.42 for Spanish. Overall breed genetic differentiation (F_{ST}) was 0.24 (Table 4.4). Reynold's pairwise genetic distance ranged from 0.34 between Brahma and Cochin to 0.72 between Lincolnshire Buff and Spanish. Average pairwise genetic distance across all breeds ranged from 0.44 for Araucana to 0.66 for Spanish and was positively correlated with average breed F_{ST} (Spearman's rank correlation = 0.99, $p \leq 0.05$) (Table 4.4). The high correlation was not surprising as Reynold's genetic distance is based on Wright's F_{ST}.

Table 4.4 Population genetic differentiation among 24 chicken breeds. Upper diagonal contains the estimates of pairwise genetic distance between breeds estimated by Reynold's genetic distance. Lower diagonal contains the pairwise genetic differentiation between breeds.

	Breed	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Reynolds	F_{ST}
1	Appenzeller		0.44	0.60	0.58	0.58	0.60	0.52	0.46	0.48	0.57	0.61	0.43	0.62	0.56	0.52	0.53	0.47	0.53	0.53	0.55	0.55	0.69	0.57	0.52	0.54	0.28
2	Araucana	0.17		0.43	0.46	0.42	0.46	0.48	0.41	0.45	0.45	0.47	0.35	0.51	0.40	0.38	0.41	0.43	0.38	0.41	0.46	0.46	0.61	0.42	0.36	0.44	0.18
3	Brahma	0.35	0.17		0.45	0.34	0.46	0.59	0.55	0.60	0.56	0.56	0.52	0.52	0.45	0.53	0.51	0.57	0.44	0.49	0.60	0.46	0.70	0.48	0.46	0.52	0.26
4	Buff Orpington	0.33	0.19	0.18		0.41	0.48	0.57	0.52	0.57	0.55	0.55	0.50	0.45	0.43	0.52	0.52	0.55	0.42	0.52	0.59	0.50	0.70	0.46	0.43	0.51	0.25
5	Cochin	0.32	0.16	0.10	0.15		0.42	0.58	0.53	0.57	0.54	0.54	0.48	0.48	0.39	0.47	0.50	0.56	0.41	0.46	0.57	0.46	0.68	0.42	0.43	0.49	0.23
6	Croad Langshans	0.35	0.19	0.20	0.21	0.16		0.58	0.51	0.57	0.53	0.49	0.52	0.56	0.40	0.47	0.52	0.55	0.43	0.43	0.59	0.48	0.66	0.40	0.45	0.50	0.24
7	Derbyshire Redcap	0.26	0.21	0.34	0.31	0.32	0.33		0.44	0.46	0.57	0.60	0.48	0.61	0.55	0.53	0.51	0.46	0.56	0.53	0.56	0.54	0.63	0.53	0.53	0.54	0.28
8	Dorking	0.19	0.15	0.29	0.25	0.26	0.25	0.17		0.38	0.47	0.55	0.42	0.56	0.48	0.46	0.46	0.40	0.49	0.47	0.53	0.51	0.62	0.47	0.48	0.49	0.22
9	Hamburg	0.21	0.18	0.34	0.31	0.31	0.30	0.20	0.12		0.54	0.62	0.43	0.61	0.54	0.47	0.47	0.40	0.54	0.49	0.53	0.51	0.57	0.54	0.52	0.52	0.25
10	Indian Game	0.31	0.18	0.29	0.28	0.28	0.27	0.31	0.20	0.28		0.54	0.50	0.59	0.49	0.52	0.51	0.51	0.51	0.50	0.57	0.53	0.69	0.48	0.50	0.53	0.27
11	Ixworth	0.36	0.21	0.30	0.29	0.28	0.23	0.35	0.29	0.36	0.28		0.53	0.60	0.51	0.53	0.60	0.58	0.54	0.50	0.59	0.55	0.72	0.45	0.50	0.55	0.30
12	Leghorn	0.17	0.10	0.26	0.23	0.22	0.26	0.21	0.15	0.16	0.23	0.27		0.55	0.46	0.45	0.44	0.46	0.47	0.44	0.46	0.50	0.60	0.46	0.40	0.47	0.21
13	Lincolnshire Buff	0.37	0.25	0.25	0.18	0.21	0.30	0.36	0.30	0.36	0.33	0.35	0.29		0.51	0.55	0.57	0.59	0.49	0.57	0.63	0.55	0.72	0.54	0.51	0.56	0.30
14	Maran	0.30	0.14	0.19	0.16	0.13	0.14	0.28	0.21	0.27	0.22	0.24	0.20	0.24		0.45	0.45	0.52	0.38	0.45	0.55	0.47	0.65	0.37	0.38	0.47	0.21
15	Marsh Daisy	0.25	0.13	0.27	0.26	0.21	0.20	0.27	0.19	0.21	0.25	0.27	0.18	0.29	0.18		0.47	0.49	0.44	0.44	0.50	0.49	0.63	0.45	0.43	0.49	0.22
16	Norfolk Grey	0.26	0.14	0.24	0.25	0.23	0.26	0.24	0.19	0.20	0.24	0.35	0.17	0.31	0.18	0.20		0.47	0.45	0.47	0.50	0.48	0.63	0.46	0.45	0.50	0.23
17	OEFP	0.21	0.17	0.31	0.29	0.30	0.29	0.20	0.14	0.14	0.24	0.32	0.20	0.34	0.25	0.22	0.20		0.52	0.50	0.53	0.50	0.59	0.52	0.50	0.51	0.24
18	Rhode Island Red	0.27	0.13	0.18	0.16	0.15	0.16	0.30	0.22	0.27	0.24	0.27	0.20	0.22	0.13	0.17	0.18	0.25		0.46	0.55	0.49	0.67	0.40	0.34	0.47	0.21
19	Scots Dumpy	0.27	0.15	0.23	0.25	0.20	0.17	0.27	0.20	0.21	0.23	0.24	0.17	0.31	0.18	0.18	0.20	0.23	0.19		0.51	0.47	0.62	0.40	0.43	0.48	0.22
20	Scots Grey	0.28	0.19	0.34	0.34	0.31	0.33	0.30	0.26	0.26	0.31	0.35	0.19	0.38	0.29	0.24	0.23	0.26	0.29	0.24		0.54	0.69	0.53	0.51	0.55	0.29
21	Silkie	0.29	0.19	0.20	0.23	0.19	0.21	0.28	0.24	0.24	0.26	0.29	0.23	0.28	0.21	0.23	0.21	0.24	0.22	0.20	0.28		0.67	0.46	0.47	0.50	0.24
22	Spanish	0.47	0.36	0.48	0.48	0.45	0.43	0.38	0.38	0.31	0.47	0.52	0.35	0.51	0.41	0.38	0.40	0.34	0.43	0.37	0.46	0.44		0.67	0.66	0.66	0.42
23	Light Sussex	0.32	0.16	0.21	0.19	0.16	0.14	0.26	0.20	0.28	0.21	0.18	0.20	0.28	0.12	0.19	0.19	0.26	0.15	0.14	0.26	0.19	0.43		0.36	0.47	0.21
24	Sussex	0.25	0.11	0.19	0.17	0.17	0.18	0.26	0.21	0.25	0.24	0.23	0.14	0.24	0.13	0.16	0.17	0.23	0.10	0.16	0.25	0.21	0.42	0.11		0.46	0.20

The phylogenetic reconstruction of the genetic relationships between the British chicken breeds is shown in Figure 4.3.

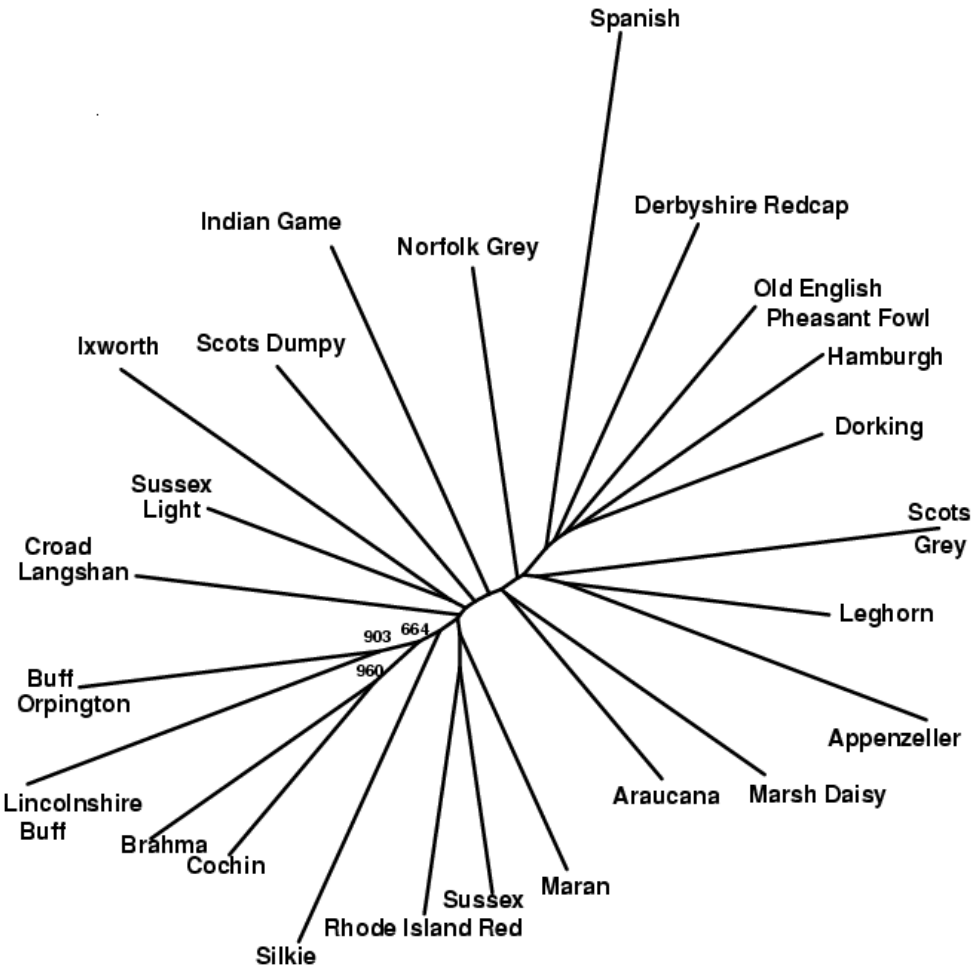


Figure 4.3 Phylogenetic reconstruction of the chicken breeds using Reynolds genetic distance. Bootstrap support values greater than 50% are indicated.

There was overall low bootstrap support of genetic relationships between the chicken breeds, with a two highly supported breed pairs: Cochin and Brahma, Lincolnshire Buff and Buff Orpington (Fig 4.3).

4.3.3 Clustering of individuals to populations

The BAPS results showed that the likelihood of the data was best described by 30 - 35 genetic clusters as can be seen, in Figure 4.4, where the log-likelihood values started to plateau from K = 30.

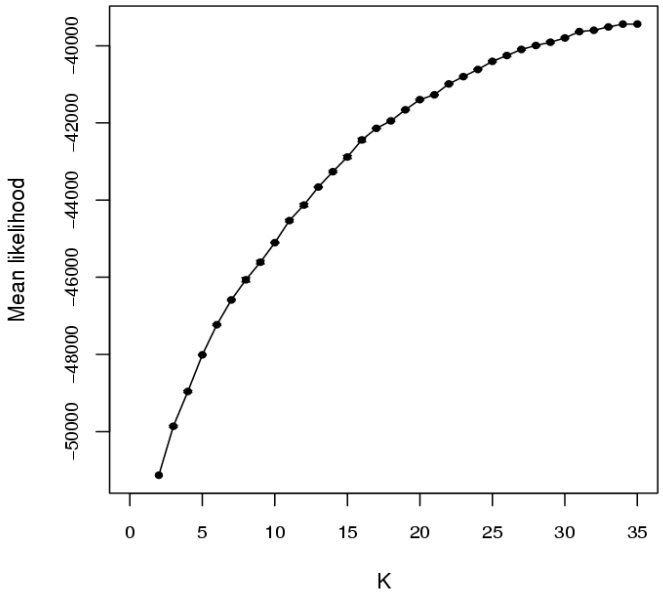


Figure 4.4 Plot of the likelihood output of BAPS with increasing K value. Points indicate the average estimated across 10 runs for each K value with standard errors.

The BAPS clustering solutions are presented in Figure 4.5. Populations that split to form independent clusters at lower K values can be interpreted as being relatively genetically distinct (Rosenberg et al. 2001). Breeds split to form their own distinct genetic clusters in the following order: Silkie, Indian Game, Scots Grey, Buff Orpington and Lincolnshire Buff, Croad Langshan, Appenzeller, Spanish, Marsh Daisy, Ixworth and Scots Dumpy at K = 6, K = 7, K = 10, K = 11, K = 12, K = 13, K = 14, K = 15, K = 16 and K = 18, respectively.

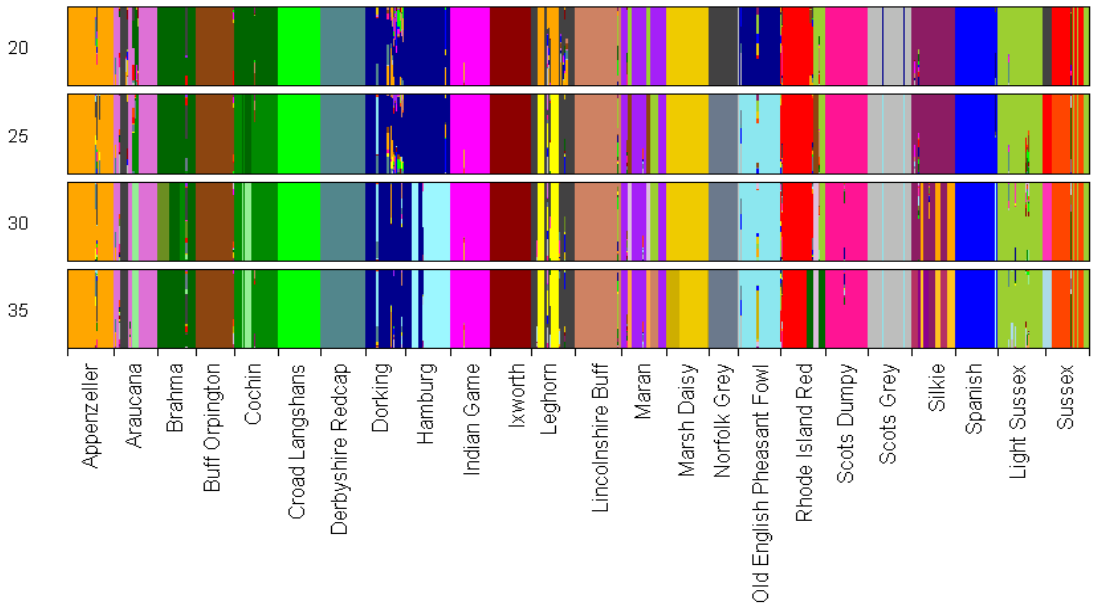


Figure 4.5 Individual assignment from Bayesian genotypic cluster analysis using BAPS at various values of K. Histograms demonstrate the proportion of each individual's genome that originated from each of 24 populations. Each individual is represented by a vertical line corresponding to its membership coefficient (q).

A number of breeds clustered together indicating genetic affinities. Brahma and Cochin formed a distinct cluster at K = 8 and remained together until K = 22. From K = 5 to 10, Buff Orpington and Lincolnshire Buff formed their own cluster, after which they split into two clusters. Derbyshire Redcap, Dorking, Hamburg and Old English Pheasant Fowl clustered together until K = 19 whereupon Derbyshire Redcap split to form an independent cluster, followed by Old English Pheasant Fowl at K = 25 and Dorking and Hamburg at K = 28. Notable genetic admixture within individuals occurred in Araucana, Dorking, Leghorn and Light Sussex breeds as evidenced by the proportion of genomes assigned to two or more clusters ($0.2 < q < 0.8$) at various K values.

The phylogenetic reconstruction of 674 individuals using the proportion of shared allele distance measure is shown in Figure 4.6.

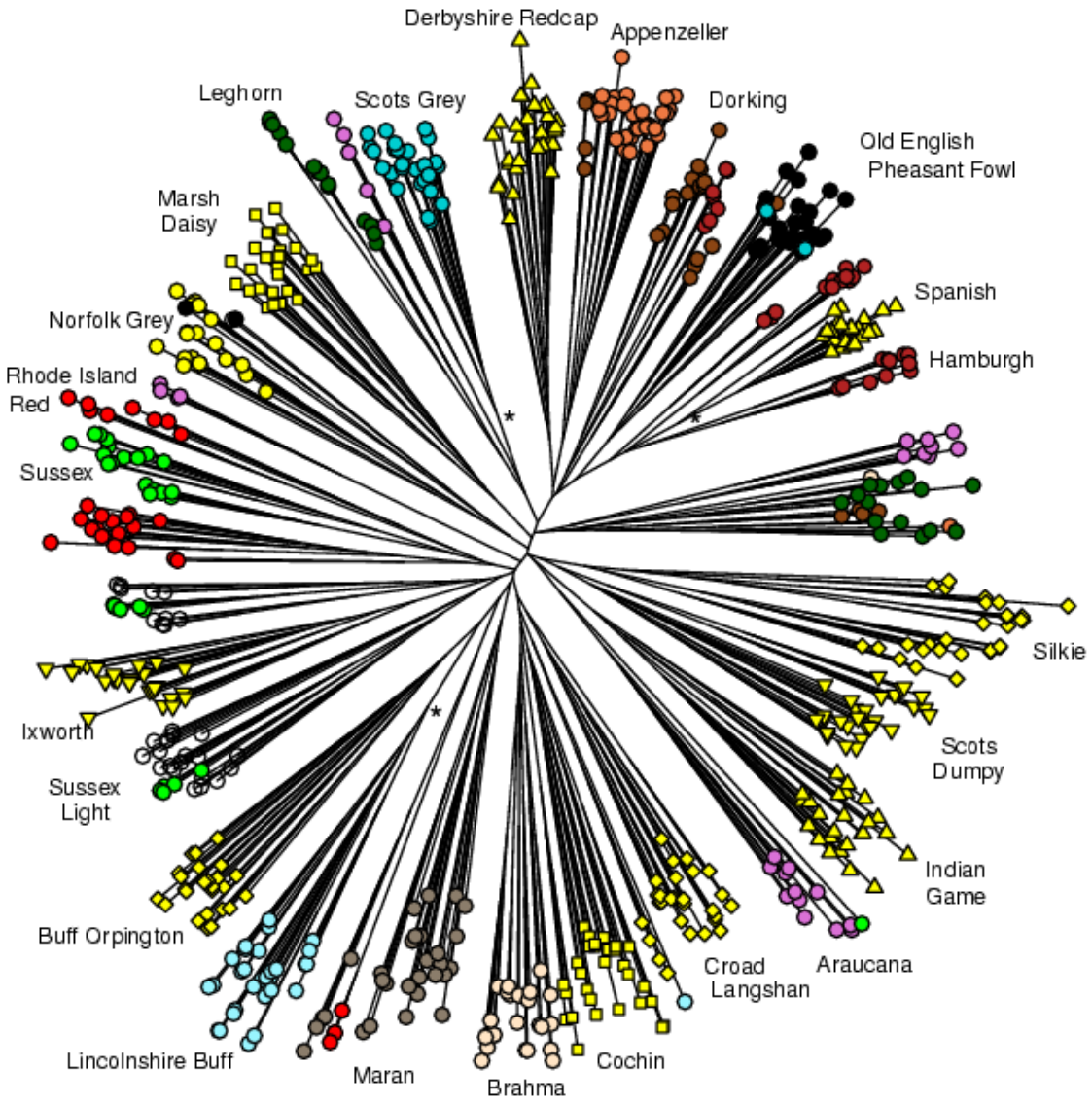


Figure 4.6 A neighbour-joining tree constructed from allele-sharing distances among all individuals. Bootstrap values greater than 500 are indicated with an asterisk. Breeds where all individuals clustered to origin are designated as yellow (various symbols). All other breeds had a specific colour such that certain individuals that did not cluster to origin can be traced on the evolutionary tree.

There was no bootstrap support for genetic relationships amongst the chicken breeds. There was high bootstrap support (> 50%) for the Lincolnshire Buff, Scots Grey and Spanish clades. Eleven chicken breeds formed single distinct genetic units where all individuals clustered to breed origin, although the bootstrap support was weak: Buff Orpington, Cochin, Croad Langshan, Derbyshire Redcap, Indian Game, Ixworth, Marsh Daisy, Norfolk Grey, Scots Dumpy, Silkie and Spanish (Fig 4.6, all designated as yellow in various shapes).

In the phylogenetic reconstruction of individuals, five other breeds formed single distinct genetic units, with the exception one or two individuals: Appenzeller (1 individual clustered with a mixed clade), Brahma (1 clustered with a mixed clade), Lincolnshire Buff (1 clustered with Croad Langshan), Old English Pheasant Fowl (2 clustered with Norfolk Grey) and Scots Grey (2 clustered with Old English Pheasant Fowl) (Fig 4.6). These mis-assignments were mirrored by the BAPS results and the individuals appeared admixed for the following: Appenzeller (1), Brahma (1), and Old English Pheasant Fowl (2) (Fig 4.5). The mis-assignments that were not admixed could be the result of mis-labelling during the data collection process.

The remaining breeds exhibited extensive genetic subdivision with 4 groups each detected in Dorking and Light Sussex, 3 groups each detected in Araucana, Hamburg, Rhode Island Red and Sussex and 2 groups each detected in Leghorn and Maran (Fig 4.6). BAPS concurred that there was extensive genetic subdivision in Aruacana, Leghorn, Maran, Rhode Island Red, and Sussex. In addition, BAPS split both Cochin and Marsh Daisy into 2 clusters and Silkie into 3 clusters (Fig 4.5).

4.3.4 Breed genetic contributions

Genetic contribution to between-breed diversity (CB) ranged from 3.17 in Cochin and Sussex to 6.71 in Spanish (Table 4.5).

Table 4.5 The genetic contribution of the 24 chicken breeds to genetic diversity ranked by aggregate diversity (D). [1] contribution to between-breed diversity; [2] contribution to within-breed diversity; [3] aggregate diversity.

Breed	CB[1]	CW[2]	D[3]
Araucana	3.28	1.13	1.67
Silkie	4.35	0.60	1.54
Maran	3.41	0.78	1.44
Rhode Island Red	3.59	0.69	1.42
Buff Orpington	4.12	0.51	1.41
Sussex	3.17	0.69	1.31
Leghorn	3.75	0.42	1.25
Cochin	3.17	0.60	1.24
Light Sussex	3.60	0.42	1.22
Indian Game	4.91	-0.02	1.21
Lincolnshire Buff	5.12	-0.12	1.19
Norfolk Grey	4.37	0.07	1.16
Scots Dumpy	4.02	0.16	1.13
Brahma	3.61	0.25	1.09
Marsh Daisy	4.37	-0.02	1.08
Croad Langshans	4.23	-0.02	1.04
Old English Pheasant Fowl	3.84	-0.11	0.88
Ixworth	5.07	-0.55	0.86
Appenzeller	4.81	-0.55	0.79
Derbyshire Redcap	4.91	-0.64	0.75
Dorking	3.51	-0.20	0.73
Scots Grey	5.15	-0.81	0.68
Hamburg	3.51	-0.64	0.40
Spanish	6.71	-2.67	-0.33

CB was negatively correlated with the genetic contribution to within-breed diversity (CW) (Spearman's rank correlation = -0.70, $p \leq 0.05$), with CW values ranging from

-2.67 in Spanish to 1.13 in Araucana. Genetic contribution to total aggregate diversity (D) ranged from -0.33 in Spanish to 1.67 in Araucana. As the calculation of D was weighted by the F_{ST} estimate, greater weight was applied to CW than CB (in this study) and, as a result, the breed estimates of D essentially mirrored those of CW (Spearman's rank correlation = 0.94, $p \leq 0.05$).

4.4 Discussion

4.4.1 Genetic diversity within breeds

The microsatellite markers revealed large genetic diversity among individuals within British chicken breeds (75%, F_{ST} = 0.25). This was evidenced by the estimated average number of alleles and expected heterozygosity within the breeds (Table 4.3), levels comparable to those reported in other European chicken breeds (Berthouly et al. 2008; Bodzsar et al. 2009; Granevitze et al. 2007). However, an exception was the Spanish breed, which exhibited a considerably lower genetic diversity than other breeds (Table 4.3). A well-established breed in Britain (Table 4.1), the Spanish has an extensive white face (Hams 2004; Roberts 1997). Very limited genetic introgression to maintain this physically distinctive characteristic and genetic drift, enhanced by the small population size (Table 4.1), may have led to the fixation of numerous loci and low genetic diversity in this breed.

4.4.2 Heterozygote deficiency within breeds

The observed heterozygosity within breeds was lower than expected, assuming random union of gametes, leading to significant deviations from Hardy-Weinberg

equilibrium (HWE) in many breeds (Table 4.3). Consequently, the estimates of inbreeding coefficients (Fig 4.2b) were considerably higher in the British chicken breeds than those reported in other European chicken breeds (Berthouly et al. 2008; Bodzsar et al. 2009; Granevitze et al. 2007; Zanetti et al. 2010).

Several non-mutually exclusive hypotheses could account for the observed heterozygote deficiency: null alleles, Wahlund effect (pooling of distinct genetic populations) and inbreeding (mating between close relatives). The presence of null alleles would cause a locus-specific heterozygote deficit for all the breed-locus combinations, which was not observed in any loci (Fig 4.2a). Moreover, the 30 genotyped microsatellites constitute a FAO-recommended panel and many of the loci have been used in earlier European chicken breed diversity studies, none of which have reported the occurrence of null alleles (Bodzsar et al. 2009; Granevitze et al. 2007; Granevitze et al. 2009).

The microsatellite markers revealed genetic substructure in many of the breeds (Fig 4.5-4.6). Indeed, the extensive genotypic linkage disequilibrium (LD) observed in certain breeds (Table 4.3) could be due to differences in allele frequencies among genetically differentiated populations within breeds (Nei and Li 1973). Further analysis of H_E and H_O for the subpopulations within breeds that were defined by the Bayesian genotypic clustering analysis revealed that the large subpopulations of Sussex (Speckled), Maran and Rhode Island Red were in HWE (Fig 4.2c). It is likely that the Wahlund effect was the cause of the heterozygote deficit for those breeds.

It was not possible to assess the influence of mating of close relatives due to the absence of a pedigree. Although this practice may be a contributing factor, the principal cause of heterozygote deficit was likely to be genetic substructuring within breeds as this was evident in many breeds (Fig 4.5-4.6).

4.4.3 Genetic substructure within breeds

In general, European chicken breeds were observed to be distinct homogenous genetic populations with little evidence of substructure within breeds (Bodzsar et al. 2009; Zanetti et al. 2010). In contrast, many of the British chicken breeds exhibited extensive genetic substructure (Fig 4.5-4.6). The patterns of genetic subdivision within the Leghorn and Sussex breeds were associated with morphological type. However, genetic subdivision could not be explained by morphological type in the Araucana, Buff Orpington, Dorking, Hamburg, Maran, Norfolk Grey, Rhode Island Red and Light Sussex. Instead, the observed genetic substructuring in these breeds reflected suppliers, whereby all individuals from a particular flock were genetically separated from the rest of the same chicken breed. Such genetic substructure associated with flock supplier has been observed in a few other chicken breeds (Rosenberg et al. 2001; Zanetti et al. 2010). Certain management practices within these breeds (e.g. restricted gene flow between smallholders) could have produced subtle genetic heterogeneity amongst flocks within breeds (Rosenberg et al. 2001).

4.4.4 Genetic distinctiveness and similarities of breeds

The genetic differentiation amongst the British chicken breeds was strikingly high (Table 4.4, Fig 4.3), with levels comparable to those found in French and Hungarian chicken breeds (Berthouly et al. 2008; Bodzsar et al. 2009). In general, chicken breeds appear to be genetically distinct populations with limited gene flow amongst the phenotypically diverse breeds. Given the small population sizes for many of the British chicken breeds (Table 4.1) and the short generation interval, random genetic drift has likely contributed to the elevated levels of genetic differentiation.

Although the British chicken breeds were highly differentiated, the overall genetic relationships between breeds were largely unresolved. Only a few pairs of breeds had high bootstrap support. This support is lower than that attained in other population genetic analysis of chickens (Vanhalal et al. 1998). Nonetheless, clustering of certain pairs and groups of breeds across the methods indicates some genetic similarities between certain breeds which can be supported by historical information on breed development. Phylogenetic reconstruction revealed two principal clades, one consisting of Derbyshire Redcap, Dorking, Hamburg, Old English Pheasant Fowl and Spanish and the other clade consisting of Buffs (Buff Orpington and Lincolnshire Buff), Asian (Cochin, Brahma and Croad Langshan), Sussex, Ixworth, Maran and Rhode Island Red breeds (Fig 4.3, 4.6). The first group (excluding the Spanish) consisted of old indigenous breeds (Table 4.1). These breeds are claimed to be closely related and may have contributed to the breed improvement of each other (Hams 2004).

Regarding the second group, it has been suggested that Asian breeds may have been used to improve the Sussex breeds (Vorwald Dohner 2001). The Asian breeds also influenced the development of Buff Orpington and Lincolnshire Buff (Table 4.1). In turn, the Sussex contributed to the development of the white-feathered Ixworth (Table 4.1). Cross-breeding of the Sussex and Rhode Island Red became the basis of commercial poultry enterprises in Britain (Hams 2004) and could have had an effect on non-commercial stocks of the breeds. Before the importation of Maran to Britain the breed was improved using, amongst other breeds, Croad Langshan stock (personal communication Andrew Sheppy, 2010).

It was not possible to explain the genetic relationships amongst the remaining breeds (Appenzeller, Araucana, Indian Game, Leghorn, Marsh Daisy, Norfolk Grey, Silkie, Scots Dumpy and Scots Grey). The low bootstrap support for these British chicken breeds suggest little shared evolutionary relationships with the other sampled breeds. In addition, the predominantly long branches depicted on the evolutionary tree indicate that these breeds are also genetically distinct (Fig 4.3).

4.4.5 Genetics related to breed conservation

The estimates of genetic diversity both within and between breeds highlighted which British traditional chicken breeds were genetically robust and those that were of potential concern (Table 4.5). The most genetically diverse breeds were generally those represented by more than one morphological colour type. For instance, Leghorn had 4 sampled types, Araucana, Brahma, Silkie and Sussex each had 3 and

Cochin and Maran each had 2 (Table 4.1). However, there were exceptions where genetically diverse breeds had only one sampled type (Buff Orpington, Rhode Island Red, Scots Dumpy and Light Sussex). The most genetically distinctive breeds were amongst the most morphologically distinctive breeds. For example, Indian Game (short legs, squat and wide), Silkie (degenerate feathers, dark skin) and Spanish (white face) possess traits not found in other breeds sampled in this study. Furthermore, the most genetically distinctive breeds also tended to have low population sizes (e.g. Appenzeller, Buff Orpington, Hamburg, Indian Game, Ixworth, Lincolnshire Buff, Scots Grey and Spanish) (Table 4.1).

Although the Spanish was the highest contributor to between breed diversity (CB, Table 4.5) due to high genetic differentiation (Table 4.4), it also possessed very low genetic variation (Table 4.3). Indeed, generally the genetically unique breeds (those with relatively high F_{ST}) had lower genetic diversity (H_E) than other breeds (Spearman's rank correlation = -0.78, $P \leq 0.05$) (Table 4.3-4.4), and, thus, these measures are not independent (Tapio et al. 2005b). This was not a surprising result as it is generally known that F_{ST} is intrinsically related to the proportion of variability present within populations due to the dependence of F_{ST} on average within population heterozygosity (Charlesworth 1998; Hedrick 1999; Jost 2008). Because F_{ST} measures the reduction in heterozygosity in the subpopulations relative to the total population ($(H_T - H_E)/H_T$), F_{ST} is constrained to be less than the average heterozgosity ($1 - H_E$). Consequently, ranking breeds on the levels of contribution to between breed diversity has received criticism because there is a greater emphasis on highly differentiated populations, which tend to be more monomorphic than less

differentiated populations (Toro and Caballero 2005; Toro et al. 2009). This is of particular relevance for livestock breeds because a large amount of genetic variation is present within breeds and, therefore, defining conservation priorities based solely on genetic distinctiveness will effectively ignore a lot of genetic diversity. Instead, a method that incorporates both between- and within-breed contributions may be a preferable alternative to rank breeds based on genetic estimates (Toro and Caballero 2005; Toro et al. 2009). By adopting an approach that incorporates both within and between breed diversity (Ollivier and Foulley 2005), the Spanish was the lowest contributor to overall genetic diversity (measured by D, Table 4.5). The highest contributors to overall diversity were the most genetically diverse breeds such as Araucana, Buff Orpington, Maran, Rhode Island Red and Sussex.

Two breeds, Buff Orpington and Silkie, were not only genetically unique (F_{ST} > 0.24, Table 4.4), but also possessed high levels of genetic diversity (H_E > 0.50, Table 4.3), and thus these breeds were high contributors to overall British chicken breed biodiversity (Table 4.5). A reverse perspective can be adopted to identify the more genetically vulnerable breeds (i.e. those that possess both low genetic diversity and uniqueness). Three breeds satisfied these criteria, Dorking, Hamburg and Old English Pheasant Fowl, and this was confirmed by the low contribution of these breeds to overall aggregate diversity (Table 4.5). In addition, these breeds possessed no private alleles (Table 4.3). The genetic results (coupled with the low population size of Old English Pheasant Fowl, Table 4.1) highlight that the future genetic viability of these breeds is of potential concern.

4.4.6 British chicken breeds as a genetic resource

As with other livestock species, the mechanisms behind agricultural industrialisation have raised concerns over genetic erosion in commercial chicken populations. A recent study using SNPs reported an absence of genetic diversity in both broiler and layer lines in comparison to ancestral and non-commercial populations (Muir et al. 2008). However, microsatellite studies have found that the amount of genetic diversity present in certain commercial chicken lines is not substantially different from that in non-commercial chicken breeds. As with the European morphologically selected breeds (managed to a breed standard) (Granevitze et al. 2007; Hillel et al. 2003), the genetic diversity found in British chicken breeds was far greater than in white egg layer lines, similar to brown egg layer lines, but less than in broiler lines. The differing levels of genetic diversity of the commercial lines are congruent with their known history. White egg layers were intensively bred from the single comb White Leghorn whilst several breeds contributed to the development of other commercial lines (Crawford 1990). Similarly, in comparison to other commercial lines, the magnitude of LD in white egg layers was greater and more extended over longer physical distances, reflecting greater long-term selection (Megens et al. 2009; Qanbari et al. 2010). The inconsistent results gleaned using the two genetic markers could be due to the fact that the study of Muir et al (2008) likely involved the sampling of more intensively selected commercial lines. The latest findings using SNPs reaffirms the need to conserve more traditional poultry resources, like the phenotypically diverse British traditional breeds, as they could be important reservoirs of genetic polymorphism.

4.5 Conclusion

The FAO-recommended panel of microsatellite markers revealed that the phenotypically diverse set of sampled British chicken breeds had high levels of genetic diversity and genetic differentiation, comparable to those reported in mainland European chicken breeds. However, the observed heterozygosity was considerably lower than what would be expected under HWE. It is likely that this was due to the sample pooling whereby flocks had subtle allelic differences thus creating a 'flock' Wahlund effect. In order to maintain and preserve the current levels of genetic diversity it is proposed that gene flow should be enhanced amongst flocks. In addition, certain breeds had low levels of both genetic diversity and uniqueness and consideration is required for the conservation and preservation of these potentially vulnerable breeds.

CHAPTER FIVE

Evaluation of approaches for selecting breed informative markers from high density assays

5.1 Introduction

Genetic markers can be used for the identification and verification the origin of individuals if there is sufficient genetic heterogeneity among populations (Paetkau et al. 1995). The task is useful in a variety of biological contexts; topical issues in population, conservation and evolutionary biology, such as the characterisation of population structure (Maudet et al. 2002; Paetkau et al. 1995) or the migratory patterns of individuals between populations, can benefit from the inference of ancestry of individuals (Davies et al. 1999; Manel et al. 2005; Waser and Strobeck 1998). In an applied context, genetic identification of individuals can shed light on issues such as the contribution of potential source populations in mixed fisheries (Manel et al. 2005; Roques et al. 1999), meat traceability or brand authentication (Ciampolini et al. 2006), illegally translocated individuals (Rannala and Mountain 1997), anthropological forensic investigations (Davies et al. 1999) and tracing illegally poached animals and trafficking routes (Manel et al. 2005).

The recent development of dense genome-wide Single Nucleotide Polymorphism (SNP) marker assays for many livestock species offers the prospect of detailed study of population, conservation and evolutionary biology (Gibbs et al. 2009). However, dense genome-wide data can be relatively costly to produce and time-consuming or computationally intensive to analyse. Therefore, it is often desirable to reduce the number of markers by screening and selecting the most informative genetic markers (Lao et al. 2006; Paschou et al. 2007). Dense genome-wide arrays provide the

opportunity to develop tailor-made panels with high information content for use in population genetics analyses.

The level of genetic information contained in markers can be estimated using methods that describe the levels of genetic differentiation between populations. Normally, where these approaches are applied in empirical studies, estimates are averaged across many markers to summarise population genetic differentiation as a single number. However, genetic markers will contain varying levels of population discriminatory information and population genetic differentiation measures can be used to identify the most genetically informative markers.

Several different population differentiation measures have been developed. The statistic delta measures the allele frequency difference between a pair of populations and is commonly used in the field of human genetics to assess marker information content (Shriver et al. 1997; Smith et al. 2001). Bowcock et al (1994) suggested that informative genetic markers could be identified using F-statistics. Wright (Wright 1943; Wright 1951) introduced F_{ST} as a way to described the proportion of genetic diversity within and among populations (Holsinger and Weir 2009). Wright's F_{ST} has been extended several times and one of the preferred statistics based on the analysis of variance of allele frequencies is Weir and Cockerham's (W&C) F_{ST} (Weir and Cockerham 1984). The multivariate statistical method of Principal Component Analysis (PCA) has also been more recently proposed as an alternative method to identify population informative SNP markers (Paschou et al. 2007).

Although these marker selection methods have been used to identify informative markers (Paschou et al. 2007; Smith et al. 2001), in the case of selecting SNPs from dense genetic arrays (i.e. SNP chips), there is little guidance as to what may be the most appropriate method to pull out informative markers. Blott et al (1999) recommended extracting microsatellites that had the greatest number of alleles and heterozygosity. However, bi-allelic SNP markers are less variable that microsatellites. Not surprisingly, researchers that have proposed a particular selection method generally advocate it, like Paschou et al (2007) introducing PCA for SNP selection and subsequently encouraging this method of choice in further population genetic studies (Drineas et al. 2010; Lewis et al. 2011; Paschou et al. 2008; Paschou et al. 2010).

The objective of this study was to evaluate methods for selecting population informative SNP loci. To achieve this, the minimum number of SNP markers from the Illumina Bovine50SNP beadchip (Illumina Inc., San Diego, CA) that would be required to discriminate a set of European cattle breeds was determined for each of several SNP selection methods. This was approached in a two-stage manner. First, the SNP selection methods were used to determine the genetic information content of each SNP marker and markers were ranked by decreasing level of informativeness for each of the methods. Second, the likelihood of assigning individual genotypes to their known breed origin was estimated by cumulatively increasing the number of SNP markers, ranked according to the estimates of each SNP marker's informativeness for each selection method.

5.2 Materials and Methods

5.2.1 Data

Allele frequencies from 17 cattle breeds representing the 'reference' populations and a total of 384 individual genotypes of known breed origin, sampled from the reference populations, were available. Information on the sampling of the reference populations is given in Table 5.1. Decker et al (2009) selected 40,843 SNPs from the Bovine SNP50 Bead Chip after a strict quality screening where "Loci selected for analysis were all located on autosomes, had a call rate of at least 80% in 36 (75%) *Bos taurus* breeds, and were not monomorphic in all breeds…" Since only *Bos taurus* breeds were used in the current study the selected set of SNP markers by Decker et al (2009) was adopted. Detailed information of the genotyping procedure can be found in Decker et al (2009).

5.2.2 Selection methods to determine the most informative markers

The breed-specific allele frequencies for the 40,483 SNPs were used to estimate the genetic information contained in each SNP marker using the following selection methods: delta, Wright's F_{ST}, Weir and Cockerham's F_{ST} and PCA. The larger the estimated value, the more informative the marker is for differentiating among populations and providing information regarding ancestry. A Spearman's rank correlation was calculated between the different estimates from the selection methods. All analyses were conducted in the R statistical environment (Team 2011).

Table 5.1 Information on the breeds. *N*, reference sample size (used to estimate the allele frequencies), [1] HapMap individuals are unrelated except where indicated by 'trio' (Gibbs et al. 2009), 'Registered' refers to animals that have been recorded by its breed registry and, *n*, the number of individuals used in assignment testing.

	Breed	*N*	Animal resources of *N*	*N*	Purpose	Historical origin	Distribution	Sample Locality
1	Angus - British	23	several Scottish farms; majority different sires	23	Beef	Scotland (UK)	Global	UK
2	Angus - American	6124	Registered bulls and steers	25	Beef	Scotland (UK)	Global	USA
3	Brown Swiss	74	24 HapMap[1] (3 trios); remaining no pedigree	24	Dairy	Switzerland	Alpine Europe, Americas	USA
4	Charolais	135	26 HapMap[1] (3 trios); remaining registered	25	Beef	France	France, USA, Brazil, RSA	USA
5	Finnish Ayrshire	444	215 unrelated; 17 paternal half-sib families with average of 13 progeny per sire	10	Dairy	Scotland (UK)	Global	Finland
6	Guernsey	23	21 HapMap[1]; remaining unrelated	21	Dairy	Island of Guernsey (UK)	USA, UK, Oceania, RSA	UK
7	Hereford	143	32 HapMap[1] (4 trios); remaining registered	25	Beef	UK	Global	USA
8	Holstein	18904	Registered	25	Dairy	Netherlands	Global	USA
9	Jersey	93	28 HapMap[1] (3 trios); remaining registered	28	Dairy	Island of Jersey (UK)	Global	USA
10	Limousin	1621	All registered	25	Beef	France	France, UK, USA	USA
11	Norwegian Red	21	HapMap[1] (1 trio)	21	Dual Purpose	Norway	Norway	Norway
12	Piedmontese	29	24 HapMap[1] (3 trios); remaining unrelated	19	Beef	Italy	Italy	Italy
13	Red Angus	15	Registered	15	Beef	Scotland (UK)	USA, Australia	USA
14	Red Poll	23	Registered, a few shared sires and dams	23	Beef	UK		UK
15	Shorthorn	108	Registered (7 trios)	25	Dual Purpose	UK	Global	USA
16	Simmental	777	104 sires; 673 steers from 24 sires	25	Beef	Switzerland	Global	USA
17	Welsh Black	32	several Welsh farms; unrelated	25	Beef	Wales (UK)		UK
	Total:	28589		384				

For a biallelic marker the delta value, the absolute allele frequency difference, is given by $| p_{Ai} - p_{Aj} |$, where p_{Ai} and p_{Aj} are the frequencies of allele A in the i^{th} and j^{th} populations, respectively. Delta could be viewed as a form of genetic distance in that it is designed to measure the genetic differences between populations, where if there are no differences the value obtained would be 0 and if populations share no alleles the value obtained would be 1. However, delta differs from the standard genetic distance measures, such as Nei's D_S, in that these models usually contain population genetic assumptions. For instance, Nei's D_S (1972) is a standard genetic distance which measures the proportion of alleles that are the same between a pair of populations. D_S assumes that initial genetic variability in a population is at equilibrium between mutation and genetic drift, with effective population sizes remaining constant. Delta can only be estimated between pairs of populations (K = 2). Since K = 17 in this study, values were averaged across all pairwise comparisons to produce an estimated value for each SNP marker.

Both Wright's (Wright 1943; Wright 1951) and Weir and Cockerham's (W&C) (Weir and Cockerham 1984) F_{ST} approaches were used to quantify the degree of population allelic variation (i.e. population genetic differentiation) contained in each SNP marker. For both methods unbiased estimates of F_{ST} were first calculated over all populations (global F_{ST}) and on a pairwise basis (pairwise F_{ST}), with the latter values being averaged over all pairs to produce an estimated information content value for each SNP marker. Wright's F_{ST} was estimated as  , where var(p_A) is the variance of the allele frequency among breeds and $\bar{p}_A$ is the mean allele frequency across the breeds. Unbiased estimates of W&C's F_{ST} were

estimated as functions of variance components as detailed in Akey et al (2002). These F_{ST} estimates can be negative if alleles drawn randomly from within a population are less similar to one another than those drawn from different populations ($F_{ST} < 0$) (Akey et al. 2002; Weir 1996). In this study the estimated F_{ST} values were left as negative.

Principal Component Analysis (PCA) was used to reduce the multidimensional dataset (for further details see section 2.2.3 in chapter 2). The coefficients ("loadings") used in the linear transformation of the original variables into new variables generate the proportion of variance that a variable (SNP marker) contributes to a given principal component. Since loadings represent the 'weight' of a variable (the amount that a variable contributes to the structuring obtained by a PC), they can be used to determine how much information is present in each variable. PCA was performed following Paschou et al (2007) but on the breed-specific allele frequency matrix rather than the individual genotypes. To determine which principal components were significant, 100 random matrices were created by sampling with replacement allele frequencies within each SNP marker across all breeds. The first eight principal components for the actual data contained more information than in the randomly generated components (i.e., their eigenvalues were greater). Therefore the loadings of the first eight principal components were used to calculate the level of genetic information contained in each marker. The loadings for each SNP marker were squared and summed over the eight significant principal components to produce an estimate of informativeness.

5.2.3 Individual assignment analysis

Several genetic assignment approaches are available (Cornuet et al. 1999; Paetkau et al. 1995; Rannala and Mountain 1997). The Bayesian implementation developed by Rannala and Mountain (1997) has been found to be more effective at individual assignment than other methods (Cornuet et al. 1999). However, the method of Paetkau et al (1995) is equally effective at individual assignment when the levels of genetic differentiation between reference populations are high (Cornuet et al. 1999). Consequently, the method of Paetkau et al (1995) was employed as it is easier to implement than that of Rannala & Mountain (1997) and is most frequently employed in empirical studies.

Allele frequencies of zero were replaced by a value of 1×10^{-5} because log(0) is not defined (Paetkau et al. 1995). Likewise, if an observed allele frequency was 1, it was replaced by a value of 0.99999. Genotype likelihoods were calculated for each individual in each reference population based on the observed allele frequencies for each marker. Let p_{ijk} denote the frequency of the k^{th} allele (k = 1, 2) at the j^{th} locus (j = 1 .. J) in the i^{th} population (I = 1 .. I). Let $g_{jkk'}$ denote an individual's diploid genotype at the j^{th} locus, and let the Mendelian transmission probability of $g_{jkk'}$ arising in the i^{th} population be $T(g_{jkk'} \mid i)$

,

where a genotype is homozygous if $k = k'$ and heterozygous otherwise, under the assumption of random union of gametes. Next, let g denote an individual's multilocus genotype. The likelihood of an individual diploid genotype occurring in a particular population, $T(g|i)$, was estimated as above, as the square of the observed allele frequency for homozygotes or twice the product of the two allele frequencies for heterozygotes. Under the assumption of independence between the J loci

$$T(g \mid i) = \prod_j T(g_{jkk'} \mid i) \text{ and } \log_{10}(T(g \mid i)) = \sum_j \log_{10}(T(g_{jkk'} \mid i)).$$

To assess the performance of the assignment procedure, log-likelihood ratios (LLR) were calculated by comparing the likelihood of an individual being assigned to its population of origin and the likelihood of it being assigned to another population


.

The log-likelihood ratio (LLR) (also termed LOD score) is often used in population genetics to determine the origin of individuals in population (Campbell et al. 2003; Roques et al. 1999) and paternity studies (Marshall et al. 1998), where the likelihood has been shown to be an efficient approach for the evaluation of alternative populations or parental relationships. Meagher (1986) criticised the use of LLR (or LOD) because it is not a valid likelihood ratio, as one hypothesis is not tested within the other. When a constrained model is nested within an unconstrained model, then for large samples -2Log(LLR) (or -2Ln(LLR)) is chi-square distributed with m-k-1 degrees of freedom where m is the number of categories in the data (e.g., cells in a

contingency table) and *k* is the number of independent parameters to be estimated from the data (e.g., allele or genotype frequencies). With individual assignment tests, *m* = 1 (since there is only one sample genotype as each individual is tested separately) and so the degrees of freedom is undefined. As a result the significance of -2Log(LLR) cannot be tested using approximation to a chi-square distribution (Meagher 1986; Marshall et al. 1998). Instead different stringency thresholds can be applied as confidence levels of assignment precision. Four stringency levels are commonly used: LLR > 0, LLR > 1, LLR > 2 and LLR > 3 (Campbell et al. 2003; Roques et al. 1999; Shriver et al. 1997; Smith et al. 2001). LLR > 1, LLR > 2 and LLR > 3 levels, respectively, mean that a multilocus genotype is required to be 10, 100 or 1000 times more likely in one population than in the other(s). The LLR > 0 level only requires the genotype to be more likely in one population relative to the other(s). The correct assignment of an individual genotype to its known origin occurred when the calculated LLR was greater than the selected stringency level. If the LLR was lower than the selected stringency level, the individual genotype failed to be assigned to its origin and was instead assigned to the reference population that yielded the highest overall log-likelihood.

To obtain an estimate of the number of SNP markers required to achieve 90%, 95% and 98% correct assignment success of the 384 individual genotypes for each of the selection methods, at each of the four threshold levels, a non-linear regression model was fitted to the curves of correct assignment percentage against cumulative markers. An asymptotic regression model ( , where parameter *a* represents the value of the asymptote, parameter *b* represents the difference between the value

of y when $x = 0$ and the upper asymptote and parameter c represents the natural logarithm of the rate of exponential increase) was found to best fit the data, with x representing the cumulative number of SNP markers and y representing the percentage of correct assignment. When $a > 0$, $b < 0$ and $c < 0$ the model represents the law of diminishing returns in which the rate of increase of y declines with successive equal increments of x.

To test whether the level of genetic differentiation of a breed corresponded to the power of assignment, a Spearman's rank correlation was calculated between the percentage of correctly assigned individuals for the 20 top ranked SNP markers for each breed (selection method = pairwise Wright's F_{ST}, LLR > 0) and the average F_{ST} for each breed (pairwise Wright's F_{ST} values across all breeds, based on 40, 843 SNP markers, averaged to provide an estimate for each breed).

A negative control to individual assignment analysis was applied by analysing 20 sets of 400 randomly selected SNPs. The average individual assignment success was estimated across the 20 random SNP sets at the stringency level LLR > 3.

In order to evaluate the power of assignment for samples of unknown origin, the individual assignment analysis was evaluated by cross-validation whereby a training sample was used to identify the informative loci and a holdout sample from each of the breeds was used to test the power of the resulting panel and the reference training sample. For breeds with a reference sample size > 50 (column *N*, Table 5.1) the holdout sample comprised all the individuals to be assigned (10 - 28; given in

column *n*). The genotypes of these 'hold-out' individuals were removed from their respective reference (defining the training sample) breed and allele frequencies of the reference breeds were re-estimated for the training sample. For breeds with a reference sample size < 50 (column *N*, Table 5.1) five random individual genotypes of the individuals assigned in the main analysis (those in column *n*) were designated as the holdout sample; these were removed from their respective reference breed (again, defining the training sample) and allele frequencies were re-estimated. The individual assignment analysis was repeated with the new training samples and the hold-out samples.

5.3 Results

5.3.1 Comparison of the marker selection methods

Frequency histograms of the level of genetic information in the SNP markers are shown for each selection method in Figure 5.1. A predominantly left-skewed distribution was produced for each selection method, except delta, which produced a fairly symmetric distribution. The majority of the markers contained low to medium levels of genetic information and a small proportion had high levels of genetic information.

The level of similarity of the estimates of genetic information contained in each SNP marker across the different selection methods was assessed. High levels of correlation were observed between delta, pairwise Wright's F_{ST}, pairwise W&C's F_{ST} and PCA (Table 5.2). Similarly, there was substantial overlap (> 200) in the top ranked 500 SNP markers between these four selection methods. In contrast, the level

of correlation was lower between global F_{ST} and the other selection methods. There was also far less overlap (< 200) in the top ranked 500 SNP markers between the global F_{ST} estimates and the other selection methods (Table 5.2).

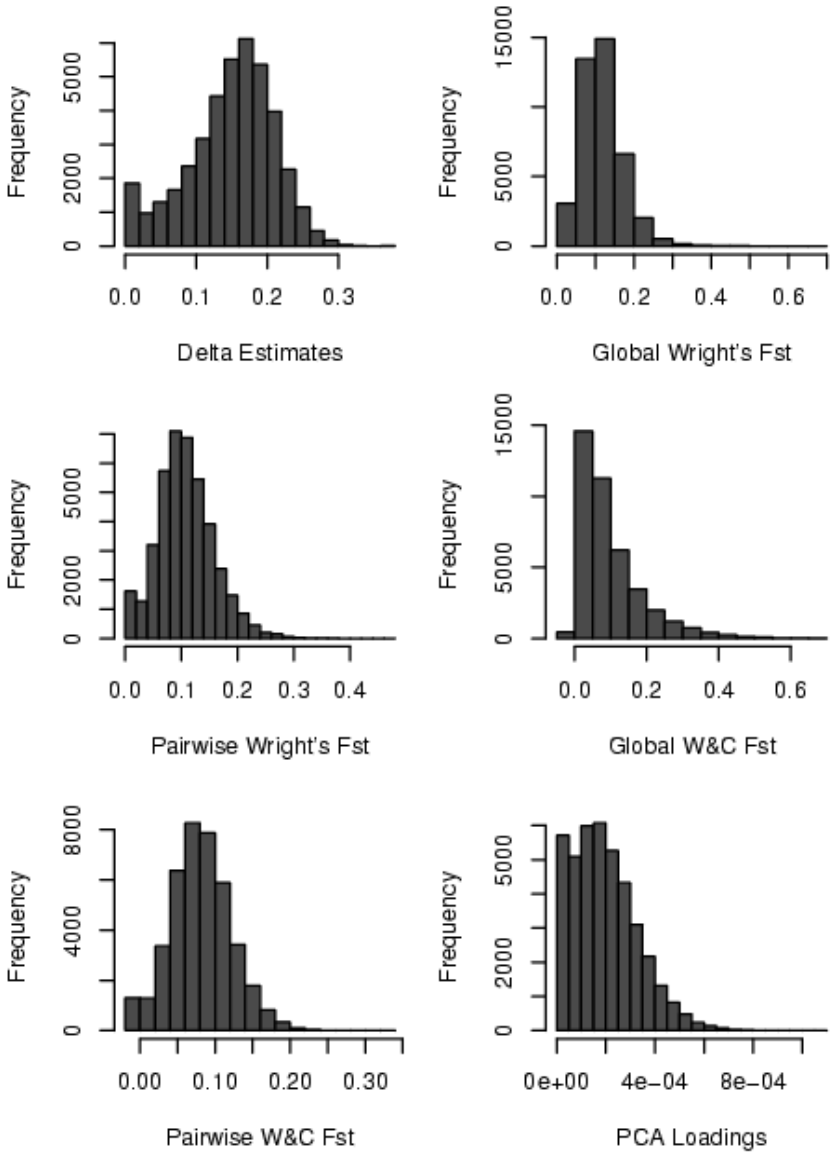


Figure 5.1 Frequency histograms of the estimates of genetic information contained in each SNP marker, for each selection method (x-axis scale is method-specific). The majority of the SNP markers display low to moderate estimates of genetic informativeness with few markers displaying high levels of population differentiation.

Table 5.2 Concordance between the SNP selection methods. The upper-triangle contains the Spearman's rank correlation of information content for each pair of selection methods (over all 40,483 SNPs). The lower-triangle contains the amount of overlap for the top 500 ranked SNP markers between each pair of selection methods.[1] the first eight principal components were significant and used to calculate marker informativeness.

	delta	Global Wright's F_{ST}	Pairwise Wright's F_{ST}	Global W&C'S F_{ST}	Pairwise W&C'S F_{ST}	PCA [1-8][1]
delta	-	0.589	0.884	0.370	0.819	0.928
global Wright's F_{ST}	98	-	0.847	0.462	0.821	0.682
pairwise Wright's F_{ST}	381	151	-	0.448	0.952	0.888
global W&C F_{ST}	59	49	63	-	0.461	0.408
pairwise W&C F_{ST}	306	156	367	67	-	0.810
PCA [1-8][1]	273	101	274	66	229	-

The conflicting results between global Wright's and global W&C's F_{ST} and the other selection methods were explored further. The observed breed allele frequencies for the top ranked 50 SNP markers for each selection method were displayed in a box-plot in Figure 5.2.

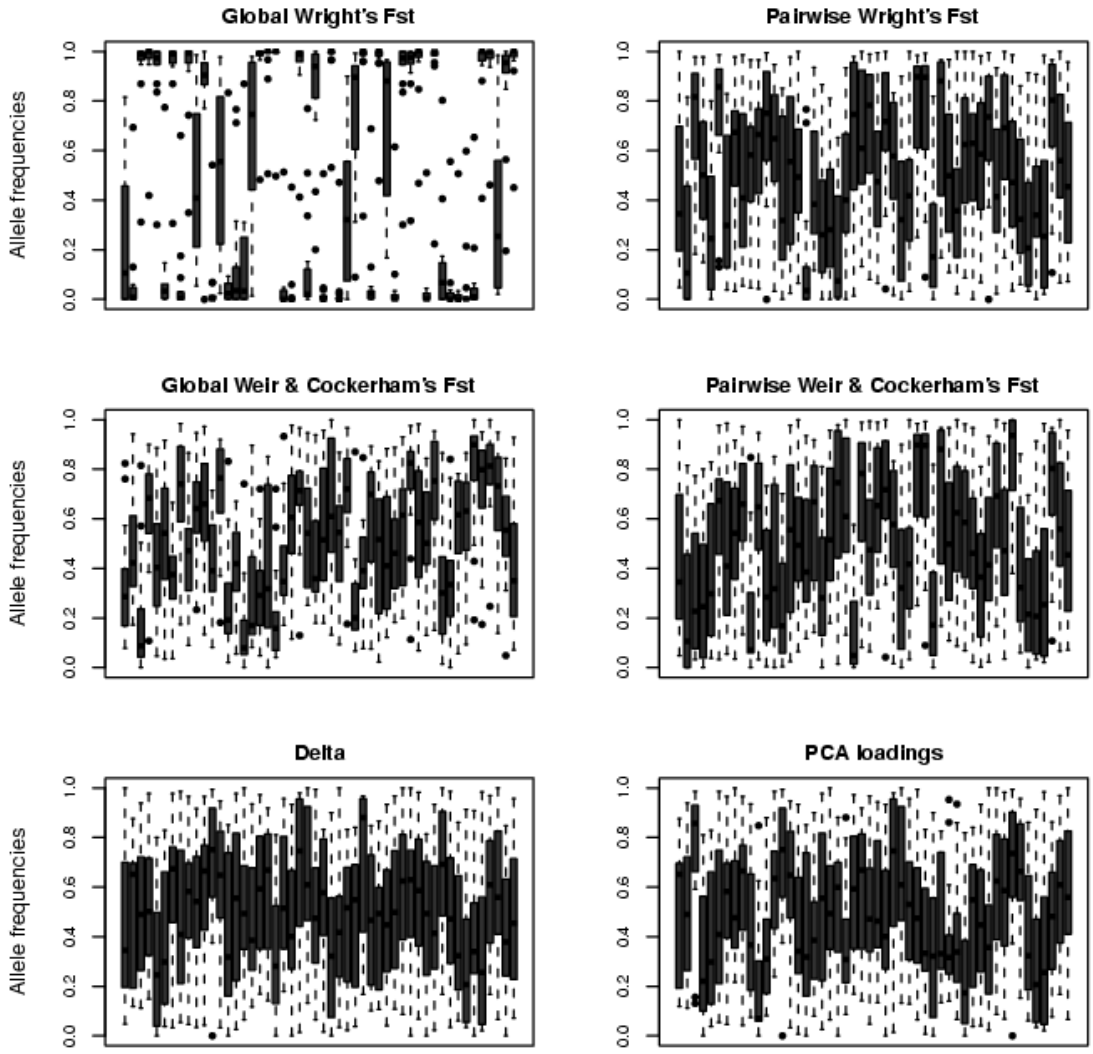


Figure 5.2 Boxplot of the observed allele frequency of the cattle breeds for the top-ranked SNPs.

The boxplot is an effective visual representation of both the central tendency and dispersion of data. As can be seen in Figure 5.2, delta, pairwise Wright's F_{ST}, pairwise W&C's F_{ST} and PCA selected SNP markers with median allele frequency between 0.2 and 0.8 and with large interquartile ranges indicating a high level of dispersion amongst the observed allele frequencies. In comparison, the majority of the top-ranked SNP markers selected by global Wright's F_{ST} had median allele frequencies near 0 or 1 and low levels of dispersion. The global W&C's F_{ST} resulted in the selection of SNPs with a higher level of dispersion amongst the observed allele frequencies than global Wright's F_{ST}, but, nonetheless, also included markers with outlying allele frequencies and smaller interquartile ranges than the other selection methods.

5.3.2 Assignment precision: overall assessment

The accuracy of assignment of individual genotypes to known breed origin was evaluated by cumulatively adding 20 markers, in descending order of estimated marker informativeness for each selection method. No population genetic differentiation was detected between the American and British Angus populations (Table 5.1), consequently the two populations were pooled together and treated as a single breed in subsequent analyses.

The success of assignment of the 384 individual genotypes to breed of origin at the four stringency level thresholds for four of the selection methods (delta, pairwise Wright's F_{ST}, pairwise W&C's F_{ST} and PCA) is presented in Figure 5.3. Strikingly, it

was immediately noticeable that > 50% assignment success for all four selection methods was achieved at stringency level LLR > 0 using just the first 20 SNP markers.

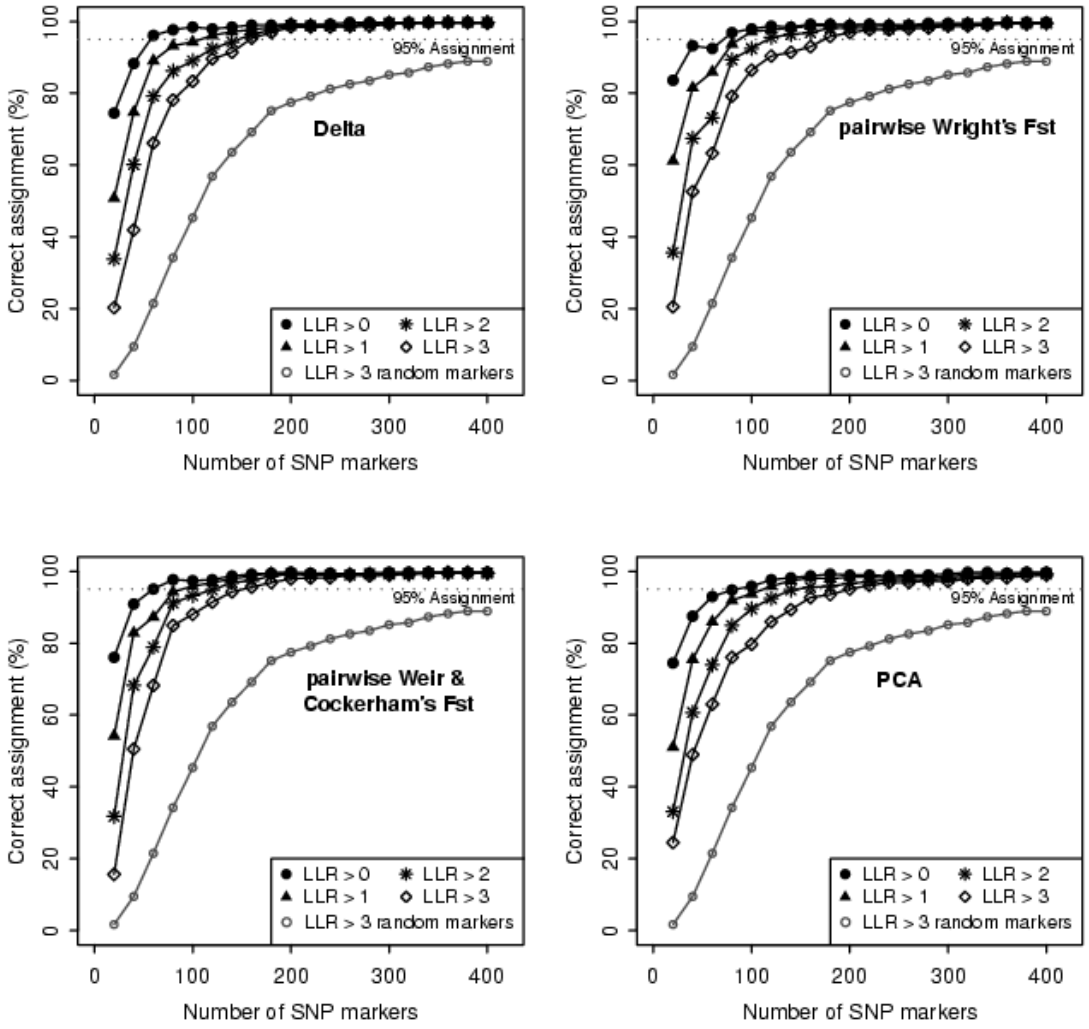


Figure 5.3 The percentage assignment success with cumulative number of top-ranked SNP markers at the four stringency threshold levels, for each selection method. 70% success was achieved with the first 20 SNP markers across all ranking methods; power of assignment did not increase beyond 200 SNP markers. Average assignment success across 20 sets of randomly selected markers is also shown for the LLR >3 stringency threshold level.

Overall, pairwise Wright's F_{ST} required the smallest number of SNP markers to reach 90%, 95% and 98% correct assignment at the four stringency threshold levels (Table 5.3). Of the four selection methods, PCA was the poorest performer, requiring > 190 SNP markers to attain 95% assignment success at the strictest stringency threshold (Fig 5.3; Table 5.3). The power of assignment using PCA as a selection method decreased considerably across all the stringency thresholds when a 98% assignment success was imposed.

Full results are not shown for assignment precision using ranked SNP markers for global F_{ST} because they performed comparatively poorly. For global Wright's F_{ST}, 90% assignment success was obtained with 230 and 380 SNP markers at the stringency levels of LLR > 0 and LLR > 3, respectively. Using up to 400 markers,95% assignment success was not achieved at any stringency level. For global W&C's F_{ST}, 90% assignment success was obtained with 80 and 230 SNP markers at the stringency levels of LLR > 0 and LLR > 3, respectively. The global W&C's F_{ST} had greater assignment accuracy over global Wright's F_{ST}, but still performed worse than the other four selection methods.

Randomly chosen SNP sets performed worse than ranked informative SNP markers in individual assignment analysis (Fig 5.3). Neither an asymptote nor 95% assignment success were reached using up to 400 markers (averaged across 20 sets of randomly chosen SNP at LLR > 3).

Table 5.3 Individual assignment performance for the four selection methods. Estimated number of SNP markers required to achieve 90%, 95% and 98% correct assignment at the four stringency thresholds for each SNP selection method (the individuals from the two Angus populations were pooled). Values estimated from asymptotic regression equation.

	delta			pairwise Wright's F_{ST}			pairwise W&C's F_{ST}			PCA		
Log_{10}	90%	95%	98%	90%	95%	98%	90%	95%	98%	90%	95%	98%
0	43	60	87	41	58	84	37	63	104	51	76	117
1	68	91	130	61	81	115	65	90	130	72	99	153
2	96	127	180	81	105	148	90	120	172	102	140	284
3	124	160	210	106	138	196	121	160	242	140	193	404

Each step in the individual assignment study was repeated using a training set and a holdout set in order to evaluate the power of assignment for samples not included in the reference population. Similar frequency histograms of estimates of genetic information, levels of correlations between the SNP selection methods, allele frequency dispersion of the top-ranked SNP markers and power of assignment were observed. The assignment power for breeds with large sample sizes N > 50 was comparable to the results of the main analysis (results not shown). Only certain breeds with a low sample size had worse assignment power in the cross-validation analysis. For example, poor assignment power was observed in Red Angus and Norwegian Red, two breeds of low sample size and for which closely related breeds were included in the dataset (Angus and Finnish Ayrshire, respectively).

5.3.3 Assignment precision: individual breeds

The SNP selection methods differed for power of assignment in individual breeds, but no one method consistently outperformed any other in all breeds (Table 5.4). No substantial further gain in power of assignment in individual breeds was observed beyond ~ 200 SNP markers. Certain breeds required relatively few SNP markers to attain > 95% assignment success (Table 5.4). For example, the Jersey breed required < 50 SNPs to achieve 100% individual assignment even when strict stringency levels were applied. In contrast, the Charolais breed required ~100 SNP markers to achieve > 95% individual assignment and power was severely reduced with increasing stringency level.

Table 5.4 Power of assignment in individual breeds. Percentage of individuals that were successfully assigned to their breed origin, at the four stringency threshold levels, for each selection method.

		Delta				pairwise Wright's F_{ST}				pairwise W&C's F_{ST}				PCA			
Breed	Markers	log0	log1	log2	log3	log0	log1	log2	log3	log0	log1	log2	log3	log0	log1	log2	log3
Angus	50	100	80	67	34	85.4	64.58	43.75	18.75	93.8	81.25	37.5	16.67	85.4	68.75	52.08	20.83
	100	100	92	78	73	97.9	91.67	89.58	87.5	100	100	95.83	91.67	89.6	79.17	77.08	60.42
	200	100	100	100	100	100	100	100	100	100	100	97.92	97.92	100	97.92	97.92	97.92
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	97.92	97.92
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Brown Swiss	50	100	96	96	96	100	100	100	96	100	100	100	92	100	100	100	100
	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	200	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Charolais	50	72	56	24	0	92	76	60	24	92	60	44	16	88	68	20	4
	100	88	76	56	24	96	96	84	60	96	88	80	44	92	92	84	52
	200	96	96	96	92	96	96	92	92	96	96	96	92	96	96	92	84
	300	100	96	96	96	96	96	96	92	96	96	96	96	96	96	96	96
	400	100	96	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Finnish Ayrshire	50	100	60	20	10	100	90	60	40	70	70	60	50	100	100	70	40
	100	100	100	90	90	100	90	80	50	90	90	80	50	100	100	100	80
	200	100	100	100	100	100	100	100	80	100	100	100	90	100	100	100	100
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Guernsey	50	100	100	96	96	96	96	96	96	100	96	96	96	96	96	96	96
	100	100	100	100	96	100	100	100	100	100	100	100	100	96	96	96	96
	200	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Table 5.4 cont.

		Delta				pairwise Wright's F_{ST}				pairwise W&C's F_{ST}				PCA			
Breed	Markers	log0	log1	log2	log3	log0	log1	log2	log3	log0	log1	log2	log3	log0	log1	log2	log3
Hereford	50	68	60	36	24	92	80	60	48	100	92	84	68	96	88	76	72
	100	100	88	88	84	100	100	96	84	100	100	100	96	100	100	100	100
	200	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Holstein	50	96	72	48	24	92	72	64	40	96	96	96	96	96	96	88	84
	100	100	96	96	92	100	100	100	100	100	100	92	88	100	100	92	92
	200	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Jersey	50	100	100	100	92.9	100	100	100	100	100	100	100	100	100	100	100	96.4
	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	200	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Limousin	50	92	84	56	40	96	92	84	48	88	80	72	44	84	60	20	12
	100	100	100	96	76	100	92	88	84	88	88	72	72	92	92	72	48
	200	100	100	100	100	100	100	100	96	92	92	92	92	100	100	100	100
	300	100	100	96	96	100	96	96	96	96	96	96	92	100	100	100	100
	400	100	100	100	100	100	96	96	96	100	96	96	96	100	100	100	100
Norwegian Red	50	90.5	71.4	61.9	33.3	90.5	71.4	57.1	28.6	90.5	81	71.4	57.1	85.7	76.2	61.9	28.6
	100	100	96	90.5	85.7	96	90.5	85.7	76.2	90.5	90.5	76.2	71.4	96	96	90.5	85.7
	200	100	100	100	100	100	100	100	96	100	100	100	96	96	100	100	96
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Piedmontese	50	100	94.7	94.7	78.9	100	100	100	94.7	100	94.7	94.7	73.7	94.7	84.2	73.7	47.4
	100	100	100	100	94.7	100	100	100	100	100	100	100	100	94.7	94.7	94.7	68.4
	200	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Table 5.4 cont.

		Delta				pairwise Wright's F_{ST}				pairwise W&C's F_{ST}				PCA			
Breed	Markers	log0	log1	log2	log3	log0	log1	log2	log3	log0	log1	log2	log3	log0	log1	log2	log3
Piedmontese	50	100	94.7	94.7	78.9	100	100	100	94.7	100	94.7	94.7	73.7	94.7	84.2	73.7	47.4
	100	100	100	100	94.7	100	100	100	100	100	100	100	100	94.7	94.7	94.7	68.4
	200	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Red Angus	50	93	73.3	46.7	26.7	86.7	53.3	33.3	20	80	53.3	46.7	13.3	93.3	53.3	46.7	20
	100	93	80	66.7	60	86.7	86.7	80	66.7	100	93.3	93.3	73.3	93.3	80	73.3	60
	200	93	93.3	93.3	93.3	100	100	100	93.3	100	100	93.3	93.3	100	93.3	80	80
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	86.7	86.7	86.7
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	93.3	93.3	93.3
Red Poll	50	88.9	88.9	83.3	72.2	100	100	83.3	77.8	94.4	88.9	77.8	66.7	100	94.4	94.4	94.4
	100	100	100	100	100	100	100	100	94.4	100	100	100	100	100	100	100	100
	200	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Shorthorn	50	80	76	68	56	92	92	92	80	92	92	92	88	96	92	80	80
	100	92	88	88	88	92	92	92	88	96	96	92	92	100	96	96	92
	200	96	96	92	92	100	96	96	96	100	96	96	96	100	100	100	100
	300	96	100	100	100	100	96	96	96	100	96	96	96	100	100	100	100
	400	100	100	100	96	100	100	100	100	100	100	100	100	100	100	100	100
Simmental	50	100	92	68	36	92	92	80	60	96	84	68	40	88	68	44	32
	100	100	92	84	80	96	100	96	96	100	100	88	76	92	92	76	56
	200	100	100	100	96	100	100	100	100	100	100	96	92	92	88	76	68
	300	100	100	100	96	100	100	100	100	100	96	96	96	96	92	88	76
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96	92
Welsh Black	50	100	100	93.3	76.7	96.7	93.3	83.3	80	100	96.7	90	83.3	96.7	96.7	90	83.3
	100	100	100	100	100	96.7	93.3	93.3	93.3	100	100	100	100	96.7	96.7	96.7	96.7
	200	100	100	100	100	100	100	100	100	100	100	100	96.7	100	100	96.7	96.7
	300	100	100	100	100	100	100	100	100	100	100	100	100	100	96.7	96.7	96.7
	400	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

There was a positive significant correlation between the percentage of correctly assigned individuals and a breed's average level of genetic differentiation (Fig 5.4; Spearman's rank correlation, rho = 0.635, p = 0. 0082).

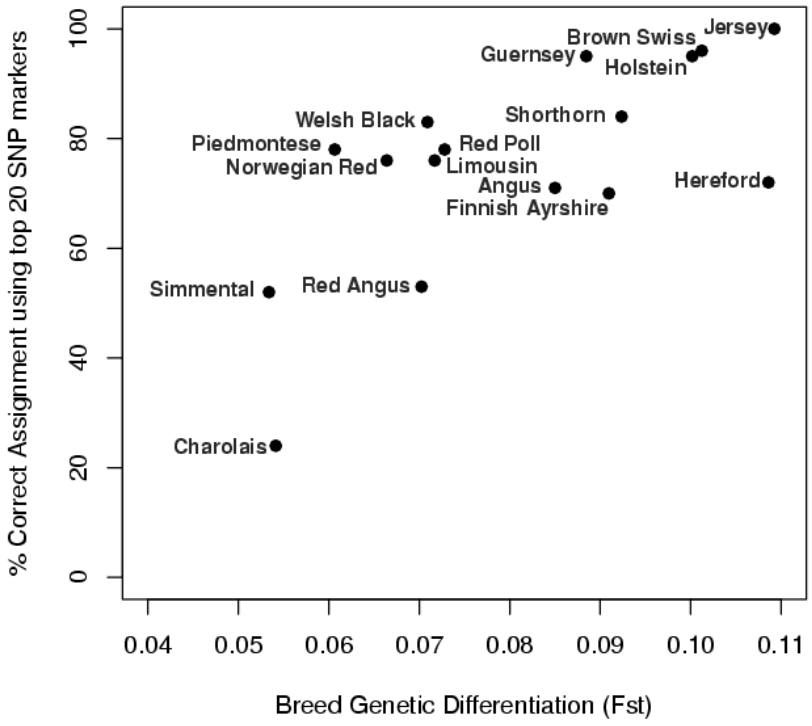


Figure 5.4 Scatterplot of percentage correct assignment using the top-ranked 20 SNP markers versus average pairwise breed genetic differentiation correlated (Wright's F_{ST} method; Spearman's rank correlation = 0.635).

Table 5.5 Error rates I and II for each breed following individual assignment analysis using SNP markers ranked by the pairwise Wright's F_{ST} selection method.

Error rate I		Breed of origin															
Allocated breed		A	BS	CH	FA	GU	HR	HO	JR	LIM	NR	PD	RA	RP	SH	SIM	WB
50 SNPs	Angus	85	0	0	0	0	0	0	0	0	0	0	13.3	0	0	0	0
	Brown Swiss	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Charolais	0	0	92	0	0	4	8	0	0	0	0	0	0	4	0	0
	Finnish Ayrshire	0	0	0	100	0	0	0	0	0	9.5	0	0	0	0	0	0
	Guernsey	0	0	0	0	95.26	0	0	0	0	0	0	0	0	0	0	0
	Hereford	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0	0
	Holstein	0	0	0	0	0	0	92	0	0	0	0	0	0	0	0	0
	Jersey	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
	Limousin	0	0	0	0	0	0	0	0	96	0	0	0	0	0	4	0
	Norwegian Red	0	0	0	0	0	0	0	0	0	90.5	0	0	0	0	0	0
	Piedmontese	0	0	0	0	4.8	0	0	0	4	0	100	0	0	0	4	3.3
	Red Angus	15	0	0	0	0	0	0	0	0	0	0	86.7	0	0	0	0
	Red Poll	0	0	0	0	0	4	0	0	0	0	0	0	100	0	0	0
	Shorthorn	0	0	0	0	0	0	0	0	0	0	0	0	0	92	0	0
	Simmental	0	0	8	0	0	0	0	0	0	0	0	0	0	4	92	0
	Welsh Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96.7
150 SNPs	Angus	100	0	0	0	0	0	0	0	0	0	0	6.7	0	0	0	0
	Brown Swiss	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Charolais	0	0	96	0	0	0	0	0	0	0	0	0	0	4	0	0
	Finnish Ayrshire	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
	Guernsey	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
	Hereford	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
	Holstein	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
	Jersey	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
	Limousin	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
	Norwegian Red	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
	Piedmontese	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0
	Red Angus	0	0	0	0	0	0	0	0	0	0	0	93.3	0	0	0	0
	Red Poll	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
	Shorthorn	0	0	0	0	0	0	0	0	0	0	0	0	0	96	0	0
	Simmental	0	0	4	0	0	0	0	0	0	0	0	0	0	0	100	0
	Welsh Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Error rate II		Breed of origin															
Allocated breed		A	BS	CH	FA	GU	HR	HO	JR	LIM	NR	PD	RA	RP	SH	SIM	WB
50 SNPs	Angus	95.4	0	0	0	0	0	0	0	0	0	0	4.6	0	0	0	0
	Brown Swiss	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Charolais	0	0	85.19	0	0	3.7	7.41	0	0	0	0	0	0	3.7	0	0
	Finnish Ayrshire	0	0	0	83.33	0	0	0	0	0	16.67	0	0	0	0	0	0
	Guernsey	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
	Hereford	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
	Holstein	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
	Jersey	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
	Limousin	0	0	0	0	0	0	0	0	96	0	0	0	0	0	4	0
	Norwegian Red	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
	Piedmontese	0	0	0	0	4.35	0	0	0	4.35	0	82.61	0	0	0	4.35	4.35
	Red Angus	35	0	0	0	0	0	0	0	0	0	0	65	0	0	0	0
	Red Poll	0	0	0	0	0	5.26	0	0	0	0	0	0	94.74	0	0	0
	Shorthorn	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
	Simmental	0	0	7.69	0	0	0	0	0	0	0	0	0	0	3.85	88.46	0
	Welsh Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
150 SNPs	Angus	98	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
	Brown Swiss	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Charolais	0	0	96	0	0	0	0	0	0	0	0	0	0	4	0	0
	Finnish Ayrshire	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
	Guernsey	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
	Hereford	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
	Holstein	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
	Jersey	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
	Limousin	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
	Norwegian Red	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
	Piedmontese	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0
	Red Angus	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
	Red Poll	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
	Shorthorn	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
	Simmental	0	0	3.85	0	0	0	0	0	0	0	0	0	0	0	96.15	0
	Welsh Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Two error rates of mis-assignment were calculated. Error rate I was calculated as the proportion of individuals of a breed that were assigned to another breed, instead of the breed of origin. Error rate II was calculated as the proportion of individuals assigned to a breed that actually originated from another breed. In Table 5.5a each column shows the error rate I indicating the proportion of individuals from a known breed origin that were allocated to each breed category (such that each column sums to 100%). In Table 5.5b each row shows error rate II indicating the proportion of individuals assigned to each breed category according to their known breed of origin (such that each row sums to 100%). Table 5.5a and 5.5b show error rates I and II that occurred in the individual assignment analysis, using pairwise Wright's F_{ST} at the lowest stringency threshold level (LLR > 0). With 50 SNP markers, five breeds (Brown Swiss, Finnish Ayrshire, Jersey, Piedmontese and Red Poll) were assigned with 100% assignment success, and the remaining breeds had an error rate I < 15%.

The error rate I was highest for Angus (14.6%), followed closely by Red Angus (13.3%); for these breeds, if an individual was not assigned to its correct origin it was assigned to the other breed. Using 50 SNP markers, eight breeds (Brown Swiss, Guernsey, Hereford, Holstein, Jersey, Norwegian Red, Shorthorn and Welsh Black) had no individuals assigned from other breeds, and the remaining breeds displayed a an error rate II of < 17% (except for the Red Angus breed, where 35% of the assigned individuals were Angus; and this may have been inflated by the relatively low sample size of Red Angus breed (15), compared to Angus (41)). By 150 markers, error rates I and II decreased to < 5% and occurred in the following breed pairs: Angus and Red Angus, Charolais and Simmental and Shorthorn and Charolais.

5.3.4 Ascertainment bias

The SNP markers on the BovineSNP50 beadchip were discovered through various breed sources. The majority of the markers were discovered from Angus, Holstein and Hereford breeds (others included Charolais, Jersey, Limousin, Norwegian Red , Red Angus and Simmental, but fewer SNPs were found through these breeds) (Matukumalli et al. 2009).

Table 5.6 Average minor allele frequency for each breed across the 40, 483 SNP markers, ranked by MAF in ascending order. [1] breed contribution to the SNP discovery process with '1' representing a 'major' contribution and '2' a 'minor' contribution.

	Breed	MAF	SNP discover process[1]
1	Jersey	0.196	2
2	Brown Swiss	0.199	
3	Guernsey	0.202	
4	Shorthorn	0.204	
5	Red Poll	0.215	
6	Red Angus	0.218	2
7	Finnish Ayrshire	0.219	
8	Welsh Black	0.221	
9	Norwegian Red	0.227	2
10	Angus	0.230	1
10	Limousin	0.230	2
10	Piedmontese	0.230	
11	Holstein	0.231	1
12	Hereford	0.236	1
13	Charolais	0.243	2
14	Simmental	0.244	2

Although Jersey was one of the breeds used for SNP discovery, it had the lowest average minor allele frequency (MAF) (Table 5.6). MAF values for Angus, Hereford

and Holstein were relatively high but lower than for Charolais and Simmental. The power of assignment at a breed level revealed that the breeds used for the SNP discovery process were not amongst those (except for Jersey) that required the fewest markers to achieve 100% assignment success (Table 5.5). The top 500 SNP markers were ranked by decreasing informativeness with their corresponding SNP discovery method (7 in total, (Matukumalli et al. 2009)). A X^2-test revealed that the proportions of SNP discovery methods represented in the pairwise Wright's F_{ST} 500 top SNP markers were not significantly different from those of the overall Bovine SNP50 set (X^2= 42, df = 36, p = 0.23).

5.4 Discussion

5.4.1 Behaviour of the marker selection methods

Similar levels of success in individual assignment of the European cattle breed genotypes were observed across the four selection methods, delta, pairwise Wright's F_{ST}, pairwise W&C's F_{ST} and PCA (Fig 5.3). This was likely due to an overall agreement between these selection methods as to which were the informative SNP markers (and uninformative markers). The resulting estimates of genetic informativeness of each SNP marker were highly correlated across the four selection methods, and there was a large degree of overlap among the top-ranked 500 SNP markers (Table 5.2). This was expected because all methods were employed using individual SNP marker allele frequencies. It has been demonstrated that delta and Wright's F_{ST} function similarly (Rosenberg et al. 2003). Wiener et al (2011), using the Bovine HapMap dataset, similarly observed a high correlation between delta and W&C's F_{ST} for the Holstein and Charolais breed pair. When more than two

populations are under consideration, PCA is certainly an attractive choice to determine marker informativeness because it provides an overall estimate for a SNP marker, as compared to other selection methods where it is necessary to estimate an average from pairwise calculations (Paschou et al. 2007). However, of the four selection methods, PCA exhibited the poorest correlation with the other methods and lowest overall individual assignment power (Fig 5.3, Table 5.3). The PCA approach developed by Paschou et al (2007) was recently applied to the Bovine HapMap dataset, consisting of 19 cattle breeds (12 European, 3 *Bos indicus* and 4 other taurine breeds), to identify ancestry informative SNPs (Lewis et al. 2011). By selecting PCA informative markers using individual genotypes, between 200 and 450 markers (1.5% of 30,000 SNPs) were needed to achieve 95% assignment accuracy for the 12 European cattle breeds in that study. This range of informative markers required for accurate individual ancestry prediction was far greater than that found using delta and the pairwise F_{ST} methods where between 55 and 160 (0.40% of 40,843 SNPs) were required to achieve the same level of accuracy in ancestry prediction for the 17 European cattle breeds (Fig 5.3, Table 5.3). PCA is an approach used to characterise the structure of a set of variables (in this case SNPs). The inferred relationships between objects (e.g., populations/breeds) are determined by the structure of the covariance matrix between the marker allele frequencies. Using the covariance matrix (instead of the correlation matrix) should select SNPs with large affects (i.e. large variance) and, thus, should select informative SNPs. However, an underlying assumption of PCA is that variables are correlated, therefore the informativeness of a given marker will depend on the other markers included in the analysis and this could influence the informative markers that PCA identified

(Paschou et al. 2007). In contrast, delta and F_{ST} do not take into account the relationships amongst markers and the level of information of each marker is estimated independently of the others.

Two other selection methods implemented here, global Wright's and W&C's F_{ST}, performed comparatively poorly in the individual assignment test. As similarly observed in another study (Kersbergen et al. 2009), global F_{ST} may not be appropriate to assess the level of genetic information in SNP markers when more than two populations are under consideration, as the method could result in the selection of SNP markers which are segregating in very few populations (Fig 5.2). As a result, the performance of individual assignment tests using global F_{ST} selected markers may be compromised compared to the other selection methods. Consequently, when more than two populations are under consideration, it is preferable to estimate F_{ST}, either Wright's or W&C's, on a population pairwise basis and then estimate the average across the pairwise comparisons to obtain an overall estimate for a marker.

5.4.2 Assignment success: individual breeds

Breaking down the accuracy of genetic identification of individuals to focus on specific breeds, it was evident that the number of markers required to achieve a high assignment success varied markedly across breeds, regardless of the selection method used (Table 5.4). For example, the Jersey, Brown Swiss, Guernsey and Piedmontese breeds achieved 100% assignment success, even at strict stringency

thresholds, using only 50 SNP markers. In contrast, the French beef breeds like the Charolais, Limousin and Simmental achieved around 90% assignment success at the lowest stringency threshold with 50 SNP markers, which fell to less than 50% with increasing stringency threshold.

Differences in the individual assignment success of breeds could be attributed to SNP ascertainment bias (Morin et al. 2004). Ascertainment bias can arise due to the inclusion of few breed sources during the SNP discovery process, such that represented breeds could show higher SNP variability and breeds not included in the SNP discovery process could have lower minor allele frequencies (MAF) (Decker et al. 2009; Kijas et al. 2009). Breeds with the lowest average MAF generally had the highest power of individual assignment, requiring the fewest markers to attain 100% assignment success (Table 5.4-5.6). The lowest average MAF values were found in breeds that were not a part of the SNP discover process (Table 5.6) suggesting that SNP ascertainment bias did have some effect on the individual assignment of certain breeds. However, the Jersey breed had the lowest MAF value yet was part of the SNP discovery process, though not central to the process. The highest average MAF was observed in breeds that contributed to the SNP discovery process, however, they were not central to it, like the three breeds, Angus, Hereford and Holstein, though their average MAF values were high. In addition, the average MAF values were variable amongst the breeds that had a minor contribution to the SNP discover process (Table 5.6). The results suggest that SNP ascertainment bias may have had some effect on the individual assignment of breeds, but the bias would certainly have been more pronounced if *Bos indicus* breeds had been included in this study (Decker

et al. 2009). In addition, no one particular SNP discovery method was over-represented in the top identified SNP markers as the discovery method proportions were similar to that represented on the Bovine SNP50 assay. Morin et al (2004) concluded that ascertainment bias may be an issue in the assessment of population size and demographic changes but that it has the least effect on individual identification and assignment tests, where the intentional selection of informative markers provides greater power than do randomly chosen markers.

It is generally considered that the number of markers required to obtain a high accuracy of individual assignment is influenced by the level of population genetic differentiation (Cornuet et al. 1999; Maudet et al. 2002). That is, the power of assignment success and variation in power of assignment between breeds depends closely on the populations under consideration and respective levels of genetic heterogeneity. As demonstrated in Figure 5.4, the level of genetic differentiation of a breed, measured by F_{ST}, was correlated with power of assignment success. Low breed genetic differentiation was observed in Charolais and Simmental, which similarly had higher error rates (Fig 5.4). False positive assignments also occurred between breeds of known recent ancestry, for example, Angus and Red Angus, and Finnish Ayrshire and Norwegian Red. In addition, cases of mistaken assignment occurred between Charolais, Simmental, Limousin and Shorthorn, where the pairwise F_{ST} values amongst these breeds were less than 0.1 (Table 5.5). Difficulty in discriminating the Charolais and Limousin breeds has even been encountered with polymorphic microsatellite markers (Ciampolini et al. 2006).

5.4.3 Informative marker panels in population genetics

Earlier studies on the population genetics of cattle breeds focused on the analysis of limited data from either microsatellites or SNPs (e.g. (Ciampolini et al. 2006; Maudet et al. 2002)). The identification of the origin of individuals using sparse markers was a means to characterise population genetic differentiation and did not focus on the extraction of informative markers. For instance, in a study of French cattle breeds, Maudet et al (2002) found that using 23 microsatellite loci, greater than 93% of individuals could be assigned to their breed origin and Negrini et al (2009) found that with 90 SNP markers genotyped in 24 European cattle breeds, 85% of individuals could be assigned to breed origin.

Dense genome-wide SNP marker arrays now allow the prospect of marker selection and the creation of marker panels. Evaluation of the selection methods revealed that only a small proportion of the markers from the BovineSNP50 beadchip were highly informative for discriminating among European cattle 17 breeds, and the majority contained medium to low levels of genetic information (Fig 5.1). This is consistent with the development of the assay in which SNPs with high MAF across *Bos taurus* breeds were preferentially selected in the assay design. Consequently, sets of randomly chosen SNP markers contained sufficient genetic information to produce moderate levels of individual assignment power (Fig 5.3). In contrast, the prior selection of informative SNP markers produced a reduced panel of highly informative markers with substantially more power thus achieving precise discrimination amongst the European cattle breeds. For instance, in a study on 6 *Bos*

taurus breeds STRUCTURE yielded the correct number of clusters in only 40% of cases when using 150 randomly chosen loci (from a dataset of 2,641 loci) (McKay et al. 2008). This is consistent with reduced assignment power for randomly-selected markers found in this chapter (Fig 5.3). The lower power was likely a direct consequence of using an insufficient number of informative loci.

Panels of informative markers are of value for studying evolutionary history and population structure. A reduced set of selected informative markers has been shown to effectively capture the genetic structure of human populations (Lao et al. 2006; Paschou et al. 2007). For instance, Lao et al (2006) found that 10 SNP markers from a 10K SNP array contained enough genetic information to differentiate individuals from Africa, Europe, Asia and America and additional loci contributed very little extra information. Indeed, it is generally considered that the inclusion of uninformative markers (i.e., monomorphic loci) may compromise performance in population genetic studies and add noise to the results (Liu et al. 2005; Smouse et al. 1982). It would also be useful to create a minimum panel of maximum power when elucidating population structure when using Bayesian genotypic clustering software such as STRUCTURE because these approaches are computationally demanding (which intensifies as the number of markers increases). In addition, genotype imputation can also be a computationally intensive exercise, especially with the high density SNP chips. Panels of informative markers could be used for genotype prediction with reduce computational cost. Consequently, it is practical and cost-effective to apply a selection method to a large-scale marker set to isolate the highly diagnostic markers and increase the power of analysis.

The more practical application of panels of informative makers ranges from tracing the origin of food products to tracking the illegal translocation of wildlife from one geographical location to another. The availability tailor-made genetic kits permit the routine regulatory use of genetic kits to expose fraudulent practices (Woolfe and Primrose 2004).

5.5 Conclusion

While the marker selection methods explored agreed to a large extent on which SNPs were the most informative and yielded reduced marker panels capable of breed identification, the power of assignment varied markedly among resulting ranked SNP panels, with delta and pairwise Wright's F_{ST} outperforming all other approaches. These results illustrate that with effective exploration of available high density genetic markers, it is possible to identify the most informative markers and produce an optimal minimum set of markers that can differentiate among populations.

CHAPTER SIX

Development of a DNA marker assay for the genetic verification of British traditional pig breed products

6.1 Introduction

Industrial consolidation of the agriculture sector as part of production improvement has led to the dominance of food products derived from very few sources. More recently, a growing appreciation and awareness of the potential diversity of food products and their origins has led the promotion of more local and less industrially derived food products. These less commercially available 'exotic' food products often attract a premium value.

In the livestock industry in Britain there has been a marked rise over the past decade in meat sold by breed name, with traditional British livestock breed products attracting a premium price. This trend is exemplified by British pork products and there are several contributing factors to explain the premium value of the traditional breed products and changing consumer preferences. The traditional pig breeds are slow growing relative to their commercial counterparts, increasing production costs. The traditional breeds also have relatively low population sizes and the rarity makes them a more valuable commodity (Table 6.1). In addition, traditional breeds possess certain meat qualities: high fat concentrations in the muscle and a fine muscle grain (Warriss et al. 1996). These physiological attributes may contribute to an enhanced eating experience and increased preference for traditional pig breed meat. The enriched quality is not going unnoticed in the food industry; it is becoming common to see pork products labelled with a traditional pig breed names on restaurant menus, in butchers, in supermarkets and at town farmers markets in Britain. For instance, Middle White pork is now a mainstay on the menus of top restaurants (BPA 2002).

The increasing population sizes of the traditional pig breeds also bear testimony to their rising popularity (RBST 2008). This trend has led to increased concerns over the authenticity of traditional breed meats, as the consumer is unlikely to be aware when substitution has taken place and fraud may therefore be perceived as a low risk crime. In addition to defrauding the consumer, breed mislabelling threatens the livelihoods of traditional breed farmers by undermining their brand and undercutting their prices through the illegal substitution with commercially mass-produced meat.

DNA-based analysis offers the possibility to identify animals and verify the origin of animal derived food products at the breed taxonomic level (Primrose et al. 2010; Teletchea et al. 2005; Woolfe and Primrose 2004). The ability to genetically authenticate the claimed origin of food products is well established and has led to its use by industry to self-regulate, by eco-labels to promote sustainability and by government authorities to monitor the food supply chain and enforce legislation (Primrose et al. 2010). A number of genetic studies have addressed the potential use of genetic markers for food authentication in livestock breeds through individual assignment analysis (Blott et al. 1999; Ciampolini et al. 2006; Negrini et al. 2009; Ramos et al. 2011), which have been important in laying the groundwork for the use of DNA analysis to expose fraudent breed-labelling practices. However, these studies in essence, have been explorative and discursive: illustrating that DNA markers can be applied to food forensics, but without leading to the actual development of specific genetic kits.

Whole genome sequencing, genotyping of genome-wide Single Nucleotide Polymorphism (SNP) markers and the availability of dense genome-wide SNP markers provided in SNP chips for many livestock species now permits the development of transferable and affordable genetic identification assays designed for regulatory purposes. The PorcineSNP60 beadchip (Ramos et al. 2009) can be exploited to verify British pig breed-labelled pork products and, in particular, samples allegedly originated from traditional pig breeds sold at a premium (Table 6.1).

With the aim of developing a genetic tool for the verification of meat from British traditional pig breeds for food authentication purposes, the objectives of this study were to: (1) select SNP markers that contain sufficient genetic information to be able to discriminate amongst the pig populations, (2) create a custom-made assay with an appropriate number of informative SNP markers, (3) demonstrate the effectiveness of the assay as a diagnostic tool, and (4) validate the application for product regulation.

6.2 Materials and Methods

6.2.1 Data

A total of 14 British pig breeds were used in this study (Table 6.1). The sample set comprehensively includes the two classification types of pig breeds (traditional and commercial) and majority of both breed types present in Britain (BPA 2002). Also included are the Meishan and Mangalica, two breeds of foreign origin that have been imported in high numbers to Britain (Table 6.1). By covering an almost complete spectrum of pig breeds present in Britain these dedicated samples have the potential

Table 6.1 The British pig breeds.[1] population status quantified by the number of breeding females: Vulnerable = 300; At Risk = 500; Minority = 1000; taken from Rare Breeds Survival Trust (http://www.rbst.org.uk/watch-list/pigs), [2] first imported to Britain from Hungary in 2006, [3] first imported to Britain from China in 1800s, [3] the sampling protocol is further described in section 6.2.1.

	Breed	Sample size	Type	Status[1]	Sampling[3]
1	Berkshire	73	Traditional	At Risk	PigBioDiv and USA
2	British Saddleback	30	Traditional	Minority	PigBioDiv
3	Duroc	31	Commercial		2 European and 2 USA populations
4	Gloucestershire Old Spots	24	Traditional	Minority	PigBioDiv
5	Hampshire	30	Commercial		PigBioDiv
6	Landrace	30	Commercial		3 European and 2 USA populations
7	Large Black	30	Traditional	Vulnerable	PigBioDiv
8	Large White	34	Commercial		3 European and 4 USA populations
9	Mangalica	26	European[2]		PigBioDiv
10	Meishan	24	Asian[3]		PigBioDiv
11	Middle White	30	Traditional	Vulnerable	PigBioDiv
12	Pietrain	21	Commercial		2 European and 1 USA population
13	Tamworth	30	Traditional	At Risk	PigBioDiv
14	Welsh	33	Traditional	At Risk	Welsh Pig Society
		446			

to be used as custom sets for future food authentication investigations and regulatory purposes in the country's porcine food industry. A total of 446 individuals were genotyped using the PorcineSNP60 beadchip (Ramos et al. 2009), which features ~60 000 SNPs with an estimated marker per 40kb across the pig genome. Breed sample sizes ranged from 24 (in Gloucestershire Old Spots and Pietrain) to 73 (in Berkshire), with an average of 32 individuals genotyped per breed (Table 6.1). The majority of the traditional breed DNA samples used in this study were previously extracted and genotyped using microsatellite loci as part of the PigBioDiv, whereby breed sampling constituted a pair of siblings from 25 litters in order to have 25 sires and 25 dams as unrelated as possible (SanCristobal et al. 2006a). Breed samples for the following commercial breeds, Duroc, Landrace, Large White and Pietrain, were obtained from several European and USA populations (Amaral et al. 2009). Additional samples in this study were collected from a separate Berkshire pig population in the U.S.A and Welsh pigs.

Loci selected for analysis had a call rate of at least 80% across all the British pig breeds and in total 59,436 SNP matched the call rate criterion. The individual multilocus genotypes were then used to identify genetically informative SNP markers and subsequently assess the genetic power of a selected panel of diagnostic markers chosen to create a custom-made genotyping multiplex assay.

6.2.2 SNP selection and assay development

The genetic informativeness of each SNP was measured by estimating delta from the allele frequency matrix (for further details see section 5.2.2 in chapter 5). The pairwise comparisons for each marker were averaged to obtain an overall estimate of the level of genetic information contained in each marker.

SNPs were subsequently ranked according to their delta value. To determine the numeric range of informative markers that would be appropriate for a custom-made GoldenGate Veracode™ multiplex assay, an individual assignment test was performed using cumulatively increasing numbers of top-ranked markers. A 'self-assignment' test, as described by Piry et al (2004), was performed in GENECLASS 2 using a partially Bayesian assignment method (Rannala and Mountain 1997). Prior to assignment testing of each individual, the observed allele frequencies of its respective reference population were re-estimated excluding the genotype in question, commonly referred to as the 'leave-one-out' validation method (Efron 1983). The likelihood of the multilocus individual genotypes occurring in each population was estimated based on their observed allele frequencies and an individual was assigned to a reference population for which it had the highest likelihood of assignment. If this was the known origin of the individual then the assignment test was deemed successful. This was a preliminary analysis to gauge the approximate number of markers that would be required and, consequently, the self-assignment test was used as it is straightforward to implement.

6.2.3 Assessment of the assay for breed genetic discrimination

The performance of the selected informative SNP markers as the diagnostic marker panel for a custom-made 96-plex assay was assessed. The extent of population genetic divergence of the reference populations based on this assay was evaluated using a combination of traditional population genetic statistics and individual-based methods.

Weir and Cockerham's (Weir and Cockerham 1984) unbiased estimator of Wright's fixation index (F_{ST}) was calculated between pairs of breeds using FSTAT 2.9.3 (Goudet 1995) (for further details see section 2.2.5 in chapter 2). Reynold's genetic distance (Reynolds et al. 1983) was calculated between pairs of breeds from the allele frequency matrix using GENDIST (Phylip v 3.67) (Felsenstein 1989) (for further details see section 4.2.2 in chapter 4). An unrooted neighbour-joining tree was constructed from the genetic distance matrix using the R package APE (Paradis et al. 2004 (for further details see section 2.2.4 in chapter 2). A total of 1000 bootstrap replicates were created in SEQBOOT, for each replicate Reynold's genetic distance was calculated between pairs of breeds in GENDIST and a consensus cladogram was calculated using CONSENSE (Phylip v 3.67) (Felsenstein 2008) (for further details see section 2.2.4 in chapter 2).

Population discrimination, group membership and levels of mixed ancestry in individuals were assessed using the Bayesian genotypic clustering method

implemented in BAPS (Corander et al. 2008) (for further details see section 2.2.2 in chapter 2).

An exclusion-simulation test using a partially Bayesian method (Rannala and Mountain 1997) was performed using GENECLASS 2 (Piry et al. 2004). For each reference population 10,000 independent individual genotypes were constructed from the observed allele frequencies. The likelihood that each simulated individual genotype was assigned to its respective reference population was calculated and a likelihood distribution for all 10, 000 simulated individuals for each reference population was constructed. The likelihoods of the individual genotypes were then compared to the distribution of likelihoods of simulated genotypes for each reference population. A critical rejection region (α) was implemented on the likelihood distribution such that an individual genotype was excluded from a population if the likelihood fell below the $\alpha * 10,000^{th}$ lowest value of the distribution. Unlike the self-assignment test, under the exclusion-simulation method an individual genotype may be excluded from all reference populations; hence, it does not require that the population of origin is sampled.

6.2.4 Power of the assay for pairwise breed discrimination

The power of breed assignment using the 96-plex assay was also assessed by calculating the probability that an animal of an assigned breed was actually from that breed rather than from another breed. This allowed an assessment of probabilities of correct assignment in specific breed comparisons and was undertaken in order to

represent a typical investigation in which there are specific claims and counter claims made concerning the breed origin of a pork product. The likely defence hypothesis that an observed individual genotype belongs to its designated breed origin (breed A) was tested against the likely prosecution hypothesis that an observed individual genotype may in actuality belong to another (breed B). If the defence hypothesis (that the observed individual genotype belongs to breed A) is rejected when it is in fact true, a Type I error has occurred (correct labelling undetected). If the defence hypothesis (that the observed individual genotype belongs to breed A) is accepted when it is in fact false, a Type II error has occurred (mislabelling undetected). Using these error rates, the posterior probability that a product is actually from breed A (its claimed breed origin) instead of from breed B can be estimated (Ciampolini et al. 2006). In brief, the log-likelihood that an individual originated from each breed was estimated in GENECLASS2 (Piry et al. 2004) as above and the log-likelihood ratio (log(LR)) of an individual originating from breed A versus breed B was calculated. The means and standard deviations of the observed log(LR) distributions were calculated and the false positives (α) and true positives ($1 - \beta$) were obtained for test values log(LR) > 0 and log(LR) > 2. Thus, the log(LR) of a positive result was estimated as the ratio between the likelihood of having a true positive result against the likelihood of having a false positive result: $(1 - \beta)/ \alpha$, which gives the odds that the claimed breed origin (breed A) is correct when a test is positive. The posterior probability that an individual actually originated from breed A given the alternative hypothesis that it originated from breed B, assuming equal priors, was calculated as follows: $(1 - \beta) / \alpha) / ((1 - \beta) / \alpha) + 1)$, which represents the proportion of individuals from claimed breed origin (breed A) correctly testing positive.

6.2.5 Validation of the assay using independent samples

Following selection of a panel of 96 SNP markers, a custom GoldenGate Veracode™ multiplex assay was designed and tested to assess its performance across a range of samples. This assay was then produced and run against a set of control samples. Three sets of independent samples were used to validate and test the assay:

i) Control DNA from 70 samples from target breeds and comparative breeds at a concentration of 50 ng/µl (Table 6.1). These were included to demonstrate the ability of the assay to correctly assign samples to their breed origin.

ii) Processed/treated meat samples. These were included to examine the performance of the assay across a range of sample types, including various cooking methods (fried, baked, boiled, grilled, cooked in sauce). Samples were obtained from the market sources (see below).

iii) Market/commercial samples sold by named breed. These were included as a final examination of how the assay would perform using market samples and to take an initial look at what breeds could be identified from a small sample of traditional breed products on sale in the UK. Samples of pork meat (pork chops unless otherwise stated) labelled by breed were purchased from 26 specialist suppliers and one supermarket by Minton Treharne & Davies Ltd, a Welsh Public Analyst involved in validating the assay. Names of individual suppliers are subject to confidentiality.

DNA from all samples was extracted using the Qiagen DNEasy Blood and Tissue kit following the manufacturer's instructions and initially normalized to 50 ng / ul as

suggested for the GoldenGate Veracode™ assay. DNA was then processed following the Illumina protocol and the data analysed using the proprietary Genome Studio software. Individual genotypes were exported for assignment analysis in GENECLASS2 (exclusion-simulation tests), as described in section 6.2.3.

6.3 Results

6.3.1 Selection of markers for a breed informative panel

The power of the individual assignment test with cumulatively increasing number of top-ranked informative SNP markers is presented in Figure 6.1.

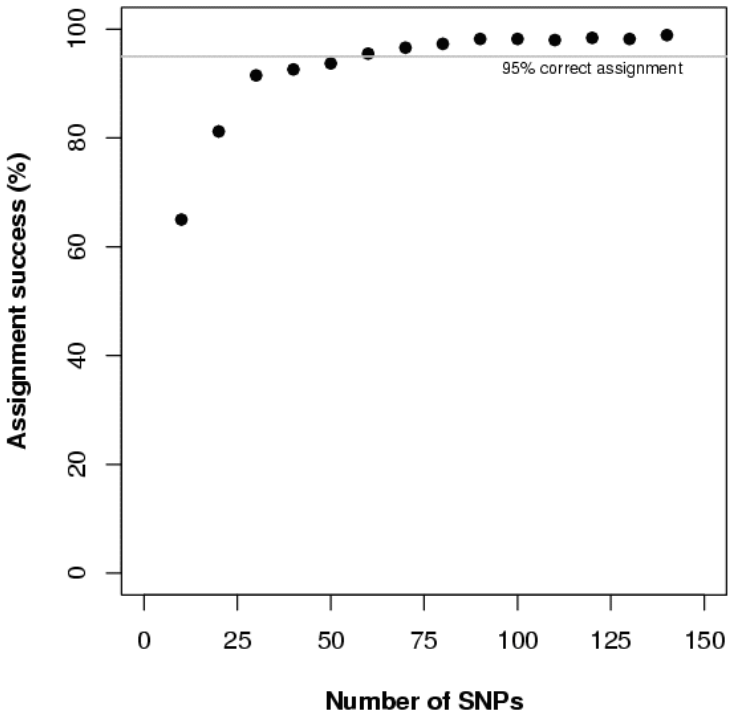


Figure 6.1 Plot of the individual assignment success for cumulatively increasing numbers of top-ranked informative SNP markers.

With the top-ranked 50 SNP markers 93.7% of the individual genotypes (418) were correctly assigned. Correct assignment increased to 95% (426) with 60 SNP markers. For 90 SNP markers, an individual assignment accuracy of 98.2% (438) was attained. The 8 incorrectly assigned individuals involved the following breed pairs: British Saddleback and Large Black (3), Landrace and Large White (1), Landrace and Welsh (3) and Middle White and Large White (1). For 140 SNPs, 98.9% (441) of the individual genotypes were correctly assigned (Fig 6.1). The 5 incorrectly assigned individuals involved Landrace and Welsh (4) and Middle White and Large White (1). Given the observed plateau of assignment success beyond 100 SNPs (Fig 6.1), the top 96 SNP markers were selected to form a marker panel for the subsequent production of a 96-plex genotyping assay.

The genomic distribution of the final 96 SNPs is given in Table 6.2. As can be seen, the informative SNP markers were located on all chromosomes except for 2, 9, 10 and 18. The number of SNP markers selected from chromosomes ranged from 1 on chromosomes 12 and 17 to 25 on chromosome 8, with an average of 4 selected SNP markers per chromosome. A remaining 20 SNP markers were selected that have yet to be mapped to the porcine genome.

A disproportionately large number of SNPs were located on chromosome 8 (Table 6.2). Paschou et al (2007) observed that panels of informative SNPs selected from genome-wide arrays tend to contain a large number of markers that are in high linkage disequlibrium (LD). This introduces redundant information into a panel because markers in complete LD will contain the same genetic information. The

extent of LD between the 25 SNPs mapped to chromosome 8 was explored using Haploview (Barrett et al. 2005). Out of 600 marker pairs, 18 pairs exhibited moderate to high levels of LD in one or more pig breeds ($r^2 > 0.4$; Figure 6.2). The high levels of LD for each of the 18 marker pairs were not present in all 14 pig breeds, indicating that though a given marker pair may contain redundant information for one breed that it is not necessarily the case for another breed.

Table 6.2 Properties of the 96 SNP panel. [1] the average distance between pairs of markers is provided with the minimum and maximum distance between pairs of markers in brackets.

Chromosome	Occurrences	distance (bp)[1]
1	9	84,849,030 (97,690-221,777,480)
2	0	n.a.
3	2	10,181,626
4	4	20,026,279 (215,501-33,004,030)
5	5	4,864,199 (47,779-12,111,040)
6	2	18,925,439
7	6	8,294,008 (142,325-23,742,753)
8	25	14,218,857 (18,524-71,017,493)
9	0	n.a.
10	0	n.a.
11	4	15,622,017 (18,731-28,458,841)
12	1	n.a.
13	4	27,645,655 (39,053-55,278,293)
14	3	522,828.7 (127,875 – 784,243)
15	5	32,778,555 (91,640-81,029,219)
16	3	654,860 (70,990 – 982,290)
17	1	n.a.
18	0	n.a.
X	2	5,045,381
Undetermined	20	n.a.

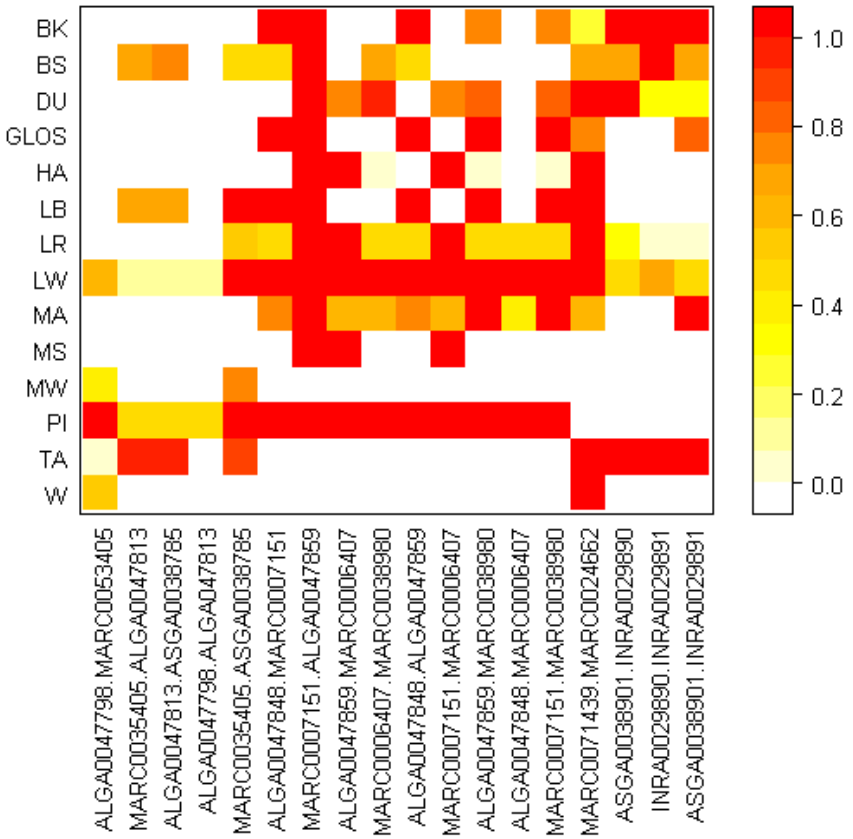


Figure 6.2 Level of linkage disequlibrium (LD), measured using r^2, between the 25 markers on chromosome 8 for each pig breed. r^2 represents the correlation of allele frequencies between two loci such that SNPs in complete LD have a value of 1. The darker the colour. the higher the LD with white indicating no LD between a pair of SNPs.

6.3.2 Assessment of the assay for breed genetic discrimination

Based on the reference data, the average pairwise breed genetic differentiation (F_{ST}) using the 96-SNP panel was 0.54 (Table 6.3). The genetic differentiation (F_{ST}) between pairs of breeds ranged from 0.10 (Landrace vs Welsh) to 0.82 (Hampshire vs Meishan), with average breed F_{ST} values ranging from 0.39 for British Saddleback to 0.71 for Meishan.

Table 6.3 Population genetic differentiation among 14 pig breeds using 96 SNP markers. The lower diagonal contains the pairwise genetic differentiation between breeds estimated using Weir & Cockerham's F_{ST} (Weir and Cockerham 1984). The column on the far-right of the table presents the average breed F_{ST} and the standard deviation.

	Breed	BK	BS	DU	GLS	HA	LR	LB	LW	MA	MS	MW	PI	TA	W	F_{ST}
1	Berkshire															0.51 (0.10)
2	British Saddleback	0.29														0.39 (0.08)
3	Duroc	0.42	0.36													0.53 (0.08)
4	Gloucestershire Old Spots	0.47	0.43	0.56												0.62 (0.12)
5	Hampshire	0.51	0.45	0.60	0.75											0.64 (0.10)
6	Landrace	0.52	0.32	0.44	0.64	0.63										0.45 (0.19)
7	Large Black	0.39	0.23	0.50	0.43	0.56	0.53									0.50 (0.11)
8	Large White	0.56	0.35	0.50	0.67	0.60	0.19	0.55								0.46 (0.18)
9	Mangalica	0.51	0.40	0.57	0.68	0.63	0.53	0.52	0.50							0.58 (0.10)
10	Meishan	0.65	0.55	0.69	0.71	0.82	0.71	0.57	0.71	0.78						0.71 (0.08)
11	Middle White	0.64	0.45	0.60	0.75	0.69	0.36	0.61	0.22	0.63	0.78					0.56 (0.16)
12	Pietrain	0.63	0.46	0.59	0.77	0.71	0.30	0.64	0.33	0.63	0.81	0.47				0.57 (0.18)
13	Tamworth	0.45	0.43	0.53	0.61	0.66	0.62	0.45	0.64	0.64	0.75	0.70	0.74			0.61 (0.11)
14	Welsh	0.55	0.37	0.48	0.67	0.65	0.10	0.57	0.23	0.56	0.73	0.43	0.33	0.65		0.49 (0.19)

Reynolds' pairwise genetic distance ranged from 0.34 between British Landrace and Welsh to 0.91 between Hampshire and Meishan. Average pairwise genetic distance across all breeds ranged from 0.63 for British Saddleback to 0.85 for Meishan. The phylogenetic reconstruction of breed relationships is shown in Figure 6.3 (bootstrap support > 50% indicated). There was high support for a clade of five white-skinned breeds (Landrace, Large White, Middle White, Pietrain and Welsh) with additional support of the depicted branching within the clade. For the remaining breeds, there was overall low bootstrap support for the depicted genetic relationships.

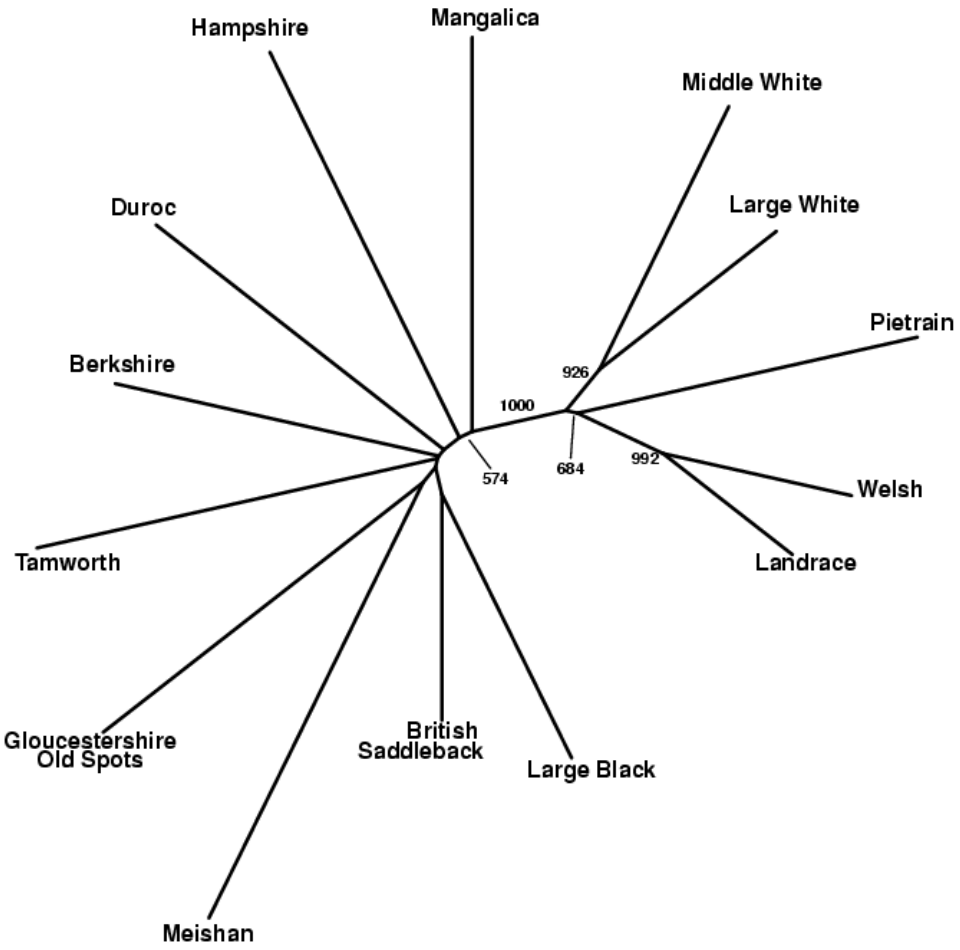


Figure 6.3 Phylogenetic reconstructions of the British pig breeds using Reynold's genetic distance. Bootstrap support values greater than 50% are indicated.

The results of the BAPS analysis are presented in Figure 6.4. Given that there are 14 pig breeds sampled in this study, if all breeds were genetically distinct entities each pig breed would form an independent homogenous cluster for K = 14.

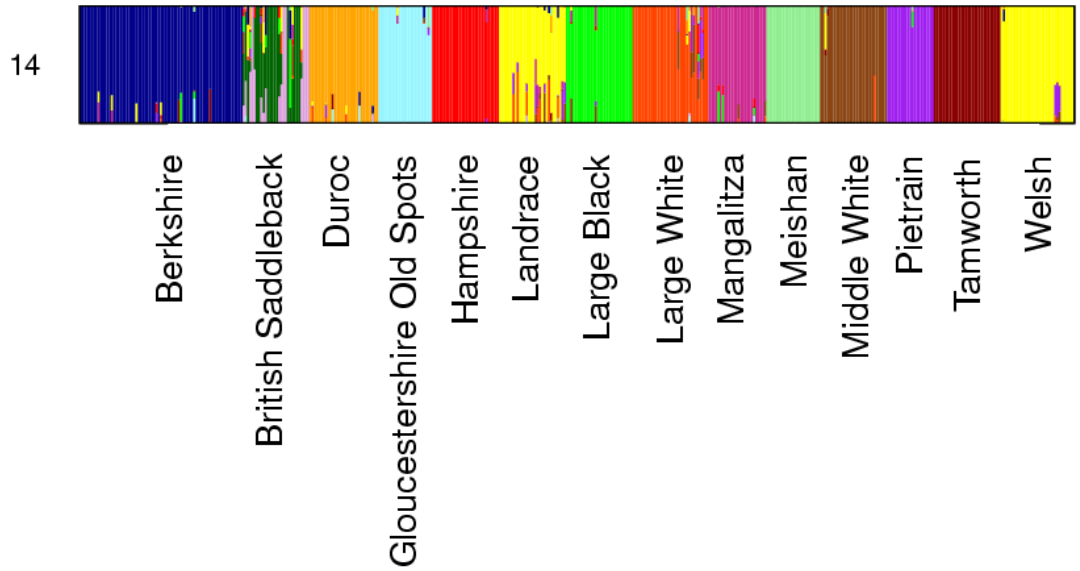


Figure 6.4 Individual assignment based on BAPS analysis at K = 14. The histogram demonstrates the proportion of each individual's genome that originated from each of populations. Each individual is represented by a vertical line corresponding to its membership coefficient (q).

However, at K = 14 the individuals of the Landrace and Welsh breeds clustered together, whilst the British Saddleback was split into two clusters. The other pig breeds were essentially distinct homogenous populations, with minimal individual genetic admixture. Large White and Middle White clustered together until K = 14, at which point they split to form separate clusters. The genetic subdivision in the

British Saddleback breed was observed from K = 9. At K = 15, the Landrace and Welsh breeds still clustered together whilst Berkshire individuals split over two groups (mirroring the sampling of two geographic origins: USA and UK). Landrace and Welsh split at K = 16 to form two distinct clusters. A plot of the posterior likelihood of K values produced an asymptotic curve with a plateau that started at K = 15 and extended to K = 20 (at K > 16 the different populations within the commercial breeds split).

The exclusion-simulation test results are presented in Table 6.4. At a critical rejection region (α) of 0.001, 99.1% (442) of the individual genotypes were not excluded from their reference population origin. The individuals excluded from their presumed origin (one each from Hampshire, Landrace, Large White and Pietrain breeds) were also excluded from all other reference populations.

Table 6.4 Exclusion-simulation analysis of the reference populations for α = 0.001.

	Breed	Number of samples excluded from claimed breed origin
1	Berkshire	0 / 73
2	British Saddleback	0 / 30
3	Duroc	0 / 31
4	Gloucestershire Old Spots	0 / 24
5	Hampshire	1 / 30
6	Landrace	1 / 30
7	Large Black	0 / 30
8	Large White	1 / 34
9	Mangalica	0 / 26
10	Meishan	0 / 24
11	Middle White	0 / 30
12	Pietrain	1 / 21
13	Tamworth	0 / 30
14	Welsh	0 / 33

6.3.3 Power of the assay for pairwise breed discrimination

The posterior probability that any individual with a log likelihood ratio greater than a given threshold originated from the claimed breed origin rather than from another specified breed, was calculated for all breed pairs at two thresholds (log(LR) > 0 and log(LR) > 2). At the test value of log(LR) > 0 the posterior probability of correct assignment was > 99.5% in 174 of the 182 and > 99.9% in 172 out of the 182 contrasts (Table 6.5a). A posterior probability of correct assignment of < 99.5% of individuals to claimed breed was only observed in 4 breeds: Landrace, Large Black, Large White and Welsh. The remaining 10 breeds had a high level of assignment evident across their 13 pairwise contrasts (Berkshire, British Saddleback, Duroc, Gloucestershire Old Spots, Hampshire, Mangalica, Meishan, Middle White, Pietrain and Tamworth). Three contrasts had a posterior probability of correct assignment < 99.0% at the test value of log(LR) > 0: Large White against Landrace (0.97), Landrace against Welsh (0.97) and Welsh against Landrace (0.91) (Table 6.5a). At the test value of log(LR) > 2 the posterior probability of correct assignment was > 99.5% and > 99.9% in 175 of the 182 contrasts and 173 out of the 182 contrasts, respectively (Table 6.5b). There were 2 contrasts with a posterior probability of < 99.0% at test value of log(LR) > 2: Large White against Landrace (0.98) and Welsh against Landrace (0.95) (Table 6.5b). The lowest posterior probability of assignment at both log(LR) test values was the Welsh against Landrace contrast (Table 6.5a,b).

Table 6.5a The posterior probability any individual with log(LR) > 0 originates from the claimed breed.

	Contrasted breed													
Claimed breed	BK	BS	DU	GLOS	HA	LR	LB	LW	MA	MS	MW	PI	TA	W
Berkshire	-	0.999946	1	1	1	1	1	1	1	1	1	1	1	1
British Saddleback	1	-	1	1	1	0.999997	0.999695	0.999612	0.999927	1	1	1	1	1
Duroc	1	0.999998	-	1	1	1	1	1	1	1	1	1	1	1
Gloucester Old Spot	1	1	0.999998	-	1	1	1	1	1	1	1	1	1	1
Hampshire	1	1	1	1	-	1	1	1	1	1	1	1	1	1
Landrace	1	0.991010	0.999570	1	1	-	1	0.998634	1	1	0.999989	0.999992	1	0.971693
Large Black	0.999968	0.990014	0.999968	1	1	1	-	1	0.999889	1	1	1	1	1
Large White	1	0.992750	0.999633	1	1	0.972057	1	-	1	1	0.994449	1	1	0.999673
Mangalica	1	0.999991	1	1	1	1	1	1	-	1	1	1	1	1
Meishan	1	1	1	1	1	1	1	1	1	-	1	1	1	1
Middle White	1	0.999747	1	1	1	0.999819	1	0.999401	1	1	-	1	1	1
Pietrain	1	0.999739	0.999967	1	1	0.999107	1	0.999805	1	1	1	-	1	0.999817
Tamworth	0.999996	1	0.999978	1	1	1	1	1	1	1	1	1	-	1
Welsh	0.999997	0.993296	0.998325	1	1	0.907609	1	0.999993	1	1	1	1	1	-

Table 6.5b The posterior probability any individual log(LR) > 2 originates from the claimed breed.

	Contrasted breed													
Claimed breed	BK	BS	DU	GLOS	HA	LR	LB	LW	MA	MS	MW	PI	TA	W
Berkshire	-	0.999976	1	1	1	1	1	1	1	1	1	1	1	1
British Saddleback	1	-	1	1	1	1	0.999910	0.999811	0.999965	1	1	1	1	1
Duroc	1	1	-	1	1	1	1	1	1	1	1	1	1	1
Gloucester Old Spot	1	1	1	-	1	1	1	1	1	1	1	1	1	1
Hampshire	1	1	1	1	-	1	1	1	1	1	1	1	1	1
Landrace	1	0.993502	0.999727	1	1	-	1	0.999662	1	1	0.999997	1	1	0.991931
Large Black	0.999985	0.993539	0.999981	1	1	1	-	1	0.999929	1	1	1	1	1
Large White	1	0.994595	0.999749	1	1	0.984738	1	-	1	1	0.998120	1	1	0.999923
Mangalica	1	0.999995	1	1	1	1	1	1	-	1	1	1	1	1
Meishan	1	1	1	1	1	1	1	1	1	-	1	1	1	1
Middle White	1	0.999824	1	1	1	0.999913	1	0.999814	1	1	-	1	1	1
Pietrain	1	0.999813	0.999977	1	1	0.999549	1	0.999912	1	1	1	-	1	0.999922
Tamworth	0.999998	1	0.999987	1	1	1	1	1	1	1	1	1	-	1
Welsh	1	0.994867	0.998787	1	1	0.945844	1	1	1	1	1	1	1	-

6.3.4 Validation of the assay using independent samples

The results of the validation analysis using three sets of independent samples are presented in Table 6.6.

Table 6.6 Exclusion-simulation analysis of the independent test samples.

		Number of samples excluded		
	Breed	Test	Cooked	Market
1	Berkshire	0 / 5	-	1 / 6
2	British Saddleback	0 / 5	-	0 / 3
3	Duroc	0 / 5	-	-
4	Gloucestershire Old Spots	0 / 5	0 / 6	2 / 10
5	Hampshire	0 / 5	-	8 / 8
6	Landrace	2 / 5 (to Welsh)	-	-
7	Large Black	0 / 5	-	0 / 1
8	Large White	1 / 5 (to Middle White)	-	-
9	Mangalica	0 / 5	-	-
10	Meishan	0 / 5	-	-
11	Middle White	0 / 5	0 / 6	0 / 3
12	Pietrain	0 / 5	-	-
13	Tamworth	0 / 5	0 / 6	0 / 4
14	Welsh	0 / 5	0 / 6	0 / 5

For each of the 70 control samples, 90 polymorphic SNP markers were unambiguously genotyped from the 96-plex assay (two SNPs failed to amplify and another four were monomorphic). In the validation analysis, 96% of the control samples were assigned to breed origin (Table 6.6). Only two breeds did not attain 100% assignment success, Landrace (2) and Middle White (1), for which test samples were assigned to Welsh and Large White, respectively.

The performance of the 96-plex assay following various cooking treatments (fried, baked, boiled, grilled, baked in sauce) showed correct assignment of all samples to their five breeds of origin, although the genotyping success rate (SNPs per sample) fell to a minimum of 88% (Table 6.6).

Out of 45 market samples, the individual assignment analysis resulted in 2 samples not assigned to claimed breed origin but assigned to another breed, indicating possibly mislabelled meat (1 claimed Gloucestershire Old Spot and 1 claimed Hampshire sample) (Table 6.6). While all 8 Hampshire samples were excluded from the Hampshire reference population, 7 out of 8 samples were not assigned to any other breed.

6.4 Discussion

6.4.1 Development of the assay

The objective of this study was to develop a custom-made diagnostic genetic tool for the authentication of products originating from traditional British pig breeds and future regulation in the British porcine food industry. The availability of robust genotyping systems, where users can design their own multiplex assays using existing genetic markers, conveniently allows the achievement of this goal. In this study the GoldenGate Veracode™ system was used to develop the assay and certain pre-defined multiplex sizes were available: 48-, 96-, 144-, 192- and 384-plex. Careful analysis of the large number of markers available from the PorcineSNP60 beadchip indicated that the 96-marker assay would be sufficient to achieve a high level of assignment power. It was our assessment that more than 96 SNP markers did

not sufficiently enhance the power of individual assignment analysis to warrant the development of a 144-plex assay for pork product authentication (Fig 6.1).

6.4.2 The genetic power and utility of the assay

It is important to establish whether the both the sampling of genetic markers for the 96-plex assay and individuals for the British pig breeds were adequate, such that the developed assay and set of reference populations can be repeatedly used for future porcine food authentication. An earlier study using a panel of 50 microsatellites showed that European pig breeds are generally highly distinct populations (SanCristobal et al. 2006a). One biological factor that could influence the levels of genetic differentiation amongst populations is hybridisation (cross-breeding). Bayesian genotypic clustering analysis indicated that very few individuals showed evidence of shared genetic ancestry in the British pig breeds (Fig 6.4). The lack of evidence of genetic admixture within most populations and the genetic homogeneity of British pig breeds is consistent with previous work using microsatellite markers presented in chapter 2. Strict breeding practices in Britain appear to maintain the genetic distinction of the pig breeds. This was further substantiated in this study where population genetic estimates demonstrated that the 96-plex assay was a highly effective selection of markers as it was able to genetically discriminate the British pig breeds. As can be seen in Figure 6.3, the predominantly long branches of breeds coupled with the high reported F_{ST} values are indicative of high breed genetic differentiation (Table 6.3). As a result of prior SNP selection, the 96-plex assay captured a large proportion of the genetic variation between the British pig breeds, with estimates of F_{ST} exceeding those previously reported using a standard diversity

panel of 50 microsatellite loci (SanCristobal et al. 2006a). Although the high F_{ST} estimates of the SNPs on the 96-plex assay could be due to the process of random genetic drift, locus-specific breed genetic differences could also be a result of past artificial selection. A large proportion of the genetically informative SNPs were found on chromosome 8 (SSC8) (Table 6.2), which harbours the *KIT* gene, a locus involved in coat colour variation in domestic pig breeds. High linkage disequilibrium (LD) between some of these markers, especially in the commercial Large White and Pietrain breeds (Fig 6.2), could be a signature reflecting positive selection. This is in agreement with a recent genome wide study of commercial pig breeds in which low nucleotide diversity was found in regions of SSC8 (Amaral et al. 2011). High bootstrap support for the clustering of the white-skinned breeds and the Pietrain using phylogenetic reconstruction in the current study (Fig 6.3) was probably due to the selection of informative SNPs that are also associated with the *KIT* gene. Markers that show high breed differentiation due to positive selection for breed-specific characteristics may also be highly informative for breed assignment analyses.

The power of the individual assignment tests provided an indication that the breadth of actual genetic variation within each of the British pig breeds has also been effectively captured. That is, with sufficient numbers of individuals sampled, the estimated allele frequencies will provide a reasonable estimate of the actual population allele frequencies and, as a result, the individual assignment tests should perform well. The vast majority of the test samples used to validate the 96-plex assay were unambiguously authenticated, supporting the notion that the sampled breed populations are good representatives of the breeds (Table 6.6). The validation step

was a vital exercise, not only to test the effectiveness of the SNP panel and the suitability of the reference population data, but also to demonstrate the application of the assay by a UK public analyst on case-type samples. It supported the accuracy and performance of the previous assignment tests and the overall low error rate indicated that the sampled British pig breed populations are genetically representative of the actual populations. The one possible exception to this was the observed lack of assignment in market samples of Hampshire. While it is not possible to determine if the failure was due to insufficient sampling of genetic variation within the reference population or mislabelled test samples, in many countries the male Hampshire is often used to sire cross-bred pigs (BPA 2002) and this practice could have altered the genetic composition of the breed to an extent that the reference Hampshire population (sampled in 1999) is not a good representative of the contemporary breed population.

Although the prior selection of genetically informative markers allowed a high rate of correct assignment there were, nonetheless, a few instances of incorrect assignment of individuals. However, this was concentrated to a few breed pairings: the majority of the incorrectly assigned individuals were between the Landrace and Welsh breeds (Table 6.5-6.6). Relatively low genetic differentiation was observed between Landrace and Welsh with the 96-plex assay (Table 6.3, Fig 6.3). It would not be surprising to the pig breeding community that a close genetic relationship was observed between these two morphologically similar breeds. Dwindling numbers of the Welsh in the mid-20th century resulted in the introduction of Landrace blood to boost the breed population size (Porter 1993) and today the two breeds look

remarkably similar. The results from this study show that the 96-plex assay does not allow differentiation of Welsh and Landrace pigs with sufficient accuracy for authenticity testing. Incorrect assignment also occurred in one case between Large White and Middle White (Table 6.6). Close genetic relationships between breeds need to be carefully considered in product authentication.

6.4.3 The British pig breed market

The diversity of British pig breeds, expanding consumer preference and disparity in price between pork products create the potential for the substitution of labelled breed names in this food market. The conceivably profitable scenario of labelling a pork product with a traditional breed name when it actually originated from another source can be readily exploited. Therefore, it is in the interests of the food industry and consumer confidence to be able to verify traditional pig breed labelled products.

The 96-plex assay has the ability to authenticate pork products labelled with traditional breed names and thus expose the mislabelled products. The levels of individual assignment accuracy were extremely high in the traditional breeds for both the reference populations and the test samples (Table 6.4, Table 6.6). More importantly, except for the Landrace/Welsh pairing, very few (commercial breed) individuals were falsely assigned to a traditional breed. Therefore, there is a high likelihood that an individual assignment test would assign a sample that was correctly labelled with a traditional pig breed name to that breed origin. Consequently, there was an extremely high probability of correct assignment for

majority of the traditional pig breeds: Berkshire, British Saddleback, Gloucestershire Old Spots, Large Black, Middle White and Tamworth, particularly when contrasted against the other breeds (Table 6.5). Given the scenario that a food product labelled with one of these traditional pig breed names is in fact derived from another source then the probability of detecting such a swap is high.

Furthermore, the validation step of this study revealed a high level of breed label conformity across a range of samples tested for the traditional British pig breeds (Table 6.6). The molecular technology of the 96-plex assay can be confidently applied to not only raw samples, but also meat subjected to various cooking treatments which is particularly relevant to verifying claims made on restaurant menus.

The power of the 96-plex assay as a genetic tool for British pig breed product authentication was only really compromised when confronted with Landrace and Welsh breed pair, as indicated by the notably reduced posterior probability of correct assignment (Table 6.5). A lower posterior probability of assignment of Welsh samples was obtained due to the relatively higher proportion of Landrace individuals falsely assigned to the former breed. These results are in concordance with the double cross-validation analysis in which two out of five Landrace individuals were assigned to the Welsh breed (Table 6.6).

This study illustrates the potential of the 96-plex assay to authenticate the origin of pork products labelled with traditional pig breed names. However, although

commercial breed types were included in this study, in general commercially produced meat does not normally originate from purebred animals. Instead, commercial pork products are usually derived from lines that represent a broader cross of multiple from, perhaps including genetic components from traditional breeds. Although the 96-plex assay may be powerful at discriminating traditional pig breed from commercial pork products, actual samples from these crosses from a range of companies would need to be incorporated. This would then conclusively demonstrate that traditional pig breeds products may be discriminated from commercial pork products and validate the applicability of this genetic tool in the pork industry.

6.5 Conclusion

The false labelling or mis-description of food is considered prevalent in the industry and the need to authenticate product origin is a long-standing challenge. The development of the 96-plex assay using markers available from the PorcineSNP60 beadchip will contribute to on-going product authentication and future regulation in the British food industry. This genetic tool provides a powerful method for authenticating products claimed to originate from traditional pig breeds.

CHAPTER SEVEN

General Discussion

7.1 Thesis motivation and objectives overview

Molecular characterisation of breeds is of primary interest for animal breeders. Aside from lending a perspective on the historical development of breeds, genetic surveying of breeds can contribute to conservation initiatives of livestock diversity. This thesis concentrated on the characterisation of genetic diversity, structure and individual genetic admixture in livestock breeds. It also focused on breed identification with aim of developing panels of informative genetic markers for breed verification.

7.2 Conclusions, relevance of findings and implications

7.2.1 Genetic diversity and structure of livestock breeds

This research first evaluated the performance of individual-based population genetic clustering tools at characterising population structure using an empirical dataset of British pig breeds. As was described in **chapter 2**, inconsistent results between the Bayesian genotypic clustering methods were observed in terms of detection of breed substructure, detection of individual genetic admixture and determining the number of underlying populations (K). Of the Bayesian genotypic clustering methods, BAPS detected known structuring not only between the British and French Meishan, but also within the British population whilst STRUCTURE detected substructure in the British Saddleback breed. Low genetic differentiation can cause the Bayesian genotypic clustering algorithms to perform poorly (Waples and Gaggiotti 2006) and variability in the inference of the more weakly differentiated subpopulations may have been because the different algorithms were operating at their extremes. With

regards to determining K, the Bayesian methods were not in agreement over the optimal K value where BAPS detected finer genetic differentiation resulting in a higher number of detected populations that were all biologically credible. In contrast, in **chapter 3** both BAPS and STRUCTURE concurred that there were two underlying populations in the British Saddleback and the two approaches identified the same two subpopulations in the breed. Li et al (2011) also found that with a dataset of 7 cattle populations the same number of populations could be determined using both BAPS and STRUCTURE. The nature of the population genetic structure is another factor that could affect the behaviour of different clustering (Waples and Gaggiotti 2006). The genetic histories of livestock breeds can be complex due to an assortment of processes that may have occurred during breed development. Hence, the variability in the ascertainment of population structure of the British pig breed with the different Bayesian genotypic clustering approaches could have been due to greater complexity in the dataset (**chapter 2**), compared to the simpler structure present in the dataset of the British Saddleback breed (**chapter 3**). Since population structure can be complex it may be more feasible and constructive to conduct Bayesian genotypic clustering analyses on identified subgroups within a large population dataset (e.g. analysis presented in **chapter 3**). For instance, the uppermost hierarchical layer of the population structure could be identified using delta K (Evanno et al. 2005) and the genetic structuring within the inferred subgroups could then be elucidated separately.

Based on the inconclusive genetic clustering results of the British Saddleback reported in **chapter 2** and its known history the genetic structure of the breed was

further characterised in **chapter 3**. Two genetically differentiated subpopulations were identified (F_{ST} = 0.084) and were found to correspond with herd, where one subpopulation consisted of individuals from the Rainbarrow herd and the other of individuals that originated from other sampled herds. In **chapter 4** a heterozygote deficit was observed in a majority of the British traditional chicken breeds and subsequent individual-based clustering analyses also revealed genetic substructure in many breeds. For two breeds, the Sussex and Leghorn, the within-breed genetic structuring was associated with morphological varieties. A prior assumption was that breeders would want to maintain distinct morphological types and this would result in breed substructure that corresponded with the different types. However, for another eight breeds the observed genetic subpopulations were associated with flock supplier. The extensive genetic substructure observed in many of the British traditional chicken breeds was in contrast to the overall consensus that livestock breeds tend to be homogenous genetic populations that rarely deviate from the Hardy-Weinberg equilibrium (HWE) proportions (Lawson Handley et al. 2007). Deviations from HWE due to a heterozygote deficit and further detection of genetic subpopulations using individual-based clustering methods tend to be limited to the odd sampled breed, as was found in French horse breeds (Glowatzki-Mullis et al. 2006), Italian chicken breeds (Zanetti et al. 2010) and the British pig breeds (**chapter 2**). Thus, overall phenotypically defined livestock breeds equate to genetic populations. This could be due to the efforts of regulatory bodies, such as individual breed societies and organisations like the British Pig Association and Rare Breeds Survival Trust, which strive to preserve and maintain genetic diversity, distinctiveness and integrity of livestock breeds. For the British chicken breeds,

however, the genetic situation appears to be more complex and the genetic substructure may be a reflection of the absence of a regulatory body commonly in place for other livestock species. It appears that chicken breeders may be implementing certain management practices within breeds, such as restricting gene flow between flocks, thus producing subtle genetic substructure within breeds. Such practices may have profound implications on levels of genetic diversity. The British Saddleback pig breed as a whole exhibited high genetic diversity, but there was an absence of allelic diversity and higher individual inbreeding in the Rainbarrow herd (**chapter 3**). Microsatellite markers revealed moderate to high within-population genetic diversity in the British traditional chicken breeds (**chapter 4**) and in the interests of preserving current levels of genetic diversity the exchange of genetic material within breeds should be encouraged.

Breed structure was surveyed in this thesis by quantifying the levels of genetic variation between breeds and substantial genetic differentiation was observed in both the British pig (**chapter 2**) and chicken breeds (**chapter 4**). Even breeds that shared a common ancestry were moderately differentiated from one another, such as Landrace and British Lop pig breeds and Brahma and Cochin chicken breeds. Although gene flow may have occurred between breeds in the past through upgrading schemes, the British pig and chicken breeds represent differentiated genetic populations. Restrictions on contemporary gene flow between breeds have possibly contributed to the genetic distinctiveness of the breeds. In addition, many livestock breeds have small population sizes, sometimes a result of population contractions, which increases the effects of random genetic drift driving alleles to fixation thereby rapidly

enhancing breed genetic differentiation (Hartl and Clark 1997). For instance, some of the British chicken breeds studied in this thesis are listed as 'Endangered' or 'Critical', such as the Ixworth, Lincolnshire Buff, Scots Grey and Spanish (DEFRA 2010) and these were found to be amongst the most genetically distinctive of the British traditional chicken breeds.

Genetic distances between breeds were used to further characterise breed structure through the identification of groups of genetically similar breeds using both population-based and individual-based phylogenetic reconstruction. Two main genetic groups were identified in the British traditional chicken breeds: an Asian cluster consisting of Brahma, Cochin, Buff Orpington, Croad Langshan, Silkie and Lincolnshire Buff and an indigenous breed group consisting of Derbyshire Redcap, Dorking, Old English Pheasant Fowl and Hamburg (**chapter 4**). In the British pig breeds genetic affinities were highlighted between Berkshire, British Saddleback and Gloucestershire Old Spots and between Landrace and British Lop (**chapter 2**). The genetic proximity of these breeds is probably due to shared ancestry and historical cross-breeding, such as the once popular Berkshire pig breed that was used to improve other breeds (BPA 2002; Porter 1993) and Asian chicken breeds used to improve Sussex, which in turn contributed to the development of the white-feathered Ixworth (Hams 2004; Vorwald Dohner 2001). However, beyond these few identified breed groups, the genetic relationships of the other breeds and the overall hierarchical structure could not be discerned. Even though the breeds were highly differentiated, phylogenetic reconstruction of both the pig and chicken breeds produced cladograms with low bootstrap support for relationships between breeds,

star-shaped topologies and short internal branches that separated breeds. As a result, the evolutionary radiation of the breeds was left largely unresolved and there was no evidence of which breeds were ancestral or first established. Although livestock breeds are genetically distinct, it has proven difficult to unravel the evolutionary relationships of modern breeds using phylogenetic reconstruction as the reproduced trees often lacks robustness (cattle (Li et al. 2007); pig (Megens et al. 2008; SanCristobal et al. 2006a); cat (Menotti-Raymond et al. 2008); sheep (Peter et al. 2007)). A possible cause could be the complex developmental histories of livestock breeds, which often involved genetic introgression or cross-breeding. These processes do not adhere to the assumptions of linear bifurcation of the tree-building methods where an ancestral population gives rise to a pair of descendant populations (Eding and Bennewitz 2007; Rosenberg et al. 2001; Toro and Caballero 2005). The effect of these human-mediated processes may be that breeds are equally related to one another due to their complex historical genetic foundations, making it difficult to genetically infer the evolutionary history of breeds. In addition, bottlenecks are known to rapidly increase the genetic distinctiveness of populations and these altered genetic distance values can distort the phylogenetic reconstruction of evolutionary relationships (Takezaki and Nei 1996). Many traditional livestock breeds have experienced substantial bottlenecks and their historic genetic relationships may be untraceable through phylogenetic reconstruction of microsatellites. Another possibility of the lack of bootstrap support of phylogenetic reconstruction of genetic relationships between breeds is insufficient sampling of loci and individuals. By focusing on maximising the probability of exactly recovering the one true tree topology, several studies conducted by Nei and colleagues have addressed the

number of microsatellite loci required in diversity studies, with recommendations varying from 30 to at least 50 (Nei et al. 29183; Takezaki and Nei 1996; Takezaki and Nei 2008). Since only 30 microsatellite loci were genotyped for the British traditional chicken breeds, a greater sample of markers may produce a more resolved topology (**chapter 4)**. In addition, since microsatellite loci constitute only a small portion of the genome, sampling greater genomic data (e.g. DNA sequences) could provide additional information on the history of livestock populations.

7.2.2 Identification of the breed of origin of individuals

Due to the maintenance of genetic integrity and distinctiveness of livestock breeds, the clustering of individuals to breed of origin using both Bayesian genotypic clustering approaches and phylogenetic reconstruction has proved to be successful, as was found in dog breeds (Koskinen 2003), horse breeds (Glowatzki-Mullis et al. 2006), and, in this thesis, British pig breeds (**chapter 2**). Although individuals clustered to breed of origin for many of the British traditional chicken breeds, other breeds were subdivided and, in effect, individuals of these breeds appeared to cluster to a subpopulation of origin (**chapter 4**). Although, from a perspective of population management, the result highlights possible breeding practices (as discussed earlier), the ease of implementation of a genetic test to verify the claimed breed origin of marketed chicken products could be affected the genetic substructure. Nonetheless, at present there does not appear to be a market for chicken products labelled with premium value breed names. However, there is an increasing market for traditional

British pig breed products and, consequently, the individual-based clustering results for the British pig breeds in **chapter 2** were promising.

In **chapters 5** and **6** of this thesis breed identification in both European cattle and British pig breeds was further explored using likelihood-based individual assignment methods and dense genome-wide assays. Following the selection of markers using different population genetic differentiation methods in **chapter 5**, the power of individual assignment varied markedly amongst marker panels, with those generated by delta and pairwise Wright's F_{ST} outperforming other panels and agreeing to a large extent on which were the most informative markers. Individual assignment success followed an asymptotic curve as a function of cumulatively increasing number of informative markers, a result often observed in breed assignment analyses (Bjornstad and Roed 2002; Blott et al. 1999; Cornuet et al. 1999). No further gain in power of assignment was achieved by sampling in excess of 200 SNP markers in the European cattle breeds. Similarly, the 96-marker panel developed in **chapter 6** was sufficient at discriminating the British pig breeds; including the next 50 informative markers did not substantially improve the power of assignment to warrant the development of a larger assay. In **chapter 5** it was shown that prior selection of informative markers produced a higher level of correct breed identification than panels of randomly selected markers. Similarly, the 96-SNP panel of breed-informative markers in **chapter 6** exhibited levels of genetic differentiation (average F_{ST} = 0.54) that far exceeded that using the FAO-recommended panel of polymorphic microsatellite markers in **chapter 2** (average F_{ST} = 0.29). However, although the individual assignment success was high in both the European cattle and

British pigs, a small proportion of individuals tended to remain mis-classified or unassigned in both analyses, regardless of the number of genetic markers. This appears to be less to do with an inadequate number of genetic markers and instead was a reflection of genetically atypical individuals (Bjornstad and Roed 2002). The power of individual assignment was variable across breeds, with more markers required to distinguish the closely related breeds. For instance, the dairy breeds, Jersey and Guernsey, and the traditional beef breeds, Red Poll and Welsh Black, required fewer than 100 markers to achieve 100% assignment success, whilst the closely related commercial beef breeds Charolais, Simmental and Limousin required more than 200 markers to achieve greater than 95% assignment success. In the British pig breeds, the 96-SNP panel could not effectively discriminate between the Welsh and Landrace breeds as indicated by a relatively low posterior probability of correct assignment. In contrast, in the other traditional British pig breeds the posterior probability of correct assignment that a sample was truly from a traditional breed, when contrasted against other breed sources, was extremely high.

The power of breed assignment was related to levels of genetic differentiation of a breed. In particular, breeds that have small population sizes, often the traditional breeds, were found to be genetically distinct and were amongst the breeds that were effectively discriminated using assignment tests. Consequently, the breed assignment results were encouraging because it is more likely that product label substitutions would involve the premium breed ‘brand’ names of the traditional livestock breeds. Furthermore, independent test samples validated the effectiveness of the 96-SNP panel at verifying the origin of the traditional pig breed products. However, there are

potential practical limitations with regards to the use of SNP panels for food authentication such as the breed verification of products like sausages and burgers, where DNA could be extensively mixed during processing (Primrose et al. 2010). Second, such tests can be expensive to implement on a widespread basis. Nonetheless, the availability or at least the knowledge that a test is being implemented can deter further potential fraudulent practice, as was found with basmati rice (Dr Rob Ogden, personal communication 2010). Third, the genetic composition of the reference populations used in the assignment test is important. As was seen in **chapter 6**, not one of the Hampshire independent market test samples could be verified using the current reference population. It is difficult to speculate as to whether the test samples were not purebred Hampshires or whether the Hampshire reference population was a poor representative of the contemporary breed gene pool.

From a conservation perspective, the panel of informative SNP markers identified in **chapters 5** and **6**, for the European cattle and British pig breeds respectively, also represent an important resource for the conservation of livestock breed genetic variation as the panels encompass some of the genetic differences between the breeds.

7.2.3 Data sampling

Prior to characterisation of genetic diversity and structure of livestock breeds, sampling criteria are generally established. The general aim is to obtain an adequate sample size from a range of breeders' flocks or herds, such that as much population

variation as possible is captured in the genetic dataset. In **chapter 2** the dataset on the British pig breeds was from the PigBioDiv project, where the aim was to collect up to 50 individuals with two siblings per litter from as many herds as possible (SanCristobal et al. 2006a). Such large sample sizes are not always easily achieved in livestock biodiversity studies as an inherent challenge is that traditional breeds can be low in population size. This can be further compounded by identifying trustworthy animal breeders and, hence, relying on samples from a restricted number of herds or flocks from breeders of known provenance. As a consequence, sampling objectives may not be fulfilled. For instance, although the aim was to collect 30 samples from at least 5 or 6 flocks for the British traditional chicken breeds, only 19 samples could be obtained for the Norfolk Grey breed (**chapter 4**). Similarly, an extremely low sample size of 15 individuals was collected for the Red Angus cattle breed (**chapter 5**).

A second issue that may arise during the data collection process is a possible mix-up of samples resulting in mis-labelling. Preliminary analysis of the microsatellite data of the British traditional chicken breeds produced evidence to surmise that this may have occurred in a few cases (**chapter 4**). Individual-based phylogenetic reconstruction resulted in one or two individuals not clustering to origin in a number of breeds. On closer examination, the suppliers of several of those mis-assigned individuals also provided samples for the breed to which they clustered with (these samples were removed). However, no evidence of a possible egg mix-up was found for several other mis-assigned individuals (these individuals were retained). Subsequent Bayesian genotypic clustering analysis indicated that some of these

individuals were admixed, which could help explain the mis-assignment found in the phylogenetic reconstruction. However, not all of the mis-assigned individuals were admixed, such as two Scot Grey individuals clustering with Old English Pheasant Fowl. Since no potential evidence of an egg mix-up was found for the two Scot Grey individuals it is difficult to speculate as to the cause. It is possible that the suppliers possessed Old English Pheasant Fowl since their full breed stock is unknown. Mis-labelling of samples was probably not as relevant for the pig and cattle breeds (**chapters 2, 3, 5** and **6**) because, unlike the chicken breeds where DNA was obtained from eggs, biological samples were obtained directly from animals.

7.3 Future work

The thesis revealed challenges posed mainly by the population genetic statistical tools and practical applications in breed identification, as well as opportunities for further research.

Although novel Bayesian genotypic clustering techniques are useful at detecting certain genetic patterns of populations, the behaviour of these sophisticated approaches can make it difficult to evaluate the reliability and correctness of results (Ball et al. 2010; Frantz and Cellina 2009; Rowe and Beebee 2007; Safner et al. 2011). The methods can be used to detect individual genetic admixture, the result of hybridisation between populations, but, as demonstrated in **chapter 2**, the detection of this biological phenomenon can vary between approaches. Although both BAPS and STRUCTURE identified genetic admixture in certain individuals, sometimes

only one approach detected genetic introgression in other individuals. Recent studies have evaluated the ability of Bayesian genotypic clustering tools at detecting hybridisation using both simulated data (Latch et al. 2006; Sanz et al. 2009) and naturally occurring animal populations (Bohling and Waits 2011), with reported variability in the detection of introgression and first and second generation hybrids between the different approaches. However, simulated conditions are sometimes unrealistic. Furthermore, the extent of hybridisation in natural populations is often unknown due to the lack of a pedigree and, consequently, it is difficult to conclude which detected levels of individual genetic admixture are more accurate. More extensive testing is required and livestock populations offer the prospect of evaluating the ability of these Bayesian tools at detecting individual admixture due to the availability of pedigrees. The individual genetic admixture compositions estimated by the Bayesian genotypic clustering tools could be compared with ancestral predictions derived from a pedigree. Although, checks on the pedigree should be conducted to assess their accuracy as pedigrees can contain errors. More complex patterns of hybridisation could also be investigated using genotyped experimental animal populations that incorporate pure-bred founder animals (F0), first generation hybrids (F1), second generation hybrids (F2) and backcrosses (F0 x F1) (e.g. Bovine Genome (RoBoGen) herd; Charolais and Holstein cattle breeds (Gutierrez-Gil et al. 2007)).

Another challenge that arose with the application of Bayesian genotypic clustering tools was the identification of the optimal number of underlying populations (K), particularly with STRUCTURE the most commonly used clustering approach in

population genetic studies. Even with delta K, an ad-hoc statistic developed explicitly to be used on STRUCTURE outputs, the identification of the optimal K was a challenge. To that end, other statistical methods could be used to identify K. The deviance information criterion (DIC), which measures model fit and complexity, is widely used to compare Bayesian models (Spiegelhalter et al. 2002). Using an empirical dataset of known population structure or simulated datasets encompassing various demographic scenarios (varying gene flow/migration rates), further analysis could determine if DIC would be a more appropriate alternative than delta K at determining K.

Chapter 3 reported that the British Saddleback breed was divided into two genetically differentiated subpopulations, one of which was composed of individuals that originated from the Rainbarrow herd. A subtle pattern of herd structuring in the British Saddleback may not be an exceptional result because breeds are generally composed of herds, which are likely to contain individuals that are more genetically similar to one another than to individuals from other herds (Blott et al. 1998b). The Rainbarrow herd individuals constituted nearly half the sample size of the British Saddleback breed. Herd sampling information is currently not available for the other British pig breeds so it is unknown if the herd representation in the other breeds was more balanced. This raises the question of whether the sampling composition of the British Saddleback breed affected the performance of the individual-based clustering approaches and whether the observed substructure was instead an artefact of an over-represented subpopulation. If the other British Saddleback herds had had a similar sample size would the same genetic substructure still have been detected? A study

using simulated data illustrated that certain sampling scenarios, such as line transect, trapper sampling and multi-generational, can affect the performance of Bayesian genotypic clustering methods (Schwartz and McKelvey 2009). More extensive testing is required to address the effect of sampling schemes on the performance of these methods. Livestock breeds offer the prospect of evaluating whether irregular or unbalanced sampling of individuals from genetically distinct breeds and subpopulations (or herds) within breeds could affect the detection of clusters using Bayesian genotypic clustering methods.

As described in **chapters 5** and **6**, using purebred cattle and pig breeds, markers can be selected from the genome-wide assays to create genetic panels for individual identification and authentication of traditional breed products. However, in practical terms the implementation would be more complicated because market meat products are not usually derived from purebred breeds. Instead, breeding companies specialise in commercial lines that are derived from 2 or 3 way breed crosses (BPA 2002). To fully validate the genetic tool developed in **chapter 6** additional samples that are representative of marketed meat products derived from commercial breeding companies, as both reference populations and independent test samples, need to be incorporated. These samples would then allow complete determination of whether purebred traditional breeds and their products can be genetically distinguished.

Another issue in food authentication is that some meat products are of mixed ancestry derived from cross-breeding, such as the highly prized Iberian pig ham which is allowed to have a genetic composition of a maximum of 50% Duroc (Garcia

et al. 2006). Similarly, the British traditional pig breeds could also be marketed as cross-bred. To tackle this issue 'dummy' mixed breed reference populations could be created and a Bayesian genotypic clustering analysis could be performed to determine the genetic composition of marketed cross-bred products. In addition, sometimes supermarkets explicitly label the breed of origin of the sire of meat, such that the named breed would attract a premium value to the product. In this situation, the dense genome wide assays offer the prospect of screening Y-chromosome SNPs to identify those that are breed-specific and a tailor-made genetic assay could be developed to address this specific food product claim. Dense genome-wide marker arrays also allow more specific points of food authentication to be addressed. Beyond developing a genetic panel to discriminate amongst a group of breeds, breed-specific genetic tests could be developed by identifying the alleles that are private to the given breed. This is particularly relevant for the verification of meat products derived from the premium value traditional livestock breeds.

With dense genome wide assays, not only can the population structure and genetic diversity of be studied, but also additional population genetic concepts like linkage disequilibrium (LD) and effective population size (N_e) can also be more fully investigated. As LD and N_e vary across chromosomes and across the genome the availability of dense genome wide assays allow a more accurate description of these properties.N_e is an important population genetic concept as it measures the number of individuals in an ideal population. Because there are factors that can affect N_e (such as inbreeding, population fluctuation, selection, population structure), its population estimation can contribute to the body of knowledge of the history of a

breed. It would be of particular interest to compare and contrast LD and N_e between commercial and traditional breeds considering that the two breed types have experienced different developmental and demographic trajectories.

7.4 Conclusion

In conclusion, this thesis demonstrated that animal breeders continue to maintain high levels of genetic diversity between, and genetic homogeneity within most livestock breeds. This ensures the protection and preservation of traditional breeds. In particular, the phenotypically distinct British pig breeds could be identified as separate populations. However, the British chicken breeds showed evidence of population structuring and a continuation of this form of breed management may be detrimental for overall breed diversity. Certain genetic differentiation methods were shown to be highly useful for screening dense genome-wide assays for information population genetic markers. Breeds could be effectively discriminated using panels of informative markers, in particular the traditional breeds, which will aid in the protection of the economic value of these breeds and their products.

BIBLIOGRAPHY

Ajmone-Marsan, P., and GlobalDiv Consortium. 2010. A global view of livestock biodiversity and conservation - GLOBALDIV. Animal Genetics 41:1-5.

Ajmone-Marsan, P., R. Negrini, P. Crepaldi, E. Milanesi, C. Gorni, A. Valentini, and M. Cicogna. 2001. Assessing genetic diversity in Italian goat populations using AFLP (R) markers. Animal Genetics 32:281-288.

Ajmone-Marsan, P., R. Negrini, E. Milanesi, R. Bozzi, I. J. Nijman, J. B. Buntjer, A. Valentini, and J. A. Lenstra. 2002. Genetic distances within and across cattle breeds as indicated by biallelic AFLP markers. Animal Genetics 33:280-286.

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Research 12:1805-1814.

Alderson, L. 2007. The Saddleback family of pig breeds. The Ark 35:22-25.

Allendorf, F. W. 1986. Genetic drift and the loss of alleles versus heterozygosity. Zoo Biology 5:181-190.

Amaral, A. J., L. Ferretti, H.-J. Megens, R. P. M. A. Crooijmans, H. Nie, S. E. Ramos-Onsins, M. Perez-Enciso, L. B. Schook, and M. A. M. Groenen. 2011. Genome-Wide Footprints of Pig Domestication and Selection Revealed through Massive Parallel Sequencing of Pooled DNA. Plos One 6

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

Ball, M., L. Finnegan, M. Manseau, and P. Wilson. 2010. Integrating multiple analytical approaches to spatially delineate and characterize genetic population structure: an application to boreal caribou (*Rangifer tarandus caribou*) in central Canada. Conservation Genetics 11:2131-2143.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265.

Beaumont, M. A., and B. Rannala. 2004. The Bayesian revolution in genetics. Nature Review Genetics 5:251-261.

Berthouly, C., B. Bed'Hom, M. Tixier-Boichard, C. F. Chen, Y. P. Lee, D. Laloe, H. Legros, E. Verrier, and X. Rognon. 2008. Using molecular markers and multivariate methods to study the genetic diversity of local European and Asian chicken breeds. Animal Genetics 39:121-129.

Bjornstad, G., and K. H. Roed. 2002. Evaluation of factors affecting individual assignment precision using microsatellite data from horse breeds and simulated breed crosses. Animal Genetics 33:264-270.

Blott, S. C., J. L. Williams, and C. S. Haley. 1998a. Genetic relationships among European cattle breeds. Animal Genetics 29:273-282.

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

Blott, S. C., J. L. Williams, and C. S. Haley. 1999. Discriminating among cattle breeds using genetic markers. Heredity 82:613-619.

Bodzsar, N., H. Eding, T. Revay, A. Hidas, and S. Weigend. 2009. Genetic diversity of Hungarian indigenous chicken breeds based on microsatellite markers. Animal Genetics 40:516-523.

Bohling, J. H., and L. P. Waits. 2011. Assessing the prevalence of hybridization between sympatric Canis species surrounding the red wolf (Canis rufus) recovery area in North Carolina. Molecular Ecology 20:2142-2156.

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

Bowling, A. T. 1994. Population genetics of great-basin feral horses. Animal Genetics 25:67-74.

BPA. 2002. British Pig Breeds. British Pig Association. Cambridge, UK.

BPA. 2008. British Pig Association Newsletter Autumn 2008. British Pig Association. Cambridge, UK.

Bray, T. C., L. Chikhi, A. J. Sheppy, and M. W. Bruford. 2009. The population genetic effects of ancestry and admixture in a subdivided cattle breed. Animal Genetics 40:393-400.

Bruford, M. W., D. G. Bradley, and G. Luikart. 2003. DNA markers reveal the complexity of livestock domestication. Nature Review Genetics 4:900-910.

Campbell, D., P. Duchesne, and L. Bernatchez. 2003. AFLP utility for population assignment studies: analytical investigation and empirical comparison with microsatellites. Molecular Ecology 12:1979-1991.

Canon, J., D. Garcia, M. A. Garcia-Atance, G. Obexer-Ruff, J. A. Lenstra, P. Ajmone-Marsan, S. Dunner, and E. Consortium. 2006. Geographical partitioning of goat diversity in Europe and the Middle East. Animal Genetics 37:327-334.

Charlesworth, B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. Molecular Biology and Evolution 15:538-543.

Chikhi, L., B. Goossens, A. Treanor, and M. W. Bruford. 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.

Ciampolini, R., V. Cetica, E. Ciani, E. Mazzanti, X. Fosella, F. Marroni, M. Biagetti, C. Sebastiani, P. Papa, G. Filippini, D. Cianci, and S. Presciuttini. 2006. Statistical analysis of individual assignment tests among four cattle breeds using fifteen STR loci. Journal of Animal Science 84:11-19.

Clarke, S. W., E. M. Tucker, and S. J. G. Hall. 1989. Genetic Polymorphisms and Their Relationships with Inbreeding and Breed Structure in Rare British Sheep: The Portland, Manx Loghtan, and Hebridean. Conservation Biology 3:381-388.

Clutton-Brock, J. 1999. A Natural History of Domesticated Mammals. Cambridge University Press. Cambridge, UK.

Coltman, D. W., J. G. Pilkington, J. A. Smith, and J. M. Pemberton. 1999. Parasite-Mediated Selection against Inbred Soay Sheep in a Free-Living Island Population. Evolution 53:1259-1267.

Corander, J., and P. Marttinen. 2006. Bayesian identification of admixture events using multi-locus molecular markers. Molecular Ecology 15:2833 - 2843.

Corander, J., P. Marttinen, J. Sirén, and J. Tang. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9:539 - 552.

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

Crawford, R. 1990. Poultry breeding and genetics. Elsevier. Cambridge, UK.

Curik, I., P. Zechner, J. Solkner, R. Achmann, I. Bodo, P. Dovc, T. Kavar, E. Marti, and G. Brem. 2003. Inbreeding, microsatellite heterozygosity, and morphological traits in Lipizzan horses. Journal of Heredity 94:125-132.

Darwin, C. 1868. The Variation of Animals and Plants under Domestication. John Murray. London, UK.

Davies, N., F. X. Villablanca, and G. K. Roderick. 1999. Determining the source of individuals: multilocus genotyping in nonequilibrium population genetics. Trends in Ecology & Evolution 14:17-21.

Decker, J. E., J. C. Pires, G. C. Conant, S. D. McKay, M. P. Heaton, K. F. Chen, A. Cooper, J. Vilkki, C. M. Seabury, A. R. Caetano, G. S. Johnson, R. A. Brenneman, O. Hanotte, L. S. Eggert, P. Wiener, J. J. Kim, K. S. Kim, T. S. Sonstegard, C. P. Van Tassell, H. L. Neibergs, J. C. McEwan, R. Brauning, L. L. Coutinho, M. E. Babar, G. A. Wilson, M. C. McClure, M. M. Rolf, J. Kim, R. D. Schnabel, and J. F. Taylor. 2009. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proceedings of the National Academy of Sciences of the United States of America 106:18644-18649.

DEFRA. 2006. UK National Action Plan on Farm Animal Genetic Resources. DEFRA. London, UK.

DEFRA. 2009. Review of molecular characterisation studies relating to UK Farm Animal Genetic Resources. London, UK.

DEFRA. 2010. Poultry in the United Kingdom. DEFRA. London, UK.

Dinno, A. 2009. Pp. paran -- Horn's Parallel Analysis of Components/Factors.

Drineas, P., J. Lewis, and P. Paschou. 2010. Inferring Geographic Coordinates of Origin for Europeans Using Small Panels of Ancestry Informative Markers. PLoS One 5

Druml, T., I. Curik, R. Baumung, K. Aberle, O. Distl, and J. Soelkner. 2007. Individual-based assessment of population structure and admixture in Austrian, Croatian and German draught horses. Heredity 98:114-122.

Eck, S. H., A. Benet-Pages, K. Flisikowski, T. Meitinger, R. Fries, and T. M. Strom. 2009. Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. Genome Biology 10:82-90.

Eding, H., and J. Bennewitz. 2007. Measuring genetic diversity in farm animals *in* K. Oldenbroek, ed. Utilisation and Conservation of Farm Animal Genetic Resources. Wageningen Academic Publishers.Wageningen, Netherlands.

Efron, B. 1983. Estimating the error rate of a predicition rule - improvement on cross-validation. Journal of the American Statistical Association 78:316-331.

Ellegren, H. 2004. Microsatellites: Simple sequences with complex evolution. Nature Reviews Genetics 5:435-445.

EPS. 2011. The Essex Pig History of Bloodlines and Breeders. The Essex Pig Society http://www.essexpigsociety.com/.

Evanno, G., and S. Regnaut. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology 14:2611-2620.

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

FAO. 2007. The State of the World's Animal Genetic Resources for Food and Agriculture – in brief. Commission on Genetic Resources for Food and Agriculture. FAO, Rome, Italy.

Felsenstein, J. 1985. Confidence of limits on phylogenies - an approach using the bootstrap Evolution 39:783-791.

Felsenstein, J. 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). Cladistics 5:164-166.

Felsenstein, J. 2008. PHYLIP (Phylogeny Inference Package) version 3.67. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Frantz, A., and S. Cellina. 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? Journal of Applied Ecology 46:493-505.

Frantz, A., J. T. Pourtois, M. Heuertz, M. Flamand, A. Krier, S. Bertouille, F. Chaumont, and T. Burke. 2006. Genetic structure and assignment tests demonstrate illegal translocation of red deer (*Cervus elaphus*) into a continuous population. Molecular Ecology 15:3191-3203.

Garcia, D., A. Martinez, S. Dunner, J. L. Vega-Pla, C. Fernandez, J. V. Delgado, and J. Canon. 2006. Estimation of the genetic admixture composition of Iberian dry-cured ham samples using DNA multilocus genotypes. Meat Science 72:560-566.

Gautier, M., D. Laloë, and K. Moazami-Goudarzi. 2010. Insights into the Genetic History of French Cattle from Dense SNP Data on 47 Worldwide Breeds. PLoS One 5:e13038.

Gautschi, B., J. P. Muller, B. Schmid, and J. A. Shykoff. 2003. Effective number of breeders and maintenance of genetic diversity in the captive bearded vulture population. Heredity 91:9-16.

Gibbs, R. A., J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversoie, C. A. Gill, R. D. Green, D. L. Hamernik, S. M. Kappes, S. Lien, L. K. Matukumalli, J. C. McEwan, L. V. Nazareth, R. D. Schnabel, G. M. Weinstock, D. A. Wheeler, P. Ajmone-Marsan, P. J. Boettcher, A. R. Caetano, J. F. Garcia, O. Hanotte, P. Mariani, L. C. Skow, J. L. Williams, B. Diallo, L. Hailemariam, M. L. Martinez, C. A. Morris, L. O. C. Silva, R. J. Spelman, W. Mulatu, K. Zhao, C. A. Abbey, M. Agaba, F. R. Araujo, R. J. Bunch, J. Burton, C. Gorni, H. Olivier, B. E. Harrison, B. Luff, M. A. Machado, J. Mwakaya, G. Plastow, W. Sim, T. Smith, T. S. Sonstegard, M. B. Thomas, A. Valentini, P. Williams, J. Womack, J. A. Wooliams, Y. Liu, X. Qin, K. C. Worley, C. Gao, H. Jiang, S. S. Moore, Y. Ren, X.-Z. Song, C. D. Bustamante, R. D. Hernandez, D. M. Muzny, S. Patil, A. S. Lucas, Q. Fu, M. P. Kent, R. Vega, A. Matukumalli, S. McWilliam, G. Sclep, K. Bryc, J. Choi, H. Gao, J. J. Grefenstette, B. Murdoch, A. Stella, R. Villa-Angulo, M. Wright, J. Aerts, O. Jann, R. Negrini, M. E. Goddard, B. J. Hayes, D. G. Bradley, M. B. da Silva, L. P. L. Lau, G. E. Liu, D. J. Lynn, F. Panzitta, and K. G. Dodds. 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324:528-532.

Giuffra, E., J. M. H. Kijas, V. Amarger, O. Carlborg, J. T. Jeon, and L. Andersson. 2000. The origin of the domestic pig: Independent domestication and subsequent introgression. Genetics 154:1785-1791.

Glowatzki-Mullis, M. L., J. Muntwyler, E. Baeumle, and C. Gaillard. 2009. Genetic diversity of Swiss sheep breeds in the focus of conservation research. Journal of Animal Breeding and Genetics 126:164-175.

Glowatzki-Mullis, M. L., J. Muntwyler, W. Pfister, E. Marti, S. Rieder, P. A. Poncet, and C. Gaillard. 2006. Genetic diversity among horse populations with a special focus on the Franches-Montagnes breed. Animal Genetics 37:33-39.

Goudet, J. 1995. FSTAT (Version 1.2): A computer program to calculate F-statistics. Journal of Heredity 86:485-486.

Granevitze, Z., J. Hillel, G. H. Chen, N. T. K. Cuc, M. Feldman, H. Eding, and S. Weigend. 2007. Genetic diversity within chicken populations from different continents and management histories. Animal Genetics 38:576-583.

Granevitze, Z., J. Hillel, M. Feldman, A. Six, H. Eding, and S. Weigend. 2009. Genetic structure of a wide-spectrum chicken gene pool. Animal Genetics 40:686-693.

Guillot, G., A. Estoup, F. Mortier, and J. C. JF. 2005. A spatial statistical model for landscape genetics. Genetics 170:1261-1280.

Gutierrez-Gil, B., P. Wiener, and J. L. Williams. 2007. Genetic effects on coat colour in cattle: dilution of eumelanin and phaeomelanin pigments in an F2-Backcross Charolais x Holstein population. BMC Genetics 8

Hall, S., and J. Clutton-Brock. 1988. Two hundred years of British farm livestock. British Museum (Natural History), London, UK.

Hall, S. J. G. 1989. Breed structure of rare pigs - implications for conservation of the Berkshire, Tamworth, Middle White, Large Black, Gloucester Old Spot, British Saddleback, and British Lop. Conservation Biology 3:30-38.

Hall, S. J. G., and D. G. Bradley. 1995. Conserving livestock breed biodiversity Trends in Ecology & Evolution 10:267-270.

Hams, F. 2004. Old Poultry Breeds. Shire Publications Ltd.

Hartl, D., and A. Clark. 1997. Principles of Population Genetics. Sinauer Associates. New York, USA.

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

Hill, W. G. 2000. Maintenance of quantitative genetic variation in animal breeding programmes. Livestock Production Science 63:99-109.

Hill, W. G., and J. Rasbash. 1986. Models of long-term artificial selection in finite population. Genetical Research 48:41-50.

Hill, W. G., and X. S. Zhang. 2004. Genetic variation within and among animal populations. In: Farm Animal Genetic Resources. Nottingham University Press. Loughborough, UK.

Hillel, J., M. A. M. Groenen, M. Tixier-Boichard, A. B. Korol, L. David, V. M. Kirzhner, T. Burke, A. B. Dirie, R. Crooijmans, K. Elo, M. Feldman, P. J. Freidlin, A. Maki-Tanila, M. Oortwijn, P. Thomson, A. Vignal, K. Wimmers, and S. Weigend. 2003. Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. Genetics Selection Evolution 35:533-557.

Holsinger, K. E., and B. S. Weir. 2009. Fundamental concepts in genetics Genetics in geographically structured populations: defining, estimating and interpreting F-ST. Nature Reviews Genetics 10:639-650.

Huelsenbeck, J. P., and P. Andolfatto. 2007. Inference of population structure under a Dirichlet process prior Genetics 175:1787 - 1802.

Jost, L. 2008. G_{ST} and its relatives do not measure differentiation. Molecular Ecology 17:4015-4026.

Kalinowski, S. T. 2011. The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. Heredity 106:625-632.

Kersbergen, P., K. van Duijn, A. D. Kloosterman, J. T. den Dunnen, M. Kayser, and P. de Knijff. 2009. Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. BMC Genetics 10

Kijas, J. W., D. Townley, B. P. Dalrymple, M. P. Heaton, J. F. Maddox, A. McGrath, P. Wilson, R. G. Ingersoll, R. McCulloch, S. McWilliam, D. Tang, J. McEwan, N. Cockett, V. H. Oddy, F. W. Nicholas, H. Raadsma, and C. for the International Sheep Genomics. 2009. A Genome Wide Survey of SNP Variation Reveals the Genetic Structure of Sheep Breeds. PLoS One 4:e4668.

Kim, S., and A. Misra. 2007. SNP genotyping: Technologies and biomedical applications. Annual Review of Biomedical Engineering 9:289-320.

Koskinen, M. T. 2003. Individual assignment using microsatellite DNA reveals unambiguous breed identification in the domestic dog. Animal Genetics 34:297-301.

Kunene, N. W., C. C. Bezuidenhout, and I. V. Nsahlai. 2009. Genetic and phenotypic diversity in Zulu sheep populations: Implications for exploitation and conservation. Small Ruminant Research 84:100-107.

Laloë, D., K. Moazami-Goudarzi, J. A. Lenstra, P. A. Marsan, P. Azor, R. Baumung, D. G. Bradley, M. W. Bruford, J. Cañón, G. Dolf, S. Dunner, G. Erhardt, G. Hewitt, J. Kantanen, G. Obexer-Ruff, I. Olsaker, C. Rodellar, A. Valentini, P. Wiener, E. C. G. D. Consortium, and E. Consortium. 2010. Spatial Trends of Genetic Variation of Domestic Ruminants in Europe. Diversity 2:932-945.

Lao, O., K. van Duijn, P. Kersbergen, P. de Knijff, and M. Kayser. 2006. Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. American Journal of Human Genetics 78:680-690.

Larson, G., K. Dobney, U. Albarella, M. Y. Fang, E. Matisoo-Smith, J. Robins, S. Lowden, H. Finlayson, T. Brand, E. Willerslev, P. Rowley-Conwy, L. Andersson, and A. Cooper. 2005. Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science 307:1618-1621.

Latch, E., G. Dharmarajan, J. Glaubitz, and O. Rhodes. 2006. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conservation Genetics 7:295-302.

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

Lawson Handley, L. J., K. Byrne, F. Santucci, S. Townsend, M. Taylor, M. W. Bruford, and G. M. Hewitt. 2007. Genetic structure of European sheep breeds. Heredity 99:620-631.

Lewis, J., Z. Abas, C. Dadousis, D. Lykidis, P. Paschou, and P. Drineas. 2011. Tracing Cattle Breeds with Principal Components Analysis Ancestry Informative SNPs. PLoS One 6

Li, M. H., J. Kantanen, A. Michelson, and U. Saarma. 2011. Genetic components of grey cattle in Estonia as revealed by microsatellite analysis using two Bayesian clustering methods. BMC Research Notes 4:37.

Li, M. H., I. Tapio, J. Vilkki, Z. Ivanova, T. Kiselyova, N. Marzanov, M. Cinkulov, S. Stojanovic, I. Ammosov, R. Popov, and J. Kantanen. 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.

Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe, M. Kamal, M. Clamp, J. L. Chang, E. J. Kulbokas, M. C. Zody, E. Mauceli, X. H. Xie, M. Breen, R. K. Wayne, E. A. Ostrander, C. P. Ponting, F. Galibert, D. R. Smith, P. J. deJong, E. Kirkness, P. Alvarez, T. Biagi, W. Brockman, J. Butler, C. W. Chin, A. Cook, J. Cuff, M. J. Daly, D. DeCaprio, S. Gnerre, M. Grabherr, M. Kellis, M. Kleber, C. Bardeleben, L. Goodstadt, A. Heger, C. Hitte, L. Kim, K. P. Koepfli, H. G. Parker, J. P. Pollinger, S. M. J. Searle, N. B. Sutter, R. Thomas, C. Webber, E. S. Lander, and P. Broad Inst Genome Sequencing. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803-819.

Liu, N., L. Chen, S. Wang, C. Oh, and H. Zhao. 2005. Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genetics 6

Liu, Y. P., G. S. Wu, Y. G. Yao, Y. W. Miao, G. Luikart, M. Baig, A. Beja-Pereira, Z. L. Ding, M. G. Palanichamy, and Y. P. Zhang. 2006. Multiple maternal origins of chickens: Out of the Asian jungles. Molecular Phylogenetics and Evolution 38:12-19.

Loftus, R. T., O. Ertugrul, A. H. Harba, M. A. A. El-Barody, D. E. Machugh, S. D. E. Park, and D. G. Bradley. 1999. A microsatellite survey of cattle from a centre of origin: the Near East. Molecular Ecology 8:2015-2022.

Loftus, R. T., D. E. Machugh, D. G. Bradley, P. M. Sharp, and P. Cunningham. 1994. Evidence for two independent domestications of cattle Proceedings of the National Academy of Sciences of the United States of America 91:2757-2761.

Luikart, G., P. R. England, D. Tallmon, S. Jordan, and P. Taberlet. 2003. The power and promise of population genomics: from genotyping to genome typing. Nature Review Genetics 4:981-994.

Luikart, G., L. Gielly, L. Excoffier, J. D. Vigne, J. Bouvet, and P. Taberlet. 2001. Multiple maternal origins and weak phylogeographic structure in domestic goats. Proceedings of the National Academy of Sciences of the United States of America 98:5927-5932.

MacHugh, D. E., R. T. Loftus, P. Cunningham, and D. G. Bradley. 1998. Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. Animal Genetics 29:333 - 340.

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

Manel, S., O. E. Gaggiotti, and R. S. Waples. 2005. Assignment methods: matching biological questions techniques with appropriate. Trends in Ecology & Evolution 20:136-142.

Mank, J. E., and J. C. Avise. 2004. Individual organisms as units of analysis: Bayesian-clustering alternatives in population genetics. Genetical Research 84:135-143.

Marshall, T. C., J. Slate, L. E. B. Kruuk, and J. M. Pemberton. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology 7:639-655.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. Plos One 4:13.

Maudet, C., G. Luikart, and P. Taberlet. 2002. Genetic diversity and assignment tests among seven French cattle breeds based on microsatellite DNA analysis. Journal of Animal Science 80:942-950.

McKay, S. D., R. D. Schnabel, B. M. Murdoch, L. K. Matukumalli, J. Aerts, W. Coppieters, D. Crews, E. Dias, C. A. Gill, C. Gao, H. Mannen, Z. Q. Wang, C. P. Van Tassell, J. L. Williams, J. F. Taylor, and S. S. Moore. 2008. An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. BMC Genetics 9:37-45.

Meadows, J. R. S., I. Cemal, O. Karaca, E. Gootwine, and J. W. Kijas. 2007. Five ovine mitochondrial lineages identified from sheep breeds of the near east. Genetics 175:1371-1379.

Meagher, T. R. 1986. Analysis of paternity within a natural population *Chamaelirium luteum*. I. Identification of most-likely male parents. The American Naturalist 128:199-215.

Megens, H.-J., R. P. M. A. Crooijmans, J. W. M. Bastiaansen, H. H. D. Kerstens, A. Coster, R. Jalving, A. Vereijken, P. Silva, W. M. Muir, H. H. Cheng, O. Hanotte, and M. A. M. Groenen. 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genetics 10:86-97.

Megens, H.-J., R. P. M. A. Crooijmans, M. S. Cristobal, X. Hui, N. Li, and M. A. M. Groenen. 2008. Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. Genetics Selection Evolution 40:103-128.

Menotti-Raymond, M., V. A. David, S. M. Pflueger, K. Lindblad-Toh, C. M. Wade, S. J. O'Brien, and W. E. Johnson. 2008. Patterns of molecular genetic variation among cat breeds. Genomics 91:1-11.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza. 1978. Synthetic maps of human gene frequencies in Europeans. Science 201:786 - 792.

Minch, E., A. Ruiz-Linares, D. Goldstein, M. Feldman, and L. Cavalli-Sforza. 1997. Pp. Microsat v.1.5d : a computer program for calculating various statistics on microsatellite allele data.

Moll, G. 1860. La connaissanse gen. du boeuf. Paris.

Morin, P. A., G. Luikart, R. K. Wayne, and S. N. P. W. Grp. 2004. SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution 19:208-216.

Muir, W. M., G. K.-S. Wong, Y. Zhang, J. Wang, M. A. M. Groenen, R. P. M. A. Crooijmans, H.-J. Megens, H. Zhang, R. Okimoto, A. Vereijken, A. Jungerius, G. A. A. Albers, C. T. Lawley, M. E. Delany, S. MacEachern, and H. H. Cheng. 2008. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proceedings of the National Academy of Sciences 105:17312-17317.

Negrini, R., L. Nicoloso, P. Crepaldi, E. Milanesi, L. Colli, F. Chegdani, L. Pariset, S. Dunner, H. Leveziel, J. L. Williams, and P. A. Marsan. 2009. Assessing SNP markers for assigning individuals to cattle populations. Animal Genetics 40:18-26.

Nei, M., and W. H. Li. 1973. Linkage disequilibrium in subdivided populations Genetics 75:213-219.

Ollivier, L., and J. L. Foulley. 2005. Aggregate diversity: New approach combining within- and between-breed genetic diversity. Livestock Production Science 95:247-254.

Paetkau, D., W. Calvert, I. Stirling, and C. Strobeck. 1995. Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology 4:347-354.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289-290.

Parker, H. G., L. V. Kim, N. B. Sutter, S. Carlson, T. D. Lorentzen, T. B. Malek, G. S. Johnson, H. B. DeFrance, E. A. Ostrander, and L. Kruglyak. 2004. Genetic structure of the purebred domestic dog. Science 304:1160-1164.

Paschou, P., P. Drineas, J. Lewis, C. M. Nievergelt, D. A. Nickerson, J. D. Smith, P. M. Ridker, D. I. Chasman, R. M. Krauss, and E. Ziv. 2008. Tracing Sub-Structure in the European American Population with PCA-Informative Markers. PloS Genetics 4:e1000114.

Paschou, P., J. Lewis, A. Javed, and P. Drineas. 2010. Ancestry informative markers for fine-scale individual assignment to worldwide populations. Journal of Medical Genetics 47:835-847.

Paschou, P., E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. PloS Genetics 3:1672-1686.

Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PloS Genetics 2:2074-2093.

Pearse, D. E., and K. A. Crandall. 2004. Beyond FST: Analysis of population genetic data for conservation. Conservation Genetics 5:585-602.

Pemberton, J. M. 2004. Measuring inbreeding depression in the wild: the old ways are the best. Trends in Ecology & Evolution 19:613-615.

Pemberton, J. M. 2008. Wild pedigrees: the way forward. Proceedings of the Royal Society B-Biological Sciences 275:613-621.

Peter, C., M. Bruford, T. Perez, S. Dalamitra, G. Hewitt, G. Erhardt, and E. Consortium. 2007. Genetic diversity and subdivision of 57 European and Middle-Eastern sheep breeds. Animal Genetics 38:37-44.

Piry, S., A. Alapetite, J. M. Cornuet, D. Paetkau, L. Baudouin, and A. Estoup. 2004. GENECLASS2: A software for genetic assignment and first-generation migrant detection. Journal of Heredity 95:536-539.

Porter, V. 1993. Pigs. A handbook to the Breeds of the World. Helm Information, Ltd.

Primmer, C. R., M. T. Koskinen, and J. Piironen. 2000. The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. Proceedings of the Royal Society of London Series B-Biological Sciences 267:1699-1704.

Primrose, S., M. Woolfe, and S. Rollinson. 2010. Food forensics: methods for determining the authenticity of foodstuffs. Trends in Food Science & Technology 21:582-590.

Pritchard, J. K., and W. Wen. 2004. Documentation for STRUCTURE software: Version 2. Department of Human Genetics, Univerity of Chicago, 920 E 58th st, CLCS 507, Chicago IL 60637, USA.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945 - 959.

Qanbari, S., M. Hansen, S. Weigend, R. Preisinger, and H. Simianer. 2010. Linkage disequilibrium reveals different demographic history in egg laying chickens. BMC Genetics 11:103-112.

Ramos, A. M., R. P. M. A. Crooijmans, N. A. Affara, A. J. Amaral, A. L. Archibald, J. E. Beever, C. Bendixen, C. Churcher, R. Clark, P. Dehais, M. S. Hansen, J. Hedegaard, Z.-L. Hu, H. H. Kerstens, A. S. Law, H.-J. Megens, D. Milan, D. J. Nonneman, G. A. Rohrer, M. F. Rothschild, T. P. L. Smith, R. D. Schnabel, C. P. Van Tassell, J. F. Taylor, R. T. Wiedmann, L. B. Schook, and M. A. M. Groenen. 2009. Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. PloS One 4:e6524.

Ramos, A. M., H. J. Megens, R. P. M. A. Crooijmans, L. B. Schook, and M. A. M. Groenen. 2011. Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. Animal Genetics:doi: 10.1111/j.1365-2052.2011.02198.x.

Rannala, B., and J. L. Mountain. 1997. Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences of the United States of America 94:9197-9201.

RBST. 2008. Pig Bloodline Survey. Ark. Rare Breeds Survival Trust. Stoneleigh, UK.

Reynolds, J., B. S. Weir, and C. C. Cockerham. 1983. Estimation of the co-ancestry coefficient - basis for a short-term genetic distance. Genetics 105:767-779.

Rice, W. R. 1989. Analyzing tables of statistical tests. Evolution 43:223-225.

Roberts, M. 1994. British Large Fowl. Domestic Fowl Research.

Roberts, V. 1997. British Poultry Standards. Wiley-Blackwell.

Roques, S., P. Duchesne, and L. Bernatchez. 1999. Potential of microsatellites for individual assignment: the North Atlantic redfish (genus Sebastes) species complex as a case study. Molecular Ecology 8:1703-1717.

Rosenberg, N. A., T. Burke, K. Elo, M. W. Feldman, P. J. Freidlin, M. A. M. Groenen, J. Hillel, A. Maki-Tanila, M. Tixier-Boichard, A. Vignal, K. Wimmers, and S. Weigend. 2001. Empirical Evaluation of Genetic Clustering Methods Using Multilocus Genotypes From 20 Chicken Breeds. Genetics 159:699-713.

Rosenberg, N. A., L. M. Li, R. Ward, and J. K. Pritchard. 2003. Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics 73:1402-1422.

Rousset, F. 2008. GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. Molecular Ecology Resources 8:103-106.

Rowe, G., and T. J. C. Beebee. 2007. Defining population boundaries: use of three Bayesian approaches with microsatellite data from British natterjack toads (*Bufo calamita*). Molecular Ecology 16:785-796.

Russell, G. A., A. L. Archibald, C. S. Haley, and A. S. Law. 2003. The pig genetic diversity database and the WWW. Archivos de Zootecnia 52:165-172.

Safner, T., M. P. Miller, B. H. McRae, M.-J. Fortin, and S. Manel. 2011. Comparison of Bayesian Clustering and Edge Detection Methods for Inferring Boundaries in Landscape Genetics. International Journal of Molecular Sciences 12:865-889.

Saitou, N., and M. Nei. 1987. The neighbor-joining method - a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425.

SanCristobal, M., C. Chevalet, C. S. Haley, R. Joosten, A. P. Rattink, B. Harlizius, M. A. M. Groenen, Y. Amigues, M.-Y. Boscher, G. Russell, A. Law, R. Davoli, V. Russo, C. Désautés, L. Alderson, E. Fimland, M. Bagga, J. V. Delgado, J. L. Vega-Pla, A. M. Martinez, M. Ramos, P. Glodek, J. N. Meyer, G. C. Gandini, D. Matassino, G. S. Plastow, K. W. Siggens, G. Laval, A. L. Archibald, D. Milan, K. Hammond, and R. Cardellino. 2006a. Genetic diversity within and between European pig breeds using microsatellite markers. Animal Genetics 37:189-198.

SanCristobal, M., C. Chevalet, J. Peleman, H. Heuven, B. Brugmans, M. van Schriek, R. Joosten, A. P. Rattink, B. Harlizius, M. A. M. Groenen, Y. Amigues, M. -Y. Boscher, G. Russell, A. Law, R. Davoli, V. Russo, C. Desautes, L. Alderson, E. Fimland, M. Bagga, J. V. Delgado, J. L. Vega-Pla, A. M. Martinez, M. Ramos, P. Glodek, J. N. Meyer, G. Gandini, D. Matassino, K. Siggens, G. Laval, A. L.

Archibald, D. Milan, K. Hammond, R. Cardellino, C. S. Haley, and G. S. Plastow. 2006b. Genetic diversity in European pigs utilizing amplified fragment length polymorphism markers. Animal Genetics 37:232-238.

Sanz, N., R. Araguas, R. Fernández, M. Vera, and J.-L. García-Marín. 2009. Efficiency of markers and methods for detecting hybrids and introgression in stocked populations. Conservation Genetics 10:225-236.

Schwartz, M. K., and K. S. McKelvey. 2009. Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. Conservation Genetics 10:441-452.

Selkoe, K. A., and R. J. Toonen. 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecology Letters 9:615-629.

Shriver, M. D., M. W. Smith, L. Jin, A. Marcini, J. M. Akey, R. Deka, and R. E. Ferrell. 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. American Journal of Human Genetics 60:957-964.

Slate, J., P. David, K. G. Dodds, B. A. Veenvliet, B. C. Glass, T. E. Broad, and J. C. McEwan. 2004. Understanding the relationship between the inbreeding coefficient and multilocus heterozygosity: theoretical expectations and empirical data. Heredity 93:255-265.

Slate, J., L. E. B. Kruuk, T. C. Marshall, J. M. Pemberton, and T. H. Clutton-Brock. 2000. Inbreeding depression influences lifetime breeding success in a wild population of red deer (*Cervus elaphus*). Proceedings of the Royal Society of London Series B-Biological Sciences 267:1657-1662.

Small, R. W. 2004. The role of rare and traditional breeds in conservation: the Grazing Animals Project. In: Farm Animal Genetic Resources. Nottingham University Press, Loughborough.

Smith, C. 1984. Genetic aspects of conservation in farm livestock. Livestock Production Science 11:37-48.

Smith, M. W., J. A. Lautenberger, H. D. Shin, J. P. Chretien, S. Shrestha, D. A. Gilbert, and S. J. O'Brien. 2001. Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. American Journal of Human Genetics 69:1080-1094.

Smouse, P. E., R. S. Spielman, and M. H. Park. 1982. Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. American Naturalist 119:445-463.

Spiegelhalter, D. J., N. G. Best, B. R. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B-Statistical Methodology 64:583-616.

Sunnucks, P. 2000. Efficient genetic markers for population biology. Trends in Ecology & Evolution 15:199-203.

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

Tapio, I., M. Tapio, Z. Grislis, L. E. Holm, S. Jeppsson, J. Kantanen, I. Miceikiene, I. Olsaker, H. Viinalass, and E. Eythorsdottir. 2005a. Unfolding of population structure in Baltic sheep breeds using microsatellite analysis. Heredity 94:448-456.

Tapio, M., I. Tapio, Z. Grislis, L. E. Holm, S. Jeppsson, J. Kantanen, I. Miceikiene, I. Olsaker, H. Viinalass, and E. Eythorsdottir. 2005b. Native breeds demonstrate high contributions to the molecular variation in northern European sheep. Molecular Ecology 14:3951-3963.

Team, R. D. C. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. .

Teletchea, F., C. Maudet, and C. Hanni. 2005. Food and forensic molecular identification: update and challenges. Trends in Biotechnology 23:359-366.

Toro, M. A., and A. Caballero. 2005. Characterization and conservation of genetic diversity in subdivided populations. Philosophical Transactions of the Royal Society B: Biological Sciences 360:1367-1378.

Toro, M. A., J. Fernandez, and A. Caballero. 2009. Molecular characterization of breeds and its use in conservation. Livestock Science 120:174-195.

Tunon, M. J., P. Gonzalez, and M. Vallejo. 1989. Genetic relationships beteween 14 native spanish breeds of goat. Animal Genetics 20:205-212.

Vanhala, T., M. Tuiskula-Haavisto, K. Elo, J. Vilkki, and A. Maki-Tanila, 1998. Evaluation of genetic variability and genetic distance between eight chicken lines using microsatellite markers. Poultry Sci. 7:783–790.

Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild, S. S. Moore, W. C. Warren, and T. S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5:247-252.

Van Zeveren, A., Y. Bouquet, A. Van De Weghe, and W. Coppieters. 1990a. A genetic blood marker study on 4 pig breeds I. Estimation and comparison of within-breed variation. Journal of Animal Breeding and Genetics 107:104-112.

Van Zeveren, A., Y. Bouquet, A. Van De Weghe, and W. Coppieters. 1990b. A genetic blood marker study on 4 pig breeds II. Genetic relationships between the populations Journal of Animal Breeding and Genetics 107:113-118.

Visscher, P. M., D. Smith, S. J. G. Hall, and J. A. Williams. 2001. A viable herd of genetically uniform cattle - Deleterious alleles seem to have been purged in a feral strain of inbred cows. Nature 409:303-303.

vonHoldt, B. M., J. P. Pollinger, K. E. Lohmueller, E. Han, H. G. Parker, P. Quignon, J. D. Degenhardt, A. R. Boyko, D. A. Earl, A. Auton, A. Reynolds, K. Bryc, A. Brisbin, J. C. Knowles, D. S. Mosher, T. C. Spady, A. Elkahloun, E. Geffen, M. Pilot, W. Jedrzejewski, C. Greco, E. Randi, D. Bannasch, A. Wilton, J. Shearman, M. Musiani, M. Cargill, P. G. Jones, Z. Qian, W. Huang, Z.-L. Ding, Y.-P. Zhang, C. D. Bustamante, E. A. Ostrander, J. Novembre, and R. K. Wayne. 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464:898-902.

Vorwald Dohner, J. 2001. The Encyclopedia of Historic and Endangered Livestock and Poultry Breeds. Yale University Press.

Waples, R., and O. Gaggiotti. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology 15:1419-1439.

Warriss, P. D., S. C. Kestin, S. N. Brown, and G. R. Nute. 1996. The quality of pork from traditional pig breeds. Meat Focus International 5:179-182.

Waser, P. M., and C. Strobeck. 1998. Genetic signatures of interpopulation dispersal. Trends in Ecology & Evolution 13:43-44.

Weir, B. S. 1996. Genetic Data Analysis II. Sinauer Associates, Inc. Publishers.

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

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

Wiener, P., M. A. Edriss, J. L. Williams, D. Waddington, A. Law, J. A. Woolliams, and B. Gutierrez-Gil. 2011. Information content in genome-wide scans: concordance between patterns of genetic differentiation and linkage mapping associations. BMC Genomics 12:65-74.

Wong, G. K. S., B. Liu, J. Wang, Y. Zhang, X. Yang, Z. J. Zhang, Q. S. Meng, J. Zhou, D. W. Li, J. J. Zhang, P. X. Ni, S. G. Li, L. H. Ran, H. Li, J. G. Zhang, R. Q. Li, S. T. Li, H. K. Zheng, W. Lin, G. Y. Li, X. L. Wang, W. M. Zhao, J. Li, C. Ye, M. T. Dai, J. Ruan, Y. Zhou, Y. Z. Li, X. M. He, Y. Z. Zhang, X. G. Huang, W. Tong, J. Chen, J. Ye, C. Chen, N. Wei, G. Q. Li, L. Dong, F. D. Lan, Y. Q. Sun, Z. P. Zhang, Z. Yang, Y. P. Yu, Y. Q. Huang, D. D. He, Y. Xi, D. Wei, Q. H. Qi, W. J. Li, J. P. Shi, M. H. Wang, F. Xie, J. J. Wang, X. W. Zhang, P. Wang, Y. Q. Zhao, N. Li, N. Yang, W. Dong, S. N. Hu, C. Q. Zeng, W. M. Zheng, B. L. Hao, L. W. Hillier, S. P. Yang, W. C. Warren, R. K. Wilson, M. Brandstrom, H. Ellegren, R. Crooijmans, J. J. van der Poel, H. Bovenhuis, M. A. M. Groenen, I. Ovcharenko, L. Gordon, L.

Stubbs, S. Lucas, T. Glavina, A. Aerts, P. Kaiser, L. Rothwell, J. R. Young, S. Rogers, B. A. Walker, A. van Hateren, J. Kaufman, N. Bumstead, S. J. Lamont, H. J. Zhou, P. M. Hocking, D. Morrice, D. J. de Koning, A. Law, N. Bartley, D. W. Burt, H. Hunt, H. H. Cheng, U. Gunnarsson, P. Wahlberg, L. Andersson, et al. 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature 432:717-722.

Woolfe, M., and S. Primrose. 2004. Food forensics: using DNA technology to combat misdescription and fraud. Trends in Biotechnology 22:222-226.

Wright, S. 1921. Systems of mating. I, 11,111, IV, V. Genetics 6:111-178.

Wright, S. 1922. Coefficients of inbreeding and relationship. American Naturalist 56:330-338.

Wright, S. 1943. Isolation by distance. Genetics 28:114-138.

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

Zanetti, E., M. De Marchi, C. Dalvit, and M. Cassandro. 2010. Genetic characterization of local Italian breeds of chickens undergoing in situ conservation. Poultry Science 89:420-427.